Title
Guarantees for a few structured statistical problems

Author
Khamaru, Koulik

Publication Date
2022

Peer reviewed|Thesis/dissertation

Guarantees for a few structured statistical problems

by

Koulik Khamaru

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael I. Jordan, Co-chair
Professor Martin J. Wainwright, Co-chair
Professor Adityanand Guntuboyina
Professor Thomas Courtade

Summer 2022

Guarantees for a few structured statistical problems

Abstract

Guarantees for a few structured statistical problems

by

Koulik Khamaru

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Michael I. Jordan, Co-chair

Professor Martin J. Wainwright, Co-chair

In recent years, we have seen a tremendous interest in applying statistics and machine learning methods in various areas of science: health, education, drug design, public policy design to name a few. This immense popularity of statistical methods comes with challenging new questions which lie in the boundary of theoretical and methodological aspects of statistics, machine learning and optimization. The aim of this dissertation is to address some of the these challenges that arise in modern reinforcement learning, and in modern data science practice and provide new insights that are helpful to practitioners. The dissertation is divided into four parts. In Part I we discuss principled way of designing fast algorithms for various reinforcement learning problems. The Part II of the dissertation is devoted to problems that arise due to model misspecification. In Part III we discuss how can we perform inference when the data set is collected in a sequential manner; i.e. the helpful iid structure is not present in the data. Finally, Part IV focuses on deriving fast algorithms for structured non-convex problems.

To Maa and Bhai.

Contents

List of Figures

List of Tables

Acknowledgments

My time at Berkeley went by quite fast, and it certainly did not feel like I spent six years here. I would like to thank all the amazing people that I met here during my PhD, and I apologize in advance if I accidentally missed anyone.

First and foremost, I would like to thank both my wonderful advisors Martin J. Wainwright and Michael I. Jordan for their constant support and encouragement throughout my PhD. I certainly had the two most amazing advisors a student could hope for. I first met Martin by taking a course on graphical models during the first semester at Berkeley. In my second semester, I took a course on modern optimization; this class piqued my interest in optimization and equipped me with the basic tools of optimization. We started working on my first research project with him that summer. Even after six years, I am constantly amazed by his ability to explain complicated concepts in a simplified way, his mathematical sharpness and deep understanding of various areas in statistics, optimization, and information theory, his humility, his caring nature, and his sense of humor. I've learned a tremendous amount from working with him, including mathematical insights, how to think about research problems, as well as guidance on a wide range of topics. I also greatly value and hope to continue the gradual integration of mentorship into elements of friendship.

During my first semester, I started to go to the SAIL group meeting and I met Mike there. I started talking with him frequently during my second year. Throughout the years I have been constantly amazed by his breadth and vision in the field. I would also like to thank him for giving me the courage to learn and work on completely new topics, and for always supporting me when I wanted to do so. I would also like to thank Mike for his immense help and constant care during every stage of my academic job interviews and decision process. Both my advisors had an enviable balance of work and life and they have a variety of interests outside of academics; over the years it has helped me think about my priorities and maintain a balance in my own life.

I would like to thank Prof. Aditya Guntuboinya for being extremely supportive throughout my PhD, to who I would always feel safe to turn to. I would also like to thank Prof. Thomas Courtade for being on my committee and to provide me with very helpful feedback during my qualifying exam and also for teaching a wonderful course on information theory. I would also like to Prof. Bin Yu and Prof. Peter Bartlett for their help during various parts of our joint work. I would like to especially thank Bin for her insightful and refreshing comments on our projects, and it has certainly taught me to think about research problems and their broader scope.

During my fourth year of my PhD I was fortunate enough to spend a wonderful summer at Amazon research. I would like specially thank my mentor, Dean Foster, for teaching me about reinforcement learning. I would also like to thank my manager Ido Rosen for making my internship an enjoyable experience. During my last two years, I had the pleasure of working with Lester Mackey. Apart from his clear thinking, the thing that amazes me the most is his ability to keep a balance between the theoretical

and practical sides of a research problem. I have tried my best to learn this from Lester and hope to do so myself in the future.

There is a saying that you do not work a day if you enjoy your work. This was certainly true for me during my PhD, and a lot of that is due to wonderful collaborators that I had: Raaz Dwevedi, Nhat Ho, Kush Bhatia, Dhruv Malik, Ashwin Panindjay, Wenlong Mou, Licong Lin, Eric Xia, Yash Deshpande, Feng Ruan, Tor Lattimore, and Wenshuo Guo. During my last two years of my PhD, I was fortunate to work with two junior students Eric Xia and Licong Lin. I will always cherish our long discussions during various parts of projects and also for wonderful after-work dinners. I have learned many things from both of you and hope that I was able to pass on some of the lessons that my advisors taught me.

It is no surprise that a place like Berkeley is full of wonderful and intelligent people, and during my PhD I have learned a lot from them. A big thank you to everyone in the SAIL and BLISS lab and the broader statistics community. Special thanks to Reese Pathak, Eric Xia, Wenlong Mou, Raaz Dwevedi, Payam Delgosha, Dong Yin, Soham Phade, Billy Fang, Joseph Borja, Jianbo Chen, Sang Min Han, Orhan Ocal, Vidya Muthukumar, Niladri Chatterji, Kush Bhatia, Avishek Ghosh, Vipul Gupta, Laura Brink, Aviva Bulow, Banghua Zhou, Efe Aras, Kuan-Yun Lee, Vignesh Subramanian, Chinmay Maheshwari, Kabir Chandrasekher, Tavor Baharav, Sivaraman Balakrishnan, Aaditya Ramdas , Tijana Zrnic, Zhuoran Yang, Yixin Wang, Nilesh Tripuraneni, Jake Soloff, Lydia Liu, Junchi Li, Wenshuo Guo, Tatjana Chavdarova, Stephen Bates, Taejoo Ahn, Mriganka Basu Roy Chowdhury, Emily Flanagan, Ella Hiesmayr, Adam Quinn Jaffe, Miyabi Ishihara, Drew Nguyen, Dan Soriano, Fangzhou Su, Alexander Tsigler, Tyler Maltba, Jake Soloff for all the wonderful conversations we had over the years. Finally I would like to thank La Shana Porlaris and the entire statistics department staff for making my PhD life so much easier.

This acknowledgment would be incomplete if I do not thank the wonderful group of friends that I had in Berkeley outside SAIL and BLISS. First, I would like to thank Milind Hegde and Nick Bahattacharya for being the best roommate I could hope for. I must also thank Milind for teaching me how to cook nice south indian dishes. I would like to thank Vipul, Aviva, Ella, Anamika, Nived, and Avishek for accompanying me on various wonderful hikes. I would like to thank Siddesh, Anushka, Jimmy, and Satyaki for all the fun discussions while watching tv series at our apartment. Satyaki and Hari for their cooking lessons and for making amazing dishes at our house. I would like to thank Zsolt for teaching me how to play squash and for giving me a wonderful tour of Budapest and for teaching me squash. I would like to especially thank Vipul for always being enthusiastic about any kind of travel plan and for joining me on almost all my trips. Finally, I would also like to thank my friends at UC Davis; Samayita Bhattacharya, Sneha Bhattacharya, Paromita Dubey, and Abhishek Roy for making amazing food every time I visited them.

Above all, I owe the most to my mother Monika Khamaru, and brother Moulik Khamaru for their unconditional support and constant care and encouragement. All

of my achievements, if any, are all because of you. Thank you.

Chapter 1

Introduction

In recent years, we have seen a tremendous interest in applying statistics and machine learning methods in various areas of science: health, education, drug design, public policy design to name a few. This immense popularity of statistical methods comes with challenging new questions which lie in the boundary of theoretical and methodological aspects of statistics, machine learning and optimization. During my Ph.D., my focus has been on addressing some of these challenges that arise in modern reinforcement learning, and in modern data science practice. A recurring theme in my research is to understand why some algorithms work well in practice, provide new methods which are both *computationally efficient*, are (often) *provably optimal*, and most importantly, provide new insights that are helpful to practitioners.

- For some context, starting with a big concern in modern reinforcement learning, it is a well observed phenomenon that popular reinforcement learning algorithms behave far better in practice than their worst case theoretical prediction.

  What are some plausible reasons behind this theory-practice gap?

  In Part I I discuss my works in reinforcement learning, where we consider a *problem-instance* specific difficulty measure and propose algorithms that adapts to this *instance-specific* difficulty measure. This explains why algorithms should behave nicely for problems with favorable structure.

- Moving on to problems in modern statistics and data science, a big problem in statistics practice is choosing a correct model for the data set.

  What are the effects of model mis-specification?

  In Part II, I discuss my work on model mis-specification in the context of Gaussian mixture models.

- Once the model is selected, which algorithm do we use to efficiently estimate the model parameters? Many problems naturally gives rise to *non-convex problems*, and simple non-convex optimization algorithms are often used in practice due to their superior performance compared to methods based on convex-relaxations. When are these non-convex methods are provably better? Part IV is devoted to my work on provable guarantees and efficient methods for non-convex problems.

- Once we have estimated the model using our favorite algorithm, can we trust this estimated model for future predictions? In Chapter 6, I discuss my work on connections between algorithmic stability and computational and statistical efficiency.

- Finally, in many modern problems in statistics, data science and machine learning, the datasets are collected in a sequential manner which violates the helpful i.i.d. structure. Problems include adaptive trails, multiarm and contextual bandit experiments, public policy design and so on. Unfortunately, classical statistical methods and inference techniques provide errornous results in these scenarios. How do we deign inference methods for *sequentially* collected data? In Part III, I discuss my work in developing inference techniques for linear models with sequentially collected data.

Part I: Instance Dependent Bounds in Reinforcement Learning

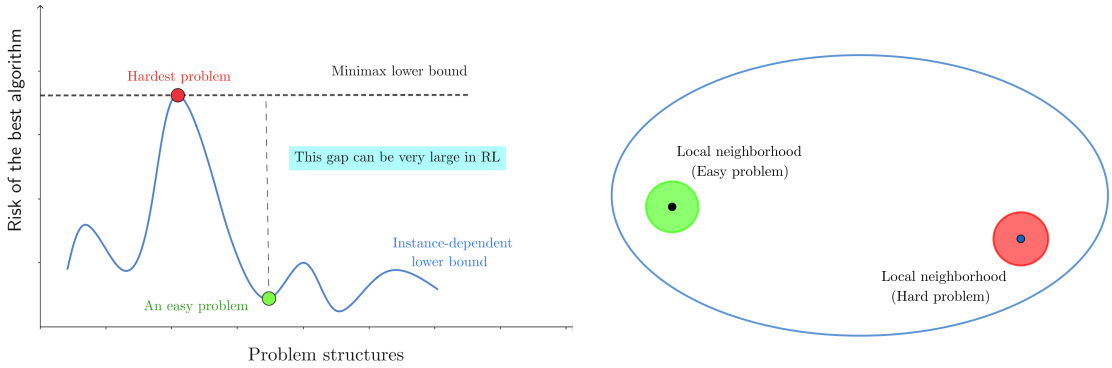
Recently, there has seen tremendous progress in the field of reinforcement learning (RL), both in theory and practice. But it is fair to say that there is a considerable gap between theory and practice: many algorithms that are used in practice behave far better than existing theory would suggest, and often they work in settings where the current worst case guarantees are completely prohibitive. Part of the thesis uncovers why worst case guarantees can heavily overestimates the difficulty of reinforcement learning problems with favorable structure. This motivates us to consider a local measure of difficulty which is problem *instance-specific*, and design algorithms that adapts to instance-specific difficulty measure. This helps explain why the algorithmic performance can be dramatically different in reinforcement learning problems within a problem class with a specific worst case performance.

In the classical minimax theory, the difficulty of a problem $\mathcal{P}$ is characterized by the global minimax risk, the risk of the best algorithm at the hardest problem in a relevant problem class containing the problem $\mathcal{P}$. Thus, the global minimax risk completely ignores the difficulty of the problem itself and overestimates the difficulty when the problem is easy. One alternative to this approach is to consider a *local measure of difficulty* by considering the risk at the hardest problem in an appropriate local neighborhood of the given problem $\mathcal{P}$. This local notion of difficulty is related to the celebrated local asymptotic minimax theory by LeCam and Háyek [Háj72; Le

72; LeC53], and provides an *instance-specific* notion of difficulty that depends on the problem itself.

In chapters 2 and 3, respectively, we discuss instance optimal almgorithms for the problem of policy evaluation — accessing the quality of a behavioral policy — and policy optimization — finding the best policy in an environment. Focusing on the problems with finite number of states and actions, we determine the functionals which characterizes the difficulty of policy evaluation and policy optimization problems in a small local neighborhood in a non-asymptotic sense. Our local measure of difficulty borrows ideas from the celebrated asymptotic local minimax framework by LeCam and Háyek [Háj72; Le 72; LeC53], and a more recent non-asymptotic local minimax framework by Cai and Low [CL15b]. The local nature of instance-specific difficulty measure allows us to distinguish an "*easy*" problem from a "*hard*" one; this also naturally provides an instance-dependent lower bound on the accuracy on any estimation procedure. Finally, via simple examples, we show that for easy problems our *instance-dependent lower bound* can be significantly better than the global minimax lower bound.

With an instance dependent lower bound for policy evaluation and policy optimization problem at hand, our next goal was to design algorithms that adapts to this instance-dependent notion of difficulty. In chapters 2 and 3 we also study the problem of *instance-optimal* algorithm design. We show, via extensive simulation, that the popular temporal difference (TD) algorithm for policy evaluation is strictly sub-optimal in a non-asymptotic sense, even when combined an iterate-averaging (Polyak-Ruppert averaging) step. Variance reduction based stochastic gradient algorithms are known to often adapt to the problem difficulty for convex optimization problems, but there is major difficulty in importing those results to reinforcement learning problems. Reinforcement learning problems do not induce convex optimization problems in general, and consequently, we cannot borrow the proof techniques from convex optimization literature. Nonetheless, we show that variance-reduction is still helpful, and it is a key ingredient for designing algorithms which adapts to problem structure. We show, via a novel analysis, that a variance reduced temporal difference algorithm adapts to the problem-difficulty and achieves the *instance-dependent* lower bound for policy evaluation and policy optimization problem.



A general principle for instance-optimal algorithm design: While reinforcement learning problems and convex optimization problems look very different on the surface, there are some similarities. In both cases the optimal solution can be thought of the unique *fixed point of an appropriate contractive operator*; in reinforcement learning the operator is the Bellman operator, in convex optimization problem the operator is the gradient update operator. problem difficulty. It is now natural to ask a more general question:

Can we design problem-adaptive algorithms for finding fixed points of contractive operators under noise? Is there a common principle?

In Chapter 4, we study the problem of finding the unique fixed point of a contractive operator under some norm. The goal is to estimate the unique fixed point using noisy observation from that operator. We propose a variance reduced stochastic approximation algorithm which provides an estimate of this unique fixed point, and we non-asymptotic bounds for the estimation error. Our bounds are instance dependent in nature, meaning that our bound adapts to the specific problem structure at hand, and are (often) attains the instance-dependent lower bound. This project provides a unified tool for understanding adaptive algorithms for fixed point finding problems under noise, including stochastic shortest path, two player zero sum Markov games.

Part II: Singularity, Stability and the Localization argument

The growth in the size and scope of modern data sets has presented the field of statistics with a number of challenges, among them is how to deal with various forms of heterogeneity. Mixture models provide a principled approach to modeling heterogeneous collections of data. In practice, it is frequently the case that the number of mixture components in the fitted model does not match the number of components in the data-generating mechanism. It is known that such mismatch can lead to substantially slower convergence rates for the maximum likelihood estimate (MLE) for the underlying parameters. In contrast, relatively less attention has been paid to the computational implications of this mismatch. In particular, the algorithm of

choice for fitting finite mixture models is the EM algorithm, a general framework that encompasses divide-and-conquer computational strategies. We seek a fundamental understanding of how EM behaves for over-fitted mixture models.

Part II of the thesis aims to shed some light on the performance of the EM algorithm for over-fitted mixtures. We do so by providing a comprehensive study of over-fitted mixture models when fit to the simplest possible (non-mixture) data-generating mechanism, that is a multivariate normal distribution $\mathcal{N}(0, \sigma^2 I_d)$ in d dimensions with known scale parameter $\sigma > 0$. This setting, despite its simplicity, suffices to uncover some rather interesting properties of EM in the over-fitted context. We show that the sample-based EM iterates converge to a Euclidean ball of radius $(d/n)^{1/4}$ around the true parameter; this is contrary to the standard $\sqrt{d/n}$ rate of parameter estimation in Gaussian mixture models [BWY17]. The $n^{-1/4}$ component of the error matches known guarantees for the global maximum of the MLE [Che95].

Our proofs are based on a novel localization-based analysis of the underlying empirical process. It turns out that the localization argument in our analysis is of independent interest, and it is quite useful in proving non-standard slower rates. Another application can be found in the paper [Dwi+20a] where we discuss the performance of the EM algorithm in estimating both mean and covariance of overfitted Gaussian mixture models. We do not include this work in the paper due to breviety. In Chapter 6, we discuss a more general application of the localization argument. We show that a general version of the localization argument can be used to understand the interplay between stability of an algoeithm (EM algorithm for instance) and the statistical rates of estimation.

Part III Inference in adaptive experiments

Adaptive experiments where the treatment assignment probabilities are adapted during the trial based on accrued responses with the aim of achieving experimental objectives while ideally preserving inferential validity. These adaptive experiments are already popular in major internet companies, in public policy design, and recently gaining traction in clinical research where adaptive experiments can be used to quickly select a few best treatments from a large number of possible ones, which are then used to design a more robust RCT. In Chapter 7 we discuss inference techniques for adaptively collected data. We focus on adaptive linear regression models, which include multi-armed and contextual bandit models and auto regressive time series model as a special case. Due to the non i.i.d. nature of the data associated with an adaptive experiment, even simple methods like ordinary least squares can exhibit non-normal asymptotic behavior. As an undesirable consequence, hypothesis tests and confidence intervals based on asymptotic normality can lead to erroneous results. To remidy this issue, we propose a family of online debiasing estimators for adaptive linear regression models that corrects the distributional anomalies in least squares estimation. These estimators are constructed by adding a linear correction term to

the least square estimator which reduces the bias of the least square estimator and increases the variance by a small logarithmic factor. We establish an asymptotic normality property for our proposed online debiasing estimators under mild conditions on the data collection process, and provide asymptotically exact confidence intervals. Finally, we show that under an adaptive data collection scheme, the performance of our estimators and width of our confidence intervals matches the performance of the minimax optimal estimator up to logarithmic factors.

Part IV: Favorable structures in non-convex problems

A large number of inference and estimation problems in statistics and machine learning are naturally formulated as non-convex problems. While convex relaxation of these non-convex problems are available in many cases, simple algorithms on the original non-convex problem work much better in practice. In what follows, I describe two instances where my research justifies the statistical accuracy and computational efficiency of simple methods by identifying structures in these non-convex problems that are favorable to simple algorithms.

In Chapter 8 we consider the problem of finding critical points of functions that are non-convex and non-smooth. Studying a fairly broad class of such problems, we analyze the behavior of three gradient-based methods (gradient descent, proximal update, and Frank-Wolfe update). For each of these methods, we establish rates of convergence for general problems, and also prove faster rates for continuous sub-analytic functions. We also show that our algorithms can escape strict saddle points for a class of non-smooth functions, thereby generalizing known results for smooth functions. Our analysis leads to a simplification of the popular CCCP algorithm, used for optimizing functions that can be written as a difference of two convex functions. Our simplified algorithm retains all the convergence properties of CCCP, along with a significantly lower cost per iteration. We illustrate our methods and theory via applications to the problems of best subset selection, robust estimation, mixture density estimation, and shape-from-shading reconstruction.

Part I

Instance dependent bounds in reinforcement learning

Chapter 2

Instance-dependent bounds for policy evaluation

In this chapter we address the problem of policy evaluation in discounted Markov decision processes, and provide instance-dependent guarantees on the ℓ_∞-error under a generative model. We establish both asymptotic and non-asymptotic versions of local minimax lower bounds for policy evaluation, thereby providing an instance-dependent baseline by which to compare algorithms. Theory-inspired simulations show that the widely-used temporal difference (TD) algorithm is strictly suboptimal when evaluated in a non-asymptotic setting, even when combined with Polyak-Ruppert iterate averaging. We remedy this issue by introducing and analyzing variance-reduced forms of stochastic approximation, showing that they achieve non-asymptotic, instance-dependent optimality up to logarithmic factors.

2.1 Introduction

Reinforcement learning (RL) is a class of methods for the optimal control of dynamical systems [Ber95a; Ber95b; BT96; SB18a] that has begun to make inroads in a wide range of applied problem domains. This empirical research has revealed the limitations of our theoretical understanding of this class of methods—popular RL algorithms exhibit a variety of behavior across domains and problem instances, and existing theoretical bounds, which are generally based on worst-case assumptions, fail to capture this variety. An important theoretical goal is to develop *instance-specific* analyses that help to reveal what aspects of a given problem make it "easy" or "hard," and allow distinctions to be drawn between ostensibly similar algorithms in terms of their performance profiles. The focus of this chapter is on developing such a theoretical understanding for a class of popular stochastic approximation algorithms used for policy evaluation.

RL methods are generally formulated in terms of a Markov decision process (MDP). An agent operates in an environment whose dynamics are described by an MDP but

are unknown: at each step, it observes the current state of the environment, and takes an action that changes the state according to some stochastic transition function. The eventual goal of the agent is to learn a *policy*—a mapping from states to actions—that optimizes the reward accrued over time. In the typical setting, rewards are assumed to be additive over time, and are also discounted over time. Within this broad context, a key sub-problem is that of *policy evaluation*, where the goal is estimate the long-term expected reward of a fixed policy based on observed state-to-state transitions and one-step rewards. It is often preferable to have ℓ_∞-norm guarantees for such an estimate, since these are particularly compatible with policy-iteration methods. In particular, policy iteration can be shown to converge at a geometric rate when combined with policy evaluation methods that are accurate in ℓ_∞-norm (see, e.g.,[AJK19; BT96]).

In this chapter, we study a class of stochastic approximation algorithms for this problem under a generative model for the underlying MDP, with a focus on developing instance-dependent bounds. Our results complement an earlier paper by a subset of the authors [PW19], which studied the least squares temporal difference (LSTD) method through such a lens.

Related work

We begin with a broad overview of related work, categorizing that work as involving asymptotic analysis, non-asymptotic analysis, or instance-dependent analysis.

Asymptotic theory: Markov reward processes have been the subject of considerable classical study [Dur99; Fel66]. In the context of reinforcement learning and stochastic control, the policy evaluation problem for such processes has been tackled by various approaches based on stochastic approximation. Here we focus on past work that studies the *temporal difference* (TD) update and its relatives; see [DNP14] for a comprehensive survey. The TD update was originally proposed by Sutton [Sut88], and is typically used in conjunction with an appropriate parameterization of value functions. Classical results on the algorithm are typically asymptotic, and include both convergence guarantees [Bor09; BM00; JJS94a] and examples of divergence [Bai95]; see the seminal work [TV97] for conditions that guarantee asymptotic convergence.

It is worth noting that the TD algorithm is a form of linear stochastic approximation, and can be fruitfully combined with the iterate-averaging procedure put forth independently by Polyak [PJ92] and Ruppert [Rup88a]. In this context, the work of Polyak and Juditsky [PJ92] deserves special mention, since it shows that under fairly mild conditions, the TD algorithm converges when combined with Polyak-Ruppert iterate averaging. To be clear, in the specific context of the policy evaluation problem, the results in the Polyak-Juditsky paper [PJ92] allow noise only in the observations of rewards (i.e., the transition function is assumed to be known). However, the underlying techniques can be extended to derive results in the setting in which we

only observe samples of transitions; for instance, see the work of Tadic [Tad04] for results of this type.

Non-asymptotic theory: Recent years have witnessed significant interest in understanding TD-type algorithms from the non-asymptotic standpoint. Bhandari et al. [BRS18b] focus on proving ℓ_2-guarantees for the TD algorithm when combined with Polyak-Ruppert iterate averaging. They consider both the generative model as well as the Markovian noise model, and provide non-asymptotic guarantees on the expected error. Their results also extend to analyses of the popular TD(λ) variant of the algorithm, as well as to Q-learning in specific MDP instances. Also noteworthy is the analysis of Lakshminarayanan and Szepesvari [LS18b], carried out in parallel with Bhandari et al. [BRS18b]; it provides similar guarantees on the TD(0) algorithm with constant stepsize and averaging. Note that both of these analyses focus on ℓ_2-guarantees (equipped with an associated inner product), and thus can directly leverage proof techniques for stochastic optimization [BM11; Nem+09a].

Other related results[1] include those of Dalal et al. [Dal+18], Doan et al. [DMR19], Korda and La [KL15], and also more contemporary papers [Wai+19; Xu+20]. The latter three of these papers introduce a variance-reduced form of temporal difference learning, a variant of which we analyze in this chapter.

Instance-dependent results: The focus on instance-dependent guarantees for TD algorithms is recent, and results are available both in the ℓ_2-norm setting [BRS18b; Dal+18; LS18b; Xu+20] and the ℓ_∞-norm settings [PW19]. In general, however, the guarantees provided by work to date are not sharp. For instance, the bounds in [Dal+18] scale exponentially in relevant parameters of the problem, whereas the papers [BRS18b; LS18b; Xu+20] do not capture the correct "variance" of the problem instance at hand. A subset of the current authors [PW19] derived ℓ_∞ bounds on policy evaluation for the plug-in estimator. These results were shown to be locally minimax optimal in certain regions of the parameter space. There has also been some recent focus on obtaining instance-dependent guarantees in online reinforcement learning settings [MMM14]. This has resulted in more practically useful algorithms that provide, for instance, horizon-independent regret bounds for certain episodic MDPs [JA18; ZB19], thereby improving upon worst-case bounds [AOM17]. Recent work has also established some instance-dependent bounds, albeit not sharp over the whole parameter space, for the problem of state-action value function estimation in Markov decision processes, for both ordinary Q-learning [Wai19c] and a variance-reduced improvement [Wai19e].

[1]There were some errors in the results of Korda and La [KL15] that were pointed out by both Lakshminarayanan and Szepesvari [LS18b] and Xu et al. [Xu+20].

Contents of this chapter

In this chapter, we study stochastic approximation algorithms for evaluating the value function of a Markov reward process in the discounted setting. Our goal is to provide a sharp characterization of performance in the ℓ_∞-norm, for procedures that are given access to state transitions and reward samples under the generative model. In practice, temporal difference learning is typically applied with an additional layer of (linear) function approximation. In the current chapter, so as to bring the instance dependence into sharp focus, we study the algorithms without this function approximation step. In this context, we tell a story with three parts, as detailed below:

Local minimax lower bounds: Global minimax analysis provides bounds that hold uniformly over large classes of models. In this chapter, we seek to gain a more refined understanding of how the difficulty of policy evaluation varies as a function of the instance. In order to do so, we undertake an analysis of the local minimax risk associated with a problem. We first prove an asymptotic statement (Proposition 1) that characterizes the local minimax risk up to a logarithmic factor; it reveals the relevance of two functionals of the instance that we define. In proving this result, we make use of the classical asymptotic minimax theorem [Háj72; Le 72; LY00]. We then refine this analysis by deriving a *non-asymptotic* local minimax bound, as stated in Theorem 1, which is derived using the non-asymptotic local minimax framework of Cai and Low [CL04], an approach that builds upon the seminal concept of hardest local alternatives that can be traced back to Stein [Ste56].

Non-asymptotic suboptimality of iterate averaging: Our local minimax lower bounds raise a natural question: Do standard procedures for policy evaluation achieve these instance-specific bounds? In Section 2.3, we address this question for the TD(0) algorithm with iterate averaging. Via a careful simulation study, we show that for many popular stepsize choices, the algorithm *fails* to achieve the correct instance-dependent rate in the non-asymptotic setting, even when the sample size is quite large. This is true for both the constant stepsize, as well as polynomial stepsizes of various orders. Notably, the algorithm with polynomial stepsizes of certain orders achieves the local risk in the asymptotic setting (see Theorem 1).

Non-asymptotic optimality of variance reduction: In order to remedy this issue with iterate averaging, we propose and analyze a variant of TD learning with variance reduction, showing both through theoretical (see Theorem 2) and numerical results (see Figure 2.3) that this algorithm achieves the correct instance-dependent rate provided the sample size is larger than an explicit threshold. Thus, this algorithm is provably better than TD(0) with iterate averaging.

Notation

For a positive integer n, let $[n] := \{1, 2, \ldots, n\}$. For a finite set S, we use $|S|$ to denote its cardinality. We use $c, C, c_1, c_2, \ldots$ to denote universal constants that may change from line to line. We let $\mathbf{1}$ denote the all-ones vector in $\mathbb{R}^D$. Let e_j denote the jth standard basis vector in $\mathbb{R}^D$. We let $v_{(i)}$ denote the i-th order statistic of a vector v, i.e., the i-th largest entry of v. For a pair of vectors (u, v) of compatible dimensions, we use the notation $u \preceq v$ to indicate that the difference vector $v - u$ is entrywise non-negative. The relation $u \succeq v$ is defined analogously. We let $|u|$ denote the entrywise absolute value of a vector $u \in \mathbb{R}^D$; squares and square-roots of vectors are, analogously, taken entrywise. Note that for a positive scalar λ, the statements $|u| \preceq \lambda \cdot \mathbf{1}$ and $\|u\|_\infty \leq \lambda$ are equivalent. Finally, we let $\|\mathbf{M}\|_{1,\infty}$ denote the maximum ℓ_1-norm of the rows of a matrix $\mathbf{M}$, and refer to it as the $(1, \infty)$-operator norm of a matrix. More generally, for scalars $q, p \geq 1$, we define $\|\mathbf{M}\|_{p,q} := \sup_{\|x\|_p \leq 1} \|\mathbf{M}x\|_q$. We let $\mathbf{M}^\dagger$ denote the Moore-Penrose pseudoinverse of a matrix $\mathbf{M}$.

2.2 Background and problem formulation

We begin by introducing the basic mathematical formulation of Markov reward processes (MRPs) and generative observation models.

Markov reward processes and value functions

We study MRPs defined on a finite set of D states, which we index by the set $[D] \equiv \{1, 2, \ldots, D\}$. The state evolution over time is determined by a set of transition functions, $\{P(\cdot|i),\ i \in [D]\}$. Note that each such transition function can be naturally associated with a D-dimensional vector; denote the i-th such vector as p_i. We let $\mathbf{P} \in [0, 1]^{D \times D}$ denote a row-stochastic (Markov) transition matrix, where row i of this matrix contains the vector p_i. Also associated with an MRP is a population reward function, $r : [D] \mapsto \mathbb{R}$, possessing the semantics that a transition from state i results in the reward $r(i)$. For convenience, we engage in a minor abuse of notation by letting r also denote a vector of length D; here r_i corresponds to the reward obtained at state i.

We formulate the long-term value of a state in the MRP in terms of the infinite-horizon, discounted reward. This value function (denoted here by the vector $\theta^* \in \mathbb{R}^D$) can be computed as the unique solution of the Bellman fixed-point relation, $\theta^* = r + \gamma \mathbf{P}\theta^*$.

Observation model

In the learning setting, the pair $(\mathbf{P}, r)$ is unknown, and we accordingly assume access to a black box that generates samples from the transition and reward functions. In

this chapter, we operate under a setting known as the synchronous or generative setting [KS99]; this setting is also often referred to as the "i.i.d. setting" in the policy evaluation literature. For a given sample index, $k \in \{1, 2, \ldots, N\}$ and for each state $j \in [D]$, we observe a random next state

$$X_{k,j} \sim P(\cdot|j) \qquad \text{for } j \in [D]. \tag{2.1a}$$

We collect these transitions in a matrix $\mathbf{Z}_k$, which by definition contains one 1 in each row: the 1 in the j-th row corresponds to the index of state $X_{k,j}$. We also observe a random reward vector $R_k \in \mathbb{R}^D$, where the rewards are generated independently across states with[2]

$$R_{k,j} \sim \mathcal{N}(r_j, \sigma_r^2). \tag{2.1b}$$

Given these samples, define the k-th (noisy) linear operator $\widehat{\mathcal{T}}_k : \mathbb{R}^D \mapsto \mathbb{R}^D$ whose evaluation at the point θ is given by

$$\widehat{\mathcal{T}}_k(\theta) = R_k + \gamma \mathbf{Z}_k \theta. \tag{2.2}$$

The construction of these operators is inspired by the fact that we are interested in computing the fixed point of the population operator,

$$\mathcal{T} : \theta \mapsto r + \gamma P \theta, \tag{2.3}$$

and a classical and natural way to do so is via a form of stochastic approximation known as temporal difference learning, which we describe next.

Temporal difference learning and its variants

Classical temporal difference (TD) learning algorithms are parametrized by a sequence of stepsizes, $\{\alpha_k\}_{k \geq 1}$, with $\alpha_k \in (0, 1]$. Starting with an initial vector $\theta_1 \in \mathbb{R}^D$, the TD updates take the form

$$\theta_{k+1} = (1 - \alpha_k)\theta_k + \alpha_k \widehat{\mathcal{T}}_k(\theta_k) \quad \text{for } k = 1, 2, \ldots. \tag{2.4}$$

In the sequel, we discuss three popular stepsize choices:

$$\texttt{Constant stepsize:} \qquad \alpha_k = \alpha, \qquad \text{where} \qquad 0 < \alpha \leq \alpha_{\max}. \tag{2.5a}$$

$$\texttt{Polynomial stepsize:} \qquad \alpha_k = \frac{1}{k^\omega} \qquad \text{for some } \omega \in (0, 1). \tag{2.5b}$$

$$\texttt{Recentered-linear stepsize:} \qquad \alpha_k = \frac{1}{1 + (1 - \gamma)k}. \tag{2.5c}$$

[2]All of our upper bounds extend with minor modifications to the sub-Gaussian reward setting.

In addition to the TD sequence (2.4), it is also natural to perform *Polyak-Ruppert averaging*, which produces a parallel sequence of averaged iterates

$$\widetilde{\theta}_k = \frac{1}{k} \sum_{j=1}^{k} \theta_j \quad \text{for } k = 1, 2, \ldots. \tag{2.6}$$

Such averaging schemes were introduced in the context of general stochastic approximation by Polyak [PJ92] and Ruppert [Rup88a]. A large body of theoretical literature demonstrates that such an averaging scheme improves the rates of convergence of stochastic approximation when run with overly "aggressive" stepsizes.

2.3 Main results

We turn to the statements of our main results and discussion of their consequences. All of our statements involve certain measures of the local complexity of a given problem, which we introduce first. We then turn to the statement of lower bounds on the ℓ_∞-norm error in policy evaluation. In Section 7.3, we prove two lower bounds. Our first result, stated as Proposition 1, is asymptotic in nature (holding as the sample size $N \to +\infty$). Our second lower bound, stated as Theorem 1, provides a result that holds for a range of finite sample sizes. Given these lower bounds, it is then natural to wonder about known algorithms that achieve them. Concretely, does the TD(0) algorithm combined with Polyak-Ruppert averaging achieve these instance-dependent bounds? In Section 2.3, we undertake a careful empirical study of this question, and show that in the non-asymptotic setting, this algorithm fails to match the instance-dependent bounds. This finding sets up the analysis in Section 2.3, where we introduce a variance-reduced version of TD(0), and prove that it does achieve the instance-dependent lower bounds from Theorem 1 up to a logarithmic factor in dimension.

Local complexity measures: Recall the generative observation model described in Section 2.2. For a transition matrix P, we write $\mathbf{Z} \sim P$ to denote a draw of a random matrix with $\{0, 1\}$ entries and a single one in each row (with the position of the one in row $\mathbf{Z}_j$ determined by sampling from the multinomial distribution specified by p_j). For a fixed vector $\theta \in \mathbb{R}^D$, note that $(\mathbf{Z} - P)\theta$ is a random vector in $\mathbb{R}^D$, and define its covariance matrix as follows:

$$\Sigma_P(\theta) = \operatorname{cov}_{\mathbf{Z} \sim P} \left((\mathbf{Z} - P)\theta \right). \tag{2.7}$$

We often use $\Sigma(\theta)$ as a shorthand for $\Sigma_P(\theta)$ when the underlying transition matrix P is clear from the context.

With these definitions in hand, define the complexity measures

$$\nu(P,\theta) := \max_{\ell \in [D]} \left(e_\ell^\top (I_d - \gamma P)^{-1} \Sigma(\theta) (I_d - \gamma P)^{-\top} e_\ell \right)^{1/2}, \quad \text{and} \tag{2.8a}$$

$$\rho(P,r) := \sigma_r \left\| (I_d - \gamma P)^{-1} \right\|_{2,\infty} \equiv \sigma_r \max_{\|u\|_2 = 1} \|(I_d - \gamma P)^{-1} u\|_\infty. \tag{2.8b}$$

Note that $\nu(P,\theta)$ corresponds to the maximal variance of the random vector $(I_d - \gamma P)^{-1}(\mathbf{Z} - P)\theta$. As we demonstrate shortly, the quantities $\nu(P,\theta^*)$ and $\rho(P,r)$ govern the local complexity of estimating the value function θ^* induced by the instance (P,r) under the observation model (2.1). A portion of our results also involve the quantity

$$b(\theta) := \frac{\|\theta\|_{\text{span}}}{1-\gamma}, \tag{2.8c}$$

where $\|\theta\|_{\text{span}} = \max_{j \in [D]} \theta_j - \min_{j \in [D]} \theta_j$ is the span seminorm.

Local minimax lower bound

Throughout this section, we use the letter $\mathcal{P}$ to denote an individual problem instance, $\mathcal{P} = (P,r)$, and use $\theta(\mathcal{P}) := \theta^* = (I_d - \gamma P)^{-1} r$ to denote the *target* of interest. The aim of this section is to establish *instance-specific* lower bounds for estimating $\theta(\mathcal{P})$ under the observation model (2.1). In order to do so, we adopt a local minimax approach.

The remainder of this the section is organized as follows. In Section 2.3, we prove an asymptotic local minimax lower bound, valid as the sample size N tends to infinity. It gives an explicit Gaussian limit for the rescaled error that can be achieved by any procedure. The asymptotic covariance in this limit law depends on the problem instance, and is very closely related to the functionals $\nu(P,\theta^*)$ and $\rho(P,r)$ that we have defined. Moreover, we show that this limit can be achieved—in the asymptotic limit—by the TD algorithm combined with Polyak-Ruppert averaging. While this provides a useful sanity check, in practice we implement estimators using a finite number of samples N, so it is important to obtain non-asymptotic lower bounds for a full understanding. With this motivation, Section 2.3 provides a new, *non-asymptotic* instance-specific lower bound for the policy evaluation problem. We show that the quantities $\nu(P,\theta^*)$ and $\rho(P,r)$ also cover the instance-specific complexity in the finite-sample setting. In proving this non-asymptotic lower bound, we build upon techniques in the statistical literature based on constructing hardest one-dimensional alternatives [Bir83a; CL15b; DL87; DL91; Ste56]. As we shall see in later sections, while the TD algorithm with averaging is instance-specific optimal in the asymptotic setting, it *fails* to achieve our non-asymptotic lower bound.

Asymptotic local minimax lower bound

Our first approach towards an instance-specific lower bound is an asymptotic one, based on classical local asymptotic minimax theory. For regular and parametric families, the Hájek–Le Cam local asymptotic minimax theorem [Háj72; Le 72; LY00] shows that the Fisher information—an instance-specific functional—characterizes a fundamental asymptotic limit. Our model class is both parametric and regular (cf. Eq. (2.1)), and so this classical theory applies to yield an asymptotic local minimax bound. Some additional work is needed to relate this statement to the more transparent complexity measures $\nu(P, \theta^*)$ and $\rho(P, r)$ that we have defined.

In order to state our result, we require some additional notation. Fix an instance $\mathcal{P} = (P, r)$. For any $\epsilon > 0$, we define an ϵ-neighborhood of problem instances by

$$\mathfrak{N}(\mathcal{P}; \epsilon) = \{\mathcal{P}' = (P', r') : \|P - P'\|_F + \|r - r'\|_2 \leq \epsilon\}.$$

Adopting the ℓ_∞-norm as the loss function, the *local asymptotic minimax risk* is given by

$$\mathfrak{M}_\infty(\mathcal{P}) \equiv \mathfrak{M}_\infty(\mathcal{P}; \|\cdot\|_\infty) = \lim_{c \to \infty} \liminf_{N \to \infty} \inf_{\hat{\theta}_N} \sup_{\mathcal{Q} \in \mathfrak{N}(\mathcal{P}; c/\sqrt{N})} \mathbb{E}_{\mathcal{Q}} \left[\sqrt{N} \left\| \hat{\theta}_N - \theta(\mathcal{Q}) \right\|_\infty \right]. \quad (2.9)$$

Here the infimum is taken over all estimators $\widehat{\theta}_N$ that are measurable functions of N i.i.d. observations drawn according to the observation model (2.1).

Our first main result characterizes the local asymptotic risk $\mathfrak{M}_\infty(\mathcal{P})$ exactly, and shows that it is attained by stochastic approximation with averaging. Recall the Polyak-Ruppert (PR) sequence $\{\widetilde{\theta}_k\}_{k \geq 1}$ defined in Eq. (2.6), and let $\{\widetilde{\theta}_k^\omega\}_{k \geq 1}$ denote this sequence when the underlying SA algorithm is the TD update with the polynomial stepsize sequence (2.5b) with exponent ω.

Proposition 1. *Let $Z \in \mathbb{R}^D$ be a multivariate Gaussian with zero mean and covariance matrix*

$$(I_d - \gamma P)^{-1}(\gamma^2 \Sigma_P(\theta(\mathcal{P})) + \sigma_r^2 I_d)(I_d - \gamma P)^{-\top}. \quad (2.10a)$$

Then the local asymptotic minimax risk at problem instance $\mathcal{P}$ is given by

$$\mathfrak{M}_\infty(\mathcal{P}) = \mathbb{E}[\|Z\|_\infty]. \quad (2.10b)$$

Furthermore, for each problem instance $\mathcal{P}$ and scalar $\omega \in (1/2, 1)$, this limit is achieved by the TD algorithm with an ω-polynomial stepsize and PR-averaging:

$$\lim_{N \to \infty} \sqrt{N} \cdot \mathbb{E}\left[\|\widetilde{\theta}_N^\omega - \theta(\mathcal{P})\|_\infty \right] = \mathbb{E}[\|Z\|_\infty]. \quad (2.10c)$$

With the convention that $\theta^* \equiv \theta(\mathcal{P})$, a short calculation bounding the maximum absolute value of sub-Gaussian random variables (see, e.g., Ex. 2.11 in Wainwright [Wai19b]) yields the sandwich relation

$$\gamma \nu(P, \theta^*) + \rho(P, r) \leq \mathbb{E}[\|Z\|_\infty] \leq \sqrt{2 \log D} \cdot (\gamma \nu(P, \theta^*) + \rho(P, r)),$$

so that Proposition 1 shows that, up to a logarithmic factor in dimension D, the local asymptotic minimax risk is entirely characterized by the functional $\gamma\nu(P,\theta^*)+\rho(P,r)$.

It should be noted that lower bounds similar to Eq. (2.10b) have been shown for specific classes of stochastic approximation algorithms [TP74]. However, to the best of our knowledge, a local minimax lower bound—one applying to any procedure that is a measurable function of the observations—is not available in the existing literature.

Furthermore, Eq. (2.10c) shows that stochastic approximation with polynomial stepsizes and averaging attains the exact local asymptotic risk. Our proof of this result essentially mirrors that of Polyak and Juditsky [PJ92], and amounts to verifying their assumptions under the policy evaluation setting. Given this result, it is natural to ask if averaging is optimal also in the non-asymptotic setting; answering this question is the focus of the next two sections of the chapter.

Non-asymptotic local minimax lower bound

Proposition 1 provides an instance-specific lower bound on $\theta(\mathcal{P})$ that holds asymptotically. In order to obtain a non-asymptotic guarantee, we borrow ideas from the non-asymptotic framework introduced by Cai and Low [CL15b] for nonparametric shape-constrained inference. Adapting their definition of local minimax risk to our problem setting, given the loss function $L(\theta-\theta^*)=\|\theta-\theta^*\|_\infty$, the (normalized) *local non-asymptotic minimax risk* for $\theta(\cdot)$ at instance $\mathcal{P}=(P,r)$ is given by

$$\mathfrak{M}_N(\mathcal{P}) = \sup_{\mathcal{P}'} \inf_{\widehat{\theta}_N} \max_{\mathcal{Q}\in\{\mathcal{P},\mathcal{P}'\}} \sqrt{N}\cdot\mathbb{E}_{\mathcal{Q}}\left[\|\widehat{\theta}_N-\theta(\mathcal{Q})\|_\infty\right]. \tag{2.11}$$

Here the infimum is taken over all estimators $\widehat{\theta}_N$ that are measurable functions of N i.i.d. observations drawn according to the observation model (2.1), and the normalization by $\sqrt{N}$ is for convenience. The definition (2.11) is motivated by the notion of the hardest one-dimensional alternative [Vaa98b, Ch. 25]. Indeed, given an instance $\mathcal{P}$, the local non-asymptotic risk $\mathfrak{M}_N(\mathcal{P})$ first looks for the hardest alternative $\mathcal{P}'$ against $\mathcal{P}$ (which should be local around $\mathcal{P}$), then measures the worst-case risk over $\mathcal{P}$ and its (local) hardest alternative $\mathcal{P}'$.

With this definition in hand, we lower bound the local non-asymptotic minimax risk using the complexity measures $\nu(P,\theta^*)$ and $\rho(P,r)$ defined in Eq. (2.8):

Theorem 1. *There exists a universal constant $c>0$ such that for any instance $\mathcal{P}=(P,r)$, the local non-asymptotic minimax risk is lower bounded as*

$$\mathfrak{M}_N(\mathcal{P}) \geq c\Big(\gamma\nu(P,\theta^*)+\rho(P,r)\Big). \tag{2.12}$$

This bound is valid for all sample sizes N that satisfy

$$N \geq N_0 := \max\left\{\frac{\gamma^2}{(1-\gamma)^2}, \frac{b^2(\theta^*)}{\nu^2(P,\theta^*)}\right\}. \tag{2.13}$$

A few comments are in order. First, it is natural to wonder about the necessity of condition (2.13) on the sample size N in our lower bound. Our past work provides upper bounds on the ℓ_∞-error of the plugin estimator [PW19], and these results also require a bound of this type. In fact, when the rewards are observed with noise (i.e., for any $\sigma_r > 0$), the condition $N \gtrsim \frac{\gamma^2}{(1-\gamma)^2}$ is natural, since it is necessary in order to obtain an estimate of the value function with $\mathcal{O}(1)$ error. On the other hand, in the special case of deterministic rewards ($\sigma_r = 0$), it is interesting to ask how the fundamental limits of the problem behave in the absence of this condition.

Second, note that the non-asymptotic lower bound (2.12) is closely connected to the asymptotic local minimax bound from Proposition 1. In particular, for any sample size N satisfying the lower bound (2.13), our non-asymptotic lower bound (2.12) coincides with the asymptotic lower bound (2.10b) up to a constant factor. Thus, it cannot be substantially sharpened. The finite-sample nature of the lower bound (2.12) is a powerful tool for assessing optimality of procedures: it provides a performance benchmark that holds over a large range of finite sample sizes N. Indeed, in the next section, we study the performance of the TD learning algorithm with Polyak-Ruppert averaging. While this procedure achieves the local minimax lower bound asymptotically, as guaranteed by Eq. (2.10c) in Proposition 1, it falls short of doing so in natural *finite-sample* scenarios.

Suboptimality of averaging

Polyak and Juditsky [PJ92] provide a general set of conditions under which a given stochastic-approximation (SA) algorithm, when combined with Polyak-Ruppert averaging, is guaranteed to have asymptotically optimal behavior. For the current problem, the bound (2.10c) in Proposition 1, which is proved using the Polyak-Juditsky framework, shows that SA with polynomial stepsizes and averaging have this favorable asymptotic property.

However, asymptotic theory of this type gives no guarantees in the finite-sample setting. In particular, suppose that we are given a sample size N that scales as $(1-\gamma)^{-2}$, as specified in our lower bounds. Does the averaged TD(0) algorithm exhibit optimal behavior in this non-asymptotic setting? In this section, we answer this question in the negative. More precisely, we describe a parameterized family of Markov reward processes, and provide careful simulations that reveal the suboptimality of TD without averaging.

A simple construction

The lower bound in Theorem 1 predicts a range of behaviors depending on the pair $\nu(P, \theta^*)$ and $\rho(P, r)$. In order to observe a large subset of these behaviors, it suffices to consider a very simple MRP, $\mathcal{P} = (P, r)$ with $D = 2$ states, as illustrated in Figure 2.1.

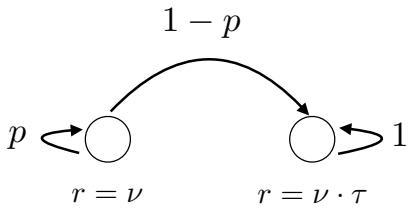


Figure 2.1. Illustration of the 2-state MRP used in the simulation. The triple of scalars (p, ν, τ), along with the discount factor γ, are parameters of the construction. The chain remains in state 1 with with probability p and transitions to state 2 with probability $1 - p$; on the other hand, state 2 is absorbing. The rewards in states 1 and 2 are deterministic, specified by ν and $\nu\tau$, respectively.

In this MRP, the transition matrix $P \in \mathbb{R}^{2\times 2}$ and reward vector $r \in \mathbb{R}^2$ take the form

$$P = \begin{bmatrix} p & 1-p \\ 0 & 1 \end{bmatrix}, \quad \text{and} \quad r = \begin{bmatrix} \nu \\ \nu\tau \end{bmatrix}.$$

Here the triple (p, ν, τ), along with the discount factor $\gamma \in [0, 1)$, are parameters of the construction.

In order to parameterize this MRP in a scalarized manner, we vary the triple (p, ν, τ) in the following way. First, we fix a scalar $\lambda \geq 0$, and then we set

$$p = \tfrac{4\gamma-1}{3\gamma}, \qquad \nu = 1 \quad \text{and} \quad \tau = 1 - (1-\gamma)^{\lambda}.$$

Note that this sub-family of MRPs is fully parametrized by the pair (γ, λ). Let us clarify why this particular scalarization is interesting. It can be shown via simple calculations that the underlying MRP satisfies

$$\nu(P, \theta^*) \sim \left(\frac{1}{1-\gamma}\right)^{1.5-\lambda}, \quad \rho(P, r) = 0 \quad \text{and} \quad b(\theta^*) \sim \left(\frac{1}{1-\gamma}\right)^{2-\lambda},$$

where $\sim$ denotes equality that holds up to a constant pre-factor. Consequently, by Theorem 1 the minimax risk, measured in terms of the ℓ_∞-norm, satisfies

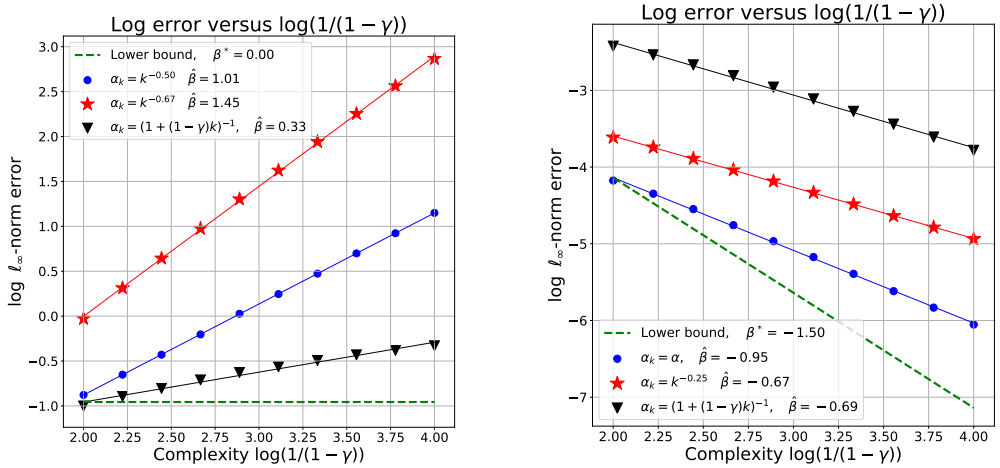
$$\mathfrak{M}_N(\mathcal{P}) \geq c \cdot \left(\frac{1}{1-\gamma}\right)^{1.5-\lambda}. \tag{2.14}$$

Thus, it is natural to study whether the TD(0) algorithm with PR averaging achieves this error.

A simulation study

In order to compare the behavior of averaged TD with the lower bound (2.14), we performed a series of experiments of the following type. For a fixed parameter λ in

the range $[0, 1.5]$, we generated a range of MRPs with different values of the discount factor γ. For each value of the discount parameter γ, we consider the problem of estimating θ^* using a sample size N set to be one of two possible values: namely, $N \in \left\{ \lceil \frac{8}{(1-\gamma)^2} \rceil, \lceil \frac{8}{(1-\gamma)^3} \rceil \right\}$.



(a) $\lambda = 0.5, \quad N = \lceil \frac{8}{(1-\gamma)^2} \rceil$. (b) $\lambda = 1.5, \quad N = \lceil \frac{8}{(1-\gamma)^3} \rceil$.

Figure 2.2. Log-log plots of the ℓ_∞-error versus the discount complexity parameter $1/(1-\gamma)$ for various algorithms. Each point represents an average over 1000 trials, with each trial simulations are for the 2-state MRP depicted in Figure 2.1 with the parameter choices $p = \frac{4\gamma - 1}{3\gamma}$, $\nu = 1$ and $\tau = 1 - (1-\gamma)^\lambda$. We have also plotted the least-squares fits through these points, and the slopes of these lines are provided in the legend. In particular, the legend contains the stepsize choice for averaged SA (denoted as α_k), the slope $\hat{\beta}$ of the least-squares line, and the ideal value β^* of the slope computed in equation 2.15. We also include the lower bound predicted by Theorem 1 for these examples as a dotted line for comparison purposes. Logarithms are to the natural base.

In Figure 2.2, we plot the ℓ_∞-error of the averaged SA, for constant stepsize (2.5a), polynomial-decay stepsize (2.5b) and recentered linear stepsize (2.5c), as a function of γ. The plots show the behavior for $\lambda \in \{0.5, 1.5\}$. Each point on each curve is obtained by averaging 1000 Monte Carlo trials of the experiment. Note that from our lower bound calculations above (2.14), the log ℓ_∞-error is related to the complexity $\log\left(\frac{1}{1-\gamma}\right)$ in a linear fashion; we use β^* to denote the slope of this idealized line. Simple algebra yields

$$\beta^* = \frac{1}{2} - \lambda \quad \text{for} \quad N = \frac{1}{(1-\gamma)^2}, \quad \text{and} \quad \beta^* = -\lambda \quad \text{for} \quad N = \frac{1}{(1-\gamma)^3}. \quad (2.15)$$

In other words, for an algorithm which achieves the lower bound predicted by our theory, we expect a linear relationship between the log ℓ_∞-error and log discount complexity $\log\left(\frac{1}{1-\gamma}\right)$, with the slope β^*.

Accordingly, for the averaged SA estimators with the stepsize choices in (2.5a)-(2.5c), we performed a linear regression to estimate the slopes between the log ℓ_∞-error and the log discount-complexity $\log\left(\frac{1}{1-\gamma}\right)$. The plot legend reports the stepsize choices α_k and the slope $\hat{\beta}$ of the fitted regression line. We also include the lower bound in the plots, as a dotted line along with its slope, for a visual comparison. We see that the slopes corresponding to the averaged SA algorithm are higher compared to the ideal slopes of the dotted lines. Stated differently, this means that the averaged SA algorithm does not achieve the lower bound with either the constant step or the polynomial-decay step. Overall, the simulations provided in this section demonstrate that the averaged SA algorithm, although guaranteed to be asymptotically optimal by Eq. (2.10c) in Proposition 1, does not yield the ideal non-asymptotic behavior.

Variance-reduced policy evaluation

In this section, we propose and analyze a variance-reduced version of the TD learning algorithm. As in standard variance-reduction schemes, such as SVRG [JZ13], our algorithm proceeds in epochs. In each epoch, we run a standard stochastic approximation scheme, but we recenter our updates in order to reduce their variance. The recentering uses an empirical approximation to the population Bellman operator $\mathcal{T}$.

We describe the behavior of the algorithm over epochs by a sequence of operators, $\{\mathcal{V}_m\}_{m\geq 1}$, which we define as follows. At epoch m, the method uses a vector $\bar{\theta}_m$ in order to recenter the update, where the vector $\bar{\theta}_m$ should be understood as the best current approximation to the unknown vector θ^*. In the ideal scenario, such a recentering would involve the quantity $\mathcal{T}(\bar{\theta}_m)$, where $\mathcal{T}$ denotes the population operator previously defined in Eq. (2.3). Since we lack direct access to the population operator $\mathcal{T}$, however, we use the Monte Carlo approximation

$$\widetilde{\mathcal{T}}_{N_m}(\bar{\theta}_m) := \frac{1}{N_m}\sum_{i\in\mathfrak{D}_m}\widehat{\mathcal{T}}_i(\bar{\theta}_m), \tag{2.16}$$

where the empirical operator $\widehat{\mathcal{T}}_i$ is defined in Eq. (2.2). Here the set $\mathfrak{D}_m$ is a collection of N_m i.i.d. samples, independent of all other randomness.

Given the pair $(\bar{\theta}_m, \widetilde{\mathcal{T}}_{N_m}(\bar{\theta}_m))$ and a stepsize $\alpha \in (0,1)$, we define the operator $\mathcal{V}_t$ on $\mathbb{R}^d$ as follows:

$$\theta \mapsto \mathcal{V}_k\left(\theta; \alpha, \bar{\theta}_m, \widetilde{\mathcal{T}}_{N_m}\right) := (1-\alpha)\theta + \alpha\left\{\widehat{\mathcal{T}}_k(\theta) - \widehat{\mathcal{T}}_k(\bar{\theta}_m) + \widetilde{\mathcal{T}}_{N_m}(\bar{\theta}_m)\right\}. \tag{2.17}$$

As defined in Eq. (2.2), the quantity $\widehat{\mathcal{T}}_t$ is a stochastic operator, where the randomness is independent of the set of samples $\mathfrak{D}_m$ used to define $\widetilde{\mathcal{T}}_{N_m}$. Consequently, the stochastic operator $\widehat{\mathcal{T}}_t$ is independent of the recentering vector $\widetilde{\mathcal{T}}_{N_m}(\bar{\theta}_m)$. Moreover, by construction, for each $\theta \in \mathbb{R}^D$, we have

$$\mathbb{E}\left[\widehat{\mathcal{T}}_k(\theta) - \widehat{\mathcal{T}}_k(\bar{\theta}_m) + \widetilde{\mathcal{T}}_{N_m}(\bar{\theta}_m)\right] = \mathcal{T}(\theta).$$

Thus, we see that $\mathcal{V}_k$ can be seen as an unbiased stochastic approximation of the population-level Bellman operator. As will be clarified in the analysis, the key effect of the recentering steps is to reduce its associated variance.

A single epoch

Based on the variance-reduced policy evaluation update defined in Eq. (2.17), we are now ready to define a single epoch of the overall algorithm. We index epochs using the integers $m = 1, 2, \ldots, M$, where M corresponds to the total number of epochs to be run. Epoch m requires as inputs the following quantities:

- a vector $\bar{\theta}$, which is chosen to be the output of the previous epoch,
- a positive integer K denoting the number of steps within the given epoch,
- a positive integer N_m denoting the number of samples used to calculate the Monte Carlo update (2.16),
- a sequence of stepsizes $\{\alpha_k\}_{k\geq 1}^{K}$ with $\alpha_k \in (0, 1)$, and
- a set of fresh samples $\{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m}$, with $|\mathfrak{E}_m| = N_m + K$. The first N_m samples are used to define the dataset $\mathfrak{D}_m$ that underlies the Monte Carlo update (2.16), whereas the remaining K samples are used in the K steps within each epoch.

We summarize the operations within a single epoch in Algorithm 1.

Algorithm 1 RunEpoch $(\bar{\theta}; K, N_m, \{\alpha_k\}_{k=1}^{K}, \{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m})$

1: Given (a) Epoch length K , (b) Recentering vector $\bar{\theta}$, (c) Recentering sample size N_m,
(d) Stepsize sequence $\{\alpha_k\}_{k\geq 1}^{K}$, (e) Samples $\{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m}$

2: Compute the recentering quantity $\widetilde{\mathcal{T}}_{N_m}(\bar{\theta}) := \frac{1}{N_m} \sum_{i\in\mathfrak{D}_m} \widehat{\mathcal{T}}_i(\bar{\theta})$

3: Initialize $\theta_1 = \bar{\theta}$

4: **for** $k = 1, 2, \ldots, K$ **do**

5: Compute the variance-reduced update:

$$\theta_{k+1} = \mathcal{V}_k\left(\theta_k; \alpha_k, \bar{\theta}, \widetilde{\mathcal{T}}_{N_m}\right)$$

6: **end for**

The choice of the stepsize sequence $\{\alpha_k\}_{k\geq 1}$ is crucial, and it also determines the epoch length K. Roughly speaking, it is sufficient to choose a large enough epoch length to ensure that the error is reduced by a constant factor in each epoch. In Section 2.3 to follow, we study three popular stepsize choices—the constant

stepsize (2.5a), the polynomial stepsize (2.5b) and the recentered linear stepsize (2.5c)—and provide lower bounds on the requisite epoch length in each case.

Overall algorithm

We are now ready to specify our variance-reduced policy-evaluation (VRPE) algorithm. The overall algorithm has five inputs: (a) an integer M, denoting the number of epochs to be run, (b) an integer K, denoting the length of each epoch, (c) a sequence of sample sizes $\{N_m\}_{m=1}^M$ denoting the number of samples used for recentering, (d) Sample batches $\{\{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m}\}_{m=1}^M$ to be used in m epochs, and (e) a sequence of stepsize $\{\alpha_k\}_{k\geq 1}$ to be used in each epoch. Given these five inputs, we summarize the overall procedure in Algorithm 2:

Algorithm 2 Variance-reduced policy evaluation (VRPE)

1: Given (a) Number of epochs M, (b) Epoch length K , (c) Recentering sample sizes $\{N_m\}_{m=1}^M$, (d) Sample batches $\{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m}$, for $m = 1, \ldots, M$, (e) Stepsize $\{\alpha_k\}_{k=1}^K$

2: Initialize at $\bar{\theta}_1$

3: **for** $m = 1, 2, \ldots, M$ **do**

4: $\quad \bar{\theta}_{m+1} = \text{RunEpoch}\left(\bar{\theta}_m; K, N_m, \{\alpha\}_{k=1}^K, \{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m}\right)$

5: **end for**

6: Return $\bar{\theta}_{M+1}$ as the final estimate

In the next section, we provide a detailed description on how to choose these input parameters for three popular stepsize choices (2.5a)–(2.5c). Finally, we reiterate that at epoch m, the algorithm uses $N_m + K$ new samples, and the samples used in the epochs are independent of each other. Accordingly, the total number of samples used in M epochs is given by $KM + \sum_{m=1}^M N_m$.

Instance-dependent guarantees

Given a desired failure probability, $\delta \in (0, 1)$, and a total sample size N, we specify the following choices of parameters in Algorithm 2:

$$\texttt{Number of epochs :} \quad M := \log_2\left(\frac{N(1-\gamma)^2}{8\log((8D/\delta)\cdot\log N)}\right) \tag{2.18a}$$

$$\texttt{Recentering sample sizes :} \quad N_m := 2^m \frac{4^2\cdot 9^2\cdot\log(8MD/\delta)}{(1-\gamma)^2} \qquad \text{for } m = 1, \ldots, M \tag{2.18b}$$

$$\texttt{Sample batches:} \quad \text{Partition the } N \text{ samples to obtain } \{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m} \text{ for } m = 1, \ldots M \tag{2.18c}$$

$$\texttt{Epoch length:} \quad K = \frac{N}{2M} \tag{2.18d}$$

In the following theorem statement, we use (c_1, c_2, c_3, c_4) to denote universal constants.

Theorem 2. *(a) Suppose that the input parameters of Algorithm 2 are chosen according to Eq. (2.18). Furthermore, suppose that the sample size N satisfies one of the following three stepsize-dependent lower bounds:*

(a) $\frac{N}{M} \geq c_1 \frac{\log(8ND/\delta)}{(1-\gamma)^3}$ *for recentered linear stepsize* $\alpha_k = \frac{1}{1+(1-\gamma)k}$,

(b) $\frac{N}{M} \geq c_2 \log(8ND/\delta) \cdot \left(\frac{1}{1-\gamma}\right)^{\left(\frac{1}{1-\omega} \vee \frac{2}{\omega}\right)}$ *for polynomial stepsize* $\alpha_k = \frac{1}{k^\omega}$ *with* $0 < \omega < 1$,

(c) $\frac{N}{M} \geq \frac{c_3}{\log\left(\frac{1}{1-\alpha(1-\gamma)}\right)}$ *for constant stepsize* $\alpha_k = \alpha \leq \frac{1}{5^2 \cdot 32^2} \cdot \frac{(1-\gamma)^2}{\log(8ND/\delta)}$.

Then for any initilization $\bar{\theta}_1$, the output $\bar{\theta}_{M+1}$ satisfies

$$\begin{aligned} \|\bar{\theta}_{M+1} - \theta^*\|_\infty \leq c_4 \cdot \left\|\bar{\theta}_1 - \theta^*\right\|_\infty \cdot \frac{\log^2((8D/\delta) \cdot \log N)}{N^2(1-\gamma)^4} \\ + c_4 \cdot \left\{ \sqrt{\frac{\log(8DM/\delta)}{N}} \cdot \Big(\gamma \cdot \nu(P, \theta^*) + \rho(P, r)\Big) + \frac{\log(8DM/\delta)}{N} \cdot b(\theta^*) \right\}, \end{aligned} \tag{2.19}$$

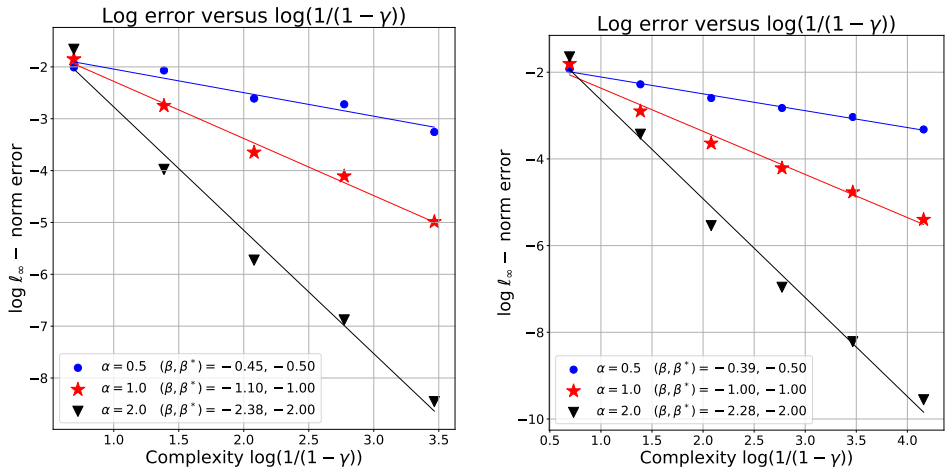
with probability exceeding $1 - \delta$.

See Section 2.4 for the proof of this theorem.

A few comments on the upper bound provided in Theorem 2 are in order. In order to facilitate a transparent discussion in this section, we use the notation $\gtrsim$ in order to denote a relation that holds up to logarithmic factors in the tuple $(N, D, (1-\gamma)^{-1})$.

Initialization dependence: The first term on the right-hand side of the upper bound (2.19) depends on the initialization $\bar{\theta}_1$. It should be noted that when viewed as a function of the sample size N, this initialization-dependent term decays at a faster rate compared to the other two terms. This indicates that the performance of Algorithm 2 does not depend on the initialization $\bar{\theta}_1$ in a significant way. A careful look at the proof (cf. Section 2.4) reveals that the coefficient of $\|\bar{\theta}_1 - \theta^*\|_\infty$ in the bound (2.19) can be made significantly smaller. In particular, for any $p \geq 1$ the first term in the right-hand side of bound (2.19) can be replaced by

$$c_4 \cdot \frac{\|\bar{\theta}_1 - \theta^*\|_\infty}{N^p} \cdot \frac{\log^p((8D/\delta) \cdot \log N)}{(1-\gamma)^{2p}},$$



(a) Recentered-linear stepsize (b) Polynomial stepsize

Figure 2.3. Log-log plots of the ℓ_∞-error versus the discount complexity parameter $1/(1-\gamma)$ for the VRPE algorithm. Each point is computed from an average over 1000 trials. Each trial entails drawing $N = \lceil \frac{8}{(1-\gamma)^3} \rceil$ samples from the 2-state MRP in Figure 2.1 with the parameter choices $p = \frac{4\gamma-1}{3\gamma}$, $\nu = 1$ and $\tau = 1 - (1-\gamma)^\lambda$. Each line on each plot represents a different value of λ, as labeled in the legend. We have also plotted the least-squares fits through these points, and the slopes of these lines are also provided in the legend. We also report the pair $(\hat{\beta}, \beta^*)$, where the coefficient $\hat{\beta}$ denotes the slope of the least-squares fit and β^* denotes the slope predicted from the lower bound calculation (2.15). (a) Performance of VRPE for the recentered linear stepsize (2.5c). (b) Performance of VRPRE with polynomially decaying stepsizes (2.5b) with $\omega = 2/3$.

by increasing the recentering sample size (2.18b) by a constant factor and changing the values of the absolute constants (c_1, c_2, c_3, c_4), with these values depending only on the value of p. We have stated and proved a version for $p = 2$. Assuming the number of samples N satisfies $N \geq (1-\gamma)^{-(2+\Delta)}$ for some $\Delta > 0$, the first term on the right-hand side of bound (2.19) can always be made smaller than the other two terms. In the sequel we show that each of the lower bound conditions (a)-(c) in the statement of Theorem 2 requires a lower bound condition $N \gtrsim (1-\gamma)^{-3}$.

Comparing the upper and lower bounds: The second and the third terms in (2.19) show the instance-dependent nature of the upper bound, and they are the dominating terms. Furthermore, assuming that the minimum sample size requirements from Theorems 1 and 2 are met, we find that the upper bound (2.19) matches the lower bound (2.12) up to logarithmic terms.

It is worthwhile to explicitly compute the minimum sample size requirements in Theorems 1 and 2. Ignoring the logarithmic terms and constant factors for the moment, unwrapping the lower bound conditions (a)-(c) in Theorem 2, we see that

for both the constant stepsize and the recentered linear stepsize the sample size needs to satisfy $N \gtrsim (1-\gamma)^{-3}$. For the polynomial stepsize $\alpha_k = \frac{1}{k^\omega}$, the sample size has to be at least $(1-\gamma)^{-\left(\frac{1}{1-\omega} \vee \frac{2}{\omega}\right)}$. Minimizing the last bound for different values of $\omega \in (0,1)$, we see that the minimum value is attained at $\omega = 2/3$, and in that case the bound (2.19) is valid when $N \gtrsim (1-\gamma)^{-3}$. Overall, for all the three stepsize choices discussed in Theorem 2 we require $N \gtrsim (1-\gamma)^{-3}$ in order to certify the upper bound. Returning to Theorem 1, from assumption (2.13) we see that in the best case scenario, Theorem 1 is valid as soon as $N \gtrsim (1-\gamma)^{-2}$. Putting together the pieces we find that the sample size requirement for Theorem 2 is more stringent than that of Theorem 1. Currently we do not know whether the minimum sample size requirements in Theorems 1 and 2 are necessary; answering this question is an interesting future research direction.

Simulation study: It is interesting to demonstrate the sharpness of our bounds via a simulation study, using the same scheme as our previous study of TD(0) with averaging. In Figure 2.3, we report the results of this study; see the figure caption for further details. At a high level, we see that the VRPE algorithm, with either the recentered linear stepsize (panel (a)) or the polynomial stepsize $t^{-2/3}$, produces errors that decay with the exponents predicted by our instance-dependent theory for $\lambda \in \{0.5, 1.0, 2.0\}$. See the figure caption for further details.

2.4 Proofs

We now turn to the proofs of our main results.

Proof of Proposition 1

Recall the definition of the matrix $\Sigma_P(\theta)$ from Eq. (2.7), and define the covariance matrix

$$V_{\mathcal{P}} = (I_d - \gamma P)^{-1}(\gamma^2 \Sigma_P(\theta) + \sigma_r^2 I_d)(I_d - \gamma P)^{-T}. \tag{2.20}$$

Recall that we use Z to denote a multivariate Gaussian random vector $Z \sim \mathcal{N}(0, V_{\mathcal{P}})$, and that the sequence $\{\widetilde{\theta}_k^\omega\}_{k \geq 1}$ is generated by averaging the iterates of stochastic approximation with polynomial stepsizes (2.5b) with exponent ω. With this notation, the two claims of the theorem are:

$$\mathfrak{M}_\infty(\mathcal{P}) = \mathbb{E}[\|Z\|_\infty], \quad \text{and} \tag{2.21a}$$

$$\lim_{N \to \infty} \mathbb{E}\left[\sqrt{N} \cdot \|\widetilde{\theta}_N^\omega - \theta^*\|_\infty\right] = \mathbb{E}[\|Z\|_\infty]. \tag{2.21b}$$

We now prove each of these claims separately.

Proof of Eq. (2.21a)

For the reader's convenience, let us state a version of the Hájek–Le Cam local asymptotic minimax theorem:

Theorem 3. *Let $\{P_{\vartheta'}\}_{\vartheta' \in \Theta}$ be a family of parametric models, quadratically mean differentiable with Fisher information matrices $J_{\vartheta'}$. Fix some parameter $\vartheta \in \Theta$, and consider a function $\psi : \Theta \to \mathbb{R}^D$ that is differentiable at ϑ. Then for any quasi-convex loss $L : \mathbb{R}^D \to \mathbb{R}$, we have:*

$$\lim_{c \to \infty} \lim_{N \to \infty} \inf_{\hat{\vartheta}_N} \sup_{\substack{\vartheta' \\ \|\vartheta' - \vartheta\|_2 \leq c/\sqrt{N}}} \mathbb{E}_{\vartheta'} \Big[L\Big(\sqrt{N} \cdot (\hat{\vartheta}_N - \vartheta') \Big) \Big] = \mathbb{E}[L(Z)], \tag{2.22}$$

where the infimum is taken over all estimators $\hat{\vartheta}_N$ that are measurable functions of N i.i.d. data points drawn from P_ϑ, and the expectation is taken over a multivariate Gaussian $Z \sim \mathcal{N}(0, \nabla\psi(\vartheta)^T J_\vartheta^\dagger \nabla\psi(\vartheta))$.

Returning to the problem at hand, let $\vartheta = (P, r)$ denote the unknown parameters of the model and let $\psi(\vartheta) = \theta(\mathcal{P}) = (I_d - \gamma P)^{-1} r$ denote the target vector. A direct application of Theorem 3 shows that

$$\mathfrak{M}_\infty(\mathcal{P}) = \mathbb{E}[\|Z\|_\infty] \quad \text{where } Z = \mathcal{N}(0, \nabla\psi(\vartheta)^T J_\vartheta^\dagger \nabla\psi(\vartheta)), \tag{2.23}$$

where J_ϑ is the Fisher information at ϑ. The following result provides a more explicit form of the covariance of Z:

Lemma 1. *We have the identity*

$$\nabla\psi(\vartheta)^T J_\vartheta^\dagger \nabla\psi(\vartheta) = (I_d - \gamma P)^{-1}(\gamma^2 \Sigma_P(\theta) + \sigma_r^2 I_d)(I_d - \gamma P)^{-T}. \tag{2.24}$$

Although the proof of this claim is relatively straightforward, it involves some lengthy and somewhat tedious calculations; we refer the reader to Section 2.6 for the proof.

Given the result from Lemma 1, the claim (2.21a) follows by substituting the relation (2.24) into (2.23).

Proof of Eq. (2.21b)

The proof of this claim follows from the results of Polyak and Juditsky [PJ92, Theorem 1], once their assumptions are verified for TD(0) with polynomial stepsizes. Recall that the TD iterates in Eq. (2.4) are given by the sequence $\{\theta_k\}_{k \geq 1}$, and that $\widetilde{\theta}_k^\omega$ denotes the k-th iterate generated by averaging.

For each $k \geq 1$, note the following equivalence between the notation in this chapter and that of Polyak and Juditsky [PJ92], or PJ for short:

$$x_k \equiv \theta_k, \qquad \gamma_k \equiv \alpha_k, \qquad A \equiv I_d - \gamma P, \quad \text{and} \quad \xi_k = (R_k - r) + (\mathbf{Z}_k - P)\theta_k.$$

Let us now verify the various assumptions in the PJ paper. Assumption 2.1 in the PJ paper holds by definition, since the matrix $I_d - \gamma P$ is Hurwitz. Assumption 2.2 in the PJ paper is also satisfied by the polynomial stepsize sequence for any exponent $\omega \in (0, 1)$.

It remains to verify the assumptions that must be satisfied by the noise sequence $\{\xi_k\}_{k \geq 1}$. In order to do so, write the k-th such iterate as

$$\xi_k = (R_k - r) + (\mathbf{Z}_k - P)\theta^* + (\mathbf{Z}_k - P)(\theta_k - \theta^*).$$

Since $\mathbf{Z}_k$ is independent of the sequence $\{\theta_i\}_{i=1}^k$, it follows that the condition

$$\lim_{N \to \infty} \mathbb{E}\left[\|\theta_N - \theta^*\|_2^2\right] \tag{2.25}$$

suffices to guarantee that Assumptions 2.3–2.5 in the PJ paper are satisfied. We now claim that for each $\omega \in (1/2, 1]$, condition (2.25) is satisfied by the TD iterates. Taking this claim as given for the moment, note that applying Theorem 1 of Polyak and Juditsky [PJ92] establishes claim (2.21b), for any exponent $\omega \in (1/2, 1)$.

It remains to establish condition (2.25). For any $\omega \in (1/2, 1]$, the sequence of stepsizes $\{\alpha_k\}_{k \geq 1}$ satisfies the conditions

$$\sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

Consequently, classical results due to Robbins and Munro [RM51, Theorem 2] guarantee ℓ^2-convergence of θ_N to θ^*.

Proof of Theorem 1

Throughout the proof, we use the notation $\mathcal{P} = (P, r)$ and $\mathcal{P}' = (P', r')$ to denote, respectively, the problem instance at hand and its alternative. Moreover, we use $\theta^* \equiv \theta(\mathcal{P})$ and $\theta(\mathcal{P}')$ to denote the associated target parameters for each of the two problems $\mathcal{P}$ and $\mathcal{P}'$. We use $\Delta_P = P - P'$ and $\Delta_r = r - r'$ to denote the differences of the parameters. For probability distributions, we use P and P' to denote the marginal distribution of a single observation under $\mathcal{P}$ and $\mathcal{P}'$, and use P^N and $(P')^N$ to denote the distribution of N i.i.d observations drawn from P or P', respectively.

Proof structure

We introduce two special classes of alternatives of interest, denoted as $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively:

$$\mathcal{S}_1 = \{\mathcal{P}' = (P', r') \mid r' = r\}, \quad \text{and} \quad \mathcal{S}_2 = \{\mathcal{P}' = (P', r') \mid P' = P\}.$$

In words, the class $\mathcal{S}_1$ consists of alternatives $\mathcal{P}'$ that have the same reward vector r as $\mathcal{P}$, but a different transition matrix P'. Similarly, the class $\mathcal{S}_2$ consists of alternatives

$\mathcal{P}'$ with the same transition matrix P, but a different reward vector. By restricting the alternative $\mathcal{P}'$ within class $\mathcal{S}_1$ and $\mathcal{S}_2$, we can define *restricted versions* of the local minimax risk, namely

$$\mathfrak{M}_N(\mathcal{P}; \mathcal{S}_1) \equiv \sup_{\mathcal{P}' \in \mathcal{S}_1} \inf_{\hat{\theta}_N} \max_{\mathcal{P} \in \{\mathcal{P}, \mathcal{P}'\}} \mathbb{E}_{\mathcal{P}} \left[\sqrt{N} \cdot \left\| \hat{\theta}_N - \theta(\mathcal{P}) \right\|_\infty \right], \quad \text{and} \tag{2.26a}$$

$$\mathfrak{M}_N(\mathcal{P}; \mathcal{S}_2) \equiv \sup_{\mathcal{P}' \in \mathcal{S}_2} \inf_{\hat{\theta}_N} \max_{\mathcal{P} \in \{\mathcal{P}, \mathcal{P}'\}} \mathbb{E}_{\mathcal{P}} \left[\sqrt{N} \cdot \left\| \hat{\theta}_N - \theta(\mathcal{P}) \right\|_\infty \right]. \tag{2.26b}$$

The main part of the proof involves showing that there is a universal constant $c > 0$ such that the lower bounds

$$\mathfrak{M}_N(\mathcal{P}; \mathcal{S}_1) \geq c \cdot \gamma \nu(P, \theta^*), \quad \text{and} \tag{2.27a}$$

$$\mathfrak{M}_N(\mathcal{P}; \mathcal{S}_2) \geq c \cdot \rho(P, r) \tag{2.27b}$$

both hold (assuming that the sample size N is sufficiently large to satisfy the condition (2.13)). Since we have $\mathfrak{M}_N(\mathcal{P}) \geq \max\{\mathfrak{M}_N(\mathcal{P}; \mathcal{S}_1), \mathfrak{M}_N(\mathcal{P}; \mathcal{S}_2)\}$, these lower bounds in conjunction imply the claim Theorem 1. The next section shows how to prove these two bounds.

Proof of the lower bounds (2.27a) and (2.27b):

Our first step is to lower bound the local minimax risk for each problem class in terms of a modulus of continuity between the Hellinger distance and the ℓ_∞-norm.

Lemma 2. *For each $\mathcal{S} \in \{\mathcal{S}_1, \mathcal{S}_2\}$, we have the lower bound $\mathfrak{M}_N(\mathcal{P}; \mathcal{S}) \geq \frac{1}{8} \cdot \underline{\mathfrak{M}}_N(\mathcal{P}; \mathcal{S})$, where we define*

$$\underline{\mathfrak{M}}_N(\mathcal{P}; \mathcal{S}) := \sup_{\mathcal{P}' \in \mathcal{S}} \left\{ \sqrt{N} \cdot \left\| \theta(\mathcal{P}) - \theta(\mathcal{P}') \right\|_\infty \mid d_{\mathrm{hel}}(P, P') \leq \frac{1}{2\sqrt{N}} \right\}. \tag{2.28}$$

The proof of Lemma 2 follows a relatively standard argument, one which reduces estimation to testing; see Section 2.7 for details.

This lemma allows us to focus our remaining attention on lower bounding the quantity $\underline{\mathfrak{M}}_N(\mathcal{P}; \mathcal{S})$. In order to do so, we need both a lower bound on the ℓ_∞-norm $\left\| \theta(\mathcal{P}) - \theta(\mathcal{P}') \right\|_\infty$ and an upper bound on the Hellinger distance $d_{\mathrm{hel}}(P, P')$. These two types of bounds are provided in the following two lemmas. We begin with lower bounds on the ℓ_∞-norm:

Lemma 3. *(a) For any $\mathcal{P}$ and for all $\mathcal{P}' \in \mathcal{S}_1$, we have*

$$\|\theta(\mathcal{P}) - \theta(\mathcal{P}')\|_\infty \geq \left(1 - \frac{\gamma}{1-\gamma} \|\Delta_P\|_\infty\right)_+ \cdot \left\| \gamma (I_d - \gamma P)^{-1} \Delta_P \theta^* \right\|_\infty. \tag{2.29a}$$

(b) For any $\mathcal{P}$ and for all $\mathcal{P}' \in \mathcal{S}_2$, we have

$$\|\theta(\mathcal{P}) - \theta(\mathcal{P}')\|_\infty \geq \left\| (I_d - \gamma P)^{-1} \Delta_r \right\|_\infty. \tag{2.29b}$$

See Section 2.7 for the proof of this claim.

Next, we require upper bounds on the Hellinger distance:

Lemma 4. *(a) For each $\mathcal{P}$ and for all $\mathcal{P}' \in \mathcal{S}_1$, we have*

$$d_{\mathrm{hel}}(P, P')^2 \leq \frac{1}{2} \sum_{i,j} \frac{\left((\Delta_P)_{i,j}\right)^2}{P_{i,j}}. \tag{2.30a}$$

(b) For each $\mathcal{P}$ and for all $\mathcal{P}' \in \mathcal{S}_2$, we have

$$d_{\mathrm{hel}}(P, P')^2 \leq \frac{1}{2\sigma_r^2} \|r_1 - r_2\|_2^2 . \tag{2.30b}$$

See Section 2.7 for the proof of this upper bound.

Using Lemmas 3 and 4, we can derive two different lower bounds. First, we have the lower bound $\underline{\mathfrak{M}}_N(\mathcal{P}; \mathcal{S}_1) \geq \underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1)$, where

$$\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1) \equiv \sup_{\mathcal{P}' \in \mathcal{S}_1} \left\{ \sqrt{N} \cdot \left(1 - \frac{\gamma \|\Delta_P\|_\infty}{1-\gamma}\right)_+ \cdot \left\|\gamma (I_d - \gamma P)^{-1} \Delta_P \theta^*\right\|_\infty \right. \\ \left. \mid \sum_{i,j} \frac{\left((\Delta_P)_{i,j}\right)^2}{P_{i,j}} \leq \frac{1}{2N} \right\}. \tag{2.31a}$$

Second, we have the lower bound $\underline{\mathfrak{M}}_N(\mathcal{P}; \mathcal{S}_2) \geq \underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_2)$, where

$$\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_2) \equiv \sup_{\mathcal{P}' \in \mathcal{S}_2} \left\{ \sqrt{N} \cdot \| \left(\mathbf{I} - \gamma \mathbf{P}\right)^{-1} \Delta_r \|_\infty \ \ \frac{1}{\sigma_r^2} \|r_1 - r_2\|_2 \leq \frac{1}{2N} \right\}. \tag{2.31b}$$

In order to complete the proofs of the two lower bounds (2.27a) and (2.27b), it suffices to show that

$$\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_2) \geq \frac{1}{\sqrt{2}} \cdot \rho(P, r), \quad \text{and} \tag{2.32a}$$

$$\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1) \geq \frac{1}{2\sqrt{2}} \cdot \gamma \nu(P, \theta^*). \tag{2.32b}$$

Proof of the bound (2.32a): This lower bound is easy to show—it follows from the definition:

$$\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_2) = \frac{\sigma_r}{\sqrt{2}} \left\|(I_d - \gamma P)^{-1} \Delta_r\right\|_\infty = \frac{1}{\sqrt{2}} \rho(P, r).$$

Proof of the bound (2.32b): The proof of this claim is much more delicate. Our strategy is to construct a special "hard" alternativ, $\overline{\mathcal{P}} \in \mathcal{S}_1$, that leads to a good lower bound on $\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1)$. Lemma 5 below is the main technical result that we require:

Lemma 5. *There exists some probability transition matrix $\bar{P}$ with the following properties:*

(a) It satisfies the constraint $\sum_{i,j} \frac{\left((\bar{P}-P)_{i,j}\right)^2}{P_{i,j}} \leq \frac{1}{2N}$.

(b) It satisfies the inequalities

$$\left\|\bar{P} - P\right\|_\infty \leq \frac{1}{\sqrt{2N}}, \quad \text{and} \quad \left\|\gamma(I_d - \gamma P)^{-1}(\bar{P} - P)\theta^*\right\|_\infty \geq \frac{\gamma}{\sqrt{2N}} \cdot \nu(P, \theta^*).$$

See Section 2.7 for the proof of this claim.

Given the matrix $\bar{P}$ guaranteed by this lemma, we consider the "hard" problem $\overline{\mathcal{P}} := (\bar{P}, r) \in \mathcal{S}_1$. From the definition of $\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1)$ in Eq. (2.31a), we have that

$$\begin{aligned}
\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1) &\geq \sqrt{N} \cdot \left(1 - \frac{\gamma}{1-\gamma}\left\|P - \bar{P}\right\|_\infty\right)_+ \cdot \left\|\gamma(I_d - \gamma\bar{P})^{-1}(P - \bar{P})\theta^*\right\|_\infty \\
&\geq \sqrt{N} \cdot \left(1 - \frac{\gamma}{1-\gamma} \cdot \frac{1}{\sqrt{2N}}\right)_+ \cdot \frac{\gamma}{\sqrt{2N}} \cdot \nu(P, \theta^*) \geq \frac{1}{2\sqrt{2}} \cdot \gamma\nu(P, \theta^*),
\end{aligned}$$

where the last inequality follows by the assumed lower bound $N \geq \frac{4\gamma^2}{(1-\gamma)^2}$. This completes the proof of the lower bound (2.32b).

Proof of Theorem 2

This section is devoted to the proof of Theorem 2, which provides the achievability results for variance-reduced policy evaluation.

Proof of part (a):

We begin with a lemma that characterizes the progress of Algorithm 2 over epochs:

Lemma 1. *Under the assumptions of Theorem 2 (a), there is an absolute constant c such that for each epoch $m = 1, \ldots, M$, we have:*

$$\begin{aligned}
\left\|\bar{\theta}_{m+1} - \theta^*\right\|_\infty \leq \frac{\|\bar{\theta}_m - \theta^*\|_\infty}{4} \\
+ c\left\{\sqrt{\frac{\log(8DM/\delta)}{N_m}}\Big(\gamma \cdot \nu(P, \theta^*) + \rho(P, r)\Big) + \frac{\log(8DM/\delta)}{N_m} \cdot b(\theta^*)\right\},
\end{aligned} \tag{2.33}$$

with probability exceeding $1 - \frac{\delta}{M}$.

Taking this lemma as given for the moment, let us complete the proof. We use the shorthand

$$\tau_m := \sqrt{\frac{\log(8DM/\delta)}{N_m}}\Big(\gamma \cdot \nu(P, \theta^*)\rho(P, r)\Big) \quad \text{and} \quad \eta_m := \frac{\log(8DM/\delta)}{N_m} \cdot b(\theta^*) \tag{2.34}$$

to ease notation, and note that $\frac{\tau_m}{\sqrt{2}} \leq \tau_{m+1}$ and $\frac{\eta_m}{2} \leq \eta_{m+1}$, for each $m \geq 1$. Using this notation and unwrapping the recursion relation from Lemma 1, we have

$$\begin{aligned}
\left\|\bar{\theta}_{M+1} - \theta^*\right\|_\infty &\leq \frac{\left\|\bar{\theta}_M - \theta^*\right\|_\infty}{4} + c(\tau_M + \eta_M) \\
&\overset{(i)}{\leq} \frac{\left\|\bar{\theta}_{M-1} - \theta^*\right\|_\infty}{4^2} + \frac{c}{2}(\tau_M + \eta_M) + c(\tau_M + \eta_M) \\
&\overset{(ii)}{\leq} \frac{\left\|\bar{\theta}_1 - \theta^*\right\|_\infty}{4^M} + 2c(\tau_M + \eta_M).
\end{aligned}$$

Here, step (i) follows by applying the one-step application of the recursion (2.33), and by using the upper bounds $\frac{\tau_m}{\sqrt{2}} \leq \tau_{m+1}$ and $\frac{\eta_m}{2} \leq \eta_{m+1}$. Step (ii) follows by repeated application of the recursion (2.33). The last inequality holds with probability at least $1 - \delta$ by a union bound over the M epochs.

It remains to express the quantities 4^M, τ_M and η_M—all of which are controlled by the recentering sample size N_M—in terms of the total number of available samples N. Towards this end, observe that the total number of samples used for recentering at M epochs is given by

$$\sum_{m=1}^{M} N_m \asymp 2^M \cdot \frac{\log(8MD/\delta)}{(1-\gamma)^2}.$$

Substituting the value of $M = \log_2\left(\frac{N(1-\gamma)^2}{8\log((8D/\delta)\cdot\log N)}\right)$ we have

$$c_1 N \leq N_M \asymp \sum_{m=1}^{M} N_m \leq \frac{N}{2},$$

where c_1 is a universal constant. Consequently, the total number of samples used by Algorithm 2 is given by

$$MK + \sum_{m=1}^{M} N_m \leq \frac{N}{2} + \frac{N}{2} = N,$$

where in the last equation we have used the fact that $MK = \frac{N}{2}$. Finally, using $M = \log_2\left(\frac{N(1-\gamma)^2}{8\log((8D/\delta)\cdot\log N)}\right)$ we have the following relation for some universal constant

c:

$$4^M = c \cdot \frac{N^2(1-\gamma)^4}{\log^2((8D/\delta) \cdot \log N)}$$

Putting together the pieces, we conclude that

$$\|\bar{\theta}_{M+1} - \theta^*\|_\infty \leq c_2 \left\|\bar{\theta}_1 - \theta^*\right\|_\infty \cdot \frac{\log^2((8D/\delta) \cdot \log N)}{N^2(1-\gamma)^4} + c_2 \left\{ \sqrt{\frac{\log(8DM/\delta)}{N}} \Big(\gamma \cdot \nu(P, \theta^*) + \rho(P, r) \Big) + \frac{\log(8DM/\delta)}{N} \cdot b(\theta^*) \right\},$$

for a suitable universal constant c_2. The last bound is valid with probability exceeding $1-\delta$ via the union bound. In order to complete the proof, it remains to prove Lemma 1, which we do in the following subsection.

Proof of Lemma 1

We now turn to the proof of the key lemma within the argument. We begin with a high-level overview in order to provide intuition. In the m-th epoch that updates the estimate from $\bar{\theta}_m$ to $\bar{\theta}_{m+1}$, the vector $\bar{\theta} \equiv \bar{\theta}_m$ is used to recenter the updates. Our analysis of the m-th epoch is based on a sequence of recentered operators $\{\mathcal{J}_k^m\}_{k \geq 1}$ and their population analogs $\mathcal{J}^m(\theta)$. The action of these operators on a point θ is given by the relations

$$\mathcal{J}_k^m(\theta) := \widehat{\mathcal{T}}_k(\theta) - \widehat{\mathcal{T}}_k(\bar{\theta}_m) + \widetilde{\mathcal{T}}_N(\bar{\theta}_m), \quad \text{and} \quad \mathcal{J}^m(\theta) := \mathcal{T}(\theta) - \mathcal{T}(\bar{\theta}_m) + \widetilde{\mathcal{T}}_N(\bar{\theta}_m). \tag{2.35a}$$

By definition, the updates within epoch m can be written as

$$\theta_{k+1} = (1 - \alpha_k)\, \theta_k + \alpha_k \mathcal{J}_k^m(\theta_k). \tag{2.35b}$$

Note that the operator $\mathcal{J}^m$ is γ-contractive in $\|\cdot\|_\infty$-norm, and as a result it has a unique fixed point, which we denote by $\widehat{\theta}_m$. Since $\mathcal{J}^m(\theta) = \mathbb{E}[\mathcal{J}_k^m(\theta)]$ by construction, when studying epoch m, it is natural to analyze the convergence of the sequence $\{\theta_k\}_{k \geq 1}$ to $\widehat{\theta}_m$.

Suppose that we have taken K steps within epoch m. Applying the triangle inequality yields the bound

$$\|\bar{\theta}_{m+1} - \theta^*\|_\infty = \|\theta_{K+1} - \theta^*\|_\infty \leq \left\|\theta_{K+1} - \widehat{\theta}_m\right\|_\infty + \left\|\widehat{\theta}_m - \theta^*\right\|_\infty. \tag{2.35c}$$

With this decomposition, our proof of Lemma 1 is based on two auxiliary lemmas that provide high-probability upper bounds on the two terms on the right-hand side of inequality (2.35c).

Lemma 2. *Let (c_1, c_2, c_3) be positive numerical constants, and suppose that the epoch length K satisfies one the following three stepsize-dependent lower bounds:*

(a) $K \geq c_1 \frac{\log(8KMD/\delta)}{(1-\gamma)^3}$ *for recentered linear stepsize* $\alpha_k = \frac{1}{1+(1-\gamma)k}$,

(b) $K \geq c_2 \log(8KMD/\delta) \cdot \left(\frac{1}{1-\gamma}\right)^{\left(\frac{1}{1-\omega} \vee \frac{2}{\omega}\right)}$ *for polynomial stepsize* $\alpha_k = \frac{1}{k^\omega}$ *with* $0 < \omega < 1$,

(c) $K \geq \frac{c_3}{\log\left(\frac{1}{1-\alpha(1-\gamma)}\right)}$ *for constant stepsize* $\alpha_k = \alpha \leq \frac{(1-\gamma)^2}{\log(8KMD/\delta)} \cdot \frac{1}{5^2 \cdot 32^2}$.

Then after K update steps with epoch m, the iterate θ_{K+1} satisfies the bound

$$\|\theta_{K+1} - \widehat{\theta}_m\|_\infty \leq \frac{1}{8}\|\bar{\theta}_m - \theta^*\|_\infty + \frac{1}{8}\|\widehat{\theta}_m - \theta^*\|_\infty \quad \text{with probability at least } 1 - \tfrac{\delta}{2M}. \tag{2.36}$$

See Section 2.8 for the proof of this claim.

Our next auxiliary result provides a high-probability bound on the difference $\|\widehat{\theta}_m - \theta^*\|_\infty$.

Lemma 3. *There is an absolute constant c_4 such that for any recentering sample size satisfying $N_m \geq 4^2 \cdot 9^2 \cdot \frac{\log(MD/\delta)}{(1-\gamma)^2}$, we have*

$$\|\widehat{\theta}_m - \theta^*\|_\infty \leq \tfrac{1}{9}\|\bar{\theta}_m - \theta^*\|_\infty + c_4 \left\{ \sqrt{\frac{\log(8DM/\delta)}{N_m}} \Big(\gamma \cdot \nu(P, \theta^*) + \rho(P, r) \Big) + \frac{\log(8DM/\delta)}{N_m} \cdot b(\theta^*) \right\},$$

with probability exceeding $1 - \frac{\delta}{2M}$.

See Section 2.8 for the proof of this claim.

With Lemmas 2 and 3 in hand, the remainder of the proof is straightforward. Recall from Eq. (2.34) the shorthand notation τ_m and η_m. Using our earlier bound (2.35c), we have that at the end of epoch m (which is also the starting point of epoch $m+1$),

$$\begin{aligned}
\left\|\bar{\theta}_{m+1} - \theta^*\right\|_\infty &\leq \|\theta_{K+1} - \widehat{\theta}_m\|_\infty + \|\widehat{\theta}_m - \theta^*\|_\infty \\
&\overset{(i)}{\leq} \left\{ \frac{\|\bar{\theta}_m - \theta^*\|_\infty}{8} + \frac{1}{8}\left\|\widehat{\theta}_m - \theta^*\right\|_\infty \right\} + \left\|\widehat{\theta}_m - \theta^*\right\|_\infty \\
&= \frac{\|\bar{\theta}_m - \theta^*\|_\infty}{8} + \frac{9}{8} \cdot \left\|\widehat{\theta}_m - \theta^*\right\|_\infty \\
&\overset{(ii)}{\leq} \frac{\|\bar{\theta}_m - \theta^*\|_\infty}{8} + \frac{1}{8}\left\{ \|\bar{\theta}_m - \theta^*\|_\infty + c_4(\tau_m + \eta_m) \right\} \\
&\leq \frac{\|\bar{\theta}_m - \theta^*\|_\infty}{4} + c_4(\tau_m + \eta_m),
\end{aligned}$$

where inequality (i) follows from Lemma 2(a), and inequality (ii) from Lemma 3. Finally, the sequence of inequalities above holds with probability at least $1 - \frac{\delta}{M}$ via a union bound. This completes the proof of Lemma 1.

Proof of Theorem 2, parts (b) and (c)

The proofs of Theorem 2 parts (b) and (c) require versions of Lemma 1 for the polynomial stepsize (2.5b) and constant stepsize (2.5a), respectively. These two versions of Lemma 1 can be obtained by simply replacing Lemma 2, part (a), by Lemma 2, parts (b) and (c), respectively, in the proof of Lemma 1.

2.5 Discussion

We have discussed the problem of policy evaluation in discounted Markov decision processes. The main contribution of this chapter is three-fold. First, we provided a non-asymptotic instance-dependent local-minimax bound on the ℓ_∞-error for the policy evaluation problem under the generative model. Next, via careful simulations, we showed that the standard TD-learning algorithm—even when combined with Polyak-Rupert iterate averaging—does not yield ideal non-asymptotic behavior as captured by our lower bound. In order to remedy this difficulty, we introduced and analyzed a variance-reduced version of the standard TD-learning algorithm which achieves our non-asymptotic instance-dependent lower bound up to logarithmic factors. Both the upper and lower bounds discussed in this chapter hold when the sample size is bigger than an explicit threshold; relaxing this minimum sample size requirement is an interesting future research direction. Finally, we point out that although we have focused on the tabular policy evaluation problem, the variance-reduced algorithm discussed in this chapter can be applied in more generality, and it would be interesting to explore applications of this algorithm to non-tabular settings.

2.6 Proofs of auxiliary lemmas for Proposition 1

In this section, we provide proofs of the auxiliary lemmas that underlie the proof of Proposition 1.

Proof of Lemma 1

The proof is basically a lengthy computation. For clarity, let us decompose the procedure into three steps. In the first step, we compute an explicit form for the inverse information matrix $J_\vartheta^\dagger$. In the second step, we evaluate the gradient $\nabla\psi(\vartheta)$. In the third and final step, we use the result in the previous two steps to prove the claim (2.24) of the lemma.

Step 1: In the first step, we evaluate $J_\vartheta^\dagger$. Recall that our data $(\mathbf{Z}, R)$ is generated as follows. We generate the matrix $\mathbf{Z}$ and the vector R independently. Each row of $\mathbf{Z}$ is generated independently. Its jth row, denoted by z_j, is sampled from a multinomial distribution with parameter p_j, where p_j denotes the jth row of P. The vector R is sampled from $\mathcal{N}(r, \sigma_r^2 I_d)$. Because of this independence structure, the Fisher information J_ϑ is a block diagonal matrix of the form

$$J_\vartheta = \begin{bmatrix} J_{p_1} & 0 & 0\dots & 0 & 0 \\ 0 & J_{p_2} & 0\dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0\dots & J_{p_D} & 0 \\ 0 & 0 & 0\dots & 0 & J_r \end{bmatrix}.$$

Here each sub-block matrix J_{p_j} is the Fisher information corresponding to the model where a single data Z_j is sampled from the multinomial distribution with parameter p_j, and J_r is the Fisher information corresponding to the model in which a single data point R is sampled from $\mathcal{N}(r, \sigma_r^2 I_d)$. Thus, the inverse Fisher information $J_\vartheta^\dagger$ is also a block diagonal matrix of the form

$$J_\vartheta^\dagger = \begin{bmatrix} J_{p_1}^\dagger & 0 & 0\dots & 0 & 0 \\ 0 & J_{p_2}^\dagger & 0\dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0\dots & J_{p_D}^\dagger & 0 \\ 0 & 0 & 0\dots & 0 & J_r^\dagger \end{bmatrix}. \tag{2.37}$$

It is easy to compute $J_{p_j}^\dagger$ and $J_r^\dagger$:

$$J_{p_j}^\dagger = \operatorname{diag}(p_j) - p_j p_j^T = \operatorname{cov}(Z_j - p_j) \qquad \text{for } j \in [D], \text{ and} \tag{2.38a}$$

$$J_r^\dagger = J_r^{-1} = \sigma_r^2 I. \tag{2.38b}$$

For a vector $q \in \mathbb{R}^D$, we use $\operatorname{diag}(q) \in \mathbb{R}^{D\times D}$ to denote the diagonal matrix with diagonal entries q_j.

Step 2: In the second step, we evaluate $\nabla\psi(\vartheta)$. Recall that $\psi(\vartheta) = (I_d - \gamma P)^{-1} r$. It is straightforward to see that

$$\nabla_r \psi(\vartheta) = (I_d - \gamma P)^{-1}. \tag{2.39}$$

Below we evaluate $\nabla_{p_j}\psi(\theta)$ for $j \in [D]$, where p_j is the jth row of the matrix P. We show that

$$\nabla_{p_j}\psi(\vartheta) = \gamma(I_d - \gamma P)^{-1} e_j \theta^T. \tag{2.40}$$

Here we recall $\theta = \psi(\vartheta) = (I_d - \gamma P)^{-1} r$.

To prove Eq. (2.40), we start with the following elementary fact: for the matrix inverse mapping $A \to A^{-1}$, we have $\frac{\partial A^{-1}}{\partial A_{jk}} = -A^{-1} e_j e_k^T A^{-1}$ for all $j, k \in [D]$. Combining this fact with chain rule, we find that

$$\frac{\partial \psi(\vartheta)}{\partial P_{jk}} = \gamma (I_d - \gamma P)^{-1} e_j e_k^T (I_d - \gamma P)^{-1} r = \gamma (I_d - \gamma P)^{-1} e_j \theta^T e_k,$$

valid for all $j, k \in [D]$. This immediately implies Eq. (2.40) since p_j is the vector with coordinates P_{jk}.

Step 3: In the third step, we evaluate $\nabla \psi(\vartheta)^T J_\vartheta^\dagger \nabla \psi(\vartheta)$. From Eq. (2.37), we observe that the inverse Fisher information $J_\vartheta^\dagger$ has a block structure. Consequently, we can write

$$\nabla \psi(\vartheta)^T J_\vartheta^\dagger \nabla \psi(\vartheta) = \sum_{j \in [D]} \nabla_{p_j} \psi(\vartheta)^T J_{p_j}^\dagger \nabla_{p_j} \psi(\vartheta) + \nabla_R \psi(\vartheta)^T J_R^\dagger \nabla_R \psi(\vartheta). \tag{2.41}$$

Combining Eqs. (2.38b) and (2.39) yields

$$\nabla_R \psi(\vartheta)^T J_R^\dagger \nabla_R \psi(\vartheta) = \sigma_r^2 (I_d - \gamma P)^{-1} (I_d - \gamma P)^{-T}. \tag{2.42}$$

Combining Eqs. (2.38a) and (2.40) yields

$$\begin{aligned} \nabla_{p_j} \psi(\vartheta)^T J_{p_j}^\dagger \nabla_{p_j} \psi(\vartheta) &= \gamma^2 (I_d - \gamma P)^{-1} e_j \theta^T \operatorname{cov}(Z_j - p_j) \theta e_j^T (I_d - \gamma P)^{-T} \\ &= \gamma^2 (I_d - \gamma P)^{-1} e_j \operatorname{cov}((Z_j - p_j)^T \theta) e_j^T (I_d - \gamma P)^{-T}, \end{aligned}$$

valid for each $j \in [D]$. Summing over $j \in [D]$ then leads to

$$\begin{aligned} \sum_{j \in [D]} \nabla_{p_j} \psi(\vartheta)^T J_{p_j}^\dagger \nabla_{p_j} \psi(\vartheta) &= \gamma^2 (I_d - \gamma P)^{-1} \Big(\sum_{j \in [D]} e_j \operatorname{cov}((Z_j - p_j)^T \theta) e_j^T \Big) (I_d - \gamma P)^{-T} \\ &= \gamma^2 (I_d - \gamma P)^{-1} \Sigma_P(\theta) (I_d - \gamma P)^{-T}, \end{aligned} \tag{2.43}$$

where the last line uses the definition of $\Sigma_P(\theta)$ in Eq. (2.7). Finally, substituting Eq. (2.42) and Eq. (2.43) into Eq. (2.41) yields the claim (2.24), which completes the proof of Lemma 1.

2.7 Proofs of auxiliary lemmas for Theorem 1

In this appendix, we detailed proofs of the auxiliary lemmas that underlie the proof of the non-asymptotic local minimax lower bound stated in Theorem 1.

Proof of Lemma 2

The proof uses the standard device of reducing estimation to testing (see, e.g., [Bir83a; Tsy09; Wai19b]). The first step is to lower bound the minimax risk over $\mathcal{P}$ and $\mathcal{P}'$ by its averaged risk:

$$\inf_{\hat{\theta}_N} \max_{\mathcal{P}\in\{\mathcal{P},\mathcal{P}'\}} \mathbb{E}_{\mathcal{P}}\left[\left\|\theta-\theta(\mathcal{P})\right\|_\infty\right] \geq \frac{1}{2}\left(\mathbb{E}_{P^N}\left[\left\|\hat{\theta}_N-\theta\right\|_\infty\right]+\mathbb{E}_{P'^N}\left[\left\|\hat{\theta}_N-\theta'\right\|_\infty\right]\right). \quad (2.44)$$

By Markov's inequality, for any $\delta \geq 0$, we have

$$\begin{aligned}\mathbb{E}_{P^N}\left[\left\|\hat{\theta}_N-\theta\right\|_\infty\right]+\mathbb{E}_{P'^N}\left[\left\|\hat{\theta}_N-\theta'\right\|_\infty\right] \\ \geq \delta\left[P^N\left(\left\|\hat{\theta}_N-\theta\right\|_\infty\geq\delta\right)+P'^N\left(\left\|\hat{\theta}_N-\theta'\right\|_\infty\geq\delta\right)\right].\end{aligned}$$

If we define $\delta_{01} := \frac{1}{2}\left\|\theta-\theta'\right\|_\infty$, then we have the implication

$$\left\|\theta-\theta\right\|_\infty < \delta_{01} \quad \Longrightarrow \quad \left\|\theta-\theta'\right\|_\infty > \delta_{01}, \quad (2.45)$$

from which it follows that

$$\begin{aligned}\mathbb{E}_{P^n}\left[\left\|\hat{\theta}_n-\theta\right\|_\infty\right]+\mathbb{E}_{P'^n}\left[\left\|\hat{\theta}_n-\theta'\right\|_\infty\right] &\geq \delta_{01}\left[1-P^n(\left\|\hat{\theta}_n-\theta\right\|_\infty<\delta_{01})\right. \\ &\quad \left.+\,P'^n(\left\|\hat{\theta}_n-\theta'\right\|_\infty\geq\delta_{01})\right] \\ &\geq \delta_{01}\left[1-P^n(\left\|\hat{\theta}_N-\theta'\right\|_\infty\geq\delta_{01})\right. \\ &\quad \left.+P'^n(\left\|\hat{\theta}_N-\theta'\right\|_\infty\geq\delta_{01})\right] \\ &\geq \delta_{01}\left[1-\left\|P^n-P'^n\right\|_{\mathrm{TV}}\right] \\ &\geq \delta_{01}\left[1-\sqrt{2}d_{\mathrm{hel}}(P^n,P'^n)^2\right].\end{aligned}$$

The tensorization property of Hellinger distance (cf. Section 15.1 in [Wai19b]) guarantees that

$$d_{\mathrm{hel}}(P^N,P'^N)^2 = 1-\left(1-d_{\mathrm{hel}}(P,P')^2\right)^N \leq N\, d_{\mathrm{hel}}(P,P')^2.$$

Thus, we have proved that

$$\inf_{\hat{\theta}_N}\max_{\mathcal{Q}\in\{\mathcal{P},\mathcal{P}'\}}\mathbb{E}_{\mathcal{Q}}\left[\left\|\theta-\theta(\mathcal{Q})\right\|_\infty\right] \geq \frac{1}{4}\left\|\theta(\mathcal{P})-\theta(\mathcal{P}')\right\|_\infty\cdot\left(1-\sqrt{2}N\cdot d_{\mathrm{hel}}(P,P')^2\right)_+.$$

Taking the supremum over all the possible alternatives $\mathcal{P}'\in\mathcal{S}$ yields

$$\mathfrak{M}_N(\mathcal{P};\mathcal{S}) \geq \sup_{\mathcal{P}'\in\mathcal{S}}\frac{1}{4}\cdot\sqrt{N}\left\|\theta(\mathcal{P})-\theta(\mathcal{P}')\right\|_\infty\cdot\left(1-\sqrt{2}N\cdot d_{\mathrm{hel}}(P,P')^2\right)_+. \quad (2.46)$$

A calculation shows that this bound implies the claim in Lemma 2.

Proof of Lemma 3

Recall the shorthand $\Delta_P = P - P'$ and $\Delta_r = r - r'$, and let $\theta^* \equiv \theta(\mathcal{P})$. We prove that $\|\theta(\mathcal{P}) - \theta(\mathcal{P}')\|_\infty$ is lower bounded by

$$\left\|\gamma(I_d - \gamma P)^{-1}\Delta_P\theta^* + (I_d - \gamma P)^{-1}\Delta_r\right\|_\infty - \left(\frac{\gamma\,\|\Delta_P\|_\infty}{(1-\gamma)}\left\|\gamma(I_d - \gamma P)^{-1}\Delta_P\theta^*\right\|_\infty + \frac{\gamma\,\|\Delta_P\|_\infty\,\|\Delta_r\|_\infty}{(1-\gamma)^2}\right). \quad (2.47)$$

Since $\theta(\mathcal{P}) = (I_d - \gamma P)^{-1}r$ and $\theta(\mathcal{P}') = (I_d - \gamma P')^{-1}r'$ by definition, if we introduce the shorthand $M_P = (I_d - \gamma P)^{-1} - (I_d - \gamma P')^{-1}$, some elementary calculation gives the identity

$$\theta(\mathcal{P}) - \theta(\mathcal{P}') = M_P r + (I_d - \gamma P)^{-1}\Delta_r + M_P\Delta_r. \quad (2.48)$$

Now we find a new expression for $M_P = (I_d - \gamma P)^{-1} - (I_d - \gamma P')^{-1}$ that is easy to control. Recall the elementary identity $A_1^{-1} = A_0^{-1} + A_1^{-1}(A_0 - A_1)A_0^{-1}$ for any matrices A_0, A_1. Thus,

$$\begin{aligned}
M_P &= (I_d - \gamma P)^{-1} - (I_d - \gamma P')^{-1} \\
&= \gamma(I_d - \gamma P')^{-1}(P - P')(I_d - \gamma P)^{-1} \\
&= \gamma(I_d - \gamma P)^{-1}(P - P')(I_d - \gamma P)^{-1} \\
&\quad + \gamma^2(I_d - \gamma P')^{-1}(P - P')(I_d - \gamma P)^{-1}(P - P')(I_d - \gamma P)^{-1} \\
&= \gamma(I_d - \gamma P)^{-1}\Delta_P(I_d - \gamma P)^{-1} \\
&\quad + \gamma^2(I_d - \gamma P')^{-1}\Delta_P(I_d - \gamma P)^{-1}\Delta_P(I_d - \gamma P)^{-1}.
\end{aligned}$$

Substituting this identity into Eq. (2.48), we obtain

$$\theta(\mathcal{P}) - \theta(\mathcal{P}') = \gamma(I_d - \gamma P)^{-1}\Delta_P\theta^* + (I_d - \gamma P)^{-1}\Delta_r + \mathcal{R}_{01}, \quad (2.49)$$

where the remainder term $\mathcal{R}_{01}$ takes the form

$$\mathcal{R}_{01} = \gamma^2(I_d - \gamma P')^{-1}\Delta_P(I_d - \gamma P)^{-1}\Delta_P\theta^* + M_P\Delta_r.$$

Since $(1-\gamma)(I_d - \gamma P')^{-1}$ is a probability transition matrix, it follows that $\|(1-\gamma)(I_d - \gamma P')^{-1}\|_\infty \leq 1$. Thus, the remainder term $\mathcal{R}_{01}$ satisfies the bound

$$\|\mathcal{R}_{01}\|_\infty \leq \frac{\gamma}{(1-\gamma)}\|\Delta_P\|_\infty\left\|\gamma(I_d - \gamma P)^{-1}\Delta_P\theta^*\right\|_\infty + \frac{\gamma}{(1-\gamma)^2}\|\Delta_P\|_\infty\|\Delta_r\|_\infty.$$

The claimed lower bound (2.47) now follows from Eq. (2.49) and the triangle inequality. It is clear that Eq. (2.47) implies the claim in the lemma statement once we restrict $\mathcal{P}' \in \mathcal{S}_1$ and $\mathcal{P}' \in \mathcal{S}_2$.

Proof of Lemma 4

Throughout the proof, we use $(\mathbf{Z}, R)$ (respectively $(\mathbf{Z}', R')$) to denote a sample drawn from the distribution P (respectively from the distribution P'). We use $P_{\mathbf{Z}}, P_R$ (respectively $P'_{\mathbf{Z}}, P'_R$) to denote the marginal distribution of $\mathbf{Z}, R$ (respectively $\mathbf{Z}'$, R'). By the independence of $\mathbf{Z}$ and R (and similarly for $(\mathbf{Z}', R')$, the joint distributions have the product form

$$P = P_{\mathbf{Z}} \otimes P_R, \quad \text{and} \quad P' = P'_{\mathbf{Z}} \otimes P'_R. \tag{2.50}$$

Proof of part (a): Let $\mathcal{P}' = (P', R') \in \mathcal{S}_1$ (so $r' = r$). Because of the independence between $\mathbf{Z}$ and R (see Eq. (2.50)) and $r = r'$, we have that

$$d_{\text{hel}}(P, P')^2 = d_{\text{hel}}(P_{\mathbf{Z}}, P_{\mathbf{Z}'})^2.$$

Note that the rows of $\mathbf{Z}$ and $\mathbf{Z}'$ are independent. Thus, if we let $\mathbf{Z}_i, \mathbf{Z}'_i$ denote the i-th rows of $\mathbf{Z}$ and $\mathbf{Z}'$, we have

$$d_{\text{hel}}(P_{\mathbf{Z}}, P_{\mathbf{Z}'})^2 = 1 - \prod_i \left(1 - d_{\text{hel}}(P_{\mathbf{Z}_i}, P_{\mathbf{Z}'_i})^2\right) \leq \sum_i d_{\text{hel}}(P_{\mathbf{Z}_i}, P_{\mathbf{Z}'_i})^2.$$

Now, note that $\mathbf{Z}_i$ and $\mathbf{Z}'_i$ have multinomial distribution with parameters P_i and P'_i, where $P_{0,i}, P_{1,i}$ are the ith row of P_0 and P_1. Thus, we have

$$d_{\text{hel}}(P_{\mathbf{Z}_i}, P_{\mathbf{Z}'_i})^2 \leq \frac{1}{2} D_{\chi^2}\left(P_{\mathbf{Z}'_i} \| P_{\mathbf{Z}_i}\right) = \frac{1}{2} \sum_j \frac{\left(P_{i,j} - P'_{i,j}\right)^2}{P_{i,j}}.$$

Putting together the pieces yields the desired upper bound (2.30a).

Proof of part (b): Let $\mathcal{P}' = (P', R') \in \mathcal{S}_2$ (so $P' = P$). Given the independence between $\mathbf{Z}$ and R (see Eq. (2.50)) and $P = P'$, we have the relation $d_{\text{hel}}(P, P')^2 = d_{\text{hel}}(P_R, P_{R'})^2$. Note that $R \sim \mathcal{N}(r, I_d)$ and $R' \sim \mathcal{N}(r', I_d)$. Thus, we have

$$d_{\text{hel}}(P_R, P_{R'})^2 \leq D_{\text{kl}}\left(P_R \| P_{R'}\right) = \frac{1}{2\sigma_r^2} \|r - r'\|_2^2,$$

as claimed.

Proof of Lemma 5

We now specify how to construct the probability matrix $\bar{P}$ that satisfies the desired properties stated in Lemma 5. We introduce the shorthand notation $\bar{\theta} = P\theta^*$, and $U = (I_d - \gamma P)^{-1}$. Let $\bar{\ell} \in [D]$ be an index such that

$$\bar{\ell} \in \operatorname*{argmax}_{\ell \in [D]} \left(e_\ell^\top (I_d - \gamma P)^{-1} \Sigma(\theta) (I_d - \gamma P)^{-\top} e_\ell\right)^{1/2} = \operatorname*{argmax}_{\ell \in [D]} \left(\sum_i U_{\ell,i}^2 \sigma_i^2(\theta^*)\right)^{1/2}$$

We construct the matrix $\bar{P}$ entrywise as follows:

$$\bar{P}_{i,j} = P_{i,j} + \frac{1}{\nu\sqrt{2N}} \cdot P_{i,j} U_{\bar{\ell},i}(\theta^*_j - \bar{\theta}_i)$$

for $\nu \equiv \nu(P, \theta^*) = \left(\sum_i U^2_{\bar{\ell},i} \sigma^2(\theta_i) \right)^{1/2}$. Now we show that $\bar{P}$ satisfy the following properties:

(P1) The matrix $\bar{P}$ is a probability transition matrix.

(P2) It satisfies the constraint $\sum_{i,j} \frac{\left((P-\bar{P})_{i,j}\right)^2}{P_{i,j}} \leq \frac{1}{2N}$.

(P3) It satisfies the inequalities

$$\left\|P - \bar{P}\right\|_\infty \leq \frac{1}{\sqrt{2N}}, \quad \text{and} \quad \left\|\gamma(I_d - \gamma P)^{-1}(P - \bar{P})\theta^*\right\|_\infty \geq \frac{\gamma}{\sqrt{2N}} \cdot \nu(P, \theta^*). \tag{2.51}$$

We prove each of these properties in turn.

Proof of (P1): For each row $i \in [D]$, we have

$$\sum_j \bar{P}_{i,j} = \sum_j P_{i,j} + \frac{1}{\nu\sqrt{2N}} U_{\bar{\ell},i} \sum_j P_{i,j}(\theta^*_j - \bar{\theta}_i) = \sum_j P_{i,j} = 1, \tag{2.52}$$

thus showing that $\bar{P}\mathbf{1} = \mathbf{1}$ as desired. Moreover, since $(1-\gamma)U = (1-\gamma)(I_d - \gamma P)^{-1}$ is a probability transition matrix, we have the bound $|U_{\bar{\ell},i}| \leq \frac{1}{1-\gamma}$. By the triangle inequality, we have

$$2\|\theta^*\|_{\text{span}} \geq |\theta^*_j - \bar{\theta}_i|.$$

Thus, our assumption on the sample size N implies that $\nu\sqrt{N} \geq \frac{2}{1-\gamma}\|\theta^*\|_{\text{span}} \geq |U_{\bar{l},i}(\theta^*_j - \bar{\theta}_j)|$, which further implies that

$$\bar{P}_{i,j} = P_{i,j}\left(1 + \frac{1}{\nu\sqrt{2N}} \cdot U_{\bar{\ell},i}(\theta^*_j - \bar{\theta}_i)\right) \geq 0.$$

In conjunction with the property $\bar{P}\mathbf{1} = \mathbf{1}$, we conclude that $\bar{P}$ is a probability transition matrix, as claimed.

Proof of (P2): We begin by observing that $(\Delta_P)_{i,j} = \frac{1}{\nu\sqrt{2N}} \cdot P_{i,j} U_{\bar{\ell},i}(\theta^*_j - \bar{\theta}_i)$. Now it is simple to check that

$$\sum_{i,j} \frac{\left((\Delta_P)_{i,j}\right)^2}{P_{i,j}} = \frac{1}{2N\nu^2} \sum_{i,j} P_{i,j} U^2_{\bar{\ell},i}(\theta^*_j - \bar{\theta}_i)^2 \overset{(i)}{=} \frac{1}{2N\nu^2} \sum_i U^2_{\bar{\ell},i} \sigma^2_i(\theta^*) = \frac{1}{2N}, \tag{2.53}$$

where in step (i), we use $\sigma^2_i(\theta^*) = \sum_j P_{i,j}(\theta^*_j - \bar{\theta}_i)^2$ for each i, as the ith row of our observation $\mathbf{Z}$ is a multinomial distribution with mean specified by the ith row of P. This proves that $\bar{P}$ satisfies the constraint, as desired.

Proof of (P3): In order to verify the first inequality, we note that for any row i,

$$\sum_j |(\Delta_P)_{i,j}| \overset{(i)}{\leq} \left(\sum_j \frac{(\Delta_P)_{i,j}^2}{P_{i,j}} \right)^{1/2} \leq \left(\sum_{i,j} \frac{(\Delta_P)_{i,j}^2}{P_{i,j}} \right)^{1/2} \overset{(ii)}{=} \frac{1}{\sqrt{2N}},$$

where step (i) follows from the Cauchy-Schwartz inequality, and step (ii) follows by the previously established Property 2. Taking the maximum over row i yields

$$\|\Delta_P\|_\infty = \max_i \left\{ \sum_j |(\Delta_P)_{i,j}| \right\} \leq \frac{1}{\sqrt{2N}},$$

thus establishing the first claimed inequality in Eq. (2.51).

In order to establish the second inequality in Eq. (2.51), our starting point is the lower bound

$$\left\| \gamma(I_d - \gamma P)^{-1} \Delta_P \theta^* \right\|_\infty \geq \left| e_{\bar{\ell}}^T \gamma (I_d - \gamma P)^{-1} \Delta_P \theta^* \right| = \gamma \cdot \left| \sum_{i,j} U_{\bar{\ell},i} (\Delta_P)_{i,j} \theta_j^* \right|.$$

It is straightforward to check that

$$\sum_{i,j} U_{\bar{\ell},i} (\Delta_P)_{i,j} \theta_j^* \overset{(i)}{=} \sum_{i,j} U_{\bar{\ell},i} (\Delta_P)_{i,j} (\theta_j^* - \bar{\theta}_i) = \frac{1}{\nu\sqrt{2N}} \sum_{i,j} P_{i,j} U_{\bar{\ell},i}^2 (\theta_j^* - \bar{\theta}_i)^2 \overset{(ii)}{=} \frac{\nu}{\sqrt{2N}}.$$

Here step (i) follows from the fact that $\sum_j (\Delta_P)_{i,j} = 0$ for all i (as $\Delta_P \mathbf{1} = \bar{P}\mathbf{1} - P\mathbf{1} = 0$); whereas step (ii) follows from our previous calculation (see Eq. (2.53)) showing that

$$\sum_{i,j} P_{i,j} U_{\bar{\ell},i}^2 (\theta_j^* - \bar{\theta}_i)^2 = \nu^2.$$

Thus, we have verified the second inequality in Eq. (2.51).

2.8 Proofs of auxiliary lemmas for Theorem 2

This appendix is devoted to the proofs of auxiliary lemmas involved in the proof of Theorem 2.

Proof of Lemma 2:

In this section, we prove all three parts of Lemma 2, which provides high-probability upper bounds on the suboptimality gap at the end of each epoch. Parts (a), (b) and (c), respectively, of Lemma 2 provides guarantees for the recentered linear stepsize, polynomially-decaying stepsizes and constant stepsize. In order to de-clutter the notation, we omit the dependence on the epoch m in the operators and epoch initialization $\bar{\theta}_m$. In order to distinguish between the total sample size N and the recentering sample size at epoch m, we retain the notation N_m for the recentering sample size.

Proof of part (a)

We begin by rewriting the update Eq, (2.35b) in a form suitable for application of general results from [Wai19c]. Subtracting off the fixed point $\widehat{\theta}$ of the operator $\mathcal{J}$, we find that

$$\theta_{k+1} - \widehat{\theta} = (1 - \alpha_k)\left(\theta_k - \widehat{\theta}\right) + \alpha_k\left\{\mathcal{J}_k(\theta_k) - \widehat{\theta}\right\}.$$

Note that the operator $\theta \mapsto \widehat{\mathcal{J}_k}(\theta)$ is γ-contractive in the ℓ_∞-norm and monotonic with respect to the orthant ordering; consequently, Corollary 1 from [Wai19c] can be applied. In applying this corollary, the effective noise term is given by

$$W_k := \mathcal{J}_k(\widehat{\theta}) - \mathcal{J}(\widehat{\theta}) \;=\; \left\{\widehat{\mathcal{T}}_k(\widehat{\theta}) - \widehat{\mathcal{T}}_k(\bar{\theta})\right\} - \left\{\mathcal{T}(\widehat{\theta}) - \mathcal{T}(\bar{\theta})\right\}.$$

With this setup, by adapting Corollary 1 from [Wai19c] we have

$$\left\|\theta_{K+1} - \widehat{\theta}\right\|_\infty \leq \frac{2}{1 + (1-\gamma)K}\left\{\|\bar{\theta} - \widehat{\theta}\|_\infty + \sum_{k=1}^{K}\|V_t\|_\infty\right\} + \|V_{K+1}\|_\infty, \tag{2.54a}$$

where the auxiliary stochastic process $\{V_k\}_{k\geq 1}$ evolves according to the recursion

$$V_{k+1} = \left(1 - \alpha_k\right)V_k + \alpha_k W_k. \tag{2.54b}$$

We claim that the ℓ_∞-norm of this process can be bounded with high probability as follows:

Lemma 4. *Consider any sequence of stepsizes $\{\alpha_k\}_{k\geq 1}$ in $(0,1)$ such that*

$$(1 - \alpha_{k+1})\alpha_k \leq \alpha_{k+1}. \tag{2.55}$$

Then for any tolerance level $\delta > 0$, we have

$$\mathbb{P}\left[\|V_{t+1}\|_\infty \geq 4\|\widehat{\theta} - \bar{\theta}\|_\infty\sqrt{\alpha_t}\sqrt{\log(8KMD/\delta)}\right] \leq \frac{\delta}{2KM}. \tag{2.56}$$

See Section 2.8 for a proof of this claim. For future reference, note that all three stepsize choices (2.5a)–(2.5c) satisfy the condition (2.55).

Substituting the bound (2.56) into the relation (2.54a) yields

$$\begin{aligned}\left\|\theta_{K+1} - \widehat{\theta}\right\|_\infty &\leq c\left\{\frac{\|\bar{\theta} - \widehat{\theta}\|_\infty}{1 + (1-\gamma)K} + \frac{\|\bar{\theta} - \widehat{\theta}\|_\infty}{(1-\gamma)^{3/2}\sqrt{K}}\right\}\sqrt{\log(8KMD/\delta)} \\ &\leq c\|\bar{\theta} - \widehat{\theta}\|_\infty\left\{\frac{\sqrt{\log(8KMD/\delta)}}{1 + (1-\gamma)K} + \frac{\sqrt{\log(8KMD/\delta)}}{(1-\gamma)^{3/2}\sqrt{K}}\right\},\end{aligned}$$

with probability at least $1 - \frac{\delta}{2M}$. Combining the last bound with the fact that $KM = \frac{N}{2}$ we find that for all $K \geq c_1\frac{\log(8ND/\delta)}{(1-\gamma)^3}$, we have

$$\|\theta_{K+1} - \widehat{\theta}\|_\infty \leq \tfrac{1}{8}\|\bar{\theta} - \widehat{\theta}\|_\infty \leq \tfrac{1}{8}\|\bar{\theta} - \theta^*\|_\infty + \tfrac{1}{8}\|\widehat{\theta} - \theta^*\|_\infty,$$

which completes the proof of part (a).

Proof of part (b):

The proof of part (b) is similar to that of part (a). In particular, adapting Corollary 2 from the paper [Wai19c] for polynomial steps, we have

$$\|\theta_{k+1} - \widehat{\theta}\|_\infty \leq e^{-\frac{1-\gamma}{1-\omega}\left(k^{1-\omega}-1\right)}\|\bar{\theta} - \widehat{\theta}\|_\infty + e^{-\frac{1-\gamma}{1-\omega}k^{1-\omega}} \sum_{\ell=1}^{k} \frac{e^{\frac{1-\gamma}{1-\omega}\ell^{1-\omega}}}{\ell^\omega}\|V_\ell\|_\infty + \|V_{k+1}\|_\infty. \tag{2.57}$$

Recall that polynomial stepsize (2.5b) satisfies the conditions of Lemma 4. Consequently, applying the bound from Lemma 4 we find that

$$\|\theta_{k+1} - \widehat{\theta}\|_\infty \leq \|\bar{\theta} - \widehat{\theta}\|_\infty \left\{ e^{-\frac{1-\gamma}{1-\omega}\left(k^{1-\omega}-1\right)} \right.$$
$$\left. +4\sqrt{\log(8KMD/\delta)} \left(e^{-\frac{1-\gamma}{1-\omega}k^{1-\omega}} \sum_{\ell=1}^{k} \frac{e^{\frac{1-\gamma}{1-\omega}\ell^{1-\omega}}}{\ell^{3\omega/2}} + \frac{1}{k^{\omega/2}} \right) \right\}.$$

It remains to bound the coefficient of $\|\bar{\theta} - \widehat{\theta}\|_\infty$ in the last equation, and we do so by using the following lemma from [Wai19c]:

Lemma 5 (Bounds on exponential-weighted sums)**.** *There is a universal constant c such that for all $\omega \in (0,1)$ and for all $k \geq \left(\frac{3\omega}{2(1-\gamma)}\right)^{\frac{1}{1-\omega}}$, we have*

$$e^{-\frac{1-\gamma}{1-\omega}k^{1-\omega}} \sum_{\ell=1}^{k} \frac{e^{\frac{1-\gamma}{1-\omega}\ell^{1-\omega}}}{\ell^{3\omega/2}} \leq c \left\{ \frac{e^{-\frac{1-\gamma}{1-\omega}(k^{1-\omega})}}{(1-\gamma)^{\frac{1}{1-\omega}}} + \frac{1}{(1-\gamma)}\frac{1}{k^{\omega/2}} \right\}.$$

Substituting the last bound in Eq. (2.57) yields

$$\left\|\theta_{k+1} - \widehat{\theta}\right\|_\infty \leq c\|\bar{\theta} - \widehat{\theta}\|_\infty \left\{ e^{-\frac{1-\gamma}{1-\omega}\left(k^{1-\omega}-1\right)} \right.$$
$$\left. +4\sqrt{\log(8KMD/\delta)} \left(\frac{e^{-\frac{1-\gamma}{1-\omega}\left(k^{1-\omega}-1\right)}}{(1-\gamma)^{\frac{1}{1-\omega}}} + \frac{1}{(1-\gamma)}\frac{1}{k^{\omega/2}} + \frac{1}{k^{\omega/2}} \right) \right\}$$
$$\leq c\|\bar{\theta} - \widehat{\theta}\|_\infty \cdot \sqrt{\log(8KMD/\delta)} \left\{ 5 \cdot \frac{e^{-\frac{1-\gamma}{1-\omega}\left(k^{1-\omega}-1\right)}}{(1-\gamma)^{\frac{1}{1-\omega}}} + \frac{2}{(1-\gamma)k^{\omega/2}} \right\}.$$

Finally, doing some algebra and using the fact that $KM = \frac{N}{2}$ we find that there is an absolute constant c such that for all K lower bounded as $K \geq c\log(4ND/\delta) \cdot \left(\frac{1}{1-\gamma}\right)^{\frac{1}{1-\omega} \vee \frac{2}{\omega}}$, we have

$$\|\theta_{K+1} - \widehat{\theta}\|_\infty \leq \frac{\|\bar{\theta} - \widehat{\theta}\|_\infty}{8} \leq \frac{1}{8}\|\bar{\theta} - \theta^*\|_\infty + \frac{1}{8}\|\widehat{\theta} - \theta^*\|_\infty.$$

The completes the proof of part (b).

Proof of part (c):

Invoking Theorem 1 from [Wai19c], we have $\|\theta_K - \widehat{\theta}\|_\infty \leq a_K + b_K + \|V_K\|_\infty$. For a constant stepsize $\alpha_k = \alpha$, the pair (a_K, b_K) is given by

$$\begin{aligned} b_K &= \left\|\bar{\theta} - \widehat{\theta}\right\|_\infty \cdot (1 - \alpha(1-\gamma))^{K-1}, \\ a_K &= \gamma\alpha \left\|V_k\right\|_\infty + \gamma\alpha \left\|V_\ell\right\|_\infty \sum_{k=1}^{K-1} \left\{(1 - (1-\gamma)\alpha)^{K-k}\right\} \\ &\overset{(i)}{\leq} \|\bar{\theta} - \widehat{\theta}\|_\infty \cdot \left(2\gamma\alpha^{\frac{3}{2}}\sqrt{\log(8KMD/\delta)} + \frac{2\gamma\alpha^{\frac{1}{2}}}{1-\gamma}\sqrt{\log(8KMD/\delta)}\right), \end{aligned}$$

where inequality (i) follows by substituting $\alpha_k = \alpha$, and using the bound on $\|V_\ell\|_\infty$ from Lemma 4.

It remains to choose the pair (α, K) such that $\|\theta_{K+1} - \widehat{\theta}\|_\infty \leq \frac{1}{8}\|\bar{\theta} - \widehat{\theta}\|_\infty$. Doing some simple algebra and using the fact that $KM = \frac{N}{2}$ we find that it is sufficient to choose the pair (α, K) satisfying the conditions

$$0 < \alpha \leq \frac{(1-\gamma)^2}{\log(4ND/\delta)} \cdot \frac{1}{5^2 \cdot 32^2}, \quad \text{and} \quad K \geq 1 + \frac{2\log 16}{\log\left(\frac{1}{1-\alpha(1-\gamma)}\right)}.$$

With this choice, we have

$$\|\theta_{K+1} - \widehat{\theta}\|_\infty \leq \frac{\|\bar{\theta} - \widehat{\theta}\|_\infty}{8} \leq \frac{1}{8}\left\|\bar{\theta} - \theta^*\right\|_\infty + \frac{1}{8}\left\|\widehat{\theta} - \theta^*\right\|_\infty,$$

which completes the proof of part (c).

Proof of Lemma 3

Recall our shorthand notation for the local complexities (2.8). The following lemma characterizes the behavior of various random variables as a function of these complexities. In stating the lemma, we let $\widehat{\mathbf{P}}_n$ be a sample transition matrix constructed as the average of n i.i.d. samples, and let $\widehat{r}_n$ denote the reward vector constructed as the average of n i.i.d. samples.

Lemma 6. *Each of the following statements holds with probability exceeding* $1 - \frac{\delta}{M}$*:*

$$\|(I_d - \gamma P)^{-1}(\widehat{\mathbf{P}}_n - P)\theta^*\|_\infty \leq 2\nu(P, \theta^*) \cdot \sqrt{\frac{\log(4DM/\delta)}{n}} + 4 \cdot b(\theta^*) \cdot \frac{\log(4DM/\delta)}{n}, \text{ and}$$

$$\|(I_d - \gamma P)^{-1}(\widehat{r}_n - r)\|_\infty \leq 2\rho(P, r) \cdot \sqrt{\frac{\log(4DM/\delta)}{n}}.$$

Proof. Entry ℓ of the vector $(I_d - \gamma P)^{-1}(\widehat{\mathbf{P}}_n - P)\theta^*$ is zero mean with variance given by the ℓ^{th} diagonal entry of the matrix $(I_d - \gamma P)^{-1}\Sigma(\theta^*)(I_d - \gamma P)^{-T}$, and is bounded

by $b(\theta^*)$ almost surely. Consequently, applying the Bernstein bound in conjunction with the union bound completes the proof of the first claim. In order to establish the second claim, note that the vector $(I_d - \gamma P)^{-1}(\widehat{r}_n - r)$ has sub-Gaussian entries, and apply the Hoeffding bound in conjunction with the union bound. $\square$

In light of Lemma 6, note that it suffices to establish the inequality

$$\Pr\Big\{\|\widehat{\theta}_m - \theta^*\|_\infty \geq \frac{\left\|\bar{\theta} - \theta^*\right\|_\infty}{9} + \|(I_d - \gamma P)^{-1}(\widehat{\mathbf{P}}_{N_m} - P)\theta^*\|_\infty + \|(I_d - \gamma P)^{-1}(\widehat{r}_{N_m} - r)\|_\infty\Big\} \leq \frac{\delta}{2M}, \quad (2.58)$$

where we have let $\widehat{\mathbf{P}}_{N_m}$ and $\widehat{r}_{N_m}$ denote the empirical mean of the observed transitions and rewards in epoch m, respectively. The proof of Lemma 3 follows from Eq. (2.58) by a union bound.

Establishing the bound (2.58): Since the epoch number m should be clear from context, let us adopt the shorthand $\widehat{\theta} \equiv \widehat{\theta}_m$, along with the shorthand $\widehat{r} \equiv \widehat{r}_{N_m}$ and $\widehat{\mathbf{P}} \equiv \widehat{\mathbf{P}}_{N_m}$. Note that $\widehat{\theta}$ is the fixed point of the following operator:

$$\mathcal{J}(\theta) := \mathcal{T}(\theta) - \mathcal{T}(\bar{\theta}) + \widetilde{\mathcal{T}}_{N_m}(\bar{\theta}) = \underbrace{\widehat{r} + \gamma\left(\widehat{P} - P\right)\bar{\theta}}_{\widetilde{r}} + \gamma P\theta,$$

where we have used the fact that $\widetilde{T}_{N_m}(\theta) = \widehat{r} + \gamma\widehat{P}\theta$.

Thus, we have $\widehat{\theta} = (I_d - \gamma P)^{-1}\widetilde{r}$, so that $\widehat{\theta} - \theta^* = (I_d - \gamma P)^{-1}(\widetilde{r} - r)$. Also note that we have

$$\widetilde{r} - r = \widehat{r} + \gamma\left(\widehat{P} - P\right)\bar{\theta} - r = \widehat{r} - r + \gamma\left(\widehat{P} - P\right)\theta^* + \gamma\left(\widehat{P} - P\right)\left(\bar{\theta} - \theta^*\right),$$

so that putting together the pieces and using the triangle inequality yields the bound

$$\begin{aligned}\|\widehat{\theta} - \theta^*\|_\infty &\leq \|(I_d - \gamma P)^{-1}(\widehat{r} - r)\|_\infty + \gamma\|(I_d - \gamma P)^{-1}\left(\widehat{P} - P\right)\theta^*\|_\infty \\ &\quad + \gamma\|(I_d - \gamma P)^{-1}\left(\widehat{P} - P\right)\left(\bar{\theta} - \theta^*\right)\|_\infty \\ &\leq \|(I_d - \gamma P)^{-1}(\widehat{r} - r)\|_\infty + \gamma\|(I_d - \gamma P)^{-1}\left(\widehat{P} - P\right)\theta^*\|_\infty + \\ &\frac{\gamma}{1-\gamma}\|\left(\widehat{P} - P\right)\left(\bar{\theta} - \theta^*\right)\|_\infty.\end{aligned}$$

Note that the random vector $\left(\widehat{P} - P\right)\left(\bar{\theta} - \theta^*\right)$ is the empirical average of N_m i.i.d. random vectors, each of which is bounded entrywise by $2\|\bar{\theta} - \theta^*\|_\infty$. Consequently, by a combination of Hoeffding's inequality and the union bound, we find that

$$\left\|\left(\widehat{P} - P\right)\left(\bar{\theta} - \theta^*\right)\right\|_\infty \leq 4\|\bar{\theta} - \theta^*\|_\infty\sqrt{\frac{\log(8DM/\delta)}{N_m}},$$

with probability at least $1 - \frac{\delta}{4M}$. Thus, provided $N_m \geq 4^2 \cdot 9^2 \cdot \frac{\gamma^2}{(1-\gamma)^2} \log(8DM/\delta)$ for a large enough constant c_1, we have

$$\frac{\gamma}{1-\gamma} \left\| \left(\widehat{P} - P\right) \left(\bar{\theta} - \theta^*\right) \right\|_\infty \leq \frac{\|\bar{\theta} - \theta^*\|_\infty}{9}.$$

This completes the proof.

Proof of Lemma 4

Recall that by definition, the stochastic process $\{V_k\}_{k \geq 1}$ evolves according to the linear recursion $V_k = (1 - \alpha_k)V_{k-1} + \alpha_k W_{k-1}$, where the effective noise sequence $\{W_k\}_{k \geq 0}$ satisfies the uniform bound

$$\|W_k\|_\infty \leq \left\| \widehat{\mathcal{T}}_k(\widehat{\theta}) - \widehat{\mathcal{T}}_k(\bar{\theta}) \right\|_\infty + \|\mathcal{T}(\widehat{\theta}) - \mathcal{T}(\bar{\theta})\|_\infty \leq \underbrace{2\|\widehat{\theta} - \bar{\theta}\|_\infty}_{:=b} \quad \text{for all } k \geq 0.$$

Moreover, we have $\mathbb{E}[W_k] = 0$ by construction so that each entry of the random vector W_k is a zero-mean sub-Gaussian random variable with sub-Gaussian parameter at most $2\|\widehat{\theta} - \bar{\theta}\|_\infty$. Consequently, by known properties of sub-Gaussian random variables (cf. Chapter 2 in [Wai19b]), we have

$$\log \mathbb{E}\left[e^{sW_k(x)}\right] \leq \frac{s^2 b^2}{8} \qquad \text{for all scalars } s \in \mathbb{R}, \text{ and states } x. \tag{2.59}$$

We complete the proof by using an inductive argument to upper bound the moment generating function of the random variable V_ℓ; given this inequality, we can then apply the Chernoff bound to obtain the stated tail bounds. Beginning with the bound on the moment generating function, we claim that

$$\log \mathbb{E}\left[e^{sV_k(x)}\right] \leq \frac{s^2 \alpha_k b^2}{8} \qquad \text{for all scalars } s \in \mathbb{R} \text{ and states } x. \tag{2.60}$$

We prove this claim via induction on k.

Base case: For $k = 1$, we have

$$\log \mathbb{E}\left[e^{sV_1(x)}\right] = \log \mathbb{E}\left[e^{s\alpha_1 W_0(x)}\right] \leq \frac{s^2 \alpha_1^2 b^2}{8},$$

where the first equality follows from the definition of V_1, and the second inequality follows by applying the bound (2.59).

Inductive step: We now assume that the bound (2.60) holds for some iteration $k \geq 1$ and prove that it holds for iteration $k + 1$. Recalling the definition of V_k, and

the independence of the random variables V_k and W_k, we have

$$\begin{aligned}
\mathbb{E}\left[e^{sV_{k+1}(x)}\right] &= \log \mathbb{E}\left[e^{s(1-\alpha_k)V_k(x)}\right] + \log \mathbb{E}\left[e^{s\alpha_k W_k(x)}\right] \\
&\leq \tfrac{s^2(1-\alpha_k)^2\alpha_{k-1}b^2}{8} + \tfrac{s^2\alpha_k^2 b^2}{8} \\
&\overset{(i)}{\leq} \tfrac{s^2(1-\alpha_k)\alpha_k b^2}{8} + \tfrac{s^2\alpha_k^2 b^2}{8} \\
&= \tfrac{s^2\alpha_k b^2}{8},
\end{aligned}$$

where inequality (i) follows from the assumed condition (2.55) on the stepsizes.

Simple algebra yields that all the stepsize choices (2.5a)–(2.5c) satisfy the condition (2.55). Finally, combining the bound (2.60) with the Chernoff bounding technique along with a union bound over iterations $k = 1, \ldots K$ yields

$$\mathbb{P}\left[\|V_\ell\|_\infty \geq 2b\sqrt{\alpha_{\ell-1}}\sqrt{\log(8KMD/\delta)}\right] \leq \frac{\delta}{8KM},$$

as claimed.

Chapter 3

Instance-optimality in optimal value estimation

Various algorithms in reinforcement learning exhibit dramatic variability in their convergence rates and ultimate accuracy as a function of the problem structure. Such instance-specific behavior is not captured by existing global minimax bounds, which are worst-case in nature. In this chapter we analyze the problem of estimating optimal Q-value functions for a discounted Markov decision process with discrete states and actions and identify an instance-dependent functional that controls the difficulty of estimation in the ℓ_∞-norm. Using a local minimax framework, we show that this functional arises in lower bounds on the accuracy on any estimation procedure. In the other direction, we establish the sharpness of our lower bounds, up to factors logarithmic in the state and action spaces, by analyzing a variance-reduced version of Q-learning. Our theory provides a precise way of distinguishing "easy" problems from "hard" ones in the context of Q-learning, as illustrated by an ensemble with a continuum of difficulty.

3.1 Introduction

The need for data-driven decision-making has fueled tremendous interest in Markov decision processes and reinforcement learning (RL). Indeed, such techniques have found use cases across a wide range of application domains [Lev+16; Sil+16; Tob+17]. An intriguing fact is that in many applications, RL algorithms behave far better than the theoretical bounds provided by worst-case analyses would suggest. This gap provides impetus for a more refined *instance-specific* analysis, one which highlights the properties of a given instance that render it "easy" or "difficult."

Instance-dependent analysis of RL algorithms has become of substantial interest in recent years [see, e.g., Kha+20a; MMM14; PW20; SJ19; ZB19; ZKB19]. By now, we have a fairly refined understanding of instance-dependence for policy evaluation problems, including work on temporal difference (TD) algorithms under the ℓ_2-

norm [BRS18a; Dal+18; LS18a], as well as bounds for the LSTD estimator in the ℓ_∞-norm [PW20]. A subset of the current authors [Kha+20a] provided a sharper instance-dependent ℓ_∞-bounds for a variance-reduced version of the TD(0) algorithm, and showed that this algorithm is optimal in a local non-asymptotic minimax sense.

For TD and LSTD methods, the underlying structure is linear in nature—in particular, it corresponds to solving a linear system—a property which greatly facilitates the analysis. In the current chapter we undertake a similar instance-dependent analysis in the more challenging setting of Q-learning, for which the underlying updates are non-linear. Our main contributions are to identify a natural functional of the problem instance and show that it controls the fundamental difficulty of estimating optimal Q-value functions. We do so by establishing non-asymptotic lower bounds within a local minimax framework and matching those bounds, up to logarithmic factors, by analyzing a version of variance-reduced Q-learning [Sid+18a; Sid+18b; Wai19d].

This work is done in the context of Markov decision processes (MDPs) with a finite set of states $\mathcal{X}$ and a finite set of possible actions $\mathcal{U}$. We proceed to provide some background and notation to be able to introduce the functional that plays a central role in our analysis, and describe our contributions in more detail.

Some background

In a Markov decision process, the state x evolves dynamically in time under the influence of the actions. More precisely, there is a collection of probability transition kernels, $\{\mathbf{P}_u(\cdot \mid x) \mid (x,u) \in \mathcal{X} \times \mathcal{U}\}$, where $\mathbf{P}_u(x' \mid x)$ denotes the transition to the state x' when the action u is taken at the current state x. In addition, an MDP is equipped with a reward function r that maps every state-action pair, (x,u), to a real number $r(x,u)$. The reward $r(x,u)$ is the reward received upon performing an action u in the state x. Overall, a given MDP is characterized by the problem pair $(\mathbf{P}, r)$, along with a discount factor $\gamma \in (0,1)$.

A deterministic policy π is a mapping $\mathcal{X} \to \mathcal{U}$: the quantity $\pi(x) \in \mathcal{U}$ indicates the action to be taken in the state x. The value of a policy is defined by the expected sum of discounted rewards in an infinite sample path. For a given policy π and discount factor $\gamma \in (0,1)$, the Q-function is given by

$$\theta^\pi(x,u) := \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(x_k, u_k) \mid x_0 = x,\, u_0 = u\right], \qquad \text{where } u_k = \pi(x_k) \text{ for all } k \geq 1. \tag{3.1}$$

When both the state space $\mathcal{X}$ and action space $\mathcal{U}$ are finite, the Q-function $\mathbb{Q}$ can be conveniently represented as an element of $\mathbb{R}^{|\mathcal{X}| \times |\mathcal{U}|}$.

There are various observation models in reinforcement learning, and in this chapter we study the *generative setting* in which we have the ability to draw next-state samples from the MDP when initialized with an arbitrary state-action pair (x,u). More

precisely, we are given a collection of n i.i.d. samples of the form $\{(\mathbf{Z}_k, R_k)\}_{k=1}^n$, where both $\mathbf{Z}_k$ and R_k are random matrices in $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{U}|}$. For each state-action pair (x,u), the entry $\mathbf{Z}_k(x,u)$ is drawn according to the transition kernel $\mathbf{P}_u(\cdot \mid x)$, whereas the entry $R_k(x,u)$ is a zero-mean random variable with mean $r(x,u)$ and σ_r-sub-Gaussian tails, corresponding to a noisy observation of the reward function. Here the rewards $\{R_k(x,u)\}_{(x,u)\in\mathcal{X}\times\mathcal{U}}$ are independent across the all state-action pairs, and the random rewards $\{R_k\}$ are independent of the randomness in $\{\mathbf{Z}_k\}$.

Based on the observations, our goal is to estimate the optimal Q-value function $\theta^\star$, along with an optimal policy $\pi^\star$. From the classical theory of MDPs [Ber09; Put14; SB18b], the optimal Q-function is a fixed point of the Bellman (optimality) operator $\mathbf{T}$, a map from $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{U}|}$ to itself given by

$$\mathbf{T}(\theta)(x,u) := r(x,u) + \gamma \sum_{x'\in\mathcal{X}} \mathbb{P}_u(x' \mid x) \max_{u'\in\mathcal{U}} \theta(x',u'), \tag{3.2}$$

and an optimal policy $\pi^\star$ can be obtained from the optimal Q-function $\theta^\star$ via the maximization $\pi^\star(x) \in \arg\max_{u\in\mathcal{U}} \theta^\star(x,u)$. In this chapter, we measure the quality of a given estimate $\widehat{\theta}$ in terms of the ℓ_∞-norm error:

$$\|\widehat{\theta} - \theta^\star\|_\infty = \max_{(x,u)} |\widehat{\theta}(x,u) - \theta^\star(x,u)|. \tag{3.3}$$

Contents of this chapter

The main contribution of this chapter is to show that for a given MDP, the difficulty of estimating the optimal Q-value function in ℓ_∞-norm is characterized by a particular functional of the problem instance $(\mathbf{P}, r)$, defined here.

An instance-dependent functional: Given a sample $(\mathbf{Z}, R)$ from our observation model, we can define the single-sample empirical Bellman operator

$$\widehat{\mathbf{T}}(\theta) := R(x,u) + \gamma \sum_{x'\in\mathcal{X}} \mathbf{Z}_u(x' \mid x) \max_{u'\in\mathcal{U}} \theta(x',u'), \tag{3.4}$$

where we have introduced $\mathbf{Z}_u(x' \mid x) := \mathbb{1}_{\mathbf{Z}(x,u)=x'}$.

Note that for any fixed Q-function θ, the difference $\widehat{\mathbf{T}}(\theta) - \mathbf{T}(\theta)$ is a zero-mean random matrix, and a key object in this chapter is the matrix $\nu \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{U}|}$ with entries

$$\nu(\pi; \mathbf{P}, r, \gamma)(x,u) := \sqrt{\operatorname{var}\Big(\big(\mathcal{I} - \gamma \mathbf{P}^\pi\big)^{-1}\big(\widehat{\mathbf{T}}(\theta^\star) - \mathbf{T}(\theta^\star)\big)\Big)}. \tag{3.5}$$

More explicitly, the quantity $\mathbf{P}^\pi$ is a right-linear mapping of $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{U}|}$ to itself, given by:

$$\mathbf{P}^\pi \mathbb{Q}(x,u) := \sum_{x'\in\mathcal{X}} \mathbf{P}_u(x' \mid x) \cdot \mathbb{Q}(x', \pi(x')) \qquad \text{for each } (x,u) \in \mathcal{X}\times\mathcal{U}, \tag{3.6}$$

and the square-root and variance operators in equation (3.5) are applied elementwise.

Let us provide some intuition as to why $\nu(\pi; \mathbf{P}, r, \gamma)$ plays a fundamental role. The appearance of the zero mean term $\widehat{\mathbf{T}}(\theta^\star) - \mathbf{T}(\theta^\star)$ is natural: it reflects the noise present in the empirical Bellman operator (3.4) as an estimate of the population Bellman operator (3.2). As for the pre-factor $(\mathcal{I} - \gamma \mathbf{P}^\pi)^{-1}$, by a von Neumann expansion we can write

$$(\mathcal{I} - \gamma \mathbf{P}^\pi)^{-1} = \sum_{k=0}^{\infty} (\gamma \mathbf{P}^\pi)^k.$$

The sum of the powers of $\gamma \mathbf{P}^\pi$ account for the compounded effect of an initial perturbation when following the Markov chain specified by the policy π.

Upper and lower bounds: With these definitions in place, the core of our work involves proving, via a combination of a lower and an upper bound (matching up to logarithmic factors), that the instance-specific difficulty of estimating the Q-function is captured by the quantity $\max_{\pi \in \mathbf{\Pi}^\star} \|\nu(\pi; \mathbf{P}, r, \gamma)\|_\infty$. Here $\mathbf{\Pi}^\star$ denotes the set of all optimal policies for the MDP instance $(\mathbf{P}, r)$. This functional exhibits a wide range of behaviors: in Example 1 to follow in Section 3.2, we exhibit a family of MDPs $(\mathbf{P}_\lambda, r_\lambda)$, parameterized by a scalar $\lambda \geq 0$ such that

$$\max_{\pi \in \mathbf{\Pi}^\star} \|\nu(\pi; \mathbf{P}_\lambda, r_\lambda, \gamma)\|_\infty \asymp \left(\frac{1}{1-\gamma}\right)^{\frac{1}{2} - \lambda}.$$

The setting $\lambda = 0$ recovers a "hard" instance, one for which the global minimax bound for estimation of Q-functions, known from past work [AMK13] on batched Q-learning, is sharp. On the other hand, as λ grows, the problems in this family become progressively easier, so that the global minimax bound is no longer sharp.

In more detail, we prove a non-asymptotic lower bound, stated as Theorem 4 to follow, by adapting a particular definition of local minimax risk studied in past work on shape-constrained estimation [CL15a]. The central challenge in this proof is that perturbations to the transition matrices of a given MDP change not only the transitions themselves, but also the structure of the optimal policies. In order to prove matching upper bounds, given the role of the empirical operators $\widehat{\mathbf{T}}$ in our lower bound, which are used in the classical Q-learning algorithm [JJS94b; Sze97; Tsi94; WD92a], a natural thought would be to analyze this operator directly. However, it is known from past work [Wai19c] that the classical Q-learning algorithm is *non-optimal*, even when assessed when using the coarser metric of global minimax. Thus, in order to obtain a sharp upper bound, we turn to variance-reduced forms of Q-learning, as introduced in past work [Sid+18a; Sid+18b; Wai19d] and shown to be optimal in a globally minimax sense. Our main contribution is to show that under certain structural conditions and lower bounds on the sample size, there is a form of variance-reduced Q-learning that achieves our local minimax lower bound up to a logarithmic

factor. These upper bounds, stated precisely in Theorem 5, confirm that our lower bound technique has extracted a useful form of instance dependence for estimating optimal Q-functions.

Notation: For a positive integer n, we use the shorthand $[n] := \{1, 2, \ldots n\}$. For a finite set S, we use $|S|$ to denote its cardinality. We use $c_1, c_2, \ldots$ to denote universal constants that may change from line to line. For any pair of vectors or matrices (v, w) with matching dimension(s), we write that $v \succeq w$ to imply $v - w$ has only positive entries, and $v \preceq w$ is defined similarly. We let $|u|$ denote the entrywise absolute value of a vector $u \in \mathbb{R}^n$ or a matrix $u \in \mathbb{R}^{m \times n}$; we use $|u|_+$ to denote the entry-wise positive part of u. For any vector or matrix u, we let $\|u\|_\infty$ denote the maximum absolute value taken over all entries of u, and $\|u\|_{\text{span}} = \max_j u_j - \min_j u_j$ denote the span seminorm. For a continuous operator $P : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$, we define its ℓ_∞-operator norm as $|\!|\!| P |\!|\!|_{\infty \to \infty} = \sup_{\|u\|_\infty = 1} \|Pu\|_\infty$. We often identify a Q-value function $\mathbb{Q}$ with its matrix representation and use $\|\mathbb{Q}\|_\infty$ to denote the infinity norm (i.e., largest entry in absolute terms). In the matrix representation of $\mathbb{Q}$, its rows and columns are indexed via an enumeration of the states and actions, respectively. We use the symbol $\gtrsim$ to denote a relation that holds up to logarithmic factors in the problem parameters.

3.2 Main results

We proceed to provide precise statements of the main results of this chapter, along with a discussion of some of their consequences. In Section 7.3, we define a notion of a local non-asymptotic minimax risk, and then state Theorem 4, which provides such a lower bound for estimating optimal Q-value functions. In Section 3.2, we turn to the complementary problem of deriving achievable results. Theorem 5 shows that under certain structural conditions on the policies, there is a form of variance-reduced Q-learning that achieves the local minimax risk up to logarithmic factors.

Instance-dependent lower bounds

In this section, we state a non-asymptotic lower bound for estimating optimal Q-function in the ℓ_∞-norm. This lower bound, to be stated in Theorem 4, is instance-dependent, meaning that it depends on the particular instance of the MDP $(\mathbf{P}, r)$ at hand. This dependence should be contrasted with classical global minimax bounds, which are oblivious to such local properties.

The starting point of our lower bound development is the two-point framework introduced by Cai and Low [CL15a] for local minimax bounds for nonparametric shape-constrained inference; here we adapt it to our current setting. Focusing on the ℓ_∞-norm error metric, the *local non-asymptotic minimax risk* for $\theta(\mathcal{P})$ at an instance

$\mathcal{P} = (\mathbf{P}, r)$ is defined as

$$\mathfrak{M}_N(\mathcal{P}) := \sup_{\mathcal{P}'} \inf_{\hat{\theta}_N} \max_{\mathcal{Q} \in \{\mathcal{P}, \mathcal{P}'\}} \sqrt{N} \cdot \mathbb{E}_{\mathcal{Q}} \left[\|\hat{\theta}_N - \theta(\mathcal{Q})\|_\infty \right]. \tag{3.7}$$

Here the infimum is taken over all estimators $\hat{\theta}_N$ that are measurable functions of the N i.i.d. observations drawn according to our observation model (see Section 3.1).

The intuition underlying the definition (3.7) is that given an instance $\mathcal{P}$, the adversary that defines the instance-dependent non-asymptotic risk $\mathfrak{M}_N(\mathcal{P})$ behaves as follows: it extracts the hardest alternative $\mathcal{P}'$ relative to $\mathcal{P}$, and then measures the worst-case risk over $\mathcal{P}$ and this alternative $\mathcal{P}'$.

Lower bounds for Q-function estimation

We now turn to the statement of some lower bounds for estimating the optimal Q-function. Recall the definition (3.6) of the operator $\mathbf{P}^\pi$, along with the functional $\nu(\pi; \mathbf{P}, r, \gamma)$ from equation (3.5). We let $\nu^2(\pi; \mathbf{P}, r, \gamma)$ denote the matrix obtained by taking squares entrywise. Our first step is to provide a decomposition of this matrix into two separate components, corresponding to the noisiness in the reward function observation and transition matrix observations, respectively.

In order to deal with the latter source of noise, with a slight abuse of notation, we use the observed matrix $\mathbf{Z}$ to define a stochastic analog of $\mathbf{P}^\pi$—namely, the (random) right-linear operator

$$(\mathbf{Z}^\pi \theta)(x, u) := \sum_{x' \in \mathcal{X}} \mathbf{Z}_u(x' \mid x) \cdot \theta(x', \pi(x')), \quad \text{where} \quad \mathbf{Z}_u(x' \mid x) := \mathbb{1}_{\mathbf{Z}(x,u) = x'}. \tag{3.8}$$

By assumption, the randomness in our observations of the reward and transitions are independent, so that for any optimal[1] policy π, we have the decomposition

$$\nu^2(\pi; \mathbf{P}, r)(x, u) = \gamma^2 \rho^2(\pi; \mathbf{P}, r)(x, u) + \sigma^2(\pi; \mathbf{P}, r)(x, u). \tag{3.9a}$$

Here we define

$$\rho^2(\pi; \mathbf{P}, r) := \operatorname{var}\left((\mathcal{I} - \gamma \mathbf{P}^\pi)^{-1} (\mathbf{Z}^\pi - \mathbf{P}^\pi) \theta^\star \right), \qquad \text{and} \tag{3.9b}$$

$$\sigma^2(\pi; \mathbf{P}, r) := \operatorname{var}\left((\mathcal{I} - \gamma \mathbf{P}^\pi)^{-1} (R - r) \right), \tag{3.9c}$$

where we compute the variances in an elementwise sense.

With this notation, we have the following guarantee:

[1] Optimality of π is required so that $\mathbf{T}(\theta^\star) = r + \gamma \mathbf{P}^\pi \theta^\star$, with a similar relation for the empirical Bellman operator.

Theorem 4. *There exists a universal constant $c > 0$ such that for any instance $\mathcal{P} = (\mathbf{P}, r)$, the local non-asymptotic minimax risk is lower bounded as*

$$\mathfrak{M}_N(\mathcal{P}) \geq c \max_{\pi \in \mathbf{\Pi}^\star} \|\nu(\pi; \mathbf{P}, r, \gamma)\|_\infty. \tag{3.10a}$$

This bound is valid for all sample sizes N that satisfy the lower bound

$$N \geq N_0 := \max \left\{ \frac{2\gamma^2}{(1-\gamma)^2}, \frac{2\|\theta^\star\|^2_{\text{span}}}{(1-\gamma)^2 \|\rho^2(\pi^\star; \mathbf{P}, r)\|_\infty} \right\}, \tag{3.10b}$$

where $\pi^\star \in \arg\max_{\pi \in \mathbf{\Pi}^\star} \|\nu(\pi; \mathbf{P}, r)\|_\infty$.

We prove this theorem in Section 3.3. The main take-away is that the functional $\max_{\pi \in \mathbf{\Pi}^\star} \|\nu(\pi; \mathbf{P}, r, \gamma)\|_\infty$ controls the local minimax risk. In order to gain intuition for this claim, it is worth exploring the range of possible behaviors exhibited by this functional.

Exploring the range of possible behaviors

One point of comparison is between the instance-dependent lower bound from Theorem 5 with the existing minimax lower bounds for Q-learning. Azar et al. [AMK13] provided a global minimax lower bound on the ℓ_∞-norm error for estimating the optimal Q-function. For a γ-discounted MDP, they showed that the ℓ_∞-error of any procedure is lower bounded by the quantity $\frac{1}{(1-\gamma)^{1.5}} \cdot \frac{1}{\sqrt{n}}$, up to logarithmic factors in dimension.

This lower bound is optimal in a globally minimax sense, and it is worthwhile understanding the properties of instances that exhibit this worst-case behavior—that is, instances for which $\max_{\pi \in \mathbf{\Pi}^\star} \|\nu(\pi; \mathbf{P}, r, \gamma)\|_\infty \asymp \frac{1}{(1-\gamma)^{1.5}}$. It is also worthwhile understanding the properties of problems that are much "easier" than this worst-case theory would suggest. The following construction, which takes inspiration from [Kha+20a; PW20], allows us to explore this continuum.

Example 1. A continuum of local minimax risks Consider an MDP with two states $\{x_1, x_2\}$, two actions $\{u_1, u_2\}$, and with transition functions and reward functions given by

$$\mathbf{P}_{u_1} = \begin{bmatrix} p & 1-p \\ 0 & 1 \end{bmatrix} \quad \mathbf{P}_{u_2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{and} \quad r = \begin{bmatrix} 1 & 0 \\ \tau & 0 \end{bmatrix}. \tag{3.11}$$

We assume that there is no randomness in the rewards. Here, the pair (p, τ) along with the discount factor γ are parameters of the construction, and we consider a sub-family of these parameters indexed by a scalar $\lambda \geq 0$. For any such λ and discount factor $\gamma \in (\frac{1}{4}, 1)$, consider the settings

$$p = \frac{4\gamma - 1}{3\gamma}, \quad \text{and} \quad \tau = 1 - (1-\gamma)^\lambda.$$

With these choices, the optimal Q-function $\theta^\star$ takes the form

$$\theta^\star = \begin{bmatrix} \frac{1}{4} \cdot \frac{3+\tau}{1-\gamma} & \frac{\gamma}{4} \cdot \frac{3+\tau}{1-\gamma} \\ \frac{\tau}{1-\gamma} & \frac{\gamma\tau}{1-\gamma} \end{bmatrix},$$

with an unique optimal policy $\pi^\star(x_1) = \pi^\star(x_2) = u_1$. We can then compute that

$$\max_{\pi^\star \in \mathbf{\Pi}^\star} \|\nu(\pi^\star; \mathbf{P}_\lambda, r_\lambda)\|_\infty = c \cdot \left(\frac{1}{1-\gamma}\right)^{1.5-\lambda}. \tag{3.12}$$

See Section 3.6 for the details of this calculation.

Substituting into equation (3.10a) yields that the local minimax risk is lower bounded as $\mathfrak{M}_N(\mathcal{P}) \geq c\frac{1}{(1-\gamma)^{1.5-\lambda}}$. Consequently, for $\lambda > 0$, our lower bounds suggest it should be possible to estimate the optimal Q-function more accurately by a factor $(1-\gamma)^\lambda$; note that this difference is particularly significant for values of the discount factor γ that are close to one.

Instance-dependent upper bounds

Thus far, we have stated some instance-dependent lower bounds on the sample complexity of estimating Q-value functions. As we saw in the preceding Example 1, these lower bounds exhibit a wide range of behavior depending on the structure of the transition functions, discount parameter and reward functions. However, these differences in the lower bounds are only interesting if we can show that they are optimal, meaning that there is a (hopefully practical) algorithm that matches the behavior predicted by the lower bounds.

In this section we close this gap, in particular via a careful analysis of variance-reduced Q-learning (or VR-QL for short). Variance-reduced forms of Q-learning have been proposed and shown to be globally minimax in previous work [Sid+18a; Sid+18b; Wai19d]; the version analyzed here is motivated from [Wai19d]. In Theorem 5, we show that the VR-QL algorithm is instance-optimal up to logarithmic factors under two different sets of assumption.

From standard to variance-reduced Q-learning

The classical Q-learning algorithm is a stochastic approximation algorithm for estimating the unique fixed point $\theta^\star$ of the Bellman operator $\mathbf{T}$. Recall the definition (3.4) of the empirical Bellman operator $\widehat{\mathbf{T}}_k$. At each iteration $k = 1, 2, \ldots$, standard Q-learning performs an update of the form

$$\theta_{k+1} = (1 - \alpha_k)\theta_k + \alpha_k \widehat{\mathbf{T}}_k(\theta_k), \tag{3.13}$$

where $\alpha_k \in (0, 1)$ is a stepsize parameter. Appropriately decaying choices of the stepsize ensure that the estimate θ_k converges to $\theta^\star$. Unfortunately, the convergence

rate is known to be non-optimal, failing to achieve the global minimax rate [Wai19c], let alone the finer-grained instance-dependent requirements in this chapter. This non-optimality has to do with the rate at which variance accumulates as the procedure is run.

Variance reduction is a general principle that can be applied to stochastic approximation schemes so as to accelerate their convergence. Here we describe the variance-reduced version of Q-learning that we analyze here. Similar to standard variance-reduced schemes for stochastic optimization [see, e.g., JZ13], the algorithm consists of a sequence of epochs. Within each epoch, we run a re-centered version of the QL update. The re-centering is done in such a way, using a Monte Carlo approximation of the population Bellman operator $\mathbf{T}$, so that the re-centered updates have lower variance. We leave the details of the epochs and Monte Carlo to Section 3.2; here let us describe the basic form of the updates within a given epoch.

Suppose that we run the algorithm using a total of M epochs. At epoch m, the algorithm uses a re-centering point $\overline{\theta}_m$ in order to re-center the update, where $\overline{\theta}_m$ acts as the current best estimate of $\theta^\star$. Ideally, we should re-center the operator $\widehat{\mathbf{T}}_k$ using the quantity $\mathbf{T}(\overline{\theta}_m)$, but we lack the access to it; instead, we use the Monte Carlo approximation

$$\overline{\mathbf{T}}_{N_m}(\overline{\theta}_m) := \frac{1}{N_m} \sum_{i \in \mathcal{D}_m} \widehat{\mathbf{T}}_i(\overline{\theta}_m). \tag{3.14}$$

Given the pair $(\overline{\theta}_m, \overline{\mathbf{T}}_{N_m}(\overline{\theta}))$ and a stepsize parameter $\alpha \in (0,1)$, we define the *variance-reduced Q-learning update* as follows:

$$\theta \mapsto \mathcal{V}_k\left(\theta; \alpha, \overline{\theta}_m, \overline{\mathbf{T}}_{N_m}\right) := (1-\alpha)\theta + \alpha\left\{\widehat{\mathbf{T}}_k(\theta) - \widehat{\mathbf{T}}_k(\overline{\theta}_m) + \overline{\mathbf{T}}_{N_m}(\overline{\theta}_m)\right\}, \tag{3.15}$$

where the operator $\widehat{\mathbf{T}}_k$ is independent of the set of operators $\{\widehat{\mathbf{T}}_i\}_{i \in \mathcal{D}_m}$, used to compute the Monte Carlo approximation $\overline{\mathbf{T}}_{N_m}$. As a result, the stochastic operator $\widehat{\mathbf{T}}_k$ is independent of the re-centering quantity $\overline{\mathbf{T}}_{N_m}(\overline{\theta}_m)$. See Section 3.2 for the details on how the epoch lengths and re-centering sample sizes $\mathcal{D}_m$ are chosen.

Non-asymptotic guarantees for variance-reduced Q-learning

In this section, we state some non-asymptotic guarantees for the VR-QL algorithm. We provide guarantees under two conditions, both of which involve the structure of the set of optimal policies. We begin by introducing some definitions that underlie these two conditions.

Given an MDP instance $(\mathbf{P}, r)$, we define the *optimality gap*

$$\Delta := \min_{\pi \in \Pi \setminus \mathbf{\Pi}^\star} \|\theta^\star - \{r + \gamma \mathbf{P}^\pi \theta^\star\}\|_\infty, \tag{3.16}$$

where $\theta^\star, \mathbf{\Pi}^\star$, and Π, respectively, denote the optimal Q-function, the set of optimal policies, and the set of all policies for MDP $(\mathbf{P}, r)$. Observe that the scalar Δ captures

the difficulty in detecting the set of optimal policies. In other words, when Δ is small, it is hard to distinguish an optimal policy from a suboptimal policy.

Our second set of conditions involves the family of right-linear operators $\{\mathbf{P}^\pi : \pi \in \Pi\}$ defined in equation (3.6). For any Q-value function θ, we say that a policy π is greedy with respect to θ if $\pi(x) \in \arg\max_{u \in \mathcal{U}} \theta(x,u)$ for all $x \in \mathcal{X}$. Note that any policy $\pi^\star$ that is greedy with respect to the optimal Q-value function $\theta^\star$ is an optimal policy. We say that the operators satisfy a *Lipschitz condition* if there is a constant L such that for any Q-value function θ and associated greedy-optimal policy π, we have

$$\|(\mathbf{P}^\pi - \mathbf{P}^{\pi^\star})\,(\theta - \theta^\star)\|_\infty \leq L\|\theta - \theta^\star\|_\infty^2. \tag{3.17}$$

Intuitively, this condition means that the operator difference $\mathbf{P}^\pi - \mathbf{P}^{\pi^\star}$ is small whenever the underlying Q-value functions that induce the policies are close. Conditions of this type were introduced by Puterman and Brumelle [PB79] in their classical analysis of the convergence rates of policy iteration algorithms for exact dynamic programming.

With these definitions in place, we can specify the two settings under which we provide upper bounds on the VR-QL algorithm:

Setting UNQ: For some $\beta > 0$, there is a unique optimal policy $\pi^\star$, and the sample size N is lower bounded as

$$\frac{N}{(\log N)^2} \geq c_2 \log(D/\delta) \cdot \frac{(1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma})^2}{(1-\gamma)^3} \cdot \max\{1, \frac{1}{\Delta^2(1-\gamma)^\beta}\}. \tag{3.18}$$

Setting LIP: For some $\beta > 0$, the Lipschitz condition (3.17) condition holds, and the sample size is lower bounded as

$$\frac{N}{(\log N)^2} \geq c_2 \log(D/\delta) \frac{(1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma})^2}{(1-\gamma)^{3+\beta}} \cdot \min\{\frac{L^2}{(1-\gamma)^2}, \frac{1}{\Delta^2}\}. \tag{3.19}$$

In all cases, we assume that we are given an initial point $\bar{\theta}_1$ such that

$$\|\bar{\theta}_1 - \theta^\star\|_\infty \leq \frac{\|r\|_\infty}{\sqrt{1-\gamma}}. \tag{3.20}$$

Such an initial condition has already been used in the literature [Wai19d], and it can be ensured by first running Algorithm VR-QL for a total of $\frac{1}{(1-\gamma)^3}$ samples (up to logarithmic factor corrections).

Theorem 5. *Under either settings (UNQ) or (LIP), there are choices of epoch parameters such that given any discount parameter $\gamma \in [\frac{1}{2}, 1)$ and an initial point*

$\bar{\theta}_1$ satisfying the initialization condition (3.20), Algorithm VR-QL run for $M := \log_4\left(\frac{N(1-\gamma)^2}{8\log((16D/\delta)\cdot\log N)}\right)$ epochs yields an estimate $\bar{\theta}_{M+1}$ such that

$$\|\bar{\theta}_{M+1} - \theta^\star\|_\infty \leq c_0 \cdot \sqrt{\frac{\log_4(8DM/\delta)}{N}} \max_{\pi^\star \in \mathbf{\Pi}^\star} \|\nu(\pi^\star; \mathbf{P}, r, \gamma\,())\,\|_\infty + c_1 \cdot \frac{\log_4(8DM/\delta)}{N} \cdot \frac{\|\theta^\star\|_{\text{span}}}{1-\gamma}, \tag{3.21}$$

with probability exceeding $1 - \delta$.

See Section 3.4 for the proof of this claim.

Comparing the upper and lower bounds: Assuming the sample size lower bound from Theorem 4 are valid, we see that the second term in the bound (3.21) is of smaller order. In this case, the upper bound from Theorem 5 and the lower bound from Theorem 4 matches, and we conclude that the VR-QL algorithm is *instance optimal*.

Although the guarantee (3.21) involves the same $1/\sqrt{n}$ rate and complexity term $\max_{\pi^\star \in \mathbf{\Pi}^\star} \|\nu(\pi^\star; \mathbf{P}, r, \gamma\,())\,\|_\infty$ as the lower bound in Theorem 4, it should be noted that the sample size lower bounds required for Theorem 5 are more stringent than that in Theorem 4. Moreover, our lower bound does not require the side conditions—either the unique optima or Lipschitz conditions—that are imposed in Theorem 5. Closing these remaining differences between the two results is a worthwhile goal for future work.

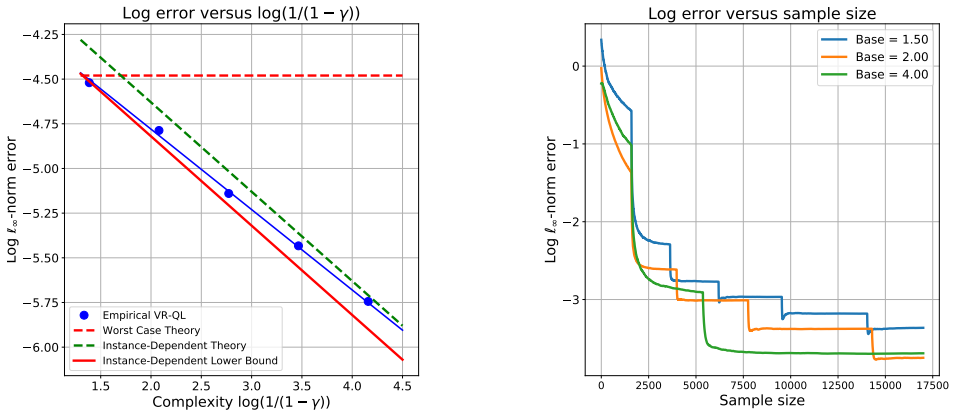
Confirming the theoretical predictions

Some numerical experiments are helpful in order to illustrate instance-adaptive behavior guaranteed by Theorem 5. Recall the family of MDPs (3.11) from Example 1. Suppose that we set $\lambda = 0.5$ and for each choice of $\gamma \in (1/2, 1)$, we collect $N = \lceil \frac{16 \cdot 32}{9} \frac{1}{(1-\gamma)^3.} \rceil$ samples, and then run the VR-QL algorithm over a range of discount parameters γ, using the settings from Theorem 5 and Section 3.2, thereby obtaining an estimate $\bar{\theta}_{M+1}$.

Figure 3.1(a) plots the evolution of log ℓ_∞-norm error of the estimate over time as the algorithm proceeds; the form of these curves show the epoch-based nature of the convergence. See Section 3.2 for more details on the parameters of the epochs, including the base parameter illustrated here. Plotted as blue circles in panel (b) of Figure 3.1 are the logarithm of the ℓ_∞-norm error of the final output; that is, $\log \|\bar{\theta}_{M+1} - \theta^\star\|_\infty$, versus the logarithm of the discount complexity, $\log(1/(1-\gamma))$. Each point in this plot represents an average over 1000 trials.

In terms of theory, with the settings given above, existing worst-case bounds [AMK13; Wai19d] predict that the log ℓ_∞-norm error remains constant as the log discount complexity grows; accordingly, we have plotted a dotted red line with slope zero to

illustrate the worst-case guarantee. On the other hand, for the MDP instance (3.11) with $\lambda = 0.5$, a simple calculation yields that for the instance (3.11) the suboptimality gap Δ satisfies $\Delta = 1 - \frac{(1-\gamma)^{\lambda}}{4} \geq \frac{3}{4}$. In our experiment, we set the sample size to be $N = \lceil \frac{32}{(1-\gamma)^3} \cdot \frac{4^2}{3^2} \rceil \geq \frac{32}{(1-\gamma)^3 \cdot \Delta^2}$; as a result, the bounds from Theorems 4 and 5 are valid.



(a) (b)

Figure 3.1. (a) $\lambda = 0.5$, $N = \lceil \frac{32}{(1-\gamma)^3} \cdot \frac{4^2}{3^2} \rceil$ $\gamma = 0.9$. Illustration of the qualitative behavior of Algorithm VR-QL applied on the MDP (1) along with instance dependent and the worst case bounds. The figure plots the log ℓ_∞-error $\|\bar{\mathbb{Q}}_{M+1} - \theta^\star\|_\infty$ against the log discount complexity factor $\log(\frac{1}{1-\gamma})$ with $\lambda = 0.5$. We have also plotted the least-squares fit through these points, and the instance-dependent lower bound from Theorem 4, the instance-dependent upper bound from Theorem 5, and the worst-case bound [Wai19d]. (b) Behavior of the VR-QL algorithm with different choices of the base b. The plot demonstrates that different choices of the base b yield similar behavior.

With the setting $\lambda = 0.5$, our calculations from Example 1 yield

$$\max_{\pi^\star \in \mathbf{\Pi}^\star} \|\nu(\pi^\star; \mathbf{P}, r, \gamma\,())\,\|_\infty \asymp \left(\frac{1}{1-\gamma}\right)^{-0.5}.$$

Thus, with the choice of sample size N given above, our theory predicts that the log ℓ_∞-norm error should exhibit the scaling

$$\log \|\bar{\theta}_{M+1} - \theta^\star\|_\infty \asymp \log\left(\frac{1}{\sqrt{N}} \max_{\pi^\star \in \mathbf{\Pi}^\star} \|\nu(\pi^\star; \mathbf{P}, r, \gamma\,())\,\|_\infty\right) \asymp c - 0.5 \log\left(\frac{1}{1-\gamma}\right),$$

where c is a constant. In Figure 3.1(b), we plot the lower bound from Theorem 4 as a solid red line, and the upper bound from Theorem 5 as a dashed green line. (While these lines both have slope -0.50, the intercept term c is different due to the additional logarithmic factors in dimension present in the upper bound.)

In order to test how the empirical behavior conforms to the theoretical prediction, we did an ordinary least-squares fit of the log ℓ_∞-norm error versus the log discount complexity; this fit yields a line with slope $\widehat{\beta} = -0.45$, and is plotted in solid blue. This test shows good agreement between the theoretical prediction and the practical behavior.

Details of the epochs and procedure

In this section, we provide the complete details of the algorithm used in our version of variance-reduced Q-learning.

A single epoch: A single epoch of the overall variance-reduced QL algorithm involves repeated applications of the basic variance-reduced update $\mathcal{V}_k$ from equation (3.15). The epochs are indexed with integers $m = 1, 2, \ldots, M$, where M corresponds to the total number of epochs to be run. Each epoch m requires the following four inputs:

- an element $\overline{\theta}$, which is chosen to be the output of the previous epoch $m-1$;
- a positive integer K denoting the number of steps within the given epoch;
- a positive integer N_m denoting the batch size used to calculate the Monte Carlo update (3.14);
- a set of fresh operators $\{\widehat{\mathbf{T}}_i\}_{i \in \mathcal{C}_m}$, with $|\mathcal{C}_m| = N_m + K$. The set $\mathcal{C}_m$ is partitioned into two subsets having sizes N_m and K, respectively. The first subset, of size N_m, which we call $\mathcal{D}_m$, is used to construct the Monte Carlo approximation (3.14). The second subset, of size K is used to run the K steps within the epoch.

We summarize a single epoch in pseudocode form in Algorithm SingleEpoch.

Overall algorithm: The overall algorithm, denoted by VR-QL for short, has five inputs: (a) an initialization $\bar{\theta}_1$, (b) an integer M, denoting the number of epochs to be run, (c) an integer K, denoting the length of each epoch, (d) a sequence of batch sizes $\{N_m\}_{m=1}^K$, denoting the number of operators used for re-centering in the M epochs, and (e) sample batches $\{\{\widehat{\mathbf{T}}_i\}_{i \in \mathcal{C}_m}\}_{m=1}^M$ to be used in the M epochs. Given these five inputs, the overall procedure can be summarized as in Algorithm VR-QL.

Settings for Theorem 5: Given a tolerance probability $\delta \in (0,1)$ and the number of available i.i.d. samples N, we run Algorithm VR-QL with a total of $M := \log_4 \left(\frac{N(1-\gamma)^2}{8 \log((16D/\delta) \cdot \log N)} \right)$ epochs, along with the following parameter choices:

`Re-centering sizes:`

$$N_m = c_1 \frac{4^m}{(1-\gamma)^2} \cdot \log_4(16MD/\delta) \tag{3.22a}$$

Algorithm SingleEpoch RunEpoch $(\overline{\theta}; K, N_m, \{\widehat{\mathbf{T}}_i\}_{i \in \mathcal{C}_m})$

1: Given (a) Epoch length K, (b) Re-centering vector $\overline{\theta}$, (c) Re-centering batch size N_m, (d) Operators $\{\widehat{\mathbf{T}}_i\}_{i \in \mathcal{C}_m}$
2: Compute the re-centering quantity

$$\overline{\mathbf{T}}_{N_m}(\overline{\theta}) := \frac{1}{N_m} \sum_{i \in \mathcal{D}_m} \widehat{\mathbf{T}}_i(\overline{\theta})$$

3: Initialize $\mathbb{Q}_1 = \bar{\theta}$
4: **for** $k = 1, 2, \ldots, K$ **do**
5: Compute the variance-reduced update:

$$\theta_{k+1} = \mathcal{V}_k(\theta_k; \alpha_k, \overline{\theta}, \overline{\mathbf{T}}_{N_m}) \quad \text{with stepsize } \alpha_k = \frac{1}{1 + (1-\gamma)k}.$$

6: **end for**
7: **return** θ_{K+1}

Algorithm VR-QL .

1: Given (a) Initialization $\bar{\theta}_1$, (b) Number of epochs, M, (c) Epoch length K, (d) Re-centering sample sizes $\{N_m\}_{m=1}^M$, (e) Sample batches $\{\widehat{\mathbf{T}}_i\}_{i \in \mathcal{C}_m}$ for $m = 1, \ldots, M$

2: Initialize at $\overline{\theta}_1$
3: **for** $m = 1, 2, \ldots, M$ **do**
4: $\overline{\theta}_{m+1} = \text{RunEpoch}(\overline{\theta}_m; K, N_m, \{\widehat{\mathbf{T}}_i\}_{i \in \mathcal{C}_m})$
5: **end for**
6: **return** $\overline{\theta}_{M+1}$ as final estimate

`Sample batches:`

Partition the N samples to obtain $\{\widehat{\mathbf{T}}_i\}_{i \in \mathcal{C}_m}$ for $m = 1, \ldots, M$ (3.22b)

`Epoch length:`

$$K = \frac{N}{2M}. \tag{3.22c}$$

3.3 Proof of Theorem 4

Given an MDP instance $\mathcal{P} = (\mathbf{P}, r)$, we start by introducing the following two classes of alternative MDPs:

$$\mathcal{S}_1 = \{\mathcal{P}' = (\mathbf{P}', r') \mid r' = r\}, \quad \text{and} \quad \mathcal{S}_2 = \{\mathcal{P}' = (\mathbf{P}', r') \mid \mathbf{P}' = \mathbf{P}\}. \tag{3.23}$$

We consider the restricted version of the local minimax risk at the instance $\mathcal{P}'$ to the classes $\mathcal{S}_i$:

$$\mathfrak{M}_N(\mathcal{P}; \mathcal{S}_i) := \sup_{\mathcal{P}' \in \mathcal{S}_i} \inf_{\hat{\theta}_N} \max_{\mathcal{Q} \in \{\mathcal{P}, \mathcal{P}'\}} \mathbb{E}_{\mathcal{P}}[\sqrt{N} \|\hat{\theta}_N - \theta(\mathcal{Q})\|_\infty], \quad i = 1, 2. \tag{3.24}$$

The main part of the proof involves showing that there exists a universal constant $c > 0$ such that

$$\mathfrak{M}_N(\mathcal{P}; \mathcal{S}_1) \geq c \cdot \max_{\pi \in \mathbf{\Pi}^\star} \|\gamma \rho(\pi; \mathbf{P}, r)\|_\infty, \quad \text{and} \tag{3.25a}$$

$$\mathfrak{M}_N(\mathcal{P}; \mathcal{S}_2) \geq c \cdot \max_{\pi \in \mathbf{\Pi}^\star} \|\sigma(\pi; \mathbf{P}, r)\|_\infty, \tag{3.25b}$$

where $\mathbf{\Pi}^\star$ denotes the optimal policy set for $(\mathbf{P}, r)$. We can then conclude

$$\begin{aligned}
\mathfrak{M}_N(\mathcal{P}) &\geq \max\{\mathfrak{M}_N(\mathcal{P}; \mathcal{S}_1), \mathfrak{M}_N(\mathcal{P}; \mathcal{S}_2)\} \\
&\geq \frac{1}{2} \left(\mathfrak{M}_N(\mathcal{P}; \mathcal{S}_1) + \mathfrak{M}_N(\mathcal{P}; \mathcal{S}_2)\right) \\
&\geq \frac{c}{2} \max_{\pi \in \mathbf{\Pi}^\star} \|\gamma \rho(\pi; \mathbf{P}, r)\|_\infty + \frac{c}{2} \max_{\pi \in \mathbf{\Pi}^\star} \|\sigma(\pi; \mathbf{P}, r)\|_\infty \\
&\geq \frac{c}{2} \max_{\pi \in \mathbf{\Pi}^\star} \|\nu(\pi; \mathbf{P}, r)\|_\infty.
\end{aligned}$$

The last inequality above follows from the decomposition (3.9a). It remains to prove the claims (3.25a) and (3.25b). More precisely, the core of our proof involves proving the following two lemmas:

Lemma 7. *For all $\mathcal{S} \in \{\mathcal{S}_1, \mathcal{S}_2\}$, we have that $\mathfrak{M}_N(\mathcal{P}; \mathcal{S}) \geq \frac{1}{8} \underline{\mathfrak{M}}_N(\mathcal{P}; \mathcal{S})$ where we define*

$$\underline{\mathfrak{M}}_N(\mathcal{P}; \mathcal{S}) := \sup_{\mathcal{P}' \in \mathcal{S}} \left\{ \sqrt{N} \cdot \|\theta(\mathcal{P}) - \theta(\mathcal{P}')\|_\infty \mid d_{\text{hel}}(\mathcal{P}, \mathcal{P}') \leq \frac{1}{2\sqrt{N}} \right\}.$$

This lemma follows as a fairly straightforward consequence of the standard reduction from estimation to testing; see Section 3.7 for the details.

Our next lemma requires more effort to prove, and leverages the specific structure of the problem at hand:

Lemma 8. *Given any MDP instance $\mathcal{P} = (\mathbf{P}, r)$:*

(a) There exists an instance $\mathcal{P}_1 = (\mathbf{P}', r) \in \mathcal{S}_1$ such that $d_{\text{hel}}(\mathcal{P}, \mathcal{P}_1) \leq \frac{1}{2\sqrt{N}}$ and

$$\sqrt{N} \cdot \|\theta(\mathcal{P}) - \theta(\mathcal{P}_1)\|_\infty \geq c \cdot \max_{\pi \in \mathbf{\Pi}^\star} \|\gamma \rho(\pi; \mathbf{P}, r)\|_\infty.$$

(b) There exists an instance $\mathcal{P}_2 = (\mathbf{P}, r') \in \mathcal{S}_2$ such that $d_{\text{hel}}(\mathcal{P}, \mathcal{P}_2) \leq \frac{1}{2\sqrt{N}}$ and

$$\sqrt{N} \cdot \|\theta(\mathcal{P}) - \theta(\mathcal{P}_2)\|_\infty \geq c \cdot \max_{\pi \in \mathbf{\Pi}^\star} \|\sigma(\pi; \mathbf{P}, r)\|_\infty.$$

Note that the bounds (3.25a)–(3.25b) stated in Theorem 4 follow by combining the claims of Lemmas 7 and 8. The remainder of our proof focuses on establishing Lemma 8.

Proof of Lemma 8

In this section, we prove the two parts of Lemma 8.

Proof of Lemma 8(a)

Throughout the proof, we use z to denote a generic element of the state-action set $\mathcal{X} \times \mathcal{U}$. Let θ be the true Q-function for the MDP $\mathcal{P} = (P, r)$. We adopt the shorthands

$$\pi_1 \in \arg\max_{\pi \in \mathbf{\Pi}^\star} \|\rho(\pi; \mathbf{P}, r)\|_\infty, \quad \bar{z} \in \arg\max_{z \in \mathcal{X} \times \mathcal{U}} \rho(\pi_1; \mathbf{P}, r), \quad \tilde{\rho}(z) := \rho(\pi_1; \mathbf{P}, r)(z), \tag{3.26a}$$

$$\mathbf{U} := (\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1} \quad \text{and} \quad \varphi^2(z) := \mathrm{Var}\left(\mathbf{Z}^{\pi_1} \theta(z)\right). \tag{3.26b}$$

To explain this notation, we chose π_1 to be the optimal policy that achieves the largest ℓ_∞-norm across $\rho(\pi^*; \mathbf{P}, r)$ for optimal policies π^*, we let $\bar{z}$ is the state-action pair index that achieves the maximal entry of $\rho(\pi_1; \mathbf{P}, r)$, and we use $\tilde{\rho}$ as convenient shorthand to refer to the values of $\rho(\pi_1; \mathbf{P}, r)$. This choice of notation implies that

$$\tilde{\rho}(\bar{z}) = \max_{\pi \in \mathbf{\Pi}^\star} \|\rho(\pi; \mathbf{P}, r)\|_\infty.$$

Additionally, note that $\mathbf{U}$ is a linear transformation from $\mathbb{R}^{|\mathcal{X}| \times |\mathcal{U}|}$ to itself, so we can express the action of $\mathbf{U}$ on θ as

$$(\mathbf{U}\theta)(z) = \sum_{z' \in \mathcal{X} \times \mathcal{U}} \mathbf{U}_{z,z'} \theta(z').$$

Note moreover that

$$\varphi^2(z) = \sum_{x'} \mathbf{P}_{x',z} (\theta(x', \pi_1(x')) - (\mathbf{P}^{\pi_1}\theta)(z))^2 \quad \text{and} \quad \tilde{\rho}^2(z) = \sum_{z'} (\mathbf{U}_{z,z'})^2 \varphi^2(z'). \tag{3.27}$$

With these definitions, we now define $\bar{\mathbf{P}}_{y,z}$ as follows (we will prove that this choice is a valid probability transition kernel shortly):

$$\bar{\mathbf{P}}_{y,z} = \mathbf{P}_{y,z} + \frac{1}{\tilde{\rho}(\bar{z})\sqrt{2N}} \mathbf{P}_{y,z} \mathbf{U}_{\bar{z},z} \cdot (\theta(y, \pi_1(y)) - (\mathbf{P}^{\pi_1}\theta)(z)). \tag{3.28}$$

Here, we have used the shorthand $\mathbf{P}_{y,z} \equiv P_u(y \mid x)$, where $z = (x, u) \in \mathcal{X} \times \mathcal{U}$. Let $\theta := \theta(\mathbf{P}, r)$, and $\bar{\theta} := \theta(\bar{\mathbf{P}}, r)$ be the optimal Q functions for MDP instances $(\mathbf{P}, r)$ and $(\bar{\mathbf{P}}, r)$ respectively. In the rest of the proof, we use the following properties of $\bar{\mathbf{P}}$.

Lemma 9. *For any MDP $\mathcal{P} = (P, r)$ and the optimal policy π_1 defined in equation (3.26), the corresponding $\bar{\mathbf{P}}$ has the following properties:*

(a) The $\bar{\mathbf{P}}$ is a probability transition kernel.

(b) The MDP instances $\mathcal{P} = (P, r)$ and $\mathcal{P}_1 = (\bar{\mathbf{P}}, r)$ satisfy $d_{\text{hel}}(\mathcal{P}, \mathcal{P}_1) \leq \frac{1}{2\sqrt{N}}$ and $|||\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}|||_{op} \leq \frac{1}{\sqrt{2N}}$.

(c) Each entry of $(\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1}[\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}]\theta$ is non-negative.

See Section 3.7 for a proof of this lemma.
Equipped with these tools, we are now ready to lower bound the norm $\|\theta - \bar{\theta}\|_\infty$. The optimal Q-functions θ and $\bar{\theta}$ satisfy the following Bellman equations:

$$\theta = r + \gamma \mathbf{P}^{\pi_1}\theta \qquad \text{and} \qquad \bar{\theta} = r + \gamma \bar{\mathbf{P}}^{\bar{\pi}}\bar{\theta}, \tag{3.29}$$

where $\pi_1 \in \mathbf{\Pi}^\star$ is the optimal policy that achieves $\max_{\pi \in \mathbf{\Pi}^\star} \|\rho(\pi; \mathbf{P}, r)\|_\infty$, and $\bar{\pi}$ is an optimal policy for $(\bar{\mathbf{P}}, r)$. By the optimality of policy $\bar{\pi}$ and the Q-function $\bar{\theta}$, we have the entrywise inequality $\bar{\mathbf{P}}^{\bar{\pi}}\bar{\theta} \succeq \bar{\mathbf{P}}^{\pi_1}\bar{\theta}$, which implies $(\mathcal{I} - \gamma \bar{\mathbf{P}}^{\pi_1})\bar{\theta} \succeq (\mathcal{I} - \gamma \bar{\mathbf{P}}^{\bar{\pi}})\bar{\theta} = r$. Thus, using the identity $A_1^{-1} - A_0^{-1} = A_1^{-1}(A_0 - A_1)A_0^{-1}$ for invertible operators A_0 and A_1, we have

$$\begin{aligned}
\bar{\theta} - \theta &\succeq \left[(\mathcal{I} - \gamma \bar{\mathbf{P}}^{\pi_1})^{-1} - (\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1}\right] r \\
&= (\mathcal{I} - \gamma \bar{\mathbf{P}}^{\pi_1})^{-1} \left[(\mathcal{I} - \gamma \mathbf{P}^{\pi_1}) - (\mathcal{I} - \gamma \bar{\mathbf{P}}^{\pi_1})\right] (\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1} r \\
&= \gamma (\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1} \left[\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}\right] (\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1} r \\
&\quad + \gamma \left((\mathcal{I} - \gamma \bar{\mathbf{P}}^{\pi_1})^{-1} - (\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1}\right) (\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})(\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1} r \\
&= \gamma (\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1} \left[\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}\right] \theta \\
&\quad + \gamma \left((\mathcal{I} - \gamma \bar{\mathbf{P}}^{\pi_1})^{-1} - (\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1}\right) (\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})\theta,
\end{aligned}$$

where the final equation follows from the Bellman optimality condition (3.29). Lemma 9(c) guarantees that the entries of $(\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1}[\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}]\theta$ are non-negative, and therefore we conclude

$$\begin{aligned}
\|\bar{\theta} - \theta\|_\infty \geq \gamma \|(\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1}[\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}]\theta\|_\infty \\
- \gamma \|((\mathcal{I} - \gamma \bar{\mathbf{P}}^{\pi_1})^{-1} - (\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1})(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})\theta\|_\infty.
\end{aligned} \tag{3.30}$$

Consider the second term $T_2 := \|((\mathcal{I} - \gamma \bar{\mathbf{P}}^{\pi_1})^{-1} - (\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1})(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})\theta\|_\infty$. We have

$$\begin{aligned}
T_2 &= \|[(\mathcal{I} - \gamma \bar{\mathbf{P}}^{\pi_1})^{-1}(\mathcal{I} - \gamma \mathbf{P}^{\pi_1}) - \mathcal{I}](\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1}(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})\theta\|_\infty \\
&\leq |||(\mathcal{I} - \gamma \bar{\mathbf{P}}^{\pi_1})^{-1}(\mathcal{I} - \gamma \mathbf{P}^{\pi_1}) - \mathcal{I}|||_{\text{op}} \cdot \|(\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1}(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})\theta\|_\infty \\
&= \gamma |||(\mathcal{I} - \gamma \bar{\mathbf{P}}^{\pi_1})^{-1}(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})|||_{\text{op}} \cdot \|(\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1}(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})\theta\|_\infty \\
&\leq \frac{\gamma}{1-\gamma} |||\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}|||_{\text{op}} \cdot \|(\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1}(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})\theta\|_\infty \\
&\leq \frac{\gamma}{2} \|(\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1}(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})\theta\|_\infty,
\end{aligned}$$

where the last inequality uses Lemma 9(b) and the first part of the minimum sample size assumption (3.10b). Combining this result with the bound (3.30) we conclude

$$\|\bar{\theta} - \theta\|_\infty \geq \frac{\gamma}{2}\|(\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1}(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})\theta\|_\infty.$$

With this result in hand, substituting the value of the transition kernel $\bar{\mathbf{P}}$ from equation (3.28) and recalling the definition of state-action pair z from equation (3.26) we have

$$\begin{aligned}
\sqrt{N} \cdot \|\bar{\theta} - \theta\|_\infty &\geq \frac{\gamma\sqrt{N}}{2} \cdot (\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1}(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})\theta(\bar{z}) \\
&= \frac{\gamma\sqrt{N}}{2\sqrt{2}} \cdot \sum_z \mathbf{U}_{\bar{z},z} \cdot (\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})\theta(z) \\
&\overset{(i)}{\geq} \frac{\gamma}{4\tilde{\rho}(\bar{z})} \sum_z (\mathbf{U}_{\bar{z},z})^2 \varphi^2(z) \\
&\overset{(ii)}{=} \frac{\gamma\tilde{\rho}(\bar{z})}{4} = \frac{1}{4} \cdot \max_{\pi \in \mathbf{\Pi}^\star} \|\gamma\rho(\pi; \mathbf{P}, r)\|_\infty,
\end{aligned}$$

where step (i) follows by substituting the value of the transition kernel $\bar{\mathbf{P}}$ (cf. Proof of Lemma 9 part (c)), and step (ii) follows by using the expression (3.27). This completes the proof of part (a) of Lemma 8.

Proof of Lemma 8(b)

Borrowing the notation from part (a) of the proof, let z denote a generic element of the state-action set $\mathcal{X} \times \mathcal{U}$. Let $\pi_2 \in \arg\max_{\pi \in \mathbf{\Pi}^\star} \|\sigma(\pi; \mathbf{P}, r)\|_\infty$. We use the shorthands

$$\sigma^2(\bar{z}) := \max_{\pi \in \mathbf{\Pi}^\star} \|\sigma(\pi; \mathbf{P}, r)\|_\infty^2 = \|\sigma(\pi_2; \mathbf{P}, r)\|_\infty^2, \qquad \text{and} \quad \mathbf{U} := (\mathcal{I} - \gamma \mathbf{P}^{\pi_2})^{-1}. \tag{3.31}$$

We define our perturbed reward function to be

$$\bar{r}(z) = r(z) + \frac{1}{\sigma(\bar{z})\sqrt{2N}} \mathbf{U}_{\bar{z},z} \sigma_r^2 \quad \text{for} \quad z \in \mathcal{X} \times \mathcal{U}. \tag{3.32}$$

For $\mathcal{P}_2 := (P, \bar{r})$, a short computation shows that the Hellinger distance between the components of the instance pair $(\mathcal{P}, \mathcal{P}_2)$ takes the form

$$d_{\text{hel}}(\mathcal{P}, \mathcal{P}_2)^2 \leq D_{KL}(\mathcal{N}(r, \sigma_r^2 \mathcal{I}) \mid \mathcal{N}(\bar{r}, \sigma_r^2 \mathcal{I})) = \frac{1}{2\sigma_r^2}\|r - \bar{r}\|_2^2.$$

Substituting the value of the reward $\bar{r}$ from equation (3.32) yields

$$d_{\text{hel}}(\mathcal{P}, \mathcal{P}_2)^2 \leq \frac{1}{2\sigma_r^2}\|\bar{r} - r\|_2^2 = \frac{1}{\sigma^2(\bar{z}) \cdot 4N} \sum_z (\mathbf{U}_{\bar{z},z})^2 \sigma_r^2 = \frac{1}{4N},$$

where the last equality uses the definition of the term $\sigma^2(\bar{z})$, i.e.,

$$\sigma^2(\bar{z}) = \sum_{z'} (\mathbf{U}_{\bar{z},z'})^2 \sigma_r^2. \tag{3.33}$$

It remains to prove a lower bound on the ℓ_∞-norm between the optimal Q-functions for instances $\mathcal{P}$ and $\mathcal{P}_2$.

Let $\theta := \theta(\mathbf{P}, r)$, and $\bar{\theta} := \theta(\mathbf{P}, \bar{r})$ be the optimal Q functions for MDP instances $\mathcal{P} := (\mathbf{P}, r)$ and $\mathcal{P}_2 := (\mathbf{P}, \bar{r})$, respectively. Note that θ and $\bar{\theta}$ satisfy the Bellman equations

$$\theta = r + \gamma \mathbf{P}^{\pi_2} \theta, \quad \text{and} \quad \bar{\theta} = \bar{r} + \gamma \mathbf{P}^{\bar{\pi}} \bar{\theta}, \tag{3.34}$$

where $\bar{\pi}$ is an optimal policy for the MDP instance $(\mathbf{P}, \bar{r})$. By the optimality of policy $\bar{\pi}$, we have the entrywise inequality $\mathbf{P}^{\bar{\pi}} \bar{\theta} \succeq \mathbf{P}^{\pi_2} \bar{\theta}$; as a result, we have

$$(\mathcal{I} - \gamma \mathbf{P}^{\pi_2}) \bar{\theta} \succeq \bar{r} \implies \bar{\theta} \succeq (\mathcal{I} - \gamma \mathbf{P}^{\pi_2})^{-1} \bar{r},$$

where the last step uses the fact that $(\mathcal{I} - \gamma \mathbf{P}^{\pi_2})^{-1}$ is entry-wise non-negative. Combining the last inequality with the Bellman equation (3.34) we have that

$$\bar{\theta} - \theta \succeq (\mathcal{I} - \gamma \mathbf{P}^{\pi_2})^{-1} (\bar{r} - r) \tag{3.35}$$

and that

$$\begin{aligned} \|(\mathcal{I} - \gamma \mathbf{P}^{\pi_2})^{-1} (\bar{r} - r)\|_\infty \geq (\mathcal{I} - \gamma \mathbf{P}^{\pi_2})^{-1} (\bar{r} - r)(\bar{z}) &= \frac{1}{\sigma(\bar{z}) \sqrt{2N}} \sum_z (\mathbf{U}_{\bar{z},z})^2 \sigma_r^2 \\ &= \frac{\sigma(\bar{z})}{\sqrt{2N}}, \end{aligned}$$

where the last equality uses the relation (3.33). Putting together the pieces, we have shown that

$$\|\bar{\theta} - \theta\|_\infty \geq \frac{\sigma(\bar{z})}{\sqrt{2N}} = \frac{1}{\sqrt{2N}} \cdot \max_{\pi \in \mathbf{\Pi}^\star} \|\sigma(\pi; \mathbf{P}, r)\|_\infty,$$

as desired.

3.4 Proof of Theorem 5

In the section, we provide a proof of the upper bounds stated in Theorem 5. Throughout the proof, we adopt the following shorthands

$$\kappa = \frac{\|\theta^\star\|_{\text{span}}}{(1-\gamma)} \cdot \log(8DM/\delta), \qquad \tau^* = \max_{\pi^\star \in \mathbf{\Pi}^\star} \|\nu(\pi^\star; \mathbf{P}, r)\|_\infty \cdot \sqrt{\log(8DM/\delta)},$$
$$\text{and} \qquad \tau_{\max} = \frac{1 + \|r\|_\infty + \sigma_r \sqrt{1-\gamma}}{(1-\gamma)^{1.5}} \cdot \sqrt{\log(8DM/\delta)}. \tag{3.36}$$

Proof of Theorem 5(a)

Our proof is based on the following two lemmas characterizing the behavior of VR-QL across epochs.

Lemma 10. *Under the assumptions of Theorem 5, for each epoch $m = 1, \ldots, M$, we have*

$$\|\bar{\theta}_{m+1} - \theta^\star\|_\infty \leq \frac{\|\bar{\theta}_m - \theta^\star\|_\infty}{16} + c\left(\frac{\tau_{\max}}{\sqrt{N_m}} + \frac{\kappa}{N_m}\right), \tag{3.37}$$

with probability at least $1 - \frac{\delta}{M}$.

Lemma 10 follows by an argument similar to that used in the proof of Theorem 1 of the paper [Wai19d], so we omit the details here. See also the proof of Lemma 11 for some relevant arguments.

Lemma 11. *Under the assumptions of Theorem 5, for epochs m such that the re-centering sample size N_m satisfies the bound $N_m \geq \log_4(8DM/\delta)\frac{(1+\|r\|_\infty+\sigma_r\sqrt{1-\gamma})^2}{\Delta^2(1-\gamma)^3}$, we have*

$$\|\bar{\theta}_{m+1} - \theta^\star\|_\infty \leq \frac{\|\bar{\theta}_m - \theta^\star\|_\infty}{16} + c \cdot \left(\frac{\tau^*}{\sqrt{N_m}} + \frac{\kappa}{N_m}\right), \tag{3.38}$$

with probability at least $1 - \frac{\delta}{M}$.

See Section 3.4 for the proof of Lemma 11.

Completing the proof

Using the two lemmas above, we can now complete the proof of Theorem 5(a). Recalling the epoch sample size formula (3.22a), we see that the bound (3.38) holds for all epochs

$$m \geq m^* := \log_2 \frac{1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma}}{\Delta\sqrt{1-\gamma}}.$$

Observe that the minimum sample size requirement from Theorem 5 ensures that $M \geq m^*$. Now, applying the recursions (3.38) and (3.37) we obtain

$$\begin{aligned}
\|\bar{\theta}_{M+1} - \theta^\star\|_\infty &\leq \frac{\|\bar{\theta}_M - \theta^\star\|_\infty}{16} + c\left(\frac{\tau^*}{\sqrt{N_M}} + \frac{\kappa}{N_M}\right) \\
&\overset{(i)}{\leq} \frac{\|\bar{\theta}_{m^*} - \theta^\star\|_\infty}{16^{M-m^*}} + c \cdot \left(\sum_{k=0}^{M-m^*} \frac{\tau^*}{16^k\sqrt{N_{M-k}}} + \frac{\kappa}{16^k \cdot N_{M-k}}\right) \\
&\overset{(ii)}{\leq} \frac{\|\bar{\theta}_1 - \theta^\star\|_\infty}{16^M} + c \cdot \left(\sum_{k=0}^{M-m^*} \frac{\tau^*}{16^k\sqrt{N_{M-k}}} + \sum_{k=M-m^*+1}^{M} \frac{\tau_{\max}}{16^k\sqrt{N_{M-k}}}\right) \\
&\qquad + c \cdot \sum_{k=0}^{M} \frac{\kappa}{16^k N_{M-k}} \\
&\overset{(iii)}{\leq} \frac{\|\bar{\theta}_1 - \theta^\star\|_\infty}{16^M} + c \cdot \left(\frac{\tau_{\max}}{8^{M-m^*} \cdot \sqrt{N_M}} + \frac{\tau^*}{\sqrt{N_M}} + \frac{\kappa}{N_M}\right).
\end{aligned}$$

Inequality (i) follows via repeated application of the recursion (3.38), inequality (ii) follows via repeated application of the recursion (3.37), and inequality (iii) utilizes the relation $N_{M-k} \cdot 4^k = N_M$ (cf. definition (3.22a)). Via the union bound, the above inequalities hold simultaneously with probability at least $1 - \delta$. Next, note that by our choice of N_m, we have the inequality $2N_M \leq N \leq \frac{8}{3}N_M$. Putting together the pieces, we conclude that

$$\begin{aligned}
\|\bar{\theta}_{M+1} - \theta^\star\|_\infty &\leq c \cdot \|\bar{\theta}_1 - \theta^\star\|_\infty \cdot \frac{\log^2((8D/\delta) \cdot \log n)}{n^2(1-\gamma)^4} \\
&\quad + c \cdot \frac{(1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma})^4}{(1-\gamma)^{1.5}\sqrt{n}} \cdot \frac{\log^2((8D/\delta) \cdot \log n)}{N^{3/2}(1-\gamma)^{\frac{9}{2}}\Delta^3} \\
&\quad + c \cdot \left(\sqrt{\frac{\log_4(8DM/\delta)}{N}} \max_{\pi^\star \in \mathbf{\Pi}^\star} \|\nu(\pi^\star; \mathbf{P}, r, \gamma\,())\,\|_\infty + \frac{\log_4(8DM/\delta)}{N} \cdot \frac{\|\theta^\star\|_{\text{span}}}{1-\gamma}\right).
\end{aligned} \tag{3.39}$$

Substituting the lower bound condition

$$\frac{N}{\log^2(N)} \geq c \log(D/\delta) \frac{(1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma})^2}{(1-\gamma)^3} \cdot \max\left\{1, \frac{1}{\Delta^2 \cdot (1-\gamma)^\beta}\right\}$$

yields the claimed bound. All that remains is to verify the choice of batch sizes $\{N_m\}_{M=1}^M$ is a valid choice, i.e., we need to verify that the algorithm VR-QL with parameter choices (3.22) uses at most N samples. Recall that the total number of samples used in the M epochs is given by $KM + \sum_{m=1}^M N_m$. Substituting the values

of N_m and M from equations (3.22) we obtain

$$KM + \sum_{M=1}^{M} N_m \leq c \cdot \log_4(8DM/\delta) \cdot \sum_{m=1}^{M} \frac{4^m}{(1-\gamma)^2} + \frac{N}{2}$$
$$\leq c' \cdot \log_4(8DM/\delta) \cdot \frac{4^M}{(1-\gamma)^2} \leq \frac{N}{2} + \frac{N}{2} \leq N.$$

This completes the proof of Theorem 5(a).

Comment on the lower-order terms: Here, we argue that the first two terms in the right-hand side of the bound (3.39) are of lower order. A careful look at the proof reveals that for any $p \geq 1$ by increasing our choice of N_m by a constant factor depending on p, we can bound the first term by

$$c_1 \cdot \frac{\|\bar{\theta}_1 - \theta^\star\|_\infty}{N^p} \cdot \frac{\log^p((8D/\delta) \cdot \log N)}{(1-\gamma)^{2p}},$$

and the second term by

$$c_2 \cdot \frac{(1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma})^{3q+1}}{(1-\gamma)^{1.5}\sqrt{N}} \cdot \frac{\log^{2q}((8D/\delta) \cdot \log N)}{N^{3q/2}(1-\gamma)^{\frac{9q}{2}}\Delta^{3q}},$$

where $q = \frac{2}{3}p - \frac{1}{3}$, and (c_1, c_2) are universal constants only depending on (p, q). The number of samples satisfies $N \gtrsim \frac{(1+\|r\|_\infty+\sigma_r\sqrt{1-\gamma})^2}{\Delta^2(1-\gamma)^{3+\beta}}$ by assumption, and consequently, the two terms can be made arbitrarily small by increasing (p, q) appropriately. The equation (3.39) displays the bound for the pair $(p, q) = (2, 1)$.

The only remaining detail is to prove Lemma 11.

Proof of Lemma 11

Recall that the update within an epoch takes the form (cf. SingleEpoch)

$$\theta_{k+1} = (1-\alpha_k)\theta_k + \alpha_k \left\{\widehat{\mathbf{T}}_k(\mathbb{Q}) - \widehat{\mathbf{T}}_k(\bar{\theta}_m) + \overline{\mathbf{T}}_{N_m}(\bar{\theta}_m)\right\},$$

where $\bar{\theta}_m$ represents the input into epoch m. We define the shifted operators and noisy shifted operators for epoch m by

$$\mathbf{J}(\theta) = \mathbf{T}(\theta) - \mathbf{T}(\bar{\theta}_m) + \overline{\mathbf{T}}_{N_m}(\bar{\theta}_m) \quad \text{and} \quad \widehat{\mathbf{J}}_k(\mathbb{Q}) = \widehat{\mathbf{T}}_k(\mathbb{Q}) - \widehat{\mathbf{T}}_k(\bar{\theta}_m) + \overline{\mathbf{T}}_{N_m}(\bar{\theta}_m). \tag{3.40}$$

Since both of the operators $\mathbf{T}$ and $\widehat{\mathbf{T}}_k$ are γ-contractive in the ℓ_∞-norm, the operators $\mathbf{J}$ and $\widehat{\mathbf{J}}_k$ are also γ-contractive operators in the same norm. Let $\widehat{\theta}_m$ denote the unique fixed point of the operator $\mathbf{J}$. The roadmap of the proof is to show that at the end of

epoch m, the estimate θ_{K+1} is close to the fixed point $\widehat{\theta}_m$ for sufficiently large value of the epoch length K and that the fixed point $\widehat{\theta}_m$ is closer to $\theta^\star$ than the epoch initialization $\bar{\theta}_m$ for sufficiently large N_m.

The proof of Lemma 11 relies on two auxiliary lemmas that formalize this intuition. Lemma 12 characterizes the progress of Algorithm VR-QL within an epoch, and Lemma 13 addresses the progress of Algorithm VR-QL over the epochs.

Lemma 12. *Given an epoch length K lower bounded as $K \geq c_2 \frac{\log(ND/\delta)}{(1-\gamma)^3}$, we have*

$$\|\theta_{K+1} - \widehat{\theta}_m\|_\infty \leq \frac{1}{33}\|\bar{\theta}_m - \theta^\star\|_\infty + \frac{1}{33}\|\widehat{\theta}_m - \theta^\star\|_\infty,$$

with probability exceeding $1 - \frac{\delta}{2M}$.

Lemma 12 is borrowed from the paper [Kha+20a]; see the proof of Lemma 2 in that paper for details.

Our next lemma bounds the difference between the epoch fixed point $\widehat{\theta}_m$ and the optimal value function $\theta^\star$.

Lemma 13. *Assume that N_m satisfies the bound $N_m \geq c \log_4(8DM/\delta) \frac{(1+\|r\|_\infty + \sigma_r\sqrt{1-\gamma})^2}{\Delta^2(1-\gamma)^3}$. Then we have*

$$\|\widehat{\theta}_m - \theta^\star\|_\infty \leq \frac{\|\bar{\theta}_m - \theta^\star\|_\infty}{33} + c_4 \left\{ \frac{\tau^*}{\sqrt{N_m}} + \frac{\kappa}{N_m} \right\},$$

with probability exceeding $1 - \frac{\delta}{2M}$.

See Section 3.8 for a proof of this lemma.

With these two auxiliary results in hand, completing the proof of Lemma 11 is relatively straightforward. By the triangle inequality, we have

$$\begin{aligned}
\|\bar{\theta}_{m+1} - \theta^\star\|_\infty \equiv \|\theta_{K+1} - \theta^\star\|_\infty &\leq \|\theta_{K+1} - \widehat{\theta}_m\|_\infty + \|\widehat{\theta}_m - \theta^\star\|_\infty \\
&\overset{(i)}{\leq} \left\{ \frac{1}{32}\|\bar{\theta}_m - \theta^\star\|_\infty + \frac{1}{32}\|\widehat{\theta}_m - \theta^\star\|_\infty \right\} + \|\widehat{\theta}_m - \theta^\star\|_\infty \\
&= \frac{1}{32}\|\bar{\theta}_m - \theta^\star\|_\infty + \frac{33}{32}\|\widehat{\theta}_m - \theta^\star\|_\infty \\
&\overset{(ii)}{\leq} \frac{1}{32}\|\bar{\theta}_m - \theta^\star\|_\infty + \frac{1}{32}\left\{ \|\bar{\theta}_m - \theta^\star\|_\infty \right\} + c\left(\frac{\tau^*}{\sqrt{N_m}} + \frac{\kappa}{N_m} \right) \\
&. \leq \frac{1}{16}\|\bar{\theta}_m - \theta^\star\|_\infty + + c\left(\frac{\tau^*}{\sqrt{N_m}} + \frac{\kappa}{N_m} \right). \qquad (3.41)
\end{aligned}$$

Here inequality (i) follows from Lemma 12, whereas inequality (ii) follows from Lemma 13. Finally, the two bounds hold jointly with probability at least $1 - \frac{\delta}{M}$ via a union bound.

Proof of Theorem 5(b)

This argument follows the same structure as the proof of part (a) of Theorem 5; we retain the same shorthands from equation (3.36). Our proof uses Lemma 10 along with the following modification of Lemma 11.

Lemma 14. *Under the conditions of Theorem 5(b), for epochs m such that the re-centering sample size N_m satisfies the bound $N_m \geq \log_4(8DM/\delta)\frac{L^2(1+\|r\|_\infty+\sigma_r\sqrt{1-\gamma})^2}{(1-\gamma)^5}$, we have*

$$\|\bar{\theta}_{m+1} - \theta^\star\|_\infty \leq \frac{\|\bar{\theta}_m - \theta^\star\|_\infty}{16} + c \cdot \left(\frac{\tau^*}{\sqrt{N_m}} + \frac{\kappa}{N_m}\right), \tag{3.42}$$

with probability at least $1 - \frac{\delta}{M}$.

See Section 3.8 for a proof of this lemma.

Observe that Lemma 14 holds for all epochs $m \geq m^* := \log_2 \frac{L(1+\|r\|_\infty+\sigma_r\sqrt{1-\gamma})}{(1-\gamma)^{3/2}}$. Invoking Lemma 10 for all $m < m^*$ and Lemma 14 for all epochs $m \geq m^*$, and doing a calculation similar to the proof of part (a), yields

$$\|\bar{\theta}_{M+1} - \theta^\star\|_\infty \leq \frac{\|\bar{\theta}_1 - \theta^\star\|_\infty}{16^M} + c \cdot \left(\frac{\tau_{\max}}{8^{M-m^*}\sqrt{N_M}} + \frac{\tau^*}{\sqrt{N_M}} + \frac{\kappa}{N_M}\right),$$

with probability exceeding $1 - \delta$. Finally, our choice of the epoch size N_m (cf. definition (3.22a)) ensures $2N_M \leq N \leq \frac{8}{3}N_M$, and substituting the values of the triple (N_m, m^*, M) we conclude that

$$\begin{aligned}
\|\bar{\theta}_{M+1} - \theta^\star\|_\infty \leq\; & c \cdot \|\bar{\theta}_1 - \theta^\star\|_\infty \cdot \frac{\log^2((8D/\delta)\cdot\log N)}{N^2(1-\gamma)^4} \\
& + c \cdot \frac{L^3(1+\|r\|_\infty+\sigma_r\sqrt{1-\gamma})^4}{(1-\gamma)^{1.5}\sqrt{N}} \cdot \frac{\log^2((8D/\delta)\cdot\log N)}{N^{1.5}(1-\gamma)^{7.5}} \\
& + c \cdot \left(\sqrt{\frac{\log_4(8DM/\delta)}{N}} \max_{\pi^\star \in \mathbf{\Pi}^\star} \|\nu(\pi^\star; \mathbf{P}, r, \gamma\,())\|_\infty \right. \\
& \qquad \left. + \frac{\log_4(8DM/\delta)}{N} \cdot \frac{\|\theta^\star\|_{\text{span}}}{1-\gamma}\right),
\end{aligned}$$

Comment on the lower-order terms: For any $p \geq 1$ by increasing our choice of N_m by a constant factor depending on p, we can bound the first term via

$$c_1 \cdot \frac{\|\bar{\theta}_1 - \theta^\star\|_\infty}{N^p} \cdot \frac{\log^p((8D/\delta)\cdot\log N)}{(1-\gamma)^{2p}},$$

and the second term by

$$c_2 \cdot \frac{L^{3q}(1+\|r\|_\infty+\sigma_r\sqrt{1-\gamma})^{3q+1}}{(1-\gamma)^{1.5}\sqrt{N}} \cdot \frac{\log^{3q/2}((8D/\delta)\cdot\log N)}{N^{3q/2}(1-\gamma)^{15q/2}},$$

with $q = \frac{2}{3}p - \frac{1}{3}$. The number of samples satisfies $N \gtrsim \frac{L^2(1+\|r\|_\infty+\sigma_r\sqrt{1-\gamma})^2}{(1-\gamma)^{5+\beta}}$ by assumption, so the two can be made arbitrarily small by increasing (p, q) appropriately. This completes the proof of Theorem 5(b).

3.5 Discussion

In this chapter, we presented an analysis of Q-learning through the instance-dependent framework in the synchronous setting. For γ-discounted MDPs with finite state space $\mathcal{X}$ and action space $\mathcal{U}$, we have proved a local non-asymptotic lower bound for estimating the Q-function dependent on a functional $\nu(\cdot; \mathbf{P}, r)$ of the MDP instance $(\mathbf{P}, r)$ that measures the variance of solving for the Q-function. In addition, we have provided an analysis of a form of variance-reduced Q-learning and obtained instance-dependent guarantees on the ℓ_∞-error for sample sizes $\frac{N}{(\log N)^2} \geq c \cdot \frac{\log(D/\delta)}{(1-\gamma)^3\Delta^{@}}$ and $\frac{N}{(\log N)^2} \geq c\frac{\log(D/\delta)}{(1-\gamma)^5}$ on Lipschitz MDPs that match the corresponding lower bound, establishing instance-optimality. We conjecture that optimality of Algorithm VR-QL still remains true for general MDPs and sample sizes $N \geq \frac{\log(D/\delta)}{(1-\gamma)^3}$, and is left for further endeavours.

3.6 Calculations for Example 1

Here we derive the bound (3.12). Letting V^* denote the value function of the optimal policy $\pi^\star$, we have

$$(\mathbf{Z}^{\pi^\star} - \mathbf{P}^{\pi^\star})\mathbb{Q} = \begin{bmatrix} | & | \\ (\mathbf{Z}_{u_1} - \mathbf{P}_{u_1})V^* & 0 \\ | & | \end{bmatrix}. \tag{3.43}$$

Letting $\mathbf{W} = (\mathcal{I}-\gamma\mathbf{P}_{u_1})^{-1}(\mathbf{Z}_{u_1}-\mathbf{P}_{u_1})\theta_{\pi^\star}$ and solving for $(\mathcal{I}-\gamma\mathbf{P}^{\pi^\star})\mathbf{Y} = \gamma(\mathbf{Z}^{\pi^\star}-\mathbf{P}^{\pi^\star})\mathbb{Q}$ gives

$$\mathbf{Y} = \gamma \cdot \begin{bmatrix} | & | \\ \mathbf{W} & \gamma\mathbf{W} \\ | & | \end{bmatrix}. \tag{3.44}$$

Thus, we have

$$\|\nu(\pi^\star; \mathbf{P}_\lambda, r_\lambda)\|_\infty := \max_{(x,u)} |\nu(\pi^\star; \mathbf{P}_\lambda, r_\lambda)(x,u)| = \max_{(x,u)} \left|\sqrt{\mathrm{var}(\mathbf{Y})(x,u)}\right| \leq c \cdot \frac{1}{(1-\gamma)^{1.5-\lambda}}$$

The second equality above follows from the definition (3.5) of the matrix $\nu(\pi^\star; \mathbf{P}_\lambda, r_\lambda)$, and the last step via some simple calculations.

3.7 Auxiliary lemmas for Theorem 4

In this section, we prove the auxiliary lemmas that are the used in the proof of Theorem 4.

Proof of Lemma 7

This proof uses standard arguments, in particular following the usual avenue of reducing estimation to testing [Bir83b; Wai19f]. For completeness, we provide the details here. We use θ and θ' to denote the optimal Q-functions for problem $\mathcal{P}$ and $\mathcal{P}'$ respectively. We first lower bound the minimax risk over $\mathcal{P}, \mathcal{P}'$ by the averaged risk as follows:

$$\inf_{\hat{\theta}_N} \max_{\mathcal{Q} \in \{\mathcal{P}, \mathcal{P}'\}} \mathbb{E}_{\mathcal{P}} \left[\|\theta - \theta(\mathcal{Q})\|_\infty \right] \geq \frac{1}{2} \left(\mathbb{E}_{\mathcal{P}^N} \left[\|\hat{\theta}_N - \theta\|_\infty \right] + \mathbb{E}_{(\mathcal{P}')^N} \left[\|\hat{\theta}_N - \theta'\|_\infty \right] \right).$$

Here $\mathcal{P}^N$ is a product measure that yields N i.i.d. samples from $\mathcal{P}$. Then, for any $\delta \geq 0$, we have by Markov's inequality

$$\mathbb{E}_{\mathcal{P}^N} \left[\|\hat{\theta}_N - \theta\|_\infty \right] + \mathbb{E}_{(\mathcal{P}')^N} \left[\|\hat{\theta}_N - \theta'\|_\infty \right] \geq \delta \left[\mathcal{P}^N \left(\|\hat{\theta}_N - \theta\|_\infty \geq \delta \right) + (\mathcal{P}')^N \left(\|\hat{\theta}_N - \theta'\|_\infty \geq \delta \right) \right].$$

Define $\delta_{01} := \frac{1}{2}\|\theta - \theta'\|_\infty$, we have

$$\|\hat{\theta}_N - \theta\|_\infty < \delta_{01} \implies \|\hat{\theta}_N - \theta'\|_\infty > \delta_{01},$$

yielding

$$\begin{aligned}
\mathbb{E}_{\mathcal{P}^N} \left[\|\hat{\theta}_N - \theta\|_\infty \right] + \mathbb{E}_{(\mathcal{P}')^N} \left[\|\hat{\theta}_N - \theta\|_\infty \right] &\geq \delta_{01} \left[1 - \mathcal{P}^N \left(\|\hat{\theta}_N - \theta\|_\infty < \delta_{01} \right) \right. \\
&\quad \left. + (\mathcal{P}')^N \left(\|\hat{\theta}_N - \theta'\|_\infty \geq \delta_{01} \right) \right] \\
&\geq \delta_{01} \left[1 - \mathcal{P}^N \left(\|\hat{\theta}_N - \theta'\|_\infty \geq \delta_{01} \right) \right. \\
&\quad \left. + (\mathcal{P}')^N \left(\|\hat{\theta}_N - \theta'\|_\infty \geq \delta_{01} \right) \right] \\
&\geq \delta_{01} \left[1 - \|\mathcal{P}^N - (\mathcal{P}')^N\|_{\mathrm{TV}} \right] \\
&\geq \delta_{01} \left[1 - \sqrt{2} d_{\mathrm{hel}}(\mathcal{P}^N, (\mathcal{P}')^N)^2 \right].
\end{aligned}$$

Via the tensorization property of the Hellinger distance for independent random variables we have

$$d_{\mathrm{hel}}(\mathcal{P}^N, (\mathcal{P}')^N)^2 = 1 - \left(1 - d_{\mathrm{hel}}(\mathcal{P}, \mathcal{P}')^2\right)^N \leq N d_{\mathrm{hel}}(\mathcal{P}, \mathcal{P}')^2.$$

Putting together the pieces, we have that

$$\inf_{\hat{\theta}_N} \max_{\mathcal{Q} \in \{\mathcal{P}, \mathcal{P}'\}} \mathbb{E}_{\mathcal{Q}} \left[\|\theta - \theta(\mathcal{Q})\|_\infty \right] \geq \frac{1}{4} \|\theta(\mathcal{P}) - \theta(\mathcal{P}')\|_\infty \cdot \left(1 - \sqrt{2} N \cdot d_{\mathrm{hel}}(\mathcal{P}, \mathcal{P}')^2\right)_+.$$

The desired result then follows from taking a supremum over all positive alternative $\mathcal{P}' \in \mathcal{S}$ and a simple calculation.

Proof of Lemma 9

We devote a subsection to each of the three parts of this lemma.

Proof of Lemma 9(a)

In order to establish that $\bar{\mathbf{P}}$ is a transition kernel, we observe that

$$\sum_{x'} \bar{\mathbf{P}}_{x',z} = 1 + \frac{1}{\tilde{\rho}(\bar{z})\sqrt{2N}} \mathbf{U}_{\bar{z},z} \cdot \left(\sum_{x'} \mathbf{P}_{x',z} (\theta(x', \pi_1(x')) - (\mathbf{P}^{\pi_1}\theta)(z)) \right) = 1,$$

where the last equality above follows by noting that $(\mathbf{P}^{\pi_1}\theta)(z) = \sum_{x'} \mathbf{P}_{x',z}\theta(x', \pi_1(x'))$. To check non-negativity of entries of $\bar{\mathbf{P}}$ note we have $|\mathbf{U}_{z,z'}| \leq \frac{1}{1-\gamma}$, and $2\|\theta\|_{\text{span}} \geq |\theta(x', \pi_1(x')) - (\mathbf{P}^{\pi_1}\theta)(z)|$. Combining the last two observation along with the sample size requirement (3.10b) implies

$$\bar{\mathbf{P}}_{x',z} \geq 1 - \frac{1}{\tilde{\rho}(\bar{z})\sqrt{2N}} \cdot \frac{\|\theta\|_{\text{span}}}{1-\gamma} \geq 0,$$

establishing that $\bar{\mathbf{P}}$ defines a valid set of transition kernels.

Proof of Lemma 9(b)

The proof of part (b) follows by first providing a general upper bound on the Hellinger distance $d_{\text{hel}}(\mathcal{P}, \mathcal{P}_1)$, and then substituting the values of instances $\mathcal{P}$ and $\mathcal{P}_1$. Concretely, we prove

$$d^2_{\text{Hel}}(\mathcal{P}, \mathcal{P}_1) \overset{\text{(a)}}{\leq} \frac{1}{2} \cdot \sum_{z,x'} \frac{(\mathbf{P}_{x',z} - \bar{\mathbf{P}}_{x',z})^2}{\mathbf{P}_{x',z}} \overset{\text{(b)}}{\leq} \frac{1}{4N}. \tag{3.45}$$

With this result in hand, the claimed bound on $\|\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}\|_{\text{op}}$ is immediate. Indeed,

$$\|\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1}\|^2_{\text{op}} \leq \sum_{z,x'} (\mathbf{P}_{x',z} - \bar{\mathbf{P}}_{x',z})^2 \leq \sum_{z,x'} \frac{(\mathbf{P}_{x',z} - \bar{\mathbf{P}}_{x',z})^2}{\mathbf{P}_{x',z}} \leq \frac{1}{2N}.$$

It remains to prove the bounds (3.45)(a) and (3.45)(b).

Proof of (3.45a): We use $(\mathbf{Z}, R)$ (respectively $(\mathbf{Z}', R')$) to denote a sample drawn from the distribution P (respectively P'), and $P_{\mathbf{Z}}, P_R$ (respectively $P'_{\mathbf{Z}}, P'_R$) to denote the marginal distribution of $\mathbf{Z}, R$ (respectively $\mathbf{Z}', R'$). By independence of $\mathbf{Z}$ and R (and likewise for $\mathbf{Z}', R'$) we have

$$P = P_{\mathbf{Z}} \otimes P_R, \quad \text{and} \quad P' = P'_{\mathbf{Z}} \otimes P'_R. \tag{3.46}$$

Let $\mathcal{P}' = (\mathbf{P}', r') \in \mathcal{S}_1$ (so $r' = r$). Via the independence between $\mathbf{Z}$ and R, we have

$$d^2_{\text{Hel}}(P, P') = d^2_{\text{Hel}}(P_{\mathbf{Z}}, P'_{\mathbf{Z}}). \tag{3.47}$$

For state-action pairs (x, u), $\mathbf{Z}(x, u)$ are independent (and likewise for $\mathbf{Z}'$) so

$$d^2_{\text{Hel}}(P_{\mathbf{Z}}, P'_{\mathbf{Z}}) = 1 - \prod_{x,u} \left(1 - d_{\text{Hel}}(P_{\mathbf{Z}(x,u)}, P_{\mathbf{Z}'(x,u)})\right)^2 \leq \sum_{x,u} d^2_{\text{Hel}}(P_{\mathbf{Z}(x,u)}, P_{\mathbf{Z}'(x,u)}).$$

Note that $\mathbf{Z}(x, u)$ and $\mathbf{Z}'(x, u)$ have multinomial distribution with parameters $\mathbf{P}_u(\cdot \mid x)$ and $\mathbf{P}'_u(\cdot \mid x)$ respectively. Therefore,

$$d^2_{\text{Hel}}(P_{\mathbf{Z}(x,u)}, P_{\mathbf{Z}'(x,u)}) \leq \frac{1}{2} D_{\chi^2}\left(P_{\mathbf{Z}'(x,u)} \| P_{\mathbf{Z}(x,u)}\right) = \frac{1}{2} \sum_{x'} \frac{(\mathbf{P}_{x',z} - \bar{\mathbf{P}}_{x',z})^2}{\mathbf{P}_{x',z}}.$$

Proof of (3.45b): We have

$$\begin{aligned}
\sum_{z,x'} \frac{(\mathbf{P}_{x',z} - \bar{\mathbf{P}}_{x',z})^2}{\mathbf{P}_{x',z}} &= \frac{1}{2N\tilde{\rho}^2(\bar{z})} \sum_z \sum_{x'} \mathbf{P}_{x',z} (\mathbf{U}_{\bar{z},z})^2 (\theta(x', \pi_1(x')) - (\mathbf{P}^{\pi_1}\theta)(z))^2 \\
&= \frac{1}{2N\tilde{\rho}^2(\bar{z})} \sum_z (\mathbf{U}_{\bar{z},z})^2 \cdot \left(\sum_{x'} \mathbf{P}_{x',z} (\theta(x', \pi_1(x')) - (\mathbf{P}^{\pi_1}\theta)(z))^2 \right) \\
&\overset{(i)}{=} \frac{1}{2N\tilde{\rho}^2(\bar{z})} \sum_z (\mathbf{U}_{\bar{z},z})^2 \varphi^2(z) \\
&\overset{(ii)}{=} \frac{1}{2N},
\end{aligned}$$

Equality (i) follows from the definition

$$\varphi^2(z) = \text{Var}\left(\mathbf{Z}^{\pi_1}\theta(z)\right) = \sum_{x'} \mathbf{P}_{x',z} (\theta(x', \pi_1(x')) - (\mathbf{P}^{\pi_1}\theta)(z))^2, \tag{3.48}$$

whereas equality (ii) follows from the definition (3.26), which ensures that

$$\tilde{\rho}^2(\bar{z}) = \text{Var}\left((\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1} \mathbf{Z}^{\pi_1}\theta(\bar{z})\right) = \sum_{z'} (\mathbf{U}_{z,z'})^2 \varphi^2(z').$$

Proof of Lemma 9(c)

The entries of the matrix $\mathbf{U} := (\mathcal{I} - \gamma \mathbf{P}^{\pi_1})^{-1}$ are positive, so that it suffices to show that the vector $(\bar{\mathbf{P}}^{\pi} - \mathbf{P}^{\pi})\theta$ is entry-wise positive. We have

$$\begin{aligned}
(\bar{\mathbf{P}}^{\pi_1} - \mathbf{P}^{\pi_1})\theta(z) &= \sum_{x'} (\bar{\mathbf{P}}_{x',z} - \mathbf{P}_{x',z}) \theta(x', \pi_1(x')) \\
&= \sum_{x'} (\bar{\mathbf{P}}_{x',z} - \mathbf{P}_{x',z}) (\theta(x', \pi_1(x')) - (\mathbf{P}^{\pi_1}\theta)(z)) \\
&= \frac{1}{\tilde{\rho}(\bar{z})\sqrt{2N}} \mathbf{U}_{\bar{z},z} \sum_{x'} \mathbf{P}_{x',z} (\theta(x', \pi_1(x')) - (\mathbf{P}^{\pi_1}\theta)(z))^2 \\
&= \frac{1}{\tilde{\rho}(\bar{z})\sqrt{2N}} \mathbf{U}_{\bar{z},z} \varphi^2(z) \geq 0,
\end{aligned}$$

where the second equality follows from the fact that $\sum_{x'} \bar{\mathbf{P}}_{x',z} = \sum_{x'} \mathbf{P}_{x',z} = 1$, the third equality follows by substituting the value of $\bar{\mathbf{P}}$ from equation (3.28), and the equality in the last line follows from the definition (3.48). This completes the proof of part (c).

3.8 Auxiliary lemmas for Theorem 5

In this section, we prove the auxiliary lemmas that are used in the proof of Theorem 5.

Proof of Lemma 11

This section is devoted to the proof of Lemma 11 which underlies the proof of Theorem 5. In order to simplify notation, we drop the epoch number m from $\widehat{\theta}_m$ and $\bar{\theta}_m$ throughout the remainder of the proof. Let $\widehat{\pi}$ and $\pi^\star$ denote the greedy policies with respect to the Q functions $\widehat{\theta}$ and $\theta^\star$, respectively. Concretely,

$$\pi^\star(x) = \arg\max_{u \in \mathcal{U}} \ \theta^\star(x,u) \qquad \widehat{\pi}(x) = \arg\max_{u \in \mathcal{U}} \ \widehat{\theta}(x,u). \tag{3.49}$$

Ties in the arg max are broken by choosing the action u with smallest index.

By the optimality of the policies $\widehat{\pi}$ and $\pi^\star$ for the Q-functions $\widehat{\theta}$ and $\theta^\star$, respectively, we have

$$\theta^\star = r + \gamma \mathbf{P}^{\pi^\star}\theta^\star \quad \text{and} \quad \widehat{\theta} = \widetilde{r} + \gamma \mathbf{P}^{\widehat{\pi}}\widehat{\theta}, \quad \text{where} \quad \widetilde{r} := r + \overline{\mathbf{T}}_{N_m}(\bar{\theta}) - \mathbf{T}(\bar{\theta}). \tag{3.50}$$

In order to complete the proof, we require the following auxiliary result.

Lemma 15. *We have*

$$\|(\mathcal{I} - \gamma \mathbf{P}^{\pi^\star})^{-1}(\widetilde{r} - r)\|_\infty \leq \frac{\|\bar{\theta} - \theta^\star\|_\infty}{33} + 4 \cdot \sqrt{\frac{\log_4(8DM/\delta)}{N_m}} \cdot \max_{\pi \in \mathbf{\Pi}^\star} \|\nu(\pi; \mathbf{P}, r)\|_\infty + 4 \cdot \frac{\log_4(8DM/\delta)}{N_m} \cdot \frac{\|\theta^\star\|_{\text{span}}}{(1-\gamma)},$$

with probability exceeding $1 - \frac{\delta}{8M}$.

See Section 3.8 for the proof.

It remains to prove that under the assumptions of Lemma 11, the following bound holds with probability $1 - \frac{\delta}{M}$:

$$\|\widehat{\theta} - \bar{\theta}\|_\infty \leq \|(\mathcal{I} - \gamma \mathbf{P}^{\pi^\star})^{-1}(\widetilde{r} - r)\|_\infty. \tag{3.51}$$

We establish this claim by first showing that the policy $\widehat{\pi}$ is an optimal policy, which is achieved in the following lemma.

Lemma 16. *Given two Q-functions $\theta^\star$ and $\widehat{\theta}$ and the associated optimal policies $\pi^\star$ and $\widehat{\pi}$, we have*

$$\mathbf{P}^{\widehat{\pi}}\theta^\star(x,u) \geq \mathbf{P}^{\pi^\star}\theta^\star(x,u) - 2\|\widehat{\theta} - \theta^\star\|_\infty \quad \textit{for all} \quad (x,u) \in \mathcal{X} \times \mathcal{U}.$$

Moreover, if the batch size satisfies the lower bound $N_m \geq c_3 \frac{(1+\|r\|_\infty + \sigma_r\sqrt{1-\gamma})^2}{(1-\gamma)^3} \cdot \frac{\log(DM^2/\delta)}{\Delta^2}$, then $\widehat{\pi}$ is an optimal policy with probability at least $1 - \frac{\delta}{M}$. Hence, under the unique optimal policy (UNQ) condition or Lipschitz (LIP) condition, we have $\mathbf{P}^{\widehat{\pi}} = \mathbf{P}^{\pi^\star}$.

We prove this lemma in Section 3.8. In order to prove the bound (3.51), it suffices to prove the following elementwise inequalities:

$$\theta^\star - \widehat{\theta} \preceq (\mathcal{I} - \gamma\mathbf{P}^{\pi^\star})^{-1}(r - \widetilde{r}) \tag{3.52a}$$

$$\widehat{\theta} - \theta^\star \preceq (\mathcal{I} - \gamma\mathbf{P}^{\widehat{\pi}})^{-1}(r - \widetilde{r}) \tag{3.52b}$$

Indeed, we have

$$\left|\theta^\star - \widehat{\theta}\right| \preceq \max\{\left|\theta^\star - \widehat{\theta}\right|_+, \left|\widehat{\theta} - \theta^\star\right|_+\} \preceq \left|(\mathcal{I} - \gamma\mathbf{P}^{\pi^\star})^{-1}(r - \widetilde{r})\right|,$$

where the maximum operator $\max\{\cdot,\cdot\}$ is applied entry-wise. Combining the last two bounds with Lemma 16, and using the lower bound assumption on the epoch sample size N_m we obtain

$$\begin{aligned}\left|\theta^\star - \widehat{\theta}\right| \preceq \max\{\left|\theta^\star - \widehat{\theta}\right|_+, \left|\widehat{\theta} - \theta^\star\right|_+\} &\preceq \left|(\mathcal{I} - \gamma\mathbf{P}^{\pi^\star})^{-1}(r - \widetilde{r}), (\mathcal{I} - \gamma\mathbf{P}^{\widehat{\pi}})^{-1}(r - \widetilde{r})\right| \\ &\preceq \left|(\mathcal{I} - \gamma\mathbf{P}^{\pi^\star})^{-1}(r - \widetilde{r})\right|,\end{aligned}$$

where the last inequality uses $\mathbf{P}^{\pi^\star} = \mathbf{P}^{\widehat{\pi}}$ (cf. Lemma 16). It remains to prove the bounds (3.52a) and (3.52b).

Proof of bounds (3.52a) and (3.52b): By optimality of policies $\widehat{\pi}$ and $\pi^\star$ for Q-functions $\widehat{\theta}$ and $\theta^\star$, respectively, we have

$$\theta^\star = r + \gamma\mathbf{P}^{\pi^\star}\theta^\star \succeq r + \gamma\mathbf{P}^{\pi^\star}\widehat{\theta} \quad \text{and} \quad \widehat{\theta} = \widetilde{r} + \gamma\mathbf{P}^{\widehat{\pi}}\widehat{\theta} \succeq \widetilde{r} + \gamma\mathbf{P}^{\widehat{\pi}}\theta^\star. \tag{3.53}$$

This yields:

$$\theta^\star - \widehat{\theta} = r - \widetilde{r} + \gamma(\mathbf{P}^{\pi^\star}\theta^\star - \mathbf{P}^{\widehat{\pi}}\widehat{\theta}) \preceq r - \widetilde{r} + \gamma\mathbf{P}^{\pi^\star}(\theta^\star - \widehat{\theta}). \tag{3.54}$$

Rearranging the last inequality, and using the non-negativity of the entries of $(\mathcal{I} - \gamma\mathbf{P}^{\pi^\star})^{-1}$ we conclude

$$(\theta^\star - \widehat{\theta}) \preceq (\mathcal{I} - \gamma\mathbf{P}^{\pi^\star})^{-1}(r - \widetilde{r}).$$

This completes the proof of the bound (3.52a). The proof of bound (3.52b) is similar.

Proof of Lemma 15

Recall the definition $\widetilde{r} := \widehat{R} + \gamma(\widehat{\mathbf{Z}}^{\overline{\pi}} - \mathbf{P}^{\overline{\pi}})\bar{\theta}$, where $\overline{\pi}$ a policy greedy with respect to $\bar{\theta}$; that is, $\overline{\pi}(x) = \arg\max_{u \in \mathcal{U}} \bar{\theta}(x,u)$, where we break ties in the arg max by choosing the action u with smallest index. We have

$$\begin{aligned}\|(\mathcal{I} - \gamma \mathbf{P}^{\pi^\star})^{-1}(\widetilde{r} - r)\|_\infty \leq \|(\mathcal{I} - \gamma \mathbf{P}^{\pi^\star})^{-1}\left\{(\widehat{R} - r) + \gamma(\widehat{\mathbf{Z}}^{\pi^\star} - \mathbf{P}^{\pi^\star})\theta^\star\right\}\|_\infty \\ + \gamma\|(\mathcal{I} - \gamma \mathbf{P}^{\pi^\star})^{-1}\left\{(\widehat{\mathbf{Z}}^{\overline{\pi}}\bar{\theta} - \widehat{\mathbf{Z}}^{\pi^\star}\theta^\star) - (\mathbf{P}^{\overline{\pi}}\bar{\theta} - \mathbf{P}^{\pi^\star}\theta^\star)\right\}\|_\infty.\end{aligned}$$

Observe that the random variable $\widehat{R}$ and $\widehat{\mathbf{Z}}$ are averages of N_m i.i.d. random variables $\{R_i\}$ and $\{\widehat{\mathbf{Z}}_i\}$, respectively. Additionally, entrywise, the random reward is a Gaussian random variable with variance σ_r^2, and by the generative model assumption, the randomness in the random rewards $\{R_i\}$ is independent of the randomness in $\{\widehat{\mathbf{Z}}_i\}$. Consequently, applying Hoeffding's bound on the term involving $\{R_i\}$, a Bernstein bound on the term involving $\{\widehat{\mathbf{Z}}_i\}$ and a union bound yields the following result which holds with probability at least $1 - \frac{\delta}{4M}$:

$$\begin{aligned}&\|(\mathcal{I} - \gamma \mathbf{P}^{\pi^\star})^{-1}\left\{(\widehat{R} - r) + \gamma(\widehat{\mathbf{Z}}^{\pi^\star} - \mathbf{P}^{\pi^\star})\theta^\star\right\}\|_\infty \\ &\quad\leq \frac{4}{\sqrt{N_m}} \cdot \|\nu(\pi^\star; \mathbf{P}, r)\|_\infty \cdot \sqrt{\log_4(8DM/\delta)} + \frac{4\|\theta^\star\|_{\text{span}}}{(1-\gamma)N_m} \cdot \log_4(8DM/\delta) \\ &\quad\leq \frac{4}{\sqrt{N_m}} \cdot \max_{\pi \in \mathbf{\Pi}^\star} \|\nu(\pi; \mathbf{P}, r)\|_\infty \cdot \sqrt{\log_4(8DM/\delta)} + \frac{4\|\theta^\star\|_{\text{span}}}{(1-\gamma)N_m} \cdot \log_4(8DM/\delta).\end{aligned}$$

Finally, for each state-action pair (x,u) the random variable $(\widehat{\mathbf{Z}}^{\overline{\pi}}\bar{\theta} - \widehat{\mathbf{Z}}^{\pi^\star}\theta^\star)(x,u)$ has expectation $(\mathbf{P}^{\overline{\pi}}\bar{\theta} - \mathbf{P}^{\pi^\star})(x,u)$ with entries uniformly bounded by $2\|\bar{\theta} - \theta^\star\|_\infty$. Consequently, by a standard application of Hoeffding's inequality combined with the lower bound $N_m \geq c_3 \frac{4^m}{(1-\gamma)^2}\log_4(8DM/\delta)$, we have

$$\frac{\gamma}{1-\gamma} \cdot \|(\widehat{\mathbf{Z}}^{\overline{\pi}}\bar{\theta} - \widehat{\mathbf{Z}}^{\pi^\star}\theta^\star) - (\mathbf{P}^{\overline{\pi}}\bar{\theta} - \mathbf{P}^{\pi^\star}\theta^\star)\|_\infty \leq \frac{\|\bar{\theta} - \theta^\star\|_\infty}{33},$$

with probability exceeding $1 - \frac{\delta}{4M}$. The statement then follows from combining these two high-probability statements with a union bound.

Proof of Lemma 14

By Lemma 12 and Lemma 15, it suffices to show

$$\|\widehat{\theta}_m - \theta^\star_m\|_\infty \leq \frac{1}{2}\|(\mathcal{I} - \gamma \mathbf{P}^{\pi^\star})^{-1}(\widetilde{r} - r)\|_\infty$$

Recalling the bounds (3.52a)–(3.52b), we have

$$\begin{aligned}\|\widehat{\theta}_m - \theta^\star\|_\infty &\leq \|(\mathcal{I} - \gamma^{\pi^\star})^{-1}(\widetilde{r} - r)\|_\infty + \gamma\|(\mathcal{I} - \gamma\mathbf{P}^{\pi^\star})^{-1}(\mathbf{P}^{\widehat{\pi}} - \mathbf{P}^{\pi^\star})(\widehat{\theta}_m - \theta^\star)\|_\infty \\ &\leq \|(\mathcal{I} - \gamma\mathbf{P}^{\pi^\star})^{-1}(\widetilde{r} - r)\|_\infty + \frac{L\gamma}{1-\gamma}\|\widehat{\theta}_m - \theta^\star\|_\infty^2\end{aligned}$$

where the last inequality uses the Lipschitz condition (3.17). If we can show $\frac{L\gamma}{1-\gamma}\|\widehat{\theta}_m - \theta^\star\|_\infty \leq \frac{1}{2}$, we are done. In order to do so, we require the following auxiliary result:

Lemma 17. *Given a batch size N_m lower bounded as $N_m \geq c_3 \frac{\log_4(8DM/\delta)}{(1-\gamma)^2}$, we have*

$$\|\widehat{\theta}_m - \theta^\star\|_\infty \leq c_1 \cdot \frac{1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma}}{\sqrt{N_m}(1-\gamma)^{1.5}} \cdot \log_4(8DM^2/\delta)$$

with probability at least $1 - \frac{\delta}{4M}$.

With the above lemma at hand and using $N_m \geq c\log_4(8DM/\delta)\frac{L^2(1+\|r\|_\infty+\sigma_r\sqrt{1-\gamma})^2}{(1-\gamma)^5}$ we conclude

$$\|\widehat{\theta}_m - \theta^\star\|_\infty \leq \frac{1-\gamma}{2L},$$

as desired. It remains to prove Lemma 17.

Proof of Lemma 17: This proof exploits the result of Lemma 10, that with probability at least $1 - \frac{\delta}{M^2}$, we have

$$\begin{aligned}\|\widehat{\theta}_m - \theta^\star\|_\infty \leq \frac{\|\bar{\theta}_m - \theta^\star\|_\infty}{33} + \frac{1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma}}{(1-\gamma)^{1.5}}\sqrt{\frac{\log(8DM^2/\delta)}{N_m}} \\ + \frac{\log(8DM^2/\delta)}{N_m} \cdot \frac{\|\theta^\star\|_{\text{span}}}{1-\gamma}. \quad (3.55)\end{aligned}$$

Following an argument similar to the proof of Theorem 5, we have

$$\begin{aligned}
\|\bar{\theta}_{m+1} - \theta^\star\|_\infty &\leq \frac{\|\bar{\theta}_m - \theta^\star\|_\infty}{16} + c\left\{\frac{1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma}}{(1-\gamma)^{1.5}}\sqrt{\frac{\log(8DM^2/\delta)}{N_m}}\right. \\
&\left. + \frac{\log(8DM^2/\delta)}{N_m} \cdot \frac{\|\theta^\star\|_{\text{span}}}{1-\gamma}\right\} \\
&\overset{(i)}{\leq} \frac{\|\bar{\theta}_1 - \theta^\star\|_\infty}{16^m} + 2c\left\{\frac{1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma}}{(1-\gamma)^{1.5}}\sqrt{\frac{\log(8DM^2/\delta)}{N_m}}\right. \\
&\left. + \frac{\log(8DM^2/\delta)}{N_m} \cdot \frac{\|\theta^\star\|_{\text{span}}}{1-\gamma}\right\} \\
&\overset{(ii)}{\leq} \frac{\|r\|_\infty}{\sqrt{1-\gamma}} \cdot \frac{1}{(1-\gamma)\sqrt{N_m}} \cdot \frac{1}{4^m} \\
&+ 2c\left\{\frac{1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma}}{(1-\gamma)^{1.5}}\sqrt{\frac{\log(8DM^2/\delta)}{N_m}}\right. \\
&\left. + \frac{\log(8DM^2/\delta)}{N_m} \cdot \frac{\|\theta^\star\|_{\text{span}}}{1-\gamma}\right\},
\end{aligned} \tag{3.56}$$

with probability at least $1 - \frac{\delta}{4M}$. Inequality (ii) follows by recursing the first inequality; the last inequality uses the initialization condition $\|\bar{\theta}_1 - \theta^\star\|_\infty \leq \frac{\|r\|_\infty}{\sqrt{1-\gamma}}$, and $N_m \geq \frac{4^m}{(1-\gamma)^2}$. Combining the bounds (3.55) and (3.56) and using the bounds $\|\theta^\star\|_\infty \leq \frac{\|r\|_\infty}{1-\gamma}$ and $\|\theta^\star\|_{\text{span}} \leq 2\|\theta^\star\|_\infty$, we find that

$$\|\widehat{\theta}_m - \theta^\star\|_\infty \leq 8c \cdot \frac{1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma}}{\sqrt{N_m}(1-\gamma)^{1.5}} \cdot \log(8DM^2/\delta),$$

with probability at least $1 - \frac{\delta}{4M}$. This completes the proof.

Proof of Lemma 16

The proof of this lemma exploits the optimality of the policies $\pi^\star$ and $\widehat{\pi}$ with respect to the Q-functions $\theta^\star$ and $\widehat{\theta}$, respectively. Accordingly, we have for all state action pair $(x, u) \in \mathcal{X} \times \mathcal{U}$

$$\begin{aligned}
\mathbf{P}^{\widehat{\pi}}\theta^\star(x,u) &= \mathbf{P}^{\widehat{\pi}}\widehat{\theta}(x,u) + \mathbf{P}^{\widehat{\pi}}\theta^\star(x,u) - \mathbf{P}^{\widehat{\pi}}\widehat{\theta}(x,u) \\
&\geq \mathbf{P}^{\pi^\star}\widehat{\theta}(x,u) - \|\theta^\star - \widehat{\theta}\|_\infty \\
&= \mathbf{P}^{\pi^\star}\theta^\star(x,u) + \mathbf{P}^{\pi^\star}\widehat{\theta}(x,u) - \mathbf{P}^{\pi^\star}\theta^\star(x,u) - \|\theta^\star - \widehat{\theta}\|_\infty \\
&\geq \mathbf{P}^{\pi^\star}\theta^\star(x,u) - 2\|\theta^\star - \widehat{\theta}\|_\infty.
\end{aligned} \tag{3.57}$$

The first inequality follows from the optimality of the policy $\widehat{\pi}$ with respect to the Q-function $\widehat{\theta}$. This completes the proof of the first part of the lemma.

Turning to the second part, invoking Lemma 17 with a batch size $N_m \geq \frac{(1+\|r\|_\infty+\sigma_r\sqrt{1-\gamma})^2}{(1-\gamma)^3} \cdot \frac{\log(DM^2/\delta)}{\Delta^2}$ guarantees that

$$2\|\theta^\star - \widehat{\theta}\|_\infty < \Delta.$$

This inequality, combined with the bound (3.57) and the definition of the optimality gap Δ, implies that $\widehat{\pi}$ is an optimal policy, and hence $\mathbf{P}^{\widehat{\pi}} = \mathbf{P}^{\pi^\star}$ under the unique policy or Lipschitz assumptions.

Chapter 4

Optimal variance-reduced stochastic approximation in Banach spaces

In Chapters 2 and 3 we derived istance optimal bounds for the policy evaluation and the policy optimization problem. It turns out that both the problems are special cases of fixed point problems, in this chapter we study the problem of fixed point of an operator in presence of noisy data. Concretely, we study the problem of estimating the fixed point of a contractive operator defined on a separable Banach space. Focusing on a stochastic query model that provides noisy evaluations of the operator, we analyze a variance-reduced stochastic approximation scheme, and establish non-asymptotic bounds for both the operator defect and the estimation error, measured in an arbitrary semi-norm. In contrast to worst-case guarantees, our bounds are instance-dependent, and achieve the local asymptotic minimax risk non-asymptotically. For linear operators, contractivity can be relaxed to multi-step contractivity, so that the theory can be applied to problems like average reward policy evaluation problem in reinforcement learning. We illustrate the theory via applications to stochastic shortest path problems, two-player zero-sum Markov games, as well as policy evaluation and Q-learning for tabular Markov decision processes.

4.1 Introduction

In this chapter, we consider a class of stochastic fixed-point problems defined in Banach spaces. In particular, let $\mathbb{V}$ be a separable Banach space with its associated norm $\|\cdot\|$, and suppose that $\mathbf{h} : \mathbb{V} \to \mathbb{V}$ is an operator on the Banach space. Of interest to us are solutions θ^* to the fixed-point equation

$$\theta^* = \mathbf{h}(\theta^*). \tag{4.1}$$

When $\mathbf{h}$ is contractive, the Banach fixed point theorem (e.g., [DG03]) ensures the existence and uniqueness of the fixed point. The bulk of our analysis focuses on this contractive case, but we also allow for weaker multi-stage contraction in certain settings.

Fixed points of this type lie at the core of many mathematical areas, including differential and integral equations [Kir11; Tes12], game theory [Sto89], optimization and variational inequalities [Nes03; RW09], as well as dynamic programming and reinforcement learning [Ber12a; Ber19]. In these settings, the contraction property not only plays an instrumental role in existence and uniqueness proofs, but also leads to efficient methods for computing fixed points. Our focus will be on the extension of such methods to problems in which the operator $\mathbf{h}$ can be observed only via a stochastic oracle that, when given a query point θ, returns a noisy version of the operator evaluation $\mathbf{h}(\theta)$. Such random observation models necessitate the use of stochastic approximation schemes. A fundamental question associated with such schemes is their *statistical complexity*: how many noisy operator evaluations are required to estimate the fixed point θ^* to a pre-specified accuracy? In this chapter, we undertake a fine-grained analysis of this question in a general setting. Our analysis captures the way in which statistical complexity depends on the geometry of the Banach space, as well as the structure of the fixed point θ^* itself.

An important sub-class of Banach spaces are Hilbert spaces, with the Euclidean case ($\mathbb{V} = \mathbb{R}^d$ with the usual inner product) being one special example. The behavior of stochastic approximation for many Hilbert spaces is relatively well understood. In this case, the space $\mathbb{V}$ is endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathbb{V}}$ that induces the norm $\|x\| = \sqrt{\langle x, x \rangle_{\mathbb{V}}}$. For example, for the Euclidean space $(\mathbb{R}^d, \| \cdot \|_2)$, if we set $\mathbf{h}(x) := x - \beta^{-1} \nabla f(x)$ for a β-smooth and strongly convex function f, then solving the fixed-point equation (4.1) is equivalent to minimizing the function f. A rich theory has been developed around this stochastic optimization problem [BCN18; Nem+09b], giving rise to the concepts of averaging [PJ92; Rup88b], acceleration [GL12; GL13], and variance reduction [JZ13; Li+20; NST21]. Optimal complexities have been established for both the computational and statistical aspects of this problem [DR16; Mou+20; MB11].

Much less is known, however, for the case of general Banach spaces. We note that non-Euclidean geometry has been studied in the literature on stochastic optimization and stochastic variational inequalities, with the method of mirror descent being a representative example [JNT11; KLL20; NY83]. Our study, however, deviates from this line of research. The difference can be observed from the formulation of the problem: the operator $\mathbf{h}$ in equation (4.1) is a mapping from $\mathbb{V}$ to itself, whereas the operators studied in variational inequalities map a Banach space to its dual. This difference requires novel methods in the analysis of our algorithms.

From a superficial perspective, it might seem that non-Euclidean geometry poses little difficulty for stochastic approximation: all norms are equivalent in the finite-dimensional case and their convergence depends ultimately on the limiting

ODE [Bor09]. From a non-asymptotic point of view, however, the picture is more nuanced: a natural desideratum is that the bounds depend on the *geometric complexity* of $\mathbb{V}$, as opposed to its (possibly much larger) ambient dimension. Our theory makes explicit how this geometric complexity should be measured, and how it differs from the ambient dimension. This kind of dependence is important, as in many real-world problems, the ambient dimensionality is substantially larger than the geometric complexity. As one concrete example, when solving fixed-point equations that arise in tabular Markov decision processes, the ambient dimension is the size of state-action space, whereas one can obtain ℓ_∞-norm bounds that have only logarithmic dependency on the dimension (see, e.g., [Kha+20b; Wai19c]). Our first goal, therefore, is to develop a unified, dimension-free theory for a certain class of stochastic approximation procedures in Banach spaces.

Our second goal is to establish bounds that are instance-dependent, and so move us beyond a classical worst-case analysis. Local asymptotic minimax theory establishes lower risk bounds over a small neighborhood of a given problem instance [Häj72; van00]. These lower bounds are universal, applicable to any estimator and valid for any bowl-shaped loss function. On the other hand, under certain regularity conditions on the space, a sum of i.i.d. random variables in Banach spaces is known to satisfy a central limit theorem (see Ledoux and Talagrand [LT13], Section 10). These two lines of classical asymptotic analysis, in conjunction, indicate that the "right" complexity for estimation in a Banach space $\mathbb{V}$ should depend on the expected norm of a Gaussian random element with a suitable covariance structure. Given this fundamental limit, a natural statistical challenge is to construct an estimator whose *non-asymptotic* risk matches this quantity, with possible higher-order terms which, again, depends only the geometric complexity of the norm $\|\cdot\|$ (and not the ambient dimension).

In order to address these goals, we analyze an extension of the ROOT-SGD algorithm, a stochastic approximation (SA) algorithm introduced in past work involving a subset of the current authors [Li+20]. We adapt the scheme to solve general fixed-point problems and establish instance-dependent non-asymptotic guarantees in general Banach spaces. More specifically, our main contributions are as follows:

- We establish sharp non-asymptotic bounds on the *operator defect*, $\|\mathbf{h}(\theta_n) - \theta_n\|$, of the iterate θ_n after n rounds. The leading-order term is defined in terms of a Gaussian complexity induced by the noisy evaluations of the operator $\mathbf{h}$, and the leading-order term in the bound matches the expected norm of the optimal Gaussian limit. The high-order term depends on the Dudley chaining integral of the dual ball, measured under a metric defined by the space in which the random observations lie. Our result only assumes that the population-level operator $\mathbf{h}$ is a γ-contraction, and it holds for any sample size $n \gtrsim (1-\gamma)^{-4}$. To the best of our knowledge, this is the first non-asymptotic bound for SA procedures with general non-Euclidean norm that depends directly on the geometric complexity of the underlying space.

- Under a local linearization assumption on the operator $\mathbf{h}$, we establish a sharp

instance-dependent upper bound on the *estimation error* $\left\|\theta_n - \theta^*\right\|_C$, measured by any semi-norm $\left\|\cdot\right\|_C$ that is dominated by $\|\cdot\|$. The leading-order term of this bound is a Gaussian complexity involving the dual ball of the semi-norm $\left\|\cdot\right\|_C$ and the class of linear operators that achieve linearization in a local neighborhood. Our result uses a weaker local linearization assumption, which allows for important problems such as solving MDPs using a generative model.

- When the operator $\mathbf{h}$ is affine, we establish an improved result that matches the leading-order term in the nonlinear case, but with sample complexity scaling more mildly as $(1-\gamma)^{-2}$ for a γ-contractive operator. This result is further generalized to the case where $\mathbf{h}$ itself is not necessarily contractive, but its m-step composition is contractive. The result is valid as long as the sample size is lower bounded as $n \gtrsim m^2$.

- We illustrate the value of our theorems via a number of concrete examples, including stochastic shortest path problems, minimax Markov games, and average-reward policy evaluation. Our analysis leads to sharp instance-dependent bounds on the estimation error for such problems, achieved by the `ROOT-SA` algorithm.

Related work

In this section, we survey existing literature on stochastic approximation and its variance-reduced analogues.

Stochastic approximation and asymptotic guarantees: The study of stochastic approximation methods dates back to the seminal work of Robbins and Monro [RM51], as well as Kiefer and Wolfowitz [KW52], who established asymptotic convergence for various classes of one-dimensional problems. Subsequent work by Ljung [Lju77a; Lju77b] and Kushner and Clark [KC78] provided general criteria for convergence to a stable limit, in particular by using an ordinary differential equation (ODE) to track the trajectory of SA procedures. The paper [Ben96] further developed the ODE method to cover the case where the limit is an invariant set. Moving beyond convergence properties, the asymptotic distribution of SA trajectories has been characterized using certain invariance principles. Khas'minskii [Kha66] developed general criteria for the rescaled trajectory to converge to a Gauss-Markov process; see also Kushner and coauthors [KS84; Kus84] for further results in this vein. We refer the reader to the monographs [BMP12; Bor09; KY03] for more background and details on these results.

The idea of improving SA schemes by averaging the iterates was proposed in independent work by Polyak and Juditsky [Pol90; PJ92] as well as Ruppert [Rup88b]. Averaging the iterates allows for the use of more aggressive stepsize choices, and Gaussian limiting behavior is achieved over a broad range. Such a limiting distribution is known to be local asymptotic minimax optimal [DR16; Häj72; van00]. The idea

of iterate averaging lies behind many important aspects of large-scale statistical learning, leading to improved algorithms in different settings [BM13; DR16; Tri+18] and laying the foundations of online statistical inference [Che+20a]. The ROOT-SGD algorithm [Li+20] that inspired our approach is motivated by the averaging scheme, but combines variance reduction with averaging of the gradient sequence (as opposed to the sequence of iterates).

Non-asymptotic guarantees for stochastic approximation: Recent years have witnessed significant interest in obtaining non-asymptotic guarantees of the standard SA scheme (see equation (4.3) in the sequel). Wainwright [Wai19c] proved non-asymptotic guarantees for stochastic approximation algorithms under a cone-contractive assumption. Qu and Wierman [QW20] directly analyzed the iterates of SA algorithm in the asynchronous setting. Chen et al. [Che+20b] derived non-asymptotic bounds on stochastic approximation methods using Lyapunov functions. Using the generalized Moreau envelope, they constructed a smooth Lyapunov function, and show that the iterates of a standard SA scheme have a negative drift with respect to this Lyapunov function. Such Lyapunov techniques have been used to derive non-asymptotic guarantees for SA schemes in variety of settings (e.g., [Che+21a; Che+21b; Che+19; ZZM21]). For general contractive fixed-point problems in Banach spaces, Gupta et al., [GJG18] developed general criteria for the asymptotic convergence of mini-batch fixed-point iterations; and recently, Borkar [Bor21] established non-asymptotic concentration inequalities for the iterates, albeit with potentially dimension-dependent pre-factors. It should be noted that the standard SA scheme (4.3), while guaranteed to converge to the fixed point, may do so at a sub-optimal rate when measured in a minimax sense; for example, the papers [Li+21; Wai19c] demonstrate the non-optimality of this approach for the Q-learning problem in reinforcement learning.

Non-asymptotic guarantees that are instance-dependent—meaning that they go beyond worst-case and are adaptive to the difficulty—have been established for several stochastic approximation procedures. For stochastic gradient (SG) methods in the Euclidean setting, such bounds have been established for Polyak-Ruppert-averaged SG [GP17; MB11] and variance-reduced SG algorithms [Fro+15; Li+20], with the sample complexity and high-order terms being improved over time. For reinforcement learning problems, such type of guarantees have been established in the $\|\cdot\|_\infty$ norm for temporal difference methods [Kha+20b] and Q-learning [Kha+21] under a generative model, as well as Markovian trajectories [LLP21; Mou+21] under the ℓ_2-norm. In the context of stochastic optimization, the paper [Li+20] provides fine-grained bound for ROOT-SGD with a *unity* pre-factor on the leading-order instance-dependent term. The bounds in this chapter, on the other hand, involve constants that need not be optimal in this sense. It is an interesting future direction of research to establish similar non-asymptotic bounds for `ROOT-SA` with the sharp unity pre-factor.

Variance-reduced stochastic approximation algorithms: In order to obtain optimal SA procedures, different forms of variance reduction have been analyzed. The idea of variance reduction in stochastic approximation is classical; in the specific context of stochastic gradient methods, the papers [DBL14; JZ13; SLB17] proposed versions of variance reduction that accelerate convergence by careful averaging and re-centering of the gradient sequence. In this special case of stochastic optimization, the fixed-point operator $\mathbf{h}$ is obtained from the gradient update operator (cf. the discussion in Section 4.1); under suitable convexity and smoothness conditions, it is contractive under the ℓ_2-norm. In more recent work, several fully online schemes for variance-reduced stochastic optimization have been developed and analyzed, including SARAH [NST21; Ngu+17], STORM [CO19] and ROOT-SGD [Li+20]. The ROOT-SGD scheme uses recursive $1/t$-averaging of gradients, and has been been shown to be optimal for various convex problems in both asymptotic and non-asymptotic settings; see the paper [Li+20] and references therein for more details.

In the context of reinforcement learning (RL) problems, the operator $\mathbf{h}$ often corresponds to some type of Bellman operator [Ber12a; Ber19], known to be contractive under the ℓ_∞-norm [Ber12b]. Unfortunately, the key techniques used to design optimal methods for RL are differ considerably from those used in the stochastic optimization literature. Concretely, in order to obtain optimal RL algorithms, it is often necessary to exploit monotonicity properties of the Bellman operator, combined with variance reduction schemes [Kha+21; Kha+20b; Sid+18a; Wai19e]. Consequently, the literature is currently lacking a more unified perspective on how to obtain optimal SA schemes in a general setting. The main contribution of this chapter is to fill this gap by proposing and analyzing a single variance-reduced stochastic approximation algorithm for finding the fixed point of *any contractive* operator. In this way, our analysis does not depend on the exact form of the contraction norm $\|\cdot\|$.

Notation: We use $\mathbb{V}^*$ to denote the dual space of the Banach space $\mathbb{V}$, i.e., the space of all bounded linear functionals on $\mathbb{V}$. We define the dual norm $\|y\|_* := \sup_{x \in \mathbb{V}\setminus\{0\}} \langle x, y\rangle / \|x\|$. We define the unit norm ball $\mathbb{B} := \left\{x \in \mathbb{V}, \|x\| \leq 1\right\}$ in $\mathbb{V}$, as well the dual norm unit ball $\mathbb{B}^* := \left\{y \in \mathbb{V}^* \mid \|y\|_* \leq 1\right\}$.

Given a bounded linear operator $A : \mathbb{V} \to \mathbb{V}$, the adjoint operator $A^* : \mathbb{V}^* \to \mathbb{V}^*$ is characterized by the property

$$\langle Ax, y\rangle = \langle x, A^* y\rangle \qquad \text{for all } x \in \mathbb{V} \text{ and } y \in \mathbb{V}^*.$$

The operator norm of a bounded linear operator A on $\mathbb{V}$ is given by $|\!|\!| A |\!|\!|_{\mathbb{V}} := \sup_{x \in \mathbb{V}\setminus\{0\}} \frac{\|Ax\|}{\|x\|}$. Similarly, we can define the operator norm $|\!|\!| \cdot |\!|\!|_{\mathbb{V}^*}$ of a bounded linear operator mapping from $\mathbb{V}^*$ to itself. It can be seen that for any bounded linear operator A that maps $\mathbb{V}$ to itself, we have the equivalence $|\!|\!| A^* |\!|\!|_{\mathbb{V}^*} = |\!|\!| A |\!|\!|_{\mathbb{V}}$.

4.2 Problem Setup and the ROOT-SA Algorithm

In this section, we begin with a precise description of the class of problems that we study, along with the assumptions imposed. We then describe the ROOT-SA algorithm analyzed in this chapter.

Problem formulation

Consider a separable Banach space $(\mathbb{V}, \|\cdot\|)$, and an operator $\mathbf{h}$ mapping from $\mathbb{V}$ to itself. Assuming sufficient regularity to guarantee the existence and uniqueness of the fixed-point θ^* of the operator $\mathbf{h}$, we study stochastic approximation procedures for estimating the fixed point, i.e., for approximately solving the equation $\mathbf{h}(\theta) = \theta$.

In many practical applications, we may not have access to the operator $\mathbf{h}$ itself; instead, at each time t, we have access to a stochastic oracle $\mathbf{H}_t$ that, when queried at some $\theta \in \mathbb{V}$, returns a noisy version $\mathbf{H}_t(\theta)$ of the operator evaluation $\mathbf{h}(\theta)$. We impose the following conditions on the stochastic operators $\{\mathbf{H}_t\}_{t\geq 1}$ and the population operator $\mathbf{h}$:

Assumptions

(A1) There is a scalar $\gamma \in [0, 1)$ such that the operator $\mathbf{h} : \mathbb{V} \to \Omega$ is γ-contractive—viz.

$$\|\mathbf{h}(\theta_1) - \mathbf{h}(\theta_2)\| \leq \gamma \|\theta_1 - \theta_2\| \quad \text{for all } \theta_1, \theta_2 \in \mathbb{V},$$

where $\Omega \subseteq \mathbb{V}$ is a symmetric, closed and convex set that contains the range of $\mathbf{h}$.

(A2) For each $t = 1, 2, \ldots$, the stochastic operator $\mathbf{H}_t : \mathbb{V} \mapsto \Omega$ is almost surely (a.s.) L-Lipschitz:

$$\|\mathbf{H}_t(\theta_1) - \mathbf{H}_t(\theta_2)\| \leq L \|\theta_1 - \theta_2\| \qquad \text{a.s. for all } \theta_1, \theta_2 \in \mathbb{V}.$$

(A3) For any fixed $\theta \in \mathbb{V}$, the noise variables $\{\varepsilon_t(\theta) := \mathbf{H}_t(\theta) - \mathbf{h}(\theta)\}_{t\geq 1}$ are zero-mean and i.i.d., and $\|\varepsilon_t(\theta^*)\| \leq b_*$ almost surely for all $t = 1, 2, \ldots$.

A few remarks are in order. By the Banach fixed point theorem (e.g., [DG03]), the contractivity condition in Assumption (A1) ensures that $\mathbf{h}$ has a unique fixed point θ^*. The bulk of our analysis imposes Assumption (A1), with the exception of Section 4.3, where it is relaxed to a multi-stage contraction assumption in the special case of linear operators. Throughout this chapter, we assume that $\gamma \geq \frac{3}{4}$ for the ease of presentation. Note that this assumption can be made without loss of generality, since an operator that is γ-contractive for some $\gamma \in [0, 3/4)$ is also $3/4$-contractive. Assumption (A2)

requires the stochastic operator $\mathbf{H}_t$ to be Lipschitz, with the associated constant L allowed to be much larger than one—that is, there is no requirement that $\mathbf{H}_t$ be contractive or non-expansive. This setup should be contrasted with past work on cone-contractive operators [Wai19c; Wai19e] or ℓ_∞-norm contractions [Kha+21; Kha+20b], in which the stochastic operator $\mathbf{H}_t$ itself is required to be contractive. In the special case of stochastic optimization in $\mathbb{R}^d$, this type of sample-level Lipschitz condition is widely used, especially for variance-reduced procedures (cf. [JZ13; Li+20; NST21]).

In Assumptions (A1) and (A2), it is always possible to choose $\Omega = \mathbb{V}$. However, in certain cases, we find that choosing Ω to be a strict subset containing $\mathbf{h}$ leads to a sharper analysis, especially when $\mathbb{V}$ is infinite-dimensional. As for Assumption (A3), it imposes bounds only on the noise function when evaluated at the fixed point θ^* of the operator $\mathbf{h}$. In conjunction with Assumption (A2), this bound implies that $\|\varepsilon_t(\theta)\| \leq b_* + (L+\gamma)\,\|\theta - \theta^*\|$, allowing the norm of the noise $\varepsilon_t(\theta)$ to grow linearly with $\|\theta - \theta^*\|$. It is worth remarking that by using slightly more involved concentration arguments, it is possible to relax the almost-sure bounds in Assumptions (A2) and (A3). More precisely, it suffices to impose a p^{th}-moment condition on all projections:

$$\sup_{u \in \mathbb{B}^*} \mathbb{E}\left[\langle u,\, \mathbf{H}_1(\theta_1) - \mathbf{H}_1(\theta_2)\rangle^p\right] \leq p! \cdot L^p \,\|\theta_1 - \theta_2\|^p \qquad \text{for all } \theta_1, \theta_2 \in \mathbb{V}, \text{ and} \tag{4.2a}$$

$$\sup_{u \in \mathbb{B}^*} \mathbb{E}\left[\langle u,\, \varepsilon_1(\theta^*)\rangle^p\right] \leq p! \cdot b_*^p, \tag{4.2b}$$

for all $p \geq 2$. In order to simplify the presentation, we use the almost-sure bounds (A2)-(A3) throughout the chapter.

The `ROOT-SA` algorithm

Stochastic approximation algorithms are methods for solving fixed-point equations based on noisy observations. The simplest of such schemes takes the following form: starting with an initial point θ_0, one follows the recursion

$$\theta_{t+1} = \theta_t + \alpha_t \left\{\mathbf{H}_t(\theta_t) - \theta_t\right\}, \tag{4.3}$$

where $\{\alpha_t\}_{t \geq 0}$ is a sequence of stepsizes, typically in the interval $(0, 1)$. At any given step t, conditioned on θ_t, the quantity $\mathbf{H}_t(\theta_t)$ is an unbiased estimate of $\mathbf{h}(\theta_t)$, and the noise in the observation model is given by $\mathbf{H}_t(\theta_t) - \mathbf{h}(\theta_t)$. Under the contractivity assumptions (A1) on the operator $\mathbf{h}$ and moment bounds on the observation noise $\{\mathbf{H}_t(\theta_t) - \mathbf{h}(\theta_t)\}_{t \geq 1}$, the sequence $\{\theta_t\}$ converges almost surely to the unique fixed point θ^*; moreover, the rate of convergence of θ_t to θ^* is governed by the conditional variance of $\mathbf{H}_t(\theta_t)$ around its conditional mean $\mathbf{h}(\theta_t)$. See the standard texts [BMP12; Bor09; KY03] for results of this type.

The goal of variance reduction is to improve the basic stochastic approximation scheme (4.3) by replacing $\mathbf{H}_t(\theta_t) - \theta_t$ with an alternative quantity v_t that has lower variance. In this chapter, we study a simple version of such a variance-reduction

scheme, as described in Algorithm 5. This algorithm is a variant of the Recursive-one-over-t SGD (ROOT-SGD) algorithm proposed and analyzed in the past work [Li+20] involving a subset of the current authors. The ROOT-SGD algorithm was developed for stochastic optimization; it exploits a two-time scale framework that averages the gradient while performing variance reduction. Our ROOT-SA algorithm extends this same idea to the more general setting of stochastic approximation for fixed-point finding in Banach spaces. While the algorithms are similar in spirit, the analysis in this chapter uses completely different techniques, since it applies to the Banach-space setting for general operators, as opposed to the Euclidean setting and gradient operators of convex functions. The key technical difficulties lie in the absence of inner product structure.

Algorithm 5 ROOT-SA : A recursive SA algorithm

Given (a) Initialization $\theta_0 \in \mathbb{V}$, (b) Burn-in $B_0 \geq 2$, and (c) stepsize $\alpha > 0$
for $t = 1, \ldots, T$ **do**
  if $t \leq B_0$ **then**
    $v_t = \frac{1}{B_0} \sum_{t=1}^{B_0} \{\mathbf{H}_t(\theta_0) - \theta_0\}$, and $\theta_t = \theta_0$.
  else
    $v_t = \mathbf{H}_t(\theta_{t-1}) - \theta_{t-1} + \frac{t-1}{t}(v_{t-1} - \mathbf{H}_t(\theta_{t-2}) + \theta_{t-2})$,
    $\theta_t = \theta_{t-1} + \alpha v_t$.
  end if
end for
return θ_T

4.3 Main Results

In this section, we state our main results regarding the performance of the ROOT-SA procedure (cf. Algorithm 5), and discuss some of the consequences of these results. We provide non-asymptotic bounds on the behavior of the ROOT-SA algorithm in a number of different (semi)-norms. In Section 4.3, we derive bounds on the *operator defect* $\|\mathbf{h}(\theta_t) - \theta_t\|$, which measures how far the t^{th}-iterate θ_t of Algorithm 5 is from being a fixed point of the population operator $\mathbf{h}$. In other settings, we are interested in bounds on the *estimation error* $\|\theta_t - \theta^*\|$; accordingly, Section 4.3 is devoted to such results, along with bounds on various kinds of semi-norms. Finally, in Section 4.3, we discuss how to obtain refined results in the special case of linear operators, for which the contractivity assumption (A1) can be relaxed.

Upper bounds on operator defect

In this section, we provide non-asymptotic upper bounds on $\|\mathbf{h}(\theta_t) - \theta_t\|$. We begin by defining some important quantities that appear in our bound.

Complexity measures: Our bounds depend on certain properties of the i.i.d. noise sequence $\{\varepsilon_t(\theta^*)\}_{t\geq 1}$ and the range Ω of the operator $\mathbf{h}$. Recall that Ω is assumed to be a convex and symmetric set. Let $W \in \mathbb{V}$ be a zero-mean Gaussian random element such that

$$\mathbb{E}\left[\langle W,\, y\rangle \cdot \langle W,\, z\rangle\right] = \mathbb{E}\left[\langle \varepsilon_1(\theta^*),\, y\rangle \cdot \langle \varepsilon_1(\theta^*),\, z\rangle\right] \quad \text{for all } \; y, z \in \mathbb{V}^*. \tag{4.4}$$

Let $\mathbb{B}$ denote the unit ball in the space $(\mathbb{V}, \|\cdot\|)$, and let $\mathbb{B}^*$ denote its dual ball, i.e., a unit ball in the dual space $(\mathbb{V}^\star, \|\cdot\|_*)$. For a given integer sample size $n \geq 1$, we define a pseudo-metric ρ_n on the dual space $\mathbb{V}^*$ via

$$\rho_n(x, y) := \sup_{e \in n\Omega \cap \mathbb{B}} \langle x - y,\, e\rangle \quad \text{for all} \quad x, y \in \mathbb{B}^*. \tag{4.5}$$

Note that the additional restriction $e \in n\Omega$ makes ρ_n a weaker pseudo-metric than the dual norm, and in particular, we have $\rho_n(x, y) \leq \|x - y\|_*$ for any $x, y \in \mathbb{V}^*$. This weakening is important in the infinite-dimensional case, where the dual ball $\mathbb{B}^*$ is not compact under the original norm $\|\cdot\|_*$. The pseudo-metric ρ_n depends on the sample size n, but this dependence is mild in many interesting cases. For instance, if Ω is itself a linear subspace, then the pseudo-metric is independent of n. When studying the sup-norm, the n-dependence in the pseudo-metric contributes at most an additional $\log n$ factor to the packing numbers, and hence the high-order terms in our ultimate bounds.

Let $N(s; \mathbb{B}^*, \rho_n)$ denote the cardinality of a maximal s-packing of the dual ball $\mathbb{B}^* \subseteq \mathbb{V}^\star$ under the pseudo-metric ρ_n. For any $q \geq 1$ we define the Dudley entropy integral

$$\mathcal{J}_q(\mathbb{B}^*, \rho_n) := \int_0^\infty \Big[\log N(s; \mathbb{B}^*, \rho_n)\Big]^{1/q}\, ds.$$

Of particular interest are the cases $q = 2$ and $q = 1$, which arise in the cases of sub-Gaussian and sub-exponential tails, respectively. Finally, we define

$$\mathscr{W} = \mathbb{E}[\|W\|] \qquad \text{and} \qquad \nu = \sqrt{\sup_{u \in \mathbb{B}^*} \mathbb{E}\Big[\langle u,\, W\rangle^2\Big]}. \tag{4.6}$$

These two terms dictate the behavior of the leading-order term in our bound in Theorem 6.

Tuning parameters: Given a desired failure probability $\delta \in (0, 1)$, and a total sample size n, we run Algorithm 5 with the following choices of parameters:

$$\texttt{stepsize:} \qquad \alpha \in \left(0, \frac{(1-\gamma)^2}{cL^2 \mathcal{J}_2^2(\mathbb{B}^*, \rho_n) \log\left(\frac{n}{\delta}\right)}\right] \tag{4.7a}$$

$$\texttt{Burn-in time:} \qquad B_0 = \frac{c}{(1-\gamma)^2 \alpha} \log\left(\frac{n}{\delta}\right), \tag{4.7b}$$

where c is a universal constant.

Theorem 6. *Suppose that Assumptions (A1)—(A3) are in force, and given a sample size $n \geq 2B_0$, we run Algorithm 5 using an initialization θ_0 and the tuning parameters specified in equation (4.7). Then there is a universal constant c such that for any given $t \in [B_0, n]$, with probability $1 - \delta$, we have:*

$$\|\mathbf{h}(\theta_t) - \theta_t\| \leq \tfrac{cB_0}{t} \cdot \|\theta_0 - \mathbf{h}(\theta_0)\| + \tfrac{c}{\sqrt{t}}\left\{\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\right\} + c\tfrac{b_*}{(1-\gamma)}\left[\tfrac{1}{t} + \tfrac{\alpha L}{\sqrt{t}}\mathcal{J}_2(\mathbb{B}^*, \rho_n)\log(\tfrac{n}{\delta})\right] \cdot \left\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta})\right\}. \quad (4.8)$$

See Section 4.5 for a proof of this theorem.

A few comments regarding Theorem 6 are in order. In order to ease the notation, we adopt the shorthand $\gtrsim$ and $\lesssim$ to denote inequalities that hold up to terms that are at most logarithmic in the pair $(n, 1/\delta)$.

Optimal stepsize choice: Note that Theorem 6 holds for a range of stepsizes, and the stepsize requirement is interleaved with the formula (4.7b) for the burn-in period, which together impose a lower bound on the sample size needed:

$$n \geq \tfrac{c}{(1-\gamma)^4} L^2 \mathcal{J}_2^2(\mathbb{B}^*, \rho_n)\log^2(\tfrac{n}{\delta}). \quad (4.9a)$$

Given a sample size n satisfying this requirement, if we choose the stepsize to be

$$\alpha = \left\{L\mathcal{J}_2(\mathbb{B}^*, \rho_n)\log\left(\tfrac{n}{\delta}\right)\sqrt{n}\right\}^{-1}, \quad (4.9b)$$

the bound (4.8) for $t = n$ takes the form

$$\|\mathbf{h}(\theta_n) - \theta_n\| \leq \tfrac{c}{\sqrt{n}}\left\{\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\right\} + c\tfrac{b_*}{(1-\gamma)n}\left\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log\left(\tfrac{1}{\delta}\right)\right\} + \tfrac{cB_0}{n} \cdot \|\theta_0 - \mathbf{h}(\theta_0)\|. \quad (4.9c)$$

Initialization dependence and restarts: The dependency of the bound on the initial gap $\|\theta_0 - \mathbf{h}(\theta_0)\|$ is sub-optimal, but can be removed using a simple restarting procedure at the cost of $\mathcal{O}(B_0 \log n)$ additional samples. Concretely, given some fixed number $R \geq 1$ of restarting epochs, we can run the `ROOT-SA` algorithm for R consecutive short epochs, each with length $2cB_0$, with the constant c being the one in Eq (4.9c). The last iterate θ_{2cB_0} of each short epoch is used as the initial point of the subsequent epoch, and in the end, the output of last short epoch is used as the initial point $\widetilde{\theta}_0$ to run a final single-epoch instantiation of `ROOT-SA` on the rest of the data stream. In total, this restarting procedure uses an additional $2cB_0R$ samples, and the initialization of the last epoch satisfies the bound

$$\left\|\widetilde{\theta}_0 - \mathbf{h}(\widetilde{\theta}_0)\right\| \leq \tfrac{c}{\sqrt{B_0}}\left(\mathscr{W} + \nu\sqrt{\log(1/\delta)}\right) + \tfrac{cb_*}{B_0(1-\gamma)}\left(\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(1/\delta)\right) + \tfrac{\|\theta_0 - \mathbf{h}(\theta_0)\|}{2^R}. \quad (4.10)$$

By choosing $R \geq \log\left(\frac{\|\mathbf{h}(\theta_0)-\theta_0\|\sqrt{n}}{\mathscr{W}}\right)$ with a restarting sample size $2cB_0R$, we can ensure that $\frac{\|\theta_0-\mathbf{h}(\theta_0)\|}{2^R} \leq \frac{\mathscr{W}}{\sqrt{n}}$. Throughout this section, we assume that that the initialization θ_0 is such that the number of restarts R satisfies

$$\texttt{Initialization:} \qquad \log\left(\frac{\|\theta_0-\mathbf{h}(\theta_0)\|\sqrt{n}}{\mathscr{W}}\right) \leq c_0 \log n, \tag{4.11a}$$

for a universal constant $c_0 > 0$. In words, the condition ensures that the operator defect $\|\mathbf{h}(\theta_0) - \theta_0\|$ for the initialization θ_0 is not exponentially large compared to $\mathscr{W}$. We set the number of restarts R as

$$\texttt{Number of restarts:} \qquad R = 2c_0 \log n \tag{4.11b}$$

These conditions ensure that performing R many restarts requires at most

$$2cB_0 \log\left(\frac{\|\mathbf{h}(\theta_0)-\theta_0\|\sqrt{n}}{\mathscr{W}}\right) \lesssim 4c_0B_0 \log(n)$$

additional samples, assuming that the original sample size is lower bounded as $n \gtrsim \frac{L^2\mathcal{J}_2(\mathbb{B}^*,\rho_n)^2}{(1-\gamma)^4}$. Substituting this bound back to the bounds from Theorem 6 with the optimal stepsize choice (4.9b), we find that

$$\|\mathbf{h}(\theta_n) - \theta_n\| \leq \frac{c}{\sqrt{n}}\left\{\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\right\} + \frac{cb_*}{(1-\gamma)n} \cdot \left\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta})\right\}. \tag{4.12}$$

For the remainder of our discussion in this section, unless otherwise specified, we assume that the initial condition satisfies (4.10) with $R = 2c_0 \log n$, so that the effect due to initialization is negligible in the bound (4.9c).

Instance-dependent terms: The first and the second terms on the right-hand side of the bound (4.12) are instance-dependent quantitites, proportional to $1/\sqrt{n}$ and $1/n$, respectively. Consequently, viewed as function of sample size n, the first term is dominant. That term is proportional to the quantities $\mathscr{W}$ and ν, which depend on the covariance structure of the noise $\varepsilon(\theta^*)$—the covariance structure of the noise at the optimum θ^*—and hence the instance dependence. Next, observe that $\mathbb{E}[Z^2] \leq \frac{\pi}{2} \cdot (\mathbb{E}[|Z|])^2$ for any mean-zero Gaussian random variable Z, and we have the upper bound

$$\nu := \sqrt{\sup_{u\in\mathbb{B}^*} \mathbb{E}\left[\langle u, W\rangle^2\right]} \leq \sqrt{\sup_{u\in\mathbb{B}^*} \tfrac{\pi}{2}(\mathbb{E}|\langle u, W\rangle|)^2} \leq \sqrt{\tfrac{\pi}{2}} \cdot \mathbb{E}\,\|W\| := \sqrt{\tfrac{\pi}{2}} \cdot \mathscr{W}.$$

As a result, we conclude that the dominant term in the bound (4.12) admits a coarse upper bound $\frac{\mathscr{W}\cdot\sqrt{\log(1/\delta)}}{\sqrt{n}}$.

Moreover, given any $\delta \in (0, 1)$, a near-optimal tail bound for the Gaussian process W is given by

$$\mathbb{P}\left[\,\|W\| \geq \mathscr{W} + c \cdot \nu \cdot \sqrt{\log(\tfrac{1}{\delta})}\,\right] \leq \delta.$$

(For instance, see Ledoux and Talagrand [LT13], Section 3.1). In this way, Theorem 6 matches the behavior of the limiting Gaussian random variable, up to constant factors and high-order terms. Finally, it is natural to question whether the leading-order term $\frac{\mathscr{W}}{\sqrt{n}}$ in Theorem 6 is optimal. As we discuss in Section 4.3, this term actually matches the asymptotic limit up to a constant factor when $\mathbb{V}$ is finite dimensional, even though we have provided a non-asymptotic guarantee.

Comments on the higher-order terms: The second term in the right-hand side of the bound (4.12) depends on the Dudley entropy integral $\mathcal{J}_1(\mathbb{B}^*, \rho_n)$, a quantity that often appears in bounding the suprema of stochastic processes. In the current context, we use a standard chaining argument to derive the bound

$$\mathscr{W} \leq b_* \mathcal{J}_2(\mathbb{B}^*, \rho_n) \leq b_* \sqrt{\mathcal{J}_1(\mathbb{B}^*, \rho_n)},$$

with high probability. Such bounds characterize the fundamental complexity of the geometry for the underlying normed space structure, without explicitly depending on the problem dimension. Nonetheless, as with most bounds obtained via Bernstein-type inequalities, this high-order term depends on a worst-case uniform upper bound on the noise, instead of the actual variance.

As discussed in the next sections (cf. Sections 4.4 and 4.4), when the norm of interest $\|\cdot\|$ is a (weighted) ℓ_∞-norm on $\mathbb{R}^d$, we can replace the entropy term $\mathcal{J}_1(\mathbb{B}^*, \rho_n)$ by $\log d$. See the papers [Ber12b; BT91] for settings in which such a norm plays a central role.

Semi-norm bounds on the operator defect

There are various practical settings in which it is of interest to obtain a bound in some semi-norm $\left\|\cdot\right\|_C$ as opposed to the original Banach space norm $\|\cdot\|$. As a simple example, in the Euclidean setting, i.e. $\mathbb{V} = \mathbb{R}^d$, one might have an operator that is contractive in the ℓ_2-norm, but be interested in deriving bounds in the ℓ_∞-norm. Another interesting family is given by the semi-norms $\left\|\theta\right\|_C := \operatorname{abs} v^\top \theta$, where $v \in \mathbb{R}^d$ is a fixed direction.

Our analysis applies to any semi-norm $\left\|\cdot\right\|_C$ generated in the following way. For any symmetric and convex subset $C \subset \mathbb{V}^\star$, consider the associated semi-norm

$$\left\|\theta\right\|_C := \sup_{v \in C} \langle v, \theta \rangle. \tag{4.13}$$

Note that a wide class of interesting semi-norms can be generated in this way.

It is always possible to derive crude bounds in this semi-norm by relating it to the Banach space norm $\|\cdot\|$. In particular, if we define the *norm domination factor* $\mathcal{D} := \sup_{v \in C} \|v\|_*$, then we have the upper bound

$$\left\|\theta\right\|_C \leq (\sup_{v \in C} \|v\|_*) \cdot \|\theta\| := \mathcal{D} \cdot \|\theta\|. \qquad \text{for all} \quad \theta \in \mathbb{V}. \tag{4.14}$$

Consequently, a direct application of Theorem 6 yields

$$\left\|\mathbf{h}(\theta_n) - \theta_n\right\|_C \leq \mathcal{D} \cdot \|\mathbf{h}(\theta_n) - \theta_n\| \lesssim \tfrac{\mathcal{D}}{\sqrt{n}} \cdot \left\{\mathscr{W} + \nu \cdot \sqrt{\log(\tfrac{1}{\delta})}\right\},$$

with probability at least $1 - \delta$. However, this bound is of little utility when the norm domination constant $\mathcal{D}$ is large and/or dependent on the ambient dimension of the problem. Accordingly, the main result of this section is to prove sharper upper bound on $\left\|\mathbf{h}(\theta_n) - \theta_n\right\|_C$ that, while similar to the bound (4.12), has a leading-order term that depends on $\left\|W\right\|_C$, as opposed to the ambient dimension.

Our improved bound involves the complexity terms

$$\mathscr{W}_C = \mathbb{E}[\left\|W\right\|_C] \qquad \text{and} \qquad \nu_C = \sqrt{\sup_{u \in C} \mathbb{E}\left[\langle u, W\rangle^2\right]}. \tag{4.15}$$

These two terms dictate the behavior of the leading-order term in our bound on $\left\|\theta_n - \theta^*\right\|_C$.

Corollary 1. *Suppose that Assumptions (A1)–(A3) are in force, the sample size satisfies the lower bound $n \geq 2B_0$, and that we run Algorithm 5 using tuning parameters satisfying conditions (4.7) and (4.11). Then the iterate θ_n satisfies the bound*

$$\begin{aligned}\left\|\mathbf{h}(\theta_n) - \theta_n\right\|_C \leq \tfrac{c}{\sqrt{n}}\left\{\mathscr{W}_C + \nu_C\sqrt{\log(\tfrac{1}{\delta})}\right\} \\ + c\tfrac{\mathcal{D}}{(1-\gamma)}\left\{L\mathcal{J}_2(\mathbb{B}^*, \rho_n)\log(\tfrac{n}{\delta})\sqrt{\tfrac{\alpha}{n}} + \tfrac{1}{n\sqrt{\alpha}}\right\}\left\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\right\} \\ + \tfrac{c\mathcal{D}Lb_*}{1-\gamma}\left\{\tfrac{\sqrt{\alpha}}{n} + \tfrac{\alpha}{\sqrt{n}}\right\}\mathcal{J}_2(\mathbb{B}^*, \rho_n)\mathcal{J}_1(\mathbb{B}^*, \rho_n)\log^2(\tfrac{n}{\delta}),\end{aligned} \tag{4.16}$$

with probability at least $1 - \delta$.

See Section 4.5 for a proof of this corollary.

If we choose the stepsize $\alpha = \left\{L\mathcal{J}_2(\mathbb{B}^*, \rho_n)\log\left(\tfrac{n}{\delta}\right)\sqrt{n}\right\}^{-1}$, then Corollary 1 implies that

$$\left\|\mathbf{h}(\theta_n) - \theta_n\right\|_C \leq \tfrac{c}{\sqrt{n}}\left\{\mathscr{W}_C + \nu_C\sqrt{\log(\tfrac{1}{\delta})}\right\} + \left\{\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\right\} \cdot \mathcal{O}\left(\tfrac{\mathcal{D}}{(1-\gamma)n^{3/4}}\right) + \mathcal{O}\left(\tfrac{\mathcal{D}}{(1-\gamma)n}\right).$$

Thus, we see that the dominating term in the bound (4.16) (viewed as a function of sample size) depends only on the quantity $\mathscr{W}_C + \nu_C\sqrt{\log(\frac{1}{\delta})}$; this dependence is optimal for a bound in the semi-norm $\left\|\cdot\right\|_C$ (cf. the discussion following Theorem 6). Any dependence on the norm domination factor $\mathcal{D}$ remains only in the higher-order terms.

Upper bounds on the estimation error

In many problems, we are interested in computing an upper bound on estimation error $\|\theta_n - \theta^*\|$. A simple calculation[1] yields the bound

$$\|\theta_n - \theta^*\| \leq \tfrac{1}{1-\gamma} \cdot \|\theta_n - \mathbf{h}(\theta_n)\|. \tag{4.17}$$

Although this bound is useful—and sharp in a worst-case sense— it can certainly be improved in general.

In this section, we develop a result (to be stated as Theorem 7) that gives a sharper bound on the estimation error $\|\theta_n - \theta^*\|$ when it is possible to construct linear approximations of the operator $\mathbf{h}$ in a neighborhood of θ^*. More precisely, we impose the following local linearity condition.

Assumption: Local linearity

(A4) For any $s > 0$, there exists a set $\mathcal{A}_s$ of bounded linear operators on $\mathbb{V}$ such that

$$\|\theta - \theta^*\| \leq \sup_{A \in \mathcal{A}_s} \left\| (I - A)^{-1} \Big(\mathbf{h}(\theta) - \theta \Big) \right\| \qquad \text{for all } \theta \in \mathbb{B}(\theta^*, s). \tag{4.18}$$

As before, let W be a centered Gaussian random variable in $\mathbb{V}$ with the same covariance structure as $\varepsilon_1(\theta^*) := \mathbf{H}_1(\theta^*) - \mathbf{h}(\theta^*)$—that is

$$\mathbb{E}\left[\langle W, y\rangle \cdot \langle W, z\rangle\right] = \mathbb{E}\left[\langle \varepsilon_1(\theta^*), y\rangle \cdot \langle \varepsilon_1(\theta^*), z\rangle\right] \qquad \text{for all } y, z \in \mathbb{V}^\star.$$

Our bounds in this section are stated in terms of the solution to a fixed-point equation involving functionals of the Gaussian noise W. For any $s > 0$, define

$$\mathcal{G}(s) := \mathbb{E}\Big[\sup_{\substack{y \in \mathbb{B}^* \\ A \in \mathcal{A}_s}} \langle W, (I - A)^{-1} y \rangle \Big], \quad \text{and} \quad \sigma_*^2(s) := \sup_{\substack{y \in \mathbb{B}^* \\ A \in \mathcal{A}_s}} \mathbb{E}\left[\langle y, (I - A)^{-1} W \rangle^2 \right]. \tag{4.19}$$

Given a stepsize α satisfying (4.7a) and a tolerance probability $\delta \in (0, \frac{1}{1+\log(1/(1-\gamma))})$, we define the function

$$\mathcal{H}_n(\alpha, \delta) := \tfrac{\log(\frac{n}{\delta})}{(1-\gamma)^2} \Big\{ \Big[\mathcal{J}_2(\mathbb{B}^*, \rho_n) L \sqrt{\tfrac{\alpha}{n}} + \tfrac{1}{n\sqrt{\alpha}} \Big] \cdot \mathscr{W} + \Big[\tfrac{\mathcal{J}_2(\mathbb{B}^*, \rho_n) L \alpha}{\sqrt{n}} + \tfrac{1}{n} \Big] \cdot b_* \Big[\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{n}{\delta}) \Big] \Big\}. \tag{4.20a}$$

[1] By the triangle inequality, we have $\|\theta_n - \theta^*\| \leq \|\theta_n - \mathbf{h}(\theta_n)\| + \|\mathbf{h}(\theta_n) - \mathbf{h}(\theta^*)\|$. From the contractivity assumption (A1), we have $\|\mathbf{h}(\theta_n) - \mathbf{h}(\theta^*)\| \leq \gamma \|\theta_n - \theta^*\|$, and rearranging yields the claim.

This quantity serves as a higher-order term in our analysis. We consider the following fixed-point equation in the variable s:

$$s = \tfrac{\mathcal{G}(2s)}{\sqrt{n}} + \sigma_*(2s)\sqrt{\tfrac{\log(1/\delta)}{n}} + \mathcal{H}_n(\alpha, \delta). \tag{4.20b}$$

It can be shown that equation (4.20b) has a non-empty and bounded set of non-negative solutions (cf. the discussion following equation (4.22)); we let s_n^* be the largest such solution.

Theorem 7. *Suppose that Assumptions (A1)– (A4) are in force, and that we run Algorithm 5 using tuning parameters satisfying conditions (4.7) and (4.11). Then the final iterate θ_n satisfies the bound*

$$\|\theta_n - \theta^*\| \leq c \cdot s_n^* \quad \textit{with probability at least } 1 - \delta, \tag{4.21}$$

where c is a universal constant.

See Section 4.5 for the proof of this theorem.

Note that our contractivity assumption implies that functions $\mathcal{G}$ and ν defined in equation (4.19) are uniformly bounded—viz.

$$\begin{aligned} \mathcal{G}(s) &= \mathbb{E}\left[\sup_{y \in \mathbb{B}^*, A \in \mathcal{A}_s} \langle W, (I - A)^{-1} y \rangle\right] \leq \frac{\mathbb{E}\left[\|W\|\right]}{1 - \gamma}, \quad \text{and} \\ \sigma_*^2(s) &:= \sup_{y \in \mathbb{B}^*, A \in \mathcal{A}_s} \mathbb{E}\left[\langle y, (I - A)^{-1} W \rangle^2\right] \leq \frac{1}{(1-\gamma)^2} \sup_{y \in \mathbb{B}^*} \mathbb{E}\Big[\langle y, W \rangle^2\Big]. \end{aligned} \tag{4.22}$$

These inequalities (4.22), in conjunction with Theorem 6, guarantee that the fixed-point equation (4.20b) has a non-empty and bounded set of solutions; consequently, the maximum solution s_n^* is well-defined. Moreover, this also shows that the bound from Theorem 7 is always superior to the naive bound (4.17).

Note that only the high-order term $\mathcal{H}_n(\alpha, \delta)$ depends on the stepsize. By taking the optimal stepsize $\alpha_n = \left\{ L \mathcal{J}_2(\mathbb{B}^*, \rho_n) \log\left(\tfrac{n}{\delta}\right) \sqrt{n} \right\}^{-1}$, this term becomes:

$$\mathcal{H}_n(\alpha_n, \delta) := \frac{\log(\frac{n}{\delta})}{(1-\gamma)^2} \left\{ \frac{\sqrt{L \mathcal{J}_2(\mathbb{B}^*, \rho_n)}}{n^{3/4}} \cdot \mathscr{W} + \frac{b_*}{n} \Big[\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{n}{\delta}) \Big] \right\}, \tag{4.23}$$

which consists of two terms: an $\mathcal{O}(n^{-3/4})$ term depending on the Gaussian process supremum $\mathscr{W}$ that captures the second moment of the noise, and an $\mathcal{O}(n^{-1})$ term depending on the worst-case upper bound on the noise, as well as the Dudley integral. Under our stepsize choice, the high-order terms not only decay at a faster rate with sample size n, but also capture the underlying complexity of the norm $\|\cdot\|$, instead of the ambient dimension of the space $\mathbb{V}$.

Asymptotic optimality: As with the results from the previous section, Theorem 7 provides a non-asymptotic bound involving the Gaussian process $(I - A)^{-1}W$ for some $A \in \mathcal{A}_s$. Once again, it is natural to ask if such a Gaussian process is locally asymptotically minimax (LAM) (cf. [Häj72; LeC53; van00]). In order to apply Le Cam's theory so as to address this question, let us assume that the vector space $\mathbb{V}$ has finite dimension. Moreover, we impose an additional assumption on the (approximate) uniqueness of the local linearization. In particular, we assume that there exists a mapping $A_0 : \mathbb{V} \to \mathbb{V}$ such that

$$\lim_{s \to 0^+} \sup_{A \in \mathcal{A}_s} |\!|\!| A - A_0 |\!|\!|_{\mathbb{V}} = 0. \tag{4.24}$$

This condition holds, for example, when the operator $\mathbf{h}$ is Fréchet differentiable in an open neighborhood of the point θ^*, in which case A_0 corresponds to the Fréchet derivative of $\mathbf{h}$ at θ^*. It is also worth noticing that the condition (4.24) implies $\lim_{s \to 0^+} \mathcal{G}(s) = \mathcal{G}(0) = \mathbb{E}\left[\|(I - A_0)^{-1}W\|\right]$.

When $\mathbb{V}$ is finite-dimensional and condition (4.24) holds, the optimality of the random variable $(I - A_0)^{-1}W$ follows from classical LAM theory [Häj72; van00]. In particular, let us adopt the "tilting" method described in the paper [DR16]. Suppose that the (i.i.d.) random operators $\{\mathbf{H}_t\}_{t \geq 1}$ follow distribution $\mathbb{P}$, and for any distribution $\mathbb{Q}$, let $\theta^*(\mathbb{Q})$ denote a solution (assuming that one exists) to the fixed-point equation $\theta = \mathbb{E}_{\mathbf{H} \sim \mathbb{Q}}\left[\widetilde{\mathbf{H}}(\theta)\right]$. Under suitable tail assumptions on the distributions $\mathbb{P}$ and $\mathbb{Q}$, for any estimator $\widetilde{\theta}_n$ that maps a sequence of observed operators $\{\mathbf{H}_t\}_{t=1}^n$ to the vector space $\mathbb{V}$, an adaptation of Theorem 1 from [DR16] yields the lower bound

$$\liminf_{\Delta \to \infty} \liminf_{n \to \infty} \sup_{\mathbb{Q} \,|\, D(\mathbb{Q} \,\|\, \mathbb{P}) \leq \frac{\Delta}{n}} \mathbb{E}\left[\mathscr{L}\left(\sqrt{n}\left(\widetilde{\theta}_n - \theta^*(\mathbb{Q})\right)\right)\right] \geq \mathbb{E}\left[\mathscr{L}\left((I - A_0)^{-1}W\right)\right], \tag{4.25}$$

valid for any quasi-convex symmetric loss function $\mathscr{L} : \mathbb{V} \to \mathbb{R}^+$. If we make the particular choice $\mathscr{L}(\cdot) = \|\cdot\|$, then we can conclude that, when estimating θ^* in the Banach norm $\|\cdot\|$, the asymptotic lower bound is given by $\mathbb{E}\left[\|(I - A_0)^{-1}W\|\right]$.

Let us compare this fundamental limit to the behavior of the `ROOT-SA` estimator. We take a sequence of stepsizes $\alpha = \alpha_n$ such that $\alpha_n \to 0^+$ and $n\alpha_n \to \infty$. With this choice, applying Theorem 7 yields that the `ROOT-SA` estimator θ_n satisfies the bound

$$\limsup_{n \to \infty} \mathbb{P}\left[\|\theta_n - \theta^*\| \geq c \cdot \mathbb{E}\left[\left\|(I - A_0)^{-1}W\right\|\right]\right] \leq \tfrac{1}{3}, \tag{4.26}$$

for some universal constant $c > 0$. Note that the constant $1/3$ in the right-hand-side can be replaced by any positive constant in $(0, 1)$.

Equation (4.26) establishes that the `ROOT-SA`estimator θ_n is asymptotically optimal up to a constant pre-factor, albeit in a high-probability sense (and not necessarily the expectation). This distinction arises since the tuning parameters underlying Theorem 7 depend on the failure probability δ. Nevertheless, the leading-order term

in Theorem 7 is optimal up to constant factors, even when the failure probability is considered. In particular, given an integer $p > 0$, taking $\mathscr{L}(\cdot) := \|\cdot\|^p$, the LAM result (4.25) implies that the risk is lower bounded as

$$\begin{aligned}\mathbb{E}\left[\left\|(I - A_0)^{-1}W\right\|^p\right] &\geq \tfrac{1}{2}\left(\left(\mathbb{E}\left[\left\|(I - A_0)^{-1}W\right\|\right]\right)^p + \sup_{x \in \mathbb{B}^*}\langle x,\, (I - A_0)^{-1}W\rangle^p\right) \\ &\geq \tfrac{1}{2}\left[\mathcal{G}(0)^p + \left(\sqrt{p}\cdot\sigma_*(0)\right)^p\right].\end{aligned}$$

On the other hand, given $\delta \in (0, 1)$, we have

$$\limsup_{n\to\infty}\ \mathbb{P}\left[\|\theta_n - \theta^*\| \geq c\cdot\left(\mathcal{G}(0) + \sigma_*(0)\sqrt{\log 1/\delta}\right)\right] \leq \delta, \tag{4.27}$$

which matches the behavior of the p-th moment lower bound.

Semi-norm bounds on the estimation error

Recall the setup of Section 4.3. We now refine these results by providing an upper bound on $\left\|\theta_n - \theta^*\right\|_C$, where $\left\|\cdot\right\|_C$ is a semi-norm of the form (4.13), assumed to satisfy the domination condition (4.14). Moreover, we assume the following modification of the local linearity condition holds.

Assumption: Local linearity in semi-norm

(A4)′ For any $s > 0$, there is a set $\mathcal{A}_s$ of bounded linear operators on $\mathbb{V}$ such that

$$\left\|\theta - \theta^*\right\|_C \leq \sup_{A\in\mathcal{A}_s}\left\|(I - A)^{-1}\Big(\mathbf{h}(\theta) - \theta\Big)\right\| \qquad \text{for all } \theta \in \mathbb{B}(\theta^*, s). \tag{4.28}$$

As a refinement of the definition (4.19), we introduce the complexity terms

$$\mathcal{G}_C(s) := \mathbb{E}\Big[\sup_{\substack{y\in C\\ A\in\mathcal{A}_s}}\langle W,\, (I - A)^{-1}y\rangle\Big], \quad \text{and} \quad \sigma^2_{*,C}(s) := \sup_{\substack{y\in C\\ A\in\mathcal{A}_s}}\mathbb{E}\Big[\langle y,\, (I - A)^{-1}W\rangle^2\Big]. \tag{4.29}$$

Given a stepsize α satisfying the bound (4.7a) and a tolerance probability $\delta \in (0, \frac{1}{\log(1/(1-\gamma))})$, we define $s^*_{C,n} > 0$ to be the largest solution to the fixed-point equation

$$s = \tfrac{\mathcal{G}_C(2s)}{\sqrt{n}} + \sigma_{*,C}(2s)\sqrt{\tfrac{\log(1/\delta)}{n}} + \mathcal{D}\cdot\mathcal{H}_n(\alpha, \delta), \tag{4.30}$$

where the higher-order term $\mathcal{H}_n(\alpha, \delta)$ was previously defined (4.20a).

Corollary 2. *Under Assumptions (A1)– (A3) and (A4)′, the estimate θ_n from Algorithm 5, obtained using tuning parameters satisfying conditions (4.7) and (4.11), satisfies the bound*

$$\left\|\theta_n - \theta^*\right\|_C \leq c\cdot s^*_{C,n} \quad \textit{with probability at least } 1 - \delta, \tag{4.31}$$

where c is a universal constant.

See Section 4.5 for the proof of this corollary.

Linear operators with multi-step contraction

In the special case where $\mathbf{h}$ is a bounded linear operator in $\mathbb{V}$, the contraction assumption (A1) can be significantly weakened. In particular, it suffices to require that a multi-step composition of the operator be contractive.

Assumption: multi-step contraction

(A1)′ For some integer $m \geq 1$, the affine operator $\mathbf{h}(\theta) = A\theta + b$ is m-stage contractive, meaning that

$$\|\!|A|\!\|_{\mathbb{V}} \leq 1 \quad \text{and} \quad \|\!|A^m|\!\|_{\mathbb{V}} \leq \tfrac{1}{2}. \tag{4.32}$$

Note that assumption (A1)′ implies that the linear operator $(I - A)$ is invertible; in particular, we have the operator norm bound

$$\|\!|(I-A)^{-1}|\!\|_{\mathbb{V}} \leq \sum_{k=0}^{\infty} \sup_{v \in \mathbb{B}} \left\|A^k v\right\| = \sum_{k=0}^{\infty} \sum_{j=0}^{m-1} \|\!|A^{mk+j}|\!\|_{\mathbb{V}} \leq \sum_{k=0}^{\infty} \sum_{j=0}^{m-1} \|\!|A^m|\!\|_{\mathbb{V}}^k \cdot \|\!|A^j|\!\|_{\mathbb{V}} \leq 2m. \tag{4.33}$$

Again, let W be a centered Gaussian variable in $\mathbb{V}$ with the same covariance structure as $\varepsilon(\theta^*) := \mathbf{H}(\theta^*) - \mathbf{h}(\theta^*)$; that is, $\mathbb{E}\left[\langle W, y\rangle \cdot \langle W, z\rangle\right] = \mathbb{E}\left[\langle \varepsilon(\theta^*), y\rangle \cdot \langle \varepsilon(\theta^*), z\rangle\right]$ for all $y, z \in \mathbb{V}^\star$. Finally, we define

$$\mathscr{W} = \mathbb{E}[\|W\|] \qquad \text{and} \qquad \nu = \sqrt{\sup_{u \in \mathbb{B}^*} \mathbb{E}\Big[\langle u, W\rangle^2\Big]}. \tag{4.34}$$

Tuning parameters: Given a desired failure probability $\delta \in (0, 1)$, and a total sample size n, we run Algorithm 5 with the following choices of parameters:

$$\texttt{Stepsize choice:} \qquad \alpha \leq \frac{c}{mL^2 \mathcal{J}_2(\mathbb{B}^*, \rho_n)^2 \cdot \log^2 \frac{n}{\delta}} \tag{4.35a}$$

$$\texttt{Burn-in time:} \qquad B_0 = \tfrac{cm}{\alpha} \log(\tfrac{n}{\delta}), \tag{4.35b}$$

where c is an universal constant.

Theorem 8. *Suppose that Assumptions (A1)′, (A2), and (A3) are in force, and given a sample size $n \geq 2B_0$, we run Algorithm 5 using tuning parameters from equation (4.35a) and (4.35b). Then for any given $t \in [B_0, n]$, with probability $1 - \delta$, the following bound holds true:*

$$\|\mathbf{h}(\theta_t) - \theta_t\| \leq \tfrac{c}{\sqrt{t}} \left\{ \mathscr{W} + \nu \sqrt{\log(\tfrac{1}{\delta})} \right\} + cb_* \left\{ \tfrac{1}{t} + \tfrac{\alpha L \mathcal{J}_2(\mathbb{B}^*, \rho_n)}{\sqrt{t}} \log(\tfrac{n}{\delta}) \right\} \left\{ \mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta}) \right\} + \frac{cB_0}{t} \|\theta_0 - \mathbf{h}(\theta_0)\|, \tag{4.36}$$

See Section 4.5 for the proof of this theorem.

A few remarks are in order. First, we note that the leading-order term $\frac{c}{\sqrt{t}}\Big\{\mathscr{W} + \nu\sqrt{\log(\frac{1}{\delta})}\Big\}$ coincides with the high-probability bounds for the limiting Gaussian random variable that captures the second moment of the noise $\varepsilon_t(\theta^*)$, as with Theorem 6. In order to obtain sharp bounds with exponentially-decaying dependency on the initial condition $\|\theta_0 - \mathbf{h}(\theta_0)\|$, we employ the re-starting scheme discussed after Theorem 6, with R consecutive short epochs, each of length $2cB_0$, and the rest of the data stream being used in the final epoch. Assuming that the initialization satisfies the condition

$$\texttt{Initialization:} \qquad \log\Big(\tfrac{\|\theta_0 - \mathbf{h}(\theta_0)\|\sqrt{n}}{\mathscr{W}}\Big) \leq c_0 \log n \tag{4.37a}$$

for some universal constant c_0, we set the number of restarts to be equal to

$$\texttt{Restarts:} \qquad R = 2c_0 \log n. \tag{4.37b}$$

Under such restarting scheme, for $n \geq 2cB_0R$ the output θ_n of the final iterate satisfies the bound

$$\begin{aligned}\|\mathbf{h}(\theta_n) - \theta_n\| \leq \tfrac{c}{\sqrt{n}}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\Big\} \\ + cb_*\Big\{\tfrac{1}{n} + \tfrac{\alpha L \mathcal{J}_2(\mathbb{B}^*, \rho_n)}{\sqrt{n}}\log(\tfrac{n}{\delta})\Big\}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta})\Big\}.\end{aligned} \tag{4.37c}$$

Taking the optimal stepsize $\alpha = \Big\{L\mathcal{J}_2(\mathbb{B}^*, \rho_n)\log\big(\tfrac{n}{\delta}\big)\sqrt{n}\Big\}^{-1}$, the bound (4.37c) becomes:

$$\|\mathbf{h}(\theta_n) - \theta_n\| \leq \tfrac{c}{\sqrt{n}}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\Big\} + \frac{cb_*}{n}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta})\Big\}. \tag{4.38}$$

Sample size requirement and high-order terms: Note that Theorem 8 requires the following sample size lower bound:

$$n \geq cL^2m^2\mathcal{J}_2(\mathbb{B}^*, \rho_n)^2 \cdot \log^3 \tfrac{n}{\delta}.$$

In the special case when the operator $\mathbf{h}$ is γ-contractive, assumption (A1)′ is satisfied with $m = \frac{1}{1-\gamma}$, and the sample size requirement of order $\mathcal{O}\big(\frac{1}{(1-\gamma)^2}\big)$ is milder than the $\mathcal{O}\big(\frac{1}{(1-\gamma)^4}\big)$ in the general case. Moreover, the high-order term in Eq (4.37c) scales as $\frac{cb_*}{n}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\frac{1}{\delta})\Big\}$, which does not depend on the contraction factor, and improves over the general nonlinear case. Intuitively, such differences come from the linear structure of the population-level operator $\mathbf{h}$, making the noises in the stochastic operators $\mathbf{H}_t$ pass through smooth transformations over the iterates and allowing for concentration arguments with more terms. Finally, the results hold true only under

the weaker multi-step contraction Assumption (A1)′. As we will see in Corollary 7, such a relaxation of the condition is crucial to stochastic approximation problems associated to the average-cost policy evaluation problems.

Observe that for a linear operator $\mathbf{h}(\theta) = A\theta + b$ that satisfies the contractivity condition (cf. Assumption (A1)′), the inverse $(\mathcal{I} - A)^{-1}$ exists, and we have

$$\theta - \theta^* = (\mathcal{I} - A)^{-1}(\mathbf{h}(\theta) - \theta^*).$$

Consequently, given any semi-norm $\|\cdot\|_C$ of the form (4.13) satisfying condition (4.14), an argument similar to Corollary 2 yields the following corollary.

Corollary 3. *Suppose that the conditions of Theorem 8 are in force. By running the* `ROOT-SA` *algorithm with the restarting scheme described in Eq (4.37), the output θ_n obtained from Algorithm 5 satisfies the bound*

$$\begin{aligned}\|\theta_n - \theta^*\|_C \le \tfrac{c}{\sqrt{n}}\Big\{\mathbb{E}\big[\|(I-A)^{-1}W\|_C\big] + \sqrt{\sup_{u\in C}\mathbb{E}\big[\langle u,\,(I-A)^{-1}W\rangle^2\big]\log(\tfrac{1}{\delta})}\Big\} \\ + cm\mathcal{D}\Big\{L\mathcal{J}_2(\mathbb{B}^*,\rho_n)\log(\tfrac{n}{\delta})\sqrt{\tfrac{\alpha m}{n}} + \tfrac{1}{n}\sqrt{\tfrac{m}{\alpha}}\Big\}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} \\ + cmb_*\mathcal{D}\Big\{\tfrac{1}{n} + \tfrac{\alpha L\mathcal{J}_2(\mathbb{B}^*,\rho_n)}{\sqrt{n}}\log(\tfrac{n}{\delta})\Big\}\Big\{\mathcal{J}_1(\mathbb{B}^*,\rho_n) + \log(\tfrac{n}{\delta})\Big\} \quad (4.39)\end{aligned}$$

with probability at least $1-\delta$.

See Section 4.5 for the proof of this corollary.

A few remarks are in order. First, as with Theorem 7 and Corollary 2, the leading-order term of the bound in Corollary 3 achieves the high-probability bound (under $\|\cdot\|_C$-norm) for the Gaussian random vector in the local asymptotic minimax limit, up to universal constant factors. Since the problem itself is linear, the class $\mathcal{A}_s$ of linear operators is singleton, and the estimation error upper bounds can be expressed directly through $\mathbb{E}\big[\|(I-A)^{-1}W\|_C\big]$, without resorting to fixed-point equations. Compared with the high-order terms defined by Eq (4.20a) in the general case, the high order terms in Eq (4.39) (the second and third line of the equation) save a factor of $\frac{1}{1-\gamma}$ in the contractive case, while generalizing to the multi-step contraction case. Such an improvement under linearity is in accordance with the operator defect bound in Theorem 8. Finally, we note that by choosing the optimal stepsize $\alpha = \Big\{L\mathcal{J}_2(\mathbb{B}^*,\rho_n)\log\big(\tfrac{n}{\delta}\big)\sqrt{n}\Big\}^{-1}$, the following upper bound holds true:

$$\begin{aligned}\|\theta_n - \theta^*\|_C \le \tfrac{c}{\sqrt{n}}\Big\{\mathbb{E}\big[\|(I-A)^{-1}W\|_C\big] + \sqrt{\sup_{u\in C}\mathbb{E}\big[\langle u,\,(I-A)^{-1}W\rangle^2\big]\log(\tfrac{1}{\delta})}\Big\} \\ + cm\mathcal{D}\frac{\sqrt{L\mathcal{J}_2(\mathbb{B}^*,\rho_n)m}}{n^{3/4}}\mathscr{W}\cdot\log(n/\delta) + \frac{cmb_*\mathcal{D}}{n}\Big\{\mathcal{J}_1(\mathbb{B}^*,\rho_n) + \log(\tfrac{n}{\delta})\Big\} \quad (4.40)\end{aligned}$$

In addition to the $\mathcal{O}\big(n^{-1/2}\big)$ optimal leading-order term, two high-order terms exists: an $\mathcal{O}\big(n^{-3/4}\big)$ term that depends on the variance of the noise $\varepsilon_t(\theta^*)$ and properties of the associated Gaussian process under the norm $\|\cdot\|$, and an $\mathcal{O}\big(n^{-1}\big)$ scaling with the almost-sure upper bounds on the noise as well as the Dudley integral. Once again, both high-order terms depends on geometry of the norm $\|\cdot\|$, instead of the ambient dimension of the problem.

4.4 Consequences for Concrete Use Cases

This section is devoted to the discussion of three classes of problems that fall within the framework of this chapter. They serve as illustrative examples for the consequences of our main results in Section 4.3.

Computing stochastic shortest paths

We begin with the problem of computing stochastic shortest paths (SSPs) [BT91; YB13]. Such SSP problems can be formulated in terms of a particular type of Markov decision process (MDP) with a finite state space $\mathcal{X}$ and action space $\mathcal{U}$. A generic MDP involves a collection of probability transition kernels, $\{\mathbf{P}_u(x' \mid x) \mid (x,u) \in \mathcal{X} \times \mathcal{U}\}$, where the transition kernel $\mathbf{P}_u(x' \mid x)$ denotes the probability of transition to the state x' when an action u is taken at the current state x. The MDP is equipped with a cost function, $c : \mathcal{X} \times \mathcal{U} \mapsto \mathbb{R}$, such that the scalar $c(x,u)$ denotes the cost received upon performing the action u in state x.

In the special case of a *stochastic shortest path* (SSP) problem, we assume that state 1 is absorbing and is cost-free, meaning that

$$c(x=1,u) = 0 \quad \text{and} \quad \mathbb{P}_u(x' \mid x=1) = \mathbf{1}_{x'=1} \quad \text{for all actions } u \in \mathcal{U}. \tag{4.41}$$

A stationary policy π is a mapping $\mathcal{X} \mapsto \mathcal{U}$ such that $\pi(x) \in \mathcal{U}$ denotes the action to be taken in the state x. We assume that the total infinite-horizon cost incurred by any stationary policy π is finite—viz.

$$\mathbb{E}_{x_0=x}\Big[\sum_{k=1}^{\infty} |c(x_k, \pi(x_k))|\Big] < \infty \quad \text{for all } x \in \mathcal{X}. \tag{4.42}$$

Any such stationary policy π is called a *proper policy*; throughout this section, if not mentioned otherwise, all policies are assumed to be proper policies. The aim is to obtain a policy $\pi^\star$ minimizing the total cost for any initial state x.

One way to estimate an optimal policy is to calculate the optimal Q-function. Associated with a proper policy π is its Q-function

$$\theta^\pi(x,u) := \mathbb{E}\Big[\sum_{k=0}^{\infty} c(x_k,u_k) \mid x_0 = x, u_0 = u\Big], \quad \text{where } u_k = \pi(x_k) \quad \text{for all } k = 1,2,\ldots.$$

The optimal Q-function is given by $\theta^\star(x,u) := \inf_{\pi\in\Pi}\ \theta^\pi(x,u)$.

Finally, we note that both finite-horizon and discounted dynamical programming problems are special cases of the stochastic shortest path formulation. In the former case, for an H-horizon problem on the state space $\mathcal{X}$, we construct the new state space $\mathcal{X}' = \Big(\mathcal{X}\times[H]\Big)\cup\{1\}$, where transitions are made from (x,h) to $(y,h+1)$ for $x,y\in\mathcal{X}$ and $h\le H-1$. At the end of H steps, the process moves directly to the absorbing state 1 and stays there afterwards. In the latter case, we let the augmented space be $\mathcal{X}' = \mathcal{X}\cup\{1\}$, and the Markov decision process is "killed" and moves to the absorbing state 1 at a rate $1-\gamma$. The resulting infinite-horizon undiscounted problems are equivalent to the original (finite-horizon or discounted) problems. On the other hand, the SSP formulation can cover much more general class of MDPs with absorbing states, beyond the simple finite-step or independent geometric killing structure.

Bellman operator and contractivity

Observe that for any policy π, the cost-free absorbing state property (4.41) ensures that $\theta^\pi(1,u)=0$, and as a result $\theta^\star(1,u)=0$ for all actions $u\in\mathcal{U}$. Using the shorthand $\mathcal{X}_{-1} := \mathcal{X}\setminus\{1\}$, it suffices to estimate $\{\theta^\star(x,u)\mid(x,u)\in\mathcal{X}_{-1}\times\mathcal{U}\}$. Under the proper policy assumption (4.42), classical theory [BT91; YB13] guarantees that the optimal Q-function restricted to the set $\mathcal{X}_{-1}\times\mathcal{U}$ is the unique fixed point of the Bellman operator

$$\mathbf{h}(\theta)(x,u) = c(x,u) + \sum_{x'\in\mathcal{X}_{-1}} \mathbf{P}_u(x'\mid x)\min_{u'\in\mathcal{U}}\theta(x',u') \qquad (x,u)\in\mathcal{X}_{-1}\times\mathcal{U}, \tag{4.43}$$

which is an operator on the space of Q-functions. For SSP problems with finite state and action spaces, any Q-function can be viewed an element of $\mathbb{R}^{|\mathcal{X}_{-1}\times\mathcal{U}|}$, in which case $\mathbf{h}$ can be viewed as an operator on $\mathbb{R}^D$ where $D := |\mathcal{X}_{-1}\times\mathcal{U}|$.

For a weight vector $\mathbf{w} := \{w_1,\dots,w_D\}\in\mathbb{R}^D_+$, we define a weighted ℓ_∞-norm on $\mathbb{R}^D$ via $\|\theta\|_w := \max\limits_{i=1,\dots,D}\frac{|\theta_i|}{w_i}$. In order to construct the norm, we let Π be the set of all stationary policies, and define the hitting times:

$$\forall x\in\mathcal{X}_{-1},\quad \tau_{\text{hit}}^{(x)} := \sup_{\pi\in\Pi}\ \mathbb{E}^\pi_{x_0=x}\left[\inf\Big\{t>0 : x_t = 1\Big\}\right]. \tag{4.44}$$

In words, the quantity $\tau_{\text{hit}}^{(x)}$ is the maximum over all the possible policies of the largest expected time for the MDP to go from the state x to the absorbing state 1. We further define

$$\tau_{\min} = \min_{x\in\mathcal{X}_{-1}}\tau_{\text{hit}}^{(x)} \qquad\text{and}\qquad \tau_{\max} = \max_{x\in\mathcal{X}_{-1}}\tau_{\text{hit}}^{(x)}.$$

Taking the weight vector as $w_x = \tau_{\text{hit}}^{(x)}$ for each $x\in\mathcal{X}_{-1}$, it is known [BT91; Tse90] that

$$\|\mathbf{h}(\theta_1)-\mathbf{h}(\theta_2)\|_w \le \left(1-\frac{1}{\tau_{\max}}\right)\cdot\|\theta_1-\theta_2\|_w \quad\text{for}\quad \theta_1,\theta_2\in\mathbb{R}^D. \tag{4.45}$$

Thus, the operator $\mathbf{h}$ is $\left(1 - \frac{1}{\tau_{\max}}\right)$-contractive in the weighted ℓ_∞-norm, so that our general theory can be applied with this choice of Banach space.

Generative observation model

We analyze the `ROOT-SA` algorithm under a stochastic oracle known as the *generative observation model* for the SSP problem. For any state-action pair (x, u), the generative model allows us to draw next-state and cost samples from the MDP $(r, \mathbf{P})$. More precisely, we have access to a collection of n i.i.d. samples of the form $\{(\mathbf{Z}_k, C_k)\}_{k=1}^n$, where both $\mathbf{Z}_k$ and C_k are random matrices in $\mathbb{R}^{|\mathcal{X}_{-1}| \cdot |\mathcal{U}|}$. For each state-action pair (x, u), the entry $\mathbf{Z}_k(x, u)$ is drawn according to the transition kernel $\mathbf{P}_u(\cdot \mid x)$, whereas the entry $C_k(x, u)$ is a zero-mean random variable with mean $c(x, u)$; this corresponds to a noisy observation of the cost function. We assume that the random cost $C_k(x, u)$ is upper bounded by $c_{\max}$ in absolute value. Here the cost samples $\{C_k(x, u)\}_{(x,u) \in \mathcal{X} \times \mathcal{U}}$ are independent across all state-action pairs, and the cost samples $\{C_k\}$ are independent of the transition samples $\{\mathbf{Z}_k\}$.

The empirical Bellman operator: Given a sample $(\mathbf{Z}, C)$ from our observation model, we define the single-sample empirical Bellman operator $\mathbf{H}(\cdot)$ on the space of Q-functions, whose action on a Q-function θ is given by

$$\mathbf{H}(\theta)(x, u) := C(x, u) + \sum_{x' \in \mathcal{X}_{-1}} \mathbf{Z}_u\left(x' \mid x\right) \min_{u' \in \mathcal{U}} \theta\left(x', u'\right). \tag{4.46}$$

Here we have introduced $\mathbf{Z}_u\left(x' \mid x\right) := \mathbf{1}_{\mathbf{Z}(x,u) = x'}$. We are ready to state our guarantees for the stochastic shortest path problem.

Guarantees for stochastic shortest path

It is easy to see that the operators $\mathbf{h}(\cdot)$ and $\mathbf{H}(\cdot)$, defined respectively in equations (4.43) and (4.46), satisfy Assumptions (A1)- (A3) with the weighted ℓ_∞-norm $\|\cdot\|_w$. In order to obtain an optimal policy from an estimate θ_n of the optimal Q function, it is natural to obtain performance bounds in the $\|\cdot\|_\infty$ norm, and we do so by invoking Corollaries 1 and 2 with $\left\|\cdot\right\|_C = \|\cdot\|_\infty$.

Accordingly, consider a Gaussian random vector W with $W \sim \mathcal{N}\Big(0, \operatorname{cov}(\mathbf{H}(\theta^*) - \theta^*)\Big)$, and define

$$\mathscr{W} = \mathbb{E}[\|W\|_\infty], \qquad \nu^2 := \sup_{x \in \mathcal{X}_{-1}, u \in \mathcal{U}} \mathbb{E}[W_{x,u}^2], \quad \text{and} \quad b_* := \frac{c_{\max}}{\tau_{\min}} + \|\theta^*\|_w. \tag{4.47}$$

For a given failure probability $\delta \in (0, 1)$, our result applies to the algorithm with parameters

$$\alpha = c_1 \Big\{\sqrt{n \log |\mathcal{X} \times \mathcal{U}|} \cdot \log(n/\delta)\Big\}^{-1}, \quad \text{and} \quad B_0 = \tfrac{c_2 \tau_{\max}^2}{\alpha} \log(\tfrac{n}{\delta}), \tag{4.48a}$$

We also choose the initialization θ_0 and the number of restarts R such that

$$\log\left(\frac{\|\theta_0 - \mathbf{h}(\theta_0)\|\sqrt{n}}{\mathscr{W}}\right) \leq c_0 \log n \qquad \text{and} \qquad R \geq 2c_0 \log n, \tag{4.48b}$$

where c_0, c_1, c_2 are appropriate universal constants. We obtain the following guarantee:

Corollary 4. *Given a sample size n such that $\frac{n}{\log n} \geq c' \log(|\mathcal{X}| \cdot |\mathcal{U}|) \cdot \tau_{\max}^4 \log(1/\delta)$, running Algorithm 5 with the tuning parameter choices (4.48) yields an estimate θ_n such that*

$$\|\mathbf{h}(\theta_n) - \theta_n\|_\infty \leq \frac{c}{\sqrt{n}} \cdot \left\{\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\right\} + cb_* \tau_{\max}^2 \tfrac{\log(|\mathcal{X}|\cdot|\mathcal{U}|)}{n} \log^2(\tfrac{n}{\delta}),$$

with probability at least $1 - \delta$.

Note that when we invoke Corollary 1 to obtain this corollary, the second term is absorbed into the leading-order term under the sample size lower bound $\frac{n}{\log n} \geq c' \log(|\mathcal{X}| \cdot |\mathcal{U}|) \cdot \tau_{\max}^4 \log(1/\delta)$. In particular, the semi-norm domination factor is $\mathcal{D} = \tau_{\max}$ in this case, and we have the following inequalities:

$$\mathcal{D} \cdot \mathbb{E}\left[\|W\|_w\right] \leq \frac{\tau_{\max}}{\tau_{\min}} \mathbb{E}\left[\|W\|_\infty\right] \leq \tau_{\max} \mathscr{W}, \quad \text{and}$$
$$\mathcal{D} \cdot \sup_{\|y\|_{w^{-1}} \leq 1} \sqrt{\mathbb{E}\Big[\langle y, W\rangle^2\Big]} \leq \frac{\tau_{\max}}{\tau_{\min}} \sup_{\|y\|_\infty \leq 1} \sqrt{\mathbb{E}\Big[\langle y, W\rangle^2\Big]} \leq \tau_{\max} \nu,$$

which makes the second term of equation (4.16) dominated by the first term.

Next, in order to obtain an upper bound on the estimation error $\|\theta_n - \theta^*\|_\infty$ we need a few more definitions. For a given Q-function $\mathbb{Q}$, we say π is a greedy policy of θ if and only if

$$\pi(x) = \arg\min_u \mathbb{Q}(x, u) \qquad \text{for all} \quad x \in \mathcal{X}_{-1},$$

and denote Π^θ as the set of all greedy policies of θ. Note that the greedy policies of a given Q-function may not be unique. Using this greedy policy, we can define the right-linear operator

$$\mathbf{P}^{\pi_\mathbb{Q}} \mathbb{Q}(x, u) = \sum_{x'} \mathbf{P}_u(x' \mid x) \mathbb{Q}(x', \pi_\mathbb{Q}(x')).$$

We also define a set $\mathcal{A}_s$ of linear operators as

$$\mathcal{A}_s = \{\mathbf{P}^{\pi_\mathbb{Q}} \mid \pi_\mathbb{Q} \text{ is a greedy policy of } \mathbb{Q} \text{ with } \mathbb{Q} \in \mathbb{B}(\theta^\star, s)\}. \tag{4.49}$$

Let $\mathbb{B}(\theta^*, s) := \{\theta \mid \|\theta - \theta^*\|_\infty \leq s\}$ denote the ℓ_∞-ball of radius s around θ^*. We use $\pi_\star$ to denote the greedy policy associated with the optimal Q-function $\theta^\star$. In Section 4.9, we show that the local linearity assumption ((A4)′) is satisfied for the

Bellman operator (4.43) with the set of operators $\mathcal{A}_s$ from equation (4.49), and with $\left\|\cdot\right\|_C = \|\cdot\|_\infty$.

Given a tolerance probability $\delta \in (0, \frac{1}{\log(1/(1-\gamma))})$, let s_n^* denotes the largest positive solution to the fixed-point equation

$$\begin{aligned} s_n = \frac{1}{\sqrt{n}}\Bigg\{ & \mathbb{E}\Big[\sup_{\substack{\theta \in \mathbb{B}_\infty(\theta^*, s_n) \\ \pi \in \Pi^\theta}} \left\| \left(\mathcal{I} - \mathbf{P}^\pi\right)^{-1} W \right\|_\infty \Big] \\ & + \sup_{\substack{\theta \in \mathbb{B}_\infty(\theta^*, s_n), \pi \in \Pi^\theta \\ (x,u) \in \mathcal{X}_{-1} \times \mathcal{U}}} \left(\mathbb{E}\Big[\delta_{x,u}^\top (\mathcal{I} - \mathbf{P}^\pi)^{-1} W \Big] \log(1/\delta) \right)^{1/2} \Bigg\} \\ & + \tau_{\max}^2 \log(\tfrac{n}{\delta}) \left\{ \tfrac{\tau_{\max}}{\tau_{\min}} \left(\tfrac{\log(|\mathcal{X}_{-1}| \cdot |\mathcal{U}|)}{n} \right)^{3/4} \mathbb{E}[\|W\|_\infty] + b_* \tau_{\max} \tfrac{\log(|\mathcal{X}_{-1}| \cdot |\mathcal{U}|)}{n} \right\}. \end{aligned} \tag{4.50}$$

Here we have defined the indicator function $\delta_{x,u} = \mathbf{1}_{(x',u')=(x,u)}$. We obtain the following corollary:

Corollary 5. *Under the setup of Corollary 4, the estimate θ_n satisfies the bound*

$$\|\theta_n - \theta^*\|_\infty \leq c \cdot s_n^* \quad \text{with probability at least } 1 - \delta, \tag{4.51}$$

where c is a universal constant.

A few remarks are in order. First, the bound depends on the size of state-action space only poly-logarithmically, and depends on the quantity $\tau_{\max}$ through two sources: the contraction parameter and the norm domination factor between $\|\cdot\|_\infty$ and $\|\cdot\|_w$. Second, let Π^* be the set of all optimal policies for the SSP problem, for sample size n large enough,[2] the ball $\mathbb{B}_\infty(\theta^*, s_n)$ will eventually shrink to the singleton θ^*, and the supremum in the fixed-point equation (4.50) is taken over $\pi \in \Pi^*$. The solution s_n to such an equation therefore takes the following form:

$$\begin{aligned} s_n = \frac{1}{\sqrt{n}}\Bigg\{ & \mathbb{E}\Big[\sup_{\pi \in \Pi^*} \left\| \left(\mathcal{I} - \mathbf{P}^\pi\right)^{-1} W \right\|_\infty \Big] \\ & + \sup_{\substack{\pi \in \Pi^* \\ (x,u) \in \mathcal{X}_{-1} \times \mathcal{U}}} \left(\mathbb{E}\Big[\delta_{x,u}^\top (\mathcal{I} - \mathbf{P}^\pi)^{-1} W \Big] \log(1/\delta) \right)^{1/2} \Bigg\} + \text{high-order terms} \\ & \leq \frac{1}{\sqrt{n}} \max_{\substack{\pi \in \Pi^* \\ (x,u) \in \mathcal{X}_{-1} \times \mathcal{U}}} \sqrt{\mathbb{E}\left[\left(\delta_{x,u}^\top (\mathcal{I} - \mathbf{P}^\pi)^{-1} (\mathbf{H}(\theta^*) - \theta^*) \right)^2 \right]} \cdot \sqrt{\log \tfrac{|\mathcal{X}| \cdot |\mathcal{U}| \cdot |\Pi^*|}{\delta}} \\ & + \text{high-order terms}. \end{aligned}$$

Up to a factor of $\sqrt{\log \frac{|\mathcal{X}| \cdot |\mathcal{U}| \cdot |\Pi^*|}{\delta}}$, this matches the two-point lower bound in the paper [Kha+21] (in the discounted MDP case). When specializing to the cases

[2] The sample size requirement may depend on the gap between the value of optimal and sub-optimal actions, as in the prior work [Kha+21].

where the optimal policy is unique, or satisfies the Lipschitz-type assumptions in the paper [Kha+21], the upper bound above also recovers the leading-order term in that paper. We conjecture that the leading-order term of the solution s_n to the fixed-point equation is actually optimal for large n. It is an important direction of future work to investigate this gap, and establish optimality results under suitably defined problem classes.

Note that when specialized to the γ-discounted MDPs, the sample size requirement in Corollary 5 becomes $\mathcal{O}(1-\gamma)^{-4}$, which can be worse than the corresponding requirements in the paper [Kha+21], at least in certain regimes. Intuitively, this is the price we pay when moving to the general case where only the contraction of the population-level operator is assumed, instead of the sample-level contraction.

Two-player zero-sum Markov games

Two-player zero-sum Markov games are a generalization of MDPs and two player zero-sum games, in which two players play multiple rounds of a zero-sum game and their goal is to maximize their expected long-term reward. Markov games are characterized by a six-tuple $\{\mathcal{X}, \mathcal{U}_1, \mathcal{U}_2, \mathbf{P}, r, \gamma\}$. Let $\mathcal{X}$ denote the state space, and let $\mathcal{U}_1$ and $\mathcal{U}_2$ denote the action sets for players one and two, respectively. Here we focus on games with finite state and action space, i.e., $|\mathcal{X} \times \mathcal{U}_1 \times \mathcal{U}_2| < \infty$.

The probability transition kernel $\{\mathbf{P}_{u_1,u_2}(x' \mid x) \mid (x, u_1, u_2) \in \mathcal{X} \times \mathcal{U}_1 \times \mathcal{U}_2\}$, encodes the transition to the next state given the actions of the players. In particular, the scalar $\mathbf{P}_u(x' \mid x)$ denotes the probability of transition to the state x', when at state x player 1 takes the action u_1 and player 2 takes the action $\mathcal{U}_2$. The MDP is equipped with a reward function $r : \mathcal{X} \times \mathcal{U}_1 \times \mathcal{U}_2 \mapsto \mathbb{R}$ such that the scalar $r(x, u_1, u_2)$ denotes the cost received at state x when player 1 takes the action u_1 and player 2 takes the action u_2. Finally, the scalar $\gamma \in (0, 1)$ is a parameter reflecting the discounting of future rewards.

For each player $i \in \{1, 2\}$, a stationary policy π_i is a mapping $\mathcal{X} \to \mathcal{P}(\mathcal{U}_i)$, where $\mathcal{P}(\mathcal{U}_i)$ denotes the set of probability distributions over the finite action set $\mathcal{U}_i$. In other words, the actions taken by the players can be random, and for any state $x \in \mathcal{X}$, the distribution $\pi_i(\cdot \mid x)$ is a probability distribution on the set of actions $\mathcal{U}_i$ to be taken by player i. We use Π_1 and Π_2 to denote the set of all policies for players 1 and 2, respectively.

Assuming player 1 is following policy π_1, and player 2 is following policy π_2, the value $V(\cdot \mid \pi_1, \pi_2) : \mathbb{R}^{|\mathcal{X}|} \mapsto \mathbb{R}$ of player 1 is defined as the expected sum of discounted rewards in an infinite sample path:

$$V(x \mid \pi_1, \pi_2) = \mathbb{E}\Big[\sum_{k=1}^{\infty} r(x_k, u_{1k}, u_{2k} \mid x_0 = x)\Big],$$
$$\text{where } u_{1k} \sim \pi_1(\cdot \mid x_k) \text{ and } u_{2k} \sim \pi_2(\cdot \mid x_k). \tag{4.52}$$

Given that the game is zero-sum, the reward for player 2 with initial state x is $-V(x \mid \pi_1, \pi_2)$. Players 1 and 2 want to choose their policies π_1 and π_2 that maximize their respective reward for all values of initial state x.

Nash equilibrium: A natural notion of equilibrium in two-player zero-sum Markov games is the Nash equilibrium. A policy pair $(\pi_1^\star, \pi_2^\star)$ is called a *Nash equilibrium* if for all initial states $x \in \mathcal{X}$

$$\begin{aligned} V(x \mid \pi_1^\star, \pi_2^\star) &\geq V(x \mid \pi_1, \pi_2^\star) \quad \text{for all policies} \quad \pi_1 \in \Pi_1, \qquad \text{and} \\ -V(x \mid \pi_1^\star, \pi_2^\star) &\geq -V(x \mid \pi_1^\star, \pi_2) \quad \text{for all policies} \quad \pi_2 \in \Pi_2. \end{aligned} \tag{4.53}$$

In words, the policy $\pi_1^\star$ is the best response for player 1 assuming player 2 is playing policy $\pi_2^\star$, and the policy $\pi_2^\star$ is the best response for player 2 assuming player 1 is playing policy $\pi_1^\star$. Thus, neither player has any incentive to deviate from the policy pair $(\pi_1^\star, \pi_2^\star)$. In two-player zero-sum Markov games, a Nash equilibrium always exists, and it is equivalent to the minimax solution [Pat97; Per+15]. Concretely, there exist policies $(\pi_1^\star, \pi_2^\star)$ such that

$$V^\star(x) = V(x \mid \pi_1^\star, \pi_2^\star) = \min_{\pi_1} \max_{\pi_2} V(x \mid \pi_1, \pi_2) = \max_{\pi_1} \min_{\pi_2} V(x \mid \pi_1, \pi_2) \quad \text{for all} \quad x \in \mathcal{X}. \tag{4.54}$$

The function $V^\star$ is known as the *value* of the game.

Q-function and the Bellman fixed-point equation

One method for finding a pair of policies $(\pi_1^\star, \pi_2^\star)$ that achieves the equilibrium (4.54) is by computing the optimal state-action value functions or the optimal Q-function $\theta^\star$. It is known [Pat97; Per+15] to be the fixed point of the Bellman operator

$$\begin{aligned} \mathbf{h}(\theta)(x, u_1, u_2) &= c(x, u_1, u_2) \\ &\quad + \gamma \cdot \sum_{x' \in \mathcal{X}} \mathbf{P}_{u_1, u_2}(x' \mid x) \max_{\pi_1} \min_{\pi_2} \sum_{u_1', u_2'} \pi_1(u_1' \mid x') \cdot \pi_2(u_2' \mid x') \cdot \theta(x', u_1', u_2'). \end{aligned} \tag{4.55}$$

Notably, when the number of states and actions are finite, the minimax problem on the right-hand side of equation (4.55) can be computed by solving the two-player zero-sum matrix game with the payoff matrix $\{\theta(x', u_1, u_2) \mid u_1 \in \mathcal{U}_1, u_2 \in \mathcal{U}_2\}$. Finally, for Markov games with finite state and action spaces, the Q-function θ can be conveniently represented as an element of $\mathbf{R}^{|\mathcal{X}| \times |\mathcal{U}_1| \times |\mathcal{U}_2|}$, and the Bellman operator $\mathbf{h}$ is an operator on $\mathbb{R}^{|\mathcal{X}| \times |\mathcal{U}_1| \times |\mathcal{U}_2|}$.

A simple calculation yields that the Bellman operator is γ-contractive in the ℓ_∞-norm [Pat97; Per+15], and as a result, the optimal Q-function is the unique fixed point of the operator $\mathbf{h}$. We can thus apply our general Banach space theory to derive bounds on the `ROOT-SA` procedure.

The generative model and empirical Bellman operator

We analyze the behavior of the `ROOT-SA` algorithm under a stochastic oracle known as the generative model. A sample from this model consists of a pair of real-valued tensors $(\mathbf{Z}, R)$, each with dimensions $|\mathcal{X}| \times |\mathcal{U}_1| \times |\mathcal{U}_2|$. For each triple (x, u_1, u_2), the entry $\mathbf{Z}_(x, u_1, u_2)$ is drawn according to the transition kernel $\mathbf{P}_{u_1,u_2}(\cdot \mid x)$, whereas the entry $R(x, u_1, u_2)$ is a zero-mean random variable with mean $r(x, u_1, u_2)$, corresponding to a noisy observation of the reward function. The transition and reward samples across entries of the tensors are independently sampled, and we assume that the rewards are bounded in absolute value by $r_{\max}$.

Given a sample $(\mathbf{Z}, R)$ from our observation model, we can define the single-sample empirical Bellman operator

$$\mathbf{H}(\theta)(x, u_1, u_2) := R(x, u_1, u_2) + \sum_{x' \in \mathcal{X}} \mathbf{Z}_{u_1,u_2}(x' \mid x) \max_{\pi_1} \min_{\pi_2} \sum_{u_1', u_2'} \pi_1(u_1' \mid x') \cdot \pi_2(u_2' \mid x') \cdot \theta(x', u_1', u_2'), \quad (4.56)$$

where we have introduced the notation $\mathbf{Z}_{u_1,u_2}(x' \mid x) := \mathbf{1}_{\mathbf{Z}(x,u_1,u_2)=x'}$. With these definitions in hand, we are now ready to state our guarantees for two-player zero-sum Markov games.

Guarantees for two-player zero-sum Markov games

Consider a Gaussian random vector W with $W \sim \mathcal{N}\Big(0, \operatorname{cov}(\mathbf{H}(\theta^*) - \theta^*)\Big)$, we define

$$\mathscr{W} = \mathbb{E}[\|W\|_\infty], \qquad \nu^2 := \sup_{x \in \mathcal{X}, u_1 \in \mathcal{U}_1, u_2 \in \mathcal{U}_2} \mathbb{E}[W^2_{x,u_1,u_2}], \quad \text{and} \quad b_* := r_{\max} + \|\theta^*\|_\infty. \quad (4.57)$$

For a given failure probability $\delta \in (0, 1)$, our result applies to the algorithm with parameters

$$\alpha = c_1 \left\{ \sqrt{n \log |\mathcal{X} \times \mathcal{U}_1 \times \mathcal{U}_2|} \cdot \log(\tfrac{n}{\delta}) \right\}^{-1}, \quad \text{and} \quad B_0 = \tfrac{c_2}{(1-\gamma)^2 \alpha} \log(\tfrac{n}{\delta}). \quad (4.58a)$$

We also choose the initialization θ_0 and the number of restarts R such that

$$\log\left(\tfrac{\|\theta_0 - \mathbf{h}(\theta_0)\|\sqrt{n}}{\mathscr{W}}\right) \leq c_0 \log n \qquad \text{and} \qquad R \geq 2 c_0 \log n \quad (4.58b)$$

for appropriate universal constants c_0, c_1 and c_2. With this setup, a direct application of Theorem 6 yields the following corollary:

Corollary 6. *Given a sample size n such that $\frac{n}{\log n} \geq \frac{c' \log(|\mathcal{X}| \cdot |\mathcal{U}_1| \cdot |\mathcal{U}_2|)}{(1-\gamma)^4} \log(\frac{1}{\delta})$, running Algorithm 5 with the tuning parameter choices (4.58) yields an estimate θ_n such that*

$$\|\mathbf{h}(\theta_n) - \theta_n\|_\infty \leq \tfrac{c}{\sqrt{n}} \cdot \left\{ \mathscr{W} + \nu \sqrt{\log(\tfrac{1}{\delta})} \right\} + \tfrac{c b_*}{1-\gamma} \cdot \tfrac{\log(|\mathcal{X}| \cdot |\mathcal{U}_1| \cdot |\mathcal{U}_2|)}{n} \log^2(\tfrac{n}{\delta}).$$

with probability at least $1 - \delta$.

Note that the bound in Corollary (6) depends on the size of state-action space $|\mathcal{X}| \cdot |\mathcal{U}_1| \cdot |\mathcal{U}_2|$ only poly-logarithmically. Moreover, one can obtain an upper bound on the estimation error $\|\theta_n - \theta^*\|_\infty$ using the bound (4.17).

A special case of interest is when the set of actions for player two is a singleton, i.e., $|\mathcal{U}_2| = 1$. Observe that in this case the optimal state-action value estimation problem for the two-player zero-sum Markov game reduces to the optimal value estimation problem of an appropriate MDP in the discounted setting [Ber19; Wai19e; WD92b]. In Section 4.9, we show that the Bellman operator associated with the optimal value estimation problem of an MDP in the discounted setting satisfies the local linearity assumption (A4). Consequently, an argument similar to Corollary 4 yields an upper bound on the estimation error $\|\theta_n - \theta^*\|_\infty$ which matches the instance dependent lower bound (up to logarithmic terms) from the paper [Kha+21] for large n.[3] Finally, it is an important direction of future work to investigate whether the local linearity assumption (A4) holds when $|\mathcal{U}_2| > 1$.

Average cost policy evaluation

Consider an undiscounted Markov reward process (MRP) with state space $\mathcal{X}$, probability transition kernel $\mathbf{P} \in \mathbb{R}^{\mathcal{X} \times \mathcal{X}}$ and cost function $c : \mathcal{X} \to \mathbb{R}$. When the Markov chain is irreducible and ergodic, there is a unique stationary distribution ξ.

Let $\mu_* := \mathbb{E}_{x \sim \xi}[c(x)]$ denote the average cost under this stationary distribution. Consider the problem of estimating the value function

$$\theta^*(x) := \sum_{n=0}^{\infty} \mathbf{P}^n \Big\{ c(x) - \mu_* \Big\}.$$

The value function θ^* and the average cost μ_* jointly satisfy the Bellman equation

$$\mu_* + \theta^*(x) - \mathbf{P}\theta^*(x) - c(x) = 0 \quad \text{for all } x \in \mathcal{X}. \tag{4.59}$$

See the sources [Der66; TV99] for more background.

In practical application of policy evaluations problems, of primary interest are the relative differences between the value function at different state-actions pairs. In such case, we focus on the estimation problem of θ^*, with μ_* being a nuisance parameter. As shown in the sequel, by considering the span semi-norm in an appropriate vector space $\mathbb{V}$, it is possible to estimate θ^* without estimating μ_*.

Observation models and relevant operators: As before, we consider a generative observation model, where we observe a collection of n i.i.d. samples of the form $\{(\mathbf{Z}_k, C_k)\}_{k=1}^n$, where both $\mathbf{Z}_k \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ and $C_k \in \mathbb{R}^{|\mathcal{X}|}$. For each state $x \in \mathcal{X}$, the row x of the matrix $\mathbf{Z}_k$ is an indicator vector $\mathbf{1}_{s'}$, where the state s' is drawn

[3] The sample size requirement for achieving the lower bound [Kha+21] may depend on the gap between the value of optimal and sub-optimal actions.

according to the transition kernel $\mathbf{P}(\cdot \mid x)$; the entry $C_k(x)$ is a random variable with mean $c(x)$ and uniformly bounded by σ_r, corresponding to a noisy observation of the reward function.

The population and empirical Bellman operators for the average-cost policy evaluation can be written as follows:

$$\mathbf{h}(\theta) := \mathbf{P}\theta + c, \quad \text{and} \quad \mathbf{H}_k(\theta) := \mathbf{Z}_k\theta + C_k.$$

It can be seen that both $\mathbf{h}$ and $\mathbf{H}_k$ are linear operators, satisfying $\mathbb{E}[\mathbf{H}_k] = \mathbf{h}$.

In the rest of this subsection, we define a semi-norm and discuss the multi-step contraction properties of the operator $\mathbf{h}$, and then present the main consequences of Theorem 8 and Corollary 3 for such models.

The semi-norm and multi-step contraction

Consider the Banach space $\mathbb{V}$ given by

$$\mathbb{V} = \mathbb{R}^{|\mathcal{X}|}/\Big\{\theta + \alpha\mathbf{1} \;\mid\; \alpha \in \mathbb{R}\Big\}, \tag{4.60}$$

where each element of $\mathbb{V}$ is an equivalence class of the form $\Big\{\theta+\alpha\mathbf{1} : \alpha \in \mathbb{R}\Big\}$, equipped with the span norm

$$\|\theta\|_{\text{span}} := \max_{x\in\mathcal{X}} \theta(s) - \min_{x\in\mathcal{X}} \theta(s) \quad \text{for all } \theta \in \mathbb{V}.$$

Note that $\|\cdot\|_{\text{span}}$ is a semi-norm on $\mathbb{R}^{\mathcal{X}}$, but a norm on the quotient space $\mathbb{V}$. For reinforcement learning problems, this choice is natural, since we often care only about the relative advantages of state-action pairs, in which case the average cost μ_* is irrelevant.

Under the norm $\|\cdot\|_{\text{span}}$ on $\mathbb{V}$, the operator $\mathbf{h}$ is non-expansive, but not necessarily a contraction. However, under suitable conditions, it can be shown to contractive in a multi-step sense. In order to do so, we impose the following mixing time condition.

Assumption: Mixing time

(MT) There exists a positive integer t_{mix} such that

$$d_{\text{TV}}(\delta_x^\top \mathbf{P}^{t_{\text{mix}}}, \delta_y^\top \mathbf{P}^{t_{\text{mix}}}) \le \tfrac{1}{2} \qquad \text{for any } x, y \in \mathcal{X}.$$

Here the vector $\delta_x \in \mathbb{R}^{\mathcal{X}}$ is the unit basis vector with a single one in entry $x \in \mathcal{X}$.

Under Assumption (MT), for any $\theta \in \mathbb{V}$, we have

$$\begin{aligned}\|\mathbf{P}^{2t_{\text{mix}}}\theta\|_{\text{span}} = \max_{x\in\mathcal{X}} \Big\{\delta_x^\top \mathbf{P}^{2t_{\text{mix}}}\theta\Big\} - \min_{x\in\mathcal{X}} \Big\{\delta_x^\top \mathbf{P}^{2t_{\text{mix}}}\theta\Big\} &\overset{(i)}{\le} 2\max_{x\in\mathcal{X}} \Big|\delta_x^\top \mathbf{P}^{2t_{\text{mix}}}\theta - \xi^\top \mathbf{P}^{2t_{\text{mix}}}\theta\Big| \\ &\le 2 d_{\text{TV}}(\delta_x \mathbf{P}^{2t_{\text{mix}}}, \xi\mathbf{P}^{2t_{\text{mix}}}) \cdot \|\theta\|_{\text{span}} \overset{(ii)}{\le} \tfrac{1}{2}\|\theta\|_{\text{span}}, \end{aligned} \tag{4.61}$$

where in step (i), we use triangular inequality, and in step (ii), we use the fact that $d_{\mathrm{TV}}(\delta_x \mathbf{P}^{2t_{\mathrm{mix}}}, \xi \mathbf{P}^{2t_{\mathrm{mix}}}) \leq \frac{1}{2} d_{\mathrm{TV}}(\delta_x \mathbf{P}^{t_{\mathrm{mix}}}, \xi \mathbf{P}^{t_{\mathrm{mix}}}) \leq \frac{1}{4}$, by applying the mixing time condition (MT) twice.

Consequently, the multi-step contraction assumption (A1)′ holds if the operator is composed $m = 2t_{\mathrm{mix}}$ times.

Estimation error upper bounds

Having defined the norm $\|\cdot\|_{\mathrm{span}}$ and the established the multi-step contraction property (4.61), we are ready to derive a guarantee for average-cost policy evaluation. This involves the Gaussian random variable

$$W \sim \mathcal{N}\left(0, \operatorname{cov}\left(\mathbf{H}(\theta^*) - \theta^*\right)\right),$$

as well as $\mathscr{W} := \mathbb{E}\left[\|W\|_{\mathrm{span}}\right]$. For a given failure probability $\delta \in (0,1)$, our result applies to the algorithm with parameters

$$\alpha = c_1 \left\{ \sqrt{n \log |\mathcal{X}|} \cdot \log(\tfrac{n}{\delta}) \right\}^{-1}, \quad \text{and} \quad B_0 = \tfrac{c t_{\mathrm{mix}}}{\alpha} \log(\tfrac{n}{\delta}). \tag{4.62a}$$

We also choose the initialization θ_0 and the number of restarts R such that

$$\log\left(\tfrac{\|\theta_0 - \mathbf{h}(\theta_0)\|\sqrt{n}}{\mathscr{W}}\right) \leq c_0 \log n \qquad \text{and} \qquad R \geq 2 c_0 \log n, \tag{4.62b}$$

where c, c_0, c_1 are appropriate universal constants. We have the following guarantee:

Corollary 7. *Suppose Assumption (MT) holds, and the sample size n is lower bounded as $\frac{n}{\log^2 n} \geq c' t_{\mathrm{mix}}^2 \log(|\mathcal{X}|) \cdot \log(1/\delta)$. Then the estimate θ_n from Algorithm 5, obtained using tuning parameters satisfying conditions (4.62), satisfies the bound*

$$\begin{aligned} \|\theta_n - \theta^*\|_{\mathrm{span}} \leq \tfrac{c}{\sqrt{n}} \Big\{ & \mathbb{E}\Big[\|(\mathcal{I} - \mathbf{P})^\dagger W\|_{\mathrm{span}}\Big] \\ & + \sqrt{\sup_{x_1, x_2 \in \mathcal{X}} \mathbb{E}\Big[((\delta_{x_1} - \delta_{x_2})(\mathcal{I} - \mathbf{P})^\dagger W)^2\Big] \log(1/\delta)} \Big\} \\ & + c t_{\mathrm{mix}} \left\{ \left[\tfrac{\log|\mathcal{X}|}{n}\right]^{3/4} \mathscr{W} + \tfrac{\log|\mathcal{X}|}{n} \Big(\sigma_r + \|\theta^*\|_{\mathrm{span}}\Big) \right\} \log^2(\tfrac{n}{\delta}), \end{aligned} \tag{4.63}$$

with probability at least $1 - \delta$.

A few remarks are in order. First, the linear operator $(\mathcal{I} - \mathbf{P})$ is not invertible in $\mathbb{R}^{\mathcal{X}}$, with the all-one vector lying in its nullspace. However, it is invertible in the quotient space $\mathbb{V}$, with the pseudo-inverse $(\mathcal{I} - \mathbf{P})^\dagger$ being a representation of its inverse in the coordinate system of $\mathbb{R}^{\mathcal{X}}$, which appears in the bound. Second, as with the previous two cases, the bound depends on the size of state space only poly-logarithmically, and depend quadratically (in the sample size requirement) on the mixing time t_{mix}.

Taking the γ-discounted MRP as a special case of the average-cost framework,[4] Corollary 7 improves the results of the previous paper [Kha+20b] in two aspects: Corollary 7 holds true whenever sample size satisfies $n \gtrsim (1-\gamma)^{-2}$ (omitting log factors), improving upon previous $(1-\gamma)^{-3}$ dependency; and the instance-dependent quantity in the paper [Kha+20b] is replaced with an optimal one matching the local asymptotic minimax limit. Such improvement is made possible by making use of the linear structure in policy evaluation problems. More importantly, Corollary 7 holds true for more general class of problems, where the mixing time t_{mix} replaces the role of effective horizon. Finally, we also note that the $O(t_{\text{mix}}^2)$ sample complexity matches that of the paper [JS20]. While we are providing more instance-dependent guarantees, their results apply to Markov decision processes with actions. It is therefore an important direction of future work to extend our instance-dependent bounds to the case of average-cost MDPs.

4.5 Proofs

This section is devoted to the proofs of our main results—namely, Theorems 6, 7 and 8—along with the associated corollaries. So as to facilitate reading of the proofs, we reproduce here the steps that define Algorithm `ROOT-SA` from Section 4.2.

Algorithm 1 `ROOT-SA` : A recursive SA algorithm

Given (a) Initialization $\theta_0 \in \mathbb{V}$, (b) Burn-in $B_0 \geq 2$, and (c) stepsize $\alpha > 0$
for $t = 1, \ldots, T$ **do**
  if $t \leq B_0$ **then**
    $v_t = \frac{1}{B_0} \sum_{t=1}^{B_0} \{\mathbf{H}_t(\theta_0) - \theta_0\}$, and $\theta_t = \theta_0$.
  else
    $v_t = \mathbf{H}_t(\theta_{t-1}) - \theta_{t-1} + \frac{t-1}{t}(v_{t-1} - \mathbf{H}_t(\theta_{t-2}) + \theta_{t-2})$,
    $\theta_t = \theta_{t-1} + \alpha v_t$.
  end if
end for
return θ_T

Proof of Theorem 6

Our proof is based on a bootstrapping argument, and can be broken down into four steps:

[4] This can be done by adding an absorbing state $\perp$ to the state space. At a rate of $(1-\gamma)$, the Markov process is killed and moved to the absorbing state. In such case, the unique stationary distribution is the singleton at $\perp$, and the mixing time assumption is satisfied with $t_{\text{mix}} = \frac{c}{1-\gamma}$ for universal constant $c > 0$.

1. First, we establish recursions that relate $\|\mathbf{h}(\theta_t) - \theta_t\|$ and $\|v_t\|$.
2. Second, we prove coarse upper bounds on $\|\mathbf{h}(\theta_t) - \theta_t\|$ and $\|v_t\|$.
3. Third, starting with the sub-optimal bounds from step 2, we iteratively refine them using a bootstrapping argument and the recursions from Step 1.
4. Finally, we improve higher-order terms in the bounds.

For the purposes of analysis, it is useful to define the auxiliary sequence

$$z_t := \Big\{\mathbf{h}(\theta_{t-1}) - \theta_{t-1}\Big\} - v_t, \quad \text{for } t = B_0, B_0+1, \cdots. \tag{4.64}$$

The idea is to prove an upper bound on $\|\mathbf{h}(\theta_t) - \theta_t\|$ by proving upper bounds on $\|z_{t+1}\|$ and $\|v_{t+1}\|$.

Let $r_\theta(t)$ and $r_v(t)$, respectively, denote high probability bounds on the quantities $\|\mathbf{h}(\theta_t) - \theta_t\|$ and $\|v_t\|$. It is useful to introduce the notion of an *admissible sequence*: for some $\kappa \geq 0$, the sequence $\{r(t)\}_{t \geq B_0}$ is said to be κ-*admissible* if

- The sequence $\{r(t)\}_{t \geq B_0}$ is non-increasing.
- The sequence $\{t^\kappa \cdot r(t)\}_{t \geq B_0}$ is non-decreasing.

We say that the sequence is admissible if it is κ-admissible for some $\kappa \geq 0$. For notational simplicity, we sometimes use the sequences with time index less than B_0, in such cases, we denote $r_v(t) := r_v(B_0)$ and $r_v(t) := r_\theta(B_0)$ for $t \in [1, B_0]$.

Observe that κ-admissible sequences are also β-admissible sequences for any $\beta > \kappa$. For sake of notational convenience we use the shorthands r_θ and r_v to denote the estimate sequences $\{r_\theta(t)\}_{t \geq B_0}$ and $\{r_v(t)\}_{t \geq B_0}$, respectively. Given an admissible pair (r_θ, r_v) and an integer $n > 0$, we define the events

$$\mathscr{E}_n^{(\theta)}(r_\theta) := \left\{ \sup_{B_0 \leq t \leq n} \frac{\|\mathbf{h}(\theta_t) - \theta_t\|}{r_\theta(t)} \leq 1 \right\}, \quad \text{and} \quad \mathscr{E}_n^{(v)}(r_v) := \left\{ \sup_{B_0 \leq t \leq n} \frac{\|v_t\|}{r_v(t)} \leq 1 \right\}. \tag{4.65}$$

A key portion of our proof involves ensuring that the estimate sequences r_θ and r_v are κ-admissible for carefully chosen values of κ. With these concepts and notation in place, we are now ready to start the main argument.

Step 1: Relation between $\|\mathbf{h}(\theta_t) - \theta_t\|$ and $\|v_t\|$

From the definition (4.64), we have the relation $\mathbf{h}(\theta_t) - \theta_t = z_{t+1} + v_{t+1}$. As mentioned before, we prove an upper bound on $\|\mathbf{h}(\theta_t) - \theta_t\|$ by proving upper bounds on $\|z_{t+1}\|$ and $\|v_{t+1}\|$. We do so using two auxiliary lemmas, the first of which depends on a stepsize α satisfying the bound (cf. the stepsize condition (4.7a))

$$\alpha \leq \frac{(1-\gamma)^2}{cL^2 \mathcal{J}_2^2(\mathbb{B}^*, \rho_n) \log\left(\frac{n}{\delta}\right)}. \tag{4.66}$$

Lemma 18. *Suppose that Assumptions (A1), (A3) and (A2) are in force, and that (r_θ, r_v) are κ-admissible sequences for some $\kappa \in [0,2]$. Then given a stepsize α satisfying the bound (4.66) and a burn-in period $B_0 \geq \frac{100}{(1-\gamma)\alpha}$ conditioned on the event $\mathscr{E}_n^{(v)}(r_v) \cap \mathscr{E}_n^{(\theta)}(r_\theta)$, for each $t \in [B_0, n]$ we have*

$$\|v_t\| \leq \tfrac{1+\gamma}{2} r_v(t) + \tfrac{8}{t\alpha} r_\theta(t) + \tfrac{c}{\sqrt{\alpha}} \Big\{ \mathscr{W} + \nu \sqrt{\log(\tfrac{1}{\delta})} \Big\} + \tfrac{cb_*}{t} \Big\{ \log(\tfrac{1}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n) \Big\} + 6(1-\gamma) \Big(\tfrac{B_0}{t} \Big)^2 \|v_{B_0}\|, \quad (4.67)$$

with probability at least $1 - \delta$.

See Section 4.5 for the proof of this lemma.

Lemma 19. *Under the same conditions as Lemma 18, for each $t \in [B_0, n]$, we have:*

$$\|z_t\| \leq \tfrac{c}{\sqrt{t}} \Big\{ \mathscr{W} + \nu \sqrt{\log(\tfrac{1}{\delta})} \Big\} + \tfrac{b_*}{t} \Big\{ \mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta}) \Big\} + \tfrac{cL}{t} \Big\{ \mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{1}{\delta})} \Big\} \Big\{ \alpha \Big(\sum_{s=B_0}^{t-1} s^2 r_v^2(s) \Big)^{1/2} + \tfrac{1}{1-\gamma} \Big(\sum_{s=1}^{t-1} r_\theta^2(s) \Big)^{1/2} \Big\}, \quad (4.68)$$

with probability $1 - \delta$.

This lemma is a special case of Lemma 22, which is proved in Section 4.5.

Note that although the two lemmas are for a single time index $t \in [B_0, n]$, it is easy to transform them to guarantees that are uniform over $t \in [B_0, n]$. In particular, applying a union bound for $t = B_0, B_0 + 1, \cdots, n$, and by replacing δ with $\delta' = \delta/n$, the bounds (4.67) and (4.68) are valid uniformly over $t \in [B_0, n]$.

We use these two lemmas in our bootstrapping argument. In particular, beginning with the relation $\mathbf{h}(\theta_t) - \theta_t = z_{t+1} + v_{t+1}$, applying the triangle inequality yields the bound $\|\mathbf{h}(\theta_t) - \theta_t\| \leq \|z_{t+1}\| + r_v(t+1)$ on the event $\mathscr{E}_n^{(v)}(r_v)$. Our analysis shows that by starting with an initial estimate $(r_\theta(t), r_v(t))$, the bounds (4.67) and (4.68) allow us to obtain an improved estimate $(r_\theta^+(t), r_v^+(t))$ such that

$$\begin{aligned} \|\mathbf{h}(\theta_t) - \theta_t\| &\leq r_\theta^+(t) < r_\theta(t), \quad \text{and} \\ \|v_t\| &\leq r_v^+(t) < r_v(t) \end{aligned}$$

with high probability. We quantify the improvement in $(r_\theta^+(t), r_v^+(t))$, and repeatedly apply this argument so as to "bootstrap" the bound and ultimately obtain sharp estimates for $r_\theta(t)$ and $r_v(t)$.

Step 2: Setup for the bootstrapping argument

Throughout this step, we require that the estimate sequences r_θ and r_v be $\frac{1}{2}$-admissible and 1-admissible, respectively. As shown in this section, these choices allow us to obtain upper bounds on $\|\mathbf{h}(\theta_t) - \theta_t\|$ and $\|v_t\|$ that decay at the rates $1/\sqrt{t}$ and $1/t$, respectively.

We assume that the pair (r_v^+, r_θ^+) satisfy the initialization condition

$$r_v^+(B_0) \geq \|v_{B_0}\|, \quad \text{and} \quad r_\theta^+(B_0) \geq \|\mathbf{h}(\theta_0) - \theta_0\|, \tag{4.69a}$$

and for each integer $t \in [B_0, n]$, the bounds

$$r_v^+(t) \geq \frac{1+\gamma}{2} r_v(t) + \tfrac{8}{\alpha t} r_\theta(t) + \tfrac{c}{t\sqrt{\alpha}} \Big\{ \mathscr{W} + \nu \sqrt{\log(\tfrac{n}{\delta})} \Big\} + \tfrac{cb_*}{t} \Big\{ \log(\tfrac{n}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n) \Big\} + 6(1-\gamma) \Big(\tfrac{B_0}{t}\Big)^2 \|v_{B_0}\|, \tag{4.69b}$$

and

$$r_\theta^+(t) \geq \Big\{ 1 + c\alpha\sqrt{t} L \Big[\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{n}{\delta})} \Big] \Big\} r_v(t) + \tfrac{2cL}{(1-\gamma)\sqrt{t}} \mathcal{J}_2(\mathbb{B}^*, \rho_n) \log(\tfrac{n}{\delta}) \cdot r_\theta(t) + \tfrac{c}{\sqrt{t}} \Big\{ \mathscr{W} + \nu \sqrt{\log(\tfrac{n}{\delta})} \Big\} + \tfrac{cb_*}{t} \Big\{ \log(\tfrac{n}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n) \Big\}. \tag{4.69c}$$

Under these conditions, by combining the bounds (4.67) and (4.68) and applying a union bound over $t \in [B_0, n]$, we find that

$$\mathbb{P}\Big[\mathscr{E}_n^{(\theta)}(r_\theta^+) \cap \mathscr{E}_n^{(v)}(r_v^+) \Big] \geq \mathbb{P}\Big[\mathscr{E}_n^{(\theta)}(r_\theta) \cap \mathscr{E}_n^{(v)}(r_v) \Big] - \delta,$$

valid for any pair (r_v, r_θ) that are $\frac{1}{2}$ and 1-admissible, respectively.

We consider sequences of a particular form $r_v^{(i)}(t) = \frac{\psi_v^{(i)}}{t\sqrt{\alpha}}$ and $r_\theta^{(i)}(t) = \frac{\psi_\theta^{(i)}}{\sqrt{t}}$, for pairs of positive reals $\big(\psi_v^{(i)}, \psi_\theta^{(i)}\big)$ independent of t. Clearly, with such forms, the sequence $r_\theta^{(i)}$ is $\frac{1}{2}$-admissible, and the sequence $r_v^{(i)}$ is 1-admissible. However, if we directly substitute the sequences $\big(r_v^{(i)}(t), r_v^{(i)}(t)\big)$ of such forms into the relations (4.69b)-(4.69c), the resulting sequences (r_θ^+, r_v^+) are no longer be of the desired form. So in order to unify the coefficients in equations (4.69b)-(4.69c) into the same time scale, given $\alpha > 0$, we define the burn-in time

$$B_0 = \tfrac{c}{(1-\gamma)^2\alpha} \log(\tfrac{n}{\delta}). \tag{4.70a}$$

For each $t = B_0, B_0 + 1 \ldots$, the coefficients in (4.69b) and (4.69c) then satisfy the bounds

$$\tfrac{8}{\alpha t} \leq \tfrac{1-\gamma}{3} \cdot \tfrac{1}{\sqrt{\alpha t}}, \quad \tfrac{1}{\sqrt{\alpha t}} \leq \tfrac{1-\gamma}{6}, \quad \text{and} \quad \tfrac{1}{(1-\gamma)\sqrt{t}} \log(\tfrac{n}{\delta}) \leq \sqrt{\alpha}, \tag{4.70b}$$

Therefore, if we construct a two-dimensional vector sequence $\psi^{(i)} = \begin{bmatrix}\psi_v^{(i)} & \psi_\theta^{(i)}\end{bmatrix}^T$ satisfying the recursive relation $\psi^{(i+1)} = Q\psi^{(i)} + b$, where

$$Q := \begin{bmatrix} \frac{1+\gamma}{2} & \frac{1-\gamma}{3} \\ \frac{1-\gamma}{6} + cL\mathcal{J}_2(\mathbb{B}^*, \rho_n)\sqrt{\alpha}\log(\frac{n}{\delta}) & 2cL\mathcal{J}_2(\mathbb{B}^*, \rho_n)\log(\frac{n}{\delta})\sqrt{\alpha} \end{bmatrix}, \quad \text{and} \tag{4.71a}$$

$$b := \begin{bmatrix} c\Big\{\mathscr{W} + \nu\sqrt{\log(\frac{n}{\delta})}\Big\} + cb_*\sqrt{\alpha}\Big\{\log(\frac{n}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\Big\} + (1-\gamma)B_0\sqrt{\alpha}\,\|v_{B_0}\| \\ \Big\{\mathscr{W} + \nu\sqrt{\log(\frac{n}{\delta})}\Big\} + cb_*\sqrt{\alpha}\Big\{\log(\frac{n}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\Big\} + \sqrt{B_0}\,\|\mathbf{h}(\theta_0) - \theta_0\| \end{bmatrix} \tag{4.71b}$$

satisfy the requirement (4.91). Thus, we are led to the probability bound

$$\mathbb{P}\Big[\mathscr{E}_n^{(\theta)}\big(r_\theta^{(i+1)}\big) \cap \mathscr{E}_n^{(v)}\big(r_v^{(i+1)}\big)\Big] \geq \mathbb{P}\Big[\mathscr{E}_n^{(\theta)}\big(r_\theta^{(i)}\big) \cap \mathscr{E}_n^{(v)}\big(r_v^{(i)}\big)\Big] - \delta, \tag{4.72}$$

for the sequences $r_\theta^{(i)}(t) = \psi_\theta^{(i)}/\sqrt{t}$ and $r_v^{(i)}(t) = \psi_v^{(i)}/(\sqrt{\alpha}t)$. In order to initialize the argument, we need a coarse bound on the pair $(\|v_t\|, \|\mathbf{h}(\theta_t) - \theta_t\|)$; the following lemma provides the requisite bound:

Lemma 20. *Under Assumptions (A3) and (A2), we have*

$$\|\theta_t - \theta^*\| + \|v_t\| \leq e^{1+L\alpha t}\left(b_* + \|\theta_0 - \theta^*\|\right),$$

almost surely for each $t = 0, 1, 2, \ldots$.

See Section 4.8 for the proof of this claim.

Based on Lemma 20, it follows that for each integer $t \in [1, n]$, we have (almost surely) the bound

$$\|v_t\| \leq r_v^{(0)}(t) := \tfrac{n}{t}e^{1+L\alpha t}\Big\{b_* + \|\theta_0 - \theta^*\|\Big\}, \quad \text{and}$$
$$\|\mathbf{h}(\theta_t) - \theta_t\| \overset{(i)}{\leq} r_\theta^{(0)}(t) := 2 \cdot \sqrt{\tfrac{n}{t}}e^{1+L\alpha t}\Big\{b_* + \|\theta_0 - \theta^*\|\Big\},$$

where step (i) follows from the bound $\|\mathbf{h}(\theta_t) - \theta_t\| \leq \|\theta_t - \theta^*\| + \|\mathbf{h}(\theta_t) - \mathbf{h}(\theta^*)\| \leq 2 \cdot \|\theta_t - \theta^*\|$.

By construction, the sequences $r_v^{(0)}$ and $r_\theta^{(0)}$ are 1-admissible and $\frac{1}{2}$-admissible, respectively, and by Lemma 20, the event $\mathscr{E}_n^{(\theta)}\big(r_\theta^{(0)}\big) \cap \mathscr{E}_n^{(v)}\big(r_v^{(0)}\big)$ happens almost surely.

Step 3: Bootstrapping step

Recursing the bound (4.72) for i steps yields

$$\mathbb{P}\Big[\mathscr{E}_n^{(v)}(r_v^{(i)}) \cap \mathscr{E}_n^{(\theta)}(r_\theta^{(i)})\Big] \geq \mathbb{P}\Big[\mathscr{E}_n^{(v)}(r_v^{(0)}) \cap \mathscr{E}_n^{(\theta)}(r_\theta^{(0)})\Big] - i\delta = 1 - i\delta.$$

It remains to analyze the sequence $\psi^{(i)} = \begin{bmatrix} \psi_v^{(i)} & \psi_\theta^{(i)} \end{bmatrix}^T$ as the number of bootstrap steps i increases. We do so by analyzing the recursion relation $\psi^{(i+1)} = Q\psi^{(i)} + b$ with the matrix Q given in equation (4.71).

Observe that the stepsize condition (4.66) ensures that

$$cL\mathcal{J}_2(\mathbb{B}^*, \rho_n) \log(\tfrac{n}{\delta}) \cdot \sqrt{\alpha} \leq \tfrac{1-\gamma}{6}. \tag{4.73}$$

Consequently, the matrix Q from equation (4.66) is entrywise upper bounded by the matrix

$$\widetilde{Q} = \begin{bmatrix} \frac{1+\gamma}{2} & \frac{1-\gamma}{3} \\ \frac{1-\gamma}{3} & \frac{1}{2} \end{bmatrix}$$

This fact implies that for any vector $u \in \mathbb{R}^2$ with non-negative entries, we have the upper bound

$$Qu \preceq_{\text{orth}} \widetilde{Q}u,$$

where $\preceq_{\text{orth}}$ denotes the orthant ordering. Straightforward calculation yields the bound $\|\widetilde{Q}\|_{\text{op}} \leq 1 - \frac{1-\gamma}{8}$. Putting together the pieces, we find that for each $N = 1, 2, \ldots$, conditioned on the event $\mathscr{E}_n^{(v)}(r_v^{(N)}) \cap \mathscr{E}_n^{(\theta)}(r_\theta^{(N)})$, we have

$$\begin{aligned} \psi^{(N)} = \Big(\sum_{i=0}^{N-1} Q^i\Big) b_\psi + Q^N \begin{bmatrix} \psi_v^{(0)} \\ \psi_\theta^{(0)} \end{bmatrix} &\preceq_{\text{orth}} \Big(\sum_{i=0}^{N-1} \widetilde{Q}^i\Big) b + \widetilde{Q}^N \begin{bmatrix} \psi_v^{(0)} \\ \psi_\theta^{(0)} \end{bmatrix} \\ &\preceq_{\text{orth}} (I - \widetilde{Q})^{-1} b + e^{\frac{(1-\gamma)}{8} N} \Big(\psi_v^{(0)} + \psi_\theta^{(0)}\Big) \mathbf{1}_2. \end{aligned}$$

We take $N = \lceil \frac{cLn}{1-\gamma} \log n \rceil$. Replacing δ with δ/N and substituting into the above inequalities then yields

$$\begin{aligned} t\sqrt{\alpha} \cdot \|v_t\| \leq \psi_v^{(N)} \leq \tfrac{c}{1-\gamma}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} + \tfrac{cb_*\sqrt{\alpha}}{1-\gamma}\Big\{\log(\tfrac{n}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\Big\} \\ + cB_0\sqrt{\alpha}\,\|v_{B_0}\| + \sqrt{B_0}\,\|\mathbf{h}(\theta_0) - \theta_0\|, \end{aligned} \tag{4.74a}$$

and

$$\begin{aligned} \sqrt{t}\,\|\mathbf{h}(\theta_t) - \theta_t\| \leq \psi_\theta^{(N)} \leq c\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} + cb_*\sqrt{\alpha}\Big\{\log(\tfrac{n}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\Big\} \\ + cB_0(1-\gamma)\sqrt{\alpha}\,\|v_{B_0}\| + \sqrt{B_0}\,\|\mathbf{h}(\theta_0) - \theta_0\|, \end{aligned} \tag{4.74b}$$

with probability at least $1 - \delta$, uniformly for each $t \in B_0, B_0 + 1, \cdots, n$.

It remains to provide upper bounds on $\|v_{B_0}\|$.

Lemma 21. *Under Assumptions (A1) and (A3), and a burn-in period given by equation (4.70a), we have*

$$\|v_{B_0}\| \leq 2\,\|\mathbf{h}(\theta_0) - \theta_0\| + \tfrac{c}{\sqrt{B_0}}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\Big\} + \tfrac{cb_*}{B_0}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta})\Big\}.$$

with probability at least $1 - \delta$.

See Section 4.8 for the proof.

Combining Lemma 21 and bound (4.74a), we find that

$$\|\mathbf{h}(\theta_t) - \theta_t\| \leq \tfrac{c}{\sqrt{t}}\Big(\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big) + \tfrac{cb_*\sqrt{\alpha}}{\sqrt{t}}\Big\{\log(\tfrac{n}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\Big\} + \|\mathbf{h}(\theta_0) - \theta_0\|\,\tfrac{\sqrt{B_0}}{\sqrt{t}}\log^{\frac{3}{2}}(\tfrac{n}{\delta}) \quad (4.75)$$

with probability at least $1 - \delta$, uniformly for all integers $t \in [B_0, n]$.

Although this bound has optimal dependence on $\mathscr{W} + \nu\sqrt{\log(\frac{n}{\delta})}$, its dependence on the terms $\|\mathbf{h}(\theta_0) - \theta_0\|$ and $\mathcal{J}_1(\mathbb{B}^*, \rho_n)$ and $\log(n/\delta)$ in the bound (4.75) can be sharpened. This motivates the second phase of the bootstrap argument.

Step 4: Improving higher-order terms

Given the pair $(\psi_v^{(N)}, \psi_\theta^{(N)})$ defined by[5] the right-hand side of (4.74), conditioned on the event $\mathscr{E}_n^{(\theta)}(r_\theta) \cap \mathscr{E}_n^{(v)}(r_v)$ with $r_v(t) = \frac{\psi_v^{(N)}}{t\sqrt{\alpha}}$ and the sequence $r_\theta(t) = \frac{\psi_\theta^{(N)}}{\sqrt{t}}$, invoking the bound (4.68) from Lemma 19 we have

$$\begin{aligned}
\|\mathbf{h}(\theta_t) - \theta_t\| &\leq \tfrac{c}{\sqrt{t}}\Big(\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\Big) + \tfrac{cb_*}{t}\Big(\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta})\Big) \\
&\quad + \Big\{\tfrac{1}{t} + c\tfrac{\alpha L\mathcal{J}_2(\mathbb{B}^*, \rho_n)}{\sqrt{t}} \cdot \log(\tfrac{n}{\delta})\Big\}\tfrac{\psi_v^{(N)}}{\sqrt{\alpha}} + 2c\tfrac{L\mathcal{J}_2(\mathbb{B}^*, \rho_n)}{(1-\gamma)t}\log(\tfrac{n}{\delta}) \cdot \psi_\theta^{(N)} \\
&\leq \tfrac{c}{\sqrt{t}}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\Big\} + c'b_*\Big\{\tfrac{1}{(1-\gamma)t} + \tfrac{\alpha L\mathcal{J}_2(\mathbb{B}^*, \rho_n)}{(1-\gamma)\sqrt{t}} \cdot \log(\tfrac{n}{\delta})\Big\}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{n}{\delta})\Big\} \\
&\quad + c'\Big\{\tfrac{1}{(1-\gamma)t\sqrt{\alpha}} + \tfrac{L\mathcal{J}_2(\mathbb{B}^*, \rho_n)\sqrt{\alpha}}{(1-\gamma)\sqrt{t}} \cdot \log(\tfrac{n}{\delta}) + \tfrac{L\mathcal{J}_2(\mathbb{B}^*, \rho_n)}{(1-\gamma)t}\log(\tfrac{n}{\delta})\Big\}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} \\
&\quad + c'\Big\{\tfrac{1}{t\sqrt{\alpha}} + \tfrac{\sqrt{\alpha}L\mathcal{J}_2(\mathbb{B}^*, \rho_n)}{\sqrt{t}} \cdot \log(\tfrac{n}{\delta}) + \tfrac{L\mathcal{J}_2(\mathbb{B}^*, \rho_n)}{(1-\gamma)t}\log(\tfrac{n}{\delta})\Big\})\sqrt{B_0}\,\|\mathbf{h}(\theta_0) - \theta_0\| \\
&\quad + c'\Big\{\tfrac{1}{t} + \tfrac{\alpha L\mathcal{J}_2(\mathbb{B}^*, \rho_n)}{\sqrt{t}} \cdot \log(\tfrac{n}{\delta}) + \tfrac{\sqrt{\alpha}L\mathcal{J}_2(\mathbb{B}^*, \rho_n)}{t}\log(\tfrac{n}{\delta})\Big\}B_0\,\|v_{B_0}\|\,,
\end{aligned}$$

which holds with probability at least $1 - \delta$. Given the burn-in period satisfying equation (4.70a) and stepsize satisfying equation (4.73), by combining with the bound

[5] We redefine $(\psi_v^{(N)}, \psi_\theta^{(N)})$ using the right-hand side of (4.74)

on $\|v_{B_0}\|$ from Lemma 21, we find that $\|\mathbf{h}(\theta_t) - \theta_t\| \leq \widetilde{r}_\theta(t)$ with at least probability $1-\delta$, uniformly for any integer $t \in [n]$, where

$$\widetilde{r}_\theta(t) := \tfrac{c_1}{\sqrt{t}}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} + \tfrac{c_2 b_*}{1-\gamma}\Big\{\tfrac{1}{t} + \tfrac{\alpha L}{\sqrt{t}} \cdot \mathcal{J}_2(\mathbb{B}^*, \rho_n)\log(\tfrac{n}{\delta})\Big\}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{n}{\delta})\Big\} + c_2\Big\{\tfrac{\alpha B_0 L}{\sqrt{t}} \cdot \mathcal{J}_2(\mathbb{B}^*, \rho_n)\log(\tfrac{n}{\delta}) + \tfrac{B_0}{t}\Big\}\|\mathbf{h}(\theta_0) - \theta_0\|. \quad (4.76)$$

By substituting our upper bound in terms of $\widetilde{r}_\theta$ into equation (4.69b), we obtain a recursive inequality that takes an admissible sequence r_v and generates a sequence r_v^+ such that

$$\mathbb{P}\Big[\mathscr{E}_n^{(v)}(r_v^+)\Big] \geq \mathbb{P}\Big[\mathscr{E}_n^{(v)}(r_v)\Big] - \delta.$$

For any positive integer N_1, we can apply the recursive inequality for N_1 times with $\delta' = \delta/N_1$; doing so yields a sharper bound for $\|v_t\|$. In particular, with probability at least $1-\delta$, we have

$$\|v_t\| \leq \tfrac{2}{1-\gamma}\Big\{\tfrac{c}{t\sqrt{\alpha}}\Big[\mathscr{W} + \nu\sqrt{\log(\tfrac{nN_1}{\delta})}\Big] + \tfrac{cb_*}{t}\Big[\log(\tfrac{nN_1}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\Big]\Big\} + \tfrac{8}{(1-\gamma)\alpha t}\widetilde{r}_\theta(t) + \Big(\tfrac{B_0}{t}\Big)^2\|v_{B_0}\| + \Big(\tfrac{1+\gamma}{2}\Big)^{N_1} \cdot \tfrac{\psi_v^{(N)}}{t\sqrt{\alpha}}.$$

We take $N_1 := \lceil \frac{10\log n}{1-\gamma} \rceil$, and a stepsize and burn-in period satisfying the conditions (4.70a) and (4.73). With these choices, some algebra yields $\|v_t\| \leq \widetilde{r_v}(t)$ holds with probability at least $1-\delta$, uniformly for each integer $t \in [B_0, n]$, where

$$\widetilde{r_v}(t) := \tfrac{c'}{1-\gamma}\Big\{\tfrac{1}{t\sqrt{\alpha}}\Big[\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big] + \tfrac{b_*}{t}\Big[\log(\tfrac{n}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\Big]\Big\} + 2\Big(\tfrac{B_0}{t}\Big)^2\|\theta_0 - \mathbf{h}(\theta_0)\|. \quad (4.77)$$

It can be seen that the sequences $\widetilde{r_v}$ and $\widetilde{r}_\theta$ are 2-admissible. Substituting their definitions into the bound (4.68) from Lemma 19 we find that the inequality

$$\|z_t\| \leq \tfrac{c}{\sqrt{t}}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\Big\} + \tfrac{cb_*}{t}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta})\Big\} + \tfrac{cL}{t}\Big\{\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\Big\}\Big\{\alpha\Big(\sum_{s=B_0}^{t-1} s^2 r_v^2(s)\Big)^{1/2} + \frac{1}{1-\gamma} \cdot \Big(\sum_{s=1}^{t-1} r_\theta^2(s)\Big)^{1/2}\Big\}$$

holds with probability at least $1-\delta$.

Under the stepsize and burn-in period conditions (4.70a) and (4.73), some algebra yields:

$$\|z_t\| \leq \tfrac{c}{\sqrt{t}}\Big(\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\Big) + cb_*\Big\{\tfrac{1}{t} + \tfrac{\alpha L}{\sqrt{t}} \cdot \mathcal{J}_2(\mathbb{B}^*, \rho_n)\log(\tfrac{n}{\delta})\Big\}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta})\Big\} + c\tfrac{B_0}{t}\|\theta_0 - \mathbf{h}(\theta_0)\|.$$

Combining with equation (4.77) yields the upper bound

$$\|\mathbf{h}(\theta_t) - \theta_t\| \le \tfrac{c}{\sqrt{t}}\Big(\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\Big) + cb_*\Big\{\tfrac{1}{t} + \tfrac{\alpha L}{\sqrt{t}}\mathcal{J}_2(\mathbb{B}^*, \rho_n)\log(\tfrac{n}{\delta})\Big\}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta})\Big\} + c\tfrac{B_0}{t}\|\theta_0 - \mathbf{h}(\theta_0)\|,$$

which completes the proof of the Theorem 6.

Besides, we also note that by taking a union bound over time steps $t \in \{B_0, B_0 + 1, \ldots, n\}$, we have the lower bound $\mathbb{P}\big[\mathscr{E}_n^{(\theta)}(r_\theta^*)\big] \ge 1 - \delta$, where

$$r_\theta^*(t) := \tfrac{c}{\sqrt{t}}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} + cb_*\Big\{\tfrac{1}{t} + \tfrac{\alpha L \mathcal{J}_2(\mathbb{B}^*, \rho_n)}{\sqrt{t}}\log(\tfrac{n}{\delta})\Big\}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{n}{\delta})\Big\} + c\tfrac{B_0}{t}\|\theta_0 - \mathbf{h}(\theta_0)\|.$$

Proof of Corollary 1

The proof of this corollary is based on a modification of Lemma 19. We introduce the shorthand

$$\mathscr{W} := \mathbb{E}[\|W\|], \qquad \mathscr{W}_C := \mathbb{E}[\big\|W\big\|_C],$$
$$\nu := \sqrt{\sup_{u \in \mathbb{B}^*} \mathbb{E}\big[\langle u, W\rangle^2\big]} \quad \text{and} \quad \nu_C := \sqrt{\sup_{u \in C} \mathbb{E}\big[\langle u, W\rangle^2\big]}.$$

We begin by stating a lemma—a generalization of Lemma 19—that bounds the supremum of an averaged process. In the proof of Corollary 1, we only use a special case of Lemma 22, but the generality is useful later.

Recall the events $\mathscr{E}_n^{(\theta)}(r_\theta)$ and $\mathscr{E}_n^{(v)}(r_v)$ defined in equation (4.65). Given a bounded symmetric convex set $\mathcal{S} \subseteq \mathbb{V}^*$, we define the dimension factor $\mathcal{D}_\mathcal{S} := \sup_{u \in \mathcal{S}} \|u\|_*$. Moreover, we assume that there exists a constant $\mu > 0$ such that

$$\|\theta - \theta^*\| \le \tfrac{1}{\mu}\|\mathbf{h}(\theta) - \theta\| \text{ for any } \theta \in \mathbb{V}. \tag{4.78}$$

We point out that under assumption (A1), the last condition is satisfied for $\mu = 1 - \gamma$. The condition (4.78) also allows us to analysis behavior of operators which satisfies a multi-step contraction assumption (A1)′ (cf. the proof of Theorem 8).

Lemma 22. *Suppose that the Assumptions (A2) and (A3) are in force, the sequences r_θ and r_v are κ-admissible for some $\kappa \in (0, 2]$, and condition (4.78) holds. Then conditioned on the event $\mathscr{E}_n^{(\theta)}(r_\theta) \cap \mathscr{E}_n^{(v)}(r_v)$, we have*

$$\sup_{u \in \mathcal{S}}\langle u, z_t\rangle \le \tfrac{c}{\sqrt{t}}\Big\{\mathbb{E}\big[\sup_{u \in \mathcal{S}}\langle u, W\rangle\big] + \Big(\sup_{u \in \mathcal{S}}\mathbb{E}\big[\langle u, W\rangle^2\big]\log(\tfrac{1}{\delta})\Big)^{1/2}\Big\} + \tfrac{\mathcal{D}_\mathcal{S} b_*}{t}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta})\Big\}$$
$$+ \tfrac{c\mathcal{D}_\mathcal{S} L}{t}\Big\{\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\Big\}\Big\{\alpha\Big(\sum_{s=B_0}^{t-1} s^2 r_v^2(s)\Big)^{1/2} + \frac{1}{\mu}\cdot\Big(\sum_{s=1}^{t-1} r_\theta^2(s)\Big)^{1/2}\Big\}, \tag{4.79}$$

with probability at least $1-\delta$*, uniformly for all integers* $t \in [B_0, n]$.

See Section 4.5 for the proof of this lemma.

Taking this lemma as given, we now proceed with proof of Corollary 1. As mentioned before, under assumption (A1)′, condition (4.78) is satisfied with $\mu = 1-\gamma$. Applying Lemma 22 with $\mathcal{S} = C$ implies that

$$\begin{aligned}\left\|z_t\right\|_C &\leq \tfrac{c}{\sqrt{t}}\Big\{\mathscr{W}_C + \nu_C\sqrt{\log(\tfrac{1}{\delta})}\Big\} + c\tfrac{\mathcal{D}b_*}{t}\Big\{\mathcal{J}_1(\mathbb{B}^*,\rho_n) + \log(\tfrac{1}{\delta})\Big\} \\ &+ \tfrac{c\mathcal{D}L}{t}\Big\{\mathcal{J}_2(\mathbb{B}^*,\rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\Big\}\Big\{\alpha\Big(\sum_{s=B_0}^{t-1} s^2 r_v^2(s)\Big)^{1/2} + \tfrac{1}{1-\gamma}\Big(\sum_{s=1}^{t-1} r_\theta^2(s)\Big)^{1/2}\Big\}. \quad (4.80)\end{aligned}$$

Now all we have to do is substitute an appropriate value of the sequences r_v and r_θ. Note that the estimate sequences $\widetilde{r_v}$ and $\widetilde{r_\theta}$ from equations (4.77) and (4.76), respectively, are 2-admissible; moreover, they provide upper bounds on the quantities $\|v_t\|$ and $\|z_t\|$ respectively. Next, using the stepsize and burn-in conditions (4.73) and (4.70a), we find that

$$\begin{aligned}\Big(\frac{1}{t}\sum_{s=B_0}^{t-1} s^2\widetilde{r_v}^2(s)\Big)^{1/2} &\leq \tfrac{c}{(1-\gamma)\sqrt{\alpha}}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} \\ &\quad + \tfrac{c\cdot b_*}{(1-\gamma)}\Big\{\log(\tfrac{n}{\delta}) + \mathcal{J}_1(\mathbb{B}^*,\rho_n)\Big\} + \tfrac{2c\,{B_0}^{3/2}\,\|\theta_0 - \mathbf{h}(\theta_0)\|}{t},\end{aligned}$$

and

$$\begin{aligned}\Big(\sum_{s=1}^{t-1}\widetilde{r_\theta}^2(s)\Big)^{1/2} &\leq c\cdot\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\}\cdot\sqrt{\log t} \\ &\quad + \tfrac{cb_*}{1-\gamma}\Big\{\tfrac{1}{\sqrt{B_0}} + \alpha L\sqrt{\log t}\Big[\mathcal{J}_2(\mathbb{B}^*,\rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\Big]\Big\}\Big\{\mathcal{J}_1(\mathbb{B}^*,\rho_n) + \log(\tfrac{n}{\delta})\Big\} \\ &\quad + c\Big\{\alpha B_0 L\sqrt{\log(t)}\Big[\mathcal{J}_2(\mathbb{B}^*,\rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\Big] + \sqrt{B_0}\Big\}\cdot\|\mathbf{h}(\theta_0) - \theta_0\|\,;\end{aligned}$$

both with probability at least $1-\delta$. Finally, substituting the last two bounds to the bound (4.80), and applying the conditions on stepsize (4.73) and burn-in period (4.70a), and using $\left\|\mathbf{h}(\theta_t) - \theta_t\right\|_C \leq \left\|z_t\right\|_C + \mathcal{D}\cdot\|v_t\|$ we have

$$\begin{aligned}\left\|\mathbf{h}(\theta_t) - \theta_t\right\|_C &\leq \tfrac{c}{\sqrt{t}}\Big\{\mathscr{W}_C + \nu_C\sqrt{\log(\tfrac{1}{\delta})}\Big\} \\ &\quad + c\tfrac{\mathcal{D}L}{(1-\gamma)}\Big\{\mathcal{J}_2(\mathbb{B}^*,\rho_n)\log(\tfrac{n}{\delta})\sqrt{\tfrac{\alpha}{t}} + \tfrac{1}{t\sqrt{\alpha}}\Big\}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} \\ &\quad + \tfrac{c\mathcal{D}Lb_*}{1-\gamma}\Big\{\tfrac{\sqrt{\alpha}}{t} + \tfrac{\alpha}{\sqrt{t}}\Big\}\mathcal{J}_2(\mathbb{B}^*,\rho_n)\mathcal{J}_1(\mathbb{B}^*,\rho_n)\log^2(\tfrac{n}{\delta}) + \tfrac{\mathcal{D}B_0}{t}\cdot\|\mathbf{h}(\theta_0) - \theta_0\|\,.\end{aligned}$$

This completes the proof of Corollary 1.

Proofs of key Lemmas for Theorem 6

In this section, we provide a detailed proofs of Lemmas 18 and 22 , which play a central role in the proofs of Theorem 6 and Corollary 1.

Proof of Lemma 18

We recursively expand the update rule for v_t from Algorithm 5, and obtain the identity:

$$\begin{aligned}
tv_t &= (t-1)\left(v_{t-1} - \theta_{t-1} + \theta_{t-2} - \mathbf{H}_t(\theta_{t-2}) + \mathbf{H}_t(\theta_{t-1})\right) + \mathbf{H}_t(\theta_{t-1}) - \theta_{t-1} \\
&= (1-\alpha)(t-1)v_{t-1} + (t-1)\Big(\mathbf{H}_t(\theta_{t-1}) - \mathbf{H}_t(\theta_{t-2})\Big) + \Big(\mathbf{H}_t(\theta_{t-1}) - \theta_{t-1}\Big) \\
&= (1-\alpha)^\tau (t-\tau) v_{t-\tau} \\
&\qquad + \sum_{j=1}^{\tau} (1-\alpha)^{j-1}\Big[(t-j)\Big(\mathbf{H}_{t-j+1}(\theta_{t-j}) - \mathbf{H}_{t-j+1}(\theta_{t-j-1})\Big) + \mathbf{H}_{t-j+1}(\theta_{t-j}) - \theta_{t-j}\Big],
\end{aligned}$$

where the positive integer τ will be chosen later.

Consequently, we have the bound

$$\begin{aligned}
t\left\|v_t\right\| &\le (1-\alpha)^\tau (t-\tau)\left\|v_{t-\tau}\right\| + \sum_{j=1}^{\tau}(1-\alpha)^{j-1}(t-j)\left\|\mathbf{h}(\theta_{t-j}) - \mathbf{h}(\theta_{t-j-1})\right\| \\
&\qquad + \Big\| \sum_{j=1}^{\tau}(1-\alpha)^{j-1}\Big((t-j)(\varepsilon_{t-j+1}(\theta_{t-j}) - \varepsilon_{t-j+1}(\theta_{t-j})) + \mathbf{H}_{t-j+1}(\theta_{t-j}) - \theta_{t-j}\Big)\Big\| \\
&\le (1-\alpha)^\tau (t-\tau)\left\|v_{t-\tau}\right\| + \sum_{j=1}^{\tau}(1-\alpha)^{j-1}\Big((t-j)\gamma\alpha\left\|v_{t-j}\right\| + \left\|\mathbf{h}(\theta_{t-j}) - \theta_{t-j}\right\|\Big) \\
&\qquad + \Big\| \sum_{j=1}^{\tau}(1-\alpha)^{j-1}\Big((t-j)(\varepsilon_{t-j+1}(\theta_{t-j}) - \varepsilon_{t-j+1}(\theta_{t-j-1})) + \varepsilon_{t-j+1}(\theta_{t-j})\Big)\Big\|.
\end{aligned}$$

In the following, we prove the case of $t \ge B_0 + 2/\alpha$ and $t \in [B_0, B_0 + 2/\alpha]$ separately. We first consider an iteration index t satisfying the lower bound $t \ge B_0 + 2/\alpha$.

The estimate sequence is r_v κ-admissible for some $\kappa \in [0,2]$, so that the map $t \mapsto t^2 \cdot r_v(t)$ is non-decreasing. Thus, on the event $\mathscr{E}_n^{(v)}(r_v)$ for a burn-in $B_0 \ge \frac{12\tau}{1-\gamma}$, we have the upper bound

$$(t-j)\left\|v_{t-j}\right\| \le (t-j) r_v(t-j) \le \tfrac{t^2}{t-j} r_v(t) \le \tfrac{1}{1-\tau/B_0} \cdot t r_v(t) \le \left\{1 + \tfrac{1-\gamma}{6}\right\} \cdot t r_v(t),$$

valid for each integer $j \in [\tau]$.

Therefore, on the event $\mathscr{E}_n^{(\theta)}(r_\theta) \cap \mathscr{E}_n^{(v)}(r_v)$, we have the bound

$$t\left\|v_t\right\| \le \left(1 + \tfrac{1-\gamma}{6}\right) \cdot \left\{(1-\alpha)^\tau + \gamma\alpha \sum_{j=1}^{\tau}(1-\alpha)^{j-1}\right\} \cdot t r_v(t) + \sum_{j=1}^{\tau} r_\theta(t-j) + T_1 + T_2, \tag{4.81}$$

where

$$T_1 := \Big\| \sum_{j=1}^{\tau} (1-\alpha)^{j-1}(t-j)\Big(\varepsilon_{t-j+1}(\theta_{t-j}) - \varepsilon_{t-j+1}(\theta_{t-j-1})\Big)\Big\|, \quad \text{and}$$
$$T_2 := \Big\| \sum_{j=1}^{\tau} (1-\alpha)^{j-1}\varepsilon_{t-j+1}(\theta_{t-j})\Big\|.$$

We simplify the first two terms on the right-hand side of bound (4.81) by appropriately choosing the triple (τ, α, B_0). The later two terms T_1 and T_2 are norms of zero-mean random vectors in Banach spaces. First, we provide upper bound on these two noise terms.

Upper bound on T_1: First, we observe that the sum consists of the $(1-\alpha)$-weighted differences $(1-\alpha)^{j-1}(t-j)\Big(\varepsilon_{t-j+1}(\theta_{t-j}) - \varepsilon_{t-j+1}(\theta_{t-j-1})\Big)$ that form a martingale difference sequence with respect to the natural filtration $(\mathcal{F}_t)_{t\geq 0}$. On the event $\mathscr{E}_n^{(v)}(r_v)$, we have that

$$\begin{aligned}\Big\|(1-\alpha)^{j-1}(t-j)\Big(\varepsilon_{t-j+1}(\theta_{t-j}) - \varepsilon_{t-j+1}(\theta_{t-j-1})\Big)\Big\| &\leq (t-j)\alpha L r_v(t-j) \\ &\leq \tfrac{t^2}{t-j}\alpha L r_v(t) \leq 2t\alpha L r_v(t), \quad \text{a.s.}\end{aligned}$$

The last inequality is due to the non-decreasing property of the function $t \mapsto t^2 r_v(t)$ and the fact that $t \geq B_0 > 2\tau$.

Since Ω is symmetric and convex by assumption, the difference $\varepsilon_{t-j+1}(\theta_{t-j}) - \varepsilon_{t-j+1}(\theta_{t-j-1})$ belongs to the set 2Ω. Conditioning on the event $\mathscr{E}_n^{(v)}(r_v)$ and invoking Lemma 25 yields

$$\begin{aligned}&\Big\|\tfrac{1}{\tau}\sum_{j=1}^{\tau}(1-\alpha)^{j-1}(t-j)\Big(\varepsilon_{t-j+1}(\theta_{t-j}) - \varepsilon_{t-j+1}(\theta_{t-j-1})\Big)\Big\| \\ &\qquad\qquad \leq \tfrac{ct\alpha L r_v(t)}{\sqrt{\tau}}\Big\{\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\Big\},\end{aligned}$$

with probability at least $1-\delta$.

Upper bound on T_2: In order to bound the last term in the decomposition (4.81), we decompose it into two parts:

$$\begin{aligned}&\sum_{j=1}^{\tau}(1-\alpha)^{j-1}\varepsilon_{t-j+1}(\theta_{t-j}) \\ &\qquad = \sum_{j=1}^{\tau}(1-\alpha)^{j-1}\varepsilon_{t-j+1}(\theta^*) + \sum_{j=1}^{\tau}(1-\alpha)^{j-1}\Big(\varepsilon_{t-j+1}(\theta_{t-j}) - \varepsilon_{t-j+1}(\theta^*)\Big).\end{aligned}$$

The former term is sum of independent random variables, while the latter is a martingale. Note that by Assumption (A1), we have $\|\varepsilon_{t-j+1}(\theta_{t-j}) - \varepsilon_{t-j+1}(\theta^*)\| \leq \frac{Lr_v(t-j+1)}{1-\gamma}$ on the event $\mathscr{E}_n^{(\theta)}(r_\theta)$. Invoking Lemma 24 yields

$$\|\tfrac{1}{\tau}\sum_{j=1}^{\tau}(1-\alpha)^{j-1}\varepsilon_{t-j+1}(\theta^*)\| \leq \tfrac{c}{\sqrt{\tau}}\Big\{\mathscr{W} + \nu\sqrt{\log(1/\delta)}\Big\} + \tfrac{cb_*}{\tau}\Big\{\log(1/\delta) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\Big\},$$

with probability at least $1-\delta$.

Using the Lipschitz assumption (A2) and the contraction assumption (A1), we have

$$\|\varepsilon_{t-j+1}(\theta_{t-j}) - \varepsilon_{t-j+1}(\theta^*)\| \leq L\,\|\theta_{t-j} - \theta^*\| \leq \tfrac{L}{1-\gamma}\,\|\mathbf{h}(\theta_{t-j}) - \theta_{t-j}\|\,.$$

Furthermore, since Ω is symmetric and convex, we have that $\varepsilon_{t-j+1}(\theta_{t-j}) - \varepsilon_{t-j+1}(\theta^*) \in 2\Omega$. Conditioning on the event $\mathscr{E}_n^{(\theta)}(r_\theta)$ and invoking Lemma 25 yields

$$\Big\|\tfrac{1}{\tau}\sum_{j=1}^{\tau}(1-\alpha)^{j-1}\Big(\varepsilon_{t-j+1}(\theta_{t-j}) - \varepsilon_{t-j+1}(\theta^*)\Big)\Big\| \leq \tfrac{c}{\sqrt{\tau}} \cdot \tfrac{Lr_\theta(t-\tau+1)}{1-\gamma}\Big(\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\Big),$$

with probability at least $1-\delta$.

Combining the pieces: Substituting the above concentration bounds into the decomposition (4.81) yields the upper bound

$$\begin{aligned}
t \cdot \|v_t\| \leq{}& \Big\{(1-\alpha)^\tau + \gamma\alpha\sum_{j=1}^{\tau}(1-\alpha)^j\Big\} \cdot \Big\{1 + \tfrac{1-\gamma}{6}\Big\} \cdot tr_v(t) + \tau \cdot r_\theta(t-\tau+1) \\
&+ c\sqrt{\tau} \cdot \tfrac{Lr_\theta(t-\tau+1)}{1-\gamma}\Big\{\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\Big\} + c\sqrt{\tau}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\Big\} \\
&+ cb_*\Big\{\log(\tfrac{1}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\Big\} + c\sqrt{\tau} \cdot t\alpha L r_v(t) \cdot \Big\{\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\Big\}.
\end{aligned}$$

Re-arranging the terms in the last bound yields

$$\begin{aligned}
t \cdot \|v_t\| \leq{}& \Big\{\gamma + (1-\gamma)\Big((1-\alpha)^\tau + \tfrac{1}{3}\Big) + cL\alpha\sqrt{\tau}\Big(\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(1/\delta)}\Big)\Big\}tr_v(t) \\
&+ \Big\{\tau + c\tfrac{L\sqrt{\tau}}{1-\gamma}\Big(\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(1/\delta)}\Big)\Big\}r_\theta(t-\tau+1) \\
&+ c\sqrt{\tau}\Big(\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\Big) + cb_*\Big\{\log(\tfrac{1}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\Big\}.
\end{aligned}$$

Case I: $t \geq B_0 + \lceil 2\alpha^{-1}\rceil$

Taking $\tau = \lceil 2\alpha^{-1}\rceil \leq t - B_0$ and given a stepsize α satisfying the bound

$$6cL\sqrt{\alpha} \cdot \Big(\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{n}{\delta})}\Big) < 1 - \gamma, \tag{4.82}$$

we have the upper bounds

$$\frac{cL}{1-\gamma}\sqrt{\tau}\Big(\mathcal{J}_2(\mathbb{B}^*,\rho_n)+\sqrt{\log(\tfrac{1}{\delta})}\Big)\leq\tfrac{4}{\alpha},\quad\text{and}$$
$$\gamma+(1-\gamma)\cdot\Big\{(1-\alpha)^\tau+\tfrac{1}{3}+cL\alpha\sqrt{\tau}\Big(\mathcal{J}_2(\mathbb{B}^*,\rho_n)+\sqrt{\log(\tfrac{1}{\delta})}\Big)\Big\}\leq\tfrac{1+\gamma}{2}.$$

Furthermore, since the function $t\mapsto t^2\cdot r_\theta(t)$ is non-decreasing, for burn-in period $B_0\geq 4\tau$, we have

$$r_\theta(t-\tau+1)\leq\tfrac{t^2}{(t-\tau+1)^2}r_\theta(t)\leq\tfrac{16}{9}r_\theta(t)\qquad\text{for all}\qquad t\geq B_0.$$

Substituting with above bounds, the recursive bound becomes:

$$t\,\|v_t\|\leq\tfrac{1+\gamma}{2}t\cdot r_v(t)+\tfrac{8}{\alpha}r_\theta(t)+\tfrac{c}{\sqrt{\alpha}}\Big(\mathscr{W}+\nu\sqrt{\log(\tfrac{1}{\delta})}\Big)+c\left(\log(\tfrac{1}{\delta})+\mathcal{J}_1(\mathbb{B}^*,\rho_n)\right),\tag{4.83}$$

which completes the proof of Lemma 18 in the case of $t\geq B_0+2/\alpha$.

Case II: $B_0\leq t\leq B_0+\lceil 2\alpha^{-1}\rceil$**:**

This case deserves a special treatment as the number τ of recursive expansion steps cannot be taken as large as $\lceil 2/\alpha\rceil$. Instead, we choose $\tau=t-B_0$, and expand the recursions backwards up to the beginning of the iterates. In particular, following the same arguments as above, on the event $\mathscr{E}_n^{(\theta)}(r_\theta)\cap\mathscr{E}_n^{(v)}(r_v)$, the error decomposition (4.81) in this case becomes:

$$t\cdot\|v_t\|\leq(1-\alpha)^\tau B_0\,\|v_{B_0}\|+\Big(1+\tfrac{1-\gamma}{6}\Big)\cdot\Big\{\gamma\alpha\sum_{j=1}^{\tau}(1-\alpha)^{j-1}\Big\}\cdot tr_v(t)$$
$$+\sum_{j=1}^{\tau}r_\theta(t-j)+T_1+T_2.$$

Substituting with the upper bounds on the terms T_1 and T_2, we obtain that:

$$t\cdot\|v_t\|\leq(1-\alpha)^\tau B_0\,\|v_{B_0}\|+\gamma\alpha\sum_{j=1}^{\tau}(1-\alpha)^j\cdot\Big\{1+\tfrac{1-\gamma}{6}\Big\}\cdot tr_v(t)+\tau\cdot r_\theta(t-\tau+1)$$
$$+c\sqrt{\tau}\cdot\tfrac{Lr_\theta(t-\tau+1)}{1-\gamma}\Big\{\mathcal{J}_2(\mathbb{B}^*,\rho_n)+\sqrt{\log(\tfrac{1}{\delta})}\Big\}+c\sqrt{\tau}\Big\{\mathscr{W}+\nu\sqrt{\log(\tfrac{1}{\delta})}\Big\}$$
$$+cb_*\Big\{\log(\tfrac{1}{\delta})+\mathcal{J}_1(\mathbb{B}^*,\rho_n)\Big\}+c\sqrt{\tau}\cdot t\alpha Lr_v(t)\cdot\Big\{\mathcal{J}_2(\mathbb{B}^*,\rho_n)+\sqrt{\log(\tfrac{1}{\delta})}\Big\}.$$

For a time index $t\in[B_0,B_0+2/\alpha]$, we have the decomposition

$$(1-\alpha)^\tau B_0\cdot\|v_{B_0}\|\leq((1-\alpha)^\tau-3(1-\gamma))\cdot q\Big\{1+\frac{2}{\alpha B_0}\Big\}t\cdot r_v(t)+3(1-\gamma)B_0\cdot\|v_{B_0}\|$$
$$\leq\Big\{(1-\alpha)^\tau-2(1-\gamma)\Big\}\cdot tr_v(t)+6(1-\gamma)\frac{{B_0}^2}{t}\cdot\|v_{B_0}\|$$

Given a stepsize α satisfying the requirement (4.66), choosing the number of steps such that $\tau = t - B_0 \leq 2/\alpha$ leads to the inequalities

$$\frac{1+\gamma}{2} \geq \Big\{(1-\alpha)^\tau - 2(1-\gamma)\Big\} + \gamma\alpha \sum_{j=1}^{\tau} (1-\alpha)^j \cdot \Big\{1 + \tfrac{1-\gamma}{6}\Big\} + c\sqrt{\tau} \cdot \alpha L \cdot \Big\{\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\Big\}$$

and

$$\frac{cL}{1-\gamma}\sqrt{\tau}\Big(\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(1/\delta)}\Big) \leq \frac{4}{\alpha} + \tau \leq \frac{8}{\alpha}.$$

Putting together these bounds completes the proof in the second case.

Proof of Lemma 22

Expanding the update rule for z_t from Algorithm 5 we obtain the three-term decomposition $t \cdot z_t = B_0 \cdot z_{B_0} + M_t + \Psi_t$, where

$$M_t := \sum_{s=B_0}^{t-1} \varepsilon_s(\theta_{s-1}), \quad \text{and} \quad \Psi_t := \sum_{s=B_0}^{t-1} (s-1)\Big\{\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2})\Big\}.$$

It suffices to control each of these three terms in the semi-norm induced by the set $\mathcal{S}$.

Beginning with the martingale $\{M_t\}_{t \geq B_0}$, we further break it down into two parts:

$$M_t = \sum_{s=B_0}^{t-1} \varepsilon_s(\theta^*) + \sum_{s=B_0}^{t-1} \Big(\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta^*)\Big) := M_t^* + \widetilde{M}_t.$$

The term M_t^* is sum of i.i.d. random variables. Invoking Lemma 24 and using the fact that the set $\mathcal{S}$ is contained within $\mathcal{D}_{\mathcal{S}}\mathbb{B}^*$, we have the bound

$$\sup_{u \in \mathcal{S}} \langle u,\, M^*(t)\rangle \leq c\sqrt{t}\Big\{\mathbb{E}\Big[\sup_{u \in \mathcal{S}} \langle u,\, W\rangle\Big] + \Big(\sup_{u \in v} \mathbb{E}\Big[\langle u,\, W\rangle^2\Big] \cdot \log(\tfrac{1}{\delta})\Big)^{1/2}\Big\} + c\mathcal{D}_{\mathcal{S}} b_* \Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta})\Big\}, \quad (4.84)$$

where W is a centered Gaussian process with covariance matching that of $\varepsilon(\theta^*)$ (cf. equation (4.4)).

Next we bound the terms $\langle u,\, \widetilde{M}(t)\rangle$ and $\langle u,\, \Psi(t)\rangle$. First, we claim that conditioned on the event $\mathscr{E}_n^{\varepsilon(\theta)}(r_\theta) \cap \mathscr{E}_n^{\varepsilon(v)}(r_v)$, we have

$$\|\widetilde{M}(t)\| \leq c\frac{L}{\mu}\Big\{\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{n}{\delta})}\Big\} \cdot \Big(\sum_{k=B_0}^{t-1} r_\theta^2(k)\Big)^{1/2}, \quad \text{and} \quad (4.85a)$$

$$\|\Psi(t)\| \leq c\alpha L\Big\{\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{n}{\delta})}\Big\} \cdot \Big(\sum_{s=B_0}^{t} s^2 r_v^2(s)\Big)^{1/2}, \quad (4.85b)$$

both bounds holding with probability at least $1-\delta$.

The proof of these two inequalities can be found at the end of this subsection. Since the set $\mathcal{S}$ is contained within $\mathcal{D}_{\mathcal{S}}\mathbb{B}^*$, it follows that

$$\sup_{u\in\mathcal{S}}\langle u,\, \widetilde{M}(t)\rangle \leq \mathcal{D}_{\mathcal{S}}\|\widetilde{M}(t)\|, \quad \text{and} \quad \sup_{u\in\mathcal{S}}\langle u,\, \Psi(t)\rangle \leq \mathcal{D}_{\mathcal{S}}\,\|\Psi(t)\|\,.$$

Finally, observe that $B_0 z_{B_0} = \sum_{t=1}^{B_0}\varepsilon_t(\theta^*) + \sum_{t=1}^{B_0}\{\varepsilon_t(\theta_0)-\varepsilon_t(\theta^*)\}$. By Lemma 24 we have

$$\sup_{u\in\mathcal{S}}\langle u,\, \sum_{t=1}^{B_0}\varepsilon_t(\theta^*)\rangle \leq c\sqrt{B_0}\Big\{\mathbb{E}\Big[\sup_{u\in\mathcal{S}}\langle u,\, W\rangle\Big] + \Big(\sup_{u\in v}\mathbb{E}\Big[\langle u,\, W\rangle^2\Big]\cdot\log(\tfrac{1}{\delta})\Big)^{1/2}\Big\} + c\mathcal{D}_{\mathcal{S}}b_*\Big\{\mathcal{J}_1(\mathbb{B}^*,\rho_n)+\log(\tfrac{1}{\delta})\Big\},$$

with probability at least $1-\delta$. On the other hand, using Lemma 25, we have

$$\begin{aligned}\sup_{u\in\mathcal{S}}\langle u,\, \sum_{t=1}^{B_0}\Big(\varepsilon_t(\theta_0)-\varepsilon_t(\theta^*)\Big)\rangle &\leq \mathcal{D}_{\mathcal{S}}\|\sum_{t=1}^{B_0}\Big\{\varepsilon_t(\theta_0)-\varepsilon_t(\theta^*)\Big\}\| \\ &\leq cL\mathcal{D}_{\mathcal{S}}\|\theta_0-\theta^*\|\cdot\sqrt{B_0}\Big\{\mathcal{J}_2(\mathbb{B}^*,\rho_n)+\sqrt{\log(\tfrac{1}{\delta})}\Big\} \\ &\leq c\mathcal{D}_{\mathcal{S}}\cdot\frac{L}{\mu}\sqrt{B_0}r_\theta(B_0)\Big\{\mathcal{J}_2(\mathbb{B}^*,\rho_n)+\sqrt{\log(\tfrac{1}{\delta})}\Big\},\end{aligned}$$

with probability at least $1-\delta$. Combining the two bounds, we conclude that

$$\begin{aligned}\sup_{u\in\mathcal{S}}\langle u,\, z_{B_0}\rangle \leq \tfrac{c}{\sqrt{B_0}}\Big\{\mathbb{E}\Big[\sup_{u\in\mathcal{S}}\langle u,\, W\rangle\Big] + \Big(\sup_{u\in v}\mathbb{E}\Big[\langle u,\, W\rangle^2\Big]\cdot\log(\tfrac{1}{\delta})\Big)^{1/2}\Big\} \\ + \tfrac{c\mathcal{D}_{\mathcal{S}}b_*}{B_0}\Big(\mathcal{J}_1(\mathbb{B}^*,\rho_n)+\log(\tfrac{1}{\delta})\Big) + \tfrac{c\mathcal{D}_{\mathcal{S}}\cdot Lr_\theta(B_0)}{\mu\sqrt{B_0}}\Big(\mathcal{J}_2(\mathbb{B}^*,\rho_n)+\Big(\log(\tfrac{1}{\delta})\Big)^{1/2}\Big), \quad (4.86)\end{aligned}$$

again with at least probability $1-\delta$.

We now put together the bounds (4.84), (4.85a) (4.85b), and (4.86). By doing so, we are guaranteed that conditioned on the event $\mathscr{E}_n^{(\theta)}(r_\theta)\cap\mathscr{E}_n^{(v)}(r_v)$, for each integer $t\in[B_0,n]$, we have

$$\begin{aligned}\sup_{u\in\mathcal{S}}\langle u,\, z_t\rangle \leq \tfrac{c}{\sqrt{t}}\Big\{\mathbb{E}\Big[\sup_{u\in\mathcal{S}}\langle u,\, W\rangle\Big] + \Big(\sup_{u\in\mathcal{S}}\mathbb{E}\Big[\langle u,\, W\rangle^2\Big]\log(\tfrac{1}{\delta})\Big)^{1/2}\Big\} \\ + \tfrac{\mathcal{D}_{\mathcal{S}}b_*}{t}\Big\{\mathcal{J}_1(\mathbb{B}^*,\rho_n)+\log(\tfrac{1}{\delta})\Big\} \\ + c\tfrac{\mathcal{D}_{\mathcal{S}}\cdot\alpha L}{\sqrt{t}}\Big\{\mathcal{J}_2(\mathbb{B}^*,\rho_n)+\sqrt{\log(\tfrac{1}{\delta})}\Big\}\Big(\frac{1}{t}\sum_{s=B_0}^{t-1}s^2r_v^2(s)\Big)^{1/2} \\ + 2c\tfrac{\mathcal{D}_{\mathcal{S}}\cdot L}{\mu t}\Big\{\mathcal{J}_2(\mathbb{B}^*,\rho_n)+\sqrt{\log(\tfrac{1}{\delta})}\Big\}\cdot\Big(\sum_{s=B_0}^{t-1}r_\theta^2(s)+B_0r_\theta^2(B_0)\Big)^{1/2}.\end{aligned}$$

The claim of Lemma 22 now follows by noting $\Big(\sum_{s=B_0}^{t-1} r_\theta^2(s) + B_0 r_\theta^2(B_0)\Big)^{1/2} = \Big(\sum_{s=1B_0}^{t-1} r_\theta^2(s)\Big)^{1/2}$. It remains to prove inequalities (4.85a) and (4.85b).

Proof of the bound (4.85a): Conditioned on the event $\mathscr{E}_n^{(\theta)}(r_\theta)$, we have the upper bounds

$$\begin{aligned}\|\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta^*)\| = \|\mathbf{H}_s(\theta_{s-1}) - \mathbf{H}_s(\theta^*) - \mathbf{h}(\theta_{s-1}) + \mathbf{h}(\theta^*)\| &\leq L\,\|\theta_{s-1} - \theta^*\| \\ &\leq \frac{L}{\mu} r_\theta(s-1),\end{aligned}$$

where the last inequality follows from the assumption $\|\theta_{s-1} - \theta^*\| \leq \frac{1}{\mu}\|\mathbf{h}(\theta_{s-1}) - \mathbf{h}(\theta^*)\|$ (cf. assumption (4.78)).

On the event $\mathscr{E}_n^{(\theta)}(r_\theta)$, we apply Lemma 25 to the martingale differences $\{\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta^*)\}_{s=B_0+1}^t$, and find that

$$\Big\| \sum_{s=B_0+1}^{t} \big(\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta^*)\big)\Big\| \leq c\Big\{\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\Big\}\,\frac{L}{\mu}\,\Big(\sum_{s=B_0}^{t-1} r_\theta^2(s)\Big)^{1/2},$$

with probability at least $1-\delta$, as claimed in inequality (4.85a).

Proof of bound (4.85b): We now control the martingale sequence $\{\Psi_t\}_{t\geq B_0}$. Conditioned on the event $\mathscr{E}_n^{(v)}(r_v)$, we have

$$\begin{aligned}\Big\|(s-1)\Big\{\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2})\Big\}\Big\| \leq (s-1)L\cdot\|\theta_{s-1} - \theta_{s-2}\| &= (s-1)\alpha L\,\|v_{s-1}\| \\ &\leq (s-1)\alpha L r_v(s-1),\end{aligned}$$

valid for any integer $s \in [B_0, t]$. By Lemma 25, on the event $\mathscr{E}_n^{(v)}(r_v)$, we have

$$\Big\| \sum_{s=B_0+1}^{t} (s-1)\Big\{\varepsilon_s(\theta_{s-1}) - \varepsilon_s(\theta_{s-2})\Big\}\Big\| \leq c\alpha\Big\{\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\Big\} L\Big(\sum_{s=B_0}^{t-1} s^2 r_v^2(s)\Big)^{1/2},$$

with probability at least $1-\delta$, which establishes the claim. This completes the proof of Lemma 19.

Proofs of Theorem 7 and Corollary 2

In this section, we prove Theorem 7 and Corollary 2. Note that Corollary 2 is a generalized version of Theorem 7. Indeed, assuming conditions (A1)- (A4) holds, Theorem 7 follows from Corollary 2 with $\|\cdot\|_C = \|\cdot\|$.

Define the pair

$$r_\theta^*(t) := \tfrac{c}{\sqrt{t}}\Big(\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big) + \tfrac{cb_*}{1-\gamma}\Big\{\tfrac{1}{t} + \tfrac{\alpha L}{\sqrt{t}} \cdot \mathcal{J}_2(\mathbb{B}^*, \rho_n)\log(\tfrac{n}{\delta})\Big\}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta})\Big\} + c\tfrac{B_0}{t}\left\|\theta_0 - \mathbf{h}(\theta_0)\right\|, \quad \text{and} \quad (4.87a)$$

$$r_v^*(t) := \tfrac{c}{1-\gamma}\Big\{\tfrac{1}{t\sqrt{\alpha}}\Big(\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big) + \tfrac{b_*}{t}\Big\{\log(\tfrac{n}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\Big\}\Big\} + 2\Big(\tfrac{B_0}{t}\Big)^2\left\|\theta_0 - \mathbf{h}(\theta_0)\right\|. \quad (4.87b)$$

Invoking Theorem 6 and applying a union bound over the iterates, we have the pair of bounds $\|\mathbf{h}(\theta_t) - \theta_t\| \leq r_\theta^*(t)$, and and $\|v_t\| \leq r_v^*(t)$, uniformly for $t = B_0, B_0+1, \cdots, n$ with probability at least $1-\delta$. Using the restarting scheme with parameter choice (4.11b), we can guarantee that the initial operator defect $\|\mathbf{h}(\theta_0) - \theta_0\|$ satisfies the upper bound:

$$\|\theta_0 - \mathbf{h}(\theta_0)\| \leq \tfrac{c}{\sqrt{B_0}}\Big(\mathscr{W} + \nu\sqrt{\log(1/\delta)}\Big) + \tfrac{cb_*}{B_0(1-\gamma)}\Big(\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(1/\delta)\Big).$$

By the linearization condition (A4)′, for any $\theta \in \mathbb{B}(\theta^*, s_0)$, we have

$$s := \left\|\theta - \theta^*\right\|_C \;\leq\; \sup_{A \in \mathcal{A}_s}\left\|(I-A)^{-1}(\mathbf{h}(\theta) - \theta)\right\|_C.$$

To obtain an upper bound on $\left\|\theta_n - \theta^*\right\|_C$, it suffices to provide an bound for the quantity $\sup_{A\in\mathcal{A}_s}\left\|(I-A)^{-1}(\mathbf{h}(\theta_{n-1}) - \theta_{n-1})\right\|_C$ for any given $s > 0$. Recall that $h(\theta_{n-1}) - \theta_{n-1} = v_n - z_n$, by definition, and in the rest of this section we provide upper bounds on $\left\|v_n\right\|_C$ and $\left\|z_n\right\|_C$

Upper bound on $\left\|v_n\right\|_C$

Observe that $\left\|(I-A)^{-1}v_n\right\|_C \leq \frac{\mathcal{D}}{1-\gamma}\|v_n\|$. Thus, if we invoke the bound $\|v_n\| \leq r_v^*(t)$, where r_v^* is defined in (4.87b), we are guaranteed that

$$\left\|(I-A)^{-1}v_n\right\|_C \leq \tfrac{c'\mathcal{D}}{(1-\gamma)^2}\Big\{\tfrac{1}{n\sqrt{\alpha}}\Big[\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big] + \tfrac{b_*}{n}\Big[\log(\tfrac{n}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\Big]\Big\} + \tfrac{c\mathcal{D}}{1-\gamma}\Big\{\tfrac{B_0}{n}\Big\}^2\left\|\theta_0 - \mathbf{h}(\theta_0)\right\|.$$

with probability at least $1-\delta$.

Upper bound on $\|z_n\|_C$

In order to establish a sharp upper bound on the term $\sup_{A\in\mathcal{A}_s}\|(I-A)^{-1}z_{n+1}\|$, we define the class of test functions $\mathcal{S} := \{(I-A^*)^{-1}u : \ A\in\mathcal{A}_s, u\in C\}$. Substituting the the bounds (4.87a) and (4.87b) in Lemma 22 with we find that for any given $s>0$, the quantity $\sup_{A\in\mathcal{A}_s}\|(I-A)^{-1}z_n\|_C$ is upper bounded as

$$\begin{aligned}\tfrac{c}{\sqrt{n}}\Big\{\mathbb{E}\big[\sup_{A\in\mathcal{A}_s}\|(I-A)^{-1}W\|_C\big]+\nu(s)\sqrt{\log(\tfrac{1}{\delta})}\Big\}+\tfrac{cb_*\mathcal{D}}{n(1-\gamma)}\Big\{\mathcal{J}_1(\mathbb{B}^*,\rho_n)+\log(\tfrac{1}{\delta})\Big\}\\ +\tfrac{cL\mathcal{D}}{(1-\gamma)^2n}\Big\{\mathcal{J}_2(\mathbb{B}^*,\rho_n)+\sqrt{\log(\tfrac{1}{\delta})}\Big\}\Big\{\alpha\Big(\sum_{s=B_0}^{n-1}s^2r_v^{*2}(s)\Big)^{1/2}+\tfrac{1}{1-\gamma}\Big(\sum_{s=1}^{n-1}r_\theta^{*2}(s)\Big)^{1/2}\Big\},\end{aligned} \tag{4.88}$$

with probability at least $1-\delta$.

Putting together the pieces

The last two bounds are valid for a fixed value of s. In order to derive the fixed-point condition in Theorem 7 and Corollary 2, however, we need a bound that holds uniformly over s in a suitable range, which we now do. Define the quantity

$$\overline{R}_n := \tfrac{c}{(1-\gamma)\sqrt{n}}\Big\{\mathbb{E}\big[\|W\|_C\big]+\Big(\sup_{u\in C}\mathbb{E}\big[\langle u,\,W\rangle^2\big]\log(\tfrac{1}{\delta})\Big)^{1/2}\Big\}+\mathcal{D}\cdot\mathcal{H}_n(\alpha,\delta),$$

and let $\underline{R} := \frac{1-\gamma}{1+\gamma}\overline{R}$.

It can be seen that the solutions to equation (4.30) all belong to the interval $[\underline{R},\overline{R}]$. In particular, contraction assumption (A1), we find that

$$\frac{1}{1+\gamma}\le\|\!|(I-A)^{-1}|\!\|_{\mathrm{op}}\le\frac{1}{1-\gamma},\qquad\text{valid for any } A\in\mathcal{A}_s,$$

which leads to the bounds $\frac{\mathscr{W}_C}{1+\gamma}\le\mathcal{G}_C(s)\le\frac{\mathscr{W}_C}{1-\gamma}$ and $\frac{\nu_C}{1+\gamma}\le\sigma_{*,C}(s)\le\frac{\nu_C}{1-\gamma}$ for any $s>0$.

By Theorem 6 and the contractive assumption (A1), we have the upper bound

$$\mathbb{P}\big[\|\theta_n-\theta^*\|_C\ge\overline{R}_n\big]\le\delta.$$

Consider the sequence $s_\ell = 2^{\ell-1}\underline{R}_n$ for $\ell=1,2,\ldots,k$, where $k := \log_2(\lceil\overline{R}_n/\underline{R}_n\rceil)$. It forms a doubling grid $\mathcal{M}_n := \{s_1,s_2,\cdots,s_k\}$ on the interval $[\underline{R}_n,\overline{R}_n]$, and it can be seen that k satisfies the upper bound

$$k\le\log(\tfrac{1+\gamma}{1-\gamma})\le 1+\log(\tfrac{1}{1-\gamma}).$$

Taking a union bound over $s\in\mathcal{M}_n$, we find that the bound (4.88) holds with probability at least $1-k\delta$, uniformly over $s\in\mathcal{M}_n$. For any $s\in\big[\underline{R}_n,\overline{R}_n\big]$, define the

index $\ell(s) := \max\{\ell \mid s_\ell \le s\}$. On the event above, we can conclude that

$$\begin{aligned}
&\sup_{A\in\mathcal{A}_s} \left\|(I-A)^{-1}z_n\right\|_C \le \sup_{A\in\mathcal{A}_{s_{\ell(s)+1}}} \left\|(I-A)^{-1}z_n\right\|_C \\
&\le \tfrac{c}{\sqrt{n}}\Big\{\mathcal{G}_C(2s)+\sigma_{*,C}(2s)\sqrt{\log(\tfrac{1}{\delta})+\log(\log n)}\Big\}+\tfrac{c\mathcal{D}b_*}{(1-\gamma)n}\Big\{\mathcal{J}_1(\mathbb{B}^*,\rho_n)+\log(\tfrac{1}{\delta})+\log(\log n)\Big\} \\
&\quad + \tfrac{c\mathcal{D}L}{(1-\gamma)^2 n}\Big\{\mathcal{J}_2(\mathbb{B}^*,\rho_n)+\sqrt{\log(\tfrac{1}{\delta})}\Big\}\Big\{\alpha\Big(\sum_{s=B_0}^{n-1} s^2 r_v^{*2}(s)\Big)^{1/2} + \tfrac{1}{1-\gamma}\Big(\sum_{s=1}^{n-1} r_\theta^{*2}(s)\Big)^{1/2}\Big\},
\end{aligned}$$

for $s \in [\underline{R}_n, \overline{R}_n]$. Here we have used the facts that $\mathcal{G}_C(\cdot)$ and $\sigma_{*,C}(\cdot)$ are non-decreasing functions. We now substitute our expressions for r_θ^* and r_v^*, and conclude that conditioned on the event $\mathscr{E}_n^{(\theta)} \cap \mathscr{E}_n^{(v)} \cap \Big\{\|\theta_n - \theta^*\| \le \overline{R}_n\Big\}$, we have

$$\begin{aligned}
s_n &\le \sup_{A\in\mathcal{A}_{s_n}} \left\|(I-A)^{-1}z_n\right\|_C + \tfrac{\mathcal{D}}{1-\gamma} r_v^*(n) \\
&\le \tfrac{c}{\sqrt{n}}\Big(\mathcal{G}(2s_n)+\nu(2s_n)\sqrt{\log(\tfrac{1}{\delta})}\Big) + \underline{R}_n + \tfrac{c\mathcal{D}b_*}{(1-\gamma)n}\Big(\mathcal{J}_1(\mathbb{B}^*,\rho_n)+\log(\tfrac{1}{\delta})\Big) \\
&\quad + \tfrac{c\mathcal{D}L}{(1-\gamma)^2\sqrt{n}}\cdot\mathcal{J}_2(\mathbb{B}^*,\rho_n)\log(\tfrac{n}{\delta})\cdot\Big(\sqrt{\alpha}\mathscr{W} + \alpha b_*\Big\{\mathcal{J}_1(\mathbb{B}^*,\rho_n)+\log(\tfrac{n}{\delta})\Big\} \\
&\qquad\qquad + \sqrt{\tfrac{B_0}{n}}\,\|\theta_0 - \mathbf{h}(\theta_0)\|\Big) \\
&\qquad + \tfrac{c'\mathcal{D}}{(1-\gamma)^2}\Big\{\tfrac{1}{n\sqrt{\alpha}}\mathscr{W}\sqrt{\log\tfrac{n}{\delta}} + \tfrac{b_*}{n}\Big[\log\tfrac{n}{\delta}+\mathcal{J}_1(\mathbb{B}^*,\rho_n)\Big]\Big\} + \tfrac{c\mathcal{D}}{1-\gamma}\Big(\tfrac{B_0}{n}\Big)^2\|\theta_0-\mathbf{h}(\theta_0)\| \\
&\le \tfrac{c}{\sqrt{n}}\Big\{\mathcal{G}_C(2s_n)+\sigma_{*,C}(2s_n)\sqrt{\log(\tfrac{1}{\delta})}\Big\} + \tfrac{c\mathcal{D}\log(n/\delta)}{(1-\gamma)^2}\Big\{\mathcal{J}_2(\mathbb{B}^*,\rho_n)L\sqrt{\tfrac{\alpha}{n}}+\tfrac{1}{n\sqrt{\alpha}}\Big\}\cdot\mathscr{W} \\
&\quad + \tfrac{c\mathcal{D}b_*\log(\frac{n}{\delta})}{(1-\gamma)^2}\Big\{\mathcal{J}_2(\mathbb{B}^*,\rho_n)L\tfrac{\alpha}{\sqrt{n}}+\tfrac{1}{n}\Big\}\cdot\Big\{\mathcal{J}_1(\mathbb{B}^*,\rho_n)+\log(\tfrac{n}{\delta})\Big\} + \tfrac{\mathcal{D}B_0}{(1-\gamma)n}\cdot\|\theta_0-\mathbf{h}(\theta_0)\|,
\end{aligned}$$

with probability at least $1-\delta$, valid for any $\delta \in (0, 1/k)$, where $k = 1 + \log\frac{1}{1-\gamma}$.

Finally, noting that $\mathbb{P}\Big[\mathscr{E}_n^{(\theta)} \cap \mathscr{E}_n^{(v)} \cap \Big\{\|\theta_n - \theta^*\| \le \overline{R}_n\Big\}\Big] \ge 1-\delta$, and using the initialization conditions (4.11), we obtain the bound that was claimed in Corollary 2.

Proof of Theorem 8

The proof of this theorem is similar to that of Theorem 6, but is based on an improved version of Lemma 18, stated as Lemma 23. At a high level, there are three main steps:

1. First, we use Lemma 19 and Lemma 23 to establish a relation between $\|\mathbf{h}(\theta_t) - \theta_t\|$ and $\|v_t\|$.
2. Second, starting with the coarse bound on $\|\mathbf{h}(\theta_t) - \theta_t\|$ and $\|v_t\|$ from Lemma 20, we iteratively refine our bounds using the relation from Step 1.
3. Finally, we improve the higher-order terms in these bounds.

Step 1: Relating $\|\mathbf{h}(\theta_t) - \theta_t\|$ and $\|v_t\|$

We first state a sharpening of Lemma 18 that holds for a multi-step contractive linear operator (see Assumption (A1)′).

Lemma 23. *Under assumptions (A1)′, (A3), and (A2), there exists a universal constant $c > 0$ such that for stepsize α satisfying the bound*

$$c\sqrt{m\alpha} \cdot L\mathcal{J}_2(\mathbb{B}^*, \rho_n) \cdot \log \tfrac{n}{\delta} \leq \tfrac{1}{3}, \tag{4.89a}$$

and the burn-in period $B_0 \geq \frac{cm}{\alpha}$, given any κ-admissible sequences $r_\theta(t)$ and $r_v(t)$ with $0 < \kappa \leq 2$, on the event $\mathscr{E}_n^{(v)}(r_v) \cap \mathscr{E}_n^{(\theta)}(r_\theta)$, the following bound holds uniformly with respect to $t \in [B_0, n]$, with probability $1 - \delta$:

$$\begin{aligned}\|v_t\| \leq \tfrac{2r_v(t)}{3} + \tfrac{cmr_\theta(t)}{t\alpha} + \tfrac{c}{t}\sqrt{\tfrac{m}{\alpha}}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} \\ + \tfrac{cb_*}{t}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{n}{\delta})\Big\} + 4\Big(\tfrac{B_0}{t}\Big)^2 \|v_{B_0}\|\,.\end{aligned} \tag{4.89b}$$

See Section 4.5 for the proof of this lemma.

In addition, by Lemma 22 and the operator norm bound on $(I - A)^{-1}$, conditioned on the event $\mathscr{E}_n^{(\theta)}(r_\theta) \cap \mathscr{E}_n^{(v)}(r_v)$, we have the following bound uniformly for $t \in [B_0, n]$,

$$\begin{aligned}\|z_t\| \leq \frac{c}{\sqrt{t}}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} + \frac{cb_*}{t}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{n}{\delta})\Big\} \\ + \frac{cL}{t}\Big\{\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{n}{\delta})}\Big\}\Big\{\alpha\Big(\sum_{s=B_0}^{t-1} s^2 r_v^2(s)\Big)^{1/2} + m\Big(\sum_{s=1}^{t-1} r_\theta^2(s)\Big)^{1/2}\Big\},\end{aligned} \tag{4.90}$$

with probability at least $1 - \delta$.

Step 2: Bounds using bootstrapping

Akin to the proof of Theorem 6, we impose the restrictions that the estimate sequences (r_θ, r_v) are $\frac{1}{2}$- and 1-admissible, respectively.

Consider a new pair (r_v^+, r_θ^+) satisfying the initial bounds $r_v^+(B_0) \geq \|v_{B_0}\|$ and $r_\theta^+(B_0) \geq \|\mathbf{h}(\theta_0) - \theta_0\|$, and such that

$$\begin{aligned}r_v^+(t) \geq \tfrac{2}{3}r_v(t) + \tfrac{1}{t\sqrt{\alpha}} \cdot \frac{cm}{\sqrt{t\alpha}} \cdot \sqrt{t} r_\theta(t) \\ + \frac{c}{t}\sqrt{\tfrac{m}{\alpha}}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} + c\frac{b_*}{t}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{n}{\delta})\Big\} + 4\Big(\tfrac{B_0}{t}\Big)^2 \|v_{B_0}\|\,,\end{aligned} \tag{4.91a}$$

and

$$\begin{aligned}r_\theta^+(t) \geq \frac{c}{\sqrt{t}}\Big\{\frac{1}{\sqrt{\alpha t}} + \sqrt{\alpha}L\mathcal{J}_2(\mathbb{B}^*, \rho_n)\log\tfrac{n}{\delta}\Big\} \cdot \Big\{t\sqrt{\alpha} r_v(t)\Big\} + 2c\tfrac{Lm}{\sqrt{t}}\mathcal{J}_2(\mathbb{B}^*, \rho_n)\log\tfrac{n}{\delta} \cdot r_\theta(t), \\ + \tfrac{c}{\sqrt{t}}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} + \tfrac{cb_*}{t}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{n}{\delta})\Big\}.\end{aligned} \tag{4.91b}$$

for each integer $t \in [B_0, n]$.

By combining the bounds (4.89b) and (4.90), we are guaranteed that

$$\mathbb{P}\Big[\mathscr{E}_n^{(\theta)}(r_\theta^+) \cap \mathscr{E}_n^{(v)}(r_v^+)\Big] \geq \mathbb{P}\Big[\mathscr{E}_n^{(\theta)}(r_\theta) \cap \mathscr{E}_n^{(v)}(r_v)\Big] - \delta,$$

Our goal is to construct two series of admissible sequences $\left(r_v^{(i)}, r_v^{(i)}\right)$ with $i = 0, 1, \cdots$, such that the pair $(\|v_t\|, \|\mathbf{h}(\theta_t) - \theta_t\|)_{t \geq B_0}$ are dominated by $\left(r_v^{(i)}(t), r_v^{(i)}(t)\right)_{t \geq 0}$, with high probability. Concretely, we consider sequences of a particular form $r_v^{(i)}(t) = \frac{\psi_v^{(i)}}{t\sqrt{\alpha}}$ and $r_\theta^{(i)}(t) = \frac{\psi_\theta^{(i)}}{\sqrt{t}}$, for pairs of positive reals $\left(\psi_v^{(i)}, \psi_\theta^{(i)}\right)$ independent of t. Apparently, with such forms, the sequence $r_\theta^{(i)}$ is $\frac{1}{2}$-admissible, and the sequence $r_v^{(i)}$ is 1-admissible. However, if we directly substitute the sequences $\left(r_v^{(i)}(t), r_v^{(i)}(t)\right)$ of such forms into the iteration (4.91), the resulting sequences (r_θ^+, r_v^+) will no longer be of the desired form. So in order to unify the coefficients in equation (4.91) into the same time scale, given a stepsize $\alpha > 0$, we define the burn-in time

$$B_0 = \tfrac{cm}{\alpha} \log(\tfrac{n}{\delta}). \tag{4.92a}$$

For each $t = B_0, B_0 + 1 \ldots$, the coefficients in (4.91) then satisfy the bounds

$$\tfrac{cm}{\sqrt{\alpha}t} \leq \tfrac{1}{6}\sqrt{m}, \quad \tfrac{c}{\sqrt{\alpha}t} \leq \tfrac{1}{12\sqrt{m}}, \quad \text{and} \quad \tfrac{m}{\sqrt{t}} \log(\tfrac{n}{\delta}) \leq \sqrt{\alpha m}. \tag{4.92b}$$

Therefore, if we construct a two-dimensional vector sequence $\psi^{(i)} = \begin{bmatrix} \psi_v^{(i)} & \psi_\theta^{(i)} \end{bmatrix}^T$ satisfying the recursive relation $\psi^{(i+1)} = Q\psi^{(i)} + b$, where

$$Q := \begin{bmatrix} 2/3 & \frac{\sqrt{m}}{6} \\ \frac{1}{12\sqrt{m}} + cL\mathcal{J}_2(\mathbb{B}^*, \rho_n)\sqrt{\alpha} \cdot \log(\tfrac{n}{\delta}) & 2cL\mathcal{J}_2(\mathbb{B}^*, \rho_n)\sqrt{\alpha m}\log(\tfrac{n}{\delta}) \end{bmatrix}, \quad \text{and}$$

$$b := c \cdot \begin{bmatrix} \sqrt{m}\Big(\mathscr{W} + \nu\sqrt{\log(n/\delta)}\Big) + b_*\sqrt{\alpha}\left(\log(\frac{1}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\right) + B_0\sqrt{\alpha}\,\|v_{B_0}\| \\ \Big(\mathscr{W} + \nu\sqrt{\log(n/\delta)}\Big) + b_*\sqrt{\frac{\alpha}{m}}\left(\log(n/\delta) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\right) + \sqrt{B_0}\,\|\mathbf{h}(\theta_0) - \theta_0\| \end{bmatrix}, \tag{4.93}$$

they will satisfy the requirement (4.91), leading to the probability bound:

$$\mathbb{P}\Big[\mathscr{E}_n^{(\theta)}\Big(r_\theta^{(i+1)}\Big) \cap \mathscr{E}_n^{(v)}\Big(r_v^{(i+1)}\Big)\Big] \geq \mathbb{P}\Big[\mathscr{E}_n^{(\theta)}\Big(r_\theta^{(i)}\Big) \cap \mathscr{E}_n^{(v)}\Big(r_v^{(i)}\Big)\Big] - \delta, \tag{4.94}$$

for the sequences $r_\theta^{(i)}(t) = \psi_\theta^{(i)}/\sqrt{t}$ and $r_v^{(i)}(t) = \psi_v^{(i)}/(\sqrt{\alpha}t)$.

It remains to specify an initial condition for the recursion above. Note that Lemma 20 implies that we have

$$\|\theta_t - \theta^*\| + \|v_t\| \leq e^{1+L\alpha t}\Big(b_* + \|\theta_0 - \theta^*\|\Big)$$

almost surely. So we can take the initialization:

$$\psi_v^{(0)} := n\sqrt{\alpha}e^{1+L\alpha n}(b_* + \|\theta_0 - \theta^*\|), \quad \text{and} \quad \psi_\theta^{(0)} := \sqrt{n}e^{1+L\alpha n}(b_* + \|\theta_0 - \theta^*\|),$$

for which the bounds $\|v_t\| \leq \frac{\psi_v^{(0)}}{t\sqrt{\alpha}}$ and $\|\theta_t - \mathbf{h}(\theta_t)\| \leq \frac{\psi_\theta^{(0)}}{\sqrt{t}}$ hold almost surely.

Given such an initial condition and the recursion (4.94), we find that

$$\mathbb{P}\Big[\mathscr{E}_n^{(v)}(r_v^{(i)}) \cap \mathscr{E}_n^{(\theta)}(r_\theta^{(i)})\Big] \geq \mathbb{P}\Big[\mathscr{E}_n^{(v)}(r_v^{(0)}) \cap \mathscr{E}_n^{(\theta)}(r_\theta^{(0)})\Big] - i\delta = 1 - i\delta.$$

It remains to understand the behavior of $\psi^{(i)}$ for large values of the index i, i.e. the after i iterations of the bootstrapping argument. We do so by solving the recursion $\psi^{(i+1)} = Q\psi^{(i)} + b$. Let us define a new matrix

$$\widetilde{Q} := \begin{bmatrix} 2/3 & \frac{\sqrt{m}}{6} \\ \frac{1}{6\sqrt{m}} & 2/3 \end{bmatrix} \overset{(i)}{=} \begin{bmatrix} \sqrt{m} & \sqrt{m} \\ 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} \frac{5}{6} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \cdot \begin{bmatrix} \sqrt{m} & \sqrt{m} \\ 1 & -1 \end{bmatrix}^{-1},$$

where the equivalence (i) follows by a direct calculation. Note that the stepsize condition (4.35a) ensures that

$$cL\mathcal{J}_2(\mathbb{B}^*, \rho_n) \log \tfrac{n}{\delta} \cdot \sqrt{m\alpha} \leq \tfrac{1}{12}, \tag{4.95}$$

then the matrix $\widetilde{Q}$ is coordinate-wise larger than the matrix Q from equation (4.93), and consequently we are guaranteed that $Qu \preceq_{\text{orth}} \widetilde{Q}u$ for any 2-dimensional vector $u \succeq_{\text{orth}} 0$. Thus, for each integer $N = 1, 2, \ldots$, we have the upper bounds

$$\begin{aligned} \begin{bmatrix} \psi_v^{(N)} \\ \psi_\theta^{(N)} \end{bmatrix} = \left(\sum_{i=0}^{N-1} Q^i\right) b + Q^N \begin{bmatrix} \psi_v^{(0)} \\ \psi_\theta^{(0)} \end{bmatrix} &\preceq_{\text{orth}} \left(\sum_{i=0}^{N-1} \widetilde{Q}^i\right) b_\psi + \widetilde{Q}^N \begin{bmatrix} \psi_v^{(0)} \\ \psi_\theta^{(0)} \end{bmatrix} \\ &\preceq_{\text{orth}} (I - \widetilde{Q})^{-1} b + e^{-N/6}\sqrt{m}\Big(\psi_v^{(0)} + \psi_\theta^{(0)}\Big)\mathbf{1}_2. \end{aligned}$$

By taking $N = cLn \log n$, replacing δ with δ/N and substituting with the above bounds, we find that

$$\begin{aligned} t\sqrt{\alpha} \cdot \|v_t\| \leq \psi_v^{(N)} &\leq c\sqrt{m}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} \\ &\quad + cb_*\sqrt{\alpha}\left(\log(\tfrac{n}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\right) + cB_0\sqrt{\alpha}\,\|v_{B_0}\| + \sqrt{B_0 m}\,\|\mathbf{h}(\theta_0) - \theta_0\|, \end{aligned} \tag{4.96a}$$

along with

$$\begin{aligned} \sqrt{t}\,\|\mathbf{h}(\theta_t) - \theta_t\| \leq \psi_\theta^{(N)} &\leq c\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} + cb_*\sqrt{\alpha}\Big\{\log(\tfrac{n}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\Big\} \\ &\quad + cB_0\sqrt{\alpha/m}\,\|v_{B_0}\| + \sqrt{B_0}\,\|\mathbf{h}(\theta_0) - \theta_0\|, \end{aligned} \tag{4.96b}$$

valid uniformly over $t \in \{B_0, B_0 + 1, \cdots, n\}$ with probability at least $1 - \delta$.

The latter bound, when combined with the Lemma 21 yields an upper bound on $\|\mathbf{h}(\theta_t) - \theta_t\|$ which has the correct leading-order term, i.e., the correct dependence on the term $\mathscr{W} + \nu\sqrt{\log \frac{n}{\delta}}$. In order to refine the dependence on the terms $\|\mathbf{h}(\theta_0) - \theta_0\|$ and $\log(\frac{n}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)$, we need do another round of bootstrapping.

Step 3: Improving the higher-order terms

With a slight abuse of notation, let the 2-vector $\psi^{(N)} := (\psi_v^{(N)}, \psi_\theta^{(N)})$ be defined by the right-hand side of equation (4.96), and consider the choices $r_v(t) := \frac{\psi_v^{(N)}}{t\sqrt{\alpha}}$ and $r_\theta(t) := \frac{\psi_\theta^{(N)}}{\sqrt{t}}$. Conditioned on the event $\mathscr{E}_n^{(\theta)}(r_\theta) \cap \mathscr{E}_n^{(v)}(r_v)$, we have

$$\begin{aligned}
\|\mathbf{h}(\theta_t) - \theta_t\| &\leq \tfrac{c}{\sqrt{t}}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\Big\} + \tfrac{cb_*}{t}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta})\Big\} \\
&\quad + \Big\{\tfrac{1}{t} + c\tfrac{\alpha L \mathcal{J}_2(\mathbb{B}^*, \rho_n)}{\sqrt{t}} \cdot \log(\tfrac{n}{\delta})\Big\}\tfrac{\psi_v^{(N)}}{\sqrt{\alpha}} + 2c\tfrac{mL\mathcal{J}_2(\mathbb{B}^*, \rho_n)}{t}\log(\tfrac{n}{\delta}) \cdot \psi_\theta^{(N)} \\
&\leq \Big\{\tfrac{c}{\sqrt{t}} + \sqrt{\tfrac{m}{\alpha}}\Big[\tfrac{1}{t} + c\tfrac{\alpha L \mathcal{J}_2(\mathbb{B}^*, \rho_n)}{\sqrt{t}} \cdot \log(\tfrac{n}{\delta})\Big]\Big\}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} \\
&\quad + c' b_* \Big\{\tfrac{1}{t} + \tfrac{\alpha L \mathcal{J}_2(\mathbb{B}^*, \rho_n)}{\sqrt{t}} \cdot \log(\tfrac{n}{\delta})\Big\}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{n}{\delta})\Big\} \\
&\quad + c'\Big\{\tfrac{1}{t}\sqrt{\tfrac{m}{\alpha}} + \tfrac{\sqrt{\alpha m} L \mathcal{J}_2(\mathbb{B}^*, \rho_n)}{\sqrt{t}} \cdot \log(\tfrac{n}{\delta})\Big\}\sqrt{B_0}\,\|\mathbf{h}(\theta_0) - \theta_0\| \\
&\quad + c'\Big\{\tfrac{mL\mathcal{J}_2(\mathbb{B}^*, \rho_n)}{t}\log(\tfrac{n}{\delta})\Big\}\sqrt{B_0}\sqrt{B_0}\,\|\mathbf{h}(\theta_0) - \theta_0\| \\
&\quad + c'\Big\{\tfrac{1}{t} + \tfrac{\alpha L \mathcal{J}_2(\mathbb{B}^*, \rho_n)}{\sqrt{t}} \cdot \log(\tfrac{n}{\delta}) + \tfrac{\sqrt{\alpha/m} L \mathcal{J}_2(\mathbb{B}^*, \rho_n)}{t}\log(\tfrac{n}{\delta})\Big\}B_0\,\|v_{B_0}\|\,,
\end{aligned}$$

with probability at least $1 - \delta$.

Given a burn-in period B_0 satisfying (4.92a) and step size satisfying (4.95), using the bound on $\|v_{B_0}\|$ from Lemma 21, we have the upper bound $\|\mathbf{h}(\theta_t) - \theta_t\| \leq \widetilde{r}_\theta(t)$, with probability at least $1 - \delta$, uniformly over all integers $t \in [n]$, where

$$\begin{aligned}
\widetilde{r}_\theta(t) := \tfrac{c_1}{\sqrt{t}}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} + c_2 b_*\Big\{\tfrac{1}{t} + \tfrac{\alpha L \mathcal{J}_2(\mathbb{B}^*, \rho_n)}{\sqrt{t}} \cdot \log^3(\tfrac{n}{\delta})\Big\}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{n}{\delta})\Big\} \\
+ c_2\Big\{\tfrac{\alpha B_0 L \mathcal{J}_2(\mathbb{B}^*, \rho_n)}{\sqrt{t}} \cdot \log(\tfrac{n}{\delta}) + \tfrac{B_0}{t}\Big\}\,\|\mathbf{h}(\theta_0) - \theta_0\|\,.
\end{aligned}$$

By substituting the upper bound $\widetilde{r}_\theta$ into equation (4.91a), we obtain a recursive inequality that takes as input an admissible sequence $r_v(t)$, and generates as output a new sequence $r_v^+(t)$ such that

$$\mathbb{P}\Big[\mathscr{E}_n^{(v)}(r_v^+)\Big] \geq \mathbb{P}\Big[\mathscr{E}_n^{(v)}(r_v)\Big] - \delta.$$

Taking any integer $N_1 > 0$, by applying the recursive inequality for N_1 times with $\delta' = \delta/N_1$, we get a sharper bound for $\|v_t\|$ with probability $1 - \delta$:

$$\begin{aligned}
\|v_t\| \leq 3c\Big[\tfrac{\sqrt{m}}{t\sqrt{\alpha}}\Big(\mathscr{W} + \nu\sqrt{\log(\tfrac{nN_1}{\delta})}\Big) + \tfrac{b_*}{t}\Big(\log(\tfrac{nN_1}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\Big)\Big] \\
+ \tfrac{cm}{\alpha t}\widetilde{r}_\theta(t) + c\Big(\tfrac{B_0}{t}\Big)^2\|v_{B_0}\| + \Big(\tfrac{1+\gamma}{2}\Big)^{N_1} \cdot \tfrac{\psi_v^{(N)}}{t\sqrt{\alpha}}.
\end{aligned}$$

Taking $N_1 := 10 \log n$, for stepsize and burn-in period satisfying the conditions (4.92a) and (4.95), some algebra yields that $\|v_t\| \leq \widetilde{r_v}(t)$ with probability at least $1 - \delta$, uniformly for each integer $t \in [B_0, n]$, where

$$\widetilde{r_v}(t) := c' \Big\{ \tfrac{1}{t} \sqrt{\tfrac{m}{\alpha}} \Big[\mathscr{W} + \nu \sqrt{\log(\tfrac{n}{\delta})} \Big] + \tfrac{b_*}{t} \Big[\log(\tfrac{n}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n) \Big] \Big\} + 2c' \Big(\tfrac{B_0}{t} \Big)^2 \|\theta_0 - \mathbf{h}(\theta_0)\| \tag{4.97}$$

for a universal constant $c' > 0$.

It can be seen that the sequences $\widetilde{r_v}$ and $\widetilde{r_\theta}$ are 2-admissible. Substituting their definitions into the bound (4.90). we find that the inequality

$$\begin{aligned} \|z_t\| \leq \tfrac{c}{\sqrt{t}} \Big\{ \mathscr{W} + \nu \sqrt{\log(\tfrac{1}{\delta})} \Big\} + \tfrac{c b_*}{t} \Big\{ \mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta}) \Big\} \\ + \tfrac{cL}{t} \Big\{ \mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{1}{\delta})} \Big\} \Big\{ \alpha \Big(\sum_{s=B_0}^{t-1} s^2 r_v^2(s) \Big)^{1/2} + m \Big(\sum_{s=1}^{t-1} r_\theta^2(s) \Big)^{1/2} \Big\} \end{aligned}$$

holds with probability at least $1 - \delta$.

Under the conditions (4.95) and (4.92a), some algebra yields:

$$\begin{aligned} \|z_t\| \leq \tfrac{c}{\sqrt{t}} \Big\{ \mathscr{W} + \nu \sqrt{\log(\tfrac{1}{\delta})} \Big\} \\ + c b_* \Big\{ \tfrac{1}{t} + \tfrac{\alpha L}{\sqrt{t}} \cdot \mathcal{J}_2(\mathbb{B}^*, \rho_n) \log(\tfrac{n}{\delta}) \Big\} \Big\{ \mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta}) \Big\} + c \tfrac{B_0}{t} \|\theta_0 - \mathbf{h}(\theta_0)\| . \end{aligned}$$

Combining with equation (4.97) yields the upper bound

$$\begin{aligned} \|\mathbf{h}(\theta_t) - \theta_t\| \leq \tfrac{c}{\sqrt{t}} \Big(\mathscr{W} + \nu \sqrt{\log(\tfrac{1}{\delta})} \Big) \\ + c b_* \Big\{ \tfrac{1}{t} + \tfrac{\alpha L}{\sqrt{t}} \mathcal{J}_2(\mathbb{B}^*, \rho_n) \log(\tfrac{n}{\delta}) \Big\} \Big\{ \mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta}) \Big\} + c \tfrac{B_0}{t} \|\theta_0 - \mathbf{h}(\theta_0)\| , \end{aligned} \tag{4.98}$$

which completes the proof of equation (4.36).

Besides, by taking a union bound over time steps $t \in \{B_0, B_0 + 1, \ldots, n\}$, we have the lower bound $\mathbb{P}\Big[\mathscr{E}_n^{(\theta)}(r_\theta^*) \Big] \geq 1 - \delta$, where

$$\begin{aligned} r_\theta^*(t) := \tfrac{c}{\sqrt{t}} \Big\{ \mathscr{W} + \nu \sqrt{\log(\tfrac{n}{\delta})} \Big\} \\ + c b_* \Big\{ \tfrac{1}{t} + \tfrac{\alpha L \mathcal{J}_2(\mathbb{B}^*, \rho_n)}{\sqrt{t}} \log(\tfrac{n}{\delta}) \Big\} \Big\{ \mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{n}{\delta}) \Big\} + c \tfrac{B_0}{t} \|\theta_0 - \mathbf{h}(\theta_0)\| . \end{aligned}$$

Proof of Lemma 23

Starting with recursion satisfied by v_t, we have

$$\begin{aligned} t \cdot v_t &= (t-1)\Big\{v_{t-1} + \theta_{t-2} - \mathbf{H}_t(\theta_{t-1}) - \theta_{t-1} + \mathbf{H}_t(\theta_{t-2})\Big\} + \Big\{\mathbf{H}(\theta_{t-1}) - \theta_{t-1}\Big\} \\ &= \Big\{(1-\alpha)I + \alpha A\Big\} \cdot (t-1)v_{t-1} - (t-1)\Big\{\varepsilon_t(\theta_{t-1}) - \varepsilon_t(\theta_{t-2})\Big\} \\ &\quad + \varepsilon_t(\theta_{t-1}) + \Big\{\mathbf{h}(\theta_{t-1}) - \theta_{t-1}\Big\}. \end{aligned}$$

For any positive integer τ, we can expand the above expression for τ steps so as to obtain

$$\begin{aligned} t \cdot v_t &= \Big((1-\alpha)I + \alpha A\Big)^{\tau}(t-\tau)v_{t-\tau} \\ &\qquad - \sum_{j=1}^{\tau}(t-j)\Big((1-\alpha)I + \alpha A\Big)^{j-1}\left(\varepsilon_{t-j+1}(\theta_{t-j}) - \varepsilon_{t-j+1}(\theta_{t-j-1})\right) \\ &\quad + \sum_{j=1}^{\tau}\Big((1-\alpha)I + \alpha A\Big)^{j-1}\varepsilon_{t-j+1}(\theta_{t-j}) + \sum_{j=1}^{\tau}\Big((1-\alpha)I + \alpha A\Big)^{j-1}\Big(\mathbf{h}(\theta_{t-j}) - \theta_{t-j}\Big). \end{aligned} \tag{4.99}$$

In addition, our analysis makes use of the following auxiliary bound

$$\Big|\Big|\Big|\Big((1-\alpha)I + \alpha A\Big)^{t}\Big|\Big|\Big|_{\mathbb{V}} \leq \min\left\{1, 2\left(1 - \tfrac{\alpha}{2m}\right)^{t}\right\}, \tag{4.100}$$

valid for all $t = 1, 2, \ldots$. See the end of this subsection for the proof of this claim.

Taking this bound as given, we proceed with the proof of this lemma. First, substituting the bound (4.100) into the decomposition (4.99) yields the bound

$$t \cdot \|v_t\| \leq 2\left(1 - \tfrac{\alpha}{2m}\right)^{\tau}(t-\tau)\,\|v_{t-\tau}\| + \|\Psi_{t-\tau,\tau}\| + \|M_{t-\tau,\tau}\| + \sum_{j=1}^{\tau}\|\mathbf{h}(\theta_{t-j}) - \theta_{t-j}\|. \tag{4.101}$$

where we define the terms

$$\Psi_{t-\tau,\tau} := \sum_{j=1}^{\tau}(t-j)\Big((1-\alpha)I + \alpha A\Big)^{j-1}\left(\varepsilon_{t-j+1}(\theta_{t-j}) - \varepsilon_{t-j+1}(\theta_{t-j-1})\right), \quad \text{and} \tag{4.102a}$$

$$M_{t-\tau,\tau} := \sum_{j=1}^{\tau}\Big((1-\alpha)I + \alpha A\Big)^{j-1}\varepsilon_{t-j+1}(\theta_{t-j}). \tag{4.102b}$$

Now we bound the terms in the decomposition (4.101). On the event $\mathscr{E}_n^{(v)}(r_v)$, each term in the summation defining $\Psi_{t-\tau,\tau}$ satisfies an almost-sure upper bound:

$$\begin{aligned} \Big((1-\alpha)I + \alpha A\Big)^{j-1}\left(\varepsilon_{t-j+1}(\theta_{t-j}) - \varepsilon_{t-j+1}(\theta_{t-j-1})\right) &\leq (t-j)L\alpha\,\|v_{t-j}\| \\ &\leq (t-j)L\alpha r_v(t-j). \end{aligned}$$

Since the sequence r_v is admissible, for burn-in time $B_0 \geq 2\tau$, we have that $(t-j)r_\theta(t-j) \leq \frac{t^2}{(t-j)} r_v(t) \leq 2tr_v(t)$. Note that the terms in $\Psi_{t-\tau,\tau}$ form a martingale difference sequence, adapted to the natural filtration $(\mathcal{F}_t)_{t\geq 0}$. Invoking the martingale concentration inequality from Lemma 25 yields the bound

$$\|\Psi_{t-\tau,\tau}\| \leq c\sqrt{\tau}\Big\{\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(1/\delta)}\Big\} \cdot L\alpha t r_v(t), \tag{4.103}$$

which holds with probability at least $1-\delta$.

As for the term $M_{t-\tau,\tau}$, we use a decomposition similar to the one used in the proof of Lemma 18:

$$\begin{aligned} M_{t-\tau,\tau} &= \sum_{j=1}^{\tau} \Big((1-\alpha)I + \alpha A\Big)^{j-1} \varepsilon_{t-j+1}(\theta^*) \\ &\quad + \sum_{j=1}^{\tau} \Big((1-\alpha)I + \alpha A\Big)^{j-1} \Big(\varepsilon_{t-j+1}(\theta_{t-j}) - \varepsilon_{t-j+1}(\theta^*)\Big) \\ &=: M^*_{t-\tau,\tau} + \widetilde{M}_{t-\tau,\tau}. \end{aligned}$$

The term $M^*_{t-\tau,\tau}$ is sum of independent random variables in $\mathbb{V}$, with each term satisfying the conditions

$$\begin{aligned} &\left\|\Big((1-\alpha)I + \alpha A\Big)^{j-1} \varepsilon_{t-j+1}(\theta^*)\right\| \leq \cdot \left\|\varepsilon_{t-j+1}(\theta^*)\right\|, \quad \text{and} \\ &\Big((1-\alpha)I + \alpha A\Big)^{j-1} \varepsilon_{t-j+1}(\theta^*) \in \Omega. \end{aligned}$$

Invoking the concentration inequality from Lemma 24 yields the bound

$$\left\|M^*_{t-\tau,\tau}\right\| \leq c\sqrt{\tau}\left(\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\right) + cb_*\left(\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta})\right),$$

which holds with probability at least $1-\delta$.

For the excess noise term $\widetilde{M}_{t-\tau,\tau}$, we note that conditioned on the event $\mathscr{E}_n^{(\theta)}(r_\theta)$, we have the upper bound

$$\left\|\Big((1-\alpha)I + \alpha A\Big)^{j-1} \Big(\varepsilon_{t-j+1}(\theta_{t-j}) - \varepsilon_{t-j+1}(\theta^*)\Big)\right\| \leq 2Lmr_\theta(t-j).$$

For an admissible sequence r_θ and burn-in period $B_0 \geq 2\tau$, we have that $r_\theta(t-j) \leq \frac{t^2}{(t-j)^2} r_\theta(t) \leq 4r_\theta(t)$ for any $j \in [\tau]$. Furthermore, the terms in $\widetilde{M}_{t-\tau,\tau}$ form a martingale difference sequence adapted to the natural filtration. By Lemma 25, on the event $\mathscr{E}_n^{(\theta)}(r_\theta)$, we have the martingale concentration inequality:

$$\left\|\widetilde{M}_{t-\tau,\tau}\right\| \leq c\sqrt{\tau}\Big(\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(1/\delta)}\Big) \cdot Lmr_\theta(t).$$

Finally, for the last term in the decomposition (4.101), we note that on the event $\mathscr{E}_n^{(\theta)}(r_\theta)$, we have the bounds:

$$\|\mathbf{h}(\theta_{t-j}) - \theta_{t-j}\| \leq r_\theta(t-j) \leq \tfrac{t^2}{(t-j)^2} r_\theta(t) \leq 4r_\theta(t).$$

In order to prove the final results, as with the proof of Lemma 18, we consider the cases of $t \geq B_0 + 2m/\alpha$ and $t \leq B_0 + 2m/\alpha$ separately.

When $t \geq B_0 + 2m/\alpha$, collecting above bounds, by taking $\tau = 2m/\alpha$, we find that

$$\begin{aligned} t \cdot \|v_t\| \leq & \left\{ \tfrac{1}{3} + c\sqrt{\tau}\left[\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\right] \cdot L\alpha \right\} t r_v(t) \\ & + \left\{ c\sqrt{\tau}\left[\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\right] \cdot Lm + \tau \right\} r_\theta(t) \\ & + c\sqrt{\tau}\left\{ \mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})} \right\} + cb_* \left\{ \mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta}) \right\}. \end{aligned}$$

Given a stepsize α such that

$$c\sqrt{m\alpha} \cdot \left\{ \mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{1}{\delta})} \right\} \cdot L \leq \tfrac{1}{3}, \tag{4.104}$$

the above inequality implies that

$$t \cdot \|v_t\| \leq \tfrac{2}{3} t r_v(t) + \tfrac{cm}{\alpha} r_\theta(t) + c\sqrt{\tfrac{m}{\alpha}} \left\{ \mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})} \right\} + cb_* \left\{ \mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta}) \right\},$$

which completes the proof of the first case.

On the other hand, when $t \leq B_0 + 2m/\alpha$, we let $\tau = t - B_0$, and find that:

$$\begin{aligned} t \cdot \|v_t\| \leq & 2B_0 \cdot \|v_{B_0}\| + c\sqrt{\tau}\left[\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\right] \cdot L\alpha t r_v(t) \\ & + \left\{ c\sqrt{\tau}\left[\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\right] \cdot Lm + \tau \right\} r_\theta(t) \\ & + c\sqrt{\tau}\left\{ \mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})} \right\} + cb_* \left\{ \mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta}) \right\}. \end{aligned}$$

Note that for $t \in [B_0, B_0 + 2m/\alpha]$, we have that $2B_0 \cdot \|v_{B_0}\| \leq 4\frac{{B_0}^2}{t} \|v_{B_0}\|$. Assuming the stepsize condition (4.104), we conclude the inequality:

$$\begin{aligned} t \cdot \|v_t\| \leq & \tfrac{2}{3} t r_v(t) + \tfrac{cm}{\alpha} r_\theta(t) + c\sqrt{\tfrac{m}{\alpha}} \left\{ \mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})} \right\} + cb_* \left\{ \mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta}) \right\} \\ & + \frac{{B_0}^2}{t} \|v_{B_0}\|, \end{aligned}$$

Proof of equation (4.100): Applying the triangle inequality yields

$$\left\| \left((1-\alpha)I + \alpha A \right)^t \right\|_{\mathbb{V}} \leq \sum_{k=0}^{t} \binom{t}{k} (1-\alpha)^k \alpha^{t-k} \left\| A^{t-k} \right\|_{\mathbb{V}}. \tag{4.105}$$

Since $\left\| A^t \right\|_{\mathbb{V}} \leq \left\| A \right\|_{\mathbb{V}}^t \leq 1$ for each $t = 0, 1, 2, \ldots$, we have

$$\left\| \left((1-\alpha)I + \alpha A \right)^t \right\|_{\mathbb{V}} \leq \sum_{k=0}^{t} \binom{t}{k} (1-\alpha)^k \alpha^{t-k} \leq 1.$$

On the other hand, we note that for any time index $i \in \mathbb{N}_+$, using the m-step contraction condition (A1)$'$, we have that:

$$|||A^i|||_{\mathbb{V}} \leq |||A^m|||_{\mathbb{V}}^{\lfloor \frac{i}{m} \rfloor} \cdot |||A^{i-m\lfloor \frac{i}{m} \rfloor}|||_{\mathbb{V}} \leq 2^{-\lfloor \frac{i}{m} \rfloor} = 2^{1-i/m}.$$

Applying this inequality with $i = t - k$ and substituting into equation (4.105), we have that:

$$\begin{aligned} |||\Big((1-\alpha)I + \alpha A\Big)^t|||_{\mathbb{V}} \leq \sum_{k=0}^{t} \binom{t}{k} (1-\alpha)^k \alpha^{t-k} \cdot 2^{1-\frac{t-k}{m}} &\leq 2\left(1 - \alpha + \alpha \cdot \left(1 - \tfrac{1}{2m}\right)\right)^t \\ &= 2\left(1 - \tfrac{\alpha}{2m}\right)^t. \end{aligned}$$

Proof of Corollary 3

Recall that by Theorem 8 and a union bound, for number of restarting epochs satisfying the given condition (4.37b), the event $\mathscr{E}_n^{(\theta)}(r_\theta^*) \cap \mathscr{E}_n^{(v)}(r_v^*)$ occurs with probability $1 - \delta$, for the function pair (r_θ^*, r_v^*) given by

$$r_v^*(t) := c\left[\tfrac{1}{t}\sqrt{\tfrac{m}{\alpha}}\left(\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\right) + \tfrac{b_*}{t}\Big\{\log(\tfrac{n}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\Big\}\right] \tag{4.106a}$$

$$r_\theta^*(t) := \tfrac{c}{\sqrt{t}}\Big\{(\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} + cb_*\Big\{\tfrac{1}{t} + \tfrac{\alpha L \mathcal{J}_2(\mathbb{B}^*, \rho_n)}{\sqrt{t}}\log^3(\tfrac{n}{\delta})\Big\}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta})\Big\}. \tag{4.106b}$$

Since $\mathbf{h}$ is an affine operator, we have the decomposition

$$\big\|\theta_n - \theta^*\big\|_C \leq \big\|(I-A)^{-1}v_{n+1}\big\|_C + \big\|(I-A)^{-1}z_{n+1}\big\|_C.$$

By the operator norm bound (4.33) and the bound (4.106a) on the norm $\|v_t\|$, we have

$$\big\|(I-A)^{-1}v_{t+1}\big\|_C \leq c'\mathcal{D}m\left[\tfrac{1}{t}\sqrt{\tfrac{m}{\alpha}}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} + \tfrac{b_*}{t}\Big\{\log(\tfrac{n}{\delta}) + \mathcal{J}_1(\mathbb{B}^*, \rho_n)\Big\}\right].$$

For the term $\|(I-A)^{-1}z_{n+1}\|$, we consider the class of test functions $\mathcal{S} := \Big\{(I - A^*)^{-1}u \mid u \in C\Big\}$. Invoking Lemma 22 with $\mu = (1-\gamma)$ yields that $\big\|(I - A^{-1})z_n\big\|_C$ is at most

$$\begin{aligned} &\tfrac{c}{\sqrt{n}}\Big\{\mathbb{E}\Big[\big\|(I-A)^{-1}W\big\|_C\Big] + \Big(\sup_{u \in C}\mathbb{E}\Big[\langle u,\, (I-A)W\rangle^2\Big]\log(\tfrac{1}{\delta})\Big)^{1/2}\Big\} \\ &\qquad + c\tfrac{\mathcal{D}mb_*}{n}\Big\{\mathcal{J}_1(\mathbb{B}^*, \rho_n) + \log(\tfrac{1}{\delta})\Big\} \\ &+ \tfrac{c\mathcal{D}mL}{t}\Big\{\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \Big(\log(\tfrac{1}{\delta})\Big)^{1/2}\Big\}\Big\{\alpha\Big(\sum_{s=B_0}^{n-1} s^2 r_v^*(s)^2\Big)^{1/2} + m\Big(\sum_{s=1}^{n-1} r_\theta^*(s)^2\Big)^{1/2}\Big\}, \end{aligned}$$

with probability at least $1-\delta$. Combining above results, some algebra yields that

$$\begin{aligned}\left\|\theta_n - \theta^*\right\|_C \leq \tfrac{c}{\sqrt{n}}\Big\{\mathbb{E}\Big[\big\|(I-A)^{-1}W\big\|_C\Big] + \Big(\sup_{u\in C}\mathbb{E}\Big[\langle u,\,(I-A)W\rangle^2\Big]\log(\tfrac{1}{\delta})\Big)^{1/2}\Big\} \\ + cm\mathcal{D}\Big\{L\mathcal{J}_2(\mathbb{B}^*,\rho_n)\log(\tfrac{n}{\delta})\sqrt{\tfrac{\alpha m}{n}} + \tfrac{1}{n}\sqrt{\tfrac{m}{\alpha}}\Big\}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{n}{\delta})}\Big\} \\ + cmb_*\mathcal{D}\Big\{\tfrac{1}{n} + \tfrac{\alpha L\mathcal{J}_2(\mathbb{B}^*,\rho_n)}{\sqrt{n}}\log(\tfrac{n}{\delta})\Big\}\Big\{\mathcal{J}_1(\mathbb{B}^*,\rho_n) + \log(\tfrac{n}{\delta})\Big\},\end{aligned}$$

with probability at least $1-\delta$. This completes the proof of Corollary 3.

4.6 Discussion

In this chapter, we have analyzed `ROOT-SA`, a variance-reduced stochastic approximation procedure, built upon the ROOT-SGD algorithm [Li+20] for stochastic optimization, and designed for solving contractive fixed-point equations in Banach spaces. Under suitable regularity assumptions, we show that the estimator outputted by the algorithm achieves non-asymptotic risk upper bounds under any semi-norm, with the leading-order term being the local minimax optimal risk. Furthermore, the sample complexity needed for such an instance-dependent optimal statistical behavior scale with the intrinsic complexity of the norm (measured in Dudley integral of the dual ball under certain metric), instead of the problem dimension. We have also shown applications to many important problems in dynamical programming and reinforcement learning, including stochastic shortest path problems, minimax Markov games, and average-cost policy evaluation. A major technical contribution of the chapter lies in the novel proof techniques, which allows us to provide sharp analysis for stochastic approximation procedures *without* requiring any inner product structure in the vector space. This work opens up the path towards optimal statistical properties and efficient algorithms for fixed-point problems of a non-Euclidean geometric structure. In the following, we list a few interesting directions of future research.

- **Optimal sample complexity for SA schemes:** One open question in our analysis concerns the $n \gtrsim (1-\gamma)^{-4}$ scaling condition needed in Theorem 6 and 7. On one hand, while our results are novel even under this sample size lower bound (and optimal for large n), it is not yet clear whether this sample size lower bound is needed, or an artifact of our proof technique. In certain special cases, this lower bound is not needed; for example, if $\mathbf{h}$ is a linear operator, it is known that $\mathcal{O}\big((1-\gamma)^{-2}\big)$ samples suffices (see Theorem 8 and Corollary 3). Furthermore, in the Euclidean setting with the gradient-update operator $\mathbf{h}: \theta \mapsto \theta - \beta^{-1}\nabla f(\theta)$ for a μ-strongly-convex and β-smooth function f, the paper [Li+20] establishes instance-optimal bounds that require only $\mathcal{O}\big((1-\gamma)^{-2}\big)$ samples. An interesting open problem, therefore, is to determine the minimum sample size for which bounds of the form stated in this chapter hold in the general Banach space setting.

- **Online statistical inference procedures:** In this chapter, we focused exclusively on computing point estimates of the fixed point. However, a natural question is the construction of confidence sets for the solution θ^* to the fixed-point equation. Ideally, such confidence set should be efficiently computable, asymptotically exact, while capturing the desirable non-asymptotic properties satisfied by our estimator. For Polyak-Ruppert-averaged SGD, the paper [Che+20a] proposed an online estimator for the covariance that partly achieves these goals for stochastic optimization in the Euclidean setting. In a concurrent piece of work involving a subset of the authors [Xia+22], confidence sets and early stopping rules are developed in the special case of policy evaluation and optimal policy estimation for discounted MRPs and MDPs. It is an interesting direction for future research to construct confidence sets with improved guarantees in the general setting, based purely on the algorithmic trajectory itself.

- **General operator equations beyond the contractive setting:** In the Euclidean setting, stochastic approximation procedures for nonlinear equations share geometric structure, giving rise to key concepts such as monotonicity and smoothness. This story becomes more complex for Banach spaces, with two different routes depending on the set up. On the one hand, if the operator $\mathbf{h}$ is mapping from the space $\mathbb{V}$ to itself, then convergence is governed by contraction properties of the operator. On the other hand, if $\mathbf{h}$ maps from the Banach space $\mathbb{V}$ to its dual space $\mathbb{V}^*$, then a monotonicity condition with respect to the Bregman divergence plays a key role (see e.g. [JNT11; KLL20]). This chapter focuses on the former case, in which $\mathbf{h}$ maps the Banach space to itself, but it is an interesting direction of future research to provide instance-dependent guarantees for various stochastic approximation procedures in the latter case, and examine their optimality properties. Even more broadly, it is interesting to consider stochastic approximation procedures for solving general non-linear equations defined on pairs of Banach spaces.

4.7 Some Concentration Inequalities in Banach Spaces

Our analysis makes use of some concentration inequalities for Banach-space-valued random variables, which we state and prove here.

Statement of the results

We begin with a bound for a sequence $\{X_i\}_{i=1}^n$ of i.i.d. zero-mean random elements. Our bound involves a zero-mean Gaussian random variable W in $\mathbb{V}$ such that

$$\mathbb{E}\left[\langle W, y\rangle \cdot \langle W, z\rangle\right] = \mathbb{E}\left[\langle X_1, y\rangle \cdot \langle X_1, z\rangle\right] \qquad \text{for all } y, z \in \mathbb{V}^*.$$

Lemma 24. *Let $\{X_i\}_{i=1}^n$ be independent zero-mean random elements taking values in $\Omega \subseteq \mathbb{V}$ with $\|X_i\| \leq 1$ almost surely for each $i = 1, 2, \cdots, n$. Then there exists a universal constant $c > 0$ such that for any $\delta \in (0,1)$ and any bounded symmetric convex set $\mathcal{S} \subseteq \mathbb{B}^*$, we have*

$$\frac{1}{n}\sup_{u\in\mathcal{S}}\langle u, \sum_{i=1}^n X_i\rangle \leq \frac{c}{\sqrt{n}}\Big\{\mathbb{E}\Big[\sup_{u\in\mathcal{S}}\langle u, W\rangle\Big] + \sqrt{\sup_{u\in\mathcal{S}}\mathbb{E}[\langle u, W\rangle^2]\cdot\log(\tfrac{1}{\delta})}\Big\} + \frac{c}{n}\Big\{\log(\tfrac{1}{\delta}) + \mathcal{J}_1(\mathcal{S}, \rho_n)\Big\}, \quad (4.107)$$

with probability at least $1-\delta$.

See Section 4.7 for the proof of this claim.

We next state a bound for the martingale case:

Lemma 25. *Let $\{X_t\}_{t=1}^n$ be a martingale in $\mathbb{V}$ adapted to the filtration $\{\mathcal{F}_t\}_{t=1}^n$. Assume that there exists a deterministic sequence $\{b_t\}_{t=1}^n$ such that $b_t \geq \frac{1}{n}$ and $\|X_t\| \leq b_t$ almost surely for each $t = 1, 2, \cdots, n$. Then there exists a universal constant $c > 0$ such that for any $\delta \in (0,1)$*

$$\|\sum_{i=1}^n X_i\| \leq c\Big(\mathcal{J}_2(\mathbb{B}^*, \rho_n) + \sqrt{\log(1/\delta)}\Big)\cdot\sqrt{\sum_{i=1}^n b_i^2}, \quad (4.108)$$

with probability at least $1-\delta$.

See Section 4.7 for the proof of this claim.

Proof of Lemma 24

Our proof is based on a combination of Talagrand's concentration inequality [Tal96], the generic chaining [Tal06] and a functional Bernstein inequality [Wai19b]. The left-hand-side of the desired inequality is the supremum of an empirical process. Define the associated Rademacher complexity $\mathscr{R}_n(\mathcal{S}) := \frac{1}{n}\mathbb{E}[\sup_{y\in\mathcal{S}} R_n(y)]$, where $R_n(y) := \mathrm{Avg}\,(\zeta_i\langle y, X_i\rangle)$ with $\{\zeta_i\}_{i=1}^n$ an i.i.d. sequence of Rademacher random variables. The expectation is taken over the randomness of both the Rademacher sequence $\{\zeta_i\}_{i=1}^n$ and the random elements $(X_i)_{i=1}^n$.

Our first lemma is a type of functional Bernstein inequality; it bounds the supremum of the empirical process by the Rademacher complexity and some additional deviation terms:

Lemma 26. *Under the assumptions of Lemma 24, we have*

$$\frac{1}{n}\sup_{u\in\mathcal{S}}\langle u,\, \sum_{i=1}^{n} X_i\rangle \leq 3\cdot\mathscr{R}_n(\mathcal{S}) + 8\sqrt{\sup_{u\in\mathcal{S}}\langle u,\, W\rangle^2\cdot\frac{\log(\frac{1}{\delta})}{n}} + c\cdot\frac{\log(\frac{1}{\delta})}{n},$$

with probability at least $1-\delta$.

See Section 4.7 for the proof of this claim.

We now use this auxiliary claim to complete the proof of Lemma 24. It suffices to upper bound the Rademacher complexity $\mathscr{R}_n(\mathcal{S})$. We define the pseudometrics

$$\rho_*(x,y) := \sqrt{\mathbb{E}\left[\langle x-y,\, X_1\rangle^2\right]} \quad \text{and} \quad \rho_n(x,y) := \sup_{e\in\Omega\cap\mathbb{B}}\langle x-y,\, e\rangle, \qquad \text{for all } x,y\in\mathbb{V}^\star.$$

Recalling that $R_n(y) = \mathrm{Avg}\,(\zeta_i\langle y,\, X_i\rangle)$, applying Bernstein's inequality yields

$$\mathbb{P}\Big[|R_n(y)-R_n(z)|>t\Big] \leq 2\exp\Big\{-\min\Big(\tfrac{n\alpha^2}{2\rho_*(y,z)^2},\tfrac{n\alpha}{\rho_n(y,z)}\Big)\Big\} \qquad \text{for any } \alpha>0.$$

For $q\geq 1$, we let γ_q denote the q^{th}-order generic chaining functional of Talagrand. With this notation, we have

$$\begin{aligned}\mathscr{R}_n(\mathcal{S}) = \mathbb{E}\left[\sup_{y\in\mathcal{S}} R_n(y)\right] &\overset{(i)}{\leq} \tfrac{c}{\sqrt{n}}\cdot\gamma_2(\mathcal{S},\rho_\star) + \tfrac{1}{n}\gamma_1(\mathcal{S},\rho_n)\\ &\overset{(ii)}{\leq} \tfrac{c}{\sqrt{n}}\cdot\mathbb{E}\left[\sup_{u\in\mathcal{S}}\langle u,\, W\rangle\right] + \tfrac{1}{n}\mathcal{J}_1(\mathcal{S},\rho_n).\end{aligned}$$

Here step (i) follows from the generic chaining theorem (see Theorem 1.2.7 [Tal06]). In step (ii), we bound the first term using the generic chaining lower bound (see Theorem 2.1.1 [Tal06]) and bound the second term using the fact that γ_1 functional is upper bounded by the Dudley entropy integral of order 1. This completes the proof of Lemma 24. It remains to prove Lemma 26.

Proof of Lemma 26

The proof of this lemma is based on Talagrand's concentration inequality for the suprema of empirical process [Tal96] and a symmetrization argument. Define the random variance $\widehat{\sigma}^2 := \frac{1}{n}\sup_{y\in\mathcal{S}}\sum_{i=1}^n\langle y,\, X_i\rangle^2$. Since the random variables X_i are bounded and $\mathcal{S}\subseteq\mathbb{B}^*$, we have $|\sup_{y\in\mathcal{S}}\langle y,\, X_i\rangle|\leq 1$. Invoking Talagrand's concentration inequality [Tal96] yields the tail bound

$$\mathbb{P}\Big[\sup_{u\in\mathcal{S}}\langle u,\, \overline{X}_n\rangle \geq \mathbb{E}\Big[\sup_{u\in\mathcal{S}}\langle u,\, \overline{X}_n\rangle\Big] + \alpha\Big] \leq \exp\left\{\tfrac{-n\alpha^2}{56\mathbb{E}[\widehat{\sigma}^2]+4\alpha}\right\}, \quad \text{valid for all } \alpha>0.$$

Consequently, for any $\delta \in (0,1)$, we have

$$\sup_{u\in\mathcal{S}}\langle u,\, \overline{X}_n\rangle \le \mathbb{E}\Big[\sup_{u\in\mathcal{S}}\langle u,\, \overline{X}_n\rangle\Big] + 8\sqrt{\tfrac{\log(\frac{1}{\delta})}{n}\mathbb{E}[\widehat{\sigma}^2]} + 4\cdot\tfrac{\log(\frac{1}{\delta})}{n}$$

with probability at least $1-\delta$.

It remains to upper bound the expected supremum $\mathbb{E}\big[\sup_{u\in\mathcal{S}}\langle u,\, \overline{X}_n\rangle\big]$ and the variance term $\widehat{\sigma}^2$. By a standard symmetrization argument, we have

$$\mathbb{E}\left[\sup_{u\in\mathcal{S}}\langle u,\, \overline{X}_n\rangle\right] \le \tfrac{2}{n}\mathbb{E}\left[\sup_{u\in\mathcal{S}}\sum_{i=1}^n \zeta_i\langle u,\, X_i\rangle\right] = 2\mathscr{R}_n(\mathcal{S}),$$

Moving onto the bound on $\widehat{\sigma}^2$, we have

$$\widehat{\sigma}^2 \le \tfrac{1}{n}\sup_{y\in\mathcal{S}}\sum_{i=1}^n\Big\{\langle y,\, X_i\rangle^2 - \mathbb{E}\left[\langle y,\, X_i\rangle^2\right]\Big\} + \tfrac{1}{n}\sup_{y\in\mathcal{S}}\sum_{i=1}^n\mathbb{E}\left[\langle y,\, X_i\rangle^2\right] \;=\; Z_n + \sup_{y\in\mathcal{S}}\mathbb{E}\Big[\langle y,\, X_i\rangle^2\Big],$$

where $Z_n := \frac{1}{n}\sup_{y\in\mathcal{S}}\sum_{i=1}^n\left(\langle y,\, X_i\rangle^2 - \mathbb{E}\left[\langle y,\, X_i\rangle^2\right]\right)$. Note that each term $|\langle y,\, X_i\rangle|$ is almost surely bounded by 1, and the map $a \mapsto a^2$ is 2-Lipschitz over the interval $[-1,1]$. Consequently, letting $\{\zeta_i\}_{i=1}^n$ denote an i.i.d. sequence of Rademacher variables, we have

$$\begin{aligned}\mathbb{E}[Z_n] \overset{(i)}{\le} \tfrac{2}{n}\mathbb{E}\left[\sup_{y\in\mathcal{S}}\sum_{i=1}^n\zeta_i\langle y,\, X_i\rangle^2\right] &\overset{(ii)}{\le} \tfrac{4}{n}\cdot\mathbb{E}\left[\sup_{y\in\mathcal{S}}\sum_{i=1}^n\zeta_i\langle y,\, X_i\rangle\right] \\ &\overset{(iii)}{\le} \tfrac{n}{64\log(\frac{1}{\delta})}\mathscr{R}_n^2(\mathcal{S}) + \tfrac{128}{n}\log(\tfrac{1}{\delta}),\end{aligned}$$

where step (i) follows from a symmetrization argument; step (ii) follows from the Ledoux–Talagrand contraction; and step (iii) follows from the Cauchy–Schwarz inequality. Overall, we have

$$8\sqrt{\tfrac{\log(\frac{1}{\delta})}{n}\mathbb{E}[\widehat{\sigma}^2]} \le \mathscr{R}_n + 8\sqrt{\sup_{u\in\mathcal{S}}\mathbb{E}\left[\langle u,\, X_1\rangle^2\right]\cdot\tfrac{\log(\frac{1}{\delta})}{n}} + \tfrac{c\log(\frac{1}{\delta})}{n}.$$

Putting together the pieces yields the bound of Lemma 26.

Proof of Lemma 25

For each vector $u \in \mathbb{V}_*$, we define the random variable $M_n(u) := \frac{1}{n}\sum_{i=1}^n\langle X_i,\, u\rangle$. Clearly, the sequence $\{M_t\}_{t\ge1}$ is a scalar martingale adapted to the filtration $(\mathcal{F}_t)_{t\ge0}$. Since $b_t \ge \frac{1}{n}$, we have

$$|\langle X_t,\, u\rangle| \le b_t\cdot\sup_{x\in\mathbb{B}\cap b_t^{-1}\Omega}\langle x,\, u\rangle \le b_t\rho_n(u,0)$$

almost surely for each $t = 1, 2, \ldots$, where $\rho_n(\cdot, \cdot)$ is a pseudo-metric on the dual space $\mathbb{V}_*$ defined in (4.5). For any $u_1, u_2 \in \mathbb{V}^*$, the Azuma-Hoeffding inequality implies that

$$\mathbb{P}\Big[\, |M_n(u_1) - M_n(u_2)| \geq \alpha \Big] \leq \exp\Big\{ - \tfrac{n\alpha^2}{\rho_n(u_1,u_2)^2 \sum_{i=1}^n b_i^2} \Big\} \qquad \text{for each } \alpha > 0.$$

Applying the Dudley chaining tail bound (see e.g. [Van14], Theorem 5.29) to the sub-Gaussian process $\{M_n(u)\}_{u \in \mathbb{B}^*}$, there exist universal constants $c, c_1 > 0$ such that

$$\mathbb{P}\left[\sup_{u \in \mathbb{B}^*} M_n(u) \geq c \sqrt{\sum_{i=1}^n b_i^2} \cdot \Big(\int_0^1 \sqrt{\log N(s; \mathbb{B}^*, \rho_n)} ds + t \Big) \right] \leq c e^{-c_1 t^2} \quad \text{for each } t > 0.$$

Setting $t = \sqrt{c_1^{-1} \log(1/\delta)}$ yields the claim.

4.8 Proofs of Auxiliary Lemmas

In this section we prove various auxiliary Lemmas that we use throughout the main proof Section 4.5.

Proof of Lemma 20

From the recursive relation, we have the upper bounds $\|\theta_t - \theta^*\| \leq \|\theta_{t-1} - \theta^*\| + \alpha \|v_t\|$, as well as

$$\begin{aligned} \|v_t\| &\leq \tfrac{t-1}{t} \|v_{t-1}\| + \tfrac{1}{t} \Big\{ \|\theta_{t-1} - \theta_{t-2}\| + \|\mathbf{H}_t(\theta_{t-1}) - \mathbf{H}_t(\theta_{t-2})\| \Big\} + \tfrac{1}{t} \|\mathbf{H}_t(\theta_{t-1}) - \theta_{t-1}\| \\ &\leq \Big\{ 1 + \tfrac{\alpha(L+1)}{t} \Big\} \|v_{t-1}\| + \tfrac{2L}{t} \|\theta_{t-1} - \theta^*\| + \tfrac{1}{t} b_*. \end{aligned}$$

Putting these two inequalities together yields the vector-based recursion

$$\begin{bmatrix} \|\theta_t - \theta^*\| \\ \|v_t\| \end{bmatrix} \leq \begin{bmatrix} 1 & \alpha \\ 0 & 1 \end{bmatrix} \cdot \left(\begin{bmatrix} 1 & 0 \\ \frac{2L}{t} & 1 + \frac{\alpha(L+1)}{t} \end{bmatrix} \begin{bmatrix} \|\theta_{t-1} - \theta^*\| \\ \|v_{t-1}\| \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{t} b_* \end{bmatrix} \right),$$

where the inequality is taken elementwise. Solving this vector recursion yields

$$\|\theta_t - \theta^*\| + \|v_t\| \leq e^{1 + \alpha L t} \left(b_* + \|\theta_0 - \theta^*\| \right),$$

valid for any $t \geq B_0 \geq \frac{1}{\alpha}$.

Proof of Lemma 21

By definition, we have

$$v_{B_0} = \tfrac{1}{B_0} \sum_{t=1}^{B_0} \Big(\mathbf{H}_t(\theta_0) - \theta_0 \Big) = \Big(\mathbf{h}(\theta_0) - \theta_0 \Big) + \tfrac{1}{B_0} \sum_{t=1}^{B_0} \varepsilon_t(\theta^*) + \tfrac{1}{B_0} \sum_{t=1}^{B_0} \Big(\varepsilon_t(\theta_0) - \varepsilon_t(\theta^*) \Big).$$

Lemma 24 guarantees that

$$\|\sum_{t=1}^{B_0}\varepsilon_t(\theta^*)\| \le c\sqrt{B_0}\Big\{\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\Big\} + c\Big\{\mathcal{J}_1(\mathbb{B}^*,\rho_n) + \log(\tfrac{1}{\delta})\Big\},$$

with probability $1-\delta$. Moreover, for each integer $t \in [B_0]$, we have:

$$\|\varepsilon_t(\theta_0) - \varepsilon_t(\theta^*)\| \le L\,\|\theta_0 - \theta^*\| \le \tfrac{L}{1-\gamma}\,\|\mathbf{h}(\theta_0) - \theta_0\|\,.$$

Lemma 25 implies that

$$\|\sum_{t=1}^{B_0}\Big(\varepsilon_t(\theta_0) - \varepsilon_t(\theta^*)\Big)\| \le \tfrac{cL}{1-\gamma}\,\|\mathbf{h}(\theta_0) - \theta_0\|\,\sqrt{B_0}\,\Big\{\mathcal{J}_2(\mathbb{B}^*,\rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\Big\}$$

with probability at least $1-\delta$.

By combining these bounds, we find that

$$\begin{aligned}\|v_{B_0}\| \le \|\mathbf{h}(\theta_0) - \theta_0\|\,\Big\{1 + \tfrac{cL}{(1-\gamma)\sqrt{B_0}}\Big[\mathcal{J}_2(\mathbb{B}^*,\rho_n) + \sqrt{\log(\tfrac{1}{\delta})}\Big]\Big\} \\ + \tfrac{c}{\sqrt{B_0}}\Big(\mathscr{W} + \nu\sqrt{\log(\tfrac{1}{\delta})}\Big) + \tfrac{c}{B_0}\Big\{\mathcal{J}_1(\mathbb{B}^*,\rho_n) + \log(\tfrac{1}{\delta})\Big\}\end{aligned}$$

with probability at least $1-\delta$. Substituting the burn-in time bound (4.70a) yields the final claim.

4.9 Comments on Theorem 7

In Section 4.9, we prove that the Bellman optimality operator associated with the optimal Q-function estimation problem satisfies the local linearity condition (A4). Using a similar argument, in Section 4.9 we show that the Bellman fixed-point operator for the stochastic shortest path problem satisfies the local linearity condition.

Verifying local linearity for Bellman optimality operator

In this section, we verify that the local linearity assumption (A4) holds for the Bellman optimality operator for Q-learning [Sze98; Wai19e; WD92b]. Consider a tabular MDP $M = (r, \mathbf{P}, \gamma)$ with state space $\mathcal{S}$ and action space $\mathcal{A}$. For any state-action pair $(x,u) \in \mathcal{S}\times\mathcal{A}$, the scalar $r(x,u)$ denotes the reward when the action u is taken at state x, and the scalar $\mathbf{P}_u(x' \mid x)$ denotes the probability of transitioning to state x' when the action u is chosen at state x.

One way to estimate an optimal policy is to calculate the optimal Q-function. Associated with a (deterministic) policy π is its *Q-function*

$$\theta^\pi(x,u) := \mathbb{E}\Big[\sum_{k=0}^{\infty} r(x_k,u_k) \mid x_0 = x, u_0 = u\Big], \quad \text{where } u_k = \pi(x_k) \quad \text{for all } k = 1,2,\ldots.$$

The optimal Q-function is given by $\theta^\star(x,u) := \sup_{\pi\in\Pi} \theta^\pi(x,u)$, and an optimal policy can be obtained as $\pi_\star(x) = \arg\max_u \theta^\star(x,u)$.

The Bellman optimality operator $\mathbf{h}$ acts on the space of Q-functions; more precisely, its action on a given Q-function $\mathbb{Q}$ is given by

$$\mathbf{h}(\mathbb{Q})(x,u) = r(x,u) + \gamma \sum_{x'} \mathbf{P}_u(x' \mid x) \cdot \max_{u'} \mathbb{Q}(x',u') \quad \text{for all } (x,u) \in \mathcal{S}\times\mathcal{A}. \tag{4.109}$$

By standard results [Ber12a], the operator $\mathbf{h}$ is γ-contractive in the ℓ_∞-norm, and the optimal state-action value function $\theta^\star$ is its unique fixed point.

For a given Q-function $\mathbb{Q}$, the associated greedy policy $\pi_{\mathbb{Q}}$ is given by

$$\pi_{\mathbb{Q}}(x) = \arg\max_u \mathbb{Q}(x,u), \tag{4.110}$$

where we break any ties by taking the smallest action (in the enumeration order) that achieves the maximum. Using this greedy policy, we can define the right linear operator

$$\mathbf{P}^{\pi_{\mathbb{Q}}}\mathbb{Q}(x,u) = \sum_{x'} \mathbf{P}_u(x' \mid x)\mathbb{Q}(x', \pi_{\mathbb{Q}}(x')).$$

Let $\mathbb{B}(\theta^*, s) := \{\theta \mid \|\theta - \theta^*\|_\infty \le s\}$ denote the ℓ_∞-ball of radius s around θ^*, and define the set

$$\mathcal{A}_s = \{\gamma \cdot \mathbf{P}^{\pi_{\mathbb{Q}}} \;\mid\; \pi_{\mathbb{Q}} \text{ is a greedy policy of } \mathbb{Q} \text{ with } \mathbb{Q} \in \mathbb{B}(\theta^\star, s)\} \tag{4.111}$$

of linear operators. We use $\pi_\star$ to denote the greedy policy associated with the optimal Q-function $\theta^\star$. By definition, the Q-functions θ and θ^* satisfy the fixed-point relations

$$\mathbf{h}(\theta) = r + \gamma \mathbf{P}^{\pi_{\mathbb{Q}}}\mathbb{Q} \qquad \text{and} \qquad \theta^\star = r + \gamma \mathbf{P}^{\pi_\star}\theta^\star.$$

Rearranging the last two equations yields

$$\mathbf{h}(\mathbb{Q}) - \mathbb{Q} = r + \gamma \mathbf{P}^{\pi_{\mathbb{Q}}}\mathbb{Q} - \mathbb{Q} = (\mathcal{I} - \gamma \mathbf{P}^{\pi_\star})(\theta^\star - \mathbb{Q}) + (\gamma \mathbf{P}^{\pi_{\mathbb{Q}}} - \gamma \mathbf{P}^{\pi_\star})\mathbb{Q} \tag{4.112a}$$

$$\mathbf{h}(\mathbb{Q}) - \mathbb{Q} = r + \gamma \mathbf{P}^{\pi_{\mathbb{Q}}}\mathbb{Q} - \mathbb{Q} = (\mathcal{I} - \gamma \mathbf{P}^{\pi_{\mathbb{Q}}})(\theta^\star - \mathbb{Q}) + (\gamma \mathbf{P}^{\pi_{\mathbb{Q}}} - \gamma \mathbf{P}^{\pi_\star})\theta^\star. \tag{4.112b}$$

Next we claim that

$$(\mathcal{I} - \gamma \mathbf{P}^{\pi_\star})^{-1}(\gamma \mathbf{P}^{\pi_{\mathbb{Q}}} - \gamma \mathbf{P}^{\pi_\star})\mathbb{Q} \overset{(a)}{\succcurlyeq} 0 \qquad \text{and} \qquad (\mathcal{I} - \gamma \mathbf{P}^{\pi_{\mathbb{Q}}})^{-1}(\gamma \mathbf{P}^{\pi_{\mathbb{Q}}} - \gamma \mathbf{P}^{\pi_\star})\theta^\star \overset{(b)}{\preccurlyeq} 0. \tag{4.113}$$

Indeed, since the policy π_θ is greedy for $\mathbb{Q}$, we have the element-wise inequality $(\gamma \mathbf{P}^{\pi_{\mathbb{Q}}} - \gamma \mathbf{P}^{\pi_\star})\mathbb{Q} \succcurlyeq 0$. The matrix $(\mathcal{I} - \gamma \mathbf{P}^{\pi_\star})^{-1}$ has non-negative entries, so that element-wise inequality (a) holds. A similar argument, using the fact that $\pi_\star$ is greedy for $\theta^\star$, yields the element-wise (b).

With the last observation in hand, combining the element-wise inequalities (4.113) with the two expressions of the Bellman defect (4.112) yields

$$|\theta^\star - \mathbb{Q}| \preccurlyeq \max\{|(\mathcal{I} - \gamma \mathbf{P}^{\pi_{\mathbb{Q}}})^{-1}(\mathbf{h}(\mathbb{Q}) - \mathbb{Q})|,\ |(\mathcal{I} - \gamma \mathbf{P}^{\pi_\star})^{-1}(\mathbf{h}(\mathbb{Q}) - \mathbb{Q})|\}.$$

Finally, note that the operator $\gamma \mathbf{P}^{\pi_{\mathbb{Q}}} \in \mathcal{A}_s$ for any $\mathbb{Q} \in \mathbb{B}(\theta^\star, s)$. Putting together the pieces we conclude that for all $\theta \in \mathbb{B}(\theta^*, s)$

$$\|\theta - \theta^*\|_\infty \leq \sup_{A \in \mathcal{A}_s} \|(\mathcal{I} - A)^{-1}(\mathbf{h}(\theta) - \theta)\|$$

Thus, we deduce that the local linearity condition (A4) is satisfied for the Bellman optimality operator $\mathbf{h}$ from equation (4.109) with $\left\| \cdot \right\|_C = \| \cdot \|_\infty$.

Verifying local linearity for the SSP operator

Recall from Section 4.4 the definition of a stochastic shortest path (SSP) problem $(r, \mathbf{P})$ with optimal-Q value $\theta^\star$. For a given Q-function $\mathbb{Q}$, we define the set of associated greedy policy $\Pi_{\mathbb{Q}}$ as

$$\Pi_{\mathbb{Q}}(x) = \arg\min_u \mathbb{Q}(x, u) \qquad \text{for all} \quad x \in \mathcal{X}_{-1}, \tag{4.114}$$

Using this greedy policy, we can define the right linear operator

$$\mathbf{P}^{\pi_{\mathbb{Q}}}\mathbb{Q}(x, u) = \sum_{x'} \mathbf{P}_u(x' \mid x)\mathbb{Q}(x', \pi_{\mathbb{Q}}(x')).$$

Letting $\mathbb{B}(\theta^*, s) := \{\theta \mid \|\theta - \theta^*\| \leq s\}$ denote the ℓ_∞-ball of radius s around θ^*, we define the set

$$\mathcal{A}_s = \{\mathbf{P}^{\pi_{\mathbb{Q}}} \mid \pi_{\mathbb{Q}} \text{ is a greedy policy of } \mathbb{Q} \text{ with } \mathbb{Q} \in \mathbb{B}(\theta^\star, s)\} \tag{4.115}$$

of linear operators. We use $\pi_\star$ to denote the greedy policy associated with the optimal Q-function $\theta^\star$.

With this setup in hand, following the same argument as Section 4.9, the local linearity assumption (A4) for the Bellman operator (4.43) can be verified with the set $\mathcal{A}_s$ of local linear operators defined in equation (4.115), and with $\left\| \cdot \right\|_C = \| \cdot \|_\infty$.

Part II

Singularity, Stability and the Localization argument

Chapter 5

Singularity, Misspecification, and the Convergence Rate of EM

A line of recent work has characterized the behavior of the EM algorithm in favorable settings in which the population likelihood is locally strongly concave around its maximizing argument. Examples include suitably separated Gaussian mixture models and mixtures of linear regressions. In this chapter we consider instead over-fitted settings in which the likelihood need not be strongly concave, or, equivalently, when the Fisher information matrix might be singular. In such settings, it is known that a global maximum of the MLE based on n samples can have a non-standard $n^{-1/4}$ rate of convergence. How does the EM algorithm behave in such settings? Focusing on the simple setting of a two-component mixture fit to a multivariate Gaussian distribution, we study the behavior of the EM algorithm both when the mixture weights are different (unbalanced case), and are equal (balanced case). Our analysis reveals a sharp distinction between these cases: in the former, the EM algorithm converges geometrically to a point at Euclidean distance $O((d/n)^{1/2})$ from the true parameter, whereas in the latter case, the convergence rate is exponentially slower, and the fixed point has a much lower $O((d/n)^{1/4})$ accuracy. The slower convergence in the balanced over-fitted case arises from the singularity of the Fisher information matrix. Analysis of this singular case requires the introduction of some novel analysis techniques, in particular we make use of a careful form of localization in the associated empirical process, and develop a recursive argument to progressively sharpen the statistical rate.

5.1 Introduction

The growth in the size and scope of modern data sets has presented the field of statistics with a number of challenges, among them is how to deal with various forms of heterogeneity. Mixture models provide a principled approach to modeling heterogeneous collections of data. In practice, it is frequently the case that the number

of mixture components in the fitted model does not match the number of components in the data-generating mechanism. It is known that such mismatch can lead to substantially slower convergence rates for the maximum likelihood estimate (MLE) for the underlying parameters. In contrast, relatively less attention has been paid to the computational implications of this mismatch. In particular, the algorithm of choice for fitting finite mixture models is the EM algorithm, a general framework that encompasses divide-and-conquer computational strategies. We seek a fundamental understanding of how EM behaves for over-fitted mixture models.

While density estimation in finite mixture models is relatively well understood [Gee00; GV01], characterizing the behavior of the MLE for the parameters has remained challenging. The main difficulty for analyzing the MLE in such settings can be attributed to the associated geometry of the parameters and the inevitable label switching between the mixtures [RG97; Ste02]. Such issues do not interfere with density estimation since standard divergence measures like the Kullback-Leibler and Hellinger distances remain invariant under permutations of labels. An important contribution to the understanding of parameter estimation in finite mixture models was made by Chen [Che95]. He considered a class of over-fitted finite mixture models; here the term "over-fitted" means that the model to be fit has more mixture components than the distribution generating the data. In an interesting contrast to the usual convergence rate $n^{-1/2}$ for the MLE based on n samples, Chen showed that for estimating scalar location parameters in a certain class of over-fitted finite mixture models, the corresponding rate slows down to $n^{-1/4}$. Such a result is of practical interest, as methods that overfit the number of mixtures are often more feasible than methods that estimate the number of components and subsequently estimate the parameters conditionally [RM11]. In subsequent work, Nguyen [Ngu13] and Heinrich et al. [HK18] have characterized the (minimax) convergence rates of parameter estimation rates for mixture models in both exact-fitted or over-fitted settings using the Wasserstein distance.

While previous works address the statistical behavior of a global maximum of the log likelihood, it does not consider the associated computational issues of obtaining such a maximum. Non-convexity of the log likelihood function makes it impossible to guarantee that a global maximum of the MLE is obtained by the iterative algorithms used in practice. Perhaps the most widely used algorithm is the expectation-maximization (EM) algorithm [DLR97]. Early work on the EM algorithm [Wu83] showed that its iterates converge to a local maximum of the log-likelihood function for a broad class of incomplete data models; this general class includes the fitting of mixture models as a special case. The EM algorithm has also been studied in the specific setting of Gaussian mixture models; here we find results both for the population EM algorithm, which is the idealized version of EM based on an infinite sample size, as well as the usual sample-based EM algorithm that is used in practice. For Gaussian mixture models, the population EM algorithm is known to exhibit various convergence rates, ranging from geometric to sub-geometric depending on the overlap between the mixture components [MXJ00; XJ96]. Recently,

Balakrishnan et al. [BWY17] provided a general theoretical framework to analyze the convergence of EM updates to a neighborhood of MLE, and to provide a non-asymptotic bounds on the Euclidean error of sample-based EM iterates. These results can be applied to the special case of well-specified two-component Gaussian location mixtures, and under suitable separation conditions, their theory guarantees that (1) population EM updates enjoy geometrical convergence rate to the true parameters when initialized in a sufficiently small neighborhood around the truth, and (2) sample-based EM updates have standard minimax convergence rate of order $(d/n)^{1/2}$, based on n i.i.d. samples of a finite mixture model in $\mathbb{R}^d$.

Further work in this vein has characterized the behavior of EM in more general settings, including global convergence of population EM [XHM16], guarantees of geometric convergence under less restrictive conditions on the mixture components [DTZ17; KYB17], analysis of EM with unknown mixtures weights and covariances [CMZar], convergence analysis with additional sparsity constraints [Hao+ar; Wan+15; YC15a], and extensions to more than two Gaussian components [Hao+ar; YYS17a]. Other works include providing optimization-theoretic guarantees for EM by viewing it in a generalized surrogate function framework [KS17], and analyzing the statistical properties of confidence intervals based on an EM estimator [Che22]. An assumption common to all of these papers is that there is no misspecification in the fitting of the Gaussian mixtures; in particular, it is assumed that the data are generated from a mixture model with the same number of components as the fitted model. As noted above, using over-fitted mixture models is common in practice, so that it is desirable to understand how the EM algorithm behaves in the over-fitted settings.

The current chapter aims to shed some light on the performance of the EM algorithm for over-fitted mixtures. We do so by providing a comprehensive study of over-fitted mixture models when fit to the simplest possible (non-mixture) data-generating mechanism, that is a multivariate normal distribution $\mathcal{N}(0, \sigma^2 I_d)$ in d dimensions with known scale parameter $\sigma > 0$. This setting, despite its simplicity, suffices to uncover some rather interesting properties of EM in the over-fitted context. In particular, considering the behavior of the EM algorithm when it is used to fit two different types of models to data of this type, we obtain the following results.

- **Two-mixture unbalanced fit:** For our first model class, we study a mixture of two location-Gaussian distributions with unknown location, known variance and unequal weights for the two components. We establish that in this case the population EM updates converge at a geometric rate to the true parameter; as an immediate consequence, the sample-based EM algorithm converges in $\mathcal{O}\log(n/d)$ steps to a ball of radius $(d/n)^{1/2}$. The fast convergence rate of EM under the unbalanced setting provides an antidote to the pessimistic belief that statistical estimators generically exhibit slow convergence for over-fitted mixtures.

- **Two-mixture balanced fit:** In the balanced version of the problem in which the mixture weights are assumed to be equal for both components, we find

that the EM algorithm behaves very differently. Beginning with the population version of the EM algorithm, we show that it converges to the true parameter from an arbitrary initialization. However, the rate of convergence varies as a function of the distance of the current iterate from the true parameter value, becoming exponentially slower as the iterates approach the true parameter. This behavior is in sharp contrast to well-specified settings settings [BWY17; DTZ17; YYS17a], where the population updates are globally contractive. We also show that our rates for population EM are tight. By combining the slow convergence of population EM with a novel localization-based analysis of the underlying empirical process, we show that the sample-based EM iterates converge to a ball of radius $(d/n)^{1/4}$ around the true parameter after $\mathcal{O}((n/d)^{1/2})$ steps. The $n^{-1/4}$ component of the Euclidean error matches known guarantees for the global maximum of the MLE [Che95]. The localization argument in our analysis is of independent interest, because such techniques are not required in analyzing the EM algorithm in well-specified settings when the population updates are globally contractive. We note that localization methods are known to be essential in deriving sharp statistical rates for M-estimators (e.g., [BBM05; Gee00; Kol06]); to the best of our knowledge, their use in analyzing the EM algorithm is novel.

The remainder of the chapter is organized as follows. In Section 5.2, we begin with a few simulations of EM in different settings that motivate the problem set-up considered later. Leading with a brief description of EM for the set-ups analyzed in the chapter, we then provide a thorough analysis of the convergence rates of EM when over-fitting Gaussian data with two components. In Section 6.5, we provide a discussion of our results and a few directions for future work. Proofs of our main results are presented in Section 5.4 with a few technical lemmas deferred to the end of this chapter.

Experimental settings: Throughout the chapter, we provide simulation experiments to demonstrate theoretical results. We summarize a few common aspects of those experiments here. For population-level computations, we computed expectations using numerical integration on a sufficiently fine grid. In the experiments for sample-based EM using n samples, we declared convergence if one of the following two criteria was met: (1) the change in the iterates was small enough, or (2) the number of iterations was too large $(100,000)$. We ran 400 trials for different settings and computed the error (which we define on a case-to-case basis) upon convergence across these experiments. In majority of the runs (for each case), criteria (1) led to convergence. Let $\widehat{m}_e, \widehat{s}_e$ denote the mean error and standard deviation across these experiments. In our plots for sample-EM, we report $\widehat{m}_e + 2\widehat{s}_e$ on the y-axis. Furthermore, whenever a slope is provided, it is the slope for the best linear fit on the log-log scale for the quantity on y-axis when fitted with the quantity reported on the x-axis. For instance, in Figure 5.1(b), for the corresponding value of n on the x-axis,

the y-axis value for a green solid dot represents the value of $|\widehat{\theta}_n - \theta^*|$ averaged over 400 experiments, accounting for the deviation across these experiments. Furthermore, the green dotted line with legend $\pi = 0.3$ and the corresponding slope -0.48 denotes the best linear fit and the respective slope for $\log |\widehat{\theta}_n - \theta^*|$ (green solid dots) with $\log n$ for the experiments corresponding to the setting $\pi = 0.3$.

5.2 Behavior of EM for over-fitted Gaussian mixtures

In this section, we explore the wide range of behavior demonstrated by EM for different settings of over-specified location Gaussian mixtures.[1] We begin with several simulations in Section 5.2 that illustrate fast and slow convergence of EM for various settings, and serve as motivation for the theoretical results derived later in the chapter. We provide basic background on EM in Section 5.2, and describe the problems to be tackled. We then present our main results and discussion of their consequences, distinguishing between the cases of unbalanced and balanced fits in Section 5.2 and Section 5.2 respectively. In Section 5.2, we provide a high-level description of the novel proof techniques that are introduced to obtain sharp guarantees for the balanced mixture case.

Motivating simulations: Fast to slow convergence of EM

In order to illustrate the diversity in the behavior of EM, we begin with some simulations of its behavior for different types of over-fitted location Gaussian mixtures. In Section 5.2, we consider effect of separation between the mixtures and observe that weak or no separation leads to slow rates. This phenomenon is actually fairly generic, as shown by the additional simulations that we present in Section 5.2.

Effect of signal strength on EM

For Gaussian location mixtures, an important aspect of the problem is the signal strength, which is measured as the separation between the means of mixture components relative to the spread in the components. For the special case of the two component mixture model

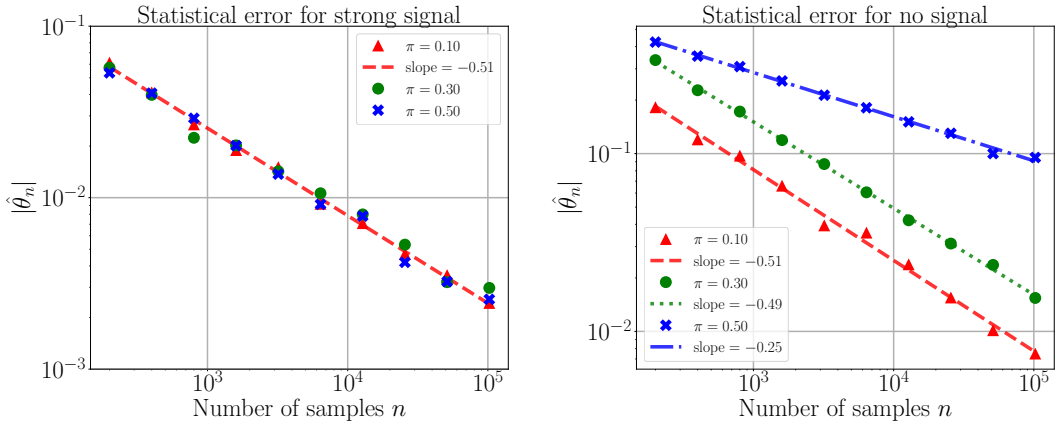
$$\pi \mathcal{N}(\theta^*, \sigma^2 I_d) + (1 - \pi)\mathcal{N}(-\theta^*, \sigma^2 I_d), \tag{5.1}$$

the signal strength is given by the ratio $\|\theta^*\|_2/\sigma$. When this ratio is large, we refer it to as the strong signal case; otherwise, it corresponds to the weak signal case.

[1] Note that in all model fits in the chapter, we assume that the scale parameter σ is known and is set equal to the true parameter. For the case when scale parameter is unknown, refer to the discussion in Section 6.5.

In Figure 5.1, we show simulations for data generated from the model (5.1) in dimension $d = 1$ and noise variance $\sigma^2 = 1$, and for three different values of the weight $\pi \in \{0.1, 0.3, 0.5\}$. In all cases, we fit a two-location Gaussian mixture with fixed weights and special structure on the location parameters—more precisely, we fit the model $\pi\mathcal{N}(\theta, 1) + (1 - \pi)\mathcal{N}(-\theta, 1)$ using EM, and take the solution as an estimate of θ^*. The two panels demonstrate the rates for EM for two distinct cases of the data-generating mechanism: (a) In the strong signal case, we set $\theta^* = 5$ so that the data has two well-separated mixture components, and (b) in the limiting case of no signal, we set $\theta^* = 0$, so that the two mixture components in the data-generating distribution collapse to one, and we are simply fitting a standard normal distribution.

In the strong signal case, it is well known [BWY17; DTZ17] that EM solutions have estimation error that achieve the classical (parametric) rate $n^{-1/2}$; the empirical results in Figure 5.1(a) are simply a confirmation of these theoretical predictions. More interesting is the case of no signal, where the simulation results shown in panel (b) of Figure 5.1 reveal a different story. Indeed, for this case the statistical rate of EM undergoes a phase transition: while it achieves the parametric rate of $n^{-1/2}$ when $\pi \neq 1/2$, it slows down to $n^{-1/4}$ when $\pi = 1/2$. We return to these cases in more detail in Section 5.2.



(a) $\theta^* = 5$ (b) $\theta^* = 0$

Figure 5.1. Plots of the error $|\widehat{\theta}_n - \theta^*|$ in the EM solution versus the sample size n, focusing on the effect of signal strength on EM solution accuracy. The true data distribution is given by $\pi\mathcal{N}(\theta^*, 1) + (1-\pi)\mathcal{N}(-\theta^*, 1)$ and we use EM to fit the model $\pi\mathcal{N}(\theta, 1) + (1-\pi)\mathcal{N}(-\theta, 1)$, generating the EM estimate $\widehat{\theta}_n$ based on n samples. (a) When the signal is strong, the estimation rate decays at the parametric rate $n^{-1/2}$, as revealed by the $-1/2$ slope in a least-square fit of the log error based on the log sample size $\log n$. (b) When there is no signal ($\theta^* = 0$), then depending on the choice of weight π in the fitted model, we observe two distinct scalings for the error: $n^{-1/2}$ when $\pi \neq 0.5$, and, $n^{-1/4}$ when $\pi = 0.5$, again as revealed by least-squares fits of the log error using the log sample size $\log n$.

Slow rate of EM with general over-fitted location mixtures

Based on simulations, we have found that the slow rate of order $n^{-1/4}$ observed in Figure 5.1(b) is not limited to fitting a mixture model of the special form $\frac{1}{2}\mathcal{N}(\theta, 1) + \frac{1}{2}\mathcal{N}(-\theta, 1)$; rather, it is a more generic phenomenon that arises in many over-fitted models. In Figure 5.2, we plot the statistical error of the EM updates when we over-fit more general location mixtures to the data generated from the standard normal distribution $\mathcal{N}(0, 1)$. For these cases, we use the second-order Wasserstein metric to quantify the convergence rates of EM updates under these settings. In our case with $\theta^* = 0$, the second-order Wasserstein metric takes the form $\widehat{W}_{2,n} = \sqrt{\sum \pi_k \widehat{\theta}_{k,n}^2}$, i.e., it corresponds to a weighted Euclidean distance of the estimated parameters from the truth. Here π_k and $\widehat{\theta}_{k,n}$, respectively, denote the mixture weight and the location parameter of the k-th component of the mixture. Let us summarize our findings for three cases[2]:

1. Unconstrained location parameters (Figure 5.2 (a)): We consider the model fit $\pi\mathcal{N}(\theta_1, 1) + (1 - \pi)\mathcal{N}(\theta_2, 1)$, where the mixture weight $\pi \in (0, 1)$ is fixed a priori. In contrast to the cases considered in Figure 5.1—in which we imposed the constraint $\theta_1 = -\theta_2 = \theta$—here we allow the location parameters θ_1 and θ_2 to be estimated independently without any coupling constraint.

2. Unknown weights (Figure 5.2(b)): In addition to estimating the location parameters independently (as in the previous case), we assume that the mixture weights are also unknown and estimate them using EM.

3. Three mixtures (also in Figure 5.2(b)): We over-fit standard Gaussian data with three mixtures: more precisely, we fit the model $\frac{1}{3}\mathcal{N}(\theta_1, 1) + \frac{1}{3}\mathcal{N}(\theta_2, 1) + \frac{1}{3}\mathcal{N}(\theta_3, 1)$, and estimate the unconstrained location parameters $\{\theta_k, k = 1, 2, 3\}$ independently using the EM algorithm.

In all these cases, we observe that EM solutions have statistical error with the slow scaling $n^{-1/4}$. Thus, these slow rates are a more generic phenomenon with over-fitted mixture models.

Problem set-up

The wide range of behavior of the EM algorithm when applied to over-fitted location Gaussian mixtures is quite intriguing, and warrants a deeper investigation,, in particular to characterize the behavior in a rigorous manner. In order to do so, we focus on the simplest instance that demonstrates the dichotomy between the fast and slow rates. As before, we assume that data is generated according to a Gaussian density in d dimensions with covariance $\sigma^2 I_d$:

$$f_*(x) = \phi(x; 0, \sigma^2 I_d), \tag{5.2}$$

[2]The exact EM updates for these settings are standard and can be found in the book [MB88].

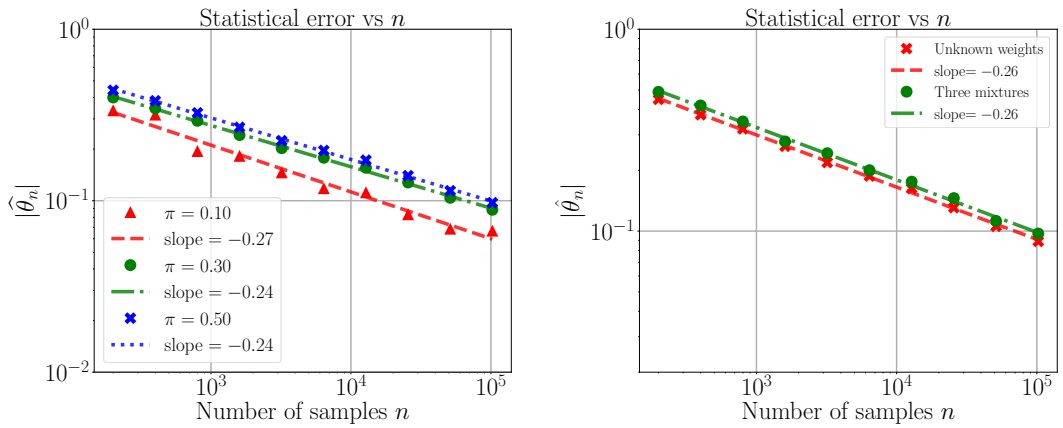


Figure 5.2. Plots of the Wasserstein error $\widehat{W}_{2,n}$ associated with EM fixed points versus the sample size for fitting various kinds of location mixture models to standard normal $\mathcal{N}(0,1)$ data. We fit mixture models with either two or three components, with all location parameters estimated in an unconstrained manner. The lines are obtained by a linear regression of the log error on the sample size n. (a) Fitting a two-mixture model with three different fixed values of weights $\pi \in \{0.1, 0.3, 0.5\}$, along with least-squares fits to the log errors. (b) Data plotted as red triangles is obtained by fitting a two-component model with unknown mixture weights, whereas green circles correspond to results fitting a three-component mixture model. In all cases, the EM solutions exhibit the slow $n^{-1/4}$ rate of convergence of the Euclidean error relative to the unknown parameter θ^*. See text for more details.

where $\phi(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-d/2} e^{-\|x-\mu\|_2^2/(2\sigma^2)}$. We study the performance of the EM algorithm when we over-fit this data-generating mechanism with a two-component isotropic location-Gaussian mixture model whose density is given by

$$f_\theta^\pi(x) = \pi\phi(x; \theta, \sigma^2 I_d) + (1-\pi)\phi(x; -\theta, \sigma^2 I_d), \tag{5.3}$$

where we assume that the mixture weight $\pi \in (0, 1/2)$ and the variance σ^2 are known and fixed apriori. Recall that we simulated this particular set-up, with our findings reported in Figure 5.1(b).

For the model fit (5.3), the expected (population) log-likelihood is given by

$$\mathcal{L}^\pi(\theta) = \mathbb{E}_X \left[\log f_\theta^\pi(X)\right] = \mathbb{E}\left[\log\left(\pi\phi\left(X; \theta, \sigma^2 I_d\right) + (1-\pi)\phi\left(X; -\theta, \sigma^2 I_d\right)\right)\right]. \tag{5.4}$$

Note that for any given π, the true model belongs to the family $\{f_\theta^\pi\}$ with $\theta^* = 0$, since $f_* = f_{\theta^*}^\pi$. Noting that

$$\arg\max_\theta \mathcal{L}^\pi(\theta) = \arg\min_\theta \mathrm{KL}(f_* \| f_\theta^\pi),$$

where $\mathrm{KL}(f_* \| f_\theta^\pi)$ denotes the Kullback-Leibler divergence between the distributions corresponding to the densities f_* and f_θ^π. As a result, finding maximizer of the

population log-likelihood would yield the true parameter θ^*. In practical situations, when one has access to only n i.i.d. samples $\{X_i\}_{i=1}^n$, the most popular choice to estimate θ^* is the maximum likelihood estimate (MLE):

$$\widehat{\theta}_n^{\mathrm{MLE}} \in \arg\max_{\theta \in \Theta} \underbrace{\frac{1}{n} \sum_{i=1}^n \log f_\theta^\pi(X_i)}_{=: \mathcal{L}_n^\pi(\theta)}. \tag{5.5}$$

In order to simplify notation, we often use the more compact notation $\mathcal{L} = \mathcal{L}^\pi$, $\mathcal{L}_n = \mathcal{L}_n^\pi$, and $f_\theta = f_\theta^\pi$ when the choice of weight π is clear from context. In general, there is no closed-form expression for the maximizer of $\mathcal{L}$ and for the estimate $\widehat{\theta}_n^{\mathrm{MLE}}$. EM attempts to circumvent this problem via a minorization-maximization scheme. While population EM is a surrogate method to compute the maximizer of the population log-likelihood $\mathcal{L}$, sample EM attempts to estimate $\widehat{\theta}_n^{\mathrm{MLE}}$. While only sample EM is a practical algorithm, several recent works have first analyzed the convergence of population EM, and then leverage the corresponding findings to understand the behavior of sample EM. In the sequel, our theoretical analysis takes a similar path and thereby to facilitate the further discussion we now describe the expressions for these updates for the model-fit (5.3).

Population versus sample EM updates

Useful for conceptual understanding, the population EM algorithm is an idealized algorithm in which all expectations are computed under the true (population) distribution, effectively taking the sample size to be infinite. We begin by describing this population algorithm, before turning to the sample-based EM algorithm that is actually used in practice.

In particular, given any point θ, the EM algorithm proceeds in two steps: (1) compute a surrogate function $Q(\cdot; \theta)$ such that $Q(\theta'; \theta) \leq \mathcal{L}(\theta')$ and $Q(\theta; \theta) = \mathcal{L}(\theta)$; and (2) compute the maximizer of $Q(\theta'; \theta)$ with respect to θ'. These steps are referred to as the E-step and the M-step, respectively. In the case of two-location mixtures, it is useful to describe a hidden variable representation of the mixture model. Consider a binary indicator variable $Z \in \{0, 1\}$ with the marginal distribution $\mathbb{P}(Z = 1) = \pi$ and $\mathbb{P}(Z = 0) = 1 - \pi$, and define the conditional distributions

$$(X \mid Z = 0) \sim \mathcal{N}(-\theta, \sigma^2 I_d), \quad \text{and} \quad (X \mid Z = 1) \sim \mathcal{N}(\theta, \sigma^2 I_d).$$

These marginal and conditional distributions define a joint distribution over the pair (X, Z), and by construction, the induced marginal distribution over X is a Gaussian mixture of the form (5.3). For EM, we first compute the conditional probability of $Z = 1$ given $X = x$:

$$w_\theta(x) = w_\theta^\pi(x) := \frac{\pi \exp\left(-\frac{\|\theta - x\|_2^2}{2\sigma^2}\right)}{\pi \exp\left(-\frac{\|\theta - x\|_2^2}{2\sigma^2}\right) + (1 - \pi) \exp\left(-\frac{\|\theta + x\|_2^2}{2\sigma^2}\right)}. \tag{5.6}$$

Then, given a vector θ, the E-step in the population EM algorithm involves computing the minorization function $\theta' \mapsto Q(\theta', \theta)$. Doing so is equivalent to *computing the expectation*

$$Q(\theta'; \theta) = -\frac{1}{2}\mathbb{E}\left[w_\theta(X)\|X - \theta'\|_2^2 + (1 - w_\theta(X))\|X + \theta'\|_2^2\right], \tag{5.7}$$

where the expectation is taken over the true density of the random variable specified in model (5.2). Next, in the M-step we maximize the function $\theta' \mapsto Q(\theta'; \theta)$, which we capture by defining the *population EM operator*, $M : \mathbb{R}^d \to \mathbb{R}^d$:

$$M(\theta) = \arg\max_{\theta' \in \mathbb{R}^d} Q(\theta', \theta) = \mathbb{E}\Big[(2w_\theta(X) - 1)X\Big], \tag{5.8}$$

where the second equality follows by computing the gradient $\nabla_{\theta'} Q$, and setting it to zero. In summary, for the two-location mixtures considered in this chapter, the population EM algorithm is defined by the sequence $\theta^{t+1} = M(\theta^t)$, where the operator M is defined in equation (5.8).

We obtain the *sample EM update* by simply replacing the expectation $\mathbb{E}$ in equations (5.7) and (5.8) by the empirical average based on an observed set of samples. In particular, given a set of i.i.d. samples $\{X_i\}_{i=1}^n$, the sample EM operator $M_n : \mathbb{R}^d \mapsto \mathbb{R}^d$ takes the form

$$M_n(\theta) := \frac{1}{n}\sum_{i=1}^n (2w_\theta(X_i) - 1)X_i. \tag{5.9}$$

Thus, the sample EM algorithm defines the sequence of iterates given by $\theta^{t+1} = M_n(\theta^t)$.

Balanced versus unbalanced mixtures: As noted in Figure 5.1(b), there is a stark difference in terms of convergence rates of sample EM between the case that $\pi \in (0, 1/2)$ and the the case when $\pi = 1/2$. In the chapter, we study theoretically the behavior of the EM algorithm in two distinct settings:

- **Unbalanced mixtures**: in this case, the mixture weights in equation (5.3) are assumed to be unequal, i.e, $\pi = \frac{1}{2}(1 - \rho)$ and $1 - \pi = \frac{1}{2}(1 + \rho)$ for some $\rho \in (0, 1)$.
- **Balanced mixtures**: in this case, the mixture weights are assumed to be equal, that is $\pi = 1 - \pi = 1/2$.

For future references, we summarize the two model-fits when the true distribution is $\mathcal{N}(0, \sigma^2)$:

$$\text{Unbalanced mixture-fit:} \quad \pi\mathcal{N}(\theta, \sigma^2) + (1 - \pi)\mathcal{N}(-\theta, \sigma^2), \quad (\pi \neq 0.5 \text{ fixed}), \tag{5.10a}$$

$$\text{Balanced mixture-fit:} \quad \frac{1}{2}\mathcal{N}(\theta, \sigma^2) + \frac{1}{2}\mathcal{N}(-\theta, \sigma^2). \tag{5.10b}$$

To be clear, in both the models above, only the parameter θ is allowed to vary; the mixture weights and scale parameter are fixed.

In the sequel, we study the convergence of EM in these two settings, both for the population EM algorithm in which the updates are given by $\theta^{t+1} = M(\theta^t)$, as well as the sample-based EM sequence given by $\theta^{t+1} = M_n(\theta^t)$. We now turn to a few interesting observations with the population EM for the two cases described above.

An interesting empirical phenomenon with population EM

In order to motivate the analysis to follow, Figure 5.3 illustrates how the population EM algorithm behaves in the unbalanced case (panel (a)) versus the balanced case (panel (b)). Each plot shows the distance of the EM iterate θ^t to the true parameter value, $\theta^* = 0$, on the vertical axis, versus the iteration number t on the horizontal axis. With the vertical axis on a log scale, a geometric convergence rate shows up as a negatively sloped line (disregarding transient effects in the first few iterations).

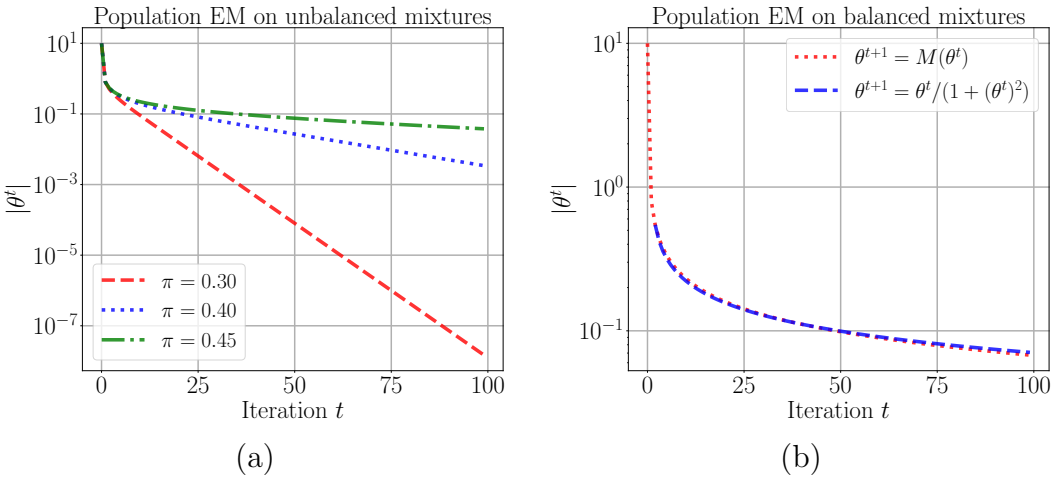


Figure 5.3. Behavior of the (numerically computed) population EM updates (5.8) when the underlying data distribution is $\mathcal{N}(0, 1)$. (a) Unbalanced mixture fits (5.10a) with weights $(\pi, 1-\pi)$: We observe geometric convergence towards $\theta^* = 0$ for all $\pi \neq 0.5$ although the rate of convergence gets slower as $\pi \to 0.5$. (b) Balanced mixture fits (5.10b) with weights $(0.5, 0.5)$: We observe two phases of convergence. First, EM quickly converges to ball of constant radius and then it exhibits slow convergence towards $\theta^* = 0$. Indeed, we see that during the slow convergence, the population EM updates track the curve given by $\theta^{t+1} = \theta^t/(1 + (\theta^t)^2)$ very closely, as predicted by our theory.

For the unbalanced mixtures in panel (a), we see that EM converges geometrically quickly, although the rate of convergence (corresponding to the slope of the line) tends towards zero as the mixture weight π tends towards $1/2$ from below. For $\pi = 1/2$, we obtain a balanced mixture, and, as shown in the plot in panel (b), the convergence

rate is now sub-geometric. In fact, the behavior of the iterates is extremely well characterized by the recursion $\theta \mapsto \frac{\theta}{1+(\theta)^2}$.

Nature of the log-likelihood and the Fisher information matrix

We now briefly discuss the fundamental difference between the unbalanced and balanced fits that result in very different behavior of EM in the two settings: the nature of the log-likelihood and the associated Fisher information matrix (FIM). For the mixture-fit (5.3) with mixture weights $(\pi, 1-\pi)$, the Fisher information matrix $\mathcal{I}^\pi(\theta) := -\nabla^2_\theta \mathcal{L}^\pi(\theta)$ can be expressed as

$$[\mathcal{I}^\pi(\theta)]_{ii} = -4\pi(1-\pi)\mathbb{E}\left[\frac{Y_i^2}{(\pi\exp(\theta^\top Y) + (1-\pi)\exp(-\theta^\top Y))^2}\right] + 1, \quad \text{for } i \in [d], \text{ and}$$

$$[\mathcal{I}^\pi(\theta)]_{ij} = -4\pi(1-\pi)\mathbb{E}\left[\frac{Y_i Y_j}{(\pi\exp(\theta^\top Y) + (1-\pi)\exp(-\theta^\top Y))^2}\right], \quad \text{for } i, j \in [d], i \neq j,$$

where the expectation is taken under the true model: $Y = (Y_1, \ldots, Y_d) \sim \mathcal{N}(0, I_d)$. Clearly, at $\theta = \theta^* = 0$, we have

$$\mathcal{I}^\pi(\theta^*) = \beta^\pi I_d, \quad \text{where} \quad \beta^\pi = -4\pi(1-\pi) + 1. \tag{5.11}$$

Note that $\beta^\pi > 0$ for any $\pi \in (0,1)$ such that $\pi \neq 1/2$. On the other hand, for $\pi = 1/2$, we have $\beta^\pi = 0$. Consequently, we find that for any unbalanced fit with $\pi \neq 1/2$, the FIM is non-singular at θ^*, and, for the balanced fit with $\pi = 1/2$, it is singular. Indeed, when we computed the population log-likelihood $\theta \mapsto \mathcal{L}^\pi(\theta)$ for different values of π, the observations were consistent with equation (5.11): the log-likelihood is strongly concave for the unbalanced fit, and weakly concave for the balanced fit. The results are plotted in Figure 5.4(a)[3], where we observe that when the mixture weights are unbalanced ($\pi < 1/2$), the population log-likelihood for the model has more curvature, and in fact is (numerically) well-approximated as $\mathcal{L}^\pi(\theta) \approx -c^\pi\theta^2$. On the other hand, for the balanced model with $\pi = \frac{1}{2}$, the likelihood is quite flat near origin and is (numerically) well-approximated as $\mathcal{L}^\pi(\theta) \approx -c\,\theta^4$. It is a folklore that the convergence rate of optimization methods has a phase transition: optimizing strongly concave functions is exponentially fast than weakly concave functions. Combining this result with the computation above serves as a good intuition as to why population EM may have fundamentally different rate of convergence in the two model fits as observed in Figure 5.3.

From population EM to sample EM: It is well established [Vaa98a] that when FIM is invertible in a neighborhood of the true parameter, MLE has the parametric

[3]Figure 5.4(b), shows the sample likelihoods $\mathcal{L}_n^\pi$ based on $n = 1000$ samples, and weights $\pi \in \{0.1, 0.5\}$. We observe that while the sample-likelihood may have more critical points, its curvature resembles very closely the curvature of the corresponding population log-likelihood.

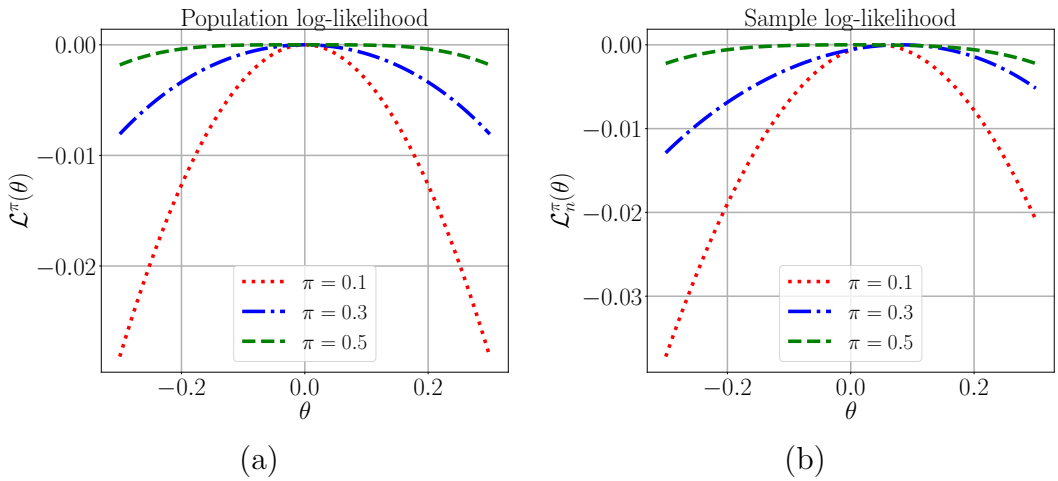


Figure 5.4. Plots of the log-likelihood for the unbalanced and balanced fit for data generated from $\mathcal{N}(0, 1)$. (a) Behavior of population log-likelihood $\mathcal{L}^{\pi}$ (5.4) (computed using numerical integration) as a function of θ for different weights $\pi \in \{0.1, 0.3, 0.5\}$. (b) Behavior of sample log-likelihood $\mathcal{L}_n^{\pi}$ (5.5) with $n = 1000$ samples for $\pi \in \{0.1, 0.3, 0.5\}$. The plots in these panels portray a stark contrast in the shapes of the log-likelihood functions in the balanced and unbalanced case, it gets flatter around $\theta^* = 0$ as $\pi \to 0.5$. More concretely, in unbalanced case we see a quadratic type behavior (strongly concave); whereas in balanced case, the log-likelihood function is flatter and depicts a fourth degree polynomial type (weakly concave) behavior.

rate of $n^{-1/2}$, and the singularity of FIM may lead to a slower than $n^{-1/2}$ rate for the MLE. As a result, for the singular case (balanced fit), we may expect a slower than parametric rate. Indeed, in the sequel, we show that the slow convergence of population EM—caused by the singularity of FIM—has a direct consequence for sample EM. Not only does the sample EM updates have a slower convergence rate in the balanced case, but this slower convergence leads to a fixed point that has lower statistical accuracy, as already illustrated in Figure 5.1(b). In particular, consider the problem of fitting a d-dimensional mixture based on $n > d$ samples: in the unbalanced case, we prove that any fixed point $\widehat{\theta}$ of the sample EM updates lies at a Euclidean distance of $\mathcal{O}(d/n)^{1/2}$ from the truth, whereas in the balanced case, we show that such fixed points lie at order $\mathcal{O}(d/n)^{1/4}$ from the truth. Thus, the EM algorithm is worse in the balanced case, both in terms of optimization speed and in terms of ultimate accuracy. Although, this slower statistical rate is in accord with existing results for the MLE in over-fitted mixture models [Che95], the theory to follow takes a different route to provide a rigorous justification, why such behaviors are to be expected with EM.

Behavior of EM for unbalanced mixtures

We now proceed to characterize the behavior of both population and sample-based EM algorithms, beginning with the setting of unbalanced mixtures (5.10a). In particular, we assume that the fitted model has known weights π and $1-\pi$, where $\pi \in (0, 1/2)$. The following result characterizes the behavior of the population EM updates for this problem:

Theorem 9. *For the unbalanced mixture model (5.10a) model fit to the true model (5.2), the population EM operator is globally strictly contractive, meaning that*

$$\|M(\theta)\|_2 \leq \left(1 - \rho^2/2\right) \|\theta\|_2 \qquad \text{for all } \theta \in \mathbb{R}^d, \tag{5.12a}$$

where $\rho = 1 - 2\pi \in (0, 1)$ denotes the measure of unbalancedness. Moreover, there are universal constants c, c' such that, given any $\delta \in (0, 1)$ and a sample size $n \geq c\, d \log(1/\delta) \frac{\sigma^2}{\rho^4}$, the sample EM sequence $\theta^{t+1} = M_n(\theta^t)$ satisfies the bound

$$\|\theta^t\|_2 \leq \|\theta^0\|_2 \left[\left(1 - \rho^2/2\right)^t + \frac{c'\sigma^2}{\rho^2} \sqrt{\frac{d \log(1/\delta)}{n}} \right], \tag{5.12b}$$

with probability at least $1 - \delta$.

See Section 5.4 for the proof of this theorem.

The bulk of the effort in proving Theorem 9 lies in establishing the guarantee (5.12a) for the population EM iterates. Once this bound has been established, we follow the analysis scheme laid out by Balakrishnan et al. [BWY17] to establish the guarantee for sample EM (5.12b). In particular, we establish a non-asymptotic uniform law of large numbers (Lemma 27 to be stated in the sequel) that allows for the translation from population to sample EM iterates. Roughly, this lemma guarantees that under suitable technical conditions, for any radius $r > 0$, tolerance $\delta \in (0, 1)$, and sufficiently large n, we have

$$\mathbb{P}\left[\sup_{\|\theta\|_2 \leq r} \|M_n(\theta) - M(\theta)\|_2 \leq c_2 \sigma^2 r \sqrt{\frac{d \log(1/\delta)}{n}} \right] \geq 1 - \delta. \tag{5.13}$$

This bound, when combined with the contractive behavior of the population EM iterates, allows us to establish the stated bound (5.12b) on the sample EM iterates.

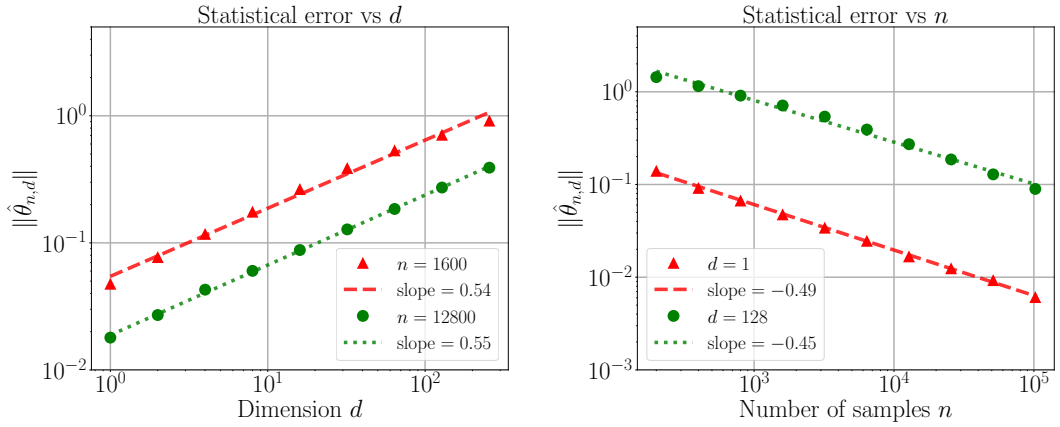
As a consequence of the bound (5.12b), we can guarantee that the sample-EM-based updates converge to an estimate with Euclidean error of the order $(d/n)^{1/2}$ after a relatively small number of steps. More precisely, if we run the algorithm for $T_\star$ iterations, where

$$T_\star := \frac{\log(\rho^2/\sigma^2) + \log(n/(d \log(1/\delta)))}{-\log(1 - \rho^2/2)},$$

then with probability at least $1 - \delta$, we are guaranteed that

$$\|\theta^{T_\star}\|_2 \leq \sqrt{\frac{d \log(1/\delta)}{n}} \cdot \|\theta^0\|_2 \cdot \left(1 + \frac{c'\sigma^2}{\rho^2}\right). \tag{5.14}$$

The above theoretical prediction (5.14) is verified by simulation study in Figure 5.1(b) for the univariate setting of unbalanced mixture-fit, i.e., $d = 1$. Let us now study the dependence on dimension d of the sample EM error for fitting unbalanced mixtures. In order to do so, we simulated 400 runs of sample EM for various settings of n and d. In Figure 5.5, we present the scaling of the radius of the final EM iterate with respect to n and d averaged over these runs. Linear fits on the log-log scale in these simulations suggest a rate close to $(d/n)^{1/2}$ as claimed in Theorem 9.



(a) (b)

Figure 5.5. Scaling of the Euclidean error $\|\widehat{\theta}_{n,d} - \theta^*\|_2$ for EM estimates $\widehat{\theta}_{n,d}$ computed using the unbalanced mixture-fit (5.10a). Here the true data distribution is $\mathcal{N}(0, I_d)$, i.e., $\theta^* = 0$, and $\widehat{\theta}_{n,d}$ denotes the EM iterate upon convergence when we fit a two-mixture model with mixture weights $(0.3, 0.7)$ using n samples in d dimensions. (a) Scaling with respect to d for $n \in \{1600, 12800\}$. (b) Scaling with respect to n for $d \in \{1, 128\}$. We ran experiments for several other pairs of (n, d) and the conclusions were the same. The empirical results here show that that our theoretical upper bound of the order $(d/n)^{0.5}$ on the EM solution is sharp in terms of n and d.

We note that this result provides some encouragement for the use of over-fitted mixtures in practice. Even though the true model is overfit with a two-component Gaussian mixture, as long as the mixture weights are unbalanced (i.e., $\pi \neq 1/2$), the EM algorithm yields an estimate with the classical minimax-optimal parametric rate $n^{-1/2}$ for estimating the true location parameter θ^*. This positive result serves as an antidote to the popular belief that the price for overestimating the number of components in the mixture is a necessarily slower rate for parameter estimation.

As revealed in Theorem 9, the extent of unbalancedness in the mixture weights plays a crucial role in the geometric rate of convergence for the population EM. Indeed, in order to obtain an ϵ-accurate estimate of $\theta^* = 0$, population EM takes $\log(\|\theta^0\|/\epsilon)/\left(-\log(1-\rho^2/2)\right)$ steps, where $\rho = |1-2\pi|$ denotes the unbalancedness of the mixtures. When the mixture weights are bounded away from $1/2$, we have that ρ is bounded away from zero, and EM converges in $\mathcal{O}\log(1/\epsilon)$ steps to an ϵ-level of accuracy. However, when the mixtures become more balanced, that is, weight π approaches $1/2$ or equivalently ρ approaches zero, the number of steps required to achieve ϵ-accuracy scales as $\mathcal{O}\log(\|\theta^0\|_2/\epsilon)/\rho^2$. Note that in the limit $\rho \to 0$, this bound degenerates to ∞ for any finite ϵ. In fact, the bound (5.12a) from Theorem 9 simply states that the population EM operator is non-expansive for balanced mixtures, and does not provide rates of convergence for this case. We now describe an exact characterization of the EM updates when fitting balanced mixtures (5.10b) for the true model (5.2).

Behavior of population EM for balanced mixtures

We begin our study of balanced mixtures by showing that the population EM operator is globally convergent, albeit with a contraction parameter that depends on θ, and degrades towards 1 as $\|\theta\|_2 \to 0$. Our statement involves the constant $p := \mathbb{P}(|X| \leq 1) + \frac{1}{2}\mathbb{P}(|X| > 1)$ where $X \sim \mathcal{N}(0,1)$ is a standard normal variate.

Theorem 10. *For the balanced mixtures setting (5.10b) of the true model (5.2), the population EM operator $\theta \mapsto M(\theta)$ has the following properties. For all non-zero θ, the operator satisfies the upper bound*

$$\frac{\|M(\theta)\|_2}{\|\theta\|_2} \leq \gamma_{up}(\theta) \; := 1 - p + \frac{p}{1 + \frac{\|\theta\|_2^2}{2\sigma^2}} \; < \; 1. \tag{5.15a}$$

Moreover, for all non-zero θ such that $\|\theta\|_2^2 \leq \frac{5\sigma^2}{8}$, it satisfies the lower bound

$$\frac{\|M(\theta)\|_2}{\|\theta\|_2} \geq \gamma_{lo}(\theta) \; := \; \frac{1}{1 + \frac{2\|\theta\|_2^2}{\sigma^2}}. \tag{5.15b}$$

See Section 5.4 for the proof of Theorem 10.

The salient feature of Theorem 10 is that the contraction coefficient $\gamma_{\mathrm{up}}(\theta)$ is not globally bounded away from 1 and in fact satisfies $\lim_{\theta \to 0} \gamma_{\mathrm{up}}(\theta) = 1$. In conjunction with the lower bound (5.15b), we see that

$$\frac{\|M(\theta)\|_2}{\|\theta\|_2} \asymp \left(1 - \frac{\|\theta\|_2^2}{\sigma^2}\right) \qquad \text{for small } \|\theta\|_2. \tag{5.16}$$

This precise contraction behavior of population EM operator is in accord with that of the simulation study in Figure 5.3(b).

Two phases of convergence: The preceding results show that the population EM updates should exhibit two phases of behavior. In the first phase, up to a relatively coarse accuracy, which is of the order σ, the iterates exhibit geometric convergence. Concretely, we are guaranteed to have $\|\theta^{T_0}\|_2 \leq \sqrt{2}\sigma$ after running the algorithm for $T_0 := \frac{\log(\|\theta^0\|_2^2/(2\sigma^2))}{\log(2/(2-p))}$ steps. In the second phase, as the error decreases from $\sqrt{2}\sigma$ to a given $\epsilon \in \left(0, \sqrt{2}\sigma\right)$, the convergence rate becomes sub-geometric: concretely, we have

$$\|\theta^{T_0+t}\|_2 \leq \epsilon \quad \text{for} \quad t \geq \frac{c\sigma^2}{\epsilon^2} \log(\sigma/\epsilon). \tag{5.17}$$

Note that the conclusion (5.16) shows that for small enough ϵ, the population EM takes $\Theta(\log(1/\epsilon)/\epsilon^2)$ steps to find ϵ-accurate estimate of θ^*. This rate is extremely slow in comparison to the geometric rate $\mathcal{O}(\log(1/\epsilon))$ established for the unbalanced mixtures in Theorem 9, and demonstrates a phase transition in the behavior of EM between the two settings.

Moreover, the sub-geometric rate derived above is also in stark contrast with the favorable behavior of EM for the exact-fitted setting established in past work. Balakrishnan et al. [BWY17] showed that when the EM algorithm is used to fit a two-component Gaussian mixture with sufficiently large separation between the means relative to the standard deviation σ (known as the signal-to-noise ratio, or SNR for short), the population EM operator is contractive in a neighborhood of the true parameter θ^*, which implies a geometric rate of convergence rate in this same neighborhood. In later work on this same model, Daskalakis et al. [DTZ17] showed that the convergence is in fact geometric for any non-zero SNR. The model considered in Theorem 10 can be seen as the degenerate limit of zero SNR, and we see that the behavior of EM breaks down, as geometric convergence no longer holds.

New techniques for analyzing sample EM

We now turn to the problem of converting the preceding guarantees on population EM to associated guarantees for the sample EM updates that are implemented in practice. A direct application of the framework developed by Balakrishnan et al. [BWY17] leads to a sub-optimal guarantee for sample EM in the case of balanced mixtures (see Section 5.2). This sub-optimality motivates the development of new methods for analyzing the behavior of the sample EM iterates, based on localization over a sequence of epochs. We sketch out these methods in Sections 5.2 and 5.2, and in Section 5.2 we draw these results together and state a theorem that provides a sharp rate of convergence for sample EM applied to balanced mixtures.

A sub-optimal guarantee

Let us recall the set-up for the procedure suggested by Balakrishnan et al. [BWY17], specializing to the case where the true parameter $\theta^* = 0$, as in our specific set-up.

Using the triangle inequality, the norm of the sample EM iterates can be upper bounded by a sum of two terms as follows:

$$\|\theta^{t+1}\|_2 = \|M_n(\theta^t)\|_2 \leq \|M_n(\theta^t) - M(\theta^t)\|_2 + \|M(\theta^t)\|_2, \quad \text{valid for all } t = 0, 1, 2, \ldots. \tag{5.18}$$

The first term on the right-hand side corresponds to the deviations between the sample and population EM operators, and can be controlled via empirical process theory. The second term corresponds to the behavior of the (deterministic) population EM operator, as applied to the sample EM iterate θ^t, and needs to be controlled via a result on population EM.

Theorem 5(a) from Balakrishnan et al. [BWY17] is based on imposing generic conditions on each of these two terms, and then using them to derive a generic bound on the sample EM iterates. In the current context, their theorem can be summarized as follows. For given tolerances $\delta \in (0, 1)$, $\epsilon > 0$ and starting radius $r > 0$, suppose that there exists a function $\varepsilon(n, \delta) > 0$, decreasing in terms of the sample size n, such that

$$\sup_{\|\theta\|_2 \geq \epsilon} \frac{\|M(\theta)\|_2}{\|\theta\|_2} \leq \kappa \quad \text{and} \quad \mathbb{P}\left[\sup_{\|\theta\|_2 \leq r} \|M_n(\theta) - M(\theta)\|_2 \leq \varepsilon(n, \delta)\right] \geq 1 - \delta. \tag{5.19a}$$

Then for a sample size n sufficiently large and ϵ sufficiently small to ensure that

$$(1 - \kappa)\epsilon \overset{(i)}{\leq} \varepsilon(n, \delta) \overset{(ii)}{\leq} (1 - \kappa)r, \tag{5.19b}$$

the sample EM iterates are guaranteed to converge to a ball of radius $\varepsilon(n, \delta)/(1 - \kappa)$ around the true parameter $\theta^* = 0$.

In order to apply this theorem to the current setting, we need to specify a choice of $\varepsilon(n, \delta)$ for which the bound on the empirical process holds. The following auxiliary result provides such control for us:

Lemma 27. *There are universal positive constants c_1, c_2 such that for any positive radius r, any threshold $\delta \in (0, 1)$, and any sample size $n \geq c_1 d \log(1/\delta)$, we have*

$$\mathbb{P}\left[\sup_{\|\theta\|_2 \leq r} \|M_n(\theta) - M(\theta)\|_2 \leq c_2 \sigma^2 r \sqrt{\frac{d \log(1/\delta)}{n}}\right] \geq 1 - \delta. \tag{5.20}$$

The proof of this lemma is based on Rademacher complexity arguments; see Section 5.5 for the details. In the proof, we establish the result with $c_1 = 1$ and $c_2 = 2777$, but make no effort to obtain "good" constants.

With the choice $r = \|\theta^0\|_2$, Lemma 27 guarantees that the second inequality in line (5.19a) holds with $\varepsilon(n, \delta) \lesssim \sigma^2 \|\theta^0\|_2 \sqrt{d/n}$. On the other hand, Theorem 10 implies that for any θ such that $\|\theta\|_2 \geq \epsilon$, we have that population EM is contractive with parameter bounded above by $\kappa(\epsilon) \asymp 1 - \epsilon^2$. In order to satisfy inequality (i)

in equation (5.19b), we solve the equation $\varepsilon(n, \delta)/(1 - \kappa(\epsilon)) = \epsilon$. Tracking only the dependency on d and n, we obtain[4]

$$\frac{\sqrt{d/n}}{\epsilon^2} = \epsilon \quad \Longrightarrow \quad \epsilon = \mathcal{O}(d/n)^{1/6}, \tag{5.21}$$

which shows that the Euclidean norm of the sample EM iterate is bounded by a term of order $(d/n)^{1/6}$.

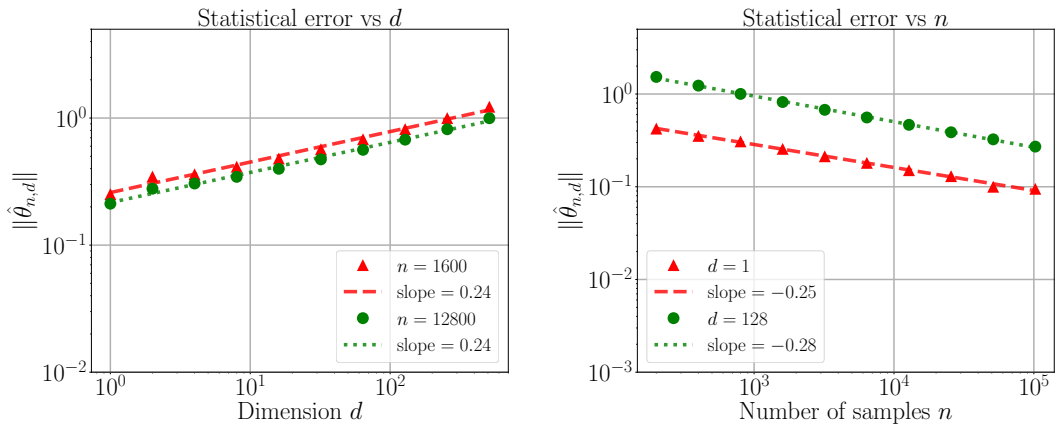
While this rate is much slower than the classical $(d/n)^{1/2}$ rate that we established in the unbalanced case, it does not coincide with the $n^{-1/4}$ rate that we obtained in Figure 5.1(b) for balanced setting with $d = 1$. Thus, the proof technique based on the framework of Balakrishnan et al. [BWY17] appears to be sub-optimal. The sub-optimality of this approach necessitates the development of a more refined technique. Before sketching this technique, we would like to quantify empirically the convergence rate of sample EM in terms of both dimension d and sample size n under balanced setting. In Figure 5.6, we summarize the results of these experiments. The two panels in the figure suggest a statistical rate of order $(d/n)^{1/4}$ for sample EM and thereby provide further numerical evidence that the preceding discussion indeed yielded a sub-optimal rate.

Localization over epochs

Let us try to understand why the preceding argument led to a sub-optimal bound. In brief, its "one-shot" nature contains two major deficiencies. First, the tolerance parameter ϵ is used both for measuring the contractivity of the updates, as in the first inequality in equation (5.19a), *and* for determining the final accuracy that we achieve. At earlier phases of the iteration, the algorithm will converge more quickly than the worst-case analysis based on the final accuracy suggests. A second deficiency is that the argument uses the radius r only once, setting it to a constant to reflect the initialization θ^0 at the start of the algorithm. This means that we failed to "localize" our bound on the empirical process in Lemma 27. At later iterations of the algorithm, the norm $\|\theta^t\|_2$ will be smaller, meaning that the empirical process can be more tightly controlled. We note that similar ideas of localizing an empirical process plays a crucial role in obtaining sharp bounds on the error of M-estimation procedures [BBM05; Gee00; Kol06].

In order to exploit this multi-phase behavior of the EM algorithm, we introduce a novel localization argument, in which the sequence of iterations is broken up into a sequence of different epochs, and we track the decay of the error carefully through each epoch. The epochs are set up in the following way:

[4]Moreover, with this choice of ϵ, inequality (ii) in equation (5.19b) is satisfied with a constant r, as long as n is sufficiently large relative to d.



(a) (b)

Figure 5.6. Scaling of the Euclidean error $\|\widehat{\theta}_{n,d} - \theta^*\|_2$ for EM estimates $\widehat{\theta}_{n,d}$ computed using the balanced mixture-fit (5.10b). Here the true data distribution is $\mathcal{N}(0, I_d)$, i.e., $\theta^* = 0$, and $\widehat{\theta}_{n,d}$ denotes the EM iterate upon convergence when we fit a balanced mixture with n samples in d dimensions. (a) Scaling with respect to d for $n \in \{1600, 12800\}$. (b) Scaling with respect to n for $d \in \{1, 128\}$. We ran experiments for several other pairs of (n, d) and the conclusions were the same. Clearly, the empirical results suggest a scaling of order $(d/n)^{1/4}$ for the final iterate of sample-based EM.

- We index epochs by the integer $\ell = 0, 1, 2, \ldots$, and associate them with a sequence $\{\alpha_\ell\}_{\ell \geq 0}$ of scalars in the interval $[0, 1/4]$. The input to epoch ℓ is the scalar α_ℓ, and the output from epoch ℓ is the scalar $\alpha_{\ell+1}$.

- For all iterations t of the sample EM algorithm contained strictly within epoch ℓ, the sample EM iterate θ^t has Euclidean norm $\|\theta^t\|_2$ and lies within the interval $\left[\left(\frac{d}{n}\right)^{\alpha_{\ell+1}}, \left(\frac{d}{n}\right)^{\alpha_\ell}\right]$.

- Upon completion of epoch ℓ at some iteration T_ℓ, the EM algorithm returns an estimate θ^{T_ℓ} such that $\|\theta^{T_\ell}\|_2 \precsim (d/n)^{\alpha_{\ell+1}}$, where

$$\alpha_{\ell+1} = \frac{1}{3}\alpha_\ell + \frac{1}{6}. \tag{5.22}$$

  Note that the new scalar $\alpha_{\ell+1}$ serves as the input to epoch $\ell + 1$.

The recursion (5.22) is crucial in our analysis: it tracks the evolution of the exponent acting upon the ratio d/n, and the rate $(d/n)^{\alpha_{\ell+1}}$ is the bound on the Euclidean norm of the sample EM iterates achieved after epoch ℓ.

A few properties of the recursion are worth noting. First, given our initialization $\alpha_0 = 0$, we see that $\alpha_1 = 1/6$, which agrees with the outcome of our one-step analysis from above. Second, as the recursion is iterated, it converges from below to the fixed

point $\alpha^* = 1/4$. Thus, our argument will allow us prove a bound arbitrarily close to $(d/n)^{1/4}$, as stated formally in Theorem 11 to follow.

How does the key recursion (5.22) arise?

Let us now sketch out how the key recursion (5.22) arises. Consider epoch ℓ specified by input $\alpha_\ell < 1/4$, and consider an iterate θ^t such that $\|\theta^t\|_2 \in (d/n)^{\alpha_\ell}$. We begin by proving that this initial condition ensures that $\|\theta^t\|_2$ is less than level $(d/n)^{\alpha_\ell}$ for all future iterations (Lemma 30 in the sequel). Given this guarantee, our second step is to apply Theorem 10 for the population EM operator, for all t such that $\|\theta^t\|_2 \geq (d/n)^{\alpha_{\ell+1}}$. Consequently, for these iterations, we have

$$\|M(\theta^t)\|_2 \leq \left(1 - p + \frac{p}{1 + \frac{\|\theta\|_2^2}{2\sigma^2}}\right)\|\theta^t\|_2 \precsim (1 - (d/n)^{2\alpha_{\ell+1}})(d/n)^{\alpha_\ell} \leq \widetilde{\gamma} \left(\frac{d}{n}\right)^{\alpha_\ell}, \tag{5.23a}$$

where $\widetilde{\gamma} := e^{-(d/n)^{2\alpha_{\ell+1}}}$. On the other hand, applying Lemma 27 for this epoch, we obtain that

$$\|M_n(\theta^t) - M(\theta^t)\|_2 \precsim \left(\frac{d}{n}\right)^{\alpha_\ell} \sqrt{\frac{d}{n}} = \left(\frac{d}{n}\right)^{\alpha_\ell + 1/2}, \tag{5.23b}$$

for all t in the epoch. Unfolding the basic triangle inequality (5.18) for T steps, we find that

$$\begin{aligned}\|\theta^{t+T}\|_2 &\leq \|M_n(\theta^t) - M(\theta^t)\|_2(1 + \widetilde{\gamma} + \ldots + \widetilde{\gamma}^{T-1}) + \widetilde{\gamma}^T\|\theta_t\|_2 \\ &\leq \frac{1}{1-\widetilde{\gamma}}\|M_n(\theta^t) - M(\theta^t)\|_2 + e^{-T(d/n)^{2\alpha_{\ell+1}}}(d/n)^{\alpha_\ell}.\end{aligned}$$

The second term decays exponentially in T, and our analysis shows that it is dominated by the first term in the relevant regime of analysis. Examining the first term, we find that θ^{t+T} has Euclidean norm of the order

$$\|\theta^{t+T}\|_2 \precsim \frac{1}{1-\widetilde{\gamma}}\|M_n(\theta^t) - M(\theta^t)\|_2 \approx \underbrace{\left(\frac{d}{n}\right)^{-2\alpha_{\ell+1}} \left(\frac{d}{n}\right)^{\alpha_\ell + 1/2}}_{=: r}. \tag{5.24}$$

The epoch is said to be complete once $\|\theta^{t+T}\|_2 \precsim \left(\frac{d}{n}\right)^{\alpha_{\ell+1}}$. Disregarding constants, this condition is satisfied when $r = \left(\frac{d}{n}\right)^{\alpha_{\ell+1}}$, or equivalently when

$$\left(\frac{d}{n}\right)^{-2\alpha_{\ell+1}} \left(\frac{d}{n}\right)^{\alpha_\ell + 1/2} = \left(\frac{d}{n}\right)^{\alpha_{\ell+1}}.$$

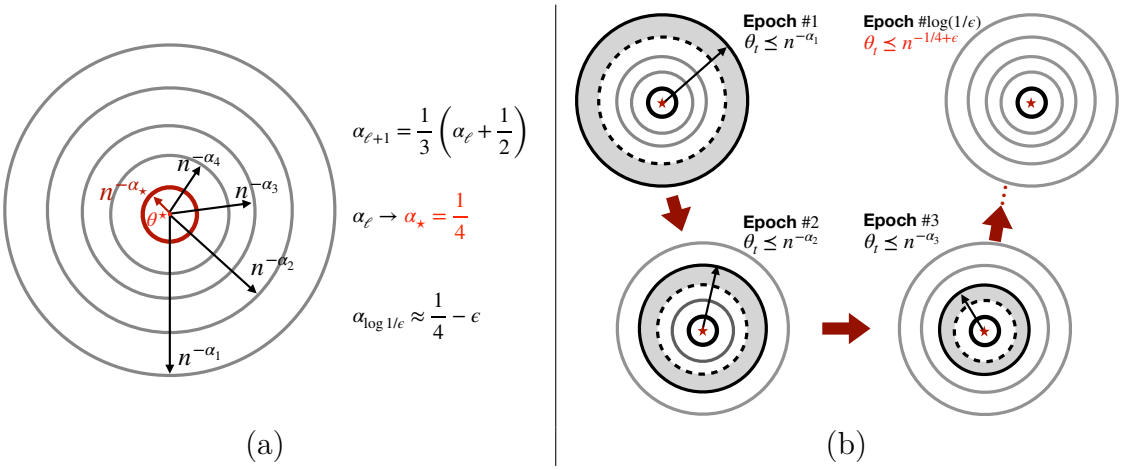


Figure 5.7. Illustration of the localization argument: Defining the epochs. (a) Radius for the ℓ-th epoch is given by $n^{-\alpha_\ell}$ (tracking dependency only on n). (b) For any given epoch ℓ, we analyze the behavior of the EM sequence $\theta^{t+1} = M_n(\theta^t)$, when θ^t lies in the disc around θ^* with inner and outer radii given by $n^{-\alpha_{\ell+1}}, n^{-\alpha_\ell}$, respectively. We prove that EM iterates move from one epoch to the next epoch (e.g. epoch ℓ to epoch $\ell+1$) after at most $\sqrt{n}$ iterations. Given the definition of α_ℓ, we see that the inner and outer radii of the aforementioned disc converges linearly to $1/4$. Consequently, after at most $\log(1/\epsilon)$ epochs (or $\sqrt{n}\log(1/\epsilon)$ iterations), the EM iterate lies in a ball of radius $n^{-1/4+\epsilon}$ around θ^*. We illustrate the one-step dynamics in any given epoch in Figure 5.8.

Viewing this equation as a function of the pair $(\alpha_{\ell+1}, \alpha_\ell)$ and solving for $\alpha_{\ell+1}$ in terms of α_ℓ yields the recursion (5.22). Refer to Figures 5.7 and 5.8 for a visual illustration of the above localization argument.

Of course, the preceding discussion is informal, and there remain many details to be addressed in order to obtain a formal proof. Nonetheless, it contains the essence of the argument used to establish a sharp rate of convergence for the sample EM iterates, which we now state.

Upper and lower bounds on sample EM

We turn to a statements of upper and lower bounds on the rate of the sample EM iterates for balanced Gaussian-location mixtures. We begin with a sharp upper bound, proved using the epoch-based localization technique that was just sketched:

Theorem 11. *There are universal constants c, c' and c'' such that for any scalars $\epsilon \in (0, 1/4)$ and $\delta \in (0,1)$, any sample size $n \geq cd\log(\log(1/\epsilon)/\delta)$ and for any iterate number larger than $t \geq c'' \log \frac{\|\theta^0\|^2 n}{\sigma^2 d} + c' \left(\frac{n}{d}\right)^{\frac{1}{2}-2\epsilon} \log(\frac{n}{d}) \log(\frac{1}{\epsilon})$, the sample-based EM*

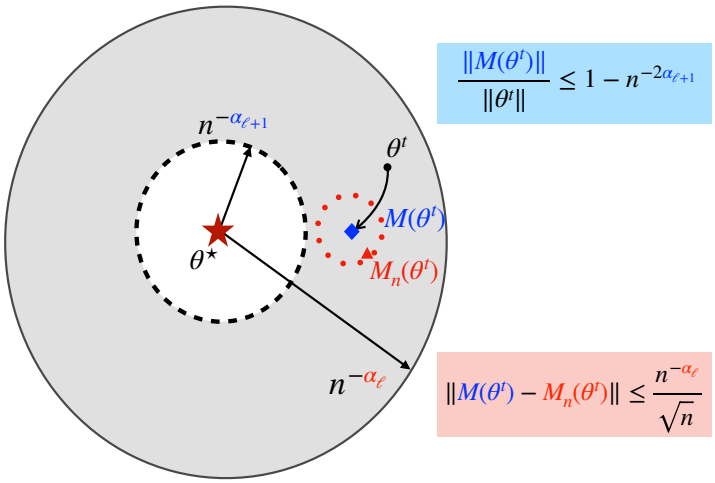


Figure 5.8. Illustration of the localization argument: Dynamics of EM in ℓ-th epoch. For a given epoch ℓ, we analyze the behavior of the EM sequence $\theta^{t+1} = M_n(\theta^t)$, when θ^t lies in the disc with inner and outer radii given by $n^{-\alpha_{\ell+1}}, n^{-\alpha_\ell}$, respectively. In this epoch, the population EM operator $M(\theta^t)$ contracts with a contraction coefficient that depends on $n^{-\alpha_{\ell+1}}$, which is the inner radius of the disc, while the perturbation error $\|M_n(\theta^t) - M(\theta)\|_2$ between the sample and population EM operators depends on $n^{-\alpha_\ell}$, which is the outer radius of the disc. Overall, we prove that M_n is non-expansive and after at most $\sqrt{n}$ steps, the sample EM updates move from epoch ℓ to epoch $\ell + 1$.

updates $\theta^t = M_n(\theta^{t-1})$ *satisfy the bound*

$$\|\theta^t\|_2 \le \left[\|\theta^0\|_2 \cdot \prod_{j=0}^{t-1} \gamma_{up}(\theta^j)\right] + \sqrt{2}\sigma \left(\frac{\sigma^2 d}{n} \log \frac{\log(4/\epsilon)}{\delta}\right)^{\frac{1}{4}-\epsilon}, \tag{5.25}$$

with probability at least $1 - \delta$.

See Section 5.4 for the detail proof of this theorem, where we also provide explicit expressions for the bounds on sample size and the number of time steps with precise values of constants c and c' that appear in the statement.

As we show in our proofs, once the iteration number t satisfies the lower bound stated in the theorem, the second term on the right-hand side of the bound (5.25) dominates the first term; therefore, from this point onwards, the the sample EM iterates have Euclidean norm of the order $(d/n)^{1/4-\epsilon}$. Note that $\epsilon \in (0, 1/4)$ can be chosen arbitrarily close to zero, so at the expense of increasing the lower bound on the number of iterations t, we can obtain rates arbitrarily close to $(d/n)^{1/4}$.

We note that earlier studies of parameter estimation for over-fitted mixtures, in both the frequentist [Che95] and Bayesian settings [IJS01; Ngu13], have derived a

rate of $n^{-1/4}$ for the global maximum of likelihood. To the best of our knowledge, Theorem 11 is the first algorithmic result that shows that such rates apply to the fixed points and dynamics of the EM algorithm, which need not converge to global optima.

The preceding discussion was devoted to an upper bound on EM. Let us now match this upper bound, at least in the univariate case $d = 1$, by showing that any non-zero fixed point of the sample EM updates has Euclidean norm of the order $n^{-1/4}$. In particular, we prove the following lower bound:

Theorem 12. *There are universal constants c, c' such that for any non-zero solution $\widehat{\theta}_n$ to the sample EM fixed-point equation $\theta = M_n(\theta)$, we have*

$$\mathbb{P}\left[|\widehat{\theta}_n| \geq c\, n^{-1/4}\right] \geq c'. \tag{5.26}$$

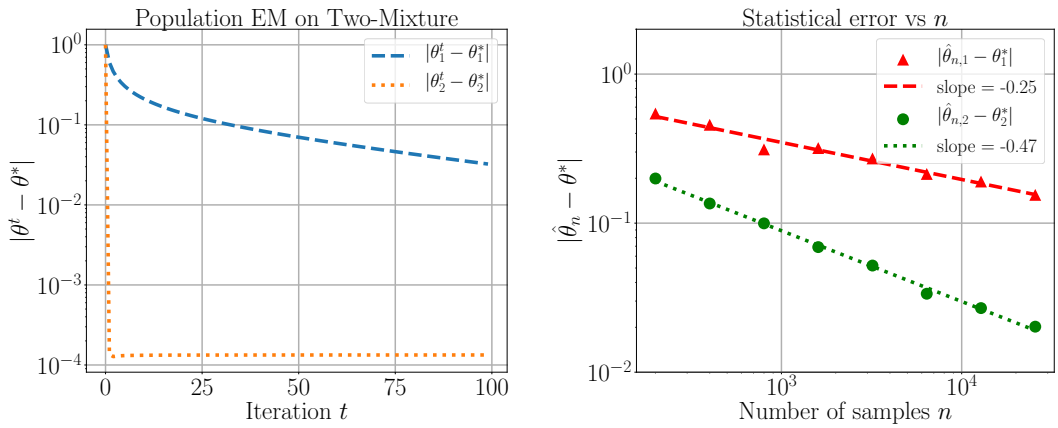
See Section 5.4 for the proof.

Since EM converges monotonically to one of its fixed points, the theorem shows that one cannot obtain a high-probability bound for any radius smaller than $n^{-1/4}$. As a consequence, the radius of convergence $n^{-1/4}$ for sample EM convergence in Theorem 11 for the univariate setting is tight with constant probability.

5.3 Discussion

In this chapter, we explored a wide range of behaviors demonstrated by the EM algorithm for different settings of over-specified location Gaussian mixtures. At a high level, our extensive simulations in Section 5.2 demonstrate that the EM algorithm can exhibit a variety of statistical error rates: starting from a parametric error rate $n^{-1/2}$ to slower statistical error rate $n^{-1/4}$. Motivated by these empirical observations, we then established rigorous statistical guarantees of EM under two particular but representative settings of over-fitted location Gaussian mixtures: the balanced and unbalanced mixture-fit. One of our key findings is that there is a phase transition with the convergence rates of sample EM between the aforementioned two settings. More concretely, in unbalanced case in dimension d, the Euclidean error in the EM solution decays at the rate $(d/n)^{1/2}$, whereas in balanced case, this same error decays at the slower rate $(d/n)^{-1/4}$. We view our results as a preliminary step in understanding and possibly improving the EM algorithm in non-regular settings. This work lays foundations for several natural questions that we now discuss.

Even though we have focused on a very simple data generating mechanism—standard multivariate Gaussian data—the EM algorithm exhibits a wide range of behavior, which provides insight that is more generally applicable. For instance, suppose that the true model is a mixture of two Gaussians with the component means/locations denoted by θ_1^* and θ_2^*, and that we fit a mixture of three Gaussian distributions. Simulations indicate that when two out of three components of the fitted Gaussian mixture are initialized close to θ_1^*, and the third component is initialized



(a) (b)

Figure 5.9. Behavior of EM for an over-fitted Gaussian mixture. True model: $\frac{1}{2}\mathcal{N}(\theta_1^*, 1) + \frac{1}{2}\mathcal{N}(\theta_2^*, 1)$ where $\theta_1^* = 0$ and $\theta_2^* = 10$. We fit a model $\frac{1}{4}\mathcal{N}(-\theta_1, 1) + \frac{1}{4}\mathcal{N}(\theta_1, 1) + \frac{1}{2}\mathcal{N}(\theta_2, 1)$, where we initialize θ_1^0 close to θ_1^* and θ_2^0 close to θ_2^*. (a) Population EM updates: We observe that while θ_1^t converges slowly to $\theta_1^* = 0$, the iterates θ_2^t converge exponentially fast to $\theta_2^* = 10$. (b) We plot the statistical error for the two parameters. While the strong signal component has a parametric $n^{-1/2}$ rate, for the no signal component EM has the slower $n^{-1/4}$ rate, which is in good agreement with the theoretical results derived in this chapter. (We remark that the error floor for θ_2^t in panel (a) is because of the error in numerical integration.)

close to θ_2^*, the estimate for θ_1^* has a slow statistical error of the order $n^{-1/4}$; the estimate for θ_2^*, on the contrary, enjoys a fast parametric rate of order $n^{-1/2}$. Refer to Figure 5.9 for more details. In more general settings, we observe similar phenomena with EM convergence rate depending on the initialization and extent of over-fitting. Such observations once again support the broader implications of our results: in mixture model fitting using the EM algorithm, the statistical error may be slow or fast, depending on whether the components are being over-fitted or not.

In our current work, we assumed that only the location parameters were unknown, meaning that the scale/variance of the underlying model are known. What if the scale parameter were also unknown? Would over-fitting with the EM algorithm exhibit an even slower statistical error in that case? We note that the MLE is known to have even slower statistical rates for such higher-order mixtures; as a result it would be interesting to determine if the EM algorithm also suffers with a similar slow down when scale is unknown. Another important direction is to analyze the behavior of EM under different data distributions. While our analysis is focused on Gaussian mixtures, the non-standard statistical rate $n^{-1/4}$ also arises in other types of over-fitted mixture models; examples include the MLE as applied to fitting Student's-t or Bernoulli distributions. We suspect that the analysis of our chapter can be generalized to a broader class of finite mixture models that includes the aforementioned models.

Finally, we discuss why the slow convergence of EM under balanced mixtures is not merely of theoretical interest. In fact, it can potentially provide a novel methodology for a classical testing problem: testing the simple null of a standard multivariate Gaussian versus the compound alternative of a two-component Gaussian mixture. This problem is known to be challenging due to the break-down of the (generalized) likelihood ratio test around the singularity of the Fisher information matrix; see the papers [CL+09; LCM09] for some past work on the problem. The results of our chapter suggest an alternative approach, which is based on monitoring the convergence rate of EM. If the EM algorithm converges slowly for a balanced fit (5.10b), then we may accept the null, whereas the opposite behavior can be used as evidence for rejecting the null. We leave the detailed development of such a methodology using the convergence rates of EM for future work.

5.4 Proofs

We now turn to the proofs of our main results, with the proofs of more technical lemmas deferred to the end of this chapter.

Proof of Theorem 9

As mentioned earlier, it remains to prove the contraction property (5.12a) for the population operator. Recall that $\theta^* = 0$ is a fixed point of the population EM update (i.e., $M(0) = 0$). This fact, combined with the definition (5.8) of the M-update, yields the relation

$$\|M(\theta)\|_2 = \|M(\theta) - M(\theta^*)\|_2 = \|\mathbb{E}\left[2(w_\theta(X) - w_0(X))X\right]\|_2.$$

Here, in the unbalanced setting, the weight function w_θ takes the form

$$w_\theta(X) = \frac{\pi}{\pi + (1-\pi)e^{-\frac{2\theta^\top X}{\sigma^2}}}, \quad \text{with gradient} \quad \nabla_\theta(w_\theta(X)) = \frac{\frac{2\pi(1-\pi)X}{\sigma^2}}{\left(\pi e^{-\frac{\theta^\top X}{\sigma^2}} + (1-\pi)e^{\frac{\theta^\top X}{\sigma^2}}\right)^2}.$$

For a scalar $u \in [0,1]$, define the function $h(u) = w_{u\theta}(X)$, and note that $h'(u) = \nabla w_{u\theta}(X)^\top \theta$. Thus, using a Taylor series expansion along the line $\theta_u = u\theta, u \in [0,1]$, we find that

$$\begin{aligned}
\|M(\theta)\|_2 &= \|\mathbb{E}\left[2X \int_0^1 h'(u)du\right]\|_2 \\
&= 4\pi(1-\pi)\| \int_0^1 \mathbb{E}\left[\frac{XX^\top}{\sigma^2\left((1-\pi)\exp\left(\frac{\theta_u^\top X}{\sigma^2}\right) + \pi\exp\left(-\frac{\theta_u^\top X}{\sigma^2}\right)\right)^2}\right]\theta du\|_2 \\
&\leq 4\pi(1-\pi)\|\theta\|_2 \max_{u\in[0,1]} \|\mathbb{E}\left[\Gamma_{\theta_u}(X)\right]\|_{\text{op}}, \qquad (5.27)
\end{aligned}$$

where we have defined the matrix

$$\Gamma_{\theta_u}(X) := \frac{XX^\top}{\sigma^2\left(\pi \exp\left(-\frac{\theta_u^\top X}{\sigma^2}\right) + (1-\pi)\exp\left(\frac{\theta_u^\top X}{\sigma^2}\right)\right)^2}.$$

Writing the mixture weight as $\pi = \frac{1}{2}(1-\rho)$, we claim that

$$\max_{u\in[0,1]} |||\mathbb{E}\left[\Gamma_{\theta_u}(X)\right]|||_{\text{op}} \leq \frac{1-\rho^2/2}{1-\rho^2}. \tag{5.28}$$

Taking this bound as given and substituting into inequality (5.27), we find that

$$\|M(\theta)\|_2 \leq 4\pi(1-\pi)\frac{1-\rho^2/2}{1-\rho^2}\|\theta\|_2 \;=\; (1-\rho^2/2)\|\theta\|_2,$$

as claimed.

Proof of claim (5.28)**:** We begin by making a convenient change of coordinates: Let R be an orthonormal matrix such that $R\theta_u = \|\theta_u\|_2 e_1$, where e_1 denotes the first canonical basis vector in dimension d. Define the random vector $V := RX/\sigma$. Since the vector $X \sim \mathcal{N}(0, \sigma^2 I_d)$ and the matrix R is orthonormal, the random vector V follows a $\mathcal{N}(0, I_d)$ distribution. Substituting $X = \sigma R^\top V$ and $R\theta_u = \|\theta_u\|_2 e_1$ in the expression for Γ_{θ_u} and using the fact that $|||R^\top B R|||_{\text{op}} = |||B|||_{\text{op}}$ for any matrix B and any orthogonal matrix R, we find that $|||\mathbb{E}\left[\Gamma_{\theta_u}(X)\right]|||_{\text{op}} = |||B_{\theta_u}|||_{\text{op}}$, where

$$B_{\theta_u} := \mathbb{E}_V\left[\frac{VV^\top}{\left(\pi\exp\left(-\|\theta_u\|_2 V_1/\sigma\right) + (1-\pi)\exp\left(\|\theta_u\|_2 V_1/\sigma\right)\right)^2}\right].$$

Note that the matrix B_{θ_u} is a diagonal matrix, with non-negative entries. Thus, in order to prove the bound (5.28), it suffices to show that

$$\max_{j\in[d]} [B_{\theta_u}]_{jj} \leq \frac{1-\rho^2/2}{1-\rho^2}. \tag{5.29}$$

When $\theta_u = 0$, the matrix $B_{\theta_u} = \mathbb{E}[VV^\top] = I_d$ and the claim holds trivially. Turning to the case $\theta_u \neq 0$, we split our analysis into two cases, depending on whether $j = 1$ or $j \neq 1$.

Bounding $[B_{\theta_u}]_{11}$**:** Denoting $\pi = \frac{1}{2}(1-\rho)$, we observe that

$$\begin{aligned}
&(\pi e^{-y} + (1-\pi)e^{y}) \in [\sqrt{(1-\rho^2)},\, 1], \quad \text{if } e^y \in \left[1, \frac{1+\rho}{1-\rho}\right], \quad \text{and} \\
&(\pi e^{-y} + (1-\pi)e^{y}) > 1, \qquad\qquad\qquad\quad \text{otherwise.}
\end{aligned}$$

Let $\mathcal{E}^c$ and $\mathbb{I}(\mathcal{E})$ respectively denote the complement and the indicator of any event $\mathcal{E}$. Let $\mathcal{E}_{\theta_u}$ denote the event such that $\mathcal{E}_{\theta_u} = \left\{e^{\|\theta_u\|_2 V_1/\sigma} \in [1, (1+\rho)/(1-\rho)]\right\}$. Using the observation above and the fact that $V_1 \sim \mathcal{N}(0,1)$, we obtain that

$$\begin{aligned}
[B_{\theta_u}]_{11} &= \mathbb{E}\left[\frac{V_1^2}{\left(\pi \exp\left(-\|\theta_u\|_2 V_1/\sigma\right) + (1-\pi)\exp\left(\|\theta_u\|_2 V_1/\sigma\right)\right)^2}\right] \\
&\leq \frac{1}{(1-\rho^2)}\mathbb{E}\left[V_1^2\,\mathbb{I}(\mathcal{E}_{\theta_u})\right] + \mathbb{E}\left[V_1^2\,\mathbb{I}(\mathcal{E}_{\theta_u}^c)\right] \\
&= \frac{1-\rho^2+\rho^2\mathbb{E}\left[V_1^2\,\mathbb{I}(\mathcal{E}_{\theta_u})\right]}{(1-\rho^2)}. \qquad (5.30)
\end{aligned}$$

Note that whenever $\theta_u \neq 0$, we have that $\mathcal{E}_{\theta_u} \subseteq \left\{V_1 \geq 0\right\}$ and consequently we obtain that

$$\mathbb{E}\left[V_1^2\,\mathbb{I}(\mathcal{E}_{\theta_u})\right] \leq \mathbb{E}\left[V_1^2\,\mathbb{I}(V_1 \geq 0)\right] = \frac{1}{2}. \qquad (5.31)$$

Putting the inequalities (5.30) and (5.31) together, we conclude that $[B_{\theta_u}]_{11} \leq (1-\rho^2/2)/(1-\rho^2)$.

Bounding $[B_{\theta_u}]_{jj},\ j \neq 1$: Using arguments similar to the previous case, and the fact that the random variables $V_i, i \in [d]$ are independent standard normal random variables, we find that

$$\begin{aligned}
[B_{\theta_u}]_{jj} &= \mathbb{E}\left[\frac{V_j^2}{\left(\pi \exp\left(-\|\theta_u\|_2 V_1/\sigma\right) + (1-\pi)\exp\left(\|\theta_u\|_2 V_1/\sigma\right)\right)^2}\right] \\
&= \mathbb{E}\left[\frac{1}{\left(\pi \exp\left(-\|\theta_u\|_2 V_1/\sigma\right) + (1-\pi)\exp\left(\|\theta_u\|_2 V_1/\sigma\right)\right)^2}\right] \\
&\leq \frac{1}{(1-\rho^2)}\mathbb{E}\left[\mathbb{I}(\mathcal{E}_{\theta_u})\right] + \mathbb{E}\left[\mathbb{I}(\mathcal{E}_{\theta_u}^c)\right] \\
&= \frac{1-\rho^2+\rho^2\mathbb{E}\left[\mathbb{I}(\mathcal{E}_{\theta_u})\right]}{(1-\rho^2)}.
\end{aligned}$$

Noting that $\mathbb{E}\left[\mathbb{I}(\mathcal{E}_{\theta_u})\right] \leq \mathbb{E}\left[\mathbb{I}(V_1 \geq 0)\right] = 1/2$ whenever $\theta_u \neq 0$, yields the claim.

Proof of Theorem 10

We split our proof into two parts, corresponding to the upper bound (5.15a) and the lower bound (5.15b) respectively.

Proof of the upper bound (5.15a)

For the balanced fit, we have

$$w_\theta(X) = \frac{1}{1+e^{-2\theta^\top X/\sigma^2}} \quad \text{and} \quad \nabla_\theta(w_\theta(X)) = \frac{2X^\top/\sigma^2}{(e^{-\theta^\top X/\sigma^2} + e^{\theta^\top X/\sigma^2})^2}.$$

Using a Taylor expansion and repeating the preliminary computations as those in the proof of Theorem 9 from the unbalanced setting, we obtain that

$$\begin{aligned}\|M(\theta)\|_2 &= \|\mathbb{E}\left[2X\int_0^1 w'_{\theta_u}(X)^\top \theta_u du\right]\|_2 \\ &= 4\left\|\int_0^1 \mathbb{E}\left[\frac{XX^\top}{\sigma^2\left(e^{-\theta_u^\top X/\sigma^2}+e^{\theta_u^\top X/\sigma^2}\right)^2}\right]\theta du\right\|_2 \\ &\leq 4\|\theta\|_2\int_0^1 |\!|\!|\mathbb{E}\left[\Gamma_{\theta_u}(X)\right]|\!|\!|_{\mathrm{op}} du,\end{aligned} \tag{5.32}$$

where $\Gamma_{\theta_u}(X) := \frac{XX^\top/\sigma^2}{(e^{-\theta_u^\top X/\sigma^2}+e^{\theta_u^\top X/\sigma^2})^2}$. Consequently, in order to prove the upper bound (5.15a), it suffices to show that

$$\int_0^1 |\!|\!|\mathbb{E}\left[\Gamma_{\theta_u}(X)\right]|\!|\!|_{\mathrm{op}} du \leq \frac{1}{4}\left(p+\frac{1-p}{1+\|\theta\|_2^2/2\sigma^2}\right) = \frac{\gamma_{\mathrm{up}}(\theta)}{4} \tag{5.33}$$

where $p := (1+\mathbb{P}_{Z\sim\mathcal{N}(0,1)}(|Z|\leq 1))/2 < 1$.

We now establish the claim (5.33). Like in proof of Theorem 9, we perform a change of coordinates using an orthogonal transformation R such that $R\theta_u = \|\theta_u\|_2 e_1$, where e_1 is the first canonical basis in dimension d. Define the random vector $V := RX/\sigma$. Since the vector $X \sim \mathcal{N}(0,\sigma^2 I_d)$ and the matrix R is orthogonal, we have that the vector $V \sim \mathcal{N}(0, I_d)$. Substituting $X = \sigma R^\top V$, $R\theta_u = \|\theta_u\|_2 e_1$ in the expression for Γ_{θ_u} and using the fact that $|\!|\!|R^\top B R|\!|\!|_{\mathrm{op}} = |\!|\!|B|\!|\!|_{\mathrm{op}}$ for any matrix B and any orthogonal matrix R, we obtain that $|\!|\!|\mathbb{E}\left[\Gamma_{\theta_u}(X)\right]|\!|\!|_{\mathrm{op}} = |\!|\!|B_{\theta_u}|\!|\!|_{\mathrm{op}}$, where

$$B_{\theta_u} := \mathbb{E}_V\left[\frac{VV^\top}{\left(\exp\left(-\|\theta_u\|_2 V_1/\sigma\right)+\exp\left(\|\theta_u\|_2 V_1/\sigma\right)\right)^2}\right].$$

Clearly, the matrix B_{θ_u} is a diagonal matrix with non-negative entries. Consequently, to obtain a bound for the operator norm of the matrix B_{θ_u}, it is sufficient to control the diagonal entries of the matrix B_{θ_u}. We introduce an auxiliary claim:

Lemma 28. *The ℓ_2-operator norm of the matrix B_{θ_u} is bounded as*

$$|\!|\!|B_{\theta_u}|\!|\!|_{op} = \max_{j\in[d]}[B_{\theta_u}]_{jj} \leq \frac{p_2}{4} + \frac{(1-p_2)}{4}\frac{1}{(1+\|\theta_u\|_2^2/(2\sigma^2))^2}, \tag{5.34}$$

where $p_2 = \mathbb{P}\left(|V_1|\leq 1\right) < 1$.

See Section 5.6 for the proof of Lemma 28.

Using Lemma 28, we now complete the proof. Integrating both sides of the inequality (5.34) with respect to $u \in [0,1]$, we find that

$$\begin{aligned}\int_0^1 \|\!| B_{\theta_u} |\!\|_{\mathrm{op}} du &\leq \int_0^1 \frac{p_2}{4} du + \int_0^1 \frac{(1-p_2)}{4} \frac{1}{(1+\|\theta_u\|_2^2/(2\sigma^2))^2} du \\ &= \frac{p_2}{4} + \frac{(1-p_2)}{4} \int_0^1 \frac{1}{(1+u^2\|\theta\|_2^2/(2\sigma^2))^2} du.\end{aligned}$$

Direct computation of the second integral yields

$$\int_0^1 \frac{1}{(1+u^2\|\theta\|_2^2/(2\sigma^2))^2} du = \frac{1}{2}\left(\frac{1}{1+\frac{\|\theta\|^2}{2\sigma^2}} + \frac{\tan^{-1}(\|\theta\|/(\sqrt{2}\sigma)}{\|\theta\|/(\sqrt{2}\sigma)}\right) \leq \frac{1}{2}\left(\frac{1}{1+\frac{\|\theta\|^2}{2\sigma^2}} + 1\right),$$

where the last inequality above follows since $\tan^{-1}(x) \leq x$, for all $x \in \mathbb{R}$. Putting together the pieces yields

$$\int_0^1 \|\!| \mathbb{E}\left[\Gamma_{\theta_u}(X)\right] |\!\|_{\mathrm{op}} du = \int_0^1 \|\!| B_{\theta_u} |\!\|_{\mathrm{op}} du \leq \frac{(1+p_2)}{8} + \frac{(1-p_2)/8}{1+\|\theta\|_2^2/(2\sigma^2)},$$

which implies the claim (5.33) with $p = \frac{1+p_2}{2}$.

Proof of the lower bound (5.15b)

We now prove the lower bound (5.15b) on the population EM operator $M(\theta)$. The argument involves Jensen's inequality and certain properties of the moment generating function (MGF) of the Gaussian distribution.

Recalling equation (5.32), we find that

$$\begin{aligned}\|M(\theta)\|_2 = \|M(\theta) - M(0)\|_2 &= 4\Big\| \underbrace{\int_0^1 \mathbb{E}\left[\frac{XX^\top}{\sigma^2\left(\exp\left(-\theta_u^\top X/\sigma^2\right) + \exp\left(\theta_u^\top X/\sigma^2\right)\right)^2}\right] du}_{=:\Gamma_\theta} \ \theta \Big\|_2 \\ &\geq 4\lambda_{\min}\left(\Gamma_\theta\right)\|\theta\|_2, \qquad (5.35)\end{aligned}$$

where $\lambda_{\min}(\Gamma_\theta)$ denotes the smallest eigenvalue of the square matrix B. Following the change of variable $V := RX/\sigma$ used in the proof of upper bound (5.15a), we obtain that

$$\lambda_{\min}\left(\Gamma_\theta\right) = \lambda_{\min}\left(\mathbb{E}_V\left[\int_0^1 \frac{VV^\top}{\left(\exp\left(-\|\theta_u\|_2 V_1/\sigma\right) + \exp\left(\|\theta_u\|_2 V_1/\sigma\right)\right)^2} du\right]\right).$$

We denote the matrix on the right-hand side by B_θ. Clearly, the matrix B_θ is a diagonal matrix with non-negative diagonal entries and consequently, we have

$$\lambda_{\min}(B_\theta) = \min_{j\in[d]} [B_\theta]_{jj}. \qquad (5.36)$$

We make the following auxiliary claim:

Lemma 29. *For all vectors $\theta \in \mathbb{R}^d$ such that $\|\theta\|_2^2 \leq \frac{5\sigma^2}{8}$, the square matrix B_θ satisfies the bounds*

$$[B_\theta]_{jj} \geq [B_\theta]_{11} \geq \frac{1}{4(1 + 2\|\theta\|_2^2/\sigma^2)} \quad \text{for all } j \in [d]. \tag{5.37}$$

See Section 5.7 for the proof of this claim.

Combining the result of Lemma 29 with bounds (5.35) and (5.36), we conclude that

$$\frac{\|M(\theta)\|_2}{\|\theta\|_2} \geq 4\lambda_{\min}(\Gamma_\theta) \;=\; 4[B_\theta]_{11} \;\geq\; \frac{1}{(1 + 2\|\theta\|_2^2/\sigma^2)} = \gamma_{\text{lo}}(\theta),$$

as claimed.

Proof of Theorem 11

The reader should recall the framework that was laid out in Section 5.2, especially Lemma 27 which was used to bound the deviation between the sample and population EM operators, as well the epoch-based localization argument sketched out in Section 5.2. The proof of Theorem 11 is based on making this sketch more precise.

Epochs and non-expansivity

Let us introduce the notation required to formalize the epoch-based analysis that leads to the recursion (5.22). For convenience, recall that this recursion generates the sequence $\{\alpha_\ell\}_{\ell \geq 0}$ given by

$$\alpha_0 = 0 \quad \text{and} \quad \alpha_{\ell+1} = \frac{\alpha_\ell}{3} + \frac{1}{6}. \tag{5.38a}$$

By inspection, this sequence is increasing and satisfies $\lim_{\ell \to \infty} \alpha_\ell = 1/4$. Furthermore, we have $\alpha_\ell \leq 1/4 - \epsilon$ for $\ell \geq \lceil \log(4/\epsilon)/\log 3 \rceil$. For any given $\delta \in (0, 1)$, define the following intermediate quantity

$$\omega = \sigma^2 \frac{d}{n} \log((2\ell_\epsilon + 1)/\delta), \qquad \text{where } \ell_\epsilon := \lceil \log(4/\epsilon)/\log 3 \rceil + 1. \tag{5.38b}$$

Note that the lower bound on the sample size stated in the theorem ensures that $\omega \leq 1$. For the proof sketch in Section 5.2, we used the rough approximation $\omega \approx d/n$, since we were tracking only the dependencies on n and d.

For $\ell = 0, 1, 2, \ldots, \ell_\epsilon - 1$, define the scalars t_ℓ and T_ℓ as

$$t_0 = \left\lceil \frac{2}{p} \log \frac{\|\theta_0\|_2}{\sqrt{2}\sigma\sqrt{\omega}} \right\rceil, \quad t_\ell = \left\lceil \frac{2}{p\omega^{2\alpha_{\ell+1}}} \log(1/\omega) \right\rceil, \quad \text{and} \quad T_\ell = \sum_{j=0}^{\ell} t_j, \tag{5.38c}$$

where we have used the constant $p \in (0,1)$ from Theorem 10, viz, $p = \mathbb{P}(|X| \leq 1) + \frac{1}{2}\mathbb{P}(|X| > 1)$ where $X \sim \mathcal{N}(0,1)$. For each $\ell = 1, 2, \ldots$, the term t_ℓ corresponds to the number of iterations for the ℓ-th epoch, whereas the quantity T_ℓ denotes the total number of iterations up to the completion of that epoch.

Recall that Lemma 27 gives a bound on the quantity $\sup_{\|\theta\|_2 \leq r} \|M_n(\theta) - M(\theta)\|_2$ for a given radius r. In the epoch-based argument, we have a sequence of such radii, so that we need to control this same quantity uniformly over all radii r in the set $\mathcal{R}$ given by

$$\mathcal{R} = \left\{ \|\theta^0\|_2, \sqrt{2}\sigma\omega^{\alpha_0}, \ldots, \sqrt{2}\sigma\omega^{\alpha_{\ell_\epsilon - 1}}, c'\sqrt{2}\sigma\omega^{\alpha_0}, \ldots, c'\sqrt{2}\sigma\omega^{\alpha_{\ell_\epsilon - 1}} \right\}. \tag{5.39}$$

Here $c' = (2c_2\sigma/p + 1)$ denotes a constant independent of n, d, δ and ϵ where c_2 is the universal constant that appeared in equation (5.20) in Lemma 27. In order to do so, we apply a standard union bound with Lemma 27 and obtain that

$$\sup_{\|\theta\|_2 \leq r} \|M_n(\theta) - M(\theta)\|_2 \leq c_2\sigma r\sqrt{\omega} \quad \text{for all } r \in \mathcal{R}, \tag{5.40}$$

with probability at least $1 - \delta$. We denote the event defined in equation (5.40) as $\mathcal{E}(n, d, \epsilon, \delta)$.

With this notation in place, we start with our first claim: the sample-based EM operator is *non-expansive* in the following sense:

Lemma 30. *Given a sample size $n \geq (2c_2\sigma/p)^{1/(2\epsilon)}\sigma^2 d \log((2\ell_\epsilon + 1)/\delta)$, suppose that there exists an index $\ell \in \{0, 1, \ldots, \ell_\epsilon - 1\}$ and an iteration number t such that the sample-based EM iteration $\theta^{t+1} = M_n(\theta^t)$ satisfies $\|\theta^t\|_2 \leq \sqrt{2}\sigma\omega^{\alpha_\ell}$. Then, conditional on the event $\mathcal{E}(n, d, \epsilon, \delta)$ from equation (5.40), we have*

$$\|\theta^{t'}\|_2 \leq \sqrt{2}\sigma\omega^{\alpha_\ell} \qquad \text{for all } t' \geq t. \tag{5.41}$$

See Section 5.8 for the proof of this claim.

Core of the argument

We now proceed to the core of the argument. Suppose that the sample size be lower bounded as

$$n \geq \max\left\{ \left(\frac{5554\sigma + p}{p}\right)^{4/\epsilon}, \left(\frac{5554\|\theta_0\|_2 + \sqrt{2}p}{\sqrt{2p}}\right)^2 \right\} \cdot \sigma^2 \cdot d \log\left(\frac{3\log(4/\epsilon)}{\delta}\right). \tag{5.42}$$

Moreover, recall that the quantity ω and the timesteps T_ℓ were defined in equations (5.38b) and (5.38c) respectively. The core of the proof consists of the following:

Key claim: For all $\ell \in \{0, 1, \ldots, \ell_\epsilon - 1\}$, we have

$$\|\theta^t\|_2 \leq \sqrt{2}\sigma\omega^{\alpha_\ell} \qquad \text{for all } t \geq T_\ell, \tag{5.43}$$

with probability at least $1 - \delta$.

Taking this claim as given, let us now show how the bounds in Theorem 11 hold for all $t \geq T_{\ell_\epsilon - 1}$. Straightforward computations yield that

$$\begin{aligned} T_{\ell_\epsilon - 1} \leq T_0 + (\ell_\epsilon - 1)t_{\ell_\epsilon - 1} &\leq \frac{4}{p}\left[\log\frac{\|\theta_0\|_2}{\sqrt{2}\sigma\sqrt{\omega}} + \log\frac{4}{\epsilon}\cdot\omega^{1/2-2\epsilon}\cdot\log\frac{n}{\sigma^2 d}\right] \\ &\leq \frac{8}{p}\left[\log\frac{\|\theta_0\|_2^2 n}{\sigma^2 d} + \left(\frac{n}{d}\right)^{\frac{1}{2}-2\epsilon}\cdot\log\left(\frac{4}{\epsilon}\right)\cdot\log\left(\frac{n}{\sigma^2 d}\right)\cdot\sigma^{4\epsilon-1}\right]. \end{aligned} \tag{5.44}$$

In other words, equations (5.42) and (5.44) provide the explicit expression for the number of samples and number of steps required by sample-based EM to converge to a ball of radius $(d/n)^{1/4-\epsilon}$ around the truth $\theta^* = 0$.

Proof of the claim (5.43): It remains to prove the key claim, and we do so by induction on the epoch index ℓ. All of the argument is performed conditioned on the event $\mathcal{E}(n, d, \epsilon, \delta)$ (5.40), which occurs with probability at least $1 - \delta$. Note that the sample size assumption (5.42) for Theorem 11 is larger than required in Lemma 30 and hence we can invoke the non-expansiveness of the sample-based EM operator in our arguments that follow.

Proof of base case $\ell = 0$: In order to simplify notation, we let $\nu = \|\theta_0\|_2/\sqrt{2}\sigma$. The non-expansiveness property of the sampled-based EM-operator (Lemma 30) ensures that it is sufficient to consider the case that $\|\theta^t\|_2 \in [\sqrt{2}\sigma, \nu\sqrt{2}\sigma]$ for all $t \leq T_0$. Applying the triangle inequality yields

$$\|\theta^{t+1}\|_2 \leq \|M_n(\theta^t) - M(\theta^t)\|_2 + \|M(\theta^t)\|_2 \tag{5.45a}$$

$$\overset{(i)}{\leq} c_2\sigma\cdot\nu\sqrt{2}\sigma\cdot\sqrt{\omega} + \gamma_{\text{up}}(\theta^t)\|\theta^t\|_2, \tag{5.45b}$$

where step (i) follows from using $r = \nu\sqrt{2}\sigma$ in the event (5.40) and Theorem 10. Noting that $\|\theta^t\|_2 \geq \sqrt{2}\sigma$, we also have that

$$\gamma_{\text{up}}(\theta^t) = 1 - p + \frac{p}{1 + \|\theta^t\|_2^2/\sigma^2} = 1 - \frac{p\|\theta^t\|_2^2}{\|\theta^t\|_2^2 + 2\sigma^2} \leq \underbrace{1 - \frac{p}{2}}_{\gamma_0}.$$

Recursing the inequalities (5.45a) and (5.45b) from $t = 0$ up to $t = T_0$, and using the fact that $\gamma_{\text{up}}(\theta^t) \leq \gamma_0$, we find that

$$\begin{aligned} \|\theta^{T_0}\|_2 &\leq c_2\sigma\cdot\nu\sqrt{2}\sigma\cdot\sqrt{\omega}(1 + \gamma_0 + \ldots + \gamma_0^{T_0-1}) + \gamma_0^{T_0}\|\theta^0\|_2 \\ &\leq \frac{c_2\sigma\cdot\nu\sqrt{2}\sigma\cdot\sqrt{\omega}}{1-\gamma_0} + \gamma_0^{T_0}\nu\sqrt{2}\sigma. \end{aligned}$$

Substituting the expressions $\gamma_0 = 1 - p/2$ and $T_0 = \lceil (2/p)\log(\nu/\sqrt{\omega})\rceil$, we obtain that

$$\|\theta^{T_0}\|_2 \leq (2\nu c_2\sigma/p + 1)\sqrt{\omega}\sqrt{2}\sigma \leq \sqrt{2}\sigma,$$

where the last inequality follows from the fact that for the assumed bound (5.42) on n, we have $(2\nu c_2\sigma/p + 1)\sqrt{\omega} \leq 1$. The base case now follows.

Proof of inductive step: Now we prove the inductive step. In particular, we assume that $\|\theta^{T_\ell}\|_2 \leq \sqrt{2}\sigma\omega^{\alpha_\ell}$ and show that $\|\theta^{T_{\ell+1}}\|_2 \leq \sqrt{2}\sigma\omega^{\alpha_{\ell+1}}$. Once again, Lemma 30 implies that we may assume without loss of generality that $\|\theta^t\|_2 \in [\omega^{\alpha_{\ell+1}}, \omega^{\alpha_\ell}]$ for all $t \in \{T_\ell, \ldots, T_{\ell+1}\}$. Under this condition, we have that

$$\gamma_{\mathrm{up}}(\theta^t) \leq 1 - \frac{p\omega^{2\alpha_{\ell+1}}}{1+\omega^{2\alpha_{\ell+1}}} \leq \underbrace{1 - \frac{p\omega^{2\alpha_{\ell+1}}}{2}}_{=:\gamma_\ell} \quad \text{for all } t \in \{T_\ell, \ldots, T_{\ell+1} - 1\}, \tag{5.46}$$

where the last step follows from the fact that $\omega \in [0,1]$ and $\alpha_\ell \geq 0$. From our earlier definition definition (5.38c), we have $T_{\ell+1} = T_\ell + t_\ell$. We split the remainder of our proof in two parts. First, we show that

$$\|\theta^{T_\ell + \lceil t_{\ell+1}/2\rceil}\|_2 \leq c'\sqrt{2}\sigma\omega^{\alpha_{\ell+1}}, \tag{5.47a}$$

where $c' = (2c_2\sigma/p + 1)$ is a constant independent of n, d, δ and ϵ. Next we use this result to show that

$$\|\theta^{T_{\ell+1}}\|_2 = \|\theta^{T_\ell + t_{\ell+1}}\|_2 \leq \sqrt{2}\sigma\omega^{\alpha_{\ell+1}}, \tag{5.47b}$$

which completes the proof of the induction step. We now prove these two claims one by one.

Proof of claim (5.47a): Applying the triangle inequality yields

$$\|\theta^{T_\ell+1}\|_2 \leq \|M_n(\theta^{T_\ell}) - M(\theta^{T_\ell})\|_2 + \|M(\theta^{T_\ell})\|_2 \overset{(i)}{\leq} c_2\sigma \cdot \sqrt{2}\sigma\omega^{\alpha_\ell} \cdot \sqrt{\omega} + \gamma_{\mathrm{up}}(\theta^{T_\ell})\|\theta^{T_\ell}\|_2, \tag{5.48}$$

where step (i) follows from Theorem 10 and using $r = \sqrt{2}\sigma\omega^{\alpha_\ell}$ in the event (5.40). Recursing the inequality (5.48) for $T \leq \lceil t_\ell/2\rceil$ steps, and invoking the bound (5.46), i.e., $\gamma_{\mathrm{up}}(\theta^t) \leq \gamma_\ell$ for all $t \in \{T_\ell, \ldots, T_\ell + T\}$, we obtain that

$$\begin{aligned}
\|\theta^{T_\ell+T}\|_2 &\leq c_2\sigma \cdot \sqrt{2}\sigma\omega^{\alpha_\ell} \cdot \sqrt{\omega} \cdot (1 + \gamma_\ell + \ldots + \gamma_\ell^{T-1}) + \gamma_\ell^T\|\theta^{T_\ell}\|_2 \\
&\leq \frac{c_2\sigma \cdot \sqrt{2}\sigma\omega^{\alpha_\ell} \cdot \sqrt{\omega}}{1-\gamma_\ell} + \sqrt{2}\sigma\gamma_\ell^T\omega^{\alpha_\ell} \\
&\overset{(i)}{\leq} c_2\sigma \cdot \sqrt{2}\sigma \cdot (2/p) \cdot \omega^{\alpha_\ell + 1/2 - 2\alpha_{\ell+1}} + e^{-Tp\omega^{2\alpha_\ell+1}/2} \cdot \sqrt{2}\sigma\omega^{\alpha_\ell} \\
&\overset{(ii)}{\leq} \sqrt{2}\sigma\omega^{\alpha_\ell + 1/2 - 2\alpha_{\ell+1}} \cdot (2c_2\sigma/p + 1) \\
&\overset{(iii)}{=} c'\sqrt{2}\sigma\omega^{\alpha_{\ell+1}},
\end{aligned}$$

where step (i) follows from the inequality (5.46) and the consequent bound $\gamma_\ell \leq e^{-p/(2\omega^{2\alpha_{\ell+1}})}$. Furthermore, in step (ii), we used the following bound

$$\gamma_\ell^T \leq e^{-Tp\omega^{2\alpha_{\ell+1}}/2} \leq \omega^{1/2-2\alpha_{\ell+1}} \quad \text{for } T \geq \frac{(1-4\alpha_{\ell+1})}{p\omega^{2\alpha_{\ell+1}}} \log \frac{1}{\omega}, \tag{5.49}$$

and in step (iii) we invoked the relation (6.38c), i.e., $3\alpha_{\ell+1} = 1/2 + \alpha_\ell$. The claim now follows from noting that $T = \lceil t_\ell/2 \rceil$ satisfies the condition of equation (5.49).

Proof of claim (5.47b)**:** The proof of this claim makes use of arguments similar to those used above in the proof of claim (5.47a). Starting at time $T_\ell + \lceil t_{\ell+1}/2 \rceil$, and applying the triangle inequality, we find that

$$\|\theta^{T_\ell + \lceil t_{\ell+1}/2 \rceil + 1}\|_2 \leq c_2\sigma \cdot c'\sqrt{2}\sigma\omega^{\alpha_{\ell+1}} \cdot \sqrt{\omega} + \gamma_\ell \|\theta^{T_\ell + \lceil t_{\ell+1}/2 \rceil}\|_2,$$

where we have used the bound (5.40) with $r = c'\sqrt{2}\sigma\omega^{\alpha_{\ell+1}}$. Repeating this inequality for $T \geq \frac{(1-4\alpha_{\ell+1})}{p\omega^{2\alpha_{\ell+1}}} \log \frac{1}{\omega}$ steps and performing computations similar to the proof above, we find that

$$\|\theta^{T_\ell + \lceil t_{\ell+1}/2 \rceil + T}\|_2 \leq \sqrt{2}\sigma\omega^{\alpha_{\ell+1}+1/2-2\alpha_{\ell+1}} \cdot c' \cdot (2c_2\sigma/p + 1) = c'^2\omega^{1/2-2\alpha_{\ell+1}} \cdot \sqrt{2}\sigma\omega^{\alpha_{\ell+1}}.$$

Observe that $2\alpha_{\ell+1} - 1/2 \leq -2\epsilon$ for all $\ell \leq \ell_\epsilon - 1$ and that the sample size given by bound (5.42) satisfies $n \geq d \log(2\ell_\epsilon/\delta)(c')^{4/\epsilon}$; together, these facts imply that $c'^2\omega^{1/2-2\alpha_{\ell+1}} \leq 1$. The claim now follows.

Proof of Theorem 12

We now turn to the proof of the lower bound on the accuracy of EM fixed points, as stated in Theorem 12.

Recalling the definition (5.9) of the sample-based EM operator M_n, the fixed point equation can be re-written as

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i \tanh\left(\frac{\widehat{\theta}_n X_i}{\sigma^2}\right), \tag{5.50}$$

where $\widehat{\theta}_n$ denotes the fixed point solution. Our proof makes use of the following elementary bounds on the hyperbolic tangent function:

$$x \cdot \tanh(\alpha x) \geq \alpha x^2 - \frac{1}{3}\alpha^3 x^4, \quad \text{for } \alpha \geq 0, \quad \text{and} \tag{5.51a}$$

$$x \cdot \tanh(\alpha x) \leq \alpha x^2 - \frac{1}{3}\alpha^3 x^4, \quad \text{for } \alpha < 0. \tag{5.51b}$$

In order to keep the proof self-contained, we prove these bounds at the end of this proof. Now plugging in $\alpha = \widehat{\theta}_n/\sigma^2$ and employing the bound (5.51a) for the case $\widehat{\theta}_n \geq 0$ and the bound (5.51b) for the case $\widehat{\theta}_n < 0$, we find that

$$|\widehat{\theta}_n| \geq \frac{|\widehat{\theta}_n|}{\sigma^2} \cdot \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{|\widehat{\theta}_n|^3}{3\sigma^6} \cdot \frac{1}{n}\sum_{i=1}^{n} X_i^4.$$

Denoting $Y_i = X_i/\sigma$ for $i \in [n]$ and re-arranging the terms yields that

$$|\widehat{\theta}_n|^3 \geq \frac{3\sigma^2 \left(\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - 1\right) |\widehat{\theta}_n|}{\frac{1}{n}\sum_{i=1}^{n} Y_i^4}. \tag{5.52}$$

Note that the random variables $Y_i \sim \mathcal{N}(0,1)$ and thereby the quantity on the RHS above is a ratio of empirical moments of Gaussian random variables. Now to obtain the lower bound for $|\widehat{\theta}_n|$ from the inequality (5.52), we employ a few standard probability bounds for the concentration of moments of standard Gaussian distribution (refer to Theorem 5.2 in Inglot [Ing10] and Theorem 6.7 in Janson [Jan97]). In particular, we have

$$\mathbb{P}\left[\frac{\sum_{i=1}^{n} Y_i^2}{n} - 1 \geq \frac{\log 17}{n} + \frac{\sqrt{\log 17}/4}{\sqrt{n}}\right] \geq \frac{1}{17}, \qquad \text{and} \tag{5.53a}$$

$$\mathbb{P}\left[\frac{\sum_{i=1}^{n} Y_i^4}{n} \leq c_1\right] \geq 1 - \frac{1}{34}, \tag{5.53b}$$

where $c_1 = (e\log(34)/2)^2\sqrt{6}$. Plugging these bounds in the inequality (5.52), we find that

$$\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - 1}{\frac{1}{n}\sum_{i=1}^{n} Y_i^4} \geq \frac{\sqrt{\log 17}/4}{c_1} \cdot \frac{1}{\sqrt{n}}, \tag{5.54}$$

with probability at least $1/34$. Now, we claim that

$$\left\{\text{There exists at least two non-zero fixed points } \widehat{\theta}_n.\right\} \subseteq \left\{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 > 1\right\}. \tag{5.55}$$

Deferring the proof of this claim to end of this section, we now complete the proof of our original claim. Note that the event $\{\sum_{i=1}^{n} Y_i^2/n > 1\}$ is implied by the event in the bound (5.53a), and hence we have non-zero fixed points under the same event. Now, for any of these non-zero fixed points, dividing both sides of inequality (5.52) by $|\widehat{\theta}_n|$ and using the bound (5.54), we conclude that

$$\mathbb{P}\left[|\widehat{\theta}_n|^2 \geq \frac{\sqrt{\log 17}/4}{c_1} \cdot \frac{1}{\sqrt{n}}\right] \geq \frac{1}{34},$$

as claimed.

We now prove our earlier claims (5.51a)-(5.51b) and (5.55).

Proof of the bounds (5.51a) and (5.51b): Note that it suffices to establish that

$$y \tanh(y) \geq y^2 - y^4/3, \quad \text{for all } y \in \mathbb{R}. \tag{5.56}$$

Indeed, a change of variable $y = \alpha x$ and dividing both sides by α yields the desired claims. Using the fact that $\tanh(y) = (e^y - e^{-y})/(e^y + e^{-y})$, it remains to verify that

$$y(e^y - e^{-y}) \geq (e^y + e^{-y}) \cdot (y^2 - y^4/3)$$

or equivalently that

$$\sum_{k=0}^{\infty} \frac{2y^{2k+2}}{(2k+1)!} \geq \sum_{k=0}^{\infty} \frac{2y^{2k}}{(2k)!} \cdot (y^2 - y^4/3) = \sum_{k=0}^{\infty} \frac{2y^{2k+2}}{(2k)!} \cdot (1 - y^2/3),$$

which simplifies to

$$\sum_{k=1}^{\infty} \frac{y^{2k+2}}{(2k+1)!} \left(\frac{1}{(2k+1)!} - \frac{1}{(2k)!} + \frac{1}{3(2k-2)!} \right) \geq 0.$$

Since only even powers of y exist on both sides in the power series, it suffices to verify that each coefficient on the LHS is non-negative. After some algebra, we find that the condition above reduces to

$$\frac{1}{2k+1} + \frac{(2k-1)2k}{3} - 1 \geq 0, \quad \text{for all } k \geq 1.$$

This elementary inequality is indeed true, and so the proof is complete.

Proof of set-inclusion (5.55): Consider the (random) function $g : \mathbb{R} \to \mathbb{R}$ such that $g(\theta) := M_n(\theta) - \theta$. Also introduce the shorthand $Z = \sum_{i=1}^n Y_i^2/n$. Note that any fixed point of the operator M_n is a zero of the function g and vice-versa. It is easy to see that the function g is twice continuously differentiable. Now for the event $\{Z > 1\}$, the function g satisfies $g(0) = 0$ and $g'(0) > 0$ and hence there exists $c > 0$ such that $g(c) > 0$. Furthermore for any sequence of Y_i's, we have that $\lim_{\theta \to \infty} g(\theta) = -\infty$. Putting the two pieces together, we obtain that under the event $\{Z > 1\}$. the function g has at least one strictly positive root. Since g is an odd function, we also have that under the same event, g has at least one strictly negative root. The claim now follows.

5.5 Proof of Lemma 27

The proof of this lemma is based on standard arguments to derive Rademacher complexity bounds. First, we reduce the supremum of random variables over an uncountable set to a finite maximum, symmetrize with Rademacher variables, and

then apply the Ledoux-Talagrand contraction inequality We then results on sub-Gaussian and sub-exponential random variables and finally perform the associated Chernoff-bound computations to obtain the desired result.

Let $\mathbb{S}^d = \left\{u \in \mathbb{R}^d : \|u\|_2 = 1\right\}$ denote the unit sphere in d-dimensions. Then, we have

$$\begin{aligned} Z := \sup_{\theta \in \mathbb{B}(0,r)} \|M_n(\theta) - M(\theta)\|_2 &= \sup_{\theta \in \mathbb{B}(0,r)} \sup_{u \in \mathbb{S}^d} (M_n(\theta) - M(\theta))^\top u \\ &= \sup_{u \in \mathbb{S}^d} \underbrace{\sup_{\theta \in \mathbb{B}(0,r)} (M_n(\theta) - M(\theta))^\top u}_{=: Z_u}. \end{aligned}$$

Note that Z is defined as the supremum over the sphere $\mathbb{S}^d$. Using a standard discretization argument, we reduce our problem to a maximum over a finite cover. In particular, we denote $\left\{u^1, \ldots, u^N\right\}$ a $1/8$-cover for the unit sphere $\mathbb{S}^d$. It is well known that we can find such a set with $N \leq 17^d$. Given the covering set, for each vector $u \in \mathbb{S}^d$ we can find an index $j \in [N]$ such that $\|u - u^j\|_2 \leq 1/8$. Furthermore, using the triangle inequality we also have that $\|u^j\|_2 \in [7/8, 9/8]$ for all $j \in [N]$. We have

$$\begin{aligned} Z_u &= \sup_{\theta \in \mathbb{B}(0,r)} (M_n(\theta) - M(\theta))^\top \left(u - u^j + u^j\right) \\ &\leq \sup_{\theta \in \mathbb{B}(0,r)} (M_n(\theta) - M(\theta))^\top \left(u - u^j\right) + \sup_{\theta \in \mathbb{B}(0,r)} (M_n(\theta) - M(\theta))^\top u^j \\ &= Z_{u-u^j} + Z_{u^j} \\ &\leq Z\|u - u^j\|_2 + Z_{u^j}, \end{aligned}$$

where the last step follows from the fact that $Z_v \leq Z\|v\|_2$ for any vector v. Re-arranging terms and using the fact that $\|u - u^j\|_2 \leq 1/8$, followed by taking a supremum over $u \in \mathbb{S}^d$ and a corresponding maximum over $j \in [N]$, we obtain that

$$Z \leq \max_{j \in [N]} \frac{8 Z_{u^j}}{7} \leq 2 \max_{j \in [N]} Z_{u^j}. \tag{5.57}$$

Consequently, it is sufficient to study the behavior of the random variables Z_{u^j} for $j \in [N]$, which we do next.

Substituting this relation into the definitions (5.8) and (5.9) of the EM operators M_n and M, respectively, we find that

$$Z_{u^k} = \sup_{\theta \in \mathbb{B}(0,r)} \left\{ \frac{1}{n} \sum_{i=1}^n (2w_\theta(X_i) - 1) X_i^\top u - \mathbb{E}\left[(2w_\theta(X) - 1) X^\top u^k\right] \right\}.$$

Employing a standard symmetrization argument [VW00] yields

$$\mathbb{E}[\exp(\lambda Z_{u^k})] \leq \mathbb{E}\left[\exp\left(\sup_{\theta \in \mathbb{B}(0,r)} \frac{2\lambda}{n} \sum_{i=1}^n \varepsilon_i (2w_\theta(X_i) - 1) X_i^\top u^k\right)\right] \quad \text{for any } \lambda > 0, \tag{5.58}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ denote i.i.d. Rademacher random variables which are independent of $\{X_i, i \in [n]\}$. We now make use of the Ledoux-Talagrand contraction inequality for Lipschitz functions of Rademacher processes [LT91]. Noting that $|2w_\theta(x) - 2w_{\theta'}(x)| \leq \left|\theta^\top x - (\theta')^\top x\right|$ for all x, we obtain that the map $\theta \mapsto w_\theta(x)$ is Lipschitz for each fixed x. Consequently, applying the Ledoux-Talagrand contraction inequality yields that

$$\mathbb{E}\left[\exp\left(\sup_{\theta\in\mathbb{B}(0,r)} \frac{2\lambda}{n}\sum_{i=1}^n \varepsilon_i(2w_\theta(X_i)-1)X_i^\top u^k\right)\right] \leq \mathbb{E}\left[\exp\left(\sup_{\theta\in\mathbb{B}(0,r)} \frac{4\lambda}{n}\sum_{i=1}^n \varepsilon_i\theta^\top X_i X_i^\top u^k\right)\right].$$

Furthermore, using the fact that $\|u^k\|_2 \leq 9/8$ and the standard bound $u^\top B v \leq \|u\|_2 |\!|\!| B |\!|\!|_{\mathrm{op}} \|v\|_2$, we obtain that

$$\begin{aligned}\mathbb{E}\left[\exp\left(\sup_{\theta\in\mathbb{B}(0,r)} \frac{4\lambda}{n}\sum_{i=1}^n \varepsilon_i\theta^\top X_i X_i^\top u^k\right)\right] &\leq \mathbb{E}\left[\exp\left(\sup_{\theta\in\mathbb{B}(0,r)} 4\lambda\|u^k\|_2\|\theta\|_2 \,|\!|\!|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i X_i^\top |\!|\!|_{\mathrm{op}}\right)\right] \\ &\leq \mathbb{E}\left[\exp\left(\frac{9}{2}\lambda r \,|\!|\!|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i X_i^\top |\!|\!|_{\mathrm{op}}\right)\right]. \qquad (5.59)\end{aligned}$$

We now make the following two claims. The operator norm of the matrix $\sum_{i=1}^n \varepsilon_i X_i X_i^\top/n$ can be bounded as follows:

$$|\!|\!|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i X_i^\top |\!|\!|_{\mathrm{op}} \leq 2\max_{j\in[N]}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top u^j)^2\right|. \qquad (5.60a)$$

For all $(i,j) \in [n]\times[N]$, we have the following bound on the moment generating function of the random variable $\varepsilon_i(X_i^\top u^j)^2$:

$$\mathbb{E}\left[\exp(t\varepsilon_i(X_i^\top u^j)^2)\right] \leq \exp(1377/2 \cdot t^2\sigma^4) \quad \text{for all } |t| \leq \tfrac{1}{9\sigma^2}. \qquad (5.60b)$$

We provide the proofs of these auxiliary claims at the end of this subsection.

Taking them as given for the moment, let us now complete the proof of the lemma. Putting together the pieces and replacing $9/2$ by 5 in bound (5.59) to simplify calculations, we find that

$$\begin{aligned}\mathbb{E}[\exp(\lambda Z_{u^k})] &\overset{\text{(bnd. (5.58), (5.59))}}{\leq} \mathbb{E}\left[\exp\left(5\lambda r \,|\!|\!|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i X_i^\top |\!|\!|_{\mathrm{op}}\right)\right] \\ &\overset{\text{(eqn. (5.60a))}}{\leq} \mathbb{E}\left[\exp\left(\max_{j\in[N]}\frac{10\lambda r}{n}\left|\sum_{i=1}^n \varepsilon_i (X_i^\top u^j)^2\right|\right)\right] \\ &\leq \mathbb{E}\left[\exp\left(\max_{j\in[N]}\frac{-10\lambda r}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top u^j)^2\right)\right] \\ &\qquad + \mathbb{E}\left[\exp\left(\max_{j\in[N]}\frac{10\lambda r}{n}\sum_{i=1}^n \varepsilon_i (X_i^\top u^j)^2\right)\right] \\ &\overset{\text{(eqn. (5.60b))}}{\leq} 2N\cdot\prod_{i=1}^n \exp\left(1377/2\cdot\frac{100\lambda^2 r^2}{n^2}\cdot\sigma^4\right)\end{aligned}$$

for any $|\lambda| \leq n/(90r\sigma^2)$. Now invoking the inequality $2N \leq 34^d \leq e^{4d}$, we find that

$$\mathbb{E}[\exp(\lambda Z_{u^k})] \leq \exp\left(68850 \cdot \lambda^2 r^2 \sigma^4/n + 4d\right) \quad \text{for any } k \in [N], \tag{5.61}$$

and sufficiently small λ. Now using our earlier inequality (5.57) obtained from reduction and the fact that $N \leq e^{3d}$, we obtain that

$$\begin{aligned}\mathbb{E}[\exp(\lambda Z)] \leq \mathbb{E}[\exp(2\lambda \max_{j \in [N]} Z_{u^j})] &\leq N \exp(68850 \cdot 4\lambda^2 r^2 \sigma^4/n + 4d) \\ &\leq \exp(275400 \cdot \lambda^2 r^2 \sigma^4/n + 7d),\end{aligned}$$

for any $|\lambda| \leq n/(180r\sigma^2) = \lambda_\star$. Now using the standard Chernoff bound, we have that

$$\mathbb{P}\left[Z \geq \alpha\right] \leq \inf_{\lambda \in [0, \lambda_\star]} \mathbb{E}[e^{\lambda Z}] e^{-\lambda \alpha} \leq \inf_{\lambda \in [0, \lambda_\star]} e^{275400 r^2 \sigma^4/n \cdot \lambda^2 - \lambda \alpha + 7d}. \tag{5.62}$$

Furthermore, note that any quadratic function $\lambda \mapsto g(\lambda) = a\lambda^2 - b\lambda + c$ where $a, b > 0$ satisfies $g(\lambda) \geq g(b/2a) = c - b^2/4a$. Using this result with $a = 275400 r^2 \sigma^4/n, b = \alpha$ and $c = 7d$, we find that

$$\inf_{\lambda \in [0, \lambda_\star]} e^{275400 r^2 \sigma^4/n \cdot \lambda^2 - \lambda \alpha + 7d} = e^{7d - \alpha^2 n/(4 \cdot 275400 r^2 \sigma^4)} \quad \text{if } \alpha \leq 3060 r\sigma^2, \tag{5.63}$$

where the condition on α is needed to ensure that the choice of $\lambda = \alpha n/(2 \cdot 275400 r^2 \sigma^4)$ is feasible in the interval $[0, \lambda_\star]$. Noting that $2777 \geq \sqrt{4 \cdot 275400 \cdot 7}$ and that for any $\delta \in (0, 1)$, we have $d \log(1/\delta) \geq d + \log(1/\delta)$, we obtain that

$$e^{7d - \alpha^2 n/(4 \cdot 275400 r^2 \sigma^4)} \leq \delta \quad \text{for} \quad \alpha \geq 2777 r\sigma^2 \cdot \sqrt{\frac{d \log(1/\delta)}{n}}. \tag{5.64}$$

It is now straightforward to see that as long as $n \geq d \log(1/\delta)$, we can put the three bounds (5.62), (5.63) and (5.64) together to conclude that

$$Z \leq 2777 \cdot r\sigma^2 \cdot \sqrt{\frac{d \log(1/\delta)}{n}}, \quad \text{with probability at least } 1 - \delta.$$

which yields the conclusion of the lemma.

We now return to prove our earlier claims (5.60a) and (5.60b).

Proof of claim (5.60a)**:** It is convenient to prove a somewhat more general result: for any matrix $B \in \mathbb{R}^{d \times d}$, we have

$$|\!|\!| B |\!|\!|_{\text{op}} = \sup_{u \in \mathbb{S}^d} \left| u^\top B u \right| \leq 2 \max_{j \in [N]} \left| \left(u^j\right)^\top B u^j \right|. \tag{5.65}$$

Note that applying this result to the matrix $B := \sum_{i=1}^n \varepsilon_i X_i X_i^\top / n$ implies our claim (5.60a). We now prove the claim (5.65). For each vector $u \in \mathbb{S}^d$ we can find an index $j \in [N]$ such that $\|u - u^j\|_2 \leq 1/8$. Furthermore, using triangle inequality we also have that $\|u_j\|_2 \in [7/8, 9/8]$ for all $j \in [N]$. By letting $\Delta = u - u^j$, we obtain that

$$u^\top B u = \left(u^j\right)^\top B u^j + \Delta^\top B \Delta + 2\Delta^\top B u^j.$$

Note that for any vectors u, v, we have that $\left|u^\top B v\right| \leq |\!|\!| B |\!|\!|_{\mathrm{op}} \|u\|_2 \|v\|_2$. Combining this observation with the facts that $\|\Delta\|_2 \leq 1/8$ and $\|u_j\|_2 \leq 9/8$, we find that

$$\begin{aligned}
\left|u^\top B u\right| &\leq \left|\left(u^j\right)^\top B u^j\right| + |\!|\!| B |\!|\!|_{\mathrm{op}} (\|\Delta\|_2^2 + 2\|\Delta\|_2 \|u^j\|_2) \\
&\leq \left|\left(u^j\right)^\top B u^j\right| + \frac{1}{2} |\!|\!| B |\!|\!|_{\mathrm{op}}.
\end{aligned}$$

Re-arranging and then taking a supremum over $u \in \mathbb{S}^d$ and the associated maximum over $j \in [N]$ yields the desired result (5.65).

Proof of claim (5.60b)**:** Note that since $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I_d)$, we have that $X_i^\top v \sim \mathcal{N}(0, \|v\|_2^2 \sigma^2)$. For any vector v with norm $\|v\|_2 \leq 3/2$, we have that

$$\mathbb{E}[\exp(X_i^\top v)] = \exp(\|v\|^2 \sigma^2 / 2) \leq \exp(9\sigma^2/4) \quad \text{for all } i \in [n].$$

In other words, the random variables $\left\{X_i^\top v, i \in [n]\right\}$ are sub-Gaussian with parameter $\gamma = 3\sigma/2$. Since a squared sub-Gaussian random variable is a sub-exponential random variable, we obtain the following inequality [Ver]:

$$\mathbb{E}\left[\exp\left(t(X_i^\top v)^2 - t\mathbb{E}(X_i^\top v)^2\right)\right] \leq e^{16 t^2 \gamma^4} \quad \text{for all } |t| \leq \frac{1}{4\gamma^2}. \tag{5.66}$$

Noting that the random variable ε_i is independent of $X_i^\top v$, we find that

$$\begin{aligned}
\mathbb{E}\left[\exp(t\varepsilon_i (X_i^\top v)^2)\right] &= \frac{1}{2}\mathbb{E}\left[\exp(t(X_i^\top v)^2)\right] + \frac{1}{2}\mathbb{E}\left[\exp(-t(X_i^\top v)^2)\right] \\
&\overset{(i)}{\leq} e^{16 t^2 \gamma^4} \cdot \frac{1}{2}\left[e^{t\gamma^2} + e^{-t\gamma^2}\right] \\
&\overset{(ii)}{\leq} e^{17 t^2 \gamma^4},
\end{aligned}$$

for all $|t| \leq \frac{1}{4\gamma^2}$. In asserting the above sequence of steps, we have applied the inequality (5.66) along with the fact that $\mathbb{E}(X_i^\top v)^2 \leq \gamma^2$ to conclude step (i), and step (ii) follows from the inequality $e^x + e^{-x} \leq 2e^{x^2}$ for all $x \in \mathbb{R}$. Noting that $\|u^j\|_2 \leq 9/8 \leq 3/2$ (the choice of $3/2$ is made to simplify constants) and putting together the pieces yields the claim.

5.6 Proof of Lemma 28

Using the inequality $\exp(y) + \exp(-y) \geq 2 + y^2$, valid for all $y \in \mathbb{R}$, we find that

$$[B_{\theta_u}]_{11} = \mathbb{E}_V \left[\frac{V_1^2}{\left(\exp\left(-\|\theta_u\|_2 V_1/\sigma\right) + \exp\left(\|\theta_u\|_2 V_1/\sigma\right)\right)^2} \right] \leq \mathbb{E}_{V_1} \left[\frac{V_1^2}{\left(2 + V_1^2 \|\theta_u\|_2^2/\sigma^2\right)^2} \right]. \tag{5.67}$$

Let $\mathbb{I}_A$ denote the indicator random variable for event A, i.e., it takes value 1 when the event A occurs and 0 otherwise. Then we have

$$\begin{aligned} \mathbb{E}\left[\frac{V_1^2}{\left(2 + V_1^2 \|\theta_u\|_2^2/\sigma^2\right)^2} \right] &= \mathbb{E}\left[\frac{V_1^2}{\left(2 + V_1^2 \|\theta_u\|_2^2/\sigma^2\right)^2} \mathbb{I}_{\{|V_1| \leq 1\}} \right] + \mathbb{E}\left[\frac{V_1^2}{\left(2 + V_1^2 \|\theta_u\|_2^2/\sigma^2\right)^2} \mathbb{I}_{\{|V_1| > 1\}} \right] \\ &\leq \frac{1}{4}\mathbb{E}\left[V_1^2 \mathbb{I}_{\{|V_1| \leq 1\}} \right] + \mathbb{E}\left[\frac{V_1^2}{\left(2 + \|\theta_u\|_2^2/\sigma^2\right)^2} \mathbb{I}_{\{|V_1| > 1\}} \right]. \end{aligned} \tag{5.68}$$

Here the final inequality follows from the following observation:

$$\frac{V_1^2}{\left(2 + V_1^2 \|\theta_u\|_2^2/\sigma^2\right)^2} \leq \begin{cases} \dfrac{V_1^2}{4} & \text{if } |V_1| \leq 1, \\ \dfrac{V_1^2}{\left(2 + \|\theta_u\|_2^2/\sigma^2\right)^2} & \text{if } |V_1| > 1. \end{cases} \tag{5.69}$$

Putting the inequalities (5.67) and (5.68) together, we conclude that

$$[B_{\theta_u}]_{11} \leq \frac{1}{4}\mathbb{E}\left[V_1^2 \mathbb{I}_{\{|V_1| \leq 1\}} \right] + \mathbb{E}\left[\frac{V_1^2}{\left(2 + \|\theta_u\|_2^2/\sigma^2\right)^2} \mathbb{I}_{\{|V_1| > 1\}} \right],$$

where $V_1 \sim \mathcal{N}(0,1)$. Define $p_1 := \mathbb{E}\left[V_1^2 \mathbb{I}_{\{|V_1| \leq 1\}} \right]$. Then we have $\mathbb{E}\left[V_1^2 \mathbb{I}_{\{|V_1| \geq 1\}} \right] = 1 - p_1$ and consequently we obtain that

$$[B_{\theta_u}]_{11} \leq \frac{p_1}{4} + \frac{(1-p_1)}{4} \frac{1}{\left(1 + \|\theta_u\|_2^2/(2\sigma^2)\right)^2}. \tag{5.70}$$

Now we bound the entries $[B_{\theta_u}]_{jj}$, $j \neq 1$. Using the inequality $\exp(y) + \exp(-y) \geq 2 + y^2$ once again and noting that the random variables are V_j are i.i.d. $\mathcal{N}(0,1)$, we find that

$$[B_{\theta_u}]_{jj} = \mathbb{E}_V \left[\frac{V_j^2}{\left(\exp\left(-\|\theta_u\|_2 V_1/\sigma\right) + \exp\left(\|\theta_u\|_2 V_1/\sigma\right)\right)^2} \right] \leq \mathbb{E}_{V_1} \left[\frac{1}{\left(2 + V_1^2 \|\theta_u\|_2^2/\sigma^2\right)^2} \right]. \tag{5.71}$$

With an observation similar to that of inequality (5.69), we also have that

$$[B_{\theta_u}]_{jj} \leq \frac{1}{4}\mathbb{E}\left[\mathbb{I}_{\{|V_1| \leq 1\}} \right] + \mathbb{E}\left[\frac{1}{\left(2 + \|\theta_u\|_2^2/\sigma^2\right)^2} \mathbb{I}_{\{|V_1| > 1\}} \right]. \tag{5.72}$$

Define $p_2 := \mathbb{P}(|V_1| \leq 1)$. Putting together the inequalities (5.71) and (5.72), we obtain that

$$[B_{\theta_u}]_{jj} \leq \frac{p_2}{4} + \frac{(1-p_2)}{4} \frac{1}{(1 + \|\theta_u\|_2^2/(2\sigma^2))^2} \quad \text{for } j = 2, \ldots, d. \tag{5.73}$$

Note that

$$p_2 = \mathbb{P}(|V_1| \leq 1) = \mathbb{E}\left[\mathbb{I}_{\{|V_1| \leq 1\}}\right] > \mathbb{E}\left[V_1^2 \mathbb{I}_{\{|V_1| \leq 1\}}\right] = p_1,$$

and consequently, the bound on the right-hand side of inequality (5.73) is larger than the right-hand side of inequality (5.70). As a result, we have

$$|||B_{\theta_u}|||_{\text{op}} = \max_{j \in [d]} [B_{\theta_u}]_{jj} \leq \frac{p_2}{4} + \frac{(1-p_2)}{4} \frac{1}{(1 + \|\theta_u\|_2^2/(2\sigma^2))^2},$$

where $p_2 = \mathbb{P}(|V_1| \leq 1)$ and the claim (5.34) follows.

5.7 Proof of Lemma 29

We now prove the claim (5.37) in two steps. First, we show that $[B_\theta]_{jj} \geq [B_\theta]_{11}$ for all $j \in [d]$. Then, we derive the claimed lower bound for $[B_\theta]_{11}$.

Proof of $[B_\theta]_{jj} \geq [B_\theta]_{11}$: For all $j \neq 1$, by changing the order of integration, we obtain that

$$\begin{aligned}
[B_\theta]_{jj} &= \int_0^1 \mathbb{E}_V \left[\frac{V_j^2}{\left(\exp\left(-\|\theta_u\|_2 V_1/\sigma\right) + \exp\left(\|\theta_u\|_2 V_1/\sigma\right)\right)^2} \right] du \\
&\overset{(i)}{=} \int_0^1 \mathbb{E}_V \left[\frac{1}{\left(\exp\left(-\|\theta_u\|_2 V_1/\sigma\right) + \exp\left(\|\theta_u\|_2 V_1/\sigma\right)\right)^2} \right] du \\
&\overset{(ii)}{\geq} \int_0^1 \mathbb{E}_V \left[\frac{V_1^2}{\left(\exp\left(-\|\theta_u\|_2 V_1/\sigma\right) + \exp\left(\|\theta_u\|_2 V_1/\sigma\right)\right)^2} \right] du = [B_\theta]_{11},
\end{aligned}$$

where step (i) follows since $\mathbb{E}[V_j^2] = 1$, and from the fact that the random variables $\{V_j, j \neq 1\}$ are independent of the random variable V_1. Finally, note that the map $|V_1| \mapsto V_1^2$ is increasing in $|V_1|$, and for any fixed value of θ_u the function $|V_1| \mapsto \frac{1}{(\exp(-\|\theta_u\|_2 V_1/\sigma) + \exp(\|\theta_u\|_2 V_1/\sigma))^2}$ is a decreasing function of $|V_1|$; consequently, step (ii) above follows from a standard application of the Harris inequality.[5]

[5]The Harris inequality: Given any pair of functions (f, g) such that the function $f : \mathbb{R} \mapsto \mathbb{R}$ is increasing, and the function $g : \mathbb{R} \mapsto \mathbb{R}$ is decreasing. Then for any real-valued random variable U we have $\mathbb{E}(f(U)g(U)) \leq \mathbb{E}(f(U))\mathbb{E}(g(U))$. Here we have assumed that all three expectations exist and are finite.

Lower bound on $[B_\theta]_{11}$: Substituting $\theta_u = u\theta$ in the expression for $[B_\theta]_{11}$, and noting that $\int_0^1 (e^{au} + e^{-au})^{-2} du = \tanh(a)/(4a)$, we obtain that

$$\begin{aligned}
[B_\theta]_{11} &= \mathbb{E}_{V_1}\left[\int_0^1 \frac{V_1^2}{\left(\exp\left(-u\|\theta\|_2 V_1/\sigma\right) + \exp\left(u\|\theta\|_2 V_1/\sigma\right)\right)^2} du\right] \\
&= \mathbb{E}_{V_1}\left[\frac{\sigma V_1}{4\|\theta\|_2} \tanh\frac{\|\theta\|_2 V_1}{\sigma}\right] \\
&\overset{(i)}{=} \frac{1}{4}\mathbb{E}_{V_1}\left[\operatorname{sech}^2\left(\frac{\|\theta\|_2 V_1}{\sigma}\right)\right] = \mathbb{E}_{V_1}\left[\frac{1}{\left(\exp\left(-\|\theta\|_2 V_1/\sigma\right) + \exp\left(\|\theta\|_2 V_1/\sigma\right)\right)^2}\right],
\end{aligned}$$

where step (i) follows from Stein's Lemma for standard Gaussian distribution[6]. Expanding the expression in the denominator, we obtain

$$\begin{aligned}
[B_\theta]_{11} &= \mathbb{E}_{V_1}\left[\frac{1}{2 + \exp\left(-2\|\theta\|_2 V_1/\sigma\right) + \exp\left(2\|\theta\|_2 V_1/\sigma\right)}\right] \\
&\geq \frac{1}{\mathbb{E}_{V_1}\left[2 + \exp\left(-2\|\theta\|_2 V_1/\sigma\right) + \exp\left(2\|\theta\|_2 V_1/\sigma\right)\right]},
\end{aligned} \tag{5.74}$$

where the last inequality follows from Jensen's inequality applied for the convex function $t \mapsto \frac{1}{t}$ on $t \in (0, \infty)$. Noting that $V_1 \sim \mathcal{N}(0,1)$ and consequently that $\mathbb{E}_{V_1}\left(\exp\left(tV_1\right)\right) = e^{t^2/2}$ for all $t \in \mathbb{R}$, we obtain that

$$\mathbb{E}_{V_1}\Big[2 + \exp\left(-2\|\theta\|_2 V_1/\sigma\right) + \exp\left(2\|\theta\|_2 V_1/\sigma\right)\Big] = 2(1 + e^{2\|\theta\|_2^2/\sigma^2}) \leq 4(1 + 2\|\theta\|_2^2/\sigma^2), \tag{5.75}$$

for all θ such that $\|\theta\|_2^2 \leq 5\sigma^2/8$. Here the last step follows from the fact that $e^t \leq 1 + 2t$, for all $t \in [0, 5/4]$. Putting the bounds (5.74) and (5.75) together yields the claimed lower bound for $[B_\theta]_{11}$.

5.8 Proof of Lemma 30

Note that it is sufficient to show that a one-step update is non-expansive. Without loss of generality, we can assume that $\|\theta^t\|_2 \geq \sqrt{2}\sigma\omega^{\alpha_{\ell+1}}$, else we can start with the assumption $\|\theta^t\|_2 \leq \sqrt{2}\sigma\omega^{\alpha_{\ell+1}}$ and repeat the arguments that follow. Applying the

[6]Stein's Lemma: For any differentiable function $g : \mathbb{R} \mapsto \mathbb{R}$, we have $\mathbb{E}\left[Yg(Y)\right] = \mathbb{E}\left[g'(Y)\right]$ where $Y \sim \mathcal{N}(0,1)$ provided that expectations $\mathbb{E}\left[g'(Y)\right]$ and $\mathbb{E}\left[Yg(Y)\right]$ exist.

triangle inequality, we find that

$$\begin{aligned}\|\theta^{t+1}\|_2 = \|M_n(\theta^t)\|_2 &\leq \|M_n(\theta^t) - M(\theta^t)\|_2 + \|M(\theta^t)\|_2 \\ &\overset{(i)}{\leq} c_2\sigma \cdot \sqrt{2}\sigma\omega^{\alpha_\ell} \cdot \sqrt{\omega} + \gamma(\theta^t)\|\theta^t\|_2 \\ &\overset{(ii)}{\leq} c_2\sigma \cdot \sqrt{2}\sigma\omega^{\alpha_\ell} \cdot \sqrt{\omega} + \left(1 - \frac{p\omega^{2\alpha_{\ell+1}}}{2}\right)\sqrt{2}\sigma\omega^{\alpha_\ell} \\ &= \left(1 - \frac{p\omega^{2\alpha_{\ell+1}}}{2} + c_2\sigma\sqrt{\omega}\right)\sqrt{2}\sigma\omega^{\alpha_\ell},\end{aligned}$$

where step (i) follows from the bound (5.40) with $r = \|\theta^t\|_2 \leq \sqrt{2}\sigma\omega^{\alpha_\ell}$, and step (ii) follows from condition that $\|\theta^t\|_2 \geq \sqrt{2}\sigma\omega^{\alpha_{\ell+1}}$ and consequently that $\gamma(\theta^t) \leq 1 - p\omega^{2\alpha_{\ell+1}}/2$. Note that $2\alpha_{\ell+1} - 1/2 \leq -2\epsilon$ for all $\ell \leq \ell_\epsilon - 1$ and $\omega \leq 1$. As a result, for $n \geq (2c_2\sigma/p)^{1/(2\epsilon)}\sigma^2 d\log(2\ell_\epsilon/\delta)$, we have that $\omega^{2\alpha_{\ell+1}-1/2} \geq \omega^{-2\epsilon} \geq 2c_2\sigma/p$ and thereby that

$$\left(1 - \frac{p\omega^{2\alpha_{\ell+1}}}{2} + c_2\sigma\sqrt{\omega}\right) \leq 1.$$

Putting all the pieces together yields the result.

Chapter 6

Instability, computational efficiency and statistical accuracy

Many statistical estimators are defined as the fixed point of a data-dependent operator, with estimators based on minimizing a cost function being an important special case. The limiting performance of such estimators depends on the properties of the population-level operator in the idealized limit of infinitely many samples. In this chapter we develop a general framework that yields bounds on statistical accuracy based on the interplay between the deterministic convergence rate of the algorithm at the population level, and its degree of (in)stability when applied to an empirical object based on n samples. The key insight in this chapter is a genralization of the localization-argument discussed in Chapter 5. Using this framework, we analyze both stable forms of gradient descent and some higher-order and unstable algorithms, including Newton's method and its cubic-regularized variant, as well as the EM algorithm. We provide applications of our general results to several concrete classes of models, including Gaussian mixture estimation, single-index models, and informative non-response models. We exhibit cases in which an unstable algorithm can achieve the same statistical accuracy as a stable algorithm in exponentially fewer steps—namely, with the number of iterations being reduced from polynomial to logarithmic in sample size n.

6.1 Introduction

The interplay between the stability and computational efficiency of optimization algorithms has long been a fundamental problem in statistics and machine learning. The stability of the algorithm, a classical desideratum, is often believed to be a necessity for obtaining efficient statistical estimators. Such a belief rules out the use of a variety of faster algorithms due to their instability. This chapter shows that this popular belief can be misleading: the situation is more subtle in that there are various settings in which unstable algorithms may be preferable to their stable counterparts.

Recent years have seen a significant body of work involving performance of various machine-learning algorithms when applied to statistical estimation problems. Examples include sparse signal recovery [BT09; BBC11; GK09; HYZ08], more general forms of M-estimation [ANW12; LW15; ZZ12], principal component analysis [AW08; Ma13; YZ13], regression with concave penalties [LW15; WLZ14], phase retrieval problems retrieval (e.g., [CLS15; CSV12; CW15; Che+18; Zha+17]), and mixture model estimation [BWY17; CMZar; YBW17; YC15b].

A unifying theme in these works is to study, in a finite-sample setting, the computational efficiency of different algorithms and the statistical accuracy of the resulting estimates. For estimators based on solving optimization problems that are convex, standard algorithms and theory can be applied. However, many modern estimators arise from non-convex optimization problems, in which case the associated algorithms become more complex to understand. But evidence is accumulating for the practical and theoretical advantages of such algorithms. For instance, the paper [ANW12] established the fast convergence of projected gradient descent (GD) for high-dimensional signal recovery in a weakly convex setting, whereas the papers [LW15; WLZ14] provided similar guarantees for a class of non-convex learning problems. Other work has demonstrated fast convergence of the truncated power method for PCA [YZ13], analyzed the behavior of projected gradient methods for low-rank matrix recovery [CW15], and characterized the behavior of gradient descent for phase-retrieval problems [Che+18]. Additionally, there is also a recent line on work on the fast convergence of EM for various types of mixture models [BWY17; CMZar; YBW17]. Finally, several works [CP18; CJY18; HRS16; KL18] provide statistical error bounds in generic machine learning problems (with certain assumptions on loss functions), for estimators obtained via iterative optimization algorithms, e.g., stochastic gradient descent (SGD).

Population-to-sample or stability-based analysis

The analysis in these works can be classified into two types. The first is a *direct analysis*, in which one directly characterizes the behavior of the iterates of the algorithm on the finite-sample objective. A long line of papers has used the direct approach (e.g., [ANW12; LW15; WLZ14; YZ13; ZZ12]) to demonstrate that certain optimization algorithms converge at geometric rates to a local neighborhood of the true parameter, with the radius proportional to the statistical minimax risk. The second kind of analysis is more indirect and can be referred to as *population-to-sample analysis* or *stability-based analysis* where one analyzes the algorithmic convergence of population-level iterates, and derives statistical errors for the sample-level updates via uniform laws for stability/perturbation bounds. These approaches have been used to analyze the performance of EM and its variants in several statistical settings, see the papers [BWY17; CMZar; Dwi+20a; Dwi+20b; YBW17; YC15b] and the references therein. In general settings, it has been used to derive statistical errors for iterates from stochastic optimization methods like SGD [CP18; CJY18; HRS16; KL18].

Since our work builds on the stability-based analysis, let us discuss it in a little more detail. Let F and F_n denote the operators that define the iterates at the population level, corresponding to the idealized limit of an infinite sample size, and sample-level based on a dataset of size n. Suppose $\theta^\star$ denotes the parameter of interest such that the population-level updates converge to it, i.e., $F^t(\theta^0) \to \theta^\star$ as $t \to \infty$. Of interest is to characterize the best possible estimate of $\theta^\star$ obtained from the sample-based (noisy) iterates $\theta_n^t = F_n^t(\theta^0)$, and possibly characterize the change in the error $\|F_n^t(\theta^0) - \theta^\star\|_2$ as a function of the iteration t and the sample size n. The population-to-sample or the stability analysis proceeds by using the following decomposition:

$$F_n^t(\theta^0) - \theta^\star = \underbrace{F^t(\theta_n^0) - \theta^\star}_{=:\varepsilon_{\text{opt}}^t} + \underbrace{F_n^t(\theta_n^0) - F^t(\theta_n^0)}_{=:\varepsilon_{\text{stab}}^t}. \tag{6.1}$$

Given this decomposition, the analysis proceeds in two steps:

- The first step is a deterministic convergence analysis of the algorithm to the true parameter at the population-level, namely, obtain a control on the *optimization error* $\varepsilon_{\text{opt}}^t$ as a function of t.
- The second step is to perform a stability analysis of the difference between the population and the sample-based iterates, namely, obtain a control on the *perturbation/stability error* $\varepsilon_{\text{stab}}^t$ as a function of t.

The ultimate convergence guarantee—what statistical error can be achieved with the sample-based operator F_n, and in how many iterations—is then derived based on the interplay between the two errors in equation (6.1), namely, $\varepsilon_{\text{opt}}^t$ and $\varepsilon_{\text{stab}}^t$.

The ERM-based approach We remark that the decomposition in equation (6.1) is different from that used when invoking the uniform laws for the empirical risk minimizer (ERM). Assuming the sample-based iterates converges to the ERM, i.e., $\lim_{t\to\infty} F_n^t(\theta^0) = \widehat{\theta}_{\text{ERM}}$, the typical decomposition in the ERM-based approach is given by

$$F_n^t(\theta^0) - \theta^\star = \underbrace{F_n^t(\theta^0) - \widehat{\theta}_{\text{ERM}}}_{=:\varepsilon_{\text{opt-sample}}^t} + \underbrace{\widehat{\theta}_{\text{ERM}} - \theta^\star}_{=:\varepsilon_{\text{unif-gen}}}.$$

Here the first term in the RHS corresponds to the *optimization error at the sample-level* at iteration t and the second term corresponds to the (iteration-independent) *uniform generalization bound.* Depending on the application, a precise characterization of either of these terms can be non-trivial; moreover, applying uniform bounds to control the term $\varepsilon_{\text{unif-gen}}$ may lead to bounds that are overly loose. In such settings, the population-to-sample or stability-based analysis can prove to be a useful alternative.

Past works focus on stable methods

Most of the past work with the population-to-sample analysis has focused on algorithms whose updates are *stable*, meaning that the perturbation error between sample-level and population-level iterates decays to zero as the iterates approach the true parameter. For example, the papers [BWY17; CMZar; YBW17; YC15b] used this framework for problems where the population updates converge at a geometric rate to the true parameter, and iterates based on n samples yield an estimate within $n^{-1/2}$ of the true parameter. On the other hand, other papers [Dwi+20a; Dwi+20b] have shown that with over-specified Gaussian mixtures, the EM algorithm, which is a stable algorithm, takes a large number of steps to find an estimate whose statistical error is of order $n^{-1/4}$ or $n^{-1/8}$. Although for those problems the larger final statistical error of EM is minimax optimal, several natural questions remained unanswered: Can an algorithm converge to a statistically optimal estimate in significantly fewer steps than EM for over-specified mixtures? Moreover, will the faster algorithm continue to be stable? Besides the analysis in recent works [Dwi+20a; Dwi+20b] relied heavily on the facts that the EM updates had closed-form analytical expressions. To our best knowledge, general statistical guarantees for a generic stable or unstable algorithm (without a closed-formed expression) when the algorithmic convergence is slow, are not present in the literature.

In past work, Chen et al. [Che+18] provided a trade-off between stability and number of iterations to converge. In particular, they showed that the minimax error of a problem class forces a trade-off between the two errors in equation (6.1), $\varepsilon^t_{\text{opt}}$ and $\varepsilon^t_{\text{stab}}$, for any iterative algorithm used for solving it. In simple words, given the minimax error, an algorithm that converges quickly is necessarily unstable[1], and conversely, a stable algorithm cannot converge quickly. Their work, however, did not address the following converse questions: Under what conditions does an algorithm, either stable or unstable, achieve a statistically optimal rate? When is an unstable algorithm to be preferred to a stable counterpart?

Such questions about the trade-off between stability, computational efficiency and the statistical error upon convergence are of special interest for singular problems in which the Fisher information matrices are degenerate. Singular problems appear in a wide range of statistical settings, including mode estimation [Che64], robust regression [Rou84], stochastic utility models [Man75], informative non-response in missing data [DK94; Hec76], high-dimensional linear regression [HTW15], and over-specified mixture models [Che95; Ngu13; RM11]. Several papers have shown that maximum likelihood estimates for singular problems have much lower accuracy than the classical parametric rate $n^{-1/2}$; problems that exhibit slow rates of this type include stochastic frontier models [Lee93; LC86], certain classes of parametric

[1]There is a subtle difference in the definition of (in)stability used in Chen et al.'s work [CJY18] compared to ours. In their work, stability refers to a *slow* growth in the error $\|F^t(\theta) - F^t_n(\theta)\|_2$ with number of iterations t, where slow is defined in a relative sense with other methods. In our case, we use stability for the settings when $\|F(\theta) - F_n(\theta)\|_2$ decreases with $\|\theta - \theta^\star\|_2$ as $\theta \to \theta^\star$.

models [Rot+00], and in strongly or weakly identifiable mixture models [Che95; HN16; Ngu13]. Nevertheless, the computational aspects of parameter estimation and the trade-offs with stability in such models have not been studied in detail.

Contents of this chapter

This chapter lays out a general framework to address the questions raised above. Making use of the population-to-sample approach and a generalization of the localization argument from our previous works [Dwi+20a; Dwi+20b], we derive tight bounds on the statistical error of the final iterate produced by an algorithm. The final error and the number of steps taken depend on two things: (i) the rate of convergence of the corresponding population-level iterates, and (ii) the (in)stability of the sample-level iterates for those at the population-level. As a first contribution, our statistical guarantees for slowly converging stable algorithms and (fast/slow converging) unstable algorithms complement the findings of Balakrishnan et al. [BWY17] for fast converging stable algorithms (Theorems 13 and 14). We provide an overview of these general results in Table 6.1.

The second contribution extends the work of Chen et al. [Che+18] by showing how the final statistical errors achieved by stable and unstable algorithms can be used to directly compare and contrast the (dis)advantages between the two (Section 6.4). Our third and final contribution is an explicit demonstration of the fact that unstable methods can converge in significantly fewer steps when compared to stable methods, while still yielding statistically optimal estimates (Corollaries 8, 9 and 10). In particular, applying our framework to three estimation problems—single-index models, informative non-response models, and Gaussian mixture models—we show that while the (unstable) Newton method converges after on the order of $\log n$ steps, there is some $q > 0$ such that gradient descent—which we show to be a stable method—takes on the order of n^q steps. We also show that both methods achieve the same final minimax statistical accuracy.

Organization The remainder of our chapter is organized as follows. We begin in Section 6.2 with simulations that illustrate the phenomena to be investigated in this chapter. We then introduce some definitions and discuss different properties of the sample and population operators. Section 6.3 is devoted to statements of our general computational and statistical guarantees with detailed proofs presented in Section 6.6. In Section 6.4, we apply our general results to demonstrate the trade-off between stable and unstable methods for several examples. We conclude with a discussion of potential future work in Section 6.5. Proofs of supporting lemmas and technical results are provided in the end of this chapter.

Notation A few remarks on notation: for a pair of sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, we write $a_n \gtrsim b_n$ to mean that there is a universal constant c such that $a_n \geq c b_n$ for

Operator Properties	Optimization Rate	Stability	Iterations for convergence	Statistical error on convergence
General expressions				
Fast, stable [BWY17]	$\texttt{FAST}(\kappa)$	$\texttt{STA}(\gamma)$	$\log(1/\varepsilon(n,\delta))$	$\varepsilon(n,\delta)$
Slow, stable (Thm. 13)	$\texttt{SLOW}(\beta)$	$\texttt{STA}(\gamma)$	$\varepsilon(n,\delta^*)^{-\frac{1}{1+\beta-\gamma\beta}}$	$[\varepsilon(n,\delta^*)]^{\frac{\beta}{1+\beta-\gamma\beta}}$
Fast, unstable (Thm. 14)	$\texttt{FAST}(\kappa)$	$\texttt{UNS}(\gamma)$	$\log(1/\varepsilon(n,\delta))$	$[\varepsilon(n,\delta)]^{\frac{1}{1+\lvert\gamma\rvert}}$
Slow, unstable (Thm. 14)	$\texttt{SLOW}(\beta)$	$\texttt{UNS}(\gamma)$	$[\varepsilon(n,\delta)]^{-\frac{1}{1+\beta}}$	$[\varepsilon(n,\delta)]^{\frac{\beta}{1+\beta+\lvert\gamma\rvert\beta}}$
Examples				
Fast, stable [BWY17]	$e^{-\kappa t}$	$\frac{r}{\sqrt{n}}$	$\log n$	$n^{-1/2}$
Slow, stable (Thm. 13)	$\frac{1}{\sqrt{t}}$	$\frac{r}{\sqrt{n}}$	$n^{1/2}$	$n^{-1/4}$
Fast, unstable (Thm. 14)	$e^{-\kappa t}$	$\frac{1}{r\sqrt{n}}$	$\log n$	$n^{-1/4}$
Slow, unstable (Thm. 14)	$\frac{1}{\sqrt{t}}$	$\frac{1}{r\sqrt{n}}$	$n^{1/3}$	$n^{-1/8}$

Table 6.1. A high-level overview of our results. The notation in the problem set-up (columns 2 and 3) is formalized in Section 6.2, and the formal results (columns 4 and 5) are discussed in Section 6.3. In the top panel, we provide general expressions from our results, and in the bottom panel, we provide some explicit expressions for few specific settings. The second and third columns respectively denote the optimization and stability properties of the operator, and the last two columns provide the expressions for iterations for convergence, and the final statistical errors of the estimate returned the sample-based (noisy) operator. For the bottom panel, we use $\beta = \frac{1}{2}, \gamma = \pm 1$ with the noise function $\varepsilon(n,\delta) = \log(1/\delta)/\sqrt{n}$. We omit certain log-factors and universal constants for brevity.

all $n \geq 1$. We write $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. We use $\lceil x \rceil$ to denote the smallest integer greater than or equal to x for any $x \in \mathbb{R}$. In the chapter, we use c, c', c_i, c_i' when $i \geq 1$ to denote the universal constants. Note that the values of universal constants may change from line to line. Finally, for our operator notation, we use the subscript n to distinguish a sample-based operator (e.g., $F_n, \mathrm{G}_n^{\mathrm{NM}}, \mathrm{M}_n^{\mathrm{GA}}$) from its corresponding population-based analog (respectively $F, \mathrm{G}^{\mathrm{NM}}, \mathrm{M}^{\mathrm{GA}}$).

6.2 Motivation and problem set-up

We begin in Section 6.2 by motivating the analysis to follow by showing and discussing the results of some computational studies for the class of single-index models. These

results demonstrate a wide range of possible convergence rates, and associated stability (or instability) of the operator to perturbations. With this intuition in hand, we then turn to Section 6.2, in which we set up the definitions that underlie our analysis. In particular, we state the (i) local Lipschitz condition, and (ii) local convergence behavior for the population-level operator F, and (iii) the stability and instability condition of the sample-level operator F_n with respect to F.

A vignette on single-index models

We first consider a certain class of statistical estimation problems in which there are interesting differences between algorithms. Here we keep the discussion very brief; see Section 6.4 for a more detailed discussion. A single-index model is based on a function $f : \mathbb{R}^d \to \mathbb{R}$ that can be written in the form $f(x) = g(\langle x, \theta \rangle)$ for some parameter vector $\theta \in \mathbb{R}^d$, and some univariate function $g : \mathbb{R} \to \mathbb{R}$. In the simplest setting, the univariate function g is known, and we have a parametric family of functions as θ ranges over $\mathbb{R}^d$; when g is unknown, we have a semi-parametric family. Now suppose that we are given a collection of pairs $\{(X_i, Y_i)\}_{i=1}^n$, generated from a noisy regression model of the form

$$Y_i = g(\langle X_i, \theta^\star \rangle) + \xi_i, \qquad \text{for } i = 1, \ldots, n. \tag{6.2}$$

Here ξ_i is a zero-mean noise variable with variance σ^2, which we assume to be independent of X_i. The single index regression model (6.2) has been studied extensively in the literature (e.g., [Car+97; Ich93]).

When g is known, a natural procedure for estimating θ is based on minimizing the least squares objective function

$$\mathcal{L}_n(\theta) := \frac{1}{n} \sum_{i=1}^n \{Y_i - g(\langle X_i, \theta \rangle)\}^2. \tag{6.3}$$

When the variables ξ_i are Gaussian, then this objective coincides (up to scaling and constant factors) with the negative log-likelihood function, so that minimizing it yields the maximum likelihood estimate.

Under suitable regularity conditions on g in a neighborhood of $\theta^\star$, it is known that it is possible to estimate $\theta^\star$ at the usual parametric rate of $n^{-1/2}$. However, problems can arise when the signal-to-noise ratio (SNR), as measured by the ratio $\|\theta^\star\|_2/\sigma$, tends to zero. In particular, consider a function g whose derivative vanishes at zero—that is, $g'(0) = 0$. For instance, the function $g(t) = t^2$, which arises in the application of the single-index framework to the problem of phase retrieval, has this property. Taking the limit of low SNR amounts to trying to estimate the vector $\theta^\star = 0$ based on observations from the model (6.2). For this type of singular statistical model, we see many interesting differences between algorithms that might be used to minimize the least-squares criterion (6.3).

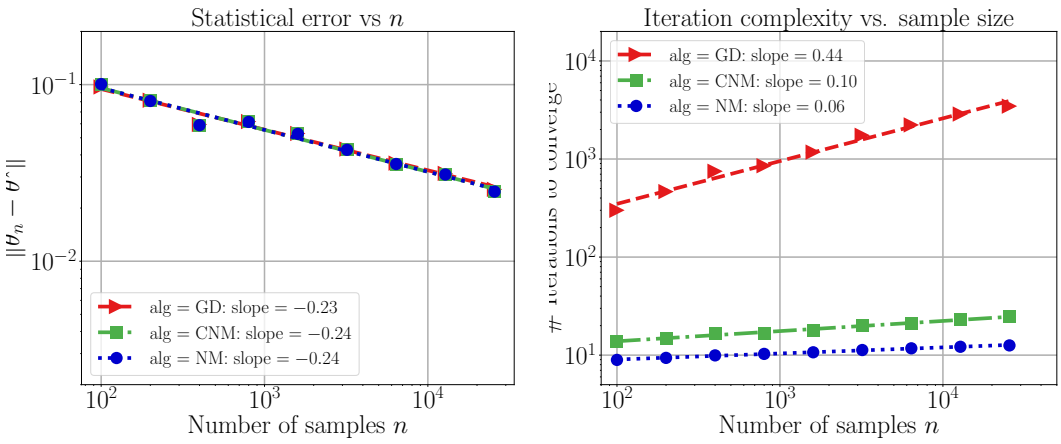


Figure 6.1. Plots characterizing the behavior of different algorithms, namely gradient descent (GD), cubic-regularized Newton's method (CNM), and the vanilla Newton's method (NM) for the single-index model when $\theta^\star = 0$. (a) Log-log plots of the Euclidean distance $\|\widehat{\theta}_n - \theta^\star\|_2$ versus the sample size. It shows that all the algorithms converge to an estimate at Euclidean distance of the order $n^{-1/4}$ from the true parameter $\theta^\star$. (b) Log-log plots for the number of iterations taken by different algorithms to converge to the final estimate. Newton's method takes the least number of steps. On the other hand, gradient descent takes significantly larger number of steps, with an empirical scaling close to $\sqrt{n}$.

More concretely, let us consider three standard optimization algorithms that might be applied to the objective (6.3): (i) gradient descent; (ii) Newton's method, and; (iii) cubic-regularized Newton's method. See Section 6.9 for a precise description of these algorithms and the associated updates in application to this model.

Statistical and iteration complexity of optimization algorithms For each procedure, we are interested both in the associated statistical error—that is, the Euclidean distance between their output and the true parameter $\theta^\star$—and their iteration complexity, meaning the number of iterations required to converge. In order to gain some understanding, we performed some simulations for single-index regression based on the function $g(t) = t^2$ in dimension $d = 1$, over a range of sample sizes n. Figure 6.1 provides some plots that summarize some results from these simulations. Panel (a) plots the Euclidean error associated with the estimate versus the sample size n on a log-log plot, along with associated least-squares fits to these data. As can be seen, all three methods lie upon a line with slope $-1/4$ on the log-log scale, showing that the statistical error decays at the rate $n^{-1/4}$. This "slow rate"—to be contrasted with the usual $n^{-1/2}$ parametric rate—is a consequence of the singularity in the model. Panel (b) plots the iteration complexity of the three algorithms versus the sample sizes, again on a log-log plot. For a given problem based on n samples, the iteration complexity is the number of iterations required for the

distance between the iterate and $\theta^\star$ to drop below $n^{-1/4}$. Here we see some interesting differences, with the gradient method having an empirical iteration complexity that grows as $\approx n^{0.44}$, based on our fits, with the two forms of Newton's method having much milder growth in iteration complexity. In the theory to follow, we will prove that iteration complexity for the gradient method scales at most like $\sqrt{n}$, that of the cubic-regularized Newton method scales as $n^{1/6}$, whereas that of Newton's method scales only as $\log n$. (See Corollary 10 for a precise statement.)

Behavior of optimization operators The plots in Figure 6.1 all concern the behavior of algorithms in practice, as applied to the empirical objective function, and our ultimate goal is to provide a theoretical explanation of phenomena of these types. In order to do so, our analysis makes use of the population-level algorithms obtained in the limit of infinite sample size; i.e., $n \to \infty$. In the special case of the single-index model considered here, we refer the Section 6.9 for the precise forms of these operators (cf. equations (6.98a)–(6.98c)). The plots in Figure 6.2 illustrate the two properties of the operators that underlie our theoretical analysis: convergence rate of the population operators (panel (a)), and the stability of the empirical operators relative to the population version (panel (b)).

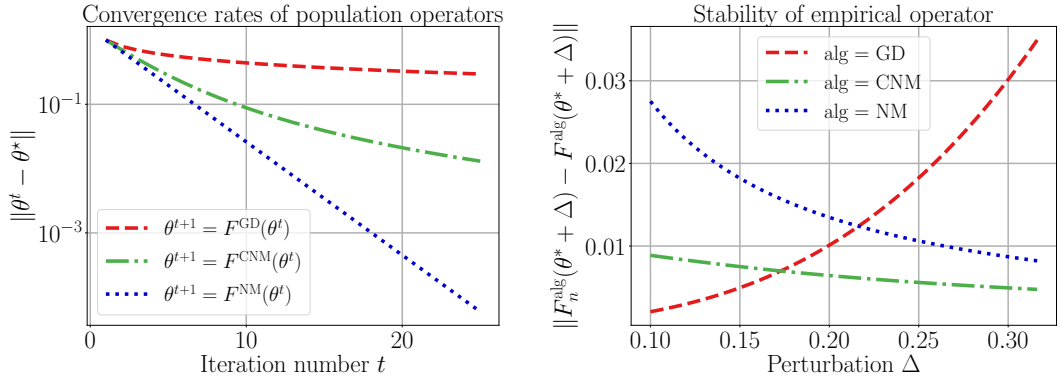


Figure 6.2. Exploration of the population level updates, and their connection to the empirical updates for the single-index problem. (a) Plots showing the convergence rate of the error $\|\theta^t - \theta^\star\|_2$ for different algorithms—namely gradient descent (GD), standard Newton's method (NM), and cubic-regularized Newton's method (CNM)—applied at the population level (limit of infinite sample size). Notice the log-scale on the y-axis. The sequence from the Newton's method converges a geometric rate to $\theta^\star$, whereas the gradient method converges at a sub-linear rate. (b) Plots showing the scaling of the perturbation error $\|F_n(\theta^\star + \Delta) - F(\theta^\star + \Delta)\|_2$ versus the perturbation Δ. For an unstable operator, the perturbation error can increase as $\|\Delta\|_2 \to 0$, with Newton's method showing a strong version of such instability. In contrast, the gradient descent method is a stable procedure in this setting.

The plots in panel (a) reveal that the three algorithms differ dramatically in their convergence rate at the population level. The ordinary Newton updates converge at a

geometric rate, with the distance to the optimum $\theta^\star$ decreasing as κ^t with the number of iterations t, where $\kappa \in (0,1)$ is a contraction coefficient. In contrast, the other two algorithms exhibit an inverse polynomial rate of convergence, with the distance to optimality decreasing at the rate $1/t^\beta$ for some exponent $\beta > 0$. In the analysis to follow, we prove that gradient descent has inverse polynomial decay with exponent $\beta = 1/2$, whereas the cubic-regularized Newton updates exhibit inverse polynomial decay with exponent $\beta = 2$.

In Corollary 10 and its proof, we characterize the optimization rate (algorithmic rate of convergence), the stability and the final statistical error obtained by these three methods. For reader's convenience, we summarize these results in Table 6.2.

Algorithm	**Optimization Rate**	**Stability**	**Iterations for convergence**	**Statistical error on convergence**
Gradient descent	$\frac{1}{\sqrt{t}}$	$\frac{r}{\sqrt{n}}$	$n^{1/2}$	$n^{-1/4}$
Newton's method	$e^{-\kappa t}$	$\frac{1}{r\sqrt{n}}$	$\log n$	$n^{-1/4}$
Cubic-regularized Newton's method	$\frac{1}{t^2}$	$\frac{1}{\sqrt{r}\sqrt{n}}$	$n^{1/6}$	$n^{-1/4}$

Table 6.2. Overview of results illustrated in Figures 6.1 and 6.2 for single-index model with the link function $g(t) = t^2$ and $\theta^\star = 0$. By characterizing the optimization rate and stability precisely, and invoking our general theory (summarized in Table 6.1), we establish that while the three methods differ significantly in terms of their optimization rate and stability, they achieve the same statistical error upon convergence, albeit by taking different number of iterations to converge. We omit logarithmic factors and universal constants for brevity. See Corollary 10 and its proof for precise details.

Problem set-up

Having provided a high-level overview of the phenomena that motivate our analysis, let us now set up the problem more abstractly, and introduce some key definitions. Consider an operator F that maps a space Θ to itself; typical examples of the space Θ that we consider are subsets of the Euclidean space $\mathbb{R}^d$, and subsets of symmetric matrices. Let $\theta^\star$ be a fixed point of the operator—i.e., an element $\theta^\star \in \Theta$ such that $F(\theta^\star) = \theta^\star$. The challenge is that we do not have access to the operator F directly, but rather can observe only a random operator F_n that can be understood as a noisy estimate of F. Throughout, we call F the *population operator* and F_n the *empirical operator*. Using the empirical operator, we generate a sequence of iterates via the fixed-point updates

$$\theta_n^{t+1} = F_n(\theta_n^t) \qquad \text{for } t = 1, 2, \ldots, \tag{6.4}$$

with a suitable initialization $\theta_n^0 \in \Theta$. Our goal is to determine conditions under which the sequence $\{\theta_n^t\}_{t \geq 0}$ approaches a suitably defined neighborhood of $\theta^\star$. More precisely, for any given triple (F, F_n, t) we provide a sharp characterization of the optimality gap $\|\theta_n^t - \theta^\star\|_2$ as a function of the iteration count t and the error $\|F - F_n\|_2$ of the empirical operator F_n.

One interesting class of problems where the operators F and F_n arise naturally is estimation problems in statistics and machine learning. More concretely, consider the problem of finding the unique minimizer $\theta^\star$ of an objective function $\mathcal{L} : \Theta \to \mathbb{R}$. In practice, we do not know the true objective function $\mathcal{L}$, instead we have access to an approximate (random) objective function $\mathcal{L}_n$, which is an unbiased estimate of the true objective function $\mathcal{L}$. Given the pair $(\mathcal{L}, \mathcal{L}_n)$, we can obtain different operators F by applying various optimization algorithms to minimize $\mathcal{L}$, including gradient methods, proximal methods, the EM algorithm and related majorization-minimization algorithms, as well as Newton and other higher-order methods. The noisy operators F_n are obtained by applying the same optimization algorithms to the approximate objective function $\mathcal{L}_n$.

Properties of the operator F

We begin by formalizing some properties of the operator F. We assume that the operator F has a unique fixed point $\theta^\star$ and we study its behavior in the local neighborhood of the Euclidean ball

$$\mathbb{B}(\theta^\star, \rho) := \Big\{ \theta \in \Theta \mid \|\theta - \theta^\star\|_2 \leq \rho \Big\} \tag{6.5}$$

centered at $\theta^\star$. Our first condition is a standard Lipschitz condition on the operator F. In particular, we say that the operator F is 1-*Lipschitz* in $\|\cdot\|$ norm over the ball $\mathbb{B}(\theta^\star, \rho)$ if

$$\|F(\theta_1) - F(\theta_2)\|_2 \leq \|\theta_1 - \theta_2\|_2 \quad \text{for all } \theta_1, \theta_2 \in \mathbb{B}(\theta^\star, \rho). \tag{6.6}$$

In words, the 1-Lipschitz condition guarantees that the operator F is non-expansive with respect to perturbations of its argument.

Our next two definitions distinguish between fast and slow rates of convergence. The first definition captures an especially favorable property of operator F; namely, it is locally contractive around the fixed point $\theta^\star$. The second definition considers a substantially slower (sub-linear) rate of convergence of the operator F.

Fast convergence For a contraction coefficient $\kappa \in (0, 1)$, the operator F is `FAST(`κ`)`-convergent on the ball $\mathbb{B}(\theta^\star, \rho)$ if

$$\|F^t(\theta_0) - \theta^\star\|_2 \leq \kappa^t \, \|\theta_0 - \theta^\star\|_2 \quad \text{for all iterations } t = 1, 2, \ldots, \tag{6.7}$$

and for all $\theta_0 \in \mathbb{B}(\theta^\star, \rho)$.

Slow convergence Given an exponent $\beta > 0$, the operator F is $\texttt{SLOW}(\beta)$-convergent over the ball $\mathbb{B}(\theta^\star, \rho)$ means that

$$\|F^t(\theta_0) - \theta^\star\|_2 \leq \frac{c}{t^\beta} \quad \text{for all iterations } t = 1, 2, \ldots, \tag{6.8}$$

and for all $\theta_0 \in \mathbb{B}(\theta^\star, \rho)$, where c is a universal constant.

Let us illustrate these definitions with a very simple example.

Example 1. Fast versus slow convergence Consider the function $\mathcal{L}(\theta) = \frac{\theta^{2p}}{2p}$ for some positive integer $p \geq 1$. Note that for any $p \geq 1$, the function $\mathcal{L}(\cdot)$ has a unique global minimum at $\theta^\star = 0$. The first two derivatives of $\mathcal{L}(\cdot)$ are given by

$$\mathcal{L}'(\theta) = \theta^{2p-1}, \quad \text{and} \quad \mathcal{L}''(\theta) = (2p-1)\theta^{2p-2}.$$

Consequently, a gradient descent update with a constant stepsize $\alpha > 0$ takes the form

$$F^{\mathrm{GRD}}(\theta) = \theta - \alpha\mathcal{L}'(\theta) \;=\; \theta\Big(1 - \alpha\theta^{2p-2}\Big). \tag{6.9}$$

Thus, when $p = 1$, for any $\alpha \in (0, 1)$, this gradient descent update is a $\texttt{FAST}(\kappa)$-convergent algorithm with $\kappa = 1 - \alpha$. On the other hand, for any $p \geq 2$, it can be shown that gradient descent is $\texttt{SLOW}(\beta)$-convergent with parameter $\beta = \frac{1}{2p-2}$ in the ball $\mathbb{B}(\theta^\star, \rho)$ with $\theta^* = 0$ and $\rho = h^{-\frac{1}{2p-2}}$.

Now, let us consider Newton's method with step size one, namely the update

$$F^{\mathrm{NWT}}(\theta) = \theta - \Big(\mathcal{L}''(\theta)\Big)^{-1}\mathcal{L}(\theta) \;=\; \theta - \frac{\theta^{2p-1}}{(2p-1)\theta^{2p-2}} \;=\; \theta\left(1 - \frac{1}{2p-1}\right). \tag{6.10}$$

For $p = 1$, this update converges in a single step (simply because the quadratic approximation that underlies Newton's method is exact in this special case). For $p \geq 2$, the pure Newton update is $\texttt{FAST}(\kappa)$-convergent with $\kappa = 1 - \frac{1}{2p-1}$ for all $\theta \in \mathbb{R}$.

From the empirical operator F_n to the population operator F

In this section, we introduce some key concepts that characterize the (in)-stability of the sample operator F_n with respect to the population operator F. Given a pair of operators (F_n, F) and a tolerance parameter $\delta \in (0, 1)$, our definitions involve a *perturbation function* $\varepsilon(\cdot)$ that maps the triple (F_n, F, δ) to a positive scalar $\varepsilon(F_n, F, \delta)$. For notational convenience, we use the shorthand $\varepsilon(n, \delta)$ instead of $\varepsilon(F_n, F, \delta)$. In general, we impose the following conditions on the perturbation function $\varepsilon(\cdot)$:

- It is decreasing in n for any fixed δ, and is monotonically increasing in the second argument δ for any fixed n.

- For any fixed $\delta \in (0,1)$, we have $\varepsilon(n,\delta) \to 0$ as $n \to \infty$, and similarly, for any fixed $n > 0$, we have $\varepsilon(n,\delta) \to \infty$ as $\delta \to 0$.

Given some choices of perturbation function, we can define our first stability condition as follows:

Stability (`STA(`γ`)` condition) For a given parameter $\gamma \geq 0$, the operator F_n is `STA(`γ`)`-stable over $\mathbb{B}(\theta^\star, \rho)$ with noise function $\varepsilon(\cdot)$ means that, for any radius $r \in (0, \rho)$ and tolerance $\delta \in (0,1)$, we have

$$\mathbb{P}\Big[\sup_{\theta \in \mathbb{B}(\theta^\star, r)} \|F_n(\theta) - F(\theta)\|_2 \leq c_2 \min\Big\{r^\gamma \varepsilon(n,\delta), r\Big\}\Big] \geq 1 - \delta, \tag{6.11}$$

for some positive universal constant c_2. Informally, the stability condition (6.11) guarantees that with high probability, the error $\|F_n(\theta) - F(\theta)\|_2$ is upper bounded by $c_2 \min\{r^\gamma \varepsilon(n,\delta), r\}$ uniformly over a disk of radius r. Note moreover that the upper bound decays to 0 as the radius $r \to 0^+$.

Next we consider the case when $\gamma < 0$, i.e., the perturbation error $\|F_n(\theta) - F(\theta)\|_2$ blows up as θ gets close to $\theta^\star$. Given radii r_1, r_2 such that $r_2 > r_1 \geq 0$, let $\mathbb{A}(\theta^\star, r_1, r_2) = \mathbb{B}(\theta^\star, r_2) \backslash \mathbb{B}(\theta^\star, r_1)$ denote the annulus around $\theta^\star$ with inner and outer radii r_1 and r_2 respectively.

Instability (`UNS(`γ`)` condition) For a given parameter $\gamma < 0$ and radii $0 < \rho_{\text{in}} < \rho_{\text{out}}$, we say that the operator F_n is `UNS(`γ`)`-unstable over the annulus $\mathbb{A}(\theta^\star, \rho_{\text{in}}, \rho_{\text{out}})$ with noise function $\varepsilon(\cdot)$ if

$$\mathbb{P}\left[\sup_{\theta \in \mathbb{A}(\theta^\star, r, \rho_{\text{out}})} \|F_n(\theta) - F(\theta)\|_2 \leq \varepsilon(n,\delta) \max\left\{\frac{1}{r^{|\gamma|}}, \rho_{\text{out}}\right\}\right] \geq 1 - \delta, \tag{6.12}$$

for any radius $r \in [\rho_{\text{in}}, \rho_{\text{out}}]$ and any tolerance $\delta \in (0,1)$. Note that the condition (6.12) defines the instability of the perturbation error $\|F_n(\theta) - F(\theta)\|_2$ in an annulus with the inner radius bounded below by ρ_{in}, and does not characterize the behavior as the distance $\|\theta - \theta^\star\|_2 \to 0$.

We illustrate these definitions by following up on Example 1.

Example 2. Stable versus unstable updates Consider an empirical function of the form

$$\mathcal{L}_n(\theta) = \frac{1}{2p}\theta^{2p} + \frac{\sigma w}{2\sqrt{n}}\theta^2, \qquad \text{where } w \sim N(0,1). \tag{6.13}$$

Here $p \geq 2$ is a positive integer. Note that $\mathbb{E}[\mathcal{L}_n(\theta)] = \frac{1}{2p}\theta^{2p}$, which is equivalent to the population likelihood function considered in Example 1.

A gradient update with stepsize $\alpha > 0$ on the empirical objective leads to the empirical gradient operator

$$F_n^{\mathrm{GRD}}(\theta) = \theta \left\{ 1 - \alpha\theta^{2p-2} - \alpha\frac{\sigma w}{\sqrt{n}} \right\}.$$

Comparing with equation (6.9), we obtain that $\left|F_n^{\mathrm{GRD}}(\theta) - F^{\mathrm{GRD}}(\theta)\right| = \frac{\sigma}{\sqrt{n}}|w|\,|\theta|$. Since $|w| \leq 4\sqrt{\log(1/\delta)}$ with probability at least $1-\delta$, we see for any $\rho > 0$ and $n \geq 16\sigma^2 \log(1/\delta)$, the operator F_n^{GRD} is $\texttt{STA}(\gamma)$-stable with parameter $\gamma = 1$, with respect to the noise function

$$\varepsilon(n,\delta) = 4\sigma\sqrt{\frac{\log(1/\delta)}{n}}.$$

As for the Newton update for the problem (6.13), we have

$$F_n^{\mathrm{NWT}}(\theta) = \theta - \frac{\theta^{2p-1} + \sigma w\theta/\sqrt{n}}{(2p-1)\theta^{2p-2} + \sigma w/\sqrt{n}},$$

and hence

$$\left|F^{\mathrm{NWT}}(\theta) - F_n^{\mathrm{NWT}}(\theta)\right| = \frac{(2p-2)}{(2p-1)} \cdot \frac{\sigma\,|w|\,|\theta|\,/\sqrt{n}}{(2p-1)\theta^{2p-2} + \sigma w/\sqrt{n}}.$$

Recall that $|w| \leq 4\sqrt{\log(1/\delta)}$ with probability at least $1-\delta$. Plugging in $w > -4\sqrt{\log(1/\delta)}$ in the denominator and $w < 4\sqrt{\log 1/\delta}$ of the RHS, and doing some algebra yields that

$$\left|F^{\mathrm{NWT}}(\theta) - F_n^{\mathrm{NWT}}(\theta)\right| \leq \frac{c_p}{|\theta|}\sqrt{\frac{\log(1/\delta)}{n}} \quad \text{for } |\theta| > \left(c_p'\sigma\sqrt{\frac{\log(1/\delta)}{n}}\right)^{\frac{1}{2p-2}},$$

where $c_p = \frac{16(p-1)}{2p-1}$ and $c_p' = \frac{8}{2p-1}$. Thus, we conclude that the operator F_n^{NWT} is $\texttt{UNS}(\gamma)$-unstable with parameter $\gamma = -1$ over the annulus $\mathbb{A}(\theta^\star, \rho_{\mathrm{in}}, \rho_{\mathrm{out}})$ with noise function ε where

$$\rho_{\mathrm{in}} = \left(c_p'\sigma\sqrt{\frac{\log(1/\delta)}{n}}\right)^{\frac{1}{2p-2}}, \quad \rho_{\mathrm{out}} = \infty, \quad \text{and} \quad \varepsilon(n,\delta) = c_p\sigma\sqrt{\frac{\log(1/\delta)}{n}}.$$

6.3 General convergence results

With the definitions from the previous section in place, we are now ready to state our main results. In Section 6.3, we consider the case when F_n is a stable perturbation of F, and in Section 6.3, we consider the case when it is an unstable perturbation of F. We summarize our findings in Table 6.1.

Results for slowly converging but stable operators

We first consider the setting in which the sample-based operator F_n is a stable perturbation of the population-level operator F. If, in addition, we assume that the operator F has fast convergence (cf. equation (6.7)), then past work is applicable. In particular, Theorem 2 of Balakrishnan et al. [BWY17] provides a precise characterization of the convergence behavior of iterates from the empirical operator F_n. Here we instead consider the more challenging setting in which the operator F exhibits slow convergence to $\theta^\star$. Analysis of this slow convergence case requires rather different techniques than those used to analyze the fast-convergent case.

Let us collect the assumptions needed to state our first result. The first two assumptions involve the Euclidean ball $\mathbb{B}(\theta^\star, \rho)$ centered at $\theta^\star$ of some fixed radius $\rho > 0$.

(A) The population operator F is 1-Lipschitz (6.6) and is $\texttt{SLOW}(\beta)$-convergent (6.8) over the ball $\mathbb{B}(\theta^\star, \rho)$.

(B) There is some $\gamma \in [0, (1+\beta)^{-1})$ such that the empirical operator F_n is $\texttt{STA}(\gamma)$-stable (6.11) over $\mathbb{B}(\theta^\star, \rho)$.

(C) The tolerance parameters $\delta \in (0,1)$ and $\epsilon \in (0, \frac{\beta}{1+\beta-\gamma\beta})$ are fixed and the sample size is large enough such that

$$\varepsilon(n, \delta^*) \leq c \quad \text{where} \quad \delta^* = \delta \cdot \frac{\log(\frac{1+\beta}{\beta\gamma})}{8\log(\frac{\beta}{\epsilon(1+\beta-\gamma\beta)})}, \tag{6.14}$$

and $c \in (0,1)$ is a sufficiently small constant.[2]

Assumptions (A) and (B) quantify, respectively, the convergence behavior of the operator F and the stability of the operator F_n; Assumption (C) is a book-keeping device needed to state our results cleanly.

Given the above conditions, we now state our first main result.

Theorem 13. *Under Assumptions (A), (B), and (C), consider the sequence $\theta_n^{t+1} = F_n(\theta_n^t)$ generated from an initialization $\theta_n^0 \in \mathbb{B}(\theta^\star, \rho/2)$. Then there is a universal constant c' such that for any fixed $\epsilon \in (0, \frac{\beta}{1+\beta-\gamma\beta})$ and uniformly for all iterations $t \geq c'\Big(1/\varepsilon(n,\delta^*)\Big)^{\frac{1}{1+\beta-\gamma\beta}} \log\frac{1}{\epsilon}$, we have*

$$\|\theta_n^t - \theta^\star\|_2 \leq 2[\varepsilon(n,\delta^*)]^{\frac{\beta}{1+\beta-\gamma\beta}-\epsilon} \qquad \text{with probability at least } 1-\delta. \tag{6.15}$$

Let us make some comments on this result and its proof. (See Section 6.6 for a detailed proof.)

[2] Refer to equation (6.70) for an explicit definition. For our examples, we typically have $\varepsilon(n,\delta) = \sqrt{\log(1/\delta)/n}$.

Tightness of Theorem 13 Disregarding the term ϵ, the bound (6.15) guarantees that the sequence $\theta_n^{t+1} = F_n(\theta_n^t)$ converges to a statistical tolerance of order $[\varepsilon(n, \delta^*)]^{\frac{\beta}{1+\beta-\gamma\beta}}$ with respect to $\theta^\star$. This guarantee turns out to be unimprovable under the given assumptions. In particular, we can construct examples of the operators F and F_n, satisfying the assumptions required to apply Theorem 13, for which there is a universal constant c_1 such that

$$\|\theta_n^t - \theta^\star\|_2 \geq c_1 \, [\varepsilon(n, \delta^*)]^{\frac{\beta}{1+\beta-\gamma\beta}} \qquad \text{for all } t = 1, 2, \ldots$$

with constant probability. As a result, we conclude that the results of Theorem 13 are tight and not improvable in general. See Section 6.7 for the details of this lower bound construction.

Outline of proof The proof of Theorem 13 involves a generalization and refinement of annulus-based localization argument introduced in our prior work on the EM algorithm [Dwi+20a; Dwi+20b]. We now summarize the proof outline. In the past work [Dwi+20a; Dwi+20b], we studied particular instantiations of the EM algorithm, for which the operators F and F_n had closed-form solutions. Here in the absence of closed-form expressions, the argument is necessarily more abstract and handles the previous analysis as a special case.

At a high level, the proof proceeds by decomposing the total collection of iterations $\{1, 2, \ldots, t\}$ into a disjoint partition of subsets $\{t_\ell\}_{\ell \geq 0}$, referred to as epochs, where the nonnegative integers ℓ and t_ℓ respectively denote the index of a given epoch and the number of iterations in that epoch. We use $T_\ell := \sum_{i=0}^{\ell} t_i$ to denote the total number of iterations up to epoch ℓ. By carefully choosing the sequence $\{t_\ell\}_{\ell \geq 0}$, we ensure that at the end of a given epoch ℓ, the error $\|\theta_n^{T_\ell} - \theta^\star\|_2$ has decreased to a prescribed threshold. More precisely, using an inductive argument, we show that

$$\|\theta_n^{T_\ell} - \theta^\star\| \leq \varepsilon(n, \delta^*)^{\alpha_\ell} \quad \text{for all epoch} \quad \ell \geq 1, \tag{6.16}$$

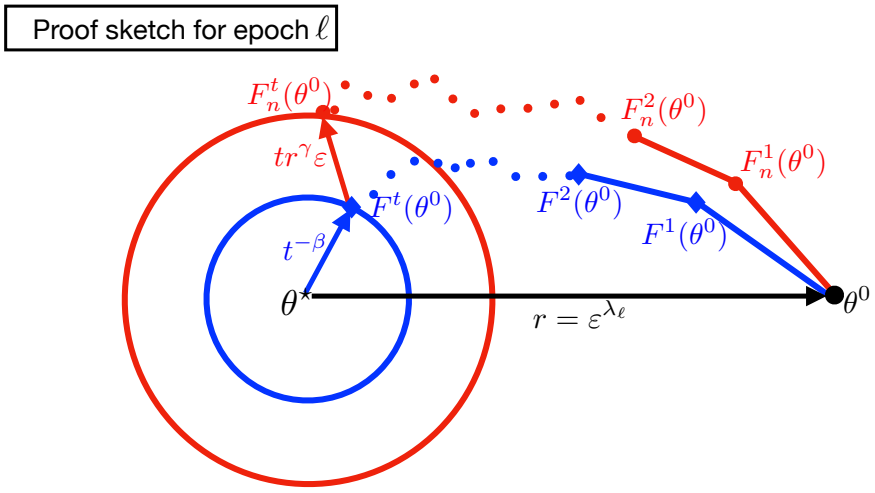
where the sequence $\{\alpha_\ell\}_{\ell \geq 0}$ is defined via the recursion

$$\alpha_0 = 0 \quad \text{and} \quad \alpha_{\ell+1} = \nu \alpha_\ell + \nu', \quad \text{for all} \quad \ell \geq 1,$$

with the scalars $\nu \in (0, 1)$ and $\nu' > 0$ determined by the problem parameters β and γ. We show that the sequence $\{\alpha_\ell\}_{\ell \geq 0}$ converges to $\nu_\star := \frac{\beta}{1+\beta-\gamma\beta}$ fast enough and we have $|\alpha_\ell - \nu_\star| \leq \epsilon$ for all $\ell \geq \mathcal{O}\log(1/\epsilon)$. Deriving a suitable upper bound on $t_{\max}$ on the epoch size t_i, we then put the pieces together to (roughly) conclude that

$$\|\theta_n^t - \theta^\star\| \leq c\varepsilon(n, \delta^*)^{\nu_\star - \epsilon} \quad \text{for} \quad t \geq c' t_{\max} \cdot \log \frac{1}{\epsilon}.$$

As expected, much of the work is required to establish the inductive step, since the base inequality $\|\theta_n^0 - \theta^\star\| \leq 1$ required to start the induction is implied by the theorem assumptions. Put simply, given that the bound (6.16) holds for epoch ℓ, several technical steps are needed to establish that it continues to hold for the next epoch $\ell + 1$. The full proof of the theorem is given in Section 6.6. We also illustrate the high-level ideas of the epoch-based localization argument in Figure 6.3.



$$\|F_n^t(\theta^0) - \theta^\star\| \le \|F_n^t(\theta^0) - F^t(\theta^0)\| + \|F^t(\theta^0) - \theta^\star\| \le tr^\gamma\varepsilon + \frac{1}{t^\beta} = t\varepsilon^{\gamma\lambda_\ell+1} + \frac{1}{t^\beta} \overset{\text{min over t}}{\asymp} \varepsilon^{\lambda_{\ell+1}}$$

$$\text{where} \quad \lambda_{\ell+1} = \nu\lambda_\ell + \nu' \Longrightarrow \lim_{\ell\to\infty} \lambda_\ell = \nu_\star = \frac{\beta}{1+\beta-\gamma\beta}, \quad \text{and} \quad \nu_\star - \lambda_\ell \le \alpha \ \text{ for all } \ \ell \ge \mathcal{O}(\log(1/\alpha))$$

Figure 6.3. An illustration of the epoch-based argument when the population operator F is $\texttt{SLOW}(\beta)$-convergent, and the noisy operator is $\texttt{STA}(\gamma)$-stable (Theorem 13). In order to simplify the visualization, we use the shorthand $\varepsilon = \varepsilon(n, \delta^*)$. Moreover, here θ^0 denotes the starting point for a given epoch ℓ (assumed to be at distance $r = \varepsilon^{\alpha_\ell}$ from $\theta^\star$), and the iterations $1, 2, \ldots, t$ denote the iteration count in that epoch. The population iterates $F^1(\theta^0), F^2(\theta^0), \ldots$ converge towards to $\theta^\star$ at the rate $t^{-\beta}$ (shown in blue), and their distance from the noisy iterates $F_n^1(\theta^0), F_n^2(\theta^0), \ldots$ grows at the rate at a distance of $tr^\gamma\varepsilon$. Trading-off the two errors, we can show that at the end of epoch ℓ (denoted by a suitable choice of t), the distance $\|F_n^t(\theta^0) - \theta^\star\|_2 \precsim \varepsilon^{\alpha_{\ell+1}}$. By establishing that α_ℓ converges to $\nu_\star$ exponentially fast, and that similar arguments can be made for sufficiently many epochs, we obtain the result in Theorem 13. See Section 6.6 for a formal argument.

Results for unstable operators

We now turn to our next main result which characterizes the convergence when the operator F_n is an unstable perturbation of the operator F. We consider two distinct cases depending on whether the operator F is (a) $\texttt{FAST}(\kappa)$-convergent or (b) $\texttt{SLOW}(\beta)$-convergent.

Theorem 14. *For a given parameter $\delta \in (0, 1)$, consider the sequence $\theta_n^{t+1} = F_n(\theta_n^t)$ for some initial point θ_n^0 in the ball $\mathbb{B}(\theta^\star, \rho/2)$. Suppose that for some $\gamma < 0$, the empirical operator F_n is $\texttt{UNS}(\gamma)$-unstable over the annulus $\mathbb{A}(\theta^\star, \widetilde{\rho}_n, \rho)$ with respect to the noise function ε.*

(a) Suppose that the operator F is $\texttt{FAST}(\kappa)$-convergent over the ball $\mathbb{B}(\theta^\star, \rho)$, and

the sample size n is sufficiently large so as to ensure that

$$[\varepsilon(n,\delta)]^{\frac{1}{1+|\gamma|}} \leq (1-\kappa)\rho. \tag{6.17a}$$

Then with probability at least $1-\delta$, for any iteration $t \geq \frac{\log(\frac{\rho}{\varepsilon(n,\delta)})}{(1+|\gamma|)\log\frac{1}{\kappa}}$, we have

$$\min_{k\in\{0,1,\ldots,t\}} \|\theta_n^k - \theta^\star\|_2 \leq \max\left\{\frac{(2-\kappa)}{(1-\kappa)} \cdot [\varepsilon(n,\delta)]^{\frac{1}{1+|\gamma|}},\ \widetilde{\rho}_n\right\}. \tag{6.17b}$$

(b) Suppose that the operator F is 1-Lipschitz and SLOW(β)-convergent for some $\beta > 0$, and that the sample size n is large enough to ensure that

$$[\varepsilon(n,\delta)]^{\frac{\beta}{1+\beta-\gamma\beta}} \leq \rho. \tag{6.18a}$$

Then with probability at least $1-\delta$, for any iteration $t \geq \frac{1}{[\varepsilon(n,\delta)]^{\frac{1}{1+\beta}}}$, we have

$$\min_{k\in\{0,1,\ldots,t\}} \|\theta_n^k - \theta^\star\|_2 \leq \max\left\{[\varepsilon(n,\delta)]^{\frac{\beta}{1+\beta-\gamma\beta}}, \widetilde{\rho}_n\right\}. \tag{6.18b}$$

Let us make a few comments about these bounds. (See Section 6.6 for a detailed proof.)

Choice of the inner radius $\widetilde{\rho}_n$ In order to obtain sharp upper bounds—ones that depend purely on the noise function ε—the inner radius $\widetilde{\rho}_n$ must be chosen suitably. Focusing on part (a), if we ensure that $\widetilde{\rho}_n \leq [\varepsilon(n,\delta)]^{\frac{1}{1+\gamma}}$, then we obtain an upper bound on the error that involves only the noise function. We show how to make such choices in our applications of this general theorem. A similar statement applies to part (b) of the theorem.

Tightness of Theorem 14 In Section 6.7, we construct examples of the operators F and F_n which satisfy the assumptions of Theorem 14, and with the inner radius satisfying the bound $\widetilde{\rho}_n \leq [\varepsilon(n,\delta)]^\tau$, $\tau = \frac{1}{1+\gamma}$ for part (a) or $\tau = \frac{\beta}{1+\beta-\gamma\beta}$ for part (b). For each of these examples, we show that the sequence $\theta_n^{t+1} = F_n(\theta_n^t)$ satisfies the lower bound

$$\|\theta_n^t - \theta^\star\|_2 \geq [\varepsilon(n,\delta)]^\tau \quad \text{for all} \quad t \geq 0,$$

with constant probability. Thus, we conclude that the results of Theorem 14 are tight and not improvable in general.

Necessity of the minimum Note that both of the bounds (6.17b) and (6.18b) apply to the minimum over all iterates $k \in \{1, 2, \ldots, t\}$, as opposed to the final iterate t. For this reason, our results only guarantee that the iterates produced by an unstable operator F_n converge at least once to a vicinity of the parameter $\theta^\star$, but *not* that they necessarily stay there for all the future iterations. In fact, such "escape" behavior for an unstable algorithm is unavoidable in the absence of any additional regularity assumptions. In particular, we provide a simple example in Section 6.7 that illustrates this unavoidability.

Additional regularity condition If we impose an additional regularity condition, then we can remove the minimum from the guarantee. In particular, consider the condition:

(D) There exists a universal constant C such that for a given initialization θ_n^0, the sequence $\theta_n^t = F_n^t(\theta_n^0)$ has the following property:

$$\|\theta_n^{t+1} - \theta^\star\|_2 \leq C\widetilde{\rho} \quad \text{whenever} \quad \|\theta_n^t - \theta^\star\|_2 \leq \widetilde{\rho}, \tag{6.19}$$

where the radius $\widetilde{\rho}$ corresponds to equation (6.17b) or (6.18b) depending on the nature of the operator F.

Under this condition, it is straightforward to modify the proof of Theorem 14 to show that the bounds in both parts (a) and (b) can be sharpened by replacing the term $\min_{k \in \{0,1,\ldots,t\}} \|\theta_n^k - \theta^\star\|_2$ with $\|\theta_n^t - \theta^\star\|_2$. In Section 6.4 to follow, we provide a number of examples for which Assumption (D) is satisfied.

6.4 Some concrete results for specific models

In this section, we study three interesting classes of statistical problems that fall within the framework of the chapter. We also discuss various consequences of Theorems 13 and Theorem 14 when applied to these problems.

Informative non-response model

In our first example, let us consider the problem of biased or informative non-response in sample surveys. In certain settings, the chance of a response to not be observed depends on the value of the response. This form of non-response introduces systematic biases in the survey and associated conclusions [Hec76]. Some examples where this issue arises include longitudinal data [DK94], housing surveys and election polls [Sha+91]. In such settings, it is common practice to estimate the non-responsive behavior in order to correct for the bias. We now describe one simple formulation of such a setting.

Suppose that we have n i.i.d. values $Y_1, \ldots, Y_n$ for the response variable $Y \sim \mathcal{N}(\mu, \sigma^2)$, where for each Y_i there is a chance that the value is not observed. To account

for such a possibility, we define $\{0,1\}$-valued random variables R_i for $i = 1, \ldots, n$ as follows:

$$R_i = 1 \quad \text{if } Y_i \text{ is observed,} \qquad \text{and} \qquad R_i = 0 \quad \text{otherwise.} \tag{6.20a}$$

We assume that the conditional distribution $R_i|Y_i$ takes the form

$$\mathbb{P}_\theta(R_i = 1|Y_i = y) = \exp\Big(H\Big(\theta(y-\mu)/\sigma\Big)\Big), \tag{6.20b}$$

where H is a known function and θ is an unknown parameter which controls the dependence of the probability of non-response on the observation $Y = y$. In a general setting, all the parameters μ, σ and θ are unknown and are estimated jointly from the data. However, to simplify our presentation, we assume that the parameters (μ, σ) are known and only θ needs to estimated. In particular, we consider the case when the response variable $Y \sim \mathcal{N}(\mu, \sigma^2) \equiv \mathcal{N}(0,1)$ and $H(x) = -x^2 - \log 2$. Under these assumptions, simple algebra yields that

$$\mathbb{P}_\theta(R_i = 1|Y_i = y) = \exp\left(-\frac{\theta^2 y^2}{2} - \log 2\right) \quad \text{and} \quad \mathbb{P}_\theta(R_i = 1) = \frac{1}{2\sqrt{\theta^2+1}}. \tag{6.20c}$$

Given n i.i.d. samples $\{R_i, Y_i\}_{i=1}^n$, where we note that Y_i is not observed when $R_i = 0$, the log-likelihood is given by

$$\bar{\mathcal{L}}_n(\theta) := \frac{1}{n}\sum_{i=1}^n -\frac{R_i\,(Y_i^2(\theta^2+1) + 2\log 2)}{2} + (1-R_i)\log\left(1 - \frac{1}{2\sqrt{\theta^2+1}}\right). \tag{6.21}$$

Note that the likelihood above does not depend on the unobserved Y_i since $R_i = 0$ makes the contribution of the corresponding term 0.

In the remainder of this section, we focus on the singular regime, i.e., when the true parameter $\theta^\star = 0$ and consequently the probability of observing any sample $Y_i = y$ is always $1/2$ (independent of the value y). For such a setting, the results of Rotnitzky et al. [Rot+00] imply that the statistical error of the MLE is larger than the parametric rate $n^{-\frac{1}{2}}$. In particular, they showed that $|\widehat{\theta}_{n,\mathrm{MLE}} - \theta^\star| = \mathcal{O}(n^{-\frac{1}{4}})$. However, with high probability, the log-likelihood $\bar{\mathcal{L}}_n$ is non-concave[3] and thereby a closed-form for the maximum-likelihood estimate is not available. Thus a theoretical analysis of the estimates obtained via different optimization algorithms (that can be used to maximize the log-likelihood $\bar{\mathcal{L}}_n$) can be of significant interest. We now apply our general theory to analyze two optimization methods: (i) gradient ascent, and (ii) Newton's method.

[3]For instance, when $\sum_{i=1}^n R_i(Y_i^2+1) < n$, the sample log-likelihood function is bimodal and symmetric around 0.

Theoretical guarantees

We now state a theoretical guarantee on the behavior of the optimization algorithms in practice with the informative non-response model (6.20)—that is, when applied to the sample log likelihood (6.21). We analyze the gradient ascent updates for a step-size $\eta \in (0, \frac{8}{3})$, and the pure Newton updates. We use $\mathrm{M}_n^{\mathrm{GA}}$ and $\mathrm{M}_n^{\mathrm{NM}}$ respectively to denote the sample-based operators for gradient ascent and Newton's method (see Section 6.9 for the precise form of these operators). The following statement also involves other universal constants c, c_i, c_i', c_i'' etc.

Corollary 8. *For the singular setting of informative non-response model ($\theta^\star = 0$) and given some $\delta \in (0,1)$, the following properties hold with probability at least $1-\delta$:*

(a) For any fixed $\epsilon \in (0, 1/4)$ and initialization $\theta^0 \in \mathbb{B}(\theta^\star, 1/2)$, the sequence $\theta^t := (\mathrm{M}_n^{\mathrm{GA}})^t(\theta^0)$ of gradient iterates satisfies the bound

$$\left|\theta^t - \theta^\star\right| \leq c_1 \left(\frac{\log(\frac{\log(1/\epsilon)}{\delta})}{n}\right)^{\frac{1}{4}-\epsilon} \qquad \text{for all iterates } t \geq c_1' \sqrt{n} \log \tfrac{1}{\epsilon}, \quad (6.22a)$$

as long as $n \geq c_1'' \log \frac{\log(1/\epsilon)}{\delta}$.

(b) For any initialization $\theta^0 \in \mathbb{A}(\theta^\star, \sqrt{2c}\,(\log(1/\delta)/n)^{1/4}, 1/2)$, the sequence of Newton iterates $\theta^t := (\mathrm{M}_n^{\mathrm{NM}})^t(\theta^0)$ satisfies the bound

$$\left|\theta^t - \theta^\star\right| \leq c_2 \left(\frac{\log(1/\delta)}{n}\right)^{\frac{1}{4}} \qquad \text{for all iterates } t \geq c_2' \log n, \quad (6.22b)$$

as long as $n \geq c_2'' \log(1/\delta)$.

See Section 6.9 for the proof of this corollary (and below for the proof sketch).

Corollary 8 shows that given n samples, (i) the final statistical errors achieved by the iterates generated by the gradient descent and the Newton's method are similar (of order $n^{-\frac{1}{4}}$), and (ii) the Newton's method takes a considerably smaller number (of order $\log n$) of steps in comparison to that taken by gradient ascent (of order $\sqrt{n}$). Finally, in Section 6.9, we show that all the non-zero fixed points of the considered operators have a magnitude of the order $n^{-\frac{1}{4}}$ with constant probability. Therefore, the statistical radius achieved by the given optimization methods are optimal.

Proof sketch for Corollary 8

Our proof of Corollary 8 starts with an analysis of the gradient ascent and Newton iterates on the population-level analog of the problem. In particular, taking

expectations in equation (6.21), we obtain the following population-level optimization problem

$$\max_{\theta \in \mathbb{R}} \bar{\mathcal{L}}(\theta) \quad \text{where} \quad \bar{\mathcal{L}}(\theta) = \frac{1}{2} \log \left(1 - \frac{1}{2\sqrt{\theta^2+1}}\right) - \frac{\theta^2+1}{4}. \tag{6.23}$$

Let M^{GA} denote the gradient update operator applied to this objective with a given step-size η, and let M^{NM} denote the Newton update. In Section 6.9 (where we also provide explicit forms of these operators), we show that with $\theta^\star = 0$, the population-level operators have the following properties:

(P1) The gradient operator M^{GA} is $\texttt{SLOW}(\beta)$-convergent with parameter $\beta = \frac{1}{2}$ over the Euclidean ball $\mathbb{B}(\theta^\star, \frac{1}{2})$, i.e., for the sequence $\theta^t = (\mathrm{M}^{\mathrm{GA}})^t(\theta^0)$ with $\theta^0 \in \mathbb{B}(\theta^\star, \frac{1}{2})$, we have $|\theta^t - \theta^\star| \leq \frac{c}{t^{1/2}}$.

(P2) The Newton operator M^{NM} is $\texttt{FAST}(\kappa)$-convergent with parameter $\kappa = \frac{4}{5}$ over the Euclidean ball $\mathbb{B}(\theta^\star, \frac{1}{2})$, i.e., for the sequence $\theta^t = (\mathrm{M}^{\mathrm{NM}})^t(\theta^0)$ with $\theta^0 \in \mathbb{B}(\theta^\star, \frac{1}{2})$, we have $|\theta^t - \theta^\star| \leq c\, e^{-\kappa t}$.

Moreover in the same Section 6.9, we show that with the noise function $\varepsilon(n, \delta) = \sqrt{\frac{\log(1/\delta)}{n}}$, the sample-level operators satisfy the following properties:

(S1) The sample-based gradient ascent operator $\mathrm{M}_n^{\mathrm{GA}}$ is $\texttt{STA}(\gamma)$-stable with parameter $\gamma = 1$ over the ball $\mathbb{B}(\theta^\star, \frac{1}{2})$, and

(S2) the operator $\mathrm{M}_n^{\mathrm{NM}}$ is $\texttt{UNS}(\gamma)$-unstable with parameter $\gamma = -1$ over the annulus $\mathbb{A}(\theta^\star, \widetilde{\rho}_n, \rho)$ with $\widetilde{\rho}_n = c[\varepsilon(n, \delta)]^{\frac{1}{2}}$ and $\rho = \frac{1}{2}$ where c denotes some universal positive constant.

Given these properties, we now show how our general theory yields the results stated in Corollary 8. To simplify the following discussion, we omit the universal constants and a few-logarithmic terms, and track the dependency only on the sample size n.

Results for gradient ascent The items (P1) and (S1) establish that the gradient operators are slow-convergent and stable, and thus we can apply our general result from Theorem 13. In particular, plugging $\beta = \frac{1}{2}$, and $\gamma = 1$ in Theorem 13, we find that the statistical error for the gradient iterates $\theta^t = (\mathrm{M}_n^{\mathrm{GA}})^t(\theta^0)$ satisfies

$$\left|\theta^t - \theta^\star\right| \precsim [\varepsilon(n, \delta)]^{\frac{\beta}{1+\beta-\gamma\beta}} \asymp [n^{-\frac{1}{2}}]^{\frac{1/2}{1+1/2-1/2}} = n^{-\frac{1}{4}}, \tag{6.24a}$$

for

$$t \succsim [\varepsilon(n, \delta)]^{-\frac{1}{1+\beta-\gamma\beta}} \asymp [n^{-\frac{1}{2}}]^{-\frac{1}{1+1/2-1/2}} = n^{\frac{1}{2}}. \tag{6.24b}$$

Results for Newton's method The items (P2) and (S2) establish that the Newton operators are fast-convergent but unstable, and as a consequence our general result from Theorem 14(a) can be applied. In particular, plugging $\gamma = -1$ in Theorem 14(a), we find that the Newton iterates $\theta^t = (\mathrm{M}_n^{\mathrm{NM}})^t(\theta^0)$ satisfy

$$\begin{aligned}\left|\theta^t - \theta^\star\right| &\precsim \max\left\{[\varepsilon(n,\delta)]^{\frac{1}{1+|\gamma|}}, \widetilde{\rho}_n\right\} \\ &\asymp [n^{-\frac{1}{2}}]^{\frac{1}{1+1}} = n^{-\frac{1}{4}} \quad \text{for} \quad t \succsim \log(1/\varepsilon(n,\delta)) \asymp \log n. \end{aligned} \tag{6.25}$$

Moreover, we show that (see the discussion around equation (6.78)) Assumption (D) holds for the Newton iterates with an initialization outside the ball $\mathbb{B}(\theta^\star, \widetilde{\rho}_n)$, and hence part (b) of the Corollary 8 states that the Newton iterates stay in a close vicinity of $\theta^\star$ for all future iterations.

Over-specified Gaussian mixture models

We now consider the problem of parameter estimation in Gaussian mixture models; and analyze the behavior of two popular algorithms namely (a) Expectation-Maximization (EM) algorithm [DLR97], and (b) Newton's method. We note that EM is arguably the most widely used algorithm for parameter estimation in mixture models and other missing data problems [DLR97]. Here we study the problem of estimating the parameters of a Gaussian mixture model given n i.i.d. samples from the model. When the number of components in the mixture is known, prior works [BWY17; CMZar; DTZ17] have shown that (i) the mixture parameters can be estimated at the parametric rate $n^{-\frac{1}{2}}$ with the EM algorithm and (ii) the algorithm takes at most $\log n$ steps to converge. In the over-specified setting, i.e., when the fitted model has more components than the true model, recent works [Dwi+20a; Dwi+20b; WZ19] have established the slow convergence of EM on both the statistical and algorithmic fronts. For example, for over-specified Gaussian-location mixtures EM takes $n^{\frac{1}{2}} \gg \log n$ steps (where $\gg$ denotes much greater than) to converge and produces an estimate for the mean parameter that has a statistical error of order $n^{-\frac{1}{4}} \gg n^{-\frac{1}{2}}$. In the sequel, we apply our general theory to study the behavior of EM and Newton's method for parameter estimation in over-specified Gaussian-location mixtures. First, we recover the slow convergence of EM as derived in prior works [Dwi+20b]. Second, we prove that the Newton's method—although an unstable algorithm in this setting—achieves a similar statistical accuracy as EM albeit in an exponentially fewer number of steps. We now formalize the details. Let $\phi(\cdot; \theta, \sigma^2)$ denote the density of $\mathcal{N}(\theta, \sigma^2)$ random variable, i.e.,

$$\phi(x; \theta, \sigma^2) = (2\pi\sigma^2)^{-1/2} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \tag{6.26a}$$

and let $X_1, \ldots, X_n$ be n i.i.d. draws from the standard normal distribution (density $\phi(\cdot; 0, 1)$). Given this data, we fit an over-specified mixture model namely, a two-component symmetric Gaussian mixture with equal fixed weights whose density is

given by

$$f_\theta(x) = \frac{1}{2}\phi(x; -\theta, 1) + \frac{1}{2}\phi(x; \theta, 1), \tag{6.26b}$$

where θ is the parameter to be estimated. In such a setting, the true parameter is unique and given by $\theta^* = 0$ since $f_0(\cdot) = \phi(\cdot; 0, 1)$. However, the fact that we fit a mixture that has one extra component than the true model (which has just one component) leads to interesting consequences as we now elaborate. Using $\mathcal{L}_n$ to denote the log-likelihood function, the MLE estimate is given by

$$\widehat{\theta}_{n,\mathrm{MLE}} \in \arg\max_{\theta \in \mathbb{R}} \mathcal{L}_n(\theta) \quad \text{where} \quad \mathcal{L}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \log f_\theta(X_i). \tag{6.26c}$$

On one hand, it is known [Che95] that the over-specification in such a setting leads to a slower than $n^{-\frac{1}{2}}$ statistical rate for the MLE, i.e., $|\widehat{\theta}_{n,\mathrm{MLE}} - \theta^\star| = \mathcal{O}(n^{-\frac{1}{4}})$. On the other hand, MLE does not admit a closed-form expression and thus it is of significant interest to understand the behavior of iterative algorithms that are used to estimate the MLE. Next, we use our general framework to provide a precise characterization of two algorithms namely, EM, and Newton's method on maximizing the log-likelihood $\mathcal{L}_n$ (6.26c).

Theoretical guarantees

The next corollary provides a precise characterization of EM and Newton's method for the over-specified setting described in the previous section. We analyze the EM updates and the pure Newton updates. Moreover, we use $\mathrm{G}_n^{\mathrm{EM}}$ and $\mathrm{G}_n^{\mathrm{NM}}$ respectively to denote the sample-based operators for EM and Newton's method (see Section 6.9 for the precise form of these operators). Finally, the scalars c, c_i, c_i', c_i'' denote some positive universal constants.

Corollary 9. *For the over-specified Gaussian mixture model (6.26) with $\theta^\star = 0$, given some $\delta \in (0, 1)$, the following properties hold with probability at least $1 - \delta$:*

(a) For any fixed $\epsilon \in (0, 1/4)$ and initialization $\theta^0 \in \mathbb{B}(\theta^\star, 1)$, the sequence $\theta^t := (\mathrm{G}_n^{\mathrm{EM}})^t(\theta^0)$ of EM iterates satisfies the bound

$$\left|\theta^t - \theta^\star\right| \leq c_1 \left(\frac{\log(\frac{\log(1/\epsilon)}{\delta})}{n}\right)^{\frac{1}{4}-\epsilon} \qquad \text{for all iterates } t \geq c_1' \sqrt{n} \log \tfrac{1}{\epsilon}, \tag{6.27a}$$

as long as $n \geq c_1'' \log \frac{\log(1/\epsilon)}{\delta}$.

(b) For any initialization $\theta^0 \in \mathbb{A}(\theta^\star, \frac{\sqrt{2c}\log^2(3n/\delta)}{n^{1/4}}, 1/3)$, the sequence of Newton iterates $\theta^t := (\mathrm{G}_n^{\mathrm{NM}})^t(\theta^0)$ satisfies the bound

$$\left|\theta^t - \theta^\star\right| \leq c_2 \left(\frac{\log(n/\delta)}{n}\right)^{\frac{1}{4}} \qquad \text{for all iterates } t \geq c_2' \log n, \tag{6.27b}$$

as long as $n \geq c_2'' \log(1/\delta)$.

See Section 6.9 for the proof (and below for the proof sketch).

Corollary 9 establishes that the Newton EM is significantly faster than EM for the model setup (6.26). More precisely, it reaches ball around $\theta^\star$ with a statistical radius of order $n^{-\frac{1}{4}}$ within $\log n$ steps, which is much smaller than the number of steps taken by EM Moreover, the updates from Newton's method do not escape this ball for future iterations. This behavior is a consequence of the fact that under the assumed initialization condition, the (cubic-regularized) Newton EM sequence satisfies assumption (D).

Proof sketch for Corollary 9

The proof strategy for this case is similar to that laid out in Section 6.4 for informative non-response model. First, to study this problem in our framework, we consider the population level objective $\mathcal{L}$ by replacing the sum over samples in equation (6.26c) with the corresponding expectation:

$$\mathcal{L}(\theta) := \mathbb{E}_{X \sim \mathcal{N}(0,1)} \left[\log f_\theta(X) \right] = \mathbb{E}_X \left[\frac{1}{2} \phi(X; -\theta, 1) + \frac{1}{2} \phi(X; \theta, 1) \right]. \tag{6.28}$$

Second, we use G^{EM} and G^{NM} respectively to denote the corresponding population-level EM and Newton's method operators (see Section 6.9 for the precise expressions).

Results for EM For the case of $\theta^\star = 0$, Theorem 2 and Lemma 1 of our prior work [Dwi+20b] show that, for any initialization θ^0, the EM operators G^{EM} and $\mathrm{G}_n^{\mathrm{EM}}$ satisfy

$$\begin{aligned} \left| (\mathrm{G}^{\mathrm{EM}})^t(\theta^0) - \theta^\star \right| &\leq \frac{c}{t^{\frac{1}{2}}} \quad \text{and,} \\ \sup_{\theta \in \mathbb{B}(\theta^\star, r)} \left| \mathrm{G}^{\mathrm{EM}}(\theta) - \mathrm{G}_n^{\mathrm{EM}}(\theta) \right| &\leq c_1 r \cdot \sqrt{\frac{\log(1/\delta)}{n}}, \end{aligned} \tag{6.29}$$

where the second bound holds with probability at least $1 - \delta$ for any fixed radius $r > 0$. In the framework of our current work, the bounds (6.29) imply that the operator G^{EM} exhibits $\texttt{SLOW}(\frac{1}{2})$-convergence, and the operator $\mathrm{G}_n^{\mathrm{EM}}$ is $\texttt{STA}(1)$-stable with the noise function $\sqrt{\frac{\log(1/\delta)}{n}}$. Thus a direct application of Theorem 13 of this chapter (in a fashion similar to that of equations (6.24a) and (6.24b)), recovers the main result of our prior work [Dwi+20b] (Theorem 3). That is, with high probability, the sequence $\theta_n^{t+1} = \mathrm{G}_n^{\mathrm{EM}}(\theta_n^t)$ satisfies

$$\left| \theta^t - \theta^\star \right| \precsim [n^{-\frac{1}{2}}]^{\frac{1/2}{1+1/2-1/2}} = n^{-\frac{1}{4}} \quad \text{for} \quad t \succsim [n^{-\frac{1}{2}}]^{-\frac{1}{1+1/2-1/2}} = n^{\frac{1}{2}}. \tag{6.30}$$

Results for Newton's method In Section 6.9, we demonstrate the following properties of Newton's method operators:

(M1) the Newton operator G^{NM} is $\texttt{FAST}(\frac{7}{9})$-convergent over the ball $\mathbb{B}(\theta^\star, \frac{1}{3})$, and

(M2) the operator $\mathrm{G}_n^{\mathrm{NM}}$ is $\texttt{UNS}(-1)$-unstable over the annulus $\mathbb{A}(\theta^\star, \widetilde{\rho}_n, 1/3)$ with noise function $\varepsilon(n,\delta) = \frac{\log(n/\delta)}{\sqrt{n}}$ where $\widetilde{\rho}_n = \frac{c\log^2(3n/\delta)}{n^{1/4}}$.

Based on the results of Theorem 14(a) with $\kappa = \frac{7}{9}$ and $\gamma = -1$, the items (M1) and (M2) suggest that the Newton updates $\theta^t = (\mathrm{M}_n^{\mathrm{NM}})^t(\theta^0)$ satisfy

$$\left|\theta^t - \theta^\star\right| \precsim \max\left\{[\varepsilon(n,\delta)]^{\frac{1}{1+1}}, \widetilde{\rho}_n\right\} \precsim n^{-\frac{1}{4}} \quad \text{for} \quad t \succsim \log(1/\varepsilon(n,\delta)) \asymp \log n. \tag{6.31}$$

Furthermore, we prove that the Newton iterates satisfy Assumption (D) (see the argument with equation (6.88)). Therefore, the Newton iterates stay in a close vicinity of $\theta^\star$ for all future iterations.

Single-index model

In our third example, we consider a single-index regression model [Car+97] with a known link function g. Models of this type have proven useful for applications in signal processing, econometrics, statistics, and machine learning [HH96; Ich93]. For simplicity, we briefly summarize the one-dimensional version of this problem. We observe the pairs of data $(X_i, Y_i) \in \mathbb{R}^2$ that are generated from the model

$$Y_i = g\left(X_i\theta^*\right) + \xi_i \quad \text{for} \qquad i = 1, \ldots, n. \tag{6.32a}$$

Here Y_i denotes the response variable, X_i corresponds to the covariate and ξ_i denotes the additive noise assumed to have a standard Gaussian distribution, i.e., $\xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.

In this example, we consider the case of random design for the covariates, i.e., the covariates $\{X_i\}_{i=1}^n$ are independent and $X_i \sim \mathcal{N}(0,1)$. Given the samples $\{(X_i, Y_i), i \in [n]\}$, we want to estimate the unknown parameter θ^*. A popular choice is the maximum-likelihood estimate (MLE):

$$\widehat{\theta}_n^{\mathrm{mle}} \in \arg\min_{\theta\in\mathbb{R}} \widetilde{\mathcal{L}}_n(\theta) \quad \text{where} \quad \widetilde{\mathcal{L}}_n := \frac{1}{2n}\sum_{i=1}^n \left(Y_i - g\left(X_i\theta\right)\right)^2. \tag{6.32b}$$

Generally, the loss-function $\widetilde{\mathcal{L}}_n$ is non-convex and hence the MLE does not admit a closed-form expression. Consequently, one needs to make use of certain optimization algorithms to compute an estimate $\widehat{\theta}_n$, which need not be the same as $\widehat{\theta}_n^{\mathrm{mle}}$.

In the remainder of this section, we study the case when the SNR degenerates to zero. Specifically, we consider $\theta^* = 0$ and a link function of the form $g(x) = x^{2p}$ with $p \geq 1$. For such a setting, the optimization problem (6.32b) takes the following form:

$$\widehat{\theta}_n \in \arg\min_{\theta\in\mathbb{R}} \widetilde{\mathcal{L}}_n(\theta) \quad \text{where} \quad \widetilde{\mathcal{L}}_n := \frac{1}{2n}\sum_{i=1}^n \left(Y_i - (X_i\theta)^{2p}\right)^2. \tag{6.32c}$$

Theoretical guarantees

For the single-index model described above with the link function $g(x) = x^{2p}$, we consider three iterative optimization methods: (a) gradient descent with a step size $\eta \in (0, \frac{1}{(4p-1)!!(2p)}]$, (b) (pure) Newton's method, and (c) cubic-regularized Newton's method with Lipschitz constant $L := (4p-1)!!(4p-1)p/3$. We denote the updates for these three methods via the operators $\mathrm{F}_n^{\mathrm{GD}}$, $\mathrm{F}_n^{\mathrm{NM}}$, and $\mathrm{F}_n^{\mathrm{CNM}}$ respectively (see Section 6.9 for the precise expressions of these operators). The next result characterizes the behavior of these three methods:

Corollary 10. *For the single-index model (6.32) with link function $g(x) = x^{2p}$ for $p \geq 1$ and true parameter $\theta^\star = 0$, given some $\delta \in (0,1)$, the following properties hold with probability at least $1 - \delta$:*

(a) For any fixed $\epsilon \in (0, 1/4)$ and initialization $\theta^0 \in \mathbb{B}(\theta^\star, 1)$, the sequence $\theta^t := (\mathrm{F}_n^{\mathrm{GD}})^t(\theta^0)$ of gradient iterates satisfies the bound

$$\left|\theta^t - \theta^\star\right| \leq c_1 \left(\frac{\log^{4p}(n\frac{\log(1/\epsilon)}{\delta})}{n}\right)^{\frac{1}{4p}-\epsilon} \quad \text{for all iterates } t \geq c_1' n^{\frac{2p-1}{2p}} \log\tfrac{1}{\epsilon}, \tag{6.33a}$$

as long as $n \geq c_1'' \log \frac{\log(1/\epsilon)}{\delta}$.

(b) For any initialization $\theta^0 \in \mathbb{A}(\theta^\star, c\frac{\log^{p/(2p-1)}(n/\delta)}{n^{1/4(2p-1)}}, 1)$, the sequence of Newton iterates $\theta^t := (\mathrm{F}_n^{\mathrm{NM}})^t(\theta^0)$ satisfies the bound

$$\left|\theta^t - \theta^\star\right| \leq c_2 \left(\frac{\log^{4p}(n/\delta)}{n}\right)^{\frac{1}{4p}} \quad \text{for all iterates } t \geq c_2' \log n, \tag{6.33b}$$

as long as $n \geq c_2'' \log(1/\delta)$.

(c) The sequence of cubic-regularized Newton iterates $\theta^t := (\mathrm{F}_n^{\mathrm{CNM}})^t(\theta^0)$ with initialization $\theta^0 \in \mathbb{A}(\theta^\star, c\frac{\log^{p/(2p-1)}(n/\delta)}{n^{1/4(2p-1)}}, 1)$ satisfies the bound

$$\left|\theta^t - \theta^\star\right| \leq c_3 \left(\frac{\log^{4p}(n/\delta)}{n}\right)^{\frac{1}{4p}} \quad \text{for all iterates } t \geq c_3' n^{\frac{4p-3}{2(4p-1)}}, \tag{6.33c}$$

as long as $n \geq c_3'' \log(1/\delta)$.

See Section 6.9 for the proof (and below for the proof sketch).

This corollary shows that the final statistical errors achieved by gradient descent and the (cubic-regularized) Newton's method have the same scaling. Moreover, Newton's method, while unstable, converges to the correct statistical radius in a significantly smaller $\log n$ number of steps when compared to gradient descent, which

takes $n^{\frac{2p-1}{2p}}$ steps and cubic-regularized Newton's method, which takes $n^{\frac{4p-3}{2(4p-1)}}$ steps. Moreover, we also show that assumption (D) holds for the iterates from the (cubic-regularized) Newton method's[4] and hence we obtain that these iterates not only converge to a ball of radius $n^{-\frac{1}{4p}}$ around $\theta^\star$, but also that they stay there for all the future iterations. Finally, in Section 6.9 (see equation (6.107)) we also establish that the statistical radius $n^{-1/(4p)}$ achieved by the considered optimization methods is tight.

When $g(x) = x^2$, the model (6.32a) corresponds to a phase retrieval problem. In the regime of large signal-to-noise ratio (SNR), i.e., $|\theta^\star| \gg 1$, and with the link function $g(x) = x^2$, there are efficient algorithms which produce an estimate $\widehat{\theta}_n$ satisfying a bound $|\widehat{\theta}_n - \theta^\star| \precsim n^{-\frac{1}{2}}$ [CLS15; EM13; TV18]. However, as the SNR approaches zero these parametric rates do not apply and precise statistical behavior of these estimates are not known.

Proof sketch for Corollary 10

In order to study these updates using our framework, we need to consider the population-level version of the optimization problem (6.32c), which is given by

$$\min_{\theta \in \mathbb{R}} \widetilde{\mathcal{L}}(\theta) \quad \text{where} \quad \widetilde{\mathcal{L}}(\theta) := \frac{1}{2}\mathbb{E}_{(X,Y)}\left[\left(Y - (X\theta)^{2p}\right)^2\right],$$

where the expectation is taken with respect to $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0,1)$ as $\theta^\star = 0$. Direct computation yields that

$$\widetilde{\mathcal{L}}(\theta) = \frac{1}{2} + \frac{(4p-1)!!\theta^{4p}}{2} \qquad \text{and} \qquad \arg\min_{\theta} \ \widetilde{\mathcal{L}}(\theta) = 0 = \theta^\star. \tag{6.34}$$

Like the previous proof sketches, we let $\mathrm{F}^{\mathrm{GD}}, \mathrm{F}^{\mathrm{NM}}$ and $\mathrm{F}^{\mathrm{CNM}}$ denote the population operators corresponding to the algorithms, gradient descent, Newton's method and cubic-regularized Newton's method, for the problem (6.34) (for a given p). See Section 6.9 for the precise definitions of these operators. In Section 6.9, we show that with $\theta^\star = 0$, these population-level operators satisfy the following properties over the ball $\mathbb{B}(\theta^\star, 1)$:

($\widetilde{\mathrm{P}}1$) the gradient operator F^{GD} is $\texttt{SLOW}(\frac{1}{4p-2})$-convergent for any step size $\eta \in (0, \frac{1}{(4p-1)!!(2p)}]$,

($\widetilde{\mathrm{P}}2$) the Newton operator F^{NM} is $\texttt{FAST}(\frac{4p-2}{4p-1})$-convergent, and

($\widetilde{\mathrm{P}}3$) the cubic-regularized Newton operator $\mathrm{F}^{\mathrm{CNM}}$ is $\texttt{SLOW}(\frac{2}{4p-3})$-convergent.

Moreover in the same Section 6.9, we also show that with the noise function $\varepsilon(n,\delta) = \sqrt{\frac{\log^{4p}(n/\delta)}{n}}$, the sample-level operators satisfy the following properties:

[4] See the proofs of equations (6.102) and (6.108) in Section 6.9 for more details.

($\widetilde{\text{S}}$1) the operator F_n^{GD} is $\texttt{STA}(2p-1)$-stable over the ball $\mathbb{B}(\theta^\star, 1)$,

($\widetilde{\text{S}}$2) the operator F_n^{NM} is $\texttt{UNS}(-(2p-1))$-unstable over the annulus $\mathbb{A}(\theta^\star, \widetilde{\rho}_n, 1)$ with inner radius $\widetilde{\rho}_n = c\ \log^{p/(2p-1)}(n/\delta)/n^{1/4(2p-1)}$, and

($\widetilde{\text{S}}$3) the operator F_n^{CNM} is $\texttt{UNS}(-\frac{1}{2})$-unstable over the annulus $\mathbb{A}(\theta^\star, \widetilde{\rho}_n, 1)$.

These properties show that the gradient descent is a slow-converging stable method and we can apply Theorem 13. On the other hand, Newton's method is a fast-converging unstable method, and Theorem 14(a) can be applied. Finally, cubic-regularized Newton's method is a slow-converging unstable method and Theorem 14(b) can be applied. In the subsequent proof-sketch, we track the dependency only on the sample size n and ignore logarithmic factors and universal constants. Moreover, since the computations here mimic the discussion from Section 6.4, we keep the discussion briefer.

Results for gradient descent Applying Theorem 13 with $\beta = \frac{1}{4p-2}$, and $\gamma = 2p-1$ (items ($\widetilde{\text{P}}$1) and ($\widetilde{\text{S}}$1) respectively), we find that the statistical error for the gradient iterates $\theta^t = (\text{F}_n^{\text{GD}})^t(\theta^0)$ satisfy

$$\left|\theta^t - \theta^\star\right| \precsim [\varepsilon(n,\delta)]^{\frac{\beta}{1+\beta-\gamma\beta}} \precsim n^{-\frac{1}{2p}} \quad \text{for} \quad t \succsim [\varepsilon(n,\delta)]^{-\frac{1}{1+\beta-\gamma\beta}} \asymp n^{\frac{2p-1}{2p}}. \tag{6.35}$$

Results for Newton's method Next applying Theorem 14(a) for the Newton's method with $\kappa = \frac{4p-2}{4p-1}$, and $\gamma = -(2p-1)$ (see items ($\widetilde{\text{P}}$2) and ($\widetilde{\text{S}}$2)), we conclude that the updates $\theta^t = (\text{F}_n^{\text{NM}})^t(\theta^0)$ from the Newton's method have the following property:

$$\left|\theta^t - \theta^\star\right| \precsim \max\left\{[\varepsilon(n,\delta)]^{\frac{1}{1+|\gamma|}}, \widetilde{\rho}_n\right\} \precsim n^{-\frac{1}{2p}} \quad \text{for} \quad t \succsim \log(1/\varepsilon(n,\delta)) \asymp \log n. \tag{6.36}$$

Results for cubic-regularized Newton's method Finally by using Theorem 14(b) for the cubic-regularized Newton's method with $\beta = \frac{2}{4p-3}$, and $\gamma = -\frac{1}{2}$ (see items ($\widetilde{\text{P}}$3) and ($\widetilde{\text{S}}$3)), the following results hold for the cubic-regularized Newton iterates $\theta^t = (\text{F}_n^{\text{CNM}})^t(\theta^0)$:

$$\begin{aligned}\left|\theta^t - \theta^\star\right| &\precsim \max\left\{[\varepsilon(n,\delta)]^{\frac{\beta}{1+\beta-\gamma\beta}}, \widetilde{\rho}_n\right\} \\ &\precsim n^{-\frac{1}{2p}} \quad \text{for} \quad t \succsim [\varepsilon(n,\delta)]^{-\frac{1}{1+\beta}} \asymp n^{\frac{4p-3}{2(4p-1)}}.\end{aligned} \tag{6.37}$$

6.5 Discussion

In this chapter, we established several results characterizing the statistical radius achieved by a sequence of updates $\{F_n^t(\theta_n^0)\}_{t\geq 0}$, induced by an operator F_n and a given initial point θ_n^0. We established these results by analyzing the interplay

between (in)-stability of the operator F_n for its population operator F and the local convergence of F around its fixed point $\theta^\star$. We then applied our general theory to derive sharp algorithmic and statistical guarantees for several iterative algorithms by analyzing the corresponding sample and population operators, in three different statistical settings. In particular, we studied the behavior of gradient methods and higher-order (cubic-regularized) Newton's method for parameter estimation—in the weak signal-to-noise ratio regime—in Gaussian mixture models, single-index models, and informative non-response models. We showed that for such models, despite instability, fast algorithms like Newton's method may still be preferred over a stable one like gradient descent since they achieve the same statistical accuracy as that of the stable counterpart in exponentially fewer steps.

We now discuss a few questions that arise naturally from our work. First, our results, as stated, are not directly applicable to the settings of accelerated optimization methods or quasi-Newton methods, e.g., accelerated gradient descent [Nes13] and L-BFGS [Fle87]. On the one hand, the updates from an accelerated gradient descent method require that the operators F_n and F to change with each iteration. On the other hand, the updates from the L-BFGS method would require additional machinery to deal with the preconditioning matrices in each step. Developing a general theory to characterize the statistical performance of algorithms associated with a time-varying operator F_n is an interesting direction for future research.

Secondly, it is desirable to understand the behavior of optimization methods to a wider range of statistical problems. In the context of mixture models, recent work by Dwivedi et al. [Dwi+20a] established that for over-specified mixtures with both location and scale parameter unknown, EM takes an $\mathcal{O}(n^{\frac{3}{4}})$ steps to return estimates with minimax statistical error of order $n^{-\frac{1}{8}}$ and $n^{-\frac{1}{4}}$ for the location and scale parameter, respectively. Whether an unstable method like (cubic-regularized) Newton's EM proves computationally advantageous (without losing statistical accuracy) in such more challenging non-convex landscapes remains an open problem.

Finally, our theory does not easily extend to the settings with dependent data, such as time series. When the samples are (time) dependent, taking the limit of infinite sample size does not yield a natural population-level operator. One possible fix is to borrow the technique of truncating the sample operator from the analysis of the Baum-Welch algorithm for hidden Markov models [YBW17]. However, even with the help of such a technique, ample technical challenges remain towards developing a general theory for such non-i.i.d. settings.

6.6 Proofs of main results

We now turn to the proofs of Theorems 13 and 14.

Proof of Theorem 13

The reader should recall the proof outline provided following the statement of the theorem. Our proof here follows this outline, making each step precise. For the remainder of the proof, we assume without loss of generality that $\theta^\star = 0$ and $r_0 = 1$. Proofs for the cases $\theta^\star \neq 0$ or $r_0 > 1$ can be reduced to this case in a straightforward fashion and are thereby omitted.

Notation for stable case

For each $\ell = 1, 2, \ldots$, let t_ℓ denote the number of iterations during the ℓ-th epoch, and let T_ℓ denote the total number of iterations up to the completion epoch ℓ. In order to define them precisely, we first introduce

$$\begin{aligned} t_\ell^{(1)} := C\varepsilon(n,\delta^*)^{-\frac{\alpha_{\ell-1}(\gamma)+1}{1+\beta}} \quad \text{and} \quad t_\ell^{(2)} := C'\varepsilon(n,\delta^*)^{-\frac{\alpha_\ell(\gamma)+1}{1+\beta}}, \\ \text{for} \quad C := (c_2 2^\gamma)^{-\frac{1}{(1+\beta)}} \qquad \text{and} \qquad C' := C(c')^{\frac{\gamma}{1+\beta}}, \end{aligned} \tag{6.38a}$$

where $c' := (c_2 2^\gamma)^{\frac{\beta}{1+\beta}} = C^{-\beta}$ and hence we have $C' = C^{\frac{1+\beta+\beta\gamma}{1+\beta}}$. Here the constant c_2 is the constant from the the stability definition (6.11). We then define $t_0 := 0$, and

$$t_\ell := \left\lceil t_\ell^{(1)} + t_\ell^{(2)} \right\rceil \quad \text{and} \quad T_\ell := \sum_{j=0}^{\ell} t_j \quad \text{for } \ell = 1, 2, \ldots. \tag{6.38b}$$

Our proof is based on studying the sequence of real-numbers $\{\alpha_\ell\}_{\ell \geq 0}$ given by

$$\alpha_0 = 0 \quad \text{and} \quad \alpha_{\ell+1} = \alpha_\ell \nu + \nu', \quad \text{where } \nu = \tfrac{\beta\gamma}{1+\beta} \text{ and } \nu' = \tfrac{\beta}{1+\beta}. \tag{6.38c}$$

Note that Assumption (B) implies that $\nu \in (0,1)$ and hence

$$\alpha_\ell = \nu_\star(1 - \nu^\ell) \uparrow \nu_\star \quad \text{where} \quad \nu_\star := \frac{\beta}{1+\beta-\gamma\beta}. \tag{6.38d}$$

In the epoch-based argument, we need to control the deviation $\sup_{\|\theta\|_2 \leq r} \|F(\theta) - F_n(\theta)\|_2$ uniformly for each radii $r \in \mathcal{R}'$. To this end, for any tolerance $\delta \in (0,1)$, we define the event $\mathcal{E}$ by

$$\mathcal{E} := \left\{ \sup_{\theta \in \mathbb{B}(\theta^\star, r)} \|F(\theta) - F_n(\theta)\|_2 \leq c_2 r^\gamma \varepsilon(n,\delta^*) \quad \text{uniformly for all } r \in \mathcal{R}' \right\}, \tag{6.39}$$

where $\delta^* = \delta \cdot \frac{\log(\frac{1+\beta}{\beta\gamma})}{8\log(\frac{\beta}{\epsilon(1+\beta-\gamma\beta)})}$ was defined in equation (6.14) and the radii-set $\mathcal{R}'$ is defined as

$$\begin{aligned} \mathcal{R}' := \mathcal{R} \cup 2\mathcal{R}, \quad \text{with} \\ \mathcal{R} := \{\varepsilon(n,\delta^*)^{\alpha_0}, \ldots, \varepsilon(n,\delta^*)^{\alpha_{\ell_\epsilon}}, c'\varepsilon(n,\delta^*)^{\alpha_0}, \ldots, c'\varepsilon(n,\delta^*)^{\alpha_{\ell_\epsilon}}\}, \\ \ell_\epsilon = \lceil \log(1/\alpha) \rceil \quad \text{and} \quad c' = (c_2 2^\gamma)^{\frac{\beta}{1+\beta}}. \end{aligned} \tag{6.40}$$

Combining the $\texttt{STA}(\gamma)$-stability assumption (6.11) with a standard application of union bound we conclude that

$$\mathbb{P}(\mathcal{E}) \geq 1 - \delta. \tag{6.41}$$

Before we start the main argument, we state a lemma useful in the proof of our theorem:

Lemma 31. *Assume that the assumptions of Theorem 13 are in force. Then conditioned on the event $\mathcal{E}$ (6.39. 6.41), for all radius r in the set $\mathcal{R}$ (6.40), we have*

$$\sup_{\theta \in \mathbb{B}(\theta^\star, r)} \|F^t(\theta) - F_n^t(\theta)\|_2 \leq c_2 (2r)^\gamma \varepsilon(n, \delta^*) \cdot t \quad \text{for all } t \leq \widetilde{\mathcal{T}}(r), \tag{6.42}$$

where $\widetilde{\mathcal{T}}(r) := \frac{r^{1-\gamma}}{2^\gamma c_2 \varepsilon(n,\delta^*)}$. Furthermore, for all $\ell \leq \ell_\epsilon$ we have

$$t_{\ell+1}^{(1)} \leq \widetilde{\mathcal{T}}(\varepsilon(n,\delta^*)^{\alpha_\ell}) \quad \text{and} \quad t_{\ell+1}^{(2)} \leq \widetilde{\mathcal{T}}(c' \varepsilon(n,\delta^*)^{\alpha_{\ell+1}}). \tag{6.43}$$

See Section 6.8 for the proof of this lemma.

Main argument

We claim that the sequence $\{\theta_n^t\}_{t \geq 1}$ satisfies

$$\|\theta_n^{T_\ell}\|_2 \leq \varepsilon(n,\delta^*)^{\alpha_\ell} \quad \text{uniformly for all } \ell \in \{0, 1, \ldots, \ell_\epsilon\}, \quad \text{and} \tag{6.44a}$$

$$\|\theta_n^{T_{\ell_\epsilon}+t}\|_2 \leq 2\varepsilon(n,\delta^*)^{\nu_\star - \epsilon} \quad \text{uniformly for all } t \in \{0, 1, 2, \ldots\}, \tag{6.44b}$$

with probability at least $1 - \delta$. The quantities α_ℓ, T_ℓ and ℓ_ϵ are defined in equations (6.38a) through equation (6.38c). With these claims at our disposal, it remains to prove an upper bound on the scalar T_{ℓ_ϵ}. Towards this end, doing some straightforward algebra we find that

$$t_\ell \leq t_{\ell_\epsilon} \leq c' \varepsilon(n,\delta^*)^{-\frac{\nu_\star}{\beta}} \quad \text{for any} \quad 0 \leq \ell \leq \ell_\epsilon. \tag{6.45}$$

Combining the above bounds on t_ℓ with the definition of T_ℓ from equation (6.38b) yields an upper bound on T_{ℓ_ϵ}. Substituting the upper bound on T_{ℓ_ϵ} in inequality (6.44b) yields the claimed bound (6.15) of Theorem 13. We now prove the claims (6.44a) and (6.44b) using induction.

Proof of claim (6.44a)

We condition on the event $\mathcal{E}$ defined in the equation (6.39), which occurs with probability at least $1 - \delta$, and establish the claim using induction on the epoch index ℓ. The base case $\ell = 0$ is immediate. We now establish the inductive step, i.e., given

$\|\theta_n^{T_\ell}\|_2 \leq \varepsilon(n,\delta^*)^{\alpha_\ell}$ for some $\ell \leq \ell_\epsilon - 1$, we show that $\|\theta_n^{T_{\ell+1}}\|_2 \leq \varepsilon(n,\delta^*)^{\alpha_{\ell+1}}$. We split the proof in two parts (primarily to handle the constants):

$$\|\theta_n^{T_\ell + t_{\ell+1}^{(1)}}\|_2 \leq c'\varepsilon(n,\delta^*)^{\alpha_{\ell+1}} \quad \text{and} \tag{6.46a}$$

$$\|\theta_n^{T_\ell + t_{\ell+1}^{(1)} + t_{\ell+1}^{(2)}}\|_2 \leq \varepsilon(n,\delta^*)^{\alpha_{\ell+1}}, \tag{6.46b}$$

where $c' > 1$ is a universal constant. These claims together imply the induction hypothesis and thereby the claim (6.44a).

Proof of claim (6.46a)**:** Inequality (6.43) implies that $t_{\ell+1}^{(1)} \leq \widetilde{\mathcal{T}}(\varepsilon(n,\delta^*)^{\alpha_\ell})$, and hence we can apply the bound (6.42) from Lemma 31 with $r = \varepsilon(n,\delta^*)^{\alpha_\ell} \in \mathcal{R}$ for any $t \leq t_{\ell+1}^{(1)}$. Applying the triangle inequality yields

$$\begin{aligned}
\|\theta_n^{t+T_\ell}\|_2 = \|F_n^t(\theta_n^{T_\ell})\|_2 &\leq \|F^t(\theta_n^{T_\ell})\|_2 + \|F^t(\theta_n^{T_\ell}) - F_n^t(\theta_n^{T_\ell})\|_2 \\
&\overset{(i)}{\leq} \frac{1}{t^\beta} + \|F^t(\theta_n^{T_\ell}) - F_n^t(\theta_n^{T_\ell})\|_2 \qquad (6.47) \\
&\overset{(ii)}{\leq} \frac{1}{t^\beta} + c_2(2\varepsilon(n,\delta^*)^{\alpha_\ell})^\gamma \varepsilon(n,\delta^*)t, \qquad (6.48)
\end{aligned}$$

for any $t \leq t_{\ell+1}^{(1)}$; where step (i) follows from the $\texttt{SLOW}(\beta)$-convergence (6.8) of the operator F along with the assumption that $\theta^\star = 0$, and step (ii) follows by using the inductive hypothesis $\|\theta_n^{T_\ell}\|_2 \leq \varepsilon(n,\delta^*)^{\alpha_\ell}$ and applying Lemma 31 with $r = \varepsilon(n,\delta^*)^{\alpha_\ell}$. Note that in the final bound (6.48) the first term decreases with iteration t while the second term increases with t. In order to trade off these two terms,[5] we set $t = t_{\ell+1}^{(1)}$ (6.38a) in the bound (6.48) and find that

$$\begin{aligned}
\|\theta_n^{T_\ell + t_{\ell+1}^{(1)}}\|_2 \leq \frac{1}{(t_{\ell+1}^{(1)})^\beta} + c_2(2\varepsilon(n,\delta^*)^{\alpha_\ell})^\gamma \varepsilon(n,\delta^*) t_{\ell+1}^{(1)} &= \underbrace{2(c_2 2^\gamma)^{\frac{\beta}{1+\beta}}}_{=:c'} \cdot \varepsilon(n,\delta^*)^{1 - \frac{\alpha_\ell\gamma+1}{1+\beta} + \alpha_\ell\gamma} \\
&= c'\varepsilon(n,\delta^*)^{\frac{\alpha_\ell(\beta\gamma)+\beta}{1+\beta}} \\
&= c'\varepsilon(n,\delta^*)^{\alpha_{\ell+1}},
\end{aligned}$$

where the last equality follows from the relation (6.38c) between α_ℓ and $\alpha_{\ell+1}$. The claim (6.46a) now follows.

Proof of claim (6.46b)**:** For any $t \leq \widetilde{\mathcal{T}}(c'\varepsilon(n,\delta^*)^{\alpha_{\ell+1}})$, we have

$$\begin{aligned}
\|\theta_n^{t+T_\ell+t_{\ell+1}^{(1)}}\|_2 &\leq \|F^t(\theta_n^{T_\ell+t_{\ell+1}^{(1)}})\|_2 + \|F^t(\theta_n^{T_\ell+t_{\ell+1}^{(1)}}) - F_n^t(\theta_n^{T_\ell+t_{\ell+1}^{(1)}})\|_2 \\
&\leq \frac{1}{t^\beta} + c_2(2c'\varepsilon(n,\delta^*)^{\alpha_{\ell+1}})^\gamma \varepsilon(n,\delta^*)t,
\end{aligned}$$

[5] We ignore the effect of the ceiling function $\lceil\cdot\rceil$ to simplify the computations

where the last inequality follows from arguments similar to those used to establish the inequalities (6.47) and (6.48) above. Next, recalling the inequality $t_{\ell+1}^{(2)} \leq \widetilde{\mathcal{T}}(c'\varepsilon(n,\delta^*)^{\alpha_{\ell+1}})$ from equation (6.43) and plugging $t = t_{\ell+1}^{(2)}$ (6.38a) in the above inequality, we find that

$$\|\theta_n^{T_{\ell+1}}\|_2 \leq \underbrace{2(c_2 2^\gamma)^{\frac{\beta}{1+\beta}} c'^{\frac{\beta\gamma}{1+\beta}}}_{=:\widetilde{C}} \cdot \varepsilon(n,\delta^*)^{\frac{\alpha_{\ell+1}\beta\gamma+\beta}{1+\beta}} = \widetilde{C}\varepsilon(n,\delta^*)^{\alpha_{\ell+2}}.$$

In order to complete the proof, it remains to show that last quantity is upper bounded by $\varepsilon(n,\delta^*)^{\alpha_{\ell+1}}$; equivalently, we need to verify the following upper bound

$$\varepsilon(n,\delta^*) \leq \frac{1}{\widetilde{C}^{\alpha_{\ell+2}-\alpha_{\ell+1}}}, \tag{6.49}$$

which is equivalent to the large sample-size assumption (C) (see condition (6.70) for a more precise statement) if we establish that

$$\alpha_{\ell+2} - \alpha_{\ell+1} \geq \epsilon_\star := \frac{\epsilon(1+\beta-\beta\gamma)}{1+\beta}. \tag{6.50}$$

In order to do so, we use the fact (6.38d) that $\alpha_\ell = \nu_\star(1-\nu^\ell)$ and obtain that

$$\alpha_\ell \leq \nu_\star - \alpha \quad \text{and consequently that} \quad \nu_\star\nu^\ell \geq \alpha$$

for all $\ell \in \{0,1,\ldots,\ell_\epsilon\}$. Putting together the pieces we have

$$\alpha_{\ell+2} - \alpha_{\ell+1} = \nu_\star\nu^{\ell+1}(1-\nu) \geq \alpha(1-\nu) = \epsilon_\star,$$

which yields the claimed bound (6.50) and we are done.

Proof of claim (6.44b)

The proof of this claim follows a similar road-map as that in the previous Section, and hence we simply sketch it. Conditional on the event $\mathcal{E}$, we claim that

$$\|\theta_n^{T_{\ell_\epsilon}+kt_{\ell_\epsilon}}\|_2 \leq \varepsilon(n,\delta^*)^{\nu_\star-\epsilon} \quad \text{uniformly for all } k \in \{0,1,2,\ldots\}. \tag{6.51}$$

Assuming this bound is given for now, we complete the proof. Invoking inequality (6.66) from the proof of Lemma 31, we obtain that

$$\|\theta_n^{T_{\ell_\epsilon}+kt_{\ell_\epsilon}+t}\|_2 \leq 2\varepsilon(n,\delta^*)^{\nu_\star-\epsilon} \quad \text{for all} \quad k \in \{1,2,\ldots\} \text{ and } t \leq \widetilde{\mathcal{T}}(\varepsilon(n,\delta^*)^{\nu_\star-\epsilon}). \tag{6.52}$$

Mimicking the arguments from claims (6.46a) and (6.46b), and using the large sample-size assumption (C) (condition (6.70)) yields the claim (6.52) for any $t \leq$

$\varepsilon(n,\delta^*)^{-\frac{\nu_\star}{\beta}}$. Putting this together with the fact (6.45) that $t_{\ell_\epsilon} \leq \varepsilon(n,\delta^*)^{-\frac{\nu_\star}{\beta}}$ implies the claim (6.44b).

Turning to the proof of claim (6.51), we note that the base case $k = 0$ follows from the claim (6.44a) by plugging in $\ell = \ell_\epsilon$. For the inductive step, assuming $\|\theta_n^{T_{\ell_\epsilon}+kt_{\ell_\epsilon}}\|_2 \leq \varepsilon(n,\delta^*)^{\nu_\star-\epsilon}$, arguments similar to that in the proof of claims (6.46a) and (6.46b) yield

$$\|\theta_n^{T_{\ell_\epsilon}+kt_{\ell_\epsilon}+t_{\ell_\epsilon}^{(1)}}\|_2 \leq c'\varepsilon(n,\delta^*)^{\nu_\star-\epsilon} \quad \text{and} \quad \|\underbrace{\theta_n^{T_{\ell_\epsilon}+kt_{\ell_\epsilon}+t_{\ell_\epsilon}^{(1)}+t_{\ell_\epsilon}^{(2)}}}_{\theta_n^{T_{\ell_\epsilon}+(k+1)t_{\ell_\epsilon}}}\|_2 \leq \varepsilon(n,\delta^*)^{\nu_\star-\epsilon},$$

thereby establishing the induction hypothesis.

Proof of Theorem 14

We divide the proof into two subsections, corresponding to parts (a) and (b) of Theorem 14.

Proof of part (a)

We introduce the shorthands $\widetilde{\varepsilon}(n,\delta) = (\varepsilon(n,\delta))^{\frac{1}{1+\gamma}}$ and $T_{\mathrm{f}} = \frac{1}{(1+\gamma)} \cdot \frac{\log(\rho/\varepsilon(n,\delta))}{\log(1/\kappa)}$. Without loss of generality, we can assume that

$$\|\theta_n^t - \theta^\star\|_2 > \frac{(2-\kappa)}{(1-\kappa)}\widetilde{\varepsilon}(n,\delta) \quad \text{for all} \quad t \in \{0,\ldots,T_{\mathrm{f}}-1\}, \tag{6.53}$$

otherwise, the claim is immediate. Given the condition (6.53), we prove the following two claims:

$$\theta_n^t \in \mathbb{A}(\theta^\star, \widetilde{\varepsilon}(n,\delta), \rho) \quad \text{for all} \quad t \in \{0,\ldots,T_{\mathrm{f}}-1\}, \tag{6.54a}$$

$$\text{and} \quad \|\theta_n^{T_{\mathrm{f}}} - \theta^\star\|_2 \leq \frac{(2-\kappa)}{(1-\kappa)}\widetilde{\varepsilon}(n,\delta). \tag{6.54b}$$

The latter claim (6.54b) completes the proof of part (a) of the theorem.

Proof of claim (6.54a): With the condition (6.53) in hand, it remains to prove that $\|\theta_n^t - \theta^\star\|_2 \leq \rho$. The base case of $t = 0$ is immediate from the initialization

conditions. For the induction step, assuming $\theta_n^t \in \mathbb{A}(\theta^\star, \widetilde{\varepsilon}(n,\delta), \rho)$, we have

$$\begin{aligned}
\|\theta_n^{t+1} - \theta^\star\|_2 = \|F_n(\theta_n^t) - \theta^\star\|_2 &\leq \|F_n(\theta_n^t) - F(\theta_n^t)\|_2 + \|F(\theta_n^t) - \theta^\star\|_2 \\
&\overset{(i)}{\leq} \sup_{\theta \in \mathbb{A}(\theta^\star, \widetilde{\varepsilon}(n,\delta), \rho)} \|F_n(\theta) - F(\theta)\|_2 + \kappa \|\theta_n^t - \theta^\star\|_2 \\
&\overset{(ii)}{\leq} \varepsilon(n,\delta) \max\left\{ \frac{1}{\widetilde{\varepsilon}(n,\delta)^\gamma}, \rho \right\} + \kappa\rho \qquad (6.55) \\
&= \frac{\varepsilon(n,\delta)}{\widetilde{\varepsilon}(n,\delta)^\gamma} + \kappa\rho \\
&= \varepsilon(n,\delta)^{\frac{1}{1+\gamma}} + \kappa\rho \overset{(iii)}{\leq} \rho,
\end{aligned}$$

where inequality (i) follows from the induction hypothesis that $\theta_n^t \in \mathbb{A}(\theta^\star, \widetilde{\varepsilon}(n,\delta), \rho)$ and the fact that operator F is κ-contractive in the ball $\mathbb{B}(\theta^\star, \rho)$; inequality (ii) follows from the first inequality from condition (6.17a) that implies that $\widetilde{\varepsilon}(n,\delta) = \varepsilon(n,\delta)^{\frac{1}{1+\gamma}} \geq \widetilde{\rho}$ and then invoking the instability condition (6.12) with $r = \widetilde{\varepsilon}(n,\delta)$ and $\rho_2 = \rho$. Finally, the last inequality (iii) follows from the second bound of the condition (6.17a). The inductive step is thus established.

Proof of claim (6.54b)**:** We observe that

$$\begin{aligned}
\|\theta_n^{T_{\mathrm{f}}} - \theta^\star\|_2 = \|F_n(\theta_n^{T_{\mathrm{f}}-1}) - \theta^\star\|_2 &\leq \|F_n(\theta_n^{T_{\mathrm{f}}-1}) - F(\theta_n^{T_{\mathrm{f}}-1})\|_2 + \|F(\theta_n^{T_{\mathrm{f}}-1}) - \theta^\star\|_2 \\
&\overset{(i)}{\leq} \sup_{\theta \in \mathbb{A}(\theta^\star, \widetilde{\varepsilon}(n,\delta), \rho)} \|F_n(\theta_n^{T_{\mathrm{f}}-1}) - F(\theta_n^{T_{\mathrm{f}}-1})\|_2 \\
&\qquad + \kappa \|\theta_n^{T_{\mathrm{f}}-1} - \theta^\star\|_2 \\
&\overset{(ii)}{\leq} \varepsilon(n,\delta) \max\left\{ \frac{1}{\widetilde{\varepsilon}(n,\delta)^\gamma}, \rho \right\} + \kappa \|\theta_n^{T_{\mathrm{f}}-1}\|_2, \qquad (6.56)
\end{aligned}$$

where inequality (i) follows from our earlier claim (6.54a) and the κ-contractivity of the operator F on the ball $\mathbb{B}(\theta^\star, \rho)$; inequality (ii) follows from an argument similar to the one used to establish the inequality (6.55). Finally, recursing equation (6.56) T_{f} times, we obtain that

$$\begin{aligned}
\|\theta_n^{T_{\mathrm{f}}} - \theta^\star\|_2 &\leq \varepsilon(n,\delta) \max\left\{ \frac{1}{\widetilde{\varepsilon}(n,\delta)^\gamma}, \rho \right\} \cdot (1 + \kappa + \ldots + \kappa^{T_{\mathrm{f}}-1}) + \kappa^{T_{\mathrm{f}}} \|\theta_n^0 - \theta^\star\|_2 \\
&\leq \frac{\varepsilon(n,\delta)}{(1-\kappa)} \max\left\{ \frac{1}{\widetilde{\varepsilon}(n,\delta)^\gamma}, \rho \right\} + \kappa^{T_{\mathrm{f}}} \rho \\
&\leq \frac{\widetilde{\varepsilon}(n,\delta)}{(1-\kappa)} + \widetilde{\varepsilon}(n,\delta) = \frac{(2-\kappa)}{(1-\kappa)} \widetilde{\varepsilon}(n,\delta),
\end{aligned}$$

where the last step follows from the upper bound on iteration T_{f}, which in turn implies that $\kappa^{T_{\mathrm{f}}} \rho \leq \widetilde{\varepsilon}(n,\delta)$. The proof is now complete.

Proof of part (b)

The proof for Theorem 14(b) borrows ideas from the proof of Theorem 13 as well as the proof of part (a) of Theorem 14. We introduce the following definitions:

$$T_{\mathrm{s}} := [\varepsilon(n,\delta)]^{-\frac{1-|\gamma|\nu_\star}{1+\beta}}, \quad \text{where} \quad \nu_\star := \frac{\beta}{1+\beta-\gamma\beta}.$$

In order to prove the result (6.18b), we can, without loss of generality, assume that

$$\|\theta_n^t - \theta^\star\|_2 > 2[\varepsilon(n,\delta)]^{\nu_\star} \quad \text{for all} \quad t \in \{0,\ldots,T_{\mathrm{s}}-1\}, \tag{6.57}$$

and show that $\|\theta_n^{T_{\mathrm{s}}} - \theta^\star\|_2 \leq 2[\varepsilon(n,\delta)]^{\nu_\star}$. We only prove the result for $\theta^\star = 0$ as the more general case can be derived in a similar fashion.

In order to proceed further, we make use of a result similar to Lemma 31 adapted to the unstable case. Given two positive scalars $r_1 < r_2$, we define

$$\widetilde{\mathcal{T}}(r_1, r_2) := \frac{r_2 r_1^{|\gamma|}}{\varepsilon(n,\delta)}. \tag{6.58}$$

Lemma 32. *Suppose that the assumptions for part (b) of Theorem 14 hold. Further, suppose that the operator F_n satisfies $\|F_n^t(\theta)\|_2 \geq r_1$ for any point θ such that $\|\theta\|_2 \in [r_1, r_2]$ and for all $t \leq \widetilde{\mathcal{T}}(r_1, r_2)$, where $\widetilde{\rho} \leq r_1 \leq r_2 \leq \rho/2$. Then with probability at least $1-\delta$, we have*

$$\sup_{\theta \in \mathbb{A}(\theta^\star, r_1, r_2)} \|F^t(\theta) - F_n^t(\theta)\|_2 \leq t \cdot \frac{\varepsilon(n,\delta)}{r_1^{|\gamma|}} \quad \text{for all} \quad t \leq \widetilde{\mathcal{T}}(r_1, r_2). \tag{6.59}$$

See Section 6.8 for its proof.

We are now ready for the main argument. We have

$$\begin{aligned}
\|\theta_n^t\|_2 = \|F_n^t(\theta_n^0)\|_2 &\leq \|F^t(\theta_n^0)\|_2 + \|F^t(\theta_n^0) - F_n^t(\theta_n^0)\|_2 \\
&\overset{(i)}{\leq} \frac{1}{t^\beta} + \|F^t(\theta_n^0) - F_n^t(\theta_n^0)\|_2 \qquad (6.60) \\
&\overset{(ii)}{\leq} \frac{1}{t^\beta} + t \cdot \frac{\varepsilon(n,\delta)}{[\varepsilon(n,\delta)]^{\nu_\star|\gamma|}}, \quad \text{for all} \quad t \leq \widetilde{\mathcal{T}}([\varepsilon(n,\delta)]^{\nu_\star}, \rho), \qquad (6.61)
\end{aligned}$$

with probability at least $1-\delta$. Here, inequality (i) follows from the SLOW(β)-convergence condition (6.8) of the operator F along with the assumptions that $\theta^\star = 0$ and $\|\theta_n^0\|_2 \leq \rho$; inequality (ii) follows by applying Lemma 32 with $r_1 = [\varepsilon(n,\delta)]^{\nu_\star}$ and $r_2 = \rho$ in light of the condition (6.57). In the final bound (6.61), the first term decreases with iteration t while the second term increases with t. In order to trade off the two terms, we plug in $t = T_{\mathrm{s}} \overset{(\dagger)}{\leq} \widetilde{\mathcal{T}}([\varepsilon(n,\delta)]^{\nu_\star}, \rho)$ (where the inequality (†) holds due to the second bound in assumption (6.18a)), and perform some algebra to obtain that

$$\|\theta_n^{T_{\mathrm{s}}}\|_2 \leq \frac{1}{T_{\mathrm{s}}^\beta} + T_{\mathrm{s}} \frac{\varepsilon(n,\delta)}{[\varepsilon(n,\delta)]^{\nu_\star|\gamma|}} \leq 2[\varepsilon(n,\delta)]^{\nu_\star},$$

which yields the claim.

6.7 Tightness of general results

In this section, we construct a simple class of problems to demonstrate that the guarantees Theorems 13 and 14 in this chapter are unimprovable in general.

Gradient descent and (cubic-regularized) Newton's methods

In order to do show that the upper bounds in the theorems are tight, it suffices to consider the following class of optimization problems

$$\min_{\theta \in \mathbb{R}^d} f_n(\theta), \qquad \text{with } f_n(\theta) = \frac{\|\theta\|_2^p}{p} - \varepsilon \frac{\|\theta\|_2^q}{q}, \tag{6.62}$$

where $p \geq 4$ and $q \geq 2$ are even numbers such that $(p+1) > 2q$, and the scalar ε is a perturbation term. We imagine that the perturbation ε goes to zero as the sample size n increases, so that the relevant population function is given by $f(\theta) := \frac{\|\theta\|_2^p}{p}$. Note that the condition $p \geq 4$ ensures that it is weakly convex, and it has global optimum $\theta^\star = 0$. Simple calculation yields that the global minima θ_n^* of f_n satisfy $r_* := \|\theta_n^*\|_2 = \varepsilon^{\frac{1}{p-q}}$. In this section, we analyze the rate at which different optimization algorithms converge to a neighborhood of $\theta^\star = 0$.

We consider the behavior of three different algorithms: (a) gradient descent method, (b) Newton's method, and (c) cubic-regularized Newton's method (for $d = 1$), with updates generated by the operators $\mathrm{Q}_n^{\mathrm{GD}}$, $\mathrm{Q}_n^{\mathrm{NM}}$, and $\mathrm{Q}_n^{\mathrm{CNM}}$, respectively. These operators take the forms

$$\mathrm{Q}_n^{\mathrm{GD}}(\theta) = \theta - \eta \nabla f_n(\theta) = \theta - \eta \left(\|\theta\|_2^{p-2} - \varepsilon \|\theta\|_2^{q-2} \right) \theta, \tag{6.63a}$$

$$\mathrm{Q}_n^{\mathrm{NM}}(\theta) = \theta - \left[\nabla^2 f_n(\theta) \right]^{-1} \nabla f_n(\theta) = \frac{(p-2)\|\theta\|_2^{p-2} - (q-2)\varepsilon \|\theta\|_2^{q-2}}{(p-1)\|\theta\|_2^{p-2} - \varepsilon(q-1)\|\theta\|_2^{q-2}} \theta, \quad \text{and} \tag{6.63b}$$

$$\mathrm{Q}_n^{\mathrm{CNM}}(\theta) = \arg\min_{y \in \mathbb{R}} \left\{ \nabla f_n(\theta)(y-\theta) + \frac{1}{2} \nabla^2 f_n(\theta)(y-\theta)^2 + \frac{(p-1)(p-2)}{6} |y-\theta|^3 \right\}. \tag{6.63c}$$

Here $\eta > 0$ denotes the step-size of gradient descent algorithm.

Theoretical guarantees

In the next corollary, we state the tight statistical properties of these operators. We consider the gradient descent updates (6.63a) with step size $\eta \in \left(0, \frac{1}{2}\right]$ and the ordinary or cubic-regularized Newton updates (6.63b) with an initialization $\theta \in (c, 1]$ for some constant $c \in (0, 1)$.

Corollary 11. *There exist universal constants* (c_1, c_2, c_3) *such that*

$$\begin{aligned}
\|(\mathrm{Q}_n^{\mathrm{GD}})^t(\theta^0) - \theta^\star\|_2 &\asymp \varepsilon^{\frac{1}{p-q}} \quad \text{for all} \quad t \geq c_1 \varepsilon^{-\frac{p-2}{q-2}}, \\
\left|(\mathrm{Q}_n^{\mathrm{CNM}})^t(\theta^0) - \theta^\star\right| &\asymp \varepsilon^{\frac{1}{p-q}} \quad \text{for all} \quad t \geq c_2 \varepsilon^{-\frac{p-3}{p-1}}, \quad \text{and} \\
\|(\mathrm{Q}_n^{\mathrm{NM}})^t(\theta^0) - \theta^\star\|_2 &\asymp \varepsilon^{\frac{1}{p-q}} \quad \text{for all} \quad t \geq c_3 \log(\varepsilon^{-1}),
\end{aligned}$$

with probability $1 - \delta$*, where* $\theta^\star = 0$ *denotes the true parameter.*

Corollary 11 demonstrates that the general convergence results of operators in Theorems 13 and 14 are tight for the class of problems (6.62). The proof of Corollary 11 follows from arguments very similar to those in Section 6.9, and so we omit the details. Here, we only sketch the main argument leading to the results in this corollary. The convergence rates of updates from gradient descent and (cubic-regularized) Newton's methods can be studied based on a minimization problem with population version of f_n, which is given by

$$\min_{\theta \in \mathbb{R}^d} f(\theta), \quad \text{where } f(\theta) = \frac{\|\theta\|_2^p}{p}. \tag{6.64}$$

It is clear that the global minimum of the above objective function is $\theta^* = 0$. As an immediate consequence, the population level operators corresponding to the operators $\mathrm{Q}_n^{\mathrm{GD}}$, $\mathrm{Q}_n^{\mathrm{NM}}$, and $\mathrm{Q}_n^{\mathrm{CNM}}$ are given by

$$\mathrm{Q}^{\mathrm{GD}}(\theta) = \theta - \eta \nabla f(\theta) = \theta \left(1 - \eta \|\theta\|_2^{q-2}\right), \tag{6.65a}$$

$$\mathrm{Q}^{\mathrm{NM}}(\theta) = \theta - \left[\nabla^2 f(\theta)\right]^{-1} \nabla f(\theta) = \left(1 - \frac{1}{p-1}\right) \theta, \quad \text{and} \tag{6.65b}$$

$$\mathrm{Q}^{\mathrm{CNM}}(\theta) = \arg\min_{y \in \mathbb{R}} \left\{ \nabla f(\theta)(y - \theta) + \frac{1}{2} \nabla^2 f(\theta)(y - \theta)^2 + \frac{(p-1)(p-2)}{6} \|y - \theta\|_2^3 \right\}. \tag{6.65c}$$

Standard algebra with the update equations (6.65a)-(6.65c) yields the following properties with the population-level operators:

($\widehat{\mathrm{P}}$1) the operator Q^{GD} is $\texttt{SLOW}(\frac{1}{q-2})$-convergent over the ball $\mathbb{B}(\theta^\star, 1)$ for a sufficiently small value of the step-size $\eta > 0$, meaning that $\|(\mathrm{Q}^{\mathrm{GD}})^t(\theta^0)\|_2 \leq \frac{c}{t^{\frac{1}{q-2}}}$,

($\widehat{\mathrm{P}}$2) the operator Q^{NM} is $\texttt{FAST}(\frac{p}{p-1})$-convergent towards $\theta^\star = 0$, and

($\widehat{\mathrm{P}}$3) the operator $\mathrm{Q}^{\mathrm{CNM}}$ can be shown to be $\texttt{SLOW}(\frac{2}{p-3})$-convergent over the ball $\mathbb{B}(\theta^\star, 1)$, meaning that $\|(\mathrm{Q}^{\mathrm{CNM}})^t(\theta^0)\|_2 \leq \frac{c_1}{t^{\frac{2}{p-3}}}$.

Moving to the (in)-stability of sample-level operators, we can verify that:

($\widehat{\text{S1}}$) the operator $\mathrm{Q}_n^{\mathrm{GD}}$ is $\texttt{STA}(\varepsilon, q-1)$-stable over the Euclidean ball $\mathbb{B}(\theta^\star, 1)$ with noise function ε, meaning that

$$\sup_{\theta \in \mathbb{B}(\theta^\star, r)} \|\mathrm{Q}_n^{\mathrm{GD}}(\theta) - \mathrm{Q}^{\mathrm{GD}}(\theta)\|_2 \leq c \cdot r^{q-1}\varepsilon \qquad \text{for } r \in [0,1],$$

($\widehat{\text{S2}}$) the operator $\mathrm{Q}_n^{\mathrm{NM}}$ is $\texttt{UNS}(-p+q+1)$-unstable over the annulus $\mathbb{A}(\theta^\star, c_1 r_*, 1)$ with noise function ε, namely, we have

$$\sup_{\theta \in \mathbb{A}(\theta^\star, r, 1)} \|\mathrm{Q}_n^{\mathrm{NM}}(\theta) - \mathrm{Q}^{\mathrm{NM}}(\theta)\|_2 \leq c \cdot \max\left\{\frac{1}{r^{p-q-1}}\varepsilon, 1\right\} \quad \text{for } r \in [c_1 r_*, 1], \quad \text{and}$$

($\widehat{\text{S3}}$) the operator $\mathrm{Q}_n^{\mathrm{CNM}}$ is $\texttt{UNS}(-\frac{p+1}{2}+q)$-unstable over the annulus $\mathbb{A}(\theta^\star, c_2 r_*, 1)$, which means that

$$\sup_{\theta \in \mathbb{A}(\theta^\star, r, 1)} \|\mathrm{Q}_n^{\mathrm{CNM}}(\theta) - \mathrm{Q}^{\mathrm{CNM}}(\theta)\|_2 \leq c \cdot \max\left\{\frac{1}{r^{(p+1)/2-q}}\varepsilon, 1\right\} \quad \text{for } r \in [c_2 r_*, 1].$$

Finally, we can show that our sequences of updates from gradient descent and (cubic-regularized) Newton's methods always converge to the global minima θ_n^* of f_n. Additionally, we also have

$$\|\mathrm{Q}_n^{\mathrm{GD}}(\theta)\|_2 \geq r_*, \quad \|\mathrm{Q}_n^{\mathrm{NM}}(\theta)\|_2 \geq r_*, \text{ and } \|\mathrm{Q}_n^{\mathrm{CNM}}(\theta)\|_2 \geq r_*$$

for all $\|\theta\|_2 \geq r_*$. It means that Assumption (D) is satisfied by these sequences of updates. In summary, for the problem (6.62), the gradient descent method is a slow converging stable method and the cubic-regularized Newton's method is a slow converging unstable method. Furthermore, the Newton's method is a fast converging unstable method.

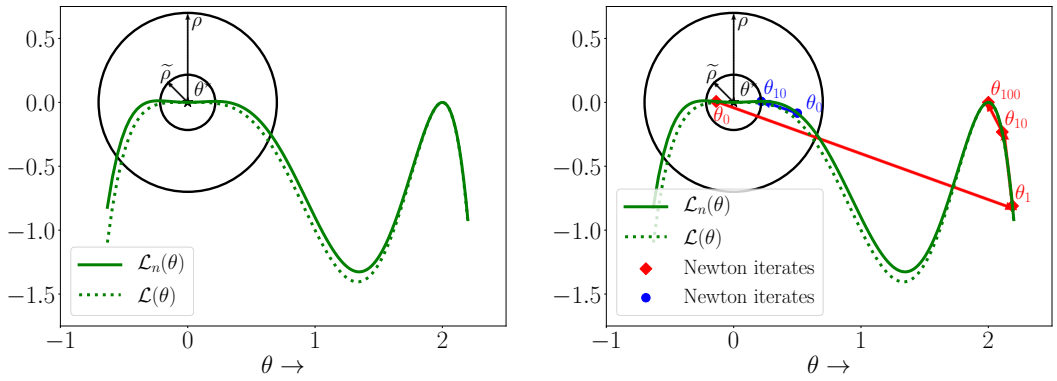
Undesirable behavior of unstable operators

In this section, we prove that the minimum over all iterates $k \in \{1, 2, \ldots, t\}$ in Theorem 14 is necessary. In particular, we consider the following example

$$\mathcal{L}(\theta) = -\theta^4(\theta-2)^2 \quad \text{and} \quad \mathcal{L}_n(\theta) = -\left(\theta^4 - \frac{\theta^2}{\sqrt{n}}\right)(\theta-2)^2.$$

We let F and F_n denote the operators corresponding to the Newton's method as applied to the functions $\mathcal{L}$ and $\mathcal{L}_n$, respectively (Consequently, the operator F has three fixed points). Following some simple algebra, it can be verified there are universal constants (c_1, c_2) such that that the operators F and F_n defined above satisfy the conditions of Theorem 14 (a) with $\theta^\star = 0$ for some $\kappa < 1$, $\gamma = -1, \varepsilon(n, \delta) = n^{-\frac{1}{2}}, \widetilde{\rho} = c_1 n^{-\frac{1}{4}}$ and $\rho = c_2$. In panel (a) of Figure 6.4, we plot the two functions $\mathcal{L}$ and $\mathcal{L}_n$ and illustrate the radii $\widetilde{\rho}, \rho$ (for a fixed n). Some additional algebra shows that there

exists $\theta_n^0 \in \mathbb{B}(\theta^\star, \widetilde{\rho})$ such that the iterates corresponding to the sequence $\theta_n^{t+1} = F_n(\theta_n^t)$ satisfy $\|\theta_n^t - \theta^\star\|_2 \geq 1 \gg n^{-\frac{1}{4}}$ for all iterations $t = 1, 2, \ldots$. See, in particular, the red (diamond) iterates in panel (b) of Figure 6.4 which are generated with a starting point $\theta_n^0 = c_3 n^{-\frac{1}{4}}$ (which is below the controlled instability threshold $\widetilde{\rho}$). Clearly, we see that the first iterate produced by Newton's method escapes the local basin of attraction and the subsequent iterates converge to a very different fixed point of the function $\mathcal{L}_n$. On the other hand, when the Newton's method is initialized in the annulus $\mathbb{A}(\theta^\star, \widetilde{\rho}, \rho)$, the sequence θ_n^t (blue circles) converges quickly to the vicinity of $\theta^\star$ as guaranteed by Theorem 14. Furthermore, the iterates do not escape this local neighborhood. Via this simple example, we have demonstrated that if no



(a) (b)

Figure 6.4. Instability of Newton's method for the example discussed above (figure best viewed in color). When the algorithm is initialized too close to $\theta^\star$ (red diamonds), the instability of Newton's method forces the iterates to jump too far away from $\theta^\star$ and converge to another fixed point. On the other hand, if the initial point is initialized in the annulus $\mathbb{A}(\theta^\star, \widetilde{\rho}, \rho)$, the Newton iterates (blue circles), do not leave this annulus and converge monotonically to a small neighborhood of $\theta^\star$.

further regularity assumptions are made, then starting an unstable algorithm from a point that is too close to $\theta^\star$, the subsequent iterates can be quite far from the true parameter.

6.8 Proofs of auxiliary results

In this section, we collect the proofs of Lemmas 31 and 32 that are central to the proofs of our main theorems.

Proof of Lemma 31

We fix a radius $r \in \mathcal{R}$. Our proof is based on the following auxiliary claim: conditioned on the event $\mathcal{E}$ from equation (6.39), we have

$$\sup_{\theta \in \mathbb{B}(\theta^\star, r)} \|F_n^t(\theta)\|_2 \leq 2r \quad \text{for all} \quad t \leq \widetilde{\mathcal{T}}(r) = \frac{r^{1-\gamma}}{2^\gamma c_2 \varepsilon(n, \delta^*)}. \tag{6.66}$$

Taking this claim as given for the moment, we now establish the bound (6.42) claimed in the lemma. We do so via induction on the iteration $t \in \{0, 1, \ldots, \widetilde{\mathcal{T}}(r)\}$. Note that the base-case $t = 0$ holds trivially, since $\|F^0(\theta) - F_n^0(\theta)\|_2 = \|\theta - \theta\|_2 = 0$. Given the induction hypothesis for t, we establish the claim for $t' = t + 1$. For any $\theta \in \mathbb{B}(\theta^\star, r)$, we have

$$\begin{aligned}
\|F^{t'}(\theta) - F_n^{t'}(\theta)\|_2 &= \|F^{t+1}(\theta) - F_n^{t+1}(\theta)\|_2 \\
&\leq \|F(F^t(\theta)) - F(F_n^t(\theta))\|_2 + \|F(F_n^t(\theta)) - F_n(F_n^t(\theta))\|_2 \\
&\overset{(i)}{\leq} \|F^t(\theta) - F_n^t(\theta)\|_2 + \sup_{\widetilde{\theta} \in \mathbb{B}(\theta^\star, 2r)} \|F(\widetilde{\theta}) - F_n(\widetilde{\theta})\|_2 \\
&\overset{(ii)}{\leq} \sup_{\theta \in \mathbb{B}(\theta^\star, r)} \|F^t(\theta) - F_n^t(\theta)\|_2 + c_2 (2r)^\gamma \varepsilon(n, \delta^*) \\
&\overset{(iii)}{\leq} c_2 (2r)^\gamma \varepsilon(n, \delta^*) t + c_2 (2r)^\gamma \varepsilon(n, \delta^*) = (t+1) c_2 (2r)^\gamma \varepsilon(n, \delta^*).
\end{aligned} \tag{6.67}$$

In the above sequence of inequalities, we have made use of the following facts. In step (i), we have used the 1-Lipschitzness (6.6) of the operator F for the first term and the bound (6.66) on $F_n^t(\theta)$ for the second term. In order to establish step (ii), we have used the fact that $\theta \in \mathbb{B}(\theta^\star, r)$ for the first term, while for the second term we have invoked the definition of the event $\mathcal{E}$ in equation (6.39) with radius $2r$ (note that $2\mathcal{R} \subset \mathcal{R}'$ and the event $\mathcal{E}$ is defined for all $r' \in \mathcal{R}'$). Finally step (iii) follows directly from the induction hypothesis. Noting that the bound (6.66) holds for any $t \leq \widetilde{\mathcal{T}}(r)$ and taking supremum over $\theta \in \mathbb{B}(\theta^\star, r)$ on the LHS of equation (6.72), we obtain the desired proof of the inductive step.

Proof of claim (6.66): We establish the claim (6.66) by proving the following stronger result: For any fixed $r \in \mathcal{R}$, and any $\theta \in \mathbb{B}(\theta^\star, r)$, we have

$$\|F_n^t(\theta)\|_2 \leq r + c_2 (2r)^\gamma \varepsilon(n, \delta^*) \cdot t \quad \text{for all iterations } t = 0, 1, \ldots, \widetilde{\mathcal{T}}(r). \tag{6.68}$$

We note that the claim (6.66) is a direct application of this result along with the definition $\widetilde{\mathcal{T}}(r) = \frac{r^{1-\gamma}}{2^\gamma c_2 \varepsilon(n, \delta^*)}$. We now use an induction argument on the iteration t (similar to the ones used in the paragraph above) to establish the claim (6.68). The base-case $t = 0$ holds trivially. Let us assume that $\|F_n^t(\theta)\|_2 \leq r + c_2 (2r)^\gamma \varepsilon(n, \delta^*) \cdot t$

and establish the claim (6.68) for $t' = t+1$. Note that since $t \leq \widetilde{\mathcal{T}}(r)$, this assumption trivially yields that $\|F_n^t(\theta)\|_2 \leq 2r$. We have

$$\begin{aligned}\|F_n^{t+1}(\theta)\|_2 &\leq \|F(F_n^t(\theta))\|_2 + \|F(F_n^t(\theta)) - F_n(F_n^t(\theta))\|_2 \\ &\overset{(i)}{\leq} \|F_n^t(\theta)\|_2 + \sup_{\widetilde{\theta} \in \mathbb{B}(\theta^\star, 2r)} \|F(\widetilde{\theta}) - F_n(\widetilde{\theta})\|_2 \\ &\overset{(ii)}{\leq} (r + c_2(2r)^\gamma \varepsilon(n, \delta^*) \cdot t) + c_2(2r)^\gamma \varepsilon(n, \delta^*) \\ &= r + c_2(2r)^\gamma \varepsilon(n, \delta^*)(t+1),\end{aligned}$$

where in step (i), we have used the 1-Lipschitzness (6.6) of the operator F for the first term and the observation that $\|F_n^t(\theta)\|_2 \leq 2r$ for the second term. On the other hand, in step (ii), we have used the induction hypothesis to bound the first term, and invoked the definition of the event $\mathcal{E}$ in equation (6.39) with radius $2r$ to bound the second term. Taking supremum over $\theta \in \mathbb{B}(\theta^\star, r)$ completes the proof.

Proof of claim (6.43): Combining the relation $\alpha_\ell = \nu_\star(1 - \nu^\ell)$ with the two inequalities in equation (6.43), we find that it suffices to prove the following two bounds:

$$\varepsilon(n, \delta^*)^{-\frac{\beta \nu^\ell}{1+\beta}} \geq (2^\gamma c_2)^{\frac{\beta}{1+\beta}} \quad \text{and} \quad \varepsilon(n, \delta^*)^{-\frac{\beta \nu^{\ell+1}}{1+\beta}} \geq (2^\gamma c_2)^{\frac{\beta}{1+\beta}} (c')^{-\frac{\beta}{\nu_\star(1+\beta)}}. \tag{6.69}$$

Observe that $\alpha_\ell \leq \nu_\star - \epsilon/4$; consequently, we find that $1/\nu^\ell \leq 4\nu_\star/\epsilon$ for all $\ell \leq \ell_\epsilon$. Finally, invoking assumption (6.14) we find that

$$\varepsilon(n, \delta^*) \leq \frac{1}{(2^\gamma c_2)^{\frac{4\nu_\star}{\epsilon}} \cdot \max\left\{1, (c')^{\frac{4}{\epsilon}}\right\}}. \tag{6.70}$$

The rest of the proof follows by noting that the upper bound (6.70) implies the bounds in equation (6.69).

Proof of Lemma 32

Fix an arbitrary pair of radii $r_1, r_2 \in \mathcal{R}$. Our proof is based on the following intermediate claim

$$\|F_n^t(\theta)\| \leq 2r_2 \quad \text{for all } t \leq \widetilde{\mathcal{T}}(r_1, r_2). \tag{6.71}$$

We prove this claim at the end of this section. Assuming that this claim is given at the moment, we now establish the bound (6.59) claimed in the lemma. We do so by using induction on the iteration $t \in \{0, 1, \ldots, \widetilde{\mathcal{T}}(r_1, r_2)\}$ where we note that the base-case $t = 0$ holds trivially, since $\|F^0(\theta) - F_n^0(\theta)\|_2 = \|\theta - \theta\|_2 = 0$. Turning to

the induction step (with $t' = t+1$), for any θ with $\|\theta\| \in [r_1, r_2]$, we have

$$\begin{aligned}
\|F^{t'}(\theta) - F_n^{t'}(\theta)\|_2 &= \|F^{t+1}(\theta) - F_n^{t+1}(\theta)\|_2 \\
&\leq \|F(F^t(\theta)) - F(F_n^t(\theta))\|_2 + \|F(F_n^t(\theta)) - F_n(F_n^t(\theta))\|_2 \\
&\overset{(i)}{\leq} \|F^t(\theta) - F_n^t(\theta)\|_2 + \sup_{r_1 \leq \|\widetilde{\theta}\| \leq 2r_2} \|F(\widetilde{\theta}) - F_n(\widetilde{\theta})\|_2 \\
&\overset{(ii)}{\leq} \sup_{r_1 \leq \|\theta\| \leq 2r_2} \|F^t(\theta) - F_n^t(\theta)\|_2 + \frac{\varepsilon(n, \delta^*)}{r_1^{|\gamma|}} \\
&\overset{(iii)}{\leq} t\frac{\varepsilon(n, \delta^*)}{r_1^{|\gamma|}} + \frac{\varepsilon(n, \delta^*)}{r_1^{|\gamma|}} = (t+1) \cdot \frac{\varepsilon(n, \delta^*)}{r_1^{|\gamma|}}.
\end{aligned} \tag{6.72}$$

In step (i), we have used the 1-Lipschitzness (6.6) of the operator F for the first term and the upper bound (6.66) on $F_n^t(\theta)$ for the second term. In step (ii), the upper bound for the first term follows from the sequence of inequalities

$$\widetilde{\rho} \leq r_1 \leq \|\theta\| \leq r_2 \leq 2r_2 \leq \rho,$$

whereas for the second term we have invoked the bound $\|\widetilde{\theta}\| := F_n^{t'}(\theta) \leq 2r_2$ (6.71) and applied the instability condition (6.12). Finally, step (iii) follows from a direct application of the induction hypothesis. Note that the bound (6.66) holds for any $t \leq \widetilde{\mathcal{T}}(r)$. By taking supremum over $\theta \in \mathbb{B}(\theta^\star, r)$ on the LHS of equation (6.72), we obtain the desired proof of the inductive step.

Proof of bound (6.71)**:** We use an inductive argument to prove the following bound:

$$\|F_n^t(\theta)\| \leq t \cdot \frac{\varepsilon(n, \delta^*)}{r_1^{|\gamma|}} + r_2 \quad \text{for all} \quad 1 \leq t \leq \widetilde{\mathcal{T}}(r_1, r_2), \tag{6.73}$$

which immediately implies the claim (6.71) once we plug in the definition of $\widetilde{\mathcal{T}}$ (6.58).

For the base-case $t = 0$, invoking the properties of the operators F and F_n we have

$$\begin{aligned}
\|F_n(\theta)\|_2 \leq \|F_n(\theta) - F(\theta)\|_2 + \|F(\theta)\|_2 &\overset{(i)}{\leq} \sup_{r_1 \leq \|\theta\| \leq r_2} \|F_n(\theta) - F(\theta)\|_2 + \|\theta\|_2 \\
&\overset{(ii)}{\leq} \frac{\varepsilon(n, \delta^*)}{r_1^{|\gamma|}} + r_2,
\end{aligned}$$

where step (i) follows since $\|\theta\|_2 \in [r_1, r_2]$ and the operator F is 1-Lipschitz, and step (ii) follows from the instability condition (6.12). This proves the base case of the induction hypothesis (6.73).

Now we prove the inductive step. In particular, we assume that the induction hypothesis (6.73) holds for $t \leq \widetilde{\mathcal{T}}(r_1, r_2) - 1$ and show that the upper bound (6.73)

holds for $t' = t+1$. Towards this end, unwrapping the expression for $\|F_n^{t+1}(\theta)\|_2$ we have

$$\begin{aligned}\|F_n^{t'}(\theta)\|_2 &\leq \|F_n^{t+1}(\theta) - F(F_n^t(\theta))\|_2 + \|F(F_n^t(\theta))\|_2 \\ &\overset{(iii)}{\leq} \sup_{r_1 \leq \|\theta\|_2 \leq 2r_2} \|F_n(\theta) - F(\theta)\|_2 + \|F_n^t(\theta)\|_2 \\ &\overset{(iv)}{\leq} \frac{\varepsilon(n,\delta^*)}{r_1^{|\gamma|}} + t\frac{\varepsilon(n,\delta^*)}{r_1^{|\gamma|}} + r_2 \\ &= (t+1)\frac{\varepsilon(n,\delta^*)}{r_1^{|\gamma|}} + r_2.\end{aligned}$$

Here, step (iii) follows from the fact that $\|F_n^t(\theta)\|_2 \geq r_1$ and the LL(ρ) condition (6.6); step (iv) stems from the instability condition (6.12) and the induction hypothesis. This completes the proof of the intermediate claim (6.73).

6.9 Proofs of corollaries

We now collect the proofs of several corollaries stated in the chapter. As a high-level summary, our analysis in all three examples in Section 6.4 involves applying Theorem 13 to analyze gradient descent/ascent and EM, both of which are stable algorithms and exhibit slow convergence for the considered examples. We invoke Theorem 14(b) to characterize the cubic-regularized Newton algorithm, a slowly convergent and unstable algorithm. Finally, the analysis of Newton's method in all the examples relies on Theorem 14(a). Appendices 6.9 and 6.9 are devoted to the proofs of Corollaries 8 and 9, respectively. We then prove Corollary 10 in Section 6.9. In this section, the values of universal constants (e.g., c, c' etc.) can change from line-to-line.

Proof of Corollary 8

In this section, we demonstrate the convergence and stability of the gradient and Newton methods. The operators for the gradient method and Newton's method take the following forms

$$\mathrm{M}^{\mathrm{GA}}(\theta) = \theta + \eta\bar{\mathcal{L}}'(\theta), \quad \text{and} \quad \mathrm{M}_n^{\mathrm{GA}}(\theta) = \theta + \eta\bar{\mathcal{L}}_n'(\theta), \tag{6.74a}$$

$$\mathrm{M}^{\mathrm{NM}}(\theta) = \theta - \left[\frac{\bar{\mathcal{L}}'(\theta)}{\bar{\mathcal{L}}''(\theta)}\right], \quad \text{and} \quad \mathrm{M}_n^{\mathrm{NM}}(\theta) = \theta - \left[\frac{\bar{\mathcal{L}}_n'(\theta)}{\bar{\mathcal{L}}_n''(\theta)}\right]. \tag{6.74b}$$

Proofs for the gradient operators

In lieu of the discussion around Corollary 8 it remains to establish that (a) the operator M^{GA} exhibits a slow convergence condition $\texttt{SLOW}(\frac{1}{2})$ over the Euclidean ball $\mathbb{B}(\theta^\star, 1/2)$

and (b) the operator $\mathrm{M}_n^{\mathrm{GA}}$ satisfies a stability condition $\texttt{STA}(1)$ over the Euclidean ball $\mathbb{B}(\theta^\star, 1/2)$ with noise function $\varepsilon(n,\delta) = \sqrt{\log(1/\delta)/n}$ when $n \geq c\log(1/\delta)$ for some universal constant $c > 0$.

Slow convergence of M^{GA}: Direct computation with the gradient of population log-likelihood function $\bar{\mathcal{L}}$ leads to

$$\bar{\mathcal{L}}'(\theta) := \frac{\theta}{2(\theta^2+1)(2\sqrt{1+\theta^2}-1)} - \frac{\theta}{2} \tag{6.75}$$
$$\implies \quad \mathrm{M}^{\mathrm{GA}}(\theta) = \theta\left[1 - \eta\left(\frac{1}{2} - \frac{1}{2(\theta^2+1)(2\sqrt{1+\theta^2}-1)}\right)\right].$$

Noting that the fixed point of the population operator is $\theta^\star = 0$ and that $\eta \leq 8/3$, we find that

$$\begin{aligned}\left|\mathrm{M}^{\mathrm{GA}}(\theta) - \theta^\star\right| &= |\theta|\left[1 - \eta\left(\frac{1}{2} - \frac{1}{2(\theta^2+1)(2\sqrt{1+\theta^2}-1)}\right)\right] \\ &\leq |\theta|\left[1 - \eta\left(\frac{1}{2} - \frac{1}{2(\theta^2+1)}\right)\right] \\ &\leq |\theta|\left(1 - \frac{\eta\theta^2}{4}\right) \quad \text{for all } |\theta| \in [0, 1/2].\end{aligned}$$

Thus the population operator M^{GA} satisfies a slow convergence condition $\texttt{SLOW}(\frac{1}{2})$ over the ball $\mathbb{B}(\theta^\star, 1/2)$.

Stability of the sample operator $\mathrm{M}_n^{\mathrm{GA}}$: We have

$$\begin{aligned}\left|\mathrm{M}_n^{\mathrm{GA}}(\theta) - \mathrm{M}^{\mathrm{GA}}(\theta)\right| &= \eta\left|\nabla\bar{\mathcal{L}}(\theta) - \nabla\bar{\mathcal{L}}_n(\theta)\right| \\ &\leq \eta\Bigg(\left|\frac{\theta}{2(\theta^2+1)\left(2\sqrt{1+\theta^2}-1\right)}\left(\frac{2}{n}\sum_{i=1}^n(1-R_i) - 1\right)\right| \\ &\qquad + \left|\theta\left(\frac{1}{2} - \frac{1}{n}\sum_{i=1}^n R_i Y_i^2\right)\right|\Bigg).\end{aligned}$$

Recall that, $R_1, \ldots, R_n$ are i.i.d. samples from Bernoulli distribution with probability $1/2$. Invoking Hoeffding's inequality yields that

$$\left|\frac{2}{n}\sum_{i=1}^n(1-R_i) - 1\right| \leq c\sqrt{\frac{\log(1/\delta)}{n}}, \tag{6.76}$$

with probability at least $1-\delta$. Additionally, as $Y_1, \ldots, Y_n$ are i.i.d. samples from standard Gaussian distribution $\mathcal{N}(0,1)$ and $R_1, \ldots, R_n$ are independent of $Y_1, \ldots, Y_n$,

by following the same argument as that in the proof of Lemma 1 from the paper [Dwi+20b], we can demonstrate that

$$\left|\frac{1}{n}\sum_{i=1}^{n} R_i Y_i^2 - \frac{1}{2}\right| \leq c_1\sqrt{\frac{\log(1/\delta)}{n}}, \tag{6.77}$$

as long as the sample size $n \geq c_2 \log(1/\delta)$ with probability at least $1-\delta$ where c_1 and c_2 are some universal constants.

Combining the inequalities (6.76) and (6.77) yields the following bound

$$\begin{aligned}\sup_{\theta\in\mathbb{B}(\theta^\star,r)}\left|\mathrm{M}_n^{\mathrm{GA}}(\theta) - \mathrm{M}^{\mathrm{GA}}(\theta)\right| &\leq c_3\sqrt{\frac{\log(1/\delta)}{n}}\sup_{\theta\in\mathbb{B}(\theta^\star,r)}\left(\frac{|\theta|}{2(\theta^2+1)\left(2\sqrt{1+\theta^2}-1\right)} + |\theta|\right)\\ &\leq \frac{3c_3 r}{2},\end{aligned}$$

with probability at least $1-2\delta$ for any $r > 0$. Here, the second inequality in the above display follows from the fact that $(\theta^2+1)\left(2\sqrt{1+\theta^2}-1\right) \geq 1$ for all $\theta \in \mathbb{R}$. Thus, the sample-level operator $\mathrm{M}_n^{\mathrm{GA}}$ is $\mathtt{STA}(1)$-stable over the Euclidean ball $\mathbb{B}(\theta^\star, 1/2)$ with noise function $\varepsilon(n,\delta) = \sqrt{\log(1/\delta)/n}$ when $n \geq c\log(1/\delta)$ for some universal constant $c > 0$.

Proof for the Newton operators

Similar to the proof for Newton operators in over-specified Gaussian mixtures (see Section 6.9), we first verify the geometric convergence of population operator M^{NM} and the instability condition of sample operator $\mathrm{M}_n^{\mathrm{NM}}$. Then, we validate Assumption (D) by showing that the Newton updates are monotone decreasing and satisfy the following lower bound

$$\left|\mathrm{M}^{\mathrm{NM}}(\theta)\right| \geq |\theta_n^*|, \tag{6.78}$$

for all $|\theta| \in [|\theta_n^*|, 1/2]$ for any global maxima θ_n^* of the sample log-likelihood function $\bar{\mathcal{L}}_n$ in equation (6.21).

Geometric convergence of M^{NM}: We can verify that $\bar{\mathcal{L}}''(\theta) < 0$ for all $\theta \in \mathbb{R}$. Additionally, we have the following equation

$$\left|\mathrm{M}^{\mathrm{NM}}(\theta) - \theta^\star\right| = |\theta - \theta^\star|\,\frac{\theta^2 T_2(\theta)}{T_1(\theta) + \theta^2 T_2(\theta)},$$

where the functions T_1 and T_2 are defined as

$$\begin{aligned}T_1(\theta) &:= \frac{1}{2} - \frac{1}{2(\theta^2+1)(2\sqrt{\theta^2+1}-1)}, \quad \text{and}\\ T_2(\theta) &:= \frac{1}{2(\theta^2+1)^2(2\sqrt{\theta^2+1}-1)}\left(3 + \frac{1}{2\sqrt{\theta^2+1}-1}\right).\end{aligned}$$

From the earlier proof argument for slow convergence of M^{GA}, we have $T_1(\theta) \geq \frac{\theta^2}{8}$ for all $|\theta| \in [0, 1/2]$. Given the above lower bound of T_1, we directly obtain that

$$\left|\mathrm{M}^{\mathrm{NM}}(\theta) - \theta^\star\right| \leq |\theta - \theta^\star| \, \frac{T_2(\theta)}{1/8 + T_2(\theta)} \leq |\theta - \theta^\star| \, \frac{T_2(1/2)}{1/8 + T_2(1/2)} \leq \frac{4}{5} \, |\theta - \theta^\star| \, ,$$

for all $|\theta| \in [0, 1/2]$ where the last inequality is due to the fact that $T_2(\theta)/(c + T_2(\theta))$ achieves its maximum value at $|\theta| = 1/2$. Therefore, the population operator M^{NM} is $\texttt{FAST}(4/5)$-convergent on the ball $\mathbb{B}(\theta^*, 1/2)$.

Instability of the sample Newton operator $\mathrm{M}_n^{\mathrm{NM}}$: Given the formulations of population operator M^{NM} and sample operator $\mathrm{M}_n^{\mathrm{NM}}$ from Newton's method, we have the following inequality

$$\left|\mathrm{M}_n^{\mathrm{NM}}(\theta) - \mathrm{M}^{\mathrm{NM}}(\theta)\right| \leq \underbrace{\left|\frac{\bar{\mathcal{L}}'(\theta) - \bar{\mathcal{L}}'_n(\theta)}{\bar{\mathcal{L}}''(\theta)}\right|}_{:=J_1} + \underbrace{\left|\bar{\mathcal{L}}'_n(\theta) \left(\frac{1}{\bar{\mathcal{L}}''(\theta)} - \frac{1}{\bar{\mathcal{L}}''_n(\theta)}\right)\right|}_{:=J_2}.$$

We claim the following upper bounds of J_1 and J_2:

$$J_1 \leq c_1 \frac{1}{|\theta|} \sqrt{\frac{\log(1/\delta)}{n}}, \tag{6.80}$$

with probability at least $1 - 2\delta$ as long as $|\theta| \in [0, 1/2]$ and $n \geq c' \log(1/\delta)$, and

$$J_2 \leq c_2 \cdot \frac{1}{|\theta|} \sqrt{\frac{\log(1/\delta)}{n}}, \tag{6.81}$$

with probability at least $1 - 6\delta$ when $|\theta| \geq \sqrt{2c} \, (\log(1/\delta)/n)^{1/4}$.

With the upper bounds (6.80) and (6.81) of J_1 and J_2 respectively, we arrive at the following inequality

$$\left|\mathrm{M}_n^{\mathrm{NM}}(\theta) - \mathrm{M}^{\mathrm{NM}}(\theta)\right| \leq c'' \, |\theta|^{-1} \sqrt{\log(1/\delta)/n},$$

with probability at least $1 - 8\delta$ as long as $\sqrt{2c} \, (\log(1/\delta)/n)^{1/4} \leq |\theta| \leq 1/2$. As a consequence, the sample operator $\mathrm{M}_n^{\mathrm{NM}}$ satisfies instability condition $\texttt{UNS}(1)$ over the annulus $\mathbb{A}(\theta^\star, \sqrt{2c} \, (\log(1/\delta)/n)^{1/4}, 1/2)$ with noise function $\varepsilon(n, \delta) = \sqrt{\frac{\log(1/\delta)}{n}}$ as long as $n \geq c' \log(1/\delta)$.

Proof for the upper bound of J_1: When $n \geq c' \log(1/\delta)$, we can validate that

$$\left|\bar{\mathcal{L}}'(\theta) - \bar{\mathcal{L}}'_n(\theta)\right| \leq c \, |\theta| \sqrt{\frac{\log(1/\delta)}{n}},$$

for any $|\theta| \in [0, 1/2]$ with probability at least $1 - 2\delta$ where c and c' are some universal constants. Furthermore, based on the computations in Section 6.9, we find that

$$\left|\bar{\mathcal{L}}''(\theta)\right| = T_1(\theta) + \theta^2 T_2(\theta) \geq \frac{\theta^2}{8} + \theta^2 T_2(1/2) \geq \frac{11\theta^2}{32}, \tag{6.82}$$

for any $|\theta| \in [0, 1/2]$. Combining the previous inequalities, we have the following upper bound with J_1:

$$J_1 \leq c_1 \frac{1}{|\theta|} \sqrt{\frac{\log(1/\delta)}{n}},$$

with probability at least $1 - 2\delta$ as long as $|\theta| \in [0, 1/2]$ and $n \geq c' \log(1/\delta)$.

Proof for the upper bound of J_2: In order to derive an upper bound for J_2, we make use of the following bounds:

$$\left|\bar{\mathcal{L}}_n'(\theta)\right| \leq c_1 \left(|\theta| \sqrt{\frac{\log(1/\delta)}{n}} + |\theta|^3 \right), \tag{6.83a}$$

$$\left|\bar{\mathcal{L}}_n''(\theta) - \bar{\mathcal{L}}''(\theta)\right| \leq c_2 \sqrt{\frac{\log(1/\delta)}{n}}, \tag{6.83b}$$

$$\left|\bar{\mathcal{L}}_n''(\theta)\right| \geq c_3 \left(\theta^2 - c \cdot \sqrt{\frac{\log(1/\delta)}{n}} \right), \tag{6.83c}$$

for all $|\theta| \in [0, 1/2]$ with probability at least $1 - 2\delta$ when $n \geq c' \log(1/\delta)$. Here, c, c_1, c_2, c_3 in the above bounds are universal constants independent of δ.

Deferring the proofs of these claims to later, we now proceed to give an upper bound for J_2 based on the given bounds in the above display. In particular, from the formulation of J_2, we achieve that

$$J_2 \leq \frac{32 c_1 c_2}{11 c_3} \left(|\theta| \sqrt{\frac{\log(1/\delta)}{n}} + |\theta|^3 \right) \frac{\sqrt{\frac{\log(1/\delta)}{n}}}{\theta^2 \left(\theta^2 - c\sqrt{\frac{\log(1/\delta)}{n}} \right)} \leq C \cdot \frac{1}{|\theta|} \sqrt{\frac{\log(1/\delta)}{n}}$$

with probability at least $1 - 6\delta$ when $|\theta| \geq \sqrt{2c} \left(\log(1/\delta)/n\right)^{1/4}$ where C is some universal constant. Here, the last inequality is due to $|\theta| \sqrt{\frac{\log(1/\delta)}{n}} + |\theta|^3 \leq |\theta|^3 \left(1 + \frac{1}{2c}\right)$ and $\theta^2 - c\sqrt{\frac{\log(1/\delta)}{n}} \geq |\theta|^2 / 2$ as long as $|\theta| \geq \sqrt{2c} \left(\log(1/\delta)/n\right)^{1/4}$.

Proof of claim (6.83a): Invoking triangle inequality, when $n \geq c' \log(1/\delta)$ we have

$$\left|\bar{\mathcal{L}}_n'(\theta)\right| \leq c\,|\theta| \left(\sqrt{\frac{\log(1/\delta)}{n}} + \frac{1}{2} - \frac{1}{2(\theta^2+1)\left(2\sqrt{\theta^2+1} - 1\right)} \right),$$

with probability at least $1-2\delta$ for any $|\theta| \in [0, 1/2]$ where the inequality in the above display is due to the inequalities (6.76) and (6.77). Furthermore, we can validate that

$$\frac{1}{2} - \frac{1}{2(\theta^2+1)\left(2\sqrt{\theta^2+1}-1\right)} \leq \frac{3\theta^2}{2}$$

for any $|\theta| \in [0, 1/2]$. In light of the previous inequalities, we arrive at the following inequality

$$\left|\bar{\mathcal{L}}_n'(\theta)\right| \leq \frac{3c\,|\theta|}{2}\left(\sqrt{\frac{\log(1/\delta)}{n}} + \theta^2\right),$$

with probability at least $1-2\delta$ for all $|\theta| \in [0, 1/2]$. As a consequence, we reach the conclusion of claim (6.83a).

Proof of claims (6.83b) and (6.83c): The proof of claim (6.83b) is a direct application of triangle inequality and the fact that $|\theta| \in [0, 1/2]$. In addition, we have

$$\left|\bar{\mathcal{L}}_n''(\theta)\right| \geq \left|\bar{\mathcal{L}}''(\theta)\right| - \left|\bar{\mathcal{L}}_n''(\theta) - \bar{\mathcal{L}}''(\theta)\right| \geq c'\left(\theta^2 - c\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

with probability at least $1-2\delta$ for any $|\theta| \in [0, 1/2]$ where c, c' are universal constants independent of δ and the last inequality in the above display is due the results from equation (6.82) and claim (6.83b). As a consequence, we achieve the conclusion of claim (6.83c).

Lower bound and monotonicity of Newton updates: Now, we proceed to verify the lower bound of Newton updates in claim (6.78). In order to ease the ensuing presentation, we denote $f(\theta) := \frac{1}{(\theta^2+1)(2\sqrt{\theta^2+1}-1)}$ for all θ. The global maxima θ_n^* of the sample log-likelihood function $\bar{\mathcal{L}}_n$ are the solutions of the following equation

$$\theta_n^* f(\theta_n^*)\left(\frac{1}{n}\sum_{i=1}^n (1-R_i)\right) = \theta_n^*\left(\frac{1}{n}\sum_{i=1}^n R_i Y_i^2\right).$$

The specific forms of θ_n^* depend on the values of R_i, Y_i for $i \in [n]$. In particular, when $\sum_{i=1}^n R_i Y_i^2 < \sum_{i=1}^n (1-R_i)$, namely, the Hessian of sample likelihood function $\bar{\mathcal{L}}_n$ at 0 is positive, the function $\bar{\mathcal{L}}_n$ is bimodal and symmetric around 0. Additionally, θ_n^* are different from 0 and become the solution of the following equation

$$f(\theta_n^*)\left(\frac{1}{n}\sum_{i=1}^n (1-R_i)\right) = \left(\frac{1}{n}\sum_{i=1}^n R_i Y_i^2\right). \tag{6.84}$$

On the other hand, when $\sum_{i=1}^n R_i Y_i^2 > \sum_{i=1}^n (1-R_i)$, the function $\bar{\mathcal{L}}_n$ is unimodal and symmetric around 0. Under this case, $\theta_n^* = 0$ is the unique global maximum.

Without loss of generality, we assume that $\theta > 0$ and the global maxima are solutions of equation (6.84). From the formulation of $\mathrm{M}_n^{\mathrm{NM}}$, the inequality $\mathrm{M}_n^{\mathrm{NM}}(\theta) > 0$ is equivalent to

$$\theta f'(\theta) + f(\theta) < f(\theta_n^*),$$

which holds for all $\theta \geq |\theta_n^*|$ since $f(\theta) < f(\theta_n^*)$ and $f'(\theta) < 0$ as $\theta \geq |\theta_n^*|$. Therefore, we have $\mathrm{M}_n^{\mathrm{NM}}(\theta) > 0$ for all $\theta \geq |\theta_n^*|$. Now, in order to demonstrate that $\mathrm{M}_n^{\mathrm{NM}}(\theta) \geq |\theta_n^*|$ for $\theta \geq |\theta_n^*|$, it is equivalent to

$$(|\theta_n^*| - \theta)\, \theta f'(\theta) + |\theta_n^*|\, (f(\theta) - f(\theta_n^*)) \geq 0. \tag{6.85}$$

Invoking mean value theorem, we can find some constant $\bar{\theta} \in (|\theta_n^*|\,, \theta)$ such that

$$f(\theta) - f(\theta_n^*) = f(\theta) - f(|\theta_n^*|) = f'(\bar{\theta})(\theta - |\theta_n^*|).$$

Given the above equation, the inequality (6.85) can be rewritten as

$$|\theta_n^*|\, f'(\bar{\theta}) \geq \theta f'(\theta) \tag{6.86}$$

for all $\theta \geq |\theta_n^*|$. Since the function $\theta f'(\theta)$ is a decreasing function in $(0, 1/2]$, we have $\theta f'(\theta) \leq \bar{\theta} f'(\bar{\theta})$ for any $\bar{\theta} < \theta$. Since $f'(\bar{\theta}) < 0$ and $\bar{\theta} > |\theta_n^*|$, we find that $\bar{\theta} f'(\bar{\theta}) \leq |\theta_n^*|\, f'(\bar{\theta})$. In light of these two inequalities, we achieve the inequality (6.86). As a consequence, we reach the conclusion of claim (6.78).

Proof of Corollary 9

Under the model (6.26b), the sample EM operator takes the following form:

$$\mathrm{G}_n^{\mathrm{EM}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} X_i \tanh(\theta X_i),$$

where $\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}$ for all $x \in \mathbb{R}$. We note that the result characterizing the behavior of sample EM operator is already proven in our prior work [Dwi+20b] (see Theorem 3 in that paper). Therefore, we only present the proof for the convergence rate of Newton updates in Section 6.9. The forms for the sample and population Newton operators are equivalent to running Newton's method on the sample and population log-likelihoods:

$$\mathrm{G}^{\mathrm{NM}}(\theta) = \theta - \left[\mathcal{L}''(\theta)\right]^{-1} \mathcal{L}'(\theta) = \theta + \frac{\mathbb{E}\left[X \tanh(X\theta)\right] - \theta}{\mathbb{E}\left[X^2 \tanh^2(X\theta)\right]}, \quad \text{and} \tag{6.87a}$$

$$\mathrm{G}_n^{\mathrm{NM}}(\theta) = \theta - \left[\mathcal{L}_n''(\theta)\right]^{-1} \mathcal{L}_n'(\theta) = \theta + \frac{\left(\frac{1}{n}\sum_{i=1}^{n} X_i \tanh(X_i\theta)\right) - \theta}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 \tanh^2(X_i\theta) + 1 - \frac{1}{n}\sum_{i=1}^{n} X_i^2}. \tag{6.87b}$$

Proofs for Newton operators

We begin by verifying the fast convergence of the operator G^{NM} and then the instability of the operator $\mathrm{G}_n^{\mathrm{NM}}$ with respect to G^{NM} in Theorem 14. Then, we demonstrate that the Newton updates satisfy Assumption (D). Noting that it can be done by establishing that the Newton updates are monotone decreasing and admit the following lower bound

$$\left|\mathrm{G}_n^{\mathrm{NM}}(\theta)\right| \geq |\theta_n^*| \tag{6.88}$$

for all $|\theta| \in [|\theta_n^*|, 1/3]$ for any global maximum θ_n^* of $\mathcal{L}_n$.

Fast convergence of the population-level operator G^{NM}: We provide the full proof for the case $\theta \in (0, \frac{1}{3}]$; the proof for the case $\theta \in [-\frac{1}{3}, 0)$ is analogous. We make use of the following known bounds [Dwi+20a] on the hyperbolic function $x \mapsto x\tanh(x)$:

$$x^2 - \frac{x^4}{3} \leq x\tanh(x) \leq x^2 - \frac{x^4}{3} + \frac{2x^6}{15} \qquad \text{for all } x \in \mathbb{R}. \tag{6.89}$$

Applying this bound, we obtain that

$$\mathbb{E}\left[X\tanh(X\theta)\right] \leq \frac{1}{\theta}\mathbb{E}\left[(X\theta)^2 - (X\theta)^4/3 + 2(X\theta)^6/15\right] = \theta - \theta^3 + 2\theta^5, \quad \text{as well as}$$
$$\mathbb{E}\left[X^2\tanh^2(X\theta)\right] \leq \frac{1}{\theta^2}\mathbb{E}\left[(X\theta)^4\right] = 3\theta^2,$$

and consequently that

$$\frac{\theta - \mathbb{E}\left[X\tanh(X\theta)\right]}{\mathbb{E}\left[X^2\tanh^2(X\theta)\right]} \geq \frac{\theta - (\theta - \theta^3 + 2\theta^5)}{3\theta^2} = \frac{\theta - 2\theta^3}{3} \overset{(\theta\in(0,\frac{1}{3}])}{\geq} \frac{2\theta}{9}.$$

Noting that $\mathrm{G}^{\mathrm{NM}}(\theta) = \theta - \frac{\theta - \mathbb{E}[X\tanh(X\theta)]}{\mathbb{E}\left[X^2\tanh^2(X\theta)\right]}$ and $\theta^\star = 0$, we conclude that the population Newton operator G^{NM} is $\texttt{FAST}(\frac{7}{9})$-convergent over the ball $\mathbb{B}(\theta^\star, \frac{1}{3})$.

Instability of the sample-level operator $\mathrm{G}_n^{\mathrm{NM}}$: Let us introduce the shorthand

$$A_n := \frac{1}{n}\sum_{i=1}^n X_i\tanh(X_i\theta), \quad \text{and} \quad B_n := \frac{1}{n}\sum_{i=1}^n X_i^2\tanh^2(X_i\theta) + 1 - \frac{1}{n}\sum_{i=1}^n X_i^2.$$

Using the definitions (6.87b) of the operators $\mathrm{G}_n^{\mathrm{NM}}$ and G^{NM}, we find that

$$\begin{aligned}\left|\mathrm{G}_n^{\mathrm{NM}}(\theta) - \mathrm{G}^{\mathrm{NM}}(\theta)\right| &= \left|\frac{\mathbb{E}\left[X\tanh(X\theta)\right] - \theta}{\mathbb{E}\left[X^2\tanh^2(X\theta)\right]} - \frac{A_n - \theta}{B_n}\right| \\ &\leq \underbrace{\frac{\left|\mathbb{E}\left[X\tanh(X\theta)\right] - A_n\right|}{\mathbb{E}\left[X^2\tanh^2(X\theta)\right]}}_{:=J_1} + \underbrace{|A_n - \theta|\left|\frac{1}{\mathbb{E}\left[X^2\tanh^2(X\theta)\right]} - \frac{1}{B_n}\right|}_{:=J_2}.\end{aligned} \tag{6.90}$$

Thus, in order to bound the difference $\left|\mathrm{G}_n^{\mathrm{NM}}(\theta) - \mathrm{G}^{\mathrm{NM}}(\theta)\right|$, it suffices to derive bounds for the terms J_1 and J_2.

Upper bound for J_1: For a given $\delta \in (0,1)$, as long as the sample size $n \geq C\log(1/\delta)$ for some universal constant C, we can apply Lemma 1 from the paper [Dwi+20b] to assert that

$$\left|\mathbb{E}\left[X\tanh(X\theta)\right] - A_n\right| \leq c\,|\theta|\sqrt{\frac{\log(1/\delta)}{n}} \quad \text{for all } |\theta| \in (0, \tfrac{1}{3}) \tag{6.91}$$

with probability $1-\delta$. Moreover, the bound (6.89) implies that

$$\mathbb{E}\left[X^2\tanh^2(X\theta)\right] \geq \frac{1}{\theta^2}\mathbb{E}\left[\left((X\theta)^2 - \frac{(X\theta)^4}{3}\right)^2\right] = 3\theta^2 - 10\theta^4 + \frac{35\theta^6}{33} \geq 2\theta^2,$$

for $\theta \in [-\frac{1}{3}, \frac{1}{3}]$. Combining the above inequalities yields

$$J_1 = \frac{\left|\mathbb{E}\left[X\tanh(X\theta)\right] - A_n\right|}{\mathbb{E}\left[X^2\tanh^2(X\theta)\right]} \leq c\frac{|\theta|\sqrt{\frac{\log(1/\delta)}{n}}}{2\theta^2} \leq c'\frac{1}{|\theta|}\sqrt{\frac{\log(1/\delta)}{n}}, \tag{6.92}$$

for all $|\theta| \in (0, 1/3)$ with probability at least $1-\delta$.

Upper bound for J_2: In order to obtain an upper bound for J_2, we claim the following key bounds appearing in its formulation:

$$|A_n - \theta| \leq c_1\left(|\theta|\sqrt{\frac{\log(1/\delta)}{n}} + |\theta|^3\right), \tag{6.93a}$$

$$|B_n| \geq c_2\left(\theta^2 - c\frac{\log^4(3n/\delta)}{\sqrt{n}}\right), \tag{6.93b}$$

$$\left|\mathbb{E}\left[X^2\tanh^2(X\theta)\right] - B_n\right| \leq c_3\frac{\log(n/\delta)}{\sqrt{n}}, \tag{6.93c}$$

for all $|\theta| \in (0, 1/3]$ with probability at least $1 - 2\delta$ as long as the sample size $n \geq c \log(1/\delta)$. Here, c, c_1, c_2, c_3 in the above probability bounds are universal constants independent of δ. Assume that the above claims are given at the moment. The results in these claims lead to

$$J_2 = |A_n| \left| \frac{\mathbb{E}\left[X^2 \tanh^2(X\theta)\right] - B_n}{B_n \mathbb{E}\left[X^2 \tanh^2(X\theta)\right]} \right| \leq c' \left(|\theta| \sqrt{\frac{\log(1/\delta)}{n}} + |\theta|^3 \right) \frac{\frac{\log(n/\delta)}{\sqrt{n}}}{\theta^2 \left(\theta^2 - c \frac{\log^4(3n/\delta)}{\sqrt{n}} \right)}$$
$$\leq c'' \frac{1}{|\theta|} \frac{\log(n/\delta)}{\sqrt{n}} \tag{6.94}$$

with probability at least $1 - 5\delta$. Here, the last inequality is due to the facts that

$$|\theta| \sqrt{\frac{\log(1/\delta)}{n}} + |\theta|^3 \leq |\theta|^3 \left(1 + \frac{1}{2c}\right) \quad \text{and} \quad \theta^2 - c \frac{\log^4(3n/\delta)}{\sqrt{n}} \geq |\theta|^2 / 2,$$

as long as $|\theta| \geq \sqrt{2c} \log^2(3n/\delta)/n^{1/4}$. Plugging the bounds (6.92) and (6.94) into equation (6.90), we conclude that the operator $\mathrm{G}_n^{\mathrm{NM}}$ is $\mathtt{UNS}(-1)$-unstable over the annulus $\mathbb{A}(\theta^\star, \frac{\sqrt{2c}\log^2(3n/\delta)}{n^{1/4}}, 1/3)$ with noise function $\varepsilon(n, \delta) = \frac{\log(n/\delta)}{\sqrt{n}}$ as long as the sample size $n \geq C \frac{\log^8(3n/\delta)}{n^{1/4}}$.

Proof of claim (6.93a)**:** Invoking the concentration bound (6.91) and applying the triangle inequality, we find that

$$|A_n - \theta| \leq \left| \frac{1}{n} \sum_{i=1}^n X_i \tanh(X_i\theta) - \mathbb{E}\left[X \tanh(X\theta)\right] \right| + \left| \mathbb{E}\left[X \tanh(X\theta)\right] - \theta \right|$$
$$\leq c \left(|\theta| \sqrt{\frac{\log(1/\delta)}{n}} + \frac{1}{|\theta|} \left| \mathbb{E}\left[X\theta \tanh(X\theta)\right] - \theta^2 \right| \right)$$

for all $|\theta| \in (0, 1/3]$ with probability $1 - \delta$. Next, taking expectation on both sides in the bounds (6.89), we find that

$$\mathbb{E}\left[X\theta \tanh(X\theta)\right] - \theta^2 \leq \mathbb{E}\left[(X\theta)^2 - \frac{(X\theta)^4}{3} + \frac{2(X\theta)^6}{15} \right] - \theta^2 = -\theta^4 + 2\theta^6 \leq -\frac{7\theta^4}{9}, \quad \text{and}$$
$$\mathbb{E}\left[X\theta \tanh(X\theta)\right] - \theta^2 \geq \mathbb{E}\left[(X\theta)^2 - \frac{(X\theta)^4}{3} \right] - \theta^2 = -\theta^4.$$

Putting these pieces together yields the claim (6.93a).

Proof of claim (6.93b)**:** Invoking standard chi-squared concentration bounds and applying triangle inequality, we obtain that

$$|B_n| \geq \frac{1}{n} \sum_{i=1}^n X_i^2 \tanh^2(X_i\theta) - \left| \frac{1}{n} \sum_{i=1}^n X_i^2 - 1 \right| \geq c \left(\frac{1}{n} \sum_{i=1}^n X_i^2 \tanh^2(X_i\theta) - \sqrt{\frac{\log(1/\delta)}{n}} \right)$$

with probability at least $1-\delta$. Using the lower bound from inequality (6.89), we find that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} X_i^2 \tanh^2(X_i\theta) &\geq \frac{1}{n}\sum_{i=1}^{n}\left(\theta X_i^2 - \frac{\theta^3 X_i^4}{3}\right)^2 \\
&= \theta^2\left(\frac{1}{n}\sum_{i=1}^{n} X_i^4\right) - \frac{2\theta^4}{3}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^6\right) + \frac{\theta^6}{9}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^8\right) \\
&\overset{(i)}{\geq} \theta^2\left(3 - c'\frac{\log^2(3n/\delta)}{\sqrt{n}}\right) - \frac{2\theta^4}{3}\left(15 + c'\frac{\log^3(3n/\delta)}{\sqrt{n}}\right) \\
&\qquad + \frac{\theta^6}{9}\left(105 - c'\frac{\log^4(3n/\delta)}{\sqrt{n}}\right) \\
&\geq \theta^2 - c'\frac{\log^4(3n/\delta)}{\sqrt{n}},
\end{aligned}$$

with probability at least $1-\delta$ for some universal constant c. Here step (i) makes use of the following concentration bound for higher moments of Gaussian random variables (Lemma 5 [Dwi+20a]):

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i^{2k} - \mathbb{E}\left[X^{2k}\right]\right| \leq c'\frac{\log^k(3n/\delta)}{n^{\frac{1}{2}}}\right] \geq 1 - \frac{\delta}{3} \quad \text{for } k \in \{2,4,6\}$$

with probability at least $1-\delta/3$ for $k \in \{2,4,6\}$. Putting together the pieces yields the claim (6.93b).

Proof of claim (6.93c)**:** Applying the triangle inequality yields

$$\begin{aligned}
\left|\mathbb{E}\left[X^2\tanh^2(X\theta)\right] - B_n\right| &\leq \left|\frac{1}{n}\sum_{i=1}^{n} X_i^2\tanh^2(X_i\theta) - \mathbb{E}\left[X^2\tanh^2(X\theta)\right]\right| + \left|\frac{1}{n}\sum_{i=1}^{n} X_i^2 - 1\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{n} X_i^2\tanh^2(X_i\theta) - \mathbb{E}\left[X^2\tanh^2(X\theta)\right]\right| + c\sqrt{\frac{\log(1/\delta)}{n}}
\end{aligned} \tag{6.95}$$

with probability at least $1-\delta$. By adapting the truncation argument from the proof of Lemma 5 in the paper [Dwi+20a] for the random variable $X\tanh(X)$ with $X \sim \mathcal{N}(0,1)$, it follows that

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i^2\tanh^2(X_i\theta) - \mathbb{E}\left[X^2\tanh^2(X\theta)\right]\right| \leq c'\frac{\log(n/\delta)}{\sqrt{n}},$$

for all $|\theta| \in (0, 1/3]$ with probability at least $1-\delta$. Putting the results together yields the claim (6.93c).

Lower bound and monotonicity of Newton updates: We first provide several insights into the landscape of sample log-likelihood function $\mathcal{L}_n$. To faciliate the proof argument, we define

$$f(\theta) = \theta - \frac{1}{n}\sum_{i=1}^{n} X_i \tanh(X_i\theta).$$

A simple calculation with the gradient and Hessian of sample log-likelihood function $\mathcal{L}_n$ indicates that when $\sum_{i=1}^n X_i^2 > n$, the function $\mathcal{L}_n$ is bimodal and symmetric around 0. Additionally, the global maxima θ_n^* of the function $\mathcal{L}_n$ are different from 0 and solutions of the equation $f(\theta) = 0$. On the other hand, when $\sum_{i=1}^n X_i^2 \leq n$, the function $\mathcal{L}_n$ is unimodal and symmetric around 0 while $\theta_n^* = 0$ is the unique global maximum of that function.

Now, we only verify the lower bound of Newton updates $\mathrm{G}_n^{\mathrm{NM}}(\theta)$ in claim (6.88); the proof of monotonicity can be argued similarly. Without loss of generality, we only consider the setting when the global maxima θ_n^* are different from 0 and $\theta > 0$. Under that case, the Hessian of the function $\mathcal{L}_n$ at $|\theta_n^*|$ is negative. A direct computation with the gradient of the function f leads to

$$\begin{aligned} f'(\theta) = 1 - \frac{1}{n}\sum_{i=1}^{n} X_i^2 \mathrm{sech}^2(X_i\theta) &= 1 - \frac{1}{n}\sum_{i=1}^{n} X_i^2 \mathrm{sech}^2(|X_i|\,|\theta|) \\ &\geq 1 - \frac{1}{n}\sum_{i=1}^{n} X_i^2 \mathrm{sech}^2(|X_i|\,|\theta_n^*|) = -\nabla^2 \mathcal{L}_n(\theta_n^*) > 0 \end{aligned}$$

for any $\theta > |\theta_n^*|$. Therefore, the function f is a strictly increasing function when $\theta > |\theta_n^*|$. It leads to the inequality $f(\theta) \geq f(\theta_n^*) = 0$ for all $\theta \geq |\theta_n^*|$. Further computation with second derivative of f yields that

$$f''(\theta) = \frac{2}{n}\sum_{i=1}^{n} X_i^3 \tanh(X_i\theta)\mathrm{sech}^2(X_i\theta) > 0$$

for all $\theta > 0$. The above inequality is due to $X_i \tanh(X_i\theta) > 0$ for all $\theta > 0$ and $i \in [n]$. Thus, the function f' is strictly increasing when $\theta > 0$.

Now the inequality $\mathrm{G}_n^{\mathrm{NM}}(\theta) \geq |\theta_n^*|$ for all $\theta \geq |\theta_n^*|$ is equivalent to

$$f'(\theta)(\theta - |\theta_n^*|) \geq f(\theta) - f(\theta_n^*). \tag{6.96}$$

Invoking the mean value theorem, we find that

$$f(\theta) - f(\theta_n^*) = f(\theta) - f(|\theta_n^*|) = f'(\bar{\theta})(\theta - |\theta_n^*|)$$

for some $\bar{\theta} \in (|\theta_n^*|, \theta)$. Given that equality, the equality (6.96) can be rewritten as $f'(\theta) \geq f'(\bar{\theta})$ for all $\theta \geq |\theta_n^*|$. This inequality is true since f' is an increasing function when $\theta > 0$. As a consequence, we achieve the conclusion of claim (6.88).

Proof of Corollary 10

In this section, we demonstrate the convergence and stability properties of operators from gradient descent and (cubic-regularized) Newton's methods in the single-index model. The sample operators of these methods take the following forms

$$\mathrm{F}_n^{\mathrm{GD}}(\theta) = \theta - \eta\widetilde{\mathcal{L}}_n'(\theta) = \theta - \eta\left(\frac{2p}{n}\sum_{i=1}^n X_i^{4p}\theta^{4p-1} - \frac{2p}{n}\sum_{i=1}^n Y_i X_i^{2p}\theta^{2p-1}\right), \tag{6.97a}$$

$$\begin{aligned}\mathrm{F}_n^{\mathrm{NM}}(\theta) &= \theta - \left[\widetilde{\mathcal{L}}_n''(\theta)\right]^{-1}\widetilde{\mathcal{L}}_n'(\theta) \\ &= \theta - \frac{\left(\frac{1}{n}\sum_{i=1}^n X_i^{4p}\right)\theta^{2p+1} - \left(\frac{1}{n}\sum_{i=1}^n Y_i X_i^{2p}\right)\theta}{\left(\frac{4p-1}{n}\sum_{i=1}^n X_i^{4p}\right)\theta^{2p} - \frac{2p-1}{n}\sum_{i=1}^n Y_i X_i^{2p}}, \quad \text{and}\end{aligned} \tag{6.97b}$$

$$\mathrm{F}_n^{\mathrm{CNM}}(\theta) = \arg\min_{y\in\mathbb{R}}\left\{\widetilde{\mathcal{L}}_n'(\theta)(y-\theta) + \frac{1}{2}\widetilde{\mathcal{L}}_n''(\theta)(y-\theta)^2 + L\left|y-\theta\right|^3\right\}, \tag{6.97c}$$

where $L := (4p-1)!!(4p-1)p/3$. Noting that the specific choice of L in the formulation of the cubic-regularized Newton operator $\mathrm{F}_n^{\mathrm{CNM}}$ arises because the second-order derivative of $\widetilde{\mathcal{L}}_n$ is Lipschitz continuous with constant L. Similarly, the population-level operators are given by

$$\mathrm{F}^{\mathrm{GD}}(\theta) = \theta - \eta\widetilde{\mathcal{L}}'(\theta) = \theta\left[1 - (4p-1)!!(2p)\eta\theta^{4p-2}\right], \tag{6.98a}$$

$$\mathrm{F}^{\mathrm{NM}}(\theta) = \theta - \left[\widetilde{\mathcal{L}}''(\theta)\right]^{-1}\widetilde{\mathcal{L}}'(\theta) = \frac{(4p-2)}{4p-1}\theta, \quad \text{and} \tag{6.98b}$$

$$\mathrm{F}^{\mathrm{CNM}}(\theta) = \arg\min_{y\in\mathbb{R}}\left\{\widetilde{\mathcal{L}}'(\theta)(y-\theta) + \frac{1}{2}\widetilde{\mathcal{L}}''(\theta)(y-\theta)^2 + L\left|y-\theta\right|^3\right\}. \tag{6.98c}$$

Proofs for the gradient descent operators

In order to achieve the conclusion of the corollary with convergence rate of updates from gradient descent method, it is sufficient to demonstrate that the sample gradient operator $\mathrm{F}_n^{\mathrm{GD}}$ is $\mathtt{STA}(2p-1)$-stable over the Euclidean ball $\mathbb{B}(\theta^\star, 1)$ with noise function $\varepsilon(n,\delta) = \frac{\log^{2p}(n/\delta)}{\sqrt{n}}$. By using the similar truncation argument as that in equation (6.95), we can verify the following concentration bound

$$\left|\frac{1}{n}\sum_{i=1}^n Y_i X_i^{2p}\right| \leq c\log^{2p}(n/\delta)/\sqrt{n}, \tag{6.99}$$

with probability $1-\delta$ where c is some universal constant. An application of triangle inequality yields

$$\left|\mathrm{F}^{\mathrm{GA}}(\theta) - \mathrm{F}_n^{\mathrm{GA}}(\theta)\right| \leq \left|\frac{1}{n}\sum_{i=1}^n X_i^{4p} - (4p-1)!!\right|\left|\theta\right|^{4p-1} + c\frac{\log^{2p}(n/\delta)}{\sqrt{n}}\left|\theta\right|^{2p-1}. \tag{6.100}$$

Based on known concentration bounds for moments of Gaussian random variables (cf. Lemma 5 in [Dwi+20a]), we have

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i^{4p} - (4p-1)!!\right| \leq c' \log^{2p}(n/\delta)/\sqrt{n} \tag{6.101}$$

with probability $1-\delta$ where c' is some universal constant. Substituting the inequality (6.101) into equation (6.100) yields the above claim with the stability of $\mathrm{F}_n^{\mathrm{GD}}$.

Proofs for the Newton operators

Moving to the convergence rates of updates from Newton's method, it is sufficient to establish the instability of $\mathrm{F}_n^{\mathrm{NM}}$ with respect to F^{NM}, and moreover that, for any global minimum θ_n^* of the sample least-squares function $\widetilde{\mathcal{L}}_n$ in equation (6.32b), we have

$$\left|\mathrm{F}_n^{\mathrm{NM}}(\theta)\right| \geq |\theta_n^*|\,, \tag{6.102}$$

for all $|\theta| \in [|\theta_n^*|\,, 1]$.

Instability of the sample Newton operator $\mathrm{F}_n^{\mathrm{NM}}$: Let us introduce the following shorthand notation:

$$\begin{aligned}
A_n &:= \left(\frac{2p}{n}\sum_{i=1}^{n} X_i^{4p}\right)\theta^{4p-1} - \left(\frac{2p}{n}\sum_{i=1}^{n} Y_i X_i^{2p}\right)\theta^{2p-1}, \\
B_n &:= \left(\frac{2p(4p-1)}{n}\sum_{i=1}^{n} X_i^{4p}\right)\theta^{4p-2} - \left(\frac{2p(2p-1)}{n}\sum_{i=1}^{n} Y_i X_i^{2p}\right)\theta^{2p-2}.
\end{aligned}$$

Applying the triangle inequality yields

$$\begin{aligned}
\left|\mathrm{F}_n^{\mathrm{NM}}(\theta) - \mathrm{F}^{\mathrm{NM}}(\theta)\right| \leq \underbrace{\frac{|(4p-1)!!(2p)\theta^{4p-1} - A_n|}{(4p-1)!!(2p)(4p-1)\theta^{4p-2}}}_{:=J_1} \\
+ \underbrace{|A_n|\left|\frac{1}{(4p-1)!!(2p)(4p-1)\theta^{4p-2}} - \frac{1}{B_n}\right|}_{:=J_2}.
\end{aligned}$$

Upper bound for J_1: Invoking triangle inequality, we obtain that

$$\begin{aligned}
\left|A_n - (4p-1)!!(2p)\theta^{4p-1}\right| &\leq 2p\left|\frac{1}{n}\sum_{i=1}^{n} X_i^{4p} - (4p-1)!!\right||\theta|^{4p-1} + \left|\frac{2p}{n}\sum_{i=1}^{n} Y_i X_i^{2p}\right||\theta|^{2p-1} \\
&\leq c\frac{\log^{2p}(n/\delta)}{\sqrt{n}}\left(|\theta|^{4p-1} + |\theta|^{2p-1}\right),
\end{aligned}$$

where the last inequality is due to concentration bounds for moments of Gaussian random variables (6.99). With the above inequality, we have

$$J_1 \leq \frac{c \log^{2p}(n/\delta) \left(|\theta|^{4p-1} + |\theta|^{2p-1} \right)}{(4p-1)!!(2p)(4p-1)\sqrt{n}\,|\theta|^{4p-2}} \leq \frac{2c}{|\theta|^{2p-1}} \frac{\log^{2p}(n/\delta)}{\sqrt{n}}, \tag{6.103}$$

for all $|\theta| \leq 1$ with probability at least $1 - 2\delta$.

Upper bound for J_2: In order to obtain an upper bound for J_2, we exploit the following concentration bounds

$$|A_n| \leq c_1 \left(|\theta|^{4p-1} + \frac{\log^{2p}(n/\delta)}{\sqrt{n}} |\theta|^{2p-1} \right), \tag{6.104a}$$

$$\left| B_n - (4p-1)!!(2p)(4p-1)\theta^{4p-2} \right| \leq c_2 \frac{\log^{2p}(n/\delta)}{\sqrt{n}}, \tag{6.104b}$$

$$|B_n| \geq c_3 \left((4p-1)!!(2p)(4p-1)\theta^{4p-2} - c\frac{\log^{2p}(n/\delta)}{\sqrt{n}} \right), \tag{6.104c}$$

for all $|\theta| \leq 1$ with probability at least $1 - 2\delta$. Here, c, c_1, c_2, c_3 are universal constants independent of δ. The proofs of the above claims are direct applications of triangle inequalities and concentration bounds we utilized earlier with gradient descent operators in Section 6.9; therefore, they are omitted. In light of the above bounds, we can bound J_2 as follows:

$$\begin{aligned} J_2 &\leq \frac{c_1 c_2}{c_3} \left(|\theta|^{4p-1} + \frac{\log^{2p}(n/\delta)}{\sqrt{n}} |\theta|^{2p-1} \right) \frac{\frac{\log^{2p}(n/\delta)}{\sqrt{n}}}{\theta^{4p-2} \left((4p-1)!!(2p)(4p-1)\theta^{4p-2} - c\frac{\log^{2p}(n/\delta)}{\sqrt{n}} \right)} \\ &\leq \frac{2c_1 c_2}{c_3 c} \frac{1}{|\theta|^{2p-1}} \frac{\log^{2p}(n/\delta)}{\sqrt{n}}, \end{aligned} \tag{6.105}$$

for all $|\theta| \in [C \cdot \log^{p/(2p-1)}(n/\delta)/n^{1/4(2p-1)}, 1]$ with probability $1 - 6\delta$ where C is solution of the equation $(4p-1)!!(2p)(4p-1)\theta^{4p-2} = 2c\frac{\log^{2p}(n/\delta)}{\sqrt{n}}$. Combining the results from equations (6.103) and (6.105), we achieve that

$$\left| \mathrm{F}_n^{\mathrm{NM}}(\theta) - \mathrm{F}^{\mathrm{NM}}(\theta) \right| \leq c' \frac{1}{|\theta|^{2p-1}} \frac{\log^{2p}(n/\delta)}{\sqrt{n}} \tag{6.106}$$

for all $|\theta| \in [C \log^{p/(2p-1)}(n/\delta)/n^{1/4(2p-1)}, 1]$ with probability $1 - 8\delta$ where c' is some universal constant.

As a consequence, the sample operator $\mathrm{F}_n^{\mathrm{NM}}$ is $\mathtt{UNS}(-2p+1)$-unstable over the annulus $\mathbb{A}(\theta^\star, c_1 \log^{p/(2p-1)}(n/\delta)/n^{1/4(2p-1)}, 1)$ with noise function $\varepsilon(n,\delta) = \dfrac{\log^{2p}(n/\delta)}{\sqrt{n}}$.

Lower bound and monotonicity of Newton updates: Moving to the claim (6.102), we first study the global minima θ_n^* of the sample least-squares function $\widetilde{\mathcal{L}}_n$ in equation (6.32b). In particular, they satisfy the equation $\nabla \widetilde{\mathcal{L}}_n(\theta_n^*) = 0$, which is equivalent to

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i^{4p}\right)(\theta_n^*)^{4p-1} - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i X_i^{2p}\right)(\theta_n^*)^{2p-1} = 0.$$

Given the above equation, the specific form of θ_n^* depends on the sign of second derivative of $\widetilde{\mathcal{L}}_n$ at 0. In particular, when $\sum_{i=1}^{n} Y_i X_i^{2p} > 0$, the function $\widetilde{\mathcal{L}}_n$ is bimodal and symmetric around 0. Additionally, global mimima θ_n^* have the form

$$(\theta_n^*)^{2p} = \left(\frac{1}{n}\sum_{i=1}^{n} Y_i X_i^{2p}\right) \Big/ \left(\frac{1}{n}\sum_{i=1}^{n} X_i^{4p}\right). \tag{6.107}$$

On the other hand, when $\frac{1}{n}\sum_{i=1}^{n} Y_i X_i^{2p} \leq 0$, the function $\widetilde{\mathcal{L}}_n$ is unimodal and symmetric around 0. Furthermore, it has only global minimum $\theta_n^* = 0$.

Now, we focus on the case $\theta > 0$ and $\sum_{i=1}^{n} Y_i X_i^{2p} > 0$, i.e., the global minima θ_n^* are different from 0 and the solutions of equation (6.107). A simple calculation demonstrates that $B_n > 0$ and $\mathrm{F}_n^{\mathrm{NM}}(\theta) > 0$ as long as $\theta > |\theta_n^*|$. Now, the inequality $\mathrm{F}_n^{\mathrm{NM}}(\theta) \geq |\theta_n^*|$ is equivalent to

$$\left(\frac{4p-2}{n}\sum_{i=1}^{n} X_i^{4p}\right)\theta^{2p+1} + \left(\frac{2p-1}{n}\sum_{i=1}^{n} Y_i X_i^{2p}\right)|\theta_n^*| \geq \left(\frac{4p-1}{n}\sum_{i=1}^{n} X_i^{4p}\right)\theta^{2p}\,|\theta_n^*| + \left(\frac{2p-2}{n}\sum_{i=1}^{n} Y_i X_i^{2p}\right)\theta$$

for $\theta \geq |\theta_n^*|$. In light of the closed form expression of $|\theta_n^*|$ in equation (6.107), a simple algebra with the above inequality leads to the inequality

$$(4p-2)\theta^{2p+1} + (2p-1)\,|\theta_n^*|^{2p+1} \geq (2p-2)\,(\theta_n^*)^{2p}\,\theta + (4p-1)\,|\theta_n^*|\,\theta^{2p},$$

which holds true due to AM-GM inequality. Thus, we have established the claim (6.102).

Proofs for the cubic-regularized Newton operators

Our proof is divided into three separate steps. First, we establish the slow convergence of operator $\mathrm{F}^{\mathrm{CNM}}$. Then, we proceed to establishing the instability of operator $\mathrm{F}_n^{\mathrm{CNM}}$. Finally, we demonstrate the monotonicity of cubic-regularized Newton updates and their lower bound

$$\left|\mathrm{F}_n^{\mathrm{CNM}}(\theta)\right| \geq |\theta_n^*|\,, \tag{6.108}$$

for all $|\theta| \in [|\theta_n^*|\,, 1]$ for any global minima θ_n^* of the sample least-squares function $\widetilde{\mathcal{L}}_n$ in equation (6.32b).

Slow convergence of $\mathrm{F}^{\mathrm{CNM}}$: Without loss of generality, we assume that $\theta \in (0, 1]$. Direct computation leads to

$$\begin{aligned}\mathrm{F}^{\mathrm{CNM}}(\theta) &= \theta + \theta^{4p-2} - \sqrt{\theta^{8p-4} + \frac{2}{4p-1}\theta^{4p-1}} \\ &= \theta - \frac{\frac{2}{4p-1}\theta^{4p-1}}{\theta^2 + \sqrt{\theta^{8p-4} + \frac{2}{4p-1}\theta^{4p-1}}} \leq \theta\left(1 - c_1\theta^{(4p-3)/2}\right),\end{aligned}$$

for any $\theta \in (0, 1]$ where $c_1 < 1$ is some universal constant. As a consequence, the operator $\mathrm{F}^{\mathrm{CNM}}$ satisfies slow convergence condition $\texttt{SLOW}(2/(4p-3))$ over the Euclidean ball $\mathbb{B}(\theta^\star, 1)$.

Instability of the sample operator $\mathrm{F}_n^{\mathrm{CNM}}$: To ease the presentation, we assume that $\theta > |\theta_n^*|$ where θ_n^* are global minima of the sample least-squares function $\widetilde{\mathcal{L}}_n$. With this condition, direct computation of $\mathrm{F}_n^{\mathrm{CNM}}(\theta)$ leads to

$$\mathrm{F}_n^{\mathrm{CNM}}(\theta) = \theta - \frac{2\widetilde{\mathcal{L}}_n'(\theta)}{\widetilde{\mathcal{L}}_n''(\theta) + \sqrt{\left(\widetilde{\mathcal{L}}_n''(\theta)\right)^2 + 12L \cdot \widetilde{\mathcal{L}}_n'(\theta)}} := \theta - \frac{2\widetilde{\mathcal{L}}_n'(\theta)}{T_n}.$$

Similar to the previous proofs with cubic-regularized Newton operators, we achieve that

$$\left|\mathrm{F}^{\mathrm{CNM}}(\theta) - \mathrm{F}_n^{\mathrm{CNM}}(\theta)\right| \leq 2\frac{\widetilde{\mathcal{L}}'(\theta)\,|T_n - T| + T\left|\widetilde{\mathcal{L}}_n'(\theta) - \widetilde{\mathcal{L}}'(\theta)\right|}{TT_n},$$

where $T := \widetilde{\mathcal{L}}''(\theta) + \sqrt{\left(\widetilde{\mathcal{L}}''(\theta)\right)^2 + 12L \cdot \widetilde{\mathcal{L}}'(\theta)} \geq \sqrt{12L\widetilde{\mathcal{L}}'(\theta)} \geq C \cdot \theta^{(4p-1)/2}$ for some universal constant $C > 0$. Additionally, we have

$$|T_n - T| \leq c' \cdot \theta^{-1/2}\frac{\log^{2p}(n/\delta)}{\sqrt{n}}$$

when $\theta \geq c \cdot \max\left\{|\theta_n^*|, \frac{\log^{p/(2p-1)}(n/\delta)}{n^{1/4(2p-1)}}\right\}$ with probability $1 - 10\delta$ for some universal constants c and c'. Furthermore, we can check that $T_n \geq \sqrt{12L \cdot \widetilde{\mathcal{L}}_n'(\theta)} \geq c''\theta^{(4p-1)/2}$ as long as $\theta \geq c \cdot \max\left\{|\theta_n^*|, \frac{\log^{p/(2p-1)}(n/\delta)}{n^{1/4(2p-1)}}\right\}$ with probability $1 - 2\delta$ for some universal constant c''. These inequalities guarantee that

$$\left|\mathrm{F}^{\mathrm{CNM}}(\theta) - \mathrm{F}_n^{\mathrm{CNM}}(\theta)\right| \leq c_1\theta^{-1/2}\frac{\log^{2p}(n/\delta)}{\sqrt{n}}$$

for all $\theta \geq c \cdot \max\left\{|\theta_n^*|, \frac{\log^{p/(2p-1)}(n/\delta)}{n^{1/4(2p-1)}}\right\}$ with probability $1 - 14\delta$. As a consequence, we conclude that the operator $\mathrm{F}_n^{\mathrm{CNM}}$ is $\texttt{UNS}(-1/2)$-unstable over the annulus $\mathbb{A}(\theta^\star, c\frac{\log^{p/(2p-1)}(n/\delta)}{n^{1/4(2p-1)}}, 1)$ with noise function $\varepsilon = \frac{\log^{2p}(n/\delta)}{\sqrt{n}}$ where c is some universal constant.

Lower bound and monotonicity of cubic-regularized Newton updates: To simplify the presentation, we only consider $\theta > 0$ and the setting when global minima θ_n^* are different from 0. As $\theta \geq |\theta_n^*|$, the inequality $\mathrm{F}_n^{\mathrm{CNM}}(\theta) \geq |\theta_n^*|$ is equivalent to

$$\widetilde{\mathcal{L}}_n''(\theta) + \sqrt{\left(\widetilde{\mathcal{L}}_n''(\theta)\right)^2 + 12L\widetilde{\mathcal{L}}_n'(\theta)} > 2\widetilde{\mathcal{L}}_n''(\widetilde{\theta})$$

for some $\widetilde{\theta} \in (|\theta_n^*|, \theta)$. This inequality holds since $\widetilde{\mathcal{L}}_n'$ and $\widetilde{\mathcal{L}}_n''$ are positive and strictly increasing when $\theta > |\theta_n^*|$, thereby completing the proof of claim (6.108).

Part III

Inference in sequential environments

Chapter 7

Near-optimal inference in adaptive linear regression

When data is collected in an adaptive manner, even simple methods like ordinary least squares can exhibit non-normal asymptotic behavior. As an undesirable consequence, hypothesis tests and confidence intervals based on asymptotic normality can lead to erroneous results. I this chapter we propose a family of online debiasing estimators to correct these distributional anomalies in least squares estimation. Our proposed methods take advantage of the covariance structure present in the dataset and provide sharper estimates in directions for which more information has accrued. We establish an asymptotic normality property for our proposed online debiasing estimators under mild conditions on the data collection process and provide asymptotically exact confidence intervals. We additionally prove a minimax lower bound for the adaptive linear regression problem, thereby providing a baseline by which to compare estimators. There are various conditions under which our proposed estimators achieve the minimax lower bounds up to logarithmic factors. We demonstrate the usefulness of our theory via applications to multi-armed bandit, autoregressive time series estimation, and active learning with exploration.

7.1 Introduction

Consider a prediction problem in which we observe n datapoints of the form $(\mathbf{x}_i, y_i) \in \mathbb{R}^D \times \mathbb{R}$ with covariate vector $\mathbf{x}_i$ and response y_i linked via the linear model

$$y_i = \mathbf{x}_i^\top \theta^* + \epsilon_i \quad \text{for } i = 1, \ldots, n. \tag{7.1}$$

Here the vector $\theta^* \in \mathbb{R}^D$ is an unknown parameter of interest, and ϵ_i is additive noise. When the datapoints are generated via some i.i.d. sampling process, this model, and in particular the behavior of the ordinary least squares (OLS) estimate $\widehat{\boldsymbol{\theta}}_{\text{LS}}$, is very well-understood. We focus here on a more challenging setting in which the covariate

vectors $\{\mathbf{x}_i\}_{i=1}^n$ have been *adaptively* collected, meaning that the choice of $\mathbf{x}_i$ can depend on the entire set of previous observations $\{\mathbf{x}_j, y_j\}_{j=1}^{i-1}$.

More precisely, given a filtration $\{\mathcal{F}_i\}_{i=1}^n$, assume that $\mathbf{x}_i$ is $\mathcal{F}_{i-1}$-measurable and that the additive error $\{\epsilon_i\}_{i=1}^n$ is a martingale difference sequence with respect to $\{\mathcal{F}_i\}_{i=1}^n$, so that the random variable ϵ_i is finite-variance martingale difference sequence, so that it is $\mathcal{F}_i$-measurable, with

$$\mathbb{E}[\epsilon_i \mid \mathcal{F}_{i-1}] = 0, \quad \text{and} \quad \mathbb{E}[\epsilon_i^2 | \mathcal{F}_{i-1}] = \sigma^2, \tag{7.2}$$

for some non-random scalar $\sigma^2 > 0$. We refer to the combination of the linear observation model (7.1) with such (potentially) adaptive collection procedures as the *adaptive linear regression model*. Instances of adaptive linear regression arise in a variety of applications, including multi-armed bandits [LS20], active learning [Fon+19], times series modeling [Box+15], stochastic control [Åst12], and adaptive stochastic approximation schemes [Des+18; LW+82].

Let us discuss some known results for the OLS estimate $\widehat{\boldsymbol{\theta}}_{\text{LS}}$. It can be expanded in the form

$$\widehat{\boldsymbol{\theta}}_{\text{LS}} = \mathbf{S}_n^{-1} X_n^\top y_n = \theta^* + \mathbf{S}_n^{-1} \sum_{i=1}^n \mathbf{x}_i \epsilon_i, \qquad \text{where} \quad \mathbf{S}_n = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top. \tag{7.3}$$

This decomposition reveals that the statistical properties of the OLS estimate depend on the martingale transform $\sum_{i=1}^n \mathbf{x}_i \epsilon_i$, along with the random matrix $\mathbf{S}_n^{-1}$. There is a lengthy literature on conditions under which the OLS estimate is consistent [Åst12; Box+15; GP77; LR79; LRW79; LW+82]. Notably, [LW+82, Thm. 1] show that the OLS estimate is strongly consistent, meaning that $\widehat{\boldsymbol{\theta}}_{\text{LS}} \xrightarrow{\text{a.s.}} \theta^*$, whenever

$$\lambda_{\min}(\mathbf{S}_n) \xrightarrow{\text{a.s.}} \infty \quad \text{and} \quad \frac{\log \lambda_{\max}(\mathbf{S}_n)}{\lambda_{\min}(\mathbf{S}_n)} \xrightarrow{\text{a.s.}} 0. \tag{7.4}$$

Arguably, these conditions for consistency are quite mild. In contrast, [LW+82, Theorem 3] also show that asymptotic normality of the least squares estimator in the adaptive linear regression model holds under a stability condition that is substantially more restrictive—namely, the existence of a sequence $\{\mathbf{B}_n\}_{n \geq 1}$ of *non-random* strictly positive definite matrices such that

$$\mathbf{B}_n^{-1} \mathbf{S}_n \xrightarrow{\text{p}} \mathcal{I}. \tag{7.5}$$

Moreover, [LW+82, Example 3] demonstrate through the example of a unit root autoregressive model, that the OLS estimator fails to be asymptotically normal in absence of the stability property (7.5). In such cases, confidence intervals and other forms of inference performed using Gaussian limit theory are no longer valid.

Contributions

In this chapter, we propose and analyze a new family of estimators for the parameter vector θ^* and its coordinates based on online debiasing techniques. We show that,

under mild conditions, our proposed estimators are both asymptotically unbiased and asymptotically normal. The underlying assumptions are less stringent than the stability condition (7.5) and are satisfied by a large class of models for data generation and protocols for choosing covariate vectors. We provide a detailed discussion of three such example classes in Section 7.4. By deriving minimax lower bounds on the performance of any estimator, we show that the confidence intervals obtained using our estimators are asymptotically near-optimal, in that they match the performance of the best possible estimator up to a logarithmic factor.

Related work

The broader literature on bandit algorithms and experimentation focuses mostly on a single statistical objective like minimizing regret or selecting an optimal arm with high probability. In the papers [VBW15; XQL13], the authors empirically observed that bandit algorithms induce bias, which can be problematic for ex-post inference. Later works [Nie+18; SRR19a; SRR19b] characterize the sign and bound the magnitude of this bias. In the paper [Had+19], the authors develop estimators that use propensity scores for the multi-armed bandit setting, a special case of the stochastic regression model (7.1) in which the covariate vectors $\mathbf{x}_i$ are restricted to standard basis vectors. However, it is not clear how to extend this approach to general designs. Also in the bandit setting, Zhang et al. [ZJM20] develop a least squares estimator that exploits an assumed batch structure, meaning that only a fixed, finite number of adaptive decisions are made. This approach, however, does not apply to more general schemes that make adaptive decisions at each round. Recently, Zhang et al. [ZJM] proposed a weighted M-estimator for contextual bandit problems where the bandit algorithm is known. It is also not clear how to generalize this approach to a more general data collection scheme or to the case when the data collection algorithm is not completely known.

There is also a parallel line of work that exploits concentration of measure results (e.g., see the papers [BLM13; Wai19b]) to develop confidence regions that are valid uniformly in time. This approach has its roots in the bandits literature [APS11; Jam+14] and has been refined in more recent work [How+; KK18]. An advantage of this approach is that it yields bounds that are uniform in time. On the flip side, it requires very strong exponential tail conditions on the error sequence in contrast to the relatively mild moment conditions that we impose. Overall, we view this line of work as being complementary to our goal of developing corrected estimators that obey asymptotic normality.

This chapter builds upon and extends past work, due to a subset of the current authors [DJM19; Des+18], using online debiasing techniques. We begin with a lower bound, stated in Theorem 2, that shows that it is the matrix sequence $\mathbf{S}_n^{-1}$ that controls the fundamental difficulty of the problem. This lower bound motivates the particular form of debiasing proposed in this chapter. The construction used in past work [DJM19; Des+18] is based on a non-adaptive upper bound of the form $\lambda_* \mathcal{I}$,

where the scalar λ_* is chosen to be much larger than $\lambda_{\max}(\mathbf{S}_n^{-1})$ with high probability. By sharp contrast, our analysis instead makes use of an adaptive upper bound that simultaneously respects the structure of $\mathbf{S}_n^{-1}$ and leads to a stable martingale transform; this particular construction and our analysis thereof allows us to obtain sharper guarantees than past work [DJM19; Des+18].

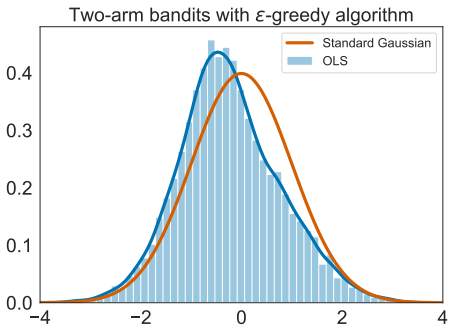
Notation

Let us summarize some notation used throughout the remainder of the chapter. For a positive integer n, we make use of the convenient shorthand $[n] := \{1, 2, \ldots, n\}$. We use $\mathbf{e}_j$ to denote the jth standard basis vector in $\mathbb{R}^D$. For a matrix $\mathbf{M}$, we use the notation $|||\mathbf{M}|||_{\text{op}}$ and $\|\mathbf{M}\|_F$ to denote the operator norm (maximum singular value) and the Frobenius norm of the matrix $\mathbf{M}$, respectively; similarly, we use the notation $\|\mathbf{M}\|_{\max}$ to denote the maximum entry in absolute value. For a square matrix $\mathbf{S}$, the quantities $\lambda_{\max}(\mathbf{S})$ and $\lambda_{\min}(\mathbf{S})$ respectively denote the maximum and minimum eigenvalue of the matrix $\mathbf{S}$. The quantity $\text{tr}(\mathbf{S})$ denotes the sum of diagonal entries of the square matrix $\mathbf{S}$. For a pair of squares matrices $(\mathbf{A}, \mathbf{B})$ of compatible dimensions, we use the notation $\mathbf{A} \succcurlyeq \mathbf{B}$ to indicate that the difference matrix $\mathbf{A} - \mathbf{B}$ is positive semidefinite; we use the notation $\mathbf{A} \succ \mathbf{B}$ when the difference matrix $\mathbf{A} - \mathbf{B}$ is positive definite. The relations $\mathbf{A} \preccurlyeq \mathbf{B}$ and $\mathbf{A} \prec \mathbf{B}$ are defined analogously. For a symmetric positive semidefinite matrix $\mathbf{S}$, we use $\mathbf{S}^{\frac{1}{2}}$ to denote a symmetric matrix square root of the matrix $\mathbf{S}$.

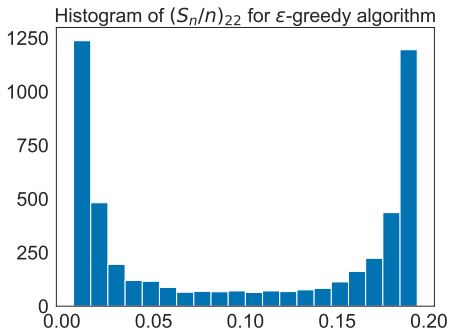
For a sequence of random variables $\{Z_n\}_{n \geq 1}$ and a random variable Z, we use the notation $Z_n \xrightarrow{\text{p}} Z$ to denote that the sequence of random variables $\{Z_n\}_{n \geq 1}$ converges to Z in probability; the notation $Z_n \xrightarrow{\text{d}} Z$ is used to denote convergence in distribution. For a sequence of real-valued random variables $\{Z_n\}_{n \geq 1}$ and a sequence of non-zero real numbers $\{a_n\}_{n \geq 1}$, we say that $Z_n = o_p(a_n)$, if the ratio $\frac{Z_n}{a_n} \xrightarrow{\text{p}} 0$. We use the notation $Z_n = O_p(a_n)$ to mean that the ratio Z_n / a_n is stochastically bounded. More precisely, for every scalar $\epsilon > 0$, there exits a positive real number C_ϵ such that $\sup_{n \geq 1} \mathbb{P}[Z_n / a_n > C_\epsilon] < \epsilon$.

7.2 From ordinary least squares to online debiasing

In this section, we begin by motivating the work by discussing how classical theory about ordinary least squares estimate can break down when data is collected in an adaptive manner. We then introduce a family of online debiasing estimators for computing alternative estimates.



(a) Distribution of $\frac{\widehat{\theta}_2 - \theta_2^*}{\sqrt{(\mathbf{S}_n^{-1})_{22}}}$



(b) Distribution of $(\mathbf{S}_n/n)_{22}$

Figure 7.1. Quantitative behavior of the first coordinate $\widehat{\theta}_1$ of the ordinary least squares estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} := (\widehat{\theta}_1, \widehat{\theta}_2)$ on a dataset drawn from the ε-greedy two-armed bandit model of Section 7.2. The results are obtained with a dataset of size $n = 1000$ and 5000 independent replications. (a) The distribution of $\frac{\widehat{\theta}_2 - \theta_2^*}{\sqrt{(\mathbf{S}_n^{-1})_{22}}}$ is far from standard Gaussian. (b) The bimodal distribution of $(\mathbf{S}_n/n)_{22}$ suggests that the scaled covariance matrix $\mathbf{S}_n/n$ does not converge to a deterministic matrix.

Breakdown of the ordinary least squares estimator

Let us begin by considering the behavior of the OLS estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}$ (7.3). When the covariates $\{\mathbf{x}_i\}_{i\geq 1}$ are either fixed or independently sampled from a fixed distribution, it has several optimality properties. Accordingly, it is natural to ask what the performance of the OLS estimator is when the covariates $\{\mathbf{x}_i\}$ are drawn in an adaptive manner.

In order to fix ideas, let us consider a two-armed bandit problem [LS20], a special case of the linear regression model (7.1) with each $\mathbf{x}_i$ chosen to be either $(1,0)^\top$ or $(0,1)^\top$ based on the prior data $\{\mathbf{x}_j, y_j \mid j \leq i-1\}$. In order to generate the covariates $\{\mathbf{x}_i\}_{i=1}^n$, suppose that we apply the ε-greedy selection algorithm, a popular choice for tackling bandit problems [LS20].

A simple simulation helps to reveal some interesting phenomena. We generated linear regression data using ε-greedy selection algorithm with the choices $\varepsilon = 0.1$, $\theta^* = (0.3, 0.3)^\top$, and noise variables $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Let $\widehat{\theta}_2$ denote the first coordinate of the OLS estimator fit to the bandit data. Figure 7.1(a) demonstrates that the distribution of $\widehat{\theta}_2$, even after proper re-centering and scaling, does not converge to a standard normal distribution. As an undesirable consequence, the confidence intervals for θ_2^*, usually constructed using the quantiles of a standard normal random variable, are not valid.

Let us try to understand why the OLS estimate fails to be asymptotically normal. Figure 7.1(b) plots a histogram of the $(2,2)$ entry of the scaled sample covariance matrix $\mathbf{S}_n/n$. The bimodal behavior suggests that $\mathbf{S}_n/n$ fails to converge to a non-random matrix $\mathbf{B}$ and indicates that the stability condition (7.5) is not satisfied. Indeed, in a recent paper [ZJM20], the authors show that when $\theta_1^* = \theta_2^*$ as in our

example, the OLS estimator, after proper centering and scaling, converges to a distribution which is *not a standard Gaussian* distribution.

It turns out that this distributional anomaly of the OLS estimator is neither specific to the two-armed bandit problem [LS20] nor to the ε-greedy algorithm used to simulate the data for Figure 7.1. The same phenomenon was documented in the time-series and forecasting literature half a century ago, dating back to the works of [DF79; Whi58] and [LW+82]. More recent work [Des+18; ZJM20] has highlighted that a similar phenomenon commonly occurs in multi-armed bandit problems when using popular selection algorithms, including Thompson sampling and the upper confidence bound (UCB) algorithm [LS20].

In Section 7.2, we rectify the distributional anomaly of the OLS estimator by proposing an estimator based on the online debiasing principles of [Des+18] and show that our online debiasing estimator exhibits asymptotic normality even in the absence of the stability condition (7.5). In Section 7.4, we demonstrate the usefulness of our theory via applications to the multi-armed bandit problems, autoregressive time series, and active learning problems with exploration.

Online debiasing estimator

In this section, we propose and analyze an estimator based on an online debiasing technique motivated by the work of [Des+18]. At a high-level, the estimator involves a specific perturbation of the ordinary least squares estimator $\widehat{\boldsymbol{\theta}}_{\text{LS}}$. This perturbation is constructed via a linear combination of the prediction errors $\{y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\theta}}_{\text{LS}}\}_{i=1}^n$ along with a carefully chosen sequence of weight vectors $\{\mathbf{w}_i\}_{i=1}^n$. The key property ensured by the construction is that the weight vector $\mathbf{w}_i$ is $\mathcal{F}_{i-1}$ measurable for each $i \in [n]$.

Concretely, for weight vectors $\{\mathbf{w}_i\}_{i=1}^n$, we compute the *online debiasing estimate*

$$\widehat{\boldsymbol{\theta}}_{\text{OD}} := \widehat{\boldsymbol{\theta}}_{\text{LS}} + \mathbf{S}_n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{w}_i (y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\theta}}_{\text{LS}}). \tag{7.6}$$

Here the reader should recall our earlier definition $\mathbf{S}_n := \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$, and throughout, we assume that the sample covariance $\mathbf{S}_n$ is invertible. The matrix $\mathbf{S}_n^{-\frac{1}{2}}$ denotes a symmetric matrix square root of $\mathbf{S}_n^{-1}$.

Of course, there is an infinite family of estimators of the form (7.6), and the key question is how to define the weight vectors. In this chapter, we propose an estimator in which the sequence $\{\mathbf{w}_i\}_{i=1}^n$ is obtained by solving an optimization problem that takes three inputs:

(i) the original data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$,

(ii) a non-random scalar $\gamma_n \in (0, 1]$, and

(iii) a sequence of symmetric positive semidefinite matrices $\{\boldsymbol{\Gamma}_i\}_{i=1}^n$ such that $\boldsymbol{\Gamma}_i \in \mathcal{F}_{i-1}$ for each $i \in [n]$.

In order to simplify notation, we adopt the shorthand $\mathbf{z}_i := \mathbf{\Gamma}_i^{-\frac{1}{2}} \mathbf{x}_i$. Moreover, for each index $i \in [n]$, we define the matrices

$$\mathbb{Z}_i^\top := \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \cdots & \mathbf{z}_i \end{bmatrix}, \quad \text{and} \quad W_i := \begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \cdots & \mathbf{w}_i \end{bmatrix}$$

We also define $W_0 = 0$ and $\mathbb{Z}_0 = 0$. With these definitions the vectors $\{\mathbf{w}_i\}_{i=1}^n$ are obtained recursively by solving the following convex program

$$\mathbf{w}_i := \arg\min_{\mathbf{w} \in \mathbb{R}^D} \left\{ \|\mathcal{I} - W_{i-1}\mathbb{Z}_{i-1} - \mathbf{w}\mathbf{z}_i^\top\|_F^2 + \frac{\gamma_n}{2}\|\mathbf{w}\|_2^2 \right\}. \tag{7.7a}$$

Conveniently, this optimization problem has the following explicit solution

$$\mathbf{w}_i = \frac{(\mathcal{I} - W_{i-1}\mathbb{Z}_{i-1})\mathbf{z}_i}{(\gamma_n/2) + \|\mathbf{z}_i\|_2^2}. \tag{7.7b}$$

7.3 Main results

Having motivated and introduced the online debiasing approach, we now turn to some theoretical guarantees that can be given for these methods.

We begin in Section 7.3 by providing sufficient conditions for the online debiasing estimator of Section 7.2 to exhibit asymptotically Gaussian behavior (Theorem 1). In Section 7.3, we provide an asymptotically exact confidence region for θ^* as well as an asymptotically exact confidence interval (Proposition 1) for $v^\top\theta^*$, where v is an arbitrary fixed direction $v \in \mathbb{R}^d$. In Section 7.3 (Theorem 2), we complement these results by providing minimax lower bounds on a family of Mahalanobis errors and the length of confidence intervals. These lower bounds apply to any estimator for the stochastic regression model which does not know the true value of the target parameter θ^* but may have the full knowledge of how the data was collected. Finally, in Section 7.3 we provide general strategies which can be used to verify the conditions of Theorem 1. All of our asymptotic statements assume that the dimension D is fixed (constant) while the sample size n grows.

Asymptotic normality guarantees

The main result of this section is an asymptotic normality guarantee for the proposed estimator (7.6), where the weight vectors are defined via the recursion (7.7).

We begin by stating our assumptions and providing some intuition about their role in the theorem.

Assumption A

(A1) There are positive scalars σ and Δ such that the noise sequence $\{\epsilon_i\}_{i=1}^n$ satisfies the conditions $\mathbb{E}[\epsilon_i \mid \mathcal{F}_{i-1}] = 0$ and $\mathbb{E}[\epsilon_i^2 \mid \mathcal{F}_{i-1}] = \sigma^2$ for all $i \in [n]$ and moreover

$$\max_{i \in [n]} \mathbb{E}[\epsilon_i^{2+\Delta} \mid \mathcal{F}_{i-1}] < \infty.$$

(A2) The sequence of matrices $\{\mathbf{S}_n\}_{n\geq 1}$ satisfy the conditions $\lambda_{\min}(\mathbf{S}_n) \xrightarrow{\text{a.s.}} \infty$ and $\frac{\log \lambda_{\max}(\mathbf{S}_n)}{\lambda_{\min}(\mathbf{S}_n)} \xrightarrow{\text{a.s.}} 0$.

(A3) For each n, the scalar $\gamma_n > 0$ and positive semidefinite matrices $\{\mathbf{\Gamma}_i\}_{i=1}^n$ with $\mathbf{\Gamma}_i \in \mathcal{F}_{i-1}$ are chosen such that:

(a) Asymptotic negligibility: $\max_{i\in[n]} \left\{ \frac{1}{\gamma_n} \mathbf{x}_i^\top \mathbf{\Gamma}_i^{-1} \mathbf{x}_i \right\} \xrightarrow{\text{p}} 0,$

(b) Vanishing bias: $\sqrt{\gamma_n \log \lambda_{\max}(\mathbf{S}_n)} \cdot \left\| \mathcal{I} - W_n X_n \mathbf{S}_n^{-\frac{1}{2}} \right\|_{\text{op}} \xrightarrow{\text{p}} 0,$ and

(c) Variance stability: $\left\| \mathcal{I} - \sum_{i=1}^n \mathbf{w}_i \mathbf{x}_i^\top \mathbf{\Gamma}_i^{-\frac{1}{2}} \right\|_{\text{op}} \xrightarrow{\text{p}} 0.$

Let us provide some intuition for the role of each of these assumptions in the theorem. First, Assumption (A1) is quite simple: it imposes relatively mild moment conditions on the noise variables. Second, as discussed in the introduction, Assumption (A2) is standard in guaranteeing the consistency of the least squares estimate. Both Assumptions (A1) and (A2) are viewed as mild conditions in the stochastic linear regression literature and are satisfied by many practical models including those studied in this work [Des+18; LR79; Lai94; LW+82]. Note that Assumptions (A1) and (A2) concern the regression model itself as opposed to the method: in particular, they do not depend on the algorithm parameters γ_n and $\{\mathbf{\Gamma}_i\}_{i=1}^n$.

The more subtle requirements for our theorem to apply, which do depend on the algorithm parameters, are stated in Assumption (A3). We discuss the technical role of these conditions in the comments after Theorem 1 to follow. In Section 7.3 to follow, we provide concrete choices of the algorithm parameters γ_n and $\{\mathbf{\Gamma}_i\}_{i=1}^n$ that ensure that Assumption (A3) holds.

With these preliminaries in place, we are now equipped to state our main theorem on the online debiasing estimator $\widehat{\boldsymbol{\theta}}_{\text{OD}}$:

Theorem 1. *Under Assumptions (A1)–(A3) and given any consistent estimator $\widehat{\sigma}^2$ of σ^2, we have*

$$\sqrt{\frac{\gamma_n}{\widehat{\sigma}^2}} \cdot \mathbf{S}_n^{\frac{1}{2}} (\widehat{\boldsymbol{\theta}}_{\text{OD}} - \theta^*) \xrightarrow{\text{d}} \mathcal{N}(0, \mathcal{I}). \tag{7.9}$$

We prove this theorem in Section 7.5.

A few comments on this theorem are in order. First, it should be noted that needing a consistent estimate for the error variance σ^2 is a mild requirement; under our conditions, it can be obtained using the training mean squared error of the OLS estimate (see Lemma 3 in the paper [LW+82] for details).

A second important fact is that Assumption (A3) is considerably weaker than the stability condition (7.5) required for asymptotic normality of the OLS estimate. To reinforce this point, Section 7.4 provides a detailed discussion of three classes of problems for which OLS fails to be asymptotically normal but the guarantee (7.9) still holds for the online debiasing estimator.

Of all the conditions of Theorem 1, verifying the variance stability condition in Assumption (A3) part (c) is the most challenging, and our arguments for doing so vary from problem to problem. In Corollaries 1 and 2, respectively, we verify the variance stability condition for multi-armed bandit problems and autoregressive time series models. In Corollary 3, we verify this condition for a large class of problems satisfying a sufficient exploration condition.

Let us discuss how Assumption (A3) enters the proof of Theorem 1. Our argument is based on the decomposition

$$\sqrt{\gamma_n} \cdot \mathbf{S}_n^{\frac{1}{2}} (\widehat{\boldsymbol{\theta}}_{\mathrm{OD}} - \theta^*) = \mathbf{b}_n + \mathbf{v}_n, \quad \text{where} \tag{7.10a}$$

$$\mathbf{b}_n := \sqrt{\gamma_n} \cdot \left(\mathcal{I} - W_n X_n \mathbf{S}_n^{-\frac{1}{2}} \right) (\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} - \theta^*) \quad \text{and} \tag{7.10b}$$

$$\mathbf{v}_n := \sqrt{\gamma_n} \cdot \sum_{i=1}^{n} \mathbf{w}_i \epsilon_i. \tag{7.10c}$$

By construction, the term $\mathbf{b}_n$ corresponds to the bias in our estimate, a quantity that must be shown to vanish. In order to do so, we first derive an upper bound on the norm $\|\mathbf{b}_n\|_2$. The "vanishing bias" condition stated in Assumption (A3)(b) enters in showing that, via our choices of the tuning parameters γ_n and $\{\boldsymbol{\Gamma}_i\}_{i=1}^n$, this upper bound converges to zero in probability.

The random vector $\mathbf{v}_n$ defines a zero-mean martingale, and our proof controls its behavior via a standard martingale central limit theorem. Doing so requires a Lindeberg type condition on the weight vectors $\{\mathbf{w}_i\}_{i=1}^n$, as given in part (a) of Assumption (A3). Moreover, it requires that the conditional covariance of the martingale behave suitably, in which context part (c) of Assumption (A3) enters.

Obtaining confidence regions and intervals

In this section, we use the online debiasing procedure to obtain asymptotically exact confidence regions and intervals.

Confidence region for θ^*

First, consider the problem of finding a confidence region for θ^*—that is, a (random) set $\mathbf{A}_{1-\alpha}$ that contains θ^* with probability at least $1-\alpha$. We would like a set that is as small as possible, asymptotically exact in the sense that its coverage converges to $1-\alpha$.

Theorem 1 allows us to construct such a set in the following straightforward way. For any $\alpha \in (0,1)$, consider the subset of $\mathbb{R}^d$ given by

$$\mathbf{A}_{1-\alpha} = \left\{ \theta \in \mathbb{R}^d \mid \frac{\gamma_n}{\widehat{\sigma}^2} \cdot (\widehat{\boldsymbol{\theta}}_{\text{OD}} - \theta)^\top \mathbf{S}_n (\widehat{\boldsymbol{\theta}}_{\text{OD}} - \theta) \leq \chi^2_{d,1-\alpha} \right\}$$

where $\chi^2_{d,1-\alpha}$ denotes the $(1-\alpha)$-quantile for a standard chi-squared distribution with degrees of freedom d. From the result of Theorem 1, we have the guarantee

$$\lim_{n \to \infty} \mathbb{P}(\theta^* \in \mathbf{A}_{1-\alpha}) = 1 - \alpha,$$

In many applications, however, instead of a confidence region for the full vector θ^*, we are instead interested in obtaining a confidence interval for the scalar quantity $v^\top \theta^*$, where $v \in \mathbb{R}^d$ is a fixed direction. It turns out that Theorem 1 no longer provides a straightforward answer to this question. In order to understand why, it is useful to begin by following a naive line of reasoning that is incorrect and then show how it can be fixed.

An incorrect argument

In order to obtain a confidence interval for $v^\top \theta^*$, it might be tempting to "directly invert" the distributional property (7.9). In particular, letting $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ denote the $1-\alpha/2$ quantile of the standard Gaussian distribution, we might claim that the interval

$$\left[v^\top \widehat{\boldsymbol{\theta}}_{\text{OD}} - \tfrac{\widehat{\sigma}}{\sqrt{\gamma_n}} (v^\top \mathbf{S}_n^{-1} v)^{\frac{1}{2}} z_{1-\alpha/2}, \qquad v^\top \widehat{\boldsymbol{\theta}}_{\text{OD}} + \tfrac{\widehat{\sigma}}{\sqrt{\gamma_n}} (v^\top \mathbf{S}_n^{-1} v)^{\frac{1}{2}} z_{1-\alpha/2} \right], \tag{7.11}$$

is an asymptotically exact $1-\alpha$ confidence interval for $v^\top \theta^*$.

Unfortunately, the conclusion (7.11) is based on faulty logic, namely that the asymptotic guarantee (7.9) allows us to write

$$\frac{\sqrt{\gamma_n}}{\widehat{\sigma}} \cdot (\widehat{\boldsymbol{\theta}}_{\text{OD}} - \theta^*) \;\approx\; \mathcal{N}(0, \mathbf{S}_n^{-1}). \tag{7.12}$$

This statement is loose in nature; in the absence of the stability condition (7.5), there is no rigorous and correct form of this statement, since the random matrix $\mathbf{S}_n$ is dependent on the estimate $\widehat{\boldsymbol{\theta}}_{\text{OD}}$. Thus, in absence of any further assumptions on the matrix $\mathbf{S}_n$ or direction vector v, the interval (7.11) is *not* a valid CI for $v^\top \theta^*$.

Nonetheless, there are certain special cases in which the interval (7.11) is a valid CI. Concretely, suppose that $v = \mathbf{e}_j$ is one of the standard coordinate basis vectors and that $\mathbf{S}_n$ is diagonal as in the multi-armed bandit setting studied in Section 7.4. In this case, the calculations of Section 7.8 show that the interval (7.11) is valid. More generally, given an arbitrary direction v, our strategy will be to run a variant of online debiasing that effectively reduces the problem to this favorable case.

Correct fixed-direction confidence intervals

Let us now describe the variant of online debiasing that can be used to obtain asymptotically correct confidence intervals for fixed directions. Let $v \in \mathbb{R}^d$ be the direction of interest; without loss of generality, we assume that $\|v\|_2 = 1$. We now form an orthonormal basis of $\mathbb{R}^d$ with v as its first element—that is, a collection of orthonormal vectors $\{v_1 = v, v_2, \ldots, v_d\}$. Let $\mathbf{V}$ be the matrix with $v_j^\top$ as its j^{th} row. Note that we have $\mathbf{V}\mathbf{V}^\top = \mathcal{I}$ and $\mathbf{V}^\top \mathbf{e}_1 = v$ by construction. Using these two properties, we can rewrite our model as

$$v^\top \theta^* = \mathbf{e}_1^\top \mathbf{V}\theta^* \;\text{ and }\; y_i = \langle \mathbf{V}\mathbf{x}_i,\, \mathbf{V}\theta^* \rangle + \epsilon_i \quad \text{for all } \; i = 1, \ldots, n.$$

Consequently, in this new basis, estimating the scalar $v^\top \theta^*$ is same as estimating the first coordinate of transformed vector $\mathbf{V}\theta^*$.

This fact allows us to define a variant of online debiasing that supports asymptotically exact confidence intervals for $v^\top \theta^*$. In particular, let us introduce the notation

$$\mathbf{x}_{v,i} = \mathbf{V}\mathbf{x}_i, \quad X_{v,n} = X_n \mathbf{V}^\top, \quad \mathbf{S}_{v,n} = (\mathbf{V}\mathbf{S}_n\mathbf{V}^\top)^{-1}, \quad \text{and} \quad \mathbf{\Omega}_{v,n} = \mathbf{S}_{v,n}^{-1}.$$

Define the block diagonal matrix $\mathbf{D}_{v,n}$ as by

$$\mathbf{D}_{v,n}^{\frac{1}{2}} = \begin{pmatrix} \omega_{11}^{-\frac{1}{2}} & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{\Omega}_{22}^{-\frac{1}{2}} \end{pmatrix} \qquad \text{where} \qquad \mathbf{\Omega}_{v,n} = \begin{pmatrix} \omega_{11} & \mathbf{\Omega}_{12} \\ \mathbf{\Omega}_{21} & \mathbf{\Omega}_{22} \end{pmatrix}, \tag{7.13}$$

where the matrix $\mathbf{\Omega}_{22}^{-\frac{1}{2}}$ denotes the symmetric square root of the matrix $\mathbf{\Omega}_{22}^{-1}$. Now consider the estimator

$$\widehat{\boldsymbol{\theta}}_{v,\text{diagOD}} := \widehat{\boldsymbol{\theta}}_{v,\text{LS}} + \beta_n \cdot \mathbf{D}_{v,n}^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{w}_i (y_i - \mathbf{x}_{v,i}^\top \widehat{\boldsymbol{\theta}}_{v,\text{LS}}), \tag{7.14}$$

where $\beta_n := \|\mathbf{D}_{v,n}^{\frac{1}{2}} \mathbf{S}_{v,n}^{-\frac{1}{2}}\|_{\text{op}}$ is a scalar which is at least 1 by definition, and $\widehat{\boldsymbol{\theta}}_{v,\text{LS}} := \mathbf{S}_{v,n}^{-1} X_{v,n}^\top y$ is the OLS estimator using the data $\{y_i, \mathbf{V}\mathbf{x}_i\}_{i=1}^n$. We analyze the behavior of $\widehat{\boldsymbol{\theta}}_{v,\text{diagOD}}$ under the following variant of Assumption (A3).

Assumption (A3)′

(A3)′ For each n, the scalar $\gamma_n > 0$ and positive semidefinite matrices $\{\mathbf{\Gamma}_i\}_{i=1}^n$ with $\mathbf{\Gamma}_i \in \mathcal{F}_{i-1}$ are chosen such that:

(a) Asymptotic negligibility: $\max_{i \in [n]} \left\{ \frac{1}{\gamma_n} \mathbf{x}_{v,i}^\top \mathbf{\Gamma}_i^{-1} \mathbf{x}_{v,i} \right\} \xrightarrow{\text{p}} 0,$

(b) Vanishing bias: $\sqrt{\gamma_n \log \lambda_{\max}(\mathbf{S}_{v,n})} \cdot \| \frac{1}{\beta_n} \cdot \mathbf{D}_{v,n}^{\frac{1}{2}} \mathbf{S}_{v,n}^{-\frac{1}{2}} - W_n X_{v,n} \mathbf{S}_{v,n}^{-\frac{1}{2}} \|_{\text{op}} \xrightarrow{\text{p}} 0,$

(c) Variance stability: $\| \mathcal{I} - \sum_{i=1}^n \mathbf{w}_i \mathbf{x}_{v,i}^\top \mathbf{\Gamma}_i^{-\frac{1}{2}} \|_{\text{op}} \xrightarrow{\text{p}} 0.$

Proposition 1. *Under Assumptions (A1),(A2), and (A3)′, given any consistent estimator $\widehat{\sigma}^2$ of σ^2, the following interval is an asymptotically exact $1-\alpha$ confidence interval for $v^\top \theta^*$*

$$\Big[\mathbf{e}_1^\top \widehat{\boldsymbol{\theta}}_{v,\mathrm{diagOD}} - \tfrac{\beta_n \widehat{\sigma}}{\sqrt{\gamma_n}}(v^\top \mathbf{S}_n^{-1} v)^{\frac{1}{2}} z_{1-\alpha/2}, \quad \mathbf{e}_1^\top \widehat{\boldsymbol{\theta}}_{v,\mathrm{diagOD}} + \tfrac{\beta_n \widehat{\sigma}}{\sqrt{\gamma_n}}(v^\top \mathbf{S}_n^{-1} v)^{\frac{1}{2}} z_{1-\alpha/2}\Big]. \quad (7.16)$$

See Section 7.8 for the proof of this claim.

A few comments regarding Proposition 1 are in order. Observe that the length of the confidence intervals (7.16) matches the length of the confidence interval (7.11) up to a multiplicative factor β_n. Thus, it is interesting to understand the value of the scalar β_n. Note that $\beta_n^2 = \lambda_{\max}(\mathbf{D}_{v,n}^{\frac{1}{2}} \mathbf{S}_{v,n}^{-1} \mathbf{D}_{v,n}^{\frac{1}{2}})$, and a little calculation yields

$$\mathbf{D}_{v,n}^{\frac{1}{2}} \mathbf{S}_{v,n}^{-1} \mathbf{D}_{v,n}^{\frac{1}{2}} = \mathcal{I}_d + \begin{pmatrix} 0 & (\boldsymbol{\Omega}_{22}^{-\frac{1}{2}} \boldsymbol{\Omega}_{21} \omega_{11}^{-\frac{1}{2}})^\top \\ \boldsymbol{\Omega}_{22}^{-\frac{1}{2}} \boldsymbol{\Omega}_{21} \omega_{11}^{-\frac{1}{2}} & \mathbf{0}_{d-1} \end{pmatrix}.$$

Thus, we have

$$1 \le \beta_n^2 = 1 + 2 \cdot \|\boldsymbol{\Omega}_{22}^{-\frac{1}{2}} \boldsymbol{\Omega}_{21} \omega_{11}^{-\frac{1}{2}}\|_2^2, \quad (7.17)$$

which yields $\beta_n \approx 1$ when the vector $\boldsymbol{\Omega}_{21} \approx 0$. To gain further intuition on when $\beta_n \approx 1$, let us assume $v = \mathbf{e}_1$, i.e., we are interested in obtaining a confidence interval for the coordinate θ_1^*. In this case, a natural choice of the basis matrix is $\mathbf{V} = \mathcal{I}$, and as a result, we have $\boldsymbol{\Omega}_{21} = (\mathbf{S}_n^{-1})_{21} = (\mathbf{S}_{21}^{-1}, \ldots, \mathbf{S}_{d1}^{-1})$. Recall that for $j \neq 1$, the entry $\mathbf{S}_{j1}^{-1}$ is proportional to the (empirical) partial correlation coefficient between the first and the j^{th} coordinate, conditioned the remaining $d-2$ coordinates of $\mathbf{x}$; meaning that $\boldsymbol{\Omega}_{21} \approx 0$ when the first coordinate of $\mathbf{x}$ has small correlation with all linear functions of the other $d-1$ coordinates of $\mathbf{x}$.

Finally, we point out that the assumption (A3)′ is not significantly stronger than the original assumption (A3). To fix ideas, we again assume $v = \mathbf{e}_1$ and $\mathbf{V} = \mathcal{I}$. In that case, assumption (A3)′ and (A3) only differ in the vanishing bias condition (b). Assuming $\sqrt{\gamma_n \log \lambda_{\max}(\mathbf{S}_n)} = o_p(1)$ and condition (A3)(b) holds, we have

$$\begin{aligned} &\sqrt{\gamma_n \cdot \log \lambda_{\max}(\mathbf{S}_{\mathbf{e}_1,n})} \cdot \|\tfrac{1}{\beta_n} \cdot \mathbf{D}_n^{\frac{1}{2}} \mathbf{S}_n^{-\frac{1}{2}} - W_n X_n \mathbf{S}_n^{-\frac{1}{2}}\|_{\mathrm{op}} \\ &\le \sqrt{\gamma_n \cdot \log \lambda_{\max}(\mathbf{S}_n)} \cdot \|\tfrac{1}{\beta_n} \cdot \mathcal{I} - W_n X_n \mathbf{S}_n^{-\frac{1}{2}}\|_{\mathrm{op}} \\ &\qquad + \sqrt{\gamma_n \cdot \log \lambda_{\max}(\mathbf{S}_n)} \cdot (1 + \tfrac{1}{\beta_n} \cdot \|\mathbf{D}_n^{\frac{1}{2}} \mathbf{S}_n^{-\frac{1}{2}}\|_{\mathrm{op}}) \\ &= o_p(1) + o_p(1) \xrightarrow{\mathrm{p}} 0. \end{aligned} \quad (7.18)$$

The first step uses the fact $\lambda_{\max}(\mathbf{S}_n) = \lambda_{\max}(\mathbf{S}_{v,n})$ for any basis matrix $\mathbf{V}$. The second step uses the vanishing bias condition (A3)(b), the fact that the dimension d is fixed, and the upper bound $\|\mathbf{D}_n^{\frac{1}{2}} \mathbf{S}_n^{-\frac{1}{2}}\|_{\mathrm{op}} = \beta_n$.

Minimax lower bounds

Thus far, we have derived two guarantees for online debiasing procedures: asymptotic normality in Theorem 1 along with confidence intervals in Proposition 1. It is natural to wonder in what sense these guarantees are optimal. Accordingly, this section is devoted to lower bounds that apply to the performance of *any* estimator $\widehat{\theta}$. These bounds are derived within the classical minimax framework and cover two particular risk measures.

Our first risk measure involves the *Mahalanobis pseudometric*: given an arbitrary positive semidefinite matrix $\mathbf{M}$, possibly random, this pseudometric[1] is given by

$$\|\widehat{\theta} - \theta^*\|_{\mathbf{M}} := \|\mathbf{M}^{\frac{1}{2}}(\widehat{\theta} - \theta^*)\|_2, \tag{7.19}$$

and we provide lower bounds on the squared form of this pseudometric in part (a) of Theorem 2, below. Notably, our analysis allows for the matrix $\mathbf{M}$ to also depend on the dataset $\{\mathbf{x}_i, y_i\}_{i=1}^n$ itself, so that for example, setting $\mathbf{M} = \mathbf{S}_n$ is a valid choice.

Our second risk measure corresponds to the length of a two-sided confidence interval. For a given vector $v \in \mathbb{R}^D$ and significance level $\alpha \in (0, 1)$, let $\mathcal{I}_{\alpha,v} \equiv [\ell_\alpha, u_\alpha] \subseteq \mathbb{R}$ be any level α confidence interval for the scalar $v^\top \theta^*$, so that by definition, we have

$$\mathbb{P}_{\theta^*}\left[v^\top \theta^* \in \mathcal{I}_{\alpha,v}\right] \geq 1 - \alpha \qquad \text{for all} \quad \theta^* \in \mathbb{R}^d. \tag{7.20}$$

We are interested in finding the smallest such confidence interval, and part (b) of Theorem 2 provides a lower bound on its length $|\mathcal{I}_{\alpha,v}| := u_\alpha - \ell_\alpha$.

Our bounds apply to any estimator $\widehat{\theta}$, meaning a measurable function of the data as well as the data collection process. The data collection process is summarized by a collection of (potentially randomized) selection algorithms, each of the form $\psi_i : (\mathbb{R} \times \mathbb{R}^D)^{i-1} \to \mathbb{R}^D$, which take the observed data $\{(\mathbf{x}_j, y_j)\}_{j=1}^{i-1}$ up to time i and output a new observation $\mathbf{x}_i$. With a slight abuse of notation, we refer to $\Psi_n := (\psi_i)_{i \in [n]}$ as the *selection algorithm* of the data collection process.

Theorem 2. *Fix any selection algorithm Ψ_n. Under the linear model (7.1) with i.i.d. Gaussian noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and data collected using Ψ_n, the following results hold.*

(a) For any choice of $\mathbf{M}$ such that $\mathbb{E}[\mathrm{tr}(\mathbf{S}_n^{-1}\mathbf{M})]$ is finite, we have

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \mathbb{R}^d} \mathbb{E}\|\widehat{\theta} - \theta^*\|_{\mathbf{M}}^2 \geq \sigma^2 \mathbb{E}[\mathrm{tr}(\mathbf{S}_n^{-1}\mathbf{M})\Big], \tag{7.21a}$$

where the infimum is taken over any estimator $\widehat{\theta}$ of θ^*, potentially depending on the selection algorithm Ψ_n.

[1] We parameterize the Mahalnobis pseudometric slightly differently than standard definitions, using $\mathbf{M}$ as opposed to its inverse for the quadratic form. This is only for notational ease when $\mathbf{M}$ has a non-trivial null space.

(b) If $\mathbb{E}[\mathbf{S}_n]$ exists and is invertible, then, for any direction $v \in \mathbb{R}^D$ and $\alpha \in (0, 1/8)$,

$$\inf_{\mathcal{I}_{\alpha,v}} \sup_{\theta^* \in \mathbb{R}^d} \mathbb{E}\Big[|\mathcal{I}_{\alpha,v}|\Big] \geq \sigma \cdot \Big(\tfrac{1}{2} - 2\alpha\Big) \cdot \Big(v^\top \Big(\mathbb{E}[\mathbf{S}_n]\Big)^{-1} v\Big)^{\frac{1}{2}}, \tag{7.21b}$$

where the infimum is taken over any valid level $1-\alpha$ confidence interval $\mathcal{I}_{\alpha,v}$ for $v^\top\theta^*$ (cf. definition (7.20)), potentially depending on the selection algorithm Ψ_n.

We provide the proofs of Theorem 2(a) and Theorem 2(b) in Sections 7.5 and 7.5, respectively.

Comments on the MSE bound

In order to gain intuition for the MSE bound in part (a), it is helpful to begin with the simplest case—that is, the non-adaptive setting. Let us consider the classical setting of fixed design linear regression, in which the covariates (and hence $\mathbf{S}_n$) are viewed as fixed, and the additive noise is zero-mean Gaussian with variance σ^2. In this case, the standard OLS estimate $\widehat{\boldsymbol{\theta}}_{\text{LS}}$ has a Gaussian distribution $\mathcal{N}(\theta^*, \sigma^2\mathbf{S}_n^{-1})$. Consequently, for any fixed matrix $\mathbf{M}$, we have

$$\mathbb{E}\|\widehat{\theta} - \theta^*\|_{\mathbf{M}}^2 = \sigma^2\mathbb{E}[\text{tr}(\mathbf{S}_n^{-1}\mathbf{M})], \tag{7.22}$$

so that the lower bound (7.21a) is sharp.

Of course, the more substantive content of Theorem 2(a) lies in the fact that it allows for adaptive data collection, along with potentially random choices of $\mathbf{M}$. One interesting choice is the random matrix $\mathbf{M} = \mathbf{S}_n$, for which the bound (7.21a) guarantees that $\mathbb{E}[\|\widehat{\theta} - \theta^*\|_{\mathbf{S}_n}^2] \geq \sigma^2 D$. It is worth comparing this lower bound to Theorem 1. From the arguments used to prove this theorem, and under a *mildly stronger* assumption (A3) where the convergence in distribution conditions are replaced by convergence in L_1, it can be shown (see Section 7.5) that

$$\lim_{n\to+\infty} \gamma_n \|\widehat{\boldsymbol{\theta}}_{\text{OD}} - \theta^*\|_{\mathbf{S}_n}^2 = \sigma^2 d. \tag{7.23}$$

As discussed in the following section (Section 7.4), in many practical problems of interest, the tuning parameter γ_n typically scales logarithmically in the sample size n, and also our choice of the tuning parameters ensure that the aforementioned stronger version of assumption (A3) is satisfied. Consequently, the result (7.23) combined with the lower bound (7.21a), shows that the online debiasing procedure is minimax optimal up to logarithmic factors.

Another interesting choice is the matrix $\mathbf{M} = e_1 e_1^\top$, for which the lower bound (7.21a) guarantees that the minimal mean-squared error for estimating the first coordinate θ_1^* is determined by $\sigma^2\mathbb{E}[(\mathbf{S}_n^{-1})_{11}]$.

Comments on the CI bound

Theorem 2(b) provides a lower bound on the length of the confidence interval $v^\top \theta^*$ which is valid when the data set is collected in an adaptive manner. To the best of our knowledge, this is the first result providing a lower bound on the confidence interval in an adaptive setting.

It is worth comparing this lower bound (7.21b) to the guarantees provided by the CI construction underlying Proposition 1. First, the pre-factor $c = (\frac{1}{2} - 2\alpha)$ is an an artifact of our proof technique. We suspect that it might be possible to remove with a more careful argument. Let us point out a more substantive issue in comparing the two results. From the result (7.21b), any confidence interval satisfies the lower bound

$$\frac{1}{\sigma^2}\mathbb{E}\Big[|\mathcal{I}_{\alpha,v}|^2\Big] \gtrsim v^T\Big(\mathbb{E}[\mathbf{S}_n]\Big)^{-1}v. \tag{7.24a}$$

On the other hand, if we ignore the difference between σ and $\widehat{\sigma}$, assume $\beta_n = 1$ (the multi-armed bandits with $v = \mathbf{e}_i$ for instance), and take a deterministic γ_n, the CI given by Proposition 1 satisfies

$$\frac{\gamma_n}{4\sigma^2 z^2_{1-(\alpha/2)}}\mathbb{E}\Big[|\mathcal{I}_{\alpha,v}|^2\Big] = v^T\mathbb{E}\Big[\mathbf{S}_n^{-1}\Big]v \overset{(i)}{\geq} v^T\Big(\mathbb{E}[\mathbf{S}_n]\Big)^{-1}v, \tag{7.24b}$$

where inequality (i) follows from Jensen's inequality. We suspect that the lower bound (7.21b) is *not sharp* and that the length of optimal CIs should depend on $v^\top\mathbb{E}\Big[\mathbf{S}_n^{-1}\Big]v$. However, this conjecture remains to be verified.

Choices of the tuning parameters

Let us now return to the practical issue of choosing the tuning parameters γ_n and $\{\mathbf{\Gamma}_i\}_{i=1}^n$ of our debiasing procedures. In particular, these parameters must be chosen appropriately so as to ensure that either Assumption (A3), or its variant in Assumption (A3)′, is satisfied. In this section, we explore a class of feasible choices of these parameters.

In particular, we analyze choices that are based upon on a deterministic matrix $\mathbf{L}_n$ that acts as a lower bound to the sample covariance matrix $\mathbf{S}_n$. Let $\{\mathbf{L}_n\}_{n\geq 1}$ be a sequence of $D \times D$ diagonal matrices with nonnegative entries satisfying the conditions

$$\|\mathbf{L}_n^{\frac{1}{2}}\operatorname{diag}(\mathbf{S}_n^{-1})\mathbf{L}_n^{\frac{1}{2}}\|_{\text{op}} = O_p(1) \quad \text{and} \quad \lambda_{\min}(\mathbf{L}_n) \xrightarrow{\text{a.s.}} \infty. \tag{7.25}$$

For a given n, we define a collection of (diagonal) scaling matrices

$$\mathbf{\Gamma}_{i,n} := \max\left\{\operatorname{diag}\left(\mathbf{S}_i^{-1}\right)^{-1}, \mathbf{L}_n\right\} \tag{7.26a}$$

where $\max\{\cdot,\cdot\}$ denotes the element-wise maximum operator.[2] Next, we choose a sequence of tuning parameters $\{\gamma_i\}_{i=1}^n$ such that

$$\max_{i\in[n]} \quad \frac{1}{\gamma_n}\mathbf{x}_i^\top \mathbf{L}_n^{-1}\mathbf{x}_i \xrightarrow{\text{p}} 0. \tag{7.26b}$$

We point out that it is relatively straightforward to find a diagonal matrix $\mathbf{L}_n$ satisfying the condition (7.25). To fix ideas, let us assume that the covariates satisfy $\|\mathbf{x}_i\|_2^2 \leq 1$ for all $i \in [n]$ and that the minimum eigenvalue of the matrix $\mathbf{S}_n$ satisfies $\lambda_{\min}(\mathbf{S}_n) \geq (\log n)^{1+2\delta}$ with high probability, for some scalar $\delta > 0$; the last condition on $\lambda_{\min}(\mathbf{S}_n)$ is only slightly stronger than the minimum eigenvalue condition in Assumption (A2). Then, with the choice $\mathbf{L}_n = (\log n)^{1+2\delta} \cdot \mathcal{I}$, the condition (7.25) is satisfied. Moreover, in this case, the condition (7.26b) is satisfied for $\gamma_n = \frac{1}{(\log n)^{1+\delta}}$.

Proposition 2. *Consider the solutions $\{\mathbf{w}_i\}_{i=1}^n$ obtained from the optimization problem (7.7a) using the tuning parameters γ_n and $\{\mathbf{\Gamma}_{i,n}\}_{i=1}^n$ defined in equations (7.26a)–(7.26b). Then we have the operator norm bound*

$$\left\|\mathcal{I} - W_n X_n \mathbf{S}_n^{-\frac{1}{2}}\right\|_{op} = O_p(D^2). \tag{7.27}$$

In particular, if $\gamma_n = o_p\left(D^2 \log \lambda_{\max}(\mathbf{S}_n)\right)$, the vanishing bias and asymptotic negligibility conditions in Assumption (A3) are satisfied.

See Section 7.8 for the proof of this claim.

Sharper bound for multi-armed bandits

The dimension dependence of the upper bound (7.27) can be removed in many concrete applications in which we have additional information about the data generating process.

As one concrete example, in the multi-armed bandit model of the sequel (Section 7.4), the upper bound can be sharpened to

$$\left\|\mathcal{I} - W_n X_n \mathbf{S}_n^{-\frac{1}{2}}\right\|_{\text{op}} = O_p(1). \tag{7.28}$$

See the end of Section 7.8 for a proof of this claim, and see the proofs of the Corollaries 1, 2, and 3 in the sequel for more details.

[2]The choice (7.26a) of scaling matrix is especially easy to understand for multi-armed bandit problems, where the scaling matrix $\mathbf{\Gamma}_{i,n}$ can be written as $\mathbf{\Gamma}_{i,n} = \max\{\mathbf{S}_i, \ \mathbf{L}_n\}$. Assuming that $\mathbf{S}_i = \max\{\mathbf{L}_n, \mathbf{S}_i\}$ for large value of i, we see that the tuning parameter $\mathbf{\Gamma}_{i,n}$ is the sample covariance matrix up to time i. This assumption indeed holds for Corollaries 1– 3 to be presented in the sequel.

7.4 Applications

We next illustrate the concrete consequences of our results in a number of common adaptive learning settings. Sections 7.4 and 7.4 are devoted to multi-armed bandit problems and autoregressive time series models respectively, while Section 7.4 discusses active learning with exploration. We end each section with an empirical evaluation of online debiasing. Specifically, we compare the confidence interval (CI) coverage and width of four methods: our online debiasing estimator (7.6), OLS (7.3) with standard but potentially invalid Gaussian intervals, the W-decorrelation estimator of [Des+18], and a valid CI based on the concentration inequality of [APS11]. We highlight that the CIs for OLS are based on the distributional assumption $\mathbf{S}_n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_{\text{LS}} - \theta^*) \sim \mathcal{N}(0, \mathcal{I})$. This property, while true when covariates are selected $\{\mathbf{x}_i\}_{i=1}^n$ in a non-adaptive manner, need not hold when the covariates $\{\mathbf{x}_i\}_{i=1}^n$ are collected adaptively [Des+18; ZJM20]; as a result, the corresponding CIs need not give the correct coverage. Meanwhile, the valid concentration inequality-based intervals [APS11] are guaranteed to provide at least the nominal coverage but are often unnecessarily wide.

Multi-armed bandits

Consider a multi-armed bandit with D arms indexed by the set $[D] := \{1, \ldots, D\}$. At each time $i \in [n]$, a bandit algorithm selects an arm $k_i \in [D]$ and observes the reward

$$y_i = e_{k_i}^\top \theta^* + \epsilon_i, \tag{7.29}$$

where e_{k_i} is the k_i^{th} basis vector in dimension D and $\theta^* \in \mathbb{R}^D$ is the vector containing the mean rewards of D arms. We assume that the noise sequence $\{\epsilon_i\}_{i=1}^n$ satisfies Assumption (A1). Notably, the multi-armed bandit model (7.29) is a special case of the adaptive linear regression model (7.1) with $\mathbf{x}_i = e_{k_i}$ for each $i \in [n]$.

Since the bandit observation model (7.29) has a simple linear form, the OLS solution $\widehat{\boldsymbol{\theta}}_{\text{LS}}$ is a standard estimate of the reward vector $\theta^* \in \mathbb{R}^D$. As we mentioned earlier, the behavior of the OLS estimate depends on the stability of the matrix $\mathbf{S}_n$; see the covariance stability condition (7.5). In the paper [Des+18], the authors conjectured based on empirical evidence that for various popular data selection algorithms, including the Upper Confidence Bound (UCB), Thompson Sampling, and ε-greedy algorithms (see the book [LS20]), the stability condition (7.5) is not satisfied when there are multiple optimal arms. In recent work, Zhang et al. [ZJM20] established the validity of this conjecture for the two-armed bandit problem: when the two means are equal, then the OLS estimate fails to have a Gaussian limiting distribution.

In sharp contrast to these negative results for OLS, Corollary 1 to follow guarantees that the online debiasing estimator (7.6) is asymptotically normal under a mild assumption on the minimum number of times that each arm is pulled. More precisely, for each arm $k \in [D]$ and round $i \in [n]$, let $N_{k,i}$ denote the number of times k is

pulled in the first i rounds, and define the minimum $N_{\min} := \min_{k \in [d]} N_{k,n}$, and maximum $N_{\max} = \max_{k \in [d]} N_{k,n}$ arm counts. Then the scaled sample covariance is a $D \times D$ diagonal matrix, in which the k^{th} diagonal entry corresponds to the number of times that arm k is pulled within the first i rounds:

$$\mathbf{S}_i = \operatorname{diag}\left(N_{1,i}, \ldots, N_{D,i}\right). \tag{7.30}$$

We assume a lower bound on the minimum number of times that each arm is pulled—namely,

$$N_{\min} \geq (\log n)^{1+2\delta} \quad \text{for some} \quad \delta > 0. \tag{7.31}$$

Moreover, we implement the debiasing estimate (7.6) with the choice of tuning parameters

$$\gamma_n = \frac{1}{(\log n)^{1+\delta}} \quad \text{and} \quad \mathbf{\Gamma}_{i,n} = \max\left\{\mathbf{S}_i, \ (\log n)^{1+2\delta} \cdot \mathcal{I}_D\right\}, \tag{7.32}$$

where $\max\{\cdot, \cdot\}$ denotes the element-wise maximum operator.

Corollary 1. *Suppose the minimum arm pull condition (7.31) and the moment condition (A1) are valid. Then, given any consistent estimate $\widehat{\sigma}^2$ of the error variance σ^2, the estimate $\widehat{\boldsymbol{\theta}}_{\text{OD}}$ obtained using the tuning parameter choices (7.32) satisfies*

$$\left(\widehat{\sigma}^2 \cdot (\log n)^{1+\delta}\right)^{-1/2} \cdot \mathbf{S}_n^{\frac{1}{2}} \left(\widehat{\boldsymbol{\theta}}_{\text{OD}} - \theta^*\right) \xrightarrow{\text{d}} \mathcal{N}(0, \mathcal{I}). \tag{7.33}$$

See Section 7.6 for the proof of this claim. Corollary 1 also enables us to construct asymptotically exact confidence regions for θ^*. Moreover, the sample covariance matrix $\mathbf{S}_n$ is diagonal, and as a result, we can also construct confidence intervals of the coordinates θ_i^*; see the proof of Proposition 1 for details. Finally, for a direction v which is not a standard basis direction, we can obtain an asymptotically exact $1 - \alpha$ confidence interval of $v^\top \theta^*$ using Proposition 1; see the comments following Corollary 3 for further details.

Numerical experiment

Figure 7.2 illustrates the performance of online debiasing with bandit tuning (7.32) and $\delta = 0.05$. Here we consider a two-armed bandit problem (7.29) with arm-mean vector $\theta^* = (0.3, 0.3)^\top$ and i.i.d. standard normal error $\{\epsilon_i\}_{i=1}^n$. The covariates $\{\mathbf{x}_i\}_{i=1}^n$ were generated using the Thompson sampling algorithm [Tho33], and we consider confidence intervals (CIs) for θ_1^*.

We observe first that online debiasing provides appropriate coverage for all confidence levels. Meanwhile, the OLS lower tail interval severely undercovers, and W-decorrelation undercovers for both tails despite having larger widths than online

debiasing. Finally, the concentration CI provides 100% coverage for all confidence levels but yields intervals uniformly larger than the online debiasing CIs. In Section 7.10, we present analogous results for two other popular multi-armed bandit algorithms, the upper confidence bound (UCB) and ε-greedy algorithms.

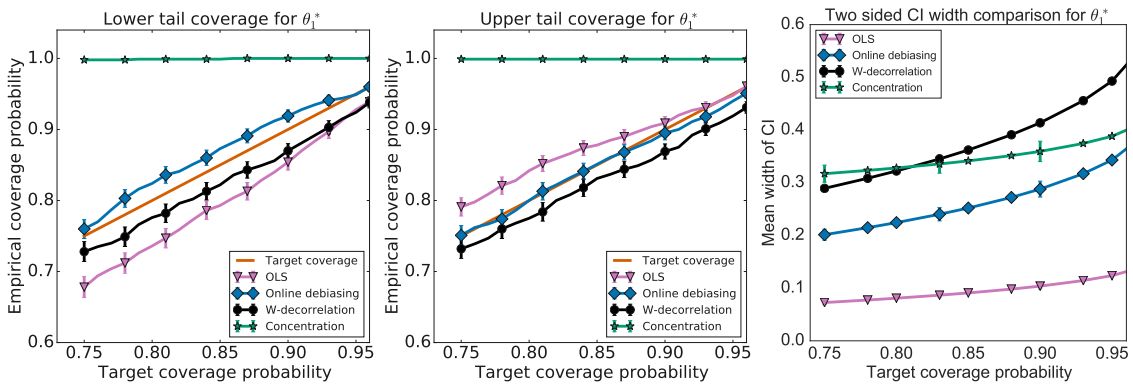


Figure 7.2. Average coverage and width of confidence intervals for θ_1^* across 1000 independent replications of a multi-armed bandit experiment (7.29) with $\theta^* \equiv (\theta_1^*, \theta_2^*) = (0.3, 0.3)^\top$. The covariates $\{\mathbf{x}_i\}_{i=1}^{1000}$ were selected using Thompson sampling [LS20], and the error bars represent ± 1 standard error. **Left** and **Center:** Coverage of one-sided $1-\alpha$ intervals for θ_1^*. **Right:** Width of two-sided $1-\alpha$ intervals for θ_1^*. See Section 7.4 for details.

Autoregressive time series model

Our next example involves estimating the parameters of an autoregressive time series model. It is well-known that the OLS estimate can exhibit non-Gaussian limit behavior for versions of such processes that are unstable [LW+82]. In order to focus attention on the key issues, we restrict ourselves here to the simple case of a scalar autoregressive process.

More precisely, given the initial point $y_0 = 0$ and an unknown scalar $\theta^* \in (-1, 1]$, consider a stochastic process generated by the first-order autoregression

$$y_i = \theta^* y_{i-1} + \epsilon_i \quad \text{for } i = 1, \ldots, n. \tag{7.34}$$

We assume that the noise sequence $\{\epsilon_i\}_{i=1}^n$ consists of i.i.d. standard normal random variables. Note that the autoregression (7.34) is a special case of the stochastic linear regression model (7.1), in particular one with $x_i = y_{i-1}$ for all $i \in [n]$. An especially interesting instantiation of the autoregression (7.34) is obtained by setting $\theta^* = 1$. Such a process is a special case of a *unit root autoregression*, a class of models that play an important role in econometric time series analysis [Box+15].

With the choice $\theta^* = 1$, the process (7.34) is a random walk and so has a variance that grows linearly with time. Moreover, by an application of Donsker's theorem (cf.

Example 3 in the paper [LW+82]), we have

$$\frac{1}{n^2}\sum_{i=1}^{n} x_i^2 := \frac{1}{n^2}\sum_{i=1}^{n} y_{i-1}^2 \xrightarrow{\mathrm{d}} \int_0^1 \mathrm{w}^2(t)dt, \qquad \text{and} \tag{7.35}$$
$$\sqrt{\sum_{i=1}^{n} y_{i-1}^2}\cdot(\theta_{\mathrm{LS}} - \theta^*) \xrightarrow{\mathrm{d}} \frac{\mathrm{w}^2(1)-1}{2\int_0^1 \mathrm{w}^2(t)dt},$$

where $\mathrm{w}(t)$ denotes the standard Wiener process (see the paper [Whi58] for details). Put simply, in the autoregressive time series model (7.34) with $\theta^* = 1$ the stability condition (7.5) is not satisfied, and the distribution of the OLS estimate θ_{LS} is not asymptotically normal.

In contrast to this negative result for the OLS estimate, we can show that the debiasing estimate $\widehat{\theta}_{\mathrm{OD}}$, after suitable centering and scaling, does indeed converge in distribution to a standard Gaussian. Our result is based on the tuning parameters and scaling matrices chosen as

$$\gamma_n = \frac{1}{(\log n)^{1+\delta}}, \quad \text{and} \quad \mathbf{\Gamma}_{i,n} = \max\left\{(\log n)^{1+2\delta} y_{i-1}^2,\ \sum_{j=1}^{i-1} y_j^2\right\}. \tag{7.36}$$

Corollary 2. *Given a sequence $\{y_i\}_{i=1}^n$ generated from the autoregressive model (7.34), the estimate $\widehat{\theta}_{\mathrm{OD}}$ (7.6) obtained with the tuning parameters (7.36) satisfies*

$$\sqrt{\frac{\sum_{i=1}^n y_{i-1}^2}{(\log n)^{1+\delta}}}\cdot(\widehat{\theta}_{\mathrm{OD}} - \theta^*) \xrightarrow{\mathrm{d}} \mathcal{N}(0,1). \tag{7.37}$$

See Section 7.6 for the proof of this claim. Corollary 2 enables us to construct asymptotically exact confidence intervals for θ^*.

Numerical experiment

Figure 7.3 illustrates the performance of online debiasing with autoregression tuning (7.36) and $\delta = 0.05$. Here, our data is generated from the time series model (7.34) with $\theta^* = 1$. We again find that online debiasing provides appropriate coverage for all confidence levels. Meanwhile, the OLS lower tail interval and the W-decorrelation upper tail interval both exhibit severe undercoverage. Finally, the concentration-based CI again provides 100% coverage for all confidence levels, at the expense of interval lengths that are uniformly longer than the online debiasing CIs.

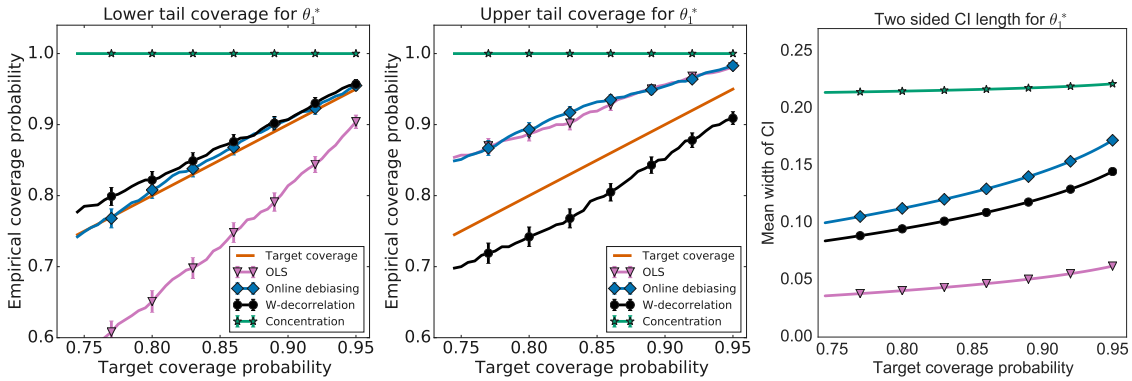


Figure 7.3. Average coverage and width of confidence intervals for $\theta^* = 1$ across 1000 independent replications of an autoregressive time series experiment (7.29). The error bars represent $\pm$ standard error. **Left** and **Center:** Coverage of one-sided $1 - \alpha$ intervals for θ^*. **Right:** Width of two-sided $1 - \alpha$ intervals for θ^*. See Section 7.4 for details.

Active learning with exploration

In our third example, we focus on the case where the covariates $\{\mathbf{x}_i\}_{i=1}^n$ are generated using any algorithm satisfying a sufficient exploration property.

Definition 1 (Selection algorithms with ε-exploration)**.** *We say that a selection algorithm $\Psi_n = \{\psi_i\}_{i=1}^n$ admits a ε-exploration property if*

$$\mathbf{x}_i := \begin{cases} u_i & \text{with probability } 1 - \varepsilon_i \text{ for some } u_i \in \mathcal{F}_{i-1}, \text{ and} \\ v_i & \text{with probability } \varepsilon_i \text{ for some } v_i \text{ independent of } \mathcal{F}_{i-1}. \end{cases} \tag{7.38}$$

Here the exploration probability sequence $\{\varepsilon_i\}_{i=1}^n$ consists of nonnegative scalars in the interval $(0, 1)$, and the vectors $\{v_i\}_{i=1}^n$ are i.i.d. random vectors such that

$$\mathbb{E}[v_i v_i^\top] \succeq \mathbf{G} \qquad \text{where} \quad \mathbf{G} \succeq \mathbf{0} \tag{7.39}$$

In words, the selection algorithm ψ_i behaves as follows: with probability $1 - \varepsilon_i$, it chooses vector u_i based on the previous data points $\{(\mathbf{x}_j, y_j)\}_{j=1}^{i-1}$, and with probability ε_i, it chooses a random direction v_i, independent of the previous data points.

Example 1. **ε-greedy linear bandits.** Let us briefly consider a concrete instance of a selection algorithm $\{\psi_i\}_{i=1}^n$ that is of the ε-greedy type. In the linearly parameterized bandit problem, at each time $i \in [n]$, an algorithm ψ_i chooses a context $\mathbf{x}_i$, usually from a bounded set $\mathcal{U}_i$, and obtains a reward $y_i = \mathbf{x}_i^\top \theta^* + \epsilon_i$. A popular and simple strategy for regret minimization is a special case of the ε-greedy selection algorithm [LS20]. In the linearly parameterized bandit setting, the selection algorithm ψ_i makes the following selection

$$\mathbf{x}_i \begin{cases} \in \arg\max_{\mathbf{x} \in \mathcal{U}_i} \mathbf{x}_i^\top \hat{\theta}_{\text{ridge}}^{(i-1)} & \text{with probability } 1 - \varepsilon_i, \\ \sim \text{Unif}(\mathcal{U}_i) & \text{with probability } \varepsilon_i. \end{cases} \tag{7.40}$$

where, $\hat{\theta}_{\text{ridge}}^{(i-1)}$ denotes a ridge regression estimator which is based on the data up to stage $i-1$, i.e. $\{\mathbf{x}_1, y_1, \ldots, \mathbf{x}_{i-1}, y_{i-1}\}$. Put simply, with probability $1 - \varepsilon_i$, the selection algorithm ψ_i chooses an optimal arm given data collected so far (exploitation), and with probability ε_i the algorithm randomizes uniformly amongst its choices (exploration). In the more general setting (7.38) considered here, it is not necessary to select the optimal arm in the exploitation step. Rather, our result holds also when an arbitrary, $\mathcal{F}_{i-1}$-measurable choice is made in the first part of equation (7.40), as in EXP3 or UCB with exploration [LS20].

Returning to our general setting (7.38), we now state a guarantee for selection algorithms with ε-exploration. As is standard in the bandit literature, we assume that the covariates are uniformly bounded, so that there exists a scalar K satisfying

$$\|\mathbf{x}_i\|_2 \leq K \quad \text{for all } i \in [n]. \tag{7.41a}$$

See our discussion following the corollary for how this condition can be relaxed. In addition, we impose a *sufficient exploration* condition, meaning a lower bound on the magnitude of the exploration probabilities, of the form

$$\sum_{i=1}^{n} \varepsilon_i \geq \frac{\mathbb{E}[\max_{i \in [n]} \|\mathbf{x}_i\|_2^2]}{\lambda_{\min}(\mathbf{G})} (\log n)^{1+2\delta} \quad \text{for some } \delta > 0, \tag{7.41b}$$

where the reader should recall that the matrix $\mathbf{G}$ was defined in equation (7.39). We implement the debiasing estimate (7.6) with the choice of tuning parameters

$$\mathbf{\Gamma}_{i,n} = \sum_{j=1}^{n} \varepsilon_j \mathbf{G} \qquad \text{and} \qquad \gamma_n = \frac{1}{(\log Kn)^{1+\delta}}. \tag{7.41c}$$

Corollary 3. *Suppose that Assumptions (A1) and (A2) hold, the covariates satisfy the bound (7.41a), and the exploration conditions (7.39) and (7.41b) both hold. Then given any consistent estimator $\widehat{\sigma}^2$ of the error variance σ^2, the estimator $\widehat{\boldsymbol{\theta}}_{\text{OD}}$ with tuning parameters (7.41c) satisfies*

$$\left(\log(Kn)^{1+\delta} \cdot \widehat{\sigma}^2 \right)^{-1/2} \cdot \mathbf{S}_n^{\frac{1}{2}} \left(\widehat{\boldsymbol{\theta}}_{\text{OD}} - \theta^* \right) \xrightarrow{\text{d}} \mathcal{N}(0, \mathcal{I}). \tag{7.42a}$$

Moreover, for any $v \in \mathbb{R}^d$, the following is an asymptotically exact $1 - \alpha$ confidence intervals for $v^\top \theta^*$

$$\left[\mathbf{e}_1^\top \widehat{\boldsymbol{\theta}}_{v,\text{diagOD}} - \tfrac{\beta_n \widehat{\sigma}}{\sqrt{\gamma_n}} (v^\top \mathbf{S}_n^{-1} v)^{\frac{1}{2}} z_{1-\alpha/2}, \qquad \mathbf{e}_1^\top \widehat{\boldsymbol{\theta}}_{v,\text{diagOD}} + \tfrac{\beta_n \widehat{\sigma}}{\sqrt{\gamma_n}} (v^\top \mathbf{S}_n^{-1} v)^{\frac{1}{2}} z_{1-\alpha/2} \right], \tag{7.42b}$$

where $\beta_n = \|\mathbf{D}_{v,n}^{-\frac{1}{2}} \mathbf{S}_{v,n}^{\frac{1}{2}}\|_{op}$ *and the estimator* $\widehat{\boldsymbol{\theta}}_{v,\text{diagOD}}$ *was calculated using* (7.14) *and with the tuning parameters* (7.41c).

See Section 7.6 for the proof.

It is worth noting that the bounded covariate condition (7.41a) can be relaxed. For instance, in absence of the condition (7.41a), one may obtain a result similar to the part (a) of Corollary 3 under the following assumptions:

$$\gamma_n = o_p(\log \lambda_{\max}(\mathbf{S}_n)), \quad \text{and} \quad \sum_{i=1}^{n} \varepsilon_i = \frac{\max_{i\in[n]} \mathbb{E}[\|\mathbf{x}_i\|_2^2]}{\lambda_{\min}(\mathbf{G}) \cdot o_p(\gamma_n)}.$$

Finally, as a special case, Corollary 3 allows us to construct confidence interval for $v^\top \theta^*$ for multi-armed bandit problems that we discussed in Section 7.4. The condition (7.41a) is readily satisfied for multi-armed bandit problems, but the conditions (7.39) and (7.41b) are mildly stronger than the analogous condition (7.32).

Numerical simulation

Figure 7.4 illustrates the performance of online debiasing with the active learning tuning (7.41c) and $\delta = 0.05$. Here we consider a linear bandits problem with $\theta^* = (0.3, 0.3)^\top$ and i.i.d. standard normal error $\{\epsilon_i\}_{i=1}^n$. The covariates $\{\mathbf{x}_i\}_{i=1}^n$ were generated using the ε-greedy linear bandits algorithm (7.40), where, for each stage, the context set $\mathcal{U}_i$ consisted of the same 50 vectors drawn and uniformly from the unit sphere in dimension 2. For this problem, the exploration lower bound equation (7.39) is satisfied with $\mathbf{G} = \frac{1}{|\mathcal{U}|} \cdot \sum_{u_i \in \mathcal{U}} u_i u_i^\top$. In this setting, [APS11] only provide concentration-based CIs based on ridge regression estimators, rather than OLS. Here we report the CIs from ridge regression with regularization parameter $\lambda_{\text{Ridge}} = 0.1$ (which closely approximates the OLS solution) and display analogous results for alternative regularization parameters in Section 7.10. We computed the confidence intervals for θ_1^* and θ_2^* using Corollary 3.

We observe once more that online debiasing provides appropriate coverage for all confidence levels, while the OLS lower tail interval severely undercovers. Meanwhile, the concentration CI provides high coverage for all confidence levels but yields intervals typically larger than the online debiasing CIs.

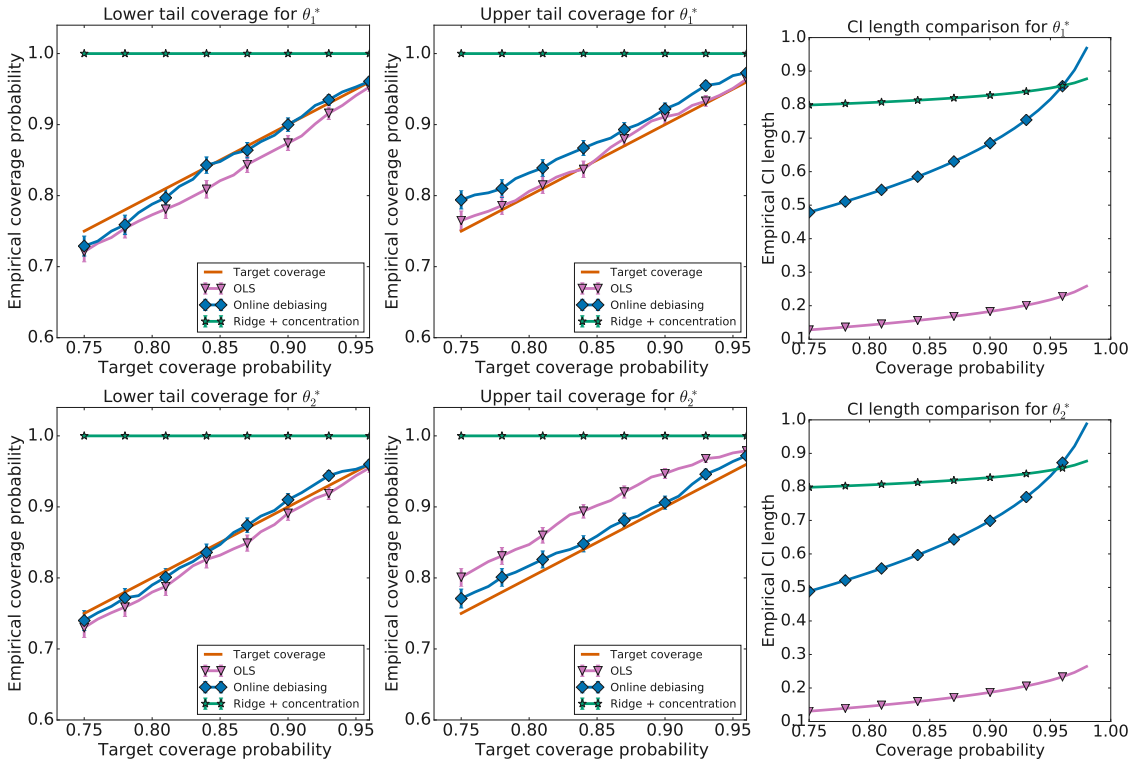


Figure 7.4. Average coverage and width of confidence intervals for θ_1^* and θ_2^* across 1000 independent replications of a linear bandits experiment (7.40) with $\theta^* \equiv (\theta_1^*, \theta_2^*) = (0.3, 0.3)^\top$. The covariates $\{\mathbf{x}_i\}_{i=1}^{1000}$ were selected using the ε-greedy linear bandits algorithm (7.40), and the error bars represent ± 1 standard error. **Left** and **Center:** Coverage of one-sided $1-\alpha$ intervals for θ_1^* and θ_2^*. **Right:** Width of two-sided $1-\alpha$ intervals for θ_1^* and θ_2^*. See Section 7.4 for details.

7.5 Proofs of the theorems

In this section, we provide the proofs of our two main results. We prove Theorem 1 in Section 7.5, and Theorem 2 in Section 7.5.

Proof of Theorem 1

Using the condition $\lambda_{\min}(\mathbf{S}_n) \xrightarrow{\text{a.s.}} \infty$ from assumption (A2), we know that, so that we may assume without loss of generality that $\mathbf{S}_n$ is invertible. We claim that it suffices to show that $\sqrt{\gamma_n} \cdot \mathbf{S}_n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_{\text{OD}} - \theta^*)$ converges in distribution to $\mathcal{N}(0, \sigma^2 \mathcal{I})$. Indeed, when this claim holds, then since $\widehat{\sigma}^2 \xrightarrow{\text{p}} \sigma^2$ by assumption, Slutsky's theorem implies the claim of the theorem.

Recall from equation (7.10) that the random vector $\sqrt{\gamma_n} \cdot \mathbf{S}_n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_{\text{OD}} - \theta^*)$ can be decomposed into the sum $\mathbf{b}_n + \mathbf{v}_n$. Based on this decomposition, we see that it is sufficient to prove that $\mathbf{b}_n \xrightarrow{\text{p}} 0$ and $\mathbf{v}_n \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2 \mathcal{I})$. The remainder of our proof is devoted to establishing these two claims.

Analysis of $\mathbf{b}_n$

By definition of the operator norm, we have the upper bound

$$\|\mathbf{b}_n\|_2 \leq \sqrt{\gamma_n}\, |\!|\!| \mathcal{I} - W_n X_n \mathbf{S}_n^{-\frac{1}{2}} |\!|\!|_{\mathrm{op}} \left\| \mathbf{S}_n^{\frac{1}{2}} (\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} - \theta^*) \right\|_2. \tag{7.43}$$

Theorem 1 from the paper [LW+82] guarantees that

$$\left\| \mathbf{S}_n^{\frac{1}{2}} (\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} - \theta^*) \right\|_2 = \mathcal{O}\left(\sqrt{\log \lambda_{\max}(\mathbf{S}_n)} \right) \quad \text{almost surely.} \tag{7.44a}$$

On the other hand, the vanishing bias condition from Assumption (A3)(b) guarantees that

$$\sqrt{\gamma_n \log \lambda_{\max}(\mathbf{S}_n)} \cdot |\!|\!| \mathcal{I} - W_n X_n \mathbf{S}_n^{-\frac{1}{2}} |\!|\!|_{\mathrm{op}} \xrightarrow{\mathrm{p}} 0. \tag{7.44b}$$

Applying the bounds (7.44a) and (7.44b) to the right-hand side of the inequality (7.43) shows that $\|\mathbf{b}_n\|_2 \xrightarrow{\mathrm{p}} 0$.

Analysis of $\mathbf{v}_n$

In order to control the second term, we seek to apply a classical martingale central limit theorem (cf. Theorem 2.2 in the paper [Dvo+72]). We begin by observing that $\{\sqrt{\gamma_n}\mathbf{w}_i \epsilon_i\}_{i=1}^n$ is a martingale difference sequence with respect to the sigma-field $\{\mathcal{F}_i\}_{i=1}^n$. Noting that the tuning parameter γ_n is non-random, it follows that the sum $\sum_{i=1}^n \sqrt{\gamma_n}\mathbf{w}_i \epsilon_i$ has zero mean and moreover that

$$\sum_{i=1}^n \mathrm{Cov}\left[\sqrt{\gamma_n}\mathbf{w}_i \epsilon_i \mid \mathcal{F}_{i-1}\right] = \gamma_n \sum_{i=1}^n \mathbf{w}_i \mathbf{w}_i^\top. \tag{7.45}$$

Consequently, in order to apply the martingale CLT so as to obtain the stated claim, we need to show that

$$\gamma_n \sum_{i=1}^n \mathbf{w}_i \mathbf{w}_i^\top \xrightarrow{\mathrm{p}} \sigma^2 \mathcal{I}_D.$$

Doing so requires the following auxiliary lemma, which characterizes the behavior of the weight vector sequence $\{\mathbf{w}_i\}_{i=1}^n$ constructed in equation (7.7b).

Lemma 6. *Under the Assumption (A3) parts (a) and (c), the sequence of vectors $\{\mathbf{w}_i\}_{i=1}^n$ obtained from equation (7.7b) has the following properties:*

$$\begin{aligned} &\textit{(Stability:)} && \gamma_n \sum_{i=1}^n \mathbf{w}_i \mathbf{w}_i^\top \xrightarrow{\mathrm{p}} \mathcal{I}_p, \quad \textit{and} \\ &\textit{(Vanishing norm:)} && \max_{i \in [n]} \sqrt{\gamma_n} \|\mathbf{w}_i\|_2 \xrightarrow{\mathrm{p}} 0. \end{aligned}$$

See Section 7.9 for the proof of this lemma.

With the above lemma in hand, we now apply a standard martingale central limit theorem[3] to conclude that

$$\sum_{i=1}^{n} \sqrt{\gamma_n}\mathbf{w}_i\epsilon_i \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2\mathcal{I}_D).$$

Putting together the pieces, we conclude that

$$\sqrt{\gamma_n} \cdot \mathbf{S}_n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_{\mathrm{OD}} - \theta) = \mathbf{b}_n + \sum_{i=1}^{n} \sqrt{\gamma_n}\mathbf{w}_i\epsilon_i \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2\mathcal{I}_D),$$

which completes the proof of Theorem 1.

Proof of claim (7.23):

For simplicity, let us assume σ is known. Recalling the decomposition (7.10) we have

$$\begin{aligned}\gamma_n \cdot \|\widehat{\boldsymbol{\theta}}_{\mathrm{OD}} - \theta^*\|_{\mathbf{S}_n}^2 &= \|\mathbf{b}_n\|_2^2 + 2\mathbf{b}_n^\top \mathbf{v}_n + \|\mathbf{v}_n\|_2^2 \\ &\leq \|\mathbf{b}_n\|_2^2 + 2\|\mathbf{b}_n\|_2 \cdot \|\mathbf{v}_n\|_2 + \|\mathbf{v}_n\|_2^2\end{aligned}$$

Invoking the condition $\sqrt{\gamma_n \log \lambda_{\max}(\mathbf{S}_n)} \cdot \|\!|\mathcal{I} - W_n X_n \mathbf{S}_n^{-\frac{1}{2}}|\!\|_{\mathrm{op}} \xrightarrow{L_1} 0$ we immediately have $\|\mathbf{b}_n\|_2^2 \xrightarrow{L_1} 0$. It suffices to show that $\|\mathbf{v}_n\|_2^2 \leq d$ and $\mathbb{E}[\|\mathbf{v}_n\|_2^2] \to d$. Observe that

$$\begin{aligned}\mathbb{E}[\|\mathbf{v}_n\|_2^2] &= \gamma_n \cdot \sum_{i=1}^{n} \mathbb{E}[\epsilon_i^2 \|\mathbf{w}_i\|_2^2] + \sum_{i \neq j} \gamma_n \cdot \mathbb{E}[\epsilon_i \epsilon_j \mathbf{w}_i^\top \mathbf{w}_j] \\ &\overset{(i)}{=} \sum_{i=1}^{n} \gamma_n \cdot \sigma^2 \cdot \mathbb{E}[\|\mathbf{w}_i\|_2^2] + \sum_{i \neq j} \gamma_n \cdot \mathbb{E}[\epsilon_i \epsilon_j \mathbf{w}_i^\top \mathbf{w}_j]\end{aligned}$$

The last line above follows from the assumption $\mathbb{E}[\epsilon_i^2 \mid \mathcal{F}_{i-1}] = \sigma^2$ and the fact (by construction) that $\mathbf{w}_i \in \mathcal{F}_{i-1}$. Taking trace on both sides of equation (7.71) we have that $\gamma_n \cdot \|\mathbf{w}_i\|_2^2 \leq d$, and using the stronger L_1 version of assumption (A3) along with the proof techniques of Lemma 6 we have $\gamma_n \cdot \sum_{i=1}^{n} \mathbb{E}[\|\mathbf{w}_i\|_2^2] \to \sigma^2 d$. It remains to show that $\mathbb{E}[\epsilon_i \epsilon_j \mathbf{w}_i^\top \mathbf{w}_j] = 0$ for all $i \neq j$. Without loss of generality, assume $i < j$. By construction of $\mathbf{w}_i$ and the martingale assumption (A1) of the noise ϵ_i, we have $\{\mathbf{w}_i, \mathbf{w}_j, \epsilon_i\} \in \mathcal{F}_{j-1}$. As a result, we conclude

$$\mathbb{E}[\epsilon_i \epsilon_j \mathbf{w}_i^\top \mathbf{w}_j] = \mathbb{E}\Big[\epsilon_i \cdot \mathbf{w}_i^\top \mathbf{w}_j \mathbb{E}[\epsilon_j \mid \mathcal{F}_{j-1}]\Big] = 0$$

This completes the proof of the claim (7.23).

[3] Concretely, by applying Theorem 2.2 from the paper [Dvo+72], we first show that for any unit vector u, the inner product $\frac{1}{\sigma} u^\top \sum_{i=1}^{n} \sqrt{\gamma_n}\mathbf{w}_i\epsilon_i$ converges to a standard Gaussian.

Proof of Theorem 2

We prove part (a) of Theorem 2 in Section 7.5 and part (b) of Theorem 2 in Section 7.5.

Proof of Theorem 2(a)

Throughout the proof, we use $\widehat{\theta}$ to denote a generic estimator for θ^*. We assume that the estimator $\widehat{\theta}$ is a function only of the n datapoints $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and the n selection algorithms selection algorithm $\Psi_n := (\psi_i)_{i\in[n]}$ and that $\widehat{\theta}$ does not know the value of the true parameter θ^*. Consider any positive semidefinite and potentially data-dependent matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, and define the nonnegative scalar loss function

$$\mathscr{L}(\widehat{\theta}, \theta^*) = (\widehat{\theta} - \theta^*)^\top \mathbf{M}(\widehat{\theta} - \theta^*). \tag{7.46}$$

From minimax to Bayes risk

In terms of the above notations, Theorem 2 (a) posits a lower bound on the minimax risk:

$$\inf_{\widehat{\theta}} \sup_{\theta^*\in\mathbb{R}^d} \mathbb{E}[\mathscr{L}(\widehat{\theta}, \theta^*) \mid \theta^*], \tag{7.47}$$

where in the above expression, we have taken an expectation of the loss $\mathscr{L}(\widehat{\theta}, \theta^*)$ over the randomness in the data $(X_n, \mathbf{y}_n)$ conditioned on θ^*. We establish the lower bound Theorem 2(a) on the minimax risk (7.47) by first lower bounding the minimax risk by the Bayes risk and then providing a lower bound for the Bayes risk. Concretely, we use the inequality

$$\inf_{\widehat{\theta}} \sup_{\theta^*} \mathbb{E}[\mathscr{L}(\widehat{\theta}, \theta^*) \mid \theta^*] \geq \inf_{\widehat{\theta}} \mathbb{E}_{\theta^*, X_n, \mathbf{y}_n} \mathscr{L}(\widehat{\theta}, \theta^*), \tag{7.48}$$

where the expectation $\mathbb{E}_{\theta^*, X_n, \mathbf{y}_n}$ above is taken with respect the joint distribution on $(\theta^*, X_n, \mathbf{y}_n)$; this joint distribution is computed by *assigning* a prior distribution on the parameter θ^*.

Main argument

We claim that it suffices to prove that for any estimator $\widehat{\theta}$

$$\mathbb{E}[\mathscr{L}(\widehat{\theta}, \theta^*) \mid X_n, \mathbf{y}_n] \geq \sigma^2 \operatorname{tr}(\mathbf{M}\mathbf{S}_n^{-1}). \tag{7.49}$$

where the expectation $\mathbb{E}[\cdot \mid X_n, y_n]$ is taken with respect to the conditional distribution $\theta^* \mid X_n, y_n$. Indeed, taking expectation over X_n, y_n yields the desired bound:

$$\begin{aligned} \mathbb{E}_{\theta^*, X_n, \mathbf{y}_n} \mathscr{L}(\widehat{\theta}, \theta^*) &= \mathbb{E}_{X_n, \mathbf{y}_n} \mathbb{E}[\mathscr{L}(\widehat{\theta}, \theta^*) \mid X_n, \mathbf{y}_n] \\ &\geq \sigma^2 \mathbb{E}_{X_n, \mathbf{y}_n} \operatorname{tr}(\mathbf{M}\mathbf{S}_n^{-1}). \end{aligned} \tag{7.50}$$

It remains to prove the bound (7.49).

Proof of bound (7.49)

We complete the proof of this bound by first computing the conditional distribution of $\theta^* \mid X_n, \mathbf{y}_n$ and then providing a lower bound on the conditional expectation of the loss $\mathscr{L}(\widehat{\theta}, \theta^*)$ given the data (X_n, y_n). Concretely, we show that under a prior distribution $\theta^* \sim \mathcal{N}(0, \rho^2 \mathcal{I}_D)$, we have:

$$\begin{aligned} \theta^* \mid X_n, \mathbf{y}_n \sim \mathcal{N}(\mu_n, \mathbf{\Sigma}_n), \quad \text{where} \\ \mu_n = \mathbf{\Sigma}_n X_n \mathbf{y}_n \quad \text{and} \quad \mathbf{\Sigma}_n = (\mathbf{S}_n/\sigma^2 + \mathcal{I}_D/\rho^2)^{-1}. \end{aligned} \tag{7.51}$$

A simple calculation using the above distributional property yields that for any positive semidefinite matrix $\mathbf{M}$ (which may depend on the data (X_n, y_n)), the conditional loss $\mathbb{E}[\mathscr{L}(\widehat{\theta}, \theta^*) \mid X_n, \mathbf{y}_n]$ is minimized[4] when $\widehat{\theta} = \mathbf{\Sigma}_n X_n \mathbf{y}_n$ with a minimum value of $\sigma^2 \operatorname{tr}(\mathbf{M}\mathbf{\Sigma}_n)$. Finally, we are free to choose the value of the prior error variance ρ^2, and letting $\rho^2 \to \infty$ yields the claim (7.49). Now let us prove the claim (7.51).

Proof of claim (7.51)

We complete the proof by induction on the number of datapoints n.

Base case

For $n = 0$, we have

$$\theta^* \mid X_0, \mathbf{y}_0 \equiv \theta^* \sim \mathcal{N}(0, \rho^2 \mathcal{I}_D). \tag{7.52}$$

The last statement above follows from the prior assumption $\theta^* \sim \mathcal{N}(0, \rho^2 \mathcal{I}_D)$, and the triple $(X_0, \mathbf{y}_0, \mathbf{S}_0)$ are defined as zeros of respective dimensions. This proves the statement (7.51) for $n = 0$, and $\mu_0 = 0$, and $\mathbf{\Sigma}_0 = \rho^2 \mathcal{I}_D$.

Inductive step

Assume that the claim (7.51) is true for $n - 1$ for some $n \geq 1$. We will prove that the statement holds true for n. Recall that, the query algorithm $\psi_i : (\mathbb{R} \times \mathbb{R}^D)^{i-1} \to \mathbb{R}^D$ is oblivious to the true value θ^*; thus, the conditional distribution $\mathbf{x}_n | \mathcal{F}_n$ is independent of θ^* (see the discussion before Theorem 2). Furthermore, from the model (7.1) we have that the conditional distribution of $y_n \mid \mathcal{F}_{n-1}, \mathbf{x}_n, \theta^*$ is $\mathcal{N}(\mathbf{x}_n^\top \theta^*, \sigma^2)$, and using the induction hypothesis (7.51) we conclude that $\theta^* \mid X_{n-1}, y_{n-1} \sim \mathcal{N}(\mu_{n-1}, \mathbf{\Sigma}_{n-1})$.

[4]Here, we have assumed that the prior distribution $\theta^* \sim \mathcal{N}(0, \rho^2)$ and the error variance σ^2 are known to the estimator $\widehat{\theta}$; this assumption is justified since without the knowledge of the prior distribution on θ^* and error-variance σ^2, the minimum value of the expected loss $\mathbb{E}[\mathscr{L}(\widehat{\theta}, \theta^*) \mid X_n, \mathbf{y}_n]$ can only increase, which yields a (possibly) stronger lower bound.

With the last three observations in hand, an application of the Bayes theorem yields

$$\begin{aligned}\frac{d\mathbb{P}(\theta^* \mid X_n, \mathbf{y}_n)}{d\mu^d(\theta^*)} &\propto \exp\left\{-\frac{1}{2}(\theta^* - \mu_{n-1})^\top \mathbf{\Sigma}_{n-1}^{-1}(\theta^* - \mu_{n-1})^\top\right\} \\ &\qquad \times \exp\left\{-\frac{1}{2\sigma^2}(y_n - \mathbf{x}_n^\top \theta^*)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\theta^* - \mu_n)^\top \mathbf{\Sigma}_n(\theta^* - \mu_n)^\top\right\}\end{aligned}$$

where $\frac{d\mathbb{P}(\theta^* | X_n, \mathbf{y}_n)}{d\mu^d(\theta^*)}$ denotes the Radon-Nikodym derivative of $\theta^* \mid X_n, y_n$ with respect to the Lebesgue measure $\mu^d(\cdot)$ on $\mathbb{R}^d$, and the pair $(\mu_n, \mathbf{\Sigma}_n)$ satisfies the following equations:

$$\mathbf{\Sigma}_n^{-1} = \mathbf{\Sigma}_{n-1}^{-1} + \frac{\mathbf{x}_n \mathbf{x}_n^\top}{\sigma^2} \quad \text{and} \quad \mu_n = \frac{1}{\sigma^2}\mathbf{\Sigma}_n \sum_{i=1}^{n} \mathbf{x}_i y_i.$$

This completes the proof of the inductive step, and putting together the pieces yields the claim of Theorem 2(a).

Proof of Theorem 2(b)

We use $\mathcal{D}_n = (\mathbf{x}_1, y_1, \ldots, \mathbf{x}_n, y_n)$ to denote the full data up to time step n, and we use $\mathbb{P}_{\theta_0}(\mathcal{D}_n)$ and $\mathbb{P}_{\theta_1}(\mathcal{D}_n)$ to denote the marginal distribution of the random variable $\mathcal{D}_n$ under $\theta^* = \theta_0$ and $\theta^* = \theta_1$. Throughout, the points θ_1, θ_0 are two fixed (non-random) points that are chosen to simplify certain calculations in the proof. Finally, the scalars $\mathrm{TV}(\cdot, \cdot)$ and $\mathrm{KL}(\cdot, \cdot)$, respectively, denote the total variation distance and the Kulback-Liebler distance between two distributions.

Confidence intervals:

We provide a lower bound on the length of $1 - \alpha$ confidence intervals $\mathcal{I}_{\alpha,v}$ for $v^\top \theta$, where $v \in \mathbb{R}^d$ is a fixed unit vector, and $\alpha \in (0, \frac{1}{8})$, is the level of the confidence interval. Concretely, the set of valid confidence intervals is

$$\mathcal{I}_{\alpha,v}(\Theta) = \left\{\mathcal{I}_{\alpha,v} = [l(\mathcal{D}_n), u(\mathcal{D}_n)] \ : \ \inf_{\theta \in \Theta} \mathbb{P}_\theta \left[l(\mathcal{D}_n) \leq v^\top \theta \leq u(\mathcal{D}_n)\right] \geq 1 - \alpha\right\}$$

Here, $\mathbb{P}_\theta(\cdot)$ denotes the conditional distribution of $\mathcal{D}_n$ under θ, and our goal is to provide lower bounds on the expected length of the confidence interval defined as

$$|\mathcal{I}_{\alpha,v}(\Theta)| := \sup_{\theta \in \Theta} \ \mathbb{E}_\theta[u(\mathcal{D}_n) - l(\mathcal{D}_n)] \tag{7.53}$$

Main argument

The proof of the theorem is based on the following lemma proved in Section 7.5.

Lemma 7. *Introduce the shorthand* $a_+ := \max\{a, 0\}$. *Then*

$$|\mathcal{I}_{\alpha,v}(\Theta)| \geq \sup_{\theta_0,\theta_1\in\Theta} \operatorname{abs} v^\top(\theta_0 - \theta_1) \cdot \left(1 - 2\alpha - \left(\frac{(\theta_0 - \theta_1)^\top \mathbb{E}(\mathbf{S}_n)(\theta_0 - \theta_1)}{4\sigma^2}\right)^{\frac{1}{2}}\right)_+$$

Let us complete the proof of Theorem 1(a) using this lemma. We do so by carefully choosing the pair of points (θ_0, θ_1) and then using those values in Lemma 7. We choose

$$\theta_0 \in \mathbb{R}^d \quad \text{and} \quad \theta_1 = \theta_0 + \sigma \cdot \frac{\mathbb{E}(\mathbf{S}_n)^{-1} v}{\|\mathbb{E}(\mathbf{S}_n)^{-\frac{1}{2}} v\|_2} \tag{7.54}$$

where, $\mathbb{E}(\mathbf{S}_n)^{-\frac{1}{2}}$ denotes a positive semidefinite matrix square-root of the matrix $\mathbb{E}(\mathbf{S}_n)^{-1}$. Thus, we conclude

$$\begin{aligned} |\mathcal{I}_{\alpha,v}(\Theta)| &\overset{(i)}{\geq} \sigma \cdot \left(v^\top \mathbb{E}(\mathbf{S}_n)^{-1} v\right)^{\frac{1}{2}} \cdot (1 - 2\alpha - 1/2) \\ &\overset{(ii)}{\geq} (1/2 - 2\alpha) \cdot \sigma \cdot \left(v^\top \mathbb{E}(\mathbf{S}_n)^{-1} v\right)^{\frac{1}{2}} \end{aligned}$$

The inequality (i) above follows by substituting the value of θ_0 and θ_1 in the Lemma 7; inequality (ii) uses the assumption $\alpha \in (0, 1/8)$. it remains to prove the Lemma 7.

Proof of Lemma 7

We claim that it suffices to prove that following two inequalities [5]

$$|\mathcal{I}_{\alpha,v}(\Theta)| \geq \sup_{\theta_0,\theta_1\in\Theta} \operatorname{abs} v^\top(\theta_0 - \theta_1) \cdot (1 - \alpha - \mathrm{TV}(\mathbb{P}_{\theta_0}(\mathcal{D}_n), \mathbb{P}_{\theta_1}(\mathcal{D}_n)))_+, \quad \text{and} \tag{7.55a}$$

$$\mathrm{KL}(\mathbb{P}_{\theta_0}(\mathcal{D}_n), \mathbb{P}_{\theta_1}(\mathcal{D}_n)) = \frac{1}{2\sigma^2} \cdot (\theta_0 - \theta_1)^\top \mathbb{E}(\mathbf{S}_n) \cdot (\theta_0 - \theta_1). \tag{7.55b}$$

Indeed, with the above two bounds at hand, the proof of Lemma 7 follows by invoking Pinsker's inequality [Wai19b]:

$$\mathrm{TV}(\mathbb{P}_{\theta_0}(\mathcal{D}_n), \mathbb{P}_{\theta_1}(\mathcal{D}_n)) \leq \frac{1}{\sqrt{2}} \sqrt{\mathrm{KL}(\mathbb{P}_{\theta_0}(\mathcal{D}_n), \mathbb{P}_{\theta_1}(\mathcal{D}_n))}.$$

It remains to prove relations (7.55a) and (7.55b).

[5] We point out that the bound (7.55a) is not new in the literature. A similar inequality can be found in the paper [CG+17] (see Lemma 1), where the authors used it to provide lower bounds on the confidence intervals for high dimensional linear regression with an *i.i.d.* dataset. To the best of our knowledge, the application of the bound (7.55a) to an non-i.i.d. data is novel.

Proof of the bound (7.55a)

Note that for each $\theta \in \Theta$, the interval $\mathcal{I}_{\alpha,v} := [l(\mathcal{D}_n), u(\mathcal{D}_n)]$ is a valid $1-\alpha$ confidence interval for $v^\top\theta$. In particular, for any fixed pair of points (θ_0, θ_1), we have

$$\begin{aligned}\mathbb{P}_{\theta_0}\ \Big\{v^\top\theta_0 \in [l(\mathcal{D}_n), u(\mathcal{D}_n)]\Big\} &\geq 1-\alpha, \quad \text{and}\\ \mathbb{P}_{\theta_1}\ \Big\{v^\top\theta_1 \in [l(\mathcal{D}_n), u(\mathcal{D}_n)]\Big\} &\geq 1-\alpha.\end{aligned}$$

Using properties of the total variance distance $\mathrm{TV}(\mathbb{P}_{\theta_0}(\mathcal{D}_n), \mathbb{P}_{\theta_1}(\mathcal{D}_n))$ we have

$$\Big|\mathbb{P}_{\theta_0}\ \Big\{v^\top\theta_1 \in [l(\mathcal{D}_n), u(\mathcal{D}_n)]\Big\} - \mathbb{P}_{\theta_1}\ \Big\{v^\top\theta_1 \in [l(\mathcal{D}_n), u(\mathcal{D}_n)]\Big\}\Big| \mathrm{TV}(\mathbb{P}_{\theta_0}(\mathcal{D}_n), \mathbb{P}_{\theta_1}(\mathcal{D}_n))$$

Combining the last two inequalities we have

$$\mathbb{P}_{\theta_0}\Big\{(v^\top\theta_0, v^\top\theta_1) \in [l(\mathcal{D}_n), u(\mathcal{D}_n)]\Big\} \geq 1 - 2\alpha - \mathrm{TV}(\mathbb{P}_{\theta_0}(\mathcal{D}_n), \mathbb{P}_{\theta_1}(\mathcal{D}_n)).$$

Recall that $[l(\mathcal{D}_n), u(\mathcal{D}_n)]$ is an interval in $\mathbb{R}$, and we have

$$\mathbb{P}_{\theta_0}\Big\{u(\mathcal{D}_n) - l(\mathcal{D}_n) \geq |v^\top\theta_0 - v^\top\theta_1|\Big\} \geq 1 - 2\alpha - \mathrm{TV}(\mathbb{P}_{\theta_0}(\mathcal{D}_n), \mathbb{P}_{\theta_1}(\mathcal{D}_n))$$

Putting together the pieces and taking supremum over the pair $(\theta_0, \theta_1) \in \Theta$ yields the bound

$$|\mathcal{I}_{\alpha,v}(\Theta)| \geq |v^\top\theta_0 - v^\top\theta_1| \cdot (1 - 2\alpha - \mathrm{TV}(\mathbb{P}_{\theta_0}(\mathcal{D}_n), \mathbb{P}_{\theta_1}(\mathcal{D}_n)))_+$$

Proof of the relation (7.55b)

The proof of the bound (7.55b) is based on a *divergence decomposition lemma* which is well known in the bandits literature [Aue+95]. Throughout, we use the shorthand $\mathcal{F}_i$ denote the σ-field generated by the data $\{\mathbf{x}_1, y_1, \ldots, \mathbf{x}_i, y_i\}$ up to time i. Let $f0(y_i \mid \mathbf{x}_i)$ and $f1(y_i \mid \mathbf{x}_i)$ denote the Radon-Nykodim derivatives of the conditional distribution $y_i \mid \mathbf{x}_i$ with respect to the Lebesgue measure[6] μ on $\mathbb{R}$, under the θ_0 and θ_1 respectively. For $i = 1, \ldots, n$, let $f0(\mathbf{x}_i \mid \mathcal{F}_{i-1})$ and $f1(\mathbf{x}_i \mid \mathcal{F}_{i-1})$, respectively, denote the Radon-Nykodim derivatives of the conditional distribution $\mathbf{x}_i \mid \mathcal{F}_{i-1}$ with respect to a dominating measure λ under θ_0 and θ_1. Clearly, $\mathbb{P}_{\theta_0}(\mathcal{D}_n)$ is dominated by the product measure $\mu^{\bigotimes n} \times \lambda^{\bigotimes n}$, and we have

$$d\mathbb{P}_{\theta_0}(\mathcal{D}_n) = \prod_{i=1}^{n} f0(y_i \mid \mathbf{x}_i) \cdot f0(\mathbf{x}_i \mid \mathcal{F}_{i-1}) \cdot d\lambda(\mathbf{x}_i) \cdot \mu(y_i)$$

[6] Recall that in the model (7.1) we have $f0(y_i \mid \mathbf{x}_i) \equiv \mathcal{N}(\theta_0^\top\mathbf{x}_i, \sigma^2)$ and $f1(y_i \mid \mathbf{x}_i) \equiv \mathcal{N}(\theta_1^\top\mathbf{x}_i, \sigma^2)$.

where, the step above uses the fact that the distribution of y_i is dependent on the history $\mathcal{F}_{i-1}$ through $\mathbf{x}_i$ only. Similarly, we also have:

$$d\mathbb{P}_{\theta_1}(\mathcal{D}_n) = \prod_{i=1}^{n} f1(y_i \mid \mathbf{x}_i) \cdot f1(\mathbf{x}_i \mid \mathcal{F}_{i-1}) \cdot d\lambda(\mathbf{x}_i) \cdot \mu(y_i)$$

Now, we *assumed* that the query algorithm $\psi_i : (\mathbb{R} \times \mathbb{R}^D)^{i-1} \to \mathbb{R}^D$ which generates $\mathbf{x}_i$ is oblivious towards the true value of θ; as a result, we have

$$f0(\mathbf{x}_i \mid \mathcal{F}_{i-1}) = f1(\mathbf{x}_i \mid \mathcal{F}_{i-1}) \quad \text{for all} \quad i = 1, \ldots, n.$$

With the above observation at hand, the KL-distance $\mathrm{KL}(\mathbb{P}_{\theta_0}(\mathcal{D}_n), \mathbb{P}_{\theta_1}(\mathcal{D}_n))$ can be simplified as follows:

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_{\theta_0}(\mathcal{D}_n), \mathbb{P}_{\theta_1}(Z_n)) &:= \mathbb{E}_{\theta_0}\left[\log \frac{d\mathbb{P}_{\theta_0}(\mathcal{D}_n)}{d\mathbb{P}_{\theta_1}(\mathcal{D}_n)}\right] \\ &= \sum_{i=1}^{n} \mathbb{E}_{\theta_0}\left[\log \frac{f0(y_i \mid \mathbf{x}_i)}{f1(y_i \mid \mathbf{x}_i)}\right] \end{aligned} \tag{7.57}$$

Now,

$$\begin{aligned} \mathbb{E}_{\theta_0}\left[\log \frac{f0(y_i \mid \mathbf{x}_i)}{f1(y_i \mid \mathbf{x}_i)}\right] &= \mathbb{E}_{\mathbf{x}_i}\left[\mathbb{E}^{\theta_0}_{y_i|\mathbf{x}_i} \log \frac{f0(y_i \mid \mathbf{x}_i)}{f1(y_i \mid \mathbf{x}_i)}\right] \\ &= \mathbb{E}_{\mathbf{x}_i}\mathbb{E}^{\theta_0}_{y_i|\mathbf{x}_i}\left[-\frac{1}{2\sigma^2} \cdot (y_i - \mathbf{x}_i^\top \theta_0)^2 + \frac{1}{2\sigma^2} \cdot (y_i - \mathbf{x}_i^\top \theta_1)^2\right] \\ &= \mathbb{E}_{\mathbf{x}_i}\left[\frac{1}{2\sigma^2} \cdot (\mathbf{x}_i^\top(\theta_0 - \theta_1))^2\right] \\ &= \mathbb{E}_{X_n}\left[\frac{1}{2\sigma^2} \cdot (\mathbf{x}_i^\top(\theta_0 - \theta_1))^2\right] \end{aligned}$$

The second equality above follows since under θ_0 we have $y_i \mid \mathbf{x}_i \sim \mathcal{N}(\mathbf{x}_i^\top \theta_0, \sigma^2)$. Substituting the last simplification in the KL distance calculation (7.57) we have

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_{\theta_0}(\mathcal{D}_n), \mathbb{P}_{\theta_1}(\mathbb{Z}_n)) &= \mathbb{E}_{X_n}\left[\frac{1}{2\sigma^2} \cdot \|X_n(\theta_0 - \theta_1)\|_2^2\right] \\ &= \frac{1}{2\sigma^2} \cdot (\theta_0 - \theta_1)^\top \mathbb{E}(\mathbf{S}_n)(\theta_0 - \theta_1). \end{aligned} \tag{7.58}$$

This completes the proof of equation (7.55b).

7.6 Proofs of the Corollaries

We now turn to the proofs of our three corollaries, with Sections 7.6, 7.6, and 7.6 devoted to the proofs of the Corollaries 1, 2 and 3, respectively.

Proof of Corollary 1

In light of Theorem 1, it suffices to verify Assumptions (A1)–(A3). The assumptions stated in Corollary 1 ensure that the error sequence $\{\epsilon_i\}_{i=1}^n$ satisfies Assumption (A1). The growth conditions in Assumption (A2) are satisfied due to the minimum arm-pull assumption (7.31). It remains to verify the three conditions in Assumption (A3).

Beginning with the asymptotic negligibility condition, we have

$$\max_{i\in[n]} \frac{1}{\gamma_n}\mathbf{x}_i^\top \mathbf{\Gamma}_i^{-1}\mathbf{x}_i \leq \frac{1}{\gamma_n}\frac{\max_{i\in[n]}\|\mathbf{x}_i^2\|}{(\log n)^{1+2\delta}} = \frac{1}{(\log n)^{\delta}} \to 0.$$

The first inequality above uses the bound $\mathbf{\Gamma}_i^{-1} \preceq \frac{1}{\log(n)^{1+2\delta}}\cdot\mathcal{I}_D$ (see the definition (7.32)); the second equality uses $\|\mathbf{x}_i\|_2^2 = 1$, and the final step follows by substituting $\gamma_n = 1/(\log n)^{1+\delta}$.

Turning to the vanishing bias condition in (A3), we invoke the operator norm bound (7.28) on the matrix $\mathcal{I}_D - W_n X_n \mathbf{S}_n^{-\frac{1}{2}}$ to find that

$$\begin{aligned}\sqrt{\gamma_n \log \lambda_{\max}(\mathbf{S}_n)}\cdot \|\mathcal{I}_D - W_n X_n \mathbf{S}_n^{-\frac{1}{2}}\|_{\text{op}} &\leq \sqrt{\gamma_n \log n}\cdot O_p(1) \\ &= \frac{1}{(\log n)^{\delta/2}}\cdot O_p(1) \xrightarrow{\text{p}} 0,\end{aligned}$$

where we have used the bound $\lambda_{\max}(\mathbf{S}_n) \leq \text{tr}(\mathbf{S}_n) = n$ in the above derivation.

Finally, we verify the variance stability condition in (A3) with the help of the following lemma

Lemma 8 (Commutative guarantee)**.** *For any collection of matrices* $\{\mathbf{\Gamma}_i^{-\frac{1}{2}}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{\Gamma}_i^{-\frac{1}{2}}\}_{i=1}^n$ *that commute with each other, we have*

$$\|\mathcal{I} - \sum_{i=1}^n \mathbf{w}_i \mathbf{x}_i^\top \mathbf{\Gamma}_i^{-\frac{1}{2}}\|_{op} \leq \exp\left(-\frac{\lambda_{\min}(\sum_{i=1}^n \mathbf{\Gamma}_i^{-\frac{1}{2}}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{\Gamma}_i^{-\frac{1}{2}})}{\gamma_n}\right).$$

See the end of this subsection for the proof of this claim.

Let us complete the proof of Corollary 1 using Lemma 8. In the multi-armed bandit setting of Corollary 1, the matrices $\{\mathbf{\Gamma}_i^{-\frac{1}{2}}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{\Gamma}_i^{-\frac{1}{2}}\}_{i=1}^n$ are all diagonal, and hence they commute. Thus, invoking the operator norm bound from Lemma 8 yields

$$\|\mathcal{I}_D - \sum_{i=1}^n \mathbf{w}_i \mathbf{x}_i^\top \mathbf{\Gamma}_i^{-\frac{1}{2}}\|_{\text{op}} \leq \exp\left(-\frac{\lambda_{\min}(\sum_{i=1}^n \mathbf{\Gamma}_i^{-\frac{1}{2}}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{\Gamma}_i^{-\frac{1}{2}})}{\gamma_n}\right).$$

Recall that in the bandits model (7.29), the matrices $\mathbf{S}_n$ and $\mathbf{x}_i\mathbf{x}_i^\top$ are diagonal. By construction (7.32) and the minimum arm-pull condition (7.31), the tuning matrix $\mathbf{\Gamma}_i$

is also diagonal with diagonal entries upper bounded by the corresponding diagonal entries of the (diagonal) matrix $\mathbf{S}_n$. Combining these two observations we have that $\mathbf{S}_n^{-\frac{1}{2}} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{S}_n^{-\frac{1}{2}} \preceq \mathbf{\Gamma}_i^{-\frac{1}{2}} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{\Gamma}_i^{-\frac{1}{2}}$. Consequently, we find that

$$-\lambda_{\min}(\sum_{i=1}^n \mathbf{\Gamma}_i^{-\frac{1}{2}} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{\Gamma}_i^{-\frac{1}{2}}) \overset{(i)}{\leq} -\lambda_{\min}(\sum_{i=1}^n \mathbf{S}_n^{-\frac{1}{2}} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{S}_n^{-\frac{1}{2}}) = -1,$$

where the final equality follows from the definition of $\mathbf{S}_n$. Substituting the value $\gamma_n = 1/(\log n)^{1+\delta}$ yields

$$\|\!|\mathcal{I}_D - \sum_{i=1}^n \mathbf{w}_i \mathbf{x}_i^\top \mathbf{\Gamma}_i^{-\frac{1}{2}}|\!\|_{\text{op}} \leq \exp(-1/\gamma_n) \;\leq\; \frac{1}{n} \to 0.$$

This verifies the variance stability condition from Assumption (A3), and applying Theorem 1 yields Corollary 1.

The only remaining detail is to prove Lemma 8.

Proof of Lemma 8

For notational convenience, we use the shorthands $\mathbf{z}_i := \mathbf{x}_i \mathbf{\Gamma}_i^{-\frac{1}{2}}$ and $\mathbb{Z}_i^\top := \begin{bmatrix} \mathbf{z}_1 & \cdots & \mathbf{z}_i \end{bmatrix}$, as previously introduced in Section 7.2. Substituting the formula for the weight vector $\mathbf{w}_i$ from equation (7.7b), and performing some algebra yields

$$\begin{aligned}
(\mathcal{I} - W_n \mathbb{Z}_n)^\top (\mathcal{I} - W_n \mathbb{Z}_n) &= \prod_{j=1}^n \left(\mathcal{I}_D - \frac{\mathbf{z}_{n+1-j} \mathbf{z}_{n+1-j}^\top}{\gamma_n + \|\mathbf{z}_{n+1-j}\|^2} \right) \prod_{i=1}^n \left(\mathcal{I}_D - \frac{\mathbf{z}_i \mathbf{z}_i^\top}{\gamma_n + \|\mathbf{z}_i\|^2} \right) \\
&= \exp\left[\sum_{j=1}^n \log\left(\mathcal{I}_D - \frac{\mathbf{z}_{n+1-j} \mathbf{z}_{n+1-j}^\top}{\gamma_n + \|\mathbf{z}_{n+1-j}\|^2} \right) \right. \\
&\qquad \left. + \sum_{i=1}^n \log\left(\mathcal{I}_D - \frac{\mathbf{z}_i \mathbf{z}_i^\top}{\gamma_n + \|\mathbf{z}_i\|^2} \right) \right] \\
&\overset{(i)}{\preccurlyeq} \exp\left(- \sum_{j=1}^n \frac{\mathbf{z}_{n+1-j} \mathbf{z}_{n+1-j}^\top}{\gamma_n + \|\mathbf{z}_i\|^2} - \sum_{i=1}^n \frac{\mathbf{z}_i \mathbf{z}_i^\top}{\gamma_n + \|\mathbf{z}_i\|^2} \right) \\
&\preccurlyeq \exp\left(-2 \cdot \sum_{i=1}^n \frac{\mathbf{z}_i \mathbf{z}_i^\top}{\gamma_n} \right),
\end{aligned}$$

where step (i) above uses the fact that $\exp(\log(1-a)) \leq \exp(-a)$ for any scalar $a < 1$ and that the matrices $\{\mathbf{z}_i \mathbf{z}_i^\top\}_{i \in [n]}$ commute. Via an inductive argument, it can be verified that the entries of the matrix $\mathcal{I}_D - \frac{\mathbf{z}_i \mathbf{z}_i^\top}{\gamma_n + \|\mathbf{z}_i\|^2}$ are all upper bounded by 1. Putting together the pieces, we conclude that the operator norm satisfies the bound

$$\|\!|\mathcal{I} - W_n \mathbb{Z}_n|\!\|_{\text{op}} \leq \exp\left(- \frac{\lambda_{\min}(\sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^\top)}{\gamma_n} \right),$$

as claimed.

Proof of Corollary 2

The proof of this claim is similar to that of Corollary 1; in particular, we need to verify Assumptions (A1)–(A3). Recall that the time series model (7.34) in Corollary 2 is a special case of the stochastic linear regression model (7.1) with $(x_i, y_i) \equiv (y_{i-1}, y_i)$; thus the covariance term based on the data $\{(x_i, y_i)\}_{i\in[n]}$ is given by $\sum_{i=1}^n x_i^2 \equiv \sum_{i=1}^n y_{i-1}^2$. Here we have used the convention $y_0 = 0$.

The moment condition (A1) is satisfied since the additive noise ϵ_i in the autoregressive model (7.34) is assumed to have a standard Gaussian distribution. Before we verify the remaining conditions, it is helpful to deduce a few bounds regarding the sample covariance term $\sum_{i=1}^n y_{i-1}^2$. In particular, we show that for any $\theta^* \in (-1, 1]$, the sample covariance term satisfies the following relations

$$\sum_{i=1}^n y_{i-1}^2 \to \infty \quad \text{almost surely,} \tag{7.59a}$$

$$\log(\sum_{i=1}^n y_{i-1}^2) = O_p(\log n), \quad \text{and} \quad (\log n)^{1+\delta} \sum_{i=1}^n y_{n-1}^2 = O_p(\sum_{i=1}^n y_{i-1}^2), \tag{7.59b}$$

where $\delta > 0$ is a fixed scalar (typically small). We prove these bounds at the end of this sub-section, but let us complete the proof of the Corollary using these bounds.

First, observe that the condition (A2) follows from the growth condition (7.59a), and the asymptotic negligibility condition in (A3) is satisfied by noting that

$$\max_{i\in[n]} \frac{1}{\gamma_n} \cdot \frac{y_{i-1}^2}{\max\{(\log n)^{1+2\delta} y_{i-1}^2, \ \sum_{j=1}^{i-1} y_j^2\}} \leq \frac{(\log n)^{1+\delta}}{(\log n)^{1+2\delta}} \to 0.$$

Next, in order to verify the vanishing bias condition in Assumption (A3), doing a calculation similar to Proposition 2 we find that (see the arguments leading up to bounds (7.66a)–(7.66b) and their proofs)

$$\begin{aligned}
\sqrt{\gamma_n \log(\sum_{i=1}^n y_{i-1}^2)} \cdot \left| 1 - \sum_{i=1}^n \frac{w_i y_{i-1}}{\mathbf{\Gamma}_i^{\frac{1}{2}}} \right| &\overset{(i)}{=} O_p((\log n)^{-\frac{\delta}{2}}) \cdot O_p\left(1 + \sqrt{\frac{\mathbf{\Gamma}_n}{\sum_{i=1}^n y_{i-1}^2}}\right) \\
&\overset{(ii)}{=} O_p((\log n)^{-\frac{\delta}{2}}) \cdot O_p(1) \\
&\xrightarrow{\text{p}} 0,
\end{aligned}$$

where step (i) follows by invoking the first part of the bound (7.59b) and step (ii) uses the second part of equation (7.59b).

Finally, we verify the variance stability condition in (A3) with the help of Lemma 8, as previously stated and proved in the proof of Corollary 1. Note that in dimension $D = 1$, the commutativity condition in Lemma 8 holds trivially. Consequently, we may apply Lemma 8 to the one-dimensional autoregressive model (7.34) so as to

obtain the bound

$$\left| 1 - \sum_{i=1}^{n} \frac{w_i y_{i-1}}{\mathbf{\Gamma}_i^{\frac{1}{2}}} \right| \leq \frac{1}{n}.$$

See the calculations following the statement of Lemma 8 in the proof of Corollary 1 for details on this step.

This verifies the variance stability condition from Assumption (A3), and applying Theorem 1 yields Corollary 2.

The only remaining detail is to prove the bounds (7.59a)–(7.59b).

Proofs of the bounds (7.59a)–(7.59b)

The proof of the first part of the bound (7.59b) follows by invoking Theorem 2 part (i) from the paper [LW+82]. Concretely, in the paper [LW+82], the authors showed that when $|\theta^*| \leq 1$, then there is some constant $a > 0$ such that $y_n = O_p(n^a)$. Thus, we have the relation $\sum_{i=1}^{n} y_{i-1}^2 = O_p(n^{2a+1})$, and first part of the bound (7.59b) follows.

We divide the proof of the remaining bounds into two parts, depending on the value of θ^*.

Case 1

First, suppose that $\theta^* = 1$. Recall that in equation (7.35) we argued that

$$\frac{1}{n^2} \sum_{i=1}^{n} y_{i-1}^2 \xrightarrow{\text{d}} \int_0^1 \mathrm{w}^2(t) dt.$$

In light of the last relation, the growth condition (7.59a) is immediate. For the remaining bounds, note that $y_n := \sum_{i \in [n-1]} \epsilon_i \sim \mathcal{N}(0, n-1)$; thus, for any $\delta > 0$, we have $\frac{1}{n^2} \cdot (\log n)^{1+\delta} y_n^2 \xrightarrow{\text{p}} 0$, and we conclude that

$$\log(n)^{1+\delta} \sum_{i=1}^{n} y_n^2 = O_p\left(\sum_{i=1}^{n} y_{i-1}^2 \right),$$

as claimed.

Case 2

Otherwise, we may assume that $|\theta^*| < 1$, in which case the term $\sum_{i=1}^{n} y_{i-1}^2$ stabilizes [LW+82]; concretely, we have

$$\frac{1}{n} \sum_{i=1}^{n} y_{i-1}^2 \xrightarrow{\text{a.s.}} c, \quad \text{where} \quad c > 0 \ \text{ is a non-random scalar.}$$

The growth condition (7.59a) follows directly from the above relation. Moreover, we have $y_{n-1} = \sum_{i=1}^{n} \theta^{*i} \epsilon_{n-i} \sim \mathcal{N}\left(0, \frac{1}{1-\theta^{*2}}\right)$. Putting these two pieces together yields $(\log n)^{1+\delta} \cdot y_{n-1}^2 = O_p\left(\sum_{i=1}^{n} y_{i-1}^2\right)$, valid for any $\delta > 0$.

Proof of Corollary 3

We obtain the first claim of the Corollary 3 by applying Theorem 1, and the second part of the Corollary 3 follows from Proposition 1. We prove these two parts separately.

Proof of claim (7.42a): In order to apply Theorem 1 to the setup of Corollary 3 it suffices to verify the Assumption (A3). Recall that our choice of scaling $\mathbf{\Gamma}_i = \sum_{j=1}^{n} \varepsilon_j \mathbf{G}$ matrices does not actually vary as a function of the round i. For this reason, we simply write $\mathbf{\Gamma}$ from here onwards.

We begin by verifying the asymptotic negligibility condition in (A3). Observe that

$$\mathbb{E}\left\{\max_{i\in[n]} \quad \frac{1}{\gamma_n} \mathbf{x}_i^\top \mathbf{\Gamma}^{-1} \mathbf{x}_i\right\} \leq \frac{1}{\gamma_n} \cdot \frac{\mathbb{E}\left[\max_{i\in[n]} \|\mathbf{x}_i\|_2^2\right]}{\lambda_{\min}(\mathbf{G}) \left(\sum_{i=1}^{n} \varepsilon_i\right)} \to 0, \tag{7.60}$$

where the first inequality above follows by substituting the value of the scaling matrix $\mathbf{\Gamma}$, and the second step follows by invoking the sufficient exploration condition (7.41b).

Next, we verify the variance stability and vanishing bias conditions in (A3). In doing, we make use of the following auxiliary result:

Lemma 9. *Under the sufficient exploration condition (7.41b), for any tuning parameter $\gamma_n \in (0, 1/(\log Kn)^{1+\delta}]$ and a sufficient large sample size n, we have*

$$\mathbb{E}\Big[\|\mathcal{I} - W_n X_n \mathbf{\Gamma}^{-\frac{1}{2}}\|_F^2\Big] \leq \frac{D}{Kn}.$$

See Section 7.6 for the proof of this lemma.

Taking Lemma 9 as given, we now complete the proof of Corollary 3. Note that the variance stability condition in (A3) follows directly from the Frobenius norm bound in Lemma 9 and by letting the number of datapoints $n \to \infty$, keeping the dimension D fixed.

In order to prove the vanishing bias condition in (A3), we first bound the operator norm of the matrix $\mathcal{I} - W_n X_n \mathbf{S}_n^{-\frac{1}{2}}$:

$$\begin{aligned} \|\mathcal{I} - W_n X_n \mathbf{S}_n^{-\frac{1}{2}}\|_{\text{op}} &\leq 1 + \|W_n X_n \mathbf{\Gamma}^{-\frac{1}{2}}\|_{\text{op}} \cdot \|\mathbf{\Gamma}^{\frac{1}{2}} \mathbf{S}_n^{-\frac{1}{2}}\|_{\text{op}} \\ &= O_p(1), \end{aligned} \tag{7.61}$$

where the derivation above uses the Frobenius norm upper bound from Lemma 9 and the fact that $\|\mathbf{\Gamma}^{\frac{1}{2}} \mathbf{S}_n^{-\frac{1}{2}}\|_{\text{op}} = O_p(1)$ by the choice of the tuning parameter $\mathbf{\Gamma}$; see the

bound (7.39) for instance. Using the last bound on $\|\!|\mathcal{I} - W_n X_n \mathbf{S}_n^{-\frac{1}{2}}|\!\|_{\mathrm{op}}$, we then find that

$$\sqrt{\gamma_n \log \lambda_{\max}(\mathbf{S}_n)} \cdot \|\!|\mathcal{I} - W_n X_n \mathbf{S}_n^{-\frac{1}{2}}|\!\|_{\mathrm{op}} \leq \sqrt{\gamma_n \log \lambda_{\max}(\mathbf{S}_n)} \cdot O_p(1) \xrightarrow{\mathrm{p}} 0,$$

where the last step above utilizes the choice $\gamma_n = o_p(\log Kn)$ and the bound $\lambda_{\max}(\mathbf{S}_n) = O_p(\log Kn)$. (Recall the uniform boundedness assumption (7.41a).) All together, we have verified the assumptions of Theorem 1, so that Corollary 3 follows.

It remains to prove Lemma 9.

Proof of Lemma 9

Throughout this proof, we use the shorthands $\mathbf{z}_i = \mathbf{\Gamma}^{-\frac{1}{2}} \mathbf{x}_i$, $\mathbb{Z}_i^\top = [\mathbf{z}_1, \mathbf{z}_2, \ldots \mathbf{z}_i]$, $W_i = [\mathbf{w}_1, \ldots, \mathbf{w}_i]$ and $\Delta_i := \mathcal{I} - W_i \mathbb{Z}_i$. Substituting the expression (7.7b) for the weight vector $\mathbf{w}_i$ we find that

$$\begin{aligned} \|\Delta_{i-1}\|_F^2 - \|\Delta_i\|_F^2 &= \frac{\gamma_n + \|\mathbf{z}_i\|_2^2}{(\gamma_n/2 + \|\mathbf{z}_i\|_2^2)^2} \operatorname{tr}\left\{\Delta_{i-1} \mathbf{z}_i \mathbf{z}_i^\top \Delta_{i-1}^\top\right\} \\ &\geq \frac{1}{\frac{\gamma_n}{2} + \|\mathbf{z}_i\|_2^2} \operatorname{tr}\left\{\Delta_{i-1} \mathbf{z}_i \mathbf{z}_i^\top \Delta_{i-1}^\top\right\}. \end{aligned} \tag{7.62}$$

In equation (7.60), we proved that the random variable $\frac{1}{\gamma_n} \max_{i \in [n]} \|\mathbf{z}_i\|_2^2$ converges to zero in probability; consequently, we may assume that

$$\mathbb{P}\Big[\max_{i \in n} \|\mathbf{z}_i\|_2^2 \leq \gamma_n/2\Big] \geq \frac{1}{2}$$

for all sufficiently large values of the sample size n. Keeping this in mind, taking expectations conditional on the sigma-field $\mathcal{F}_{i-1}$ on both sides in the inequality (7.62), and using the fact that $\Delta_i \in \mathcal{F}_{i-1}$, we have

$$\mathbb{E}\left[\|\Delta_{i-1}\|_F^2 \mid \mathcal{F}_{i-1}\right] - \mathbb{E}\left[\|\Delta_i\|_F^2 \mid \mathcal{F}_{i-1}\right] \geq \frac{\varepsilon_i}{2\gamma_n \sum_{i=1}^n \varepsilon_i} \mathbb{E}\left[\|\Delta_{i-1}\|_F^2 \mid \mathcal{F}_{i-1}\right].$$

Rearranging the last inequality and using the upper bound $(1-t) \leq \exp(-t)$ for $t \geq 0$ we obtain

$$\mathbb{E}\left[\|\Delta_i\|_F^2 \mid \mathcal{F}_{i-1}\right] \leq \exp\left(\frac{-\varepsilon_i}{2\gamma_n \sum_{i=1}^n \varepsilon_i}\right) \|\Delta_{i-1}\|_F^2.$$

Iterating the last bound n times and removing the conditioning on the sigma filed $\mathcal{F}_{i-1}$, we find that

$$\mathbb{E}\left\{\|\Delta_n\|_F^2\right\} \leq D \exp\left(-\frac{1}{2\gamma_n}\right) \overset{(i)}{\leq} D \exp\left(-\frac{1}{2} \cdot (\log Kn)^{1+\delta}\right) \overset{(ii)}{\leq} \frac{D}{Kn}.$$

Here step (i) follows by using $\gamma_n \leq \frac{1}{(\log Kn)^{1+\delta}}$, and step (ii) holds since the sample size is assumed n to be sufficiently large. This completes the proof of the claim (7.42a).

Proof of claim (7.42b): In order to apply Proposition 1, it suffices to verify condition (A3)′ for $\{\mathbf{V}\mathbf{x}_i\}_{i=1}^n$ with the choice of tuning parameters (7.41c). Note that in the proof of (7.42a), we already verified that conditions (7.39), (7.41a), and (7.41b) ensures that assumption (A3) is satisfied. Fortunately, these three conditions are not affected by the change of basis transformation, and are readily satisfied by the regressors $\{\mathbf{V}\mathbf{x}_i\}_{i=1}^n$.

Indeed, for any orthonormal basis matrix $\mathbf{V}$, via linearity of expectation, we have

$$\mathbb{E}[\mathbf{V}v_i v_i^\top \mathbf{V}] = \mathbf{V}\mathbb{E}[v_i v_i^\top]\mathbf{V}^\top \succeq \mathbf{V}\mathbf{G}\mathbf{V}^\top.$$

Moreover, for any orthonormal basis matrix $\mathbf{V}$ we have

$$\lambda_{\min}(\mathbf{V}\mathbf{G}\mathbf{V}^\top) = \lambda_{\min}(\mathbf{G}) \qquad \text{and} \qquad \|\mathbf{V}\mathbf{x}_i\|_2 = \|\mathbf{x}_i\|_2 \leq K$$

Thus, following a proof similar to (7.42a) we have that the assumption (A3)′ parts (a), (c), and condition (A3) part (b), modified for $\{\mathbf{V}\mathbf{x}_i\}_{i=1}^n$, are satisfied. Finally, from the bounded covariates condition (7.41a) we have $\lambda_{\max}(\mathbf{S}_{v,n}) \leq Kn$, and as a result, $\gamma_n \cdot \log(\lambda_{\max}(\mathbf{S}_n)) = o_p(1)$. Combining this observation with a calculation similar to (7.18) we deduce that the condition (A3)′(b) holds. This completes the proof of the claim (7.42b).

7.7 Discussion

In this chapter, we propose a family of online debiasing estimators for adaptive linear regression and analyze their asymptotic properties. We show that the online debiasing estimator admits a Gaussian limit under considerably weaker conditions than the OLS estimator and highlight practical examples from multi-armed bandits, time series modeling, and active learning in which online debiasing yields asymptotic normality while OLS does not. We also prove a minimax lower bound for the adaptive linear regression model and show that the performance of the online debiasing estimators is optimal up to factor logarithmic in the sample size.

In future work, we would like to more precisely describe the non-asymptotic behavior of the online debiasing estimators; concretely, we would like to investigate the rate of distributional convergence of the online debiasing estimators to the appropriate Gaussian distributions. Finally, the performance of the online debiasing estimators matches the minimax optimal performance up to a logarithmic factor. An open question is whether this logarithmic gap can be closed either by providing a sharper minimax lower bound or by proposing better estimators.

7.8 Proofs of the propositions

In this section, we provide the proofs of our two propositions. The section 7.8 is devoted to the proof of Proposition 1, and the Section 7.8 is devoted to the proof of Proposition 2.

Proof of Proposition 1

The proof of Proposition 1 is based on the following asymptotic normality statement regarding $\widehat{\boldsymbol{\theta}}_{v,\text{diagOD}}$.

Lemma 10. *Under Assumptions (A1),(A2), and (A3)′, given any consistent estimator $\widehat{\sigma}^2$ of σ^2, we have*

$$\sqrt{\frac{\gamma_n}{\beta_n^2 \widehat{\sigma}^2}} \cdot \mathbf{D}_n^{\frac{1}{2}} (\widehat{\boldsymbol{\theta}}_{v,\text{diagOD}} - \mathbf{V}\theta^*) \xrightarrow{\text{d}} \mathcal{N}(0, \mathcal{I}). \tag{7.63}$$

We prove this Lemma shortly, but let us complete the proof of Proposition 1 using Lemma 10. Now, by construction we have $\mathbf{V}\mathbf{V}^\top = \mathcal{I}$ and $\mathbf{V}^\top \mathbf{e}_1 = v$. Using these two properties, we can write

$$\mathbf{e}_1^\top \mathbf{V}\theta^* = v^\top \theta^* \;\text{ and }\; y_i = \langle \mathbf{V}\mathbf{x}_i,\, \mathbf{V}\theta^* \rangle + \epsilon_i \quad \text{for all } \; i = 1, \ldots, n.$$

Consequently, in this new basis, estimating the scalar $v^\top \theta^*$ is same as estimating the first coordinate of transformed vector $\mathbf{V}\theta^*$. Next, by construction of the matrix $\mathbf{V}$, we have

$$\mathbf{e}_1^\top \mathbf{D}_{v,n}^{-1} \mathbf{e}_1 = v^\top \mathbf{S}_n^{-1} v \qquad \text{and} \qquad \beta_n = \|\mathbf{D}_{v,n}^{-\frac{1}{2}} \mathbf{S}_{v,n}^{\frac{1}{2}}\|_{\text{op}}. \tag{7.64}$$

Thus, we deduce

$$\begin{aligned}
\mathbf{e}_1^\top \mathbf{D}_{v,n}^{\frac{1}{2}} (\widehat{\boldsymbol{\theta}}_{v,\text{diagOD}} - \mathbf{V}\theta^*) &= \sqrt{(\mathbf{D}_n)_{11}} \cdot ((\widehat{\boldsymbol{\theta}}_{v,\text{diagOD}})_1 - v^\top \theta^*) \\
&= \sqrt{\frac{1}{(\mathbf{D}_n^{-1})_{11}}} \cdot (\mathbf{e}_1^\top \widehat{\boldsymbol{\theta}}_{v,\text{diagOD}} - v^\top \theta^*) \\
&= \sqrt{\frac{1}{v^\top \mathbf{S}_n^{-1} v}} \cdot (\mathbf{e}_1^\top \widehat{\boldsymbol{\theta}}_{v,\text{diagOD}} - v^\top \theta^*),
\end{aligned}$$

The first equality above follows since the first row of the matrix $\mathbf{D}_{v,n}$ is proportional to $\mathbf{e}_1$ by construction and the fact that $\mathbf{e}_1 \mathbf{V}\theta^* = v^\top \theta^*$. The last line follows from the relation (7.64). Thus, from property (7.63) we deduce

$$\sqrt{\frac{\gamma_n}{\beta_n^2 \sigma^2}} \cdot \sqrt{\frac{1}{v^\top \mathbf{S}_n^{-1} v}} \cdot (\mathbf{e}_1^\top \widehat{\boldsymbol{\theta}}_{v,\text{diagOD}} - v^\top \theta^*) \xrightarrow{\text{d}} \mathcal{N}(0, 1). \tag{7.65}$$

Define, the set $\mathbf{A}_{v,1-\alpha} \subseteq \mathbb{R}$ as

$$\begin{aligned}\mathbf{A}_{v,1-\alpha} &:= \left\{\theta \in \mathbb{R} \mid -z_{1-\alpha/2} \leq \sqrt{\frac{\gamma_n}{\beta_n^2 \widehat{\sigma}^2}} \cdot \sqrt{\frac{1}{v^\top \mathbf{S}_n^{-1} v}} \cdot (\mathbf{e}_1^\top \widehat{\boldsymbol{\theta}}_{v,\text{diagOD}} - \theta) \leq z_{1-\alpha/2}\right\} \\ &\equiv \left[\mathbf{e}_1^\top \widehat{\boldsymbol{\theta}}_{v,\text{diagOD}} - \tfrac{\beta_n \cdot \widehat{\sigma}}{\sqrt{\gamma_n}} (\mathbf{e}_1^\top \mathbf{S}_n^{-1} \mathbf{e}_1)^{\frac{1}{2}} z_{1-\alpha/2}, \quad \mathbf{e}_1^\top \widehat{\boldsymbol{\theta}}_{v,\text{diagOD}} + \tfrac{\beta_n \cdot \widehat{\sigma}}{\sqrt{\gamma_n}} (\mathbf{e}_1^\top \mathbf{S}_n^{-1} \mathbf{e}_1)^{\frac{1}{2}} z_{1-\alpha/2}\right]\end{aligned}$$

where, $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard Gaussian random variable. From the equation (7.65) we have $\lim_{n\to\infty} \mathbb{P}(v^\top \theta^* \in \mathbf{A}_{v,1-\alpha}) = 1-\alpha$, i.e., $\mathbf{A}_{v,1-\alpha}$ is an asymptotically exact $1-\alpha$ confidence intervals for $v^\top \theta^*$. This completes the proof of the Proposition 1. It remains to prove Lemma 10.

Proof of Lemma 10

Observe that

$$y_i = \langle \mathbf{V}\mathbf{x}_i,\, \mathbf{V}\theta^* \rangle + \epsilon_i \quad \text{for all} \;\; i = 1, \ldots, n.$$

The proof of the Lemma 10 is similar to the proof of Theorem 1 but modified for the data $\{\mathbf{V}\mathbf{x}_i, y_i\}_{i=1}^n$ and with θ^* replaced by $\mathbf{V}\theta^*$. Without loss of generality, we assume that σ is known; thus, it suffices to prove $\frac{\sqrt{\gamma_n}}{\beta_n} \cdot \mathbf{D}_{v,n}^{\frac{1}{2}} (\widehat{\boldsymbol{\theta}}_{v,\text{diagOD}} - \theta^*) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2 \mathcal{I})$. Recalling the expression for $\widehat{\boldsymbol{\theta}}_{v,\text{diagOD}}$ from the definition (7.14) we have

$$\begin{aligned}\frac{\sqrt{\gamma_n}}{\beta_n} \cdot \mathbf{D}_{v,n}^{\frac{1}{2}} (\widehat{\boldsymbol{\theta}}_{v,\text{diagOD}} - \mathbf{V}\theta^*) &= \sqrt{\gamma_n} \cdot \sum_{i=1}^n \mathbf{w}_i \epsilon_i \\ &\quad + \sqrt{\gamma_n} \cdot \left(\tfrac{1}{\beta_n} \cdot \mathbf{D}_{v,n}^{\frac{1}{2}} \mathbf{S}_{v,n}^{-\frac{1}{2}} - W_n X_{v,n} \mathbf{S}_{v,n}^{-\frac{1}{2}}\right) \mathbf{S}_{v,n}^{\frac{1}{2}} (\widehat{\boldsymbol{\theta}}_{v,\text{LS}} - \mathbf{V}\theta^*) \\ &= \mathbf{v}_n + \mathbf{b}_n\end{aligned}$$

It remains to prove $\mathbf{b}_n \xrightarrow{\text{p}} 0$ and $\mathbf{v}_n \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2 \mathcal{I})$. Observe that

$$\begin{aligned}\|\mathbf{b}_n\|_2 &\leq \sqrt{\gamma_n} \cdot \|\tfrac{1}{\beta_n} \cdot \mathbf{D}_{v,n}^{\frac{1}{2}} \mathbf{S}_{v,n}^{-\frac{1}{2}} - W_n X_{v,n} \mathbf{S}_{v,n}^{-\frac{1}{2}}\|_{\text{op}} \cdot \|\mathbf{S}_{v,n}^{\frac{1}{2}} (\widehat{\boldsymbol{\theta}}_{v,\text{LS}} - \mathbf{V}\theta^*)\|_2 \\ &\leq \sqrt{\gamma_n \log \lambda_{\max}(\mathbf{S}_{v,n})} \cdot \|\tfrac{1}{\beta_n} \cdot \mathbf{D}_n^{\frac{1}{2}} \mathbf{S}_{v,n}^{-\frac{1}{2}} - W_n X_{v,n} \mathbf{S}_{v,n}^{-\frac{1}{2}}\|_{\text{op}} \\ &\xrightarrow{\text{p}} 0,\end{aligned}$$

where, the second inequality uses Theorem 1 from the paper [LW+82], and the last step uses the vanishing bias condition (A3)′(b).

The analysis of the martingale term $\mathbf{v}_n$ is exactly same as that of Theorem 1 proof. This completes the proof of the Lemma 10.

Proof of Proposition 2

Recalling that $\|\mathbf{M}\|_{\max} := \max_{i,j} |\mathbf{M}_{ij}|$ denotes the maximum absolute entry of a matrix, we claim that it suffices to show that $\|W_n X_n \mathbf{\Gamma}_n^{-\frac{1}{2}}\|_{\max} \leq 4$. Indeed, when this claim holds, we have

$$\begin{aligned}
\|\!|\mathcal{I} - W_n X_n \mathbf{S}_n^{-\frac{1}{2}}|\!\|_{\text{op}} &\leq 1 + |\!|\!| W_n X_n \mathbf{\Gamma}_n^{-\frac{1}{2}} |\!|\!|_{\text{op}} |\!|\!| \mathbf{\Gamma}_n^{\frac{1}{2}} \mathbf{S}_n^{-\frac{1}{2}} |\!|\!|_{\text{op}} \\
&\leq 1 + O_p(\sqrt{D}) |\!|\!| W_n X_n \mathbf{\Gamma}_n^{-\frac{1}{2}} |\!|\!|_{\text{op}} \\
&\leq 1 + O_p(D^2) \|W_n X_n \mathbf{\Gamma}_n^{-\frac{1}{2}}\|_{\max}
\end{aligned}$$

The second last inequality above follows by noting that the diagonal entries of the matrix $\mathbf{\Gamma}_n^{\frac{1}{2}} \mathbf{S}_n^{-1} \mathbf{\Gamma}_n^{\frac{1}{2}}$ is of the order $O_p(1)$; this bound uses the expression of the scaling matrix $\mathbf{\Gamma}_n$ from the definition (7.26a), and the operator-norm bound $\|\mathbf{L}_n^{\frac{1}{2}} \operatorname{diag}(\mathbf{S}_n^{-1}) \mathbf{L}_n^{\frac{1}{2}}\|_{op} = O_p(1)$ from assumption (7.25). The last inequality above follows from the fact that $|\!|\!| \mathbf{A} |\!|\!|_{\text{op}} \leq D^{\frac{3}{2}} \|\mathbf{A}\|_{\max}$, for any D-dimensional matrix $\mathbf{A}$. This completes the proof of Proposition 2. The remainder of the proof is devoted to establishing an upper-bound on the max-norm of the matrix $W_n X_n \mathbf{\Gamma}_n^{-\frac{1}{2}}$. We do so by proving the following upper bounds

$$\|\sum_{i=1}^{k} \mathbf{w}_i \mathbf{x}_i^\top \mathbf{\Gamma}_i^{-\frac{1}{2}}\|_{\max} \leq 2 \quad \text{for} \;\; k = 1, \ldots n, \qquad \text{and} \tag{7.66a}$$

$$\|\sum_{i=1}^{n} \mathbf{w}_i \mathbf{x}_i^\top \mathbf{\Gamma}_i^{-\frac{1}{2}} (\mathcal{I} - \mathbf{\Gamma}_i^{\frac{1}{2}} \mathbf{\Gamma}_n^{-\frac{1}{2}})\|_{\max} \leq 2. \tag{7.66b}$$

Note that a combination of these two bounds implies that $\|W_n X_n \mathbf{\Gamma}_n^{-\frac{1}{2}}\|_{\max} \leq 4$.

Accordingly, the remainder of our proof is devoted to establishing the bounds (7.66a) and (7.66b).

Proof of bound (7.66a)

Using the expression for the weight vector $\mathbf{w}_i$ from equation (7.7b), we have

$$\mathcal{I} - \sum_{i=1}^{k} \mathbf{w}_i \mathbf{x}_i \mathbf{\Gamma}_i^{-\frac{1}{2}} = \prod_{i=1}^{k} \left(\mathcal{I} - \frac{\mathbf{\Gamma}_i^{-\frac{1}{2}} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{\Gamma}_i^{-\frac{1}{2}}}{\frac{\gamma_n}{2} + \|\mathbf{\Gamma}_i^{-\frac{1}{2}} \mathbf{x}_i\|^2} \right).$$

Invoking the lower bound assumption (7.26b) and doing simple algebra, we find that

$$\left|\!\!\left| \mathcal{I} - \frac{\mathbf{\Gamma}_i^{-\frac{1}{2}} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{\Gamma}_i^{-\frac{1}{2}}}{\gamma_n/2 + \|\mathbf{\Gamma}_i^{-\frac{1}{2}} \mathbf{x}_i\|_2^2} \right|\!\!\right|_{\text{op}} \leq 1 \quad \text{for all } i \in [n].$$

Consequently, for all $k \in [n]$ we have the bound

$$\|\sum_{i=1}^{k} \mathbf{w}_i \mathbf{x}_i \mathbf{\Gamma}_i^{-\frac{1}{2}}\|_{\max} \leq \|\!|\sum_{i=1}^{k} \mathbf{w}_i \mathbf{x}_i \mathbf{\Gamma}_i^{-\frac{1}{2}}|\!\|_{\mathrm{op}} \leq 2, \tag{7.67}$$

where, in the last derivation we used the fact that the max-norm of a matrix is upper bounded by the operator norm of that matrix. This completes the proof of the bound (7.66a).

Proof of bound (7.66b)

The proof is this bound exploits the following auxiliary lemma:

Lemma 11. *Consider a non-increasing sequence of nonnegative real numbers* $\{\delta_i\}_{i=1}^n$ *and a sequence of real numbers* $\{a_i\}_{i=1}^n$ *for which there exists a constant* C *such that* $\max_{k\in[n]} |\sum_{i=1}^k a_i| \leq C$. *Then we have*

$$|\sum_{i=1}^{n} a_i \delta_i| \leq C\delta_1. \tag{7.68}$$

We prove this lemma at the end of this subsection.

Taking Lemma 11 as given, let us prove the bound (7.66b). The bounds (7.66a) guarantee that

$$\|\sum_{i=1}^{k} \mathbf{w}_i \mathbf{x}_i^\top \mathbf{\Gamma}_i^{-\frac{1}{2}}\|_{\max} \leq 2 \quad \text{for each } k \in [n].$$

Moreover, by construction, the diagonal entries of the matrix $(\mathcal{I} - \mathbf{\Gamma}_i^{-\frac{1}{2}} \mathbf{\Gamma}_n^{\frac{1}{2}})$, for $i = 1, \ldots, n$, are positive and non-increasing. Thus, we can apply Lemma 11 with the sequence $\{a_i\}_{i=1}^n$ as the entries of the matrix $\mathbf{w}_i \mathbf{x}_i^\top \mathbf{\Gamma}_i^{-\frac{1}{2}}$ and δ_i as the diagonal entries of the (diagonal) matrix $\mathcal{I} - \mathbf{\Gamma}_i^{\frac{1}{2}} \mathbf{\Gamma}_n^{-\frac{1}{2}}$. Invoking Lemma 11 yields

$$\|\sum_{i=1}^{n} \mathbf{w}_i \mathbf{x}_i^\top \mathbf{\Gamma}_i^{-\frac{1}{2}} (\mathcal{I} - \mathbf{\Gamma}_i^{\frac{1}{2}} \mathbf{\Gamma}_n^{-\frac{1}{2}})\|_{\max} \leq 2 \cdot \|\mathcal{I} - \mathbf{\Gamma}_1^{\frac{1}{2}} \mathbf{\Gamma}_n^{-\frac{1}{2}}\|_{\max} \leq 2,$$

where, the last inequality above uses the property that the diagonal matrices $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_n$, by construction, satisfy a positive semidefinite ordering $\mathbf{\Gamma}_1 \preceq \mathbf{\Gamma}_n$. This concludes the proof of bound (7.66b).

It remains to prove the Lemma 11.

Proof of Lemma 11

Let $s_k := \sum_{i=1}^{k} a_i$ denote the k^{th} partial sum of the sequence $\{a_i\}_{i=1}^n$. The sum $\sum_{i=1}^n a_i \delta_i$ can be represented in terms of these partial sums as

$$\begin{aligned} |\sum_{i=1}^{n} a_i \delta_i| &= |\sum_{i=1}^{n-1} q(\delta_{n-i} - \delta_{n-i+1}) s_{n-i} + \delta_n s_n| \\ &\overset{(i)}{\leq} C \cdot [\delta_n + \sum_{i=1}^{n-1} (\delta_{n-i} - \delta_{n-i+1})] \\ &= C\delta_1, \end{aligned}$$

where inequality (i) uses the bound $|s_{n-i}| \leq C$ and the ordering $\delta_{n-i} \geq \delta_{n-i+1}$. This completes the proof of Lemma 11.

Proof of bound (7.28)

Note that in the setting of multi-armed bandits (cf. Section 7.4), the covariance matrix $\mathbf{S}_n$ is diagonal, and consequently, the definition (7.26a) simplifies to $\mathbf{\Gamma}_i = \max\{\mathbf{S}_i, \mathbf{L}_n\}$. Moreover, a simple argument, using the method of induction on the integer index i, reveals that the matrix $W_i \mathbb{Z}_i$ is a diagonal matrix with nonnegative entries. In particular, we have $\|W_n \mathbb{Z}_n\|_{\max} = |||W_n \mathbb{Z}_n|||_{\rm op}$. By combining these facts, we see that

$$\begin{aligned} |||\mathcal{I} - W_n X_n \mathbf{S}_n^{-\frac{1}{2}}|||_{\rm op} &\overset{(i)}{\leq} 1 + |||W_n X_n \mathbf{\Gamma}_n^{-\frac{1}{2}}|||_{\rm op} \cdot |||\max\{\mathcal{I}, \ \mathbf{L}_n^{\frac{1}{2}} \mathbf{S}_n^{-1} \mathbf{L}_n^{\frac{1}{2}}\}|||_{\rm op} \\ &\overset{(ii)}{\leq} 1 + \|W_n X_n \mathbf{\Gamma}_n^{-\frac{1}{2}}\|_{\max} \cdot O_p(1), \end{aligned} \tag{7.69}$$

where step (i) uses the fact that in multi-armed bandit problems the covariance matrix $\mathbf{S}_n$ is diagonal, and the matrix takes the form $\mathbf{\Gamma}_n = \max\{\mathbf{S}_n, \mathbf{L}_n\}$; and step (ii) follows from assumption (7.25) on the matrix $\mathbf{L}_n$ and the fact that max-norm equals the operator norm for diagonal matrices.

By combining the bounds (7.66a) and (7.66b), we see that $\|W_n X_n \mathbf{\Gamma}_n^{-\frac{1}{2}}\|_{\max} \leq 4$. Combining this bound with inequality (7.69) yields

$$|||\mathcal{I} - W_n X_n \mathbf{S}_n^{-\frac{1}{2}}|||_{\rm op} = O_p(1),$$

as claimed in the bound (7.28).

7.9 Proof of stability Lemma 6

For notational convenience, we use the shorthand notation

$$\mathbf{z}_i := \mathbf{x}_i \mathbf{\Gamma}_i^{-\frac{1}{2}}, \quad \mathbb{Z}_i^\top := [\mathbf{z}_1, \ldots, \mathbf{z}_i], \quad \text{and} \quad W_i = [\mathbf{w}_1, \ldots, \mathbf{w}_i],$$

as previously introduced in Section 7.2.

Verifying the stability condition

The proof of the stability condition is based on a recursion relation that connects the terms $\Delta_i := \mathcal{I} - W_i \mathbb{Z}_i$ and $\Delta_{i-1} := \mathcal{I} - W_{i-1}\mathbb{Z}_{i-1}$. Substituting the expression (7.7b) for the vector $\mathbf{w}_i$ yields

$$\begin{aligned}\Delta_i \Delta_i^\top &= (\Delta_{i-1} - \mathbf{w}_i \mathbf{z}_i^\top)(\Delta_{i-1} - \mathbf{w}_i \mathbf{z}_i^\top)^\top \\ &= \Delta_{i-1}\Delta_{i-1}^\top - \Delta_{i-1}(\mathbf{w}_i \mathbf{z}_i^\top)^\top - \mathbf{w}_i \mathbf{z}_i^\top \Delta_{i-1}^\top + \mathbf{w}_i \mathbf{z}_i^\top \mathbf{z}_i \mathbf{w}_i^\top \\ &= \Delta_{i-1}\Delta_{i-1}^\top - (\gamma_n + \|\mathbf{z}_i\|^2)\mathbf{w}_i \mathbf{w}_i^\top, \end{aligned} \tag{7.70}$$

Summing the last recursion from $i = 1$ to $i = n$ and using the initial condition $W_0 = 0$ yields

$$\mathcal{I} - \sum_{i=1}^n \gamma_n \mathbf{w}_i \mathbf{w}_i^\top = \underbrace{\sum_{i=1}^n \|\mathbf{z}_i\|_2^2 \mathbf{w}_i \mathbf{w}_i^\top}_{1+\|r\|_\infty + \sigma_r\sqrt{1-\gamma_n}} + \underbrace{(\mathcal{I} - W_n \mathbb{Z}_n)(\mathcal{I} - W_n \mathbb{Z}_n)^\top}_{\mathbf{B}_n}. \tag{7.71}$$

Equipped with the last relation, it suffices to verify $\|\!|1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma_n}|\!\|_{\mathrm{op}} \xrightarrow{\mathrm{p}} 0$ and $\|\!|\mathbf{B}_n|\!\|_{\mathrm{op}} \xrightarrow{\mathrm{p}} 0$. We begin by observing that

$$\begin{aligned}\mathbb{P}[\|\!|1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma_n}|\!\|_{\mathrm{op}} > \epsilon] &\leq \mathbb{P}[\mathrm{tr}(1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma_n}) > \epsilon] \\ &\leq \mathbb{P}\left[\frac{\max_{i\in[n]} \|\mathbf{z}_i\|_2^2}{\gamma_n} \sum_{i=1}^n \gamma_n \|\mathbf{w}_i\|_2^2 > \epsilon\right].\end{aligned}$$

Now from equation (7.71), we have the upper bound

$$\sum_{i=1}^n \gamma_n \|\mathbf{w}_i\|_2^2 = D - \mathrm{tr}(1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma_n}) - \mathrm{tr}(\mathbf{B}_n) \leq D.$$

Thus, we have

$$\mathbb{P}[\|\!|1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma_n}|\!\|_{\mathrm{op}} > \epsilon] \leq \mathbb{P}\left[\frac{\max_{i\in[n]} \|\mathbf{z}_i\|_2^2}{\gamma_n} > \frac{\epsilon}{D}\right].$$

Combined with the asymptotic negligibility assumption in (A3), this bound implies that $\|\!|1 + \|r\|_\infty + \sigma_r\sqrt{1-\gamma_n}|\!\|_{\mathrm{op}} \xrightarrow{\mathrm{p}} 0$, as desired.

On the other hand, using the operator-norm bound on the matrix $\mathcal{I} - W_n\mathbb{Z}_n$ from the variance stability condition in (A3), we have

$$\|\!|\mathbf{B}_n|\!\|_{\mathrm{op}} = \|\!|\mathcal{I} - W_n\mathbb{Z}_n|\!\|_{\mathrm{op}}^2 \xrightarrow{\mathrm{p}} 0.$$

Putting together the pieces we conclude $\gamma_n \sum_{i=1}^n \mathbf{w}_i \mathbf{w}_i^\top \xrightarrow{\mathrm{p}} \mathcal{I}$ as claimed.

Verifying the vanishing norm condition

Using the expression (7.7b) for the weight vector $\mathbf{w}_i$, we find that

$$\|\mathbf{w}_i\|_2^2 \leq \frac{1}{(\gamma_n/2 + \|\mathbf{z}_i\|_2^2)^2} \cdot \|\mathcal{I} - W_{i-1}\mathbb{Z}_{i-1}\|_{\text{op}}^2 \, \|\mathbf{z}_i\|_2^2.$$

Doing a calculation similar to the derivation (7.70) we have $\|\mathbf{w}_i\|_{\text{op}}^2 \leq \|\mathbf{w}_0\|_{\text{op}}^2 = 1$. Combining the last two observations with the asymptotic negligibility assumption $\frac{1}{\gamma_n} \max_{i\in[n]} \|\mathbf{z}_i\|_2^2 \xrightarrow{\text{p}} 0$ yields

$$\gamma_n \max_{i\in[n]} \|\mathbf{w}_i\|_2^2 \leq \frac{4}{\gamma_n} \cdot \max_{i\in[n]} \|\mathbf{z}_i\|_2^2 \xrightarrow{\text{p}} 0,$$

as claimed. This completes the proof of Lemma 6.

7.10 Numerical experiment supplement

In this section, we present the results of additional experiments complementing those in Section 7.4.

Multi-armed bandits:

In this section, we repeat the experiment of Section 7.10 using covariates $\{\mathbf{x}_i\}_{i=1}^n$ generated by each of three popular multi-armed bandit algorithms:

(a) Thompson sampling algorithm [Tho33]

(b) an ε-greedy algorithm [LS20]

(c) an upper confidence bound (UCB) strategy based on the paper [Jam+14]

We observe in Figures 7.5 and 7.6 that online debiasing provides appropriate coverage for all confidence levels, all bandit algorithms, and both parameters θ_1^* and θ_2^*. Meanwhile, the OLS lower tail intervals severely undercover for all bandit algorithms and parameters, and W-decorrelation undercovers for several configurations despite having uniformly larger widths than online debiasing in all experiments. Finally, the concentration CI provides 100% coverage for all confidence levels but yields intervals uniformly larger than the online debiasing CIs.

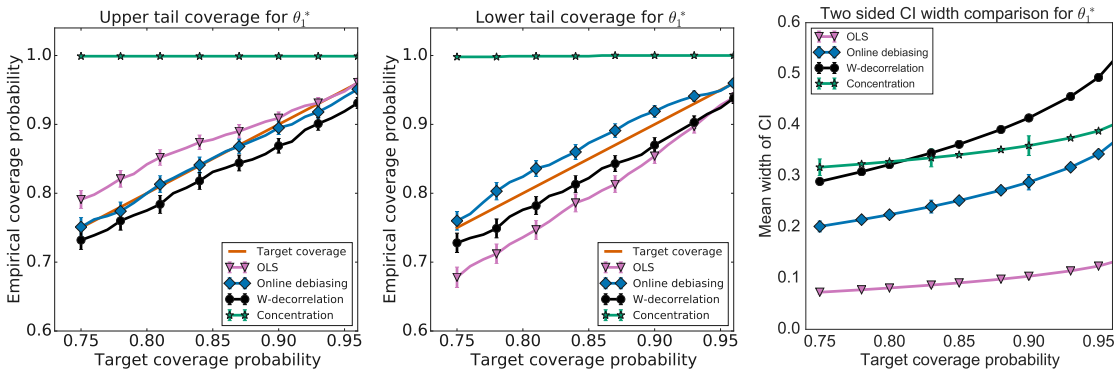
Linear bandits

In this section, we repeat the experiment of Section 7.4 with alternative settings of the ridge regression regularization parameter $\lambda_{\text{ridge}} \in \{1, 10\}$ for the concentration

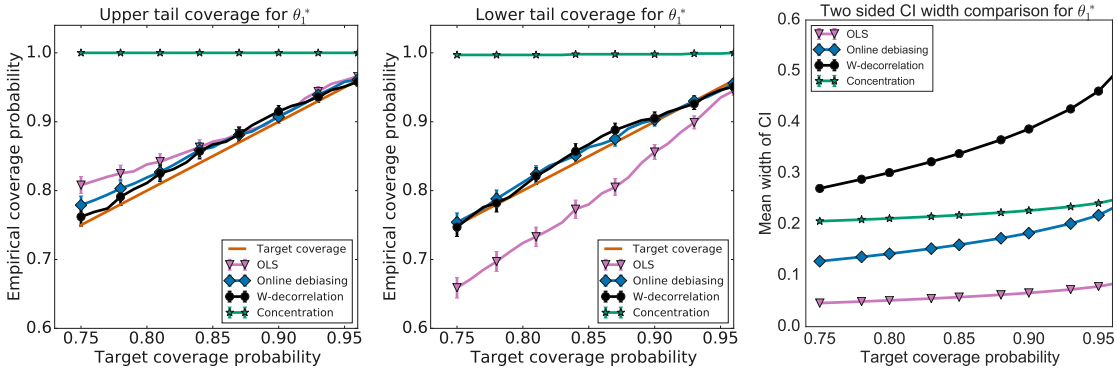
inequality CIs. Recall that given a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^n$ from the model (7.1), the ridge regression estimate $\hat{\theta}_{\text{ridge}}$is defined as

$$\hat{\theta}_{\text{ridge}} \in \arg\max_{\theta} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \theta)^2 + \lambda_{\text{ridge}} \cdot \|\theta\|_2^2 \right\} \tag{7.72}$$

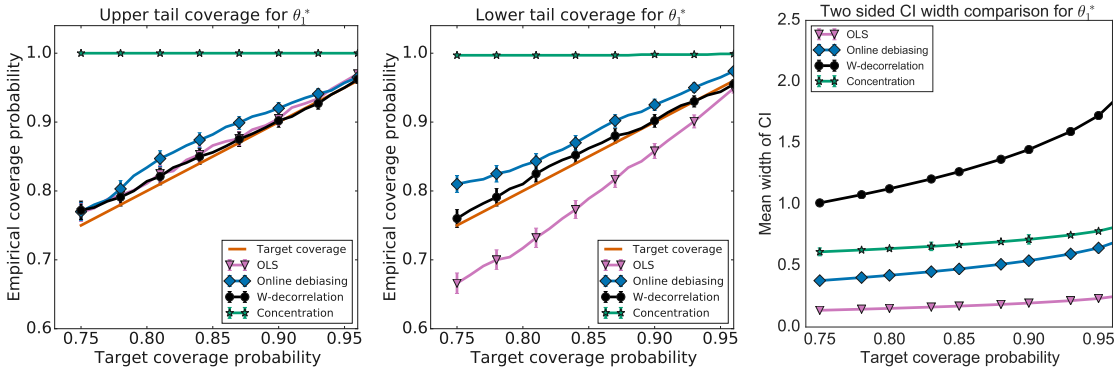
Here, $\lambda_{\text{ridge}} > 0$ is the regularization parameter for the ridge regression, and $\|\theta\|_2$ denotes the ℓ_2 norm of the vector θ. In Figure 7.7, we observe that the concentration based CIs always provide appropriate coverage but are uniformly larger than the online debiasing CIs for both $\lambda_{\text{ridge}} = 1$ and $\lambda_{\text{ridge}} = 10$ and for both parameters θ_1^* and θ_2^*.



(a) Thompson sampling algorithm

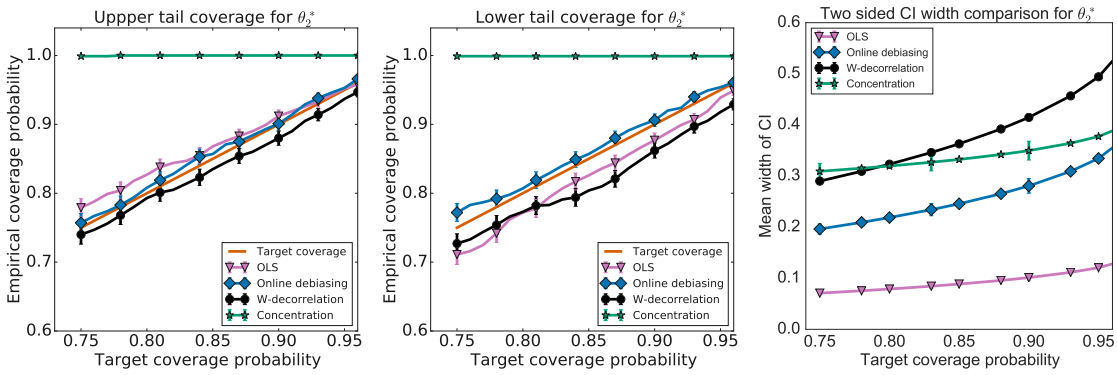


(b) ε-greedy algorithm

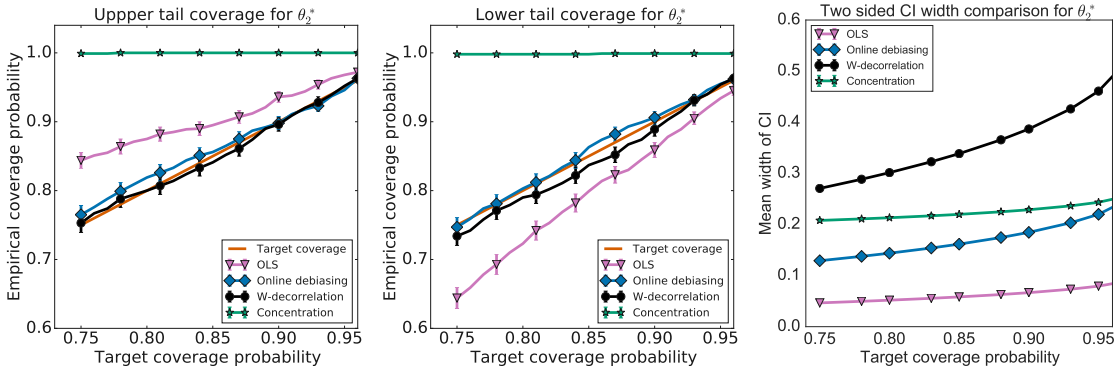


(c) Upper confidence bound (UCB) algorithm

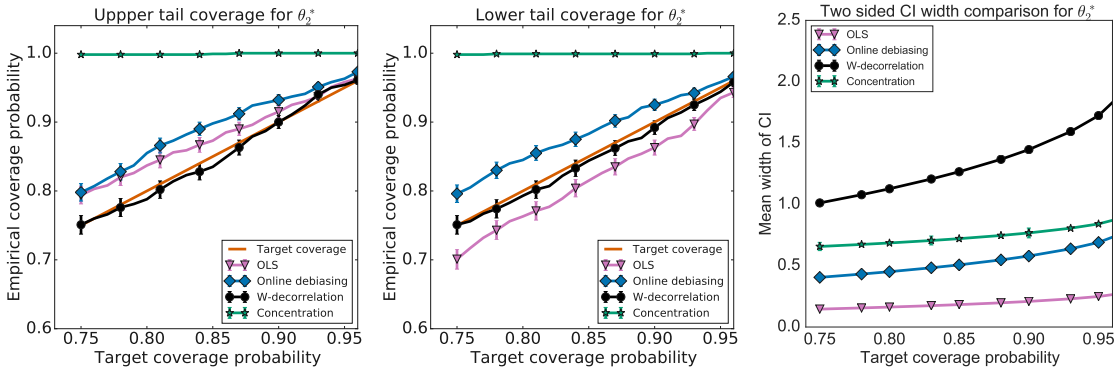
Figure 7.5. Average coverage and width of confidence intervals for θ_1^* across 1000 independent replications of a multi-armed bandit experiment (7.29) with $\theta^* \equiv (\theta_1^*, \theta_2^*) = (0.3, 0.3)^\top$. The covariates $\{\mathbf{x}_i\}_{i=1}^{1000}$ were selected using a) Thompson sampling [Tho33], (b) the ε-greedy algorithm [LS20], and (c) the upper confidence bound algorithm (UCB) [Jam+14]. The error bars represent ± 1 standard error. **Left** and **Center:** Coverage of one-sided $1-\alpha$ intervals for θ_1^*. **Right:** Width of two-sided $1-\alpha$ intervals for θ_1^*. See Section 7.10 for details.



(a) Thompson sampling algorithm

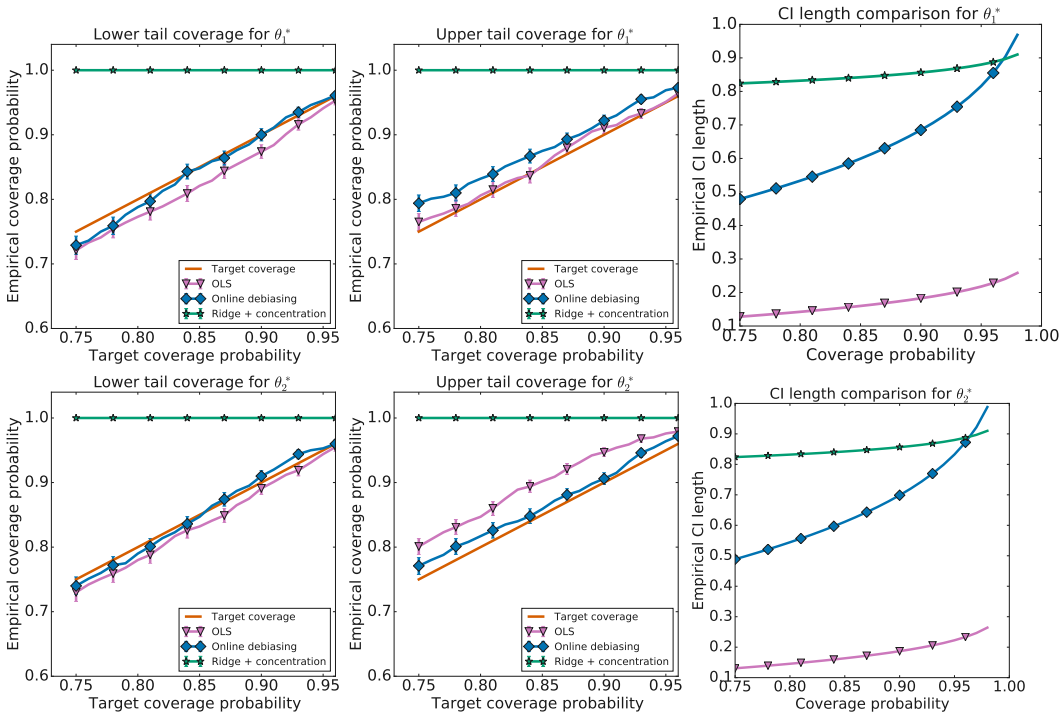


(b) ε-greedy algorithm

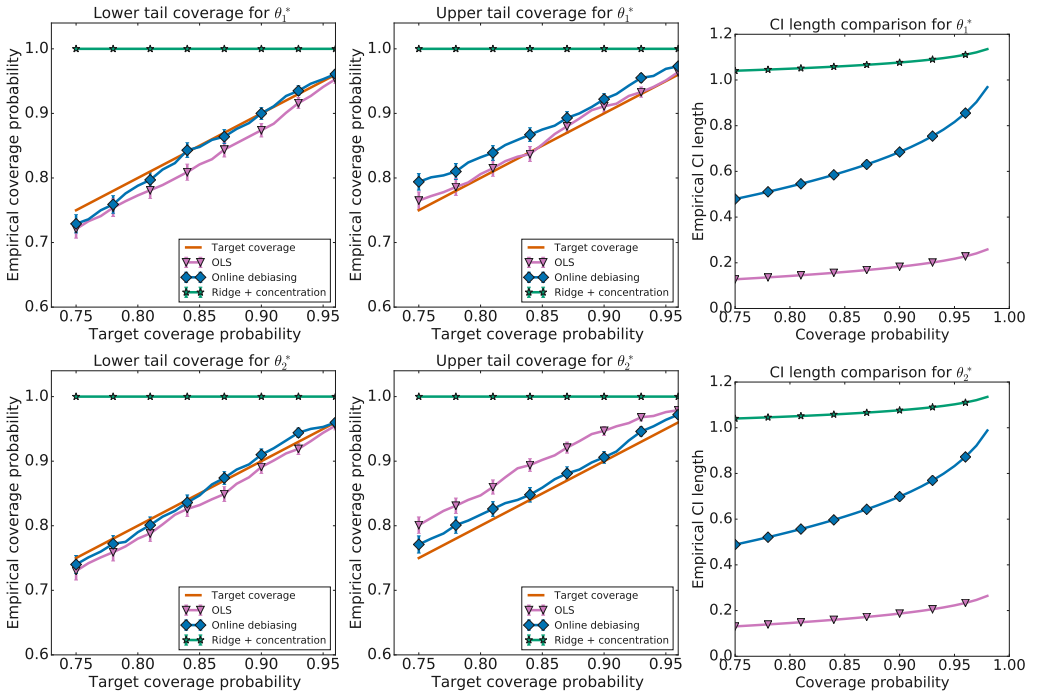


(c) Upper confidence bound (UCB) algorithm

Figure 7.6. Average coverage and width of confidence intervals for θ_2^* across 1000 independent replications of a multi-armed bandit experiment (7.29) with $\theta^* \equiv (\theta_1^*, \theta_2^*) = (0.3, 0.3)^\top$. The covariates $\{\mathbf{x}_i\}_{i=1}^{1000}$ were selected using a) Thompson sampling [Tho33], (b) the ε-greedy algorithm [LS20], and (c) the upper confidence bound algorithm (UCB) [Jam+14]. The error bars represent ± 1 standard error. **Left** and **Center:** Coverage of one-sided $1-\alpha$ intervals for θ_2^*. **Right:** Width of two-sided $1-\alpha$ intervals for θ_2^*. See Section 7.10 for details.



(a) Plots for $\lambda_{\text{ridge}} = 1$



(b) Plots for $\lambda_{\text{ridge}} = 10$

Figure 7.7. Average coverage and width of confidence intervals for θ_1^* and θ_2^* across 1000 independent replications of linear bandits experiment (7.40) with $\theta^* \equiv (\theta_1^*, \theta_2^*) = (0.3, 0.3)^\top$. The covariates $\{\mathbf{x}_i\}_{i=1}^{2500}$ were selected using the ε-greedy linear bandits algorithm (7.40), and the error bars represent ± 1 standard error. **Left** and **Center:** Coverage of one-sided $1-\alpha$ intervals **Right:** Width of two-sided $1-\alpha$ intervals. See Section 7.10 for details.

Part IV

Guarntees for structured non-convex problems

Chapter 8

Convergence Guarantees for a Class of Non-convex and Non-smooth Optimization Problems.

In this chapter we consider the problem of finding critical points of functions that are non-convex and non-smooth. Studying a fairly broad class of such problems, we analyze the behavior of three gradient-based methods (gradient descent, proximal update, and Frank-Wolfe update). For each of these methods, we establish rates of convergence for general problems, and also prove faster rates for continuous sub-analytic functions. We also show that our algorithms can escape strict saddle points for a class of non-smooth functions, thereby generalizing known results for smooth functions. Our analysis leads to a simplification of the popular CCCP algorithm, used for optimizing functions that can be written as a difference of two convex functions. Our simplified algorithm retains all the convergence properties of CCCP, along with a significantly lower cost per iteration. We illustrate our methods and theory via applications to the problems of best subset selection, robust estimation, mixture density estimation, and shape-from-shading reconstruction.

8.1 Introduction

Non-convex optimization problems arise frequently in statistical machine learning; examples include the use of non-convex penalties for enforcing sparsity [FL01; LW13; Wai19a], non-convexity in likelihoods in mixture modeling [YYS17b], and non-convexity in neural network training [LY17]. Of course, minimizing a non-convex problem is NP-hard in general, but problems that arise in machine learning applications are not constructed in an adversarial manner. Moreover, there have been a number of recent papers demonstrating that all first (and/or second) order critical points have desirable properties for certain statistical problems (e.g. [GJZ17; LW13]). Given results of this type, it is often sufficient to find critical points that are first-order

(and possibly second-order) stationary. Accordingly, recent years have witnessed an explosion of research on different algorithms for non-convex problems, with the goal of trying to characterize the nature of their fixed points, and their convergence properties.

There is a lengthy literature on non-convex optimization, dating back more than six decades, and rapidly evolving in the present (e.g., see [Att+10; BST14; CGT10; GTT17; Har59; HPT00; LS09; Lee+16; LB16; PP16; Tuy95; YR03]). Perhaps the most straightforward approach to obtaining a first-order critical point is via gradient descent. Under suitable regularity conditions and step size choices, it can be shown that gradient descent can be used to compute first-order critical points. Moreover, with a random initialization and additional regularity conditions, gradient descent converges almost surely to a second-order stationary point (e.g., [Lee+16; PP16]). These results, like much of the currently available theory for (sub)-gradient methods for non-convex problems, involve smoothness conditions on the underlying objectives. In practice, many machine learning problems have non-smooth components; examples include the hinge loss in support vector machines, the rectified linear unit in neural networks, and various types of matrix regularizers in collaborative filtering and recommender systems. Accordingly, a natural goal is to develop subgradient-based techniques that apply to a broader class of non-convex functions, allowing for non-smoothness.

The main contribution of this chapter is to provide precisely such a set of techniques, along with non-asymptotic guarantees on their convergence rates. In particular, we study algorithms that can be used to obtain first-order (and in some cases, also second-order) optimal solutions to a relatively broad class of non-convex functions, allowing for non-smoothness in certain portions of the problem. For each sequence $\{x^k\}_{k\geq 0}$ generated by one of our algorithms, we provide non-asymptotic bounds on the convergence rate of the gradient sequence $\{\|\nabla f(x^k)\|_2\}_{k\geq 0}$. Moreover, for functions that satisfy a form of the Kurdaya-Łojasiewicz inequality, we show that our methods achieve faster rates.

Our work has important points of contact with a recent line of papers on algorithms for non-convex and non-smooth problems, and we discuss a few of them here. [BST14] developed a proximal-type algorithm applicable to objective functions formed as a sum of smooth (possibly non-convex) and a convex (possibly non-differentiable) function. Some recent work by [XY17] extended these ideas and provided analysis for block co-ordinate descent methods for non-convex functions. [HLR16] analyzed the ADMM method for non-convex problems, whereas in other recent work [AN17; WCP18], the authors proposed a proximal-type method for non-convex functions that can be written as a sum of a smooth function, a concave continuous function and a convex lower semi-continuous function; we also analyze this class in one of our results (Theorem 4).

Our results also relate to another interesting sub-area of non-convex optimization, namely functions that can be represented as a difference of two convex functions, popularly known as DC functions. We refer the reader to the papers [Har59; LS09;

Tuy95; YR03] for more details on DC functions and their properties. One of the most popular DC optimization algorithms is the Convex Concave Procedure, or CCCP for short; see the papers [LB16; YR03] for further details. This is a double loop algorithm that minimizes a convex relaxation of the non-convex objective function at each iteration. While the CCCP algorithm has some attractive convergence properties [LS09], it can be slow in many situations due to its double loop structure. One outcome of the analysis in this chapter is a single-loop proximal-method that retains all the convergence guarantees of CCCP while—as shown in our experimental results—being much faster to run.

Problem setup

In this chapter, we study the problem of minimizing a non-convex and possibly non-smooth function over a closed convex set. More precisely, we consider optimization problems of the form

$$\min_{x \in \Omega} \Big\{ \underbrace{g(x) - h(x) + \varphi(x)}_{f(x)} \Big\}, \tag{8.1}$$

where the domain Ω is a closed convex set. In all cases, we assume the function f is bounded below over domain Ω, and that the function h is continuous and convex. Our aim is to derive algorithms for problem (8.1) for various types of functions g and φ.

Structural assumption on functions g and h

(a) Theorems 3 and 6 are based on the assumption that the function g is continuously differentiable and smooth, and that the function $\varphi \equiv 0$.

(b) In Theorems 4 and 7, we assume that the function g is continuously differentiable and smooth, and that the function φ is convex, proper and lower semi-continuous.[1]

(c) Theorem 5 focuses on the case in which the function g is continuously differentiable, and the function $\varphi \equiv 0$.

The class of non-convex functions covered in part (a) includes, as a special case, the class of differences of convex (DC) functions, for which the first convex function is smooth and the second convex function is continuous. Note that we only put a mild assumption of continuity on the convex function h, meaning that the difference

[1]Taking the function $\varphi \equiv 0$ yields part (a) as a special case, but it is worthwhile to point out that the assumptions in Theorem 3 are weaker than the assumptions of Theorem 4. Furthermore, we can prove some interesting results about saddle points when the function $\varphi \equiv 0$; see Corollary 6.

function $g - h$ can be non-smooth and non-differentiable in general. In particular, for any continuously differentiable function h and any smooth function g, the difference function $f = g - h$ is non-smooth. Furthermore, if we take the function $h \equiv 0$, then we recover the class of smooth functions as a special case.

Overview of our results

- Our first main result (Theorem 3) provides guarantees for a subgradient algorithm as applied to the minimization problem (8.2), to be defined in the sequel, when constrained to a closed convex set Ω. We provide convergence bounds in terms of the Euclidean norm of the subgradient and show that our rates are unimprovable in general. We also illustrate some consequences of Theorem 3 by deriving a convergence rate for our algorithm when applied to non-smooth coercive functions; this result has interesting implications for polynomial programming. We also provide a simplification of the CCCP algorithm, along with convergence guarantees. In Corollary 6, we argue that our algorithm can escape strict saddle points for a large class of non-smooth functions, thereby generalizing known results for smooth functions.
- Our second main result (Theorem 4) provides convergence rates for a proximal-type algorithm for problem (8.1). In Section 8.4, we demonstrate how this proximal-type algorithm can be used to minimize a smooth convex function subject to a sparsity constraint. We demonstrate the performance of this algorithm through the example of best subset selection.
- In Theorem 5, we provide a Frank-Wolfe type algorithm for solving optimization problem (8.17), and we provide a rate of convergence in terms of the associated Frank-Wolfe gap.
- Finally, in Theorems 6 and 7, we prove that Algorithms 2 and 3, when applied to functions that satisfy a variant of the Kurdaya-Łojasiewicz inequality, have faster convergence rates. In particular, the convergence rate in terms of gradient norm is at least $\mathcal{O}(1/k)$ – whereas the worst case rate for general non-convex functions is $\mathcal{O}(\frac{1}{\sqrt{k}})$. We also provide examples of functions for which the convergence rate is $\mathcal{O}(1/k^r)$ with $r > 1$. In Theorem 6, we characterize the class of functions that can be written as a difference of a smooth function and a differentiable convex function.

Section 8.4 is devoted to an illustration of our methods and theory via applications to the problems of best subset selection, robust estimation, mixture density estimation and shape-from-shading reconstruction.

Notation: Given a set $\Omega \subset \mathbb{R}^d$, we use $\text{int}(\Omega)$ to denote its interior. We use $\|x\|_2$, $\|x\|_1$ and $\|x\|_0$ to denote the Euclidean norm, ℓ_1-norm and ℓ_0 norms, respectively, of a vector $x \in \mathbb{R}^d$. We say that a continuously differentiable function g is M_g-smooth if the gradient ∇g is M_g-Lipschitz continuous. In many examples considered in this

chapter, the objective function f is a linear combination of a differentiable function g and one or more convex functions h and φ. With a slight abuse of notation, for a function $f = g - h + \varphi$, we refer to a vector of the form $\nabla g(x) - u(x) + v(x)$, where $u(x) \in \partial h(x)$ and $v(x) \in \partial \varphi(x)$, as a gradient of the function f at point x — and we denote it by $\nabla f(x)$; here, $\partial h(\cdot)$ and $\partial \varphi(\cdot)$ denote the subgradient sets of the convex functions h and φ respectively. We say a point x is a *critical* point of the function f if $0 \in \nabla f(x)$. For a sequence $\left\{a^k\right\}_{k\geq 0}$, we define the running arithmetic mean $\mathrm{Avg}\left(a^k\right)$ as $\mathrm{Avg}\left(a^k\right) := \frac{1}{k}\sum_{\ell=0}^{\ell=k+1} a^\ell$. Similarly, for a non-negative sequence $\left\{a^k\right\}_{k\geq 0}$, we use $\mathrm{GAvg}\left(a^k\right) := (\prod_{\ell=0}^{k} a^\ell)^{\frac{1}{k+1}}$ to denote the running geometric mean. Finally, for real-valued sequences $\{a^k\}_{k\geq 0}$ and $\{b^k\}$, we say $a^k = \mathcal{O}\left(b^k\right)$, if there exists a positive constant C, which is independent of k, such that $a^k \leq Cb^k$ for all $k \geq 0$. We say $a^k = \Omega(b^k)$ if $a^k = \mathcal{O}\left(b^k\right)$ and $b^k = \mathcal{O}\left(a^k\right)$.

8.2 Main results

Our main results are analyses of three algorithms for this class of non-convex non-smooth problems; in particular, we derive non-asymptotic bounds on their rates of convergence. The first algorithm is a (sub)-gradient-type method, and it is mainly suited for unconstrained optimization; the second algorithm is based on a proximal operator and can be applied to constrained optimization problems. The third algorithm is a Frank-Wolfe-type algorithm, which is also suitable for constrained optimization problems, but it applies to a more general class of non-convex optimization problems.

Gradient-type method

In this section, we analyze a (sub)-gradient-based method for solving a certain class of non-convex optimization problems. In particular, consider a pair of functions (g, h) such that:

Assumption GR:

(a) The function g is continuously differentiable and M_g-smooth.

(b) The function h is continuous and convex.

(c) There is a closed convex set Ω such that the difference function $f := g - h$ is bounded below on the set Ω.

Under these conditions, we then analyze the behavior of a (sub)-gradient method in application to the following problem

$$f^* = \min_{x\in\Omega} f(x) \;=\; \min_{x\in\Omega} \left\{g(x) - h(x)\right\}. \tag{8.2}$$

Let $\partial h(x)$ denote the subdifferential of the convex function h at the point x. With a slight abuse of notation, we refer to a vector of the form $\nabla g(x) - u(x)$ with $u(x) \in \partial h(x)$ as a gradient of the function f at the point x.

Algorithm 2 Subgradient-type method

1: Given an initial point $\mathbf{x}_0 \in \text{int}(\Omega)$ and step size $\alpha \in (0, \frac{1}{M_g}]$:
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Choose subgradient $u^k \in \partial h(\mathbf{x}_k)$.
4: Update $x^{k+1} = x^k - \alpha\Big(\nabla g(\mathbf{x}_k) - u^k\Big)$.
5: **end for**

In our analysis, we assume that the initial vector $\mathbf{x}_0 \in \text{int}(\Omega)$ is chosen such that the associated level set

$$\mathcal{L}(f(\mathbf{x}_0)) := \Big\{x \in \mathbb{R}^d \mid f(x) \leq f(\mathbf{x}_0)\Big\}$$

is contained within $\text{int}(\Omega)$. This condition is standard in the analysis of non-convex optimization methods (e.g., see [NP06]). When $\Omega = \mathbb{R}^d$, it holds trivially. With this set-up, we have the following guarantees on the convergence rate of Algorithm 2.

Theorem 3. *Under Assumption GR, any sequence $\{x^k\}_{k\geq 0}$ produced by Algorithm 2 has the following properties:*

(a) Any limit point is a critical point of the function f, and the sequence of function values $\{f(\mathbf{x}_k)\}_{k\geq 0}$ is strictly decreasing and convergent.

(b) For all $k = 0, 1, 2, \ldots$, we have

$$\text{Avg}\left(\|\nabla f(\mathbf{x}_k)\|_2^2\right) \leq \frac{2\Big(f(\mathbf{x}_0) - f^*\Big)}{\alpha(k+1)}. \tag{8.3}$$

See Section 8.7 for a proof of this theorem.

Comments on convergence rates

Note that the bound (8.3) guarantees that the gradient norm sequence $\min_{j\leq k} \|\nabla f(\mathbf{x}_j)\|_2$ converges to zero at the rate $\mathcal{O}(1/\sqrt{k})$. It is natural to wonder whether this convergence rate can be improved. Interestingly, the answer is no, at least for the general class of functions covered by Theorem 3. Indeed, note that the class of M-smooth functions is contained within the class of functions covered by Theorem 3. It follows from past work by [CGT10] that for gradient descent on M-smooth functions, with a step size chosen according to the Goldstein-Armijo rule, the convergence rate of the gradient sequence $\{\|\nabla f(x^k)\|_2\}_{k\geq 0}$ can be lower bounded—for appropriate choices of the function f—as $\Omega(1/\sqrt{k})$. It is not very difficult to see that the same construction

also provides a lower bound of $\Omega(1/\sqrt{k})$ for gradient descent with a constant step size. We also note that very recently, [Car+17] proved an even stronger result: more precisely, for the class of smooth functions, the rate of convergence of any algorithm given access to only the function gradients and function values cannot be faster than $\Omega(1/\sqrt{k})$. Finally, observe that in the special case $h \equiv 0$, Algorithm 2 reduces to the ordinary gradient descent with fixed step size α. Putting together the pieces, we conclude that for the class of functions which can be written as a difference of smooth and a continuous convex function, Algorithm 1 is *optimal* among all algorithms that have access to the gradients (and/or the sub-gradients) and the function values.

Consequences for differentiable functions

In the special case when the function h is convex and differentiable, Algorithm 2 reduces to an ordinary gradient descent on the difference function $f = g - h$. However, note that the step size choice required in Algorithm 2 does *not* depend on the smoothness of the function h; consequently, the algorithm can be applied to objective functions f that are not smooth. As a simple but concrete example, suppose that we wish to apply gradient descent to minimize the function $f(x) := g(x) - \|x\|_2^q$, where g is any μ-strongly convex and M_g-smooth function, and $q \in (1, 2)$ is a given parameter. Classical guarantees on gradient descent, which require the smoothness of the function f, would not apply here since the function f itself is not smooth. However, Theorem 3 guarantees that standard gradient descent would converge for any step size $\alpha \in \left(0, \frac{1}{M_g}\right]$.

More generally, given an arbitrary continuously differentiable function f, we can define its *effective smoothness constant* as

$$M_f^* := \inf_h \Big\{ L \;\mid\; (f+h) \text{ is } L\text{-smooth} \Big\}, \tag{8.4}$$

where the infimum ranges over all convex and continuously differentiable functions h. Suppose that this infimum is achieved by some function h^*, then gradient descent on the function f can be viewed as applying Algorithm 2 to the decomposition $f = g^* - h^*$, where the function $g^* := f + h^*$ is guaranteed to be M_f^*-smooth. To be clear, the algorithm itself does *not* need to know the decomposition (g^*, h^*), but the existence of the decomposition ensures the success of a backtracking procedure. Putting together the pieces, we arrive at the following consequence of Theorem 3:

Corollary 4. *Given a closed convex set Ω, consider a continuously differentiable function f with effective smoothness $M_f^* < \infty$ that is bounded below on Ω. Then for any sequence $\{\mathbf{x}_k\}_{k \geq 0}$ obtained by applying the gradient update with step size $\alpha \in \left(0, \frac{1}{M_f^*}\right)$, we have:*

$$\text{Avg}\left(\|\nabla f(\mathbf{x}_k)\|_2^2 \right) \leq \frac{2\Big(f(\mathbf{x}_0) - f^*\Big)}{\alpha(k+1)}. \tag{8.5a}$$

Moreover, if we choose step size by backtracking[2] *with parameter* $\beta \in (0,1)$*, then for all* $k = 0, 1, 2, \ldots$*, we have*

$$\text{Avg}\left(\|\nabla f(\mathbf{x}_k)\|_2^2\right) \leq \frac{2\max\left\{1, M_f^*\right\}\left(f(\mathbf{x}_0) - f^*\right)}{\beta^2(k+1)}. \tag{8.5b}$$

See Section 8.7 for proof of the above corollary.

Let us reiterate that the advantage of backtracking gradient descent is that it works without knowledge of the scalar M_f^*. The parameter β mentioned in equation (8.5b) is the user-defined backtracking parameter (see Algorithm 5 for details). In particular, substituting $\beta = \frac{1}{\sqrt{2}}$ in equation (8.5b) yields

$$\text{Avg}\left(\|\nabla f(\mathbf{x}_k)\|_2^2\right) \leq \frac{4\max\left\{1, M_f^*\right\}\left(f(\mathbf{x}_0) - f^*\right)}{(k+1)},$$

which differs from the rate obtained in equation (8.5a) only by a factor of two, and a possible multiple of M_f^*.

Consequences for coercive functions

As a consequence of Corollary 4, we can obtain a rate of convergence of the backtracking gradient descent algorithm (Algorithm 5) for a class of non-smooth coercive functions. Consider any twice continuously differentiable coercive function $f : \mathbb{R}^d \mapsto \mathbb{R}$, which is bounded below. Recall that a function f is *coercive* if

$$f(x^\ell) \overset{\ell\to\infty}{\to} \infty \quad \text{for any sequence } \{x^\ell\}_{\ell\geq 0} \text{ such that } \quad \|x^\ell\|_2 \to \infty. \tag{8.6}$$

Let $\mathcal{L}(f(\mathbf{x}_0)) := \left\{x \in \mathbb{R}^d : f(x) \leq f(\mathbf{x}_0)\right\}$ denote the level set of the function f at point $\mathbf{x}_0$. It can be verified that for any coercive function f, the set $\mathcal{L}(f(\mathbf{x}_0))$ is bounded above for all $\mathbf{x}_0 \in \mathbb{R}^d$. This property ensures that for any descent algorithm and any starting point $\mathbf{x}_0$, the set of iterates $\left\{\mathbf{x}_k\right\}_{k\geq 0}$ obtained from the algorithm remains within a bounded set—viz. the level set $\mathcal{L}(f(\mathbf{x}_0))$ in this case. Since the function f is twice continuously differentiable, we have that f is smooth over bounded set $\mathcal{L}(f(\mathbf{x}_0))$; this fact ensures that f has a finite effective smoothness constant in the set $\mathcal{L}(f(\mathbf{x}_0))$, which we denote by $M_{f,\mathbf{x}_0}^*$. Finally, note that Algorithm 5 is a descent algorithm; as a result, a simple application of Corollary 4 yields the following rate of convergence for the backtracking gradient descent algorithm (Algorithm 5):

Corollary 5. *Consider the unconstrained minimization problem of a twice continuously differentiable coercive function* f *that is bounded below on* $\mathbb{R}^d$*. Then for any*

[2] A detailed description of gradient descent with backtracking is provided in Algorithm 5.

initial point $\mathbf{x}_0$*, the sequence* $\{\mathbf{x}_k\}_{k\geq 0}$ *obtained by applying Algorithm 5 satisfies the following property:*

$$\text{Avg}\left(\|\nabla f(\mathbf{x}_k)\|_2^2\right) \leq \frac{2\max\left\{1, M^*_{f,\mathbf{x}_0}\right\}\left(f(\mathbf{x}_0) - f^*\right)}{\beta^2(k+1)} \quad \textit{for all } k = 0, 1, 2, \ldots, \tag{8.7}$$

where $\beta \in (0, 1)$ *is the backtracking parameter.*

Implications for polynomial programming: Corollary 5 has useful implications for problems that involve minimizing polynomials. Such problems of polynomial programming arise in various applications, including phase retrieval and shape-from-shading [WSU14], and we illustrate our algorithms for the latter application in Section 8.4. For minimization of a coercive polynomial, Corollary 5 shows that Algorithm 5 achieves a near-optimal rate.

It is worth noting that any even degree polynomial can be represented as a difference of convex (DC) function; hence, such problems are amenable to DC optimization techniques like CCCP, which we discuss at more length in Section 8.2. However, obtaining a good DC decomposition, which is crucial to the success of CCCP, is often a formidable task. In particular, obtaining an optimal decomposition for a polynomial with degree greater than four is NP-hard; indeed, deciding the convexity of an even degree polynomial with degree greater than four is NP-hard [Ahm+13; WSU14]. Even for a fourth degree polynomial with dimension larger than three, there is no known algorithm for finding an optimal DC decomposition [AP13]. An advantage of Algorithm 5 is that it obviates the need to find a DC decomposition.

Escaping strict saddle points

One of the obstacles with gradient-based continuous optimization method is possible convergence to saddle points. Here we show that with a random initialization this undesirable outcome does not occur for the class of strict saddle points. Recall that for a twice differentiable function f, a point x is called a strict saddle point of the function f if $\lambda_{\min}(\nabla^2 f(x)) < 0$, where $\lambda_{\min}(\nabla^2 f(x))$ denotes the minimum eigenvalue of the Hessian matrix $\nabla^2 f(x)$. The following corollary shows that such saddle points are *not* troublesome:

Corollary 6. *Suppose that, in addition to the conditions on* (g, h, Ω) *from Theorem 3, the functions* (g, h) *are twice continuously differentiable. If Algorithm 2 is applied with step size* $\alpha \in \left(0, \frac{1}{M_g}\right)$*, then the set of initial points for which it converges to a strict saddle point has measure zero.*

See Section 8.7 for the proof of this corollary.

We note that similar guarantees of avoidance of strict saddle-points are known when the function $f = g - h$ is twice continuously differentiable and M-smooth

(e.g., [Lee+16; PP16]). The novelty of Corollary 6 is that the same guarantee holds without imposing a smoothness condition on the entire function f.

Connections to the convex-concave procedure

As a consequence of Algorithm 2, we show that one can obtain a convergence rate of the Euclidean norm of the gradient for CCCP (convex-concave procedure), which is a heavily used algorithm in Difference of Convex (DC) optimization problems. Before doing so, let us provide a brief description of DC functions and the CCCP algorithm.

DC functions: Given a convex set $\Omega \subseteq \mathbb{R}^d$, we say that a function $f : \Omega \mapsto \mathbb{R}$ is DC if there exist convex functions g and h with domain Ω such that $f = g - h$. Note that the DC representation $f = g - h$ mentioned in the definition is not unique. In particular, for any convex function p, we can write $f = (g + p) - (h + p)$. The class of DC functions includes a large number of non-convex problems encountered in practice. Both convex and concave functions are DC in a trivial sense, and the class of DC functions remains closed under addition and subtraction. More interestingly, under mild restrictions on the domain, the class of non-zero DC functions is also closed under multiplication, division, and composition (e.g., [Har59; Tuy95]). The maximum and minimum of a finite collection of DC functions are also DC functions.

Convex-concave procedure: An interesting class of problems are those that involve minimizing a DC function over a closed convex set $\Omega \subseteq \mathbb{R}^d$, i.e.

$$f^* := \min_{x \in \Omega} f(x) \;=\; \min_{x \in \Omega} \Big\{ g(x) - h(x) \Big\}, \tag{8.8}$$

where g and h are proper convex functions. The above problem has been studied intensively, and there are various methods for solving it; for instance, see the papers [LB16; PNL13; Tuy95] and references therein for details. One of the most popular algorithms to solve problem (8.8) is the Convex-concave Procedure (CCCP), which was introduced by [YR03]. The CCCP algorithm is a special case of a Majorization-Minimization algorithm, one which uses the DC structure of the objective function in problem (8.8) to construct a convex majorant of the objective function f at each step. We start with a feasible point $x^0 \in \mathrm{int}(\Omega)$. Let $\mathbf{x}_k$ denote the iterate at k^{th} iteration; at the $(k+1)^{\text{th}}$ iteration we construct a convex majorant $q(\cdot,\, x^k)$ of the function f via

$$f(x) \leq \underbrace{g(x) - h(x^k) - \langle u^k,\, x - x^k \rangle}_{=:\, q(x, x^k)}, \tag{8.9}$$

where $u^k \in \partial h(x^k)$, the subgradient set of the convex function h at point $\mathbf{x}_k$. The next iterate x^{k+1} is obtained by solving the convex program

$$x^{k+1} \in \arg\min_{x \in \Omega} q(x, x^k). \tag{8.10}$$

The CCCP algorithm has some attractive convergence properties. For instance, it is a descent algorithm; when the function g is strongly convex differentiable and the function h is continuously differentiable, it can be shown [LS09] that any limit point of the sequence $\left\{x^k\right\}_{k\geq 0}$ obtained from CCCP is stationary. Under the same assumptions, one can also verify that $\lim_{k\to\infty} \|x^k - x^{k+1}\|_2 = 0$.

We now turn to an analysis of CCCP using the techniques that underlie Theorem 3. In the next proposition, we derive a rate of convergence of the gradient sequence and show that all limit points of the sequence $\left\{\mathbf{x}_k\right\}_{k\geq 0}$ are stationary. Earlier analyses of CCCP, including the papers [LS09; YR03], are mainly based on the assumption of strong convexity of the function g, whereas in the next proposition, we only assume that the function g is M_g-smooth. When the function g is strongly convex, our analysis recovers the well-known convergence result in past work [LS09]. In particular, we show that CCCP enjoys the same rate of convergence as that of Algorithm 2.

Proposition 3. *Under Assumption GR and with the function g being convex, the CCCP sequence (8.10) has the following properties:*

(a) Any limit point of the sequence $\left\{\mathbf{x}_k\right\}_{k\geq 0}$ is a critical point, and the sequence of function values $\left\{f(\mathbf{x}_k)\right\}_{k\geq 0}$ is strictly decreasing and convergent.

(b) Furthermore, for all $k = 1, 2, \ldots$, we have

$$\operatorname{Avg}\left(\|\nabla f(\mathbf{x}_k)\|_2^2\right) \leq \frac{2M_g\left(f(\mathbf{x}_0) - f^*\right)}{(k+1)}, \tag{8.11a}$$

and assuming moreover that g is μ-strongly convex,

$$\operatorname{Avg}\left(\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^2\right) \leq \frac{2\left(f(\mathbf{x}_0) - f^*\right)}{\mu(k+1)}. \tag{8.11b}$$

The proof of this proposition builds on the argument used for Theorem 3; see Section 8.7 for details.

Simplifying CCCP

Algorithm 2 provides us an alternative procedure for minimizing a difference of convex functions when the first convex function is smooth. The benefit of Algorithm 2 over standard CCCP is that Algorithm 2 is a single loop algorithm and is expected to be faster than standard double loop CCCP algorithm in many situations. Furthermore, Algorithm 2 shares convergence guarantees similar to a standard CCCP algorithm.

Proximal-type method

We now turn to a more general class of optimization problems of the form

$$f^* := \min_{x\in\mathbb{R}^d} f(x) \;=\; \min_{x\in\mathbb{R}^d} \Big\{\Big(g(x) - h(x)\Big) + \varphi(x)\Big\}. \tag{8.12}$$

We assume that the functions g, h and φ satisfy the following conditions:

Assumption PR

(a) The function $f = g - h + \varphi$ is bounded below on $\mathbb{R}^d$.

(b) The function g is continuously differentiable and M_g-smooth; the function h is continuous and convex; and the function φ is proper, convex and lower semi-continuous.

Typical examples of the function φ include $\varphi(x) = \|x\|_1$, or the indicator of a closed convex convex set $\mathcal{X}$. Since for a general lower semi-continuous function φ, the sum-function $g + \varphi$ is neither differentiable nor smooth, a gradient-based method cannot be applied. One way to minimize such functions is via a proximal-type algorithm, of which the following is an instance.

Algorithm 3 Proximal-type algorithm

1: Given an initial vector $\mathbf{x}_0 \in \mathrm{dom}(f)$ and step size $\alpha \in \Big(0, \frac{1}{M_g}\Big]$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Update $\mathbf{x}_{k+1} = \mathrm{prox}^{\varphi}_{1/\alpha}\Big(\mathbf{x}_k - \alpha\Big(\nabla g(\mathbf{x}_k) - u^k\Big)\Big)$ for some $u^k \in \partial h(\mathbf{x}_k)$.
4: **end for**

The proximal update in line 3 of Algorithm 3 is very easy to compute and often has a closed form solution (see [PB+14]). Let us now derive the rate of convergence result of Algorithm 3.

Theorem 4. *Under Assumption PR, any sequence* $\Big\{\mathbf{x}_k\Big\}_{k\geq 0}$ *obtained from Algorithm 3 has the following properties:*

(a) Any limit point of the sequence $\Big\{\mathbf{x}_k\Big\}_{k\geq 0}$ *is a critical point, and the sequence of function values* $\Big\{f(\mathbf{x}_k)\Big\}_{k\geq 0}$ *is strictly decreasing and convergent.*

(b) For all $k = 1, 2, \ldots$, *we have*

$$\mathrm{Avg}\left(\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2^2\right) \leq \frac{2\alpha\Big(f(\mathbf{x}_0) - f^*\Big)}{(k+1)}. \tag{8.13a}$$

If moreover the function h is M_h-smooth, then

$$\text{Avg}\left(\|\nabla f(\mathbf{x}_k)\|_2^2\right) \leq \frac{2\alpha C_{M,\alpha}\left(f(\mathbf{x}_0) - f^*\right)}{(k+1)}, \tag{8.13b}$$

where $C_{M,\alpha} = \left(M_g + M_h + \frac{1}{\alpha}\right)^2$.

See Section 8.8 for the proof of the theorem.

Comments: The proof of Theorem 4 reveals that the smoothness condition on the function h in Theorem 4 can be replaced by the local smoothness of h, when the sequence $\left\{\mathbf{x}_k\right\}_{k\geq 0}$ is bounded. Note that the local smoothness condition is weaker than the global smoothness condition. For instance, any twice continuously differentiable function is locally smooth. The boundedness assumption on the iterates $\left\{\mathbf{x}_k\right\}_{k\geq 0}$ holds in many situations. For instance, if the function f is coercive (8.6), then it follows that the iterates $\left\{\mathbf{x}_k\right\}_{k\geq 0}$ remain bounded. Another instance is when the function φ is the indicator function of a compact convex set. Finally, we point out that when the function h is non-smooth but the proximal-function φ is smooth, the existing proof can be easily modified to obtain a rate of convergence of the gradient-norm $\|\nabla f(\mathbf{x}_k)\|_2$.

Projected Gradient Descent: A special case of the Algorithm 3 is when φ is equal to the indicator function $\mathbb{1}_{\mathcal{X}}$ of a closed convex set $\mathcal{X}$. Consider the following constrained optimization problem

$$f^* := \min_{x\in\mathcal{X}} \quad \Big\{\underbrace{g(x) - h(x)}_{f(x)}\Big\}, \tag{8.14}$$

where $\mathcal{X}$ is a closed convex set, the function g is M_g-smooth, and the function h is convex continuous. Using Algorithm 3, the update equation in this case is given by

$$\mathbf{x}_{k+1} = \Pi_{\mathcal{X}}\Big(\mathbf{x}_k - \alpha(\nabla g(\mathbf{x}_k) - u^k)\Big). \tag{8.15}$$

In projected-gradient-type methods, we should not expect a rate in terms of the gradient. In such cases, the projected gradient step may not be aligned with the gradient direction, or the step size may be arbitrarily small due to projection. Rather, an appropriate analogue of the gradient in this case is as follows:

$$\nabla f_{\mathcal{X}}(\mathbf{x}_k) = \frac{1}{\alpha}\Big(\mathbf{x}_k - \Pi_{\mathcal{X}}(\mathbf{x}_k - \alpha(\nabla g(\mathbf{x}_k) - u^k))\Big). \tag{8.16}$$

The analysis of the projected gradient method using $\nabla f_{\mathcal{X}}(\mathbf{x}_k)$ is standard in the optimization literature [Bub+15]. It is worth pointing out that the quantity $\nabla f_{\mathcal{X}}(\mathbf{x}_k)$ is the analogue of the gradient in the constrained optimization setup, and coincides

with the gradient in the unconstrained setup. Concretely, we have $\nabla f_{\mathcal{X}}(\mathbf{x}_k) = \nabla f(\mathbf{x}_k)$ where $f := g - h$, and $\mathcal{X} = \mathbb{R}^d$. Combining equations (8.15) and (8.16) and applying the bound (8.13b) from Theorem 4, we find that

$$\text{Avg}\left(\|\nabla f_{\mathcal{X}}(\mathbf{x}_k)\|_2^2\right) \leq \frac{2\Big(f(\mathbf{x}_0) - f^*\Big)}{\alpha(k+1)}.$$

Frank-Wolfe type method

In our analysis of the previous two algorithms, we assumed that the objective function f has a smooth component g, and we leveraged the smoothness property of g to establish convergence rates. In many situations, the objective function may not have a smooth component; consequently, neither the gradient-type algorithm nor the prox-type algorithm provides any theoretical guarantee. In this section, we analyze a Frank-Wolfe-type algorithm for solving such optimization problems. In particular, consider an optimization problem of the form

$$f^* := \min_{x \in \Omega} f(x) \;\; = \min_{x \in \Omega} \Big\{ g(x) - h(x) \Big\}, \tag{8.17}$$

where Ω is a closed convex set, and the functions (g, h) satisfy the following conditions:

Assumption FW:

(a) The difference function $f = g - h$ is bounded below over range Ω.

(b) The function g is continuously differentiable, whereas the function h is convex and continuous.

The analysis of the Frank-Wolfe algorithm for a convex problem is based on the *curvature constant* $\mathcal{C}_f$ of the convex objective function with respect to the closed convex set Ω. This curvature constant can be defined for any differentiable function, which need not be convex [Lac16].

Here we define a slight generalization of this notion, applicable to a non-differentiable function $f = g - h$ that can be written as a difference of a differentiable function g and a continuous convex function h (which may be non-differentiable). Define the set

$$S_\gamma := \Big\{ x, y \in \Omega \mid \text{there exist } \gamma \in (0, 1] \text{ and } u \in \Omega \text{ with } y = x + \gamma(u - x) \Big\},$$

and the curvature constant

$$\mathcal{C}_f = \sup_{\substack{x,y \in S_\gamma \\ u \in \partial h(x)}} \frac{2}{\gamma^2} \Big[f(y) - f(x) - \langle y - x,\, \nabla g(x) - u \rangle \Big]. \tag{8.18}$$

Note that in the special case $h \equiv 0$, we recover the curvature constant of the differentiable function g used by Lacoste-Julien [Lac16]. We refer to the scalar $\mathcal{C}_f$

Algorithm 4 Frank-Wolfe type method

1: Given initial vector $\mathbf{x}_0 \in \int(\Omega)$:
2: **for** $k = 1, \ldots, K$ **do**
3: Choose any $u^k \in \partial h(\mathbf{x}_k)$.
4: Compute $s^k := \arg\min_{s\in\Omega}\langle s,\, \nabla g(\mathbf{x}_k) - u^k\rangle$.
5: Define $d^k := s^k - \mathbf{x}_k$ and $g^k := -\langle d^k,\, \nabla g(\mathbf{x}_k) - u^k\rangle$. *(Frank-Wolfe gap)*
6: Set $\gamma^k = \min\left\{\frac{g^k}{C_0}, 1\right\}$ for some $C_0 \geq \mathcal{C}_f$.
7: Update $\mathbf{x}_{k+1} = \mathbf{x}_k + \gamma^k d^k$.
8: **end for**

as the generalized curvature constant of the function f with respect to the closed convex set Ω.

Next, we provide an analysis of Algorithm 4 in terms of the Frank-Wolfe (FW) gap g^k defined Step 5. We show that the minimum FW gap $\{g^k\}_{k\geq 0}$ defined in Algorithm 4 converges to zero at the rate $\frac{1}{\sqrt{k+1}}$.

Theorem 5. *Under Assumption FW, the Frank-Wolfe gap sequence* $\{g^k\}_{k\geq 0}$ *from Algorithm 4 satisfies the following property:*

$$\min_{0\leq j\leq k} g^j \leq \frac{\max\left\{2\Big(f(\mathbf{x}_0) - f^*\Big), C_0\right\}}{\sqrt{k+1}} \qquad \text{for all } k = 0, 1, 2, \ldots.$$

See Section 8.9 for the proof of this theorem.

Comments: The FW gap appearing in Theorem 5 is standard in the analysis of Frank-Wolfe algorithm; note that it is invariant to an affine transformation of the set Ω. Similar convergence guarantees for the minimum FW-gap are available for differentiable functions; for instance, see the paper [Lac16]. The novelty of the above theorem is that it provides convergence guarantees of minimum FW-gap for a class of non-differentiable functions.

Upper bound on generalized curvature constant: It is worth mentioning that Algorithm 4 only requires an upper bound of the generalized curvature constant $\mathcal{C}_{g-h}$. Consequently, it is interesting to obtain an upper bound for the scalar $\mathcal{C}_{g-h}$. For a M_g-smooth function g, one well-known upper bound of the curvature constant $\mathcal{C}_g$ is $M_g \times \Big(\operatorname{diam}_{\|\cdot\|_2}(\Omega)\Big)^2$; see also [Jag13]. A similar upper bound also holds for the generalized curvature constant defined in equation (8.59). In particular, we prove that for a difference function $f = g - h$, with the function h being convex continuous, the scalar $\mathcal{C}_{g-h}$ is always upper bounded by $\mathcal{C}_g$, the curvature constant of the function g (see Lemma 17).

8.3 Faster rate under KL-inequality

In the preceding sections, we have derived rates of convergence for the gradient norms for various classes of problems. It is natural to wonder if faster convergence rates are possible when the objective function is equipped with some additional structure. Based on Theorems 3 and 4, we see that both Algorithms 2 and 3 ensure that $\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 \to 0$, meaning that the successive differences between the iterates converge to zero. Although we proved that any limit point of the sequence $\{\mathbf{x}_k\}_{k\geq 0}$ has desirable properties, the condition $\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 \to 0$ is not sufficient—at least in general—to prove convergence[3] of the sequence $\{\mathbf{x}_k\}_{k\geq 0}$. In this section, we provide a sufficient condition under which Algorithm 2 and Algorithm 3 yield convergent sequences of iterates $\{\mathbf{x}_k\}_{k\geq 0}$, and we establish that the gradient sequences $\{\|\nabla f(x)\|_2\}_{k\geq 0}$ converge at faster rates.

Kurdaya-Łojasiewicz inequality

Let us now establish a faster local rate of convergence of Algorithms 2 and 3 for functions that satisfy a form of the Kurdaya-Łojasiewicz (KL) inequality. More precisely, suppose that there exists a constant $\theta \in [0, 1)$ such that the ratio $\frac{(f(x)-f(\bar{x}))^\theta}{\|\nabla f(x)\|_2}$ is bounded above in a neighborhood of every point $\bar{x} \in \mathrm{dom}(f)$. This type of inequality is known as a Kurdaya-Łojasiewicz inequality, and the exponent θ is known as the Kurdaya-Łojasiewicz exponent (*KL-exponent*) of the function f at the point $\bar{x}$. These type of inequalities were first proved by [Łoj63] for real analytic functions; in later work, [Kur98] and [BDL07] proved similar inequalities for non-smooth functions, and the authors also provided examples of many functions that satisfy a form of the KL inequality. See Section 8.6 for further details on functions of the KL type.

Assumption KL: For any point[4] $\bar{x} \in \mathrm{dom}(f)$, there exists a scalar $\theta \in [0, 1)$ such that the ratio $\frac{|f(x)-f(\bar{x})|^\theta}{\|\nabla f(x)\|_2}$ is bounded above in a neighborhood of $\bar{x}$.

Convergence guarantees

Theorem 6. *Under Assumptions GR and KL, any bounded sequence $\{\mathbf{x}_k\}_{k\geq 0}$ obtained from Algorithm 2 satisfies the following properties:*

[3]The convergence of the sequence $\{\mathbf{x}_k\}_{k\geq 0}$ for Algorithm 3 was studied in the papers [AN17; WCP18]. We provide the proof under a weaker set of assumptions.

[4]It can be shown that such an inequality would hold at non-critical point of a continuous function f; see Remark 3.2 of [BDL07]. Note that the parameter θ and the neighborhood mentioned in Assumption KL above may depend on the point $\bar{x}$.

(a) The sequence $\{\mathbf{x}_k\}_{k\geq 0}$ *converges to a critical point* $\bar{x}$*, and for all* $k = 1, 2, \ldots$

$$\text{Avg}\left(\|\nabla f(\mathbf{x}_k)\|_2\right) \leq \frac{c_1}{k},$$

(b) Suppose that at the point $\bar{x}$*, the function* f *has a KL exponent* $\bar{\theta} \in \left[\frac{1}{2}, \frac{r}{2r-1}\right)$ *for some* $r > 1$*. Then we have*

$$\text{GAvg}\left(\|\nabla f(\mathbf{x}_k)\|_2\right) \leq \frac{c_2}{k^r} \qquad \text{for all} \quad k = 1, 2, \ldots,$$

where the constants (c_1, c_2) *are independent of* k*, but they may depend on the KL parameters at the point* $\bar{x}$*.*

See Section 8.10 for proof of this theorem.

Comments: It is worth pointing out that Theorem 6 does *not* require the function h to satisfy any smoothness assumption. Such conditions are needed for applying Algorithm 3, so that Theorem 6 is based on milder conditions than Theorem 7.

Our next result is to exhibit a faster convergence rate for Algorithm 3 under the KL assumption:

Theorem 7. *Suppose that, in addition to Assumptions PR & KL, the function* h *in Algorithm 3 is locally smooth. Then any bounded sequence* $\{\mathbf{x}_k\}_{k\geq 0}$ *obtained from Algorithm 3 satisfy the following properties:*

(a) The sequence $\{\mathbf{x}_k\}_{k\geq 0}$ *converges to a critical point* $\bar{x}$*, and for all* $k = 1, 2 \ldots$

$$\text{Avg}\left(\|\nabla f(\mathbf{x}_k)\|_2\right) \leq \frac{c_1}{k}.$$

(b) Given some $r > 1$*, suppose that at the point* $\bar{x}$ *the function* f *has a KL exponent* $\bar{\theta} \in \left[\frac{1}{2}, \frac{r}{2r-1}\right)$*. Then*

$$\text{GAvg}\left(\|\nabla f(\mathbf{x}_k)\|_2\right) \leq \frac{c_2}{k^r} \qquad \text{for all} \quad k = 1, 2, \ldots,$$

where the constants (c_1, c_2) *are independent of* k*, but they may depend on the KL parameters at the point* $\bar{x}$*.*

See Section 8.10 for the proof of this theorem.

Comments: Note that $\min_{1\leq i\leq k} \|\nabla f(\mathbf{x}_k)\|_2$ is upper bounded by the quantities $\text{Avg}\left(\|\nabla f(\mathbf{x}_k)\|_2\right)$ and $\text{GAvg}\left(\|\nabla f(\mathbf{x}_k)\|_2\right)$. It thus follows that the sequence $\{\|\nabla f(\mathbf{x}_k)\|_2\}_{k\geq 0}$ converges to zero at a rate of at least $1/k$, thereby improving the rate of convergence of $\|\nabla f(x)\|_2$ obtained in Theorems 3 and 4. When $\theta < \frac{1}{2}$, a simple modification of the proof (using $\gamma = 2$) shows that, Algorithms 2 and 3 converge in a finite number of steps. Finally, we point out that when the function h is non-smooth but the proximal-function φ is smooth, the existing proof can be easily modified to obtain a rate of convergence of the gradient-norm $\|\nabla f(\mathbf{x}_k)\|_2$.

8.4 Some illustrative applications

In this section, we study four interesting classes of non-convex problems that fall within the framework of this chapter. We also discuss various consequences of Theorems 3—7 as well as Corollaries 4—6 when applied to these problems.

Shape from shading

The problem of shape from shading is to reconstruct the three-dimensional (3D) shape of an object based on observing a two-dimensional (2D) image of intensities, along with some information about the light source direction. It is assumed that the observed 2D image intensity is determined by the angle between the light source direction and the surface normals of the object [EJ10].

In more detail, suppose that both the object and its 2D image are supported on a rectangular grid of size $r \times c$. We introduce the shorthand notation $[r] = \{1, 2, \ldots, r\}$ and $[c] = \{1, 2, \ldots, c\}$ for the rows and columns of this grid. For each pair $(i, j) \in [r] \times [c]$, we let $I_{ij} \in \mathbb{R}$ denote the observed intensity at location (i, j) in the image, and we let $\mathcal{N}ij \in \mathbb{R}^3$ denote the surface normal at the vertex $v_{ij} := (x_{ij}, y_{ij}, z_{ij})$ of the object. Based on observing the 2-dimensional image, both the intensity I_{ij} and co-ordinate pair (x_{ij}, y_{ij}) are known for each pair $(i, j) \in [r] \times [c]$. The goal of shape from shading is to estimate the unknown coordinate z_{ij}, which corresponds to the height of the object at location (i, j). Knowledge of these z-coordinates allows us to generate a 3D representation of the object, as illustrated in Figure 8.1.

Lambertian lighting model: In order to reconstruct the z-coordinates, we require a model that relates the observed intensity I_{ij} to the surface normal. In a Lambertian model, for a given light source direction $L := (\ell_1, \ell_2, \ell_2)^\top \in \mathbb{R}^3$, it is assumed that the surface normal $\mathcal{N}ij$ and intensity I_{ij} are related via the relation

$$I_{ij} = \frac{\langle L, \mathcal{N}ij \rangle}{\|\mathcal{N}ij\|_2}. \tag{8.19}$$

In one standard model [WSU14], the surface normal $\mathcal{N}ij := (p_{ij}, q_{ij}, 1)^\top$ is assumed to be determined by the triplet of vertices $(v_{ij}, v_{i+1,j}, v_{i,j+1})$ via the equations

$$p_{ij} = \frac{(y_{i,j+1} - y_{i,j})(z_{i+1,j} - z_{ij}) - (y_{i+1,j} - y_{i,j})(z_{i,j+1} - z_{ij})}{(x_{i,j+1} - x_{ij})(y_{i+1,j} - y_{ij}) - (x_{i+1,j} - x_{ij})(y_{i,j+1} - y_{ij})},$$
$$q_{ij} = \frac{(x_{i,j+1} - x_{i,j})(z_{i+1,j} - z_{ij}) - (x_{i+1,j} - x_{i,j})(z_{i,j+1} - z_{ij})}{(x_{i,j+1} - x_{ij})(y_{i+1,j} - y_{ij}) - (x_{i+1,j} - x_{ij})(y_{i,j+1} - y_{ij})}.$$

Squaring both sides of equation (8.19) and substituting the expression for surface normal $\mathcal{N}ij$ yields the polynomial equation

$$\Big(p_{ij}^2 + q_{ij}^2 + 1\Big) I_{ij} - (\ell_1 p_{ij} + \ell_2 q_{ij} + \ell_3)^2 = 0, \tag{8.20}$$

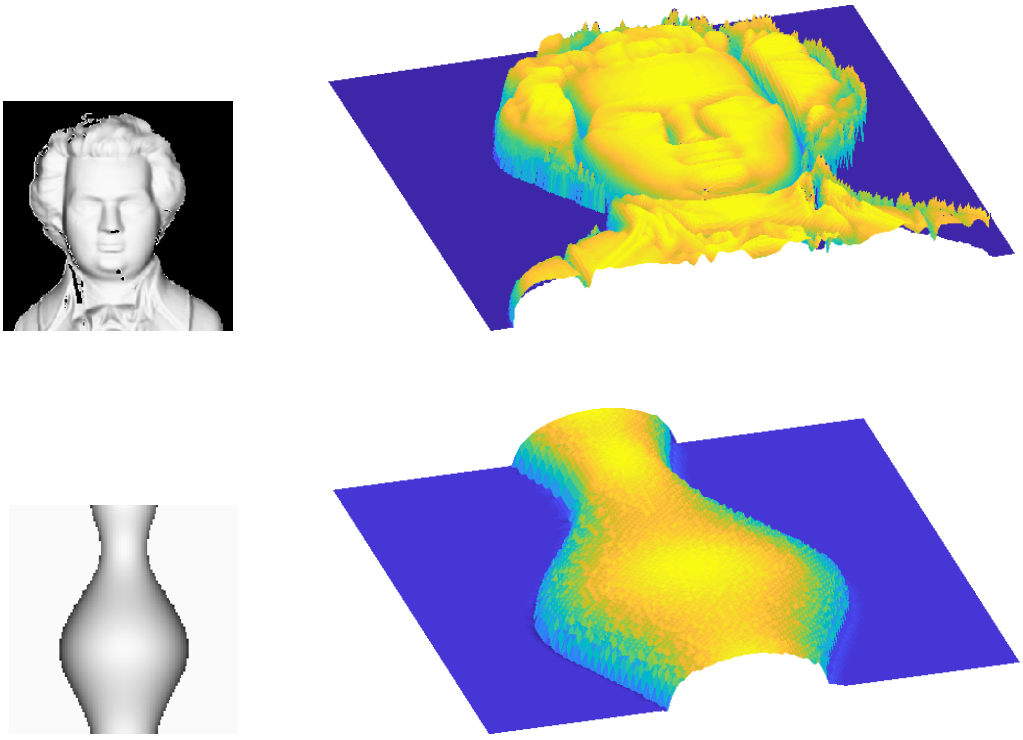


Figure 8.1. Figure shows 3D shape reconstruction of *Mozart* (first row) and *Vase* (second row) from corresponding 2D images. The gray-scale images in the left column are the 2D input images; the two colored images in the right column are the reconstructed 3D shapes. The 3D shapes are constructed by solving the problem (8.21) using Algorithm 5.

which should be satisfied under the assumed model.
In practice, this equality will not be exactly satisfied, but we can estimate the z-coordinates by solving the following non-convex optimization problem in the $r \times c$ matrix z with entries $\big\{z_{ij} \mid (i,j) \in [r] \times [c]\big\}$:

$$\min_{z \in \mathbb{R}^{r \times c}} \underbrace{\Big\{ \sum_{i=1}^{r} \sum_{j=1}^{c} \Big((1 + p_{ij}^2 + q_{ij}^2) I_{ij}^2 - (\ell_1 p_{ij} + \ell_2 q_{ij} + \ell_3)^2 \Big)^2 \Big\}}_{P(z)}. \tag{8.21}$$

Some reconstruction experiments: In order to illustrate the behavior of our method for this problem, we considered two synthetic images for simulated experiments. The first one is a 256×256 image of *Mozart* [Zha+99], and the second one is a 128×128 image of *Vase*. The 3D shapes were constructed from the 2D images by solving optimization problem (8.21) using the backtracking gradient descent algorithm 5. The reconstructed surfaces for *Vase* and *Mozart* are provided in

Figure 8.1. We ran 500 iterations of Algorithm 5 for both the images. The runtime for *Mozart*-example was 87 seconds, whereas the runtime for *Vase*-example was 39 seconds. The implementation of Algorithm 5 for Problem (8.21) is parallelizable; hence, the runtime can be much lower than our runtime with a parallel implementation. It is worth mentioning that the polynomial P is a fourth-degree polynomial with dimension $r \times c$; polynomial P is coercive and bounded below by zero. Consequently, we can apply Corollary 5 to the problem (8.21) which guarantees that average of the squared gradient norm $\text{Avg}\left(\|\nabla P\|_2^2\right)$ converges to zero at a rate $\frac{1}{k}$.

One might also consider applying the CCCP method to this problem. In a recent paper, [WSU14] provided a DC decomposition of the polynomial P using a sum of square (SOS) optimization technique. However, it is crucial to note that the DC decomposition of polynomial P obtained from the SOS-optimization method need not be optimal. In order to see this, note that the dimension of the polynomial P is much larger than three. In particular, the variable z_{ij} is used in the computation of surface normals $\mathcal{N}ij, \mathcal{N}i, j-1$ and $\mathcal{N}i-1, j$, hence is related to variables $(z_{i,j+1}, z_{i+1,j}, z_{i-1,j}, z_{i,j-1})$—which are again related to the other variables. [AP13] showed that SOS techniques for deriving a DC decomposition are sub-optimal for a fourth-degree polynomial when the dimension of the polynomial is greater than three. Consequently, deriving an optimal DC decomposition for the polynomial P will be computationally intensive.

Robust regression using Tukey's bi-weight

Next, we turn to the problem of robust regression with Tukey's bi-weight penalty function. Suppose that we observe pairs $(yi, z_i) \in \mathbb{R} \times \mathbb{R}^d$ linked via the noisy linear model

$$yi = \langle z_i,\, \mu^* \rangle + \varepsilon_i \qquad \text{for } i = 1, \ldots, n.$$

Here the vector $\mu^* \in \mathbb{R}^d$ is the unknown parameter of interest, whereas the variables $\{\varepsilon_i\}_{i=1}^n$ correspond to additive noise. In robust regression, we obtain an estimate of the parameter vector μ^* by computing

$$\min_{\mu \in \mathbb{R}^d} \underbrace{\Big\{ \frac{1}{n} \sum_{i=1}^{n} \Psi\Big(y_i - \langle z_i,\, \mu \rangle\Big) \Big\}}_{=: f(\mu)} \tag{8.22}$$

where Ψ is a known loss function with some robustness properties. One popular example of the loss function Ψ is Tukey's bi-weight function, which is given by

$$\Psi(t) = \begin{cases} 1 - (1 - (t/\lambda)^2)^3 & \text{if } |t| \leq \lambda \\ 1 & \text{otherwise} \end{cases}, \tag{8.23}$$

where $\lambda > 0$ is a tuning parameter. Note that Ψ is a smooth function, whence the function f in the objective (8.22) is also smooth, implying that Algorithm 2 is suitable

for the problem.

With this set-up, applying Theorem 3, Theorem 6 and Corollary 6, we obtain the following guarantee:

Corollary 7. *Given a random initialization, any bounded sequence $\{\mu^k\}_{k\geq 0}$ obtained by applying Algorithm 2 to the objective (8.22) has the following properties:*

(a) Almost surely with respect to the random initialization, the sequence $\{\mu^k\}_{k\geq 0}$ converges to a point $\bar{\mu}$ such that $\nabla f(\bar{\mu}) = 0$ and $\nabla^2 f(\bar{\mu}) \succeq 0$.

(b) There is a universal constant c_1 such that

$$\text{Avg}\Big(\|\nabla f(\mu^k)\|_2\Big) \leq \frac{c_1}{k} \qquad \text{for all } k = 1, 2, \ldots.$$

We provide the proof in Section 8.11.

Smooth function minimization with sparsity constraints

Moving beyond the robust regression problem, we now discuss another interesting problem of minimizing a smooth function subject to sparsity penalty. Consider the following optimization problem

$$\min_{\substack{x\in\mathbb{R}^d \\ \|x\|_0\leq s}} g(x), \tag{8.24}$$

where g is a smooth function, the ℓ_0-"norm" $\|x\|_0$ counts the number of non-zero entries in the vector x, and $s \in \{1, \ldots, d\}$ is a sparsity parameter. The constraint set $\Big\{x \in \mathbb{R}^d \mid \|x\|_0 \leq s\Big\}$ is non-convex, and consequently, the optimization problem (8.24) is non-convex. However, the constraint set can be expressed as the level set of a certain DC function [GTT17]. In particular, let $|x|_{(d)} \geq |x|_{(d-1)} \geq \cdots \geq |x|_{(1)}$ denote the values of $x \in \mathbb{R}^d$ re-ordered in terms of their absolute magnitudes. In terms of this notation, we have $\|x\|_1 \geq \sum_{i=d-s+1}^{d} |x|_{(i)}$ for all $x \in \mathbb{R}^d$, with equality holding if and only if x is s–sparse. This fact ensures that

$$\Big\{x \in \mathbb{R}^d : \|x\|_0 \leq s\Big\} = \Big\{x \in \mathbb{R}^d : \|x\|_1 - \sum_{i=d-s+1}^{d} |x|_{(i)} \leq 0\Big\}. \tag{8.25}$$

Since both of the functions $x \mapsto \|x\|_1$ and $x \mapsto \sum_{i=d-s+1}^{d} |x|_{(i)}$ are convex [BV04], this level set formulation is a DC constraint. Now using the representation (8.25), we can rewrite problem (8.24) as $\min_{x\in\mathbb{R}^d} g(x)$ such that $\|x\|_1 - \sum_{i=d-s+1}^{d} |x|_{(i)} \leq 0$. For our experiments, it is more convenient to solve the penalized analogue of the last problem, given by

$$\min_{x\in\mathbb{R}^d} \Big\{g(x) + \lambda\Big(\|x\|_1 - \sum_{i=d-s+1}^{d} |x|_{(i)}\Big)\Big\}, \tag{8.26}$$

where $\lambda > 0$ is a tuning parameter. The optimization problem (8.26) can be solved using Algorithm 3 with $g(x) = g(x)$, $\varphi(x) = \lambda\|x\|_1$ and $h(x) = \lambda \sum_{i=d-s+1}^{d} |x|_{(i)}$. For the non-smooth component $\varphi(x) = \lambda\|x\|_1$, there is a closed form expression of the proximal update in Algorithm 3, so that the method is especially efficient in this case.

Best subset selection

A special case of problem (8.26) arises from best subset selection in linear regression. Suppose that we observe a vector $y \in \mathbb{R}^n$ and a matrix $B \in \mathbb{R}^{n\times d}$ that are linked via the standard linear model $y = Bx^* + \varepsilon$. Here the vector $\varepsilon \in \mathbb{R}^n$ corresponds to additive noise, whereas $x^* \in \mathbb{R}^d$ is the unknown regression vector. We wish to estimate the unknown parameter vector $\mathbf{x}_*$ subject to a sparsity constraint, and we do so by solving the following optimization problem:

$$\min_{\substack{x\in\mathbb{R}^d \\ \|x\|_0 \le s}} \|y - Bx\|_2^2. \tag{8.27}$$

Here the non-negative integer s is a tuning parameter that controls maximum number of allowable non-zero entries in the vector x. Following the development leading to the formulation (8.26), let us consider instead the problem of minimizing the function

$$f(x) := \|y - Bx\|_2^2 + \lambda\Big\{\|x\|_1 - \sum_{i=d-s+1}^{d} |x|_{(i)}\Big)\Big\}. \tag{8.28}$$

Note that the function f can be decomposed as a difference of two convex functions as follows:

$$f(x) = \underbrace{\|y - Bx\|_2^2 + \lambda\|x\|_1}_{\text{convex}} - \underbrace{\lambda \sum_{i=d-s+1}^{d} |x|_{(i)}}_{\text{convex}}. \tag{8.29}$$

Consequently, problem (8.28) is a DC optimization problem; hence, it is amenable to standard DC optimization techniques like CCCP. We can also apply Algorithm 3 on problem (8.28) with $g(x) = \|y - Bx\|_2^2$, $\varphi(x) = \lambda\|x\|_1$ and $h(x) = \lambda \sum_{i=d-s+1}^{d} |x|_{(i)}$.

Comparison of Algorithm 3 and CCCP

Let us compare the performance of our Algorithm 2 (prox-type method) with the popular convex-concave procedure (CCCP) for minimizing differences of convex functions. We apply both algorithms to the best subset selection problem (8.28).

Let us reiterate that problem (8.28) can be written as a difference of two convex functions, and one can apply CCCP update (8.10) to the decomposition (8.29). The inner convex optimization problem in update (8.10) is solved by proximal methods for minimizing the sum of a smooth convex function and a ℓ_1 regularizer. We also apply Algorithm 3 on problem (8.28) with $g(x) = \|y - Bx\|_2^2$, $h(x) = \lambda \sum_{i=d-s+1}^{d} |x|_{(i)}$ and $\varphi(x) = \lambda\|x\|_1$.

Synthetic data generation: We generated the rows of the $n \times d$ matrix B from a d-dimensional Gaussian distribution with zero mean and an equicovariance matrix Σ, where $\Sigma_{ii} = 1$ for all i, and $\Sigma_{ij} = 0.7$ for all $i \neq j$. The regression vector $x^* \in \mathbb{R}^d$ (true value) was chosen to be a binary vector with sparsity s ($s \ll d$). The location of the nonzero entries of the vector x^* was chosen uniformly without replacement form the set $\Big\{1, \ldots, d\Big\}$.

Performance measures: We use the following two criteria to compare the performance of the prox-type method and CCCP.

(a) *Total runtime:* Firstly, we compare the algorithms in terms of their total runtime. The runtime was measured in units of seconds.

(b) *Estimation error:* Secondly, we use average estimation error of the algorithms as a measure of performance. Let us recall that if $\bar{x} \in \mathbb{R}^d$ is the estimated value of the unknown regression vector x^*, then the average estimation error is defined as $\frac{\|\bar{x}-x^*\|_2}{\sqrt{p}\|\bar{x}\|_2}$. Note that the average estimation error used here is invariant under scaling.

Comparison results: Figure 8.2 shows the performances of the prox-type method and CCCP for synthetic data simulated as above, with problem parameters $(n, p) = (190, 300)$ and $(n, p) = (380, 600)$ and different choices of sparsity s.

For both the algorithms, the tolerance level η was set to $\eta = 10^{-8}$, whereas the maximum number of iterations was 1000. Figure 8.2 suggests that total runtime of the prox-type method is significantly smaller than the runtime of CCCP. Furthermore, the estimation error for the prox-type method is lower compared to CCCP, which possibly suggests that prox-type method is finding better local minima compared to CCCP for the non-convex optimization problem (8.28). In all our simulations we used same initializations for both the algorithms. The simulation results shown in Figure 8.2 are average over 100 replications, and we also provide the pointwise error bar in the plots.

Some theoretical guarantees

Interestingly, it turns out that when applied to problem (8.28), the convergence behavior of Algorithm 3 to a given stationary point $\bar{x}$ depends on the behavior of a certain convex program defined in terms of $\bar{x}$. More precisely, for any point $\bar{x} \in \mathbb{R}^d$ with $|\bar{x}|_{(r)} > |\bar{x}|_{(r+1)}$, consider the following convex relaxation of problem (8.28):

$$\mathcal{P}(\bar{x}) := \min_{x \in \mathbb{R}^d} \Big\{ \|y - Bx\|_2^2 + \lambda \|x\|_1 - \lambda \langle \nabla h(\bar{x}),\, x - \bar{x} \rangle \Big\}. \tag{8.30}$$

Note that $|\bar{x}|_{(r)} > |\bar{x}|_{(r+1)}$ implies the differentiability of the function $h := \lambda \sum_{i=d-s+1}^{d} |x|_{(i)}$ which ensures that the above problem is well-defined.

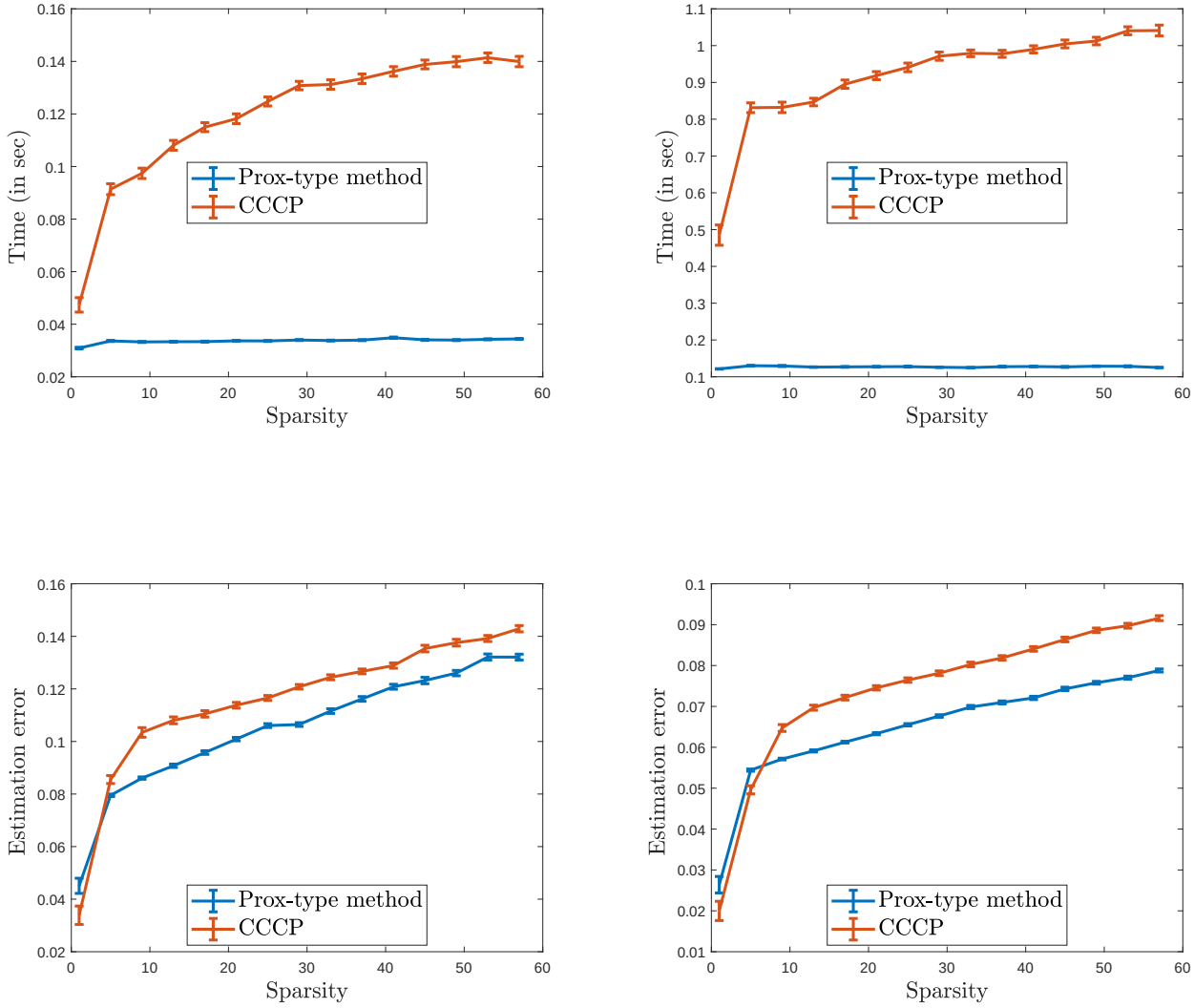


Figure 8.2. Performance of CCCP compared to that of Algorithm 3 on the best subset selection problem for synthetic data for different values of (n, p). The left columns correspond to $(n, p) = (190, 300)$, whereas the right columns correspond to $(n, p) = (380, 600)$. Plots in the first row compare the performance in terms of total runtime, those in the second row compare algorithms in terms of estimation error. We see that Algorithm 3 outperforms CCCP in terms of runtime. The performance of Algorithm 3 and CCCP in terms of estimation error are similar for low values of sparsity, whereas Algorithm 3 outperforms CCCP when sparsity is moderate to large. We initialized both the algorithms from the same starting point. Results shown above are averaged over 100 replications, and we also provide point-wise error bars in the plots.

Corollary 8. *Let* $\left\{\mathbf{x}_k\right\}_{k\geq 0}$ *be any bounded sequence obtained by applying Algorithm 3 on problem (8.28). Suppose there exists a limit point* $\bar{x}$ *of the sequence* $\left\{\mathbf{x}_k\right\}_{k\geq 0}$ *satisfying* $|\bar{x}|_{(r)} > |\bar{x}|_{(r+1)}$*, and the convex problem (8.30) has unique solution. Then the sequence* $\left\{\mathbf{x}_k\right\}_{k\geq 0}$ *converges to the point* $\bar{x}$*, and for all* $k = 1, 2, \ldots$*, we have*

$$\text{Avg}\left(\|\nabla f(\mathbf{x}_k)\|_2\right) \leq \frac{c_1}{k}, \quad \textit{and} \quad \|\mathbf{x}_k - \bar{x}\|_2 \leq cq^k,$$

where $q \in (0, 1)$*, and* (c, c_1) *are positive constants independent of* k*.*

Comments on problem (8.30): It can be shown that when the matrix B is of full rank, the objective function in problem (8.30) is strictly convex, and as a result, the problem (8.30) has unique solution. In the proof of Corollary 8, we show that the point $\bar{x}$ is always a minimizer of the convex problem (8.30), so that the uniqueness assumption implies that $\bar{x}$ is in fact the unique solution.

Mixture density estimation

As a final example, we consider the problem of estimating a two-component mixture density, where each of the constituent densities belong to an exponential family. The density of an exponential family (with respect to a fixed base measure, typically counting or Lebesgue) takes the form

$$p(y; \eta) = g(y) \exp\Big\{\langle \eta,\, T(y)\rangle - A(\eta)\Big\}. \tag{8.31}$$

Here the function $T : \mathcal{Y} \to \mathbb{R}^d$ is a vector of sufficient statistics, whereas the log-partition function

$$A(\eta) := \log\left(\int_{\mathcal{Y}} g(y) \exp\{\langle \eta,\, T(y)\rangle\} dy\right)$$

serves to normalize the density. The parameter vector $\eta \in \mathbb{R}^d$ determines the choice of density within the family. See Table 8.1 for some examples of 1-dimensional exponential families of this type. It includes various familiar examples, such as the Gaussian, Poisson and Beta families.

In the problem of mixture density estimation, one is interested in densities of the form

$$\zeta(y; \underbrace{\pi, \eta_0, \eta_1}_{\theta}) = \pi\, p(y; \eta_0) + (1 - \pi) p(y; \eta_1), \tag{8.32}$$

[5]The shape parameter$k$ is known

Distribution Name	η	$A(\eta)$	Twice continuously differentiable and sub-analytic
Poisson (λ)	$\ln(\lambda)$	$\exp \eta$	✓
Geometric (p)	$\ln(p)$	$-\ln(1-\exp \eta)$	✓
Gaussian(μ, σ^2)	$\left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^\top$	$-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2)$	✓
Exponential (λ)	$-\lambda$	$-\ln(-\eta)$	✓
Gamma (α, β)	$(\alpha-1, \beta)^\top$	$\ln \Gamma(\eta_1+1) - (\eta_1+1)\ln(\eta_2)$	✓
Weibull $(\lambda, k$[5]$)$	$-\frac{1}{\lambda^k}$	$\ln(-\eta) - \ln(k)$	✓
Beta (α, β)	$(\alpha, \beta)^\top$	$\ln \Gamma(\eta_1) + \ln \Gamma(\eta_2) - \ln \Gamma(\eta_1 + \eta_2)$	✓

Table 8.1. Table showing the natural parameter η and the log-partition function A for different densities of exponential family, which are twice continuously differentiable and sub-analytic. In Section 8.11 we prove the log-partition functions A mentioned in the above table are sub-analytic.

where $\pi \in (0,1)$ is an unknown mixing proportion, and (η_0, η_1) are the unknown parameters of the two underlying densities.

Given n i.i.d. samples $\{y_i\}_{i=1}^n$ drawn from a mixture density of the form (8.32), a standard goal is to estimate the unknown parameter vector $\theta := (\pi, \eta_0, \eta_1)$. One way to do so is by computing the maximum likelihood estimate (MLE), obtained via minimizing the negative log-likelihood of parameter θ given by the data. Frequently, a regularized form of the MLE is used, say of the form

$$\min_{\theta} \underbrace{\Big\{ -\sum_{i=1}^{n} \log\Big(\zeta(y_i;\theta)\Big) \Big\}}_{g(\theta)} \qquad \text{with } \eta_0, \eta_1 \in \mathbb{R}^d,\ \pi \in [0,1], \text{ and } \|\eta_0\|_2 \leq R_0,\ \|\eta_1\|_2 \leq R_1. \tag{8.33}$$

Here $R_0 > 0$ and $R_1 > 0$ are tuning parameters providing upper bound on the ℓ_2-norms of the parameters η_0 and η_1 respectively, often chosen by a data-dependent procedure (such as cross-validation).

By inspection, the objective function g in problem (8.33) is non-convex. By standard theory on exponential families, the function A is always infinitely differentiable on its domain, so that the objective function g is infinitely differentiable on the convex set

$$\mathcal{X} = \Big\{ \theta = (\eta_0, \eta_1, \pi) \ \mid \ \eta_j \in \mathrm{dom}(A), \pi \in [0,1], \|\eta_j\|_2 \leq R_j \quad \text{for } j = 0, 1 \Big\}.$$

Consequently, we may apply Algorithm 3 with $g(\cdot) = -\sum_{i=1}^{n} \log\Big(\zeta(\cdot; y_i)\Big)$, $h \equiv 0$ and $\varphi(\cdot) = \mathbb{1}_{\mathcal{X}}(\cdot)$ and $f = g - h + \varphi$. Interestingly, the log-partition function A is sub-analytic for many exponential family densities (see Table 8.1), which ensures

that the function g is also sub-analytic. In Section 8.6, we show that continuous sub-analytic functions satisfy Assumption KL so that we can apply Theorem 7 to obtain the following:

Corollary 9. *Any sequence* $\{\theta^k\}_{k\geq 0} = \left\{\eta_0^k, \eta_1^k, \pi^k\right\}_{k\geq 0}$ *obtained by applying Algorithm 3 to problem* (8.33) *satisfies the following properties:*

(a) It converges to a first order stationary point.

(b) For all $k = 1, 2, \ldots$, *we have* $\text{Avg}\left(\|\nabla f(\theta^k)\|_2\right) \leq \frac{c_1}{k}$, *where* c_1 *is a universal constant independent of* k.

See Section 8.11 for the proof of this corollary.

8.5 Discussion

In this chapter, we analyzed the behavior of three gradient-based algorithms—namely gradient descent, a proximal method, and an algorithm of the Frank-Wolfe type—for finding critical points of a class of non-convex non-smooth optimization problems. For each of the three algorithms, we provided non-asymptotic bounds on the rate of convergence to a first-order stationary point. We showed that our algorithm can escape strict saddle point for a class of non-smooth functions, thereby generalizing existing results for smooth functions. As a consequence of our theory, we obtained a simplification of the popular CCCP algorithm, and the simplified algorithm retains all the convergence properties of CCCP. Finally, we showed that for a large subclass of functions, which include continuous sub-analytic functions as a special case, we can have a significant improvement in the rate of convergence.

Our work leaves open a number of questions for future research. For instance, it would be interesting to characterize the class of DC-based functions mentioned in problem (8.2) when the convex function h is non-differentiable. Indeed, we then obtain a larger non-class of non-differentiable functions, and we suspect that Theorem 8 can be suitably generalized. Finally, we suspect that the proof techniques used here can be leveraged in order to establish sharper results for other forms of non-convex optimization problems.

Acknowledgments

This work was partially supported by the Office of Naval Research Grant DOD ONR-N00014 and National Science Foundation Grant NSF-DMS-1612948.

8.6 Technical background

In this section, we collect some technical background on subdifferentials and sub-analytic functions.

Fréchet and limiting subdifferential

We first recall the definitions and some useful properties of sub-differentials, which will be useful in subsequent sections.

Definition 1. *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a lower semicontinuous function. For any $x \in \mathrm{dom}(f)$, the Fréchet subgradient of the function f at point x is defined as*

$$\widehat{\partial} f(x) = \left\{ u \;\middle|\; \liminf_{y \neq x, y \to x} \frac{f(y) - f(x) - \langle u, y - x \rangle}{\|y - x\|_2} \geq 0 \right\}.$$

Definition 2. *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a lower semi-continuous function. For any $x \in \mathrm{dom}(f)$, the limiting subdifferential of the function f at point x is defined as*

$$\partial_L f(x) = \left\{ u \;\middle|\; \exists\ \mathbf{x}_k \to x, u^k \to u \text{ with } f(x^k) \to f(x) \text{ and } u^k \in \widehat{\partial} f(\mathbf{x}_k) \text{ as } k \to \infty \right\}.$$

Properties: The following properties of Fréchet and limiting sub-differential are provided in Chapter 8 of [RW09].

(a) For any proper convex function h, we have $\partial_L h(x) = \widehat{\partial} h(x)$ for all $x \in \mathrm{dom}(h)$, and both quantities agree with the usual subgradient of the convex function h.

(b) If a function g is smooth in a neighborhood of a point x, then $\partial_L f(x) = \nabla f(x)$.

(c) Consider a function f of the form $f = g + \varphi$, where the function g is smooth in a neighborhood of a point x, and the function φ is proper convex and finite at the point x. Then the limiting sub-differential of the function f at the point x is given by $\partial_L f(x) = \nabla g(x) + \partial \varphi(x)$.

(d) *(Graph continuity:)* Consider a sequence $\left\{ \left(\mathbf{x}_k, u^k \right) \right\}_{k \geq 1}$ in $\mathrm{graph}(\partial_L f)$ such that the sequence $\{(\mathbf{x}_k, u^k, f(\mathbf{x}_k)\}_{k \geq 0}$ converges to a point $(x, u, f(x))$. Then $(x, u) \in \mathrm{graph}(\partial_L f)$. Recall that $\mathrm{graph}(\partial_L f) := \left\{ (x, u) \in \mathbb{R}^d \times \mathbb{R} \mid u \in \partial_L f(x) \right\}$.

.

Sub-analytic functions satisfy KL-assumption

In this section, we show that continuous sub-analytic functions satisfy the KL-inequality. We also provide examples of functions which are sub-analytic.

Comments on limiting sub-differential: In order to facilitate our discussion, we mention some simple facts on limiting subdifferential of a function f, where f is of the form $f = g - h$ (Theorems 3 and 6) or $f = g + \varphi - h$ (Theorems 4 and 7). The following properties are direct consequences of properties of the limiting subdifferential mentioned in Section 8.6.

- Suppose that the difference function $f = g - h$ satisfies parts (a) and (b) of Assumption GR. Then we have

$$\partial_L(-f)(x) = \partial h(x) - \nabla g(x), \quad \text{and moreover}$$
$$\|\nabla f(x)\|_2 := \|\nabla g(x) - \partial h(x)\|_2 = \|\partial_L(-f)(x)\|_2$$

- Suppose that the function $f = g + \varphi - h$, where the function h is locally smooth, and the function f satisfies Assumption PR part (b). Then $\partial_L f(x) = \nabla g(x) - \nabla h(x) + \partial\varphi(x)$. Consequently, we have that $\|\nabla f(x)\|_2 = \|\partial_L f(x)\|_2$.

We prove that continuous sub-analytic functions satisfy Assumption KL by exploiting results due to [BDL07]. Let us introduce some notation used in this chapter. We use $m_f(x)$ to denote the ℓ_2 distance of the set $\partial_L f(x)$ from zero; concretely, $m_f(x) := \text{dist}_{\|\cdot\|_2}\left(0, \partial_L f(x)\right)$. In Theorem 3.1 (for critical points of the function f) and Remark 3.2 (for non-critical points of the function f), [BDL07] proved the following fact about sub-analytic functions.

Lemma 12. ([BDL07]): *Let $f : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ be a sub-analytic function with closed domain, and assume that $f|_{\text{dom}(f)}$ is continuous. Then for any $a \in \text{dom}(f)$, there exists an exponent $\theta \in [0, 1)$ such that, the function $\frac{|f-f(a)|^\theta}{m_f}$ is bounded above in a neighborhood of a.*

Using Lemma 12, we now argue that sub-analytic functions, under the conditions of Theorem 6 or Theorem 7, satisfy Assumption KL.

Lemma 13. *Any sub-analytic function f satisfying Assumption GR also satisfies Assumption KL.*

Proof. First, note that the function f is continuous by Assumption GR; suppose f is sub-analytic, then from properties of sub-analytic functions, we have that the function $-f$ is also sub-analytic. Furthermore, the function $-f$ is continuous in the closed domain Ω —which by Lemma 12 guarantees that, for any $a \in \Omega$, there exists $\theta \in [0, 1)$ such that the ratio $\frac{|-f-(-f(a))|^\theta}{m_{(-f)}}$ is bounded above in a neighborhood of the point a. Since $|-f-(-f(a))| = |f - f(a)|$, proving satisfiability of Assumption KL reduces to showing that $m_{(-f)}(x)$ is upper bounded by $\|\nabla f(x)\|_2$. To this end, note that from the discussion about limiting subdifferential in the paragraph above Lemma 12, we have

$$\|\nabla f(x)\|_2 = \|\partial_L(-f)(x)\|_2 \overset{(i)}{\geq} m_{(-f)}(x), \tag{8.34}$$

where step (i) follows from the definition of $m_{(-f)}(x)$. Putting together the pieces, we conclude that any sub-analytic function f which satisfies Assumption GR, also satisfies Assumption KL. □

Lemma 14. *Suppose that, in addition to the conditions on the functions (g, h, φ) from Theorem 4, the function $f := g - h + \varphi$ is continuous and sub-analytic in its domain $\mathrm{dom}(f)$, and the domain $\mathrm{dom}(f)$ is closed. Then the function f satisfies Assumption KL.*

Proof. Since the function $f|_{\mathrm{dom}(f)}$ is continuous and sub-analytic by assumption, from Lemma 12, we have that for any $a \in \mathrm{dom}(f)$ there exists a $\theta \in [0, 1)$ such that, the ratio $\frac{|f - f(a)|^\theta}{m_f}$ is bounded above in a neighborhood of the point a. In order to justify satisfiability of Assumption KL, it suffices to prove that $m_f(x)$ is upper bounded by $\|\nabla f(x)\|_2$. To this end, note that the function h is locally smooth by assumptions of Theorem 4 part (b). Hence, from the discussion about limiting subdifferential in the paragraph above Lemma 12, we have

$$\|\nabla f(x)\|_2 = \|\partial_L f(x)\|_2 \overset{(i)}{\geq} m_f(x), \tag{8.35}$$

where step (i) follows from the definition of $m_f(x)$. Putting together the pieces, guarantees that the function f satisfies Assumption KL. □

Instances of sub-analytic functions

In Section 8.6, we proved that continuous sub-analytic functions satisfy Assumption KL, and in those cases,—by Theorems 6 and 7—we have a faster rate of convergence of Algorithms 2 and 3. In this section, we provide examples of functions which are sub-analytic. We start by providing definitions of sub-analytic functions following the definition of [BDL07].

A subset $S \subset \mathbb{R}^d$ is called *semi-analytic* if each point of $\mathbb{R}^d$ admits a neighborhood V such that the set $S \cap V$ has the form

$$S \cap V = \cup_{i=1}^{p} \cap_{j=1}^{q} \Big\{ x \in V \mid h_{ij} = 0, g_{ij} > 0 \Big\},$$

where the functions $h_{ij}, g_{ij} : V \mapsto \mathbb{R}$ are real-analytic.

A set S is called *sub-analytic*, if each point of $\mathbb{R}^d$ admits a neighborhood V such that

$$S \cap V = \Big\{ x \in \mathbb{R}^d : \Big(x, y\Big) \in B \Big\},$$

where B is a bounded semi-analytic subset of $\mathbb{R}^d \times \mathbb{R}^m$ for some $m \geq 1$. A function f is called sub-analytic if the graph of f, defined by $\mathrm{graph}(f) := \Big\{(x, y) \in \mathbb{R}^d \times \mathbb{R} : f(x) = y\Big\}$, is sub-analytic.

The class of sub-analytic functions is quite large. In order to motivate the reader, we provide few examples here. The following results can be found in [BST14] and Chapter 6 in the book by [FP07].

(a) Any real-valued polynomial or analytic function is sub-analytic.

(b) Any real-valued semi-algebraic or semi-analytic function is sub-analytic.

(c) Indicator function of a semi-algebraic set is sub-analytic.

(d) Sub-analytic functions are closed under finite linear combinations, and the product of two sub-analytic functions is sub-analytic.

(e) Point-wise maximum and minimum of a finite collection of sub-analytic functions are sub-analytic.

(f) *Composition rule:* If g_1 and g_2 are two sub-analytic functions with the function g_1 being continuous, then the composition function $g_2 \circ g_1$ is sub-analytic. In fact, the class of continuous sub-analytic functions are *closed under algebraic operations*.

8.7 Proofs related to Algorithm 2

In this section, we collect the proofs of various results related to the gradient-based Algorithm 2, including Theorem 3, Corollaries 4 and 6, and Proposition 3.

Proof of Theorem 3

Our proof of this theorem, as well as subsequent ones, depends on the following descent lemma:

Lemma 15. *Under the conditions of Theorem 3, we have*

$$\mathbf{x}_k \in \text{int}(\Omega) \quad \textit{and} \quad f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{\alpha}{2}\|\nabla f(\mathbf{x}_k)\|_2^2 \quad \textit{for all } k = 0, 1, 2, \ldots. \tag{8.36}$$

See Section 8.7 for the proof of this lemma.

We now prove Theorem 3 using Lemma 15.

Convergence of function values: We first prove that the function value sequence $\{f(\mathbf{x}_k)\}_{k\geq 0}$ is convergent. Since $f^* := \min_{x\in\Omega} f(x)$ is finite by assumption, and $\mathbf{x}_k \in \text{int}(\Omega)$ for all $k \geq 0$ by Lemma 15, the sequence $\{f(\mathbf{x}_k)\}_{k\geq 0}$ is bounded below. For any non-stationary $\mathbf{x}_k$, inequality (8.36) also ensures that $f(\mathbf{x}_k) > f(\mathbf{x}_{k+1})$; hence, there must exist some scalar $\bar{f}$ such that $\lim_{k\to\infty} f(\mathbf{x}_k) = \bar{f}$.

Stationarity of limit points: Next, we establish that any limit point of the sequence $\{\mathbf{x}_k\}_{k\geq 0}$ must be stationary. Consider a subsequence $\{\mathbf{x}_{k_j}\}_{j\geq 0}$ of $\left\{\mathbf{x}_k\right\}_{k\geq 0}$ such that $\mathbf{x}_{k_j} \to \bar{x}$, and let $\{u^{k_j}\}_{j\geq 0}$ be the associated sequence of subgradients. It suffices to exhibit a sub-gradient $\bar{u} \in \partial h(\bar{x})$ such that $\nabla g(\bar{x}) - \bar{u} = 0$. Since the sequence $\{\mathbf{x}_{k_j}\}_{j\geq 0}$ converges to $\bar{x}$, we must have

$$\|\nabla f(\mathbf{x}_{k_j})\|_2 = \|\nabla g(\mathbf{x}_{k_j}) - u^{k_j}\|_2 \to 0.$$

The function g is continuously differentiable by assumption, and we have $\nabla g(\mathbf{x}_{k_j}) \to \nabla g(\bar{x})$. Combining these we find that $u^{k_j} \to \nabla g(\bar{x})$. Furthermore, by continuity of the function g, we have $g(\mathbf{x}_{k_j}) \to g(\bar{x})$. Putting together the pieces we have established above that $\left(\mathbf{x}_{k_j}, u^{k_j}, g(\mathbf{x}_{k_j})\right) \to \left(\bar{x}, \bar{u}, g(\bar{x})\right)$, where $\bar{u} := \nabla g(\bar{x})$. Consequently, the graph continuity of limiting-sub-differentials (see Section 8.6) guarantees that $\bar{u} = \nabla g(\bar{x}) \in \partial h(\bar{x})$. Overall, we conclude that $\nabla f(\bar{x}) := \nabla g(x) - \bar{u} = 0$, so that $\bar{x}$ is a stationary point as claimed.

Establishing the bound (8.3): Finally, we prove the claimed bound (8.3) on the averaged squared gradient. Recalling that $f^* := \min_{x\in\Omega} f(x)$ is finite, we have

$$\begin{aligned} f(\mathbf{x}_0) - f^* \geq f(\mathbf{x}_0) - f(x^{k+1}) &= \sum_{j=0}^{k} f(x^j) - f(x^{j+1}) \\ &\overset{(i)}{\geq} \frac{\alpha}{2} \sum_{j=0}^{k} \|\nabla f(\mathbf{x}_k)\|_2^2 \\ &= \frac{\alpha(k+1)}{2} \text{Avg}\left(\|\nabla f(\mathbf{x}_k)\|_2^2\right), \end{aligned}$$

where step (i) follows from equation (8.36). Rearranging yields the claimed bound (8.3) on the averaged squared gradient.

Proof of Lemma 15

Recall that by assumption, the function g is continuously differentiable and M_g-smooth, and the function h is convex. As a consequence, for any vector $\mathbf{x}_k \in \Omega$ and subgradient $u^k \in \partial h(\mathbf{x}_k)$, we have

$$g(x) \leq g(\mathbf{x}_k) + \langle \nabla g(\mathbf{x}_k),\, x - \mathbf{x}_k \rangle + \frac{M_g}{2}\|x - \mathbf{x}_k\|_2^2 \tag{8.37a}$$

$$h(x) \geq h(\mathbf{x}_k) + \langle u^k,\, x - \mathbf{x}_k \rangle. \tag{8.37b}$$

Combining inequalities (8.37a) and (8.37b) yield

$$f(x) = g(x) - h(x) \leq f(\mathbf{x}_k) + \langle \nabla g(\mathbf{x}_k) - u^k,\, x - \mathbf{x}_k \rangle + \frac{M_g}{2}\|x - \mathbf{x}_k\|_2^2. \tag{8.38}$$

Substituting $x = \mathbf{x}_{k+1} := \mathbf{x}_k - \alpha\Big(\nabla g(\mathbf{x}_k) - u^k\Big)$ in equation (8.38) and simplifying yields

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \Big(\frac{1}{\alpha} - \frac{M_g}{2}\Big)\|\mathbf{x}_{k+1} - x^k\|_2^2 = \alpha\Big(1 - \frac{\alpha M_g}{2}\Big)\|\nabla g(\mathbf{x}_k) - u^k\|_2^2$$
$$\overset{(i)}{\geq} \frac{\alpha}{2}\|\nabla f(\mathbf{x}_k)\|_2^2,$$

where inequality (i) follows from the upper bound $\alpha \leq \frac{1}{M_g}$. This proves the second part of the stated lemma. As for the claim that the sequence remains in the interior of the set Ω, note that $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) \leq f(\mathbf{x}_0)$, which ensures that $\mathbf{x}_{k+1} \in \mathcal{L}(f(\mathbf{x}0)) \subset \text{int}(\Omega)$, as claimed.

Proof of Corollary 4

The first part of the proof builds on a simple application of Theorem 3 and the definition of effective smoothness constant M_f^*. The second part of the proof utilizes a relation between the backtracking step size and the effective smoothness constant. For sake of completeness, we first describe the gradient descent backtracking algorithm.

Algorithm 5 Gradient descent with backtracking

1: Given an initial point $\mathbf{x}_0 \in \text{int}(\Omega)$ and parameter $\beta \in (0, 1)$:
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Choose the smallest nonnegative integer i_k such that the step size $hk := \beta^{i_k}$ satisfies:

$$f\Big(\mathbf{x}_k - hk\nabla f(\mathbf{x}_k)\Big) \leq f(\mathbf{x}_k) - \frac{hk}{2}\|\nabla f(\mathbf{x}_k)\|^2. \tag{8.39}$$

4: Update $x^{k+1} = x^k - hk\nabla f(\mathbf{x}_k)$.
5: **end for**

Establishing the bound in (8.5a): For any step size α in the interval $\Big(0, \frac{1}{M_{f^*}}\Big)$, the definition of the effective smoothness constant M_{f^*} ensures the following property. There exists a M_g-smooth function g and a convex-differentiable function h with $f = g - h$, and the scalar M_g satisfies $\alpha < \frac{1}{M_g} \leq \frac{1}{M_{f^*}}$. Since the function f is differentiable, applying Algorithm 2 on the function f with the decomposition $f = g - h$ is equivalent to applying gradient descent on f. Furthermore, the step size α satisfies the upper bound $\alpha \leq \frac{1}{M_g}$, and applying the bound (8.3) from Theorem 3 yields:

$$\text{Avg}\left(\|\nabla f(\mathbf{x}_k)\|_2^2\right) \leq \frac{2\Big(f(\mathbf{x}_0) - f^*\Big)}{\alpha(k+1)}. \tag{8.40}$$

Establishing the backtracking bound (8.5b)**:** For any fraction $\beta \in (0,1)$, the definition of the effective smoothness constant M_{f^*} guarantees the following. There exists a M_g-smooth function g and a convex and differentiable function h with $f = g - h$, and the scalar M_g satisfies $\beta M_g \leq M_{f^*} \leq M_g$. Comparing the descent step (8.36) from Lemma 15 and step (8.39) in Algorithm 5, we conclude that the step size hk satisfies the lower bound $hk \geq \min\left\{1, \frac{\beta}{M_g}\right\} \geq \min\left\{1, \frac{\beta^2}{M_f^*}\right\}$. Applying the descent step (8.39) in Algorithm 5 repeatedly and then utilizing the last lower bound on step size hk, we find that for all $k = 0, 1, 2 \ldots$

$$f(\mathbf{x}_0) - f(\mathbf{x}_{k+1}) \geq \sum_{i=0}^{k} \frac{hk}{2} \|\nabla f(\mathbf{x}_k)\|^2 \geq \min\left\{\frac{1}{2}, \frac{\beta^2}{2M_f^*}\right\} \sum_{i=0}^{k} \|\nabla f(\mathbf{x}_k)\|^2.$$

Rearranging the last inequality yields:

$$\begin{aligned} \text{Avg}\left(\|\nabla f(\mathbf{x}_k)\|^2\right) &\leq \frac{2\max\left\{1, \frac{M_f^*}{\beta^2}\right\}\left(f(\mathbf{x}_0) - f(\mathbf{x}_{k+1})\right)}{(k+1)} \\ &\overset{(i)}{\leq} \frac{2\max\left\{1, M_f^*\right\}\left(f(\mathbf{x}_0) - f^*\right)}{\beta^2(k+1)}, \end{aligned} \tag{8.41}$$

where step (i) follows since $\beta \in (0,1)$, along with the lower bound $f(\mathbf{x}_{k+1}) \geq f^*$.

Proof of Corollary 6

Based on Theorem 4 of [Lee+16], it suffices to show that the gradient map $G(x) := x - \alpha\nabla f(x)$ is a diffeomorphism for any step size $\alpha \in \left(0, \frac{1}{M_g}\right)$. Recall that a map $G : \mathbb{R}^d \mapsto \mathbb{R}^d$ is a diffeomorphism if the map G is a bijection, and both the maps G and G^{-1} are continuously differentiable.

Injectivity: We first prove that G is an injective map. Consider a pair of vectors x, y such that $G(x) = G(y)$; our aim is to prove that $x = y$. The condition $G(x) = G(y)$ is equivalent to $x - y = \alpha\Big(\nabla f(x) - \nabla f(y)\Big)$, and we have that

$$\begin{aligned} \|x - y\|_2^2 &= \alpha\langle x - y, \nabla f(x) - \nabla f(y)\rangle \\ &= \alpha\langle x - y, \nabla g(x) - \nabla g(y)\rangle - \alpha\langle x - y, \nabla h(x) - \nabla h(y)\rangle \\ &\overset{(i)}{\leq} \alpha M_g \|x - y\|_2^2 - \alpha\langle x - y, \nabla h(x) - \nabla h(y)\rangle \\ &\overset{(ii)}{\leq} \alpha M_g \|x - y\|_2^2. \end{aligned}$$

Here inequality (i) follows because the gradient ∇g is M_g-Lipschitz by assumption; inequality (ii) follows from the convexity of the function h, which implies the monotonicity of the gradient ∇h. Finally, since the step size $\alpha < \frac{1}{M_g}$ by assumption, the inequality $\|x - y\|_2^2 \leq \alpha M_g \|x - y\|_2^2$ can hold only when $x = y$.

Surjectivity: For any fixed vector $y \in \mathbb{R}^d$, consider the following problem

$$\arg\min_{x \in \mathbb{R}^d} \Big\{ \frac{1}{2}\|x - y\|_2^2 - \alpha g(x) + \alpha h(x) \Big\}. \tag{8.42}$$

Observe that for any step size $\alpha \in \left(0, \frac{1}{M_g}\right)$ and any fixed vector $y \in \mathbb{R}^d$, the map $x \mapsto \frac{1}{2}\|x-y\|_2^2 - \alpha g(x)$ is strongly convex, whence the map $x \mapsto \frac{1}{2}\|x - y\|_2^2 - \alpha g(x) + \alpha h(x)$ is also strongly convex. Consequently, the convex problem (8.42) has a unique minimizer, and we denote it by x_y. In order to prove surjectivity of the map G, it suffices to show the point x_y is mapped to the point y. Recalling the KKT conditions of the problem (8.42), we have that

$$y = x_y - \alpha \nabla f(x_y) = G(x_y),$$

which completes the proof of surjectivity of the map G.

Combining the injectivity and the surjectivity of the map G, we conclude that the inverse map G^{-1} exists. Next, let $DG(\cdot)$ denote the Jacobian of the map G, then $DG(x) = \mathrm{I} - \alpha \nabla^2 g(x) + \alpha \nabla^2 h(x)$. Since the function g is M_g-smooth, and the map G is continuously differentiable, standard application of the inverse-function theorem guarantees that for all step size $\alpha < \frac{1}{M_g}$, the inverse map G^{-1} is continuously differentiable. Putting together the pieces, we conclude that map G^{-1} exists, and both the maps (G, G^{-1}) are continuously differentiable. Overall, we have established that the map G is a diffeomorphism, as claimed.

Proof of Proposition 3

The CCCP update at step $(k+1)$ is given by $\mathbf{x}_{k+1} = \arg\min_{x \in \Omega} q(x, \mathbf{x}_k)$, where

$$q(x, \mathbf{x}_k) := g(x) - h(\mathbf{x}_k) - \langle \nabla h(\mathbf{x}_k),\, x - \mathbf{x}_k \rangle. \tag{8.43}$$

Observe that step $(k+1)$ of Algorithm 2 is equivalent to a gradient descent update with step size α on the map $x \mapsto q(x, \mathbf{x}_k)$. Accordingly, if we define $y^{k+1} = \mathbf{x}_k - \alpha \nabla q(x, \mathbf{x}_k)$, then we have $q(y^{k+1}, \mathbf{x}_k) \geq q(\mathbf{x}_{k+1}, \mathbf{x}_k)$; moreover

$$\begin{aligned} f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) &\overset{(i)}{\geq} q(\mathbf{x}_k, \mathbf{x}_k) - q(\mathbf{x}_{k+1}, \mathbf{x}_k) \\ &\overset{(ii)}{\geq} q(\mathbf{x}_k, \mathbf{x}_k) - q(y^{k+1}, \mathbf{x}_k) \\ &\overset{(iii)}{\geq} \frac{1}{2M_g} \|\nabla f(\mathbf{x}_k)\|_2^2. \end{aligned} \tag{8.44}$$

Here inequality (i) follows from the equality $q(\mathbf{x}_k, \mathbf{x}_k) = f(\mathbf{x}_k)$ combined with the lower bound $q(x, \mathbf{x}_k) \geq f(x)$. Inequality (ii) follows since $q(y^{k+1}, \mathbf{x}_k) \geq q(\mathbf{x}_{k+1}, \mathbf{x}_k)$,

and inequality (iii) follows from Lemma 15 with step size $\alpha = \frac{1}{M_g}$. Note that equation (8.44) guarantees that the function value sequence $\{f(\mathbf{x}_k)\}_{k\geq 0}$ is decreasing. Since the function f is bounded below, we have that the sequence $\{f(\mathbf{x}_k)\}_{k\geq 0}$ converges. In order to prove that all limit points of the sequence $\left\{\mathbf{x}_k\right\}_{k\geq 0}$ are critical points, we follow the corresponding argument in proof of Theorem 3. This completes the proof of part (a) in Proposition 3.

Turning to part (b), unwrapping the recursive lower bound (8.44) and re-arranging yields inequality (8.11a). Finally, we turn to the proof of inequality (8.11b) under the additional strong convexity condition. Under this condition, the map $x \mapsto q(x, \mathbf{x}_k)$ in equation (8.43) is μ-strongly convex, so that

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq q(\mathbf{x}_k, \mathbf{x}_k) - q(\mathbf{x}_{k+1}, \mathbf{x}_k) \stackrel{(i)}{\geq} \frac{\mu}{2}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^2, \tag{8.45}$$

where inequality (i) follows from the strong convexity of the map $x \mapsto q(x, \mathbf{x}_k)$ and the fact that $\nabla q(\mathbf{x}_{k+1}, \mathbf{x}_k) = 0$. Using this equation repeatedly, we find that

$$\begin{aligned} f(\mathbf{x}_0) - f^* \geq f(\mathbf{x}_0) - f(\mathbf{x}_{k+1}) &= \sum_{j=0}^{k} \left\{f(\mathbf{x}_j) - f(\mathbf{x}_{j+1})\right\} \\ &\geq \frac{\mu}{2}\sum_{j=0}^{k} \|\mathbf{x}_j - \mathbf{x}_{j+1}\|_2^2 \\ &= \frac{\mu(k+1)}{2}\text{Avg}\left(\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^2\right). \end{aligned}$$

Rearranging the last inequality yields the bound (8.11b). Finally, let us reiterate that bounds similar to (8.11b) are known in the literature; see the paper [LS09] for example. We provide the proof of bound (8.11b) for completeness.

8.8 Proof of Theorem 4

This proof shares some important steps with Theorem 3, but it requires a more refined argument due to the presence of a non-smooth and non-continuous function φ. We start by stating an auxiliary lemma that underlies the proof of Theorem 4. In the proof, the subgradients of the convex functions h and φ at a point $\mathbf{x}_k$ are denoted by u^k and v^k, respectively.

Lemma 16. *Under the conditions of Theorem 4, we have*

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha(\nabla g(\mathbf{x}_k) + v^{k+1} - u^k), \qquad \textit{and} \tag{8.46a}$$

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2\alpha}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^2, \tag{8.46b}$$

valid for all $k = 0, 1, 2, \ldots$. *Furthermore, for any convergent subsequence* $\left\{\mathbf{x}_{k_j}\right\}_{j\geq 0}$ *of the sequence* $\left\{\mathbf{x}_k\right\}_{k\geq 0}$ *with* $\mathbf{x}_{k_j} \to \bar{x}$, *we have*

$$\lim_{j\to\infty} \varphi(\mathbf{x}_{k_j+1}) = \varphi(\bar{x}).$$

See Section 8.8 for the proof of this lemma.

We now prove Theorem 4 using Lemma 16.

Convergence of function value: We first prove that the sequence of function values $\{f(\mathbf{x}_k)\}_{k\geq 0}$ is convergent. Since $f^* := \min_{x\in\mathbb{R}^d} f(x)$ is finite by assumption, the sequence $\{f(\mathbf{x}_k)\}_{k\geq 0}$ is bounded below. If $\mathbf{x}_k = \mathbf{x}_{k+1}$ for some k, the convergence of the sequence $\left\{f(\mathbf{x}_k)\right\}_{k\geq 0}$ is trivial. Hence, we may assume without loss of generality that $\mathbf{x}_k \neq \mathbf{x}_{k+1}$ for all $k = 0, 1, 2, ...$ In that case, inequality (8.46b) ensures that $f(\mathbf{x}_k) > f(\mathbf{x}_{k+1})$, and consequently, there must exist some scalar $\bar{f}$ such that $\lim_{k\to\infty} f(\mathbf{x}_k) = \bar{f}$.

Stationarity of limit points: Next, we establish that any limit point of the sequence $\left\{\mathbf{x}_k\right\}_{k\geq 0}$ must be stationary. Consider a subsequence $\left\{\mathbf{x}_{k_j}\right\}_{j\geq 0}$ such that $\mathbf{x}_{k_j} \to \bar{x}$. Let $\left\{v^{k_j}\right\}_{j\geq 0}$ and $\left\{u^{k_j}\right\}_{j\geq 0}$ be the associated sequence of subgradients. It suffices to exhibit subgradients $\bar{v} \in \partial\varphi(\bar{x})$ and $\bar{u} \in \partial h(\bar{x})$ such that, $\nabla g(\bar{x}) + \bar{v} - \bar{u} = 0$.

Step 1: *Existence of subgradient* $\bar{u}$: Since the sequence $\left\{\mathbf{x}_{k_j}\right\}_{j\geq 0}$ is convergent, we may assume that the sequence $\left\{\mathbf{x}_{k_j}\right\}_{j\geq 0}$ is bounded, and it lies in a compact set S. The function h is convex continuous, and we have that $h(\mathbf{x}_{k_j}) \to h(\bar{x})$, and the subgradient sequence $\left\{u^{k_j}\right\}_{j\geq 0}$ is bounded; see example 9.14 in the book [RW09]. Passing to a subsequence if necessary, we may assume that the sequence $\left\{u^{k_j}\right\}_{j\geq 0}$ converges to $\bar{u}$. Putting together these pieces, we conclude that $(\mathbf{x}_{k_j}, u^{k_j}, h(\mathbf{x}_{k_j})) \to (\bar{x}, \bar{u}, h(\bar{x}))$ as $j \to \infty$; consequently, the graph continuity of limiting sub-differentials guarantees that $\bar{u} \in \partial h(\bar{x})$ (see Section 8.6 for graph continuity).

Step 2: *Existence of subgradient* $\bar{v}$: In order to complete the proof, it suffices to show that the vector $\bar{v} := -\nabla g(\bar{x}) + \bar{u}$ belongs to the subgradient set $\partial\varphi(\bar{x})$. Since the norm of successive difference $\|\mathbf{x}_{k_j} - \mathbf{x}_{k_j+1}\|_2$ converges to zero, Lemma 16 yields $\|\nabla g(\mathbf{x}_{k_j}) + v^{k_j+1} - u^{k_j}\|_2 \to 0$, and $\mathbf{x}_{k_j+1} \to \bar{x}$. Furthermore, continuity of the gradient ∇g yields $\nabla g(\mathbf{x}_{k_j}) \to \nabla g(\bar{x})$, and step 1 above guarantees $u^{k_j} \to \bar{u}$. Combining these two facts with $\|\nabla g(\mathbf{x}_{k_j}) + v^{k_j+1} - u^{k_j}\|_2 \to 0$, we obtain $v^{k_j+1} \to \bar{v} := -\nabla g(\bar{x}) + \bar{u}$, and by Lemma 16, we have $\varphi(\mathbf{x}_{k_j+1}) \to \varphi(\bar{x})$. Putting together the pieces, we conclude that $(\mathbf{x}_{k_j+1}, v^{k_j+1}, \varphi(\mathbf{x}_{k_j+1})) \to (\bar{x}, \bar{v}, \varphi(\bar{x}))$. Consequently, the graph continuity of limiting subdifferentials guarantees that $\bar{v} \in \partial\varphi(\bar{x})$ (see Section 8.6 for graph

continuity).

Finally, the subgradients $\bar{u} \in \partial h(\bar{x})$ and $\bar{v} \in \partial \varphi(\bar{x})$ obtained from steps 1 and and 2 respectively satisfy the relation $\nabla g(\bar{x}) + \bar{v} - \bar{u} = 0$, which establishes the claimed stationarity of $\bar{x}$.

Establishing the bound (8.13a)**:** Next, we establish the claimed bound (8.13a) on the averaged squared successive difference. Recalling that $f^* := \min_{x \in \mathbb{R}^d} f(x)$ is finite, we have

$$\begin{aligned} f(\mathbf{x}_0) - f^* \geq f(\mathbf{x}_0) - f(x^{k+1}) &= \sum_{j=0}^{k} f(x^j) - f(x^{j+1}) \\ &\overset{(i)}{\geq} \frac{1}{2\alpha} \sum_{j=0}^{k} \|\mathbf{x}_j - \mathbf{x}_{j+1}\|_2^2 \\ &= \frac{(k+1)}{2\alpha} \text{Avg} \left(\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^2 \right), \end{aligned} \tag{8.47}$$

where step (i) follows from equation (8.46b). Rearranging the last inequality yields the claimed bound (8.13a) on the averaged squared successive difference.

Establishing the bound (8.13b)**:** In order to establish the bound (8.13b) on the averaged squared gradient, we start by establishing the following upper bound on the gradient-norm $\|\nabla f(\mathbf{x}_{k+1})\|_2$:

$$\|\nabla f(\mathbf{x}_{k+1})\|_2 \leq \Big(M_g + M_h + \frac{1}{\alpha} \Big) \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2. \tag{8.48}$$

Recall that the function h is M_h smooth by assumption, and we have

$$\begin{aligned} \|\nabla g(\mathbf{x}_{k+1}) - \nabla h(\mathbf{x}_{k+1}) + v^{k+1}\|_2 &\overset{(i)}{=} \|\nabla g(\mathbf{x}_{k+1}) - \nabla h(\mathbf{x}_{k+1}) \\ &\qquad + \Big(\nabla h(\mathbf{x}_k) - \nabla g(\mathbf{x}_k) + \frac{1}{\alpha} \big(\mathbf{x}_k - \mathbf{x}_{k+1} \big) \Big) \|_2 \\ &\overset{(ii)}{\leq} \|\nabla g(\mathbf{x}_k) - \nabla g(\mathbf{x}_{k+1})\|_2 \\ &\qquad + \|\nabla h(\mathbf{x}_k) - \nabla h(\mathbf{x}_{k+1})\|_2 + \frac{1}{\alpha} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 \\ &\overset{(iii)}{\leq} \Big(M_g + M_h + \frac{1}{\alpha} \Big) \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2. \end{aligned}$$

Here step (i) follows from the update equation of $\mathbf{x}_{k+1}$ in Lemma 16 and from differentiability of the function g; step (ii) follows from triangle inequality, and step (iii) follows from the smoothness of the functions g and h. Putting together the bounds (8.48) and (8.47), we obtain the desired bound (8.13b).

Proof of Lemma 16

Here we prove the claims of Lemma 16.

Establishing update equation (8.46a): Recalling the convex majorant defined in equation (8.38), we define a convex majorant $q(\cdot, \mathbf{x}_k)$ of the function f as follows:

$$q(x, \mathbf{x}_k) = g(\mathbf{x}_k) - h(\mathbf{x}_k) + \langle \nabla g(\mathbf{x}_k) - u^k,\, x - \mathbf{x}_k \rangle + \frac{1}{2\alpha}\|x - \mathbf{x}_k\|_2^2 + \varphi(x), \quad (8.49)$$

where subgradient $u^k \in \partial h(\mathbf{x}_k)$, and the step size α satisfies $0 < \alpha \leq \frac{1}{M_g}$. Observe that minimizer of the convex function $x \mapsto q(x, \mathbf{x}_k)$ over $x \in \mathbb{R}^d$ is same as $\operatorname{prox}_{1/\alpha}^{\varphi}\Big(\mathbf{x}_k - \alpha(\nabla g(\mathbf{x}_k) - u^k)\Big)$, which implies that $\mathbf{x}_{k+1}$ is a minimizer of the convex function $x \mapsto q(x, \mathbf{x}_k)$ over $x \in \mathbb{R}^d$. Consequently, the optimality condition of $\mathbf{x}_{k+1}$ guarantees that there exists subgradient $v^{k+1} \in \partial g(\mathbf{x}_{k+1})$ satisfying the following equation:

$$\nabla g(\mathbf{x}_k) - u^k + v^{k+1} + \frac{1}{\alpha}\Big(\mathbf{x}_{k+1} - \mathbf{x}_k\Big) = 0. \quad (8.50)$$

Rewriting the above equation yields the update equation (8.46a).

Establishing the descent step (8.46b): Note that

$$\begin{aligned} f(\mathbf{x}_k) - q(\mathbf{x}_{k+1}, \mathbf{x}_k) &\overset{(i)}{\geq} g(\mathbf{x}_k) - h(\mathbf{x}_k) + \varphi(\mathbf{x}_{k+1}) + \langle v^{k+1},\, \mathbf{x}_k - \mathbf{x}_{k+1} \rangle - q(\mathbf{x}_{k+1}, \mathbf{x}_k) \\ &\overset{(ii)}{\geq} \langle \nabla g(\mathbf{x}_k) - u^k + v^{k+1},\, \mathbf{x}_k - \mathbf{x}_{k+1} \rangle - \frac{1}{2\alpha}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^2 \\ &\overset{(iii)}{\geq} \frac{1}{2\alpha}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^2. \end{aligned} \quad (8.51)$$

Here step (i) follows from the convexity of the function φ; step (ii) follows by substituting $q(\mathbf{x}_{k+1}, \mathbf{x}_k)$ from equation (8.49). In step (iii), we use the relation $\nabla g(\mathbf{x}_k) - u^k + v^{k+1} = \frac{1}{\alpha}\Big(\mathbf{x}_k - \mathbf{x}_{k+1}\Big)$, which follows from equation (8.50). Finally, recall that the function $x \mapsto q(x, \mathbf{x}_k)$ is a majorant for the function f, and we deduce that

$$\begin{aligned} f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) &\geq f(\mathbf{x}_k) - q(\mathbf{x}_{k+1}, \mathbf{x}_k) \\ &\geq \frac{1}{2\alpha}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^2. \end{aligned} \quad (8.52)$$

Limit of the sequence $\Big\{\varphi(\mathbf{x}_{k_j+1})\Big\}_{j\geq 0}$: Consider any convergent subsequence $\Big\{\mathbf{x}_{k_j}\Big\}_{j\geq 0}$ of the sequence $\Big\{\mathbf{x}_k\Big\}_{k\geq 0}$ with $\mathbf{x}_{k_j} \to \bar{x}$. Recall that $f^* = \inf_{x\in\mathbb{R}^d} f(x)$ is finite by assumption; combining this with step (8.46b) in Lemma 16, we have that

$\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 \to 0$, and that $\mathbf{x}_{k_j+1} \to \bar{x}$. The function φ is lower semi-continuous, and we have

$$\liminf_{j\to\infty} \ \varphi(\mathbf{x}_{k_j+1}) \geq \varphi(\bar{x}). \tag{8.53}$$

Since we already proved $\mathbf{x}_{k_j+1}$ is a minimizer of the convex function $x \mapsto q(x, \mathbf{x}_{k_j})$, we have $q(\mathbf{x}_{k_j+1}, \mathbf{x}_{k_j}) \leq q(\bar{x}, \mathbf{x}_{k_j})$. Unwrapping the last inequality and taking lim sup yields

$$\begin{aligned}\limsup_{j\to\infty} \ \varphi(\mathbf{x}_{k_j+1}) &\overset{(i)}{\leq} \varphi(\bar{x}) + \limsup_{j\to\infty} \ \left(\langle \bar{x} - \mathbf{x}_{k_j}, \nabla g(\mathbf{x}_{k_j}) - u^{k_j}\rangle + \frac{1}{2\alpha}\|\mathbf{x}_{k_j} - \bar{x}\|_2^2\right) \\ &\overset{(ii)}{=} \varphi(\bar{x}).\end{aligned} \tag{8.54}$$

Here step (i) holds since $\|\mathbf{x}_{k_j} - \mathbf{x}_{k_j+1}\|_2 \to 0$, and the sequence $\left\{\nabla g(\mathbf{x}_{k_j})\right\} - u^{k_j}\Big\}_{j\geq 0}$ is bounded—which we prove shortly; step (ii) above follows from $\mathbf{x}_{k_j} \to \bar{x}$ and boundedness of the sequence $\left\{\nabla g(\mathbf{x}_{k_j}) - u^{k_j}\right\}_{j\geq 0}$. Combining equations (8.53) and (8.54) we obtain the claimed result.

Boundedness of the sequence $\left\{\nabla g(\mathbf{x}_{k_j}) - u^{k_j}\right\}_{j\geq 0}$: In order to prove the boundedness of the sequence $\left\{\nabla g(\mathbf{x}_{k_j}) - u^{k_j}\right\}_{j\geq 0}$, it suffices to show that the gradient sequence $\left\{\nabla g(\mathbf{x}_{k_j})\right\}_{j\geq 0}$ and the sub-gradient sequence $\left\{u^{k_j}\right\}_{j\geq 0}$ are bounded. Recall that $\mathbf{x}_{k_j} \to \bar{x}$, and we have that the sequence $\{\mathbf{x}_{k_j}\}_{j\geq 0}$ is bounded. Consequently, from the smoothness of the function g, we find that the sequence $\left\{\nabla g(\mathbf{x}_{k_j})\right\}_{j\geq 0}$ is bounded. Finally, note that the function h is convex continuous, and we already argued that the sequence $\{\mathbf{x}_{k_j}\}_{j\geq 0}$ is bounded. Combining this with example 9.14 in the book [RW09], we conclude that the subgradient sequence $\left\{u^{k_j}\right\}_{j\geq 0}$ bounded.

8.9 Proofs related to Algorithm 4

In this section, we provide the proof of Theorem 5, which applies to the Frank-Wolfe based method (Algorithm 4). We also provide an upper bound on the generalized curvature constant $\mathcal{C}_f$, which is stated in Lemma 17.

Proof of Theorem 5

Let $\mathbf{x}_\gamma := \mathbf{x}_k + \gamma d^k$, where the difference d^k is defined as $d^k := s^k - \mathbf{x}_k$, and the vector s^k is the Frank-Wolfe direction defined in Algorithm 4. Unpacking the definition (8.18) of the generalized curvature constant $\mathcal{C}_f$, we find that for any scalar $\gamma \in (0, 1)$ and

subgradient $u^k \in \partial h(\mathbf{x}_k)$, we have the following:

$$\begin{aligned} f(\mathbf{x}_\gamma) &\leq f(\mathbf{x}_k) + \gamma\langle \nabla g(\mathbf{x}_k) - u^k,\, d^k\rangle + \frac{\gamma^2}{2}\mathcal{C}_f \\ &\overset{(i)}{\leq} f(\mathbf{x}_k) - \gamma g^k + \frac{\gamma^2}{2}C_0. \end{aligned} \tag{8.55}$$

Here inequality (i) is obtained by substituting $g^k = \langle d^k,\, u^k - \nabla g(\mathbf{x}_k)\rangle$ and using $C_0 \geq \mathcal{C}_f$. Substituting $\gamma = \gamma^k := \min\left\{\frac{g^k}{C_0}, 1\right\}$ in equation (8.55) yields

$$f(x^{k+1}) \leq f(x^k) - \min\left\{\frac{(g^k)^2}{2C_0}, g^k - \frac{C_0}{2}\mathbb{1}_{\left\{g^k > C_0\right\}}\right\}, \tag{8.56}$$

where $\mathbf{x}_{k+1} = \mathbf{x}_k + \gamma^k d^k$. Let $\bar{g}^k := \min_{0\leq j\leq k} g^j$ denote the minimum FW gap up to iteration k, then repeated application of equation (8.56) yields

$$\begin{aligned} f(\mathbf{x}_0) - f(\mathbf{x}_{k+1}) &\geq \sum_{j=0}^{k} \min\left\{\frac{(g^j)^2}{2C_0}, g^j - \frac{C_0}{2}\mathbb{1}_{\left\{g^j > C_0\right\}}\right\} \\ &\geq (k+1)\min\left\{\frac{(\bar{g}^k)^2}{2C_0}, \bar{g}^k - \frac{C_0}{2}\mathbb{1}_{\left\{\bar{g}^k > C_0\right\}}\right\}. \end{aligned} \tag{8.57}$$

Rewriting the last equation yields the following upper bound

$$\min\left\{\frac{(\bar{g}^k)^2}{2C_0}, \bar{g}^k - \frac{C_0}{2}\mathbb{1}_{\left\{\bar{g}^k > C_0\right\}}\right\} \overset{(i)}{\leq} \frac{f(\mathbf{x}_0) - f^*}{k+1},$$

where step (i) follows from the lower bound $f(\mathbf{x}_{k+1}) \geq f^* := \min_{x\in\Omega}\ f(x)$. Considering the cases where $\bar{g}^k \leq C_0$ and $\bar{g}^k > C_0$ separately, it can be shown following [Lac16] that

$$\bar{g}^k \leq \begin{cases} \frac{2(f(\mathbf{x}_0)-f^*)}{\sqrt{k+1}} & \text{for } \ k+1 \leq \frac{2(f(\mathbf{x}_0)-f^*)}{C_0} \\ \sqrt{\frac{2C(f(\mathbf{x}_0)-f^*)}{k+1}} & \text{otherwise .} \end{cases}$$

Finally, note that $\sqrt{2C_0(f(\mathbf{x}_0) - f^*)} \leq \max\{2(f(\mathbf{x}_0) - f^*), C_0\}$ and we conclude that

$$\bar{g}^k \leq \frac{\max\left\{2\Big(f(\mathbf{x}_0) - f^*\Big), C_0\right\}}{\sqrt{k+1}}.$$

Upper bound on generalized curvature constant

In this section, we provide an upper bound on the generalized curvature constant $\mathcal{C}_f$, where the function f is a difference of a differentiable function g and a continuous function h. For better readability, we use $\mathcal{C}_{g-h}$ instead of $\mathcal{C}_f$ in the following lemma.

Lemma 17. *Suppose that the function g is continuously differentiable and function h is convex, then we have $\mathcal{C}_{g-h} \leq \mathcal{C}_g$. Furthermore, if the function g is M_g-smooth, and the function h is a μ strongly convex function with $0 \leq \mu < M$, then*

$$\mathcal{C}_{g-h} \leq \Big(M - \mu\Big) \times \Big(\operatorname{diam}_{\|\cdot\|_2}(\Omega)\Big)^2, \tag{8.58}$$

where $\operatorname{diam}_{\|\cdot\|_2}$ denote the diameter of the set Ω, measured in ℓ_2 norm.

Comments: The first upper bound on $\mathcal{C}_{g-h}$ in Lemma 17 posits that the curvature constant of the difference function $g - h$ is upper bounded by curvature constant of the function g, whenever the second function h is convex. Let us try to understand an implication of this result through an example. One of the well-known upper bound of curvature constant for M_g-smooth function g is $M_g \times \Big(\operatorname{diam}_{\|\cdot\|_2}(\Omega)\Big)^2$; see the paper by [Jag13]. Now consider continuously differentiable functions g and h such that the function g is M_g-smooth and the function h is non-smooth and convex. It can be verified that the difference function $g - h$ is *not* smooth in this case; consequently, the earlier bound on curvature constant $\mathcal{C}_{g-h}$ is ∞, whereas Lemma 17 ensures that

$$\mathcal{C}_{g-h} \leq \mathcal{C}_g \leq M_g \times \Big(\operatorname{diam}_{\|\cdot\|_2}(\Omega)\Big)^2.$$

Proof of the upper bound $\mathcal{C}_{g-h} \leq \mathcal{C}_g$: Unwrapping the definition of $\mathcal{C}_{g-h}$, we have

$$\begin{aligned}
\mathcal{C}_{g-h} &= \sup_{\substack{x,y\in c_\gamma \\ u\in\partial h(x)}} \frac{2}{\gamma^2}\Big[f(y) - f(x) - \langle y - x,\, \nabla g(x) - u\rangle\Big] \\
&= \sup_{\substack{x,y\in c_\gamma \\ u\in\partial h(x)}} \frac{2}{\gamma^2}\Big[g(y) - g(x) - \langle y - x,\, \nabla g(x)\rangle - \Delta_h(y,x,u)\Big] \qquad (8.59) \\
&\overset{(i)}{\leq} \underbrace{\sup_{x,y\in c_\gamma} \frac{2}{\gamma^2}\Big[f(y) - f(x) - \langle y - x,\, \nabla g(x)\rangle\Big]}_{\mathcal{C}_g},
\end{aligned}$$

where $\Delta_h(y,x,u) := h(y) - h(x) - \langle y - x,\, u\rangle$. Here inequality (i) follows by noting that, for any pair of points $x, y \in \Omega$, and for any convex function h with $u \in \partial h(x)$, we have $\Delta_h(y,x,u) \geq 0$.

Proof of upper bound (8.58): Suppose in addition, the function g is M_g-smooth, and the function h is μ-strongly convex with $\mu \geq 0$. Then we have $\Delta_h(y,x,u) \geq \frac{\mu}{2}\|x - y\|_2^2$, and equation (8.59) yields

$$\begin{aligned}
\mathcal{C}_{g-h} &\leq \sup_{x,y\in c_\gamma} \frac{2}{\gamma^2}\Big[g(y) - g(x) - \langle y - x,\, \nabla g(x)\rangle - \frac{\mu}{2}\|x - y\|_2^2\Big] \\
&\overset{(i)}{\leq} \sup_{x,y\in c_\gamma} \frac{2}{\gamma^2}\left[\frac{M_g - \mu}{2}\|x - y\|_2^2\right],
\end{aligned}$$

where step (i) follows since the function g is M_g-smooth. Substituting $y - x = \gamma s$ with $s \in \Omega$, we obtain the claimed upper bound

$$\mathcal{C}_{g-h} \leq (M_g - \mu) \times \Big(\operatorname{diam}_{\|\cdot\|_2}(\Omega) \Big)^2.$$

8.10 Proofs of faster rates under Assumption KL

In this section, we prove our results on improved convergence rates for functions which satisfy Assumption KL—as stated in Theorems 6 and 7. We begin by stating an auxiliary lemma that underlies the proofs of Theorems 6 and 7.

Lemma 18. *Under assumptions of either Theorem 6 or Theorem 7, there exists constants $\theta \in [0, 1)$, $C > 0$ and positive integer k_1 such that for all $k \geq k_1$, we have*

$$|f(\mathbf{x}_k) - \bar{f}|^\theta \leq C \|\nabla f(\mathbf{x}_k)\|_2,$$

where $f(\mathbf{x}_k) \downarrow \bar{f}$. Furthermore, if $\mathbf{x}_k \to \bar{x}$, then the parameters (θ, C), obtained from KL-inequality of the function f at the point $\bar{x}$, satisfy the above inequality.

See Section 8.10 for the proof of this lemma.

Proof of Theorem 6

Now we prove Theorem 6 using Lemma 18.

Convergence of the sequence $\{\mathbf{x}_k\}_{k \geq 0}$: We demonstrate the convergence of the sequence $\{\mathbf{x}_k\}_{k \geq 0}$ by proving that the sequence has finite length property; more precisely, we show that $\sum_{k=0}^{\infty} \|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 < \infty$. First, note that for any scalar $0 \leq \theta < 1$, the function $t \mapsto t^{1-\gamma\theta}$ is concave for $0 < \gamma < \frac{1}{\theta}$; consequently, for iteration $k \geq k_1$ we have

$$\begin{aligned}
\Big(f(\mathbf{x}_k) - \bar{f}\Big)^{1-\gamma\theta} - \Big(f(\mathbf{x}_{k+1}) - \bar{f}\Big)^{1-\gamma\theta} &\geq \Big(1 - \gamma\theta\Big)\Big(f(\mathbf{x}_k) - \bar{f}\Big)^{-\gamma\theta}\Big(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})\Big) \\
&\overset{(i)}{\geq} \Big(1 - \gamma\theta\Big)\Big(|f(\mathbf{x}_k) - \bar{f}|\Big)^{-\gamma\theta} \times \frac{1}{2\alpha}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^2 \\
&\overset{(ii)}{\geq} \frac{(1 - \gamma\theta)}{C\|\nabla f(\mathbf{x}_k)\|_2^\gamma} \times \frac{1}{2\alpha}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^2 \\
&\overset{(iii)}{=} \frac{(1 - \gamma\theta)}{2C\alpha^{1-\gamma}}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^{2-\gamma}. \qquad (8.60)
\end{aligned}$$

Here inequality (i) follows from the descent property in equation (8.36) and from the fact that $f(\mathbf{x}_k) \downarrow \bar{f}$. Inequality (ii) follows from Lemma 18, and equality (iii) follows

from the relation $\mathbf{x}_k - \mathbf{x}_{k+1} = \alpha\Big(\nabla g(\mathbf{x}_k) - u^k\Big) = \alpha\nabla f(\mathbf{x}_k)$. Substituting $\gamma = 1$ and summing both side of inequality (8.60) from index $k = k_1$ to $k = \infty$, we obtain

$$\begin{aligned}\Big(f\Big(\mathbf{x}_{k_1}\Big) - \bar{f}\Big)^{1-\theta} &= \sum_{k=k_1}^{\infty} \Big(f\Big(\mathbf{x}_k\Big) - \bar{f}\Big)^{1-\theta} - \Big(f\Big(\mathbf{x}_{k+1}\Big) - \bar{f}\Big)^{1-\theta} \\ &\geq \sum_{k=k_1}^{\infty} \frac{(1-\theta)}{2C}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2,\end{aligned}$$

which proves the finite length property of the sequence $\Big\{\mathbf{x}_k\Big\}_{k\geq 0}$. Consequently, we are guaranteed to have a vector $\bar{x}$ such that $\mathbf{x}_k \to \bar{x}$ as $k \to \infty$.

Rate of convergence of Avg $(\|\nabla f(\mathbf{x}_k)\|_2)$: Rewriting equation (8.60), we have the following:

$$\begin{aligned}C_\gamma &:= \sum_{\ell=0}^{k_1} \frac{(1-\gamma\theta)}{2C\alpha^{1-\gamma}}\|\mathbf{x}_\ell - \mathbf{x}_{\ell+1}\|_2^{2-\gamma} + \Big(f\Big(\mathbf{x}_{k_1}\Big) - \bar{f}\Big)^{(1-\gamma\theta)} \\ &\overset{(i)}{\geq} \sum_{\ell=0}^{k-1} \frac{(1-\gamma\theta)}{2C\alpha^{1-\gamma}}\|\mathbf{x}_\ell - \mathbf{x}_{\ell+1}\|_2^{2-\gamma} \\ &= \frac{k(1-\gamma\theta)}{2C\alpha^{1-\gamma}}\text{Avg}\left(\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^{2-\gamma}\right), \end{aligned} \tag{8.61}$$

where step (i) above follows from equation (8.60), and $\text{Avg}\left(\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^{2-\gamma}\right) := \frac{1}{k}\sum_{\ell=0}^{k-1}\|\mathbf{x}_\ell - \mathbf{x}_{\ell+1}\|_2^{2-\gamma}$ denote the running arithmetic average. Since $0 \leq \theta < 1$, we can take $\gamma = 1$ in equation (8.61), and we obtain the following rate:

$$\text{Avg}\left(\|\nabla f(\mathbf{x}_k)\|_2\right) = \frac{1}{\alpha}\text{Avg}\left(\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2\right) \leq \frac{c_1}{k},$$

where $c_1 = \frac{2CC_\gamma}{\alpha(1-\theta)}$. Finally, note that the last equality holds trivially for iteration $k \leq k_1$ with the given choice of the constant c_1.

Rate of convergence of GAvg $(\|\nabla f(\mathbf{x}_k)\|_2)$: Since we proved that the sequence $\Big\{\mathbf{x}_k\Big\}_{k\geq 0}$ is convergent to the point $\bar{x}$, we have that the parameter θ in Lemma 18 can be taken to be the KL-exponent of the function f at point $\bar{x}$. Suppose $\frac{1}{2} \leq \theta < \frac{r}{2r-1}$, then substituting $\gamma = \frac{2r-1}{r}$ in equation (8.61) yields,

$$\begin{aligned}\text{GAvg}\left(\|\nabla f(\mathbf{x}_k)\|_2\right) &= \frac{1}{\alpha}\text{GAvg}\left(\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2\right) \\ &\overset{(i)}{\leq} \frac{1}{\alpha}\Big\{\text{Avg}\left(\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^{\frac{1}{r}}\right)\Big\}^r \\ &\overset{(ii)}{\leq} \frac{c_2}{k^r},\end{aligned}$$

where $c_2 = \frac{1}{\alpha}\Big(\frac{2CC_\gamma \alpha^{1-\gamma\theta}}{1-\gamma\theta}\Big)^r$ with $\gamma = \frac{2r-1}{r}$, and $\text{GAvg}\left(\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^{2-\gamma}\right) := \prod_{\ell=0}^{k-1}\left(\|\mathbf{x}_\ell - \mathbf{x}_{\ell+1}\|_2\right)^{\frac{1}{k}}$, the geometric average of the sequence $\left\{\|\mathbf{x}_\ell - \mathbf{x}_{\ell+1}\|_2\right\}_{l=0}^{k-1}$. Here step (i) above follows from arithmetic-geometric mean (AM/GM) inequality; step (ii) follows from the bound in equation (8.61) and from the fact that $\gamma = \frac{2r-1}{r}$. Finally, note that the last equality holds trivially for iteration $k \leq k_1$ with the given choice of constant c_2.

Proof of Theorem 7

The proof of Theorem 7 builds on the techniques used in the proof of Theorem 6 but requires additional technical care due to the presence of possibly non-continuous function φ.

Convergence of the sequence $\left\{\mathbf{x}_k\right\}_{k\geq 0}$: The proof of Theorem 7 has two steps. First, we prove a descent condition similar to equation (8.60). We then leverage this descent condition and weighted AM-GM inequality to obtain the desired result.

Step 1: Following the proof of Theorem 6, we prove the convergence of the sequence $\left\{\mathbf{x}_k\right\}_{k\geq 0}$ by showing that the sequence $\left\{\mathbf{x}_k\right\}_{k\geq 0}$ has finite length property. First, note that for scalars $0 \leq \theta < 1$ and $0 < \gamma < \frac{1}{\theta}$, the function $t \mapsto t^{1-\gamma\theta}$ is concave. Consequently, for iteration $k \geq k_1$, from Lemma 18 we have

$$\begin{aligned}\Big(f(\mathbf{x}_k) - \bar{f}\Big)^{1-\gamma\theta} - \Big(f(\mathbf{x}_{k+1}) - \bar{f}\Big)^{1-\gamma\theta} &\geq \Big(1-\gamma\theta\Big)\Big(f(\mathbf{x}_k) - \bar{f}\Big)^{-\gamma\theta}\Big(f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})\Big) \\ &\overset{(i)}{\geq} \Big(1-\gamma\theta\Big)\Big(|f(\mathbf{x}_k) - \bar{f}|\Big)^{-\gamma\theta} \times \frac{1}{2\alpha}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^2 \\ &\overset{(ii)}{\geq} \frac{(1-\gamma\theta)}{C\|\nabla f(\mathbf{x}_k)\|_2^\gamma} \times \frac{1}{2\alpha}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^2. \end{aligned} \tag{8.62}$$

Here step (i) follows from the descent property in equation (8.52) and from the fact that $f(\mathbf{x}_k) \downarrow \bar{f}$; step (ii) follows from Lemma 18. The function h is locally smooth by assumption; as a result, we have that the difference function $g - h$ is locally smooth. We also assumed that the sequence $\left\{\mathbf{x}_k\right\}_{k\geq 0}$ is bounded (lies in a compact set S); consequently, we may assume that the difference function $g - h$ is smooth in the compact set S with a smoothness parameter M_{g-h}(say). Borrowing the argument of Theorem 4 part(b), it follows that:

$$\|\nabla g(\mathbf{x}_k) - \nabla h(\mathbf{x}_k) + v^k\|_2 \leq \Big(M_{g-h} + \frac{1}{\alpha}\Big)\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2. \tag{8.63}$$

Combining the last inequality with inequality (8.62) yields the following descent property

$$\Big(f(\mathbf{x}_k)-\bar{f}\Big)^{1-\gamma\theta}-\Big(f(\mathbf{x}_{k+1})-\bar{f}\Big)^{1-\gamma\theta}\geq\frac{(1-\gamma\theta)}{2\alpha C\Big(M_{g-h}+\frac{1}{\alpha}\Big)^{\gamma}}\times\frac{\|\mathbf{x}_k-\mathbf{x}_{k+1}\|_2^2}{\|\mathbf{x}_k-\mathbf{x}_{k-1}\|_2^{\gamma}}. \quad (8.64)$$

Step 2: We now leverage the descent condition obtained from step 1 to prove finite length property of the sequence $\Big\{\mathbf{x}_k\Big\}_{k\geq 0}$. In order to facilitate further discussion, we use Δ_γ^k to denote the following:

$$\Delta_\gamma^k:=C_3\bigg(\Big(f(\mathbf{x}_k)-\bar{f}\Big)^{1-\gamma\theta}-\Big(f(\mathbf{x}_{k+1})-\bar{f}\Big)^{1-\gamma\theta}\bigg),$$

where the constant $C_3:=\frac{2\alpha C\Big(M_{g-h}+\frac{1}{\alpha}\Big)^{\gamma}}{(1-\gamma\theta)}$. With this notation, we can rewrite the equation (8.64) as

$$\Delta_\gamma^k\|\mathbf{x}_{k-1}-\mathbf{x}_k\|_2^{\gamma}\geq\|\mathbf{x}_k-\mathbf{x}_{k+1}\|_2^2. \quad (8.65)$$

Combining equation (8.65) with the weighted AM-GM inequality, we obtain

$$\begin{aligned}\Big(1+\frac{\gamma}{2-\gamma}\Big)\times\sum_{j=k_1+1}^{k}\|\mathbf{x}_j-\mathbf{x}_{j+1}\|_2^{2-\gamma}&\overset{(i)}{\leq}\Big(1+\frac{\gamma}{2-\gamma}\Big)\times\sum_{k=k_1+1}^{k}\Big(\sqrt{\Delta_\gamma^j\,\|\mathbf{x}_{j-1}-\mathbf{x}_j\|_2^{\gamma}}\Big)^{\frac{2-\gamma}{2}}\\&\overset{(ii)}{\leq}\sum_{j=k_1+1}^{k}\Big(\Delta_\gamma^j+\frac{\gamma}{2-\gamma}\|\mathbf{x}_{j-1}-\mathbf{x}_j\|_2^{2-\gamma}\Big)\\&\overset{(iii)}{\leq}C_3\Big(f(\mathbf{x}_{k_1})-\bar{f}\Big)^{1-\gamma\theta}\\&\quad+\sum_{j=k_1+1}^{k}\frac{\gamma}{2-\gamma}\|\mathbf{x}_{j-1}-\mathbf{x}_j\|_2^{2-\gamma}. \quad (8.66)\end{aligned}$$

Here step (i) follows from equation (8.65), and step (ii) is implied by applying weighted AM-GM inequality as follows:

$$\frac{\Delta_\gamma^j+\frac{\gamma}{2-\gamma}\|\mathbf{x}_{j-1}-\mathbf{x}_j\|_2^{2-\gamma}}{1+\frac{\gamma}{2-\gamma}}\geq\Big(\Delta_\gamma^j\|\mathbf{x}_{j-1}-\mathbf{x}_j\|_2^{\gamma}\Big)^{\frac{1}{1+\frac{\gamma}{2-\gamma}}}.$$

Step (iii) in equation (8.66) follows from the following observation

$$\begin{aligned}\sum_{j=k_1}^{k}\Delta_\gamma^j&=C_3\sum_{j=k_1}^{k}\Big(f(\mathbf{x}_j)-\bar{f}\Big)^{1-\gamma\theta}-\Big(f(\mathbf{x}_{j+1})-\bar{f}\Big)^{1-\gamma\theta}\\&\leq C_3\Big(f(\mathbf{x}_{k_1})-\bar{f}\Big)^{1-\gamma\theta}.\end{aligned}$$

Rewriting inequality (8.66), we have for all $k \geq k_1 + 2$

$$\sum_{j=k_1+1}^{k-1} \|\mathbf{x}_j - \mathbf{x}_{j+1}\|_2^{2-\gamma} \leq C_3\Big(f(\mathbf{x}_{k_1}) - \bar{f}\Big)^{1-\gamma\theta} + \frac{\gamma}{2-\gamma}\|\mathbf{x}_{k_1} - \mathbf{x}_{k_1+1}\|_2^{2-\gamma} \tag{8.67}$$

$$- \Big(1 + \frac{\gamma}{2-\gamma}\Big)\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2^{2-\gamma}$$

$$\leq C_3\Big(f(\mathbf{x}_{k_1}) - \bar{f}\Big)^{1-\gamma\theta} + \frac{\gamma}{2-\gamma}\|\mathbf{x}_{k_1} - \mathbf{x}_{k_1+1}\|_2^{2-\gamma} < \infty. \tag{8.68}$$

Finally, by substituting $\gamma = 1$ and letting $k \to \infty$ in the last equation, we deduce the finite length property of the sequence $\Big\{\mathbf{x}_k\Big\}_{k\geq 0}$.

Rate of convergence of Avg $(\|\nabla f(\mathbf{x}_k)\|_2)$ and GAvg $(\|\nabla f(\mathbf{x}_k)\|_2)$: The proof of this part follows from the corresponding proof in Theorem 6 and using the inequality (8.67) and upper bound (8.63).

Proof of Lemma 18

Since the sequence $\Big\{\mathbf{x}_k\Big\}_{k\geq 0}$ is bounded by assumption, without loss of generality, we may assume that the set of limit points of the sequence $\Big\{\mathbf{x}_k\Big\}_{k\geq 0}$ — which we denote by $\bar{\mathcal{X}}$ — is a compact set. From Theorem 3 (respectively Theorem 4), we have that all the limit points of the sequence $\Big\{\mathbf{x}_k\Big\}_{k\geq 0}$ are critical points of the function f; furthermore, since $f(\mathbf{x}_k) \downarrow \bar{f}$, we also have that the function f is constant on the set of limit points $\bar{\mathcal{X}}$, and the function value on $\bar{\mathcal{X}}$ equals $\bar{f}$. Combining this with Assumption KL, we have for all $z \in \bar{\mathcal{X}}$, there exists constants $\theta(z) \in [0, 1)$, $r_z > 0$ and $C(z) > 0$ such that, $\mid f(x) - \bar{f} \mid^{\theta(z)} \leq C(z) \times \|\nabla f(x)\|_2$ for all $x \in B(z, r_z)$. Now, consider the open cover $\{B(z, r_z) : z \in \bar{\mathcal{X}}\}$ of the set $\bar{\mathcal{X}}$. From compactness of the set $\bar{\mathcal{X}}$, we are guaranteed to have a finite subcover; more precisely, there exists $\{z_1, \ldots z_p\} \subseteq \bar{\mathcal{X}}$ such that $\bar{\mathcal{X}} \subseteq \bigcup_{i=1}^{p} B(z_i, r_{z_i})$. Define constants $\theta := \max\{\theta(z_i) : 1 \leq i \leq p\}$, $C := \max\{C(z_i) : 1 \leq i \leq p\}$, and $r := \min\Big\{\frac{r_{z_i}}{2} : 1 \leq i \leq p\Big\}$. Utilizing the result $\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 \to 0$ from Theorem 3 (respectively Theorem 4), one can show that, there exists positive integer k_1 such that for all $k \geq k_1$ we have $\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 < \frac{r}{2}$, and $x^k \in \bigcup_{i=1}^{p} B(z_i, r_{z_i})$. Putting together these pieces, we conclude that for all $k \geq k_1$

$$x^k \in \bigcup_{i=1}^{p} B(z_i, r_{z_i}), \quad \text{and} \quad \mid f(x^k) - \bar{f} \mid^{\theta} \leq C\|\nabla f\|_2,$$

which proves the first part of claimed lemma. Now suppose the sequence $\Big\{\mathbf{x}_k\Big\}_{k\geq 0}$ converges to a point $\bar{x}$, then we have that the set of limit points $\bar{\mathcal{X}} = \{\bar{x}\}$, is a singleton set. The rest of the proof is immediate by repeating the argument so far, with the additional information that $\bar{\mathcal{X}} = \{\bar{x}\}$.

8.11 Proofs of Corollaries

In this section, we collect the proofs of Corollaries 7, 8 and 9 from Section 8.4.

Proof of Corollary 7

First, note that in order to apply Theorem 3 and Theorem 6 to Corollary 7, it is enough to show that the function $\mu \mapsto f(\mu)$ is M_f-smooth (in this example, function $h \equiv 0$, and hence $f \equiv g$), and the function f satisfies Assumption KL. We verify that Assumption KL is satisfied by proving that the objective function f in problem (8.22) is continuous sub-analytic (see Section 8.6). For proving sub-analyticity, we heavily use the properties mentioned in Section 8.6. In the following proof, we assume without loss of generality that $\lambda = 1$.

The function f is continuous sub-analytic: First, we show that the function Ψ is sub-analytic. We begin by observing that Ψ is piecewise polynomial. Polynomials are analytic functions and intervals are semi-analytic sets. Since piecewise analytic functions with semi-analytic pieces are semi-analytic (hence sub-analytic), we conclude that the function Ψ is sub-analytic. Now, the function $\mu \mapsto y_i - \langle z_i,\, \mu \rangle$ is linear, and hence continuous sub-analytic. Furthermore, since continuous sub-analytic functions are closed under composition, we have that the function $\mu \mapsto \Psi\Big(y_i - \langle z_i,\, \mu \rangle\Big)$ is sub-analytic. Finally, note that sub-analytic functions are closed under linear combination, and we conclude that the function f is sub-analytic. The continuity of the function f is immediate by inspection.

The function f is smooth: Since the vectors $\Big\{(z_i, y_i)\Big\}_{i=1}^{n}$ are fixed, it suffices to prove that the function Ψ is smooth. A straightforward calculation shows that Ψ is continuously differentiable and smooth; in particular, it has a smoothness parameter 36 when $\lambda = 1$.

Putting together the pieces, we conclude that Theorem 3 and Theorem 6 are applicable for problem (8.22). Convergence of the sequence $\Big\{\mu^k\Big\}_{k \geq 0}$ to a point $\bar{\mu}$ and the convergence rate of gradient norms follows from Theorem 6, and the stationary condition $\nabla f(\bar{\mu}) = 0$ follows from Theorem 3.

Escaping strict saddle points: Note that the functions (g, h) are twice continuously differentiable, and the function g is smooth. Consequently, from Corollary 6, it follows that with random initializations, Algorithm 2 avoids strict saddle points almost surely.

Proof of Corollary 8

We begin by providing a high-level outline of the proof. First, note that from Theorem 4, we have the successive difference $\|\mathbf{x}_k - \mathbf{x}_{k+1}\|_2 \to 0$, and as a result, the set of limit point of the sequence $\left\{\mathbf{x}_k\right\}_{k\geq 0}$—call it $\bar{\mathcal{X}}$—is a connected set [Ost16]. We prove that the connected-set $\bar{\mathcal{X}}$ is singleton by showing that the set $\bar{\mathcal{X}}$ has an isolated point — this also proves that sequence $\{\mathbf{x}_k\}_{k\geq 0}$ is convergent. Next, we show that the objective-function f, in the problem (8.28), satisfies Assumption KL with exponent $\theta = \frac{1}{2}$. Finally, we show that condition $|\bar{x}|_{(r)} > |\bar{x}|_{(r+1)} \geq 0$ implies that function $x \mapsto h(x) := \sum_{i=d-s+1}^{d} |x|_{(i)}$ is smooth in a neighborhood of point $\bar{x}$, and we use the proof techniques of Theorem 7 to establish the convergence rate of the gradient sequence. In order to obtain the rate of convergence of the sequence $\left\{\mathbf{x}_k\right\}_{k\geq 0}$, we use ideas similar to those in the paper [Lee+16].

Convergence of the sequence $\{\mathbf{x}_k\}_{k\geq 0}$: For notational convenience, let us use $g(x) := \|y - Bx\|_2^2$, $\varphi(x) := \lambda\|x\|_1$, and $h(x) := \lambda \sum_{i=d-s+1}^{d} |x|_{(i)}$. Since the point $\bar{x}$ satisfies the condition $|\bar{x}|_{(r)} > |\bar{x}|_{(r+1)} \geq 0$ by assumption, there must exist a neighborhood $B(\bar{x}, r)$ such that the function h is differentiable in the neighborhood $B(\bar{x}, r)$, and all points $x \in B(\bar{x}, r)$ satisfy $\operatorname{sign}(x_{(i)}) = \operatorname{sign}(\bar{x}_{(i)})$ for $1 \leq i \leq r$. We show that, in a neighborhood of the point $\bar{x}$, it is the only critical point, thereby proving that the point $\bar{x}$ is an isolated critical point. To this end consider the convex sub-problem mentioned in Corollary 8

$$\mathcal{P}(\bar{x}) := \min_{x\in\mathbb{R}^d} g(x) + \lambda\varphi(x) - \lambda\langle\nabla h(\bar{x}),\, x\rangle. \tag{8.69}$$

For any point x^* such that $x^* \in B(\bar{x}, r) \cap \bar{\mathcal{X}}$, from Theorem 4, we know that

$$\nabla g(\bar{x}) + \lambda\bar{u} - \lambda\nabla h(\bar{x}) = 0 \quad \text{and} \quad \nabla g(\mathbf{x}_*) + \lambda u^* - \lambda\nabla h(\mathbf{x}_*) = 0, \tag{8.70}$$

where subgradients $u^* \in \partial\varphi(\mathbf{x}_*)$ and $\bar{u} \in \partial\varphi(\bar{x})$. Next, note that from the choice of neighborhood $B(\bar{x}, r)$, it follows that for all $x \in B(\bar{x}, r)$ we have $\nabla h(\mathbf{x}) = \nabla h(\bar{x})$, and in particular, we deduce $\nabla h(\mathbf{x}_*) = \nabla h(\bar{\mathbf{x}})$. Combining this relation with equation (8.70) yields:

$$\nabla g(\bar{x}) + \lambda\bar{u} - \lambda\nabla h(\bar{x}) = 0 \quad \text{and} \quad \nabla g(\mathbf{x}_*) + \lambda u^* - \lambda\nabla h(\bar{x}) = 0,$$

which implies both the points $\mathbf{x}_*$ and $\bar{x}$ are zero sub-gradient points of convex problem (8.69); this contradicts the assumption that problem (8.69) has an unique solution. Hence, we conclude that $\mathbf{x}_* = \bar{x}$, and the point $\bar{x}$ is an isolated critical point of the sequence $\left\{\mathbf{x}_k\right\}_{k\geq 0}$, and $\bar{\mathcal{X}}$. Putting together the pieces, we conclude that $\mathbf{x}_k \to \bar{x}$.

Smoothness of function h in a neighborhood of $\bar{x}$: We already argued above that for all $x \in B(\bar{x}, r)$, the function h is differentiable and $\nabla h(\mathbf{x}) = \nabla h(\bar{x})$. Consequently, we have that in the neighborhood $B(\bar{x}, r)$, the function h is smooth with a smoothness parameter $M_h = 0$.

The function f satisfies Assumption KL with exponent $\theta = \frac{1}{2}$: Recently, in the paper [LP16] (Corollaries 5.1 and 5.2), the authors showed that if the functions $f_1, f_2, \ldots, f_T$ satisfy the KL-inequality with an exponent $\theta = \frac{1}{2}$, then the function $f := \min\Big\{f_1, f_2, \ldots, f_T\Big\}$ also satisfies KL-inequality with the exponent $\theta = \frac{1}{2}$. Interestingly, the function f can be represented as is minimum of finitely many functions as follows:

$$f(x) = \min_{a \in \mathcal{A}} \Big\{\|y - Bx\|_2^2 + \lambda \|x\|_1 - \lambda a^\top x\Big\}, \tag{8.71}$$

where $\mathcal{A} := \Big\{a \in \Big\{-1, 0, 1\Big\}^d : \sum_{i=1}^d |a_i| = r\Big\}$. Note that the set $\mathcal{A}$ has cardinality at most 3^d. It is known that functions of the form $x \mapsto \frac{1}{2} x^\top A x + P(x) + b^\top x$ satisfy the KL-inequality with exponent $\theta = \frac{1}{2}$, where P is a proper closed polyhedral function, and A is a positive semi-definite matrix; see Corollaries 5.1 and 5.2 in the paper [LP16]. Putting together these two observations, we conclude that the function f satisfies KL-assumption with KL-exponent $\theta = \frac{1}{2}$.

Combining the pieces: Since we proved $\mathbf{x}_k \to \bar{x}$, we have that for a suitable choice of k_1, the tail sequence $\Big\{\mathbf{x}_k\Big\}_{k \geq k_1}$ lies in the neighborhood $B(\bar{x}, r)$. Now, the function f satisfies Assumption KL with exponent $\theta = \frac{1}{2}$, and the function h is smooth in the neighborhood $B(\bar{x}, r)$; hence, following the argument in proof of Theorem 7 part(b), we conclude that:

$$\text{Avg}\left(\|\nabla f(\mathbf{x}_k)\|_2\right) \leq \frac{c_1}{k}.$$

Rate of convergence of sequence $\Big\{\mathbf{x}_k\Big\}_{k \geq 0}$: The KL-exponent for the function f is $\theta = \frac{1}{2}$, and we may use $\gamma = 1$ in equation (8.67) which yields

$$\sum_{\ell = k_1 + 1}^{\infty} \|\mathbf{x}_\ell - \mathbf{x}_{\ell+1}\|_2 \leq \|\mathbf{x}_{k_1} - \mathbf{x}_{k_1+1}\|_2 + C_3 \Big(f(\mathbf{x}_{k_1}) - \bar{f}\Big)^{\frac{1}{2}}, \tag{8.72}$$

for some constant C_3. From Lemma 18 and equation (8.48), we have

$$\Big(f(\mathbf{x}_{k_1}) - \bar{f}\Big)^{\frac{1}{2}} \leq C \|\nabla f(\mathbf{x}_{k_1})\|_2 \leq C(M + M_h + \frac{1}{\alpha}) \|\mathbf{x}_{k_1} - \mathbf{x}_{k_1 - 1}\|_2. \tag{8.73}$$

Combining equations (8.72) and (8.73) we have

$$\begin{aligned}\sum_{\ell=k_1}^{\infty} \|\mathbf{x}_\ell - \mathbf{x}_{\ell+1}\|_2 &\leq 2\|\mathbf{x}_{k_1} - \mathbf{x}_{k_1+1}\|_2 + C_3\Big(f(\mathbf{x}_{k_1}) - \bar{f}\Big)^{\frac{1}{2}} \\ &\overset{(i)}{\leq} 2\|\mathbf{x}_{k_1} - \mathbf{x}_{k_1+1}\|_2 + CC_3(M + M_h + \frac{1}{\alpha})\|\mathbf{x}_{k_1} - \mathbf{x}_{k_1-1}\|_2 \\ &\overset{(ii)}{\leq} \bar{C}\|\mathbf{x}_{k_1} - \mathbf{x}_{k_1-1}\|_2, \end{aligned} \tag{8.74}$$

where $\bar{C}$ is a constant depending on M, M_h, α, C_3 and C, and step (i) above follows from equation (8.73). We justify step (ii) shortly, but let us first derive the linear rate of convergence of the sequence $\Big\{\mathbf{x}_k\Big\}_{k\geq 0}$ using the derivation in equation (8.74). Denote $e_k = \sum_{\ell=k}^{\infty} \|\mathbf{x}_\ell - \mathbf{x}_{\ell+1}\|_2$. Then equation (8.74) provides the following recursion

$$e_{k_1} \leq \bar{C}(e_{k_1-1} - e_{k_1}).$$

Simple inspection of proof of Theorem 7 and derivations so far ensure that we can derive the equations (8.72) and (8.73) for all $k \geq k_1$; this provides us a recursion relation as above with k_1 replaced by k. Furthermore, by choosing a larger value of the constant $\bar{C}$ if necessary, we may conclude that for all $k \geq 1$ we have

$$e_k \leq \bar{C}(e_{k-1} - e_k).$$

Rearranging the above inequality yields $e_k \leq \frac{\bar{C}}{\bar{C}+1} e_{k-1}$, which guarantees that the sequence $\Big\{e_k\Big\}_{k\geq 0}$ converges to zero at a linear rate. Finally, observe that $\|\mathbf{x}_k - \mathbf{x}_*\|_2 \leq \sum_{\ell=k}^{\infty} \|\mathbf{x}_\ell - \mathbf{x}_{\ell+1}\|_2 = e_k$, and the linear rate of convergence of the sequence $\Big\{\|\mathbf{x}_k - \mathbf{x}_*\|_2\Big\}_{k\geq 0}$ to zero follows.

Justification for step (ii) in equation (8.74): Note that it suffices to show that the object $\|\mathbf{x}_{k_1} - \mathbf{x}_{k_1+1}\|_2$ is upper bounded by a constant multiple of $\|\mathbf{x}_{k_1} - \mathbf{x}_{k_1-1}\|_2$, where the constant depends only on M, M_h, α and C. Recalling the decent property proved in equation (8.52) we have:

$$\Big(f(\mathbf{x}_{k_1}) - \bar{f}\Big)^{\frac{1}{2}} \geq \Big(f(\mathbf{x}_{k_1}) - f(\mathbf{x}_{k_1+1})\Big)^{\frac{1}{2}} \geq \frac{1}{\sqrt{2\alpha}}\|\mathbf{x}_{k_1} - \mathbf{x}_{k_1+1}\|_2. \tag{8.75}$$

Combining equations (8.75) and (8.73) we obtain the following upper and lower bound of $\Big(f(\mathbf{x}_{k_1}) - \bar{f}\Big)^{\frac{1}{2}}$:

$$\frac{1}{\sqrt{2\alpha}}\|\mathbf{x}_{k_1} - \mathbf{x}_{k_1+1}\|_2 \leq \Big(f(\mathbf{x}_{k_1}) - \bar{f}\Big)^{\frac{1}{2}} \leq C(M + M_h + 1/\alpha)\|\mathbf{x}_{k_1} - \mathbf{x}_{k_1-1}\|_2.$$

Rearranging the last equality proves the desired upper bound. Finally, we reiterate that the above justification also hold for any iterate k with $k \geq k_1$.

Proof of Corollary 9

The proof of this corollary is based on application of Theorems 4 and 7. We verify the assumptions of Theorems 4 and 7 with $g(\theta) = -\sum_{i=1}^{n} \log\Big(\zeta(y_i;\theta)\Big)$, $h \equiv 0$, $\varphi = \mathbb{1}_{\mathcal{X}}$ and function $f := g - \varphi + h$. Note that the domain $\mathrm{dom}(f) = \mathcal{X}$ is compact, which guarantees that the iterate sequence $\{\theta^k\}_{k\geq 0}$ obtained from Algorithm 3 is bounded. The function $h \equiv 0$ is smooth. The log-partition function A is twice continuously differentiable by assumption, which guarantees that the function g is also twice continuously differentiable, whence smooth in the compact domain $\mathcal{X}$. Finally, we verify that the function f satisfies Assumption KL by proving that f is continuous sub-analytic in its domain $\mathcal{X}$, and the domain $\mathcal{X}$ is closed; see Lemma 14. Clearly, $\mathrm{dom}(f) = \mathcal{X}$ is closed, and the function f is continuous in $\mathrm{dom}(f)$. Finally, we show that the functions (g, φ) are sub-analytic, and invoking the property (d) of sub-analytic functions form Section 8.6, we conclude that the function $f := g + \varphi$ is sub-analytic.

The function φ is sub-analytic: Here, we use a simple result by [Att+10], which states that the indicator function of a semi-algebraic set is a semi-algebraic function (hence a sub-analytic function). In order to show that the set $\mathcal{X}$ is semi-algebraic, we note the following representation of the set $\mathcal{X}$

$$\mathcal{X} = \Big\{\sum_{i=1}^{d} \theta_i^2 > R_1^2\Big\}^c \bigcap \Big\{\sum_{i=d+1}^{2d} \theta_i^2 > R_2^2\Big\}^c \bigcap \Big\{\theta_{2d+1} > 1\Big\}^c \bigcap \Big\{-\theta_{2d+1} > 0\Big\}^c. \quad (8.76)$$

Each of the four sets in representation (8.76) are semi-algebraic by definition, and semi-algebraic sets are closed under finite intersection and complements; see the book by [Cos02]. Putting together these two observations, we conclude that the set $\mathcal{X}$ is semi-algebraic, and that $\mathbb{1}_{\mathcal{X}}$ is a sub-analytic function.

The function g is sub-analytic: The log-partition function A is sub-analytic by assumption. For a fixed vector y, the map $\eta \mapsto \eta^\top T(y)$ is linear, and hence sub-analytic. Since sub-analytic functions are closed under a finite linear combination, we conclude that the map $\eta \mapsto \eta^\top T(y) - A(\eta)$ is sub-analytic. Continuous sub-analytic functions are closed under multiplication and composition; since the $\exp(\cdot)$ function is continuous sub-analytic, we have for every fixed vector y the following map

$$(\eta_0, \eta_1, p) \mapsto \zeta(y; \eta_0, \eta_1, p) := p \exp(\eta_0^\top T(y) - A(\eta_0)) + (1-p)\exp(\eta_1^\top T(y) - A(\eta_1))$$

is sub-analytic. Furthermore, the $\log(\cdot)$ function analytic on the interval $(0, \infty)$, and using the composition rule for continuous sub-analytic functions, we obtain that the map $\theta \mapsto \log(\zeta(y_i;\theta))$ is sub-analytic, where $\theta := (\eta_0, \eta_1, p)$. Finally, the target function g is a linear combination of sub-analytic functions $\log(\zeta(y_i;\theta))$, and we conclude that the map $\theta \mapsto g(\theta)$ is sub-analytic.

Combining the pieces: Putting together the pieces, we conclude that the function f is sub-analytic, with the function f being continuous in $\mathrm{dom}(f)$, whereas $\mathrm{dom}(f)$ is closed; furthermore, the functions g and h are smooth. This allows us to apply Theorem 4 and Theorem 7 and the corollary follows.

Sub-analyticity of the log-partition functions A in Table 1: The sub-analyticity of the log-partition function A mentioned in Table 1 follows from the following two observations. First, note that the functions $\exp, \ln$ and Γ are continuous and analytic (hence sub-analytic). Given two continuous sub-analytic functions g_1 and g_2, the composition function $g_2 \circ g_1$ is also continuous sub-analytic. Secondly, any linear combination of sub-analytic functions is also sub-analytic function. See Section 8.6 for properties of sub-analytic functions.

8.12 Characterizing "smooth - convex" function class

In Theorem 3 and Theorem 4 we discussed a class of non-smooth non-convex functions, where a gradient or a prox-type algorithm provides satisfactory convergence to a critical point. One possible deficiency of the theory discussed so far is that, in Algorithm 2 (respectively Algorithm 3), we need to specify a decomposition of the objective function f as a difference of a smooth and a convex function (respectively, smooth + convex - convex). Consequently, it is natural to wonder if we can characterize the class of functions which has a decomposition needed in Algorithms 2 and 3. Furthermore, if a function has this a decomposition, how can we obtain such a decomposition easily. It is worth pointing out that for the case of Algorithm 3, the convex function φ is known in many cases. For instance, in the case of constrained optimization, the function φ is the indicator of the constraint set; in many statistical estimation problems, φ is a penalty function on the parameters; a well-known example of such penalty function is the ℓ_1 penalty, which is used to obtain sparse solutions. Hence, for all practical purposes, the task of characterizing the function class mentioned in Theorems 3 and 3 reduces to characterizing functions which can be decomposed as a difference of a smooth function(g) and a convex function (h). In the next theorem, we characterize the class of of continuously differentiable functions that can be written as a difference of a smooth function and a convex function.

Theorem 8. *Given any continuously differentiable function $f : \mathbb{R}^d \mapsto \mathbb{R}$, the following two properties are equivalent:*

(a) There exists a M-smooth function g, and a convex continuously differentiable function h such that:

$$f(x) = g(x) - h(x) \quad \textit{for all } \; x \in \mathbb{R}^d.$$

(b) The gradient of the function f satisfies the following inequality:

$$\Big\langle \nabla f(x) - \nabla f(y), x - y \Big\rangle \leq M\|x-y\|^2 \quad \textit{for all } \; x, y \in \mathbb{R}^d.$$

Proof. We establish the equivalence by proving the circle of implications $(a) \implies (b) \implies (a)$.

Implication (a) $\implies$ (b): For any M-smooth function g, we have the following:

$$\begin{aligned}\Big\langle \nabla g(x) - g(y), x - y \Big\rangle &\leq \|\nabla g(x) - g(y)\|_2 \times \|x-y\|_2 \\ &\overset{(i)}{\leq} M\|x-y\|_2^2, \quad \text{for all } \; x, y \in \mathbb{R}^d, \end{aligned} \tag{8.77}$$

where step (i) follows since the gradient ∇g is M Lipschitz. Next note that the gradient of a differentiable convex function is a monotone operator, and we have that for all $x, y \in \mathbb{R}^d$:

$$\Big\langle \nabla h(x) - \nabla h(y), x - y \Big\rangle \geq 0. \tag{8.78}$$

Subtracting equation (8.78) from equation (8.77), we obtain the desired upper bound in part (b).

Implication (b) $\implies$ (a): We prove this implication by finding a M–smooth function g and a convex differentiable function h such that $f = g - h$. To this end, we fix any $x_0 \in \mathbb{R}^d$ and consider the following two functions:

$$g(x) := f(x_0) + \Big\langle \nabla f(x_0), x - x_0 \Big\rangle + \frac{M}{2}\|x - x_0\|_2^2 \tag{8.79a}$$

$$h(x) := g(x) - f(x). \tag{8.79b}$$

The function g in definition (8.79a) is M-smooth by inspection. Since both the functions f and g are continuously differentiable, the function h is continuously differentiable by construction. In order to complete the proof, it suffices to show that the function h is convex. To this end, the first order Taylor series expansion of the function h yields

$$\begin{aligned} h(x) &= h(y) + \Big\langle \nabla h(y + t(x-y)), x - y \Big\rangle \quad \text{for some } t \; \in [0,1] \\ &= h(y) + \Big\langle \nabla h(y), x - y \Big\rangle + \Big\langle \nabla h(y + t(x-y)) - \nabla h(y), x - y \Big\rangle. \end{aligned} \tag{8.80}$$

Expanding the term $\Big\langle \nabla h(y + t(x-y)) - \nabla h(y), x - y \Big\rangle$ above yields,

$$\begin{aligned} \Big\langle \nabla h(y + t(x-y)) - \nabla h(y), x - y \Big\rangle &\overset{(i)}{=} M\|x-y\|_2^2 \\ &\quad - \frac{\Big\langle \nabla f(y + t(x-y)) - \nabla f(y), t(x-y) \Big\rangle}{t} \\ &\overset{(ii)}{\geq} M\|x-y\|_2^2 - Mt\|x-y\|_2^2 \\ &\overset{(iii)}{\geq} 0. \end{aligned}$$

Here step (i) follows by substituting the expression of the function h; step (ii) follows from the gradient inequality of part (b), and step (iii) follows from the inequality $0 \leq t \leq 1$. Since the vectors $x, y \in \mathbb{R}^d$ were arbitrary, the inequality $\Big\langle \nabla h(y + t(x - y)) - \nabla h(y), x - y \Big\rangle \geq 0$ combined with equation (8.80) proves the convexity of the function h, thereby proving the claimed result in part (a).

□

Comments: It would be interesting to characterize the class of DC-based functions mentioned in problem (8.2) when the convex function h is non-differentiable. Indeed, we obtain a larger and more interesting non-differentiable class of functions. It would interesting to see whether Theorem 8 can be suitably generalized in this setting.

Bibliography

[APS11] Yasin Abbasi-Yadkori, David Pál, and Csaba Szepesvári. "Online least squares estimation with self-normalized processes: An application to bandit problems". In: *arXiv preprint arXiv:1102.2670* (2011) (cit. on pp. 264, 278, 284).

[ANW12] A. Agarwal, S. Negahban, and M. J. Wainwright. "Fast global convergence of gradient methods for high-dimensional statistical recovery". In: *Annals of Statistics* 40.5 (2012), pp. 2452–2482 (cit. on p. 200).

[AJK19] Alekh Agarwal, Nan Jiang, and Sham M Kakade. "Reinforcement Learning: Theory and Algorithms". In: *Technical Report, Department of Computer Science, University of Washington* (2019) (cit. on p. 9).

[AP13] Amir A Ahmadi and Pablo A Parrilo. "A complete characterization of the gap between convexity and SOS-convexity". In: *SIAM Journal on Optimization* 23.2 (2013), pp. 811–833 (cit. on pp. 321, 332).

[Ahm+13] Amir A Ahmadi et al. "NP-hardness of deciding convexity of quartic polynomials and related problems". In: *Mathematical Programming* 137.1-2 (2013), pp. 453–476 (cit. on p. 321).

[AW08] Arash A Amini and Martin J Wainwright. "High-dimensional analysis of semidefinite relaxations for sparse principal components". In: *2008 IEEE International Symposium on Information Theory*. IEEE. 2008, pp. 2454–2458 (cit. on p. 200).

[AN17] Nguyen Thai An and Nguyen Mau Nam. "Convergence analysis of a proximal point algorithm for minimizing differences of functions". In: *Optimization* 66.1 (2017), pp. 129–147 (cit. on pp. 314, 328).

[Åst12] Karl J Åström. *Introduction to stochastic control theory*. Courier Corporation, 2012 (cit. on p. 263).

[Att+10] Hédy Attouch et al. "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality". In: *Mathematics of Operations Research* 35.2 (2010), pp. 438–457 (cit. on pp. 314, 364).

[Aue+95] P. Auer et al. “Gambling in a rigged casino: The adversarial multi-armed bandit problem”. In: *Proceedings of IEEE 36th Annual Foundations of Computer Science*. 1995, pp. 322–331. DOI: 10.1109/SFCS.1995.492488 (cit. on p. 292).

[AOM17] M. G. Azar, I. Osband, and R. Munos. “Minimax regret bounds for reinforcement learning”. In: *Proceedings of the International Conference on Machine Learning*. 2017 (cit. on p. 10).

[AMK13] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. “Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model”. In: *Machine Learning* 91.3 (2013), pp. 325–349 (cit. on pp. 52, 55, 59).

[BM11] F. Bach and E. Moulines. “Non-asymptotic analysis of stochastic optimization algorithms for machine learning”. In: *Advances in Neural Information Processing Systems*. Dec. 2011 (cit. on p. 10).

[BM13] Francis Bach and Eric Moulines. “Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$”. In: *arXiv preprint arXiv:1306.2119* (2013) (cit. on p. 87).

[Bai95] Leemon Baird. “Residual algorithms: Reinforcement learning with function approximation”. In: *Machine Learning Proceedings 1995*. Elsevier, 1995, pp. 30–37 (cit. on p. 9).

[BWY17] S. Balakrishnan, M. J. Wainwright, and B. Yu. “Statistical guarantees for the EM algorithm: From population to sample-based analysis”. In: *Annals of Statistics* 45 (2017), pp. 77–120 (cit. on pp. 5, 156, 157, 159, 167, 170–172, 200, 202–204, 213, 221).

[BBM05] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. “Local rademacher complexities”. In: *The Annals of Statistics* 33.4 (2005), pp. 1497–1537 (cit. on pp. 157, 172).

[BT09] Amir Beck and Marc Teboulle. “A fast iterative shrinkage-thresholding algorithm for linear inverse problems”. In: *SIAM Journal on Imaging Sciences* 2.1 (2009), pp. 183–202 (cit. on p. 200).

[BBC11] Stephen Becker, Jérôme Bobin, and Emmanuel J Candès. “NESTA: A fast and accurate first-order method for sparse recovery”. In: *SIAM Journal on Imaging Sciences* 4.1 (2011), pp. 1–39 (cit. on p. 200).

[Ben96] Michel Benaim. “A dynamical system approach to stochastic approximations”. In: *SIAM Journal on Control and Optimization* 34.2 (1996), pp. 437–472 (cit. on p. 86).

[BMP12] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations*. Vol. 22. Springer Science & Business Media, 2012 (cit. on pp. 86, 90).

[Ber95a] D. P. Bertsekas. *Dynamic Programming and Stochastic Control*. Vol. 1. Belmont, MA: Athena Scientific, 1995 (cit. on p. 8).

[Ber95b] D.P. Bertsekas. *Dynamic Programming and Stochastic Control*. Vol. 2. Belmont, MA: Athena Scientific, 1995 (cit. on p. 8).

[Ber12a] Dimitri P Bertsekas. *Approximate Dynamic Programming*. Athena Scientific Belmont, 2012 (cit. on pp. 84, 88, 151).

[Ber19] Dimitri P Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific Belmont, MA, 2019 (cit. on pp. 84, 88, 112).

[Ber12b] Dimitri P Bertsekas. "Weighted sup-norm contractions in dynamic programming: A review and some new applications". In: *Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. LIDS-P-2884* (2012) (cit. on pp. 88, 95).

[BT91] Dimitri P Bertsekas and John N Tsitsiklis. "An analysis of stochastic shortest path problems". In: *Mathematics of Operations Research* 16.3 (1991), pp. 580–595 (cit. on pp. 95, 104, 105).

[BT96] Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996 (cit. on pp. 8, 9).

[Ber09] Dimitri P. Bertsekas. *Neuro-Dynamic Programming*. Boston, MA: Springer US, 2009. ISBN: 978-0-387-74759-0 (cit. on p. 51).

[BRS18a] Jalaj Bhandari, Daniel Russo, and Raghav Singal. "A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation". In: *Conference On Learning Theory*. PMLR. 2018, pp. 1691–1692 (cit. on p. 50).

[BRS18b] Jalaj Bhandari, Daniel Russo, and Raghav Singal. "A finite time analysis of temporal difference learning with linear function approximation". In: *arXiv preprint arXiv:1806.02450* (2018) (cit. on p. 10).

[Bir83a] Lucien Birgé. "Approximation dans les espaces métriques et théorie de l'estimation". In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwebte Gebiet* 65 (1983), pp. 181–238 (cit. on pp. 15, 38).

[Bir83b] Lucien Birgé. "Approximation dans les espaces métriques et théorie de l'estimation". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 65.2 (1983), pp. 181–237 (cit. on p. 74).

[BDL07] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems". In: *SIAM Journal on Optimization* 17.4 (2007), pp. 1205–1223 (cit. on pp. 328, 341, 342).

[BST14] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. "Proximal alternating linearized minimization for nonconvex and nonsmooth problems". In: *Mathematical Programming* 146.1-2 (2014), pp. 459–494 (cit. on pp. 314, 342).

[Bor21] Vivek S Borkar. "A concentration bound for contractive stochastic approximation". In: *Systems & Control Letters* 153 (2021), p. 104947 (cit. on p. 87).

[Bor09] Vivek S Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Vol. 48. Springer, 2009 (cit. on pp. 9, 85, 86, 90).

[BM00] Vivek S Borkar and Sean P Meyn. "The ODE method for convergence of stochastic approximation and reinforcement learning". In: *SIAM Journal on Control and Optimization* 38.2 (2000), pp. 447–469 (cit. on p. 9).

[BCN18] Léon Bottou, Frank E Curtis, and Jorge Nocedal. "Optimization methods for large-scale machine learning". In: *SIAM Review* 60.2 (2018), pp. 223–311 (cit. on p. 84).

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford, UK: Oxford University Press, 2013 (cit. on p. 264).

[Box+15] George EP Box et al. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015 (cit. on pp. 263, 280).

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004 (cit. on p. 333).

[Bub+15] Sébastien Bubeck et al. "Convex optimization: Algorithms and complexity". In: *Foundations and Trends® in Machine Learning* 8.3-4 (2015), pp. 231–357 (cit. on p. 325).

[CL15a] T Cai and Mark Low. "A Framework For Estimation Of Convex Functions". In: *Statistica Sinica* 25 (Apr. 2015), pp. 423–456 (cit. on pp. 52, 53).

[CG+17] T Tony Cai, Zijian Guo, et al. "Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity". In: *The Annals of statistics* 45.2 (2017), pp. 615–646 (cit. on p. 291).

[CL04] T Tony Cai and Mark G Low. "An adaptation theory for nonparametric confidence intervals". In: *Annals of Statistics* 32.5 (2004), pp. 1805–1840 (cit. on p. 11).

[CMZar] T. T. Cai, J. Ma, and L. Zhang. "CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality". In: *Annals of Statistics* (To Appear) (cit. on pp. 156, 200, 202, 221).

[CL15b] Tony Cai and Mark Low. “A framework for estimating convex functions”. In: *Statistica Sinica* 25 (2015), pp. 423–456 (cit. on pp. 3, 15, 17).

[CLS15] E. J. Candès, X. Li, and M. Soltanolkotabi. “Phase Retrieval via Wirtinger Flow: Theory and Algorithms”. In: *IEEE Transactions on Information Theory* 61 (2015), pp. 1985–2007 (cit. on pp. 200, 226).

[CSV12] E. J. Candès, T. Strohmer, and V. Voroninski. “PhaseLift: Exact and Stable Signal Recovery from Magnitude Measurements via Convex Programming”. In: *Communications on Pure and Applied Mathematics* 66 (2012), pp. 1241–1274 (cit. on p. 200).

[Car+17] Yair Carmon et al. “Lower bounds for finding stationary points i”. In: *arXiv preprint arXiv:1710.11606* (2017) (cit. on p. 319).

[Car+97] R. J. Carroll et al. “Generalized Partially Linear Single-Index Models”. In: *Journal of the American Statistical Association* 92 (1997), pp. 477–489 (cit. on pp. 205, 224).

[CGT10] Coralia Cartis, Nicholas IM Gould, and Ph L Toint. “On the complexity of steepest descent, Newton’s and regularized Newton’s methods for nonconvex unconstrained optimization problems”. In: *Siam journal on optimization* 20.6 (2010), pp. 2833–2852 (cit. on pp. 314, 318).

[CP18] Zachary Charles and Dimitris Papailiopoulos. “Stability and Generalization of Learning Algorithms that Converge to Global Optima”. In: *International Conference on Machine Learning*. 2018, pp. 745–754 (cit. on p. 200).

[Che95] J. Chen. “Optimal rate of convergence for finite mixture models”. In: *Annals of Statistics* 23.1 (1995), pp. 221–233 (cit. on pp. 5, 155, 157, 166, 176, 202, 203, 222).

[CL+09] Jiahua Chen, Pengfei Li, et al. “Hypothesis test for normal mixture models: The EM approach”. In: *The Annals of Statistics* 37.5A (2009), pp. 2523–2542 (cit. on p. 179).

[Che+20a] Xi Chen et al. “Statistical inference for model parameters in stochastic gradient descent”. In: *The Annals of Statistics* 48.1 (2020), pp. 251–273 (cit. on pp. 87, 145).

[CW15] Y. Chen and M. J. Wainwright. *Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees*. Tech. rep. arxiv:1509.03025.pdf. UC Berkeley, Sept. 2015 (cit. on p. 200).

[Che22] Yen-Chi Chen. “Statistical inference with local optima”. In: *Journal of the American Statistical Association* (2022), pp. 1–13 (cit. on p. 156).

[CJY18] Yuansi Chen, Chi Jin, and Bin Yu. “Stability and Convergence Trade-off of Iterative Optimization Algorithms”. In: *arXiv preprint arXiv:1804.01619* (2018) (cit. on pp. 200, 202).

[Che+18] Yuxin Chen et al. “Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval”. In: *Mathematical Programming* (2018), pp. 1–33 (cit. on pp. 200, 202, 203).

[Che+21a] Zaiwei Chen et al. “A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants”. In: *arXiv preprint arXiv:2102.01567* (2021) (cit. on p. 87).

[Che+21b] Zaiwei Chen et al. “Finite-Sample Analysis of Off-Policy TD-Learning via Generalized Bellman Operators”. In: *Advances in Neural Information Processing Systems* 34 (2021) (cit. on p. 87).

[Che+20b] Zaiwei Chen et al. “Finite-sample analysis of stochastic approximation using smooth convex envelopes”. In: *arXiv e-prints* (2020), arXiv–2002 (cit. on p. 87).

[Che+19] Zaiwei Chen et al. “Performance of Q-learning with linear function approximation: Stability and finite-time analysis”. In: *arXiv preprint arXiv:1905.11425* (2019) (cit. on p. 87).

[Che64] H. Chernoff. “Estimation of the mode”. In: *Annals of the Institute of Statistical Mathematics* 16 (1964), pp. 31–41 (cit. on p. 202).

[Cos02] Michel Coste. “An introduction to semialgebraic geometry”. In: *RAAG network school* 145 (2002), p. 30 (cit. on p. 364).

[CO19] Ashok Cutkosky and Francesco Orabona. “Momentum-based variance reduction in non-convex SGD”. In: *arXiv preprint arXiv:1905.10018* (2019) (cit. on p. 88).

[Dal+18] Gal Dalal et al. “Finite sample analyses for TD(0) with function approximation”. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018 (cit. on pp. 10, 50).

[DNP14] Christoph Dann, Gerhard Neumann, and Jan Peters. “Policy evaluation with temporal differences: A survey and comparison”. In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 809–883 (cit. on p. 9).

[DTZ17] C. Daskalakis, C. Tzamos, and M. Zampetakis. “Ten Steps of EM Suffice for Mixtures of Two Gaussians”. In: *Proceedings of the 2017 Conference on Learning Theory*. 2017 (cit. on pp. 156, 157, 159, 170, 221).

[DBL14] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. “SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives”. In: *Advances in Neural Information Processing Systems*. 2014, pp. 1646–1654 (cit. on p. 88).

[DLR97] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum Likelihood from Incomplete Data via the EM Algorithm”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 39 (1997), pp. 1–38 (cit. on pp. 155, 221).

[Der66] Cyrus Derman. "Denumerable state Markovian decision processes-average cost criterion". In: *The Annals of Mathematical Statistics* 37.6 (1966), pp. 1545–1553 (cit. on p. 112).

[DJM19] Yash Deshpande, Adel Javanmard, and Mohammad Mehrabi. "Online debiasing for adaptively collected high-dimensional data". In: *arXiv preprint arXiv:1911.01040* (2019) (cit. on pp. 264, 265).

[Des+18] Yash Deshpande et al. "Accurate Inference for Adaptive Linear Models". In: *International Conference on Machine Learning*. Vol. 35. 2018, pp. 1194–1203 (cit. on pp. 263–265, 267, 269, 278).

[DF79] David A Dickey and Wayne A Fuller. "Distribution of the estimators for autoregressive time series with a unit root". In: *Journal of the American statistical association* 74.366a (1979), pp. 427–431 (cit. on p. 267).

[DK94] P. Diggle and M. G. Kenward. "Informative Drop-Out in Longitudinal Data Analysis". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 43 (1994), pp. 49–93 (cit. on pp. 202, 217).

[DMR19] Thinh T Doan, Siva Theja Maguluri, and Justin Romberg. "Finite-time performance of distributed temporal difference learning with linear function approximation". In: *arXiv preprint arXiv:1907.12530* (2019) (cit. on p. 10).

[DL87] David L. Donoho and Richard C. Liu. *Geometrizing Rates of Convergence I*. Tech. rep. 137. University of California, Berkeley, Department of Statistics, 1987 (cit. on p. 15).

[DL91] David L. Donoho and Richard C. Liu. "Geometrizing Rates of Convergence II". In: *Annals of Statistics* 19.2 (1991), pp. 633–667 (cit. on p. 15).

[DR16] John Duchi and Feng Ruan. "Asymptotic optimality in stochastic optimization". In: *arXiv preprint arXiv:1612.05612* (2016) (cit. on pp. 84, 86, 87, 99).

[DG03] James Dugundji and Andrzej Granas. *Fixed Point Theory*. Springer Verlag, 2003 (cit. on pp. 84, 89).

[Dur99] Richard Durrett. *Essentials of Stochastic Processes*. Springer, 1999 (cit. on p. 9).

[Dvo+72] Aryeh Dvoretzky et al. "Asymptotic normality for sums of dependent random variables". In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California. 1972 (cit. on pp. 286, 287).

[Dwi+20a] R. Dwivedi et al. "Sharp analysis of Expectation-Maximization for weakly identifiable models". In: *AISTATS* (2020) (cit. on pp. 5, 200, 202, 203, 214, 221, 228, 250, 253, 256).

[Dwi+20b] R. Dwivedi et al. “Singularity, misspecification, and the convergence rate of EM”. In: *To appear, Annals of Statistics* (2020) (cit. on pp. 200, 202, 203, 214, 221, 223, 245, 249, 251).

[EJ10] Ady Ecker and Allan D Jepson. “Polynomial shape from shading”. In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE. 2010, pp. 145–152 (cit. on p. 330).

[EM13] Y. C. Eldar and S. Mendelson. “Phase retrieval: Stability and recovery guarantees”. In: *Applied and Computational Harmonic Analysis* 36 (2013), pp. 473–494 (cit. on p. 226).

[FP07] Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007 (cit. on p. 342).

[FL01] Jianqing Fan and Runz Li. “Variable selection via non-concave penalized likelihood and its oracle properties”. In: *Jour. Amer. Stat. Ass.* 96.456 (Dec. 2001), pp. 1348–1360 (cit. on p. 313).

[Fel66] W. Feller. *An Introduction to Probability Theory and its Applications: Volume II*. New York: John Wiley and Sons, 1966 (cit. on p. 9).

[Fle87] R. Fletcher. *Practical methods of optimization*. John Wiley & Sons, 1987 (cit. on p. 228).

[Fon+19] Xavier Fontaine et al. “Online A-Optimal Design and Active Linear Regression”. In: *arXiv preprint arXiv:1906.08509* (2019) (cit. on p. 263).

[Fro+15] Roy Frostig et al. “Competing with the empirical risk minimizer in a single pass”. In: *Conference on Learning Theory*. PMLR. 2015, pp. 728–763 (cit. on p. 87).

[GP17] Sébastien Gadat and Fabien Panloup. “Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity”. In: *arXiv preprint arXiv:1709.03342* (2017) (cit. on p. 87).

[GK09] Rahul Garg and Rohit Khandekar. “Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property.” In: *ICML*. Vol. 9. 2009, pp. 337–344 (cit. on p. 200).

[GJZ17] Rong Ge, Chi Jin, and Yi Zheng. “No spurious local minima in nonconvex low rank problems: A unified geometric analysis”. In: *arXiv preprint arXiv:1704.00708* (2017) (cit. on p. 313).

[Gee00] S. van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000 (cit. on pp. 155, 157, 172).

[GL12] Saeed Ghadimi and Guanghui Lan. “Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework”. In: *SIAM Journal on Optimization* 22.4 (2012), pp. 1469–1492 (cit. on p. 84).

[GL13] Saeed Ghadimi and Guanghui Lan. "Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms". In: *SIAM Journal on Optimization* 23.4 (2013), pp. 2061–2089 (cit. on p. 84).

[GV01] S. Ghosal and A. van der Vaart. "Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities". In: *Annals of Statistics* 29 (2001), pp. 1233–1263 (cit. on p. 155).

[GP77] GC Goodwin and RL Payne. *Dynamic system identification experiment design and data analysis.* Mathematics in Science and Engineering, Volume 136. Elsevier, 1977 (cit. on p. 263).

[GTT17] Jun-Ya Gotoh, Akiko Takeda, and Katsuya Tono. "DC formulations and algorithms for sparse optimization problems". In: *Mathematical Programming* (2017), pp. 1–36 (cit. on pp. 314, 333).

[GJG18] Abhishek Gupta, Rahul Jain, and Peter Glynn. "Probabilistic contraction analysis of iterated random operators". In: *arXiv preprint arXiv:1804.01195* (2018) (cit. on p. 87).

[Had+19] Vitor Hadad et al. "Confidence intervals for policy evaluation in adaptive experiments". In: *arXiv preprint arXiv:1911.02768* (2019) (cit. on p. 264).

[Háj72] Jaroslav Hájek. "Local asymptotic minimax and admissibility in estimation". In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*. 1972, pp. 175–194 (cit. on pp. 2, 3, 11, 16).

[Häj72] Jaroslav Häjek. "Local asymptotic minimax and admissibility in estimation". In: *Theory of Statistics*. University of California Press, 1972, pp. 175–194 (cit. on pp. 85, 86, 99).

[HYZ08] Elaine T Hale, Wotao Yin, and Yin Zhang. "Fixed-point continuation for ℓ_1-minimization: Methodology and convergence". In: *SIAM Journal on Optimization* 19.3 (2008), pp. 1107–1130 (cit. on p. 200).

[Hao+ar] B. Hao et al. "Simultaneous Clustering and Estimation of Heterogeneous Graphical Models". In: *Journal of Machine Learning Research* (To Appear) (cit. on p. 156).

[HRS16] Moritz Hardt, Ben Recht, and Yoram Singer. "Train faster, generalize better: Stability of stochastic gradient descent". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1225–1234. URL: http://proceedings.mlr.press/v48/hardt16.html (cit. on p. 200).

[Har59] Philip Hartman. "On functions representable as a difference of convex functions". In: *Pacific Journal of Mathematics* 9.3 (1959), pp. 707–713 (cit. on pp. 314, 322).

[HTW15] T. Hastie, R. Tibshrani, and M. J. Wainwright. *Statistical Learning with Sparsity: The Lasso and generalizations*. CRC Press, 2015 (cit. on p. 202).

[Hec76] J. J. Heckman. "The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models". In: *Annals of Economic and Social Measurement* 5 (1976), pp. 475–492 (cit. on pp. 202, 217).

[HK18] P. Heinrich and J. Kahn. "Strong identifiability and optimal minimax rates for finite mixture estimation". In: *Annals of Statistics* 46 (2018), pp. 2844–2870 (cit. on p. 155).

[HN16] N. Ho and X. Nguyen. "Convergence rates of parameter estimation for some weakly identifiable finite mixtures". In: *Annals of Statistics* 44 (2016), pp. 2726–2755 (cit. on p. 203).

[HLR16] Mingyi Hong, Zhi-Quan Luo, and Meisam Razaviyayn. "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems". In: *SIAM Journal on Optimization* 26.1 (2016), pp. 337–364 (cit. on p. 314).

[HH96] Joel Horowitz and Wolfgang Härdle. "Direct semiparametric estimation of single-index models with discrete covariates". In: *Journal of the American Statistical Association* 91.436 (1996), pp. 1632–1640 (cit. on p. 224).

[HPT00] Reiner Horst, Panos M. Pardalos, and Nguyen Van Thoai. *Introduction to Global Optimization*. Vol. 9. Nonconvex Optimization and its Applications. Elsevier, 2000 (cit. on p. 314).

[How+] Steven R Howard et al. "Time-uniform, nonparametric, non-asymptotic confidence sequences". In: *The Annals of Statistics* (), To appear (cit. on p. 264).

[Ich93] H. Ichimura. "Semiparametric least squares (SLS) and weighted (SLS) estimation of single index models". In: *Journal of Econometrics* 58 (1993), pp. 71–120 (cit. on pp. 205, 224).

[Ing10] Tadeusz Inglot. "Inequalities for quantiles of the chi-square distribution". In: *Probability and Mathematical Statistics* 30.2 (2010), pp. 339–351 (cit. on p. 189).

[IJS01] H. Ishwaran, L. F. James, and J. Sun. "Bayesian model selection in finite mixtures by marginal density decompositions". In: *Journal of the American Statistical Association* 96 (2001), pp. 1316–1332 (cit. on p. 176).

[JJS94a] Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. "Convergence of stochastic iterative dynamic programming algorithms". In: *Advances in Neural Information Processing Systems*. 1994, pp. 703–710 (cit. on p. 9).

[JJS94b] Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. "On the Convergence of Stochastic Iterative Dynamic Programming Algorithms". In: *Neural Computation* 6.6 (1994), pp. 1185–1201 (cit. on p. 52).

[Jag13] Martin Jaggi. "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization." In: *ICML (1)*. 2013, pp. 427–435 (cit. on pp. 327, 354).

[Jam+14] Kevin Jamieson et al. "lil'ucb: An optimal exploration algorithm for multi-armed bandits". In: *Conference on Learning Theory*. Vol. 27. 2014, pp. 423–439 (cit. on pp. 264, 307, 309, 310).

[Jan97] Svante Janson. *Gaussian Hilbert Spaces*. Vol. 129. Cambridge university press, 1997 (cit. on p. 189).

[JA18] Nan Jiang and Alekh Agarwal. "Open problem: The dependence of sample complexity lower bounds on planning horizon". In: *Proceedings of the Conference On Learning Theory*. 2018, pp. 3395–3398 (cit. on p. 10).

[JS20] Yujia Jin and Aaron Sidford. "Efficiently solving MDPs with stochastic mirror descent". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 4890–4900 (cit. on p. 115).

[JZ13] Rie Johnson and Tong Zhang. "Accelerating stochastic gradient descent using predictive variance reduction". In: *Advances in Neural Information Processing Systems* 26 (2013), pp. 315–323 (cit. on pp. 21, 57, 84, 88, 90).

[JNT11] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. "Solving variational inequalities with stochastic mirror-prox algorithm". In: *Stochastic Systems* 1.1 (2011), pp. 17–58 (cit. on pp. 84, 145).

[KK18] Emilie Kaufmann and Wouter Koolen. "Mixture martingales revisited with applications to sequential tests and confidence intervals". In: *arXiv preprint arXiv:1811.11419* (2018) (cit. on p. 264).

[KS99] M. Kearns and S. Singh. "Finite-sample convergence rates for Q-learning and indirect algorithms". In: *Advances in Neural Information Processing Systems*. 1999 (cit. on p. 13).

[Kha+21] Koulik Khamaru et al. "Instance-optimality in optimal value estimation: Adaptivity via variance-reduced Q-learning". In: *arXiv preprint arXiv:2106.14352* (2021) (cit. on pp. 87, 88, 90, 108, 109, 112).

[Kha+20a] Koulik Khamaru et al. “Is Temporal Difference Learning Optimal? An Instance-Dependent Analysis”. In: *arXiv preprint arXiv:2003.07337* (2020), pp. 1–38 (cit. on pp. 49, 50, 55, 71).

[Kha+20b] Koulik Khamaru et al. “Is temporal difference learning optimal? An instance-dependent analysis”. In: *arXiv preprint arXiv:2003.07337* (2020) (cit. on pp. 85, 87, 88, 90, 115).

[Kha66] Rafail Khas’minskii. “On stochastic processes defined by differential equations with a small parameter”. In: *Theory of Probability and Its Applications* 11.2 (1966), pp. 211–228 (cit. on p. 86).

[KW52] Jack Kiefer and Jacob Wolfowitz. “Stochastic estimation of the maximum of a regression function”. In: *The Annals of Mathematical Statistics* (1952), pp. 462–466 (cit. on p. 86).

[Kir11] Andreas Kirsch. *An Introduction to the Mathematical Theory of Inverse Problems*. Vol. 120. Springer, 2011 (cit. on p. 84).

[KYB17] Jason M Klusowski, Dana Yang, and WD Brinda. “Estimating the coefficients of a mixture of two linear regressions by expectation maximization”. In: *arXiv preprint arXiv:1704.08231* (2017) (cit. on p. 156).

[Kol06] Vladimir Koltchinskii. “Local Rademacher complexities and oracle inequalities in risk minimization”. In: *The Annals of Statistics* 34.6 (2006), pp. 2593–2656 (cit. on pp. 157, 172).

[KL15] Nathaniel Korda and Prashanth La. “On TD(0) with function approximation: Concentration bounds and a centered variant with exponential convergence”. In: *International Conference on Machine Learning*. 2015, pp. 626–634 (cit. on p. 10).

[KLL20] Georgios Kotsalis, Guanghui Lan, and Tianjiao Li. “Simple and optimal methods for stochastic variational inequalities, II: Markovian noise and policy evaluation in reinforcement learning”. In: *arXiv preprint arXiv:2011.08434* (2020) (cit. on pp. 84, 145).

[KS17] Raunak Kumar and Mark Schmidt. “Convergence rate of expectation-maximization”. In: *10th NIPS Workshop on Optimization for Machine Learning*. 2017, p. 98 (cit. on p. 156).

[Kur98] Krzysztof Kurdyka. “On gradients of functions definable in o-minimal structures”. In: *Annales de l’institut Fourier*. Vol. 48. Chartres: L’Institut, 1950-. 1998, pp. 769–784 (cit. on p. 328).

[KY03] Harold Kushner and G George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Vol. 35. Springer Science & Business Media, 2003 (cit. on pp. 86, 90).

[KS84] Harold J Kushner and Adam Shwartz. "An invariant measure approach to the convergence of stochastic approximations with state dependent noise". In: *SIAM Journal on Control and Optimization* 22.1 (1984), pp. 13–27 (cit. on p. 86).

[KC78] Harold J. Kushner and Dean S. Clark. *Stochastic approximation methods for constrained and unconstrained systems*. Vol. 26. Applied Mathematical Sciences. Springer-Verlag, New York-Berlin, 1978, pp. x+261. ISBN: 0-387-90341-0 (cit. on p. 86).

[Kus84] Harold Joseph Kushner. *Approximation and weak convergence methods for random processes, with applications to stochastic systems theory*. Vol. 6. MIT press, 1984 (cit. on p. 86).

[KL18] Ilja Kuzborskij and Christoph Lampert. "Data-Dependent Stability of Stochastic Gradient Descent". In: *International Conference on Machine Learning*. 2018, pp. 2815–2824 (cit. on p. 200).

[Lac16] Simon Lacoste-Julien. "Convergence rate of Frank-Wolfe for non-convex objectives". In: *arXiv preprint arXiv:1607.00345* (2016) (cit. on pp. 326, 327, 353).

[LR79] T L Lai and Herbert Robbins. "Adaptive design and stochastic approximation". In: *The Annals of Statistics* 7 (1979), pp. 1196–1221 (cit. on pp. 263, 269).

[LRW79] T L_etc Lai, Herbert Robbins, and C Zi Wei. "Strong consistency of least squares estimates in multiple regression II". In: *Journal of Multivariate Analysis* 9.3 (1979), pp. 343–361 (cit. on p. 263).

[Lai94] Tze Leung Lai. "Asymptotic properties of nonlinear least squares estimates in stochastic regression models". In: *The Annals of Statistics* 22.4 (1994), pp. 1917–1930 (cit. on p. 269).

[LW+82] Tze Leung Lai, Ching Zong Wei, et al. "Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems". In: *The Annals of Statistics* 10.1 (1982), pp. 154–166 (cit. on pp. 263, 267, 269, 280, 281, 286, 297, 302).

[LS18a] Chandrashekar Lakshminarayanan and Csaba Szepesvari. "Linear Stochastic Approximation: How Far Does Constant Step-Size and Iterate Averaging Go?" In: *AISTATS: Conference on AI and Statistics*. Vol. 21. PMLR, 2018, pp. 1347–1355 (cit. on p. 50).

[LS18b] Chandrashekar Lakshminarayanan and Csaba Szepesvári. "Linear stochastic approximation: How far does constant step-size and iterate averaging go?" In: *International Conference on Artificial Intelligence and Statistics*. 2018, pp. 1347–1355 (cit. on p. 10).

[LS09] Gert R Lanckriet and Bharath K Sriperumbudur. "On the convergence of the concave-convex procedure". In: *Advances in neural information processing systems*. 2009, pp. 1759–1767 (cit. on pp. 314, 315, 323, 348).

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020 (cit. on pp. 263, 266, 267, 278, 280, 282, 283, 307, 309, 310).

[Le 72] Lucien Le Cam. "Limits of experiments". In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*. 1972, pp. 245–261 (cit. on pp. 2, 3, 11, 16).

[LY00] Lucien Le Cam and Grace Lo Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2000 (cit. on pp. 11, 16).

[LeC53] Lucien LeCam. "On some asymptotic properties of maximum likelihood estimates and related Bayes estimates". In: *Univ. California Pub. Statist.* 1 (1953), pp. 277–330 (cit. on pp. 3, 99).

[LT91] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991 (cit. on p. 192).

[LT13] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Science & Business Media, 2013 (cit. on pp. 85, 95).

[Lee+16] Jason D Lee et al. "Gradient descent only converges to minimizers". In: *Conference on Learning Theory*. 2016, pp. 1246–1257 (cit. on pp. 314, 322, 346, 361).

[Lee93] L. F. Lee. "Asymptotic distribution of the maximum likelihood estimator for a stochastic frontier function model with a singular information matrix". In: *Econometric Theory* 9 (1993), pp. 413–430 (cit. on p. 202).

[LC86] L. F. Lee and A. Chesher. "Specification testing when score test statistic are identically zero". In: *Journal of Econometrics* 31 (1986), pp. 121–149 (cit. on p. 202).

[Lev+16] Sergey Levine et al. "End-to-End Training of Deep Visuomotor Policies". In: *Journal of Machine Learning Research* 17.1 (2016), pp. 1334–1373 (cit. on p. 49).

[Li+20] Chris Junchi Li et al. "Root-SGD: Sharp nonasymptotics and asymptotic efficiency in a single algorithm". In: *arXiv preprint arXiv:2008.12690* (2020) (cit. on pp. 84, 85, 87, 88, 90, 91, 144).

[Li+21] Gen Li et al. "Is Q-learning minimax optimal? a tight sample complexity analysis". In: *arXiv preprint arXiv:2102.06548* (2021) (cit. on p. 87).

[LP16] Guoyin Li and Ting Kei Pong. “Calculus of the exponent of Kurdya-Łojasiewicz inequality and its applications to linear convergence of first-order methods”. In: *arXiv preprint arXiv:1602.02915* (2016) (cit. on p. 362).

[LCM09] P Li, J Chen, and P Marriott. “Non-finite Fisher information and homogeneity: an EM approach”. In: *Biometrika* 96.2 (2009), pp. 411–426 (cit. on p. 179).

[LLP21] Tianjiao Li, Guanghui Lan, and Ashwin Pananjady. “Accelerated and instance-optimal policy evaluation with linear function approximation”. In: *arXiv preprint arXiv:2112.13109* (2021) (cit. on p. 87).

[LY17] Yuanzhi Li and Yang Yuan. “Convergence analysis of two-layer neural networks with relu activation”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 597–607 (cit. on p. 313).

[LB16] Thomas Lipp and Stephen Boyd. “Variations and extension of the convex–concave procedure”. In: *Optimization and Engineering* 17.2 (2016), pp. 263–287 (cit. on pp. 314, 315, 322).

[Lju77a] Lennart Ljung. “Analysis of recursive stochastic algorithms”. In: *IEEE Transactions on Automatic Control* 22.4 (1977), pp. 551–575 (cit. on p. 86).

[Lju77b] Lennart Ljung. “On positive real transfer functions and the convergence of some recursive schemes”. In: *IEEE Transactions on Automatic Control* 22.4 (1977), pp. 539–551 (cit. on p. 86).

[LW15] Po-Ling Loh and Martin J Wainwright. “Regularized M-estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima”. In: *Journal of Machine Learning Research* 16 (2015), pp. 559–616 (cit. on p. 200).

[LW13] Po-Ling Loh and Martin J Wainwright. “Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima”. In: *Advances in Neural Information Processing Systems*. 2013, pp. 476–484 (cit. on p. 313).

[Łoj63] Stanislaw Łojasiewicz. “Une propriété topologique des sous-ensembles analytiques réels”. In: *Les équations aux dérivées partielles* 117 (1963), pp. 87–89 (cit. on p. 328).

[MXJ00] J. Ma, L. Xu, and M. I. Jordan. “Asymptotic convergence rate of the EM algorithm for Gaussian mixtures”. In: *Neural Computation* 12 (2000), pp. 2881–2907 (cit. on p. 155).

[Ma13] Z. Ma. “Sparse principal component analysis and iterative thresholding”. In: *Annals of Statistics* 41.2 (2013), pp. 772–801 (cit. on p. 200).

[MMM14] Odalric-Ambrym Maillard, Timothy A Mann, and Shie Mannor. "How hard is my MDP? The distribution-norm to the rescue". In: *Advances in Neural Information Processing Systems*. 2014, pp. 1835–1843 (cit. on pp. 10, 49).

[Man75] C. F. Manski. "Maximum score estimation of the stochastic utility model of choice". In: *Journal of Econometrics* 3 (1975), pp. 205–228 (cit. on p. 202).

[MB88] G. J. McLachlan and K. E. Basford. *Mixture Models: Inference and Applications to Clustering. Statistics: Textbooks and Monographs*. New York, 1988 (cit. on p. 160).

[Mou+20] Wenlong Mou et al. "On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration". In: *Conference on Learning Theory*. PMLR. 2020, pp. 2947–2997 (cit. on p. 84).

[Mou+21] Wenlong Mou et al. "Optimal and instance-dependent guarantees for Markovian linear stochastic approximation". In: *arXiv preprint arXiv:2112.12770* (2021) (cit. on p. 87).

[MB11] Éric Moulines and Francis R Bach. "Non-asymptotic analysis of stochastic approximation algorithms for machine learning". In: *Advances in Neural Information Processing Systems*. 2011, pp. 451–459 (cit. on pp. 84, 87).

[Nem+09a] A. Nemirovski et al. "Robust stochastic approximation approach to stochastic programming". In: *SIAM Journal of Optimization* 19.4 (2009), pp. 1574–1609 (cit. on p. 10).

[Nem+09b] Arkadi Nemirovski et al. "Robust stochastic approximation approach to stochastic programming". In: *SIAM Journal on Optimization* 19.4 (2009), pp. 1574–1609 (cit. on p. 84).

[NY83] Arkadi Semenovi**c***Nemirovski* and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983 (cit. on p. 84).

[Nes13] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2013 (cit. on p. 228).

[Nes03] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2003 (cit. on p. 84).

[NP06] Yurii Nesterov and Boris T Polyak. "Cubic regularization of Newton method and its global performance". In: *Mathematical Programming* 108.1 (2006), pp. 177–205 (cit. on p. 318).

[NST21] Lam M Nguyen, Katya Scheinberg, and Martin Taká**c**. "Inexact SARAH algorithm for stochastic optimization". In: *Optimization Methods and Software* 36.1 (2021), pp. 237–258 (cit. on pp. 84, 88, 90).

[Ngu+17] Lam M Nguyen et al. “SARAH: A novel method for machine learning problems using stochastic recursive gradient”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 2613–2621 (cit. on p. 88).

[Ngu13] X. Nguyen. “Convergence of latent mixing measures in finite and infinite mixture models”. In: *Annals of Statistics* 4.1 (2013), pp. 370–400 (cit. on pp. 155, 176, 202, 203).

[Nie+18] Xinkun Nie et al. “Why adaptively collected data have negative bias and how to correct for it”. In: *International Conference on Artificial Intelligence and Statistics*. Vol. 84. 2018, pp. 1261–1269 (cit. on p. 264).

[Ost16] Alexander M Ostrowski. *Solution of equations and systems of equations: Pure and applied mathematics: A Series of monographs and textbooks*. Vol. 9. Elsevier, 2016 (cit. on p. 361).

[PP16] Ioannis Panageas and Georgios Piliouras. “Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions”. In: *arXiv preprint arXiv:1605.00405* (2016) (cit. on pp. 314, 322).

[PW20] A. Pananjady and M. J. Wainwright. “Instance-Dependent ℓ_∞-Bounds for Policy Evaluation in Tabular Reinforcement Learning”. In: *IEEE Transactions on Information Theory* 67.1 (2020), pp. 566–585 (cit. on pp. 49, 50, 55).

[PW19] Ashwin Pananjady and Martin J Wainwright. “Value function estimation in Markov reward processes: Instance-dependent ℓ_∞ -bounds for policy evaluation”. In: *arXiv preprint arXiv:1909.08749* (2019) (cit. on pp. 9, 10, 18).

[PB+14] Neal Parikh, Stephen Boyd, et al. “Proximal algorithms”. In: *Foundations and Trends® in Optimization* 1.3 (2014), pp. 127–239 (cit. on p. 324).

[Pat97] Stephen David Patek. “Stochastic and shortest path games: theory and algorithms”. PhD thesis. Massachusetts Institute of Technology, 1997 (cit. on p. 110).

[Per+15] Julien Perolat et al. “Approximate dynamic programming for two-player zero-sum Markov games”. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 1321–1329 (cit. on p. 110).

[PNL13] T Pham Dinh, HV Ngai, and HA Le Thi. “Convergence analysis of the DC algorithm for DC programming with subanalytic data”. In: *preprint* (2013) (cit. on p. 322).

[Pol90] Boris T Polyak. “A new method of stochastic approximation type”. In: *Automat. i Telemekh* 7.98-107 (1990), p. 2 (cit. on p. 86).

[PJ92] Boris T Polyak and Anatoli B Juditsky. “Acceleration of stochastic approximation by averaging”. In: *SIAM Journal on Control and Optimization* 30.4 (1992), pp. 838–855 (cit. on pp. 9, 14, 17, 18, 27, 28, 84, 86).

[Put14] Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014 (cit. on p. 51).

[PB79] Martin L. Puterman and Shelby L. Brumelle. “On the Convergence of Policy Iteration in Stationary Dynamic Programming”. In: *Mathematics of Operations Research* 4.1 (1979), pp. 60–69 (cit. on p. 58).

[QW20] Guannan Qu and Adam Wierman. “Finite-Time Analysis of Asynchronous Stochastic Approximation and Q-Learning”. In: *Conference on Learning Theory*. PMLR. 2020, pp. 3185–3205 (cit. on p. 87).

[RG97] S. Richardson and P. J. Green. “On Bayesian Analysis of Mixtures with an Unknown Number of Components”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59 (1997), pp. 731–792 (cit. on p. 155).

[RM51] Herbert Robbins and Sutton Monro. “A stochastic approximation method”. In: *The Annals of Mathematical Statistics* (1951), pp. 400–407 (cit. on pp. 28, 86).

[RW09] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*. Vol. 317. Springer Science & Business Media, 2009 (cit. on pp. 84, 340, 349, 352).

[Rot+00] A. Rotnitzky et al. “Likelihood-based inference with singular information matrix”. In: *Bernoulli* 6 (2000), pp. 243–284 (cit. on pp. 203, 218).

[RM11] J. Rousseau and K. Mengersen. “Asymptotic behaviour of the posterior distribution in overfitted mixture models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (2011), pp. 689–710 (cit. on pp. 155, 202).

[Rou84] P. J. Rousseeuw. “Least Median of Squares Regression”. In: *Journal of the American Statistical Association* 79 (1984), pp. 871–880 (cit. on p. 202).

[Rup88a] D. Ruppert. *Efficient estimators from a slowly convergent Robbins-Monro process*. Tech. rep. 781. Cornell University, 1988 (cit. on pp. 9, 14).

[Rup88b] David Ruppert. *Efficient estimations from a slowly convergent Robbins-Monro process*. Tech. rep. Cornell University Operations Research and Industrial Engineering, 1988 (cit. on pp. 84, 86).

[SLB17] Mark Schmidt, Nicolas Le Roux, and Francis Bach. “Minimizing finite sums with the stochastic average gradient”. In: *Mathematical Programming* 162.1-2 (2017), pp. 83–112 (cit. on p. 88).

[Sha+91] Ronald G Shaiko et al. “Pre-election political polling and the non-response bias issue”. In: *International Journal of Public Opinion Research* 3.1 (1991), pp. 86–99 (cit. on p. 217).

[SRR19a] Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. “Are sample means in multi-armed bandits positively or negatively biased?” In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019, pp. 7102–7111 (cit. on p. 264).

[SRR19b] Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. “On the bias, risk and consistency of sample means in multi-armed bandits”. In: *arXiv preprint arXiv:1902.00746* (2019) (cit. on p. 264).

[Sid+18a] Aaron Sidford et al. “Near-optimal time and sample complexities for solving discounted Markov decision process with a generative model”. In: *arXiv preprint arXiv:1806.01492* (2018) (cit. on pp. 50, 52, 56, 88).

[Sid+18b] Aaron Sidford et al. “Variance Reduced Value Iteration and Faster Algorithms for Solving Markov Decision Processes”. In: *ACM-SIAM Symposium on Discrete Algorithms*. Vol. 29. SIAM. 2018, pp. 770–787 (cit. on pp. 50, 52, 56).

[Sil+16] David Silver et al. “Mastering the game of Go with deep neural networks and tree search”. In: *Nature* 529.7587 (2016), pp. 484–489 (cit. on p. 49).

[SJ19] Max Simchowitz and Kevin Jamieson. “Non-Asymptotic Gap-Dependent Regret Bounds for Tabular MDPs”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2019, pp. 1153–1162 (cit. on p. 49).

[Ste56] Charles Stein. “Efficient nonparametric testing and estimation”. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. 1956, pp. 187–195 (cit. on pp. 11, 15).

[Ste02] M. Stephens. “Dealing with label switching in mixture models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62 (2002), pp. 795–809 (cit. on p. 155).

[Sto89] Nancy L Stokey. *Recursive Methods in Economic Dynamics*. Harvard University Press, 1989 (cit. on p. 84).

[SB18a] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. 2nd. Cambridge, MA: MIT Press, 2018 (cit. on p. 8).

[Sut88] Richard S Sutton. “Learning to predict by the methods of temporal differences”. In: *Machine learning* 3.1 (1988), pp. 9–44 (cit. on p. 9).

[SB18b] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018 (cit. on p. 51).

[Sze97] Csaba Szepesvári. “The Asymptotic Convergence-Rate of Q-learning”. In: *Advances in Neural Information Processing Systems*. Vol. 10. 1997, pp. 1064–1070 (cit. on p. 52).

[Sze98] Csaba Szepesvári. "The asymptotic convergence-rate of Q-learning". In: *Advances in Neural Information Processing Systems* (1998), pp. 1064–1070 (cit. on p. 150).

[Tad04] Vladislav B Tadic. "On the almost sure rate of convergence of linear stochastic approximation algorithms". In: *IEEE Transactions on Information Theory* 50.2 (2004), pp. 401–409 (cit. on p. 10).

[Tal96] Michel Talagrand. "New concentration inequalities in product spaces". In: *Inventiones Mathematicae* 126.3 (1996), pp. 505–563 (cit. on pp. 146, 147).

[Tal06] Michel Talagrand. *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*. Springer Science and Business Media, 2006 (cit. on pp. 146, 147).

[TV18] Y. S. Tan and R. Vershynin. "Phase Retrieval via Randomized Kaczmarz: Theoretical Guarantees". In: *Information and Inference: A journal of the IMA* (2018) (cit. on p. 226).

[Tes12] Gerald Teschl. *Ordinary differential equations and dynamical systems*. Vol. 140. American Mathematical Soc., 2012 (cit. on p. 84).

[Tho33] William R Thompson. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". In: *Biometrika* 25.3/4 (1933), pp. 285–294 (cit. on pp. 279, 307, 309, 310).

[Tob+17] Josh Tobin et al. "Domain randomization for transferring deep neural networks from simulation to the real world". In: *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*. IEEE. 2017, pp. 23–30 (cit. on p. 49).

[Tri+18] Nilesh Tripuraneni et al. "Averaging stochastic gradient descent on Riemannian manifolds". In: *Conference On Learning Theory*. PMLR. 2018, pp. 650–687 (cit. on p. 87).

[Tse90] Paul Tseng. "Solving H-horizon, stationary Markov decision problems in time proportional to $\log(H)$". In: *Operations Research Letters* 9.5 (1990), pp. 287–297 (cit. on p. 105).

[Tsi94] John N Tsitsiklis. "Asynchronous Stochastic Approximation and Q-Learning". In: *Machine Learning* 16.3 (1994), pp. 185–202 (cit. on p. 52).

[TV97] John N Tsitsiklis and Benjamin Van Roy. "Analysis of temporal-diffference learning with function approximation". In: *Advances in Neural Information Processing Systems*. 1997, pp. 1075–1081 (cit. on p. 9).

[TV99] John N Tsitsiklis and Benjamin Van Roy. "Average cost temporal-difference learning". In: *Automatica* 35.11 (1999), pp. 1799–1808 (cit. on p. 112).

[Tsy09] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. New York: Springer, 2009 (cit. on p. 38).

[TP74] Yakov Zalmanovich Tsypkin and Boris Teodorovich Polyak. "Attainable accuracy of adaptation algorithms". In: *Doklady Akademii Nauk*. Vol. 218. Russian Academy of Sciences. 1974, pp. 532–535 (cit. on p. 17).

[Tuy95] Hoang Tuy. "DC optimization: theory, methods and algorithms". In: *Handbook of global optimization*. Springer, 1995, pp. 149–216 (cit. on pp. 314, 315, 322).

[Vaa98a] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998 (cit. on p. 165).

[VW00] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York, NY, 2000 (cit. on p. 191).

[Vaa98b] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998 (cit. on p. 17).

[van00] Aad W van der Vaart. *Asymptotic Statistics*. Vol. 3. Cambridge University Press, 2000 (cit. on pp. 85, 86, 99).

[Van14] Ramon Van Handel. *Probability in high dimension*. Tech. rep. PRINCETON UNIV NJ, 2014 (cit. on p. 149).

[Ver] R. Vershynin. "Introduction to the non-asymptotic analysis of random matrices". In: *arXiv:1011.3027v7* () (cit. on p. 194).

[VBW15] Sofia S Villar, Jack Bowden, and James Wason. "Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges". In: *Statistical science: a review journal of the Institute of Mathematical Statistics* 30.2 (2015), p. 199 (cit. on p. 264).

[Wai+19] Hoi-To Wai et al. "Variance Reduced Policy Evaluation with Smooth Function Approximation". In: *Advances in Neural Information Processing Systems*. 2019, pp. 5776–5787 (cit. on p. 10).

[Wai19a] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge, UK: Cambridge University Press, 2019 (cit. on p. 313).

[Wai19b] Martin J Wainwright. *High-dimensional Statistics: A Non-asymptotic Viewpoint*. Cambridge University Press, 2019 (cit. on pp. 16, 38, 47, 146, 264, 291).

[Wai19c] Martin J Wainwright. "Stochastic approximation with cone-contractive operators: Sharper ℓ_∞-bounds for Q-learning". In: *arXiv preprint arXiv:1905.06265* (2019) (cit. on pp. 10, 43–45, 52, 57, 85, 87, 90).

[Wai19d] Martin J Wainwright. *Variance-reduced Q-learning is minimax optimal*. Tech. rep. arXiv preprint arXiv:1906.04697. 2019 (cit. on pp. 50, 52, 56, 58–60, 68).

[Wai19e] Martin J Wainwright. "Variance-reduced Q-learning is minimax optimal". In: *arXiv preprint arXiv:1906.04697* (2019) (cit. on pp. 10, 88, 90, 112, 150).

[Wai19f] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. DOI: 10.1017/9781108627771 (cit. on p. 74).

[WSU14] Shenlong Wang, Alex Schwing, and Raquel Urtasun. "Efficient inference of continuous Markov random fields with polynomial potentials". In: *Advances in neural information processing systems*. 2014, pp. 936–944 (cit. on pp. 321, 330, 332).

[Wan+15] Z. Wang et al. "High-Dimensional Expectation-Maximization Algorithm: Statistical Optimization and Asymptotic Normality". In: *Advances in Neural Information Processing Systems 28*. 2015 (cit. on p. 156).

[WLZ14] Zhaoran Wang, Han Liu, and Tong Zhang. "Optimal computational and statistical rates of convergence for sparse nonconvex learning problems". In: *Annals of statistics* 42.6 (2014), p. 2164 (cit. on p. 200).

[WD92a] Christopher JCH Watkins and Peter Dayan. "Q-learning". In: *Machine Learning* 8.3-4 (1992), pp. 279–292 (cit. on p. 52).

[WD92b] Christopher JCH Watkins and Peter Dayan. "Q-learning". In: *Machine Learning* 8.3-4 (1992), pp. 279–292 (cit. on pp. 112, 150).

[WCP18] Bo Wen, Xiaojun Chen, and Ting Kei Pong. "A proximal difference-of-convex algorithm with extrapolation". In: *Computational Optimization and Applications* 69.2 (2018), pp. 297–324 (cit. on pp. 314, 328).

[Whi58] John S White. "The limiting distribution of the serial correlation coefficient in the explosive case". In: *The Annals of Mathematical Statistics* (1958), pp. 1188–1197 (cit. on pp. 267, 281).

[Wu83] C. F. Jeff Wu. "On the Convergence Properties of the EM Algorithm". In: *Annals of Statistics* 11 (1983), pp. 95–103 (cit. on p. 155).

[WZ19] Yihong Wu and Harrison H Zhou. "Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations". In: *arXiv preprint arXiv:1908.10935* (2019) (cit. on p. 221).

[Xia+22] Eric Xia et al. “Instance-dependent confidence and early stopping for reinforcement learning”. In: *arXiv preprint* (2022) (cit. on p. 145).

[XHM16] J. Xu, D. Hsu, and A. Maleki. “Global analysis of Expectation Maximization for mixtures of two Gaussians”. In: *Advances in Neural Information Processing Systems 29*. 2016 (cit. on p. 156).

[XJ96] L. Xu and M. I. Jordan. “On convergence properties of the EM Algorithm for Gaussian mixtures”. In: *Neural Computation* 8 (1996), pp. 129–151 (cit. on p. 155).

[XQL13] Min Xu, Tao Qin, and Tie-Yan Liu. “Estimation bias in multi-armed bandit algorithms for search advertising”. In: *Advances in Neural Information Processing Systems*. Vol. 26. 2013, pp. 2400–2408 (cit. on p. 264).

[Xu+20] Tengyu Xu et al. “Reanalysis of Variance Reduced Temporal Difference Learning”. In: *arXiv preprint arXiv:2001.01898* (2020) (cit. on p. 10).

[XY17] Yangyang Xu and Wotao Yin. “A globally convergent algorithm for nonconvex optimization based on block coordinate update”. In: *Journal of Scientific Computing* 72.2 (2017), pp. 700–734 (cit. on p. 314).

[YYS17a] B. Yan, M. Yin, and P. Sarkar. “Convergence of Gradient EM on Multi-component Mixture of Gaussians”. In: *Advances in Neural Information Processing Systems 30*. 2017 (cit. on pp. 156, 157).

[YYS17b] Bowei Yan, Mingzhang Yin, and Purnamrita Sarkar. “Convergence of Gradient EM on Multi-component Mixture of Gaussians”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 6959–6969 (cit. on p. 313).

[YBW17] F. Yang, S. Balakrishnan, and M. Wainwright. “Statistical and Computational Guarantees for the Baum-Welch algorithm”. In: *Journal of Machine Learning Research* 18 (2017), pp. 1–53 (cit. on pp. 200, 202, 228).

[YC15a] X. Yi and C. Caramanis. “Regularized EM Algorithms: A Unified Framework and Statistical Guarantees”. In: *Advances in Neural Information Processing Systems 28*. 2015 (cit. on p. 156).

[YC15b] Xinyang Yi and Constantine Caramanis. “Regularized EM algorithms: A unified framework and statistical guarantees”. In: *Advances in Neural Information Processing Systems*. 2015, pp. 1567–1575 (cit. on pp. 200, 202).

[YB13] Huizhen Yu and Dimitri P Bertsekas. “Q-learning and policy iteration algorithms for stochastic shortest path problems”. In: *Annals of Operations Research* 208.1 (2013), pp. 95–132 (cit. on pp. 104, 105).

[YZ13] Xiao-Tong Yuan and Tong Zhang. “Truncated power method for sparse eigenvalue problems”. In: *Journal of Machine Learning Research* 14.Apr (2013), pp. 899–925 (cit. on p. 200).

[YR03] Alan L Yuille and Anand Rangarajan. “The concave-convex procedure”. In: *Neural computation* 15.4 (2003), pp. 915–936 (cit. on pp. 314, 315, 322, 323).

[ZB19] Andrea Zanette and Emma Brunskill. “Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds”. In: *arXiv preprint arXiv:1901.00210* (2019) (cit. on pp. 10, 49).

[ZKB19] Andrea Zanette, Mykel J Kochenderfer, and Emma Brunskill. “Almost Horizon-Free Structure-Aware Best Policy Identification with a Generative Model”. In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019, pp. 5625–5634 (cit. on p. 49).

[ZZ12] C.-H Zhang and T. Zhang. “A general theory of concave regularization for high-dimensional sparse estimation problems”. In: *Statistical Science* 27.4 (2012), pp. 576–593 (cit. on p. 200).

[Zha+17] Huishuai Zhang et al. “A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms”. In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 5164–5198 (cit. on p. 200).

[ZJM20] Kelly W Zhang, Lucas Janson, and Susan A Murphy. “Inference for Batched Bandits”. In: *arXiv preprint arXiv:2002.03217* (2020) (cit. on pp. 264, 266, 267, 278).

[ZJM] Kelly W Zhang, Lucas Janson, and Susan A Murphy. “Statistical Inference with M-Estimators on Adaptively Collected Data”. In: () (cit. on p. 264).

[Zha+99] Ruo Zhang et al. “Shape-from-shading: a survey”. In: *IEEE transactions on pattern analysis and machine intelligence* 21.8 (1999), pp. 690–706 (cit. on p. 331).

[ZZM21] Sheng Zhang, Zhe Zhang, and Siva Theja Maguluri. “Finite Sample Analysis of Average-Reward TD Learning and Q-Learning”. In: *Advances in Neural Information Processing Systems* 34 (2021) (cit. on p. 87).