Principled Statistical Approaches For Sampling and Inference in High Dimensions

by

Raaz Dwivedi

A dissertation submitted in partial satisfaction of the

requirements for the degree of

in

Engineering—Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Martin Wainwright, Co-chair
Professor Bin Yu, Co-chair
Professor David Aldous
Professor Peter Bartlett

Summer 2021

Principled Statistical Approaches For Sampling and Inference in High Dimensions

Abstract

Principled Statistical Approaches For Sampling and Inference in High Dimensions

by

Raaz Dwivedi

in Engineering—Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Martin Wainwright, Co-chair

Professor Bin Yu, Co-chair

The growth in the number of algorithms to identify patterns in modern large-scale datasets has introduced a new dilemma for practitioners: How does one choose between the numerous methods? In supervised machine learning, accuracy on a hold-out dataset is the flagship for choice making. This dissertation presents research that can provide principled guidance for making choices in three popular settings where such a flagship measure is not readily available. (I) Convergence of Markov chain Monte Carlo sampling algorithms, used commonly in Bayesian inference, Monte Carlo integration, and stochastic simulation: We provide explicit non-asymptotic guarantees for state-of-the-art sampling algorithms in high dimensions that can help the user pick a sampling method and the number of iterations based on the computational budget at hand. (II) Statistical-computational challenges with mixture model estimation used commonly with heterogeneous data: We provide non-asymptotic guarantees with Expectation-Maximization for parameter estimation when the number of components is not known, and characterize the number of samples and iterations needed for the desired accuracy, that can inform the user of the potential two-edged price when dealing with noisy data in high dimensions. (III) Reliable estimation of heterogeneous treatment effects (HTE) in causal inference, crucial for decision making in medicine and public policy: We introduce a data-driven methodology StaDISC that is useful for validating commonly used models for estimating HTE, and for discovering interpretable and stable subgroups with HTE using calibration. While we illustrate its usefulness in precision medicine, we believe the methodology to be of general interest in randomized experiments.

To Mom, Rama Dwivedi,

To Dad, Sabha Kant Dwivedi,

To Didi, Resham Dwivedi,

And *To Nana*, Hiramani Sharma,

For Their Unfaltering Support, Unconditional Love & Unwavering Belief in me.

Contents

List of Figures

List of Tables

Acknowledgments

Time travels fast, and certainly, the last six years of my Ph.D. have been yet another example. I owe a sincere thanks to many many amazing people for this wonderful journey, that I acknowledge here. I would like to apologize in advance to the people whose support I really appreciate but have accidentally missed out here. I must first thank two super smart and amazing people Martin Wainwright and Bin Yu, who have been the most wonderful advisors and mentors to me, and through uncountable encouragements and challenges, have not only shaped me into the researcher that I am today, but also made me a more responsible, and socially aware human being.

I first met Martin during his visit to India when I was deciding about graduate school, and the first impression I had was of Martin's funny side. I got to know his true self later in my first year at Berkeley. What followed was a truly blissful relationship, minus his non-responsiveness over emails. All my meetings and correspondences with him have been filled with almost equal amounts of amazement and amusement due to his profound mathematical sharpness, and an absolutely great sense of humor. His mentoring over the years about interesting theoretical problems, and how to find, and solve them has been invaluable in making me a better researcher and thinker. His course on high dimensional statistics was my first real introduction to modern statistics, and his book on it made me fall in love with the subject. I must also thank him for the numerous hours invested in making me a better writer and presenter. Besides, I will surely miss Martin's interest in my love life, his marriage pranks, and the super-fun times with his lovely and cute kids—"Hana-Mina-Kento Power".

I first met Bin after arriving at Berkeley, and in the very first meeting, her passion for solving real problems, emphasis on human well-being, social justice, and the wholesome development of her students, left a deep impression on me. Her immense curiosity for unresolved mysteries in scientific problems, the never-ending drive to unravel them, and the incredibly sharp intellect to develop connections in very unrelated areas have been very inspiring, and have often provided me with the tools, and enthusiasm needed to overcome many barriers in my research. It took me a while to fully understand the acute wisdom that Bin possesses with applied statistics, but I am really fortunate to have received constant encouragement from her to get out of my comfort zone and embrace interdisciplinary research. Working on the applied projects with her, especially the COVID-19 project, and witnessing firsthand the beauty of real problem solving guided by the inspiring breadth of her expertise, has made me fall in love with applied statistics. Bin's mentoring for finding the most impactful theoretical problems, how to formulate and solve interesting applied problems in collaboration with domain experts, and her research philosophy has helped me cultivate one of my own towards becoming a wholesome researcher. I must also thank her for the critical feedback in improving my writing and communication skills. Finally, I am also really grateful to have a truly amazing and an utmost caring friend in Bin, whom I can approach anytime for advice about career, life, health, or otherwise.

I also thank my thesis committee members David Aldous and Peter Bartlett for serving on my committee and sharing their perspectives on what defines an interesting research problem. It was a delight learning while collaborating with some of the most brilliant minds I have met, Lester Mackey, David Madigan, Michael Jordan, and Kevin Horgan. I must specially thank Lester for being the most amazing host at Microsoft Research, and since being a wonderful collaborator, and for being incredibly generous with his words of praise and encouragement, and time spent patiently teaching me about kernels, and code large-scale experiments. I also owe a considerable portion of my technical knowledge to the very stimulating group meetings of Kannan Ramchandran, Peter Bartlett, and Michael Jordan, and I thank them for keeping their group meetings easily accessible to all the students. I am also thankful to have learned from Philip Stark, his outlook for good applied research, and what harm bad statistics can create in the real world. I was fortunate to have assisted Anant Sahai and Jennifer Listgarten in teaching the machine learning class and learn about how to effectively teach meaningful content even to very big classes.

A place like UC Berkeley is naturally full of brilliant minds all around, and I was lucky to be a part of this crowd where amazing ideas are created and exchanged every minute. I have had the fortune of closely working with several super-sharp and smart people at Berkeley. I must start by thanking Yuansi Chen—One of the sharpest, persevering, and responsible workaholics that I have met. Besides his great technical capabilities, I perhaps admire the most his humility, and the sense of humor, which makes him a very pleasant person to work with. I should thank him for a great deal of mathematics, and statistics, that I learned while working with him—part I of this thesis presents our joint work on MCMC sampling. Over the years our collegiality evolved into a beautiful and caring friendship, that I am very proud of. I must also thank Yuansi's wife, Pan, for inviting me for Chinese—yet vegetarian— three-course gala dinner which included nine most yummy dishes. There are still days when I dream about that once-in-a-lifetime dinner. I was also fortunate to work with Nhat Ho, another workaholic and sharp thinker, and the person who knew it all about mixture models. His superb organization skills and tireless efforts were crucial in pushing our joint adventures about the black holes of statistics, mixture models, to new dimensions. I am also really grateful to him for being a caring friend, and for spending long hours helping me with my job market-related questionnaires. Part II of this thesis presents some of my joint work with Nhat, and another amazing person, Koulik Khamaru. I admire Koulik's magical ability to pull out almost always the right inequality needed in some very delicate proofs. It was a very enjoyable experience working together as a team with them and making several memories of our races on the whiteboard in Cory Hall as to who gets to the correct proof first. Besides, the trip to Niagara Falls also made many beautiful memories, and I must thank Koulik for the same. I am also grateful to Yan Shuo Tan for our numerous long discussions about various research problems, the research trends in statistics, and for teaching me a great deal about causal inference, being an excellent collaborator, and a dear friend who can bear many of my academics related rants—part III of this thesis presents some of our joint work on a data-driven methodology for causal inference. I also thank Chandan Singh for sharing his sharp thinking about good applied work, and for being an excellent and responsible collaborator on our joint adventure into the crowded landscape of deep learning theory (our joint work is not included in this thesis). I am also grateful to him for teaching me coding and organization skills, and his family for hosting me very warmly at their home and very generously satisfying my craving for delicious homemade Indian food.

I was also fortunate to have learned a great deal from many other wonderful friends at Berkeley. I am really grateful to Orhan Ocal for being a caring friend, helping me improve my latex, coding, dancing, and trading skills, for the beautiful series of course projects by team RaHan, and for sharing the curse that these projects never made to a conference. I thank Yuting Wei, for teaching me several topics in statistics, gossiping and laughing together about our advisor, and for the fun trip to Oberwolfach, and Prague. I thank Fanny Yang for our joint short yet very insightful adventure into kernels, for pushing me into a healthier lifestyle, running, and some beautiful bike rides together. I thank Ashwin Pananjady and Rebecca Barter for making me a better communicator and speaker via valuable feedback on many of the practice talks. I am grateful to Aaditya Ramdas for being an amazing mentor in the early days of my Ph.D. and sharing his valuable philosophy about career and life.

Being a member of the BLISS and BAIR labs, Wainwright, and Yu groups has their privileges. I am thankful to Wooseok Ha, Yu Wang, Merle Behr, Christine Kuang, Sujayam Saha, Xiao Li, Briton Park, Mian Wei, Tiffany Tang, Robert Netzorg, Nikhil Ghosh, Omer Ronen, Xiao Li, Nick Altieri, Simon Walter from the Yu Group for the amazing time and valuable memories that we made throughout group meetings, lunches, dinners, picnics, and pot-lucks. I am also grateful to Billy Fang, Wenlong Mou, Jianbo Chen, Joseph Borja, Nihar Shah, Max Rabionovich, Horia Mania, for the several useful discussions we had over the years. I thank the physical education department at Berkeley, which allowed me to meet Toni Mar and Lucy Yoo—the people responsible for adding fitness and dance routines to my everyday life. I was lucky to have found many great friends who I must thank for sharing their knowledge and showering kindness, and support over the years: Sang Min Han for the late-night hangouts in early years, Vidya Muthukumar for always being very thorough in her feedback, Aviva Bulow for the Yellowstone trip, and her family for inviting me for Shabbat, Vipul Gupta for organizing fun events, including Frisbee evenings. I also want to thank for the blissful company of Avishek Ghosh, Armin Askari, Ludwig Schmidt, Efe Aras, Kuan-Yun Lee, Vignesh Subramanian, Banghua Zhou, Nived Rajaraman, Chinmay Maheshwari, Aditi Raghunathan, Kabir Chandrasekher, Alvin Wan, Reinhard Heckel, Ilan Shomorony, Laura Brink, Cong Ma, Kang Wook Lee, Mert Pilanci, Ramtin Pedarsani, Rashmi Vinayak, Vasuki Narasimha Swamy, Reza Abbasi-Asl, Payam Delgosha, Dong Yin, Leah Dickstein, Tavor Baharav, Dimitris Papailiopoulos, Leila Wehbe, Yun Yang, Nhat Ho, Swanand Kadhe, for the valuable time we spent together. I have been lucky to develop amazing friendships with Kush Bhatia, Ashish Kumar, and Yeshwanth Cherapanamjeri, I have cherished our philosophical discussion about life, career, and money, and our fitness training sessions over the years. Special thanks to, Soham Phade, and Sri Krishna Vadlamani, for being the amazing roommates and the close friendship that we have built together. I enjoyed doing many of our "first" things together, especially with Soham, including the first visit to People's park. A big shout out to the very friendly EECS and statistics staff: Shirley Salanio is one of the most amazing people, and her reliable promptness and humor made me happy numerous times over the years. I must also thank La Shana, Keyla Gomez, and Mary Melin for their incredible help in the Statistics department. I would also like to thank Toni Mar for encouraging me to focus more on my physical health and making physical training so much more fun. I was very grateful for taking dance classes from Lucy Yoo and adding Salsa and Bachata skills (although poor) to my Ph.D. armor. I am fortunate for having found in Jane Yu, my best friend with whom I can not wait to share my happiness and sorrow, success and failure, together. I can not thank her enough for adding cheer and joy to my life daily, teaching me the importance of fitness, and sleep, for sacrificing her time bearing my annoying habits, for broadening my philosophical and research outlook, which enabled me to not only become a better researcher but also a better human being.

I owe my undergraduate advisor at IIT Bombay, Vivek Borkar, a sincere thanks for convincing me to join a Ph.D. program and then to join UC Berkeley. I thank him, and Gopala Srinivasan from IIT Bombay for teaching me a range of topics in mathematics during my undergrad program, and that has undoubtedly been the backbone of the researcher I am today. I am also grateful to Gopala for not only being an amazing teacher and mentor but also for the kind and caring friendship that we have cultivated since my days at IITB. I am really thankful to have been constantly mentored by Prakash Singh, my physics teacher in high school, who was not only the most important pillar during my IIT-JEE preparation days, but also the first person to really challenge and redefine my ambitions in my life, and thereby setting me on the path of never-ending academic curiosity. I also thank my school teachers Ranjan Singh, Shivani Saxena, Amin Akhtar, and Ram Yash Sharma for their teachings and kindness throughout. I am also grateful to my friends from IIT, and high schools, Scholars' Home, and Sagar Public School for providing me with happy and fun-filled breaks from research whenever I visited India.

Finally, I must thank my family for their support throughout the ups and downs of my life. I am really grateful to my grandparents, especially my *Nana*, for showering words of encouragement timely. I also thank my cousins, Nivedita Shukla and Saurabh Mishra for adding joy to my life, and being the pillar of support whenever I needed them. Finally, I owe all my achievements to my mom Rama Dwivedi and dad Sabha Kant Dwivedi. My gratitude to them for their immeasurable sacrifices can not be expressed in words. I have been extremely fortunate to receive their unconditional love and support, and I thank them for pushing me to become a better person over time. I always go back to their teachings and values to overcome various psychological barriers that are omnipresent as one transitions to becoming an independent and responsible person. I am extremely lucky to have my sister, Resham Dwivedi, as the torchbearer of my life. Her unparalleled willpower to overcome the struggles, and dealing with them head-on has made my life relatively a cakewalk. I can not thank her enough for her incredible spirit towards taking on new challenges, which in turn inspires me to push the horizons of my own ambitions. My life would not be filled with so much happiness, rich and colorful experiences, without my sister, mom, and dad, who are there for me, *always*.

Chapter 1

Introduction

Recent decades have seen a surge in statistics, and machine learning research. The number of learning algorithms available for solving various data-driven problems has grown many folds. Such rapid advancements often present a practitioner with the *choice dilemma*: *Which of the many models or algorithms shall I use?* A principled choice making, necessary for reliable learning, should either have a valid theoretical backing or be guided by sufficient empirical evidence. A poor choice might waste computational resources, lead to unsatisfactory answers, and lack of theoretical or empirical checks might lead the scientists to draw false conclusions from the limited data at hand.

In the case of supervised machine learning, the user is typically blessed with the notion of *accuracy on a hold-out dataset*. Its simplicity and effectiveness in practice, especially when combined with cross-validation, has made it perhaps the most popular criterion for choosing from, and tuning various models and algorithms. However, several areas, e.g., unsupervised learning, statistical inference, and causal inference, suffer from the lack of such a simple yet powerful (or commonly accepted) tool for choice navigation.

In part I, we provide theoretical guarantees for various Markov chain Monte Carlo (MCMC) algorithms. MCMC methods serve as the numerical engine of Bayesian inference, and Monte Carlo integration, and are most commonly used for drawing random samples from a given target probability distribution. Nevertheless, they are notorious for slow convergence, and poor theoretical understanding especially in high dimensions. While there is rich literature about asymptotic convergence, and abundant empirical wisdom, fundamental results on non-asymptotic rate of convergence can provide insight into how the different methods compare for the same task, and under what settings, they are *provably* provide reliable estimates.

Part II deals with the reverse aspect of statistical problems: learning about target distributions given draws of random samples. We study a class of challenging mixture models, which are commonly used when the data has lot of heterogeneity, and characterize the requirements of (a) the sample size, and (b) the computational budget, to estimate the unknown parameters to a desired accuracy. We establish guarantees for the method of choice in settings, Expectation-Maximization (EM). While EM is known to have favorable performance with low noise in the data, we study several models with high noise, and provide several results which *provably* establish the slow down of EM on both the sample size required, and the number of iterations needed for providing a reliable estimate. Our results can provide insight on how to select the number of mixture components when doing unsupervised learning with noisy data.

Finally, in part III, we tackle the problem of reliable estimation of heterogeneous treatment effects in causal studies. Causal inference suffers from the fundamental problem of missing data, since only some of the potential outcomes for each unit in the sample are observed—thus validating individual-level treatment effect is impossible. We introduce a data-driven methodology StaDISC, designed for reliable heterogeneity treatment effect discovery and estimation at subgroup-level. StaDISC is immediately relevant for informing about the performance of different from conditional average treatment effect (CATE) models as it provides calibration-based *predictive checks* to select. Furthermore, it automatically *discovers* interpretable and stable subgroups with heterogeneous treatment effects (HTE). We illustrate the usefulness StaDISC in the context of precision medicine via re-analysis of two randomized clinical trials, and discovering subgroups that are disproportionately affected by the drug under investigation.

1.1 Part I: Non-asymptotic mixing time analysis

Random sampling, i.e., drawing random samples from a probability distribution, is a crucial computational challenge common to many disciplines, with applications in machine learning, statistics, probability, operations research, and other areas involving stochastic modeling [95, 28, 209, 107]. In statistics, these methods are useful for both estimation of unknown quantities and their inference. Markov chain Monte Carlo are the method of choice for such tasks. Given a target distribution $\Pi^\star$, an MCMC method broadly comprises of two steps: First, setting up a Markov chain whose stationary distribution is the target distribution or a good approximation of it, and second, simulating the chain for several number of steps so that the chain has mixed to generate an approximate sample from $\Pi^\star$.

Many algorithms have been proposed for sampling from probability distributions with a (general) density on a continuous state space. Generally speaking, these methods can be categorized in two broad categories: *zeroth-order methods* and *first-order methods*. On one hand, a zeroth-order method is based on querying the density of the distribution (up to a proportionality constant) at a point in each iteration, and popular examples include random walk, ball walk, hit-and-run. By contrast, a first-order method makes use of additional gradient information about the density, and the most common examples include Langevin algorithms, and Hamiltonian Monte Carlo. Several natural questions arise given this broad distinction: When does a first order method outperform a zeroth-order method? Can the gains of one method be characterized in a non-asymptotic sense?

Furthermore, there is a broad class of Markov chains that make use of a two-step simulation for each iteration: First, draw a candidate state using a proposal distribution, and then perform an accept-reject step also known as Metropolis-Hastings correction. The latter step ensures that the stationary distribution of the chain is the target distribution. Such a design is clearly advantageous in the asymptotic limit, since the chain is asymptotically unbiased and converges to the target distribution under mild conditions. But, does it really benefit in finite number of iterations? Can we simulate an unadjusted chain and get a better finite time performance?

Part I of this thesis answers these questions for a range of algorithms by proving rigorous non-asymptotic mixing time guarantees. In particular, for several popular algorithms, we derive the number of iterations needed for the algorithm. Our results are user-friendly as they provide explicit choices of hyper-parameters, and establish the mixing time guarantee as an explicit function of the problem dimension, target distribution's smoothness and curvature, and the desired target accuracy. We start with a background and setting up notation for MCMC in Chapter 2, and then discuss the mixing time bounds for random walk, and Langevin algorithms in Chapter 3, where we also show that the accept-reject step can provide significant speed-up in mixing time. In Chapter 4, we establish general machinery for proving mixing time bounds that significantly improve mixing time guarantees for a class of Markov chains when the starting distribution is far from the target. Chapter 5 provides a thorough theoretical investigation of Hamiltonian Monte Carlo, the default sampling algorithm in many softwares, and often termed as the state-of-the-art sampling method. Finally, in Chapter 6, we leverage the advancements in the interior point methods to design two new sampling algorithms, Vaidya and John walks, that achieve state-of-the-art mixing times for uniform sampling on convex polytopes. This part is based on joint work with Yuansi Chen, Martin Wainwright and Bin Yu [46, 47, 78, 48].

1.2 Part II: EM and over-specified Gaussian mixtures

The growth in the size and scope of modern data sets has presented the field of statistics with a number of challenges, one of them being how to deal with various forms of heterogeneity. Mixture models provide a principled approach to modeling heterogeneous collections of data (that are usually assumed i.i.d.). In particular, Gaussian mixture models [201] have been used widely to model heterogeneous data in many applications arising from physical and the biological sciences. However, estimating the parameters of mixture models is a challenging task, due to the non-convexity of the log likelihood function. As shown by classical work, the maximum likelihood estimate (MLE) often has good properties for mixture models, but its computation can be non-trivial. One of the most popular algorithms used to compute the MLE (approximately) is the expectation maximization (EM) algorithm. Although EM is widely used in practice, it does not always converge to the MLE, and its convergence rate can vary as a function of the problem. Classical results provide guarantees about the convergence rates of EM to local maxima [68, 251]. In the specific setting of Gaussian mixtures, population EM (idealized EM with infinite samples) was shown to have a range of behavior from super-linear convergence to slow convergence like a first-order method depending on the overlap between the mixtures [172, 254]. More recently, there has been a renewed interest in providing explicit and non-asymptotic guarantees on the convergence of EM [256, 248, 132, 253, 64, 255, 105, 35]; as a consequence, our understanding of EM in such cases is now relatively mature.

A shared assumption common to this body of past work is that either the true distribution of each subpopulations is known, or that the number of components is exactly known; in practice, both of these conditions are often violated. In several scenarios, the data has a large number of sub-populations and the mixture components in the data may not be well-separated. In such settings, estimating the true number of components may be difficult, so that one may end up fitting a mixture model with a number of components larger than that present in the data. Such mixture fits, referred to as *over-specified mixture distributions*, are commonly used by practitioners in order to deal with uncertainty in the number of components in the data [222, 108]. However, a deficiency of such models is that they are *singular*, meaning that their Fisher information matrices are degenerate. 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.

Part II tries to bridge this gap by providing several fundamental results regarding the behavior of EM when used to fit over-specified mixture models. We provide a sharp and non-asymptotic guarantees of EM with several over-specified mixture models on both statistical and computational fronts. Our results show that over-specification costs the user on two ends: Compared to the well specified models, there is not only a significant degradation in the statistical accuracy with a given number of samples, but also EM requires significantly many more iterations to converge to its final estimate. In Chapter 7, we analyze over-specified Gaussian mixture models with unknown mean and known covariance, while Chapter 8 deals with models when both mean and covariance of the model is unknown. This part is based on joint work with Nhat Ho, Koulik Khamaru, Michael Jordan, Martin Wainwright and Bin Yu [82, 81].

1.3 Part III: Reliable subgroup discovery with HTE

Understanding heterogeneous treatment effects (HTE) is at the cutting edge of causal inference and the past decade in particular has witnessed a wave of innovation in the modeling and estimation of them. Underlying the hot topic of *precision medicine* [56] is a realization that how a patient responds to a particular drug or treatment depends on the patient's genetics, lifestyle and environment, and that consequently, accounting for these differences will allow doctors to deliver better and more targeted care. Moreover, this emphasis on understanding and exploiting heterogeneity is not unique to the biomedical sciences, and has also arisen in economics [125], political sciences [96, 89], online advertising [180], and many other fields [89].

HTE is often estimated using conditional average treatment effect (CATE) models, and the last several years have seen numerous new methods proposed for CATE estimation [125, 96, 89, 34, 92, 234, 21]. With such a diverse range of estimators, most of which come with hyperparameters, model choice becomes a primary concern, and missing potential outcomes renders empirical validation of CATE models difficult since there is no direct analog of validation accuracy at individual-level. On the other hand, the existing theoretical consistency results require strong hard-to-check assumptions. Furthermore, it is well-known now that in supervised learning, no single machine learning method universally dominates all others, but instead, the inductive bias of each algorithm lends it a competitive advantage on certain classes of problems [198]. We expect the same to hold for CATE estimation, which therefore calls for data-driven model selection.

This part of the dissertation tackles these challenges with CATE estimation using a data-driven methodology. In Chapter 9 showcasing the proposed research via re-analysis of two randomized clinical trials using 18 popular CATE models. Building on the PCS framework of Yu and Kumbier [259], we mitigate the missing information problem by devising calibration-based pseduo-R^2 scores for checking the CATE model fit on the data. We often find that CATE models indeed have poor global fit, but can be locally well-calibrated, a conclusion that is stable to various model and data perturbations—thereby a reliable one to draw according to the stability principle. Overall, building on the recent CATE literature and the PCS framework, we develop a new methodology, which we call Stable Discovery of Interpretable Subgroups via Calibration (StaDISC) that we expect to be of general interest for discovering subgroups with disproportionate treatment effect compared to the average. This part is based on joint work with Yan Shuo Tan, Briton Park, Mian Wei, Kevin Horgan, David Madigan, and Bin Yu [84].

1.4 Content not included in this thesis

We now briefly summarize few relevant papers that are not included in this thesis.

Related to Part I: Monte Carlo samples typically provide a slow rate of $n^{-1/2}$ for estimating function integrals via sample mean of n samples. In several settings, it is desirable to obtain a *faster than Monte Carlo* rate. In joint work, with Ohad Feldheim, Ori Gurel-Gurevich, and Aaditya Ramdas [79], we design a computationally efficient algorithm that provides a scaling of n^{-1} up to logarithmic factors for estimating integrals with respect to the uniform distribution on the unit cube. In another work, joint with Lester Mackey [83], we provide similar guarantees for functions in reproducing Hilbert spaces and wide range of distributions with support on $\mathbb{R}^d$.

Related to Part II: In other joint work with Nhat Ho, Koulik Khamaru, Michael Jordan, Martin Wainwright, and Bin Yu, we characterize the performance of EM when less than true number of components are fitted (underspecified settings) [80] that complement the analysis of EM with overspecified settings in Part II. Furthermore, we generalize the localization proof technique, to develop a general theory for studying tradeoffs between stability and convergence rate of iteration methods. We then apply the general theory to obtain a range of statistical and computational surprises for many algorithms other than EM, e.g., gradient descent, and Newton's method, for several statistical settings with noisy data [111].

About COVID-19 forecasting: In other joint work with Bin Yu, and Yu Group, we worked on curating a COVID-19 repository, and making county-level predictions about COVID severity in the US [5]. This undertaking was a large collaboration between Yu Group, and a non-profit organization response4life `response4life.org`, aimed at providing PPE support to those who needed it the most in the hours of crisis. The paper [5] summarizes some the technical contributions of this large project, ranging from how the repository was curated and organized, and how the many forecasting models were built, and ensembled. We provided a thorough validation of our models, that showed that our predictions were fairly accurate for 7-14 days horizon, which was crucial for our end task of PPE allocation. Interactive visualization of our predictions, historical performance, and several other features of our models were provided on a daily basis from April 2020, until March 2021, on `covidseverity.com`.

Part I

Theory of High-dimensional Random Sampling

Chapter 2

Background on Markov Chain Monte Carlo

Sampling procedures are the workhorse in Bayesian statistics, used for exploring posterior distributions, obtaining credible intervals, and solving inverse problems. Under the frequentist framework, samples drawn from a suitable distribution can form confidence intervals for a point estimate, such as those obtained by maximum likelihood. Estimating the mean, posterior mean in a Bayesian setting, expectations of desired quantities, probabilities of rare events and volumes of particular sets are settings in which Monte Carlo estimates are commonly used. Let us motivate via a concrete example.

Consider a distribution $\Pi^\star$ which admits a density $\pi^\star : \mathcal{X} \to \mathbb{R}_+$, specified explicitly up to a normalization constant as follows

$$\pi^\star(x) \propto e^{-f(x)}. \tag{2.1}$$

A standard computational task is to estimate the expectation of some function $g : \mathcal{X} \to \mathbb{R}$—that is, to approximate $\Pi^\star(g) = \mathbb{E}_{\pi^\star}[g(X)] = \int_{\mathcal{X}} g(x)\pi^\star(x)dx$. In general, analytical computation of this integral is infeasible. In high dimensions, numerical integration is not feasible either, due to the well-known curse of dimensionality.

A Monte Carlo approximation to $\Pi^\star(g)$ is based on access to a sampling algorithm that can generate i.i.d. random variables $Z_i \sim \pi^\star$ for $i = 1, \ldots, N$. Given such samples, the random variable $\widehat{\Pi^\star}(g) \coloneqq \frac{1}{N}\sum_{i=1}^N g(Z_i)$ is an unbiased estimate of the quantity $\Pi^\star(g)$, and has its variance proportional to $1/N$. The challenge of implementing such a method is drawing the i.i.d. samples Z_i. If $\pi^\star$ has a complicated form and the dimension d is large, it is difficult to generate i.i.d. samples from $\pi^\star$. For example, rejection sampling [98], which works well in low dimensions, fails due to the curse of dimensionality. In such settings and more generally, one turns to the class of Markov Chain Monte Carlo (MCMC) methods.

The origin of MCMC methods can be dated back as early as the seminal work of Metropolis et al. [177], and recent decades have seen tremendous empirical success with these methods, including more recent applications in simulation-based methods for reinforcement learning, and in image synthesis in computer vision, among other areas; for instance, see the handbook [30] and references therein. In a broad sense, these methods involve two steps. First, we construct a Markov chain, that is relatively easy to simulate, whose stationary distribution is either equal to the target distribution or close to it in a suitable metric. Given this chain, the second step is to draw samples by simulating the chain for a certain number of steps. An advantage of MCMC algorithms is that they only require knowledge of the target density up to a proportionality constant. Two natural questions that arise are: (i) how do we design an easy to simulate Markov chain; and (ii) how many steps will the Markov chain take to converge close enough to the stationary distribution? Over the years, these questions have been the subject of considerable research; for instance, see the reviews [235, 227, 213] and references therein. Nevertheless, a thorough theoretical understanding of MCMC algorithms used in practice is far from complete, and the last several years have seen a renewed interest in the non-asymptotic *mixing time* analysis of these methods, meaning that deriving the number of iterations—as a function of the error tolerance δ, problem dimension d and other parameters—for a given MCMC algorithm to arrive at a distribution within distance δ of the target.

We now provide a brief background on some terminologies about Markov chains in Section 2.1, and the general recipe of constructing MCMC algorithms in Section 2.2.

2.1 Basics of Markov Chain and Mixing

Here we consider the task of drawing random samples from a *target distribution* $\Pi^\star$ with its density denoted by $\Pi^\star$, via setting up of an irreducible and aperiodic discrete-time Markov chain whose stationary distribution is equal to or close to the target distribution $\Pi^\star$ in certain metric, e.g., total variation (TV) norm. To obtain a δ-accurate sample, one simulates the Markov chain for a certain number of steps k which is determined by a mixing time analysis. Going forward, we assume familiarity of the reader with a basic background in Markov chains, and refer them to the book [179] for a formal introduction or sections 1 and 2 of the papers [167, 242] for a quick and gentle introduction to the basics of Markov chains. Here we collect some basic definitions related to Markov chains.

In this work, we work with *time-homogeneous*Markov chains defined on a measurable state space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ with a transition kernel $\Theta : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to \mathbb{R}_+$, by definition, the transition kernel satisfies the following properties:

$$\Theta(x, dy) \geq 0, \quad \text{for all } x \in \mathcal{X}, \quad \text{and} \quad \int_{y\in\mathcal{X}} \Theta(x, dy) dy = 1 \quad \text{for all } x \in \mathcal{X}. \tag{2.2}$$

The k-step transition kernel Θ^k is defined recursively as $\Theta^{k+1}(x, dy) = \int_{z\in\mathcal{X}} \Theta^k(x, dz)\Theta(z, dy)dz$. The Markov chain is *irreducible* means that for all $x, y \in \mathcal{X}$, there is a natural number $k > 0$ such that $\Theta^k(x, dy) > 0$. We say that a Markov chain satisfies the *detailed balance condition* if

$$\pi^\star(x)\Theta(x, dy)dx = \pi^\star(y)\Theta(y, dx)dy \quad \text{for all } x, y \in \mathcal{X}. \tag{2.3}$$

Such a Markov chain is also called *reversible.* Finally, we say that a probability measure $\Pi^\star$ with density $\pi^\star$ on $\mathcal{X}$ is *stationary* (or *invariant*) for a Markov chain with the transition kernel Θ if

$$\int_{x\in\mathcal{X}} \pi^\star(x)\Theta(y, dx) = \pi^\star(y) \quad \text{for all } y \in \mathcal{X}.$$

Transition operator: We use $\mathcal{T}$ to denote the transition operator of the Markov chain on the space of probability measures with state space $\mathcal{X}$. In simple words, given a distribution μ_0 on the current state of the Markov chain, $\mathcal{T}(\mu_0)$ denotes the distribution of the next state of the chain. Mathematically, we have

$$\mathcal{T}(\mu_0)(A) = \int_{\mathcal{X}} \Theta(x, A)\mu_0(x)dx, \tag{2.4}$$

for any $A \in \mathcal{B}(\mathcal{X})$. In an analogous fashion, $\mathcal{T}^k$ stands for the k-step transition operator. We use $\mathcal{T}_x$ as the shorthand for $\mathcal{T}(\delta_x)$, the *transition distribution at x*; here δ_x denotes the Dirac delta distribution at $x \in \mathcal{X}$. Note that by definition $\mathcal{T}_x = \Theta(x, \cdot)$.

In order to quantify the convergence of the Markov chain, we study the mixing time for a class of distances denoted by $\mathcal{L}_\mathfrak{p}$ for $\mathfrak{p} \geq 1$. Letting Π be a distribution with density π, its $\mathcal{L}_\mathfrak{p}$-divergence with respect to the distribution $\Pi^\star$ with positive density $\pi^\star$ is defined as

$$d_\mathfrak{p}(\Pi, \Pi^\star) := \left(\int_{\mathcal{X}} \left|\frac{\pi(x)}{\pi^\star(x)} - 1\right|^\mathfrak{p} \pi^\star(x)dx\right)^{\frac{1}{\mathfrak{p}}}. \tag{2.5a}$$

Note that for $\mathfrak{p} = 2$, we recover the χ^2-divergence. For $\mathfrak{p} = 1$, the distance $d_1(\Pi, \Pi^\star)$ represents two times the total variation (TV) distance $d_{\mathrm{TV}}\big(\Pi, \Pi^\star\big)$ between Π and $\Pi^\star$:

$$d_{\mathrm{TV}}\big(\Pi, \Pi^\star\big) := \sup_{S\in\mathcal{B}(\mathcal{X})} |\Pi(S) - \Pi^\star(S)| = \frac{1}{2}\int_{\mathcal{X}} |\pi(x) - \pi^\star x|\, dx.$$

For clarity, we continue to use $d_{\mathrm{TV}}\big(\Pi, \Pi^\star\big)$ to denote the total variation distance.

Definition 2.1 (Mixing time of a Markov chain). *For an error tolerance $\delta > 0$, and a Markov chain with initial distribution μ_0, transition operator $\mathcal{T}$ and a target distribution $\Pi^\star$ with density $\pi^\star$, its $\mathcal{L}_\mathfrak{p}$ and TV mixing time with respect to $\Pi^\star$ are defined as follows:*

$$\tau_\mathfrak{p}(\delta; (\mu_0, \Pi^\star)) := \inf\left\{k \in \mathbb{N} \mid d_\mathfrak{p}\left(\mathcal{T}^k(\mu_0), \Pi^\star\right) \leq \delta\right\}, \quad \textit{and} \tag{2.5b}$$

$$\tau_{\mathrm{TV}}(\delta; (\mu_0, \Pi^\star)) := \inf\left\{k \in \mathbb{N} \mid d_{\mathrm{TV}}\big(\mathcal{T}^k(\mu_0), \Pi^\star\big) \leq \delta\right\}. \tag{2.5c}$$

In simple words, the δ-$\mathcal{L}_\mathfrak{p}$ (TV) mixing time denotes the minimum number of steps that the chain takes to reach within δ-$\mathcal{L}_\mathfrak{p}$ (TV) distance to the target distribution, given that it starts with distribution μ_0. Going forward, when $\Pi^\star$ is clear from the context (generally the stationary distribution of the Markov chain), we often use the simplified notations:

$$\tau_{\mathfrak{p}}(\delta; \mu_0) := \tau_{\mathfrak{p}}(\delta; (\mu_0, \Pi^\star)), \quad \text{and} \quad \tau_{\mathrm{TV}}(\delta; \mu_0) := \tau_{\mathrm{TV}}(\delta; (\mu_0, \Pi^\star)).$$

We note that since distance $d_{\mathfrak{p}}(Q, \Pi^\star)$ increases as $\mathfrak{p}$ increases, we have the following useful relation between mixing times:

$$\tau_{\mathfrak{p}}(\delta; (\mu_0, \Pi^\star)) \leq \tau_{\mathfrak{p}'}(\delta; (\mu_0, \Pi^\star)) \quad \text{for any} \quad \mathfrak{p}' \geq \mathfrak{p} \geq 1. \tag{2.5d}$$

Furthermore, the relation (2.5d) also implies that $\tau_{\mathrm{TV}}(\delta; (\mu_0, \Pi^\star)) = \tau_1(\frac{\delta}{2}; (\mu_0, \Pi^\star))$.

To quantify mixing time, it is convenient to have a rough measure of the distance between the initial distribution μ_0 and the stationary distribution. As in several past work, we adopt the following notion of *warmness*:

Definition 2.2 (Warm start)**.** *For a Markov chain with state space $\mathcal{X}$ and stationary distribution $\Pi^\star$ has a β-warm start if its initial distribution μ_0 satisfies , the initial distribution μ_0 is said to be β-warm with respect to the stationary distribution $\Pi^\star$ if*

$$\sup_{S \in \mathcal{B}(\mathcal{X})} \frac{\mu_0(S)}{\Pi^\star(S)} \leq \beta, \tag{2.6}$$

for a finite scalar $\beta > 0$, where $\mathcal{B}(\mathcal{X})$ denotes the Borel σ-algebra of the state space $\mathcal{X}$.

For simplicity, we say that μ_0 is a warm start if the warmness parameter β is a small constant (e.g., β does not scale with dimension d). We are interested in establishing the results that provide a precise scaling of the mixing time bounds as a function of the problem parameters. In particular, the results to follow establish bounds on the quantity

$$\sup_{\mu_0 \in \mathfrak{P}_\beta(\Pi^\star)} \tau_{\mathfrak{p}}(\delta; (\mu_0, \Pi^\star)),$$

as a function of dimension d, parameters β, δ, and other parameters related to the target distribution; here $\mathfrak{P}_\beta(\Pi^\star)$ denotes the set of all distributions that are β-warm with respect to $\Pi^\star$. Naturally, as the value of β decreases, the task of generating samples from the target distribution becomes easier.[1] However, access to a good "warm" distribution (small β) may not be feasible for many applications, and thus deriving bounds on mixing time of the Markov chain from non-warm starts is also desirable. Several of our results directly tackle this challenge by provide practical initialization methods and polynomial-time mixing time guarantees from such starts.

[1] For instance, $\beta = 1$ implies that the chain starts at the stationary distribution and has already mixed.

Lazy chain: We say that the Markov chain is ζ-*lazy* if at each iteration the chain is forced to stay at the previous iterate with probability ζ. Unless otherwise specified, we study $\frac{1}{2}$-lazy chains in this thesis. In practice, one is not likely to use a lazy chain (since the lazy steps slow down the convergence rate by a constant factor); rather, it is a convenient assumption for theoretical analysis of the mixing rate up to constant factors since any lazy (time-reversible) chain is always aperiodic and admits a unique stationary distribution. For more details, see the survey [242] and references therein.

2.2 Metropolis-Hastings Algorithms

We now describe, arguably the most popular class of MCMC algorithms (also the focus of this thesis), known as Metropolis-Hastings algorithms—named after the authors of the earliest works on MCMC [177, 107]. Our discussion here aims to only provide a refresher on the topic on subsets of Euclidean spaces, and we refer the reader to the books [210, 30] and references therein for further background.

Starting at a given initial positive density μ_0 over $\mathcal{X} \subset \mathbb{R}^d$, any such Markov chain is simulated in two steps: (1) proposal step, and (2) accept-reject step. For the proposal step, we make use of a *proposal function* $p : \mathcal{X} \times \mathcal{X} \in \mathbb{R}_+$, where $p(x, \cdot)$ is a density function for each $x \in \mathcal{X}$. At each iteration, given a current state $x \in \mathcal{X}$ of the chain, the algorithm proposes a new vector $z \in \mathcal{X}$ by sampling from the proposal density $p(x, \cdot)$. In the second step, the algorithm accepts $z \in \mathbb{R}^d$ as the new state of the Markov chain with probability

$$\alpha(x, z) := \min \left\{ 1,\ \frac{\Pi^\star(z) p(z, x)}{\Pi^\star(x) p(x, z)} \right\}. \tag{2.7}$$

Otherwise, with probability equal to $1 - \alpha(x, z)$, the chain stays at x. Thus, the overall transition kernel q for the Markov chain is defined by the function

$$q(x, z) := p(x, z)\alpha(x, z) \qquad \text{for } z \neq x,$$

and a probability mass at x with weight $1 - \int_{\mathcal{X}} q(x, z) dz$. The step (2.7) is commonly known as the *Metropolis-Hastings* correction/adjustment or the *accept-reject* step as it ensures that the target density $\Pi^\star$ is stationary for the Markov chain associated with the overall transition kernel for the following reason: The overall transition kernel satisfies *detailed balance condition* with the target distribution, meaning that

$$q(y, x)\pi^\star(x) = q(x, y)\pi^\star(y) \text{for all} x, y \in \mathcal{X}, \tag{2.8}$$

under mild conditions on $\Pi^\star$, and p, e.g., $\pi^\star$ and $p(x, \cdot)$ being positive on $\mathcal{X}$, for all $x \in \mathcal{X}$. It is straightforward to verify that the detailed balance condition (2.8) implies that the target density $\Pi^\star$ is stationary for the Markov chain.

Overall, this set-up defines the transition operator $\mathcal{T}_p$ as a function of p and $\Pi^\star$, on the space of probability distributions: given the distribution μ_k of the chain at time k, the distribution at time $k+1$ is given by $\mathcal{T}_p(\mu_k)$. In fact, with the starting distribution μ_0, the distribution of the chain at kth step is given by $\mathcal{T}_p^k(\mu_0)$. Given the detailed balance (2.8), , it is known that the chain converges to target distribution, asymptotically, i.e., in the limit of infinite steps: $\lim_{k\to\infty} \mathcal{T}_p^k(\mu_0) = \Pi^\star$. However, the focus of this part of our thesis is non-asymptotic analysis, namely, the number of the chain sufficient to ensure that the distribution of the chain is δ-close to the target $\Pi^\star$ in appropriate metric.

From the set-up above, one can easily note that given p, the user only requires the knowledge of target density up to proportionality to run the algorithm. In terms of the notation (2.1), we only need to know the function f. Given this knowledge, there are several generic schemes—described completely by specifying just the proposal function p—used to construct wide variety of sampling algorithms. Let $\mathcal{P}_x$ denote the proposal distribution at x corresponding to the proposal density $p(x,\cdot)$. Some illustrative examples of the commonly used algorithms are as follows:

- **Independence sampler**: the proposal function does not depend on the current state of the chain, e.g., rejection sampling or when $\mathcal{P}_x = \mathcal{N}(0,\Sigma)$, where Σ is a hyper-parameter;

- **Symmetric Metropolis algorithm**: the proposal function satisfies $p(x,y) = p(y,x)$ (independence sampler is a special case); for example, Ball Walk [93] with $\mathcal{P}_x = \mathcal{U}(\mathbb{B}(x,r))$ (uniform distribution), where r is a hyper-parameter;

- **Random walk**: the proposal function satisfies $p(x,y) = q(y-x)$ for some probability density q, e.g., Metropolis random walk with $\mathcal{P}_x = \mathcal{N}(x, 2\eta\mathbb{I}_d)$, where η is a hyper-parameter;

- **Langevin algorithm**: the proposal distribution is shaped according to the target distribution and is given by $\mathcal{P}_x = \mathcal{N}(x - \eta\nabla f(x), 2\eta\mathbb{I}_d)$, where η is a hyper-parameter (random walk is a special case with f is a constant function). This class of algorithm requires additional knowledge about the target density since it assumes access to the gradient ∇f information.

Naturally the convergence rate of these algorithms would depend on the properties of the target density $\Pi^\star$ and how well suited are the proposal function p for the task at hand. In the chapters to follow, we provide additional examples, and further details on the known results about the algorithms in relevant sections .

Chapter 3

Mixing Times for Random Walk and Langevin Algorithms

In this chapter, we study sampling algorithms for sampling from a log-concave distribution $\Pi^\star$ equipped with a density $\pi^\star$. A log-concave density takes the form

$$\pi^\star(x) = \frac{e^{-f(x)}}{\displaystyle\int_{\mathbb{R}^d} e^{-f(y)} dy} \quad \text{for all } x \in \mathbb{R}^d, \tag{3.1}$$

where f is a convex function on $\mathbb{R}^d$. Up to an additive constant, the function $-f$ corresponds to the log-likelihood defined by the density. Standard examples of log-concave distributions include the normal distribution, exponential distribution and Laplace distribution.

This chapter deals with both zeroth-order and first-order sampling methods, and focus on the Metropolis random walk, and Langevin algorithms. Our goal is to derive non-asymptotic mixing time for these methods. We start with background material in Sections 3.1 and 3.2 to provide some more context before summarizing our contributions in Section 3.2.4, and the organization of the remainder of the chapter in Section 3.2.5.

3.1 Metropolis Random Walk

Perhaps, one of the easiest algorithm to implement is the *Metropolis random walk* with Gaussian proposals. (This algorithm appears with several names in the literature including Random walk Metropolized and Random walk Metropolis-Hastings.) When the chain is at state x_k, a proposal is drawn as follows

$$z_{k+1} = x_k + \sqrt{2\eta}\, \xi_{k+1}, \tag{3.2}$$

where the noise term $\xi_{k+1} \sim \mathcal{N}(0, \mathbb{I}_d)$ is independent of all past iterates. The chain then makes the transition according to an accept-reject step with respect to $\Pi^\star$. Since the proposal distribution is symmetric, this step can be described as

$$x_{k+1} = \begin{cases} z_{k+1} & \text{with probability} \min\left\{1, \dfrac{\pi^\star(z_{k+1})}{\pi^\star(x_k)}\right\} \\ x_k & \text{otherwise.} \end{cases}$$

This sampling algorithm is an instance of a zeroth-order method, since it makes use of only the function values of the density $\pi^\star$. We refer to this algorithm as MRW in the sequel. It is easy to see that the chain has positive density of jumping from any state x to y in $\mathbb{R}^d$ and hence is strongly $\Pi^\star$-irreducible and aperiodic. Consequently, Theorem 1 by Diaconis et al. [69] implies that the chain has a unique stationary distribution $\Pi^\star$ and converges to in the limit of infinite steps. Roberts and Tweedie [217] established sufficient conditions on the proposal function p and the target distribution $\Pi^\star$ for the geometric convergence of several random walk Metropolis-Hastings algorithms, including MRW. Other related work with results on ergodicity, optimal scaling of asymptotic variance and central limit theorems include [217, 130, 212, 214], the survey [213] and the references therein. Such results are crucial for gaining helpful insight into the convergence properties of the MRW algorithm, but still they do not easily provide a user-friendly rate of convergence. In other words, from these results, it is not easy to determine the computational complexity of MRW (or other MCMC algorithms) as a function of the problem dimension d and desired accuracy δ. With this context, one of the key results in this chapter establishes explicit non-asymptotic mixing time guarantee for MRW (see Theorem 3.2).

Algorithm 1: Metropolized Random Walk (MRW)

```
Input: Step size $\eta > 0$ and a sample $x_0$ from a starting distribution $\mu_0$
Output: Sequence $x_1, x_2, \ldots$
for $i = 0, 1, \ldots$ do
    Proposal step: Draw $z_{i+1} \sim \mathcal{N}(x_i, 2\eta\mathbb{I}_d)$
    Accept-reject step:
        compute $\alpha_{i+1} \leftarrow \min\left\{1, \frac{\exp(-f(z_{i+1}))}{\exp(-f(x_i))}\right\}$
        With probability $\alpha_{i+1}$ accept the proposal: $x_{i+1} \leftarrow z_{i+1}$
        With probability $1 - \alpha_{i+1}$ reject the proposal: $x_{i+1} \leftarrow x_i$
end
```

Other instances of zeroth-order algorithms include the ball walk [166, 85, 167] and the hit-and-run algorithm [16, 137, 163, 169, 171]. While there are several mixing time guarantees for Ball Walk and Hit-and-run results, these methods are designed specifically for compactly supported distributions and do not immediately apply to distributions with support $\mathbb{R}^d$ which is the focus of this chapter (equation (2.1)). We defer further discussion about them to Chapter 6.

3.2 Langevin Algorithms

A number of first-order methods are based on the Langevin diffusion, a stochastic process whose evolution is characterized by the stochastic differential equation (SDE):

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}\, dW_t, \tag{3.3}$$

where $\{W_t \mid t \geq 0\}$ is the standard Brownian motion on $\mathbb{R}^d$. Under fairly mild conditions on f, it is known that the diffusion (3.3) has a unique strong solution $\{X_t, t \geq 0\}$ that is a Markov process [216, 179]. Furthermore, it can be shown that the distribution of X_t converges as $t \to +\infty$ to the invariant distribution $\Pi^\star$ with density $\pi^\star \propto \exp(-f)$ given as in equation (2.1). See Roberts and Tweedie [216] or Meyn and Tweedie [179] for further details on such an asymptotic guarantee. In practice, one can neither simulate the diffusion (3.3) exactly, nor can one run the algorithm for infinite horizon—and one can resort to some discrete-time time algorithm that approximates the diffusion.

3.2.1 Unadjusted Langevin Algorithm

A natural way to simulate the Langevin diffusion (3.3) is to consider its forward Euler discretization, given by

$$x_{k+1} = x_k - \eta\nabla f(x_k) + \sqrt{2\eta}\xi_{k+1}, \tag{3.4}$$

where the driving noise $\xi_{k+1} \sim \mathcal{N}(0, \mathbb{I}_d)$ is drawn independently at each time step. Note that while the algorithm (3.4) is an MCMC algorithm since the updates do form a Markov chain, it is not of Metropolis-Hastings type due to the lack of accept-reject step (2.7). Nonetheless, given the nice asymptotic convergence property of the diffusion (3.3), usage of such an *unadjusted* algorithm can be traced back at least to Parisi in 1981 [200] for computing correlations as noted by Besag in his commentary on the paper by Grenander and Miller [100].

However, even when the Langevin diffusion (3.3) is well behaved, the iterates defined by the discretization (3.4) can have mixed behavior. When step size η is large, the distribution of the iterates defined by equation (3.4) converges to a stationary distribution that is no longer equal to $\Pi^\star$. In fact, Roberts and Tweedie [216] showed that if one does not choose the step size η carefully, the Markov chain defined by equation (3.4) can become transient and have no stationary distribution. Nevertheless, a series of recent work [61, 74, 51] establish that with a careful choice of step-size η and iteration count K, running the chain (3.4) for exactly K steps yields an iterate x_K whose distribution is close to $\Pi^\star$. This more recent body of work provides non-asymptotic bounds that explicitly quantify the rate of convergence for this chain. The lack of the Metropolis-Hastings correction (adjustment) yields it the name of *unadjusted Langevin algorithm*, or ULA for short. Some works also refer to it as the Langevin Monte Carlo.

Durmus and Moulines [74] show that for an appropriate decaying step size schedule, the distribution of the iterates from the ULA algorithm does converge to $\Pi^\star$, when $\Pi^\star$ is strongly log-concave target. However, their results, albeit non-asymptotic, are hard to quantify. In the sequel, we limit our discussion to Metropolis random walk and Langevin algorithms based on constant step sizes, for which there are a number of explicit quantitative bounds on the mixing time.

3.2.2 Metropolis Adjusted Langevin Algorithm

An alternative approach to handling the discretization error is to use the iterates on the RHS of equation (3.4) as the proposals in a Metropolis-Hastings type algorithm. In other words, one can use the distribution $\mathcal{N}(x_k - \eta\nabla f(x_k), 2\eta\mathbb{I}_d)$ as the proposal distribution, and perform the Metropolis-Hastings accept-reject step, thereby yielding the *Metropolis-adjusted Langevin algorithm*, or MALA for short. Consequently, this algorithm has been the focus of several work in the past [216, 215, 25]. We describe the different steps of MALA in Algorithm 2. As mentioned in Section 2.2, the Metropolis-Hastings correction ensures that the distribution of the MALA iterates $\{x_k\}$ converges to the correct distribution $\Pi^\star$ as $k \to \infty$. Indeed, since at each step the chain can reach any state $x \in \mathbb{R}^d$, it is strongly $\Pi^\star$-irreducible and thereby ergodic [179, 69].

Both MALA and ULA are instances of first order sampling methods since they make use of the function and the gradient values of f.[1] While MALA is clearly superior to ULA asymptotically due to the lack of the bias, the main question of practical relevance is whether employing the accept-reject step for the discretization (3.4) provides any gain in the convergence rate—Which of the two converge take lesser number of iterations (computational budget) to converge to a desired accuracy? Our analysis to follow provides a precise answer to this question under certain assumptions on the target distribution, establishing that MALA has a superior finite time convergence guarantees compared to ULA.

Related work on MALA: Several works [233, 178, 215, 205] characterize asymptotic limiting behaviors (in time or dimension) of the Langevin diffusion or provide non-explicit discretization error guarantees, that do not immediately yield a user-friendly mixing-time guarantee. Roberts and Tweedie [216] derived sufficient conditions for exponential convergence of the Langevin diffusion and its discretizations, with and without Metropolis-adjustment. However, they considered the distributions with $f(x) = \|x\|_2^\alpha$ and proved geometric convergence of ULA and MALA under some specific conditions; however they did not provide a precise quantification of the asymptotic bias of ULA. In a more general setting, Bou-Rabee and Hairer [25] derived non-asymptotic mixing time bounds for MALA, which can potentially be compared to the recent works on ULA. However, these bounds are non-explicit, and so makes it difficult to extract an explicit dependence in terms of the dimension d and error tolerance δ. A precise characterization of this dependence is needed if one wants to

[1] While ULA only uses gradient information, the computational complexity per step for MALA and ULA are still typically of the same order, since computing the function is often a cheaper operation, and/or a pre-requisite to computing the gradients.

make quantitative comparisons with ULA or any other sampling algorithm. Along this note, Eberle [86] derived mixing time bounds for MALA albeit in a more restricted setting compared to the one considered in this chapter. In particular, Eberle's convergence guarantees are in terms of a modified Wasserstein distance, truncated so as to be upper bounded by a constant, for a subset of strongly concave measures which are four-times continuously differentiable and satisfy certain bounds on the derivatives up to order four. With this context, one of the main contributions of this chapter is to provide an explicit upper bound on the mixing time bounds in total variation distance of the MALA algorithm for general log-concave distributions.

Algorithm 2: Metropolis adjusted Langevin algorithm (MALA)

```
Input: Step size $\eta$ and a sample $x_0$ from a starting distribution $\mu_0$
Output: Sequence $x_1, x_2, \ldots$
for $i = 0, 1, \ldots$ do
    $z_{i+1} \sim \mathcal{N}(x_i - \eta\nabla f(x_i), 2\eta\mathbb{I}_d)$   % propose a new state
    $\alpha_{i+1} = \min\left\{1, \frac{\exp\left(-f(z_{i+1}) - \left\|x_i - z_{i+1} + \eta\nabla f(z_{i+1})\right\|_2^2/4\eta\right)}{\exp\left(-f(x_i) - \left\|z_{i+1} - x_i + \eta\nabla f(x_i)\right\|_2^2/4\eta\right)}\right\}$
    $U_{i+1} \sim U[0,1]$
    if $U_{i+1} \leq \alpha_{i+1}$ then $x_{i+1} \leftarrow z_{i+1}$   % accept the proposal
    else $x_{i+1} \leftarrow x_i$   % reject the proposal
end
```

3.2.3 Other Langevin-type Algorithms

Over the years, numerous practical algorithms related to the Langevin diffusion have been proposed besides ULA and MALA. These algorithms include the underdamped Langevin MCMC [52] also called as kinetic Langevin Monte Carlo [63], second-order Langevin Monte Carlo [61], Riemannian MALA [252], Proximal-MALA [204, 75], Metropolis adjusted Langevin truncated algorithm [216], and Projected ULA [31]. Some of these work establish non-asymptotic mixing time bounds for sampling from a log-concave density; e.g., it is now well-known that both the ULA updates [61, 74, 51] as well as underdamped Langevin MCMC [52] have mixing times that scale polynomially in the dimension d, as well the inverse of the error tolerance $1/\delta$ for strongly log-concave distributions. There is now a rich body of work on these methods, and we do not attempt to provide a comprehensive summary here; see the aforementioned references, and the survey [213] for related details.

3.2.4 Overview of our contributions

This chapter provides two main results, both having to do with the upper bounds on mixing times of MCMC methods for sampling. As described above, our first and primary contribution is an explicit analysis of the mixing time of Metropolis adjusted Langevin Algorithm (MALA). A second contribution is to use similar techniques to analyze a zeroth-order method called Metropolized random walk (MRW) and derive an explicit non-asymptotic mixing time bound for it. Unlike the ULA, these methods make use of the Metropolis-Hastings accept-reject step and consequently converge to the target distributions in the limit of infinite steps. Here we provide explicit non-asymptotic mixing time bounds for MALA and MRW, thereby showing that MALA converges significantly faster than ULA, at least in terms of the best known upper bounds on their respective mixing times.[2] In particular, we show that if the density is strongly log-concave and smooth, the δ-mixing time for MALA scales as $\kappa d \log(1/\delta)$ which is significantly faster than ULA's convergence rate of order $\kappa^2 d/\delta^2$. On the other hand, we also show that MRW mixes $\mathcal{O}(\kappa)$ slower when compared to MALA. Furthermore, if the density is weakly log-concave, we show that (a modified version of) MALA converges in $\mathcal{O}(d^2/\delta^{1.5})$ time in comparison to the $\mathcal{O}(d^3/\delta^4)$ mixing time for ULA. As alluded to earlier, such a speed-up for MALA is possible since we can choose a large step size for it which in turn is possible due to its unbiasedness in the limit of infinite steps. In contrast, for ULA the step-size has to be small enough to control the bias of the distribution of the ULA iterates in the limit of infinite steps, leading to a relative slow down when compared to MALA.

3.2.5 Organization

The remainder of the chapter is organized as follows. Section 3.3 is devoted to the statement of our mixing time bounds for MALA and MRW, along with a discussion of some consequences of these results. Section 3.4 is devoted to numerical experiments that further illustrate our guarantees. We provide the proofs of our main results in Section 3.5, with certain more technical arguments deferred to the appendices. We conclude with a discussion in Section 3.6.

3.3 Main results

In this section, we state our main results on the non-asymptotic mixing time guarantees for MALA and MRW, and compare it to ULA. Before stating these results, we first state the regularity conditions assumed on the target distribution $\Pi^\star$ for deriving our results.

3.3.1 Assumptions on the target distribution

We now describe the regularity conditions on the target distributions that our results in the next section rely on. We analyze MALA and MRW and contrast their performance with existing algorithms for log-concave targets, i.e., for the case when the negative log density

[2] Throughout the chapter, we make comparisons between sampling algorithms based on known upper bounds on respective mixing times; obtaining matching lower bounds is also of interest.

$f(x) := -\log \pi^\star(x) + c$ is smooth and convex. We collect some standard definitions assuming $f : \mathbb{R}^d \to \mathbb{R}$ is a differentiable function.

A function f is said to be $\mathcal{L}$-*smooth* if

$$f(y) - f(x) - \nabla f(x)^\top (y - x) \leq \frac{\mathcal{L}}{2} \|x - y\|_2^2 \quad \text{for all } x, y \in \mathbb{R}^d. \tag{3.5a}$$

In the other direction, a convex and differentiable function f is said to be m-*strongly convex* if

$$f(y) - f(x) - \nabla f(x)^\top (y - x) \geq \frac{m}{2} \|x - y\|_2^2 \qquad \text{for all } x, y \in \mathbb{R}^d. \tag{3.5b}$$

In Appendix A.5, we state some well-known properties of smooth and strongly convex functions, that we later use for our results.

The rates derived in this chapter apply to two different settings of log-concave target distributions (defined via (3.1)).

(3A) We say that the target distribution $\Pi^\star$ is $(\mathcal{L}, m)$-*strongly log-concave* distribution if the negative log density function f is both $\mathcal{L}$-smooth (3.5a) and m-strongly convex (3.5b). For this case, we also use the notation $\kappa := \mathcal{L}/m$, and call it the condition number of the target $\Pi^\star$. Moreover, we use $x^\star$ to denote the unique mode of $\Pi^\star$ whenever f is strongly convex.

(3B) We say that the target distribution $\Pi^\star$ is $\mathcal{L}$-*weakly log-concave* distribution when the negative log density function We assume that the function f is $\mathcal{L}$-smooth, and convex (but not necessarily strongly convex, i.e., $m = 0$).

Common examples of strongly log-concave targets include multivariate Gaussian distribution[3], $f(x) = x^\top B x + g(x)$ for any convex g and any positive definite matrix B, or posterior distributions in Bayesian logistic regression with Gaussian prior. Examples of weakly log-concave target include $f(x) = \|x\|_2^4$, or $f(x) = \log(1 + e^{-\theta^\top x})$ which would arise, e.g., in Bayesian logistic regression with flat prior.

Organization of results: We discuss the guarantees for strongly log-concave target from a warm start in Section 3.3.2, and from certain feasible starting distributions in Section 3.3.3, and then we consider the case of weakly log-concave target in Section 3.3.4. An overview of results is summarized in Tables 3.1 and 3.2, as a function of the dimension d, the error-tolerance δ, the condition number κ (for the strongly log-concave target), and the smoothness parameter $\mathcal{L}$ (for the weakly log-concave target). In Table 3.1, we state the results when the chain has a warm-start (i.e., β a fixed constant, c.f. Definition 2.2). On the other hand, Table 3.2 summarizes mixing time bounds from a particular distribution $\mu_\star$ for the strongly log-concave target.

[3] For the Gaussian target $\mathcal{N}(z, \Sigma)$, the condition number of the target is equal to the condition number of the covariance matrix Σ, i.e., κ can be bounded by the ratio of maximum and minimum eigenvalues of Σ.

Random walk	Strongly log-concave	Weakly log-concave
ULA [51]	$\mathcal{O}\left(\frac{d\kappa^2\log((\log\beta)/\delta)}{\delta^2}\right)$	$\widetilde{\mathcal{O}}\left(\frac{d\mathcal{L}^2}{\delta^6}\right)$
ULA [61]	$\mathcal{O}\left(\frac{d\kappa^2\log^2(\beta/\delta)}{\delta^2}\right)$	$\widetilde{\mathcal{O}}\left(\frac{d^3\mathcal{L}^2}{\delta^4}\right)$
MRW [Thm. 3.2, Cor. 3.2]	$\mathcal{O}\left(d\kappa^2\log\left(\frac{\beta}{\delta}\right)\right)$	$\widetilde{\mathcal{O}}\left(\frac{d^3\,\mathcal{L}^2}{\delta^2}\right)$
MALA [Thm. 3.1, Cor. 3.2]	$\mathcal{O}\left(\max\left\{d\kappa, d^{0.5}\kappa^{1.5}\right\}\log\left(\frac{\beta}{\delta}\right)\right)$	$\widetilde{\mathcal{O}}\left(\frac{d^2\,\mathcal{L}^{1.5}}{\delta^{1.5}}\right)$

Table 3.1: Scalings of upper bounds on δ-mixing time for different random walks in $\mathbb{R}^d$ with target $\pi^\star \propto e^{-f}$. In the second column, we consider smooth and strongly log-concave target (Assumption (3A)), and report the bounds from a β-warm start for densities such that $m\mathbb{I}_d \preceq \nabla^2 f(x) \preceq \mathcal{L}\mathbb{I}_d$ for any $x \in \mathbb{R}^d$ and use $\kappa := \mathcal{L}/m$ to denote the condition number of the density. The big-$\mathcal{O}$ notation hides universal constants. In the last column, we summarize the scaling for weakly log-concave smooth densities: $0 \preceq \nabla^2 f(x) \preceq \mathcal{L}\mathbb{I}_d$ for all $x \in \mathbb{R}^d$. For this case, the $\widetilde{\mathcal{O}}$ notation is used to track scaling only with respect to d, δ and $\mathcal{L}$ and ignore dependence on the starting distribution and a few other parameters.

Remark: We note that our techniques yield sharper guarantees under more idealized assumptions, e.g., when f is Lipschitz, i.e., has bounded gradients, and the target distribution satisfies certain isoperimetry inequality. For such settings, the target need not even be log-concave, i.e., f need not be convex. Nonetheless, we restrict our attention in this chapter to log-concave target distributions, and refer the interested readers for further discussion for mixing times for such class of distributions to Chapter 5.

3.3.2 Mixing time for strongly log-concave target: Warm start

Given the parameters m and $\mathcal{L}$, our results involve the functions $\mathfrak{a}$ and $\mathfrak{t}$ given by

$$\mathfrak{a}(s) = 2 + 2\cdot\max\left\{\frac{1}{d^{0.25}}\log^{0.25}\left(\frac{1}{s}\right), \frac{1}{d^{0.5}}\log^{0.5}\left(\frac{1}{s}\right)\right\}, \quad \text{and} \tag{3.6a}$$

$$\mathfrak{t}(s) = \min\left\{\frac{\sqrt{m}}{\mathfrak{a}(s)\cdot\mathcal{L}\sqrt{d\mathcal{L}}}, \frac{1}{\mathcal{L}d}\right\} \qquad \text{for } s \in \left(0, \tfrac{1}{2}\right). \tag{3.6b}$$

Let $\mathcal{T}_{\mathrm{MALA}(\eta)}$ denote the transition operator (2.4) on probability distributions induced by one step of MALA. The next result our first guarantee for MALA.

Random walk	$\mu_\star$	$t_{\mathrm{mix}}(\delta;\mu_0)$
ULA [51]	$\mathcal{N}(x^\star, m^{-1}\mathbb{I}_d)$	$\mathcal{O}\left(\frac{d\kappa^2\log(d\kappa/\delta)}{\delta^2}\right)$
ULA [61]	$\mathcal{N}(x^\star, \mathcal{L}^{-1}\mathbb{I}_d)$	$\mathcal{O}\left(\frac{(d^3+d\log^2(1/\delta))\kappa^2}{\delta^2}\right)$
MRW	$\mathcal{N}(x^\star, \mathcal{L}^{-1}\mathbb{I}_d)$	$\mathcal{O}\left(d^2\kappa^2\log^{1.5}\left(\frac{\kappa}{\delta}\right)\right)$
MALA	$\mathcal{N}(x^\star, \mathcal{L}^{-1}\mathbb{I}_d)$	$\mathcal{O}\left(d^2\kappa\log\left(\frac{\kappa}{\delta}\right)\right)$

Table 3.2: Scalings of upper bounds on δ-mixing time, from the starting distribution $\mu_\star$ given in column two, for different random walks in $\mathbb{R}^d$ with target $\pi^\star \propto e^{-f}$ such that $m\mathbb{I}_d \preceq \nabla^2 f(x) \preceq \mathcal{L}\mathbb{I}_d$ for any $x\in\mathbb{R}^d$ and $\kappa := \mathcal{L}/m$. Here $x^\star$ denotes the unique mode of the target density $\pi^\star$.

Theorem 3.1. *For an $(\mathcal{L}, m)$-strongly log-concave target distribution $\Pi^\star$ (Assumption (3A)), given any β-warm initial distribution μ_0, and any error tolerance $\delta\in(0,1]$, the Metropolis adjusted Langevin algorithm with step size $\eta = c\,\mathfrak{r}(\delta/(2\beta))$ satisfies*

$$d_{\mathrm{TV}}\left(\mathcal{T}^{\ell}_{\mathrm{MALA}(\eta)}(\mu_0), \Pi^\star\right) \leq \delta \quad \textit{for all} \quad \ell \geq c'\log\left(\frac{2\beta}{\delta}\right)\max\left\{d\kappa,\ d^{0.5}\kappa^{1.5}\mathfrak{a}\left(\frac{\delta}{2\beta}\right)\right\}, \tag{3.7}$$

where $\kappa = \mathcal{L}/m$, and c, c' denote universal constants.

See Section 3.5.3 for the proof.

Noting that $\mathfrak{a}(s) \leq 4$ for $s \geq e^{-d}$, we can treat $\mathfrak{a}(\delta/2\beta)$ as a small constant for most interesting values of δ assuming β is not too large (hence the name warm start). Treating the term corresponding to $\mathfrak{a}$ as a constant, we obtain that if $\kappa = o(d)$, the mixing time of MALA scales as $\mathcal{O}\left(d\kappa\log(1/\delta)\right)$ which is exponentially better in the tolerance-δ compared to $\mathcal{O}\left(d\kappa^2\log^2(1/\delta)/\delta^2\right)$ mixing time of ULA, and has better dependence on κ while still maintaining linear dependence on d. In fact, for any setting of κ, d and δ, MALA always has a better mixing time bound compared to ULA. A limitation of our analysis is that the constant c' is not small. However, we demonstrate in Section 3.4 that in practice small constants provide performance that match the scalings suggested by our theoretical bounds.

Let $\mathcal{T}_{\mathrm{MRW}(\eta)}$ denote the transition operator (2.4) on the space of probability distributions induced by one step of MRW. We now state our first mixing time bound for Metropolized random walk.

Theorem 3.2. *For an $(\mathcal{L}, m)$-strongly log-concave target distribution $\Pi^\star$ (Assumption (3A)), given any β-warm initial distribution μ_0, and any error tolerance $\delta \in (0, 1]$, the Metropolized random walk with step size $\eta = \frac{cm}{d\mathcal{L}^2 \mathfrak{a}(\delta/(2\beta))}$ satisfies*

$$d_{\mathrm{TV}}\left(\mathcal{T}^{\ell}_{\mathrm{MRW}(\eta)}(\mu_0), \Pi^\star\right) \leq \delta \quad \text{for all} \quad \ell \geq c'\, d\kappa^2 \mathfrak{a}\left(\frac{\delta}{2\beta}\right) \log\left(\frac{2\beta}{\delta}\right), \tag{3.8}$$

where c, c' denote universal constants.

See Section 3.5.4 for the proof.

Again treating $\mathfrak{a}(\delta/2\beta)$ as a small constant, we find that the mixing time of MRW scales as $\mathcal{O}\left(d\kappa^2 \log(1/\delta)\right)$ which has an exponential factor in δ better than ULA. Compared to the mixing time bound for MALA, the bound in Theorem 3.2 has an extra factor of $\mathcal{O}(\kappa)$. While such a factor is conceivable given that MALA's proposal distribution uses first order information about the target distribution, and MRW uses only the function values, it would be interesting to determine if this gap can be improved.

3.3.3 Mixing time for strongly log-concave target: Feasible start

In many cases, a good warm start may not be readily known, and thus it would be useful to derive mixing time guarantees from an initial distribution that can be implemented easily. The next lemma (with proof in Appendix A.1) provides a bound on the warmness of a feasible distribution.

Lemma 3.1. *The distribution $\mu_\star = \mathcal{N}(x^\star, \mathcal{L}^{-1}\mathbb{I}_d)$ is $\beta_\star = \kappa^{d/2}$-warm with respect to an $(\mathcal{L}, m)$ target distribution, where $x^\star$ denotes the unique mode of the target distribution $\Pi^\star$.*

When the gradient ∇f is available, finding $x^\star$ comes at nominal additional cost: in particular, standard optimization algorithms such as gradient descent be used to compute a δ-approximation of $x^\star$ in $\mathcal{O}\left(\kappa \log(1/\delta)\right)$ steps (e.g., see the monograph [32]). where the function $\mathfrak{t}$ was previously defined in equation (3.6b). The next result provides the mixing time guarantees when MALA and MRW are initialized with the feasible distribution $\mu_\star$.

Corollary 3.1. *For any threshold $\delta \in (0, 1]$, define $\eta_1 = c'\mathfrak{t}(\delta/(2\beta_\star))$, and $\eta_2 = \frac{c'm}{d\mathcal{L}^2 \cdot \mathfrak{a}(\delta/(2\beta_\star))}$, where $\beta_\star = \kappa^{d/2}$. Then with $\mu_\star = \mathcal{N}(x^\star, \mathcal{L}^{-1}\mathbb{I}_d)$ as the starting distribution, we have*

$$d_{\mathrm{TV}}\left(\mathcal{T}^{\ell}_{\mathrm{MRW}(\eta_2)}(\mu_\star), \Pi^\star\right) \leq \delta \quad \text{for all } \ell \geq c\ d^2\kappa^2 \log^{1.5}\left(\frac{\kappa}{\delta^{1/d}}\right), \quad \text{and}$$

$$d_{\mathrm{TV}}\left(\mathcal{T}^{k}_{\mathrm{MALA}(\eta_1)}(\mu_\star), \Pi^\star\right) \leq \delta \quad \text{for all } \ell \geq c\ d^2\kappa \log\left(\frac{\kappa}{\delta^{1/d}}\right) \max\left\{1, \sqrt{\frac{\kappa}{d}} \log\left(\frac{\kappa}{\delta^{1/d}}\right)\right\}.$$

Remark: The proof of Corollary 3.1 follows by plugging the warmness bound from Lemma 3.1 in Theorem 3.1 and 3.2 and is thereby omitted. We note that compared to the mixing time from a warm-start with constant β, the bounds in Corollary 3.1 are roughly $\mathcal{O}(d)$ worse since the warmness parameter of $\mu_\star$ is exponential in d. Such a worsening is caused by the $\log\beta$ scaling of the mixing time bound in Theorems 3.1 and 3.2. In the next chapter, we sharpen the results of this corollary by a factor of $\mathcal{O}(d/\log d)$ for different choices of step sizes using a more refined analysis. See Section 4.3 and Corollary 4.2 for further discussion.

Inexact parameters: Our techniques can also establish mixing time guarantees when the mode $x^\star$ and the smoothness parameter $\mathcal{L}$ are only approximately known—a situation occurring quite often in practice. See section 3.2.1 of the full paper [78] for further details, where we show that the mixing time bounds are not severely affected if the parameters $x^\star$ and $\mathcal{L}$ are known to a reasonable perturbation error from the exact parameters.

3.3.4 Mixing time for weakly log-concave target

In this section, we show that MRW and MALA can also be used for approximate sampling from a density which is $\mathcal{L}$-smooth but not necessarily strongly log-concave (also referred to as weakly log-concave [61]). In simple words, the negative log-density f satisfies the condition (3.5a) with parameter $\mathcal{L}$ and satisfies the condition (3.5b) with parameter $m = 0$. Equivalently, we have $\mathcal{L}\mathbb{I}_d \succeq \nabla^2 f(x) \succeq 0$ for all $x \in \mathbb{R}^d$ (see Appendix A.5).

In order to make use of our previous machinery for such a case, we approximate the given log-concave density $\Pi^\star$ with a strongly log-concave density $\widetilde{\Pi}^\star$ such that $d_{\mathrm{TV}}\big(\widetilde{\Pi}^\star, \Pi^\star\big)$ is small. Next, we use MRW or MALA to sample from $\widetilde{\Pi}^\star$ and consequently obtain an approximate sample from $\Pi^\star$. In order to construct $\widetilde{\Pi}^\star$, we use a scheme previously suggested by Dalalyan [61]. With λ as a tuning parameter, consider the distribution $\widetilde{\Pi}^\star$ given by the density

$$\widetilde{\pi}^\star(x) = \frac{1}{\displaystyle\int_{\mathbb{R}^d} e^{-\widetilde{f}(y)}dy} e^{-\widetilde{f}(x)} \quad \text{where} \quad \widetilde{f}(x) = f(x) + \frac{\lambda}{2}\left\|x - x^\star\right\|_2^2. \tag{3.10}$$

Dalalyan (Lemma 3 in the paper [61]) showed that that the total variation distance between $\Pi^\star$ and $\widetilde{\Pi}^\star$ is bounded as follows:

$$d_{\mathrm{TV}}\big(\widetilde{\Pi}^\star, \Pi^\star\big) \leq \frac{1}{2}\|\widetilde{f} - f\|_{L^2(\pi^\star)} \leq \frac{\lambda}{4}\left(\int_{\mathbb{R}^d} \left\|x - x^\star\right\|_2^4 \pi^\star(x)dx\right)^{1/2}. \tag{3.11}$$

Suppose that the original distribution $\Pi^\star$ has its fourth moment bounded as

$$\int_{\mathbb{R}^d} \left\|x - x^\star\right\|_2^4 \pi^\star(x)dx \leq \frac{d^2\omega^2}{\mathcal{L}^2}. \tag{3.12}$$

We now set $\lambda := 2\mathcal{L}\delta/(d\omega)$ to obtain $d_{\mathrm{TV}}\big(\widetilde{\Pi}^{\star}, \Pi^{\star}\big) \leq \delta/2$. Since $\widetilde{f}$ is $\lambda/2$-strongly convex and $\mathcal{L} + \lambda/2$-smooth, the condition number of $\widetilde{\Pi}^{\star}$ is given by $\widetilde{\kappa} = 1 + d\omega/\delta$. We substitute $\widetilde{\kappa} = d\omega/\delta$ to obtain simplified expressions for mixing time bounds in the results that follow. Since now the target distribution is $\widetilde{\Pi}^{\star}$, we suitably modify the step size for MALA as follows:

$$\mathfrak{t}_{\mathrm{lc}}(s) = \frac{1}{\mathcal{L}d} \min\left\{\frac{\sqrt{s}}{\mathfrak{a}(s)\,\sqrt{\omega}},\ 1\right\}, \tag{3.13}$$

where the function $\mathfrak{a}$ was previously defined in equation (3.6a). We refer to this new set-up with a modified target distribution $\widetilde{\Pi}^{\star}$ as the *modified MALA method*. Similarly, we call the algorithm MRW with target $\widetilde{\Pi}^{\star}$ as *modified MRW*. To keep our results simple to state, we assume that we have a warm start with respect to $\widetilde{\Pi}^{\star}$.

Corollary 3.2. *For an $\mathcal{L}$-weakly log-concave $\Pi^{\star}$ (Assumption (3B)) that satisfies* (3.12)*, and any fixed error-tolerance $\delta \in (0, 1)$, any starting distribution μ_0 that is β-warm with respect to $\widetilde{\Pi}^{\star}$* (3.10)*, the modified MRW with step size $\eta_1 = \frac{c_1\delta}{d^2\mathcal{L}\omega\mathfrak{a}(\delta/(2\beta))}$, and the modified MALA method with step size $\eta_2 = c_2\mathfrak{t}_{lc}(\delta/(2\beta))$* (3.13) *satisfy*

$$d_{\mathrm{TV}}\big(\mathcal{T}^{\ell}_{\mathrm{MRW}(\eta_2)}(\mu_0), \Pi^{\star}\big) \leq \delta \quad \textit{for all} \quad \ell \geq c' \log\left(\frac{4\beta}{\delta}\right)\frac{d^3\omega^2}{\delta^2}, \quad \textit{and}$$

$$d_{\mathrm{TV}}\big(\mathcal{T}^{\ell}_{\mathrm{MALA}(\eta_1)}(\mu_0), \Pi^{\star}\big) \leq \delta \quad \textit{for all} \quad \ell \geq c'' \log\left(\frac{4\beta}{\delta}\right) \max\left\{\frac{d^2\omega}{\delta},\ \frac{d^2\omega^{1.5}}{\delta^{1.5}}\mathfrak{a}\left(\frac{\delta}{4\beta}\right)\right\}.$$

The proof follows by combining the triangle inequality, as applied to the TV norm, along with the bounds from Theorems 3.1 and 3.2. Thus, for weakly log-concave densities, modified MALA mixes in $\mathcal{O}\left(d^2/\delta^{1.5}\right)$, which improves upon the $\mathcal{O}\left(d^3/\delta^4\right)$ mixing time bound for a ULA scheme on $\widetilde{\Pi}^{\star}$, as established by Dalalyan [61]. Moreover, the modified MRW admits a mixing time bound of $\mathcal{O}\left(d^3/\delta^2\right)$ for the weakly log-concave target.

3.4 Numerical experiments

In this section, we compare MALA with ULA and MRW in various simulation settings. The step-size choice of ULA follows from [61] in the case of warm start. The step-size choice of MALA and MRW used in our experiments in our results are summarized in Table 3.3.

Summary of experiment set-ups and diagnostic tools: We consider four different experiments: (i) sampling a multivariate Gaussian (Section 3.4.1), (ii) sampling a Gaussian mixture (Section 3.4.2), (iii) estimating the MAP with credible intervals in a Bayesian logistic regression set-up (Section 3.4.3) and (iv) studying the effect of step-size on the accept reject step (Section 3.4.4). Since TV distance for continuous measures is hard to estimate, we use several proxy measures for convergence diagnostics: (a) errors in quantiles, (b) ℓ_1-distance in histograms (which we refer to as discrete tv-error), (c) error in sample MAP estimate, (d) trace-plot along different coordinates and (e) autocorrelation plot. While the first three measures (a-c) are useful for diagnosing the convergence of random walks over several independent runs, the last two measures (d-e) are useful for diagnosing the rate of convergence of the Markov chain in a single long run.

3.4.1 Dimension dependence for multivariate Gaussian

The goal of this simulation is to demonstrate the dimension dependence in experiments, for mixing time of ULA, MALA and MRW when the target is non-isotropic multivariate Gaussian. Note that Theorem 3.1 and 3.2 imply that the dimension dependency for both MALA and MRW is d. We consider sampling from multivariate Gaussian with density $\pi^\star$ defined by

$$x \mapsto \pi^\star(x) \propto e^{-\frac{1}{2}x^\top \Sigma^{-1} x}, \tag{3.14}$$

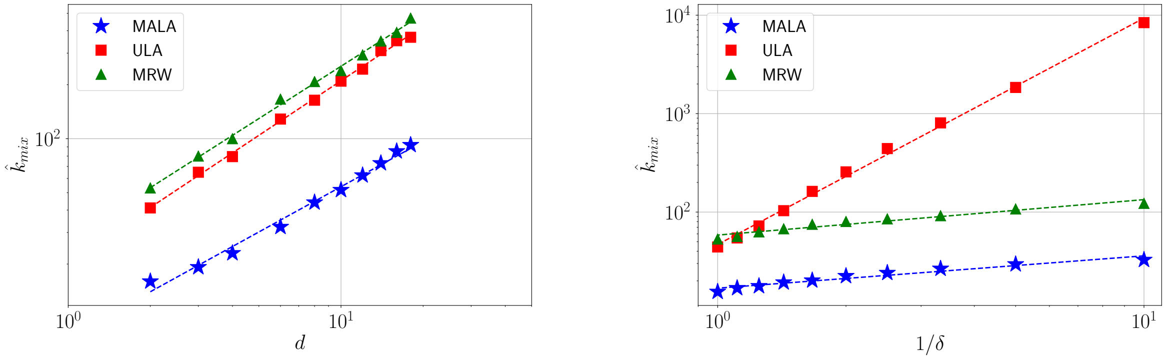
where $\Sigma \in \mathbb{R}^{d\times d}$ the covariance matrix to be specified. For this target distribution, the function f, its derivatives are given by

$$f(x) = \frac{1}{2}x^\top \Sigma^{-1} x, \quad \nabla f(x) = \Sigma^{-1}x, \quad \text{and} \quad \nabla^2 f(x) = \Sigma^{-1}.$$

Consequently, the function f is strongly convex with parameter $m = 1/\lambda_{\max}(\Sigma)$ and smooth with parameter $\mathcal{L} = 1/\lambda_{\min}(\Sigma)$. For convergence diagnostics, we use the error in quantiles along different directions. Using the exact quantile information for each direction for Gaussian, we measure the error in the 75% quantile of the sample distribution and the true distribution in the *least favorable direction*, i.e., along the eigenvector of Σ corresponding to the eigenvalue $\lambda_{\max}(\Sigma)$. The *approximate mixing time* $\hat{k}_{\text{mix}}(\delta)$ is defined as the smallest iteration when this error falls below δ. We use $\mu_\star$ as the initial distribution where $\mu_\star = \mathcal{N}\left(0, \mathcal{L}^{-1}\mathbb{I}_d\right)$.

3.4.1.1 Strongly log-concave density

The step-sizes are chosen according to Table 3.3. For ULA, the error-tolerance δ is chosen to be 0.2. We set Σ as a diagonal matrix with the largest eigenvalue 4.0 and the smallest eigenvalue 1.0 so that the $\kappa = 4$ is fixed across different settings. For a fixed dimension d, we simulate 10 independent runs of the three chains each with $N = 10,000$ samples to determine the approximate mixing time. The final approximate mixing time for each walk is the average of that over these 10 independent runs. Figure 3.1(a) shows the dependency of the approximate mixing time as a function of dimension d for the three random walks in log-log scale. To examine the dimension dependency, we perform linear regression for approximate mixing time with respect to dimensions in the log-log scale. The computations reveal that the dimension dependency of MALA, ULA and MRW are all close to order d (slope 0.84, 1.01 and 0.97). Figure 3.1(b) shows the dependency of the approximate mixing time on the inverse error $1/\delta$ for the three random walks in log-log scale. For ULA, the step-size is error-dependent, precisely chosen to be 10 times of δ. A linear regression of the



(a) (b)

Figure 3.1: Scaling of the approximate mixing time $\hat{k}_{\text{mix}}$ (refer to the discussion after equation (5.13) for the definition) for the multivariate Gaussian target (5.13) where the covariance has condition number $\kappa = 4$. (a) Dimension dependency. (b) Error-tolerance dependency.

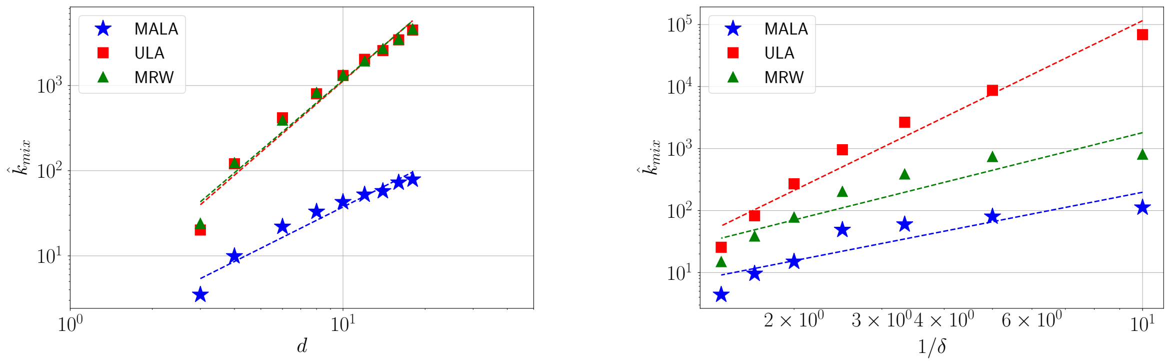
approximate mixing time on the inverse error $1/\delta$ yields a slope of 2.23 suggesting the error dependency of order $1/\delta^2$ for ULA. A similar computation for MALA and MRW yields a slope of 0.33 for both the cases which not only suggests a significantly better error dependency for these two chains but also partly verifies their theoretical mixing time bounds of order $\log(1/\delta)$.

Random walk	ULA	MALA	MRW
Step size	$\dfrac{\delta^2}{d\kappa\mathcal{L}}$	$\dfrac{1}{\mathcal{L}}\min\left\{\dfrac{1}{\sqrt{d\kappa}}, \dfrac{1}{d}\right\}$	$\dfrac{1}{d\kappa\mathcal{L}}$

Table 3.3: Step size used in simulations to obtain δ-accuracy for different random walks in $\mathbb{R}^d$ with target $\pi^\star \propto e^{-f}$ such that $m\mathbb{I}_d \preceq \nabla^2 f(x) \preceq \mathcal{L}\mathbb{I}_d$ for any $x \in \mathbb{R}^d$ and $\kappa := \mathcal{L}/m$.

3.4.1.2 Weakly log-concave density

We now discuss the convergence of the random walks when the Gaussian is flat along a direction. In particular, we consider the Gaussian distribution such that $\lambda_{\max}(\Sigma) = 1000$ and $\lambda_{\min}(\Sigma) = 1$. Such a setting implies that the strong convexity parameter $m = 0.001 \approx 0$ and hence our target density mimics a weakly log-concave density. For convergence diagnostics, we use the error in quantiles along one direction other than the ones which correspond to $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$. Using the exact quantile information for each direction for Gaussian,



(a)

(b)

Figure 3.2: Scaling of the approximate mixing time $\hat{k}_{\text{mix}}$ (refer to the discussion after equation (5.13) for the definition) for a close to weakly log-concave Gaussian density. (a) Dimension dependency. (b) Error-tolerance dependency for fixed dimension .

we measure the error between the 75% quantile of the sample distribution and the true distribution in that direction. The approximate mixing time is defined as the smallest iteration when this error falls below δ. We use $\mu_\star$ as the initial distribution where $\mu_\star = \mathcal{N}(0, \mathcal{L}^{-1}\mathbb{I}_d)$. The step-sizes are chosen according to Table 3.3 where m is chosen to be $\delta/(d\mathcal{L})$. For dimension dependence experiments, we fix the error-tolerance δ as 0.2. For a fixed dimension d, we simulate 10 independent runs of the three chains each with $N = 10,000$ samples to determine the approximate mixing time. The final approximate mixing time for each walk is the average of that over these 10 independent runs. Figure 3.2(a) and 3.2(b) show the dependency of the approximate mixing time as a function of dimension d and the inverse error $1/\delta$ respectively, for the three random walks on this weakly log-concave density (log-log scale). Linear fits on the log-log scale reveal that the dimension dependence of mixing time for MALA is close to d^2 (slope 1.61), and that for ULA is close to d^3 (slope 2.78) and for MRW it is approximately of order d^3 (slope 2.73). Linear fits of the approximate mixing time on the inverse error $1/\delta$ yield a slope of 3.92 for ULA thereby suggesting an error dependence of order $1/\delta^4$, while for MALA and MRW this dependence is of order $1/\delta^{1.5}$ (slope 1.56) and of order $1/\delta^2$ (slope 2.01), respectively. These scalings partly verify the rates derived in Corollary 3.2 and demonstrate the gains of MALA over ULA for the weakly log-concave densities.

3.4.1.3 Warmness in simulations

Strictly speaking, for both the cases considered above, the starting distribution was not warm, since we used $\mu_\star$ as the starting distribution and the corresponding warmness $\beta =$

$\mathcal{O}(e^d)$ scales exponentially with dimension d. However, the mixing time observed in the simulations, albeit with a heuristic measure, are d times faster than those stated with $\mu_\star$ as the starting distribution in Corollary 3.1, and are in fact consistent with the results for the warm-start which are stated in Theorems 3.1 and 3.2. We believe that the results stated in Corollary 3.1, with $\mu_\star$ as the starting distribution, can be improved by a factor of d. However, our current proof techniques do not close this gap and we leave further investigation of this question for future work.

3.4.2 Behavior for Gaussian mixture distribution

We now consider the task of sampling from a two component Gaussian mixture distribution, as previously considered by Dalalyan [61] for illustrating the behavior of ULA. Here compare the behavior of MALA to ULA for this case. The target density is given by

$$x \mapsto \pi^\star(x) = \frac{1}{2\,(2\pi)^{d/2}}\left(e^{-\|x-a\|_2^2/2} + e^{-\|x+a\|_2^2/2}\right),$$

where $a \in \mathbb{R}^d$ is a fixed vector. This density corresponds to the two-mixture of equal weighted Gaussian $\mathcal{N}(a, \mathbb{I}_d)$ and $\mathcal{N}(-a, \mathbb{I}_d)$. In our notation, the function f and its derivatives are given by: $f(x) = \frac{1}{2}\|x-a\|_2^2 - \log(1+e^{-2x^\top a})$,

$$\nabla f(x) = x - a + 2a(1+e^{2x^\top a})^{-1}, \text{ and }, \nabla^2 f(x) = \mathbb{I}_d - 4aa^\top \frac{e^{2x^\top a}}{\left(1+e^{2x^\top a}\right)^2}.$$

From examination of the Hessian, we see that the function f is smooth with parameter $\mathcal{L} = 1$, and whenever $\|a\|_2 < 1$, it is also strongly convex with parameter $m = 1 - \|a\|_2^2$.

For dimension $d = 2$, setting $a = \left(\frac{1}{2}, \frac{1}{2}\right)$ yields the parameters $m = \frac{1}{2}$ and $\mathcal{L} = 1$. Figure 3.3 shows the level sets of the density of this 2D-Gaussian mixture. The initial distribution is chosen as $\mu_\star = \mathcal{N}\left(0, \mathcal{L}^{-1}\mathbb{I}_d\right)$ and the step-sizes are chosen according to Table 3.1, where for ULA, we set three different choices of $\delta = 0.2$ (ULA), $\delta = 0.1$ (small-step ULA) and $\delta = 1.0$ (large-step ULA). Note that choosing a smaller threshold δ implies that the ULA has a smaller step size and consequently the chain takes larger to converge. However, the asymptotic TV error with respect to the target distribution $\Pi^\star$ for ULA also decreases with decrease in step size. These different choices of step sizes are made to demonstrate the fundamental trade-off between the rate of convergence and asymptotic error for ULA and its inability to mix faster than MALA for different settings.

Note that one can sample directly from the mixture of Gaussian in consideration by drawing independently a Bernoulli(1/2) random variable y and a standard normal variable $z \sim \mathcal{N}(0, \mathbb{I}_d)$, and by computing

$$x = y \cdot (z - a) + (1 - y) \cdot (z + a)$$

This observation makes it easy to diagnose the convergence of our Markov chains with target $\pi^\star$. In order to estimate the total variation distance, we discretize the distribution

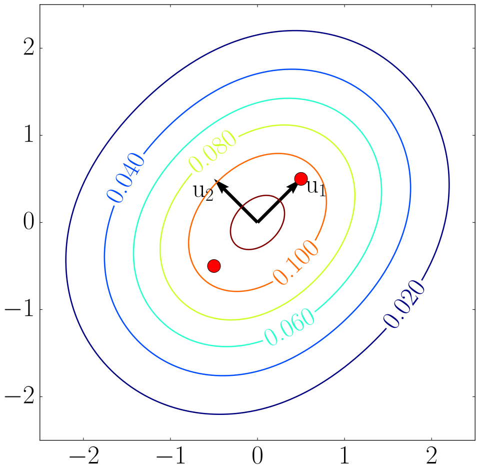


Figure 3.3: Level set of the density of the two-dimensional Gaussian mixture target. The red dots are the location of the means a and $-a$, where a is chosen such that $\|a\|_2^2 = \frac{1}{2}$. The arrows indicate the two principal directions u_1 and u_2 along which the TV error is measured.

of $N = 250,000$ samples from $\pi^\star$ over a set of bins, and consider the total variation of this discrete distribution from the empirical distribution of the Markov chain over these bins. We refer to this measure as the discretized TV error. We measure the sum of two discrete TV errors of $250,000$ samples from $\pi^\star$ with the empirical distribution obtained by simulating the chains ULA, MALA or MRW, projected on two principal directions (u_1 and u_2), over a discrete grid of size $B = 100$. Figure 3.4 shows the sum of the discretized TV errors along u_1 and u_2, as a function of iterations. The true total variation distance between the distribution of the iterate and the target distribution is upper bounded by the sum of (A) the discretized TV error and (B) the error caused by discretization. To obtain an idea of how large is the error (B) due to discretization, we simulate 100 runs of the discrete TV error between two independent drawings from the true distribution $\pi^\star$. The two black lines in Figure 3.4 are the maximum and minimum of these 100 values. The sample distribution at convergence is expected to lie between the two black lines.

Figure 3.4(a) shows that ULA converges significantly slower than MALA to the right distribution. Figure 3.4(b) illustrates this point further and shows that when compared to the ULA, the small-step ULA ($\delta = 0.1$) converges at a much slower rate and large-step ULA ($\delta = 1.0$) has a larger approximation error (asymptotic bias).

We accompany the study based on exact TV error computation with two classical convergence diagnostic plots for general MCMC algorithms. Figure 3.5 shows the traceplots of the three sampling algorithms in 10 runs. Comparing the three plots (Figure 3.5 (a), (b),

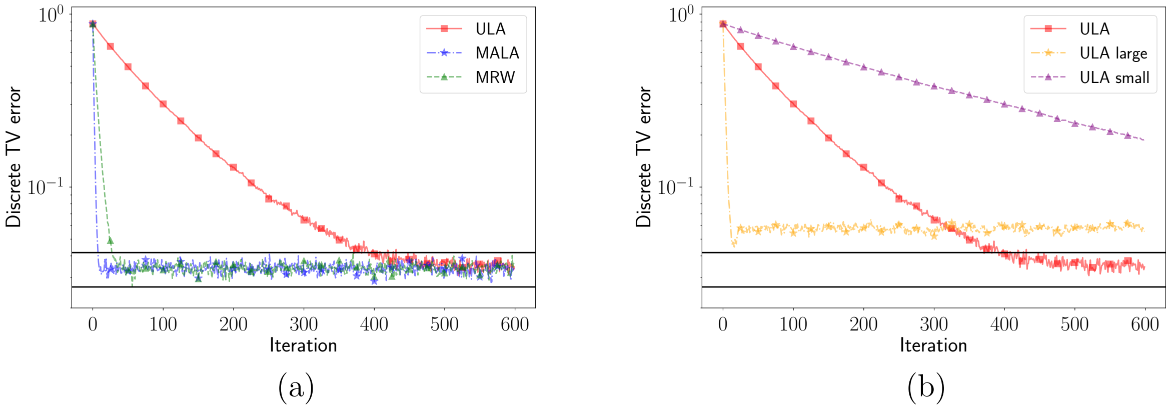


Figure 3.4: Discrete TV error on a two component Gaussian mixture target. (a) Behavior of three different random walks. (b) Behavior of ULA with different choices of step sizes.

(c)), we observe that the traceplot of MALA stabilizes much faster than that of ULA and MRW. Furthermore, to compare the efficiency of the chains in stationarity, Figure 3.6 shows the autocorrelation function of the three chains. To make sure that the computation is done in stationarity, we set in practice the burn-in period to be 300 iterations. Again, we observe that MALA is clearly significantly more efficient than ULA and MRW.

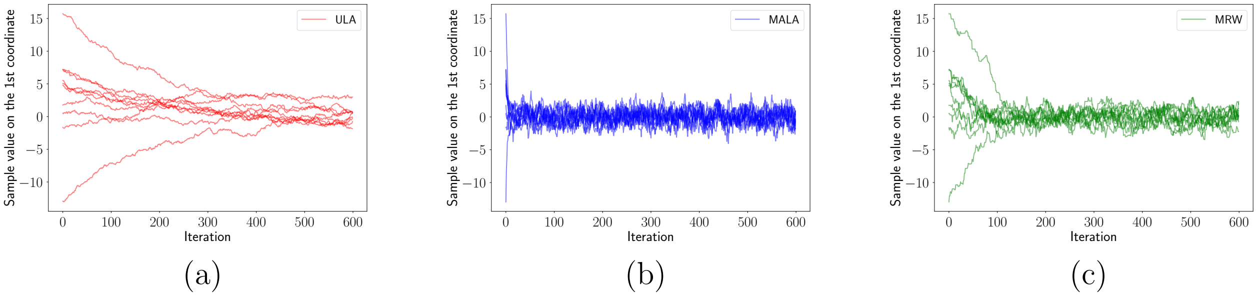


Figure 3.5: Trace-plot of the first coordinate on a two component Gaussian mixture target. (a) Trace-plot of ULA. (b) Trace-plot of MALA. (c) Trace-plot of MRW.

3.4.3 Bayesian Logistic Regression

We now consider the problem of logistic regression in a frequentist-Bayesian setting, similar to that considered by Dalalyan [61]. Once again, we establish that MALA has superior

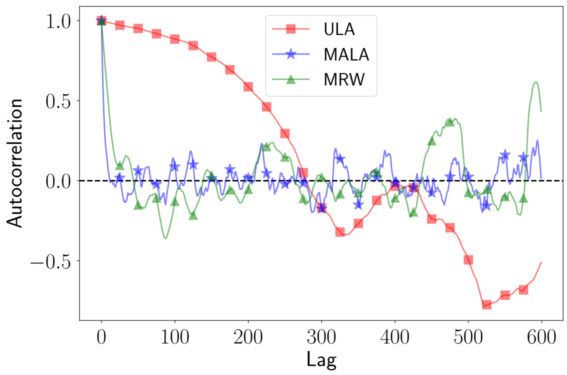


Figure 3.6: Markov chain autocorrelation function plot for two mixture of Gaussian target. The burn-in time for the plot is set to 300 iterations.

performance relative to ULA. Given a binary variable $y \in \{0, 1\}$ and a covariate $x \in \mathbb{R}^d$, the logistic model for the conditional distribution of y given x takes the form

$$\mathbb{P}(y = 1|x; \theta) = \frac{e^{\theta^\top x}}{1 + e^{\theta^\top x}}, \tag{3.15}$$

for some parameter $\theta \in \mathbb{R}^d$.

In a Bayesian framework, we model the parameter θ in the logistic equation as a random variable with a prior distribution $\pi_0^\star$. Suppose that we observe a set of independent samples $\{(x_i, y_i)\}_{i=1}^n$ with $(x_i, y_i) \in \mathbb{R}^d \times \{0, 1\}$, with each y_i conditioned on x_i drawn from a logistic distribution with some unknown parameter θ^*. Using Bayes' rule, we can then compute the posterior distribution of the parameter θ given the data. Drawing samples from this posterior distribution allows us to estimate and draw inferences about the unknown parameter. Under mild conditions, the Bernstein-von-Mises theorem guarantees that the posterior distribution will concentrate around the true parameter θ^*, in which case we expect that the credible intervals formed by sampling from the posterior should contain θ^* with high probability. This fact provides a lens for us to assess the accuracy of our sampling procedure.

Define the vector $Y = (y_1, \ldots, y_n)^\top \in \{0, 1\}^n$ and let X be the $n \times d$ matrix with x_i as i^{th}-row. We choose the prior π_0 to be a Gaussian distribution with zero mean and covariance matrix proportional to the inverse of the sample covariance matrix $\Sigma_X = \frac{1}{n} X^\top X$. Plugging in the formulas for the prior and likelihood, we find that the the posterior density is given by

$$\pi^\star(\theta) = \pi^\star(\theta|X, Y) \propto \exp\left\{ Y^\top X\theta - \sum_{i=1}^n \log\left(1 + e^{\theta^\top x_i}\right) - \alpha \left\| \Sigma_X^{1/2} \theta \right\|_2^2 \right\},$$

where $\alpha > 0$ is a user-specified parameter. Writing $\pi^\star \propto e^{-f}$, we observe that the function f and its derivatives are given by

$$\begin{aligned} f(\theta) &= -Y^\top X\theta + \sum_{i=1}^n \log\left(1 + e^{\theta^\top x_i}\right) + \alpha \left\|\Sigma_X^{1/2}\theta\right\|_2^2, \\ \nabla f(\theta) &= -X^\top Y + \sum_{i=1}^n \frac{x_i}{1 + e^{-\theta^\top x_i}} + \alpha\Sigma_X\theta, \quad \text{and,} \\ \nabla^2 f(\theta) &= \sum_{i=1}^n \frac{e^{-\theta^\top x_i}}{\left(1 + e^{-\theta^\top x_i}\right)^2} x_i x_i^\top + \alpha\Sigma_X. \end{aligned}$$

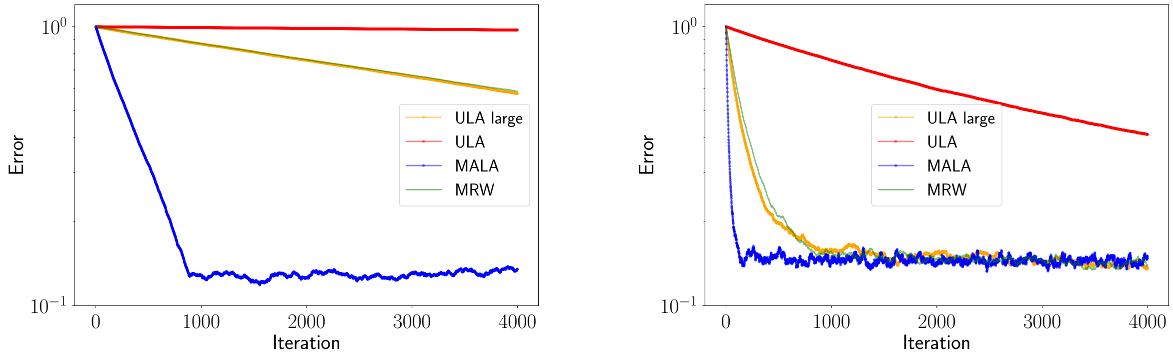
With some algebra, we can deduce that the eigenvalues of the Hessian $\nabla^2 f$ are bounded between $\mathcal{L} := (0.25n + \alpha)\, \lambda_{\max}(\Sigma_X)$ and $m := \alpha\, \lambda_{\min}(\Sigma_X)$ where $\lambda_{\max}(\Sigma_X)$ and $\lambda_{\min}(\Sigma_X)$ denote the largest and smallest eigenvalues of the matrix Σ_X. We make use of these bounds in our experiments.

As in the paper [61], we also consider a preconditioned version of the method; more precisely, we first sample from $\pi_g^\star \propto e^{-g}$ where $g(\theta) = f(\Sigma_X^{-1/2}\theta)$, and then transform the obtained random samples $\theta_i \mapsto \Sigma_X^{1/2}\theta_i$ to obtain samples from $\pi^\star$. Sampling based on the preconditioned distribution improves the condition number of the problem. After the preconditioning, we have the bounds $\mathcal{L}_g \leq 0.25n + \alpha$ and $m_g \geq \alpha$, so that the new condition number is now independent of the eigenvalues of Σ_X.

We randomly draw i.i.d. samples (x_i, y_i) as follows. Each vector $x_i \in \mathbb{R}^d$ is sampled i.i.d. Rademacher components, and then renormalized to Euclidean norm. given x_i, the response y_i is drawn from the logistic model (3.15) with $\theta = \theta^* = \mathbf{1}_d = (1, \ldots, 1)^\top$. We fix $d = 2, n = 50$ and perform $N = 1000$ experiments. To sample from the posterior, we start with the initial distribution as $\mu_0 = \mathcal{N}(0, \mathcal{L}^{-1}\mathbb{I}_d)$. As the first error metric, we measure the ℓ_1 distance between the true parameter θ^* and the sample mean $\hat{\theta}_k$ of the random samples obtained from simulating the Markov chains for k iterations:

$$e_k = \frac{1}{d}\|\hat{\theta}_k - \theta^*\|_1.$$

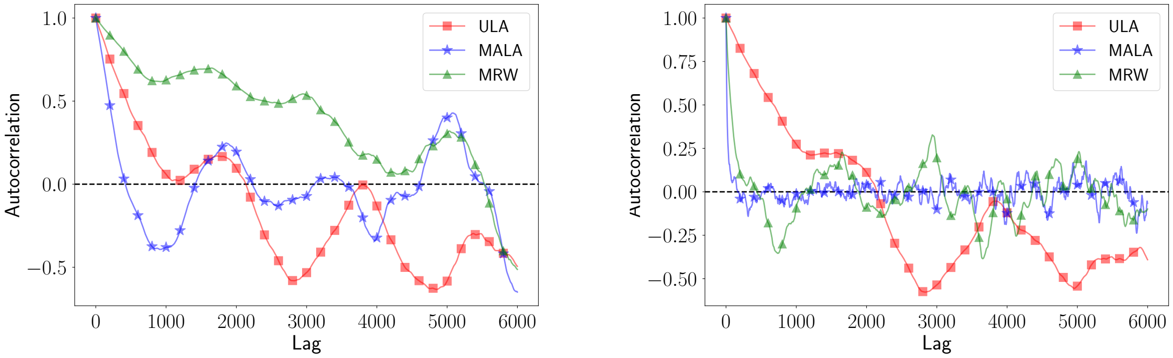
Figure 3.7 shows this error as a function of iteration number in logarithmic scale. Since there is always an approximation error caused by the prior distribution, ULA with large step-size ($\delta = 1.0$) can be used. However, our simulation shows that it is still slower than MALA. Also, the condition number κ has a significant effect on the mixing time of ULA and MRW. Their convergence in the preconditioned case is significantly better. Furthermore, the autocorrelation plots in Figure 3.8 and the plots in Figure 3.9 of the sample (across experiments) mean and 25% and 75% quantiles, with θ^* subtracted, as a function of iterations suggest a similar story: MALA converges faster than ULA and is less affected by conditioning of the problem.



(a)

(b)

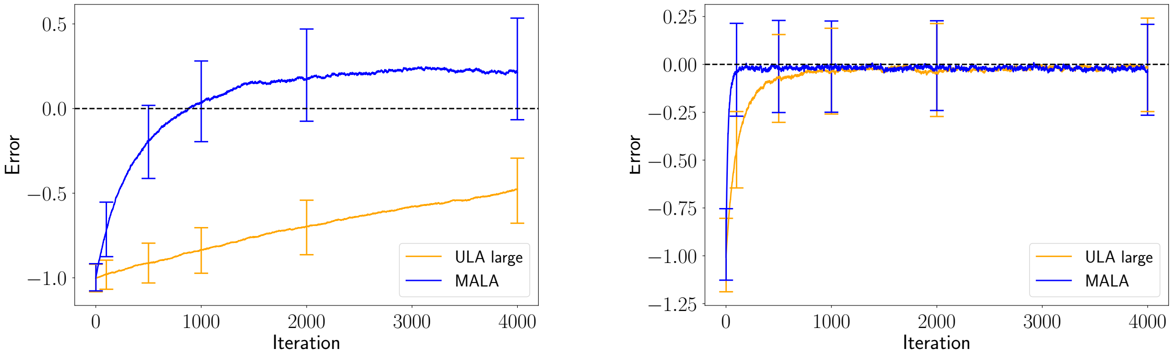
Figure 3.7: Mean error for Bayesian logistic regression as a function of iteration number. (a) Without preconditioning. (b) With preconditioning.



(a)

(b)

Figure 3.8: Markov chain autocorrelation plot of the first coordinate of the Bayesian logistic regression parameter estimate as a function of lag. The burn-in time for the plot is set to 300 iterations. (a) Without preconditioning. (b) With preconditioning.



(a) (b)

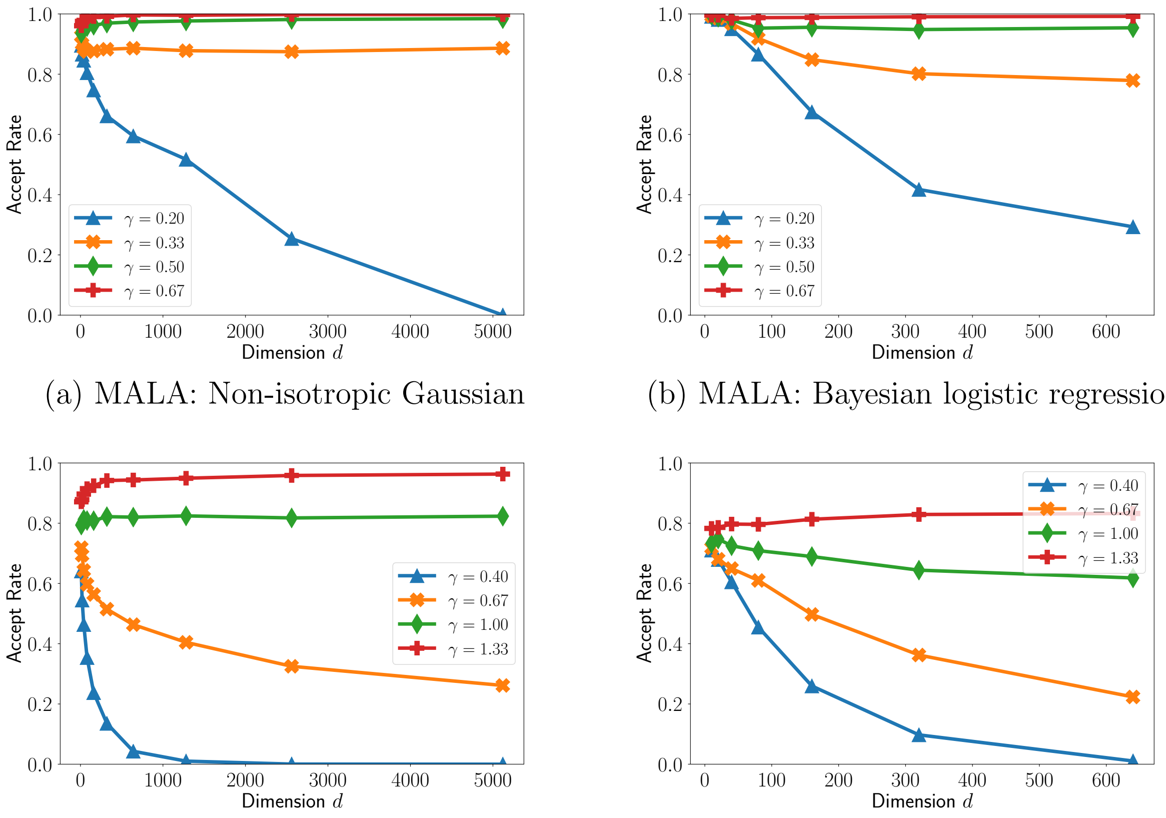
Figure 3.9: Visualizing the mean (solid lines) and 25% and 75% quantiles (as error bars) for the first coordinate of the Bayesian logistic regression parameter estimate, with θ^* subtracted, as a function of iteration number. (a) Without preconditioning. (b) With preconditioning.

3.4.4 Step size vs accept-reject rate

In this section, we provide a few simulations that highlight the effect of step size for MALA and MRW. Note that our bounds from Theorem 3.1 and 3.2 suggest a step size choice of order d^{-1} for both MALA and MRW, which in turn led to the mixing time bounds of $\mathcal{O}(d)$. These choices of step sizes arise when we try to provide a worst-case control on the accept-reject step of these algorithms. In particular, these choices ensure that the Markov chains do not get stuck at a given state x, or equivalently, that the proposals at any given state are accepted with constant probability. If instead, one chooses a very large step size, the (worst-case) probability of acceptance may decay exponentially with dimensions. Nonetheless, these worst case bounds may not hold, which would imply a faster mixing time for these chains if a larger step size were to be used.

To check the validity of larger step sizes, we repeated a few experiments discussed above, albeit with a larger step size. In particular, we simulated the random walks for a wide-range of step sizes $d^{-\gamma}$ for $\gamma \in \{0.2, 0.33, 0.5, 0.67\}$ for MALA, and $\gamma \in \{0.4, 0.67, 1, 1.33\}$ for MRW. We ran these chains for two different cases: (a) Sampling from non-isotropic Gaussian density, discussed in Section 3.4.1, and, (b) Posterior sampling in Bayesian logistic regression, discussed in Section 3.4.3). In Figure 3.10, we plot the average acceptance probability for different step sizes $d^{-\gamma}$ as the dimension d increases. These probabilities were computed as the average number of proposals accepted over 100 iterations after a manually tuned burn-in period, and further averaged across 50 independent runs.

We now remark on the observations from Figure 3.10. We see that for MALA the acceptance probability for the step size choice of $d^{-0.2}$ vanishes as d increases. Indeed, the choice of $d^{-0.5}$ appears to be a safe choice for both cases. In contrast, for MRW, we need a



(a) MALA: Non-isotropic Gaussian

(b) MALA: Bayesian logistic regression

(c) MRW: Non-isotropic Gaussian

(d) MRW: Bayesian logistic regression

Figure 3.10: Effect of large step size for accept-reject ratio for MALA and MRW. From panels (a) and (b), we see that for MALA the step size choice of $d^{-0.5}$ has a non-vanishing acceptance probability rate for both cases. On the other hand, panels (c) and (d) show that for MRW d^{-1} is a good choice for the step size.

smaller step size. From panels (c) and (d), we see that d^{-1} appears to be the correct choice to ensure that the proposal are accepted with a constant probability when the dimension d is large.

Informally, if a step size choice of $d^{-\gamma}$ were to guarantee a non-vanishing acceptance probability for MALA or MRW, our proof techniques imply a mixing time bound of $\mathcal{O}\left(d^{\gamma}\right)$. Combining this argument with the observations above, we suspect that the bounds for MALA from Theorem 3.1 may not be tight, and a $\sqrt{d}$-scaling is plausible, while for MRW the bounds from Theorem 3.2 are very likely to be tight. Deriving a faster mixing time for MALA or establishing that the current dimension dependency for MRW is tight, are interesting research directions; see [54] for a very recent work showing that $\sqrt{d}$ mixing time for MALA may be tight for certain class of distributions.

3.5 Proofs

We now turn to the proofs of our main results—which is based on conductance-based proof techniques for establishing mixing time bounds. In Section 3.5.1, we provide a brief background on conductance based proof techniques, followed by several auxiliary lemmas in Section 3.5.2 which then enable us to easily prove Theorems 3.1 and 3.2 in Sections 3.5.3 and 3.5.4. Proofs of the auxiliary results is deferred to Appendix A.

3.5.1 Conductance-based mixing time bounds

Our proofs exploit standard conductance-based arguments for controlling mixing times. Consider an ergodic Markov chain defined by a transition operator $\mathcal{T}$, and let $\Pi^\star$ be its stationary distribution. For each scalar $s \in (0, 1/2)$, we define the s-conductance

$$\Phi_s := \inf_{\Pi^\star(\mathcal{S}) \in (s, 1-s)} \frac{\int_{\mathcal{S}} \mathcal{T}_u(\mathcal{S}^c)\pi^\star(u)du}{\min\left\{\Pi^\star(\mathcal{S}) - s, \Pi^\star(\mathcal{S}^c) - s\right\}}. \tag{3.16}$$

In this formula, the notation $\mathcal{T}_u$ is shorthand for the distribution $\mathcal{T}(\boldsymbol{\delta}_u)$ obtained by applying the transition operator to a dirac distribution concentrated on u. In words, the s-conductance measures how much probability mass flows across disjoint sets relative to their stationary mass. By a continuity argument, it can be seen that limiting conductance of the chain is equal to the limiting value of s-conductance—that is, $\Phi = \lim_{s \to 0} \Phi_s$.

Lemma 3.2 (**Lovász** [**137**, **163**])**.** *A reversible lazy Markov chain with stationary distribution $\Pi^\star$ and a β-start μ_0 satisfies*

$$d_{\mathrm{TV}}\left(\mathcal{T}^\ell(\mu_0), \Pi^\star\right) \leq \beta s + \beta\left(1 - \frac{\Phi_s^2}{2}\right)^\ell \leq \beta s + \beta e^{-\ell\Phi_s^2/2} \quad \textit{for any} \quad s \in \left(0, \frac{1}{2}\right), \tag{3.17}$$

where Φ_s denotes the s-conductance (3.16) of the chain.

Thus, it suffices to lower bound the s-conductance Φ_s, and then substitute a suitable parameter s so as to optimize the tradeoff between the two terms in the bound. In particular, Lemma 3.2 implies that the δ-TV mixing time of a chain is bounded above from

$$\frac{2\log(2\beta/\delta)}{\Phi_s^2} \quad \text{with} \quad s = \frac{\delta}{2\beta}. \tag{3.18}$$

3.5.2 Auxiliary results for MALA and MRW

We now state some intermediate lemmas that are useful to establish on Φ_s for MALA and MRW. For the remainder of this section, we assert the assumption (3A) on the target, namely that $\Pi^\star$ denotes an $(\mathcal{L}, m)$-strongly log-concave distribution.

We start with a result that shows that the probability mass of any strongly log-concave distributions is concentrated in a Euclidean ball around the mode. For each $s \in (0,1)$, we introduce the Euclidean ball

$$\mathcal{R}_s = \mathbb{B}\left(x^\star, \mathfrak{a}(s)\sqrt{\frac{d}{m}}\right) \tag{3.19}$$

where the function $\mathfrak{a}$ was previously defined in equation (3.6a), and $x^\star := \arg\max_{x\in\mathbb{R}^d} \pi^\star(x)$ denotes the unique mode.

Lemma 3.3. *For any $s \in \left(0, \frac{1}{2}\right)$, we have $\Pi^\star(\mathcal{R}_s) \geq 1 - s$.*

See Section A.2 for the proof of this claim.

In order to establish the conductance bounds inside this ball, we prove an extension of a result by Lovász [163]. In particular, the next result provides a lower bound on the flow of Markov chain with transition distribution $\mathcal{T}_x$ and strongly log-concave target distributions $\Pi^\star$. Similar results have been used in several prior works to establish fast mixing of several random walks like ball walk, Hit and run [163, 169, 171], Dikin walk [187], and later in Chapter 6 for the analysis of Vaidya and John walks.

Lemma 3.4. *Let $\mathcal{S}$ be a convex set such that $d_{\mathrm{TV}}\left(\mathcal{T}_x, \mathcal{T}_y\right) \leq 1 - \rho$ whenever $x, y \in \mathcal{S}$ and $\left\|x - y\right\|_2 \leq \Delta$. Then for any measurable partition $\mathcal{S}_1$ and $\mathcal{S}_2$ of $\mathbb{R}^d$, we have*

$$\int_{\mathcal{S}_1} \mathcal{T}_u(\mathcal{S}_2)\pi^\star(u)du \geq \frac{\rho}{4}\min\left\{1, \frac{\log 2 \cdot \Delta \cdot (\Pi^\star(\mathcal{S}))^2 \cdot \sqrt{m}}{8}\right\}\min\left\{\Pi^\star(\mathcal{S}_1 \cap \mathcal{S}), \Pi^\star(\mathcal{S}_2 \cap \mathcal{S})\right\}. \tag{3.20}$$

See Section A.3 for the proof of this lemma.

Define the function $\widetilde{\mathfrak{t}} : (0,1) \times (0,1) \to \mathbb{R}_+$ as follows:

$$\widetilde{\mathfrak{t}}(s, \varepsilon) := \min\left\{\frac{\sqrt{\varepsilon}}{8\sqrt{2}\mathfrak{a}(s)}\frac{\sqrt{m}}{\mathcal{L}\sqrt{d\mathcal{L}}}, \frac{\varepsilon}{64\alpha_\varepsilon}\frac{1}{\mathcal{L}d}, \frac{\varepsilon^{2/3}}{26(\alpha_\varepsilon \mathfrak{a}^2(s))^{1/3}}\frac{1}{\mathcal{L}}\left(\frac{m}{\mathcal{L}d^2}\right)^{1/3}\right\}, \tag{3.21a}$$

$$\text{where} \quad \alpha_\varepsilon := 1 + 2\sqrt{\log(16/\varepsilon)} + 2\log(16/\varepsilon). \tag{3.21b}$$

The next lemma shows two important properties for MALA: (1) the proposal distributions at two different points are close if the two points are close, and (2) the accept-reject step is well behaved inside the ball $\mathcal{R}_s$ provided the step size is chosen carefully. Note that for MALA, the proposal distribution of the chain at x is given by

$$\mathcal{P}_x^{\mathrm{MALA}(\eta)} := \mathcal{N}(\mu_x, 2\eta\mathbb{I}_d), \quad \text{where} \quad \mu_x = x - \eta\nabla f(x). \tag{3.22}$$

We use $\mathcal{T}_x^{\mathrm{MALA}(\eta)}$ to denote the transition distribution of MALA (obtained after performing the accept-reject step (2.7)).

Lemma 3.5. *For any step size $\eta \in \left(0, \frac{2}{\mathcal{L}}\right]$, the MALA proposal distribution satisfies*

$$d_{\mathrm{TV}}\left(\mathcal{P}_x^{\mathrm{MALA}(\eta)}, \mathcal{P}_y^{\mathrm{MALA}(\eta)}\right) \leq \frac{\|x-y\|_2}{\sqrt{2\eta}}, \quad \text{for all} \quad x, y \in \mathbb{R}^d. \tag{3.23a}$$

Furthermore, for any $s \in (0, 1/2)$ and $\varepsilon \in (0, 1)$, the MALA proposal and transition distributions satisfy

$$\sup_{\eta \in [0, \widetilde{\mathfrak{t}}(s,\varepsilon)]} \sup_{x \in \mathcal{R}_s} d_{\mathrm{TV}}\left(\mathcal{P}_x^{\mathrm{MALA}(\eta)}, \mathcal{T}_x^{\mathrm{MALA}(\eta)}\right) \leq \frac{\varepsilon}{8}, \tag{3.23b}$$

where the truncated ball $\mathcal{R}_s$ was defined in equation (3.19).

See Section A.4 for the proof.

With these results in hand, we now prove the mixing time bound for MALA.

3.5.3 Proof of Theorem 3.1

At a high level, the proof involves three key steps. Our first step is to use Lemma 3.5 to establish that for an appropriate choice of step size, the MALA update has nice properties inside the region $\mathcal{R}_s$, which admits a high probability under $\Pi^\star$ thanks to Lemma 3.3. The second step is to apply Lemma 3.4 to obtain a lower bound on the s-conductance Φ_s for the MALA update. Finally, by choosing s as in equation (3.18), we establish the claimed convergence rate.

We drop the superscripts MALA(η) from our notation—that is, we use $\mathcal{T}_x$ and $\mathcal{P}_x$, respectively, to denote the transition and proposal distributions at x for MALA, each with step size η. By applying the triangle inequality, we obtain the upper bound

$$d_{\mathrm{TV}}\left(\mathcal{T}_x, \mathcal{T}_y\right) \leq d_{\mathrm{TV}}\left(\mathcal{P}_x, \mathcal{T}_x\right) + d_{\mathrm{TV}}\left(\mathcal{P}_x, \mathcal{P}_y\right) + d_{\mathrm{TV}}\left(\mathcal{P}_y, \mathcal{T}_y\right). \tag{3.24}$$

Now applying claim (3.23a) from Lemma 3.5 guarantees that

$$d_{\mathrm{TV}}\left(\mathcal{P}_x, \mathcal{P}_y\right) \leq \varepsilon/\sqrt{2} \qquad \text{for all } x, y \in \mathbb{R}^d \text{ such that } \|x-y\|_2 \leq \varepsilon\sqrt{\eta}.$$

Furthermore, for any $\eta \leq \widetilde{\mathfrak{t}}(s, \varepsilon)$, the bound (3.23b) from Lemma 3.5 implies that $d_{\mathrm{TV}}\left(\mathcal{P}_x, \mathcal{T}_x\right) \leq \varepsilon/8$ for any $x \in \mathcal{R}_s$. Plugging in these bounds in the inequality (3.24), we find that

$$d_{\mathrm{TV}}\left(\mathcal{T}_x, \mathcal{T}_y\right) \leq 1 - (1-\varepsilon) \quad \forall\, x, y \in \mathcal{R}_s \text{ such that } \|x-y\|_2 \leq \varepsilon\sqrt{\eta}.$$

Thus, the transition distribution $\mathcal{T}_x$ satisfies the assumptions of Lemma 3.4 for

$$\mathcal{S} = \mathcal{R}_s, \quad \rho = (1-\varepsilon) \quad \text{and} \quad \Delta = \varepsilon\sqrt{\eta}. \tag{3.25}$$

We now derive a lower bound on the s-conductance of MALA. Choosing a measurable set $\mathcal{S}$ such that $\Pi^\star(\mathcal{S}) > s$ and substituting the terms from equation (3.25) in the inequality (3.20), we find that

$$\begin{aligned}\int_{\mathcal{S}} \mathcal{T}_u(\mathcal{S}^c)\pi^\star(u)du &\geq \frac{(1-\varepsilon)}{4} \min\left\{1, \frac{\log 2 \cdot \varepsilon\sqrt{\eta} \cdot (\Pi^\star(\mathcal{R}_s))^2 \cdot \sqrt{m}}{8}\right\} \cdot \\ &\quad \cdot \min\left\{\Pi^\star(\mathcal{S} \cap \mathcal{R}_s), \Pi^\star(\mathcal{S}^c \cap \mathcal{R}_s)\right\} \\ &\overset{(i)}{\geq} \frac{(1-\varepsilon)\varepsilon\sqrt{\eta} \cdot (\Pi^\star(\mathcal{R}_s))^2 \cdot \sqrt{m}}{64} \min\left\{\Pi^\star(\mathcal{S}) - s, \Pi^\star(\mathcal{S}^c) - s\right\}.\end{aligned}$$

In this argument, inequality (i) follows from the facts that $\log 2 \geq 1/2$ and $\Pi^\star(\mathcal{S}), \Pi^\star(\mathcal{S}^c) > s$. Moreover, we have applied Lemma 3.3 to find that $\Pi^\star(\mathcal{R}_s) \geq 1 - s$ and hence

$$\Pi^\star(\mathcal{X} \cap \mathcal{R}_s) = \Pi^\star(\mathcal{X}) - \Pi^\star(\mathcal{X} \cap \mathcal{R}_s^c) \geq \Pi^\star(\mathcal{X}) - s \quad \text{for } \mathcal{X} \in \{\mathcal{S}, \mathcal{S}^c\}.$$

We have also assumed that the second argument of the minimum is less than 1. Applying the definition (3.16) of Φ_s for MALA, we find that

$$\Phi_s^{\mathrm{MALA}(\eta)} \geq \frac{(1-\varepsilon)\varepsilon \cdot (\Pi^\star(\mathcal{R}_s))^2 \cdot \sqrt{\eta m}}{64}, \qquad \text{for any } \eta \leq \widetilde{\mathfrak{t}}(s, \varepsilon). \tag{3.26}$$

Finally, Using Lemma 3.3, we have that $\Pi^\star(\mathcal{R}_{\delta/2}) \geq 1 - \delta/2 \geq 1/2$ for any $\delta \in (0, 1)$. Applying the definition (3.21b) of α_ε, we obtain that $\alpha_{1/2} \leq 12$. Using this fact and the definitions (3.6b) and (3.21a) for the functions $\mathfrak{t}(\cdot)$ and $\widetilde{\mathfrak{t}}(\cdot, \cdot)$, it is straightforward to verify that $c\mathfrak{t}(\delta/(2\beta)) \leq \widetilde{\mathfrak{t}}(\delta/(2\beta), 1/2)$, for an appropriate choice of universal constant c. Substituting in $s = \delta/(2\beta)$, $\varepsilon = 1/2$, and $\eta = c\mathfrak{t}(\delta/(2\beta))$, and also making use of the lower bound $\Pi^\star(\mathcal{R}_{\delta/2\beta}) \geq 1/2$ in the bound (3.26), we find that $\Phi_{\delta/2\beta}^{\mathrm{MALA}(\eta)} \geq c'\sqrt{m\eta}$ for some universal constant c'. Substituting the pieces together in the bound (3.18) yields the claim.

3.5.4 Proof of Theorem 3.2

The proof of this theorem is similar to the proof of Theorem 3.1. We begin by claiming that

$$d_{\mathrm{TV}}\big(\mathcal{P}_x^{\mathrm{MRW}(\eta)}, \mathcal{P}_y^{\mathrm{MRW}(\eta)}\big) = \frac{\varepsilon}{\sqrt{2}} \qquad \text{for all } x, y \text{ such that } \|x - y\|_2 \leq \varepsilon\sqrt{\eta}, \tag{3.27a}$$

$$d_{\mathrm{TV}}\big(\mathcal{P}_x^{\mathrm{MRW}(\eta)}, \mathcal{T}_x^{\mathrm{MRW}(\eta)}\big) = \frac{\varepsilon}{8} \qquad \text{for all } x \in \mathcal{R}_s, \tag{3.27b}$$

for any $\eta \leq c\varepsilon^2 m/(\alpha_\varepsilon d\mathcal{L}^2\mathfrak{a}(s))$ for some universal constant c. Plugging $s = \delta/(2\beta)$, $\varepsilon = 1/2$ and arguing as in Section 3.5.3, we find that $\Phi_{\delta/2\beta}^{\mathrm{MRW}(\eta)} \geq c'\sqrt{m\eta}$ for some universal constant c'. Using the bound (3.18) yields the claimed bound on the mixing time of MRW.

Next, we prove claims (3.27a) and (3.27b). Note that $\mathcal{P}_x^{\mathrm{MRW}(\eta)} = \mathcal{N}(x, 2\eta\mathbb{I}_d)$. For brevity, we drop the superscripts from our notations. Using the expression (A.13) for the KL-divergence and applying Pinsker's inequality leads to the upper bound

$$d_{\mathrm{TV}}\left(\mathcal{P}_x, \mathcal{P}_y\right) \leq \sqrt{2\,\mathrm{KL}(\mathcal{P}_x \| \mathcal{P}_y)} = \frac{\|x-y\|_2}{\sqrt{2\eta}},$$

which implies the claim (3.27a).

For the other bound (3.27b), letting p_x to denote the density of the proposal distribution $\mathcal{P}_x$ and using the bounds (A.16) and (E.16), it suffices to prove that

$$\mathbb{P}_{z\sim\mathcal{P}_x}\left[\frac{\pi^\star(z)}{\pi^\star(x)} \geq \exp\left(-\frac{\varepsilon}{16}\right)\right] \overset{(i)}{=} \mathbb{P}_{z\sim\mathcal{P}_x}\left[f(x) - f(z) \geq -\frac{\varepsilon}{16}\right] \geq (1 - \varepsilon/16), \tag{3.28}$$

where step (i) follows from the fact that $\pi^\star(x) \propto e^{-f(x)}$. We have

$$\begin{aligned} f(x) - f(z) &\overset{(i)}{\geq} \nabla f(z)^\top (x-z) = \left(\nabla f(z) - \nabla f(x)\right)^\top (x-z) + \nabla f(x)^\top (x-z) \\ &\overset{(ii)}{\geq} -\mathcal{L}\left\|x-z\right\|_2^2 + \nabla f(x)^\top (x-z) \\ &= -2\mathcal{L}\eta\left\|\xi\right\|_2^2 + \sqrt{2\eta}\nabla f(x)^\top \xi, \end{aligned} \tag{3.29}$$

where the step (i) follows from the convexity of the function f, step (ii) the smoothness of the function f (Lemma A.2(e)). Note that the random variable $\chi := \nabla f(x)^\top \xi \sim \mathcal{N}(0, \|\nabla f(x)\|_2^2)$ and that $\|\nabla f(x)\|_2 \leq \mathcal{D}_s$ for any $x \in \mathcal{R}_s$. Consequently, we have $\chi \geq -\mathcal{D}_s \cdot 2\sqrt{\log(32/\varepsilon)}$ with probability at least $1 - \varepsilon/32$. On the other hand, using the standard tail bound for a Chi-squared random variable, we obtain that $\mathbb{P}\left[\|\xi\|_2^2 \geq d\alpha_\varepsilon\right] \leq \varepsilon/32$ for $\alpha_\varepsilon = 1 + 2\sqrt{\log(32/\varepsilon)} + 2\log(32/\varepsilon)$. Recalling that $\mathcal{D}_s = \mathcal{L}\sqrt{\frac{d}{m}}\mathfrak{a}(s)$ and doing straightforward calculation reveals that for $\eta \leq \frac{\varepsilon^2}{(8192\alpha_\varepsilon d\frac{\mathcal{L}^2}{m}\mathfrak{a}(s))}$, we have

$$2\mathcal{L}\eta d\alpha_\varepsilon \leq \frac{\varepsilon}{64}, \quad \text{and} \quad \sqrt{2\eta}\mathcal{D}_s 2\sqrt{\log(32/\varepsilon)} \leq \frac{3\varepsilon}{64}.$$

Combining these bounds with the high probability statements above and plugging in the inequality (3.29), we find that $f(x) - f(z) \geq -\varepsilon/16$ with probability at least $1 - \varepsilon/16$, which yields the claim (3.28).

3.6 Conclusion and future directions

In this chapter, we derived non-asymptotic bounds on the mixing time of the Metropolis adjusted Langevin algorithm and Metropolized random walk for log-concave distributions. These algorithms are based on a two-phase scheme: (1) a proposal step followed by (2) an accept-reject step. Our results show that the accept-reject step while leading to significant complications in the analysis is practically very useful: algorithms applying this step mix significantly faster than the ones without it. In particular, we showed that for a strongly log-concave distribution in $\mathbb{R}^d$ with condition number κ, the δ-mixing time for MALA is of $\mathcal{O}\left(d\kappa\log(1/\delta)\right)$. This guarantee significantly better than the $\mathcal{O}\left(d\kappa^2/\delta^2\right)$ mixing time for ULA established in the literature. We also proposed a modified version of MALA to sample from non-strongly log-concave distributions and showed that it mixes in $\mathcal{O}\left(d^2/\delta^{1.5}\right)$; thus, this algorithm dependency on the desired accuracy δ when compared to the $\mathcal{O}\left(d^3/\delta^4\right)$ mixing time for ULA for the same task. Furthermore, we established $\mathcal{O}\left(d\kappa^2\log(1/\delta)\right)$ mixing time bound for the Metropolized random walk for log-concave sampling.

Several fundamental questions arise from our work. All of our results are upper bounds on mixing time, and our simulation results suggest that they are tight for the choice of step size used in the Theorem statements. Simulations from **??** suggest that MALA might mix faster with a larger step size, so as to admit a $\sqrt{d}$ scaling of mixing time under certain settings; some very recent work [54] show that the $\sqrt{d}$ scaling observed from these experiments is tight for a class of distributions.

Simulations from **??** suggest that the warmness parameter β does not affect the choice of step size too much and hence potentially larger choices of step sizes (and thereby faster mixing) are possible. To this end, the theory developed in the next chapter provides significant improvements (Corollary 4.2), primarily by showing that for certain class of target distributions the mixing time dependence on the warmness parameter can be improved from $\log\beta$ to $\log\log\beta$. Moreover, for a deterministic start, one may consider running ULA for a few steps run to obtain moderate accuracy, and then run MALA initialized with the ULA iterates (thereby providing a warm start to MALA). In practice, we find that this hybrid procedure can generate highly accurate samples in reasonably few number of iterations. A formal analysis of such a method

Another open question is to sharply delineate the fundamental gap between the mixing times of first-order sampling methods and that of zeroth-order sampling methods. Noting that MALA is a first-order method while MRW is a zeroth-order method, from our work, we obtain that two class of methods differ in a factor of the condition number κ of the target distribution. It is an interesting question to determine whether this represents a sharp gap between these two classes of sampling methods.

Chapter 4

Proof Techniques for Improving Mixing Time Guarantees

The analysis in Chapter 3 made use of a fairly standard approach to controlling mixing times, namely, via worst-case conductance bounds. This method was introduced by Jerrum and Sinclair [131] for discrete space chains and then extended to the continuous space settings by Lovász and Simonovits [167], and has been thoroughly studied. Interested readers can refer to the survey [242] and the references therein for a detailed discussion of conductance based methods for continuous space Markov chains.

Indeed, many mixing time proof techniques for the convergence of continuous-state Markov chains are inspired by the large body of work on discrete-state Markov chains; for instance, see the surveys [164, 4] and references therein. Historically, much work has been devoted to improving the mixing time dependency on the initial distribution. For discrete-state Markov chains, Diaconis and Saloff-Coste [70] were the first to show that the logarithmic dependency of the mixing time of a Markov chain on the warmness parameter β (2.6) of the starting distribution can be improved to double-logarithmic. This improvement—from logarithmic to doubly logarithmic—allows for a good bound on the mixing time even when starting distribution is not available. The innovation underlying this improvement is the use of log-Sobolev inequalities in place of the usual isoperimetric inequality. Later, closely related ideas such as average conductance [165, 136], evolving sets [183] and spectral profile [99] were shown to be effective for reducing dependence on initial conditions for discrete space chains. Thus far, only the notion of average conductance [165, 136] has been adapted to continuous-state Markov chains so as to sharpen mixing time analysis of the Ball walk [166].

The goal of this chapter to build on the discrete state space Markov chain literature, and establish refined conductance based results for continuous state space Markov chains, which in turn can then provide mixing time guarantees that scale doubly logarithmic in the warmness parameter β (2.6). In particular, we extend one of the conductance profile techniques from the paper [99] from discrete state to continuous state chains, albeit with several appropriate modifications suited for the general setting.

Our contributions and organization: In this chapter, we derive three main results. First, in Proposition 4.1, we establish a general mixing time bound in terms of the conductance profile—a more refined quantity than the worst-case conductance (3.16). see Section 4.1). Then, we state Proposition 4.2 that lower bounds the conductance profile for an isoperimetric target (D.46) in terms of the transition overlaps for a Markov chain in Section 4.2. Doing so involves non-trivial extensions of ideas from discrete state Markov chains to those in continuous state spaces. Our results enable us to obtain simultaneous improvements on mixing time bounds of several Markov chains (for general continuous-state space) when the starting distribution is far from the stationary distribution. Consequentially, we improve upon the previous mixing time bounds for MRW and MALA from Chapter 3, when the starting distribution is not warm with respect to the target distribution in Section 4.3; e.g., compare Corollary 3.1 and Corollary 4.2. The machinery developed in this chapter is later used in Chapter 5 to establish fast mixing time bounds for the Hamiltonian Monte Carlo for a range of target distributions.

4.1 Mixing Time Via Conductance Profile

We start by setting up some notation, and a brief background on conductance profile.

Given a Markov chain with transition probability $\Theta : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to \mathbb{R}$, its stationary *flow* $\phi : \mathcal{B}(\mathcal{X}) \to \mathbb{R}$ is defined as

$$\phi(S) = \int_{x \in S} \Theta(x, S^c)\pi^\star(x)dx \quad \text{for any } S \in \mathcal{B}(\mathcal{X}). \tag{4.1}$$

Given a set $\Omega \subset \mathcal{X}$, the Ω*-restricted conductance profile* is given by

$$\Phi_\Omega(v) = \inf_{\Pi^\star(S\cap\Omega)\in(0,v]} \frac{\phi(S)}{\Pi^\star(S\cap\Omega)} \qquad \text{for any } v \in \big(0,\ \Pi^\star(\Omega)/2\big]. \tag{4.2}$$

The classical conductance constant Φ (also defined as $\Phi = \Phi_0$ in the notation from equation (3.16)) is a special case; it can be expressed as $\Phi = \Phi_{\mathcal{X}}(\frac{1}{2})$. In fact, we can see conductance profile as a size-wise conductance for varying sizes in terms of the target measure of the set.

Next, we define the *truncated extension* $\widetilde{\Phi}_\Omega$ of the function Φ_Ω to the positive real line as

$$\widetilde{\Phi}_\Omega(v) = \begin{cases} \Phi_\Omega(v), & v \in \left(0, \frac{\Pi^\star(\Omega)}{2}\right] \\ \Phi_\Omega(\Pi^\star(\Omega)/2), & v \in \left[\frac{\Pi^\star(\Omega)}{2}, \infty\right). \end{cases} \tag{4.3}$$

The set Ω is chosen suitably in the discussion to follow.

Smooth chain assumption: We say that the Markov chain satisfies the *smooth chain assumption* if its transition probability function $\Theta : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to \mathbb{R}_+$ can be expressed in the form

$$\Theta(x, dy) = \theta(x, y)dy + \alpha_x \delta_x(dy) \quad \text{for all } x, y \in \mathcal{X}, \tag{4.4}$$

where θ is the transition kernel satisfying $\theta(x, y) \geq 0$ for all $x, y \in \mathcal{X}$. Here δ_x denotes the Dirac-delta function at x and consequently, α_x denotes the one-step probability of the chain to stay at its current state x. Note that the three algorithms discussed in this chapter (MRW, MALA and HMC) all satisfy the smooth chain assumption (4.4). Throughout this chapter, when dealing with a general Markov chain, we assume that it satisfies the smooth chain assumption.

We now state our first main result that provides a control on the mixing time of a Markov chain with continuous-state space in terms of its restricted conductance profile. We show that this control (based on conductance profile) allows us to have a better initialization dependency than the usual conductance based control (see [166, 167, 78]). This method for sharpening the dependence is known for discrete-state Markov chains; to the best of our knowledge, the following lemma is the first statement and proof of an analogous sharpening for continuous state space chains:

Proposition 4.1. *Consider a reversible, irreducible, ζ-lazy and smooth Markov chain* (4.4) *with stationary distribution $\Pi^\star$. Then for any error tolerance δ, and a β-warm distribution μ_0, given a set Ω such that $\Pi^\star(\Omega) \geq 1 - \frac{\delta^2}{3\beta^2}$, the δ-$\mathcal{L}_2$ mixing time of the chain is bounded as*

$$\tau_2(\delta; \mu_0) \leq \int_{4/\beta}^{8/\delta^2} \frac{8\, dv}{\zeta \cdot v \widetilde{\Phi}_\Omega^2(v)}, \tag{4.5}$$

where $\widetilde{\Phi}_\Omega$ denotes the truncated Ω-restricted conductance profile (4.3).

See Appendix B.1 for the proof, which is based on an appropriate generalization of the ideas used by [99] for discrete state chains.

The standard conductance based analysis makes use of the worst-case conductance bound for the chain. In contrast, Proposition 4.1 relates the mixing time to the conductance profile, which can be seen as size-wise conductance. We use the Ω-restricted conductance profile to state our bounds, because often a Markov chain has poor conductance only in regions that have very small probability under the target distribution. Such a behavior is not disastrous as it does not really affect the mixing of the chain up to a suitable tolerance. Given the bound (4.5), we can derive mixing time bound for a Markov chain readily if we have a bound on the Ω-restricted conductance profile Φ_Ω for a suitable Ω.

Corollary 4.1. *For a ζ-lazy Markov chain with a β-warm start μ_0, if the Ω-restricted conductance profile Φ_Ω satisfies,*

$$\Phi_\Omega(v) \geq \sqrt{B \log\left(\frac{1}{v}\right)} \quad \textit{for } v \in \left[\frac{4}{\beta}, \frac{1}{2}\right], \tag{4.6}$$

then the mixing time of the Markov chain satisfies

$$\tau_2(\delta; \mu_0) \leq \frac{64}{\zeta B} \log\left(\frac{\log \beta}{2\delta}\right). \tag{4.7}$$

The proof follows immediately by substituting the bound (4.6) on the RHS of equation (4.5), and integrating.

Prior work on conductance profile: We now situate Proposition 4.1 in the context prior work based on conductance profile. For discrete state chains, a result similar to Proposition 4.1 was already proposed by Lovász and Kannan (Theorem 2.3 in the paper [165]). Later on, Morris and Perres [183] and Goel et al. [99] used the notion of evolving sets and spectral profile respectively to sharpen the mixing time bounds based on average conductance for discrete-state space chains. In the context of continuous state space chains, Lovász and Kannan claimed in their original paper [165] that a similar result should hold for general state space chain as well, although we were unable to find any proof of such a general result in that or any subsequent work. Nonetheless, in a later work an average conductance based bound was used by Kannan et al. to derive faster mixing time guarantees for uniform sampling on bounded convex sets for ball walk (see Section 4.3 in the paper [136]). Their proof technique is not easily extendable to more general distributions including the general log-concave distributions in $\mathbb{R}^d$. Instead, our proof of Proposition 4.1 for general state space chains proceeds by an appropriate generalization of the ideas based on the spectral profile by Goel et al. [99] (for discrete state chains).

4.2 Lower Bound on Conductance Profile

To invoke guarantee Corollary 4.1, one needs to derive a lower bound on the conductance profile Φ_Ω of the Markov chain with a suitable choice of the set Ω. We now state a lower bound for the restricted-conductance profile of a general state space Markov chain whose target distribution satisfies isoperimetry conditions. We note that a closely related logarithmic-Cheeger inequality was used for sampling from uniform distribution of a convex body [136] and for sampling from log-concave distributions [158] without explicit constants. Since we would like to derive a non-asymptotic mixing rate, we re-derive an explicit form of their result.

Isoperimetry conditions: A distribution Π with support $\mathcal{X} \subset \mathbb{R}^d$ is said to satisfy the *isoperimetric inequality* ($\mathfrak{e} = 0$) or the *log-isoperimetric inequality* ($\mathfrak{e} = \frac{1}{2}$) with constant $\psi_\mathfrak{e}$ if given any partition $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ of $\mathcal{X}$, we have

$$\Pi(\mathcal{S}_3) \geq \frac{1}{2\psi_\mathfrak{e}} \cdot d(\mathcal{S}_1, \mathcal{S}_2) \cdot \min\left\{\Pi(\mathcal{S}_1), \Pi(\mathcal{S}_2)\right\} \cdot \log^\mathfrak{e}\left(1 + \frac{1}{\min\left\{\Pi(\mathcal{S}_1), \Pi(\mathcal{S}_2\right\}}\right), \tag{4.8}$$

where the distance between two sets $\mathcal{S}_1, \mathcal{S}_2$ is defined as $d(\mathcal{S}_1, \mathcal{S}_2) = \inf_{x \in \mathcal{S}_1, y \in \mathcal{S}_2} \{\|x-y\|_2\}$. For a distribution Π with density π and a given set Ω, its restriction to Ω is the distribution Π_Ω with the density $\pi_\Omega(x) = \frac{\pi(x)\mathbf{1}_\Omega(x)}{\Pi(\Omega)}$.

Let scalars $s \in (0, 1/2]$, $\rho \in (0,1)$ and $\Delta > 0$ be given and let $\mathcal{T}_x$ denote the one-step transition distribution of the Markov chain at point x. Suppose that that chain satisfies

$$d_{\mathrm{TV}}\big(\mathcal{T}_x, \mathcal{T}_y\big) \leq 1 - \rho \qquad \text{whenever } x, y \in \Omega \text{ and } \|x-y\|_2 \leq \Delta. \tag{4.9}$$

Proposition 4.2. *Consider a Markov chain with stationary distribution $\Pi^\star$, and let Ω be a convex measurable set such that $\Pi^\star_\Omega$ satisfies the isoperimetry (or log-isoperimetry) condition (4.8) with $\mathfrak{e} = 0$ (or $\mathfrak{e} = \frac{1}{2}$ respectively). When the Markov chain satisfies the condition (4.9), then we have*

$$\Phi_\Omega(v) \geq \frac{\rho}{4} \cdot \min\left\{1, \frac{\Delta}{16\psi_{\mathfrak{e}}} \cdot \log^{\mathfrak{e}}\left(1 + \frac{1}{v}\right)\right\}, \quad \textit{for any } v \in \left[0, \frac{\Pi^\star(\Omega)}{2}\right]. \tag{4.10}$$

See Appendix B.2 for the proof. We note that compared to the typical bounds on conductance, we gain an extra logarithmic term for target satisfying the logarithmic isoperimetric inequality ($\mathfrak{e} = \frac{1}{2}$). For any target distribution satisfying a logarithmic isoperimetric inequality (including the case of a strongly log-concave distribution), Proposition 4.2 is a strict improvement of the conductance bounds derived in previous works [163, 78].

4.3 Improved Mixing Time Guarantees

Suppose that we can find a convex set Ω such that $\Pi^\star(\Omega) \approx 1$ and the conditions of Proposition 4.2 are met, then with a β-warm start μ_0, a direct application of Corollary 4.1 along with Proposition 4.2 implies the following bound:

$$\tau_2(\delta; \mu_0) \leq \mathcal{O}\left(\frac{1}{\rho^2\Delta^2} \log \frac{\log \beta}{\delta}\right). \tag{4.11}$$

Mixing time bounds from previous work for continuous state Markov chains (including our earlier guarantees in Theorems 3.1 and 3.2) scale like $\frac{\log(\beta/\delta)}{\rho^2\Delta^2}$; for instance, see Proposition 6.1. In contrast, the bound (4.11) provides an additional logarithmic factor improvement in the factor β when the target satisfies log-isoperimetry. Lemma C.5 shows that an m-strongly log-concave target also satisfies log-isoperimetric inequality (D.46) with $\psi_{\frac{1}{2}} = 1/\sqrt{m}$. Such an improvement allows us to derive a sharper dependency on dimension d for the mixing time from non-warm starting distributions. As examples, we illustrate these improvements for MALA and MRW.

Improved Guarantees for MALA and MRW: For an $(\mathcal{L}, m)$-strongly log-concave target $\Pi^\star$ with $x^\star$ as the mode, Lemma 3.1 shows that $\mu_\star = \mathcal{N}(x^*, \frac{1}{\mathcal{L}}\mathbb{I}_d)$ is $\kappa^{d/2}$-warm with respect to $\Pi^\star$. The discussion above immediately implies that when $\beta = \mathcal{O}(e^d)$, the bound equation (4.11) implies an improvement of $\mathcal{O}(\frac{d}{\log d})$ in mixing time bounds from Theorems 3.1 and 3.2. Thus, we readily obtain the following corollary which improves upon the mixing time bounds from Corollary 3.1:

Corollary 4.2. *Given an $(\mathcal{L}, m)$-strongly log-concave target $\Pi^\star$ (Assumption (3A)), an error threshold $\delta \in (0, 1]$, and the initial distribution $\mu_\star = \mathcal{N}(x^*, \frac{1}{\mathcal{L}}\mathbb{I}_d)$, the $\frac{1}{2}$-lazy versions of MRW and MALA (Algorithms 1 and 2) with step sizes*

$$\eta_{\mathrm{MRW}} = c_1 \cdot \frac{1}{\mathcal{L} d\kappa}, \quad \textit{and} \quad \eta_{\mathrm{MALA}} = c_2 \cdot \frac{1}{\mathcal{L} d \cdot \max\left\{1, \sqrt{\kappa/d}\right\}}, \tag{4.12}$$

respectively, satisfy the mixing time bounds

$$\tau_2^{\mathrm{MRW}}(\delta; \mu_0) = \mathcal{O}\left(d\kappa^2 \log \frac{d}{\delta}\right), \quad \textit{and} \tag{4.13a}$$

$$\tau_2^{\mathrm{MALA}}(\delta; \mu_0) = \mathcal{O}\left(d\kappa \log \frac{d}{\delta} \cdot \max\left\{1, \sqrt{\frac{\kappa}{d}}\right\}\right). \tag{4.13b}$$

The proof is omitted as it directly follows from the conductance profile based mixing time bound in Proposition 4.1, Proposition 4.2 and the overlap bounds for MALA (Lemma 3.5) and MRW (proof of Theorem 3.2) from Section 3.5. Corollary 4.2 states that the mixing time bounds for MALA and MRW with the feasible distribution $\mu_\star$ as the initial distribution scale as $\widetilde{\mathcal{O}}(d\kappa \log(1/\delta))$ and $\widetilde{\mathcal{O}}(d\kappa^2 \log(1/\delta))$. In light of the inequality (2.5d), we obtain the same bounds for the number of steps taken by these algorithms to mix within δ total-variation distance of the target distribution $\Pi^\star$. Consequently, our results improve upon the previously guarantees (Corollary 3.1) mixing time bounds for MALA and MRW [78] for strongly log-concave distributions.

4.4 Conclusion and future directions

In this chapter, we provided refined results for establishing mixing time bounds on continuous state space chains using conductance profile. In particular, we studied the mixing time bounds of target that satisfy the isoperimetry or log-isoperimetry condition (D.46). We also applied these results to sharpen the mixing time bounds for MALA and MRW with a non-warm start for a strongly log-concave target.

With this machinery, one can potentially establish fast mixing time bounds for a wide range of target distributions; identifying some interesting such classes, and establishing the isoperimetry condition can be an interesting future direction. One possibility is the weakly log-concave distributions for which the KLS conjecture implies that the isoperimetry constant is bounded by the operator norm of the inverse covariance matrix. Given that the KLS conjecture has been almost proven [45] (upto poly-logartihmic factors in dimensions), our theory already provides an easily accessible mixing time bound for wide range of Markov chains for which one establish the condition (4.9).

Another interesting direction is to determine if one can further refine the conductance based proof techniques to obtain even sharper mixing time bounds for existing sampling algorithms, like MALA, and HMC (analyzed in the next chapter).

Chapter 5

Non-Asymptotic Analysis of Hamiltonian Monte Carlo

As noted earlier, there are a variety of MCMC methods for sampling from target distributions with smooth densities [211, 213, 215, 30]. One method often stands out in practice and is often considered state-of-the-art: Hamiltonian Monte Carlo (HMC). It is the default sampler for sampling from complex distributions in many popular software packages, including Stan [37], Mamba [228], and Tensorflow [1].

This chapter provides a thorough discussion and some new results on the non-asymptotic analysis of Hamiltonian Monte Carlo. We start by providing some historical context on HMC, followed by an introduction to the algorithm in Section 5.1, and past related work in Section 5.2. With the basic context in place, we then summarize our contributions in Section 5.2.1, and provide the organization of the remainder of the chapter in Section 5.2.2.

Origins of HMC: The idea of using Hamiltonian dynamics in simulation can be traced back to Alder and Wainwright [3] in the physics literature. The method is inspired by Hamiltonian dynamics, which describe the evolution of a state vector $\mathbf{q}(t) \in \mathbb{R}^d$ and its momentum $\mathbf{p}(t) \in \mathbb{R}^d$ over time t based on a Hamiltonian function $\mathcal{H} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ via Hamilton's equations:

$$\frac{d\mathbf{q}}{dt}(t) = \frac{\partial \mathcal{H}}{\partial \mathbf{p}}(\mathbf{p}(t), \mathbf{q}(t)), \quad \text{and} \quad \frac{d\mathbf{p}}{dt}(t) = -\frac{\partial \mathcal{H}}{\partial \mathbf{q}}(\mathbf{p}(t), \mathbf{q}(t)). \tag{5.1}$$

A straightforward calculation using the chain rule shows that the Hamiltonian remains invariant under these dynamics—that is, $\mathcal{H}(\mathbf{p}(t), \mathbf{q}(t)) = C$ for all $t \in \mathbb{R}$. A typical choice of the Hamiltonian $\mathcal{H} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is given by

$$\mathcal{H}(\mathbf{p}, \mathbf{q}) = f(\mathbf{q}) + \frac{1}{2} \left\| \mathbf{p} \right\|_2^2. \tag{5.2}$$

Duane et al. [73] introduced MCMC with Hamiltonian dynamics, and referred to it as the *hybrid Monte Carlo*. The algorithm was further refined by Neal [189], and later re-christened in the statistics community as Hamiltonian Monte Carlo. We refer the reader to Neal [190] for an illuminating overview of the history of HMC and a discussion of contemporary work.

5.1 Introduction to the HMC Algorithm

The ideal HMC algorithm for sampling is based on the continuous Hamiltonian dynamics (5.1), and as such, it is not implementable in practice, but instead a useful algorithm for understanding. For a given time $T > 0$ and vectors $u, v \in \mathbb{R}^d$, let $\mathbf{q}_T(u, v)$ denote the $\mathbf{q}$-solution to Hamilton's equations at time T and with initial conditions $(\mathbf{p}(0), \mathbf{q}(0)) = (u, v)$. At iteration k, given the current iterate X_k, the ideal HMC algorithm generates the next iterate X_{k+1} via the update rule $X_{k+1} = \mathbf{q}_T(\mathbf{p}_k, X_k)$ where $p_k \sim N(0, \mathbb{I}_d)$ is a standard normal random vector, independent of X_k and all past iterates. It can be shown that with an appropriately chosen T, the ideal HMC algorithm converges to the stationary distribution $\pi^\star$ without a Metropolis-Hastings adjustment (see [190, 176] for the existence of such solution and its convergence).

However, in practice, it is impossible to compute an exact solution to Hamilton's equations. Rather, one must approximate the solution $\mathbf{q}_T(\mathbf{p}_k, X_k)$ via some discrete process. There are many ways to discretize Hamilton's equations other than the simple Euler discretization; see Neal [190] for a discussion. In particular, using the leapfrog or Störmer-Verlet method for integrating Hamilton's equations leads to the Hamiltonian Monte Carlo (HMC) algorithm. It simulates the Hamiltonian dynamics for $\mathcal{K}$ steps via the leapfrog integrator. At each iteration, given previous state $\mathbf{q}_0$ and fresh $\mathbf{p}_0 \sim \mathcal{N}(0, \mathbb{I}_d)$, it runs the following updates for $\mathcal{K}$ times, for $0 \le k \le \mathcal{K} - 1$,

$$\mathbf{p}_{k+\frac{1}{2}} = \mathbf{p}_k - \frac{\eta}{2}\nabla f(\mathbf{q}_k) \tag{5.3a}$$

$$\mathbf{q}_{k+1} = \mathbf{q}_k + \eta \mathbf{p}_{k+\frac{1}{2}} \tag{5.3b}$$

$$\mathbf{p}_{k+1} = \mathbf{p}_{k+\frac{1}{2}} - \frac{\eta}{2}\nabla f(\mathbf{q}_{k+1}). \tag{5.3c}$$

Since discretizing the dynamics generates discretization error at each iteration, it is followed by a Metropolis-Hastings adjustment where the proposal $(\mathbf{p}_\mathcal{K}, \mathbf{q}_\mathcal{K})$ is accepted with probability

$$\min\left\{1, \frac{\exp\left(-\mathcal{H}(\mathbf{p}_\mathcal{K}, \mathbf{q}_\mathcal{K})\right)}{\exp\left(-\mathcal{H}(\mathbf{p}_0, \mathbf{q}_0)\right)}\right\}. \tag{5.4}$$

See Algorithm 3 for a detailed description of the HMC algorithm with leapfrog integrator. In practice, one also uses the HMC algorithm with a modified Hamiltonian, in which the quadratic term $\|\mathbf{p}\|_2^2$ is replaced by a more general quadratic form $\mathbf{p}^T\Sigma\mathbf{p}$, for a symmetric positive definite matrix Σ chosen by the user; see **??** C.3.1.1 for further discussion. Here, we restrict our analysis to the case $\Sigma = I$.

Algorithm 3: Metropolized HMC with leapfrog integrator

```
Input: Step size $\eta$, number of internal leapfrog updates $\mathcal{K}$,
and a sample $x_0$ from a starting distribution $\mu_0$
Output: Sequence $x_1, x_2, \ldots$
for $i = 0, 1, \ldots$ do
    Proposal step:
    $\mathbf{q}_0 \leftarrow x_i$
    Draw $\mathbf{p}_0 \sim \mathcal{N}(0, \mathbb{I}_d)$
    for $k = 1, \ldots, \mathcal{K}$ do
        $(\mathbf{p}_k, \mathbf{q}_k) \leftarrow \mathrm{Leapfrog}(\mathbf{p}_{k-1}, \mathbf{q}_{k-1}, \eta)$
    end
    % $\mathbf{q}_K$ is now the new proposed state
    Accept-reject step:
      compute $\alpha_{i+1} \leftarrow \min\left\{1, \frac{\exp\left(-\mathcal{H}(\mathbf{p}_K, \mathbf{q}_K)\right)}{\exp\left(-\mathcal{H}(\mathbf{p}_0, \mathbf{q}_0)\right)}\right\}$
      With probability $\alpha_{i+1}$ accept the proposal: $x_{i+1} \leftarrow \mathbf{q}_K$
      With probability $1 - \alpha_{i+1}$ reject the proposal: $x_{i+1} \leftarrow x_i$
end
Program Leapfrog($\mathbf{p}$, $\mathbf{q}$, $\eta$):
    $\widetilde{\mathbf{p}} \leftarrow \mathbf{p} - \frac{\eta}{2}\nabla f(\mathbf{q})$
    $\widetilde{\mathbf{q}} \leftarrow \mathbf{q} + \eta\widetilde{\mathbf{p}}$
    $\widetilde{\mathbf{p}} \leftarrow \widetilde{\mathbf{p}} - \frac{\eta}{2}\nabla f(\widetilde{\mathbf{q}})$
return $(\widetilde{\mathbf{p}}, \widetilde{\mathbf{q}})$
```

We refer the reader to the papers [190, 115, 76] for further examples and discussion of the HMC method. There are a number of variants of HMC, but the most popular choice involves a combination of the leapfrog integrator with Metropolis-Hastings correction. Throughout this chapter, we reserve the terminology HMC to refer to this particular Metropolized algorithm.

Connection with MALA: The HMC with leapfrog integrator can be seen as a multi-step version of Langevin algorithm. In fact, running the HMC algorithm with $\mathcal{K} = 1$ is equivalent to the MALA (Algorithm 2) after a re-parametrization of the step-size η.

5.2 Past work on HMC

While HMC enjoys fast convergence in practice, a theoretical understanding of this behavior remains incomplete. Some intuitive explanations are based on its ability to maintain a constant asymptotic accept-reject rate with large step-size [60]. Others [190] suggest, based on intuition from the continuous-time limit of the Hamiltonian dynamics, that HMC can suppress random walk behavior using momentum. However, these intuitive arguments do not provide rigorous or quantitative justification for the fast convergence of the discrete-time HMC used in practice.

More recently, general asymptotic conditions under which HMC will or will not be geometrically ergodic have been established in some recent papers [76, 162]. Other work has yielded some insight into the mixing properties of different variants of HMC, but it has focused mainly on *unadjusted* versions of the algorithm. Mangoubi et al. [175, 176] study versions of unadjusted HMC based on Euler discretization or leapfrog integrator (but omitting the Metropolis-Hastings step), and provide explicit bounds on the mixing time as a function of dimension d, condition number κ and error tolerance $\delta > 0$. Lee and Vempala [153] studied an extended version of HMC that involves applying an ordinary differential equation (ODE) solver; they established bounds with sublinear dimension dependence, and even polylogarithmic for certain densities (e.g., those arising in Bayesian logistic regression). The mixing time for the same algorithm is further refined in the recent work by Chen and Vempala [50]. In a similar spirit, Lee and Vempala [156] studied the Riemannian variant of HMC (RHMC) with an ODE solver focusing on sampling uniformly from a polytope. While their result could be extended to log-concave sampling, the practical implementation for log-concave sampling of their ODE solver is unclear, and moreover requires a regularity condition on all the derivatives of density. It should be noted that such unadjusted HMC methods behave differently from the Metropolized version most commonly used in practice. In the absence of the Metropolis-Hastings correction, the resulting Markov chain no longer converges to the correct target distribution, but instead exhibits a persistent bias, even in the limit of infinite iterations. Consequently, the analysis of such sampling methods requires controlling this bias; doing so leads to mixing times that scale polynomially in $1/\delta$, in sharp contrast with the $\log(1/\delta)$ that is typical for Metropolis-Hastings corrected methods.

Most closely related to the guarantees in this chapter is the recent work by Bou-Rabee et al. [24], which studies the same Metropolized HMC algorithm that we analyze in this chapter. They use coupling methods to analyze HMC for a class of distributions that are strongly log-concave outside of a compact set. In the strongly log-concave case, they prove a mixing time bound that scales at least as $d^{3/2}$ in the dimension d. It should be noted that with a "warm" initialization, this dimension dependence grows more quickly than known bounds for the MALA algorithm [78, 86], and so does not explain the superiority of HMC in practice.

Tradeoff between hyperparameters: In practice, it is known that Metropolized HMC is fairly sensitive to the choice of its parameters, namely the step-size η used in the discretization scheme, and the number of leapfrog steps $\mathcal{K}$. At one extreme, taking a single leapfrog step $\mathcal{K} = 1$, the algorithm reduces to the Metropolis adjusted Langevin algorithm (MALA). More generally, if too few leapfrog steps are taken, that of HMC is likely to exhibit a random walk behavior similar to MALA. At the other extreme, if $\mathcal{K}$ is too large, the leapfrog steps tend to wander back to a neighborhood of the initial state, which leads to wasted computation as well as slower mixing [19]. In terms of the step size η, choosing an overly large step size makes the discretization diverge from the underlying continuous dynamics, and causes the Metropolis acceptance probability to drop, hence slowing down the algorithm.

On the other hand, an overly small choice of η does not allow the algorithm to explore the state space rapidly enough. While it is difficult to characterize the necessary and sufficient conditions on $\mathcal{K}$ and η to ensure fast convergence, many work suggest the choice of $\mathcal{K}$ and η based on the necessary conditions such as maintaining a constant acceptance rate [44]. For instance, Beskos et al. [18] showed that in the simplified scenario of target density with independent, identically distributed components, the number of leapfrog steps should scale as $d^{1/4}$ to achieve a constant acceptance rate. Besides, instead of setting the two parameters explicitly, various automatic strategies for tuning these two parameters have been proposed [249, 115, 250]. Despite being introduced via heuristic arguments and with additional computational cost, these methods, such as the No-U-Turn (NUTS) sampler [115], have shown promising empirical evidence of its effectiveness on a wide range of simple target distributions.

5.2.1 Overview of our contributions

We provide a non-asymptotic upper bound on the mixing time of the Metropolized HMC algorithm for smooth densities (see Theorem 5.1). This theorem applies to the form of Metropolized HMC (based on the leapfrog integrator) that is most widely used in practice. To the best of our knowledge, Theorem 5.1 is the first rigorous confirmation of the faster non-asymptotic convergence of the Metropolized HMC as compared to MALA and other simpler Metropolized algorithms.[1] Other related works on HMC consider either its unadjusted version (without accept-reject step) with different integrators [175, 176] or the HMC based on an ODE solver [153, 156]. While the dimension dependency for these algorithms is usually better than MALA, they have polynomial dependence on the inverse error tolerance $1/\delta$ while MALA's mixing time scales as $\log(1/\delta)$. Moreover, our direct analysis of the Metropolized HMC with a leapfrog integrator provides explicit choices of the hyper-parameters for the sampler, namely, the step-size and the number of leapfrog updates in each step. Our theoretical choices of the hyper-parameters could potentially provide guidelines for parameter tuning in practical HMC implementations.

The proof makes use of the refined conductance profiles based techniques from Chapter 4. A number of technical challenges arise en route in particular in controlling the conductance profile of HMC: The use of multiple gradient steps in each iteration of HMC helps it mix faster but also complicates the analysis. In fact, a key step is to control the overlap between the transition distributions of HMC chain at two nearby points; doing so requires a delicate argument (see Lemma 5.1 and Appendix C.1 for further details).

Table 5.1 provides an informal summary of our mixing time bounds of HMC and how they compare with known bounds for MALA when applied to log-concave target distributions. From the table, we see that Metropolized HMC takes fewer gradient evaluations than MALA to mix to the same accuracy for log-concave distributions. Note that our current analysis

[1] As noted earlier, previous results by Bou-Rabee et al. [24] on Metropolized HMC do not establish that it mixes more rapidly than MALA.

establishes logarithmic dependence on the target error δ for strongly-log-concave as well as for a sub-class of weakly log-concave distributions. For a comparison with previous results on unadjusted HMC or ODE based HMC refer to the discussion after Corollary 5.1, and Table C.4 in Appendix C.3.2.

	Strongly log-concave	**Weakly log-concave**	
Sampling algorithm	**Assumption (5B)** ($\kappa \ll d$)	**Assumption (5C)**	**Assumption (5D)**
MALA	$d\kappa \log \frac{1}{\delta}$ [Corollary 4.2]	$\frac{d^2}{\delta^{\frac{3}{2}}} \log \frac{1}{\delta}$ [Corollary 3.2]	$d^{\frac{3}{2}} \log \frac{1}{\delta}$ [?]
Metropolized HMC	$d^{\frac{11}{12}} \kappa \log \frac{1}{\delta}$ [Corollary 5.1]	$\frac{d^{\frac{11}{6}}}{\delta} \log \frac{1}{\delta}$ [Corollary 5.1]	$d^{\frac{4}{3}} \log \frac{1}{\delta}$ [Corollary 5.2]

Table 5.1: Comparisons of the number of gradient evaluations needed by MALA and Metropolized HMC with leapfrog integrator from a *warm start* to obtain an δ-accurate sample in TV distance from a log-concave target distribution on $\mathbb{R}^d$. The second column corresponds to strongly log-concave densities with condition number κ, and the third and fourth column correspond to weakly log-concave densities satisfying certain regularity conditions.

5.2.2 Organization

The remainder of the chapter is organized as follows. Section 5.3 contains our main results on mixing time of HMC in Section 5.3.2 (with overview in Tables 5.1 and 5.2). In Section 5.4, we describe some numerical experiments that we performed to explore the sharpness of our theoretical predictions in some simple scenarios. In Section 5.5, we prove our main result, Theorem 5.1, and defer the proofs of technical lemmas and other results to the appendices. We conclude in Section 5.6 with a discussion of our results and future directions.

5.3 Main results

We now turn to the statement of our main results. We remind the readers that HMC refers to Metropolized HMC with leapfrog integrator, unless otherwise specified. We collect the set of assumptions for the target distribution in Section 5.3.1, followed by general mixing time result as Theorem 5.1 in Section 5.3.2. Then we apply Theorem 5.1 to derive guarantees for strongly log-concave target as Corollary 5.1 in Section 5.3.3, and weakly log-concave and non log-concave distributions in Section 5.3.4.

5.3.1 Assumptions on the target distribution

In this section, we introduce some regularity notions and state the assumptions on the target distribution that our results in the next section rely on.

A function f is said to be $\mathcal{L}_H$-Hessian Lipschitz if

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_{\mathrm{op}} \leq \mathcal{L}_H \left\|x-y\right\|_2, \quad \text{for all} \quad x, y \in \mathbb{R}^d, \tag{5.5}$$

where $\left\|B\right\|_{\mathrm{op}}$ denotes the operator-norm of the matrix B.

We introduce two sets of assumptions for the target distribution:

(5A) We say that the target distribution $\Pi^\star$ is $(\mathcal{L}, \mathcal{L}_H, s, \psi_{\mathfrak{e}}, \mathcal{M})$-*regular* if the negative log density f is $\mathcal{L}$-smooth (3.5a) and has $\mathcal{L}_H$-Lipschitz Hessian (5.5), and there exists a convex measurable set Ω such that the distribution $\Pi^\star_\Omega$ is $\psi_{\mathfrak{e}}$-isoperimetric (4.8), and the following conditions hold:

$$\Pi^\star(\Omega) \geq 1 - s \quad \text{and} \quad \left\|\nabla f(x)\right\|_2 \leq \mathcal{M}, \quad \text{for all } x \in \Omega. \tag{5.6}$$

(5B) We say that the target distribution $\Pi^\star$ is $(\mathcal{L}, \mathcal{L}_H, m)$-*strongly log-concave* if the negative log density is $\mathcal{L}$-smooth (3.5a), m-strongly convex (3.5b), and $\mathcal{L}_H$-Hessian-Lipschitz (5.5). Moreover, we use $x^\star$ to denote the unique mode of $\Pi^\star$ whenever f is strongly convex.

Assumption (5B) has appeared in several past papers on Langevin algorithms [61, 78, 51] and the Lipschitz-Hessian condition (5.5) has been used in analyzing Langevin algorithms with inaccurate gradients [62] as well as the unadjusted HMC algorithm [176]. It is worth noting Assumption (5A) is strictly weaker than Assumption (5B), since it allows for distributions that are not log-concave. In Appendix C.2 (see Lemma C.5), we show how Assumption (5B) implies a version of Assumption (5A).

5.3.2 Mixing time bounds for HMC

We start with the mixing time bound for HMC applied to any target distribution $\Pi^\star$ satisfying Assumption (5A). Let HMC-$(\mathcal{K}, \eta)$ denote the $\frac{1}{2}$-lazy Metropolized HMC (Algorithm 3) with η step size and $\mathcal{K}$ leapfrog steps in each iteration. Let $\tau_2^{\mathrm{HMC}}(\delta; \mu_0)$ denote the $\mathcal{L}_2$-mixing time (2.5b) for this chain with the starting distribution μ_0.

Theorem 5.1 (General bound on HMC mixing time). *Consider an $(\mathcal{L}, \mathcal{L}_H, s, \psi_{\mathfrak{e}}, \mathcal{M})$-regular target $\Pi^\star$ (Assumption (5A)) and a β-warm start μ_0. Then for any fixed target error $\delta \in (0, 1)$ such that $\delta^2 \geq 2\beta s$, there exist choices of the parameters $(\mathcal{K}, \eta)$ such that HMC-$(\mathcal{K}, \eta)$chain with μ_0 start satisfies*

$$\tau_2^{HMC}(\delta; \mu_0) \leq \begin{cases} c \cdot \max\left\{\log \beta, \dfrac{\psi_{\mathfrak{e}}^2}{\mathcal{K}^2\eta^2} \log\left(\dfrac{\log \beta}{\delta}\right)\right\} & \textit{if } \mathfrak{e} = \frac{1}{2} \ \textit{(log-isoperimetric target),} \\ c \cdot \dfrac{\psi_{\mathfrak{e}}^2}{\mathcal{K}^2\eta^2} \log\left(\dfrac{\beta}{\delta}\right) & \textit{if } \mathfrak{e} = 0 \ \textit{(isoperimetric target).} \end{cases}$$

See Section 5.5.2 for the proof, where we also provide explicit conditions on η and $\mathcal{K}$ in terms of the other parameters (cf. equation (5.15b)).

Theorem 5.1 covers mixing time bounds for distributions that satisfy isoperimetric or log-isoperimetric inequality provided that: (a) both the gradient and Hessian of the negative log-density are Lipschitz; and (b) there is a convex set that contains a large mass $(1-s)$ of the distribution. The mixing time only depends on two quantities: the log-isoperimetric (or isoperimetric) constant of the target distribution and the effective step-size $\mathcal{K}^2\eta^2$. As shown in the sequel, these conditions hold for log-concave distributions as well as certain perturbations of them. If the distribution satisfies a log-isoperimetric inequality, then the mixing time dependency on the initialization warmness parameter β is relatively weak $\mathcal{O}(\log\log\beta)$. On the other hand, when only an isoperimetric inequality (but not log-isoperimetric) is available, the dependency is relatively larger $\mathcal{O}(\log\beta)$. In our current analysis, we can establish the δ-mixing time bounds up-to an error δ such that $\delta^2 \geq 2\beta s$. If mixing time bounds up to an arbitrary accuracy are desired, then the distribution needs to satisfy (5.6) for arbitrary small s. For example, as we later show in Lemma C.5, arbitary small s can be imposed for strongly log-concave densities (i.e., satisfying Assumption (5B)).

Let us now derive several corollaries of Theorem 5.1. We begin with non-asymptotic mixing time bounds for HMC-$(\mathcal{K},\eta)$ chain for strongly-log-concave target distributions. Then we also discuss the corollaries for weakly log-concave target and non-log-concave target distributions. These results also provide a basis for comparison of our results with prior work.

5.3.3 Mixing time for strongly log-concave target

We now state an explicit mixing time bound of HMC for a strongly log-concave distribution. We consider an $(\mathcal{L},\mathcal{L}_H,m)$-strongly log-concave distribution (assumption (5B)). We use $\kappa = \mathcal{L}/m$ to denote the condition number of the distribution. Our result makes use of the function $\mathfrak{a}$ defined earlier in (3.6a) and reproduced here, and two choices of step-sizes

$$\mathfrak{a}(s) = 2 + 2\max\left\{\left(\frac{\log(1/s)}{d}\right)^{\frac{1}{4}}, \left(\frac{\log(1/s)}{d}\right)^{\frac{1}{2}}\right\}, \quad \text{for} \quad s>0, \quad \text{and} \tag{5.7a}$$

$$\eta_{\text{warm}} = \sqrt{\frac{1}{c\mathcal{L}\cdot\mathfrak{a}(\frac{\delta^2}{2\beta})d^{\frac{7}{6}}}}, \quad \text{and} \quad \eta_{\text{feas}} = \sqrt{\frac{1}{c\mathcal{L}\cdot\mathfrak{a}(\frac{\delta^2}{2\kappa^d})}\min\left\{\frac{1}{d\kappa^{\frac{1}{2}}}, \frac{1}{d^{\frac{2}{3}}\kappa^{\frac{5}{6}}}, \frac{1}{d^{\frac{1}{2}}\kappa^{\frac{3}{2}}}\right\}}. \tag{5.7b}$$

With these definitions, we have the following:

Corollary 5.1 (HMC mixing for strongly-log-concave target). *Consider an $(\mathcal{L},\mathcal{L}_H,m)$-strongly log-concave target distribution $\Pi^\star$ (Assumption (5B)) such that $\mathcal{L}_H^{2/3} = \mathcal{O}(\mathcal{L})$, and any error tolerance $\delta \in (0,1)$.*

(c) *Suppose that* $\kappa = \mathcal{O}(d^{\frac{2}{3}})$ *and* $\beta = \mathcal{O}(\exp(d^{\frac{2}{3}}))$. *Then with any* β*-warm initial distribution* μ_0, *hyper-parameters* $\mathcal{K} = d^{\frac{1}{4}}$ *and* $\eta = \eta_{warm}$, *the HMC-*$(\mathcal{K}, \eta)$ *chain satisfies*

$$\tau_2^{HMC}(\delta; \mu_0) \leq c\, d^{\frac{2}{3}}\ \kappa\, \mathfrak{a}\left(\frac{\delta^2}{2\beta}\right)\ \log\left(\frac{\log\beta}{\delta}\right). \tag{5.8a}$$

(d) *With the initial distribution* $\mu_\star = \mathcal{N}(x^\star, \frac{1}{\mathcal{L}}\mathbb{I}_d)$, *hyper-parameters* $\mathcal{K} = \kappa^{\frac{3}{4}}$ *and* $\eta = \eta_{feas}$, *the HMC-*$(\mathcal{K}, \eta)$ *chain satisfies*

$$\tau_2^{HMC}(\delta; \mu_\star) \leq c\ \mathfrak{a}\left(\frac{\delta^2}{2\kappa^d}\right)\ \max\left\{d\log\kappa, \max\left[d, d^{\frac{2}{3}}\kappa^{\frac{1}{3}}, d^{\frac{1}{2}}\kappa\right]\log\left(\frac{d\log\kappa}{\delta}\right)\right\}. \tag{5.8b}$$

See Appendix C.2 for the proof, which proceeds by showing that an $(\mathcal{L}, \mathcal{L}_H, m)$-strongly log-concave distribution is in fact an $(\mathcal{L}, \mathcal{L}_H, s, \psi_{1/2}, \mathcal{M}_s)$-regular distribution for any $s \in (0, 1)$. Here $\psi_{1/2} = 1/\sqrt{m}$ is fixed and the bound on the gradient $\mathcal{M}_s = \mathfrak{a}(s)\sqrt{d/m}$ depends on the choice of s. In the same appendix, we also provide a more refined mixing time of the HMC chain for a more general choice of hyper-parameters (see Corollary C.1). In fact, as shown in the proof, the assumption $\mathcal{L}_H^{2/3} = \mathcal{O}(\mathcal{L})$ is not necessary in order to control mixing; rather, we adopted it above to simplify the statement of our bounds. Moreover, for a refined and detailed discussion on the optimal choice for step size η, we refer the reader to **??**.

Metropolized HMC vs Unadjusted HMC: There are many recent results on the 1-Wasserstein distance mixing of unadjusted versions of HMC (for instance, see the papers [176, 153]). A direct comparisons of these different results is tricky for two reasons: (a) The 1-Wasserstein distance and the total variation distance are not strictly comparable, and, (b) the unadjusted HMC results always have a polynomial dependence on the error parameter δ while our results for Metropolized HMC have a superior logarithmic dependence on δ. For a thorough discussion, we refer the readers to Table C.4 in Appendix C.3.2. A key take away from that discussion is that the unadjusted chains have better mixing time in terms of scaling with d, if we fix δ or view it as independent of d. On the other hand, when such chains are used to estimate certain higher-order moments, the polynomial dependence on δ might become the bottleneck and Metropolis-adjusted chains would become the method of choice. We now focus on a direct comparison of the guarantees for HMC with MALA.

MALA vs HMC—Warm start: Corollary 5.1 provides mixing time bounds for two cases. The first result (5.8a) implies that given a warm start for a well-conditioned strongly log-concave distribution, i.e., with constant β and $\kappa \ll d$, the δ-$\mathcal{L}_2$-mixing time[2] of HMC scales $\widetilde{\mathcal{O}}(d^{\frac{2}{3}}\log(1/\delta))$. It is interesting to compare this guarantee with known bounds for

[2] Note that $\mathfrak{a}(\delta^2) \leq 6$ for $\delta \geq \frac{2}{e^{d/2}}$ and thus we can treat $\mathfrak{a}$ as a small constant for a large range of δ. Otherwise, if δ needs to be extremely small, the results still hold with an extra $\log^{\frac{1}{2}}\left(\frac{1}{\delta}\right)$ dependency.

the MALA algorithm. However since each iteration of MALA uses only a single gradient evaluation, a fair comparison would require us to track the total number of gradient evaluations required by the HMC-$(\mathcal{K}, \eta)$ chain to mix. For HMC to achieve accuracy δ, the total number of gradient evaluations is given by $\mathcal{K} \cdot \tau_2^{\mathrm{HMC}}(\delta; \mu_0)$, which in the above setting, scales as $\widetilde{\mathcal{O}}(d^{\frac{11}{12}}\kappa \log(1/\delta))$. This rate was also summarized in Table 5.1. On the other hand, Theorem 3.1 shows that the corresponding number of gradient evaluations for MALA is $\widetilde{\mathcal{O}}(d\kappa \log(1/\delta))$. As a result, we conclude that the upper bound for HMC is $d^{\frac{1}{12}}$ better than the known upper bound for MALA with a warm start for a well-conditioned strongly log-concave target distribution. We summarize these rates in Table 5.2. Note that MRW is a zeroth order algorithm, which makes use of function evaluations but not gradient information.

Sampling algorithm	**Mixing time**	**#Gradient evaluations**
MRW [Theorem 3.2]	$d\kappa^2 \cdot \log \frac{1}{\delta}$	NA
MALA [Theorem 3.1]	$d\kappa \cdot \log \frac{1}{\delta}$	$d\kappa \cdot \log \frac{1}{\delta}$
HMC-$(\mathcal{K}, \eta)$[Corollary 5.1]	$d^{\frac{2}{3}}\kappa \cdot \log \frac{1}{\delta}$	$d^{\frac{11}{12}}\kappa \cdot \log \frac{1}{\delta}$

Table 5.2: Summary of the δ-TV mixing time $\tau_{\mathrm{TV}}(\delta; \mu_0)$ (2.5b)and the corresponding number of gradient evaluations for MRW, MALA and HMC from a *warm start* with an $(\mathcal{L}, \mathcal{L}_H, m)$-strongly-log-concave target. These statements hold under the assumption $\mathcal{L}_H^{2/3} = \mathcal{O}(\mathcal{L})$, $\kappa = \frac{\mathcal{L}}{m} \ll d$, and omit logarithmic terms in dimension.

MALA vs HMC—Feasible start: In the second result (5.8b), we cover the case when a warm start is not available. In particular, we analyze the HMC chain with the feasible initial distribution $\mu_\star = \mathcal{N}(x^\star, \frac{1}{\mathcal{L}}\mathbb{I}_d)$, which was also used in Corollary 4.2. Once again, it is of interest to determine whether HMC takes fewer gradient steps when compared to MALA to obtain an δ-accurate sample. We summarize the results in Table 5.3, with log factors hidden, and note that HMC with $\mathcal{K} = \kappa^{3/4}$ is faster than MALA for as long as κ is not too large. From the last column, we find that when $\kappa \ll d^{\frac{1}{2}}$, HMC is faster than MALA by a factor of $\kappa^{\frac{1}{4}}$ in terms of number of gradient evaluations.

Ill-conditioned target distributions: In order to keep the statement of Corollary 5.1 simple, we stated the mixing time bounds of HMC-$(\mathcal{K}, \eta)$-chain only for a particular choice of $(\mathcal{K}, \eta)$. In our analysis, this choice ensures that HMC is better than MALA only when condition number κ is small. For Ill-conditioned distributions, i.e., when κ is large, finer tuning of HMC-$(\mathcal{K}, \eta)$-chain is required. For further discussion, we refer the reader to the

Sampling algorithm	**Mixing time**	**# Gradient Evaluations**	
	$\tau_{\mathrm{TV}}(\delta;\mu_0)$	general κ	$\kappa \ll d^{\frac{1}{2}}$
MRW [Corollary 4.2]	$d\kappa^2$	NA	NA
MALA [Corollary 4.2]	$\max\left\{d\kappa, d^{\frac{1}{2}}\kappa^{\frac{3}{2}}\right\}$	$\max\left\{d\kappa, d^{\frac{1}{2}}\kappa^{\frac{3}{2}}\right\}$	$d\kappa$
HMC-$(\mathcal{K},\eta)$ [Corollary 5.1]	$\max\left\{d, d^{\frac{2}{3}}\kappa^{\frac{1}{3}}, d^{\frac{1}{2}}\kappa\right\}$	$\max\left\{d\kappa^{\frac{3}{4}}, d^{\frac{2}{3}}\kappa^{\frac{13}{12}}, d^{\frac{1}{2}}\kappa^{\frac{7}{4}}\right\}$	$d\kappa^{\frac{3}{4}}$

Table 5.3: Summary of the δ-TV mixing time and the corresponding number of gradient evaluations for MRW, MALA and HMC from the *feasible start* $\mu_\star = \mathcal{N}(x^\star, \frac{1}{\mathcal{L}}\mathbb{I}_d)$ for an $(\mathcal{L}, \mathcal{L}_H, m)$-strongly-log-concave target. Here $x^\star$ denotes the unique mode of the target distribution. These statements hold uner the assumption $\mathcal{L}_H = \mathcal{O}(\mathcal{L}^{\frac{3}{2}})$, and hide the logarithmic factors in δ, d and $\kappa = \mathcal{L}/m$.

Appendices C.2 (see Table C.1), where we we show that HMC is strictly better than MALA as long as $\kappa \leq d$ and as good as MALA when $\kappa \geq d$.

5.3.4 Beyond strongly log-concave target distributions

We now discuss the mixing time bounds for Metropolized HMC when the target is not strongly log-concave. In the next two sections, we discuss the case when the target is weakly log-concave distribution or a perturbation of log-concave distribution, respectively.

5.3.4.1 Weakly log-concave target

The mixing rate in the weakly log-concave case differs depends on further structural assumptions on the density. We now consider two different scenarios where either a bound on fourth moment is known or the covariance of the distribution is well-behaved:

(5C) The negative log density of the target distribution is $\mathcal{L}$-smooth (3.5a) and has $\mathcal{L}_H$-Lipschitz Hessian (5.5). Additionally for some point $x^\star$, its fourth moment satisfies the bound

$$\int_{\mathbb{R}^d} \left\|x - x^\star\right\|_2^4 \pi^\star(x)dx \leq \frac{d^2\omega^2}{\mathcal{L}^2}. \tag{5.9}$$

(This condition is identical to the condition (3.12).)

(5D) The negative log density of the target distribution is $\mathcal{L}$-smooth (3.5a) and has $\mathcal{L}_H$-Lipschitz Hessian (5.5). Additionally, its covariance matrix satisfies

$$\left\| \int_{x\in\mathbb{R}^d} (x - \mathbb{E}[x])(x - \mathbb{E}[x])^\top \pi^\star(x)dx \right\|_{\mathrm{op}} \leq 1, \tag{5.10}$$

and the norm of the gradient of the negative log density f is bounded by a constant in the ball $\mathbb{B}\left(\mathbb{E}\left[x\right], \log\left(\frac{1}{s}\right) d^{3/4}\right)$ for small enough $s \geq s_0$.

When the distribution satisfies assumption (5C) we consider HMC chain with slightly modified target and assume that the μ_0 is β-warm with respect to this modified target distribution (see the discussion after Corollary 5.2 for details). Moreover, in order to simplify the bounds in the next result, we assume that $\mathcal{L}_H^{2/3} = \mathcal{O}(\mathcal{L})$. (A more general result without this condition can be derived in a similar fashion.)

Corollary 5.2 (HMC mixing for weakly log-concave). *Let μ_0 be a β-warm start, $\delta \in (0,1)$ be fixed and consider $\frac{1}{2}$-lazy HMC chain with leapfrog steps $\mathcal{K} = d^{\frac{1}{2}}$ and step size $\eta^2 = \frac{1}{c\mathcal{L}d^{\frac{4}{3}}}$.*

(a) If the distribution satisfies assumption (5C), then we have

$$\tau_{\mathrm{TV}}^{HMC}(\delta; \mu_0) \leq c \cdot \max\left\{\log\beta, \frac{d^{\frac{4}{3}}\omega}{\delta}\log\left(\frac{\log\beta}{\delta}\right)\right\}. \tag{5.11}$$

(b) If the distribution satisfies assumption (5D) such that $s_0 \leq \frac{\delta^2}{2\beta}$, then we have

$$\tau_2^{HMC}(\delta; \mu_0) \leq c \cdot d^{\frac{5}{6}}\log\left(\frac{\log\beta}{\delta}\right). \tag{5.12}$$

As an immediate consequence, we obtain that the number of gradient evaluations in the two cases is bounded as

$$\mathcal{B}_1 = \max\left\{d^{\frac{1}{2}}\log\beta, \frac{d^{\frac{11}{6}}\omega}{\delta}\log\left(\frac{\log\beta}{\delta}\right)\right\} \quad \text{and} \quad \mathcal{B}_2 = d^{\frac{4}{3}}\log\left(\frac{\log\beta}{\delta}\right).$$

We remark that the bound $\mathcal{B}_1$ for HMC chain improves upon the bound for number of gradient evaluations required by MALA to mix in a similar set-up. In the previous chapter, we showed in Corollary 3.2 that under assumption (5C) (without the Lipschitz-Hessian condition), MALA takes $\mathcal{O}(\frac{d^2\omega}{\delta}\log\frac{\beta}{\delta})$ steps to mix. Since each step of MALA uses one gradient evaluation, our result shows that HMC takes $\mathcal{O}(d^{\frac{1}{6}})$ fewer gradient evaluations. On the other hand, when the target satisfies assumption (5D), Mangoubi et al. [?] showed that MALA takes $\mathcal{O}(d^{\frac{3}{2}}\log\frac{\beta}{\delta})$ steps.[3] Thus even for this case, our result shows that HMC takes $\mathcal{O}(d^{\frac{1}{6}})$ fewer gradient evaluations when compared to MALA.

[3] Note that the authors of the paper [?] assume an infinity-norm third order smoothness which is a stronger assumption than the $\mathcal{L}_H$-Lipschitz Hessian assumption that we made here. Under our setting, the infinity norm third order smoothness is upper bounded by $\sqrt{d}\mathcal{L}_H$ and plugging in this bound changes their rate of MALA from $d^{7/6}$ to $d^{3/2}$.

Proof sketch with Assumption (5C): When the target distribution has a bounded fourth moment (assumption (5C)), proceeding as in the discussion in Section 3.3.4, we can approximate the target distribution $\Pi^\star$ by a strongly log-concave distribution $\widetilde{\Pi}$ with density given by

$$\widetilde{\pi}(x) = \frac{1}{\int_{\mathbb{R}^d} e^{-\widetilde{f}(y)} dy} e^{-\widetilde{f}(x)} \quad \text{where } \widetilde{f}(x) = f(x) + \frac{\lambda}{2} \left\| x - x^\star \right\|_2^2 .$$

Setting $\lambda := \frac{2\mathcal{L}\delta}{d\omega}$ yields that $\widetilde{f}$ is $\lambda/2$-strongly convex, $\mathcal{L} + \lambda/2$ smooth and $\mathcal{L}_H$-Hessian Lipschitz and that the TV distance $d_{\mathrm{TV}}\big(\Pi^\star, \widetilde{\Pi}\big) \leq \delta/2$ is small. The new condition number becomes $\widetilde{\kappa} := 1 + d\omega/\delta$. The new logarithmic-isoperimetric constant is $\widetilde{\psi}_{1/2} = \sqrt{2/\lambda} = (d\omega/(\mathcal{L}\delta))^{1/2}$ (Lemma C.5). Thus, in order to obtain an δ-accurate sample with respect to $\Pi^\star$, it is sufficient to run HMC chain on the new strongly log-concave distribution $\widetilde{\Pi}$ upto $\delta/2$-accuracy. Invoking Corollary 5.1 for $\widetilde{\Pi}$ and doing some algebra yields the bound (5.11).

Proof sketch with Assumption (5D): Lee et al. [155] showed that when the covariance of $\Pi^\star$ has a bounded operator norm, it satisfies isoperimetry inequality (4.8) with $\psi_0 \leq \mathcal{O}(d^{\frac{1}{4}})$. Moreover, using the Lipschitz concentration [101], we have

$$\mathbb{P}_{x \sim \Pi^\star}\left(\left\| x - \mathbb{E}_{\Pi^\star}\left[x\right] \right\|_2 \geq t\psi_0 \cdot \sqrt{d} \right) \leq e^{-ct},$$

which implies that for $\Omega_s = \mathbb{B}\left(\mathbb{E}_{\Pi^\star}\left[x\right], \frac{1}{c}\log\left(\frac{1}{s}\right)\psi_0 \cdot \sqrt{d} \right)$, we have $\Pi^\star(\Omega_s) \geq 1 - s$. In addition, assuming that the gradient is bounded in this ball Ω_s for $s = \frac{\delta^2}{2\beta}$ enables us to invoke Theorem 5.1 and obtain the bound (5.12) after plugging in the values of $\psi_0, \mathcal{K}$ and η.

5.3.4.2 Non-log-concave target

We now briefly discuss how our mixing time bounds in Theorem 5.1 can be applied for distributions whose negative log density may be non-convex. Let $\Pi^\star$ be a log-concave distribution with negative log density as f and isoperimetric constant ψ_0. Suppose that the target distribution $\widetilde{\Pi}^\star$ is a perturbation of $\Pi^\star$ with target density $\widetilde{\pi}^\star$ such that $\widetilde{\pi}^\star(x) \propto e^{-f(x) - \xi(x)}$, where the perturbation $\xi : \mathbb{R}^d \to \mathbb{R}$ is uniformly lower bounded by some constant $-b$ with $b \geq 0$. Then it can be shown that the distribution $\widetilde{\Pi}^\star$ satisfies isoperimetric inequality (4.8) with a constant $\widetilde{\psi}_0 \geq e^{-2b}\psi_0$. For example, such type of a non-log-concave distribution distribution arises when the target distribution is that of a Gaussian mixture model with several components where all the means of different components are close to each other (see e.g. the paper [173]). If a bound on the gradient is also known, Theorem 5.1 can be applied to obtain a suitable mixing time bound. However deriving explicit bounds in such settings is not our main focus, and thereby we omit the details here.

5.4 Numerical experiments

In this section, we numerically compare HMC with MALA and MRW to verify that our suggested step-size and leapfrog steps lead to faster convergence for the HMC algorithm. We use the step-size choices for MALA and MRW given in Table 3.3, whereas the choices for step-size and leapfrog steps for HMC are taken from Corollary 5.1 in this chapter. When the Hessian-Lipschitz constant $\mathcal{L}_H$ is small, our theoretical results suggest that HMC can be run with much larger step-size and much larger number of leapfrog steps (Appendix C.3.1.1). Since our experiments make use of multivariate Gaussian target distribution, the Hessian-Lipschitz constant $\mathcal{L}_H$ is always zero. Consequently we also perform experiments with a more *aggressive* choice of parameters, i.e., larger step-size and number of leapfrog steps. We denote this choice by HMCagg.

In this simulation, we check the dimension d dependency and condition number κ dependency in the multivariate Gaussian case under our step-size choices. We consider sampling from the multivariate Gaussian distribution with density

$$\Pi^{\star}(x) \propto e^{-\frac{1}{2}x^{\top}\Sigma^{-1}x}, \tag{5.13}$$

for some covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. The log density (disregarding constants) and its deriviatives are given by

$$f(x) = \frac{1}{2}x^{\top}\Sigma^{-1}x, \quad \nabla f(x) = \Sigma^{-1}x, \quad \text{and} \quad \nabla^2 f(x) = \Sigma^{-1}.$$

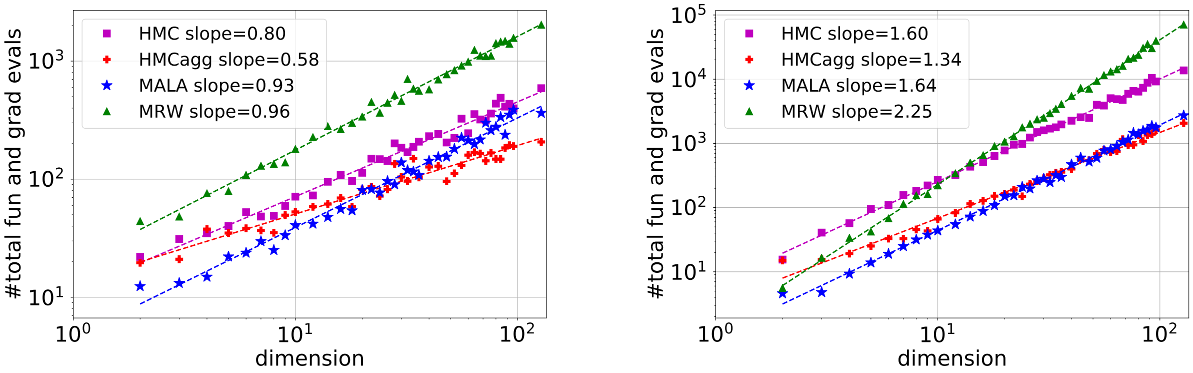
Consequently, the function f is strongly convex with parameter $m = 1/\lambda_{\max}(\Sigma)$ and smooth with parameter $\mathcal{L} = 1/\lambda_{\min}(\Sigma)$. Since $\mathcal{L}_{\mathfrak{p}}$-divergence can not be measure with finitely many samples, we use the error in quantiles along different directions for convergence diagnostics. Using the exact quantile information for each direction for Gaussian, we measure the error in the 75% quantile of the relative sample distribution and the true distribution in the *least favorable direction*, i.e., along the eigenvector of Σ corresponding to the eigenvalue $\lambda_{\max}(\Sigma)$. The *quantile mixing time* is defined as the smallest iteration when this relative error falls below a constant $\delta = 0.04$. We use $\mu_0 = \mathcal{N}\left(0, \mathcal{L}^{-1}\mathbb{I}_d\right)$ as the initial distribution. To make the comparison with MRW and MALA fair, we compare the number of total function and gradient evaluations instead of number of iterations. For HMC, the number of gradient evaluations is $\mathcal{K}$ times the number of outer-loop iterations.

For every case of simulation, the parameters for HMC-$(\mathcal{K}, \eta)$ are chosen according to the warm start case in Corollary 5.1 with $\mathcal{K} = 4 \cdot d^{1/4}$, and for MRW and MALA are chosen according to Table 3.3. As alluded to earlier, we also run the HMC chain a more aggressive choice of parameters, and denote this chain by HMCagg. For HMCagg, both the step-size and leapfrog steps are larger (Appendix C.3.1.1): $\mathcal{K} = 4 \cdot d^{1/8}\kappa^{1/4}$ where we take into account that $\mathcal{L}_H$ is zero for Gaussian distribution. We simulate 100 independent runs of the four chains, MRW, MALA, HMC, HMCagg, and for each chain at every iteration we compute the quantile error across the 100 samples from 100 independent runs of that chain. We compute the minimum number of total function and gradient evaluations required for the relative quantile error to fall below $\delta = 0.04$. We repeat this computation 10 times and report the averaged number of total function and gradient evaluations in Figure 5.1. To examine the scaling of the number of evaluations with the dimension d, we vary $d \in \{2, 4, \ldots, 128\}$. For each chain, we also fit a least squares line for the number of total function and gradient evaluations with respect to dimension d on the log-log scale, and report the slope in the figure. Note that a slope of α would denote that the number of evaluations scales as d^{α}.

(a) Dimension dependency for fixed κ: First, we consider the case of fixed condition number. We fix $\kappa = 4$ while we vary the dimensionality d of the target distribution is varied over $\{2, 4, \ldots, 128\}$. The Hessian Σ in the multivariate Gaussian distribution is chosen to be diagonal and the square roots of its eigenvalues are linearly spaced between 1.0 to 2.0. Figure 5.1(a) shows the dependency of the number of total function and gradient evaluations as a function of dimension d for the four Markov chains on the log-log scale. The least-squares fits of the slopes for HMC, HMCagg, MALA and MRW are $0.80(\pm 0.12)$, $0.58(\pm 0.15)$, $0.93(\pm 0.13)$ and $0.96(\pm 0.10)$, respectively, where standard errors of the regression coefficient is reported in the parentheses. These numbers indicate close correspondence to the theoretical slopes (reported in Table 5.2 and Appendix C.3.1.1) are $0.92, 0.63, 1.0, 1.0$ respectively.

(b) Dimension dependency for $\kappa = d^{2/3}$: Next, we consider target distributions such that their condition number varies with d as $\kappa = d^{2/3}$, where d is varied from 2 to 128. To ensure such a scaling for κ, we choose the Hessian Σ for the multivariate Gaussian distribution to be diagonal and set the square roots of its eigenvalues linearly spaced between 1.0 to $d^{1/3}$. Figure 5.1(b) shows the dependency of the number of total function and gradient evaluations as a function of dimension d for the four random walks on the log-log scale. The least squares fits yield the slopes as $1.60(\pm 0.09)$, $1.34(\pm 0.17)$, $1.64(\pm 0.11)$ and $2.25(\pm 0.08)$ for HMC, HMCagg, MALA and MRW, respectively, where standard errors of the regression coefficient are reported in the parentheses. Recall that the theoretical guarantees for HMC (Table C.2), HMCagg (Table C.3), MALA and MRW (Table 5.2) yield that these exponent should be close to 1.58, 1.46, 1.67 and 2.33 respectively. Once again, we observe a good agreement of the numerical results with that of our theoretical results.

Remark: We would like to caution that the aggressive parameter choices for HMCagg are well-informed only when the Hessian-Lipschitz constant $\mathcal{L}_H$ is small—which indeed is the case for the Gaussian target distributions considered above. When general log-concave distributions are considered, one may use the more general choices recommended in Corollary C.1. See Appendix C.3 for an in-depth discussion on different scenarios and the optimal parameter choices derived from our theory.



(a) $\kappa = 4$

(b) $\kappa = d^{\frac{2}{3}}$

Figure 5.1: Average number of total function and gradient evaluations as a function of dimension for four random walks on multivariate Gaussian density (5.13) where the covariance has a condition number κ that is (a) constant 4 and (b) scales with dimension d. With suggested step-size and leapfrog steps in Corollary 5.1, the number of total function and gradient evaluations of HMC has a smaller dimension dependency than that of MALA or MRW. Since the target distributon is Gaussian and the Hessian-Lipschitz constant $\mathcal{L}_H$ is zero, larger step-size and larger number of leapfrog steps can be chosen according to Appendix C.3.1.1. The plots does show that HMCagg with larger step-size and larger number of leapfrog steps uses smaller number of total function and gradient evaluations to achieve the same quantile mixing.

5.5 Proofs

We establish Theorem 5.1 using conductance profile based bounds using Propositions 4.1 and 4.2. However, to lower bound the conductance profile as in Proposition 4.2, we need to derive bound on the overlap between the transition distributions of HMC. We show that by (i) bounding the overlap between proposal distributions of HMC at two nearby points are close, and (ii) showing that the Metropolis-Hastings step only modifies the proposal distribution by a relatively small amount. This control is provided by Lemma 5.1. Putting the pieces together yields the proof of Theorem 5.1 in Section 5.5.2. However, proving Lemma 5.1 requires fairly technical analysis and is deferred to Appendix C.

5.5.1 Overlap bounds for HMC

In this subsection, we derive two important bounds for the Metropolized HMC chain: (1) first, we quantify the overlap between proposal distributions of the chain for nearby points, and, (2) second, we show that the distortion in the proposal distribution introduced by the Metropolis-Hastings accept-reject step can be controlled if an appropriate step-size is chosen.

Putting the two pieces together enables us to invoke Proposition 4.2 to prove Theorem 5.1.

In order to do so, we begin with some notation. Let $\mathcal{T}$ denote the transition operator of the HMC chain with leapfrog integrator taking step-size η and number of leapfrog updates $\mathcal{K}$. Let $\mathcal{P}_x$ denote the proposal distribution at $x \in \mathcal{X}$ for the chain before the accept-reject step and the lazy step. Let $\mathcal{T}_x^{\text{before-lazy}}$ denote the corresponding transition distribution after the proposal and the accept-reject step, before the lazy step. By definition, we have

$$\mathcal{T}_x(A) = \zeta\delta_x(A) + (1-\zeta)\mathcal{T}_x^{\text{before-lazy}}(A) \qquad \text{for any measurable set } A \in \mathcal{B}(\mathcal{X}). \tag{5.14}$$

Our proofs make use of the Euclidean ball $\mathcal{R}_s$ defined in equation (5.18). At a high level, the HMC chain has bounded gradient inside the ball $\mathcal{R}_s$ for a suitable choice of s, and the gradient of the log-density gets too large outside such a ball making the chain unstable in that region. However, since the target distribution has low mass in that region, the chain's visit to the region outside the ball is a rare event and thus we can focus on the chain's behavior inside the ball to analyze its mixing time.

In the next lemma, we state the overlap bounds for the transition distributions of the HMC chain. For a fixed univeral constant c, we require

$$\mathcal{K}^2\eta^2 \leq \frac{1}{4\max\left\{d^{\frac{1}{2}}\mathcal{L}, d^{\frac{2}{3}}\mathcal{L}_H^{\frac{2}{3}}\right\}}, \qquad \text{and} \tag{5.15a}$$

$$\eta^2 \leq \frac{1}{c\mathcal{L}}\min\left\{\frac{1}{\mathcal{K}^2}, \frac{1}{\mathcal{K}d^{\frac{1}{2}}}, \frac{1}{\mathcal{K}^{\frac{2}{3}}d^{\frac{1}{3}}\left(\frac{\mathcal{M}^2}{\mathcal{L}}\right)^{\frac{1}{3}}}, \frac{1}{\mathcal{K}\frac{\mathcal{M}}{\mathcal{L}^{\frac{1}{2}}}}, \frac{1}{\mathcal{K}^{\frac{2}{3}}d}\frac{\mathcal{L}}{\mathcal{L}_H^{\frac{2}{3}}}, \frac{1}{\mathcal{K}^{\frac{4}{3}}\frac{\mathcal{M}}{\mathcal{L}^{\frac{1}{2}}}}\left(\frac{\mathcal{L}}{\mathcal{L}_H^{\frac{2}{3}}}\right)^{\frac{1}{2}}\right\}. \tag{5.15b}$$

Lemma 5.1. *Consider a $(\mathcal{L}, \mathcal{L}_H, s, \psi_{\mathfrak{e}}, \mathcal{M})$-regular target distribution (cf. Assumption (5A)) with Ω the convex measurable set satisfying (5.6). Then with the parameters $(\mathcal{K}, \eta)$ satisfying $\mathcal{K}\eta \leq \frac{1}{4\mathcal{L}}$ and condition (5.15a), the HMC-$(\mathcal{K}, \eta)$ chain satisfies*

$$\sup_{\|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0\|_2 \leq \frac{\mathcal{K}\eta}{4}} d_{\mathrm{TV}}\left(\mathcal{P}_{\mathbf{q}_0}, \mathcal{P}_{\widetilde{\mathbf{q}}_0}\right) \leq \frac{1}{2}. \tag{5.16a}$$

If, in addition, condition (5.15b) holds, then we have

$$\sup_{x\in\Omega} d_{\mathrm{TV}}\left(\mathcal{P}_x, \mathcal{T}_x^{\textit{before-lazy}}\right) \leq \frac{1}{8}. \tag{5.16b}$$

See Appendix C.1 for the proof.

Lemma 5.1 is crucial to the analysis of HMC as it enables us to apply the conductance profile based bounds discussed in Section **??**. It reveals two important properties of the Metropolized HMC. First, from equation (5.16a), we see that proposal distributions of HMC at two different points are close if the two points are close. This is proved by controlling the KL-divergence of the two proposal distributions of HMC via change of variable formula. Second, equation (5.16b) shows that the accept-reject step of HMC is well behaved inside Ω provided the gradient is bounded by $\mathcal{M}$.

5.5.2 Proof of Theorem 5.1

We are now equipped to prove our main theorem. In order to prove Theorem 5.1, we begin by using Proposition 4.2 and Lemma 5.1 to derive an explicit bound for on the HMC conductance profile. Given the assumptions of Theorem 5.1, conditions (5.15a) and (5.15b) hold, enabling us to invoke Lemma 5.1 in the proof.

Define the function $\Psi_\Omega : [0,1] \mapsto \mathbb{R}_+$ as

$$\Psi_\Omega(v) = \begin{cases} \dfrac{1}{32} \cdot \min\left\{1, \dfrac{\mathcal{K}\eta}{64\psi_{\mathfrak{e}}} \log^{\mathfrak{e}}\left(\dfrac{1}{v}\right)\right\} & \text{if } v \in \left[0, \frac{1-s}{2}\right]. \\ \dfrac{\mathcal{K}\eta}{2048\psi_{\mathfrak{e}}}, & \text{if } v \in \left(\frac{1-s}{2}, 1\right]. \end{cases} \tag{5.17}$$

This function acts as a lower bound on the truncated conductance profile. Define the Euclidean ball

$$\mathcal{R}_s = \mathbb{B}\left(x^\star, \mathfrak{a}(s)\sqrt{\frac{d}{m}}\right), \tag{5.18}$$

and consider a pair $(x, y) \in \mathcal{R}_s$ such that $\|x-y\|_2 \le \frac{1}{4}\mathcal{K}\eta$. Invoking the decomposition (5.14) and applying triangle inequality for ζ-lazy HMC, we have

$$\begin{aligned} d_{\mathrm{TV}}\big(\mathcal{T}_x, \mathcal{T}_y\big) &\le \zeta + (1-\zeta)\, d_{\mathrm{TV}}\big(\mathcal{T}_x^{\text{before-lazy}}, \mathcal{T}_y^{\text{before-lazy}}\big) \\ &\le \zeta + (1-\zeta)\left(d_{\mathrm{TV}}\big(\mathcal{T}_x^{\text{before-lazy}}, \mathcal{P}_y\big) + d_{\mathrm{TV}}\big(\mathcal{P}_x, \mathcal{P}_y\big) + d_{\mathrm{TV}}\big(\mathcal{P}_x, \mathcal{T}_y^{\text{before-lazy}}\big)\right) \\ &\overset{(i)}{\le} \zeta + (1-\zeta)\left(\frac{1}{4} + \frac{1}{2} + \frac{1}{4}\right) \\ &= 1 - \frac{1-\zeta}{4}, \end{aligned}$$

where step (i) follows from the bounds (5.16a) and (5.16b) from Lemma 5.1. For $\zeta = \frac{1}{2}$, substituting $\rho = \frac{1}{8}$, $\Delta = \frac{1}{4}\mathcal{K}\eta$ and the convex set $\Omega = \mathcal{R}_s$ into Proposition 4.2, we obtain that

$$\Phi_\Omega(v) \ge \frac{1}{32} \cdot \min\left\{1, \frac{\mathcal{K}\eta}{64\psi_{\mathfrak{e}}} \log^{\mathfrak{e}}\left(1 + \frac{1}{v}\right)\right\}, \quad \text{for } v \in \left[0, \frac{1-s}{2}\right].$$

Here $\mathfrak{e}$ equals to $\frac{1}{2}$ or 0, depending on the assumption (4.8). By the definition of the truncated conductance profile (4.3), we have that $\widetilde{\Phi}_\Omega(v) \ge \frac{\mathcal{K}\eta}{2048\psi_{\mathfrak{e}}}$ for $v \in \left[\frac{1-s}{2}, 1\right]$. As a consequence, Ψ_Ω is effectively a lower bound on the truncated conductance profile. Note that the assumption (5A) ensures the existence of Ω such that $\Pi^\star(\Omega) \ge 1 - s$ for $s = \frac{\delta^2}{2\beta^2}$. Putting the pieces together and applying Proposition 4.1 with the convex set Ω concludes the proof of the theorem.

5.6 Conclusion and future directions

In this chapter, we derived non-asymptotic bounds on mixing time of Metropolized Hamiltonian Monte Carlo for log-concave distributions. By choosing appropriate step-size and number of leapfrog steps, we obtain mixing-time bounds for HMC that are smaller than the mixing-time bounds for MALA from Chapter 3. This improvement can be seen as the benefit of using multi-step gradients in HMC. An interesting open problem is to determine whether our HMC mixing-time bounds are tight for log-concave sampling under the assumptions made in this chapter. Even though, we focused on the problem of sampling only from strongly and weakly log-concave distribution, our Theorem 5.1 can be applied to general distributions including nearly log-concave distributions as mentioned in Section 5.3.4.2. It would be interesting to determine the explicit expressions for mixing-time of HMC for more general target distributions. Finally, defining a analysis of HMC with stochastic gradients for large data settings can be another interesting future direction.

Chapter 6

Novel Algorithms for Constrained Sampling on Convex Sets

So far we discussed the problems of unconstrained sampling, meaning that the target distribution was supported on $\mathbb{R}^d$. In this chapter, we turn our attention to sampling under constraints, or, equivalently, dealing with distributions supported on a proper subset of $\mathbb{R}^d$. In particular, we study here a certain class of MCMC algorithms designed for the problem of drawing samples from the uniform distribution over a polytope. We consider polytopes specified as

$$\mathcal{K} := \{x \in \mathbb{R}^d \mid Ax \leq b\}, \tag{6.1}$$

parameterized by the known matrix-vector pair $(A, b) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$, where $A \in \mathbb{R}^{n\times d}$ with $n \geq d$ is assumed to be a full-rank matrix. The target distribution $\Pi^\star$ is defined as the uniform distribution over $\mathcal{K}$:

$$\pi^\star(x) = \frac{1}{\text{vol}(\mathcal{K})}\mathbf{1}(x \in \mathcal{K}), \tag{6.2}$$

where $\text{vol}(\mathcal{K})$ denotes the volume of the set $\mathcal{K}$. Life previous chapters, we are interested in a thorough understanding ofnon-asymptotic mixing time for obtaining δ-accurate samples from the target $\Pi^\star$, and for this set-up we are specially interested in the scaling of the mixing time bounds as a function of the pair (n, d). We start with a basic introduction in Section 6.1, before summarizing our contributions in Section 6.1.1, and the organization of the remainder of the chapter in Section 6.1.2.

6.1 Introduction

The problem of sampling uniformly from a polytope is important in various applications and methodologies. For instance, it underlies various methods for computing randomized approximations to polytope volumes. There is a long line of work on sampling methods being used to obtain randomized approximations to the volumes of polytopes and other convex bodies (e.g., [166, 149, 16, 163, 57]). Polytope sampling is also useful in developing fast randomized algorithms for convex optimization [17] and sampling contingency tables [139], as well as in randomized methods for approximately solving mixed integer convex programs [120, 121]. Sampling from polytopes is also related to simlations of the hard-disk model in statistical physics [140], as well as to simulations of error events for linear programming in communication [88].

Many MCMC algorithms have been studied for sampling from polytopes, and more generally, from convex bodies. Some early examples include the Ball Walk [166] and the Hit-and-Run algorithm [16, 163], which apply to sampling from general convex bodies. Although these algorithms can be applied to polytopes, they do not exploit any special structure of the problem. In contrast, the Dikin walk introduced by Kannan and Narayanan [139] is specialized to polytopes, and thus can achieve faster convergence rates than generic algorithms. The Dikin walk was the first sampling algorithm based on a connection to interior point methods for solving linear programs. More specifically, as we discuss in detail below, it constructs proposal distributions based on the standard logarithmic barrier for a polytope. In a later paper, Narayanan [187] extended the Dikin walk to general convex sets equipped with self-concordant barriers.

For a polytope defined by n constraints, Kannan and Narayanan [139] proved an upper bound on the mixing time of the Dikin walk that scales linearly with n. In many applications, the number of constraints n can be much larger than the number of variables d. It is also possible that for a given problem, various constraints are redundant or repeated. For such problems, linear dependence on the number of constraints is not desirable. Consequently, it is natural to ask if it is possible to design a sampling algorithm whose mixing time scales in a sub-linear manner with the number of constraints. Our main contribution is to investigate and answer this question in affirmative—in particular, by designing and analyzing two sampling algorithms with provably faster convergence rates than the the Dikin walk while retaining its advantages over the ball walk and the hit-and-run methods.

6.1.1 Our contributions

We introduce and analyze a new random walk, which we refer to as the *Vaidya walk* since it is based on the *volumetric-logarithmic barrier* introduced by [237]. We show that for a polytope in $\mathbb{R}^d$ defined by n-constraints, the Vaidya walk mixes in $\mathcal{O}\left(n^{1/2}d^{3/2}\right)$ steps, whereas the Dikin walk [139] has mixing time bounded as $\mathcal{O}\left(nd\right)$. So the Vaidya walk is better in the regime $n \gg d$. We also propose the *John walk*, which is based on the *John ellipsoidal algorithm* in optimization. We show that the John walk has a mixing time of $\mathcal{O}\left(d^{2.5} \cdot \log^4(n/d)\right)$ and conjecture that a variant of it could achieve $\mathcal{O}\left(d^2 \cdot \text{poly-log}(n/d)\right)$ mixing time. We show that when compared to the Dikin walk, the per-iteration computational complexities of the Vaidya walk and the John walk are within a constant factor and a poly-logarithmic in n/d factor respectively. Thus, in the regime $n \gg d$, the overall upper bound on the complexity of generating an approximately uniform sample follows the order

Dikin walk $\gg$ Vaidya walk $\gg$ John walk.

6.1.2 Organization

The remainder of the chapter is organized as follows. In Section 6.2, we discuss many polynomial-time random walks on convex sets and polytopes, and motivate the starting point for the new random walks. We introduce the new random walks in Section 6.3, and then provide the main mixing time results in Section 6.4 (with an overview of results in Table 3.1). We discuss the computational complexity of the different random walks and demonstrate the contrast between the random walks for several illustrative examples in Section 6.5. We present the proof of the mixing time for the Vaidya walk in Section 6.6 and defer the analysis of the John walk to the appendix. We conclude with possible extensions of our work in Section 6.7.

6.2 Related Work

There are various algorithms to sample a vector from the uniform distribution over $\mathcal{K}$, including the ball walk [166] and hit-and-run algorithms [163]. To be clear, these two algorithms apply to the more general problem of sampling from a convex set; when applied to the polytope $\mathcal{K}$, Table 6.1 shows their complexity relative to the Vaidya walk analyzed in this chapter. Most closely related to our chapter is the Dikin walk proposed by Kannan and Narayanan [139], and a more general random walk on a Riemannian manifold studied by Narayanan [187]. Both of these random walks, as with the Vaidya and John walks, can be viewed as randomized versions of the interior point methods used to solve linear programs, and more generally convex programs equipped with suitable barrier functions.

In order to motivate the form of the Vaidya and John walks proposed in this chapter, we begin by discussing the ball walk, and then the Dikin walk.[1]

Ball walk: The ball walk of [166] is simple to describe: when at a point $x \in \mathcal{K}$, it draws a new point u from a Euclidean ball of radius $r > 0$ centered at x. Here the radius r is a step size parameter in the algorithm. If the proposed point u belongs to the polytope $\mathcal{K}$, then the walk moves to u; otherwise, the walk stays at x. On the one hand, unlike the walks analyzed in this chapter, the ball walk applies to any convex set, but on the other, its mixing time depends on the condition number $\gamma_{\mathcal{K}}$ of the set $\mathcal{K}$, given by

$$\gamma_{\mathcal{K}} = \inf_{R_{\text{in}}, R_{\text{out}} > 0} \Big\{ \frac{R_{\text{out}}}{R_{\text{in}}} \quad | \quad \mathbb{B}(x, R_{\text{in}}) \subseteq \mathcal{K} \subseteq \mathbb{B}(y, R_{\text{out}}) \quad \text{for some } x, y \in \mathcal{K} \Big\}. \tag{6.3}$$

Mixing time of the ball walk has been improved greatly since it was introduced [138, 136, 158]. Nonetheless, as shown in Table 6.1, the mixing time of the ball walk gets slower when the

[1] For discussion of hit-and-run, please refer to the full paper [47].

condition of the set is large; for instance, it scales[2] as d^6 for a set with condition number $\gamma_{\mathcal{K}} = d^2$. One approach to tackle bad conditioning is to use rounding as a pre-processing step, where the set is rounded to bring it in a near-isotropic position, i.e., reduce the condition $\gamma_{\mathcal{K}}$ to near-constant before sampling from it. Nonetheless, these algorithms are themselves based on several rounds of sampling algorithms and the current best algorithm by [170] puts a convex body into approximately isotropic position, i.e., $\widetilde{\mathcal{O}}(\sqrt{d})$ rounding with a running time of $\widetilde{\mathcal{O}}(d^4)$ where we have omitted the dependence on log-factors. If one has more information about the structure of the convex set (and not just oracle access as required by the ball walk), one can potentially exploit it to design fast sampling algorithms which are unaffected by the conditioning of the set thereby reducing the need of the (expensive) pre-processing step. One such algorithm is the Dikin walk for polytopes which we describe next.

Dikin walk: The Dikin walk [139] is similar in spirit, except that it proposes a point drawn uniformly from a *state-dependent* ellipsoid known as the Dikin ellipsoid [71, 192]. It then applies an accept-reject step to adjust for the difference in the volumes of these ellipsoids at different states. The state-dependent choice of the ellipsoid allows the Dikin walk to adapt to the boundary structure. A key property of the Dikin ellipsoid of unit radius—in contrast to the Euclidean ball that underlies the ball walk—is that it is always contained within $\mathcal{K}$, as is known from classic results on interior point methods [192]. Furthermore, the Dikin walk is affine invariant, meaning that its behavior does not change under linear transformations of the problem. As a consequence, the Dikin mixing time does not depend on the condition number $\gamma_{\mathcal{K}}$. In a variant of this random walk [187], uniform proposals in the ellipsoid are replaced by Gaussian proposals with covariance specified by the ellipsoid, and it is shown that with high probability, the proposal falls within the polytope.

The Dikin walk is closely related to the interior point methods for solving linear programs. In order to understand the Vaidya and John walks, it is useful to understand this connection in more detail. Suppose that our goal is to optimize a convex function over the polytope $\mathcal{K}$. A barrier method is based on converting this constrained optimization problem to a sequence of unconstrained ones, in particular by using a barrier to enforce the linear constraints defining the polytope. Letting $a_i^\top$ denote the i-th row vector of matrix A, the *logarithmic-barrier* for the polytope $\mathcal{K}$ given by the function

$$\mathcal{F}(x) \coloneqq -\sum_{i=1}^{n} \log(b_i - a_i^T x). \tag{6.4}$$

For each $i \in [n]$, we define the scalar $s_{x,i} \coloneqq (b_i - a_i^T x)$, and we refer to the vector $s_x \coloneqq (s_{x,1}, \ldots, s_{x,n})^\top$ as the *slackness at* x.

[2] Although, very recently [158] improved the mixing time of the ball walk for isotropic sets which have $\gamma_{\mathcal{K}} = \mathcal{O}(\sqrt{d})$ improved from $\mathcal{O}\left(d^3\right)$ to $\mathcal{O}\left(d^{2.5}\right)$.

Each step of an interior point algorithm [26] involves (approximately) solving a linear system involving the Hessian of the barrier function, which is given by

$$\nabla^2 \mathcal{F}(x) \coloneqq \sum_{i=1}^{n} \frac{a_i a_i^\top}{s_{x,i}^2}. \tag{6.5}$$

In the Dikin walk [139], given a current iterate x, the algorithm chooses a point uniformly at random from the ellipsoid

$$\{u \in \mathbb{R}^d \mid (u-x)^\top D_x (u-x) \leq R\}, \tag{6.6}$$

where $D_x \coloneqq \nabla^2 \mathcal{F}(x)$ is the Hessian of the log barrier function, and $R > 0$ is a user-defined radius. In an alternative form of the Dikin walk [224], the proposal vector $u \in \mathbb{R}^d$ is drawn randomly from a Gaussian centered at x, and with covariance equal to a scaled copy of $(D_x)^{-1}$. Note that in contrast to the ball walk, the proposal distribution now depends on the current state.

6.3 Two New Random Walks

Let us now define the two walks introduced in this chapter.

6.3.1 Vaidya walk

For the *Vaidya walk* analyzed in this chapter, we generate proposals from the ellipsoids defined, for each $x \in \operatorname{int}(\mathcal{K})$, by the positive definite matrix

$$V_x \coloneqq \sum_{i=1}^{n} (\sigma_{x,i} + \beta_{\mathrm{V}}) \frac{a_i a_i^\top}{s_{x,i}^2}, \qquad \text{where} \tag{6.7a}$$

$$\beta_{\mathrm{V}} \coloneqq d/n \quad \text{and} \quad \sigma_x \coloneqq \left(\frac{a_1^\top (\nabla^2 \mathcal{F}_x)^{-1} a_1}{s_{x,1}^2}, \ldots, \frac{a_n^\top (\nabla^2 \mathcal{F}_x)^{-1} a_n}{s_{x,n}^2} \right)^\top. \tag{6.7b}$$

The entries of the the vector σ_x are known as the leverage scores assciated with the matrix $\nabla^2 \mathcal{F}_x$ (6.5), and are commonly used to measure the importance of rows in a linear system [174]. The matrix V_x is related to the Hessian of the function $x \mapsto \mathcal{V}_x$ given by

$$\mathcal{V}_x \coloneqq \log\det \nabla^2 \mathcal{F}_x + \beta_{\mathrm{V}} \mathcal{F}_x. \tag{6.8}$$

This particular combination of the *volumetric barrier* and the *logarithmic barrier* was introduced by Vaidya et al. [237, 238] in the context of interior point methods, hence our name for the resulting random walk.

More concretely, the Vaidya walk with radius parameter $r > 0$, denoted by $\text{VW}(r)$ for short, is defined by a Gaussian proposal distribution: given a current state $x \in \text{int}(\mathcal{K})$, it proposes a new point by sampling from the multivariate Gaussian distribution $\mathcal{N}\left(x, \frac{r^2}{\sqrt{nd}} V_x{}^{-1}\right)$. In analytic terms, the proposal density at x is given by

$$p_x^{\text{V}}(z) := p_{\text{Vaidya}}(x, z) = \sqrt{\det V_x} \left(\frac{\sqrt{nd}}{2\pi r^2}\right)^{d/2} \exp\left(-\frac{\sqrt{nd}}{2r^2}\,(z-x)^\top V_x (z-x)\right). \qquad (6.9)$$

The proposal step is then followed by an accept-reject step (equation 2.7). Thus, the overall transition distribution for the walk at state x is defined by a density given by

$$q_{\text{Vaidya}}(x, z) = \begin{cases} \min\left\{p_x^{\text{V}}(z), p_z^{\text{V}}(x)\right\}, & z \in \mathcal{K} \text{ and } z \neq x, \\ 0, & z \notin \mathcal{K}, \end{cases}$$

and a probability mass at x, given by $1 - \int_{z \in \mathcal{K}} \min\{p_x(z), p_z(x)\}\, dz$. In Algorithm 4, we summarize the different steps of the Vaidya walk.

Algorithm 4: Vaidya Walk with parameter r ($\text{VW}(r)$)

```
Input: Parameter $r$ and $x_0 \in \text{int}(\mathcal{K})$
Output: Sequence $x_1, x_2, \ldots$
for $i = 0, 1, \ldots$ do
    $C_i \sim$ Fair Coin
    if $C_i = Heads$ then $x_{i+1} \leftarrow x_i$   // lazy step
    else
        $\xi_{i+1} \sim \mathcal{N}(0, \mathbb{I}_d)$
        $z_{i+1} = x_i + \frac{r}{(nd)^{1/4}} V_{x_i}^{-1/2} \xi_{i+1}$ // propose a new state
        if $z_{i+1} \notin \mathcal{K}$ then $x_{i+1} \leftarrow x_i$   // reject an infeasible proposal
        else
            $\alpha_{i+1} = \min\left\{1, \frac{p_{\text{Vaidya}}(z_{i+1}, x_i)}{p_{\text{Vaidya}}(x_i, z_{i+1})}\right\}$
            $U_{i+1} \sim U[0, 1]$
            if $U_{i+1} \geq \alpha_{i+1}$ then $x_{i+1} \leftarrow x_i$   // reject even a valid proposal
            else $x_{i+1} \leftarrow z_{i+1}$      // accept the proposal
        end
    end
end
```

6.3.2 John walk

Finally, let us describe the John walk. For any vector $w \in \mathbb{R}^n$, let $W := \text{diag}(w)$ denote the diagonal matrix with $W_{ii} = w_i$ for each $i \in [n]$. Let $S_x = \text{diag}(s_x)$ denote the slackness matrix at x. It is easy to see that S_x is positive semidefinite for all $x \in \mathcal{K}$, and strictly positive definite for all $x \in \text{int}(\mathcal{K})$. The (scaled) inverse covariance matrix underlying the John walk is given by

$$J_x \coloneqq \sum_{i=1}^{n} \zeta_{x,i} \frac{a_i a_i^\top}{s_{x,i}^2}, \tag{6.10}$$

where for each $x \in \text{int}(\mathcal{K})$, the weight vector $\zeta_x \in \mathbb{R}^n$ is obtained by solving the convex program

$$\zeta_x \coloneqq \arg\min_{w \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} w_i - \frac{1}{\alpha_{\text{J}}} \log\det(A^\top S_x^{-1} W^{\alpha_{\text{J}}} S_x^{-1} A) - \beta_{\text{J}} \sum_{i=1}^{n} \log w_i \right\}, \tag{6.11}$$

with $\beta_{\text{J}} \coloneqq d/2n$ and $\alpha_{\text{J}} \coloneqq 1 - 1/\log_2(1/\beta_{\text{J}})$. Lee et al. [152] proposed the convex program (6.11) associated with the *John weights* ζ_x, with the aim of searching for the best member of a family of volumetric barrier functions. They analyzed the use of the John weights in the context of speeding up interior point methods for solving linear programs; here we consider them for improving the mixing time of a sampling algorithm. Note that the roots of the optimization problem (6.11) lie in the electric network inverse problem [231], in which weights are optimized so as to obtain an electric network with minimum effective resistance.

More precisely, the John walk is similar to the Vaidya walk except that the proposals at state $x \in \text{int}(\mathcal{K})$ are generated from the multivariate Gaussian distribution $\mathcal{N}\left(x, \frac{r^2}{d^{3/2} \cdot \log_2^4(2n/d)} J_x{}^{-1}\right)$, where the matrix J_x is defined by equation (6.10), and $r > 0$ is a constant. The proposal step is then followed by an accept-reject step (equation 2.7). We use $\mathcal{T}_{\text{John}}$ to denote the resulting transition operator for the John walk. See Algorithm 5 for an overview of the algorithm, where we also use the notation

$$p_{\text{John}}(x, z) = \sqrt{\det J_x} \left(\frac{d^{3/2} \cdot \log_2^4(2n/d)}{2\pi r^2} \right)^{d/2} \exp\left(-\frac{d^{3/2} \cdot \log_2^4(2n/d)}{2r^2} \; (z - x)^\top J_x (z - x) \right).$$

6.3.3 Mixing time comparisons of all walks

Table 6.1 provides a summary of the mixing time bounds and per step complexity and the effective per sample complexity for various random walks, including the Vaidya and John walks analyzed in this chapter. In addition to the Ball Walk, Hit-and-Run, Dikin, Vaidya and John walks, we also show scalings for the recently introduced Riemannian Hamiltonian Monte Carlo (RHMC) on polytopes by [154] and the John's walk based on exact John ellipsoids studied by [103]. The details of per iteration cost for the new random walks is discussed in Section 6.5.1. We now compare and contrast the complexities of these random walks.

For the Dikin, Vaidya and John walks, the transition probability from a point x to y does not change under an affine transformation B, i.e., $\Theta(x, dy) = \Theta(Bx, Bdy)$ where Θ denotes

Algorithm 5: John Walk with parameter r (JWr)

```
Input: Parameter $r$ and $x_0 \in \mathrm{int}(\mathcal{K})$
Output: Sequence $x_1, x_2, \ldots$
for $i = 0, 1, \ldots$ do
    $C_i \sim$ Fair Coin
    if $C_i = \mathit{Heads}$ then $x_{i+1} \leftarrow x_i$  // lazy step
    else
        $\xi_{i+1} \sim \mathcal{N}(0, \mathbb{I}_d)$
        $z_{i+1} = x_i + \frac{r}{d^{3/4} \cdot \log_2^2(2n/d)} J_{x_i}^{-1/2} \xi_{i+1}$ // propose a new state
        if $z_{i+1} \notin \mathcal{K}$ then $x_{i+1} \leftarrow x_i$  // reject an infeasible proposal
        else
            $\alpha_{i+1} = \min\left\{1, \frac{p_{\mathrm{John}}(z_{i+1}, x_i)}{p_{\mathrm{John}}(x_i, z_{i+1})}\right\}$
            $U_{i+1} \sim U[0, 1]$
            if $U_{i+1} \geq \alpha_{i+1}$ then $x_{i+1} \leftarrow x_i$  // reject even a valid proposal
            else $x_{i+1} \leftarrow z_{i+1}$      // accept the proposal
        end
    end
end
```

the transition kernel (2.2) for the random walk. Consequently, the mixing time bounds for these random walks have no dependence on the condition number $\gamma_{\mathcal{K}}$ of the set $\mathcal{K}$. Such an affine invariance comes in handy for many polytopes for which the value of $\gamma_{\mathcal{K}}$ scales with dimension d. We can see from Table 6.1, that compared to ball walk and hit-and-run, Vaidya walk mixes significantly faster if the warmness parameter β (2.6) is large or $n \ll d\gamma_{\mathcal{K}}^4$. The condition number $\gamma_{\mathcal{K}}$ of polytopes with polynomially many faces can not be $\mathcal{O}(d^{\frac{1}{2}-\epsilon})$ for any $\epsilon > 0$ but can be arbitrarily larger, even exponential in dimension d [139]. For such polytopes, Vaidya walk mixes faster as long as $n \ll d^3$ (and even for larger n when $\gamma_{\mathcal{K}}$ is large). It takes $\mathcal{O}(\sqrt{n/d})$ fewer steps compared to Dikin walk and thus provides a practical speed up over all range of d.

From a warm start, the geodesic random walk of Lee and Vempala [154] has $\mathcal{O}(nd^{3/4})$ mixing time, and thus mixes faster (up to constants) compared than the Vaidya walk (respectively the John walk) when the number of constraints n is is bounded as $n \ll d^{3/2}$ (respectively $n \ll d^{7/4}$). For larger numbers of constraints, the Vaidya and John walks exhibit faster mixing. More generally, it is clear that the rate of John walk has *almost* the best order across all the walks for reasonably large values of $n \gg d^2$.

From a warm start, the Riemannian Hamiltonian Monte Carlo on polytopes introduced by [154] has $\mathcal{O}\left(nd^{2/3}\right)$ mixing time, and thus mixes faster (up to constants) compared than the Vaidya walk (respectively the John walk) when the number of constraints n is is bounded as $n \ll d^{5/3}$ (respectively $n \ll d^{11/6}$). For larger numbers of constraints, the Vaidya and John walks exhibit faster mixing. More generally, it is clear that the rate of John walk has *almost* the best order across all the walks for reasonably large values of $n \gg d^2$.

Let us also compare the (exact) John walk due to [103] with the (approximate) John walk studied in this chapter. A notable feature of their random walk is that its mixing time is independent of the number of constraints and the per iteration cost also depends linearly on the number of constraints. Nonetheless, the dependence on d, for both the mixing time (d^7) and the per iteration cost ($nd^4 + d^8$) is quite poor. In contrast, the per iteration cost for our John walk is nd^2 and the mixing time has only a poly-logarithmic dependence on n.

Random walk	$\tau_{\mathrm{TV}}(\delta;\mu_0)$	**Iteration cost**	**Per sample cost**
Ball walk$^{\#}$ [136]	$d^2\gamma_{\mathcal{K}}^2$	nd	$nd^3\gamma_{\mathcal{K}}^2$
Hit-and-Run [169]	$d^2\gamma_{\mathcal{K}}^2$	nd	$nd^3\gamma_{\mathcal{K}}^2$
Dikin walk [139]	nd	nd^2	n^2d^3
RHMC walk [157]	$nd^{2/3}$	nd^2	$n^2d^{2.67}$
John's walk† [103]	d^7	$nd^4 + d^8$	$nd^{11} + d^{15}$
Vaidya walk (this chapter)	$n^{1/2}d^{3/2}$	nd^2	$n^{1.5}d^{3.5}$
John walk (this chapter)	$d^{5/2}\log^4\left(\frac{2n}{d}\right)$	$nd^2\log^2 n$	$nd^{4.5}$
Improved John walk‡ (this chapter)	$d^2\,\omega_{n,d}$	$nd^2\log^2 n$	nd^4

Table 6.1: Upper bounds on computational complexity of random walks on the polytope $\mathcal{K} = \{x \in \mathbb{R}^d | Ax \leq b\}$ defined by the matrix-vector pair $(A, b) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$ with a warm-start. For simplicity, here we ignore the logarithmic dependence on the warmness parameter and the tolerance δ. The iteration cost terms of order nd^2 arise from linear system solving, using standard and numerically stable algorithms, for n equations in d dimensions; algorithms with best possible theoretical complexity nd^{ν} for $\nu < 1.373$ are not numerically stable enough for practical use. $^{\#}$Mixing time of the Ball walk has been recently improved to $\mathcal{O}\left(d^2\gamma_{\mathcal{K}}\right)$ for near isotropic convex bodies by [158]. While ball walk, Hit-and-run are affected by the condition number $\gamma_{\mathcal{K}}$ of the set, the Dikin and RHMC walks have quadratic dependence on the number of constraints n. †John's walk by [103] (based on the exact John ellipsoids) has linear dependence on n but poor dependence on d. In contrast, the Vaidya walk has sub-quadratic dependence on n and significantly better dependence on d. Furthermore, the John walk (based on approximate John's ellipsoids) analyzed in this chapter has linear dependence with reasonable dependence on the dimensions d. ‡The mixing time bound for the improved John walk with poly-logarithmic factor $\omega_{n,d}$ is conjectured (Section 6.4.3).

6.3.4 Visualization of Dikin, Vaidya and John walks

In order to gain intuition about the three interior point based methods—namely, the Dikin, Vaidya and John walks—it is helpful to discuss how their underlying proposal distributions change as a function of the current point x. All three walks are based on Gaussian proposal distributions with inverse covariance matrices of the general form

$$\sum_{i=1}^{n} w_{x,i} \frac{a_i a_i^\top}{s_{x,i}^2},$$

where $w_{x,i} > 0$ corresponds to a state-dependent weight associated with the i-th constraint. The Dikin walk uses the weights $w_{x,i} = 1$; the Vaidya walk uses the weights $w_{x,i} = \sigma_{x,i} + \beta_{\mathrm{V}}$; and the John walk uses the weights $w_{x,i} = \zeta_{x,i}$. For simplicity, we refer to these weights as the Dikin, Vaidya and John weights. The i-th weight characterize the importance of the i-th linear constraint in constructing the inverse covariance matrix. A larger value of the weight $w_{x,i}$ relative to the total weight $\sum_{i=1}^{n} w_{x,i}$ signifies more importance for the i-th linear constraint for the point x.

Figure 6.1a illustrates the difference in three weights as we move points inside the polytope $[-1, 1]^2$. When the point x is in the middle of the unit square formed by the four constraints, all walks exhibit equal weight for every constraint. When the point x is closer to the bottom-left boundary, the Vaidya and John weights assign larger weights to the bottom and the left constraints, while the weights for top and right constraints decrease. Note that the total sum of Vaidya weights and that of John weights remains constant independent of the position of the point x.

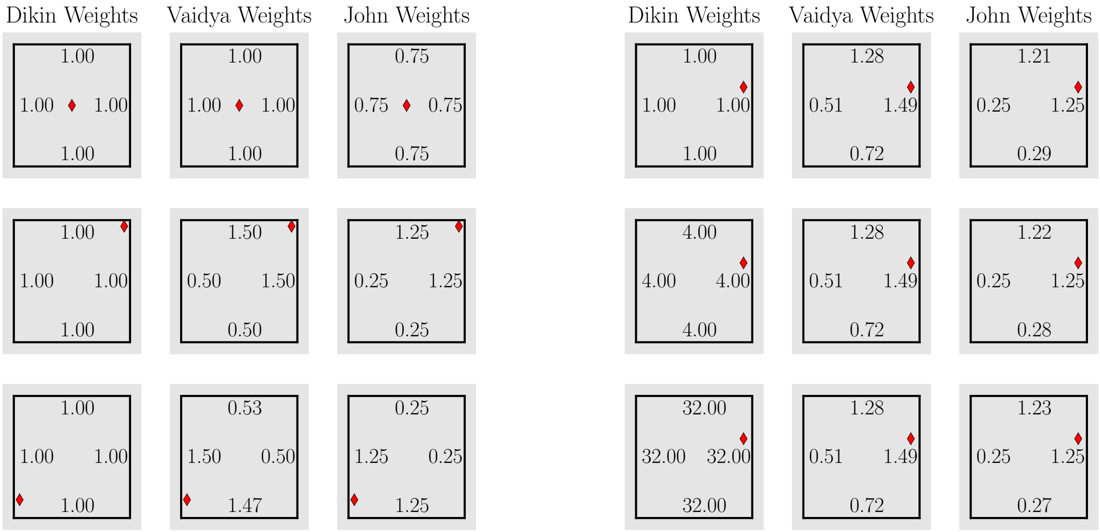
In Figure 6.1b-6.2b, we demonstrate that the Vaidya walk and the John walk are better at handling repeated constraints. Note that we can define the square $[-1, 1]^2$ as

$$[-1,1]^2 = \left\{ x \in \mathbb{R}^2 \middle| Ax \leq b, A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}, b = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}. \tag{6.12}$$

Simply repeating the rows of the matrix A several times changes the mathematical formulatiton of the polytope, but does not change the shape of the polytope. We define the square with constraints repeated $n/4$ times $\mathcal{S}_{n/4}$ as

$$\mathcal{S}_{n/4} = \left\{ x \in \mathbb{R}^2 \middle| A_{n/4} x \leq b_{n/4}, A_{n/4} = \begin{bmatrix} A \\ \vdots \\ {\scriptstyle \times(n/4)} \end{bmatrix}, b_{n/4} = \begin{bmatrix} b \\ \vdots \\ {\scriptstyle \times(n/4)} \end{bmatrix}, \right\} \tag{6.13}$$

where A and b were defined above. We denote effective weight for each distinct constraint as the sum of weights corresponding to the same constraint. Using this definition, the effective Dikin weight, which is $n/4$, is thus affected by the repeating of constraints. Consequently,

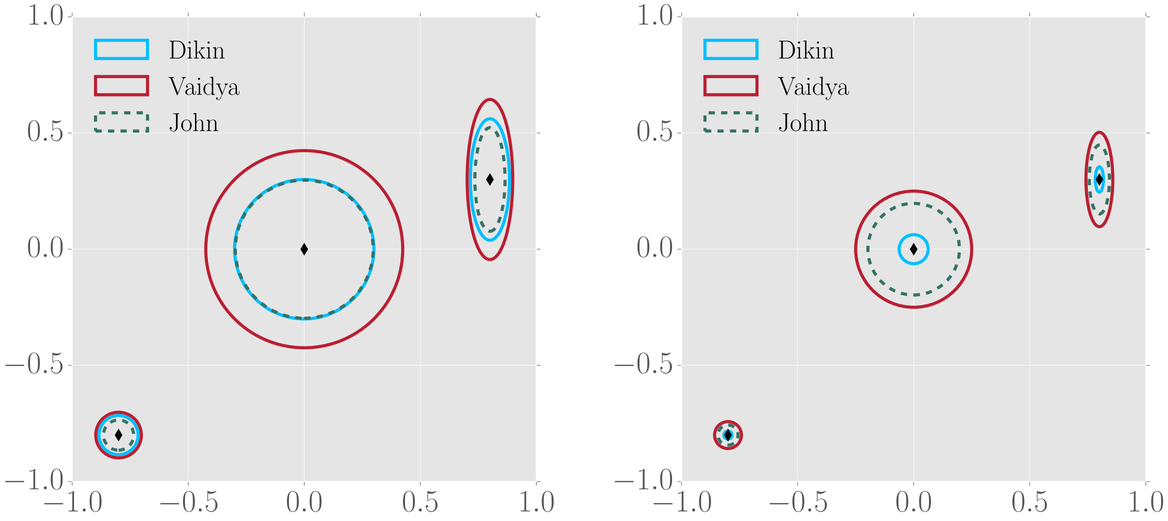


(a) Weights for different locations and a fixed number of constraints n.

(b) Effective weights for a fixed location and different number of constraints n

Figure 6.1: Visualization of the weights on the square with repeated constraints $\mathcal{S}_{n/4}$ for the different random walks. The number mentioned next to the boundary lines denotes the effective weight for the location x (denoted by diamond) for the corresponding constraint. **(a)** $n = 4$ is common across rows and $x = (0, 0)$ for the top row, $(0.9, 0.9)$ for the middle and $(-0.9, -0.7)$ for the bottom row. The Dikin weights are independent of x, the Vaidya and the John weights for a constraint increase if the location x is closer to it. **(b)** $x = (0.85, 0.30)$ is common across rows, and $n{=}4$ for the top row, $n = 16$ for the middle and $n{=}128$ for the bottom row. The effective Dikin weight for each constraint increases linearly with n but for the Vaidya and John walk adaptively, the weights get adjusted such that the sum of their weights is always of the order of the dimension d.

the Dikin ellipsoid is much smaller for polytopes with repeated constraints. However, the Vaidya and John weights do not change as observed in the Figure 6.1b. Such a property of these two weights implies that the Vaidya and John ellipsoids are not too small even for very large number of constraints. And we observe such a phenomenon in Figures 6.2a-6.2b where the repetition of rows in the matrix A leads to very small Dikin ellipsoid but large Vaidya and John ellipsoid. A few other numerical computations also suggest that the Vaidya and John ellipsoids are moder adaptive when compared to Dikin ellipsoids when the number of constraints is large. Nonetheless, such a claim is only based on heuristics and is presented simply to provide an intuition that the new ellipsoids are better behaved than Dikin ellipsoids and thereby motivated the design of the new random walks.



(a) $n = 32$

(b) $n = 2048$

Figure 6.2: Visualization of the proposal distribution on the square with repeated constraints $\mathcal{S}_{n/4}$ for the different random walks. **(a, b)** Unit ellipsoids associated with the covariances of the random walks at different states x on the square with repeated constraints $\mathcal{S}_{n/4}$. Clearly, all these ellipsoids adapt to the boundary but increasing n has a profound impact on the volume of the Dikin ellipsoids and comparatively less impact on the Vaidya and John ellipsoids.

6.4 Main results

With the basic background in place, we now state our main results with a warm start in Section 6.4.1, and a deterministic start in Section 6.4.2. In Section 6.4.3, we propose a variant of the John walk, known as the *improved John walk*, and conjecture that it has a better mixing time bound than that of the John walk.

6.4.1 Mixing time bounds for warm start

We use $\mathcal{T}_{\text{Vaidya}(r)}$ to denote the transition operator (2.4) associated with the Vaidya walk with parameter r.

Theorem 6.1. *For any $\delta \in (0, 1]$, the Vaidya walk with parameter $r = 10^{-4}$ and a β-warm starting distribution μ_0 satisfies*

$$d_{\text{TV}}\big(\mathcal{T}^{\ell}_{\textit{Vaidya}(r)}(\mu_0), \Pi^{\star}\big) \leq \delta \qquad \textit{for all } \ \ell \geq cn^{1/2}d^{3/2} \log\left(\frac{\sqrt{\beta}}{\delta}\right). \tag{6.14}$$

We provide the proof of Theorem 6.1 in Section **??**. Theorem 6.1 quantifies the dependence of mixing time of the Vaidya walk on different aspects of the sampling problem at hand. The specific choice $r = 10^{-4}$ is for theoretical purposes; in practice, we find that substantially larger values can be used.[3]

The mixing time of the Dikin walk is $\mathcal{O}(nd)$ and thereby the speed up for the Vaidya walk is $\mathcal{O}(\sqrt{n/d})$ number of steps. In Section 6.5.1, we show that the per iteration cost for the two walks is of the same order. Since $n \geq d$ for closed polytopes in $\mathbb{R}^d$, the effective cost until convergence (iteration complexity multiplied by number of iterations required) for the Vaidya walk is at least of the same order as of the Dikin walk, and significantly smaller when $n \gg d$. Thus, the Vaidya walk has a clear advantage for the problems where the number of constraints is significantly larger than the number of variables involved.

We use $\mathcal{T}_{\text{John}(r)}$ to denote the transition operator (2.4) associated with the John walk with parameter r. Let us now state our result for the mixing time of the John walk.

Theorem 6.2. *Suppose that $n \leq \exp(\sqrt{d})$. Then, for any $\delta \in (0, 1]$, the John walk with parameter $r = 10^{-5}$ and a β-warm starting distribution μ_0 satisfies*

$$d_{\text{TV}}\left(\mathcal{T}^{\ell}_{John(r)}(\mu_0), \Pi^\star\right) \leq \delta \qquad \text{for all } \ell \geq c\, d^{2.5} \log^4\left(\frac{n}{d}\right) \log\left(\frac{\sqrt{\beta}}{\delta}\right). \tag{6.15}$$

We prove this theorem in Appendix E. Note that the mixing time bound for the John walk depends only on the number of constraints n via a logarithmic factor, and so is almost independent of n. Consequently, it has a mixing time that is polynomial in d even if the number of constraints n scales exponentially in $\sqrt{d}$. Further, we show in Section 6.5.1 that the cost to execute one step of the John walk is of the same order as of the Dikin walk up to a poly-logarithmic factor in n. Thus, John walk is more efficient than the Dikin and Vaidya walks when $n \gg d^2$.

6.4.2 Mixing time bounds from deterministic start

The mixing time bounds in Theorem 6.1 and 6.2 depend on the warmness β of the initial distribution. In some applications, it may not be easy to find an β-warm initial distribution. In such cases, we can consider starting the random walk from a deterministic point $x_0 \in \text{int}(\mathcal{K})$ that is not too close to the boundary $\partial\mathcal{K}$. Indeed, such a point can be found using standard optimization methods—e.g., using a Phase-I method for Newton's algorithm. (See Section 11.5.4 in the book [26] for more discussion.)

Given such a deterministic initialization, our mixing time guarantees depend on the distance of the starting point from the boundary. This dependence involves the following notion of $\mathfrak{m}$-centrality:

[3] A larger than optimal r leads to an undesirable high rejection rate. In practice, we can fine tune r by performing a binary search over the interval $[10^{-4}, 1]$ and keeping track of the rejection rate of the samples during the run of the Markov chain for a given choice of r. A choice of $r > 1$ is obviously bad because then the Vaidya ellipsoid will have poor overlap with polytopes near the boundary, causing high rejection rate and slow down of the chain.

Definition 6.1. *A point $x \in \text{int}(\mathcal{K})$ is called* $\mathfrak{m}$-central *if for any chord $\overline{ef}$ with end points $e, f \in \partial\mathcal{K}$ passing through x, we have $\|e - x\|_2 / \|f - x\|_2 \leq \mathfrak{m}$.*

Assuming that it is started at an $\mathfrak{m}$-central point x_0, the Dikin walk (Algorithm 1 in the paper [139]) has polynomial mixing time. The authors showed that when the walk moves to a new state for the first time, the distribution of the iterate is $\mathcal{O}\left((\sqrt{n}\mathfrak{m})^d\right)$-warm with respect to the distribution $\Pi^\star$. Thus, for a deterministic start, we can use the Dikin walk in the beginning to provide a warm start to the Vaidya (or John) walk. This motivates us to define the following hybrid walk.

Given an $\mathfrak{m}$-central point x_0, simulate the Dikin walk until we observe a new state. Note that due to *lazyness* and the accept-reject step, the chain can stay at the starting point for several steps before making the first move a new state. Let k_1 denote the (random) number of steps taken to make the first move to a new state. After k_1 steps, we run the walk $\text{VW}(r)$ with x_{k_1} as the initial point. We call such a walk as "$\mathfrak{m}$-central Dikin-start-Vaidya-walk" with parameter r. Let $\mathcal{T}_{\text{Dikin}}$ denote the transition operator of the Dikin walk stated above. Then, we have the following mixing time bound for this hybrid walk.

Corollary 6.1. *Any $\mathfrak{m}$-central Dikin-start-Vaidya-walk with parameter $r = 10^{-4}$ satisfies*

$$d_{\text{TV}}\left(\mathcal{T}^{\ell}_{\textit{Vaidya}(r)}\left(\mathcal{T}^{\ell_1}_{\textit{Dikin}}(\delta_{x_0})\right), \Pi^\star\right) \leq \delta \qquad \textit{for all } \ell \geq cn^{1/2}d^{5/2}\log\left(\frac{n\mathfrak{m}}{\delta}\right),$$

where ℓ_1 is a geometric random variable with $\mathbb{E}\left[\ell_1\right] \leq c'$, and $c, c' > 0$ are universal constants.

The mixing rate is logarithmic in $n\mathfrak{m}$ and has an extra factor of d compared to the bounds in Theorem 6.1. However, guaranteeing a warm start for a general polytope is hard but obtaining a central point involves only a few steps of optimization. Consequently, the hybrid walk and the guarantees from Corollary 6.1 come in handy for all such cases. Once again we observe that the mixing time bounds are improved by a factor of $\mathcal{O}(\sqrt{n/d})$ when compared to Dikin walk from an $\mathfrak{m}$-central start [139, 187] which had a mixing time of $\mathcal{O}(nd^2)$. The proof follows immediately from Theorem 1 by Kannan et al. [139] and Theorem 6.1 and is thereby omitted.

In a similar fashion, we can provide a polynomial time guarantee for a modified John walk from a deterministic start. We can consider a hybrid random walk that starts at an $\mathfrak{m}$-central point, simulates the Dikin walk until it makes the first move to a new state, and from there onwards simulates the John walk. Such a chain would have a mixing time of $\mathcal{O}\left(d^{3.5}\text{poly-log}(n, d, \mathfrak{m})\right)$. For brevity, we omit a formal statement of this result.

6.4.3 Conjecture on improved John walk

From our analysis, we suspect that it is possible to improve the mixing time bound of $\mathcal{O}\left(d^{2.5}\text{poly-log}(n/d)\right)$ in Theorem 6.2 by considering a variant of the John walk. In particular, we conjecture that a random walk with proposal distribution given by $\mathcal{N}\left(x, \frac{r^2}{d\cdot\text{poly-log}(n/d)}{J_x}^{-1}\right)$

for a suitable choice of r has an $\mathcal{O}\left(d^2\text{poly-log}(n/d)\right)$ mixing time from a warm start. We refer to this random walk as the *improved John walk*, and denote its transition operator by $\mathcal{T}_{\text{John}^+}$. Let us now give a formal statement of our conjecture on its mixing rate.

Conjecture 6.1. *For any $\delta \in (0, 1]$, the improved John walk with parameter r_0 and a β-warm starting distribution μ_0 satisfies*

$$d_{\text{TV}}\left(\mathcal{T}^{\ell}_{John^+(r)}(\mu_0), \Pi^\star\right) \leq \delta \quad \textit{for all} \quad \ell \geq c\, d^2 \, \log_2^{c'}\left(\frac{2n}{d}\right) \log\left(\frac{\sqrt{\beta}}{\delta}\right). \tag{6.16}$$

Note that this conjecture involves quadratic (degree two) scaling in d; this exponent of two matches the sum of exponents for d and n in the mixing time bounds for both the Dikin and Vaidya walks from a warm-start. Consquently, the improved John walk would have better performance than the Dikin, Vaidya and John walks for almost all ranges of (n, d), apart from possible poly-logarithmic factors in the ratio n/d.

6.5 Numerical experiments

In this section, we first analyze the per-iteration cost to implement of three walks. We show that while the Dikin walk has the best per-iteration cost, the per-iteration cost of the Vaidya walk is only twice of that of Dikin walk and the per-iteration cost of the John walk is only of order $\log_2(2n/d)$ larger. Second, we demonstrate the speed-up gained by the Vaidya walk over the Dikin walk for a warm start on different polytopes.

6.5.1 Per iteration cost

We now show that the per iteration cost of the Dikin, Vaidya and John walks is of the same order. The proposal step of Vaidya walk requires matrix operations like matrix inversion, matrix multiplication and singular value decomposition (SVD). The accept-reject step requires computation of matrix determinants, besides a few matrix inverses and matrix-vector products. The complexity of all aforementioned operations is $\mathcal{O}\left(nd^2\right)$. Thus, per iteration computational complexity for the Vaidya walk is $\mathcal{O}\left(nd^2\right)$. In theory, the matrix computations for the Dikin walk can be carried out in time nd^ν for an exponent $\nu < 2$, but such algorithms are not stable enough for practical use.

Both the Dikin and Vaidya walks requires an SVD computation for inverting the Hessian of Dikin barrier $\nabla^2\mathcal{F}_x$. In addition for the Vaidya walk, we have to invert the matrix V_x, which leads to almost twice the computation time of the Dikin walk per step. This difference can be observed in practice.

For the John walk we need to compute the weights ζ_x at each point which involves solving the program (6.11). Lee et al. [152] argued that the convex program (6.11) for obtaining John walk's weights is strongly convex under appropriate norm. They proved that solving this program requires $\log^2 n$ number of gradient steps where each gradient step has the computational complexity of a linear system solve ($\mathcal{O}\left(nd^2\right)$ using a numerically stable routine). Thus, the overall cost for the John walk is of the same order as of the Dikin walk up to a poly-logarithmic factor in the pair (n, d).

In practice, for the John walk, the combined effect of logarithmic factors in the number of steps and the cost to implement each step is pretty significant. This extra factor becomes a bottleneck for the overall run time for the convergence of the Markov chain. Consequently, the John walk is not suitable for polytopes with moderate values of n and d, and its mixing time bounds are computationally superior to the Dikin and Vaidya walks only for the polytopes with $n \gg d \gg 1$.

6.5.2 Simulations

We now present simulation results for the random walks in $\mathbb{R}^d$ for $d = 2, 10$ and 50 with initial distribution $\mu_0 = \mathcal{N}(0, \sigma_d^2 \mathbb{I}_d)$ and target distribution being uniform, on the following polytopes:

Set-up 1 : The set $[-1, 1]^2$ defined by different number of constraints.

Set-up 2 : The set $[-1, 1]^d$ for $d \in \{2, 3, 4, 5, 6, 7\}$ for $n = \{2d, 2d^2, 2d^3\}$ constraints.

Set-up 3 : Symmetric polytopes in $\mathbb{R}^2$ with n-randomly-generated-constraints.

Set-up 4 : The interior of regular n-polygons on the unit circle.

Set-up 5 : Hyper cube $[-1, 1]^d$ for $d = 10$ and 50.

We choose σ_d such that the warmness parameter β is bounded by 100. We provide implementations of the Dikin, Vaidya and John walks in python and a jupyter notebook at the github repository `https://github.com/rzrsk/vaidya-walk`.

We use the following three ways to compare the convergence rate of the Dikin and the Vaidya walks: (1) comparing the approximate mixing time of a particular subset of the polytope—smaller value is associated with a faster mixing chain; (2) comparing the plot of the empirical distribution of samples from multiple runs of the Markov chain after k steps—if it appears *more uniform* for smaller k, the chain is deemed to be faster; and (3) contrasting the sequential plots of one dimensional projection of samples for a single long run of the chain—*less smooth* plot is associated with effective and fast exploration leading to a faster mixing [260]. Note that MCMC convergence diagnostics is a hard problem, especially in high dimensions, and since the methods outlined above are heuristic in nature we expect our experiments to not fully match our theoretical results.

In **Set-up 1**, we consider the polytope $[-1, 1]^2$ which can be represented by exactly 4 linear constraints (see Section 6.3.4). Suppose that we repeat the rows of the matrix A, and then run the Dikin and Vaidya walks with the new A. Given the larger number of constraints, our theory predicts that the random walks should mix more slowly. In Figure 6.3c and 6.3d, we plot the empirical distribution obtained by the Dikin walk and Vaidya walk, starting from 200 i.i.d initial samples, for $n = 64$ and 2048. The empirical distribution plot shows that having large n significantly slows the mixing rate of the Dikin walk, while the effect on the Vaidya walk is much less. Further, we also plot the scaling of the approximate mixing time $\hat{k}_{\text{mix}}$ (defined below) for this simulation as a function of the number of constraints n in Figure 6.3b. For **Set-up 2**, we plot $\hat{k}_{\text{mix}}$ as a function of the dimensions d in Figures 6.3e-6.3g, for the random walks on $[-1, 1]^d$ where the hypercube is parametrized by different number of constraints $n \in \{2d, 2d^2, 2d^3\}$. The approximate mixing time is defined with respect to the set $\mathcal{S}_d = \{x \in \mathbb{R}^d | \, |x_i| \geq c_d \; \forall i \in [d]\}$ where c_d is chosen such that $\Pi^\star(\mathcal{S}_d) = 1/2$. In particular, for a fixed value of n, let $\hat{\mathcal{T}}^k$ denote the empirical measure after k-iterations across 2000 experiments. The approximate mixing time $\hat{k}_{\text{mix}}$ is defined as

$$\hat{k}_{\text{mix}} := \min \left\{ k \middle| \Pi^\star(\mathcal{S}_d) - \hat{\mathcal{T}}^k(\mathcal{S}_d) \leq \frac{1}{20} \right\}, \tag{6.17}$$

We choose such a set since the set covers the regions near to the boundary of the polytope which are not covered well by the chosen initial distribution. We make the following observations:

(6F) The slopes of the best-fit lines, for $\hat{k}_{\text{mix}}$ versus n in the log-log plot in Figure 6.3b, are 0.88 and 0.45 for Dikin and Vaidya walks respectively. This observation reflects a near-linear and sub-linear dependence on n for a fixed d for the mixing time of the Dikin walk and the Vaidya walk respectively.

(6G) In Figures 6.3e-6.3g, once again we observe a more significant effect of increasing the number of constraints on the approximate mixing time $\hat{k}_{\text{mix}}$. We list the slopes of the best fit lines on these log-log plots in Table 6.2. These slopes correspond to the exponents for d for the approximate mixing time. From the table, we can observe that these experiments agree with the mixing time bounds of $\mathcal{O}(nd)$ for the Dikin walk and $\mathcal{O}(n^{0.5}d^{1.5})$ for the Vaidya walk.

In **Set-up 3**, we compare the plots of the empirical distribution of 200 runs of the Dikin walk and the Vaidya walk for different values of k, for symmetric polytopes in $\mathbb{R}^2$ with n-randomly-generated-constraints. We fix $b_i = 1$. To generate a_i, first we draw two uniform random variables from $[0, 1]$ and then flip the sign of both of them with probability 1/2 and assign these values to the vector a_i. The resulting polytope is always a subset of the square $\mathcal{K} = [-1, 1]^2$ and contains the diagonal line connecting the points $(-1, 1)$ and $(1, -1)$. From Figure 6.4a-6.4b, we observe that while there is no clear winner for the case $n = 64$, the Vaidya walk mixes mixes significantly faster than the Dikin walk for the polytope defined by 2048 constraints.

In **Set-up 4**, the constraint set is the regular n-polygons inscribed in the unit circle. A similar observation as in **Set-up 3** can be made from Figure 6.4c-6.4d: the Vaidya walk mixes at least as fast as the Dikin walk and mixes significantly faster for large n.

In **Set-up 5**, we examine the performance of the Dikin walk and the Vaidya walk on hyper-cube $[-1, 1]^d$ for $d = 10, 50$. We plot the one dimensional projections onto a random

No. of Constraints	DW Theoretical	VW Theoretical	DW Experiments	VW Experiments
$n = 2d$	2.0	2.0	1.58	1.72
$n = 2d^2$	3.0	2.5	2.80	2.48
$n = 2d^3$	4.0	3.0	3.84	2.75

Table 6.2: Value of the exponent of dimensions d for the theoretical bounds on mixing time and the observed approximate mixing time of the Dikin walk (DW) and the Vaidya walk (VW) for $[-1, 1]^d$ described by $n = 2d, 2d^2, 2d^3$ constraints. The theoretical exponents are based on the mixing time bounds of $\mathcal{O}(nd)$ for the Dikin walk and $\mathcal{O}(n^{0.5}d^{1.5})$ for the Vaidya walk. The experimental exponents are based on the results from the simulations described in **Set-up 2** in Section 6.5.2. Clearly, the exponents observed in practice are in agreement with the theoretical rates and imply the faster convergence of the Vaidya walk compared to the Dikin walk for large number of constraints.

normal direction of all the samples from a single run up to $10,000$ steps. The Vaidya sequential plot looks more jagged than that of the Dikin walk for $d = 10, n = 5120$. For other cases, we do not have a clear winner. Such an observation is consistent with the $\mathcal{O}(\sqrt{n/d})$ speed up of the Vaidya walk which is apparent when the ratio n/d is large.

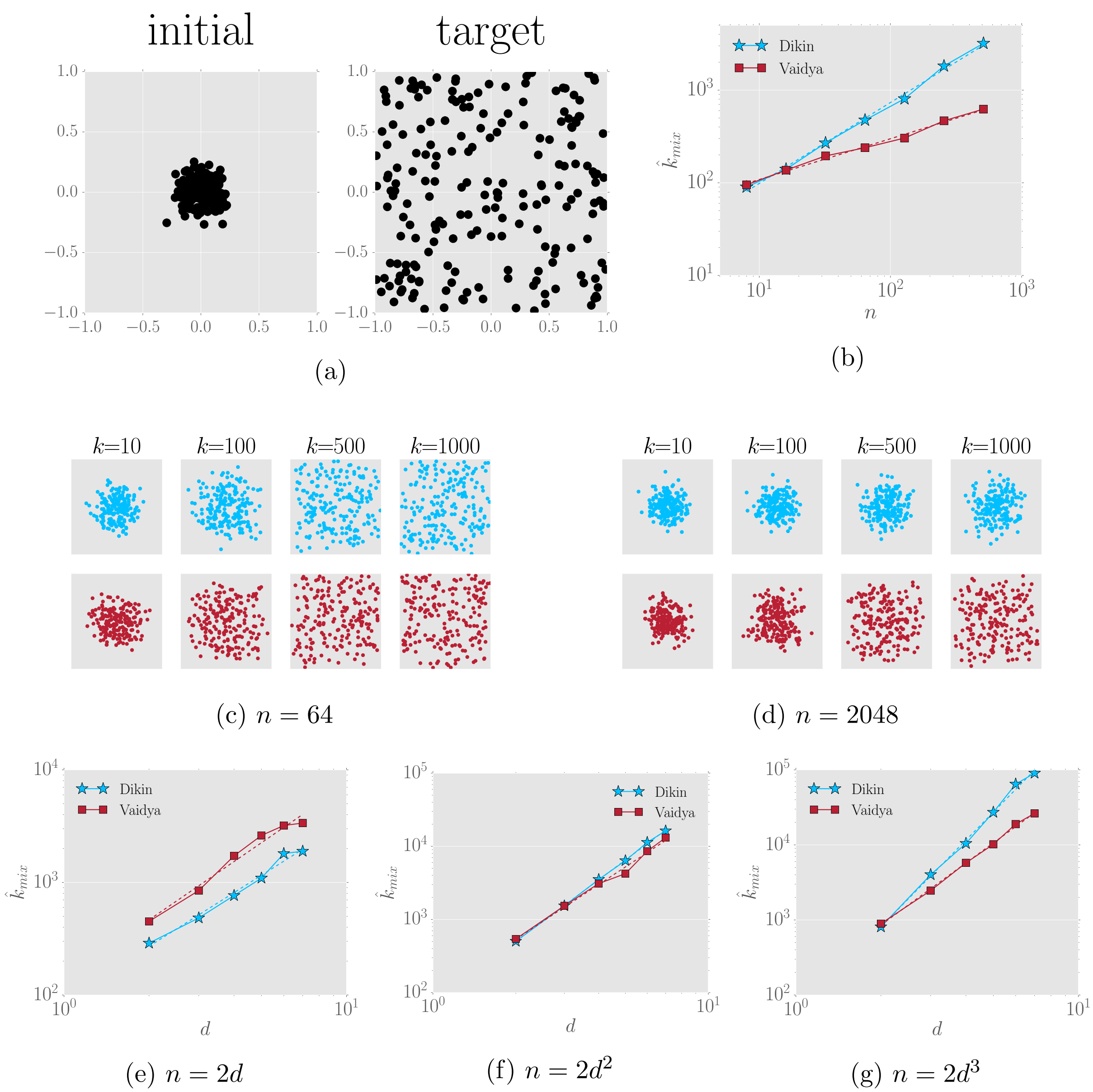
6.6 Proofs

We start with a proof sketch for Vaidya walk in Section 6.6.1, followed by some auxiliary results in Section 6.6.2 which we then use to prove Theorem 6.1 in Section 6.6.3. We provide a proof sketch for John walk's mixing time in Section 6.6.4 and defer its proof to Appendix E. Proofs of the auxiliary results from this section are provided in Appendix D.

6.6.1 Proof sketch for the Vaidya walk

In this subsection, we provide a high-level sketch of the main ingredients of the main proof. Like the proofs in earlier chapters, we use conductance-based mixing time bounds. Our main proof relies on Lovász [163]'s work that characterizes the conductance of Markov chains on a convex set using Hilbert metric. Precisely, Lovász [163] showed that a Markov chain has good conductance if it makes jumps to regions with large overlaps from two nearby points and the mixing time depends inversely on the maximum Hilbert metric between such nearby points. Using this argument, it remains to make sure that the ellipsoid radius is chosen properly such that the ellipsoids remain inside the polytope and the ellipsoids corresponding to two different points x and y overlap a lot even if the points x and y are relatively far apart.

The conductance-based argument has been used for analyzing the ball walk [166, 167], Hit-and-run [163, 169] and the Dikin walk [187, 139, 224]. We refer the reader to the survey



(a)

(b)

(c) $n = 64$

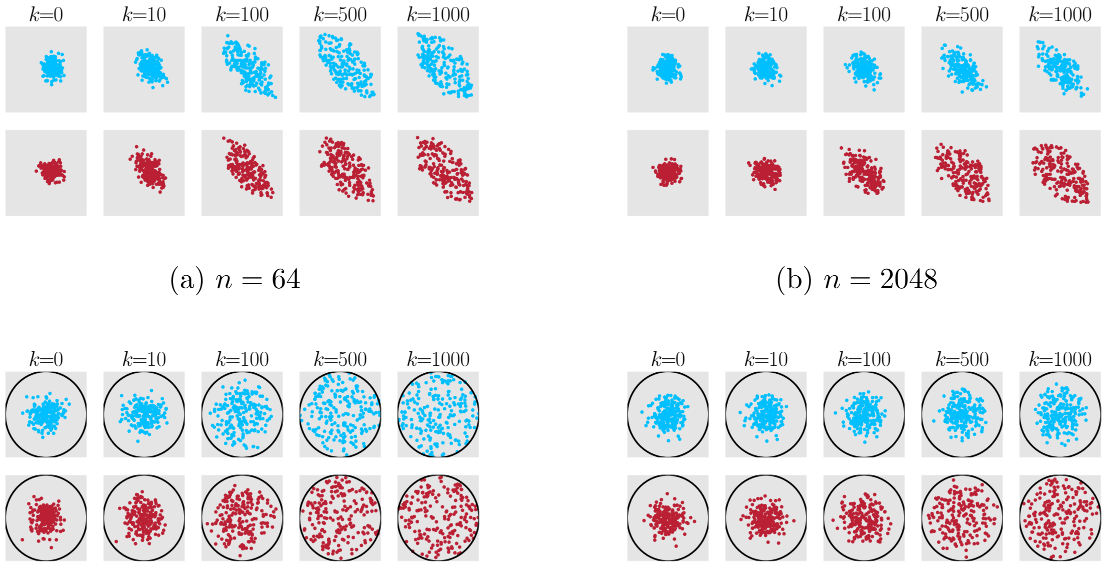
(d) $n = 2048$

(e) $n = 2d$

(f) $n = 2d^2$

(g) $n = 2d^3$

Figure 6.3: Comparison of the Dikin and Vaidya walks on the polytope $\mathcal{K} = [-1, 1]^2$. **(a)** Samples from the initial distribution $\mu_0 = \mathcal{N}(0, 0.04\,\mathbb{I}_2)$ and the uniform distribution on $[-1, 1]^2$. **(b)** Log-log plot of $\hat{k}_{\text{mix}}$ (6.17) versus the number of constraints (n) for a fixed dimension $d = 2$. **(c, d)** Empirical distribution of the samples for the Dikin walk (blue/top rows) and the Vaidya walk (red/bottom rows) for different values of n at iteration k = 10, 100, 500 and 1000. **(e, f, g)** Log-log plot of $\hat{k}_{\text{mix}}$ vs the dimension d, for $n \in \{2d, 2d^2, 2d^3\}$ for $d \in \{2, 3, 4, 5, 6, 7\}$. The exponents from these plots are summarized in Table 6.2. Note that increasing the number of constraints n has more profound effect on the Dikin walk in almost all the cases.



(a) $n = 64$

(b) $n = 2048$

(c) $n = 64$

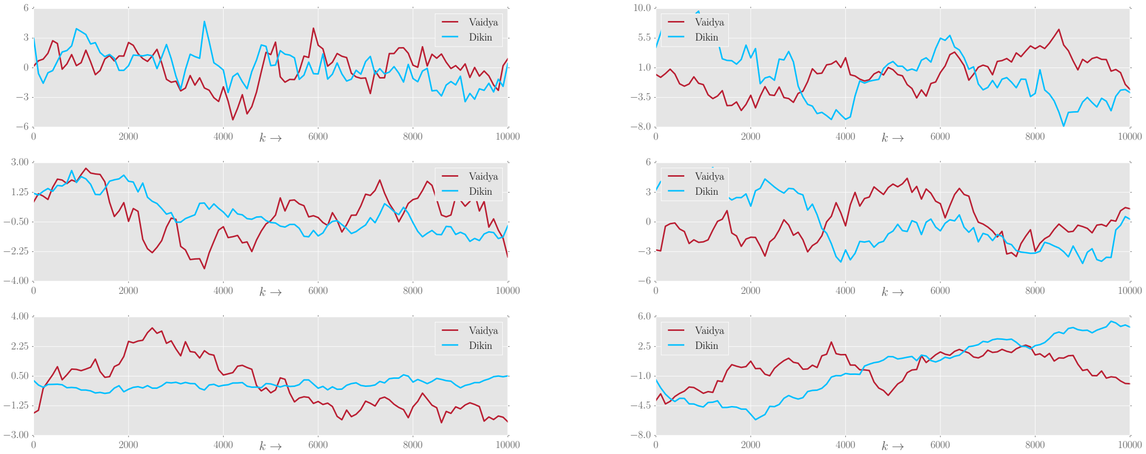
(d) $n = 2048$

Figure 6.4: Empirical distribution of the samples from 200 runs for the Dikin walk (blue/top rows) and the Vaidya walk (red/bottom rows) at different iterations k. The 2-dimensional polytopes considered are: **(a, b)** random polytopes with n-constraints, and **(c, d)** regular n-polygons inscribed in the unit circle. For both sets of cases, we observe that higher n slows down the walks, with visibly more effect on the Dikin walk compared to the Vaidya walk.

by Vempala [242] for a thorough discussion about the relation between the conductance and mixing time for Markov chains. Our proof techniques share a few features with the recent analyses of the Dikin walk by Kannan and Narayanan [139] and Sachdeva and Vishnoi [224]. However, new technical ideas are needed in order to handle the state-dependent weights σ_x (6.7b) and ζ_x (6.11) that underlie the proposal distributions for the Vaidya and John walks. Note that these techniques are not present in the analysis of the Dikin walk, which is based on constant weights.

6.6.2 Auxiliary results

Our proof proceeds by formally establishing the following property for the Vaidya walk: if two points are close, then their one-step transition distribution are also close. Consequently, we need to quantify the closeness between two points and the associated transition distributions. We measure the distance between two points in terms of the cross ratio that we define next. For a given pair of points $x, y \in \mathcal{K}$, let $e(x), e(y) \in \partial\mathcal{K}$ denote the intersection of the chord joining x and y with $\mathcal{K}$ such that $e(x), x, y, e(y)$ are in order (see Figure 6.6a). The cross-ratio



(a) $d = 10$

(b) $d = 50$

Figure 6.5: Sequential plots of a one-dimensional random projection of the samples on the hyperbox $\mathcal{K} = [-1, 1]^d$, defined by n constraints. Each plot corresponds to one long run of the Dikin and Vaidya walks, and the projection is taken in a direction chosen randomly from the sphere. **(a)** Plots for $d = 10$ and $n \in \{20, 640, 5120\}$. **(b)** Plots for $d = 50$ and $n \in \{100, 400, 1600\}$. Relative to the Dikin walk, the Vaidya walk has a more jagged plot for pairs (n, d) in which the ratio n/d is relatively large: for instance, see the plots corresponding to $(n, d) = (640, 10)$ and $(5120, 10)$. The same claim cannot be made for pairs (n, d) for which the ratio n/d is relatively small; e.g., the plot with $(n, d) = (20, 10)$. These observations are consistent with our results that the Vaidya walk mixes more quickly by a factor of order $\mathcal{O}(\sqrt{n/d})$ over the Dikin walk.

$\mathfrak{b}_{\mathcal{K}}(x, y)$ is given by

$$\mathfrak{b}_{\mathcal{K}}(x, y) := \frac{\left\|e(x) - e(y)\right\|_2 \left\|x - y\right\|_2}{\left\|e(x) - x\right\|_2 \left\|e(y) - y\right\|_2}. \tag{6.18}$$

The ratio $\mathfrak{b}_{\mathcal{K}}(x, y)$ is related to the Hilbert metric on $\mathcal{K}$, which is given by $\log\left(1 + \mathfrak{b}_{\mathcal{K}}(x, y)\right)$; see the paper [33] for more details.

Consider a lazy reversible random walk on a bounded convex set $\mathcal{K}$ with transition operator $\mathcal{T}$ defined via the mapping $\mu_0 \mapsto \mu_0/2 + \widetilde{\mathcal{T}}(\mu_0)/2$ and stationary with respect to the uniform distribution $\Pi^\star$ on $\mathcal{K}$. Recall that $\boldsymbol{\delta}_x$ denote the dirac-delta distribution with unit mass at x. The next result provides a bound on the mixing-time of the Markov chain, when the transition operator $\widetilde{\mathcal{T}}$ certain smoothness condition.

Proposition 6.1 (Lovász [163]). *Suppose that the transition operator* $\widetilde{\mathcal{T}}$ *admits* $\Pi^\star$ *as the*

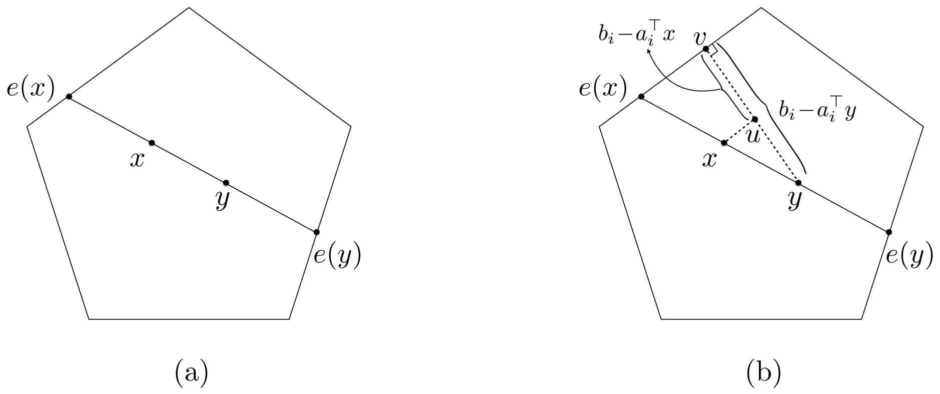


Figure 6.6: Polytope $\mathcal{K} = \{x \in \mathbb{R}^d | Ax \leq b\}$. (a) The points $e(x)$ and $e(y)$ denote the intersection points of the chord joining x and y with $\mathcal{K}$ such that $e(x), x, y, e(y)$ are in order. (b) A geometric illustration of the argument (6.23). It is straightforward to observe that $\|x - y\|_2 / \|e(x) - x\|_2 = \|u - y\|_2 / \|u - v\|_2 = \left|a_i^\top (y - x)\right| / \left(b_i - a_i^\top x\right)$.

stationary distribution, and there exist scalars $\rho, \Delta \in (0, 1)$ such that

$$d_{\mathrm{TV}}\big(\widetilde{\mathcal{T}}(\boldsymbol{\delta}_x), \widetilde{\mathcal{T}}(\boldsymbol{\delta}_y)\big) \leq 1 - \rho \qquad \textit{for all } x, y \in \operatorname{int}(\mathcal{K}) \textit{ with } \mathfrak{d}_{\mathcal{K}}(x, y) < \Delta. \tag{6.19a}$$

Then the Markov chain with lazy transition operator $\mathcal{T}$ and a β-warm start μ_0 satisfies

$$d_{\mathrm{TV}}\big(\mathcal{T}^\ell(\mu_0), \Pi^\star\big) \leq \sqrt{\beta} \exp\left(-\ell \, \frac{\Delta^2 \rho^2}{4096}\right) \quad \forall \;\; \ell = 1, 2, \ldots. \tag{6.19b}$$

Consequently, the mixing time of the chain satisfies

$$\tau_{\mathrm{TV}}(\delta; \mu_0) \leq \frac{4096}{\Delta^2 \rho^2} \log\left(\frac{\sqrt{\beta}}{\delta}\right). \tag{6.19c}$$

This result is implicit in the paper of Lovázs [163], though not explicitly stated. In order to keep the our work self-contained, we provide a proof of this result in Appendix D.4.

Our proof of Theorem 6.1 is based on applying Lovász's Lemma; the main challenge in our work is to establish that our random walks satisfy the condition (6.19a) with suitable choices of Δ and ρ. In order to proceed with the proof, we require a few additional notations. Recall that the slackness at x was defined as $s_x := (b_1 - a_1^\top x, \ldots, b_n - a_n^\top x)^\top$. For all $x \in \operatorname{int}(\mathcal{K})$, define the *Vaidya local norm of v at x* as

$$\|v\|_{V_x} := \left\|V_x^{1/2} v\right\|_2 = \sqrt{\sum_{i=1}^n (\sigma_{x,i} + \beta_{\mathrm{V}}) \frac{(a_i^\top v)^2}{s_{x,i}^2}}, \tag{6.20a}$$

and the *Vaidya slack sensitivity at x* as

$$\theta_{V_x} := \left(\left\|\frac{a_1}{s_{x,1}}\right\|_{V_x}^2, \ldots, \left\|\frac{a_n}{s_{x,n}}\right\|_{V_x}^2\right)^\top = \left(\frac{a_1^\top V_x^{-1} a_1}{s_{x,1}^2}, \ldots, \frac{a_n^\top V_x^{-1} a_n}{s_{x,n}^2}\right)^\top. \tag{6.20b}$$

Similarly, we define the *John local norm of v at x* and the *John slack sensitivity at x* as

$$\|v\|_{J_x} := \left\|J_x^{1/2} v\right\|_2 \quad \text{and} \quad \theta_{J_x} := \left(\left\|\frac{a_1}{s_{x,1}}\right\|_{J_x}^2, \ldots, \left\|\frac{a_n}{s_{x,n}}\right\|_{J_x}^2\right)^\top. \tag{6.20c}$$

The following lemma provides useful properties of the leverage scores σ_x from equation (6.7b), the weights ζ_x obtained from solving the program (6.11), and the slack sensitivities θ_{V_x} and θ_{J_x}.

Lemma 6.1. *For any $x \in \text{int}(\mathcal{K})$, the following properties hold:*

(h) $\sigma_{x,i} \in [0, 1]$ *for all* $i \in [n]$,

(i) $\sum_{i=1}^n \sigma_{x,i} = d$,

(j) $\theta_{V_x,i} \in \left[0, \sqrt{n/d}\right]$ *for all* $i \in [n]$,

(k) $\zeta_{x,i} \in [\beta_J, 1 + \beta_J]$ *for all* $i \in [n]$,

(l) $\sum_{i=1}^n \zeta_{x,i} = 3d/2$, *and*

(m) $\theta_{J_x,i} \in [0, 4]$ *for all* $i \in [n]$.

We prove this lemma in Appendix D.1.

Since the Vaidya walk is lazy with probability 1/2, for any distribution μ_0 we can write $\mathcal{T}_{\text{Vaidya}}(\mu_0) = \mu_0/2 + \widetilde{\mathcal{T}}_{\text{Vaidya}}(\mu_0)/2$ for a valid (non-lazy) transition operator $\widetilde{\mathcal{T}}_{\text{Vaidya}}$. Let $\mathcal{P}_x^{\text{V}}$ to denote the proposal distribution of the random walk VW(r) at state x.

Next, we state a lemma that shows that if two points $x, y \in \text{int}(\mathcal{K})$ are close in Vaidya local norm at x, then for a suitable choice of the parameter r, the proposal distributions $\mathcal{P}_x^{\text{V}}$ and $\mathcal{P}_y^{\text{V}}$ are close. In addition, we show that the proposals are accepted with high probability at any point $x \in \text{int}(\mathcal{K})$. We establish this result by showing that the one-step transition distribution $\widetilde{\mathcal{T}}_{\text{Vaidya}}(\boldsymbol{\delta}_x)$ is close to the proposal distribution $\mathcal{P}_x^{\text{V}}$.

Lemma 6.2. *There exists a continuous non-decreasing function $g_V : [0, 1/4] \to \mathbb{R}_+$ with $f(1/15) \geq 10^{-4}$ such that for any $\epsilon \in (0, 1/15]$, the random walk VW(r) with $r \in [0, g_V(\epsilon)]$ satisfies*

$$d_{\text{TV}}\left(\mathcal{P}_x^V, \mathcal{P}_y^V\right) \leq \epsilon \quad \forall\, x, y \in \text{int}(\mathcal{K}) \;\; s.t. \;\; \|x - y\|_{V_x} \leq \frac{\epsilon r}{2(nd)^{1/4}}, \quad \textit{and} \tag{6.21a}$$

$$d_{\text{TV}}\left(\widetilde{\mathcal{T}}_{Vaidya}(\boldsymbol{\delta}_x), \mathcal{P}_x^V\right) \leq 5\epsilon \;\; \forall\, x \in \text{int}(\mathcal{K}). \tag{6.21b}$$

See Appendix D.2 for the proof of this lemma.

With these lemmas in hand, we are now equipped to prove Theorem 6.1.

6.6.3 Proof of Theorem 6.1

To simplify notation, for the rest of this section, we adopt the shorthands $\mathcal{T}_x = \widetilde{\mathcal{T}}_{\text{Vaidya}}(\boldsymbol{\delta}_x)$, $\mathcal{P}_x = \mathcal{P}_x^{\text{V}}$ and $\left\|\cdot\right\|_{V_x} = \left\|\cdot\right\|_x$.

In order to apply Proposition 6.1 for the random walk VW(10^{-4}), we need to verify the condition (6.19a) for suitable choices of ρ and Δ. Doing so involves two main steps:

(A): First, we relate the cross-ratio $\mathfrak{b}_{\mathcal{K}}(x, y)$ to the local norm (6.20a) at x.

(B): Second, we use Lemma 6.2 to show that if $x, y \in \operatorname{int}(\mathcal{K})$ are close in local-norm, then the transition distributions $\mathcal{T}_x$ and $\mathcal{T}_y$ are close in TV-distance.

Step (A): We claim that for all $x, y \in \operatorname{int}(\mathcal{K})$, the cross-ratio can be lower bounded as

$$\mathfrak{b}_{\mathcal{K}}(x, y) \geq \frac{1}{\sqrt{2d}} \left\|x - y\right\|_x . \tag{6.22}$$

Note that we have

$$\begin{aligned} \mathfrak{b}_{\mathcal{K}}(x, y) = \frac{\left\|e(x) - e(y)\right\|_2 \left\|x - y\right\|_2}{\left\|e(x) - x\right\|_2 \left\|e(y) - y\right\|_2} &\overset{(i)}{\geq} \max\left\{\frac{\left\|x - y\right\|_2}{\left\|e(x) - x\right\|_2}, \frac{\left\|x - y\right\|_2}{\left\|e(y) - y\right\|_2}\right\} \\ &\overset{(ii)}{\geq} \max\left\{\frac{\left\|x - y\right\|_2}{\left\|e(x) - x\right\|_2}, \frac{\left\|x - y\right\|_2}{\left\|e(y) - x\right\|_2}\right\}, \end{aligned}$$

where step (i) follows from the inequality $\left\|e(x) - e(y)\right\|_2 \geq \max\left\{\left\|e(y) - y\right\|_2, \left\|e(x) - x\right\|_2\right\}$; and step (ii) follows from the inequality $\left\|e(x) - x\right\|_2 \leq \left\|e(y) - x\right\|_2$. Furthermore, from Figure 6.6b, we observe that

$$\max\left\{\frac{\left\|x - y\right\|_2}{\left\|e(x) - x\right\|_2}, \frac{\left\|x - y\right\|_2}{\left\|e(y) - x\right\|_2}\right\} = \max_{i \in [n]} \left|\frac{a_i^\top (x - y)}{s_{x,i}}\right| . \tag{6.23}$$

Note that maximum of a set of non-negative numbers is greater than the mean of the numbers. Combining this fact with properties (h) and (i) from Lemma 6.1, we find that

$$\mathfrak{b}_{\mathcal{K}}(x, y) \geq \sqrt{\frac{1}{\sum_{i=1}^n (\sigma_{x,i} + \beta_{\text{V}})} \sum_{i=1}^n (\sigma_{x,i} + \beta_{\text{V}}) \frac{(a_i^\top (x - y))^2}{s_{x,i}^2}} = \frac{\left\|x - y\right\|_x}{\sqrt{2d}},$$

thereby proving the claim (6.22).

Step (B): By the triangle inequality, we have

$$d_{\mathrm{TV}}\big(\mathcal{T}_x, \mathcal{T}_y\big) \leq d_{\mathrm{TV}}\big(\mathcal{T}_x, \mathcal{P}_x\big) + d_{\mathrm{TV}}\big(\mathcal{P}_x, \mathcal{P}_y\big) + d_{\mathrm{TV}}\big(\mathcal{P}_y, \mathcal{T}_y\big).$$

Thus, for any (r, ϵ) such that $\epsilon \in [0, 1/15]$ and $r \leq g_{\mathrm{V}}(\epsilon)$, Lemma 6.2 implies that

$$d_{\mathrm{TV}}\big(\mathcal{T}_x, \mathcal{T}_y\big) \leq 11\epsilon, \quad \forall x, y \in \mathrm{int}\,(\mathcal{K}) \text{ such that } \|x - y\|_x \leq \frac{r\epsilon}{2(nd)^{1/4}}.$$

Consequently, the walk $\mathrm{VW}(r)$ satisfies the assumptions of Lovász's Lemma with

$$\Delta := \frac{1}{\sqrt{2d}} \cdot \frac{r\epsilon}{2(nd)^{1/4}} \quad \text{and} \quad \rho := 1 - 11\epsilon.$$

Since $g_{\mathrm{V}}(1/15) \geq 10^{-4}$, we can set $\epsilon = 1/15$ and $r = 10^{-4}$, whence

$$\Delta^2 \rho^2 = \frac{(1 - 11\epsilon)^2 \epsilon^2 r^2}{8d\sqrt{nd}} = \frac{4^2}{15^2} \frac{1}{15^2} \frac{1}{10^{-8}} \cdot \frac{1}{d\sqrt{nd}} \geq 10^{-12} \frac{1}{d\sqrt{nd}}.$$

Observing that $\Delta < 1$ yields the claimed upper bound for the mixing time of Vaidya Walk.

6.6.4 Proof sketch for the John walk

Here we provide a brief outline the key steps in the analysis of John walk, leaving the details to Section 6.6.4. The proof of Theorem 6.2 is also decomposed in two steps analogous to the step (A) and (B) in section 6.6.3. From Lemma 6.1, we can see that while the sum of weights ζ_x is within a factor of $3/2$ to the sum of leverage scores σ_x, the John slack sensitivity $\theta_{J_x,i}$ (6.20c) is bounded by a constant as compared to the $\sqrt{n/d}$ bound for Vaidya slack sensitivity $\theta_{V_x,i}$ (6.20b). Along the lines of step (A) in Section 6.6.3, the former property directly establishes that $\mathfrak{b}_{\mathcal{K}}(x, y) \geq \|x - y\|_{J_x} / \sqrt{3d/2}$. To complete the proof, we establish an analog of Lemma 6.2 for the John walk which can then be used to derive a bound for its mixing time. We believe that our analysis while deriving the Lemma 6.2 analog for the John walk is loose and possibly a tighter analysis would allow us to prove Conjecture 6.1.

6.7 Conclusion and future directions

In this chapter, we focused on improving mixing rate of MCMC sampling algorithms for polytopes by building on the advancements in the field of interior point methods. We proposed and analyzed two different barrier based MCMC sampling algorithms for polytopes that outperforms the existing sampling algorithms like the ball walk, the hit-and-run and the Dikin walk for a large class of polytopes. We provably demonstrated the fast mixing of the Vaidya walk, $\mathcal{O}\left(n^{0.5}d^{1.5}\right)$ and the John walk, $\mathcal{O}\left(d^{2.5}\text{poly-log}(n/d)\right)$ from a warm start. Our numerical experiments, albeit simple, corroborated with our theoretical claims: the Vaidya walk mixes at least as fast the Dikin walk and significantly faster when the number of constraints is quite large compared to the dimension of the underlying space. For the John walk, the logarithmic factors were dominant in all our experiments and thereby we deemed the result of importance only for set-ups with polytopes in very high dimensions with number of constraints overwhelmingly larger than the dimensions. Besides, proving the mixing time guarantees for the improved John walk (Conjecture 6.1) is still an open question.

[187] analyzed a generalized version of the Dikin walk for arbitrary convex sets equipped with self-concordant barrier. From his results, we were able to derive mixing time bounds of $\mathcal{O}\left(nd^4\right)$ and $\mathcal{O}\left(d^5\text{poly-log}(n/d)\right)$ from a warm start for the Vaidya walk and the John walk respectively. Our proof takes advantage of the specific structure of the Vaidya and John walk, resulting a better mixing rate upper bound the the general analysis provided by [187].

While our paper has mainly focused on sampling algorithms on polytopes, the idea of using logarithmic barrier to guide sampling can be extended to more general convex sets. The self-concordance property of the logarithmic barrier for polytopes is extended by [7] to more general convex sets defined by semidefinite constraints, namely, linear matrix inequality (LMI) constraints. Moreover, [187] showed that for a convex set in $\mathbb{R}^d$ defined by n LMI constraints and equipped with the log-determinant barrier—the semidefinite analog of the logarithmic barrier for polytopes—the mixing time of the Dikin walk from a warm start is $\mathcal{O}\left(nd^2\right)$. It is possible that an appropriate Vaidya walk on such sets would have a speed-up over the Dikin walk. Another work [188] used the Dikin walk to generate samples from time varying log-concave distributions with appropriate scaling of the radius for different class of distributions. We believe that suitable adaptations of the Vaidya and John walks for such cases would provide significant gains.

Part II

Statistical-Computational Challenges With Mixture Models

Chapter 7

Two-Faced Slowdown of EM with Singular Mixture Models

The focus of this part of the thesis is the intersection of statistical and computational issues associated with fitting the parameters of *overspecified* mixture models, namely when the number of components in the fitted model are more than that in the true model. The algorithm of choice for fitting finite mixture models is the Expectation-Maximization (EM) algorithm [68, 251, 207], and is arguably the most popular algorithm for computing (approximate) MLEs in the mixture models. EM represents a general framework that encompasses various types of divide-and-conquer computational strategies. It is also an instance of a minorization-maximization algorithm, in which at each step, a suitably chosen lower bound of the log-likelihood is maximized.

There is now a lengthy line of work on the behavior of EM when applied to regular models. The classical papers [251, 236, 55] establish the asymptotic convergence of EM to a local maximum of the log-likelihood function for a general class of incomplete data models. Other papers [134, 254, 172] characterized the rate of convergence of EM for regular Gaussian mixtures. More recent years have witnessed a flurry of work on the behavior of EM for various kinds of regular mixture models [10, 248, 256, 253, 64, 255, 105, 35]; as a consequence, our understanding of EM in such cases is now relatively mature. More precisely, it is known that for Gaussian mixtures, EM converges in $\mathcal{O}(\log(n/d))$-steps to parameter estimates that lie within Euclidean distance $\mathcal{O}((d/n)^{1/2})$ of the true location parameters, assuming minimal separation between the mixture components. With this context in mind, the primary goal of this chapter and the next is to gain a fundamental understanding of the behavior of EM when used to fit over-specified mixture models. We start with an introduction to the common issues associated with mixture models, and EM in Section 7.1, and discuss the related prior work in Section 7.2. With this context, we summarize the contributions of this chapter in Section 7.2.1, and the organization of the remainder of this chapter in Section 7.2.2.

7.1 Introduction and prior work

While density estimation in finite mixture models is relatively well understood [239, 97], characterizing the behavior of maximum likelihood for parameter estimation has remained challenging. The main difficulty for analyzing the MLE in such settings arises from label switching between the mixtures [208, 230], and lack of strong concavity in the likelihood. Such issues do not interfere with density estimation, since the standard divergence measures like the Kullback-Leibler and Hellinger distances remain invariant under permutations of labels, and strong concavity is not required.

We now provide a brief flavor of the known statistical challenges that arise due to over-specification, besides the already existing computational challenges with mixture models. We then discuss prior work on Expectation-Maximization, and then summarize our contributions.

7.1.1 Statistical issues with singularity and over-specification

An important contribution to the understanding of parameter estimation in finite mixture models was made by Chen [42]. He considered a class of *over-specified* finite mixture models; here the term "over-specified" means that the model to be fit has more mixture components than the distribution generating the data. In an interesting contrast to the usual $n^{-\frac{1}{2}}$ convergence rate for the MLE based on n samples, Chen showed that for estimating scalar location parameters in a certain class of over-specified finite mixture models, the corresponding rate slows down to $n^{-\frac{1}{4}}$ due to the *singularity* of the Fisher information matrix at the true parameter, meaning that the Fisher information matrix is degenerate (or not full rank). This theoretical result has practical significance, since methods that over-specify the number of mixtures are often more feasible than methods that first attempt to estimate the number of components, and then estimate the parameters using the estimated number of components [222]. In subsequent work, Nguyen [194] and Heinrich et al. [109] have characterized the (minimax) convergence rates of parameter estimation rates for mixture models in both exactly-fitted or over-specified settings in terms of the Wasserstein distance.

In the context of singular mixture models, an important distinction is between those that are *strongly* versus *weakly* identifiable. Chen's work [42] studies the class of strongly identifiable models in which, while the Fisher information matrix may be degenerate at a point, and it is not degenerate over a larger set. When the degeneracy occurs over a larger set, we call the model weakly identifiable. In this chapter, we discuss strongly identifiable singular models that are obtained, e.g., when fitting over-specified Gaussian mixtures models with unknown mean and known covariance. The next chapter deals with the weakly identifiable models, which arise, e.g., when fitting over-specified Gaussian mixtures models with both unknown mean and covariance.

7.1.2 Computational concerns with mixture models

While the papers discussed above address the statistical behavior of a global maximum of the log-likelihood, they do not consider the associated computational issues of obtaining such a maximum. In general settings, non-convexity of the log-likelihood makes it impossible to guarantee that the iterative algorithms used in practice converge to the global optimum, or equivalently the MLE. Perhaps the most widely used algorithm for computing the MLE is the expectation-maximization (EM) algorithm [68]. Early work on the EM algorithm [251] showed that its iterates converge asymptotically 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 linear to super-linear (quasi-Newton like) convergence if the overlap between the mixture components tends to zero [254, 172]. It has also been noted in several papers [207, 172] that the convergence of EM can be prohibitively slow when the mixtures are not well separated.

7.2 Prior work on Expectation-Maximization

Balakrishnan et al. [10] laid out a general theoretical framework for analysis of the EM algorithm, and in particular how to prove non-asymptotic bounds on the Euclidean distance between sample-based EM iterates and the true parameter. When applied to the special case of two-component Gaussian location mixtures, assumed to be well-specified and suitably separated, their theory guarantees that (1) population EM updates enjoy a geometric rate of convergence to the true parameter when initialized in a sufficiently small neighborhood around the truth, and (2) sample-based EM updates converge to an estimate at Euclidean distance of order $(d/n)^{\frac{1}{2}}$, based on n i.i.d. draws from a finite mixture model in $\mathbb{R}^d$. Further work in this vein has characterized the behavior of EM in a variety of settings for two Gaussian mixtures, including convergence analysis with additional sparsity constraints [248, 256, 105], global convergence of population EM [253], guarantees of geometric convergence under less restrictive conditions on the two mixture components [142, 64], analysis of EM with unknown mixture weights, means and covariances for two mixtures [35], and the analysis of EM to more than two Gaussian components [255, 105]. Other related work has provided optimization-theoretic guarantees for EM by viewing it in a generalized surrogate function framework [146], and analyzed the statistical properties of confidence intervals based on an EM estimator [49].

An assumption common to all of this previous work is that there is no misspecification in the fitting of the Gaussian mixtures; in particular, it is assumed that the data is generated from a mixture model with the same number of components as the fitted model. A portion of our recent work [80] has shown that EM retains its fast convergence behavior—albeit to a biased estimate—in *under-specified* settings where the number of components in the fitted model are less than that in the true model. However, as noted above, in practice, it is most common to use over-specified mixture models. For these reasons, it is desirable to understand how the EM algorithm behaves in the over-specified settings.

7.2.1 Our contributions

The goal of this chapter is to shed some light on the non-asymptotic performance of the EM algorithm for over-specified mixtures. We provide a comprehensive study of over-specified mixture models when fit to a particularly simple (non-mixture) data-generating mechanism; a multivariate normal distribution $\mathcal{N}(0, \sigma^2 I_d)$ in d dimensions with known scale parameter $\sigma > 0$. Such a model, while being singular, is strongly identifiable [114]. This setting, despite its simplicity, suffices to reveal some rather interesting properties of EM in the over-specified context. In particular, 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 known unequal weights for the two components. For this case, we establish that the population EM updates converge at a geometric rate to the true parameter; as an immediate consequence, the sample-based EM algorithm converges in $\mathcal{O}\left(\log(n/d)\right)$ steps to a ball of radius $(d/n)^{\frac{1}{2}}$. 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-specified mixtures.

- **Two-mixture balanced fit:** In the balanced version of the problem in which the mixture weights are equal to $\frac{1}{2}$ 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 [10, 64, 255], where the population updates converge at a geometric rate. We also show that our rates for population EM are tight. By combining the slow convergence of population EM with a novel localization argument, one involving the empirical process restricted to an annulus, we show that the sample-based EM iterates converge to a ball of radius $(d/n)^{\frac{1}{4}}$ around the true parameter after $\mathcal{O}((n/d)^{\frac{1}{2}})$ steps. The $n^{-\frac{1}{4}}$ component of the Euclidean error matches known guarantees for the global maximum of the MLE [42]. 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 ball-based localization methods are known to be essential in deriving sharp statistical rates for M-estimators (e.g., [239, 14, 143]); to the best of our knowledge, the use of an annulus-based localization argument in analyzing an algorithm is novel.

Moreover, we show via extensive numerical experiments that the fast convergence of EM for the unbalanced fit is a special case; and that the slow behavior of EM proven for the balanced fit (in particular the rate of order $n^{-\frac{1}{4}}$) arises in several general (including more than two components) over-specified Gaussian mixtures with known variance, known or unknown weights, and unknown location parameters.

7.2.2 Organization

The remainder of the chapter is organized as follows. In Section 7.3 we provide illustrative simulations of EM in different settings in order to motivate the settings analyzed later. We then provide a thorough analysis of the convergence rates of EM when over-fitting Gaussian data with two components in Section 7.4 and the key ideas of the novel proof techniques in Section 7.5. We provide a thorough discussion of our results in Section 7.6, exploring their general applicability, and presenting further simulations that substantiate the value of our theoretical framework. Detailed proofs of our results and discussion of certain additional technical aspects of our results are provided in the appendix.

7.2.2.1 Notation

For any two sequences a_n and b_n, the notation $a_n \precsim b_n$ or $a_n = \mathcal{O}\left(b_n\right)$ means that $a_n \leq C b_n$ for some universal constant C. Similarly, the notation $a_n \asymp b_n$ or $a_n = \Theta(b_n)$ denotes that both the conditions, $a_n \precsim b_n$ and $b_n \precsim a_n$, hold. Throughout this chapter, $\mathrm{\pi}$ denotes a variable and π denotes the mathematical constant "pi".

7.2.2.2 Experimental settings

We summarize a few common aspects of the numerical experiments presented in this chapter. Population-level computations were done using numerical integration on a sufficiently fine grid. With finite samples, the stopping criteria for the convergence of EM were: (1) the change in the iterates was small enough ($< 0.01/n$), or (2) the number of iterations was too large (greater than $100,000$). Experiments were averaged over several repetitions (ranging from 25 to 400). In majority of the runs, for each case, criteria (1) led to convergence. In our plots for sample EM, we report $\widehat{m}_e + 2\widehat{s}_e$ on the y-axis, where $\widehat{m}_e, \widehat{s}_e$ respectively denote the mean and standard deviation across the experiments for the metric under consideration, e.g., the parameter estimation error. Furthermore, whenever a slope is provided, it is the slope for the least-squares 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 7.1(b), we plot $|\widehat{\theta}_n - \theta^*|$ on the y-axis value versus the sample size n on the x-axis, averaged over 400 experiments, accounting for the deviation across these experiments. Furthermore, the green dotted line with legend $\mathrm{\pi} = 0.3$ and the corresponding slope -0.48 denote the least-squares fit and the respective slope for $\log|\widehat{\theta}_n - \theta^*|$ (green solid dots) with $\log n$ for the experiments corresponding to the setting $\pi = 0.3$.

7.3 Motivating simulations and problem set-up

In this section, we explore a wide range of behavior demonstrated by EM for certain settings of over-specified location Gaussian mixtures. We begin with several simulations that illustrate fast and slow convergence of EM for various settings, and serve as a motivation for the theoretical results derived later in the chapter. We provide basic background on EM in Section 7.3.3, and describe the problems to be tackled.

7.3.1 Problem set-up

Let $\phi(\cdot\,; \mu, \Sigma)$ denote the density of a Gaussian random vector with mean μ and covariance Σ. Consider the two component Gaussian mixture model with density

$$f(x; \theta^*, \sigma, \pi) := \pi\phi(x; \theta^*, \sigma^2 I_d) + (1 - \pi)\phi(x; -\theta^*, \sigma^2 I_d). \tag{7.1}$$

Given n samples from the distribution (7.1), suppose that we use the EM algorithm to fit a two-component location Gaussian mixture with fixed weights and variance[1] and special structure on the location parameters—more precisely, we fit the model with density

$$f(x; \theta, \sigma, \pi) := \pi\phi(x; \theta, \sigma^2 I_d) + (1 - \pi)\phi(x; -\theta, \sigma^2 I_d) \tag{7.2}$$

using the EM algorithm, and take the solution[2] as an estimate of θ^*. An important aspect of the problem at hand 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 model (7.1), the signal strength is given by the ratio $\|\theta^*\|_2/\sigma$. When this ratio is large, we refer to it as the *strong signal* case; otherwise, it corresponds to the *weak signal* case. Of particular interest to us is the behavior of EM in the limit of weak signal when there is no separation; i.e., $\|\theta^*\|_2 = 0$. For such cases, we call the fit (7.2) an *unbalanced* fit when $\pi \neq \frac{1}{2}$ and a *balanced* fit when $\pi = \frac{1}{2}$. Note that the setting of $\theta^* = 0$ corresponds to the simplest case of over-specified fit, since the true model has just one component (standard normal distribution irrespective of the parameter π) but the fitted model has two (one extra) component (unless the EM estimate is also 0). We now present the empirical behavior of EM for these models and defer the derivation of EM updates to Section 7.3.3.

[1] Refer to Section 7.6 for a discussion for the case of unknown weights and variances.

[2] Strictly speaking, different initialization of EM may converge to different estimates. For the settings analyzed theoretically in this work, the EM always converges towards the same estimate in the limit of infinite steps, and we use a stopping criterion to determine the final estimate. See the discussion on experimental settings in Section **??** for more details.

7.3.2 Numerical Experiments: Fast to slow convergence of EM

We begin with a numerical study of the effect of separation among the mixtures on the statistical behavior of the estimates returned by EM. Our main observation is that weak or no separation leads to relatively low accuracy estimates. Additional simulations for more general mixtures, including more than two components, are provided in Section 7.6.3. Next, via numerical integration on a grid with sufficiently small discretization width, we simulate the behavior of the population EM algorithm width—an idealized version of EM in the limit of infinite samples—in order to understand the effect of signal strength on EM's algorithmic rate of convergence, i.e., the number of steps needed for population EM to converge to a desired accuracy. We observe a slow down of EM on the algorithmic front when the signal strength approaches zero.

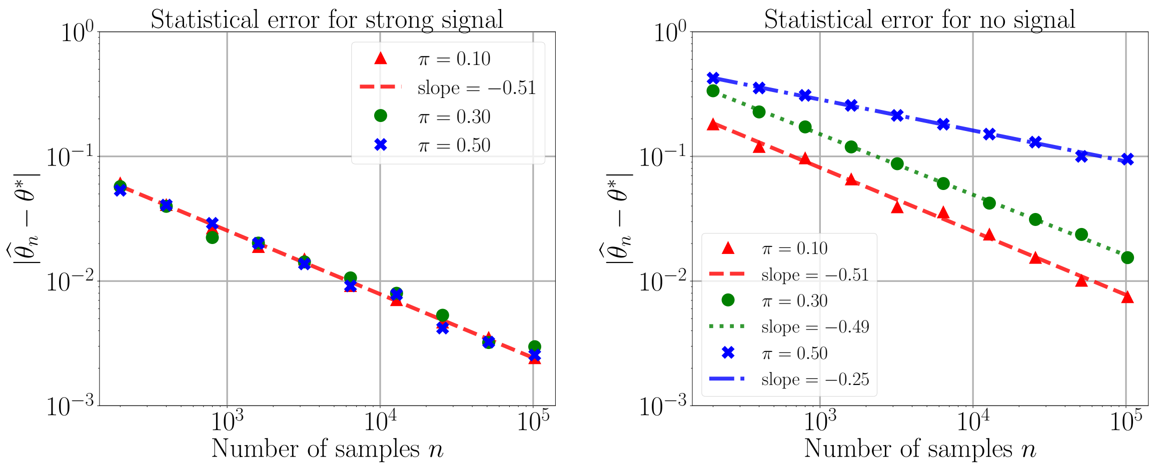
7.3.2.1 Effect of signal strength on sample EM

In Figure 7.1, we show simulation results for data generated from the model (7.1) 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 variance as specified by equation (7.2). The two panels show the estimation error of the EM solution as a function of n 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) to obtain 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 the data from a standard normal distribution.

In the strong signal case, it is well known [10, 64] that EM solutions have an estimation error (measured by the Euclidean distance between the EM estimate and the true parameter θ^*) that achieves the classical (parametric) rate $n^{-\frac{1}{2}}$; the empirical results in Figure 7.1(a) are simply a confirmation of these theoretical predictions. More interesting is the case of no signal (which is the limiting case with weak signal), where the simulation results shown in panel (b) of Figure 7.1 reveal a different story. In this case, whereas the EM solution (with random standard normal initialization) has an error that decays as $n^{-\frac{1}{2}}$ when $\pi \neq 1/2$, its error decays at the considerably slower rate $n^{-\frac{1}{4}}$ when $\pi = 1/2$. We return to these cases in further detail in Section 7.4.

7.3.2.2 Interesting behavior of population EM

The intriguing behavior of the sample EM algorithm in the "no signal" case motivated us to examine the behavior of population EM for this case. To be clear, while sample EM is the practical algorithm that can actually be applied, it can be insightful for theoretical purposes to first analyze the convergence of the population EM updates, and then leverage these findings to understand the behavior of sample EM [10]. Our analysis follows a similar road-map. Interestingly, for the case with $\theta^* = 0$, the population EM algorithm behaves



(a) $\theta^* = 5$

(b) $\theta^* = 0$

Figure 7.1: 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 $n^{-\frac{1}{2}}$, 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: $n^{-\frac{1}{2}}$ when $\pi \neq 0.5$, and, $n^{-\frac{1}{4}}$ when $\pi = 0.5$, again as revealed by least-squares fits of the log error using the log sample size $\log n$.

significantly differently for the unbalanced fit ($\pi \neq \frac{1}{2}$) as compared to the balanced fit ($\pi = \frac{1}{2}$) (equation (7.2)). In Figure 7.2, we plot the distance of the population 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 of the algorithm shows up as a negatively sloped line (disregarding transient effects in the first few iterations).

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 $\theta \mapsto \frac{\theta}{1+\theta^2}$.

The theory to follow provides a precise characterization of the behavior seen in Figures 7.1(b) and 7.2. Furthermore, in Section 7.6, we provide further support for relevannce of our theoretical results in explaining the behavior of EM for other classes of over-specified models, including Gaussian mixture models with unknown weights as well as mixtures of

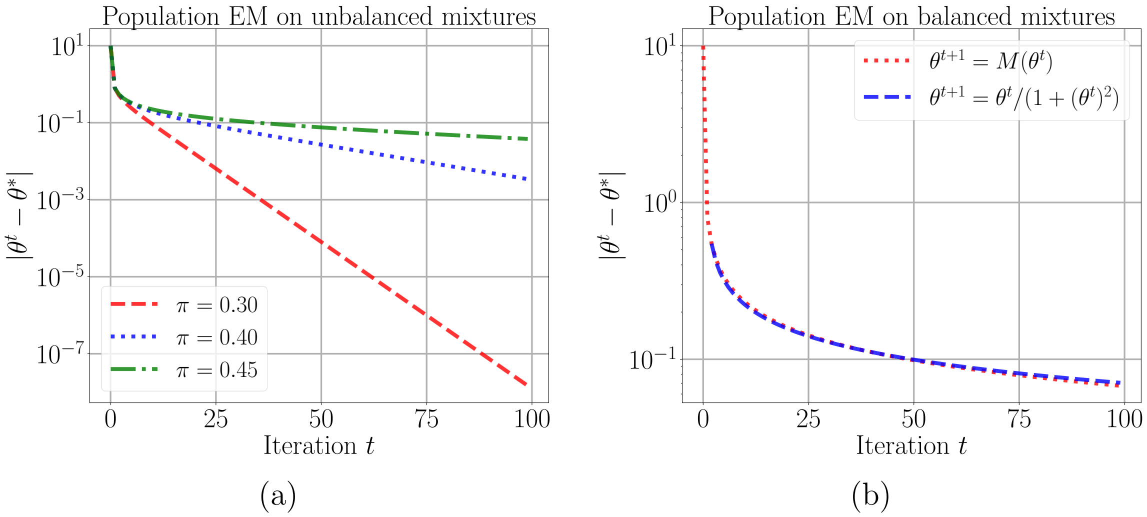


Figure 7.2: Behavior of the (numerically computed) population EM updates (7.8) when the underlying data distribution is $\mathcal{N}(0, 1)$. (a) Unbalanced mixture fits (7.2) with weights $(\pi, 1-\pi)$: We observe geometric convergence towards $\theta^* = 0$ for all $\pi \neq 0.5$ although the rate of convergence gets slower as $\pi \to 0.5$. (b) Balanced mixture fits (7.2) 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.

linear regressions.

7.3.3 EM updates for the model fit (7.2)

In this section, we provide a quick introduction to the EM updates. Readers familiar with the literature can skip directly to the main results in Section 7.4. Recall that the two-component model fit is based on the density

$$\pi\phi(x; \theta, \sigma^2 I_d) + (1 - \pi)\phi(x; -\theta, \sigma^2 I_d). \tag{7.3}$$

From now on we assume that the data is drawn from the zero-mean Gaussian distribution $\mathcal{N}(0, \sigma^2 I_d)$. Note that the model fit described above contains the true model with $\theta^* = 0$ and it is referred to as an over-specified fit since for any non-zero θ, the fitted model has two components.

The maximum likelihood estimate is obtained by solving the following optimization problem

$$\widehat{\theta}_n^{\mathrm{MLE}} \in \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \left\{ \log(\pi \phi(x_i; \theta, \sigma^2 I_d) + (1-\pi)\phi(x_i; -\theta, \sigma^2 I_d)) \right\}. \tag{7.4}$$

In general, there is no closed-form expression for $\widehat{\theta}_n^{\mathrm{MLE}}$. The EM algorithm circumvents this problem via a minorization-maximization scheme. Indeed, population EM is a surrogate method to compute the maximizer of the population log-likelihood

$$\mathcal{L}(\theta) := \mathbb{E}_X \left[\log(\pi \phi(X; \theta, \sigma^2 I_d) + (1-\pi)\phi(X; -\theta, \sigma^2 I_d) \right], \tag{7.5}$$

where the expectation is taken over the true distribution. On the other hand, sample EM attempts to estimate $\widehat{\theta}_n^{\mathrm{MLE}}$. We now describe the expressions for both the sample and population EM updates for the model-fit (7.3).

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-component location Gaussian 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 (7.3). For EM, we first compute the conditional probability of $Z = 1$ given $X = x$:

$$w_\theta(x) = w_\theta^\pi(x) := \frac{\pi \exp\left(-\frac{\|\theta - x\|_2^2}{2\sigma^2}\right)}{\pi \exp\left(-\frac{\|\theta - x\|_2^2}{2\sigma^2}\right) + (1-\pi)\exp\left(-\frac{\|\theta + x\|_2^2}{2\sigma^2}\right)}. \tag{7.6}$$

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

$$Q(\theta'; \theta) = -\frac{1}{2}\mathbb{E}\left[w_\theta(X) \|X - \theta'\|_2^2 + (1 - w_\theta(X)) \|X + \theta'\|_2^2 \right], \tag{7.7}$$

where the expectation is taken over the true distribution (here $\mathcal{N}(0, \sigma^2 I_d)$. In the M-step, we maximize the function $\theta' \mapsto Q(\theta'; \theta)$. Doing so defines a mapping $M : \mathbb{R}^d \to \mathbb{R}^d$, known as the *population EM operator*, given by

$$M(\theta) = \arg\max_{\theta' \in \mathbb{R}^d} Q(\theta', \theta) = \mathbb{E}\Big[(2w_\theta(X) - 1)X\Big]. \tag{7.8}$$

In this definition, the second equality follows by computing the gradient $\nabla_{\theta'} Q$, and setting it to zero. In summary, for the two-component 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 (7.8).

We obtain the *sample EM update* by simply replacing the expectation $\mathbb{E}$ in equations (7.7) and (7.8) 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

$$M_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} (2w_\theta(X_i) - 1) X_i. \tag{7.9}$$

Overall, the sample EM algorithm generates the sequence of iterates given by $\theta^{t+1} = M_n(\theta^t)$.

In the sequel, we study the convergence of EM both for the population EM algorithm in which the updates are given by $\theta^{t+1} = M(\theta^t)$, and the sample-based EM sequence given by $\theta^{t+1} = M_n(\theta^t)$. With this notation in place, we now turn to the main results of this chapter.

7.4 Main results

In this section, we state our main results for the convergence rates of the EM updates under the unbalanced and balanced mixture fit. We start with the easier case of unbalanced mixture fit in Section 7.4.1 followed by the more delicate (and interesting) case of the balanced fit in Section 7.4.2.

7.4.1 Behavior of EM for unbalanced mixtures

We begin with a characterization of both the population and sample-based EM updates in the setting of unbalanced mixtures. In particular, we assume that the fitted two-components mixture model (7.3) has known weights π and $1-\pi$, where $\pi \in (0, 1/2)$. The following result characterizes the behavior of the EM updates for this set-up.

Theorem 7.1. *Suppose that we fit an unbalanced instance (i.e., $\pi \neq \frac{1}{2}$) of the mixture model (7.3) to $\mathcal{N}(0, \sigma^2 I_d)$ data. Then:*

(a) The population EM operator (7.8) is globally strictly contractive, meaning that

$$\|M(\theta)\|_2 \leq \left(1 - \rho^2/2\right) \|\theta\|_2 \qquad \text{for all } \theta \in \mathbb{R}^d, \tag{7.10a}$$

where $\rho := |1 - 2\pi| \in (0, 1)$.

(b) There are universal constants c, c' such that given any $\delta \in (0,1)$ and a sample size $n \geq c\frac{\sigma^2}{\rho^4}(d + \log(1/\delta))$, the sample EM sequence $\theta^{t+1} = M_n(\theta^t)$ generated by the update (7.9) satisfies the upper bound

$$\|\theta^t\|_2 \leq \|\theta^0\|_2 \left(1 - \frac{\rho^2}{2}\right)^t + \frac{c'(\|\theta^0\|_2 \sigma^2 + \rho\sigma)}{\rho^2} \sqrt{\frac{d + \log(1/\delta)}{n}}\,, \tag{7.10b}$$

with probability at least $1 - \delta$.

See Appendix F.1 for the proof of this theorem.

Fast convergence of population EM: The bulk of the effort in proving Theorem 7.1 lies in establishing the guarantee (7.10a) for the population EM iterates. Such a contraction bound immediately yields the exponential fast convergence of the population EM updates $\theta^{t+1} = M(\theta^t)$ to $\theta^* = 0$:

$$\left\|\theta^T\right\|_2 \leq \epsilon \quad \text{for} \quad T \geq \frac{1}{\log \frac{1}{(1-\rho^2/2)}} \cdot \log\left(\frac{\|\theta^0\|_2}{\epsilon}\right). \tag{7.11}$$

Since the mixture weights $(\pi, 1-\pi)$ are bounded away from $1/2$, we have that $\rho = |1-2\pi|$ is bounded away from zero, and thus population EM iterates converge in $\mathcal{O}\left(\log(1/\epsilon)\right)$ steps to an ϵ-ball around $\theta^* = 0$. This result is equivalent to showing that in the unbalanced instance $(\pi \neq 1/2)$, the log-likelihood is strongly concave around the true parameter.

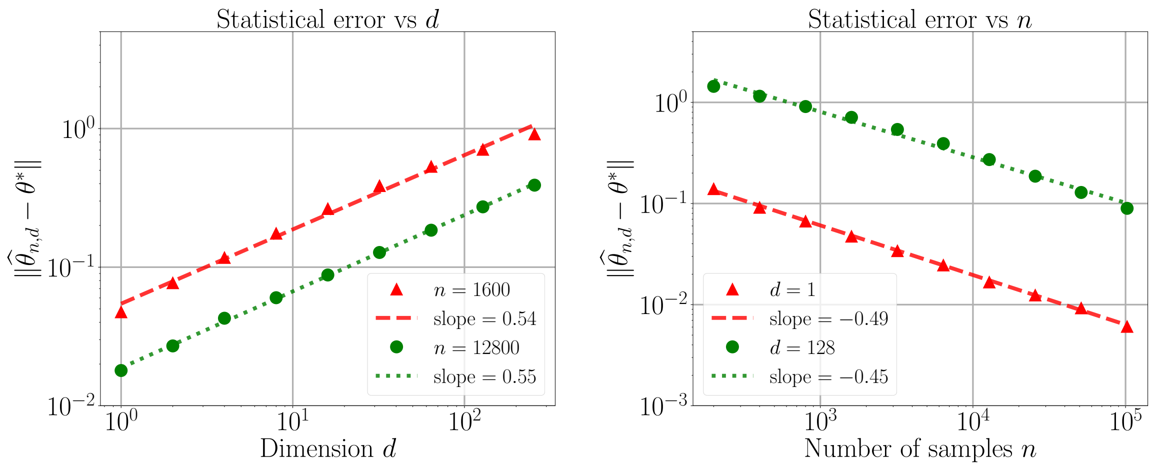
Statistical rate of sample EM: Once the bounds (7.10a) and (7.11)) have been established, the proof of the statistical rate (7.10b) for sample EM utilizes the scheme laid out by Balakrishnan et al. [10]. In particular, we prove a non-asymptotic uniform law of large numbers (Lemma 7.1 stated in Section 7.5.1) that allows for the translation from population to sample EM iterates. Roughly speaking, Lemma 7.1 guarantees that for any radius $r > 0$, tolerance $\delta \in (0, 1)$, and sufficiently large n, we have

$$\mathbb{P}\left[\sup_{\|\theta\|_2 \leq r} \|M_n(\theta) - M(\theta)\|_2 \leq c'\sigma(\sigma r + \rho)\sqrt{\frac{d + \log(1/\delta)}{n}}\right] \geq 1 - \delta. \tag{7.12}$$

This bound, when combined with the contractive behavior (7.10a) or equivalently the exponentially fast convergence (7.11) of the population EM iterates allows us to establish the stated bound (7.10b). (See, e.g., Theorem 2 in the paper [10].)

Putting the pieces together, we conclude that the sample EM updates converge to an estimate of θ^*—that has Euclidean error of the order $(d/n)^{\frac{1}{2}}$—after a relatively small number of steps that are of the order $\log(n/d)$. Note that this theoretical prediction is verified by the simulation study in Figure 7.1(b) for the univariate setting $(d = 1)$ of the unbalanced mixture-fit. In Figure 7.3, we present the scaling of the radius of the final EM iterate[3] with respect to the sample size n and the dimension d, averaged over 400 runs of sample EM for various settings of (n, d). Linear fits on the log-log scale in these simulations suggest a rate close to $(d/n)^{\frac{1}{2}}$ as claimed in Theorem 7.1.

[3] Refer to the discussion before Section 7.3 for details on the stopping rule for EM.



(a) (b)

Figure 7.3: Scaling of the Euclidean error $\|\widehat{\theta}_{n,d} - \theta^*\|_2$ for EM estimates $\widehat{\theta}_{n,d}$ computed using the unbalanced ($\pi \neq \frac{1}{2}$) mixture-fit (7.3). 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 $(d/n)^{\frac{1}{2}}$ on the EM solution is sharp in terms of n and d.

Remark: We make two comments in passing. First, the value of $\|\theta^0\|_2$ in the convergence rate of sample EM updates in Theorem 7.1 can be assumed to be of constant order; this assumption stems from the fact the population EM operator maps any θ^0 to a vector with norm smaller than $\sqrt{2/\pi}$ (cf. Lemma F.4 in Appendix F.7.1). Second, when the weight parameter π is assumed to be unknown in the model fit (7.3), the EM algorithm exhibits fast convergence when π is initialized sufficiently away from $\frac{1}{2}$; see Section 7.6.1 for more details.

From unbalanced to balanced fit: The bound (7.11) shows that the extent of unbalancedness in the mixture weights plays a crucial role in the geometric rate of convergence for the population EM. When the mixtures become more balanced, that is, weight π approaches $1/2$ or equivalently ρ approaches zero, the number of steps T required to achieve ϵ-accuracy scales as $\mathcal{O}\left(\log(\|\theta^0\|_2/\epsilon)/\rho^2\right)$ and in the limit $\rho \to 0$, this bound degenerates to ∞ for any finite ϵ. Indeed, the bound (7.10a) from Theorem 7.1 simply states that the population EM operator is non-expansive for balanced mixtures ($\rho = 0$), and does not provide any particular rate of convergence for this case. It turns out that the EM algorithm is worse in the balanced case, both in terms of the optimization speed and in terms of the statistical rate. This slower statistical rate is in accord with existing results for the MLE in over-specified mixture models [42]; the novel contribution here is the rigorous analysis of the analogous behavior for the EM algorithm.

7.4.2 Behavior of EM for balanced mixtures

We start with a sharp characterization of the algorithmic rate of convergence of the population EM update for the balanced fit in Section 7.4.2.1, followed by a sharp analysis for the statistical rate for the sample EM updates in Section 7.4.2.2.

7.4.2.1 Slow convergence of population EM

We show that the population EM operator for the balanced fit 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 $X \sim \mathcal{N}(0, 1)$ denotes a standard normal variate. (Note that $p < 1$.)

Theorem 7.2. *Suppose that we fit a balanced instance ($\pi = \frac{1}{2}$) of the mixture model (7.3) to $\mathcal{N}(0, \sigma^2 I_d)$ data. Then the population EM operator (7.8) $\theta \mapsto M(\theta)$ has the following properties:*

(a) For all non-zero θ, we have

$$\frac{\|M(\theta)\|_2}{\|\theta\|_2} \leq \gamma_{up}(\theta) \; := 1 - p + \frac{p}{1 + \frac{\|\theta\|_2^2}{2\sigma^2}} \; < \; 1. \tag{7.13a}$$

(b) For all non-zero θ such that $\|\theta\|_2^2 \leq \frac{5\sigma^2}{8}$, we have

$$\frac{\|M(\theta)\|_2}{\|\theta\|_2} \geq \gamma_{low}(\theta) \; := \; \frac{1}{1 + \frac{2\|\theta\|_2^2}{\sigma^2}}. \tag{7.13b}$$

See Appendix F.2 for the proof of Theorem 7.2.

The salient feature of Theorem 7.2 is that the contraction coefficient $\gamma_{\text{up}}(\theta)$ is not globally bounded away from 1 and in fact satisfies $\lim_{\theta \to 0} \gamma_{\text{up}}(\theta) = 1$. In conjunction with the lower bound (7.13b), we see that

$$\frac{\|M(\theta)\|_2}{\|\theta\|_2} \asymp \left(1 - \frac{\|\theta\|_2^2}{\sigma^2}\right) \qquad \text{for small } \|\theta\|_2. \tag{7.14}$$

This precise contraction behavior of the population EM operator is in accord with that of the simulation study in Figure 7.2(b).

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 of the order σ, the iterates exhibit geometric convergence. Concretely, we are guaranteed to have $\left\|\theta^{T_0}\right\|_2 \leq \sqrt{2}\sigma$ after running the algorithm for $T_0 := \frac{\log(\left\|\theta^0\right\|_2^2/(2\sigma^2))}{\log(2/(2-p))}$ 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

$$\left\|\theta^{T_0+t}\right\|_2 \leq \epsilon \quad \text{for} \quad t \geq \frac{c\sigma^2}{\epsilon^2} \log(\sigma/\epsilon). \tag{7.15}$$

Note that the conclusion (7.15) shows that for small enough ϵ, the population EM takes $\Theta(\log(1/\epsilon)/\epsilon^2)$ steps to find ϵ-accurate estimate of $\theta^* = 0$. This rate is extremely slow compared to the geometric rate $\mathcal{O}(\log(1/\epsilon))$ derived for the unbalanced mixtures in Theorem 7.1. Hence, the slow rate establishes a qualitative difference in the behavior of the EM algorithm between the balanced and unbalanced setting.

Moreover, the sub-geometric rate of EM in the balanced case is also in stark contrast with the favorable behavior of EM for the exact-fitted settings analyzed in past work. Balakrishnan et al. [10] showed that when the EM algorithm is used to fit a two-component Gaussian mixture with sufficiently large value of $\frac{\|\theta^*\|_2}{\sigma}$ (known as the high signal-to-noise ratio, or high SNR for short), the population EM operator is contractive, and hence geometrically convergent, within a neighborhood of the true parameter θ^*. In a later work on the two-component balanced mixture fit model, Daskalakis et al. [64] showed that the convergence is in fact geometric for any non-zero value of the SNR. The model considered in Theorem 7.2 can be seen as the limiting case of weak signal for a two mixture model—which degenerates to the Gaussian distribution when the SNR becomes exactly zero. For such a limit, we observe that the fast convergence of population EM sequence no longer holds.

7.4.2.2 Upper and lower bounds on sample EM

We now turn to the statements of upper and lower bounds on the rate of the sample EM iterates for the balanced fit on Gaussian data. We begin with an upper bound, which involves the previously defined function $\gamma_{\text{up}}(\theta) := 1 - p + p/\left(1 + \frac{\|\theta\|_2^2}{2\sigma^2}\right)$.

Theorem 7.3. *Consider the sample EM updates $\theta^t = M_n(\theta^{t-1})$ for the balanced instance ($\pi = \frac{1}{2}$) of the mixture model (7.3) based on n i.i.d. $\mathcal{N}(0, \sigma^2 I_d)$ samples. Then, there exist universal constants $\{c_k'\}_{k=1}^4$ such that for any scalars $\alpha \in (0, \frac{1}{4})$ and $\delta \in (0, 1)$, any sample size $n \geq c_1'(d + \log(\log(1/\alpha)/\delta))$ and any iterate number $t \geq c_2' \log \frac{\|\theta^0\|^2 n}{\sigma^2 d} + c_3' \left(\frac{n}{d}\right)^{\frac{1}{2}-2\alpha} \log(\frac{n}{d}) \log(\frac{1}{\alpha})$, we have*

$$\|\theta^t\|_2 \leq \left[\left\|\theta^0\right\|_2 \cdot \prod_{j=0}^{t-1} \gamma_{up}(\theta^j)\right] + c_4'\sigma \left(\frac{\sigma^2(d + \log \frac{\log(4/\epsilon)}{\delta})}{n}\right)^{\frac{1}{4}-\alpha}, \tag{7.16}$$

with probability at least $1 - \delta$.

See Section 7.5 for a discussion of the techniques employed to prove this theorem. The detailed proof is provided in Appendix F.3, where we also provide some more details on the definitions of these constants.

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 (7.16) dominates the first term; therefore, from this point onwards, the the sample EM iterates have Euclidean norm of the order $(d/n)^{\frac{1}{4}-\alpha}$. Note that $\alpha \in (0, \frac{1}{4})$ can be chosen arbitrarily close to zero, so at the expense of increasing the lower bound on the number of iterations t by a logarithmic factor $\log(1/\alpha)$, we can obtain rates arbitrarily close to $(d/n)^{\frac{1}{4}}$.

We note that earlier studies of parameter estimation for over-specified mixtures, in both the frequentist [42] and Bayesian settings [127, 194], have derived a rate of $n^{-\frac{1}{4}}$ for the global maximum of the log likelihood. To the best of our knowledge, Theorem 7.3 is the first non-asymptotic algorithmic result that shows that such rates apply to the fixed points and dynamics of the EM algorithm, which need not converge to the global optima.

The preceding discussion was devoted to an upper bound on sample EM for the balanced fit. 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 $n^{-\frac{1}{4}}$. In particular, we prove the following lower bound.

Theorem 7.4. *There are universal positive constants c, c' such that for any non-zero solution $\widehat{\theta}_n$ to the sample EM fixed-point equation $\theta = M_n(\theta)$ for the balanced mixture fit, we have*

$$\mathbb{P}\left[|\widehat{\theta}_n| \geq c\, n^{-\frac{1}{4}}\right] \geq c'. \tag{7.17}$$

See Appendix F.4 for the proof of this theorem.

Since the iterative EM scheme converges only to one of its fixed points, the theorem shows that one cannot obtain a high-probability bound for any radius smaller than $n^{-\frac{1}{4}}$. As a consequence, with constant probability, the radius of convergence $n^{-\frac{1}{4}}$ for sample EM convergence in Theorem 7.3 for the univariate setting is tight.

7.5 New techniques for sharp analysis of sample EM

In this section, we highlight the new proof techniques introduced in this work that are required to obtain the sharp characterization of the sample EM updates in the balanced case (Theorem 7.3). We begin in Section 7.5.1 by elaborating that a direct application of the previous frameworks leads to sub-optimal statistical rates for sample EM in the balanced fit. This sub-optimality motivates the development of new methods for analyzing the behavior of the sample EM iterates, based on an *annulus-based localization argument* over a sequence of epochs, which we sketch out in Sections 7.5.2 and 7.5.3. We remark that our novel techniques, introduced here for analyzing EM with the balanced fit, are likely to be of independent interest. We believe that they can potentially be extended to derive sharp statistical rates in other settings when the algorithm under consideration does not exhibit an geometrically fast convergence.

7.5.1 A sub-optimal guarantee

Let us recall the set-up for the procedure suggested by Balakrishnan et al. [10], 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 $\theta^{t+1} = M_n(\theta^t)$ can be upper bounded by a sum of two terms as follows:

$$\left\|\theta^{t+1}\right\|_2 = \left\|M_n(\theta^t)\right\|_2 \leq \left\|M_n(\theta^t) - M(\theta^t)\right\|_2 + \left\|M(\theta^t)\right\|_2 \tag{7.18}$$

for all $t \geq 0$. 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 2 from Balakrishnan et al. [10] 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, and a contraction coefficient $\kappa \in (0, 1)$ such that

$$\sup_{\|\theta\|_2 \geq \epsilon} \frac{\|M(\theta)\|_2}{\|\theta\|_2} \leq \kappa \text{ and } \mathbb{P}\left[\sup_{\|\theta\|_2 \leq r} \|M_n(\theta) - M(\theta)\|_2 \leq \varepsilon(n, \delta)\right] \geq 1 - \delta. \tag{7.19a}$$

Then for a sample size n sufficiently large and ϵ sufficiently small to ensure that

$$\epsilon \overset{(i)}{\leq} \frac{\varepsilon(n, \delta)}{1 - \kappa} \overset{(ii)}{\leq} r, \tag{7.19b}$$

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 7.1. *There exists universal positive constants c_1 and c_2 such that for any positive radius r, any threshold $\delta \in (0, 1)$, and any sample size $n \geq c_2 d \log(1/\delta)$, we have*

$$\mathbb{P}\left[\sup_{\|\theta\|_2 \leq r} \|M_n(\theta) - M(\theta)\|_2 \leq c_1 \sigma(\sigma r + \rho)\sqrt{\frac{d + \log(1/\delta)}{n}}\right] \geq 1 - \delta, \tag{7.20}$$

where $\rho = |1 - 2\pi|$ denotes the imbalance in the mixture fit (7.3).

The proof of this lemma is based on Rademacher complexity arguments; see Appendix B.1 for the details.

With the choice $r = \|\theta^0\|_2$, Lemma 7.1 guarantees that the second inequality in line (7.19a) holds with $\varepsilon(n, \delta) \lesssim \sigma^2 \|\theta^0\|_2 \sqrt{d/n}$. On the other hand, Theorem 7.2 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 (7.19b), we solve the equation $\varepsilon(n, \delta)/(1 - \kappa(\epsilon)) = \epsilon$. Tracking only the dependency on d and n, we obtain[4]

$$\frac{\sqrt{d/n}}{\epsilon^2} = \epsilon \quad \implies \quad \epsilon = \mathcal{O}\left((d/n)^{\frac{1}{6}}\right), \tag{7.21}$$

which shows that the Euclidean norm of the sample EM iterate is bounded by a term of order $(d/n)^{\frac{1}{6}}$.

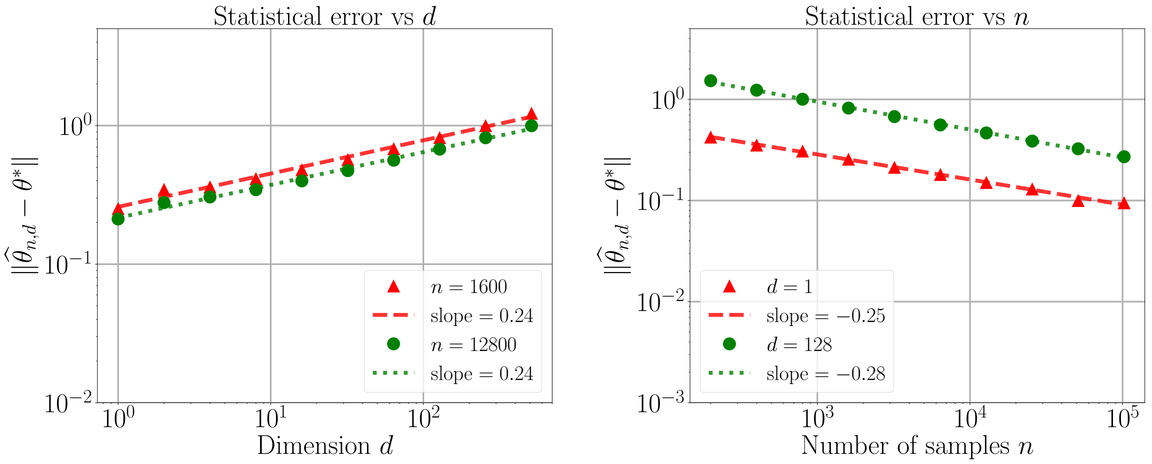
While this rate is much slower than the classical $(d/n)^{\frac{1}{2}}$ rate that we established in the unbalanced case, it does not coincide with the $n^{-\frac{1}{4}}$ rate that we obtained in Figure 7.1(b) for balanced setting with $d = 1$. Thus, the proof technique based on the framework of Balakrishnan et al. [10] appears to be non-optimal. The sub-optimality of this approach necessitates the development of a more refined technique. Before sketching this technique, we now quantify empirically the convergence rate of sample EM in terms of both dimension d and sample size n for the balanced mixture fit. In Figure 7.4, we summarize the results of these experiments. The two panels in the figure exhibit that the error in the sample EM estimate scales as $(d/n)^{\frac{1}{4}}$, thereby providing further numerical evidence that the preceding approach indeed led to a sub-optimal result.

7.5.2 Annulus-based 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 (a) for measuring the contractivity of the updates, as in the first inequality in equation (7.19a), *and* (b) for determining the final accuracy that we achieve. At earlier phases of the iteration, the algorithm will converge *more quickly* than suggested by the worst-case analysis based on the final accuracy. 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 7.1. 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 ideas of localizing the radius r for an empirical process plays a crucial role in obtaining sharp bounds on the error of M-estimation procedures [239, 14, 143, 246].

A novel aspect of the localization argument in our setting is the use of an annulus instead of a ball. In particular, we analyze the iterates from the EM algorithm assuming that they

[4] Moreover, with this choice of ϵ, inequality (ii) in equation (7.19b) is satisfied with a constant r, as long as n is sufficiently large relative to d.



(a) (b)

Figure 7.4: Scaling of the Euclidean error $\|\widehat{\theta}_{n,d} - \theta^*\|_2$ for EM estimates $\widehat{\theta}_{n,d}$ computed using the balanced ($\pi = \frac{1}{2}$) mixture-fit (7.2). 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 $(d/n)^{\frac{1}{4}}$ for the final iterate of sample-based EM.

lie within a pre-specicied annulus, defined by an inner and an outer radius. On one hand, the outer radius of the annulus helps to provide a sharp control on the perturbation bounds between the population and sample operators. On the other hand, the inner radius of the annulus is used to tightly control the algorithmic rate of convergence.

We now summarize our key arguments. The entire sequence of sample EM iterations is broken up into a sequence of different epochs. During each epoch, we localize the EM iterates to an annulus. In more detail:

- 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 $[0, \frac{1}{4}]$. The input to epoch ℓ is the scalar α_ℓ, and the output from epoch ℓ is the scalar $\alpha_{\ell+1}$.

- The ℓ-th epoch is defined to be the set of all iterations t of the sample EM algorithm such that the sample EM iterate θ^t lies in the following annulus:

$$\left(\frac{d}{n}\right)^{\alpha_{\ell+1}} \leq \|\theta^t - \theta^*\|_2 \leq \left(\frac{d}{n}\right)^{\alpha_\ell}. \tag{7.22}$$

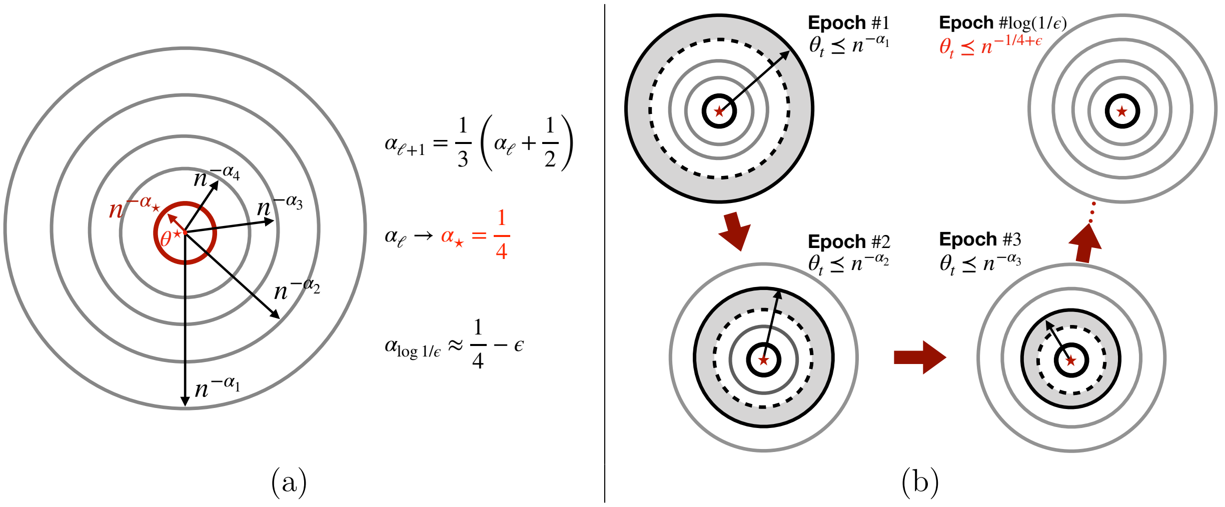

$n^{-\alpha_\star}$
θ^*
$n^{-\alpha_4}$
$n^{-\alpha_3}$
$n^{-\alpha_2}$
$n^{-\alpha_1}$
$\alpha_{\ell+1} = \frac{1}{3}\left(\alpha_\ell + \frac{1}{2}\right)$
$\alpha_\ell \to \alpha_\star = \frac{1}{4}$
$\alpha_{\log 1/\epsilon} \approx \frac{1}{4} - \epsilon$
(a)
Epoch #1
$\theta_t \precsim n^{-\alpha_1}$
Epoch #log(1/ε)
$\theta_t \precsim n^{-1/4+\epsilon}$
Epoch #2
$\theta_t \precsim n^{-\alpha_2}$
Epoch #3
$\theta_t \precsim n^{-\alpha_3}$
(b)

Figure 7.5: Illustration of the annulus-based-localization argument part (I): Defining the epochs or equivalently the annuli. (a) Outer 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 annulus around θ^* with inner and outer radii given by $n^{-\alpha_{\ell+1}}$, and $n^{-\alpha_\ell}$, 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 annulus converges linearly to $n^{-\frac{1}{4}}$. Consequently, after at most $\log(1/\alpha)$ epochs (or $\sqrt{n}\log(1/\alpha)$ iterations), the EM iterate lies in a ball of radius $n^{-1/4+\alpha}$ around θ^*. We illustrate the one-step dynamics in any given annulus in Figure 7.6.

We establish that the sample-EM operator is non-expansive so that each epoch is well-defined (and that subsequent iterations can only correspond to subsequent epochs).

- Upon completion of epoch ℓ at 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

$$\alpha_{\ell+1} = \frac{1}{3}\alpha_\ell + \frac{1}{6}. \tag{7.23}$$

  Note that the new scalar $\alpha_{\ell+1}$ serves as the input to epoch $\ell+1$.

The recursion (7.23) 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 at the end of epoch ℓ.

A few properties of the recursion (7.23) are worth noting. First, given our initialization $\alpha_0 = 0$, we see that $\alpha_1 = \frac{1}{6}$, 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

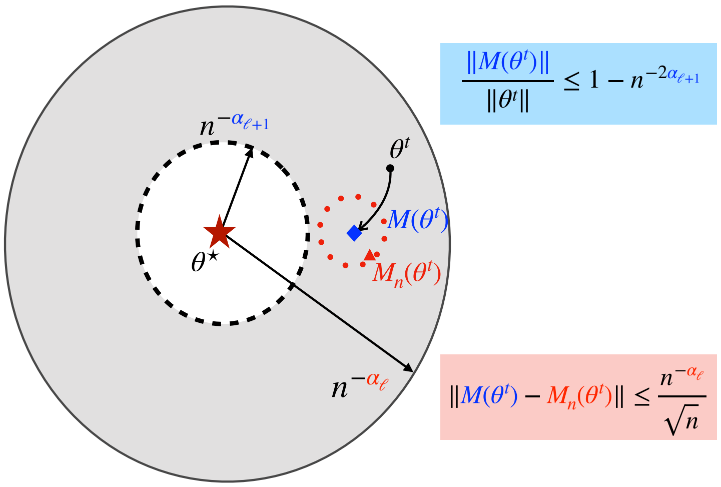

$\frac{\|M(\theta^t)\|}{\|\theta^t\|} \leq 1 - n^{-2\alpha_{\ell+1}}$
$n^{-\alpha_{\ell+1}}$
θ^t
$M(\theta^t)$
$\theta^\star$
$M_n(\theta^t)$
$n^{-\alpha_\ell}$
$\|M(\theta^t) - M_n(\theta^t)\| \leq \frac{n^{-\alpha_\ell}}{\sqrt{n}}$

Figure 7.6: Illustration of the annulus-based-localization argument part (II): Dynamics of EM in the ℓ-th epoch or equivalently the annulus $n^{-\alpha_{\ell+1}} \leq \|\theta^t - \theta^*\|_2 \leq n^{-\alpha_\ell}$. 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 annulus with inner and outer radii given by $n^{-\alpha_{\ell+1}}$, and $n^{-\alpha_\ell}$, 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 $\ell + 1$.

$\alpha^* = \frac{1}{4}$. Thus, our argument will allow us to prove a bound arbitrarily close to $(d/n)^{\frac{1}{4}}$, as stated formally in Theorem 7.3 to follow. Refer to Figures 7.5 and 7.6 for an illustration of the definition of these annuli, epochs and the associated conclusions.

7.5.3 How does the key recursion (7.23) arise?

Let us now sketch out how the key recursion (7.23) arises. Consider epoch ℓ specified by input $\alpha_\ell < \frac{1}{4}$, and consider an iterate θ^t in the following annulus: $\|\theta^t\|_2 \in [(d/n)^{\alpha_{\ell+1}}, (d/n)^{\alpha_\ell}]$. 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; for details, see Lemma F.3 stated in the Appendix. Given this guarantee, our second step is to make use of the inner radius of the considered annulus to apply Theorem 7.2 for the population EM operator, for all iterations t such that $\|\theta^t\|_2 \geq$

$(d/n)^{\alpha_{\ell+1}}$. Consequently, for these iterations, we have

$$\begin{aligned}\left\|M(\theta^t)\right\|_2 &\leq \left(1 - p + \frac{p}{1 + \frac{\|\theta\|_2^2}{2\sigma^2}}\right) \left\|\theta^t\right\|_2 \\ &\precsim (1 - (d/n)^{2\alpha_{\ell+1}})(d/n)^{\alpha_\ell} \leq \widetilde{\gamma} \left(\frac{d}{n}\right)^{\alpha_\ell}, \end{aligned} \tag{7.24a}$$

where $\widetilde{\gamma} \coloneqq e^{-(d/n)^{2\alpha_{\ell+1}}}$. On the other hand, using the outer radii of the annulus and applying Lemma 7.1 for this epoch, we obtain that

$$\left\|M_n(\theta^t) - M(\theta^t)\right\|_2 \precsim \left(\frac{d}{n}\right)^{\alpha_\ell} \sqrt{\frac{d}{n}} = \left(\frac{d}{n}\right)^{\alpha_\ell + 1/2}, \tag{7.24b}$$

for all t in the epoch. Unfolding the basic triangle inequality (7.18) for T steps, we find that

$$\begin{aligned}\left\|\theta^{t+T}\right\|_2 &\leq \left\|M_n(\theta^t) - M(\theta^t)\right\|_2 (1 + \widetilde{\gamma} + \ldots + \widetilde{\gamma}^{T-1}) + \widetilde{\gamma}^T \left\|\theta_t\right\|_2 \\ &\leq \frac{1}{1 - \widetilde{\gamma}} \left\|M_n(\theta^t) - M(\theta^t)\right\|_2 + e^{-T(d/n)^{2\alpha_{\ell+1}}} (d/n)^{\alpha_\ell}.\end{aligned}$$

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

$$\left\|\theta^{t+T}\right\|_2 \precsim \frac{1}{1 - \widetilde{\gamma}} \left\|M_n(\theta^t) - M(\theta^t)\right\|_2 \approx \underbrace{\left(\frac{d}{n}\right)^{-2\alpha_{\ell+1}} \left(\frac{d}{n}\right)^{\alpha_\ell + 1/2}}_{=:\, r}. \tag{7.25}$$

The epoch is said to be complete once $\left\|\theta^{t+T}\right\|_2 \precsim \left(\frac{d}{n}\right)^{\alpha_{\ell+1}}$. Disregarding constants, this condition is satisfied when $r = \left(\frac{d}{n}\right)^{\alpha_{\ell+1}}$, or equivalently when

$$\left(\frac{d}{n}\right)^{-2\alpha_{\ell+1}} \left(\frac{d}{n}\right)^{\alpha_\ell + 1/2} = \left(\frac{d}{n}\right)^{\alpha_{\ell+1}}.$$

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 (7.23). Refer to Figure 7.6 for a visual illustration of the localization argument summarized above for a given epoch.

Of course, the preceding discussion is informal, and there remain many details to be addressed in order to obtain a formal proof. We refer the reader to Appendix F.3 for the complete argument.

7.6 Generality of results

Thus far, we have characterized the behavior of the EM algorithm for different settings of over-specified location Gaussian mixtures. We established rigorous statistical guarantees of EM under two particular but representative settings of over-specified location Gaussian mixtures: the balanced and unbalanced mixture-fit. The log-likelihood for the unbalanced fit remains strongly log-concave[5] (due to the fixed weights and location parameters being sign flips) and hence the Euclidean error of the final iterate of EM decays at the usual rate $(d/n)^{\frac{1}{2}}$ with n samples in d dimensions. However, in the balanced case, the log-likelihood is no longer strongly log-concave and the error decays at the slower rate $(d/n)^{\frac{1}{4}}$. We view our results as the first step in understanding and possibly improving the EM algorithm in non-regular settings. We now provide a detailed discussion that sheds light on the general applicability of our results. In particular, we discuss the behavior of EM under the following settings: (i) over-specified mixture models with unknown weight parameters (Section 7.6.1), (ii) over-specified mixture of linear regression (Section 7.6.2), and (iii) more general settings with over-specified mixture models (Section 7.6.3). We conclude the chapter with a discussion of several future directions that arise from the previous settings in Section 7.7.

7.6.1 When the weights are unknown

Our theoretical analysis so far assumed that the weights were fixed, an assumption common to a number of previous papers in the area [10, 64, 146]. In Appendix F.7.2, we consider the case of unknown weights for the model fit (7.3). In this context, our main contribution is to show that if the weights are initialized far away from $\frac{1}{2}$—meaning that the initial mixture is highly unbalanced—then the EM algorithm converges quickly, and the results from Theorem 7.1 are valid. (See Lemma F.5 in Appendix F.7.2 for the details.) On the other hand, if the initial mixture is not heavily imbalanced, we observe the slow convergence of EM consistent with Theorems 7.2 and 7.3.

7.6.2 Slow rates for mixture of regressions

Thus far, we have considered the behavior of the EM algorithm in application to parameter estimation in mixture models. Our findings turn out to hold somewhat more generally, with Theorems 7.2 and 7.3 having analogues when the EM algorithm is used to fit a mixture of linear regressions in over-specified settings. Concretely, suppose that $(Y_1, X_1), \ldots, (Y_n, X_n) \in \mathbb{R} \times \mathbb{R}^d$ are i.i.d. samples generated from the model

$$Y_i = X_i^\top \theta^* + \sigma\sqrt{2\eta}\xi_i, \qquad \text{for } i = 1, \ldots, n, \tag{7.26}$$

where $\{\sqrt{2\eta}\xi_i\}_{i=1}^n$ are i.i.d. standard Gaussian variates, and the covariate vectors $X_i \in \mathbb{R}^d$ are also i.i.d. samples from the standard multivariate Gaussian $\mathcal{N}(0, I_d)$. Of interest is to estimate the parameter θ^* using these samples and EM is a popular method for doing so. When θ^* has sufficiently large Euclidean norm, a setting referred to as the strong signal case, Balakrishnan et al. [10] showed that the estimate returned by EM is at a distance $(d/n)^{\frac{1}{2}}$

[5] Moreover, in Appendix F.8 we differentiate the unbalanced and balanced fit based on the log-likelihood and the Fisher matrix and provide a heuristic justification for the different rates between the two cases.

from the true parameter θ^* with high probability. On the other hand, our analysis shows that when $\|\theta^*\|_2$ decays to zero—leading to an over-specified setting—the convergence of EM becomes slow. In particular, the EM algorithm takes significantly more steps and returns an estimate that is statistically worse, lying at Euclidean distance of the order $(d/n)^{\frac{1}{4}}$ from the true parameter. While the EM operators in this case are slightly different when compared to the over-specified Gaussian mixture analyzed before, the proof techniques remain similar. More concretely, we first show that the convergence of population EM is slow (similar to Theorem 7.2) and then use the annulus-based localization argument (similar to the proof of Theorem 7.3 from Section 7.5) to derive a sharp rate. For completeness, we present these results formally in Lemma F.6 and Corollary F.2 in Appendix F.9.

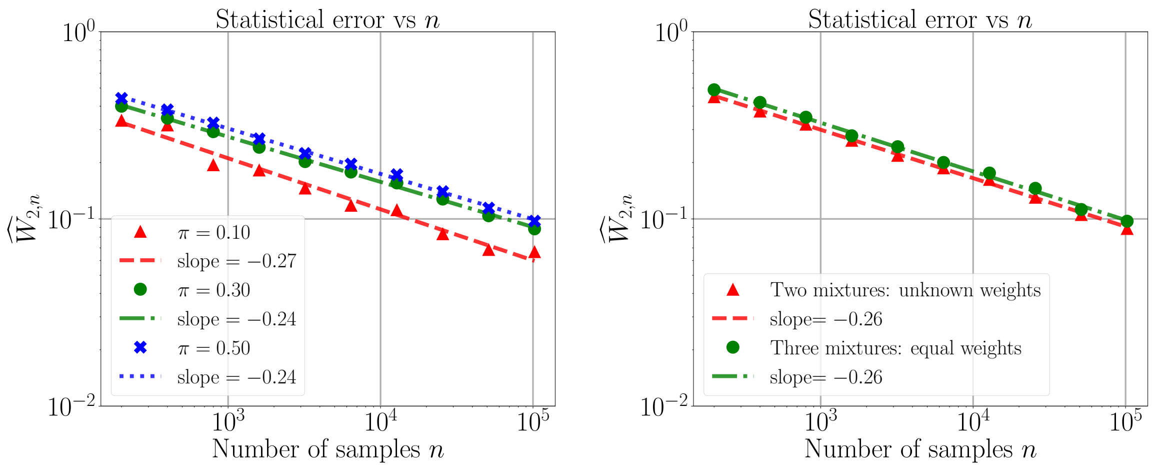
7.6.3 Slow rates for general mixtures

We now present several experiments that provide numerical backing to the claim that the slow rate of order $n^{-\frac{1}{4}}$ is not merely an artifact of the special balanced fit ((7.3) with $\pi = \frac{1}{2}$). We demonstrate that the slow convergence of EM is very likely to arise while fitting general over-specified location Gaussian mixtures with unknown weights (and known covariance). We consider three settings: (A) fitting several general over-specified location Gaussian mixture fits to Gaussian data (Figure 7.7), (B) fitting a special three-component mixture fit to a two mixture of Gaussian (Figure 7.8), and (C) fitting mixtures with unknown weights and location parameters when the number of components in the fitted model is over-specified by two (Figure 7.9). We now turn to the details of these settings.

General over-specified mixture fits on Gaussian data: First, we remark that the fast convergence in the unbalanced fit (Theorem 7.1) was a joint result of the facts that (a) the weights were fixed and unequal, and (b) the parameters were constrained to be a sign flip. If either of these conditions is violated, the EM algorithm exhibits slow convergence on both algorithmic and statistical fronts. Theorems 7.2, 7.3 and 7.4 provide rigorous details for the case of equal and fixed weights (balanced fit). When the weights are unknown, EM can exhibit slow rate (see Section 7.6.1 and Appendix F.7.2 for further details). When the weights are fixed and unequal, but the location parameters are estimated freely—that is, with the model $\pi\phi(x;\theta_1,1)+(1-\pi)\phi(x;\theta_2,1)$, as illustrated in Figure 7.7(a)—then the EM estimates have error[6] of order $n^{-\frac{1}{4}}$. In such cases, the parameter estimates approximately satisfy the relation $\sum_k \pi_k\widehat{\theta}_{k,n} \approx 0$, since the mean of the data is close to zero; moreover, for a two-components mixture model, the location estimates become weighted sign flips of each other. The features are the intuitive reason underlying the similarity of behavior of EM between this fit and the balanced fit. Finally, when we fit a two mixture model with

[6] For more general cases, we measure the error of parameter estimation using the Wasserstein metric of second order $\widehat{W}_{2,n}$ to account for label-switching between the components. When the true model is standard Gaussian this metric is simply the weighted Euclidean error: $(\sum \pi_k\widehat{\theta}_{k,n}^2)^{\frac{1}{2}}$, where π_k and $\widehat{\theta}_{k,n}$, respectively, denote the mixture weight and the location parameter of the k-th component of the mixture.

unknown weight parameter and free location parameters, the final error also has a scaling of order $n^{-\frac{1}{4}}$. Refer to Figure 7.7 for a numerical validation of these results.

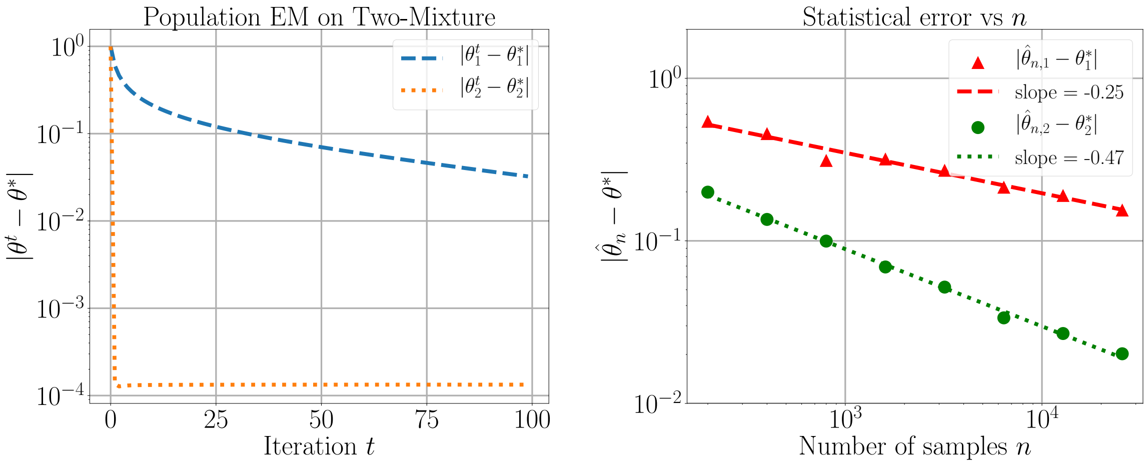


(a) (b)

Figure 7.7: 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 $\pi\mathcal{N}(\theta_1,1)+(1-\pi)\mathcal{N}(\theta_2,1)$ with three different *fixed* values of weights $\pi \in \{0.1, 0.3, 0.5\}$ and two (unconstrained) location parameters, 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 and two location parameters $\pi\mathcal{N}(\theta_1,1)+(1-\pi)\mathcal{N}(\theta_2,1)$, whereas green circles correspond to results fitting a three-component mixture model $\sum_{i=1}^{3}\frac{1}{3}\mathcal{N}(\theta_i,1)$. In all cases, the EM solutions exhibit the slow $n^{-\frac{1}{4}}$ statistical rate for the error in parameter estimation. Also see Figure 7.9.

Over-specified fits for mixtures of Gaussian data: Using similar reasoning as above, let us sketch out how our theoretical results also yield usable predictions for more general over-specified models. Roughly speaking, whenever there are extra number of components to be estimated, parameters of some of them are likely to end up satisfying certain form of local constraint. More concretely, suppose that we are given data generated from a k-component mixture, and we use the EM algorithm to fit the location parameters of a mixture model with $k+1$ components. Loosely speaking, the EM estimates corresponding to a set of $k-1$ components are likely to converge quickly, leaving the two remaining components to fit a single component in the true model. If the other components are far away, the EM updates for the parameters of these two components are unaffected by them and start to behave like the balanced case. See Figure 7.8 for a numerical illustration of this intuition in an idealized setting where we use $k+1=3$ components to fit data generated from a $k=2$ component model. In this idealized setting, the error for one of the parameter scales at the fast rate of order $n^{-\frac{1}{2}}$, and that of the parameter that is locally over-fitted exhibits a slow rate of order $n^{-\frac{1}{4}}$. Finally, we see that the statistical error of order $n^{-\frac{1}{4}}$ also arises when we over-specify the number of components by more than one. In particular, we observe in Figure 7.7(b) (green dashed dotted line with solid circles) and Figure 7.9 (both curves) that a similar scaling of order $n^{-\frac{1}{4}}$ arises when we over-specify the number of components by 2 and estimate the weight and location parameters.

Besides formally analyzing EM in these general cases, several other future directions arise from our work which we now discuss.



(a) (b)

Figure 7.8: Behavior of EM for an over-specified Gaussian mixture with more than two components. 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 $n^{-\frac{1}{2}}$ rate, for the no signal component EM has the slower $n^{-\frac{1}{4}}$ rate, which is in good agreement with the theoretical results derived in the chapter. (We remark that the error floor for θ_2^t in panel (a) arises from the finite precision inherent to numerical integration.)

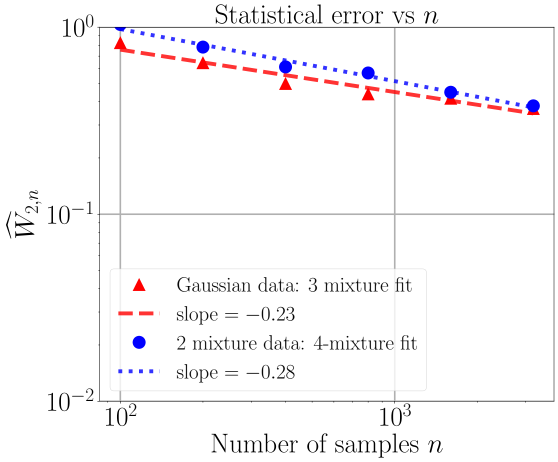


Figure 7.9: Plots of Wasserstein error when both weights and location parameters are unknown and estimated using EM and the fitted multivariate mixture model is over-specified. (a) True model: $\mathcal{N}([0,0]^\top, I_2)$, and fitted model $\sum_{i=1}^{3} w_i \mathcal{N}(\theta_i, I_2)$ and (b) True model: $\frac{2}{5}\mathcal{N}([0,0]^\top, I_2) + \frac{3}{5}\mathcal{N}([4,4]^\top, I_2)$ and fitted model: $\sum_{i=1}^{4} w_i \mathcal{N}(\theta_i, I_2)$. In both cases, once again we see the scaling of order $n^{-\frac{1}{4}}$ for the final error (similar to results in Figure 7.7 and 7.8).

7.7 Conclusion and future directions

In this chapter, we assumed that only the location parameters were unknown and that the scale parameters of the underlying model are known. Nevertheless in practice, this assumption is rather restrictive and it is natural to ask what happens if the scale parameters were also unknown. We note that the MLE is known to have even slower statistical rates for the estimation error with such higher-order mixtures; therefore, apriori it is an interesting question to determine if the EM algorithm also suffers from a similar slow down when the scale parameters are unknown. The next chapter investigates this question in great detail.

Another important direction is to analyze the behavior of EM under different models for generating the data. While our analysis is focused on Gaussian mixtures, the non-standard statistical rate $n^{-\frac{1}{4}}$ also arises in other types of over-specified mixture models, such as those involving mixtures with other exponential family members, or Student-t distributions, suitable for heavy-tailed data. We believe that the analysis of this chapter can be generalized to a broader class of finite mixture models that includes the aforementioned models.

A final direction of interest is whether the behavior of EM—slow versus fast convergence—can be used as a statistic in a classical testing problemma: 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, due the singularity of the Fisher information matrix; see the papers [159, 43] for some past work on the problem. The results of this 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, then we may accept the null, whereas the opposite behavior can be used as an evidence for rejecting the null. Analyzing such a testing procedure based on the convergence rates of EM remains open.

Chapter 8

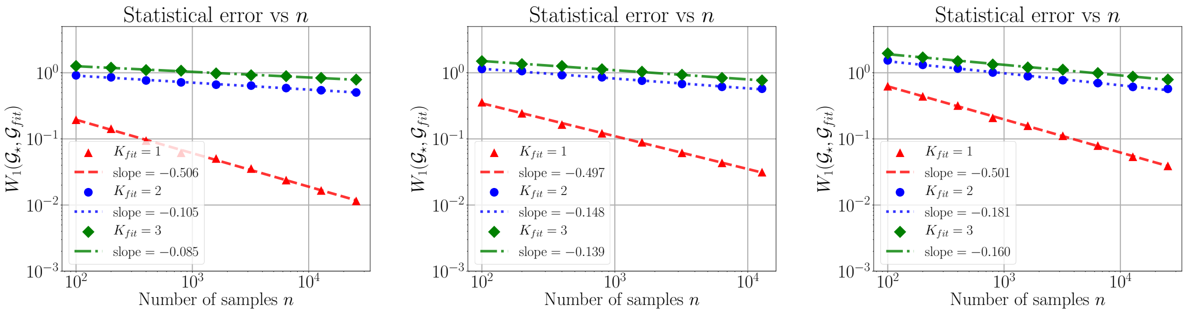
Sharp Analysis of EM for Weakly Identifiable Mixture Models

A more challenging class of mixture models are those that are only *weakly identifiable*, meaning that the Fisher information is degenerate over some larger set. This stronger form of singularity arises, for instance, when the scale (covariance) parameter in an over-specified Gaussian mixture is also unknown [40, 43]. Ho et al. [112] characterized the behavior of MLE for a class of weakly identifiable models. They showed that the convergence rates of MLE in these models could be very slow, with the precise rates determined by algebraic relations among the partial derivatives. However, this past work has not addressed the computational complexity of computing the MLE in a weakly identifiable model.

We start with an introduction contrasting this chapter with the work from the previous chapter, and discussion of the related work in Section 8.1. We then present some intriguing simulations in Section 8.2 that motivate the problem set-up considered in this chapter. With this context in place, we summarize the contributions of this chapter in Section 8.2.2, and discuss the organization of the remainder of the chapter in Section 8.2.3.

8.1 Introduction

In the previous chapter 7, we studied the behavior of EM for fitting a class of *non-regular* mixture models, namely those in which the Fisher information is degenerate at a point, but the model remains strongly identifiable. One such class of models are Gaussian location mixtures with known scale parameters that are *over-specified*, meaning that the number of components in the mixture-fit exceeds the number of components in the data generating distribution. For such non-regular but strongly identifiable mixture models, we showed that the EM algorithm takes $\mathcal{O}((n/d)^{\frac{1}{2}})$ steps to converge to a Euclidean ball of radius $\mathcal{O}((d/n)^{\frac{1}{4}})$ around the true location parameter. Recall that for such models, the MLE is known to lie at a distance $\mathcal{O}(n^{-\frac{1}{4}})$ from the true parameter [42], so that even though its convergence rate as an optimization algorithm is slow; the EM algorithm nonetheless produces a solution with



(a) $d = 1$ (b) $d = 2$ (c) $d = 4$

Figure 8.1: Scaling of the Wasserstein error between the true parameters and the EM estimates, when EM is used to fit a Gaussian mixture model with $K_{\text{fit}} \in \{1, 2, 3\}$ components, i.e., $\mathcal{G}_{\text{fit}} = \sum_{i=1}^{K_{\text{fit}}} w_i \mathcal{N}(\mu_i, \Sigma_i)$ with all parameters treated as unknown and estimated by EM, on an n sample-dataset generated from standard Gaussian distribution $\mathcal{G}_* = \mathcal{N}(0, I_d)$. In all three examples, when the fitted model is over-specified, meaning that the fitted model has more components than the true model ($K_{\text{fit}} \in \{2, 3\}$ in these examples), we observe a significant increase in the Wasserstein error. Stated differently, the simulations suggest that the estimation accuracy of the EM algorithm degrades dramatically when the fitted model is over-specified.

a statistical error of the same order as the MLE.

The previous chapter does not consider the more realistic setting in which both the location and scale parameters are unknown, and the EM algorithm is used to fit both simultaneously. Indeed, as mentioned earlier, such models may become weakly identifiable due to algebraic relations among the partial derivatives [43]. Thus, analyzing EM in the case of weakly identifiable mixtures is challenging for two reasons: (i) the weak separation between the mixture components, and (ii) the algebraic interdependence of the partial derivatives of the log-likelihood. The main contributions of this work are (a) to highlight the dramatic differences in the convergence behavior of the EM algorithm, depending on the structure of the fitted model relative to the data-generating distribution; and (b) to analyze the EM algorithm under a few specific yet representative settings of weakly identifiable models, giving a precise analytical characterization of its convergence behavior.

8.2 Illustrative examples and problem set-up

We note that the experimental settings in this chapter share same format as that in previous chapter (see Section 7.2.2.2).

To begin with, we consider the simplest case of over-specification with Gaussian mixture models—when the true data is generated from a zero-mean standard Gaussian distribution in

d dimensions and EM is used to fit a general multi-component mixture model with different number of mixtures. (We note that fitting by one mixture model is simply a Gaussian fit.) Given the estimates for the mixture weights, location and scale parameters returned by EM, we compute the first order Wasserstein distance[1] between the true and estimated parameters. Results for $d \in \{1, 2, 4\}$ and for various amount of over-specification are plotted in Figure 8.1. From these results, we notice that the decay in statistical error is $n^{-1/2}$ when the fitted number of components is well-specified and equal to the true number of components but has a much slower rate whenever the number of fitted components is two or more. Moreover, in Section 8.5 (see Figure 8.3) we show that such a phenomenon occurs more generally in mixture models.

While a rigorous theoretical analysis of EM under over-specification in general mixture models is desirable, it remains beyond the scope of this chapter. Instead, here we provide a full characterization of EM when it is used to fit the following class of models to the data drawn from standard Gaussian $\mathcal{N}(0, I_d)$:

$$\mathcal{G}_{\text{symm}}((\theta, \sigma^2)) = \frac{1}{2}\mathcal{N}(\theta, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(-\theta, \sigma^2 I_d). \tag{8.1}$$

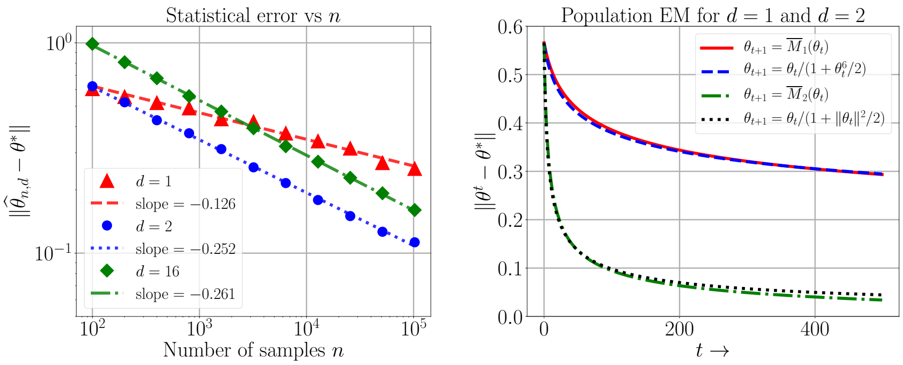
In particular, in this symmetric fit, we fix the mixture weights to be equal to $\frac{1}{2}$ and require that the two components have same scale parameter. Given the estimates $\widehat{\theta}, \widehat{\sigma}$, the Wasserstein error (see equation (G.40) in Appendix G.7) in this case can be simplified as $\|\widehat{\theta}\|_2 + \sqrt{d}\sqrt{|\widehat{\sigma}^2 - 1|}$. In our results to be stated later, we show that the two terms are of the same order (equations (8.6), (8.9)) and hence we primarily focus on the error $\|\widehat{\theta} - \theta_\star\|_2$ going forward to simplify the exposition. We consider our set-up as a simple yet first step towards understanding the behavior of EM in over-specified mixtures when *both* location and scale parameter are unknown. In the previous chapter, we studied the slow down of EM with over-specified mixtures for estimating only the location parameter, but they assumed that the scale parameter was known and fixed. Here a more general setting is considered.

Model (8.1) is weakly identifiable: We now elaborate the choice of our class of models (8.1) that may appear a bit restrictive at first glance. This model turns out to be the simplest example of a weakly identifiable model in $d = 1$. Let ϕ denote the density of a Gaussian distribution with mean θ and variance σ^2, then we have

$$\frac{\partial^2 \phi}{\partial \theta^2}(x; \mu, \sigma^2) = 2\frac{\partial \phi}{\partial \sigma^2}(x; \mu, \sigma^2), \tag{8.2}$$

valid for all $x \in \mathbb{R}$, $\mu \in \mathbb{R}$ and $\sigma > 0$. As alluded to earlier, models with algebraic dependence between partial derivatives lead to weak identifiability and slow statistical estimation with MLE. However, in the multivariate setting when the same parameter σ is shared across multiple dimensions, this algebraic relation does not hold and the model is strongly identifiable

[1] First-order Wasserstein distance has been used in prior works to characterize the error between the estimated and true parameters. See section 1.1 [113].



(a) (b)

Figure 8.2: Behavior of the EM algorithm for the fitted model (8.1), where the data is being generated from $\mathcal{N}(0, I_d)$. (a) Scaling of the Euclidean error $\|\widehat{\theta}_{n,d} - \theta^*\|_2$ with respect to the sample size n for dimension $d \in \{1, 2, 16\}$. Here, $\widehat{\theta}_{n,d}$ denotes the EM algorithm estimate of the mean parameter θ based on n samples. Note that the simulations indicate two distinct error scaling for $d = 1$ and $d > 1$. (b) Convergence behavior of the population-like EM sequence $\theta^{t+1} = \overline{M}_d(\theta^t)$ (8.12b) in dimensions $d = 1$ and 2. The rate of convergence in dimension $d = 1$ is significantly slower compared to the rate in dimension $d = 2$. Overall, both the plots provide strong empirical evidence towards two distinct behaviors of the EM algorithm for dimension $d = 1$ and dimensions $d > 1$. See the Theorems 8.1-8.2, and Lemmas 8.1 and 8.2 for a theoretical justification of trends in panels (a) and (b) respectively.

(since the Fisher information matrix is singular at $(\theta^*, \sigma^*) := (0, 1)$). For this reason, we believe that analysis of EM for the special fit (8.1) may provide important insight for more general over-specified weakly identifiable models.

Population EM: Given n samples from a d-dimensional standard Gaussian distribution, the sample EM algorithm for location and scale parameters generates a sequence of the form $\theta^{t+1} = M_{n,d}(\theta^t)$ and σ^{t+1}, which is some function of $\|\theta^{t+1}\|_2^2$; see equation (8.3b) for a precise definition. An abstract counterpart of the sample EM algorithm—not useful in practice but rather for theoretical understanding—is the population EM algorithm $\overline{M}_d$, obtained in the limit of an infinite sample size (cf. equation (8.12b)).

In practice, running the sample EM algorithm yields an estimate $\widehat{\theta}_{n,d}$ of the unknown location parameter θ^*. Panel (a) in Figure 8.2 shows the scaling of the statistical estimation error $\|\widehat{\theta}_{n,d} - \theta^*\|_2$ of this sample EM estimate versus the sample size n on a log-log scale. The three curves correspond to dimensions $d \in \{1, 2, 16\}$, along with least-squares fits (on the log-log scale) to the data. In panel (b), we plot the Euclidean norm $\|\theta^t\|_2$ of the population

EM iterate[2] versus the iteration number t, with solid red line corresponding to $d = 1$ and the dash-dotted green line corresponding to $d = 2$. Observe that the algorithm converges far more slowly in the univariate case than the multivariate case. The theory to follow in this chapter (see Theorems 8.1, 8.2 and Lemmas 8.1 and 8.2) provides explicit predictions for the rate at which different quantities plotted in Figure 8.2 should decay. We now summarize our theoretical results that are also consistent with the trends observed in Figure 8.2.

8.2.1 EM updates for symmetric fit $\mathcal{G}_{\text{symm}}$

The EM updates for Gaussian mixture models are standard, so we simply state them here. In terms of the shorthand notation $\eta := (\theta, \sigma)$, the E-step in the EM algorithm involves computing the function

$$Q_n(\eta'; \eta) := \frac{1}{n} \sum_{i=1}^{n} \left[w_{\theta,\sigma}(X_i) \log \left(\phi(X_i; \mu', (\sigma')^2 I_d) \right) (1 - w_{\theta,\sigma}(X_i)) \log \left(\phi(X_i; -\mu', (\sigma')^2 I_d) \right) \right],$$

where the weight function is given by $w_{\theta,\sigma}(x) = (1 + e^{\frac{-2\theta^\top x}{\sigma^2}})^{-1}$. The M-step involves maximizing the Q_n-function over the pair (θ', σ') with η fixed, which yields

$$\mu' = \frac{1}{n} \sum_{i=1}^{n} (2 w_{\mu,\sigma}(X_i) - 1) X_i, \quad \text{and} \quad (\sigma')^2 = \frac{1}{d} \left(\frac{\sum_{i=1}^{n} \|X_i\|_2^2}{n} - \|\mu'\|_2^2 \right). \tag{8.3a}$$

Doing some straightforward algebra, the EM updates (θ_n^t, σ_n^t) can be succinctly defined as

$$\begin{aligned} \theta_n^{t+1} &= \frac{1}{n} \sum_{i=1}^{n} \tanh \left(\frac{X_i^\top \theta_n^t}{\sum_{i=1}^{n} \|X_i\|_2^2 / (nd) - \|\theta_n^t\|_2^2 / d} \right) \\ &=: M_{n,d}(\theta_n^t), \end{aligned} \tag{8.3b}$$

and $\sigma_n^{t+1} = \sum_{i=1}^{n} \|X_i\|_2^2 / (nd) - \|\theta_n^{t+1}\|_2^2 / d$. For simplicity in presentation, we refer to the operator $M_{n,d}$ as the *sample EM operator.*

8.2.2 Our contributions

The main contribution of this chapter is to provide a precise analytical characterization of the behavior of the EM algorithm for certain special cases of over-specified mixture models (8.1).

[2] In fact, our analysis makes use of two slightly different population-level operators $\widetilde{M}_{n,d}$ and $\overline{M}_d$ defined in equations (8.16) and (8.12b) respectively. Figure 8.2(b) shows plots for the operator $\overline{M}_d$, but the results are qualitatively similar for the operator $\widetilde{M}_{n,d}$.

Univariate over-specified Gaussian mixtures: In the univariate setting ($d = 1$) of $\mathcal{G}_{\text{symm}}$ in (8.1), we prove that the EM estimate has statistical estimation error of the order $n^{-\frac{1}{8}}$ and $n^{-\frac{1}{4}}$ after order $n^{\frac{3}{4}}$ steps for the location and scale parameters respectively. In particular, Theorem 8.1 provides a theoretical justification for the slow rate observed in Figure 8.2 (a) for $d = 1$ (red dotted line with star marks). Proving these rates requires a novel analysis, and herein lies the main technical contribution of this chapter. Indeed, we show that all the analysis techniques introduced in past work on EM, including work on both the regular [10] and strongly identifiable cases [82], lead to sub-optimal rates. Our novel method is a *two-stage approach* that makes use of two different population level EM operators. Moreover, we also prove a matching lower bound (see Appendix G.1) which ensures that the upper bound of order $n^{-\frac{1}{8}}$ for the statistical error of sample EM from Theorem 8.1 is tight up to constant factors.

Multivariate setting with shared covariance: Given the technical challenges even in the simple univariate case, the symmetric spherical fit $\mathcal{G}_{\text{symm}}$ in (8.1) serves as a special case for the multivariate setting $d \geq 2$. In this case, we establish that the sharing of scale parameter proves beneficial in the convergence of EM. Theorem 8.2 shows that sample EM algorithm takes $\mathcal{O}((n/d)^{1/2})$ steps in order to converge to estimates, of the location and scale parameters respectively, that lie within distances $\mathcal{O}(d/n)^{1/4}$ and $\mathcal{O}(nd)^{-\frac{1}{2}}$ of the true location and scale parameters, respectively.

General multivariate setting: We want to remind the readers that we expect the Wasserstein error to scale much slowly than $n^{-\frac{1}{4}}$ (the rate mentioned in the previous paragraph) while estimating over-specified mixtures with no shared covariance. When the fitted variance parameters are not shared across dimensions our simulations under general multi-component fits in Figure 8.1 demonstrate a much slower convergence of EM (for which a rigorous justification is beyond the scope of this chapter).

8.2.3 Organization

The remainder of the chapter is organized as follows. We present our main results in Section 8.3, with Section 8.3.1 devoted to the univariate case, Section 8.3.2 to the multivariate case and Section 8.3.3 to the simulations with more general mixtures. Our proof ideas are summarized in Section 8.4 and we conclude with a discussion in Section 8.5. The detailed proofs of all our results are deferred to the Appendices.

Notation: In this chapter, the expressions $a_n \precsim b_n$ or $a_n \leq \mathcal{O}(b_n)$ will be used to denote $a_n \leq c b_n$ for some positive universal constant c that does not change with n. Additionally, we write $a_n \asymp b_n$ if both $a_n \precsim b_n$ and $b_n \precsim a_n$ hold. Furthermore, we denote $[n]$ as the set $\{1, \ldots, n\}$ for any $n \geq 1$. We define $\lceil x \rceil$ as the smallest integer greater than or equal to x for any $x \in \mathbb{R}$. The notation $\|x\|_2$ stands for the ℓ_2 norm of vector $x \in \mathbb{R}^d$. We use c, c', c_1

etc. to denote some universal constants independent of problem parameters (which might change in value each time they appear).

8.3 Main results

In this section, we provide our main results for the behavior of EM with the singular (symmetric) mixtures fit $\mathcal{G}_{\mathrm{symm}}$ (8.1). Theorem 8.1 discusses the result for the univariate case, Theorem 8.2 discusses the result for multivariate case. In Section 8.3.3 we discuss some simulated experiments for general multivariate location-scale Gaussian mixtures.

8.3.1 Results for the univariate case

As discussed before, due to the relationship between the location and scale parameter, namely the updates (8.3b), it suffices to analyze the sample EM operator for the location parameter. For the univariate Gaussian mixtures, given n samples $\{X_i, i \in [n]\}$, the sample EM operator is given by

$$M_{n,1}(\theta) := \frac{1}{n}\sum_{i=1}^{n} X_i \tanh\left[\frac{X_i\theta}{\sum_{j=1}^{n} X_j^2/n - \theta^2}\right]. \tag{8.4}$$

We now state our first main result that characterizes the guarantees for EM under the univariate setting. Let I'_ε denote the interval $[cn^{-\frac{1}{12}+\varepsilon}, 1/10]$ where c is a positive universal constant.

Theorem 8.1. *Fix $\delta \in (0,1)$, $\varepsilon \in (0,1/8]$, and let $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ for $i = 1,\ldots,n$ such that $n \gtrsim \log\frac{\log(1/\varepsilon)}{\delta}$. Then for any initialization μ_n^0 that satisfies $|\mu_n^0| \in I'_\varepsilon$, the sample EM sequence $\mu_n^{t+1} = M_{n,1}(\mu_n^t)$, satisfies*

$$|\mu_n^t - \theta^*| \leq c_1 \frac{1}{n^{1/8-\varepsilon}} \log^{5/4}\left(\frac{10n\log(8/\varepsilon)}{\delta}\right), \tag{8.5}$$

for all $t \geq c_2 n^{\frac{3}{4}-6\varepsilon} \cdot \log n \log\frac{1}{\varepsilon}$ with probability at least $1-\delta$.

See Appendix G.2 for the proof.

The bound (8.5) shows that with high probability after $\mathcal{O}(n^{3/4})$ steps the sample EM iterates converge to a ball around θ^* whose radius is arbitrarily close to $n^{-1/8}$. Moreover, as a direct consequence of the relation (8.3a), we conclude that the EM estimate for the scale parameter is of order $n^{-\frac{1}{4}}$ with high probability:

$$\begin{aligned}\left|(\sigma_n^t)^2 - (\sigma^*)^2\right| &= \left|\frac{\sum_{i=1}^{n} X_i^2}{n} - \left(\theta_n^t - \theta^*\right)^2 - (\sigma^*)^2\right| \\ &\lesssim n^{-\frac{1}{2}} + n^{-\frac{1}{4}} = \mathcal{O}(n^{-\frac{1}{4}})\end{aligned} \tag{8.6}$$

where we have used the standard chi-squared concentration for the sum $\sum_{i=1}^{n} X_i^2/n$.

Matching lower bound: In Appendix G.1, we prove a matching lower bound and thereby conclude that the upper bound of order $n^{-\frac{1}{8}}$ for the statistical error of sample EM from Theorem 8.1 is tight up to constant factors. In Section 8.3.3, we provide further evidence (cf. Figure 8.3) that the slow statistical rates of EM with location parameter that we derived in Theorem 8.1 might appear in more general settings of location-scale Gaussian mixtures as well.

8.3.2 Results for the multivariate case

Analyzing the general EM updates for higher dimensions turns out to be challenging. However, for the symmetric fit in higher dimensions given by

$$\mathcal{G}_{\text{symm}}((\theta,\sigma^2))=\frac{1}{2}\mathcal{N}(\theta,\sigma^2 I_d)+\frac{1}{2}\mathcal{N}(-\theta,\sigma^2 I_d), \tag{8.7}$$

the sample EM operator $M_{n,d}(\theta)$ has a closed form as already noted in the updates (8.3a) and (8.3b). Note that for the fit (8.7), we have assumed the same scale parameter for all dimensions. Such a fit is over-specified for data drawn from Gaussian distribution $\mathcal{N}(0, I_d)$. We now show that the sharing of scale parameter in the model fit across dimensions (8.7), leads to a faster convergence of EM in $d \geq 2$—both in terms of number of steps and the final statistical accuracy. In the following result, we denote $I_\varepsilon := [5\left(\frac{d}{n}\right)^{\frac{1}{4}+\varepsilon}, \frac{1}{8}]$.

Theorem 8.2. *Fix $\delta \in (0,1)$, $\varepsilon \in (0, 1/4]$, and let $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$ for $i = 1, \ldots, n$ such that $d \geq 2$ and $n \gtrsim d \log^{\frac{1}{4\varepsilon}}(\log \frac{1/\varepsilon}{\delta})$. Then with any starting point θ_n^0 such that $\|\theta_n^0\|_2 \in I_\varepsilon$, the sample EM sequence $\theta_n^{t+1} = M_{n,d}(\theta_n^t)$ satisfies*

$$\left\|\theta_n^t - \theta^*\right\|_2 \leq c_1 \left(\frac{d}{n} \log \frac{\log(1/\varepsilon)}{\delta}\right)^{\frac{1}{4}-\varepsilon}, \tag{8.8}$$

for all $t \geq c_2 \left(\frac{n}{d}\right)^{\frac{1}{2}-2\varepsilon} \log \frac{n}{d} \log \frac{1}{\varepsilon}$ with probability at least $1-\delta$.

See Appendix 8.4.2 for the proof.

The results in Theorem 8.2 show that the that the sample EM updates converge to a ball around $\theta^* = 0$ with radius arbitrarily close to $(d/n)^{\frac{1}{4}}$ when $d \geq 2$. At first sight, the initialization condition $\|\theta_n^0\|_2 \leq 1/8$, assumed in Theorem 8.2, might seem pretty restrictive but Lemma G.4 (in Appendix G.6) shows that for any θ_n^0 satisfying $\|\theta_n^0\|_2 \leq \sqrt{d}$, we have $\widetilde{M}_{n,d}(\theta_n^0) \leq \sqrt{2/\pi}$, with high probability. In light of this result, we may conclude that the initialization condition is Theorem 8.2 is not overly restrictive.

Guarantees for the scale parameter σ_n^t: Noting that $(\theta^*, \sigma^*) = (0, 1)$, we obtain the following relation

$$\left|(\sigma_n^t)^2 - (\sigma^*)^2\right| = \left|\frac{\sum_{i=1}^n \|X_i\|_2^2}{dn} - (\sigma^*)^2 - \frac{\|\theta_n^t - \theta^*\|_2^2}{d}\right|.$$

Using standard chi-squared bounds, we obtain that

$$\left|\frac{\sum_{i=1}^{n}\|X_i\|_2^2}{dn}-(\sigma^*)^2\right|\precsim(nd)^{-\frac{1}{2}},$$

with high probability. From the bound (8.8), we also have $\|\theta_n^t-\theta^*\|_2^2/d\precsim(nd)^{-\frac{1}{2}}$. Putting the pieces together, we conclude that the statistical error for the scale parameter satisfies

$$|(\sigma_n^t)^2-(\sigma^*)^2|\precsim(nd)^{-\frac{1}{2}}\quad\text{for all } t\succsim\left(\tfrac{n}{d}\right)^{\frac{1}{2}}, \tag{8.9}$$

with high probability. Consequently, in the sequel, we focus primarily on the convergence rate for the EM estimates θ_n^t of the location parameter, as the corresponding guarantee for the scale parameter σ_n^t is readily implied by it.

Comparison with Theorem 8.1: The scaling of order $n^{-\frac{1}{4}}$ with n from equation (8.8) is significantly better than the univariate case ($n^{-\frac{1}{8}}$) stated in Theorem 8.1. We note that this faster statistical rate is a consequence of the sharing of the scale parameter across dimensions, and does not hold when the fit (8.7) has different variance parameters. Indeed, as we demonstrated in Figure 8.1, when the fitted components have freely varying scale parameter, the statistical rate slows down (and can be of the order $n^{-\frac{1}{8}}$ in higher dimensions).

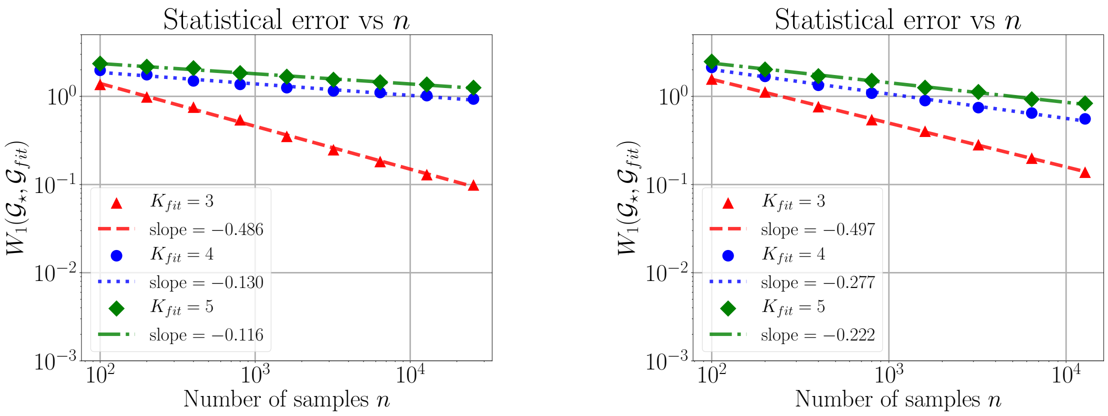
8.3.3 Simulations with general cases

We now present preliminary evidence that the slow statistical rates of EM with location parameter that we derived in Theorem 8.1 might appear in more general settings. In Figure 8.3, we plot the statistical error of estimates returned by sample EM when estimating *all* the parameters (namely weights, location and scale) simultaneously, as a function of sample size n, for the following two cases:

$$\mathcal{G}_\star^{d=1}=\frac{1}{6}\mathcal{N}(-5,1)+\frac{1}{2}\mathcal{N}(1,3)+\frac{1}{3}\mathcal{N}(7,2); \tag{8.10}$$

$$\mathcal{G}_\star^{d=2}=\frac{1}{2}\mathcal{N}(\begin{bmatrix}0\\0\end{bmatrix},I)+\frac{1}{6}\mathcal{N}(\begin{bmatrix}7\\5\end{bmatrix},2I)\frac{1}{3}\mathcal{N}(\begin{bmatrix}-4\\-7\end{bmatrix},3I). \tag{8.11}$$

We plot the results for a $K_{\text{fit}}\in\{3,4,5\}$-mixture Gaussian model fit. When K_{fit} is equal to the number of components ($=3$) in the true mixture the statistical rate is $n^{-1/2}$. When it is larger, i.e., $K_{\text{fit}}\in\{4,5\}$, the statistical rate of EM is much larger, $n^{-0.12}$ in panel (a) (for $K_{\text{fit}}=5$) and $n^{-0.20}$ in panel (b) (for $K_{\text{fit}}=5$) of Figure 8.3. These simulations suggest that the statistical rates slower than $n^{-\frac{1}{4}}$ and of order $n^{-\frac{1}{8}}$ may arise in more general settings, and moreover that the rates get slower as the over-specification of the number of mixtures increases. See Section 8.5 for possible future work in this direction.



(a) $d = 1$ with $\mathcal{G}_*$ given by equation (8.10) (a) $d = 2$ with $\mathcal{G}_*$ given by equation (8.11)

Figure 8.3: Scaling of the first-order Wasserstein error for EM estimates when fitting a Gaussian mixture with $K_{\text{fit}} \in \{3, 4, 5\}$, i.e., $\mathcal{G}_{\text{fit}} = \sum_{i=1}^{K_{\text{fit}}} w_i \mathcal{N}(\mu_i, \Sigma_i)$, on n i.i.d. samples from a 3-Gaussian mixture model (equations (8.10) and (8.11)). In the case of no over-specification, i.e., $K_{\text{fit}} = K_{\text{true}} = 3$, the error scales as $n^{-1/2}$, but when the fitted model is over-specified ($K_{\text{fit}} \in \{4, 5\}$), the scaling is much worse (and degrades further for any given n as K_{fit} gets large). See Section 8.3.3 for further details.

8.4 Two-staged argument for sharp analysis of EM

Deriving a sharp rate for univariate case (Theorem 8.1) turns out be pretty challenging and requires a thorough discussion. On the other hand, the multivariate-case considered in this chapter (Theorem 8.2) is *relatively* easy due to the shared scale parameter given the localization argument developed in Chapter 7. The proof of Theorem 8.1 is fairly technical and is thereby deferred to Appendix G.2. We collect a high-level overview of the novel two-staged localization argument in Section 8.4.1 required to establish the sharp guarantees of Theorem 8.1. We provide the proof of Theorem 8.2 in Section 8.4.2. (One may want to read the next two subsections in reverse order for a better understanding.)

8.4.1 Proof sketch for Theorem 8.1

Our proof makes use of the population-to-sample analysis framework of Balakrishnan et al. [10] albeit with several new ideas, with the highlight being a two-staged analysis with two different population-level EM operators.

Let $Y \sim \mathcal{N}(0, 1)$, then the population-level analog of the operator (8.3b) can be defined in two ways:

$$\widetilde{M}_{n,1}(\theta) := \mathbb{E}_Y \left[Y \tanh\left(\frac{Y\theta}{\sum_{j=1}^n X_j^2/n - \theta^2} \right) \right], \tag{8.12a}$$

$$\overline{M}_1(\theta) := \mathbb{E}_Y \left[Y \tanh\left(\frac{Y\theta}{1-\theta} \right) \right]. \tag{8.12b}$$

The particular choice of the population-like operator $\widetilde{M}_{n,1}$ in equation (8.12a) was motivated by the previous works [35] with the location-scale Gaussian mixtures. We refer to this operator as the *pseudo-population operator* since it depends on the samples $\{X_i, i = 1, \ldots n\}$ and involves an expectation. Nonetheless, as we show in the sequel, analyzing $\widetilde{M}_{n,1}$ is not enough to derive sharp rates for sample EM in the over-specified setting considered in Theorem 8.1. A careful inspection reveals that a "better" choice of the population operator is required, which leads us to define the operator $\overline{M}_1$ in equation (8.12b). Unlike the pseudo-population operator $\widetilde{M}_{n,1}$, the operator $\overline{M}_1$ is indeed a population operator as it does not depend on samples $X_1, \ldots, X_n$. Note that, this operator is obtained when we replace the sum $\sum_{j=1}^n {X_j}^2/n$ in the definition (8.12a) of the operator $\widetilde{M}_{n,1}$ by its corresponding expectation $\mathbb{E}[\|X\|_2^2] = 1$. For this reason, we also refer to this operator $\overline{M}_1$ as the *corrected population operator*. In the next lemma, we state the properties of the operators defined above (here I'_ε denotes the interval $[cn^{-\frac{1}{12}+\varepsilon}, 1/10]$).

Lemma 8.1. *The operators $\widetilde{M}_{n,1}$ and $\overline{M}_1$ satisfy*

$$\left(1 - \frac{3\theta^6}{2}\right)|\theta| \leq \left|\widetilde{M}_{n,1}(\theta)\right| \leq \left(1 - \frac{\theta^6}{5}\right)|\theta|, \tag{8.13a}$$

$$\left(1 - \frac{\theta^6}{2}\right)|\theta| \leq \left|\overline{M}_1(\theta)\right| \leq \left(1 - \frac{\theta^6}{5}\right)|\theta|, \tag{8.13b}$$

where bound (8.13a) holds for all $|\theta| \in I'_\varepsilon$ with high probability[3] *and the bound (8.13b) is deterministic and holds for all $|\theta| \in \left[0, \frac{3}{20}\right]$. Furthermore, for any fixed $\delta \in (0,1)$ and any fixed $r \geq \mathcal{O}(n^{-\frac{1}{12}})$, we have that*

$$\mathbb{P}\left[\sup_{\theta \in \mathbb{B}(0,r)} \left| M_{n,1}(\theta) - \widetilde{M}_{n,1} \right| \leq cr\sqrt{\frac{\log(1/\delta)}{n}} \right] \geq 1 - \delta. \tag{8.13c}$$

On the other hand, for any fixed $r \leq \mathcal{O}(n^{-\frac{1}{16}})$, we have

$$\mathbb{P}\left[\sup_{\theta \in \mathbb{B}(0,r)} \left| M_{n,1}(\theta) - \overline{M}_1(\theta) \right| \leq c' r^3 \sqrt{\frac{\log^{10}(5n/\delta)}{n}} \right] \geq 1 - \delta. \tag{8.13d}$$

[3] Since the operator $\widetilde{M}_{n,1}$ depends on the samples $\{X_j, j \in [n]\}$, only a high probability bound (and not a deterministic one) is possible.

See Appendix G.3 for its proof where we also numerically verify the sharpness of the results above (see Figure G.1). Lemma 8.1 establishes that, as $\theta \to 0$, both the operators have similar contraction coefficient $\gamma(\theta) \asymp 1 - c\theta^6$; thereby justifying the rates observed for $d = 1$ in Figure 8.2(b). However, their perturbation bounds are significantly different: while the error $\sup_{\theta \in \mathbb{B}(0,r)} \left| M_{n,1}(\theta) - \widetilde{M}_{n,1}(\theta) \right|$ scales linearly with the radius r, the deviation error $\sup_{\theta \in \mathbb{B}(0,r)} \left| M_{n,1}(\theta) - \overline{M}_1(\theta) \right|$ has a cubic scaling r^3.

Remark: A notable difference between the two bounds (8.13c) and (8.13d) is the range of radius r over which we *prove* the validity of the bounds (8.13c) and (8.13d). With our tools, we establish that the perturbation bound (8.13c) for the operator $\widetilde{M}_{n,1}$ is valid for any $r \gtrsim n^{-\frac{1}{12}}$. On the other hand, the corresponding bound (8.13d) for the operator $\overline{M}_1$ is valid for any $r \precsim n^{-\frac{1}{16}}$. We now elaborate why these different ranges of radii are helpful and make both the operators crucial to in the analysis to follow.

8.4.1.1 A sub-optimal analysis

Using the properties of the operator $\widetilde{M}_{n,1}$ from Lemma 8.1, we now sketch the statistical rates for the sample EM sequence, $\theta_n^{t+1} = M_{n,1}(\theta_n^t)$, that can be obtained using (a) the generic procedure outlined by Balakrishnan et al. [10] and (b) the localization argument introduced in Chapter 7. As we show, both these arguments end up being *sub-optimal* as they do not provide us the rate of order $n^{-\frac{1}{8}}$ stated in Theorem 8.1. We use the notation:

$$\sup_{|\theta| \geq \epsilon} \left| \widetilde{M}_{n,1}(\theta) \right| / |\theta| \precsim 1 - \epsilon^6 =: \gamma(\epsilon).$$

Sub-optimal rate I: The eventual radius of convergence obtained using Theorem 5(a) from the paper [10] can be determined by (see (7.21))

$$\frac{r/\sqrt{n}}{1 - \gamma(\epsilon)} = \epsilon \quad \Longrightarrow \quad \epsilon \sim n^{-1/14}, \tag{8.14a}$$

where r denotes the bound on the initialization radius $|\theta^0|$ but we have tracked dependency only on n. This informal computation suggests that the the sample EM iterates for location parameter are bounded by a term of order $n^{-1/14}$. This rate is clearly sub-optimal when compared to the EM rate of order $n^{-\frac{1}{8}}$ from Theorem 8.1.

Sub-optimal rate II: Next we apply the more sophisticated localization argument from Section 7.5.2 in order to obtain a sharper rate. In contrast to the computation (8.14a), this argument leads to solving the equation

$$\frac{\epsilon \cdot r/\sqrt{n}}{1 - \gamma(\epsilon)} = \epsilon \implies \frac{\epsilon r/\sqrt{n}}{\epsilon^6} = \epsilon \implies \epsilon \sim n^{-\frac{1}{12}}, \tag{8.14b}$$

where, as before, we have only tracked dependency on n. This calculation allows us to conclude that the EM algorithm converges to an estimate which is at a distance of order $n^{-\frac{1}{12}}$ from the true parameter, which is again sub-optimal compared to the $n^{-\frac{1}{8}}$ rate of EM from Theorem 8.1.

Indeed both the conclusions above can be made rigorous (See Corollary G.1 for a formal statement) to conclude that, with high probability for any $\varepsilon \in (0, \frac{1}{12}]$:

$$\left|\mu_n^t - \theta^*\right| \leq \mathcal{O}(n^{-\frac{1}{12}+\varepsilon}) \text{ for } t \geq \mathcal{O}(n^{\frac{1}{2}-6\varepsilon}). \tag{8.15}$$

8.4.1.2 A two-staged analysis for sharp rates

In lieu of the above observations, the proof of the sharp upper bound (8.5) in Theorem 8.1 proceeds in two stages. In the first stage, invoking Corollary G.1 with $\varepsilon = \frac{1}{48}$, we conclude that with high probability the sample EM iterates converge to a ball of radius at most r after $\sqrt{n}$ steps, where $r \ll n^{-1/16}$. Consequently, the sample EM iterates after $\sqrt{n}$ steps satisfy the assumptions required to invoke the perturbation bounds for the operator $\overline{M}_1$ from Lemma 8.1. Thereby, in the second stage of the proof, we apply the $1 - c\theta^6$ contraction bound (8.13b) of the operator $\overline{M}_1$ in conjunction with the cubic perturbation bound (8.13d). Using localization argument for this stage, we establish that the EM iterates obtain a statistical error of order $n^{-1/8}$ in $\mathcal{O}\left(n^{3/4}\right)$ steps as stated in Theorem 8.1. See Appendix G.2 for a detailed proof.

8.4.2 Proof of Theorem 8.2

This proof is based on the population-to-sample analysis and follows a similar road-map as of the proof of Theorem 7.3.

We first analyze the population-level EM operator and then using epoch-based-localization argument derive the statistical rates (8.8). We make use of the following d-dimensional analog of the pseudo-population operator (cf. equation (8.12a)):

$$\widetilde{M}_{n,d}(\theta) := \mathbb{E}_{Y\sim\mathcal{N}(0,I_d)}\left[Y \tanh\left(\frac{Y^\top\theta}{\sum_{j=1}^n \left\|X_j\right\|_2^2/(nd) - \|\theta\|^2/d}\right)\right]. \tag{8.16}$$

The next lemma establishes the contraction properties and the perturbation bounds for $\widetilde{M}_{n,d}$.

Lemma 8.2. *The operator $\widetilde{M}_{n,d}$ satisfies*

$$\left(1 - \frac{3\left\|\theta\right\|_2^2}{4}\right) \leq \frac{\left\|\widetilde{M}_{n,d}(\theta)\right\|_2}{\left\|\theta\right\|_2} \leq \left(1 - \frac{(1-1/d)\left\|\theta\right\|_2^2}{4}\right), \quad \textit{for all } \left\|\theta\right\|_2 \in I_\varepsilon, \tag{8.17a}$$

with probability at least $1-\delta$. *Moreover, there exists a universal constant* c' *such that for any fixed* $\delta \in (0,1)$, $\varepsilon \in (0, \frac{1}{4}]$, *and* $r \in (0, \frac{1}{8})$ *we have*

$$\mathbb{P}\left[\sup_{\theta \in \mathbb{B}(0,r)} \left\| M_{n,d}(\theta) - \widetilde{M}_{n,d}(\theta) \right\|_2 \leq c' r \sqrt{\frac{d \log(1/\delta)}{n}}\right] \geq 1 - \delta - e^{-(nd)^{4\varepsilon}/8}. \tag{8.17b}$$

See Appendix G.4 for the proof.

Lemma 8.2 shows that the operator $\widetilde{M}_{n,d}$ has a faster contraction (order $1-\|\theta\|_2^2$) towards zero, when compared to its univariate-version (order $1-\theta^6$ cf. (8.13a)). This difference between the univariate and the multivariate case had already been highlighted in Section 8.2 in Figure 8.2. Indeed substituting $d=1$ in the bound (8.17a) gives us a vacuous bound for the univariate case, providing further evidence for the benefit of sharing variance among different dimensions in multivariate setting of symmetric fit (8.1). With Lemma 8.2 at hand, the proof of Theorem 8.2 follows by using the localization argument from Section 7.5.2. Mimicking the arguments similar to equation (8.14b), we obtain the following statistical rate:[4]

$$\frac{\epsilon \cdot r/\sqrt{n}}{1-\gamma(\epsilon)} = \epsilon \quad \Longrightarrow \frac{\epsilon r/\sqrt{n}}{\epsilon^2} = \epsilon \quad \Longrightarrow \quad \epsilon \sim n^{-\frac{1}{4}}. \tag{8.18}$$

Much of the work in the proof of Theorem 8.2 is to establish Lemma 8.2. With the bounds (8.17a) and (8.17b) at hand, using the localization argument (in a manner similar to the proof of Theorem 7.3), easily leads to the statistical rate of order $(d/n)^{1/4}$ as claimed in Theorem 8.2. The detailed proof is thereby omitted.

8.5 Conclusion and future directions

In this chapter, we established several results characterizing the convergence behavior of EM algorithm for over-specified location-scale Gaussian mixtures. We view our analysis of EM for the symmetric singular Gaussian mixtures as the first step toward a rigorous understanding of EM for a broader class of weakly identifiable mixture models. Such a study would provide a better understanding of the singular models with weak identifiability which do arise in practice since: (a) over-specification is a common phenomenon in fitting mixture models due to weak separation between mixture components, and, (b) the parameters being estimated are often inherently dependent due to the algebraic structures of the class of kernel densities being fitted and the associated partial derivatives. We now discuss a few other directions that can serve as a natural follow-up of our work.

[4] Moreover, similar to the arguments made in Section 7.5, the localization argument is necessary to derive a sharp rate. Indeed, a direct application of the framework introduced by Balakrishnan et al. [10] for our setting implies a sub-optimal rate of order $(d/n)^{1/6}$ for the Euclidean error $\|\theta_n^t - \theta^*\|$ (cf. (8.14a) and (8.14b)).

The slow rate of order $n^{-\frac{1}{8}}$ for EM updates with location parameter is in a sense a worst-case guarantee. In the univariate case, for the entire class of two mixture Gaussian fits, MLE exhibits the slowest known statistical rate $n^{-\frac{1}{8}}$ for the settings that we analyzed. More precisely, for certain asymmetric Gaussian mixture fits, the MLE convergence rate for the location parameter is faster than that of the symmetric equal-weighted mixture considered in this chapter E.g., for the fit $1/3\mathcal{N}(-2\theta, \sigma^2) + 2/3\mathcal{N}(\theta, \sigma^2)$ on $\mathcal{N}(0, 1)$ data, the MLE converges at the rate $n^{-1/6}$ and $n^{-1/3}$ respectively [113]. It is interesting to understand the effect of such a geometric structure of the global maxima on the convergence of the EM algorithm.

Our work analyzed over-specified mixtures with a specific structure and only one extra component. As demonstrated above, the statistical rates for EM appear to be slow for general covariance fits and further appear to slow down as the number of over-specified components increases. The convergence rate of the MLE for such over-specified models is known to further deteriorate as a function of the number of extra components. It remains to understand how the EM algorithm responds to these more severe—and practically relevant—instances of over-specification.

Part III

Data-Driven Methodologies For Causal Inference

Chapter 9

Discovery of Interpretable and Stable Subgroups

Since its inception, the field of statistics has aimed to produce tools to help scientists seek scientific truth. Scientific truth, however, is not of a singular quality. While some relations in physics like Hooke's law are made apparent using simple linear regression, questions dealing with complex, emergent phenomena such as the efficacy of drugs or job training programs seem to have more contingent answers. It was the urge to formalize and investigate such questions that begot and nurtured the field of causal inference in statistics over the past century. One of the two most influential frameworks for causal inference, the Neyman-Rubin causal model [116], has its roots in Fisher and Neyman's [90, 229, 193] work on randomized experiments for agriculture, and was later codified by Rubin [223], who was then interested in psychometrics.[1]

9.1 Introduction

Historically, causal inference researchers have used traditional regression methods in their analyses, with econometricians in particular developing a comprehensive theory of drawing inference from linear models [6]. This is rapidly changing, however, with recent works [8, 147, 53, 182] bringing in machine learning tools to tackle causal inference problems, one genre of which has been the investigation of heterogeneous treatment effects.

9.1.1 Heterogeneous treatment effects

In both randomized experiments as well as observational studies, apart from the treatment and response variables, additional pre-treatment information is often known about the study subjects. For instance, information on medical risk factors is collected in clinical trials, while

[1] With important extensions also by Cox [59].

demographic and socioeconomic data is collected in social science studies. Such side information has always been important because it allows us to adjust for confounding in observational studies, and also to create more efficient estimators in randomized experiments [160, 124]. In addition to these uses, researchers are also increasingly interested in drawing inference about how the effect of a treatment varies depending on an individual's observed covariates.

Broadly speaking, methodological research on heterogeneous treatment effects can be put into two categories: (i) conditional average treatment effect (CATE) function estimation [125, 96, 89, 34, 92, 234, 21], and (ii) subgroup analysis, [247, 202, 9, 161] with the latter having a longer history. Here we attempt a brief review of the existing literature, and refer the readers to referenced papers for further background.

CATE Estimation: For a binary treatment, the CATE is defined to be the expected difference between the potential outcome under treatment and that under no treatment, conditional on a subject's observed covariates (see Section 9.3 for formal definitions). While the average treatment effect (ATE) is a scalar quantity, the CATE is a function and thus far more challenging to estimate. Because one observes only one of the two potential outcomes for every individual—an issue referred to as the fundamental problem of missing data in causal inference [116]—one cannot directly solve this problem using the conventional supervised learning techniques.

Over the past decade or so, researchers have made tremendous progress with CATE estimation and proposed numerous methods for it [125, 96, 89, 34, 92, 234, 21]. A large fraction of these [125, 89, 34, 21] fall under the framework of meta-learners. These are "meta-algorithms [that] decompose estimating the CATE into several regression sub-problems that can be solved with any regression or supervised learning method" [147]. Some of these meta-algorithms are fairly obvious. For instance, the T-learner strategy [92] comprises fitting models for the two response functions (the conditional expectation of each potential outcome), and then taking their difference. Others, such as the X-learner [147] and R-learner [196] strategies, are more sophisticated, and require more notation to explain (see Section 9.4.1 for further details). Not all proposed algorithms follow a meta-learner strategy, the popular causal tree and causal forest algorithms [9, 245] being prominent examples.

Concerns with model choice for CATE Estimation: With such a diverse range of estimators, most of which come with hyperparameters, model choice becomes a primary concern. Some researchers have used asymptotic efficiency [196, 141] to establish when certain estimators can be definitely favored under (uncheckable) generative models. Such arguments, however, rely on smoothness assumptions and asymptotic data regimes that are typically hard to verify for the problems typically considered by causal inference researchers. Meanwhile, plug-in prediction accuracy on holdout test sets is frequently used to do model selection in supervised learning, but this is infeasible for CATE estimation due to the data missingness we alluded to earlier. To circumvent this issue, researchers have formulated proxy loss functions [225] for data-driven model choice, with ideas including using nearest neighbor matching [218], kernel-based local linear squares fit [34], and influence functions [2]. These model choice methods, however, have only been justified using simulations often in strong signal regime, a scenario that does not hold in many if not most real data problems (including the one considered in this work).

Concerns with model validation for CATE Estimation: Before deciding which estimator to choose for a given task, we would first like to know whether there is even enough signal in the data to fit a generalizable model. Again, data missingness means that there is no clear answer to this problem. The proxy loss functions are not good substitutes for quantities like R^2 or ROC AUC scores because they can be noisy, and furthermore they do not have an easily interpretable scale. This is especially concerning because randomized experiments often have low signal strength.[2]

Subgroup analysis: An older approach to investigating heterogeneity is through "subgroup analysis". The goal here is to identify subgroups of subjects in the study over which the treatment effect is significantly larger or smaller than that the population average. Such a conception of heterogeneity has two advantages over CATE estimation: (a) It is less ambitious, and thus promises to be more tractable given the low data regime in real settings, and (b) it is often more aligned with the downstream tasks involving decision-making (e.g., identifying which subgroup of individuals to treat).

Traditionally, for subgroup analysis, researchers check the treatment effect over a predetermined list of subgroups which are suggested by prior domain knowledge. Doing this, however, ignores potential unforeseen heterogeneity in the data, and there has been much recent work on how to conduct a data-driven search for subgroups. Naive searching can quickly overfit[3], so any search method has to balance aggressiveness of searching with the need to account for multiple testing. Proposed methods include using recursive partitioning [232, 9], Cox modeling [191], controlled partitioning with significance checks using data splits [161], and several variants [77, 11]. Unfortunately, systematic analyses of these methods have usually provided unsatisfactory results in real datasettings and in low-signal simulations [199, 122]. We refer the readers to the book [36] (Chapter 8), and the review papers [199, 122] for further discussion on these methods.

Finally, we note that some researchers have proposed using CATE estimation as a stepping stone to finding subgroups. Such a strategy was proposed by Foster et al. [92] with their Virtual Twins method, namely the T-learner with random forests, while Chernozhukov et al. [53] recapitulate this idea in the context of a broader call to perform inference on features of the CATE function rather than the function itself. In another line of work, Shahn et al. [226] integrate (linear) CATE modeling with latent class mixture modeling in a Bayesian

[2] Budget constraints would dictate that they be only sufficiently powered to detect the ATE.

[3] More importantly, investigating subgroups in this manner is particularly sensitive to human failures. It opens the door to p-value hacking [261], while Gelman has argued that even when researchers try to be honest, they nonetheless have a hard time accounting for "researcher degrees of freedom" [94].

framework to allow for treatment effect heterogeneity in discrete levels. They then use the feature importance from the latent (logistic) model and the posteriors for the CATE, to estimate qualitatively, subgroups with large treatment effect.

9.1.2 The PCS framework for veridical data science

As argued in the previous section, obtaining reliable conclusions with respect to heterogeneous treatment effects is fraught with difficulty. On the one hand, poor signal and weak priors are prevalent, and on the other hand, missing potential outcomes means that test-set validation is not directly feasible. Methods validated on simulation studies may not work well for real data problems since their performance are often misleading. Furthermore, empirical evidence tells us that the relative and absolute performance of estimation algorithms is highly data and context-dependent [198].[4] Given these problems, it is puzzling to see that much new methodology is being developed that is detached from solving real data problems.

In this chapter, we re-analyzed the 1999-2000 VIGOR study (a 8076 patient randomized clinical trial), and had to face precisely these challenges. To overcome them, we take advantage of the recent works on CATE estimation [21, 147, 9, 196, 245] and build on the PCS framework for veridical data science recently introduced by Yu and Kumbier [259]. As a result, we develop a methodology called Stable Discovery of Interpretable Subgroups via Calibration (StaDISC) that is generally applicable beyond this dataset. We now briefly review the PCS framework, before turning to the overview of our contributions and StaDISC in **??**.

The PCS framework bridges, unifies, and expands on ideas from machine learning and statistics for the entire data science life cycle. The letters in PCS stand for the three core principles of data science, namely Predictability, Computability, and Stability. In a nutshell, the PCS framework advocates using both predictability and stability analysis, argued and documented in a PCS documentation, for reliable and reproducible scientific investigations, thereby providing a way for bridging Breiman's Two Cultures [27]. More specifically, predictability emphasizes reality checks for the modeling stage, by integrating the use of data-driven validation such as out-of-sample testing favored by machine learning, and that of goodness-of-fit measures that have a rich history in traditional statistics. Stability, besides encompassing sampling variability, expands to other stability or robustness concerns of the contingency of modeling conclusions to researcher "judgment calls". These calls include the choices made by the researcher at various stages of the data science life cycle, including data cleaning in addition to the modeling decisions such as model choices and data perturbations. Computability reflects the need to keep computational feasibility and efficiency in mind when constructing any modern data analysis pipeline, especially those that subscribe to the first two principles, which are usually more demanding computationally.

[4] In fact, different methods and research groups sometimes reach different conclusions on the same datasets, see the paper [38] and the references therein.

The PCS framework addresses to a certain extent Professor Efron's concern [87] that machine learning methods (or pure prediction algorithms) are not ready to be used on scientific problems.[5] The PCS framework adds a paramount consideration of stability to predictability and computability that are hallmarks of machine learning. It guides researchers in validating machine learning and statistical methods with respect to the specific task they are to be applied and extracting data conclusions that can be relied upon. As one of us has previously discussed [258], even though 100% truth is beyond reach, a useful goal is an "accurate approximation for a particular domain, and relative to a particular performance metric," which is a more precise articulation of George Box's belief that "all models are wrong, but some are useful."

9.1.3 Our contributions

This chapter makes three main contributions. First, we seek subgroups with demonstrable heterogeneous treatment effects in the dataset from the 1999-2000 VIGOR study. Complementary analyses with the 2001-2004 APPPROVe study provides additional evidence for the heterogeneity in treatment effect for the found subgroups. Enroute, building on the recent CATE literature and the PCS framework, we develop a new methodology, which we call Stable Discovery of Interpretable Subgroups via Calibration (StaDISC). We provide an overview of this methodology toward the end of this section. Finally, this work also serves as the first articulation of the PCS framework in the context of causal inference, with StaDISC providing a template for more informative understanding of heterogeneous outcomes.

9.1.4 Overview of StaDISC

First of all, a given data set (deemed approximately iid) is divided into a holdout test set $\mathbf{S}_{\text{TEST}}$ and a training set $\mathbf{S}_{\text{TRAIN}}$ (per outcome). For hyperparameter tuning, we use 4-fold cross validation with the training data $\mathbf{S}_{\text{TRAIN}}$.[6] For any set of training folds, we refer to the leftout fold as the corresponding validation fold. The test set is used only once at the final step of checking the significance of the interpretable subgroups found by our methodology. See Section 9.2.3 for more details on data splitting and Section 9.4.1 for the fitting of CATE estimators. With this set-up at hand, StaDISC can be summarized in three steps: a predictive reality check in Section 9.4 based on calibration, stability-driven ranking and aggregation of CATE estimators in Section 9.5, and finally the `CellSearch` procedure for finding interpretable subgroups in Section 9.6. In Section 9.4, we introduce a novel calibration-based

[5] In Professor Efron's timely and thought-provoking revisiting [87] of the *Two Cultures* debate [27], it is argued that contrasting philosophies on scientific truth is a clear line that separates traditional regression methods from modern machine learning methods (or pure prediction algorithms). While the former aims at an eternal scientific truth, the latter is truth-agnostic and instead content to exploit contingent and ephemeral patterns.

[6] Due to the low signal in data, we decided not to split the data into training and validation sets, and instead use 4-fold cross validation on the training data.

pseudo-R^2 score for CATE estimators denoted by $\mathcal{R}^2_{\mathrm{C}}$, which involves placing individuals (in both training and validation folds) into equally-sized bins based on their predicted CATE value, with quantiles of the predicted CATE distribution on the training folds as thresholds for the CATE estimators. Using such a binning and the $\mathcal{R}^2_{\mathrm{C}}$-scores, we show that 18 popular CATE estimators generalize poorly for the VIGOR data on the validation folds of the training data. However, we find that certain quantile-based bins (referred to as quantile-based top subgroups) do generalize well in the sense of having significantly stronger subgroup CATE on both training and validation folds. This provides the starting point of the next step. In Section 9.5, we use the t-statistics of the treatment effect over the quantile-based top subgroups and its stability over 7 different appropriate data perturbations to rank, screen, and finally average the screened CATE estimators (the ensemble CATE estimator). Section 9.6 details the last step of StaDISC, where we introduce the `CellSearch` procedure to find a stable and interpretable representation of the quantile-based top subgroup of the ensemble from the previous step, and then check its performance on the holdout test set (which was used only for final testing).

As a final overview remark, we note that we use poor performance and good/bad generalization in a slightly loose sense throughout the chapter. We only use the holdout test set at the final stage, for verifying the CATE estimates of discovered subgroups. Nonetheless, we use the phrase *poor generalization* to refer to worse-than-expected-performance, where the performance metric varies across results, on the validation folds.

9.1.5 Organization

The rest of the chapter is organized as follows. In Section 9.2, we start with a brief history of the VIGOR study, and then describe the dataset and data engineering, and splitting done by us. Section 9.3 reviews the Neymann-Rubin model briefly with basic notations introduced. The development of the StaDISC methodology (overviewed below) is carried out in Sections 9.4 to 9.6 with the final subgroups reported in Section 9.6.3. Results for the complementary analyses of the found subgroups with the APPROVe study are presented in Section 9.7. We conclude in Section 9.8 with a recap of our results, a discussion of the relevance of our discoveries in medicine, and discuss several directions for future work with StaDISC. Most of the figures and tables are deferred to the appendix. Moreover, in accordance with the PCS framework's requirement for clear and careful documentation, we provide our code, data cleaning, and statistical analyses in the form of Jupyter notebooks on GitHub (`https://github.com/Yu-Group/stadisc`).

9.2 Dataset from the VIGOR study

In this chapter, we are interested in finding subgroups of patients that benefit from the treatment in the dataset from the Vioxx gastro-intestinal outcomes research (VIGOR) study [23]. In the process of seeking such subgroups, we develop the new StaDISC methodology. In this section, we provide an overview of this study and the dataset, and also explain our data pre-processing and feature engineering.

9.2.1 VIGOR study history and description

The VIGOR study was a randomized head-to-head trial comparing two drugs used to alleviate pain and inflammation for patients with rheumatoid arthritis: a "new" cyclooxygenase-2 (COX-2) inhibitor drug Rofecoxib (Vioxx) recently approved and developed by Merck, and Naproxen, a standard nonsteroidal anti-inflammatory drug (NSAID) already in routine clinical use for many years. NSAIDs, though effective for treating pain and inflammation, cause serious gastrointestinal side effects in a small proportion of patients with frequent use. The rationale for the development of COX-2 inhibitors, such as Vioxx, was reduced gastrointestinal toxicity as compared with traditional NSAIDs. Previously conducted short term clinical studies were supportive of this hypothesis although concerns about potential cardiovascular toxicity associated with Vioxx had also been raised.

Aim of the study: The VIGOR study was designed to provide more conclusive evidence of the superior gastrointestinal safety of Vioxx. The study was conducted in the years 1999-2000 by Merck with the primary hypothesis that its drug Vioxx would have fewer gastrointestinal side effects than Naproxen for the treatment of rheumatoid arthritis. The study population comprised of 8076 patients "with rheumatoid arthritis who were at least 50 years old (or at least 40 years old and receiving long-term glucocorticoid therapy) and who were expected to require NSAIDs for at least one year". This population was known to be at relatively high risk of gastrointestinal side effects with NSAIDs.[7] The patients in the control arm were assigned the drug Naproxen, while the patients in the active treatment arm were assigned Vioxx.

Details and findings of the study: Patients were followed for a median time of 9 months, and the primary end point was time to first occurrence of a confirmed clinical upper gastrointestinal (GI) event defined as "gastroduodenal perforation or obstruction, upper gastrointestinal bleeding, and symptomatic gastroduodenal ulcers". The original study report [23] performed a survival analysis using a Cox proportional hazard model, and estimated the relative risk for patients in the treatment arm compared with those in the control arm to be 0.5, with a confidence interval of 0.3 to 0.6.[8]

The study authors also conducted a subgroup analysis for the GI events, analyzing subgroups defined by gender, age, nationality, steroids, PUB history (prior history of GI events), and presence of H. pylori antibodies. The rationale was that certain patients were known to

[7] However, the study was conducted with a safety monitoring board: an independent committee whose purpose is to monitor the results of an ongoing trial to ensure the safety of trial participants).

[8] This estimate and the other estimates reported in this chapter are based on an intention-to-treat analysis. The study also performed per-protocol and sensitivity analyses and obtained similar results.

be at increased risk of GI events, and they wanted to see if the benefit of Vioxx extended to these high-risk patients. The conclusion from the subgroup analysis was that the risk ratio for every subgroup remained significant, while differences of the ratios between subgroups were not significant.

However, VIGOR demonstrated that Vioxx was associated with an increase risk of thrombotic cardiovascular events (henceforth referred to as CVT events), an aspect that was not emphasized in the original report of the study [23]. The study authors suggested that apparent association of Vioxx with CVT events was actually the result of Naproxen preventing CVT events. However, placebo controlled studies confirmed that Vioxx did indeed cause CVT events, and this ultimately led to the withdrawal of Vioxx from the market. We refer the reader to the articles [145, 219] for more context on the VIGOR study and its consequences thereafter.

Goal of our investigation into the VIGOR dataset: In this work, we perform analysis for both the GI and CVT events. While the GI event was an infrequent event (experienced by around 2% patients) in the study, the less common CVT event (around 0.6% were reported to have a confirmed CVT event) was considered to be more significant medically. Since the earlier works already established that Vioxx led to an overall decrease in the GI risk but an increase in the cardio risk on the overall population of the study, an important by-product of this work is finding clinically relevant and interpretable subgroups of interest for which Vioxx provided a significant decrease in the risk for the GI event but did not increase the risk for the CVT event. Interpretability of the subgroup, as well as the transparency of the search procedure is important from a clinical view point, as the doctors can then better justify their choice to favor prescribing the drug for patients in the discovered subgroup.

We present detailed results both for the GI and CVT events throughout this chapter, while occasionally deferring some details to the appendix. To perform our analysis, we created a dataset with the two outcomes—GI and CVT event—as discussed above, a treatment indicator, and 16 binary features. The data processing necessary to create this dataset is the topic of the next section.

9.2.2 Feature selection and engineering

The VIGOR study collected an extensive range of patient data, including demographic details, prior medical history, as well as the timing and details of adverse events during the clinical experiment. From this, we extracted sixteen clinically relevant binary features, which we report in Table 9.1 together with covariate balance details. We now describe some of the decisions we took with respect to feature engineering, as well as the meaning the selected features.

The medical history risk factors and drug use information were all already binary, and were selected by the VIGOR study designers as being medically relevant. For instance, it is known that use of glucorticoids predisposes patients to GI events in the context of concomitant NSAID administration [110]. One feature that deserves special interest is ASPFDA.

This was an indicator for patients in the study who “met the criteria of the Food and Drug Administration (FDA) for the use of aspirin for secondary cardiovascular prophylaxis but were not taking low-dose aspirin therapy” [23], and was thought to be an especially strong risk factor for cardiovascular events. Patients who were actually undergoing aspirin therapy were excluded from the study.

On the other hand, some of the demographic and lifestyle risk factors required some engineering. The goal of the feature engineering was to simplify the data using prior information, so as to avoid overfitting and to simplify downstream data analysis. While the study collected more precise data on the patient’s country of residence and their race, in both cases, a single level (“US” and “white” respectively) contained a large fraction of the data, and we used these to binarize the two features. We also applied a similar logic to the smoking and alcohol lifestyle risk factors. We used height and weight information to calculate the body-mass-index (BMI) for every patient, and then used a threshold value of 30 to obtain an indicator for obesity.[9] Finally, we calculated the adjusted age for every patient (by multiplying their numerical age by the ratio of the life expectancy in the US to that in their country of residence), and then used a threshold value of 65 to define an indicator for being elderly. Finally, there was no direct indicator for patients with a prior history of GI event, so we made use of the medical history files to impute this. See Appendix H.2 for more details.

The dataset was fairly complete (as is the case for most RCTs), with only a single patient missing an entry for each lifestyle risk factor (we filled in this with a 1), while 35 patients were missing entries for either height or weight, leading to a missing entry for the obesity indicator (we filled this in with a 0). Furthermore, the features also have weak pairwise correlations except for the fact that the subgroup with ASPFDA=1 (321 patients) is a subset of that with ASCGRP=1 (454 patients).

9.2.3 Data splitting

As a known best practice included in the PCS framework, for each outcome, we created a holdout test set comprising 20% of the individuals, which we did not touch in our further investigations until the very last stage of our analysis, i.e. when we wanted to verify our results. Because of the rarity of events for both outcomes, we stratified the split by both the treatment and the outcome simultaneously; such a stratification ensures that the outcome remains balanced across the test-train splits. Let Y denote the binary outcome of interest (GI or CVT event), and T denote the treatment indicator. Then such a stratification (implemented as `model_selection.train_test_split` function in the sklearn library [203]) is done by first categorizing the study subjects in 4 categories $\{\{T = 0, Y = 0\}, \{T = 1, Y = 0\}, \{T = 0, Y = 1\}, \{T = 1, Y$ once with Y denoting the GI event, and once with Y denoting the CVT event. Then we select a randomly sampled (without replacement) 20% of the subjects from each category together as the test set $\mathbf{S}_{\text{TEST}}$, with the remaining subjects form the training set $\mathbf{S}_{\text{TRAIN}}$.

[9] https://www.cdc.gov/obesity/adult/defining.html, last accessed on August 11, 2020.

Covariate (ABBRV)	**Control No. (%)**	**Treatment No. (%)**
Overall population	4029 (49.9)	4047 (50.1)
Demographics		
Whether *gender* is male (MALE=1)	814 (20.2)	824 (20.4)
Whether *race* is white (WHITE=1)	2752 (68.3)	2764 (68.3)
Whether *country* is US (US=1)	1750 (43.4)	1748 (43.2)
Whether *adjusted age*[†] > 65 (ELDERLY=1)	1172 (29.1)	1136 (28.1)
Whether *body-mass-index* > 30 (OBESE=1)	1060 (26.3)	1106 (27.3)
Lifestyle		
Whether patient *smokes* ≥ 1 cig./day (SMOKE=1)	1879 (46.6)	1919 (47.4)
Whether patient has ≥ 1 *alcoholic drinks*/week (DRINK=1)	1045 (25.9)	1053 (26.0)
Prior medical history		
of *GI PUB events*[*] (PPH=1)	317 (7.9)	313 (7.7)
of *hypertension* (HYPGRP=1)	1168 (29.0)	1217 (30.1)
of *hypercholesterolemia* (CHLGRP=1)	293 (7.3)	343 (8.5)
of *diabetes* (DBTGRP=1)	254 (6.3)	240 (5.9)
of *atherosclerotic cardiovascular disease* (ASCGRP=1)	216 (5.4)	238 (5.9)
indicating use of *aspirin* under FDA guidelines (ASPFDA=1)	151 (3.7)	170 (4.2)
Prior usage of drugs		
Whether used *glucocorticoids/steroids* (PSTRDS=1)	2253 (55.9)	2244 (55.4)
Whether used *Naproxen* (PNAPRXN=1)	747 (18.5)	759 (18.8)
Whether used *NSAIDs* (PNASIDS=1)	3341 (82.9)	3344 (82.6)
Outcomes		
Whether *GI event* occurred (GI=1)	121 (3.0)	56 (1.4)
Whether *CVT event* occurred (CVT=1)	18 (0.4)	41 (1.0)

Table 9.1: Overview of the baseline covariates in the control and treatment arm of the VIGOR study. [†]Adjusted age denotes age multiplied by the ratio of the life expectancy in the US to that in the individual's country of residence. [*]PUB stands for perforations, ulcers and bleeding.

Also, keeping in mind the rarity of the signals, we do not create an additional validation set, and instead we use the training data via a stratified 4-fold cross validation, where the folds are split uniformly at random, again stratified jointly according to T and Y. For such a split, each fold has around 35 GI events and 11 CVT events among the 1615 patients. We note that for a given outcome (say GI event), we use the same 4-fold CV split—referred to as the *original split* and denoted as `cv_orig`—for tuning the hyperparameters for all the CATE estimators via cross-validation. We also use two *additional* stratified 4-fold cross-validation (random) splits in several results throughout the chapter, and denote them by {`cv_0`, `cv_1`}. No hyperparameter tuning is done on these additional splits, and we simply use the tuned parameters from the `cv_orig` split for fitting the estimators on different sets of training folds of these additional splits. Note that for any 4-fold CV split, there are 4 possible pairs of training-validation folds, denoted generically by $\mathbf{S}_{\mathrm{TF}}$ and $\mathbf{S}_{\mathrm{VF}}$ respectively.

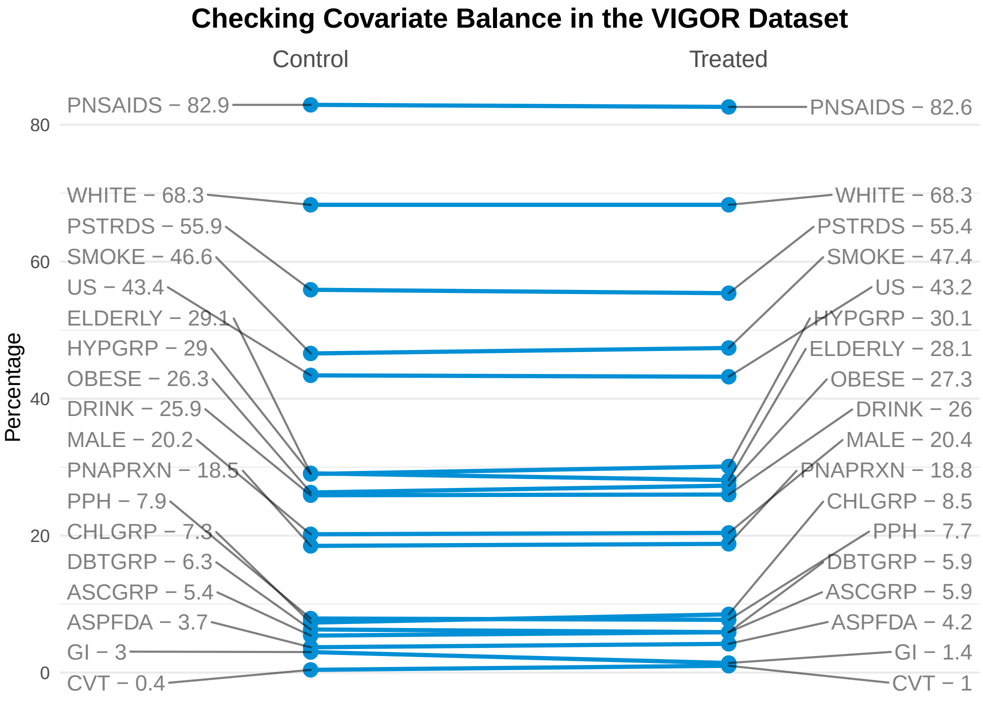


Figure 9.1: A visual illustration showing the covariate balance, and the outcome imbalance (GI and CVT) between the control and treatment population. The abbreviations are detailed in Table 9.1, the number next to the abbreviation (ABBRV) denotes the % of the study size taking value 1 for that ABBRV in the respective arm. Note that the study size was 8076 total patients, and treatment and control arms comprise of 4029 (49.9%) and 4047 (50.1%) individuals respectively.

Mathematically, given disjoint folds from one 4-fold CV split, namely $\{\mathbf{S}_{\mathfrak{f}}\}_{\mathfrak{f}=1}^{4}$ of the training data $\mathbf{S}_{\text{TRAIN}}$ such that $\mathbf{S}_{\text{TRAIN}} = \cup_{\mathfrak{f}=1}^{4}\mathbf{S}_{\mathfrak{f}}$, the 4-pairs of training-validation folds are be denoted by $\{(\mathbf{S}_{\text{TF}} = \mathbf{S}_{\text{TRAIN}} \backslash \mathbf{S}_{\mathfrak{f}}, \mathbf{S}_{\text{VF}} = \mathbf{S}_{\mathfrak{f}}), \mathfrak{f} = 1, 2, 3, 4\}$.

9.3 Review on Neyman-Rubin model and notation

Throughout this chapter, we will assume the standard set up for a completely randomized experiment under the Neyman-Rubin counterfactual framework. We assume that we observe a population of size N, in which the treatment variable T is completely randomized. For each individual i, there are two *potential outcomes*: $Y_i(0)$ when the individual i is assigned to the control arm $T_i = 0$, and $Y_i(1)$ when they are assigned to the treatment arm, $T_i = 1$. The Individual Treatment Effect (ITE) for individual i is defined as the difference of the two potential outcomes $\tau_i = Y_i(1) - Y_i(0)$. But this quantity is unobservable since for each individual we only observe one outcome corresponding to the arm that they are assigned to, i.e, $Y_{i,\text{obs}} = Y_i(T_i)$ which we denote by Y_i for brevity. For each individual i, we also observe a vector of covariates $X_i \in \mathcal{X}$. As is convention with other research into heterogeneous treatment effects, we perform inference by assuming that the samples are drawn i.i.d. from an infinite population.[10]

We now define the various quantities of interest studied throughout this chapter. Let $\mathbf{G}$ be a measurable subset of the feature space $\mathcal{X}$. The average treatment effect (ATE), conditional average treatment effect (CATE) and the subgroup CATE are respectively defined as

$$\text{ATE} : \tau_{\text{ATE}} \coloneqq \mathbb{E}\left[Y(1)\right] - \mathbb{E}\left[Y(0)\right], \tag{9.1a}$$

$$\text{CATE} : \tau(x) \coloneqq \mathbb{E}\left[Y(1) \mid X = x\right] - \mathbb{E}\left[Y(0) \mid X = x\right], \quad \text{for any} \quad x \in \mathcal{X} \tag{9.1b}$$

$$\text{sub-group CATE} : \tau_{\mathbf{G}} \coloneqq \mathbb{E}\left[\tau(X) \mid X \in \mathbf{G}\right], \quad \text{for measurable subset} \quad \mathbf{G} \subset \mathcal{X}, \tag{9.1c}$$

where the expectation is taken with respect to the iid draws from the infinite population.

At a high-level, the goal of this work is to provide a systematic framework to find subgroups $\mathbf{G} \subset \mathcal{X}$, which (i) include non-trivial fraction of the observed data, (ii) are relevant and interpretable relevant for the domain problem at hand, and (iii) most importantly have significant sub-group CATE, i.e., $\tau_{\mathbf{G}}$ has significantly larger magnitude than τ_{ATE}.

Neyman difference-in-means estimates for finite samples: We will often use the classical Neyman difference-in-means estimator to provide plug-in estimates for the ATE and sub-group CATE values. Formally, we denote the two study arms by

$$(\text{Treatment arm})\ \mathbf{T} \coloneqq \{i \in [n] : T_i = 1\} \quad \text{and} \quad (\text{Control arm})\ \mathbf{C} \coloneqq \{i \in [n] : T_i = 0\}, \tag{9.2a}$$

Throughout this chapter, we will abuse notation: for any group $\mathbf{G} \subset \mathcal{X}$, we will use the same symbol to refer the subpopulation of individuals that belong to it. This allows us to denote the restriction of the two arms of the study to the subgroup as follows:

$$\mathbf{T} \cap \mathbf{G} \coloneqq \mathbf{T} \cap \{i \in [n] : X_i \in \mathbf{G}\} \quad \text{and} \quad \mathbf{C} \cap \mathbf{G} \coloneqq \mathbf{C} \cap \{i \in [n] : X_i \in \mathbf{G}\}. \tag{9.2b}$$

For a finite set $\mathcal{A}$, let abs $\mathcal{A}$ denote the number of elements in the set. With this notation at hand, the plug-in estimators for the average treatment effect τ_{ATE} and the sub-group average treatment effect $\tau_{\mathbf{G}}$ are given by

$$\widehat{\tau}_{\text{ATE}} = \frac{1}{\text{abs}\,\mathbf{T}} \sum_{i \in \mathbf{T}} Y_i(1) - \frac{1}{\text{abs}\,\mathbf{C}} \sum_{i \in \mathbf{C}} Y_i(0), \qquad \text{and} \tag{9.3a}$$

$$\widehat{\tau}_{\mathbf{G}} = \frac{1}{\text{abs}\,\mathbf{T} \cap \mathbf{G}} \sum_{i \in \mathbf{T} \cap \mathbf{G}} Y_i(1) - \frac{1}{\text{abs}\,\mathbf{C} \cap \mathbf{G}} \sum_{i \in \mathbf{C} \cap \mathbf{G}} Y_i(0). \tag{9.3b}$$

[10] Note that the standard variance estimates reported using this perspective can be taken as conservative estimates of the finite-sample variances defined in Neyman's repeated sampling framework [72].

For randomized experiments, both estimates $\widehat{\tau}_{\mathrm{ATE}}$ and $\widehat{\tau}_{\mathbf{G}}$ are unbiased [229], and standard error estimates are available for it [124]. On the other hand, the precision of $\widehat{\tau}_{\mathbf{G}}$ degrades as the size of the subgroup shrinks. For the same reason, a direct difference-in-means estimator for CATE (9.1b) is almost never feasible, as for most values of $x \in \mathcal{X}$ (e.g., when $\mathcal{X}$ is continuous, or combinatorially very large), there might not exist any sample with covariate equal to x.

9.4 Calibration as a prediction (reality) check for CATE estimators

Following the Predictability principle of the PCS framework, any statistical model must pass a test of out-of-sample prediction accuracy before we should have any trust in it. This principle is in line with the ethos of the scientific method, which correlates the strength of a hypothesis with the rigor of prior attempts to falsify it [206]. As discussed in Section 9.1.1, however, no such test currently exists for CATE models. The missing potential outcomes mean we do not have a plug-in estimate for any risk function $R(\tau, \hat{\tau}) = \mathbb{E}\left[l(\tau(X), \hat{\tau}(X))\right]$. Furthermore, unlike R^2 and ROC AUC scores, the proxy loss functions proposed for model choice (see Section 9.1.1 and the references therein) do not have interpretable scales.

To mitigate this problem, we develop a prediction accuracy check that can be applied to any CATE estimator. This check makes use of the ideas from the calibration literature [65, 67, 102], and while passing the check is not a sufficient condition for a CATE estimator to have good performance, it is at least a necessary one. Even though our StaDISC approach is motivated by and grounded in the analysis of CATE estimators fitted to the VIGOR study data, we believe it is a general methodology useful for other causal inference problems.

The rest of this section is organized as follows. We discuss the 18 CATE estimators used in our analysis of the VIGOR data in Section 9.4.1. We then introduce the calibration-based scores for prediction checks in Section 9.4.2, and apply it to the CATE estimators trained with VIGOR data in Section 9.4.3. Finally, in Section 9.4.4 we show how despite the poor performance on the overall data, the CATE estimators have good generalization locally, thereby setting the stage for identifying subgroups with subgroup CATE significantly larger than ATE in Section 9.5.

9.4.1 CATE estimators applied on the VIGOR dataset

We now describe the 18 popular CATE estimators used in this work, 14 of which follow meta-learner strategies. Descriptions of the meta-learner strategies can be found in [147] and [196]. Here, we simply list our choices of base learners for each meta-learner. The base learners are all drawn from a pool comprising lasso, logistic regression, random forest (RF), and gradient-boosted trees (GB). In our statistical analyses, we used implementations of the former three algorithms from the `scikit-learn` package [203] and the `XGBoost` implementation of the latter [234]. Furthermore, for code cleanliness, we made use of the meta-learner interface provided by the `causalml` package [41]. In additional to estimators based on meta-learners, we also considered two versions each of causal tree [9] and causal forest [245]. The versions differ in terms of their hyperparameter choices. We used `causalml`'s implementation of the former. For the latter, we were not able to find a well-documented python implementation of the algorithm, so we built one around `causalml`'s causal tree implementation.

(9A) *S-learners* (2 estimators): We used RF and GB as the base learners, denoted by. These are denoted as `s_rf` and `s_xgb`.

(9B) *T-learners* (4 estimators): We used lasso, logistic regression, RF and GB as base learners. These are denoted as `t_lasso`, `t_logistic`, `t_rf` and `t_xgb`.

(9C) *X-learners* (4 estimators): We used lasso, logistic regression, RF and GB as base learners for the first stage, and lasso as the only base learner for the second stage. These are denoted as `x_lasso`, `x_logistic`, `x_rf` and `x_xgb`.

(9D) *R-learners* (4 estimators): In the case of randomized experiments, the R-learner requires a choice of base learner for the conditional expectation of the response with the treatment variable partialed out, and a choice of base learner for the treatment effect. We use four such pairs, each member of which was chosen uniformly at random from the base learners (with logistic regression excluded due to its similarity to lasso). Doing this, we got {lasso, lasso}, {lasso, GB}, {RF, lassso}, and {RF, RF}. These are denoted as `r_lassolasso`, `r_lassoxgb`, `r_rflasso` and `r_rfrf`.

(9E) *Causal Tree and Causal Forest* (4 estimators): We used 2 versions each of the causal tree and causal forest algorithms, which we have denoted as `causal_tree_1`, `causal_tree_2`, `causal_forest_1`, and `causal_forest_2`. Each pair of estimators differ in their hyperparameter choices. Specifically, `causal_tree_1` and `causal_forest_1` both use a minimum of 50 samples per leaf node, whereas `causal_tree_2` and `causal_forest_2` both use a minimum of 200 samples per leaf node. All other hyperparameter choices are standard and can be found in our documentation on GitHub.

Here, we briefly justify our choice of the 18 CATE estimators listed above. First, we chose our pool of base learners because they are representative of the most popular supervised learning algorithms in use today, with neural networks omitted because of the poor signal and small size of the data set. The T-learner framework is perhaps the simplest way of fitting a CATE model and has been used and studied by many different authors. Using lasso as the base learners was proposed and analyzed by Bloniarz et al. [21] and Imai and Ratkovic [123]. Meanwhile, [92] proposed using RF as the base learner. The X-learner [147] and R-learner [196] frameworks have both been used by many recent works. The former has demonstrated favorable performance over other estimators in data challenges organized by the Atlantic Causal Inference Conference, while the latter has optimality guarantees under some assumptions, and has been further supported by some follow up work [225]. We included two S-learner estimators for completion, since all four meta-learner frameworks are supported by the `causalml` package. The causal tree [9] and causal forest [245] estimators have similarly been used in much recent work, with the latter attaining the status of being a benchmark of sorts for CATE estimation methods in many simulations.

All CATE estimators based on meta-learners had the hyperparameters of their component base learners tuned via 4-fold CV using `cv_orig`. A common hyperparameter grid was used for each base learner type, with details deferred to our documentation on GitHub.

9.4.2 A calibration-based score for CATE estimators

To develop a reality check scheme for CATE estimators, we now build on the literature of calibration of probability scores.

A binary classifier is said to be well-calibrated if the class probabilities that it predicts for each sample point is close to the true class probabilities. This property is desirable in many situations, such as weather-forecasting, where we would like it to rain on close to 40% of the days on which a 40% chance of rain is forecast. Unfortunately, machine learning models are often not naturally calibrated, with neural networks in particular being overconfident in their estimated class probabilities [102]. Furthermore, because class probabilities are unobserved, we cannot directly train a model to predict these values using supervised learning. While researchers have proposed various solutions to this problem, the common theme is to *bin* the observations by their *predicted class probabilities*, and then use the observed class distribution over the bin to obtain plug-in estimates of the true class probabilities.

The concept of calibration has a long history [65, 67], and it has also been referred to as validity [181] or reliability [185]. Starting for evaluation of weather forecasts in the 1950s [29], calibration has been widely used as a generic scheme to compare several forecasters [67]. Related ideas have been used to calibrate a wide range of methods, including Bayesian models [65], SVMs, boosted trees, random forests [195, 186], and more recently deep neural networks [102].

Binning via estimated CATE values: We now begin to define our calibration-based prediction accuracy measure for CATE estimators. While our scores—to be defined below—are easy to interpret, defining them formally requires a bit of notation which we now describe.

Consider the training set $\mathbf{S}_{\text{TRAIN}}$ and let $\mathbf{S}_{\mathfrak{f}}, \mathfrak{f} = 1, 2, 3, 4$ denote its 4-fold (random) CV split. Fix a fold $\mathfrak{f}$ and let $\mathbf{S}_{\text{TF}} = \mathbf{S}_{\text{TRAIN}} \backslash \mathbf{S}_{\mathfrak{f}}$ denote the training folds used to fit the CATE estimator $\mathbf{M} : \mathcal{X} \to \mathbb{R}$, and let $\mathbf{S}_{\text{VF}} = \mathbf{S}_{\mathfrak{f}}$ denote the left-out fold, which we also call as validation fold, for the estimator $\mathbf{M}$. Let $\mathfrak{m}_{\mathfrak{q}}$ denote the q-th quantiles of the CATE estimator $\mathbf{M}$ on the training folds of the data:

$$\mathfrak{m}_{\mathfrak{q}} = \min \left\{ c \;\middle|\; \frac{\#\{i \in \mathbf{S}_{\text{TF}} : \mathbf{M}(x_i) \leq c\}}{\text{abs}\, \mathbf{S}_{\text{TF}}} \geq \mathfrak{q} \right\}, \quad \text{for any} \quad \mathfrak{q} \in (0, 1), \tag{9.4}$$

where by convention we set $\mathfrak{m}_0 = -\infty$ and $\mathfrak{m}_1 = \infty$. Then given a grid of q-values denoted by $\{\mathfrak{q}_1 \leq \mathfrak{q}_2 \leq \cdots \leq \mathfrak{q}_{K-1}\}$ in the interval $(0, 1)$, we split the real line into K bins as follows:

$$\mathfrak{m}_0 < \mathfrak{m}_{\mathfrak{q}_1} \leq \mathfrak{m}_{\mathfrak{q}_2} \leq \ldots \leq \mathfrak{m}_{\mathfrak{q}_{K-1}} < \mathfrak{m}_1.$$

We use this binning to induce a partition of $\mathcal{X}$ into K *quantile-based subgroups* given by

$$\mathbf{G}_j := \mathbf{G}_j(\mathbf{M}) = \left\{ x \in \mathcal{X} \;\middle|\; \mathbf{M}(x) \in [\mathfrak{m}_{\mathfrak{q}_j}, \mathfrak{m}_{\mathfrak{q}_{j+1}}] \right\} \quad \text{for} \quad j = 0, 1, \ldots K-1, \tag{9.5a}$$

Given a set of individuals $\mathbf{S}$ (say, training folds $\mathbf{S}_{\mathrm{TF}}$ or validation fold $\mathbf{S}_{\mathrm{VF}}$), let $\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}}$ denote the mean of the predicted CATE from the estimator $\mathbf{M}$ on the subgroups $\mathbf{G}_j \cap \mathbf{S}$:

$$\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}} := \frac{1}{\mathrm{abs}\, \mathbf{G}_j \cap \mathbf{S}} \sum_{i \in \mathbf{G}_j \cap \mathbf{S}} \mathbf{M}(X_i), \quad \text{where} \quad \mathbf{G}_j \cap \mathbf{S} = \{ i \in \mathbf{S} | X_i \in \mathbf{G}_j \}\,, \tag{9.5b}$$

Similarly, recall that $\widehat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}$ denotes the plug-in estimate for the subgroup CATE for the subgroup $\mathbf{G}_j$.

$$\widehat{\tau}_{\mathbf{G}_j \cap \mathbf{S}} := \frac{1}{\mathrm{abs}\, \mathbf{T} \cap \mathbf{G}_j \cap \mathbf{S}} \sum_{i \in \mathbf{T} \cap \mathbf{G}_j \cap \mathbf{S}} Y_i(1) - \frac{1}{\mathrm{abs}\, \mathbf{C} \cap \mathbf{G}_j \cap \mathbf{S}} \sum_{i \in \mathbf{C} \cap \mathbf{G}_j \cap \mathbf{S}} Y_i(0). \tag{9.5c}$$

Score definitions: With these definitions of the sub-groups, we are now ready to define the calibration score:

$$\text{Cal-Score}(\mathbf{S}; \mathbf{M}) := \sum_{j=1}^{K} \frac{\mathrm{abs}\, \mathbf{G}_j \cap \mathbf{S}}{\mathrm{abs}\, \mathbf{S}} \cdot \mathrm{abs}\, \overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}} - \widehat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}, \tag{9.6a}$$

where we use absolute difference (and not squared difference) since the scale of the quantities $\{\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}}, \widehat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}\}$ is pretty small for our dataset. Nonetheless, it is still hard to interpret the absolute scale of Cal-Score($\mathbf{M}$), and hence we normalize these scores by a baseline to define a pseudo-R^2 score. More precisely, we consider a baseline calibration-score Cal-Score($\mathbf{S}; \widehat{\tau}_{\mathrm{ATE}}$), obtained by replacing the the CATE estimator average $\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}}$ with that of the (constant) ATE estimate $\widehat{\tau}_{\mathrm{ATE}}$ in equation (9.6a):

$$\text{Cal-Score}(\mathbf{S}; \widehat{\tau}_{\mathrm{ATE}}) := \sum_{j=1}^{K} \frac{\mathrm{abs}\, \mathbf{G}_j \cap \mathbf{S}}{\mathrm{abs}\, \mathbf{S}} \cdot \mathrm{abs}\, \widehat{\tau}_{\mathrm{ATE}} - \widehat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}. \tag{9.6b}$$

With equations (9.6a) and (9.6b) in place, we define the $\mathcal{R}^2_{\mathrm{C}}$ score as follows:

$$\mathcal{R}^2_{\mathrm{C}}(\mathbf{S}; \mathbf{M}) := 1 - \frac{\text{Cal-Score}(\mathbf{S}; \mathbf{M})}{\text{Cal-Score}(\mathbf{S}; \widehat{\tau}_{\mathrm{ATE}})}. \tag{9.6c}$$

Just like the usual R^2-score[11], the score $\mathcal{R}^2_{\mathrm{C}}(\mathbf{S};\mathbf{M})$ can take any value between $(-\infty,1]$, and a model can be deemed a good fit if this score is close to 1. We interpret the score as measuring, conditioned on the partition of the feature space into bins, the degree to which the CATE estimator explains the variability of the CATE with respect to the partition, in comparison to the best constant model.

Since different models induce different partitions, the scores are not necessarily comparable across models. Furthermore, similar to how calibrated classification algorithms need not have good prediction accuracy, it is possible for a CATE model to have a good $\mathcal{R}^2_{\mathrm{C}}$ score and yet have poor overall prediction accuracy for the CATE. Nonetheless, having $\mathcal{R}^2_{\mathrm{C}}$-scores that are reasonably close to 1 across a range of data perturbations is *necessary* albeit not sufficient for the CATE model to have good prediction performance. Moreover, the variability of the score between the choices $\mathbf{S}=\mathbf{S}_{\mathrm{TF}}$ and $\mathbf{S}=\mathbf{S}_{\mathrm{VF}}$ also provides a check on the *overfitting* of the CATE estimator.

To conclude, the $\mathcal{R}^2_{\mathrm{C}}$ provides two predictive checks for the CATE estimators. On the one hand, when $\mathcal{R}^2_{\mathrm{C}}(\mathbf{S}_{\mathrm{TF}};\mathbf{M})$ is much smaller than 1, we conclude that the estimator $\mathbf{M}$ has a poor fit on the training data. On the other hand, a high value (close to 1) value for $\mathcal{R}^2_{\mathrm{C}}(\mathbf{S}_{\mathrm{TF}};\mathbf{M})$, and a relatively lower value (close to 0 or negative) for $\mathcal{R}^2_{\mathrm{C}}(\mathbf{S}_{\mathrm{VF}};\mathbf{M})$ would necessarily indicate overfitting of the estimator $\mathbf{M}$.

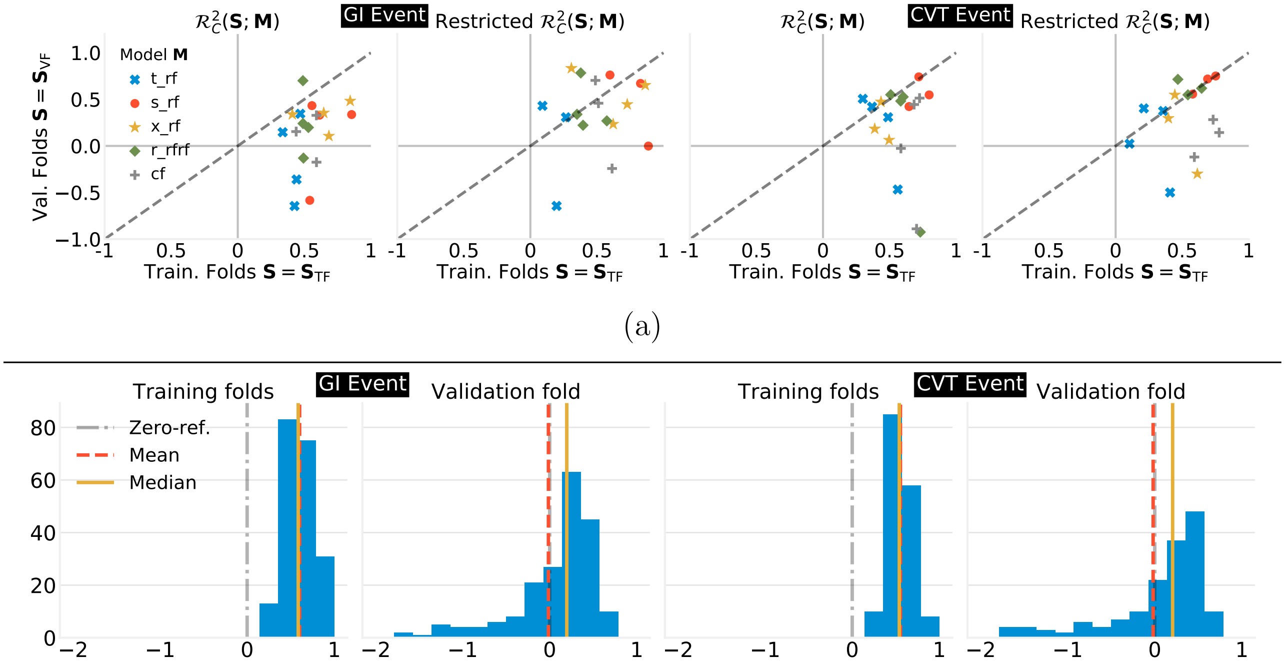
9.4.3 Calibration-based predictive check on CATE estimators for VIGOR dataset

We now compute the scores defined in the previous section for the 18 popular CATE estimators when applied to the VIGOR dataset. We use the evenly-spaced quantile grid $\{0.2, 0.4, 0.6, 0.8\}$ and compute the $\mathcal{R}^2_{\mathrm{C}}$-scores using the $K=5$ bins it induces. We also consider a restricted $\mathcal{R}^2_{\mathrm{C}}$-score to measure the predictive performance of the estimators for the bottom-2 bins for the GI event, and top-2 bins for the CVT event. To compute this *restricted* $\mathcal{R}^2_{\mathrm{C}}$-score, we simply replace the sum over the index $j\in\{1,2,\ldots,5\}$ in equations (9.6a) and (9.6b) with $j\in\{1,2\}$ for the GI event and $j\in\{4,5\}$ for the CVT event, and then plug this restricted sum in equation (9.6c).

In the previous section, we described how, given a CATE estimator and a fixed fold $\mathfrak{f}$, we obtain two (restricted) $\mathcal{R}^2_{\mathrm{C}}$-scores—one on the training folds $\mathbf{S}_{\mathrm{TRAIN}}\backslash\mathbf{S}_{\mathfrak{f}}$ and one on the validation fold $\mathbf{S}_{\mathfrak{f}}$. Repeating this over 4 folds provides us with 4 pairs of such scores. And iterating over M different types of CATE estimators yields $M\times 4$ such pairs. Furthermore, if we consider L different 4-folds splits, we get $M\times 4\times L$ such pairs of scores.

[11] While R^2-score was originally introduced for linear regression, several similar measures have been proposed for providing an interpretable scale to measure the model fit. The R^2 for linear regression takes value in [0,1] for training data, and $(-\infty,1]$ for test data. Close to 1 value suggests a good fit, and a smaller score implies a poor fit. Note that unlike the R^2 for linear regression, for CATE estimators, the pseudo-score $\mathcal{R}^2_{\mathrm{C}}$ is not guaranteed to take value in $[0,1]$ even on the training data, i.e., $\mathcal{R}^2_{\mathrm{C}}(\mathbf{S}_{\mathrm{TF}};\mathbf{M})\in(-\infty,1]$. Nonetheless, in Fig. 9.2, we observe that for all the CATE estimators, this score lies in [0, 1] on the training folds, i.e., $\mathcal{R}^2_{\mathrm{C}}(\mathbf{S}_{\mathrm{TF}};\mathbf{M})\in[0,1]$.

We trained 18 different CATE estimators for both the outcomes, namely the GI and CVT events. However, after fitting, the following estimators learned a zero CATE function: R-learner with XGBoost for the GI event, and S-learner with XGBoost, Causal Tree with a particular choice of hyperparameters, and R-learner with XGBoost for the CVT event. Thus, going forward we report results for the remaining 17 CATE estimators for the GI event and 15 CATE estimators for the CVT event. See Section 9.4.1 for more details on all the estimators. We now first discuss the details of scores presented in various plots in Fig. 9.2 and then discuss the conclusions in a separate paragraph.



(a)

(b)

Figure 9.2: Plots with the calibation-based $\mathcal{R}^2_C$-scores (9.6c) for various CATE estimators. **(a)** Scatter plot of $\mathcal{R}^2_C$-scores on the training and validation folds for 5 CATE estimators on the original 4-fold split `cv_orig` on which hyperparameters were tuned via cross-validation. Refer to the text for definition of restricted $\mathcal{R}^2_C$-scores. **(b)** Histogram of the $\mathcal{R}^2_C$-scores on the 12 training and validation folds, 4 each from the 3 different CV splits, namely {`cv_orig`, `cv_0`,`cv_1`} for 17 CATE estimators for GI event, and for 15 CATE estimators for CVT event.

Details of Fig. 9.2: In Fig. 9.2(a), we provide a scatter plot of $\mathcal{R}^2_C(\mathbf{S}_{TF}, \mathbf{M})$ (training score) and $\mathcal{R}^2_C(\mathbf{S}_{VF}, \mathbf{M})$ (validation score) for 5 different estimators for each fold of original CV split `cv_orig`on the VIGOR data both for GI and CVT events. These estimators are T_RF, S_RF, X_RF, R_RFRF and CF_1 which denote T, S, X, R-learners with random forest as base learners, and (one of the two) Causal Forest respectively. In addition, in the right two figures in Fig. 9.2(a), we also provide the scatter plot of the corresponding restricted $\mathcal{R}^2_{\mathrm{C}}$-scores (see the first paragraph of this section for its definition) on the training and validation folds for the 5 estimators and both events.

Next, to check the *stability* of our conclusion, we compute these scores for all 17 CATE estimators for the GI event, and all 15 CATE estimators for the CVT Eventon all 3 random CV splits {`cv_orig`,`cv_0`,`cv_1`}. That is, we obtain a total of 204 and 180 (training and validation) pairs of $\mathcal{R}^2_{\mathrm{C}}$-scores respectively for the GI and CVT events. In Fig. 9.2(b), we plot the histogram of these scores.

Conclusions from Fig. 9.2: Inspecting the scatter plots in Fig. 9.2(a), we see clear evidence of overfitting, as the validation fold $\mathcal{R}^2_{\mathrm{C}}$-scores (computed as $\mathcal{R}^2_{\mathrm{C}}(\mathbf{S}_{\mathrm{VF}}, \mathbf{M})$ in equation (9.6c)) are systematically much smaller, and often negative, than those on the training folds (computed as $\mathcal{R}^2_{\mathrm{C}}(\mathbf{S}_{\mathrm{TF}}, \mathbf{M})$ in equation (9.6c)). Furthermore, there is substantial variability across different folds. For instance, one dot corresponding to S_RF for GI events was not even plotted because the validation fold $\mathcal{R}^2_{\mathrm{C}}$ score exceeded the lower y-limit of the plot. These findings are supported by the histograms in Fig. 9.2(b), which show that the mean of the validation fold $\mathcal{R}^2_{\mathrm{C}}$-scores is in fact a negative number for both GI and CVT events. While we presented histograms of the aggregated scores over all the CATE estimators, the general behavior was also true when looking at individual CATE estimators. Next, we also note that the bottom-2-restricted $\mathcal{R}^2_{\mathrm{C}}$-score for the GI event and top-2-restricted $\mathcal{R}^2_{\mathrm{C}}$-score have slightly better generalization since the validation scores are generally positive albeit with the caveat of larger variability across the training folds. (We revisit this aspect in more detail in Section 9.4.4.)

The poor performance on average as well as the high variability of performance both lead us to be skeptical of the conclusions from any CATE estimator on the VIGOR study data. Here, we remark that the variability of the scores stems from both fluctuations in the trained model as well as low SNR in the validation fold (leading to Cal-Score deviating from its expected value). We remind the reader that in total there are 177 GI events and 59 total CVT events, and this fact implies that for each quantile-based subgroup, we should expect to see around 7.1 and 2.3 GI and CVT events respectively in the validation fold, under the assumption of no heterogeneity. The poor performance is hence entirely to be expected, and in fact could be a general theme for RCTs, as they are often sufficiently powered for only computing the ATE.

9.4.4 Extracting data conclusions that can be relied upon

While we conclude that we cannot trust the CATE models in their entirety, it remains to be seen if we can isolate data conclusions from them that we can rely on. To this end, we take a closer look the relative ordering of scores $\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}}$ (9.5b) and $\widehat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}$ equation (9.5c) across the quantile-based subgroups $\{\mathbf{G}_j\}_{j=1}^5$ considered in the previous section. Given the quantile-based definition of the groups, it is natural to test whether we have

$$\overline{\mathbf{M}}_{\mathbf{G}_1\cap\mathbf{S}} \leq \overline{\mathbf{M}}_{\mathbf{G}_2\cap\mathbf{S}} \leq \ldots \leq \overline{\mathbf{M}}_{\mathbf{G}_5\cap\mathbf{S}}, \quad \text{(estimator CATEs)} \quad \text{and} \tag{9.7a}$$

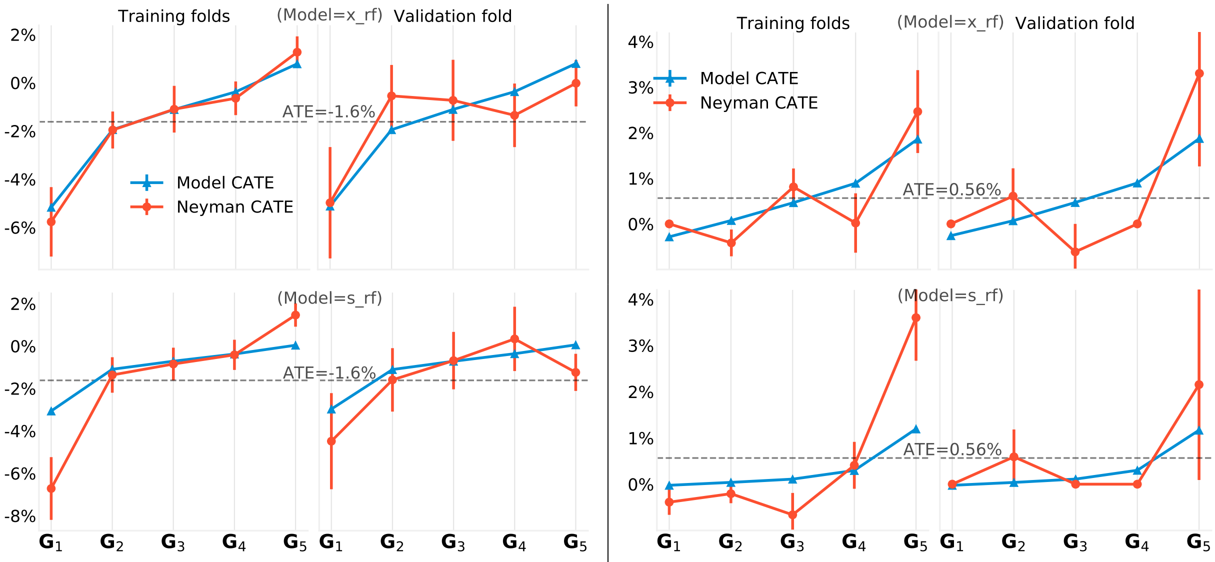
$$\widehat{\tau}_{\mathbf{G}_1\cap\mathbf{S}} \leq \widehat{\tau}_{\mathbf{G}_2\cap\mathbf{S}} \leq \ldots \leq \widehat{\tau}_{\mathbf{G}_5\cap\mathbf{S}}, \quad \text{(subgroup CATE estimates)} \tag{9.7b}$$

for a set of individuals $\mathbf{S}$ comprising either the training folds or the validation fold. In Fig. 9.3, we plot these estimates for two estimators X_RF and T_RF for the GI event in panel (a) and the CVT event in panel (b) for one set of training and validation folds from the original split. In each plot, the blue error bars denotes the sample standard deviation estimate for the sample mean $\overline{\mathbf{M}}_{\mathbf{G}_j\cap\mathbf{S}}$ computed from $\{\mathbf{M}(X_i), i\in\mathbf{G}_2\cap\mathbf{S}\}$, and the red error bars denote the standard error estimate for $\widehat{\tau}_{\mathbf{G}_j\cap\mathbf{S}_{\text{TF}}}$ given by equation (9.11b). We observe that generally the model CATE estimates $\left\{\overline{\mathbf{M}}_{\mathbf{G}_1\cap\mathbf{S}}\right\}_{j=1}^{5}$ are monotonic for both events on both training folds and validation fold. However, the story with the plug-in subgroup CATE estimates $\left\{\widehat{\tau}_{\mathbf{G}_j\cap\mathbf{S}}\right\}_{j=1}^{5}$ is—not unexpectedly—mixed. For the GI event, while these estimates are monotonic on the training folds ($\mathbf{S}=\mathbf{S}_{\text{TF}}$), they are not monotonic on the validation fold ($\mathbf{S}=\mathbf{S}_{\text{VF}}$). For the rarer CVT event, the estimates $\left\{\widehat{\tau}_{\mathbf{G}_j\cap\mathbf{S}}\right\}_{j=1}^{5}$ are not even monotonic on the training folds. This non-monotonic behavior is far from unique to the two estimators presented here. Instead, the plots are representative of what we observe for all other estimators as well, even when using alternate data splits into training and validation folds.

Pairwise comparisons: To summarize this phenomenon, we do a pairwise comparison of successive quantile-based subgroups and measure the frequency with which the ordering of their CATE values generalizes to the validation fold, and summarize our results in Fig. 9.4(a). More precisely, for a given estimator $\mathbf{M}$, we define the boolean indicators:

$$A_{j,j+1} = \mathbb{I}(\widehat{\tau}_{\mathbf{G}_j\cap\mathbf{S}_{\text{VF}}} \leq \widehat{\tau}_{\mathbf{G}_{j+1}\cap\mathbf{S}_{\text{VF}}}) \quad \text{for} \quad j=1,2,3,4. \tag{9.8a}$$

We then compute how often we have $A_{j,j+1}=1$ over the 12 validation folds 4 each from the 3 CV splits {`cv_orig`,`cv_0`,`cv_1`}, and denote this value by $\overline{A}_{j,j+1}$. Finally, we provide a boxplot of the distribution of the values $\left\{\overline{A}_{j,j+1}, j=1,2,3,4\right\}$ across all 17 CATE estimators for the GI event, and 15 CATE estimators for the CVT event in panel (a) of the Fig. 9.4. A value close to 1 suggests good generalization, and conversely, a value close to 0 reflect poor generalization. On the one hand, we see that the pairwise ordering does not generalize well for most pairs of successive quantile-based subgroups as the frequency of generalization $\overline{A}_{j,j+1}$ concentrates around values ≤ 0.5 for $j=2,3,4$ for the GI event, and $j=1,2,3$ for the CVT event. On the other hand, we see that values of $\overline{A}_{1,2}$ for the GI event, and those of $\overline{A}_{4,5}$ for the CVT event are pretty close to 1 (we present more precise numerical values in Table H.1.) This observation suggests that the ordering does generalize well for the subgroup with the strongest negative treatment effect for the GI event, and the strongest positive treatment effect for the CVT event.



(a) GI event

(b) CVT event

Figure 9.3: Investigating the monotonicty trend (equation (9.7)) for two CATE estimators X_RF and S_RF on one set of three training folds and one validation fold of the original 4-fold split `cv_orig`, for **(a)** the GI Event, and **(b)** the CVT Event. Here "Model CATE" refers to the quantity $\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}}$, and Neyman CATE refers to the quantity $\widehat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}$. In our notation, for training folds $\mathbf{S} = \mathbf{S}_{\text{TF}}$, and for validation fold $\mathbf{S} = \mathbf{S}_{\text{VF}}$. The error bars for Model CATE are the sample standard deviation for the estimated CATE values from the model, for each subgroup. For the Neyman CATE, the error bar denotes the square-root of the estimated variance (9.11b). Note that the subgroups $\{\mathbf{G}_j\}$ are defined by the CATE estimator via the training folds.

Investigating the quantile-based "top" subgroups: We call the subgroups induced by $\mathbf{G}_1$ for the GI event, and $\mathbf{G}_5$ for the CVT event, the *quantile-based top subgroup*. Note that each subgroup is specific to a choice of estimator, a choice of training-validation split, and a choice of quantile-grid. To further analyze the good generalization of ordering for these top subgroups, we also compare them to the other quantile-based subgroups via two boolean variables as follows:

$$\text{for GI event:} \quad A_{1,\min} := \mathbb{I}(\widehat{\tau}_{\mathbf{G}_1 \cap \mathbf{S}_{\text{VF}}} = \min_j \widehat{\tau}_{\mathbf{G}_j \cap \mathbf{S}_{\text{VF}}}), \text{and} \tag{9.8b}$$

$$\text{for CVT event:} \quad A_{5,\max} := \mathbb{I}(\widehat{\tau}_{\mathbf{G}_5 \cap \mathbf{S}_{\text{VF}}} = \max_j \widehat{\tau}_{\mathbf{G}_j \cap \mathbf{S}_{\text{VF}}}). \tag{9.8c}$$

We report the distribution of the frequency of generalization $\overline{A}_{1,\min}$ (mean computed over the 12 validation folds) across the 17 CATE estimators for the GI event, and $\overline{A}_{5,\max}$ across the 15 CATE estimators for the CVT event as the rightmost entry of the corresponding figure in Fig. 9.4(a). The plots show that, on the validation fold, the *quantile-based top subgroup* has the strongest treatment effect 90% of the time for the GI outcome, and about 80% of the time for the CVT outcome.

Next, to better investigate the performance of quantile-based top subgroups, we compare these top subgroups directly against their complement, reporting the results in Fig. 9.4(b). In this plot, we also vary the $\mathfrak{q}$-value threshold used to define the quantile-based top subgroup. In particular, we consider groups of the form

$$\widetilde{\mathbf{G}}_{\mathfrak{q}} = \{x \in \mathcal{X} | \mathbf{M}(x) \in (-\infty, \mathfrak{m}_{\mathfrak{q}}]\} \tag{9.9}$$

where $\mathfrak{m}_{\mathfrak{q}}$ denotes the q-th quantile of the CATE estimator $\mathbf{M}$ on the training folds (see equation (9.4) for the mathematical expression). Note that with this notation, $\widetilde{\mathbf{G}}^c_{\mathfrak{q}} = \{x \in \mathcal{X} | \mathbf{M}(x) \in (\mathfrak{m}_{\mathfrak{q}}, \infty\}$. In simple words, the subgroup $\widetilde{\mathbf{G}}_{\mathfrak{q}}$ is based on the quantile range $[0, \mathfrak{q}]$, and its complement subgroup $\widetilde{\mathbf{G}}^c_{\mathfrak{q}}$ is based on the quantile-range $[\mathfrak{q}, 1]$. Then we check the ordering for between these subgroups via the following boolean indicators:

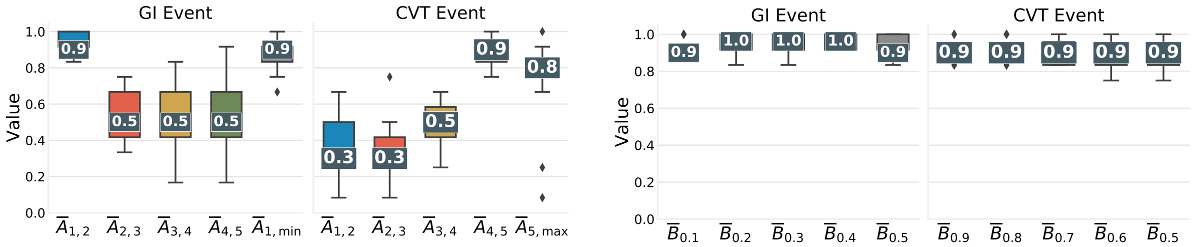
$$B_{\mathfrak{q}} = \mathbb{I}\left(\widehat{\tau}_{\widetilde{\mathbf{G}}_{\mathfrak{q}} \cap \mathbf{S}_{\text{VF}}} \leq \widehat{\tau}_{\widetilde{\mathbf{G}}^c_{\mathfrak{q}} \cap \mathbf{S}_{\text{VF}}}\right), \quad \text{for} \quad \begin{cases} \mathfrak{q} \in \{0.1, 0.2, \ldots, 0.5\} & \text{for GI event} \\ \mathfrak{q} \in \{0.9, 0.8, \ldots, 0.5\} & \text{for CVT event.} \end{cases} \tag{9.10}$$

Note that the subgroup of interest is $\mathbf{G}_{\mathfrak{q}}$ for the GI event and $\mathbf{G}^c_{\mathfrak{q}}$ for the CVT event. Moreover, in this new notation, the earlier subgroups (from Fig. 9.4(a)) would be represented as $\mathbf{G}_1 = \widetilde{\mathbf{G}}_{0.2}$ and $\mathbf{G}_5 = \widetilde{\mathbf{G}}^c_{0.8}$. We notice that the ordering (9.10) holds much more frequently (compared to the pairwise ordering in Fig. 9.4(a)). We also note from this figure that $\mathfrak{q} = 0.2$ and $\mathfrak{q} = 0.8$ provide the best generalization performance for the GI and CVT events respectively.

In summary, we have found that at least some of the CATE estimators yield quantile-based top subgroups that have subgroup CATE that is demonstrably stronger than that of the rest of the population. Thus, in the following sections, we use these quantile-based top subgroups, namely the subgroups $\{\mathbf{G}_{\mathfrak{q}}, \mathfrak{q} = 0.1, 0.2, \ldots, 0.5\}$ for the GI event, and $\left\{\mathbf{G}^c_{\mathfrak{q}}, \mathfrak{q} = 0.9, 0.8, \ldots, 0.5\right\}$ for the CVT event for further analysis.

9.5 Stability-driven ranking and aggregation of CATE estimators

Based on the discussion at the end of the last section, we believe that we can use a subcollection of the CATE estimators to find subgroups with highly negative (in the case of the GI outcome) or positive (in the case of the CVT outcome) subgroup CATE, in the form of a quantile-based top subgroup. This observation brings us back to the question of estimator screening and choice: We seek to define a more stringent predictive test, and furthermore, out of all CATE estimators we considered, we would like to select those that are able to give us the best subgroups. While the overall goal of StaDISC is to find subgroups that are



(a) (b)

Figure 9.4: Box plots for pairwise comparisons of the subgroup CATE estimates for the 5 quantile-based subgroups based on the quantile grid $\{0.2, 0.4, 0.6, 0.8\}$. The boxplots in panel **(a)**, denote the distribution for the mean fraction $\overline{A}_{j,j+1}$ (9.8a) (where the mean is computed over the 12 validation folds, 4 each from the 3 random CV splits {`cv_orig`,`cv_0`,`cv_1`}) across various CATE estimators, for the GI event on the left, and CVT event on the right. In addition, we also show the boxplot of the distribution of the boolean variables $\overline{A}_{1,\text{min}}$ (9.8b) for the GI event, and $\overline{A}_{5,\text{max}}$ (9.8c) in the rightmost column of respective plot. In panel **(b)**, we provide boxplots for the distribution of the mean value of boolean indicators $\{\overline{B}_{\mathtt{q}}$ (9.10) across all CATE estimators, for $\mathtt{q} \in \{0.1, 0.2, \ldots, 0.5\}$ for the GI event, and $\mathtt{q} \in \{0.9, 0.8, \ldots, 0.5\}$ for the GI event, where the mean is computed over the and the distribution is plotted across all the CATE estimators. Refer to Table H.1 for estimator-wise results.

both statistically significant and interpretable, we focus in this part of chapter on selecting estimators that yield the most significant subgroups, and only address interpretability in Section 9.6.

9.5.1 Comparing estimators using t-statistics

We compare different CATE estimators using the statistical significance of their quantile-based top subgroup, measured via using standardized scores, namely t-statistics. Given a subgroup $\mathbf{G}$, its corresponding t-statistic is given by:

$$\mathbb{T}_{\mathbf{G}} := \frac{\widehat{\tau}_{\mathbf{G}} - \widehat{\tau}_{\text{ATE}}}{\sqrt{\widehat{\text{Var}}\left[\widehat{\tau}_{\mathbf{G}} - \widehat{\tau}_{\text{ATE}} \mid \mathcal{F}\right]}}. \tag{9.11a}$$

Here, the term in the denominator is a plug-in estimate of a conditional variance, where the conditioning is over a σ-algebra $\mathcal{F}$ comprising knowledge of the group labels and treatment labels for all individuals in the sample population. More precisely, the variance estimate is given by

$$\widehat{\mathrm{Var}}\left[\widehat{\tau}_{\mathbf{G}} - \widehat{\tau}_{\mathrm{ATE}} \;\middle|\; \mathcal{F}\right] := \left(\frac{\operatorname{abs}\mathbf{G}^c \cap \mathbf{C}}{\operatorname{abs}\mathbf{C}}\right)^2 \cdot \left(\frac{\widehat{\mathrm{Var}}\left[Y(0) \;\middle|\; \mathbf{G} \cap \mathbf{C}\right]}{\operatorname{abs}\mathbf{G} \cap \mathbf{C}} + \frac{\widehat{\mathrm{Var}}\left[Y(0) \;\middle|\; \mathbf{G}^c \cap \mathbf{C}\right]}{\operatorname{abs}\mathbf{G}^c \cap \mathbf{C}}\right)$$
$$+ \left(\frac{\operatorname{abs}\mathbf{G}^c \cap \mathbf{T}}{\operatorname{abs}\mathbf{T}}\right)^2 \cdot \left(\frac{\widehat{\mathrm{Var}}\left[Y(1) \;\middle|\; \mathbf{G} \cap \mathbf{T}\right]}{\operatorname{abs}\mathbf{G} \cap \mathbf{T}} + \frac{\widehat{\mathrm{Var}}\left[Y(1) \;\middle|\; \mathbf{G}^c \cap \mathbf{T}\right]}{\operatorname{abs}\mathbf{G}^c \cap \mathbf{T}}\right), \tag{9.11b}$$

where for a given set $\mathcal{A} \subset \mathcal{S}$, the quantity $\widehat{\mathrm{Var}}\left[Y(t) \;\middle|\; \mathcal{A}\right]$ denotes the sample variance:

$$\widehat{\mathrm{Var}}\left[Y(t)\middle|\mathcal{A}\right] = \frac{1}{\operatorname{abs}\mathcal{A} - 1} \sum_{i \in \mathcal{A}} \left(Y_i(t) - \frac{1}{\operatorname{abs}\mathcal{A}} \sum_{j \in \mathcal{A}} Y_j(t)\right)^2 \quad \text{for} \quad t = 0, 1. \tag{9.11c}$$

We show in Appendix H.1 that the estimator (9.11b) is an unbiased estimator of the conditional variance of $\widehat{\tau}_{\mathbf{G}} - \widehat{\tau}_{\mathrm{ATE}}$, and from the proof, it also easily follows that the estimator is consistent. As such, under the null hypothesis that $\tau_{\mathbf{G}} - \tau_{\mathrm{ATE}} = 0$, the t-statistic yields an asymptotically valid p-value.

In this chapter, we deliberately choose not to use p-values to report the results, so as to avoid their susceptibility to misinterpretation. For interested readers, however, we mention the mapping between p-values and t-statistics ($\mathbb{T}$). The t-statistics presented throughout this work can be associated with one-sided p-values. In particular, a negative t-statistic with magnitude $1.65, 1.96$, and 2.33 can be mapped to a left one-sided p-value of 0.05, 0.025 and 0.01 respectively. The same mapping exists between positive t-statistics and right one-sided p-values.

9.5.2 Defining appropriate perturbations

In order to guard against spurious and unreliable discoveries, the Stability principle of the PCS framework requires conclusions to be stable to reasonable or appropriate perturbations at various stages of the data science life cycle. These include modeling and data perturbations familiar to statisticians which are appropriate under the Neyman-Rubin model assumptions, and also "judgment call" perturbations where we reproduce or at least approximate the conclusions that would have been reached had various contingent choices been made differently. Examples of these choices include those made during data cleaning and feature engineering.[12]

As mentioned earlier, we have used a random CV split in order to fit and analyze our CATE models for the VIGOR data. In line with our prior discussion, we do not just evaluate each estimator based on the 3 CV splits {`cv_orig`,`cv_0`,`cv_1`}, but also perform concurrent analyses of the estimator fitted and validated using four-fold splits of the data under 4

[12] This concern is similar to that expressed by Gelman in his influential paper on *The Garden of Forking Paths* [94].

additional perturbations. Overall, we denote the set of all 7 perturbations by {`cv_orig, cv_0, cv_1, cv_time, elderly_60, overweight, pert_outcome`}, where the 3 (random) CV splits {`cv_orig`,`cv_0`,`cv_1`} have already been used multiple times in the previous results of this chapter. For completeness and to put them in context here, we revisit them while introducing the *new* perturbations {`cv_time, elderly_60, overweight, pert_outcome`} that we make use of in our subsequent analysis of the VIGOR dataset. We remind the reader that for each perturbation, we perform the same 4-fold split for all the CATE estimators. Moreover, we continue to use the tuned hyperparameters from `cv_orig` for all other perturbations.

Sampling perturbations (cv_0, cv_1, cv_time): The additional CV (random) splits {`cv_0, cv_1`}, used earlier and also in the sequel, help to account for sampling variability and are pretty commonly used in statistics and machine learning. Nonetheless, we also share Efron's concern that the use of random splits [87] does not play well with possible covariate shift, and may lead researchers to be overly optimistic about conclusions that do not have external validity. To address this, we also split the training data into four equally-sized folds by binning based on enrollment-time, denoted by {`cv_time`}. This simulates possible variability in the sample population due to human choices (i.e. the date of the RCT)[13], and can also be seen more generally as making use of an a priori irrelevant variable to create heterogeneous folds and thus penalize ephemeral predictors.

Feature engineering perturbations (elderly_60, overweight, pert_outcome): We use alternative thresholds to create perturbed versions of the ELDERLY and OBESE features. Instead of thresholding the adjusted age at 65, we create an ELDERLY_60 feature by thresholding it at 60, and instead of thresholding BMI at 30, we instead threshold it at 25 to define the feature OVERWEIGHT. In this way, we create two perturbed datasets, denoted by {`elderly_60, overweight`}. Finally, for both the GI and CVT outcomes, the VIGOR study recorded for each patient both whether an event occurred, and also whether the occurred event was confirmed (meaning that it met the stringent criteria of an independent panel). In the original study, and thus far in this chapter, we have used the confirmed events as the response of interest, but we now make use of the unconfirmed events to create a new response variable tracking all events. This increases the number of GI events from 177 to 190 and the number of CVT events from 59 to 84. Replacing the original responses with these one creates a further perturbed dataset for each outcome, which we denote by {`pert_outcome`}. For the three perturbations {`elderly_60, overweight, pert_outcome`}, we use the original 4-fold split `cv_orig` of the patients (albeit with the perturbed features or outcomes in the data).

[13] In fact, such a time-based split would be even more relevant for studies based on RCTs that are *online* in nature, meaning that during the trial, results from earlier stages of the trial are used to guide whether the trial would be continued further or concluded.

Performing our analyses on these perturbed datasets reveals to us what would have happened had we, or the original study authors, made different contingent decisions in feature engineering or problem formulation. Although models fit on these datasets no longer have exactly the same meaning as those fit on the original data, we still expect the estimators that perform well on the original data to also perform well on these perturbed datasets.

9.5.3 Ranking and aggregation of CATE estimators

In this section, we first rank the CATE estimators based on their performance across all data perturbations elaborated in the previous section. And, then we select the estimators that are ranked in Top-10 estimators across all the perturbations. Finally, we build a single "ensemble CATE estimator" by taking a simple average (equal weights) of all the selected CATE estimators. Quantile-based top subgroups of the ensemble estimator form the starting point of finding interpretable subgroups in Section 9.6. We now describe the details of our ranking procedure.

Mean t-statistic per data perturbation: For a CATE estimator $\mathbf{M}$, for each data perturbation $\mathfrak{D} \in \{$`cv_orig, cv_0, cv_1, cv_time, elderly_60, overweight, pert_outcome`$\}$, we compute the mean t-statistic averaged across all quantiles across the corresponding 4 validation folds. In our notation, for the GI event, this mean t-statistic is given by

$$\overline{\mathbb{T}}_{\mathrm{GI}}(\mathfrak{D}) = \frac{1}{20}\sum_{\mathfrak{q}\in\mathcal{Q}}\sum_{\mathbf{S}_{\mathrm{VF}}\in\mathcal{F}} \mathbb{T}_{\widetilde{\mathbf{G}}_{\mathfrak{q}} \cap \mathbf{S}_{\mathrm{VF}}} \text{ where } \mathcal{Q} = \{0.1, 0.2, \ldots, 0.5\}\,, \mathcal{F} = \{\mathbf{S}_{\mathfrak{f}}, \mathfrak{f} = 1, 2, 3, 4\}\,, \tag{9.12a}$$

where the quantile-based top subgroup $\widetilde{\mathbf{G}}_{\mathfrak{q}}$ was defined in equation (9.9). Moreover, we remind the reader that the quantiles that define the subgroup $\widetilde{\mathbf{G}}_{\mathfrak{q}}$ (see equations (9.4) and (9.5a)) are computed based on the CATE estimates from the fitted $\mathbf{M}$ on its training folds $\mathbf{S}_{\mathrm{TF}} = \mathbf{S}_{\mathrm{TRAIN}} \backslash \mathbf{S}_{\mathrm{VF}}$. On the other hand, the t-statistic on the RHS of equation (9.12a) is computed on the validation fold $\mathbf{S}_{\mathrm{VF}}$. For the CVT event, the corresponding mean t-statistic is given by

$$\overline{\mathbb{T}}_{\mathrm{CVT}}(\mathfrak{D}) = \frac{1}{20}\sum_{\mathfrak{q}\in\mathcal{Q}}\sum_{\mathbf{S}_{\mathrm{VF}}\in\mathcal{F}} \mathbb{T}_{\widetilde{\mathbf{G}}^{c}_{\mathfrak{q}} \cap \mathbf{S}_{\mathrm{VF}}} \text{ where } \mathcal{Q} = \{0.9, 0.8, \ldots, 0.5\}\,, \mathcal{F} = \{\mathbf{S}_{\mathfrak{f}}, \mathfrak{f} = 1, 2, 3, 4\}\,, \tag{9.12b}$$

We report the mean t-statistic $\overline{\mathbb{T}}(\mathfrak{D})$ for each CATE estimator and all 7 data perturbations in Table 9.2(a) for the GI event, and Table 9.2(b) for the CVT event. We also provide a visual summary of the 7 mean t-statistic for each estimator in the form of boxplot in Fig. 9.5 in panel (a) for the GI event, and panel (b) for the CVT event.

Ranking the CATE estimators: Next, for each category $\mathfrak{D}$, we rank the mean t-statistic from lowest to highest for the GI event, and highest to lowest for the CVT event. In accordance with the Stability principle of the PCS framework, we screen for estimators that perform well across perturbations, and thereby select all estimators that rank in Top-10 across all data perturbations $\mathfrak{D}$. We provide the visual illustration of these ranks also in Fig. 9.5 for the two events. In fact, the estimators in the Fig. 9.5 are sorted based on their worst rank across the perturbations. This criterion selects (i) 2 T-learners and 4 X-learners {t_lasso, x_rf, t_rf, x_xgb, x_lasso, x_logistic} for the GI event, and (ii) 1 S-learner, 3 T-learners, and 1 X-learners {s_rf, t_lasso, t_rf, x_xgb, t_logistic} for the CVT event. The selected list can also be verified by a simple inspection of the rank plots from Fig. 9.5.

Final step before interpreting: Keeping in mind the computational aspects of the next step (finding interpretable subgroups), and to increase stability, we decided to build an ensemble CATE estimator by using a simple average of the selected CATE estimators. Moreover, we also investigate the performance of the quantile-based top subgroups for this ensemble, and report the mean t-statistic across the 12 validation folds from {`cv_orig`,`cv_0`,`cv_1`} for $\widetilde{\mathbf{G}}_{\mathfrak{q}}$ (9.9) for the GI event, and $\widetilde{\mathbf{G}}^c_{\mathfrak{q}}$ for the CVT event in Table 9.3. We report the standard deviation of the t-statistic across these folds in parentheses. In addition, we also report the mean percentage overlap computed pairwise across the entire training set $\mathbf{S}_{\text{TRAIN}}$ for the 12 ensemble estimators, 4 each from the 3 CV splits {`cv_orig`,`cv_0`,`cv_1`}. We observe that for the GI event the subgroups corresponding to $\mathfrak{q} \in \{0.2, 0.3\}$ have relatively higher $\mathbb{T}$, and for the CVT event $\mathfrak{q} \in \{0.9, 0.8\}$ are the top 2 choices. The trends for overlap are as expected, with the increase in size of the group, the overlap generally increases; and remains $> 70\%$ across all choices. In the next section, we discuss our methodology to find an interpretable representation of the quantile-based top subgroups using the ensemble CATE estimator. As a final decision before that step, we choose the groups $\widetilde{\mathbf{G}}_{0.2}$ and $\widetilde{\mathbf{G}}_{0.3}$ for the GI event, and $\widetilde{\mathbf{G}}^c_{0.9}$ for the CVT event, based on their high t-statistic. We also include the group $\widetilde{\mathbf{G}}^c_{0.8}$ for the CVT event keeping in mind the fact that the CVT event is very rare, and thus the low signal in the subgroup $\mathbf{G}^c_{0.9}$ (having only 10% of the training data) may become a bottleneck for any reasonable inference task.

9.6 Finding interpretable subgroups

The next and final step of our investigation is to make our findings interpretable. Recall that the end goal in investigating the heterogeneous treatment effects in the VIGOR study is to inform treating physicians which subgroup of patients are likely to benefit from the reduced risk of GI events, without simultaneously incurring an increased risk of CVT events. Physicians may then favor prescribing the drug for patients in this subgroup. In situations involving high stakes decision-making such as this one, decision-makers are usually not comfortable with black-box decision rules, but instead ideally require rules to be transparent

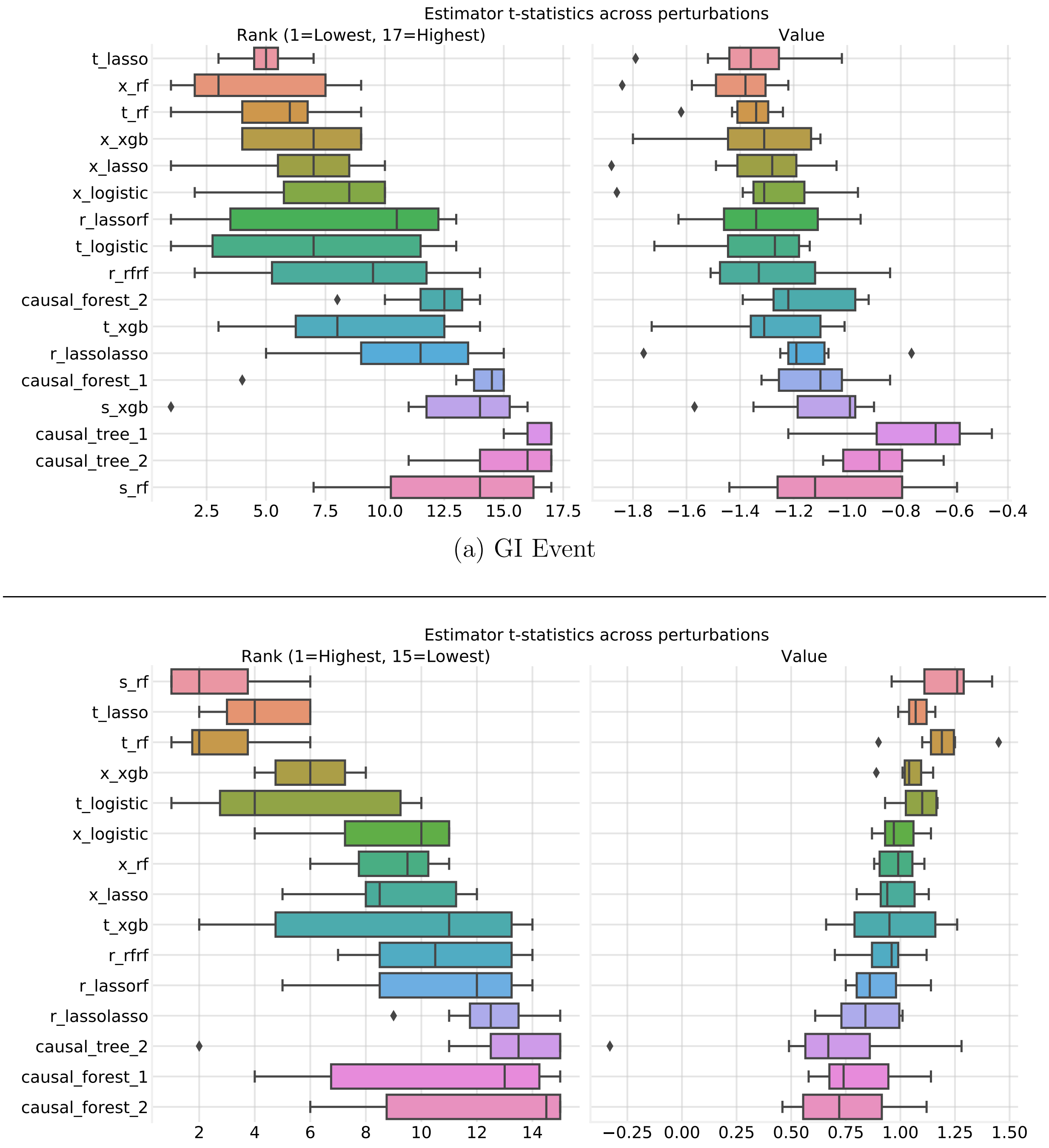
Perturbation $\mathfrak{D}$	cv_orig	cv_0	cv_1	cv_time	elderly_60	overweight	pert_outcome
Estimator $\mathbf{M}$					$\overline{\mathbb{T}}_{\mathrm{GI}}(\mathfrak{D})$		
t_lasso	-1.27	-1.79	**-1.52**	-1.36	-1.36	-1.02	-1.24
x_rf	-1.24	-1.84	-1.37	**-1.58**	-1.40	-1.22	-1.38
t_rf	-1.25	-1.62	-1.39	-1.34	-1.34	**-1.24**	**-1.43**
x_xgb	-1.16	-1.80	-1.44	-1.45	-1.31	-1.11	-1.10
x_lasso	-1.23	**-1.88**	-1.49	-1.33	-1.28	-1.04	-1.15
x_logistic	-1.31	-1.86	-1.39	-1.26	-1.31	-0.96	-1.06
r_lassorf	-1.26	-1.34	-1.36	-1.56	**-1.63**	-0.95	-0.96
t_logistic	**-1.33**	-1.72	-1.56	-1.14	-1.27	-1.17	-1.19
r_rfrf	-1.24	-1.45	-1.33	-1.51	-1.50	-1.00	-0.84
causal_forest_2	-1.00	-1.32	-1.39	-1.23	-1.22	-0.94	-0.92
t_xgb	-1.02	-1.73	-1.18	-1.31	-1.38	-1.01	-1.34
r_lassolasso	-1.10	-1.76	-1.25	-1.19	-1.19	-1.07	-0.76
causal_forest_1	-0.97	-1.26	-1.25	-1.10	-1.07	-0.84	-1.32
s_xgb	-0.95	-1.35	-1.57	-0.99	-1.02	-0.90	-0.99
causal_tree_1	-0.67	-1.22	-0.98	-0.50	-0.66	-0.80	-0.46
causal_tree_2	-1.07	-0.87	-0.72	-0.96	-1.09	-0.88	-0.64
s_rf	-0.78	-1.44	-0.81	-1.19	-1.33	-0.59	-1.12

(a) GI Event

Perturbation $\mathfrak{D}$	cv_orig	cv_0	cv_1	cv_time	elderly_60	overweight	pert_outcome
Estimator $\mathbf{M}$					$\overline{\mathbb{T}}_{\mathrm{CVT}}(\mathfrak{D})$		
s_rf	0.96	**1.29**	**1.17**	**1.42**	**1.29**	1.05	1.26
t_lasso	1.06	1.16	0.99	1.02	1.10	1.07	1.14
t_rf	**1.10**	1.19	0.90	1.25	1.24	**1.18**	**1.45**
x_xgb	1.01	1.15	0.89	1.03	1.08	1.04	1.11
t_logistic	**1.10**	1.16	1.03	1.17	1.17	0.93	1.02
x_logistic	0.97	1.11	0.87	0.94	1.14	0.92	1.01
x_rf	0.90	1.11	0.88	0.91	1.09	0.99	1.02
x_lasso	0.92	1.13	0.80	0.90	1.10	0.94	1.03
t_xgb	0.66	1.06	0.92	1.26	0.95	0.66	1.26
r_rfrf	0.86	1.12	0.70	1.01	0.88	0.96	0.97
r_lassorf	0.79	1.14	0.75	0.93	0.86	1.03	0.81
r_lassolasso	0.81	1.01	0.65	0.61	1.01	0.84	0.98
causal_tree_2	0.67	0.88	0.84	-0.33	0.64	0.49	1.28
causal_forest_1	0.93	1.14	0.96	0.74	0.58	0.64	0.71
causal_forest_2	0.46	0.72	0.87	0.55	0.56	0.96	1.12

(b) CVT Event

Table 9.2: Estimator- *and* perturbation-wise t-statistic $\overline{\mathbb{T}}_{\mathrm{GI}}(\mathfrak{D})$ (9.12a) for the GI event in panel **(a)**, and $\overline{\mathbb{T}}_{\mathrm{CVT}}(\mathfrak{D})$ (9.12b) for the CVT event in panel **(b)**. In each column the best (lowest for GI event, highest for CVT event) t-statistic is highlighted in bold. The order of the estimators in panel (a) and (b) is the same order as that in Fig. 9.5(a) and Fig. 9.5(b) respectively.



(a) GI Event

(b) CVT Event

Figure 9.5: Box plots of the rank and value of mean t-statistic scores $\overline{\mathbb{T}}_{\mathrm{GI}}(\mathfrak{D})$ (9.12a), and $\overline{\mathbb{T}}_{\mathrm{CVT}}(\mathfrak{D})$ (9.12b), where the distribution is over the 7 data perturbations $\mathfrak{D} \in$ {`cv_orig`, `cv_0`, `cv_1`, `cv_time`, `elderly_60`, `overweight`, `pert_outcome`}. Here rank for the mean t-statistic score is computed per perturbation $\mathfrak{D}$, and all CATE estimators are ranked lowest to highest for the GI event, and highest to lowest for the CVT event. The estimator- and perturbation-wise numbers for both panels are reported in Table 9.2.

Bottom quantile	**GI Event**	
based subgroup $\widetilde{\mathbf{G}}_{\mathsf{q}}$	$\mathbb{T}_{\widetilde{\mathbf{G}}_{\mathsf{q}}}$	Overlap
$\mathsf{q} = 0.1$	-1.32 (0.20)	73%
$\mathsf{q} = 0.2$	**-1.58** (0.19)	77%
$\mathsf{q} = 0.3$	-1.47 (0.16)	82%
$\mathsf{q} = 0.4$	-1.02 (0.12)	83%
$\mathsf{q} = 0.5$	-0.81 (0.12)	**87**%

Top quantile	**CVT Event**	
based subgroup $\widetilde{\mathbf{G}}^{c}_{\mathsf{q}}$	$\mathbb{T}_{\widetilde{\mathbf{G}}^{c}_{\mathsf{q}}}$	Overlap
$\mathsf{q} = 0.9$	**1.28** (0.22)	77%
$\mathsf{q} = 0.8$	1.03 (0.12)	75%
$\mathsf{q} = 0.7$	0.85 (0.12)	77%
$\mathsf{q} = 0.6$	0.71 (0.09)	79%
$\mathsf{q} = 0.5$	0.57 (0.13)	**82**%

Table 9.3: t-statistic for different quantile-based top subgroups of the ensemble CATE estimator. "Overlap" column reports the average % pairwise overlap between the 12 quantile-based top subgroups on the entire training data, namely $\widetilde{\mathbf{G}}_{\mathsf{q}} \cap \mathbf{S}_{\text{TRAIN}}$ for the GI event, and $\widetilde{\mathbf{G}}^{c}_{\mathsf{q}} \cap \mathbf{S}_{\text{TRAIN}}$ for the CVT event. The 12 subgroups correspond 4 each to the 3 CV splits $\{$`cv_orig`, `cv_0`, `cv_1`$\}$.

and interpretable, so as to align them with their own knowledge base, and justify them to patients and regulators.

9.6.1 Interpreting using "cells"

In the work by Murdoch et al. [184], one of us has argued that a key element of interpretability is the notion of relevance. Interpretations need to provide "insight for a particular audience into a chosen domain problem." Since clinical decision rules usually take the form of decision trees, a decision tree is the gold standard for our problem at hand. Each leaf of a decision tree constitutes a subset of the feature space defined by constraining the values of the features occuring along the root-to-leaf path. We call such a subset of a feature space a *cell*[14], and propose to make our quantile-based top subgroups interpretable by approximating it with a union of a few cells, which we call a *cell cover.*[15]

Two remarks are in order. First, we find empirically that no single cell gives a good approximation of quantile-based top subgroups, so we require the additional flexibility of a union of multiple cells. Furthermore, reporting a union of cells is more flexible than reporting a decision tree, because it is not always possible to construct a tree with a given collection of cells as its leaf nodes.[16] Second, by focusing on cells, we recognize the importance of interactions, or in other words, nonlinear dependence of treatment effect on the covariates. Chernozhukov et al. [53] proposed interpreting quantile-based top subgroups by estimating the differences in the "observed characteristics" between the quantile-based top subgroup and the subgroup that is defined to be least affected by the treatment, but this only considers the marginal importance of each feature.

[14] This term is motivated by the geometric interpretation of such subsets as subcubes of the hypercube that comprises the entire feature space.

[15] One may also think of this as a disjunction of conjunctions.

[16] For instance, leaf nodes will always involve the feature that splits the root node.

9.6.2 Cell-search methodology

In this section, we demonstrate a general framework for how to search for a cell cover that contains most of the individuals in the quantile-based top subgroup, but does not include too many individuals from outside it.

Feature selection: We start by selecting up to 10 features from the original list of 16 features. This is both to make the subsequent steps of cell search more computationally tractable, and also to act as a form of regularization.[17] To do this, we compute feature importance scores in two different ways. (i) Following Chernozhukov et al. [53], we make use of the difference between the mean of the feature values over the quantile-based top subgroup and that over its complement. We refer to this score as the "Logistic" feature importance score. (ii) We train a logistic classifier to predict membership in the quantile-based top subgroup, and make use of the coefficients. In either case, we normalize so that the absolute values of the scores sum to one. We refer to this score as the "Difference" feature importance score. We compute these two types of scores for the ensemble CATE estimators' quantile-based top subgroups selected at the end of Section 9.5.3, namely $\widetilde{\mathbf{G}}_{0.2}$ and $\widetilde{\mathbf{G}}_{0.3}$ for the GI outcome, and $\widetilde{\mathbf{G}}^c_{0.9}$ and $\widetilde{\mathbf{G}}^c_{0.8}$, across the twelve random training-validation splits ($\{$`cv_orig`, `cv_0`, `cv_1`$\}$). For each outcome, we average the feature importance scores across the different splits as well as both choices of the quantile-based top subgroups. The final results are shown in Fig. 9.6.

Ranking the 16 features according to the two measures of feature importance, we select the features that rank among the top 8 under either measure. Note that we choose to make use of both feature importance measures because they have different meanings: While the first score measures the marginal importance of each feature, the second measures its conditional importance. However, the choice of "top 8" was also selected keeping in mind the fact that the top features for the two measures have a high overlap, and we end up selecting 9 and 10 features respectively for the GI and CVT events listed (alphabetically) below:

GI event: CHLGRP, HYPGRP, PNAPRXN, PNSAIDS, PSTRDS, PPH, ELDERLY, OBESE, WHITE

CVT event: ASCGRP, ASPFDA, CHLGRP, PPH, US, ELDERLY, MALE, OBESE, SMOKE, WHITE

Readers may refer to Table 9.1 to remind themselves about the definitions of all the features.

Iterative procedure: We now describe the `CellSearch` procedure for finding the cell cover for a quantile-based top subgroup one cell at a time, with Fig. 9.7 also providing a pictorial explanation. For clarity, we introduce some notation, denoting the quantile-based top subgroup by $\mathbf{G}_{\text{top}}$, and the cell found at the i-th step by $\mathbb{C}_i$. For GI event $\mathbf{G}_{\text{top}}$ takes the form $\widetilde{\mathbf{G}}_{\mathtt{q}}$, and for the CVT event $\widetilde{\mathbf{G}}^c_{\mathtt{q}}$ for suitable choices of $\mathtt{q}$. As before, we will abuse notation, using these symbols to refer to the subgroups and cells as subsets of the feature

[17] The iterative Random Forest [15] algorithm for finding higher-order interactions in genomics data does soft feature selection for precisely these reasons.

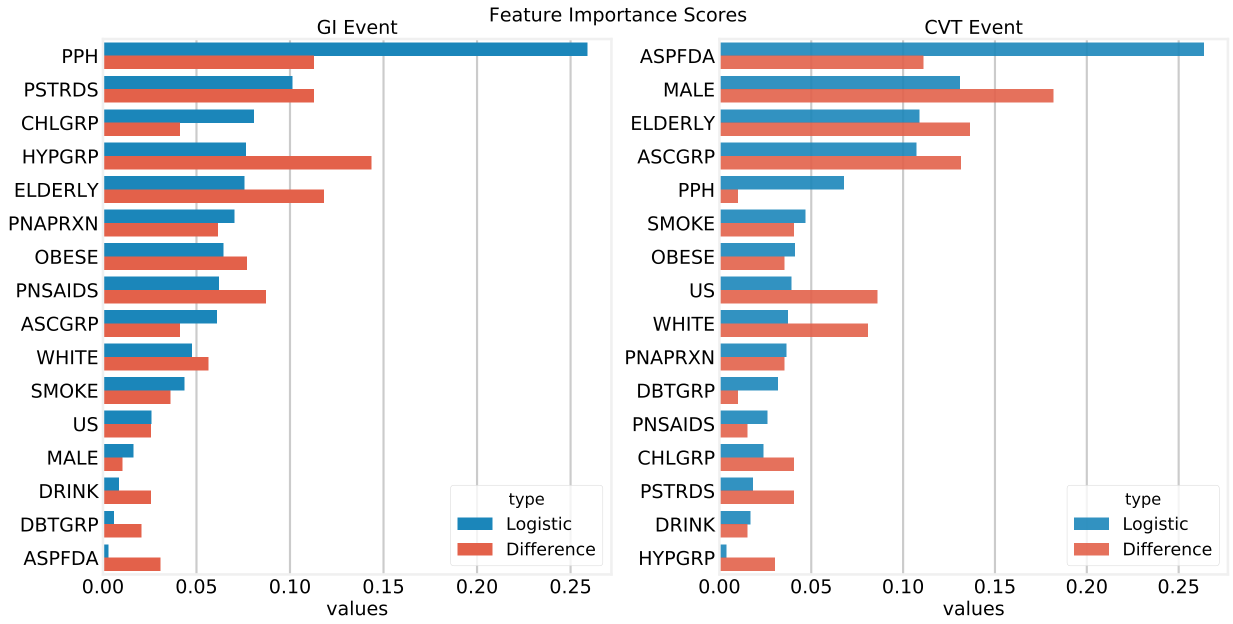


Figure 9.6: Mean feature importance scores for the quantile-based top subgroups from the ensemble CATE estimator. Best seen in color. We plot both the scores next to each other for each feature with the order (top, bottom) = (logistic, difference), but separately for each outcome. The blue bars and red bars respectively denote the "Logistic" and "Difference" feature importance scores described in the text.

space, as well as the subpopulation of individuals that belong to them. At the first step, we consider every possible cell $\mathbb{C}$ defined with m features or less, where m is a user-specified tuning parameter, and compute its "true positive" (TP) and "false positive" (FP) values with respect to $\mathbf{G}_{\text{top}}$ as follows:

$$\texttt{TP}(\mathbb{C}, \mathbf{G}_{\text{top}}) := \operatorname{abs} \mathbb{C} \cap \mathbf{G}_{\text{top}}, \text{ and } \texttt{FP}(\mathbb{C}, \mathbf{G}_{\text{top}}) := \operatorname{abs} \mathbb{C} \cap \mathbf{G}^{c}_{\text{top}}, \tag{9.13}$$

which we are able to compute efficiently using the FPGrowth algorithm [104]. Moreover, let $\Delta(\mathbb{C}, \mathbf{G}_{\text{top}}) := \texttt{TP}(\mathbb{C}, \mathbf{G}_{\text{top}}) - \texttt{FP}(\mathbb{C}, \mathbf{G}_{\text{top}})$ denote the difference of these values.

We rank the cells based on their difference score $\Delta(\mathbb{C}, \mathbf{G}_{\text{top}})$, but instead of simply picking the cell achieving the largest positive value $\Delta_{\max}$, we first create a candidate list of cells for which $\Delta(\mathbb{C}, \mathbf{G}_{\text{top}}) \geq \max(0, \Delta_{\max} - 0.05 \operatorname{abs} \mathbf{G}_{\text{top}})$, remove from cells any that are sub-cells[18] of other cells on this list, and then choose one of remaining cells uniformly at random. The returns on adding this layer of complexity are to favor simpler, more interpretable cells, and also (by running the procedure multiple times) to discover if two or more cells have comparable performance.[19]

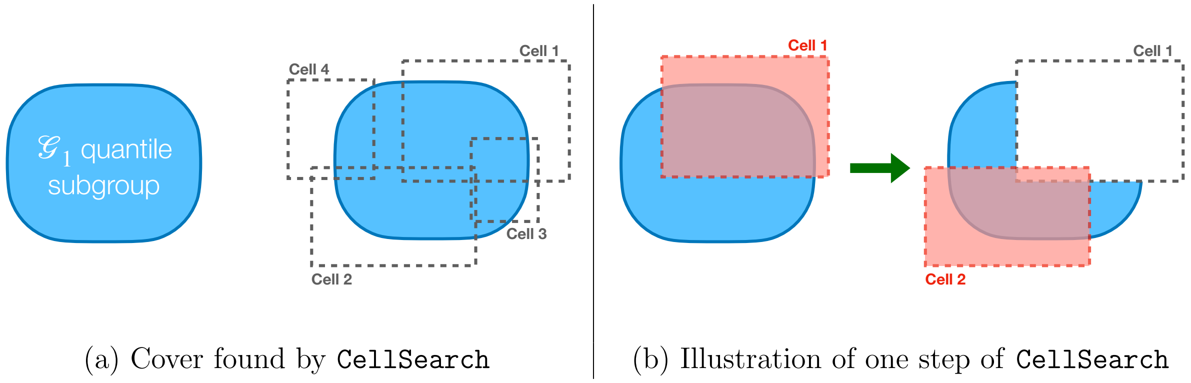
[18] We say that Cell A is a sub-cell of Cell B if it is contained in Cell A when both are though as subsets of the feature space.

[19] A user may wish to simply follow the greedy procedure.

In each subsequent step of the algorithm, to find the next cell in the cell cover, we first remove from the study population all individuals belonging to the cells already found, and then repeat the above process. More rigorously, suppose cells $\mathbb{C}_1, \ldots, \mathbb{C}_{i-1}$ have already been determined. The true and false positive scores are now defined by

$$\texttt{TP}(\mathbb{C}, \mathbf{G}_{\text{top}}; \cup_{j=1}^{i-1}\mathbb{C}_j) := \operatorname{abs}\mathbb{C} \cap \mathbf{G}_{\text{top}} \setminus \cup_{j=1}^{i-1} \mathbb{C}_j, \text{ and } \texttt{FP}(\mathbb{C}, \mathbf{G}_{\text{top}}; \cup_{j=1}^{i-1}\mathbb{C}_j) := \operatorname{abs}\mathbb{C} \cap \mathbf{G}_{\text{top}}^c \setminus \cup_{j=1}^{i-1} \mathbb{C}_j, \tag{9.14}$$

while $\Delta_{\max}$ and the threshold are also modified accordingly. Finally, the procedure terminates if $\Delta_{\max}$ at any iteration is less than or equal to 0 or if the number of iterations has reached a pre-specified threshold (default value 3).



(a) Cover found by `CellSearch`

(b) Illustration of one step of `CellSearch`

Figure 9.7: A simplified illustration of `CellSearch` methodology for finding a cell-based cover for a given (quantile-based) subgroup.

Aggregating results over multiple runs: In accordance with the Stability principle, we run `CellSearch` multiple times, and check whether the same cell cover is found. In our case, we ran it five times on each top quantile subgroup arising from 12 random training-validation splits, for a total of 60 runs. While the cell cover did not turn out to be stable, we found that certain cells or their sub-cells frequently re-appeared within each run. We thus turn our focus to individual cells, and aggregate the results over the multiple runs, calling this procedure `StabilizedCellSearch`.

To describe how we aggregate the results, we first use $\mathcal{B}$ to denote the collection of all 60 runs, and for each run $b \in \mathcal{B}$, we let $\mathfrak{C}_b$ denote the cover returned by the procedure, while the collection of all cells found is denoted $\mathfrak{C} := \cup_{b\in\mathfrak{B}}\mathfrak{C}_b$. For each cell $\mathbb{C} \in \mathfrak{C}_b$, we define its stability score as follows:

$$\texttt{Stab}(\mathbb{C}) = \frac{1}{\operatorname{abs}\mathcal{B}} \sum_{b\in\mathcal{B}} \sum_{\mathbb{C}'\in\mathcal{C}} \mathbf{1}(\mathbb{C}' \in \mathfrak{C}_b \text{ and } \mathbb{C}' \text{ is sub-cell of } \mathbb{C}) \frac{\operatorname{abs}\mathbb{C}'}{\operatorname{abs}\mathbb{C}}. \tag{9.15}$$

This score measures how frequently cell $\mathbb{C}$ and its proper sub-cells are found across the different runs, with each occurrence weighted by the relative size of the sub-cell.

Finally, we rank the cells according to their stability scores, and output those for which the score exceeds a user-defined threshold. In our case, we chose the threshold to be 1/3 which results in finding 3 cells each for the GI and CVT outcomes. We discuss these cells in the next section, while the full results obtained by running `StabilizedCellSearch` on the VIGOR data with respect to both the GI and CVT outcomes is shown in Table H.2.

9.6.3 Discussion of cells found and performance on test set

In this section, we discuss the statistical significance of the cells found for both GI and CVT outcomes. First, we list the top 3 cells found for each outcome, where detailed results for top 20 cells (sroted by `Stab`-scores) are reported in Table H.2. For the GI outcome, the top 3 stable cells are:

(vi) $\mathbb{C}_1$: Patients with prior history of GI Event denoted as {PPH=1},

(vii) $\mathbb{C}_2$: patients who (self) reported a prior (to the experiment) usage of steroids, and a history of hypertension denoted as {PSTRDS=1, HYPGRP=1}, and

(viii) $\mathbb{C}_3$: Elderly patients who reported a prior usage of steroid drugs denoted as {PSTRDS=1, ELDERLY=1}.

For the CVT outcome, they are:

(ix) $\widetilde{\mathbb{C}}_1$: Patients for which use of Aspirin has been indicated as per FDA guidelnes {ASPFDA=1},

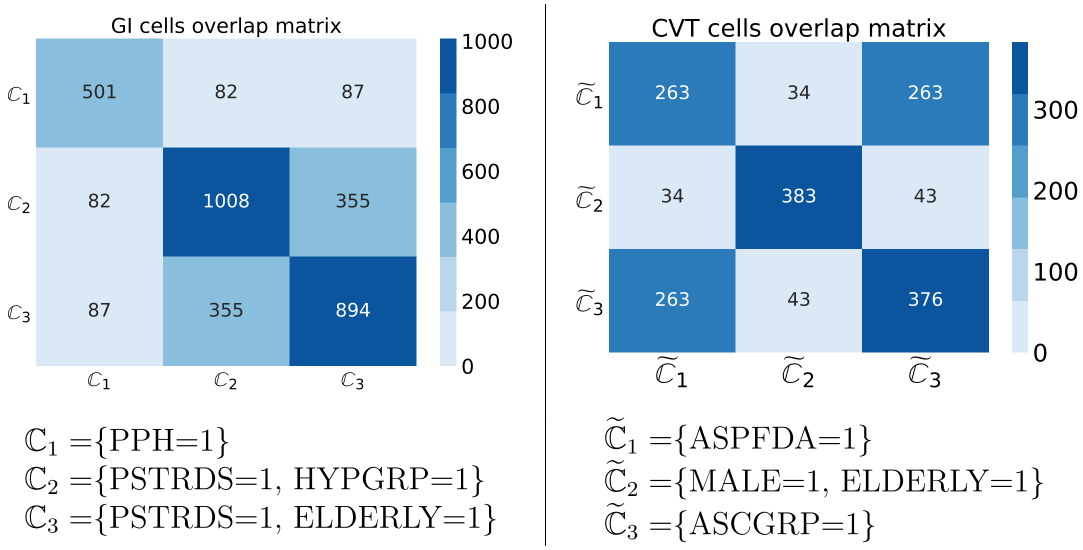
(x) $\widetilde{\mathbb{C}}_2$: Male elderly patients {MALE=1,ELDERLY=1}, and

(xi) $\widetilde{\mathbb{C}}_3$: Patients that have reported prior history {ASCGRP=1}.

For further details on the features appearing above, please refer back to Section 9.2.2. In Fig. 9.8, we plot the overlap between these cells.

Conclusions from Fig. 9.8: As can be seen in Fig. 9.8(a), there is little to moderate overlap among the cells $\mathbb{C}_1$ and $\mathbb{C}_3$, which shows that they are meaningfully different. On the other hand, there is significant overlap among the cells $\widetilde{\mathbb{C}}_1, \widetilde{\mathbb{C}}_3$ in Fig. 9.8(b). In particular, $\widetilde{\mathbb{C}}_1$ is a subset (but not a sub-cell) of $\widetilde{\mathbb{C}}_3$. The reason we report both cells is because of the suspected multi-scale nature of treatment effect variation for the CVT outcome, with $\widetilde{\mathbb{C}}_1$ found more often for $\mathfrak{q} = 0.9$, and $\widetilde{\mathbb{C}}_3$ found more often for $\mathfrak{q} = 0.8$.

We now compute and report several quantities for each of these 6 cells, finally making use of the holdout test dataset (20% of the study size) for the *very first time*. For cells $\mathbb{C}_1, \mathbb{C}_2$ and $\mathbb{C}_3$, as well as the union $\cup_{j=1}^{3}\mathbb{C}_j$ of these 3 cells, the results are reported in Table 9.4. Similar results for the cells $\widetilde{\mathbb{C}}_1, \widetilde{\mathbb{C}}_2$, and $\widetilde{\mathbb{C}}_3$ and their union $\cup_{j=1}^{3}\widetilde{\mathbb{C}}_j$ are reported in Table 9.5. We now discuss the results from Tables 9.4 and 9.5 one by one.



	$\mathbb{C}_1$	$\mathbb{C}_2$	$\mathbb{C}_3$
$\mathbb{C}_1$	501	82	87
$\mathbb{C}_2$	82	1008	355
$\mathbb{C}_3$	87	355	894

$\mathbb{C}_1 = \{\text{PPH=1}\}$
$\mathbb{C}_2 = \{\text{PSTRDS=1, HYPGRP=1}\}$
$\mathbb{C}_3 = \{\text{PSTRDS=1, ELDERLY=1}\}$

	$\widetilde{\mathbb{C}}_1$	$\widetilde{\mathbb{C}}_2$	$\widetilde{\mathbb{C}}_3$
$\widetilde{\mathbb{C}}_1$	263	34	263
$\widetilde{\mathbb{C}}_2$	34	383	43
$\widetilde{\mathbb{C}}_3$	263	43	376

$\widetilde{\mathbb{C}}_1 = \{\text{ASPFDA=1}\}$
$\widetilde{\mathbb{C}}_2 = \{\text{MALE=1, ELDERLY=1}\}$
$\widetilde{\mathbb{C}}_3 = \{\text{ASCGRP=1}\}$

(a) GI cells

(b) CVT cells

Figure 9.8: Overlap matrix for final discovered cells on the training data $\mathbf{S}_{\text{TRAIN}}$. For panel **(a)** the data split is stratified on the treatment indicator and the GI outcome, and that for **(b)** is stratified on on the treatment indicator and the CVT outcome. For instance, the number 82 for the entry corresponding to $\mathbb{C}_1$ and $\mathbb{C}_2$ in panel (a) represents that the two cells had 82 patients in common on the training data.

Results from Table 9.4: In the first three rows of Table 9.4, we examine the subgroup treatment effect for these cells with respect to the GI outcome. In the second and third columns, we report two versions of the Neyman estimate for the cell CATE $\widehat{\tau}_{\mathbb{C}\cap\mathbf{S}}$, one computed on the training set $\mathbf{S}_{\text{TRAIN}}$ as well as one computed on the test set $\mathbf{S}_{\text{TEST}}$. Likewise, in the next two columns, we report the t-statistic $\mathbb{T}_{\mathbf{G}\cap\mathbf{S}}$, one computed on the training set $\mathbf{S}_{\text{TRAIN}}$, and on the test set $\mathbf{S}_{\text{TEST}}$. Finally, in the last column with header $^{\dagger}\mathbf{S}_{\text{VAL}}$, we report the mean (and standard deviation in parenthesis) of the t-statistics $\mathbb{T}$ computed on the 12 different folds of $\mathbf{S}_{\text{TRAIN}}$ from the 3 random CV splits $\{$`cv_orig`, `cv_0`, `cv_1`$\}$. Overall, the test set results are promising, with test set CATE estimates being much more negative than the estimated ATE, and comparable to their training set counterparts. While we do not report p-values because they can be easily misunderstood, we note that the test set t-statistic values for the GI outcome are $\mathbb{C}_3$, and the union $\cup_{j=1}^{3}\mathbb{C}_j$, are both significant at the 0.025 level for a one-sided z-test.

The starting point of our investigation of the VIGOR dataset was the hope to identify a subgroup for which Vioxx simultaneously has a strong negative treatment effect for GI risk and a low positive treatment effect for CVT risk. Consequently, in the last three rows of Table 9.4, we report the treatment effect results for the cells $\{\mathbb{C}_j\}_{j=1}^{3}$ and their union, with respect to the CVT outcome. While $\mathbb{C}_2$ and $\mathbb{C}_3$ experience increased CVT risk, $\mathbb{C}_1 =$

	#evts/size		**CATE Est.** $\widehat{\tau}_{\mathbb{C}\cap\mathbf{S}}$ (std)		**t-statistic** $\mathbb{T}_{\mathbb{C}\cap\mathbf{S}}$		
Dataset S	$\mathbf{S}_{\text{TRAIN}}$	$\mathbf{S}_{\text{TEST}}$	$\mathbf{S}_{\text{TRAIN}}$	$\mathbf{S}_{\text{TEST}}$	$\mathbf{S}_{\text{TRAIN}}$	$\mathbf{S}_{\text{TEST}}$	$^{\dagger}\mathbf{S}_{\text{VAL}}$
Cell $\mathbb{C}$							
GI Event (GI-stratified split)							
PPH=1	36/501	8/129	-0.057 (0.023)	-0.055 (0.042)	-1.89	-1.01	-0.99 (0.27)
PSTRDS=1, HYPGRP=1	39/1008	6/238	-0.050 (0.012)	-0.037 (0.021)	-3.17	-1.06	-1.57 (0.22)
PSTRDS=1, ELDERLY=1	46/894	9/227	-0.051 (0.015)	-0.063 (0.026)	-2.74	-2.00	-1.38 (0.17)
Union	79/1905	19/471	-0.038 (0.009)	-0.047 (0.018)	-3.15	-2.22	-1.59 (0.20)
All	142/6460	35/1616	-0.016 (0.004)	-0.016 (0.007)	-	-	-
CVT Event (entire data)							
PPH=1	2/630		-0.006 (0.004)		-2.66		
PSTRDS=1, HYPGRP=1	11/1246		0.008 (0.005)		0.44		
PSTRDS=1, ELDERLY=1	16/1121		0.015 (0.007)		1.42		
Union	21/2376		0.007 (0.004)		0.55		
All	59/8076		0.006 (0.002)		-		

Table 9.4: Results for the final cells selected after `StabilizedCellSearch` for the GI event, namely $\mathbb{C}_1 = \{\text{PPH=1}\}$, $\mathbb{C}_2 = \{\text{PSTRDS=1,HYPGRP=1}\}$ and $\mathbb{C}_3 = \{\text{PSTRDS=1,ELDERLY=1}\}$ from Section 9.6.3. We also report the results for the other outcome, namely CVT event, on the entire data (all 8076 patients). In the column $^{\dagger}\mathbf{S}_{\text{VAL}}$, we report the mean t-statistics and standard deviation in parentheses, across the 12 different folds of the training data $\mathbf{S}_{\text{TRAIN}}$ obtained from the 3 random CV splits {`cv_orig`, `cv_0`, `cv_1`}.

$\{\text{PPH} = 1\}$ in fact shows reduced CVT risk, which makes it especially promising for further clinical investigation. We note that for the CVT outcome we report the CATE estimates and the t-statistic on the entire data as this outcome had no role to play in the entire StaDISC pipeline with the GI outcome, and hence the entire data can be treated as a "valid" test set for estimating heterogeneous treatment effect of Vioxx with the CVT outcome.

	#evts/size		**CATE Est.** $\widehat{\tau}_{\mathbb{C}\cap\mathbf{S}}$ (std)		**t-statistic** $\mathbb{T}_{\mathbb{C}\cap\mathbf{S}}$		
Dataset S	$\mathbf{S}_{\text{TRAIN}}$	$\mathbf{S}_{\text{TEST}}$	$\mathbf{S}_{\text{TRAIN}}$	$\mathbf{S}_{\text{TEST}}$	$\mathbf{S}_{\text{TRAIN}}$	$\mathbf{S}_{\text{TEST}}$	$^{\dagger}\mathbf{S}_{\text{VAL}}$
Cell $\mathbb{C}$							
CVT Event (CVT-stratifed split)							
ASPFDA=1	13/263	5/58	0.062 (0.025)	0.103 (0.074)	2.28	1.38	1.09 (0.20)
MALE=1, ELDERLY=1	12/383	0/111	0.040 (0.017)	0 (0)	2.09	-1.16	0.85 (0.24)
ASCGRP=1	15/376	6/78	0.044 (0.020)	0.047 (0.060)	2.05	0.74	1.04 (0.23)
Union	24/716	6/175	0.042 (0.013)	0.024 (0.028)	3.09	0.77	1.55 (0.13)
All	**47/6460**	12/1616	0.006 (0.002)	0.005 (0.004)	-	-	-
GI Event (entire data)							
ASPFDA=1	6/321		-0.027 (0.016)		-0.71		
MALE=1, ELDERLY=1	17/494		-0.045 (0.016)		-1.85		
ASCGRP=1	8/454		-0.028 (0.013)		-0.96		
Union	25/891		-0.040 (0.011)		-2.27		
All	177/8076		-0.016 (0.003)		-		

Table 9.5: Results for the final cells selected after `StabilizedCellSearch` for the CVT event, namely $\widetilde{\mathbb{C}}_1 = \{\text{ASPFDA=1}\}$, $\widetilde{\mathbb{C}}_2 = \{\text{MALE=1,ELDERLY=1}\}$ and $\widetilde{\mathbb{C}}_3 = \{\text{ASCGRP=1}\}$ from Section 9.6.3. We also report the results for the other outcome, namely GI event, on the entire data (all 8076 patients). In the column $^{\dagger}\mathbf{S}_{\text{VAL}}$, we report the mean t-statistics and standard deviation in parentheses, across the 12 different folds of the training data $\mathbf{S}_{\text{TRAIN}}$ obtained 4 each from the 3 random CV splits {`cv_orig`, `cv_0`, `cv_1`}.

Results from Table 9.5: In Table 9.5, we report the analogous results for cells $\widetilde{\mathbb{C}}_1, \widetilde{\mathbb{C}}_2$, and $\widetilde{\mathbb{C}}_3$, and their union $\cup_{j=1}^{3}\widetilde{\mathbb{C}}_j$, first for the CVT outcome, and then the GI outcome. For these cells, the generalization to the holdout test set is weaker, with only $\widetilde{\mathbb{C}}_1$ and $\widetilde{\mathbb{C}}_3$ having test set CATE values that remain substantially positive. Furthermore, the test set t-statistic values are smaller. All these observations are unsurprising given the rarity of the CVT outcome—in particular, only 12/1616 individuals in the test set $\mathbf{S}_{\text{TEST}}$ experienced an event. Nonetheless, the test set CVT-CATE estimates for $\widetilde{\mathbb{C}}_1$ and $\widetilde{\mathbb{C}}_3$ support the view that the treatment effect is stronger on these subgroups, while the GI-CATE estimates do not suggest that these subgroups benefit especially strongly from the treatment with Vioxx.

9.7 Complementary analysis with the APPROVe study

It is well-documented that RCTs have problems with *external validity* [135, 220, 91, 144], which is defined by Rothwell to be "whether the results can be reasonably applied to a definable group of patients in a particular clinical setting in routine practice" [220]. This phenomenon arises primarily because RCTs have carefully defined enrollment criteria, so conclusions in such studies may not apply to patients who do not conform to these criteria. In more mathematical language, the ATE, subgroup CATE, and other estimands of interest are all defined in terms of expectations with respect to a particular distribution of patients, a particular outcome, and a particular treatment, and hence do not directly apply when any of these change. We refer the interested reader to the excellent articles by Rothwell [220, 221] for a further discussion on these topics.

Despite its importance for clinical relevance, external validity has been relatively neglected by researchers and institutions overseeing the conduct of RCTs) [220, 144]. One way to argue for external validity is to attempt *external validation*, i.e. to reproduce the results obtained on one data set on a different but related data set. Recent voices that urge the community to give external valideation a higher priority across many domains [66, 144, 197] are very much in accordance with Yu and Kumbier's [259] call to statisticians to broaden the scope of their concern from data-modeling to the entire data science life cycle as part of the PCS framework. This can be seen not only as one more predictive and stability check under the PCS framework, but also as a special case of "transfer learning" where the desiderata is the transferability of the conclusions or findings from one dataset to other related datasets.

These reasons motivate the following complementary analysis of the APPROVe study [12], another RCT investigating Vioxx. More precisely, we compute the subgroup CATEs with respect to both the GI and CVT outcomes over this new data set, and show that the *qualitative* conclusions obtained by applying StaDISC to the VIGOR study also generalize to this data set for four out of the six subgroups from Fig. 9.8; the other two subgroups were too small in size and did not have any GI events. We now start with a background on the APPROVe study followed by a discussion of the results on subgroup CATEs.

9.7.1 Background for the APPROVe study

In this section, we provide only a brief background for the APPROVe study and refer the readers to the original paper [12] for additional details.

The Adenomatous Polyp Prevention on Vioxx (APPROVe) study was another randomized trial sponsored by Merck, but unlike VIGOR, it was placebo-controlled. Conducted in 2001-2004, it was designed to assess whether Vioxx could "reduce the risk of adenomatous polyps in individuals with a recent history of these tumours" [12]. The study population comprised 2587 patients who had colon adenomatous polyps removed during a 12 week period before being entered into the study, and who had no known polyps remaining. After discovering that Vioxx had significant cardiovascular toxicity, the study was terminated two months early in September 2004, but all individuals were followed-up for at least a year afterwards off-treatment.

The data files of the APPROVe study followed a very similar format to that of the VIGOR study albeit with two major differences: (i) GI event was not directly labeled in the dataset, and (ii) the risk factor file was not available. As a result, outcomes related to the GI event, and features (including but not limited to) ASPFDA, ASCGRP, HYPGRP, PSTRDS—which were used to define the final subgroups obtained in the previous section—were not directly available for APPROVe. However, with the data available to us, we were able to impute the GI outcome and the missing relevant features (used for the cells reported in Tables 9.4 and 9.5). The data cleaning and imputation were done before looking at the final results. The details for this data cleaning are provided in Appendix H.2, and the distribution of the selected features and the two outcomes is reported in Table H.3. Once we have the features and the outcomes, we compute the subgroup CATE (9.3b) and t-statistics (9.11a) and report the results in Table 9.6.

9.7.2 Results with the APPROVe study

Before presenting the quantitative results, we make a few remarks. In direct analogy with the problems with external validity mentioned earlier, there are several ways in which the causal estimands in APPROVe differ from those in VIGOR. First, the "control" arm of both studies were of entirely different natures: while VIGOR was a comparison between Vioxx and Naproxen, APPROVe compared Vioxx with a placebo. Second, the lengths of both studies were different, which is important because our estimands are defined in terms of accumulated risk over the duration of the study. Patients in VIGOR were followed for a median time of 9 months, whereas most patients in APPROVe were tracked for at least 4 years. Furthermore, while GI events were adjudicated in VIGOR, this was not the case for APPROVe. Lastly, the study populations are different. As elaborated earlier in Section 9.2, the VIGOR study comprised patients who were diagnosed with rheumatoid arthritis. On the other hand, APPROVe comprised patients with a recent history of colon polyps. Furthermore, unlike VIGOR, APPROVe excluded patients likely needing regular NSAID treatment, but allowed for concomitant low-dose aspirin therapy.

Table 9.6 describes the quantitative results for the final subgroups (from Section 9.6.3) for the APPROVe study. For the reasons explained in the previous paragraph, we do not expect the subgroup CATE estimands to be the same across the two studies. However, comparing the results across Tables 9.4 to 9.6, it is reassuring that the subgroups we found for the VIGOR study continue to be meaningful for APPROVe in illustrating the heterogeneity of treatment effects. We now discuss the results first for the CVT outcome followed by that for the GI outcome as the interpretation of the results for the latter is a bit more subtle.

Cell $\mathbb{C}$	**#evts/size**	**CATE Est.** $\widehat{\tau}_{\mathbb{C}\cap\mathbf{S}}$ (std)	**t-statistic** $\mathbb{T}_{\mathbb{C}\cap\mathbf{S}}$
GI Event with $\mathbf{S}$ = *all data*			
PPH=1	6/184	0.066 (0.026)	2.012
PSTRDS=1, HYPGRP=1	0/30	-	-
PSTRDS=1, ELDERLY=1	0/21	-	-
All	33/2587	0.016 (0.004)	-
CVT Event with $\mathbf{S}$ = *all data*			
ASPFDA=1	13/151	0.107 (0.043)	2.128
MALE=1, ELDERLY=1	30/416	0.069 (0.025)	2.251
ASCGRP=1	17/250	0.068 (0.031)	1.664
Union (of 3 cells above)	41/588	0.065 (0.021)	2.650
PPH=1	4/184	0.022 (0.022)	0.119
All	89/2587	0.020 (0.007)	-

Table 9.6: Results for the subgroups found with StaDISC on VIGOR, for the APPROVe dataset. Note that unlike VIGOR, the patients in the control arm for the APPROVe study were treated with a placebo, which makes the quantitative results reported here not directly comparable with that reported in Tables 9.4 and 9.5. Refer to the text for further discussion. The armwise statistics of the features and outcomes for the APPROVe study are provided in Table H.3.

Results for the CVT outcome: We note that the three subgroups {ASPFDA=1}, {MALE=1, ELDERLY=1}, {ASCGRP=1}, and the union of these 3 subgroups all had subgroup CATEs that were much larger than the ATE, with t-statistics that were significant at the 0.05 level for a one-sided z-test, even after accounting for multiple-testing (refer to end of this section for further discussions related to multiple-testing.) Overall these results provide evidence for the heterogeneous treatment effects of Vioxx for the CVT outcomes over these subgroups, namely that Vioxx disproportionately increases the CVT event risk for these subgroups when compared to either Naproxen or a placebo. To be consistent with the earlier results in Table 9.4, we also computed the subgroup CATE for {PPH=1} for the CVT outcome and (like the VIGOR study) did not find any evidence for a disproportionate increase in the risk for the CVT event compared with the entire population.

Recall that, the found increase in risk for VIGOR was relative to Naproxen. This observation alone may suggest a possibility that Vioxx was not the cause of the observed increase in

CVT events, and the positive ATE could have resulted due to a protective effect of Naproxen reducing them. Merck, the manufacturer of Vioxx, interpreted the CVT signal in VIGOR as being a consequence of a hitherto unknown protective effect of Naproxen, rather than a deleterious consequence of Vioxx. The CVT signal in the APPROVe study associated with Vioxx relative to placebo conclusively confirmed that Vioxx can have deleterious consequences. Moreover, both VIGOR and APPROVe study suggest that Vioxx has significant heterogeneity in how it increases the risk for CVT events for different subgroups.

Results for the GI outcome: As noted above, additional care is required to interpret the CATE results for the GI outcome. Whereas Naproxen was known to have GI toxicity, and was shown in VIGOR to increase the risk of GI events more than Vioxx, a placebo by definition does not have any toxicity. As such, our finding that treatment with Vioxx had a positive estimated ATE (1.6%) with the GI outcome in the APPROVe study does not contradict our earlier reporting of a negative ATE with respect to the GI outcome (-1.6%) in the VIGOR study. In fact, this discovery is surprising insofar as Vioxx was initially believed to have minimal if any, GI toxicity whatsoever [148].

We found the subgroup {PPH=1} to have a large positive estimated subgroup CATE (6.6%) resulting in a t-statistic score significant at the 0.025 level for a one-sided z-test (without correcting for multiple-testing.) As discussed above, this result does not contradict the negative CATE value of -5.7% (or -5.5% for the test set) estimated for the VIGOR study (see Table 9.4). We furthermore note that the GI event rates over both arms in VIGOR, and the Vioxx arm in APPROVe were all elevated compared to the entire population. The corresponding rates for the placebo in the APPROVe study were fairly similar (0% for {PPH=1} and 0.4% on average.)

We summarize our finding across the two studies as follows. (i) VIGOR study: Vioxx, in comparison to Naproxen, reduced the GI Toxicity disproportionately for the subgroup {PPH=1} when compared to the the average. (ii) APPROVe study: Vioxx, in comparison to the placebo, increases the GI Toxicity disproportionately for the subgroup {PPH=1} when compared to the average. Nonetheless, the conclusion that the estimated subgroup CATE for {PPH=1} was significantly different than the estimated ATE is *consistent* across the two studies.

Finally, due to the difference in the study population, two out of the three subgroups for the GI event reported in Table 9.4, namely {PSTRDS=1, HYPGRP=1} and {PSTRDS=1, ELDERLY=1}, were too small in size and had no GI events.[20] Consequently, it does not make sense to quantify the subgroup CATE for these subgroups.

[20] Indeed, comparing Tables 9.1 and H.3, we can attribute the discrepancy in these subgroups' sizes between the two studies to the smaller population of patients (74/2587) with a history of using glucocorticoids (PSTRDS = 1) in the APPROVe study versus that of the much larger population of such patients (4479/8076) in the VIGOR study.

Multiple-testing with FWER control: Given enough data points in the APPROVe study, we also perform corrected multiple hypothesis testing using Holm-Bonforreni procedure controlling family wise error rate (FWER) at level 0.05. Overall, we test 5 null hypotheses, that the subgroup CATE is equal to the average treatment effect for the following cases: (i) $\mathbb{C}_1$ = {PPH=1} for the GI event, (ii) $\widetilde{\mathbb{C}}_1$ ={ASPFDA=1}, (iii) $\widetilde{\mathbb{C}}_2$ ={MALE=1, ELDERLY=1}, (iv) $\widetilde{\mathbb{C}}_3$ ={ASCGRP=1}, and (v) the union $\cup_{i=1}^{3}\widetilde{\mathbb{C}}_i$—where the treatment effect in subgroups (ii)-(v) corresponds to the CVT event. The t-statistics for these hypotheses (sorted by magnitude) as reported in Table 9.6 are 2.650, 2.251, 2.128, 2.012 and 1.664, and thereby the corresponding one-sided p-values are 0.004, 0.012, 0.0167, 0.022 and 0.048. The corrected procedure for significance level 0.05 compares these sorted p-values with the cut-offs 0.01, 0.0125, 0.0167, 0.025 and 0.05. On doing so we find that all five hypotheses are rejected, and thus we conclude all the subgroups (i)-(v) have statistically significant heterogeneous treatment effect.[21]

9.8 Conclusion and future directions

In this chapter, we have made three major contributions: (I) We have re-analyzed a dataset from the 1999-2000 VIGOR study, an RCT of 8076 patients, and found three clinically relevant subgroups each for the GI outcome (total size 29.4%), and the CVT outcome (total size 11.0%), for which the treatment drug Vioxx has significantly large estimated treatment effect when compared to that from the estimated ATE. We provided external evidence for the significance of the heterogeneous treatment effects for four out of the six subgroups through a complementary analysis of the 2001-2004 APPROVe study, another RCT of 2587 patients. (II) Our work is an illustration of how clinical trial data can be analyzed to provide a basis for differential treatment decisions in subgroups in order to optimize outcomes, and how the findings can be validated with another study. We call this novel methodology StaDISC, and develop it by building on the PCS framework [259], the calibration literature, and recent developments in CATE estimation. (III) Our work introduces the PCS framework to the causal inference community, and provides a template for a more informative understanding of heterogeneous treatment effects.

An important point to note is that the notions of estimated treatment effects ATE, CATE and subgroup CATE (defined in equation (9.1)) used in this work and more broadly in CATE estimation, measure the *difference* in the adverse event risk in the treatment group to that in the control group. However, when investigating the efficacy of medical interventions, medical professionals are often more interested in relative risk, which measures the *ratio* of

[21] Note that, for the APPROVe study, we did not test for heterogeneity in the subgroups {PSTRDS=1, HYPGRP=1}, and {PSTRDS=1, ELDERLY=1} due to their small size in this study. Since the size of the subgroup can only be observed once we know the group membership of the patients, our testing procedure and the associated discoveries can be considered as being conditional on observing the group membership, and treatment variable for all the patients in the APPROVe study. In other words, the statistical significance is over the randomness in the outcome, and the conditional randomness in the covariates given the group membership indicators.

the two risks. This alternate conception of treatment effect in terms of relative risk changes the meaning of heterogeneity. For instance, the subgroup $\mathbb{C}_1$ {PPH=1} has a relative risk of 0.43 with respect to GI events, which is barely any different than the population relative risk of 0.46. On the other hand, because the baseline risk of individuals in this subgroup is far higher than the rest of the population, the subgroup CATE is similarly inflated.

We do not attempt to debate which notion of heterogeneity is better since it is context-dependent. Nevertheless, given the popularity of relative risk in the medical literature, in our future work we plan to develop a formal framework for subgroup discovery with respect to relative risk by adapting generic CATE estimation methods, and consequently extend StaDISC for relative risk estimation.

There are several other extensions of StaDISC that remain interesting future directions. First, StaDISC is currently motivated and defined for randomized experiments. We intend to formulate a statistical framework that would also make it applicable to observational studies. Second, the cell search step of StaDISC only works with binary features. One can either propose to incorporate continuous features through either careful binary encoding using quantile-thresholding, or through amending the cell search procedure. Third, we have thus far applied StaDISC to the GI and CVT outcomes in the VIGOR study one at a time and a joint investigation with multiple outcomes, even more generally, is an interesting future direction.

Bibliography

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. (Cited on page 50.)

[2] A. Alaa and M. Van Der Schaar. Validating causal inference models via influence functions. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 191–201, Long Beach, California, USA, 09–15 Jun 2019. PMLR. (Cited on page 142.)

[3] B. J. Alder and T. E. Wainwright. Studies in molecular dynamics. I. General method. *The Journal of Chemical Physics*, 31(2):459–466, 1959. (Cited on page 50.)

[4] D. Aldous and J. A. Fill. Reversible Markov chains and random walks on graphs, 2002. Unfinished monograph, recompiled 2014, available at `http://www.stat.berkeley.edu/~aldous/RWG/book.html`. (Cited on page 43.)

[5] N. Altieri, R. L. Barter, J. Duncan, R. Dwivedi, K. Kumbier, X. Li, R. Netzorg, B. Park, C. Singh, Y. S. Tan, T. Tang, Y. Wang, C. Zhang, and B. Yu. Curating a covid-19 data repository and forecasting county-level death counts in the united states. *Harvard Data Science Review*, 2 2021. https://hdsr.mitpress.mit.edu/pub/p6isyf0g. (Cited on page 6.)

[6] J. Angrist and J. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008. (Cited on page 140.)

[7] K. M. Anstreicher. The volumetric barrier for semidefinite programming. *Mathematics of Operations Research*, 25(3):365–380, 2000. (Cited on page 94.)

[8] S. Athey. *The Impact of Machine Learning on Economics*, pages 507–547. University of Chicago Press, January 2018. (Cited on page 140.)

[9] S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27):7353–7360, 2016. (Cited on pages 141, 142, 143, 153, and 154.)

[10] S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, 45:77–120, 2017. (Cited on pages 96, 98, 99, 102, 107, 110, 112, 113, 118, 129, 133, 135, and 137.)

[11] N. M. Ballarini, G. K. Rosenkranz, T. Jaki, F. König, and M. Posch. Subgroup identification in clinical trials via the predicted individual treatment effect. *PloS one*, 13(10):e0205971, 2018. (Cited on page 142.)

[12] J. A. Baron, R. S. Sandler, R. S. Bresalier, A. Lanas, D. G. Morton, R. Riddell, E. R. Iverson, and D. L. DeMets. Cardiovascular events associated with Rofecoxib: Final analysis of the APPROVe trial. *The Lancet*, 372(9651):1756–1764, 2008. (Cited on pages 177, 381, and 382.)

[13] F. Barthe and B. Maurey. Some remarks on isoperimetry of Gaussian type. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 36, pages 419–434. Elsevier, 2000. (Cited on page 251.)

[14] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. (Cited on pages 100 and 113.)

[15] S. Basu, K. Kumbier, J. B. Brown, and B. Yu. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, 115(8):1943–1948, 2018. (Cited on page 170.)

[16] C. J. Bélisle, H. E. Romeijn, and R. L. Smith. Hit-and-run algorithms for generating multivariate distributions. *Mathematics of Operations Research*, 18(2):255–266, 1993. (Cited on pages 15 and 70.)

[17] D. Bertsimas and S. Vempala. Solving convex programs by random walks. *Journal of the ACM (JACM)*, 51(4):540–556, 2004. (Cited on page 70.)

[18] A. Beskos, N. Pillai, G. Roberts, J.-M. Sanz-Serna, and A. Stuart. Optimal tuning of the hybrid monte carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 2013. (Cited on page 54.)

[19] M. Betancourt, S. Byrne, and M. Girolami. Optimizing the integrator step size for Hamiltonian Monte Carlo. *arXiv preprint arXiv:1411.6669*, 2014. (Cited on page 53.)

[20] R. Bhatia. *Matrix Analysis*, volume 169. Springer Science & Business Media, 2013. (Cited on page 268.)

[21] A. Bloniarz, H. Liu, C.-H. Zhang, J. S. Sekhon, and B. Yu. Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27):7383–7390, 2016. (Cited on pages 5, 141, 143, and 153.)

[22] S. G. Bobkov. Isoperimetric and analytic inequalities for log-concave probability measures. *The Annals of Probability*, 27(4):1903–1921, 1999. (Cited on page 250.)

[23] C. Bombardier, L. Laine, A. Reicin, D. Shapiro, R. Burgos-Vargas, B. Davis, R. Day, M. B. Ferraz, C. J. Hawkey, M. C. Hochberg, T. K. Kvien, and T. J. Schnitzer. Comparison of upper gastrointestinal toxicity of Rofecoxib and Naproxen in patients with rheumatoid arthritis. VIGOR study group. *New England Journal of Medicine*, 343(21):1520–1528, 2000. (Cited on pages 145, 146, 147, 148, and 381.)

[24] N. Bou-Rabee, A. Eberle, and R. Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *arXiv preprint arXiv:1805.00452*, 2018. (Cited on pages 53 and 54.)

[25] N. Bou-Rabee and M. Hairer. Nonasymptotic mixing of the MALA algorithm. *IMA Journal of Numerical Analysis*, 33(1):80–110, 2012. (Cited on page 17.)

[26] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. (Cited on pages 73, 81, and 214.)

[27] L. Breiman. Statistical modeling: The two cultures (with discussion). *Statist. Sci*, 16(3):16199–231, 2001. (Cited on pages 143 and 144.)

[28] P. Brémaud. *Markov Chains, Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, 1991. (Cited on page 2.)

[29] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950. (Cited on page 154.)

[30] S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011. (Cited on pages 9, 12, and 50.)

[31] S. Bubeck, R. Eldan, and J. Lehec. Sampling from a log-concave distribution with projected Langevin Monte Carlo. *arXiv preprint arXiv:1507.02564*, 2015. (Cited on page 18.)

[32] S. Bubeck et al. Convex optimization: algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015. (Cited on page 23.)

[33] P. J. Bushell. Hilbert's metric and positive contraction mappings in a Banach space. *Archive for Rational Mechanics and Analysis*, 52(4):330–338, 1973. (Cited on page 89.)

[34] T. Cai, L. Tian, P. H. Wong, and L. Wei. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12(2):270–282, 2011. (Cited on pages 5, 141, and 142.)

[35] T. T. Cai, J. Ma, and L. Zhang. CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267, 2019. (Cited on pages 4, 96, 98, and 134.)

[36] C. Carini, S. M. Menon, and M. Chang. *Clinical and Statistical Considerations in Personalized Medicine*. CRC Press, 2014. (Cited on page 142.)

[37] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017. (Cited on page 50.)

[38] C. Carvalho, A. Feller, J. Murray, S. Woody, and D. Yeager. Assessing treatment effect variation in observational studies: Results from a data challenge. *Observational Studies*, 5(2):21–35, 2019. (Cited on page 143.)

[39] J. Cheeger. A lower bound for the smallest eigenvalue of the Laplacian. In *Proceedings of the Princeton Conference in honor of Professor S. Bochner*, 1969. (Cited on page 224.)

[40] H. Chen, J. Chen, and J. D. Kalbfleisch. A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63:19–29, 2001. (Cited on page 124.)

[41] H. Chen, T. Harinen, J.-Y. Lee, M. Yung, and Z. Zhao. Causalml: Python package for causal machine learning, 2020. (Cited on page 153.)

[42] J. Chen. Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23(1):221–233, 1995. (Cited on pages 97, 99, 109, 111, 124, and 348.)

[43] J. Chen and P. Li. Hypothesis test for normal mixture models: The EM approach. *Annals of Statistics*, 37:2523–2542, 2009. (Cited on pages 123, 124, and 125.)

[44] L. Chen, Z. Qin, and J. S. Liu. Exploring hybrid monte carlo in bayesian computation. *sigma*, 2:2–5, 2001. (Cited on page 54.)

[45] Y. Chen. An almost constant lower bound of the isoperimetric coefficient in the kls conjecture. *Geometric and Functional Analysis*, 31(1):34–61, 2021. (Cited on page 48.)

[46] Y. Chen, R. Dwivedi, M. J. Wainwright, and B. Yu. Vaidya walk: A sampling algorithm based on the volumetric barrier. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1220–1227. IEEE, 2017. (Cited on page 3.)

[47] Y. Chen, R. Dwivedi, M. J. Wainwright, and B. Yu. Fast MCMC sampling algorithms on polytopes. *The Journal of Machine Learning Research*, 19(1):2146–2231, 2018. (Cited on pages 3, 71, and 207.)

[48] Y. Chen, R. Dwivedi, M. J. Wainwright, and B. Yu. Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *The Journal of Machine Learning Research*, 21:92–1, 2020. (Cited on page 3.)

[49] Y. C. Chen. Statistical inference with local optima. *arXiv preprint arXiv:1807.04431*, 2018. (Cited on page 98.)

[50] Z. Chen and S. S. Vempala. Optimal convergence rate of Hamiltonian Monte Carlo for strongly logconcave distributions. *arXiv preprint arXiv:1905.02313*, 2019. (Cited on page 53.)

[51] X. Cheng and P. Bartlett. Convergence of Langevin MCMC in KL-divergence. pages 186–211, 2018. (Cited on pages 16, 18, 21, 22, and 56.)

[52] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. *arXiv preprint arXiv:1707.03663*, 2017. (Cited on pages 18 and 259.)

[53] V. Chernozhukov, M. Demirer, E. Duflo, and I. Fernandez-Val. Generic machine learning inference on heterogenous treatment effects in randomized experiments. Technical report, National Bureau of Economic Research, 2018. (Cited on pages 140, 142, 169, and 170.)

[54] S. Chewi, C. Lu, K. Ahn, X. Cheng, T. L. Gouic, and P. Rigollet. Optimal dimension dependence of the metropolis-adjusted langevin algorithm. *arXiv preprint arXiv:2012.12810*, 2020. (Cited on pages 36 and 42.)

[55] S. Chrétien and A. O. Hero. On EM algorithms and their proximal generalizations. *ESAIM: Probability and Statistics*, 12:308–326, 2008. (Cited on page 96.)

[56] F. S. Collins and H. Varmus. A new initiative on precision medicine. *New England Journal of Medicine*, 372(9):793–795, 2015. PMID: 25635347. (Cited on page 4.)

[57] B. Cousins and S. Vempala. A cubic algorithm for computing Gaussian volume. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on discrete algorithms*, pages 1215–1228. Society for Industrial and Applied Mathematics, 2014. (Cited on pages 70 and 207.)

[58] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991. (Cited on pages 210 and 233.)

[59] D. R. Cox. Planning of experiments. 1958. (Cited on page 140.)

[60] M. Creutz. Global Monte Carlo algorithms for many-fermion systems. *Physical Review D*, 38(4):1228, 1988. (Cited on page 52.)

[61] A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016. (Cited on pages 16, 18, 21, 22, 24, 25, 29, 31, 33, and 56.)

[62] A. S. Dalalyan and A. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 2019. (Cited on page 56.)

[63] A. S. Dalalyan and L. Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020. (Cited on page 18.)

[64] 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. (Cited on pages 4, 96, 98, 99, 102, 110, and 118.)

[65] A. P. Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982. (Cited on pages 152 and 154.)

[66] T. P. Debray, Y. Vergouwe, H. Koffijberg, D. Nieboer, E. W. Steyerberg, and K. G. Moons. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of clinical epidemiology*, 68(3):279–289, 2015. (Cited on page 177.)

[67] M. H. DeGroot and S. E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983. (Cited on pages 152 and 154.)

[68] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39:1–38, 1997. (Cited on pages 3, 96, and 98.)

[69] P. Diaconis and D. Freedman. On Markov chains with continuous state space. Technical report, Technical Report, 1997. (Cited on pages 15 and 17.)

[70] P. Diaconis, L. Saloff-Coste, et al. Logarithmic Sobolev inequalities for finite Markov chains. *The Annals of Applied Probability*, 6(3):695–750, 1996. (Cited on page 43.)

[71] I. Dikin. Iterative solution to problems of linear and quadratic programming. *Doklady Akademii Nauk SSSR*, 174(4):747, 1967. (Cited on page 72.)

[72] P. Ding, X. Li, and L. W. Miratrix. Bridging finite and super population causal inference. *Journal of Causal Inference*, 5(2), 2017. (Cited on page 151.)

[73] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987. (Cited on page 50.)

[74] A. Durmus and E. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019. (Cited on pages 16, 18, and 206.)

[75] A. Durmus, E. Moulines, and M. Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: When Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018. (Cited on page 18.)

[76] A. Durmus, E. Moulines, and E. Saksman. On the convergence of Hamiltonian Monte Carlo. *arXiv preprint arXiv:1705.00166*, 2017. (Cited on pages 52 and 53.)

[77] E. Dusseldorp and I. Van Mechelen. Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Statistics in medicine*, 33(2):219–237, 2014. (Cited on page 142.)

[78] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183):1–42, 2019. (Cited on pages 3, 24, 45, 47, 48, 53, 56, and 247.)

[79] R. Dwivedi, O. N. Feldheim, O. Gurel-Gurevich, and A. Ramdas. The power of online thinning in reducing discrepancy. *Probability Theory and Related Fields*, 174(1):103–131, 2019. (Cited on page 5.)

[80] R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, and M. I. Jordan. Theoretical guarantees for EM under misspecified Gaussian mixture models. In *NeurIPS 31*, 2018. (Cited on pages 5 and 99.)

[81] R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu. Sharp analysis of expectation-maximization for weakly identifiable models. *AISTATS*, 2020. (Cited on page 4.)

[82] R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu. Singularity, misspecification, and the convergence rate of EM. *The Annals of Statistics*, 48(6):3161–3182, 2020. (Cited on pages 4 and 129.)

[83] R. Dwivedi and L. Mackey. Kernel thinning. *arXiv preprint arXiv:2105.05842*, 2021. (Cited on page 5.)

[84] R. Dwivedi, Y. S. Tan, B. Park, M. Wei, K. Horgan, D. Madigan, and B. Yu. Stable discovery of interpretable subgroups via calibration in causal studies. *International Statistical Review*, 88:S135–S178, 2020. (Cited on page 5.)

[85] M. Dyer, A. Frieze, and R. Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM (JACM)*, 38(1):1–17, 1991. (Cited on page 15.)

[86] A. Eberle. Error bounds for Metropolis-Hastings algorithms applied to perturbations of Gaussian measures in high dimensions. *The Annals of Applied Probability*, 24(1):337–377, 2014. (Cited on pages 18 and 53.)

[87] B. Efron. Prediction, estimation, and attribution. *Journal of the American Statistical Association*, 115(530):636–655, 2020. (Cited on pages 144 and 164.)

[88] J. Feldman, M. J. Wainwright, and D. R. Karger. Using linear programming to decode binary linear codes. *IEEE Transactions on Information Theory*, 51(3):954–972, 2005. (Cited on page 70.)

[89] A. Feller and C. C. Holmes. Beyond toplines: Heterogeneous treatment effects in randomized experiments. *Unpublished manuscript, Oxford University*, 2009. (Cited on pages 4, 5, and 141.)

[90] R. A. Fisher. Design of experiments. *Br Med J*, 1(3923):554–554, 1936. (Cited on page 140.)

[91] M. Fortin, J. Dionne, G. Pinho, J. Gignac, J. Almirall, and L. Lapointe. Randomized controlled trials: Do they have external validity for patients with multiple comorbidities? *The Annals of Family Medicine*, 4(2):104–108, 2006. (Cited on page 177.)

[92] J. C. Foster, J. M. Taylor, and S. J. Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011. (Cited on pages 5, 141, 142, and 153.)

[93] A. Frieze, R. Kannan, and N. Polson. Sampling from log-concave distributions. *The Annals of Applied Probability*, pages 812–837, 1994. (Cited on page 13.)

[94] A. Gelman and E. Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking'. *Unpublished draft*, 2013. (Cited on pages 142 and 163.)

[95] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. PAMI*, 6:721–741, 1984. (Cited on page 2.)

[96] A. S. Gerber, D. P. Green, and C. W. Larimer. Social pressure and voter turnout: Evidence from a large-scale field experiment. *American political Science review*, pages 33–48, 2008. (Cited on pages 4, 5, and 141.)

[97] S. Ghosal and A. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Annals of Statistics*, 29:1233–1263, 2001. (Cited on page 97.)

[98] W. R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, pages 337–348, 1992. (Cited on page 8.)

[99] S. Goel, R. Montenegro, and P. Tetali. Mixing time bounds via the spectral profile. *Electronic Journal of Probability*, 11:1–26, 2006. (Cited on pages 43, 45, 46, 216, 217, 218, 222, 223, 224, and 226.)

[100] U. Grenander and M. I. Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 549–603, 1994. (Cited on page 16.)

[101] M. Gromov and V. D. Milman. A topological application of the isoperimetric inequality. *American Journal of Mathematics*, 105(4):843–854, 1983. (Cited on page 62.)

[102] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. *34th International Conference on Machine Learning, ICML 2017*, 3:2130–2143, 2017. (Cited on pages 152 and 154.)

[103] A. Gustafson and H. Narayanan. John's walk. *arXiv preprint arXiv:1803.02032*, 2018. (Cited on pages 75 and 77.)

[104] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. *SIGMOD Rec.*, 29(2):1–12, May 2000. (Cited on page 171.)

[105] B. Hao, W. W. Sun, Y. Liu, and G. Cheng. Simultaneous clustering and estimation of heterogeneous graphical models. *Journal of Machine Learning Research*, 18(217):1–58, 2018. (Cited on pages 4, 96, and 98.)

[106] G. Hargé. A convex/log-concave correlation inequality for Gaussian measure and an application to abstract Wiener spaces. *Probability theory and related fields*, 130(3):415–440, 2004. (Cited on page 206.)

[107] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. (Cited on pages 2 and 12.)

[108] Z. V. Havre, N. White, J. Rousseau, and K. Mengersen. Overfitting Bayesian mixture models with an unknown number of components. *PLOS One*, 10, 2015. (Cited on page 4.)

[109] P. Heinrich and J. Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *Annals of Statistics*, 46:2844–2870, 2018. (Cited on page 97.)

[110] S. Hernández-Díaz and L. A. G. Rodríguez. Steroids and risk of upper gastrointestinal complications. *American journal of epidemiology*, 153(11):1089–1093, 2001. (Cited on page 147.)

[111] N. Ho, K. Khamaru, R. Dwivedi, M. J. Wainwright, M. I. Jordan, and B. Yu. Instability, computational efficiency and statistical accuracy. *arXiv preprint arXiv:2005.11411*, 2020. (Cited on page 6.)

[112] N. Ho and X. Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44:2726–2755, 2016. (Cited on page 124.)

[113] N. Ho and X. Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 2016. (Cited on pages 126, 138, and 376.)

[114] N. Ho and X. Nguyen. Singularity structures and impacts on parameter estimation in finite mixtures of distributions. Technical Report 540, Department of Statistics, University of Michigan, 2016. (Cited on page 99.)

[115] M. D. Hoffman and A. Gelman. The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014. (Cited on pages 52 and 54.)

[116] P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986. (Cited on pages 140 and 141.)

[117] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2012. (Cited on page 268.)

[118] C. Houdré. Mixed and isoperimetric estimates on the log-Sobolev constants of graphs and Markov chains. *Combinatorica*, 21(4):489–513, 2001. (Cited on page 224.)

[119] D. Hsu, S. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012. (Cited on page 206.)

[120] K.-L. Huang and S. Mehrotra. An empirical evaluation of walk-and-round heuristics for mixed integer linear programs. *Computational Optimization and Applications*, 55(3):545–570, 2013. (Cited on page 70.)

[121] K.-L. Huang and S. Mehrotra. An empirical evaluation of a walk-relax-round heuristic for mixed integer convex programs. *Computational Optimization and Applications*, 60(3):559–585, 2015. (Cited on page 70.)

[122] C. Huber, N. Benda, and T. Friede. A comparison of subgroup identification methods in clinical drug development: Simulation study and regulatory considerations. *Pharmaceutical statistics*, 18(5):600–626, 2019. (Cited on page 142.)

[123] K. Imai and M. Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics*, 7(1):443–470, 2013. (Cited on page 153.)

[124] G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015. (Cited on pages 141 and 152.)

[125] G. W. Imbens and J. M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009. (Cited on pages 4, 5, and 141.)

[126] T. Inglot. Inequalities for quantiles of the chi-square distribution. *Probability and Mathematical Statistics*, 30(2):339–351, 2010. (Cited on page 333.)

[127] H. Ishwaran, L. F. James, and J. Sun. Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96:1316–1332, 2001. (Cited on page 111.)

[128] L. Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918. (Cited on pages 283 and 319.)

[129] S. Janson. *Gaussian Hilbert Spaces*, volume 129. Cambridge university press, 1997. (Cited on pages 282 and 333.)

[130] S. F. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic processes and their applications*, 85(2):341–361, 2000. (Cited on page 15.)

[131] M. Jerrum and A. Sinclair. Conductance and the rapid mixing property for Markov chains: The approximation of permanent resolved. In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, pages 235–244. ACM, 1988. (Cited on page 43.)

[132] C. Jin, Y. Zhang, S. Balakrishnan, M. J. Wainwright, and M. I. Jordan. Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems 29*, 2016. (Cited on page 4.)

[133] F. John. Extremum problems with inequalities as subsidiary conditions. In *Traces and emergence of nonlinear programming*, pages 197–215. Springer, 2014. (Cited on page 290.)

[134] M. I. Jordan and L. Xu. Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks*, 8, 1995. (Cited on page 96.)

[135] P. Jüni, D. G. Altman, and M. Egger. Assessing the quality of controlled clinical trials. *Bmj*, 323(7303):42–46, 2001. (Cited on page 177.)

[136] R. Kannan, L. Lovász, and R. Montenegro. Blocking conductance and mixing in random walks. *Combinatorics, Probability and Computing*, 15(4):541–570, 2006. (Cited on pages 43, 46, 71, 77, 247, 248, 250, and 252.)

[137] R. Kannan, L. Lovász, and M. Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete & Computational Geometry*, 13(3-4):541–559, 1995. (Cited on pages 15, 37, and 249.)

[138] R. Kannan, L. Lovász, and M. Simonovits. Random walks and an o*(n5) volume algorithm for convex bodies. *Random Structures & Algorithms*, 11(1):1–50, 1997. (Cited on page 71.)

[139] R. Kannan and H. Narayanan. Random walks on polytopes and an affine interior point method for linear programming. *Mathematics of Operations Research*, 37(1):1–20, 2012. (Cited on pages 70, 71, 72, 73, 76, 77, 82, 86, and 88.)

[140] S. C. Kapfer and W. Krauth. Sampling from a polytope and hard-disk Monte Carlo, 2013. (Cited on page 70.)

[141] E. H. Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020. (Cited on page 141.)

[142] J. M. Klusowski, D. Yang, and W. Brinda. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Transactions on Information Theory*, 65(6):3515–3524, 2019. (Cited on page 98.)

[143] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006. (Cited on pages 100 and 113.)

[144] A. Krauss. Why all randomised controlled trials produce biased results. *Annals of medicine*, 50(4):312–322, 2018. (Cited on page 177.)

[145] H. M. Krumholz, J. S. Ross, A. H. Presler, and D. S. Egilman. What have we learnt from Vioxx? *Bmj*, 334(7585):120–123, 2007. (Cited on page 147.)

[146] R. Kumar and M. Schmidt. Convergence rate of Expectation-Maximization. *10th NIPS Workshop on Optimization for Machine Learning*, 2017. (Cited on pages 98 and 118.)

[147] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116(10):4156–4165, 2019. (Cited on pages 140, 141, 143, 152, and 153.)

[148] L. Laine, S. Harper, T. Simon, R. Bath, J. Johanson, H. Schwartz, S. Stern, H. Quan, J. Bolognese, R. O. E. S. Group, et al. A randomized trial comparing the effect of rofecoxib, a cyclooxygenase 2–specific inhibitor, with that of ibuprofen on the gastroduodenal mucosa of patients with osteoarthritis. *Gastroenterology*, 117(4):776–783, 1999. (Cited on page 180.)

[149] J. Lawrence. Polytope volume computation. *Mathematics of Computation*, 57(195):259–271, 1991. (Cited on page 70.)

[150] M. Ledoux. Concentration of measure and logarithmic Sobolev inequalities. In *Seminaire de Probabilites XXXIII*, pages 120–216. Springer, 1999. (Cited on page 248.)

[151] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991. (Cited on pages 336 and 365.)

[152] Y. T. Lee and A. Sidford. Path finding methods for linear programming: Solving linear programs in $\tilde{O}(\sqrt{\text{rank}})$ iterations and faster algorithms for maximum flow. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 424–433. IEEE, 2014. (Cited on pages 75, 83, 263, 290, 292, 296, 301, and 310.)

[153] Y. T. Lee, Z. Song, and S. S. Vempala. Algorithmic theory of ODEs and sampling from well-conditioned logconcave densities. *arXiv preprint arXiv:1812.06243*, 2018. (Cited on pages 53, 54, 58, and 259.)

[154] Y. T. Lee and S. S. Vempala. Geodesic walks in polytopes. *arXiv preprint arXiv:1606.04696*, 2016. (Cited on pages 75 and 76.)

[155] Y. T. Lee and S. S. Vempala. Eldan's stochastic localization and the KLS hyperplane conjecture: An improved lower bound for expansion. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 998–1007. IEEE, 2017. (Cited on page 62.)

[156] Y. T. Lee and S. S. Vempala. Convergence rate of riemannian hamiltonian monte carlo and faster polytope volume computation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 1115–1121, 2018. (Cited on pages 53 and 54.)

[157] Y. T. Lee and S. S. Vempala. Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1115–1121. ACM, 2018. (Cited on page 77.)

[158] Y. T. Lee and S. S. Vempala. Stochastic localization+ Stieltjes barrier= tight bound for log-Sobolev. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1122–1129. ACM, 2018. (Cited on pages 46, 71, 72, 77, 247, and 248.)

[159] P. Li, J. Chen, and P. Marriott. Non-finite Fisher information and homogeneity: an EM approach. *Biometrika*, 96:411–426, 2009. (Cited on page 123.)

[160] W. Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics*, 7(1):295–318, 2013. (Cited on page 141.)

[161] I. Lipkovich, A. Dmitrienko, J. Denne, and G. Enas. Subgroup identification based on differential effect search: A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 30(21):2601–2621, 2011. (Cited on pages 141 and 142.)

[162] S. Livingstone, M. Betancourt, S. Byrne, and M. Girolami. On the geometric ergodicity of hamiltonian monte carlo. *Bernoulli*, 25(4A):3109–3138, 2019. (Cited on page 53.)

[163] L. Lovász. Hit-and-run mixes fast. *Mathematical Programming*, 86(3):443–461, 1999. (Cited on pages 15, 37, 38, 47, 70, 71, 86, 89, 90, 207, and 287.)

[164] L. Lovász et al. Random walks on graphs: A survey. *Combinatorics, Paul erdos is eighty*, 2(1):1–46, 1993. (Cited on page 43.)

[165] L. Lovász and R. Kannan. Faster mixing via average conductance. In *Proceedings of the 31st annual ACM Symposium on Theory of Computing*, pages 282–287. ACM, 1999. (Cited on pages 43 and 46.)

[166] L. Lovász and M. Simonovits. The mixing rate of Markov chains, an isoperimetric inequality, and computing the volume. In *Proceedings of 31st Annual Symposium on Foundations of Computer Science, 1990*, pages 346–354. IEEE, 1990. (Cited on pages 15, 43, 45, 70, 71, and 86.)

[167] L. Lovász and M. Simonovits. Random walks in a convex body and an improved volume algorithm. *Random Structures & Algorithms*, 4(4):359–412, 1993. (Cited on pages 9, 15, 43, 45, 86, 248, 249, 286, and 287.)

[168] L. Lovász and S. Vempala. Hit-and-run is fast and fun. *Tehnical Report, Microsoft Research*, 2003. (Cited on page 287.)

[169] L. Lovász and S. Vempala. Hit-and-run from a corner. *SIAM Journal on Computing*, 35(4):985–1005, 2006. (Cited on pages 15, 38, 77, and 86.)

[170] L. Lovász and S. Vempala. Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithm. *Journal of Computer and System Sciences*, 72(2):392–417, 2006. (Cited on page 72.)

[171] L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007. (Cited on pages 15, 38, and 207.)

[172] J. Ma, L. Xu, and M. I. Jordan. Asymptotic convergence rate of the EM algorithm for Gaussian mixtures. *Neural Computation*, 12:2881–2907, 2000. (Cited on pages 4, 96, and 98.)

[173] Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan. Sampling can be faster than optimization. *arXiv preprint arXiv:1811.08413*, 2018. (Cited on page 62.)

[174] M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011. (Cited on page 73.)

[175] O. Mangoubi and A. Smith. Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. *arXiv preprint arXiv:1708.07114*, 2017. (Cited on pages 53 and 54.)

[176] O. Mangoubi and N. K. Vishnoi. Dimensionally tight bounds for second-order Hamiltonian Monte Carlo. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6030–6040, 2018. (Cited on pages 51, 53, 54, 56, 58, 257, and 259.)

[177] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. (Cited on pages 8 and 12.)

[178] S. P. Meyn and R. L. Tweedie. Computable bounds for geometric convergence rates of Markov chains. *The Annals of Applied Probability*, pages 981–1011, 1994. (Cited on page 17.)

[179] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, 2012. (Cited on pages 9, 16, and 17.)

[180] R. Michel, I. Schnakenburg, and T. von Martens. *Targeting Uplift: An Introduction to Net Scores*. Springer International Publishing, 2019. (Cited on page 4.)

[181] R. G. Miller. Statistical prediction by discriminant analysis. In *Statistical Prediction by Discriminant Analysis*, pages 1–54. Springer, 1962. (Cited on page 154.)

[182] M. Molina and F. Garip. Machine learning for sociology. *Annual Review of Sociology*, 45(1):27–45, 2019. (Cited on page 140.)

[183] B. Morris and Y. Peres. Evolving sets, mixing and heat kernel bounds. *Probability Theory and Related Fields*, 133(2):245–266, 2005. (Cited on pages 43 and 46.)

[184] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019. (Cited on page 169.)

[185] A. H. Murphy. A new vector partition of the probability score. *Journal of applied Meteorology*, 12(4):595–600, 1973. (Cited on page 154.)

[186] M. P. Naeini, G. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. (Cited on page 154.)

[187] H. Narayanan. Randomized interior point methods for sampling and optimization. *The Annals of Applied Probability*, 26(1):597–641, 2016. (Cited on pages 38, 70, 71, 72, 82, 86, 94, and 207.)

[188] H. Narayanan and A. Rakhlin. Efficient sampling from time-varying log-concave distributions. *arXiv preprint arXiv:1309.5977*, 2013. (Cited on page 94.)

[189] R. M. Neal. An improved acceptance procedure for the hybrid Monte Carlo algorithm. *Journal of Computational Physics*, 111:194–203, 1994. (Cited on page 50.)

[190] R. M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011. (Cited on pages 51, 52, and 258.)

[191] A. Negassa, A. Ciampi, M. Abrahamowicz, S. Shapiro, and J.-F. Boivin. Tree-structured subgroup analysis for censored survival data: Validation of computationally inexpensive model selection criteria. *Statistics and computing*, 15(3):231–239, 2005. (Cited on page 142.)

[192] Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. Siam, 1994. (Cited on page 72.)

[193] J. Neyman and K. Iwaszkiewicz. Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society*, 2(2):107–180, 1935. (Cited on page 140.)

[194] X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 4(1):370–400, 2013. (Cited on pages 97, 111, and 348.)

[195] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005. (Cited on page 154.)

[196] X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021. (Cited on pages 141, 143, 152, and 153.)

[197] B. Norgeot, G. Quer, B. K. Beaulieu-Jones, A. Torkamani, R. Dias, M. Gianfrancesco, R. Arnaout, I. S. Kohane, S. Saria, E. Topol, et al. Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. *Nature Medicine*, 26(9):1320–1324, 2020. (Cited on page 177.)

[198] R. S. Olson, W. L. Cava, Z. Mustahsan, A. Varik, and J. H. Moore. Data-driven advice for applying machine learning to bioinformatics problems. *Biocomputing 2018*, Nov 2017. (Cited on pages 5 and 143.)

[199] T. Ondra, A. Dmitrienko, T. Friede, A. Graf, F. Miller, N. Stallard, and M. Posch. Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. *Journal of biopharmaceutical statistics*, 26(1):99–119, 2016. (Cited on page 142.)

[200] G. Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180(3):378–384, 1981. (Cited on page 16.)

[201] K. Pearson. Contributions to the theory of mathematical evolution. *Philosophical Transactions of the Royal Society of London A*, 185:71–110, 1894. (Cited on page 3.)

[202] L. R. Peck. Subgroup analysis in social experiments: Measuring program impacts based on post-treatment choice. *American Journal of Evaluation*, 24(2):157–187, 2003. (Cited on page 141.)

[203] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. (Cited on pages 148 and 152.)

[204] M. Pereyra. Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*, 26(4):745–760, 2016. (Cited on page 18.)

[205] N. S. Pillai, A. M. Stuart, and A. H. Thiéry. Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *The Annals of Applied Probability*, 22(6):2320–2356, 2012. (Cited on page 17.)

[206] K. R. Popper. *The Logic of Scientific Discovery.* Basic Books, Oxford, England, 1959. (Cited on page 152.)

[207] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26(2):195–239, 1984. (Cited on pages 96 and 98.)

[208] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59:731–792, 1997. (Cited on page 97.)

[209] B. D. Ripley. *Stochastic Simulation.* Wiley, New York, 1987. (Cited on page 2.)

[210] C. P. Robert. *Monte Carlo methods.* Wiley Online Library, 2004. (Cited on page 12.)

[211] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer texts in statistics. Springer-Verlag, New York, NY, 1999. (Cited on page 50.)

[212] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001. (Cited on page 15.)

[213] G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004. (Cited on pages 9, 15, 18, and 50.)

[214] G. O. Roberts and J. S. Rosenthal. Complexity bounds for Markov chain Monte Carlo algorithms via diffusion limits. *Journal of Applied Probability*, 53(2):410 – 420, 2016. (Cited on page 15.)

[215] G. O. Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357, 2002. (Cited on pages 17 and 50.)

[216] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996. (Cited on pages 16, 17, and 18.)

[217] G. O. Roberts and R. L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996. (Cited on page 15.)

[218] C. A. Rolling and Y. Yang. Model selection for estimating treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):749–769, 2014. (Cited on page 142.)

[219] J. S. Ross, D. Madigan, K. P. Hill, D. S. Egilman, Y. Wang, and H. M. Krumholz. Pooled analysis of Rofecoxib placebo-controlled clinical trial data: Lessons for postmarket pharmaceutical safety surveillance. *Archives of Internal Medicine*, 169(21):1976–1985, 2009. (Cited on page 147.)

[220] P. M. Rothwell. External validity of randomised controlled trials:"to whom do the results of this trial apply?". *The Lancet*, 365(9453):82–93, 2005. (Cited on page 177.)

[221] P. M. Rothwell. Factors that can affect the external validity of randomised controlled trials. *PLOS Clin Trial*, 1(1):e9, 2006. (Cited on page 177.)

[222] J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:689–710, 2011. (Cited on pages 4 and 97.)

[223] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974. (Cited on page 140.)

[224] S. Sachdeva and N. K. Vishnoi. The mixing time of the Dikin walk in a polytope—a simple proof. *Operations Research Letters*, 44(5):630–634, 2016. (Cited on pages 73, 86, and 88.)

[225] A. Schuler, M. Baiocchi, R. Tibshirani, and N. Shah. A comparison of methods for model selection when estimating individual treatment effects, 2018. (Cited on pages 141 and 153.)

[226] Z. Shahn, D. Madigan, et al. Latent class mixture models of treatment effect heterogeneity. *Bayesian Analysis*, 12(3):831–854, 2017. (Cited on page 142.)

[227] A. F. Smith and G. O. Roberts. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(1):3–23, 1993. (Cited on page 9.)

[228] B. Smith. Mamba: Markov chain Monte Carlo (MCMC) for Bayesian analysis in julia, 2014. Software available at mambajl.readthedocs.io. (Cited on page 50.)

[229] J. Splawa-Neyman, D. M. Dabrowska, and T. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990. (Cited on pages 140 and 152.)

[230] M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62:795–809, 2002. (Cited on page 97.)

[231] G. Strang. Inverse problems and derivatives of determinants. *Archive for Rational Mechanics and Analysis*, 114(3):255–265, 1991. (Cited on page 75.)

[232] X. Su, C.-L. Tsai, H. Wang, D. M. Nickerson, and B. Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(2), 2009. (Cited on page 142.)

[233] D. Talay and L. Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic analysis and applications*, 8(4):483–509, 1990. (Cited on page 17.)

[234] L. Tian, A. A. Alizadeh, A. Gelman, and R. Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014. (Cited on pages 5, 141, and 152.)

[235] L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, pages 1701–1728, 1994. (Cited on page 9.)

[236] P. Tseng. An analysis of the EM algorithm and entropy-like proximal point methods. *Mathematics of Operations Research*, 29(1):27–44, 2004. (Cited on page 96.)

[237] P. M. Vaidya. A new algorithm for minimizing convex functions over convex sets. In *30th Annual Symposium on Foundations of Computer Science, 1989*, pages 338–343. IEEE, 1989. (Cited on pages 70 and 73.)

[238] P. M. Vaidya and D. S. Atkinson. A technique for bounding the number of iterations in path following algorithms. In *Complexity in Numerical Optimization*, pages 462–489. World Scientific, 1993. (Cited on page 73.)

[239] S. van de Geer. *Empirical Processes in M-estimation.* Cambridge University Press, 2000. (Cited on pages 97, 100, and 113.)

[240] A. W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, 1998. (Cited on page 348.)

[241] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer-Verlag, New York, NY, 2000. (Cited on pages 335, 336, and 365.)

[242] S. Vempala. Geometric random walks: A survey. *Combinatorial and Computational Geometry*, 52(573-612):2, 2005. (Cited on pages 9, 12, 43, 88, and 207.)

[243] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027v7.* (Cited on page 338.)

[244] C. Villani. *Optimal transport: Old and New.* Springer, 2008. (Cited on page 376.)

[245] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. (Cited on pages 141, 143, 153, and 154.)

[246] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint.* Cambridge University Press, Cambridge, UK, 2019. (Cited on pages 113, 335, 336, and 337.)

[247] R. Wang, S. W. Lagakos, J. H. Ware, D. J. Hunter, and J. M. Drazen. Statistics in medicine — reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, 357(21):2189–2194, 2007. PMID: 18032770. (Cited on page 141.)

[248] Z. Wang, Q. Gu, Y. Ning, and H. Liu. High-dimensional Expectation-Maximization algorithm: Statistical optimization and asymptotic normality. In *Advances in Neural Information Processing Systems 28*, 2015. (Cited on pages 4, 96, and 98.)

[249] Z. Wang, S. Mohamed, and N. Freitas. Adaptive Hamiltonian and Riemann manifold Monte Carlo. In *International Conference on Machine Learning*, pages 1462–1470, 2013. (Cited on page 54.)

[250] C. Wu, J. Stoehr, and C. P. Robert. Faster Hamiltonian Monte Carlo by learning leapfrog scale. *arXiv preprint arXiv:1810.04449*, 2018. (Cited on page 54.)

[251] C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103, 1983. (Cited on pages 3, 96, and 98.)

[252] T. Xifara, C. Sherlock, S. Livingstone, S. Byrne, and M. Girolami. Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statistics & Probability Letters*, 91:14–19, 2014. (Cited on page 18.)

[253] 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. (Cited on pages 4, 96, and 98.)

[254] L. Xu and M. I. Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8:129–151, 1996. (Cited on pages 4, 96, and 98.)

[255] 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. (Cited on pages 4, 96, 98, and 99.)

[256] X. Yi and C. Caramanis. Regularized EM algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems 28*, 2015. (Cited on pages 4, 96, and 98.)

[257] B. Yu. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*, pages 423–435, 1997. (Cited on page 353.)

[258] B. Yu and R. Barter. The data science process: One culture. *Journal of the American Statistical Association*, 115(530):672–674, 2020. (Cited on page 144.)

[259] B. Yu and K. Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 117(8):3920–3929, 2020. (Cited on pages 5, 143, 177, and 181.)

[260] B. Yu and P. Mykland. Looking at Markov samplers through cusum path plots: a simple diagnostic idea. *Statistics and Computing*, 8(3):275–286, 1998. (Cited on page 84.)

[261] P. Zettler. U.S. v. Harkonen: Should scientists worry about being prosecuted for how they interpret their research results? Accessed: June 29, 2020. (Cited on page 142.)

Part IV

Appendices

Appendix A

Content Deferred From Chapter 3

In this appendix, we collect the technical content deferred from the main text. Appendices A.1 to A.4 respectively contain the proofs of Lemmas 3.1 and 3.3 to 3.5. Finally, we collect some basic properties of convex and smooth functions in Appendix A.5.

A.1 Proof of Lemma 3.1

Without loss of generality, we can assume that $f(x^\star) = 0$. Such an assumption is possible because substituting $f(\cdot)$ by $f(\cdot) + \alpha$ for any scalar α leaves the distribution $\Pi^\star$ unchanged. Since f is m-strongly convex and $\mathcal{L}$-smooth, applying Lemma A.1(c) and Lemma A.2(c), we obtain that

$$\frac{\mathcal{L}}{2} \left\| x - x^\star \right\|_2^2 \geq f(x) \geq \frac{m}{2} \left\| x - x^\star \right\|_2^2, \quad \forall x \in \mathbb{R}^d.$$

Consequently, we find that $\int_{\mathbb{R}^d} e^{-f(x)} dx \leq (2\pi/m)^{d/2}$. Making note of the lower bound

$$\pi^\star(x) \geq \frac{e^{-\frac{\mathcal{L}}{2}\left\| x - x^\star \right\|_2^2}}{(2\pi m^{-1})^{d/2}}, \tag{A.1}$$

and plugging in the expression for the density of $\mu_\star$ yields the claim.

A.2 Proof of Lemma 3.3

The proof consists of two main steps. First, we establish that the distribution $\Pi^\star$ is sub-Gaussian, which then guarantees concentration around the mean. Second, we show that the mean and the mode of the distribution $\Pi^\star$ are not far apart. Combining these two claims yields a high probability region around the mode $x^\star$.

Let x denote the random variable with distribution $\Pi^\star$ and mean $\bar{x} = \mathbb{E}_{x \sim \Pi^\star}[x]$. We claim that $x - \bar{x}$ is a sub-Gaussian random vector with parameter $1/\sqrt{m}$, meaning that

$$\mathbb{E}_x \left[e^{u^\top (x - \bar{x})} \right] \leq e^{\|u\|_2^2/(2m)} \quad \text{for any vector } u \in \mathbb{R}^d.$$

In order to prove this claim, we make use of a result due to Hargé (Theorem 1.1 [106]), which we restate here. Let $y \sim \mathcal{N}(\mu, \Sigma)$ with density e and x be a random variable with density function $q \cdot e$ where q is a log-concave function. Then for any convex function $g : \mathbb{R}^d \mapsto \mathbb{R}$ we have

$$\mathbb{E}_x \left[g(x - \mathbb{E}[x])\right] \leq \mathbb{E}_y \left[g(y - \mathbb{E}[y])\right]. \tag{A.2}$$

From Lemma A.1(b) we have that $x \mapsto f(x) - \frac{m}{2} \left\|x - x^\star\right\|_2^2$ is a convex function. Thus we can express the density $\pi^\star$ as the product of a log-concave function and the density of a random variable with distribution $\mathcal{N}(x^\star, \mathbb{I}_d/m)$. Letting $y \sim \mathcal{N}(x^\star, \mathbb{I}_d/m)$ and noting that $g(z) := e^{u^\top z}$ is a convex function for each fixed vector u, applying the Hargé bound (A.2) yields

$$\mathbb{E}_x \left[e^{u^\top (x - \bar{x})}\right] \leq \mathbb{E}_y \left[e^{u^\top (y - x^*)}\right] \overset{(i)}{\leq} e^{\|u\|_2^2/2m}.$$

Here inequality (i) follows from the fact that the random vector $y - x^\star$ is sub-Gaussian with parameter $1/\sqrt{m}$.

Using the standard tail bounds for quadratic forms for sub-Gaussian random vectors (e.g., Theorem 1 [119]), we find that

$$\mathbb{P}_{x \sim \Pi^\star} \left[\left\|x - \bar{x}\right\|_2^2 > \frac{d}{m} \left(1 + 2\sqrt{\frac{t}{d}} + 2\frac{t}{d}\right)\right] \leq e^{-t}. \tag{A.3}$$

Define $\mathcal{B}_1 := \mathbb{B}\left(\bar{x}, \sqrt{\frac{d}{m}} \cdot \widetilde{\mathfrak{a}}(s)\right)$ where $\widetilde{\mathfrak{a}}(s) = 1 + 2\max\left\{\left(\frac{\log(1/s)}{d}\right)^{0.25}, \sqrt{\frac{\log(1/s)}{d}}\right\}$. Observe that $\widetilde{\mathfrak{a}}(s)^2 \geq 1 + 2\sqrt{\frac{\log(1/s)}{d}} + 2\frac{\log(1/s)}{d}$ and consequently the bound (A.3) implies that $\Pi^\star\left(\mathcal{B}_1\right) = \mathbb{P}_{x \sim \Pi^\star}\left[x \in \mathcal{B}_1\right] \geq 1 - s$. Now applying triangle inequality, we obtain that

$$\mathcal{B}_1 \subseteq \mathbb{B}\left(x^\star, \left\|\bar{x} - x^\star\right\|_2 + \sqrt{\frac{d}{m}} \cdot \widetilde{\mathfrak{a}}(s)\right) =: \mathcal{B}_2$$

From Theorem 1 by Durmus et al. [74], we have that $\mathbb{E}_{x \sim \Pi^\star} \left\|x - x^\star\right\|_2^2 \leq d/m$. Using Jensen inequality twice, we find that

$$\left\|\bar{x} - x^\star\right\|_2 = \left\|\mathbb{E}_{x \sim \Pi^\star}[x] - x^\star\right\|_2 \leq \mathbb{E}_{x \sim \Pi^\star} \left\|x - x^\star\right\|_2 \leq \sqrt{\mathbb{E}_{x \sim \Pi^\star} \left\|x - x^\star\right\|_2^2} \leq \sqrt{\frac{d}{m}}. \tag{A.4}$$

Noting the relation $\mathfrak{a}(s) = 1 + \widetilde{\mathfrak{a}}(s)$, we thus obtain that $\left\|\bar{x} - x^\star\right\|_2 + \sqrt{\frac{d}{m}} \cdot \widetilde{\mathfrak{a}}(s) \leq \mathfrak{a}(s)\sqrt{\frac{d}{m}}$ and consequently $\mathcal{B}_1 \subseteq \mathcal{B}_2 \subseteq \mathcal{R}_s$. As a result, we obtain $\Pi^\star\left(\mathcal{R}_s\right) \geq \Pi^\star\left(\mathcal{B}_1\right) \geq 1 - s$ as claimed.

A.3 Proof of Lemma 3.4

The proof of this lemma makes use of ideas used to establish conductance bounds, first for Hit-and-run [163], and since then for several other walks [171, 187, 47]. See the survey [242] for further details.

For our setting a key ingredient is the following isoperimetric inequality for log-concave distributions. Let $\mathbb{R}^d = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3$ be a partition. Let $y \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$ with density e and let $\Pi^\star$ be a distribution with a density given by $q \cdot e$ where q is a log-concave function. Then Cousins and Vempala (Theorem 4.4 [57]) proved that

$$\Pi^\star(\mathcal{S}_3) \geq \frac{\log 2 \cdot d(\mathcal{S}_1, \mathcal{S}_2)}{\sigma} \Pi^\star(\mathcal{S}_1)\Pi^\star(\mathcal{S}_2) \tag{A.5}$$

where $d(\mathcal{S}_1, \mathcal{S}_2) := \inf \left\{ \|x - y\|_2 \,\middle|\, x \in \mathcal{S}_1, y \in \mathcal{S}_2 \right\}$.

We invoke this result for the truncated distribution $\Pi^\star_{\mathcal{S}}$ with the density $\Pi^\star_{\mathcal{S}}$ defined as

$$\Pi^\star_{\mathcal{S}}(x) := \frac{1}{\displaystyle\int_{\mathcal{S}} \pi^\star(y) dy} \pi^\star(x) \mathbf{1}_{\mathcal{S}}(x) = \frac{1}{\displaystyle\int_{\mathcal{S}} e^{-f(y)} dy} e^{-f(x)} \mathbf{1}_{\mathcal{S}}(x), \tag{A.6}$$

where $\mathbf{1}_{\mathcal{S}}(\cdot)$ denotes the indicator function for the set $\mathcal{S}$, i.e., we have $\mathbf{1}_{\mathcal{S}}(x) = 1$ if $x \in \mathcal{S}$, and 0 otherwise. Let $x^\star = \arg\max \pi^\star(x) = \arg\min f(x)$. Observe that m-strong-convexity of f implies that $x \mapsto f(x) - \frac{m}{2} \|x - x^\star\|_2^2$ is a convex function (Lemma A.1(b)). Noting that the function $\mathbf{1}_{\mathcal{S}}(\cdot)$ is log-concave and that log-concavity is closed under multiplication, we conclude that $\Pi^\star_{\mathcal{S}}$ can be expressed as a product of log-concave function and density of the Gaussian distribution $\mathcal{N}\left(x^\star, \frac{1}{m}\mathbb{I}_d\right)$. Consequently, we can apply the result (D.46) with $\Pi^\star$ replaced by $\Pi^\star_{\mathcal{S}}$ and $\sigma = 1/\sqrt{m}$.

We now prove the claim of the lemma. Define the sets

$$\mathcal{S}_1' := \left\{ u \in \mathcal{S}_1 \cap \mathcal{S} \;\middle|\; \mathcal{T}_u(\mathcal{S}_2) < \frac{\rho}{2} \right\}, \quad \mathcal{S}_2' := \left\{ v \in \mathcal{S}_2 \cap \mathcal{S} \;\middle|\; \mathcal{T}_v(\mathcal{S}_1) < \frac{\rho}{2} \right\}, \tag{A.7}$$

along with the complement $\mathcal{S}_3' := \mathcal{S} \backslash (\mathcal{S}_1' \cup \mathcal{S}_2')$. See Figure A.1 for an illustration. Based on these three sets, we split our proof of the claim (3.20) into two distinct cases:

- Case 1: $\Pi^\star(\mathcal{S}_1') \leq \Pi^\star(\mathcal{S}_1 \cap \mathcal{S})/2$ or $\Pi^\star(\mathcal{S}_2') \leq \Pi^\star(\mathcal{S}_2 \cap \mathcal{S})/2$.
- Case 2: $\Pi^\star(\mathcal{S}_i') \geq \Pi^\star(\mathcal{S}_i \cap \mathcal{S})/2$ for $i = 1, 2$.

Note that these cases are mutually exclusive, and cover all possibilities.

Case 1 We have $\Pi^\star(\mathcal{S}_1 \cap \mathcal{S} \backslash \mathcal{S}_1') \geq \Pi^\star(\mathcal{S}_1 \cap \mathcal{S})/2$, then

$$\begin{aligned} \int_{\mathcal{S}_1} \mathcal{T}_u(\mathcal{S}_2) \pi^\star(u) du \overset{(i)}{\geq} \int_{\mathcal{S}_1 \cap \mathcal{S} \backslash \mathcal{S}_1'} \mathcal{T}_u(\mathcal{S}_2) \pi^\star(u) du &\overset{(ii)}{\geq} \frac{\rho}{2} \Pi^\star(\mathcal{S}_1 \cap \mathcal{S} \backslash \mathcal{S}_1') \\ &\overset{(iii)}{\geq} \frac{\rho}{4} \Pi^\star(\mathcal{S}_1 \cap \mathcal{S}), \end{aligned}$$

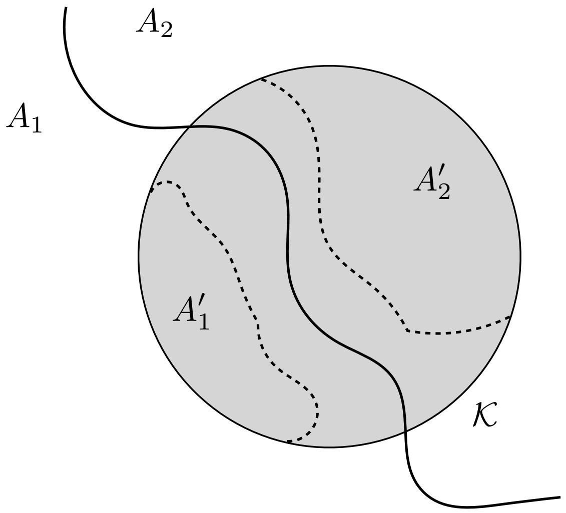


Figure A.1: The sets $\mathcal{S}_1$ and $\mathcal{S}_2$ form a partition of $\mathbb{R}^d$, and we use $\mathbb{S}$ to denote a compact convex subset. The sets $\mathcal{S}_1'$ and $\mathcal{S}_2'$ are defined in equation (D.49).

which implies the claim (3.20). In the above sequence of inequalities, step (i) is trivially true; step (ii) from the definition (D.49) of the set $\mathcal{S}_1'$, and step (iii) from the assumption for this case.

A similar argument with the roles of $\mathcal{S}_1$ and $\mathcal{S}_2$ switched, establishes the claim when $\Pi^\star(\mathcal{S}_2') \leq \Pi^\star(\mathcal{S}_2 \cap \mathbb{S})/2$.

Case 2 We have $\Pi^\star(\mathcal{S}_i') \geq \Pi^\star(\mathcal{S}_i \cap \mathbb{S})/2$ for both $i = 1$ and 2. For any $u \in \mathcal{S}_1'$ and $v \in \mathcal{S}_2'$, we have that

$$d_{\mathrm{TV}}\big(\mathcal{T}_u, \mathcal{T}_v\big) \geq \mathcal{T}_u(\mathcal{S}_1) - \mathcal{T}_v(\mathcal{S}_1) \overset{(i)}{=} 1 - \mathcal{T}_u(\mathcal{S}_2) - \mathcal{T}_v(\mathcal{S}_1) > 1 - \rho,$$

where step (i) follows from the fact that $\mathcal{S}_1 = \mathbb{R}^d \backslash \mathcal{S}_2$ and thereby $\mathcal{T}_u(\mathcal{S}_1) = 1 - \mathcal{T}_u(\mathcal{S}_2)$. Since $u, v \in \mathbb{S}$, the assumption of the lemma implies that $\|u - v\|_2 \geq \Delta$ and consequently

$$d(\mathcal{S}_1', \mathcal{S}_2') \geq \Delta. \tag{A.8}$$

We claim that

$$\int_{\mathcal{S}_1} \mathcal{T}_u(\mathcal{S}_2)\pi^\star(u)du = \int_{\mathcal{S}_2} \mathcal{T}_v(\mathcal{S}_1)\pi^\star(v)dv \tag{A.9}$$

We provide the proof of this claim at the end. Assuming this claim as given, we now complete the proof. Using equation (D.52), we have

$$\begin{aligned}\int_{\mathcal{S}_1} \mathcal{T}_u(\mathcal{S}_2)\pi^\star(u)du &= \frac{1}{2}\left(\int_{\mathcal{S}_1} \mathcal{T}_u(\mathcal{S}_2)\pi^\star(u)du + \int_{\mathcal{S}_2} \mathcal{T}_v(\mathcal{S}_1)\pi^\star(v)dv\right)\\ &\geq \frac{1}{4}\left(\int_{\mathcal{S}_1\cap\mathbb{S}\backslash\mathcal{S}_1'} \mathcal{T}_u(\mathcal{S}_2)\pi^\star(u)du + \int_{\mathcal{S}_2\cap\mathbb{S}\backslash\mathcal{S}_2'} \mathcal{T}_v(\mathcal{S}_1)\pi^\star(v)dv\right)\\ &\overset{(i)}{\geq} \frac{\rho}{8}\Pi^\star(\mathbb{S}\backslash(\mathcal{S}_1'\cup\mathcal{S}_2')),\end{aligned} \tag{A.10}$$

where step (i) follows from the definition (D.49) of the set $\mathcal{S}_3' = \mathbb{S}\backslash(\mathcal{S}_1'\cup\mathcal{S}_2')$. Further, we have

$$\begin{aligned}\Pi^\star(\mathbb{S}\backslash(\mathcal{S}_1'\cup\mathcal{S}_2')) &\overset{(i)}{=} \Pi^\star(\mathbb{S})\cdot\Pi^\star_{\mathbb{S}}(\mathbb{S}\backslash\mathcal{S}_1'\backslash\mathcal{S}_2')\\ &\overset{(ii)}{\geq} \Pi^\star(\mathbb{S})\cdot\frac{\log 2\cdot d(\mathcal{S}_1',\mathcal{S}_2')}{1/\sqrt{m}}\cdot\Pi^\star_{\mathbb{S}}(\mathcal{S}_1')\cdot\Pi^\star_{\mathbb{S}}(\mathcal{S}_2')\\ &\overset{(iii)}{\geq} \Pi^\star(\mathbb{S})\cdot\log 2\cdot d(\mathcal{S}_1',\mathcal{S}_2')\cdot\sqrt{m}\cdot\Pi^\star(\mathcal{S}_1')\cdot\Pi^\star(\mathcal{S}_2')\\ &\overset{(iv)}{\geq} \Pi^\star(\mathbb{S})\cdot\log 2\cdot\Delta\cdot\sqrt{m}\cdot\frac{1}{4}\cdot\Pi^\star(\mathcal{S}_1\cap\mathbb{S})\cdot\Pi^\star(\mathcal{S}_2\cap\mathbb{S}).\end{aligned} \tag{A.11}$$

where step (i) follows from the definition (A.6) of the truncated distribution $\Pi^\star_{\mathbb{S}}$, step (ii) follows from applying the isoperimetry (D.46) for the distribution $\Pi^\star_{\mathbb{S}}$ with $\sigma = 1/\sqrt{m}$, step (iii) from the definition of $\Pi^\star_{\mathbb{S}}$ and step (iv) from inequality (A.8) and the assumption for this case. Let $\alpha \coloneqq \Pi^\star(\mathcal{S}_1\cap\mathbb{S})/\Pi^\star(\mathbb{S})$. Note that $\alpha\in[0,1]$ and $\Pi^\star(\mathcal{S}_2\cap\mathbb{S})/\Pi^\star(\mathbb{S}) = 1-\alpha$. We have

$$\begin{aligned}\Pi^\star(\mathcal{S}_1\cap\mathbb{S})\cdot\Pi^\star(\mathcal{S}_2\cap\mathbb{S}) &= (\Pi^\star(\mathbb{S}))^2\cdot\alpha(1-\alpha)\\ &\geq (\Pi^\star(\mathbb{S}))^2\cdot\frac{1}{2}\min\{\alpha,1-\alpha\}\\ &= \Pi^\star(\mathbb{S})\cdot\frac{1}{2}\min\{\Pi^\star(\mathcal{S}_1\cap\mathbb{S}),\Pi^\star(\mathcal{S}_2\cap\mathbb{S})\}\end{aligned} \tag{A.12}$$

Putting the inequalities (A.10), (A.11) and (A.12) together, establishes the claim (3.20) of the lemma for this case.

We now prove our earlier claim (D.52). Note that it suffices to prove that

$$\int_{\mathcal{S}_1} \mathcal{T}_u(\mathcal{S}_2)\pi^\star(u)du = \int_{\mathcal{S}_2} \mathcal{T}_v(\mathcal{S}_1)\pi^\star(v)dv.$$

We have

$$\begin{aligned}
\int_{\mathcal{S}_2} \mathcal{T}_u(\mathcal{S}_1)\pi^\star(u)du &\overset{(i)}{=} \int_{\mathbb{R}^d} \mathcal{T}_u(\mathcal{S}_1)\pi^\star(u)du - \int_{\mathcal{S}_1} \mathcal{T}_u(\mathcal{S}_1)\pi^\star(u)du \\
&\overset{(ii)}{=} \Pi^\star(\mathcal{S}_1) - \int_{\mathcal{S}_1} \mathcal{T}_u(\mathcal{S}_1)\pi^\star(u)du \\
&= \int_{\mathcal{S}_1} \pi^\star(u)du - \int_{\mathcal{S}_1} \mathcal{T}_u(\mathcal{S}_1)\pi^\star(u)du \\
&\overset{(iii)}{=} \int_{\mathcal{S}_1} \mathcal{T}_u(\mathcal{S}_2)\pi^\star(u)du,
\end{aligned}$$

where steps (i) and (iii) (respectively) follow from the fact that $\mathcal{S}_1 = \mathbb{R}^d \backslash \mathcal{S}_2$ and the consequent fact that $1 - \mathcal{T}_u(\mathcal{S}_1) = \mathcal{T}_u(\mathcal{S}_2)$, and step (ii) follows from the fact that $\pi^\star$ is the stationary density for the transition distribution $\mathcal{T}_x$ and thereby $\int_{\mathbb{R}^d} \mathcal{T}_u(\mathcal{S}_1)\pi^\star(u)du = \Pi^\star(\mathcal{S}_1)$.

A.4 Proof of Lemma 3.5

We prove each claim of the lemma separately. To simplify notation, we drop the superscript from our notations of distributions $\mathcal{T}_x^{\mathrm{MALA}(\eta)}$ and $\mathcal{P}_x^{\mathrm{MALA}(\eta)}$.

A.4.0.1 Proof of claim (3.23a)

In order to bound the total variation distance $d_{\mathrm{TV}}\big(\mathcal{P}_x, \mathcal{P}_y\big)$, we apply Pinsker's inequality [58], which guarantees that $d_{\mathrm{TV}}\big(\mathcal{P}_x, \mathcal{P}_y\big) \leq \sqrt{2\,\mathrm{KL}(\mathcal{P}_x \| \mathcal{P}_y)}$. Given multivariate normal distributions $\mathcal{G}_1 = \mathcal{N}(\mu_1, \Sigma)$ and $\mathcal{G}_2 = \mathcal{N}(\mu_2, \Sigma)$, the Kullback-Leibler divergence between the two is given by

$$\mathrm{KL}(\mathcal{G}_1 \| \mathcal{G}_2) = \frac{1}{2}\left(\mu_1 - \mu_2\right)^\top \Sigma^{-1}\left(\mu_1 - \mu_2\right). \tag{A.13}$$

Substituting $\mathcal{G}_1 = \mathcal{P}_x$ and $\mathcal{G}_2 = \mathcal{P}_y$ into the above expression and applying Pinsker's inequality, we find that

$$\begin{aligned}
d_{\mathrm{TV}}\big(\mathcal{P}_x, \mathcal{P}_y\big) \leq \sqrt{2\,\mathrm{KL}(\mathcal{P}_x \| \mathcal{P}_y)} &= \frac{\|\mu_x - \mu_y\|_2}{\sqrt{2\eta}} \\
&\overset{(i)}{=} \frac{\|(x - \eta\nabla f(x)) - (y - \eta\nabla f(y))\|_2}{\sqrt{2\eta}},
\end{aligned}$$

where step (i) follows from the definition (3.22) of the mean μ_x. Consequently, in order to establish the claim (3.23a), it suffices to show that

$$\|(x - \eta\nabla f(x)) - (y - \eta\nabla f(y))\|_2 \leq \|x - y\|_2\,.$$

Recalling that $\lVert B \rVert_{\text{op}}$ denotes the ℓ_2-operator norm of a matrix B (equal to the maximum singular value), we have

$$\begin{aligned}\lVert (x - \eta \nabla f(x)) - (y - \eta \nabla f(y)) \rVert_2 &= \left\lVert \int_0^1 \left[\mathbb{I} - \eta \nabla^2 f(x + t(x-y)) \right] (x-y) dt \right\rVert_2 \\ &\leq \int_0^1 \left\lVert \left[\mathbb{I} - \eta \nabla^2 f(x + t(x-y)) \right] (x-y) \right\rVert_2 dt \\ &\overset{(i)}{\leq} \sup_{z \in \mathbb{R}^d} \lVert \mathbb{I}_d - \eta \nabla^2 f(z) \rVert_{\text{op}} \ \lVert x - y \rVert_2 ,\end{aligned}$$

where step (i) follows from the definition of the operator norm. Lemma A.1(f) and Lemma A.2(f) guarantee that the Hessian is sandwiched as $m \mathbb{I}_d \preceq \nabla^2 f(z) \preceq \mathcal{L} \mathbb{I}_d$ for all $z \in \mathbb{R}^d$, where $\mathbb{I}_d$ denotes the d-dimensional identity matrix. From this Hessian sandwich, it follows that

$$\lVert \mathbb{I}_d - \eta \nabla^2 f(x) \rVert_{\text{op}} = \max \left\{ \lvert 1 - \eta \mathcal{L} \rvert , \lvert 1 - \eta m \rvert \right\} \ < \ 1.$$

Putting together the pieces yields the claim.

A.4.0.2 Proof of claim (3.23b)

Let $\mathcal{P}_1$ be a distribution admitting a density p_1 on $\mathbb{R}^d$, and let $\mathcal{P}_2$ be a distribution which has an atom at x and admitting a density p_2 on $\mathbb{R}^d \backslash \{x\}$. The total variation distance between the distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ is given by

$$d_{\text{TV}}(\mathcal{P}_1, \mathcal{P}_2) = \frac{1}{2} \left(\mathcal{P}_2(\{x\}) + \int_{\mathbb{R}^d} \lvert p_1(z) - p_2(z) \rvert \, dz \right). \tag{A.14}$$

The accept-reject step for MALA implies that

$$\mathcal{T}_x(\{x\}) = 1 - \int_{\mathbb{R}^d} \min \left\{ 1, \frac{\pi^\star(z) \cdot p_z(x)}{\pi^\star(x) \cdot p_x(z)} \right\} p_x(z) dz, \tag{A.15}$$

where p_x denotes the density corresponding to the proposal distribution $\mathcal{P}_x = \mathcal{N}(x - \eta \nabla f(x), 2\eta \mathbb{I}_d)$. From this fact and the formula (A.14), we find that

$$\begin{aligned} d_{\text{TV}}(\mathcal{P}_x, \mathcal{T}_x) &= \frac{1}{2} \left(\mathcal{T}_x(\{x\}) + \int_{\mathbb{R}^d} p_x(z) dz - \int_{\mathbb{R}^d} \min \left\{ 1, \frac{\pi^\star(z) \cdot p_z(x)}{\pi^\star(x) \cdot p_x(z)} \right\} p_x(z) dz \right) \\ &= \frac{1}{2} \left(2 - 2 \int_{\mathbb{R}^d} \min \left\{ 1, \frac{\pi^\star(z) \cdot p_z(x)}{\pi^\star(x) \cdot p_x(z)} \right\} p_x(z) dz \right) \\ &= 1 - \mathbb{E}_{z \sim \mathcal{P}_x} \left[\min \left\{ 1, \frac{\pi^\star(z) \cdot p_z(x)}{\pi^\star(x) \cdot p_x(z)} \right\} \right]. \end{aligned} \tag{A.16}$$

By applying Markov's inequality, we obtain

$$\mathbb{E}_{z \sim \mathcal{P}_x} \left[\min \left\{ 1, \frac{\pi^\star(z) \cdot p_z(x)}{\pi^\star(x) \cdot p_x(z)} \right\} \right] \geq \alpha \, \mathbb{P} \left[\frac{\pi^\star(z) \cdot p_z(x)}{\pi^\star(x) \cdot p_x(z)} \geq \alpha \right] \quad \text{for all } \alpha \in (0,1]. \tag{A.17}$$

We now derive a high probability lower bound for the ratio $[\pi^\star(z)p_z(x)]\,/\,[\pi^\star(x)p_x(z)]$. Noting that $\pi^\star(x) \propto \exp(-f(x))$ and $p_x(z) \propto \exp\big(-\left\|x - \eta\nabla f(x) - z\right\|_2^2/(4\eta)\big)$, we have

$$\frac{\pi^\star(z)\cdot p_z(x)}{\pi^\star(x)\cdot p_x(z)} = \frac{\exp\left(-f(z) - \frac{\|x-z+\eta\nabla f(z)\|_2^2}{4\eta}\right)}{\exp\left(-f(x) - \frac{\|z-x+\eta\nabla f(x)\|_2^2}{4\eta}\right)}$$
$$= \exp\left(\frac{4\eta(f(x)-f(z)) + \left\|z-x+\eta\nabla f(x)\right\|_2^2 - \left\|x-z+\eta\nabla f(z)\right\|_2^2}{4\eta}\right). \tag{A.18}$$

Keeping track of the numerator of this exponent, we find that

$$\begin{aligned}
&4\eta(f(x)-f(z)) + \left\|z-x+\eta\nabla f(x)\right\|_2^2 - \left\|x-z+\eta\nabla f(z)\right\|_2^2 \\
&\quad = 4\eta(f(x)-f(z)) + \left\|z-x\right\|_2^2 + \left\|\eta\nabla f(x)\right\|_2^2 + 2\eta(z-x)^\top\nabla f(x) \\
&\qquad - \left\|x-z\right\|_2^2 - \left\|\eta\nabla f(z)\right\|_2^2 - 2\eta(x-z)^\top\nabla f(z) \\
&\quad = 2\eta\underbrace{\left(f(x)-f(z)-(x-z)^\top\nabla f(x)\right)}_{M_1} + 2\eta\underbrace{\left(f(x)-f(z)-(x-z)^\top\nabla f(z)\right)}_{M_2} \\
&\qquad + \eta^2\underbrace{\left(\left\|\nabla f(x)\right\|_2^2 - \left\|\nabla f(z)\right\|_2^2\right)}_{M_3}.
\end{aligned} \tag{A.19}$$

Now we provide lower bounds for the terms M_i, $i = 1,2,3$ defined in the above display. Since f is strongly convex and smooth, applying Lemma A.1(c) and Lemma A.2(c) yields

$$M_1 \geq -\frac{\mathcal{L}}{2}\left\|x-z\right\|_2^2, \quad \text{and} \quad M_2 \geq \frac{m}{2}\left\|x-z\right\|_2^2. \tag{A.20}$$

In order to lower bound M_3, we observe that

$$\begin{aligned}
M_3 = \left\|\nabla f(x)\right\|_2^2 - \left\|\nabla f(z)\right\|_2^2 &= \langle \nabla f(x) + \nabla f(z),\, \nabla f(x) - \nabla f(z)\rangle \\
&\overset{(i)}{\geq} -\left\|\nabla f(x) + \nabla f(z)\right\|_2 \left\|\nabla f(x) - \nabla f(z)\right\|_2 \\
&\overset{(ii)}{\geq} -\left(2\left\|\nabla f(x)\right\|_2 + \mathcal{L}\left\|x-z\right\|_2\right)\mathcal{L}\left\|x-z\right\|_2,
\end{aligned} \tag{A.21}$$

where step (i) follows from the Cauchy-Schwarz's inequality and step (ii) from the triangle inequality and $\mathcal{L}$-smoothness of the function f (cf. Lemma A.2(d)).

Combining the bounds (A.20) and (A.21) with equations (A.19) and (A.18), we have established that

$$\frac{\pi^\star(z)\cdot p_z(x)}{\pi^\star(x)\cdot p_x(z)} \geq \exp\left(\underbrace{-\frac{1}{4}(\mathcal{L}-m)\left\|x-z\right\|_2^2 - \frac{\eta}{4}\left(2\mathcal{L}\left\|x-z\right\|_2\left\|\nabla f(x)\right\|_2 + \mathcal{L}^2\left\|x-z\right\|_2^2\right)}_{=:T}\right). \tag{A.22}$$

Now to provide a high probability lower bound for the term T, we make use of the standard chi-squared tail bounds and the following relation between x and z:

$$z \overset{(d)}{=} x - \eta \nabla f(x) + \sqrt{2\eta}\xi,$$

where $\xi \sim \mathcal{N}(0, \mathbb{I}_d)$ and $\overset{(d)}{=}$ denotes equality in distribution. We have

$$\|x - z\|_2 = \left\|\eta \nabla f(x) + \sqrt{2\eta}\xi\right\|_2 \leq \eta \left\|\nabla f(x)\right\|_2 + \sqrt{2\eta} \left\|\xi\right\|_2,$$

which also implies

$$\|x - z\|_2^2 \leq 2\eta^2 \left\|\nabla f(x)\right\|_2^2 + 4\eta \left\|\xi\right\|_2^2.$$

Using these two inequalities, we find that

$$T \geq -\frac{(\mathcal{L} - m)\eta^2}{2} \left\|\nabla f(x)\right\|_2^2 - (\mathcal{L} - m)\eta \left\|\xi\right\|_2^2 - \frac{\mathcal{L}\eta^2}{2} \left\|\nabla f(x)\right\|_2^2 \\ - \frac{\mathcal{L}\eta\sqrt{\eta}}{\sqrt{2}} \left\|\nabla f(x)\right\|_2 \left\|\xi\right\|_2 - \frac{\mathcal{L}^2\eta^3}{2} \left\|\nabla f(x)\right\|_2^2 - \mathcal{L}^2\eta^2 \left\|\xi\right\|_2^2.$$

Simplifying and using the fact that $\mathcal{L}\eta \leq 1$, we obtain that

$$T \geq -2\left(\mathcal{L}\eta^2 \left\|\nabla f(x)\right\|_2^2 + \mathcal{L}\eta \left\|\xi\right\|_2^2 + \mathcal{L}\eta\sqrt{\eta} \left\|\nabla f(x)\right\|_2 \left\|\xi\right\|_2\right).$$

Since $x \in \mathcal{R}_s$, we have

$$\left\|\nabla f(x)\right\|_2 = \left\|\nabla f(x) - \nabla f(x^\star)\right\|_2 \overset{(i)}{\leq} \mathcal{L} \left\|x - x^\star\right\|_2 \leq \mathcal{L}\sqrt{\frac{d}{m}}\mathfrak{a}(s) =: \mathcal{D}_s, \tag{A.23}$$

where inequality (i) follows from the property (d) of Lemma A.2. Thus, we have shown that

$$T \geq -2\left(\mathcal{L}\eta^2\mathcal{D}_s^2 + \mathcal{L}\eta \left\|\xi\right\|_2^2 + \mathcal{L}\eta\sqrt{\eta}\mathcal{D}_s \left\|\xi\right\|_2\right). \tag{A.24}$$

Standard tail bounds for χ^2-variables guarantee that

$$\mathbb{P}\left[\left\|\xi\right\|_2^2 \leq d\alpha_\varepsilon\right] \geq (1 - \varepsilon/16) \text{ for } \alpha_\varepsilon = 1 + 2\sqrt{\log(16/\varepsilon)} + 2\log(16/\varepsilon).$$

A simple observation reveals that the function $\widetilde{\mathfrak{t}}$ defined in equation (3.21a) was chosen such that for any $\eta \leq \widetilde{\mathfrak{t}}(s, \varepsilon)$, we have

$$\mathcal{L}\eta^2\mathcal{D}_s^2 \leq \frac{\varepsilon}{128}, \quad \mathcal{L}\eta d\alpha_\varepsilon \leq \frac{\varepsilon}{64}, \quad \text{and,} \quad \mathcal{L}\eta\sqrt{\eta}\mathcal{D}_s\sqrt{d\alpha_\varepsilon} \leq \frac{\varepsilon}{128}.$$

Combining this observation with the high probability bound on $\left\|\xi\right\|_2$ and using the inequality (A.24) we obtain that $T \geq -\varepsilon/16$ with probability at least $1 - \varepsilon/16$. Plugging this bound in the inequality (A.22), we find that

$$\mathbb{P}\left[\frac{\pi^\star(z) \cdot p_z(x)}{\pi^\star(x) \cdot p_x(z)} \geq \exp\left(-\frac{\varepsilon}{16}\right)\right] \geq (1 - \varepsilon/16).$$

Thus, we have derived a desirable high probability lower bound on the accept-reject ratio. Substituting $\alpha = \exp(-\varepsilon/16)$ in the inequality (E.16) and using the fact that $e^{-\varepsilon/16} \geq 1 - \varepsilon/16$ for any scalar $\varepsilon > 0$, we find that

$$\mathbb{E}_{z \sim \mathcal{P}_x}\left[\min\left\{1, \frac{\pi^\star(z)\cdot p_z(x)}{\pi^\star(x)\cdot p_x(z)}\right\}\right] \geq 1 - \frac{\varepsilon}{8}, \qquad \text{for any } \varepsilon \in (0,1) \text{ and } \eta \leq \widetilde{\mathfrak{t}}(s,\varepsilon).$$

Substituting this bound in the inequality (A.16) completes the proof.

A.5 Basic properties of convex and smooth functions

In this appendix, we state a few basic properties of strongly-convex and smooth functions that we use in our proofs. See the book [26] for more details.

Lemma A.1 (Equivalent characterizations of strong convexity)**.** *For a twice differentiable convex function $f : \mathbb{R}^d \mapsto \mathbb{R}$, the following statements are equivalent:*

(a) The function f is m-strongly-convex.

(b) The function $x \mapsto f(x) - \frac{m}{2}\left\|x - x^\star\right\|_2^2$ is convex (for any fixed point $x^\star$).

(c) For any $x, y \in \mathbb{R}^d$, we have

$$f(y) \geq f(x) + \nabla f(x)^\top (y-x) + \frac{m}{2}\left\|x-y\right\|_2^2.$$

(d) For any $x, y \in \mathbb{R}^d$, we have

$$\left\|\nabla f(x) - \nabla f(y)\right\|_2 \geq m\left\|x-y\right\|_2.$$

(e) For any $x, y \in \mathbb{R}^d$, we have

$$\left(\nabla f(x) - \nabla f(y)\right)^\top (x-y) \geq m\left\|x-y\right\|_2^2.$$

(f) For any $x \in \mathbb{R}^d$, the Hessian is lower bounded as $\nabla^2 f(x) \succeq m\mathbb{I}_d$.

Lemma A.2 (Equivalent characterizations of smoothness)**.** *For a twice differentiable convex function $f : \mathbb{R}^d \mapsto \mathbb{R}$, the following statements are equivalent:*

(a) The function f is $\mathcal{L}$-smooth.

(b) The function $x \mapsto \frac{\mathcal{L}}{2}\left\|x - x^\star\right\|_2^2 - f(x)$ is convex (for any fixed point $x^\star$).

(c) For any $x, y \in \mathbb{R}^d$*, we have*

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{\mathcal{L}}{2} \left\| x - y \right\|_2^2 .$$

(d) For any $x, y \in \mathbb{R}^d$*, we have*

$$\left\| \nabla f(x) - \nabla f(y) \right\|_2 \leq \mathcal{L} \left\| x - y \right\|_2 .$$

(e) For any $x, y \in \mathbb{R}^d$*, we have*

$$\left(\nabla f(x) - \nabla f(y) \right)^\top (x - y) \leq \mathcal{L} \left\| x - y \right\|_2^2 .$$

(f) For any $x \in \mathbb{R}^d$*, the Hessian is upper bounded as* $\nabla^2 f(x) \preceq \mathcal{L} \mathbb{I}_d$*.*

Appendix B

Content Deferred from Chapter 4

Here we collect the proofs of Propositions 4.1 and 4.2 in Appendices B.1 and B.2 respectively.

B.1 Proof of Proposition 4.1

We begin by adapting the spectral profile technique [99] to the continuous state setting, and next we relate conductance profile with the spectral profile.

First, we briefly recall some of the key notation. Let $\Theta : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to \mathbb{R}_+$ denote the transition probability function for the Markov chain and let $\mathcal{T}$ be the corresponding transition operator, which maps a probability measure to another according to the transition probability Θ. Note that for a Markov chain satisfying the smooth chain assumption (4.4), if the distribution μ admits a density then the distribution $\mathcal{T}(\mu)$ would also admits a density. We use $\mathcal{T}_x$ as the shorthand for $\mathcal{T}(\delta_x)$, the transition distribution of the Markov chain at x.

Let $L_2(\pi^\star)$ be the space of square integrable functions under function $\pi^\star$. The *Dirichlet form* $\mathcal{E} : L_2(\pi^\star) \times L_2(\pi^\star) \to \mathbb{R}$ associated with the transition probability Θ is given by

$$\mathcal{E}(g,h) = \frac{1}{2} \int_{(x,y)\in\mathcal{X}^2} \left(g(x) - \mathcal{H}(y)\right)^2 \Theta(x,dy)\pi^\star(x)dx. \tag{B.1}$$

The expectation $\mathbb{E}_{\pi^\star} : L_2(\pi^\star) \to \mathbb{R}$ and the variance $\mathrm{Var}_{\pi^\star} : L_2(\pi^\star) \to \mathbb{R}$ with respect to the density $\pi^\star$ are given by

$$\mathbb{E}_{\pi^\star}(g) = \int_{x\in\mathcal{X}} g(x)\pi^\star(x)dx \quad \text{and} \quad \mathrm{Var}_{\pi^\star}(g) = \int_{x\in\mathcal{X}} \left(g(x) - \mathbb{E}_{\pi^\star}(g)\right)^2 \pi^\star(x)dx. \tag{B.2a}$$

Furthermore, for a pair of measurable sets $(S,\Omega) \subset \mathcal{X}^2$, the Ω-*restricted spectral gap for the set* S is defined as

$$\lambda_\Omega(S) = \inf_{g\in c_0^+(S\cap\Omega)} \frac{\mathcal{E}(g,g)}{\mathrm{Var}_{\pi^\star}(g)}, \tag{B.3a}$$

$$\text{where } c_0^+(S\cap\Omega) = \left\{g \in L_2(\pi^\star) \mid \mathrm{supp}(g) \subset S\cap\Omega, g \geq 0, g \neq \text{constant}\right\}. \tag{B.3b}$$

Finally, the Ω-*restricted spectral profile* Λ_Ω is defined as

$$\Lambda_\Omega(v) = \inf_{\Pi^\star(S\cap\Omega)\in[0,v]} \lambda_\Omega(S\cap\Omega), \qquad \text{for all } v \in \left[0,\infty\right). \tag{B.4}$$

Note that we restrict the spectral profile to the set Ω. Taking Ω to be $\mathcal{X}$, our definition agrees with the standard definition definitions of the restricted spectral gap and spectral profile in the paper [99] for finite state space Markov chains to continuous state space Markov chains. We are now ready to state a mixing time bound using spectral profile.

Lemma B.1. *Consider a reversible irreducible ζ-lazy Markov chain with stationary distribution $\Pi^\star$ satifying the smooth chain assumption (4.4). Given a β-warm start μ_0, an error tolerance $\delta \in (0,1)$ and a set $\Omega \subset \mathcal{X}$ with $\Pi^\star(\Omega) \geq 1 - \frac{\delta^2}{3\beta^2}$, the L_2-mixing time is bounded as*

$$\tau_2(\delta;\mu_0) \leq \left\lceil \int_{4/\beta}^{8/\delta^2} \frac{dv}{\zeta\cdot v\Lambda_\Omega(v)} \right\rceil, \tag{B.5}$$

where Λ_Ω denotes the Ω-restricted spectral profile (B.4) of the chain.

See Appendix B.1.1 for the proof.

In the next lemma, we state the relationship between the Ω-restricted spectral profile (B.4) of the Markov chain to its Ω-restricted conductance profile (4.2).

Lemma B.2. *For a Markov chain with state space $\mathcal{X}$ and stationary distribution $\Pi^\star$, given any measurable set $\Omega \subset \mathcal{X}$, its Ω-restricted spectral profile (B.4) and Ω-restricted conductance profile (4.2) are related as*

$$\Lambda_\Omega(v) \geq \begin{cases} \dfrac{\Phi_\Omega^2(v)}{4} & \textit{for all } v \in \left[0, \frac{\Pi^\star(\Omega)}{2}\right] \\ \dfrac{\Phi_\Omega^2(\Pi^\star(\Omega)/2)}{8} & \textit{for all } v \in \left(\frac{\Pi^\star(\Omega)}{2}, \infty\right). \end{cases} \tag{B.6}$$

See Appendix B.1.2 for the proof.

Proposition 4.1 now follows from Lemmas B.1 and B.2 as well as the definition (4.3) of $\widetilde{\Phi}_\Omega$.

B.1.1 Proof of Lemma B.1

We need the following lemma, proved in for the case of finite state Markov chains in [99], which lower bounds the Dirichlet form in terms of the spectral profile.

Lemma B.3. *For any measurable set $\Omega \subset \mathcal{X}$, any function $g : \mathcal{X} \to \mathbb{R}_+$ such that $g \in L_2(\pi^\star)$, $g \cdot \mathbf{1}_\Omega$ is not constant and $\mathbb{E}[g^2 \cdot \mathbf{1}_\Omega] \geq 2\mathbb{E}[g^2 \cdot \mathbf{1}_{\Omega^c}]$, we have*

$$\frac{\mathcal{E}(g,g)}{\mathrm{Var}_{\pi^\star}(g\cdot\mathbf{1}_\Omega)} \geq \frac{1}{2}\Lambda_\Omega\left(\frac{4\left(\mathbb{E}_{\pi^\star}(g\cdot\mathbf{1}_\Omega)\right)^2}{\mathrm{Var}_{\pi^\star}(g\cdot\mathbf{1}_\Omega)}\right). \tag{B.7}$$

The proof of Lemma B.3 is a straightforward extension of Lemma 2.1 from [99], which deals with finite state spaces, to the continuous state Markov chain. See the end of Section B.1.1.1 for the proof.

We are now equipped to prove Lemma B.1.

Proof of Lemma B.1: We begin by introducing some notations. Recall that for any Markov chain satisfying the smooth chain assumption (4.4), given an initial distribution μ_0 that admits a density, the distribution of the chain at any step $\mathfrak{n}$ also admits a density. As a result, we can define the ratio of the density of the Markov chain at the $\mathfrak{n}$-th iteration $\mathcal{H}_{\mu_0,\mathfrak{n}} : \mathcal{X} \to \mathbb{R}$ with respect to the target density $\pi^\star$ via the following recursion

$$\mathcal{H}_{\mu_0,0}(x) = \frac{\mu_0(x)}{\pi^\star(x)} \quad \text{and} \quad \mathcal{H}_{\mu_0,\mathfrak{n}+1}(x) = \frac{\mathcal{T}\left(\pi^\star \cdot \mathcal{H}_{\mu_0,\mathfrak{n}}\right)(x)}{\pi^\star(x)},$$

where we have used the notation $\mathcal{T}(\mu)(x)$ to denote the density of the distribution $\mathcal{T}(\mu)$ at x. Note that

$$\mathbb{E}_{\pi^\star}(\mathcal{H}_{\mu_0,\mathfrak{n}}) = 1 \quad \text{and} \quad \mathbb{E}_{\pi^\star}(\mathcal{H}_{\mu_0,\mathfrak{n}} \cdot \mathbf{1}_\Omega) \leq 1 \quad \text{for all } \mathfrak{n} \geq 0, \tag{B.8}$$

where $\Omega \subset \mathcal{X}$ is a measurable set.

We also define the quantity $\mathcal{J}(\mathfrak{n}) := \mathrm{Var}_{\pi^\star}(\mathcal{H}_{\mu_0,\mathfrak{n}})$ (we prove the existence of this variance below in Step (1)) and also $\widetilde{\mathcal{J}}(\mathfrak{n}) := \mathrm{Var}_{\pi^\star}(\mathcal{H}_{\mu_0,\mathfrak{n}} \cdot \mathbf{1}_\Omega)$. Note that the L_2-distance between the distribution of the chain at step $\mathfrak{n}$ and the target distribution is given by

$$d_{2,\pi^\star}(\mathcal{T}^\mathfrak{n}(\mu_0), \Pi^\star) = \left(\int_{x\in\mathbb{R}^d} \left(\mathcal{H}_{\mu_0,\mathfrak{n}}(x) - 1\right)^2 \pi^\star(x)dx\right)^{1/2} = \mathrm{Var}_{\pi^\star}(\mathcal{H}_{\mu_0,\mathfrak{n}}).$$

Consequently, to prove the δ-L_2 mixing time bound (B.5), it suffices to show that for any measurable set $\Omega \subset \mathcal{X}$, with $\Pi^\star(\Omega) \geq 1 - \frac{\delta^2}{3\beta^2}$, we have

$$\mathcal{J}(\mathfrak{n}) \leq \delta^2 \quad \text{for } \mathfrak{n} \geq \left\lceil \int_{4/\beta}^{8/\delta^2} \frac{dv}{\zeta \cdot v\Lambda_\Omega(v)} \right\rceil \tag{B.9}$$

We now establish the claim (B.9) via a three-step argument: (1) we prove the existence of the variance $\mathcal{J}(\mathfrak{n})$ for all $\mathfrak{n} \in \mathbb{N}$ and relate $\mathcal{J}(\mathfrak{n})$ with $\widetilde{\mathcal{J}}(\mathfrak{n})$. (2) then we derive a recurrence relation for the difference $\mathcal{J}(\mathfrak{n}+1) - \mathcal{J}(\mathfrak{n})$ in terms of Dirichlet forms that shows the $\mathcal{J}$ is a decreasing function, and (3) finally, using an extension of the variance $\mathcal{J}$ from natural indices to real numbers, we derive an explicit upper bound on the number of steps taken by the chain until $\mathcal{J}$ lies below the required threshold.

Step (1): Using the reversibility (2.3) of the chain, we find that

$$\mathcal{H}_{\mu_0,\mathfrak{n}+1}(x)dx = \frac{\int_{y\in\mathcal{X}} \Theta(y,dx)\mathcal{H}_{\mu_0,\mathfrak{n}}(y)\pi^\star(y)dy}{\pi^\star(x)} = \frac{\int_{y\in\mathcal{X}} \Theta(x,dy)\mathcal{H}_{\mu_0,\mathfrak{n}}(y)\pi^\star(x)dx}{\pi^\star(x)}$$
$$= \int_{y\in\mathcal{X}} \Theta(x,dy)\mathcal{H}_{\mu_0,\mathfrak{n}}(y)dx \tag{B.10}$$

Applying an induction argument along with the relationship (B.10) and the initial condition $\mathcal{H}_{\mu_0,0}(x) \leq \beta$, we obtain that

$$\mathcal{H}_{\mu_0,\mathfrak{n}}(x) \leq \beta, \quad \text{for all } \mathfrak{n} \geq 0. \tag{B.11}$$

As a result, the variances of the functions $\mathcal{H}_{\mu_0,0}$ and $\mathcal{H}_{\mu_0,\mathfrak{n}} \cdot \mathbf{1}_\Omega$ under the target density $\pi^\star$ are well-defined and

$$\mathcal{J}(\mathfrak{n}) = \int_{\mathcal{X}} \mathcal{H}^2_{\mu_0,\mathfrak{n}}(x)\pi^\star(x)dx - 1. \tag{B.12}$$

Then we show that $\mathcal{J}$ can be upper bounded via $\widetilde{\mathcal{J}}$ as follows

$$\begin{aligned}\mathcal{J}(\mathfrak{n}) - \widetilde{\mathcal{J}}(\mathfrak{n}) &= \mathrm{Var}_{\pi^\star}(\mathcal{H}_{\mu_0,\mathfrak{n}}) - \mathrm{Var}_{\pi^\star}(\mathcal{H}_{\mu_0,\mathfrak{n}} \cdot \mathbf{1}_\Omega) \\ &= \int_{x\in\mathcal{X}\setminus\Omega} \mathcal{H}^2_{\mu_0,\mathfrak{n}}(x)\pi^\star(x)dx - \left(\int_{x\in\mathcal{X}} \mathcal{H}_{\mu_0,\mathfrak{n}}(x)\pi^\star(x)dx\right)^2 \\ &\qquad + \left(\int_{x\in\Omega} \mathcal{H}_{\mu_0,\mathfrak{n}}(x)\pi^\star(x)dx\right)^2 \\ &\leq \beta^2\left(1 - \Pi^\star(\Omega)\right) \leq \frac{\delta^2}{3} =: B,\end{aligned} \tag{B.13}$$

where the last inequality follows from the fact that Ω satisfies $\Pi^\star(\Omega) \geq 1 - \delta^2/(3\beta^2)$. Similarly, we have

$$\mathbb{E}_{\pi^\star}(\mathcal{H}^2_{\mu_0,\mathfrak{n}}) - \mathbb{E}_{\pi^\star}(\mathcal{H}^2_{\mu_0,\mathfrak{n}} \cdot \mathbf{1}_\Omega) \leq B \tag{B.14}$$

Step (2): First, note the following bound on $\mathcal{J}(0)$:

$$\mathcal{J}(0) = \int_{x\in\mathcal{X}} \frac{\mu_0(x)^2}{\pi^\star(x)}dx - 1 \leq \beta \int_{x\in\mathcal{X}} \mu_0(x)dx - 1 \leq \beta - 1. \tag{B.15}$$

Define the two step transition kernel $\Theta \circ \Theta$ as

$$\Theta \circ \Theta(y,dz) = \int_{x\in\mathcal{X}} \Theta(y,dx)\Theta(x,dz).$$

We have

$$\begin{aligned}\mathcal{J}(\mathfrak{n}+1) \coloneqq \operatorname{Var}_{\pi^\star}(\mathcal{H}_{\mu_0,\mathfrak{n}+1}) &= \int_{x\in\mathcal{X}} \mathcal{H}^2_{\mu_0,\mathfrak{n}+1}(x)\pi^\star(x)dx - 1\\ &\overset{(i)}{=} \int_{x\in\mathcal{X}}\int_{y\in\mathcal{X}} \Theta(y,dx)\mathcal{H}_{\mu_0,\mathfrak{n}}(y)\pi^\star(y)dy \int_{z\in\mathcal{X}} \Theta(x,dz)\mathcal{H}_{\mu_0,\mathfrak{n}}(z) - 1\\ &= \int_{y,z\in\mathcal{X}^2} \Theta\circ\Theta(y,dz)\mathcal{H}_{\mu_0,\mathfrak{n}}(y)\mathcal{H}_{\mu_0,\mathfrak{n}}(z)\pi^\star(y)dy - 1,\end{aligned}$$

where step (i) follows from the relation (B.10). Using the above expression for $\mathcal{J}(\mathfrak{n}+1)$ and the expression from equation (B.12) for $\mathcal{J}(\mathfrak{n})$, we find that

$$\begin{aligned}\mathcal{J}(\mathfrak{n}+1) - \mathcal{J}(\mathfrak{n}) &= \int_{\mathcal{X}^2} \Theta\circ\Theta(y,dz)\mathcal{H}_{\mu_0,\mathfrak{n}}(y)\mathcal{H}_{\mu_0,\mathfrak{n}}(z)\pi^\star(y)dy - \int_{\mathcal{X}} \mathcal{H}^2_{\mu_0,\mathfrak{n}}(x)\pi^\star(x)dx,\\ &\overset{(a)}{=} -\mathcal{E}_{\Theta\circ\Theta}(\mathcal{H}_{\mu_0,\mathfrak{n}},\mathcal{H}_{\mu_0,\mathfrak{n}}), \qquad \text{(B.16)}\end{aligned}$$

where $\mathcal{E}_{\Theta\circ\Theta}$ is the Dirichlet form (B.1) with transition probability Θ being replaced by $\Theta\circ\Theta$. We come back to the proof of equality (a) at the end of this paragraph. Assuming it as given at the moment, we proceed further. Since the Markov chain is ζ-lazy, we can relate the two Dirichlet forms $\mathcal{E}_{\Theta\circ\Theta}$ and $\mathcal{E}_\Theta$ as follows: For any $y,z\in\mathcal{X}$ such that $y\neq z$, we have

$$\begin{aligned}\Theta\circ\Theta(y,dz) = \int_{x\in\mathcal{X}} \Theta(y,dx)\Theta(x,dz) &\geq \Theta(y,dy)\Theta(y,dz) + \Theta(y,dz)\Theta(z,dz)\\ &\geq 2\zeta\Theta(y,dz). \qquad \text{(B.17)}\end{aligned}$$

If $\mathbb{E}_{\pi^\star}(\mathcal{H}^2_{\mu_0,\mathfrak{n}}\cdot\mathbf{1}_\Omega) < 2\mathbb{E}_{\pi^\star}(\mathcal{H}^2_{\mu_0,\mathfrak{n}}\cdot\mathbf{1}_{\Omega^c})$, then according to Equation (B.14), we have

$$\mathcal{J}(\mathfrak{n}) \leq \mathbb{E}(\mathcal{H}^2_{\mu_0,\mathfrak{n}}) \leq \mathbb{E}(\mathcal{H}^2_{\mu_0,\mathfrak{n}}\cdot\mathbf{1}_\Omega) + B \leq 3B \leq \delta^2, \qquad \text{(B.18)}$$

and we are done. If $\mathcal{H}^2_{\mu_0,\mathfrak{n}}\cdot\mathbf{1}_\Omega$ is constant, then $\operatorname{Var}_{\pi^\star}(\mathcal{H}^2_{\mu_0,\mathfrak{n}}\cdot\mathbf{1}_\Omega) = 0$ and

$$\mathcal{J}(\mathfrak{n}) \leq \operatorname{Var}_{\pi^\star}(\mathcal{H}^2_{\mu_0,\mathfrak{n}}) \leq 0 + B \leq \delta^2.$$

Otherwise, we meet the assumptions of Lemma B.3 and we have

$$\begin{aligned}\mathcal{J}(\mathfrak{n}+1) - \mathcal{J}(\mathfrak{n}) = -\mathcal{E}_{\Theta\circ\Theta}(\mathcal{H}_{\mu_0,\mathfrak{n}},\mathcal{H}_{\mu_0,\mathfrak{n}}) &\overset{(i)}{\leq} -2\zeta\mathcal{E}_\Theta(\mathcal{H}_{\mu_0,\mathfrak{n}},\mathcal{H}_{\mu_0,\mathfrak{n}})\\ &\overset{(ii)}{\leq} -\zeta\operatorname{Var}_{\pi^\star}(\mathcal{H}_{\mu_0,\mathfrak{n}}\cdot\mathbf{1}_\Omega)\Lambda_\Omega\left(\frac{4\left[\mathbb{E}_{\pi^\star}(\mathcal{H}_{\mu_0,\mathfrak{n}}\cdot\mathbf{1}_\Omega)\right]^2}{\operatorname{Var}_{\pi^\star}(\mathcal{H}_{\mu_0,\mathfrak{n}}\cdot\mathbf{1}_\Omega)}\right)\\ &\overset{(iii)}{\leq} -\zeta\cdot(\mathcal{J}(\mathfrak{n})-B)\,\Lambda_\Omega\left(\frac{4}{\mathcal{J}(\mathfrak{n})-B}\right). \qquad \text{(B.19)}\end{aligned}$$

where step (i) follows from inequality (B.17), step (ii) follows from Lemma B.3, and finally step (iii) follows from inequality (B.13) which implies that $\operatorname{Var}_{\pi^\star}(\mathcal{H}_{\mu_0,\mathfrak{n}}\cdot\mathbf{1}_\Omega) \geq \mathcal{J}(\mathfrak{n}) - B$, and the fact that the spectral profile Λ_Ω is a non-increasing function.

Proof of equality (a) in equation (B.16): Since the distribution $\Pi^\star$ is stationary with respect to the kernel Θ, it is also stationary with respect to the two step kernel $\Theta \circ \Theta$. We now prove a more general claim: For any transition kernel K which has stationary distribution $\Pi^\star$ and any measurable function $\mathcal{H}$, the Dirichlet form $\mathcal{E}_K$, defined by replacing Θ with K in equation (B.1), we have

$$\mathcal{E}_K(\mathcal{H}, \mathcal{H}) = \int_{\mathcal{X}} \mathcal{H}^2(x)\pi^\star(x)dx - \int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{H}(x)\mathcal{H}(y)K(x, dy)\pi^\star(x)dx. \tag{B.20}$$

Note that invoking this claim with $K = \Theta \circ \Theta$ and $\mathcal{H} = \mathcal{H}_{\mu_0,\mathfrak{n}}$ implies step (a) in equation (B.16). We now establish the claim (B.20). Expanding the square in the definition (B.1), we obtain that

$$\begin{aligned}
\mathcal{E}_K(\mathcal{H}, \mathcal{H}) &= \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{H}^2(x)K(x, dy)\pi^\star(x)dx + \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{H}^2(y)K(x, dy)\pi^\star(x)dx \\
&\qquad - \int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{H}(x)\mathcal{H}(y)K(x, dy)\pi^\star(x)dx \\
&\overset{(i)}{=} \frac{1}{2}\int_{\mathcal{X}} \mathcal{H}^2(x)\pi^\star(x)dx + \frac{1}{2}\int_{\mathcal{X}} \mathcal{H}^2(x)\pi^\star(x)dx - \int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{H}(x)\mathcal{H}(y)K(x, dy)\pi^\star(x)dx,
\end{aligned}$$

where equality (i) follows from the following facts: For the first term, we use the fact that $\int_{\mathcal{X}} K(x, dy) = 1$ since K is a transition kernel, and, for the second term we use the fact that $\int_{\mathcal{X}} K(x, dy)\pi^\star(x)dx = \pi^\star(y)dy$, since $\Pi^\star$ is the stationary distribution for the kernel K. The claim now follows.

Step (3): Consider the domain extension of the function $\mathcal{J}$ from $\mathbb{N}$ to the set of non-negative real numbers $\mathbb{R}_+$ by piecewise linear interpolation. We abuse notation and denote this extension also by $\mathcal{J}$. The extended function $\mathcal{J}$ is continuous and is differentiable on the set $\mathbb{R}_+ \backslash \mathbb{N}$. Let $\mathfrak{n}^* \in \mathbb{R}_+ \cup \{\infty\}$ denote the first index such that $\mathcal{J}(\mathfrak{n}^*) \leq 3B$. Since Λ_Ω is non-increasing and $\mathcal{J}$ is non-increasing, we have

$$\mathcal{J}'(t) \leq -\zeta \cdot (\mathcal{J}(t) - B)\,\Lambda_\Omega\left(\frac{4}{\mathcal{J}(t) - B}\right) \quad \text{for all } t \in \mathbb{R}_+\backslash\mathbb{N} \text{ such that } t \leq \mathfrak{n}^*. \tag{B.21}$$

Moving the $\mathcal{J}$ terms on one side and integrating for $t \leq \mathfrak{n}^*$, we obtain

$$\int_{\mathcal{J}(0)}^{\mathcal{J}(t)} \frac{d\mathcal{J}}{(\mathcal{J} - B)\cdot \Lambda_\Omega\left(\frac{4}{\mathcal{J}-B}\right)} \leq -\zeta t.$$

Using the change of variable $v = 4/\left(\mathcal{J} - B\right)$, we obtain

$$\zeta t \leq \int_{4/(\mathcal{J}(0)-B)}^{4/(\mathcal{J}(t)-B)} \frac{dv}{v\Lambda_\Omega(v)} \tag{B.22}$$

Furthermore, equation (B.22) implies that for $T \geq \frac{1}{\zeta}\int_{4/\beta}^{8/\delta^2} \frac{dv}{v\Lambda_\Omega(v)}$, we have

$$\int_{4/\beta}^{8/\delta^2} \frac{dv}{v\Lambda_\Omega(v)} \leq \int_{4/(\mathfrak{J}(0)-B)}^{4/(\mathfrak{J}(T)-B)} \frac{dv}{v\Lambda_\Omega(v)}.$$

The bound (B.15) and the fact that $B = \delta^2/3$ imply that $4/(\mathfrak{J}(0) - B) > 4/\beta$. Using this observation, the fact that $0 \leq \Lambda_\Omega(v) < \infty$ for $v \geq 4/\beta$ and combining with the case in Equation (B.18), we conclude that $\mathfrak{J}$ satisfies

$$\mathfrak{J}(T) \leq 3B = \delta^2 \text{ or } \frac{4}{\mathfrak{J}(T) - B} \geq \frac{8}{\delta^2} \quad \text{for } T \geq \frac{1}{\zeta}\int_{4/\beta}^{8/\delta^2} \frac{dv}{v\Lambda(v)},$$

which implies the claimed bound (B.9).

Finally, we turn to the proof of Lemma B.3.

B.1.1.1 Proof of Lemma B.3:

Fix a function $g : \mathcal{X} \to \mathbb{R}_+$ such that $g \in L_2(\pi^\star)$ and $g \cdot \mathbf{1}_\Omega$ is not constant and $\mathbb{E}[g^2 \cdot \mathbf{1}_\Omega] \geq 2\mathbb{E}[g^2 \cdot \mathbf{1}_{\Omega^c}]$. Note that for any constant $\gamma > 0$, we have

$$\begin{aligned}
\mathcal{E}(g, g) &= \frac{1}{2}\int_{(x,y)\in\mathcal{X}^2} (g(x) - g(y))^2\, \Theta(x, dy)\Pi^\star(x)dx \\
&= \frac{1}{2}\int_{(x,y)\in\mathcal{X}^2} ((g(x) - \gamma) - (g(y) - \gamma))^2\, \Theta(x, dy)\Pi^\star(x)dx \\
&= \mathcal{E}\left((g - \gamma), (g - \gamma)\right).
\end{aligned}$$

Let $\gamma = \mathrm{Var}_{\pi^\star}(g \cdot \mathbf{1}_\Omega)/4\mathbb{E}_{\pi^\star}(g \cdot \mathbf{1}_\Omega)$. We have

$$\begin{aligned}
\mathcal{E}(g, g) = \mathcal{E}\left((g - \gamma), (g - \gamma)\right) &\overset{(i)}{\geq} \mathcal{E}\left((g - \gamma)_+, (g - \gamma)_+\right) \\
&\overset{(ii)}{\geq} \mathrm{Var}_{\pi^\star}\left((g - \gamma)_+ \cdot \mathbf{1}_\Omega\right) \inf_{f \in c^+_{0,\Omega}(\{g>\gamma\})} \frac{\mathcal{E}(f, f)}{\mathrm{Var}_{\pi^\star}(f \cdot \mathbf{1}_\Omega)} \\
&\overset{(iii)}{\geq} \mathrm{Var}_{\pi^\star}\left((g - \gamma)_+ \cdot \mathbf{1}_\Omega\right) \cdot \Lambda_\Omega(\Pi^\star(\{g > \gamma\} \cap \Omega)).
\end{aligned} \tag{B.23}$$

Here $(x)_+ = \max\{0, x\}$ denotes the positive part of x. Inequality (i) follows from Lemma 2.3 in [99]. Inequality (iii) follows from the definition (B.4) of Ω-restricted spectral profile. Inequality (ii) follows from the definition of infimum and we need to verify that $\mathbb{E}_{\pi^\star}[(g -$

$c)_+^2\mathbf{1}_{\Omega^c}] \leq \mathbb{E}_{\pi^\star}[(g-c)_+^2\mathbf{1}_\Omega]$. It follow because

$$\begin{aligned}
\mathbb{E}_{\pi^\star}[(g-c)_+^2\mathbf{1}_\Omega] &\overset{(iv)}{\geq} \mathbb{E}_{\pi^\star}[(g\cdot\mathbf{1}_\Omega)^2] - 2\gamma\cdot\mathbb{E}_{\pi^\star}(g\cdot\mathbf{1}_\Omega)\\
&= \frac{\mathbb{E}_{\pi^\star}[(g\cdot\mathbf{1}_\Omega)^2] + (\mathbb{E}_{\pi^\star}[g\cdot\mathbf{1}_\Omega])^2}{2}\\
&\overset{(v)}{\geq} \mathbb{E}_{\pi^\star}[(g\cdot\mathbf{1}_{\Omega^c})^2]\\
&\overset{(vi)}{\geq} \mathbb{E}_{\pi^\star}[((g-c)_+\cdot\mathbf{1}_{\Omega^c})^2],
\end{aligned}$$

where inequality (iv) and (vi) follows from the fact that

$$(a-b)_+^2 \geq a^2 - 2ab \quad \text{and} \quad (a-b)_+ \leq a, \quad \text{for scalars } a, b \geq 0, \tag{B.24}$$

inequality (v) follows from the assumption in Lemma B.3 that $\mathbb{E}_{\pi^\star}[g^2\cdot\mathbf{1}_\Omega] \geq 2\mathbb{E}_{\pi^\star}[g^2\cdot\mathbf{1}_{\Omega^c}]$.

Additionally, we have

$$\begin{aligned}
\mathrm{Var}_{\pi^\star}((g-\gamma)_+\cdot\mathbf{1}_\Omega) &= \mathbb{E}_{\pi^\star}((g-\gamma)_+\cdot\mathbf{1}_\Omega)^2 - [\mathbb{E}_{\pi^\star}((g-\gamma)_+\cdot\mathbf{1}_\Omega)]^2\\
&\overset{(i)}{\geq} \mathbb{E}_{\pi^\star}(g\cdot\mathbf{1}_\Omega)^2 - 2\gamma\cdot\mathbb{E}_{\pi^\star}(g\cdot\mathbf{1}_\Omega) - [\mathbb{E}_{\pi^\star}(g\cdot\mathbf{1}_\Omega)]^2\\
&\geq \mathrm{Var}_{\pi^\star}(g\cdot\mathbf{1}_\Omega) - 2\gamma\cdot\mathbb{E}_{\pi^\star}(g\cdot\mathbf{1}_\Omega),
\end{aligned} \tag{B.25}$$

where inequality (i) follows from the facts in Equation (B.24). Together the choice of γ, we obtain from equation (B.25) that

$$\mathrm{Var}_{\pi^\star}((g-\gamma)_+\mathbf{1}_\Omega) \geq \frac{1}{2}\mathrm{Var}_{\pi^\star}(g\cdot\mathbf{1}_\Omega) \tag{B.26}$$

Furthermore applying Markov's inequality for the non-negative function $g\cdot\mathbf{1}_\Omega$, we also have $\Pi^\star(\{g>\gamma\}\cap\Omega) = \Pi^\star(\{g\cdot\mathbf{1}_\Omega>\gamma\}) \leq [\mathbb{E}_{\pi^\star}(g\cdot\mathbf{1}_\Omega)]/\gamma$. Combing equation (B.23) and (B.26), together with the fact that Λ_Ω is non-increasing, we obtain

$$\mathcal{E}(g,g) \geq \frac{1}{2}\mathrm{Var}_{\pi^\star}(g\cdot\mathbf{1}_\Omega)\cdot\Lambda_\Omega\left(\frac{4\left(\mathbb{E}_{\pi^\star}(g\cdot\mathbf{1}_\Omega)\right)^2}{\mathrm{Var}_{\pi^\star}(g\cdot\mathbf{1}_\Omega)}\right),$$

as claimed in the lemma.

B.1.2 Proof of Lemma B.2

The proof of the Lemma B.2 follows along the lines of Lemma 2.4 in [99], except that we have to deal with continuous-state transition probability. This technical challenge is the main reason for introducing the restricted conductance profile. At a high level, our argument is based on reducing the problem on general functions to a problem on indicator functions, and then using the definition of the conductance. Similar ideas have appeared in the proof of the

Cheeger's inequality [39] and the modified log-Sobolev constants [118].

We split the proof of Lemma B.2 in two cases based on whether $v \in [\frac{4}{\beta}, \frac{\Pi^\star(\Omega)}{2}]$, referred to as Case 1, or $v \geq \frac{\Pi^\star(\Omega)}{2}$, referred to as Case 2.

Case 1: First we consider the case when $v \in [\frac{4}{\beta}, \frac{\Pi^\star(\Omega)}{2}]$. Define $D^+ : L_2(\pi^\star) \to L_2(\pi^\star)$ as

$$D^+(h)(x) = \int_{y\in\mathcal{X}} \left(\mathcal{H}(x) - \mathcal{H}(y)\right)_+ \Theta(x, dy) \text{ and } D^-(h)(x) = \int_{y\in\mathcal{X}} \left(\mathcal{H}(x) - \mathcal{H}(y)\right)_- \Theta(x, dy),$$

where $(x)_+ = \max\{0, x\}$ and (resp. $(\cdot)_-$) denote the positive and negative part of x respectively. We note that D^+ and D^- satisfy the following co-area formula:

$$\mathbb{E}_{\pi^\star} D^+(h) = \int_{-\infty}^{+\infty} \mathbb{E}_{\pi^\star} D^+\left(\mathbf{1}_{h>t}\right) dt. \tag{B.27a}$$

See Lemma 1 in [118] or Lemma 2.4 in [99] for a proof of the equality (B.27a). Moreover, given any measurable set $A \subset \mathcal{X}$, scalar t, and function $h \in c^+_{0,\Omega}(A)$, we note that the term $\mathbb{E}_{\pi^\star} D^+(\mathbf{1}_{h>t})(x)$ is equal to the flow ϕ (defined in equation (4.1)) of the level set $H_t = \{x \in \mathcal{X} \mid \mathcal{H}(x) > t\}$:

$$\mathbb{E}_{\pi^\star} D^+(\mathbf{1}_{h>t}) = \int_{x\in H_t} \Theta(x, H_t^c)\pi^\star(x)dx = \phi(H_t). \tag{B.27b}$$

By the definition of infimum, we have

$$\phi(H_t) \geq \Pi^\star(H_t \cap \Omega) \cdot \inf_{0\leq\Pi^\star(S\cap\Omega)\leq\Pi^\star(A\cap\Omega)} \frac{\phi(S)}{\Pi^\star(S\cap\Omega)}. \tag{B.27c}$$

Combining the previous three equations, we obtain[1]

$$\begin{aligned}\mathbb{E}_{\pi^\star} D^+(h) = \int_{-\infty}^{+\infty} \mathbb{E}_{\pi^\star} D^+\left(\mathbf{1}_{h>t}\right) dt \geq \int_{-\infty}^{+\infty} \Pi^\star(H_t \cap \Omega)dt \cdot \inf_{\substack{S\subset\mathcal{X}\\ 0\leq\Pi^\star(S\cap\Omega)\leq\Pi^\star(A\cap\Omega)}} \frac{\phi(S)}{\Pi^\star(S\cap\Omega)} \\ = \mathbb{E}_{\pi^\star}(h \cdot \mathbf{1}_\Omega) \cdot \Phi_\Omega(\Pi^\star(A\cap\Omega)),\end{aligned}$$

where the last equality follows from that $h \geq 0$ and the definition of the restricted conductance. In a similar fashion, using the fact that $\phi(H_t) = \phi(H_t^c)$, we obtain that

$$\mathbb{E}_{\pi^\star} D^-(h) \geq \mathbb{E}_{\pi^\star}(h \cdot \mathbf{1}_\Omega) \cdot \Phi_\Omega(\Pi^\star(A\cap\Omega)).$$

[1] Note that this step demonstrates that the continuous state-space treatment is different from the discrete state-space one in Lemma 2.4 of [99].

Combining the bounds on $\mathbb{E}_{\pi^\star} D^+(h)$ and $\mathbb{E}_{\pi^\star} D^-(h)$, we obtain

$$\int_{\mathcal{X}} \int_{\mathcal{X}} |\mathcal{H}(x) - \mathcal{H}(y)| \, \Theta(x, dy)\pi^\star(x)dx = \mathbb{E}_{\pi^\star} D^+(h) + \mathbb{E}_{\pi^\star} D^-(h) \geq 2\mathbb{E}_{\pi^\star}(h \cdot \mathbf{1}_\Omega) \cdot \Phi_\Omega(\Pi^\star(A \cap \Omega)).$$

Given any function $g \in c^+_{0,\Omega}(A)$, applying the above inequality by replacing h with g^2, we have

$$\begin{aligned}
&2\mathbb{E}_{\pi^\star}(g^2 \cdot \mathbf{1}_\Omega) \cdot \Phi_\Omega(\Pi^\star(A \cap \Omega)) \\
&\leq \int_{\mathcal{X}} \int_{\mathcal{X}} \left|g^2(x) - g^2(y)\right| \Theta(x, dy)\pi^\star(x)dx \\
&\overset{(i)}{\leq} \left(\int_{\mathcal{X}} \int_{\mathcal{X}} |g(x) - g(y)|^2 \, \Theta(x, dy)\pi^\star(x)dx\right)^{1/2} \cdot \left(\int_{\Omega} \int_{\Omega} |g(x) + g(y)|^2 \, \Theta(x, dy)\pi^\star(x)dx\right)^{1/2} \\
&\overset{(ii)}{\leq} (2\mathcal{E}(g,g))^{1/2} \cdot \left(\int_{\mathcal{X}} \int_{\mathcal{X}} 2\left(g(x)^2 + g(y)^2\right) \Theta(x, dy)\pi^\star(x)dx\right)^{1/2} \\
&= (2\mathcal{E}(g,g))^{1/2} \left(4\mathbb{E}_{\pi^\star}(g^2)\right)^{1/2} \\
&\overset{(iii)}{\leq} (2\mathcal{E}(g,g))^{1/2} \left(8\mathbb{E}_{\pi^\star}(g^2 \cdot \mathbf{1}_\Omega)\right)^{1/2}.
\end{aligned}$$

Rearranging the last equation, we obtain

$$\frac{\mathcal{E}(g,g)}{\mathbb{E}_{\pi^\star}(g^2 \cdot \mathbf{1}_\Omega)} \geq \frac{\Phi^2_\Omega(\Pi^\star(A \cap \Omega))}{4}. \tag{B.28}$$

In the above sequence of steps, inequality (i) follows from the Cauchy-Schwarz inequality, and inequality (ii) from the definition (B.1) and the fact that $(a+b)^2 \leq 2(a^2+b^2)$. Inequality (iii) follows from the assumption in the definition of the spectral profile (B.3a) that $\mathbb{E}_{\pi^\star}[g^2 \cdot \mathbf{1}_{\Omega^c}] \leq \mathbb{E}_{\pi^\star}[g^2 \cdot \mathbf{1}_\Omega]$. Taking infimum over $g \in c^+_{0,\Omega}(A)$ in equation (B.28), we obtain

$$\lambda_\Omega(A) = \inf_{g \in c^+_{0,\Omega}(A)} \frac{\mathcal{E}(g,g)}{\mathrm{Var}_{\pi^\star}(g \cdot \mathbf{1}_\Omega)} \geq \inf_{g \in c^+_{0,\Omega}(A)} \frac{\mathcal{E}(g,g)}{\mathbb{E}_{\pi^\star}(g^2 \cdot \mathbf{1}_\Omega)} \geq \frac{\Phi^2_\Omega(\Pi^\star(A \cap \Omega))}{4},$$

where the first inequality follows from the fact that $\mathbb{E}_{\pi^\star}(g^2 \cdot \mathbf{1}_\Omega) \geq \mathrm{Var}_{\pi^\star}(g \cdot \mathbf{1}_\Omega)$. Given $v \in [0, \frac{\Pi^\star(\Omega)}{2}]$, taking infimum over $\Pi^\star(A \cap \Omega) \leq v$ on both sides, we conclude the claimed bound for this case:

$$\Lambda_\Omega(v) = \inf_{\Pi^\star(A\cap\Omega) \in [0,v]} \lambda_\Omega(A) \geq \inf_{\Pi^\star(A\cap\Omega) \in [0,v]} \frac{\Phi^2_\Omega(\Pi^\star(A \cap \Omega))}{4} = \frac{\Phi^2_\Omega(v)}{4},$$

where the last equality follows from the fact that the conductance profile Φ_Ω defined in equation (4.2) is non-increasing over its domain $[0, \frac{\Pi^\star(\Omega)}{2}]$.

Case 2: Next, we consider the case when $v \geq \frac{\Pi^\star(\Omega)}{2}$. We claim that

$$\Lambda_\Omega(v) \overset{(i)}{\geq} \Lambda_\Omega(\Pi^\star(\Omega)) \overset{(ii)}{\geq} \frac{\Lambda_\Omega(\Pi^\star(\Omega)/2)}{2} \overset{(iii)}{\geq} \frac{\Phi_\Omega(\Pi^\star(\Omega)/2)^2}{8}, \tag{B.29}$$

where step (i) follows from the fact that the spectral profile Λ is a non-increasing function, and step (iii) from the result of Case 1. Note that the bound from Lemma B.2 for this case follows from the bound above. It remains to establish inequality (ii), which we now prove.

Note that given the definition (B.4), it suffices to establish that

$$\frac{\mathcal{E}\left(g, g\right)}{\mathrm{Var}_{\pi^\star}(g \cdot \mathbf{1}_\Omega)} \geq \frac{\Lambda_\Omega(\Pi^\star(\Omega)/2)}{2} \quad \text{for all functions } g \in L_2(\pi^\star). \tag{B.30}$$

Consider any $g \in L_2(\pi^\star)$ and let $\nu \in \mathbb{R}$ be such that

$$\Pi^\star(\{g > \nu\} \cap \Omega) = \Pi^\star(\{g < \nu\} \cap \Omega) = \frac{\Pi^\star(\Omega)}{2}.$$

Using the same argument as in the proof of Lemma B.3 and Lemma 2.3 in [99], we have

$$\begin{aligned} \mathcal{E}\left(g, g\right) &= \mathcal{E}\left((g-\nu), (g-\nu)\right) \\ &\geq \mathcal{E}\left((g-\nu)_+, (g-\nu)_+\right) + \mathcal{E}\left((g-\nu)_-, (g-\nu)_-\right). \end{aligned} \tag{B.31}$$

For the two terms above, we have

$$\mathcal{E}\left((g-\nu)_+, (g-\nu)_+\right) \geq \mathbb{E}_{\pi^\star}\left((g-\nu)_+^2 \cdot \mathbf{1}_\Omega\right) \cdot \inf_{f \in c_{0,\Omega}^+(\{g>\nu\})} \frac{\mathcal{E}\left(f, f\right)}{\mathbb{E}_{\pi^\star}\left(f \cdot \mathbf{1}_\Omega\right)^2}, \tag{B.32}$$

and similarly

$$\mathcal{E}\left((g-\nu)_-, (g-\nu)_-\right) \geq \mathbb{E}_{\pi^\star}\left((g-\nu)_-^2 \cdot \mathbf{1}_\Omega\right) \cdot \inf_{f \in c_{0,\Omega}^+(\{g<\nu\})} \frac{\mathcal{E}\left(f, f\right)}{\mathbb{E}_{\pi^\star}\left(f \cdot \mathbf{1}_\Omega\right)^2}. \tag{B.33}$$

For $f \in c_{0,\Omega}^+(\{g > \nu\})$, we have $f \cdot \mathbf{1}_\Omega \in c_{0,\Omega}^+(\{g > \nu\} \cap \Omega)$. Using Cauchy-Schwarz inequality, we have

$$\mathbb{E}_{\pi^\star}\left(f \cdot \mathbf{1}_\Omega\right)^2 = \int_{x \in \{g>\nu\}\cap\Omega} f(x)^2 \Pi^\star(x) dx \geq \frac{\left(\int_{x\in\{g>\nu\}\cap\Omega} |f(x)|\, \Pi^\star(x) dx\right)^2}{\Pi^\star(\{g > \nu\} \cap \Omega)} \geq \frac{(\mathbb{E}_{\pi^\star} f \cdot \mathbf{1}_\Omega)^2}{\Pi^\star(\{g > \nu\} \cap \Omega)}$$

Using this bound and noting the ν is chosen such that $\Pi^\star(\{g > \nu\} \cap \Omega) = \Pi^\star(\Omega)/2$, for $f \in c_{0,\Omega}^+(\{g > \nu\})$, we have

$$\mathrm{Var}_{\pi^\star}(f \cdot \mathbf{1}_\Omega) = \mathbb{E}_{\pi^\star}\left(f \cdot \mathbf{1}_\Omega\right)^2 - \left(\mathbb{E}_{\pi^\star} f \cdot \mathbf{1}_\Omega\right)^2 \geq \mathbb{E}_{\pi^\star}\left(f \cdot \mathbf{1}_\Omega\right)^2 \cdot \left(1 - \frac{\Pi^\star(\Omega)}{2}\right). \tag{B.34}$$

Putting the equations (B.31), (B.32), (B.33) and (B.34) together, we obtain

$$\begin{aligned} \mathcal{E}\left(g, g\right) &\geq \mathbb{E}_{\pi^\star}\left((g-\nu)^2 \cdot \mathbf{1}_\Omega\right) \cdot \left(1 - \frac{\Pi^\star(\Omega)}{2}\right) \cdot \inf_{\Pi^\star(S) \in [0, \frac{\Pi^\star(\Omega)}{2}]} \inf_{f \in c_{0,\Omega}^+(S)} \frac{\mathcal{E}(f, f)}{\mathrm{Var}_{\pi^\star}(f \cdot \mathbf{1}_\Omega)} \\ &\geq \mathrm{Var}_{\pi^\star}(g \cdot \mathbf{1}_\Omega) \cdot \frac{1}{2} \cdot \Lambda_\Omega(\Pi^\star(\Omega)/2). \end{aligned}$$

which implies the claim (B.30) and we conclude Case 2 of Lemma B.2.

B.2 Proof of Proposition 4.2

The proof of this proposition is partly similar to the conductance-based proof of Lemma 3.4. In addition to it, we have to deal with the case when target distribution satisfies the logarithmic isoperimetric inequality.

For any set A_1 such that $\Pi^\star(A_1 \cap \Omega) \leq \frac{\Pi^\star(\Omega)}{2}$, with its complement denoted by $A_2 = \mathcal{X} \backslash A_1$, we have $\Pi^\star(A_2 \cap \Omega) \geq \frac{\Pi^\star(\Omega)}{2} \geq \Pi^\star(A_1 \cap \Omega)$, since $\Pi^\star(A_1 \cap \Omega) + \Pi^\star(A_2 \cap \Omega) = \Pi^\star(\Omega)$. We claim that

$$\int_{x \in A_1} \Theta(x, A_2) \pi^\star(x) dx \geq \Pi^\star(A_1 \cap \Omega) \cdot \frac{\rho}{4} \cdot \min\left\{1, \frac{\Delta}{16\psi_{\mathfrak{c}}} \cdot \log^{\mathfrak{c}}\left(1 + \frac{1}{\Pi^\star(A_1 \cap \Omega)}\right)\right\}. \tag{B.35}$$

Note that the claim (4.10) of Proposition 4.2 can be directly obtained from the claim (B.35), by dividing both sides by $\Pi^\star(A_1 \cap \Omega)$, taking infimum with respect to A_1 such $\Pi^\star(A_1 \cap \Omega) \in (0, v]$ and noting that $\inf_{t \in (0,v]} \log^{\frac{1}{2}}(1 + 1/t) = \log^{\frac{1}{2}}(1 + 1/v)$.

We now prove the claim (B.35).

Define the following sets,

$$A_1' := \left\{x \in A_1 \cap \Omega \mid \Theta(x, A_2) < \frac{\rho}{2}\right\}, \quad A_2' := \left\{x \in A_2 \cap \Omega \mid \Theta(x, A_1) < \frac{\rho}{2}\right\}, \tag{B.36}$$

along with the complement $A_3' := \Omega \setminus (A_1' \cup A_2')$. Note that $A_i' \subset \Omega$ for $i = 1, 2, 3$. We split the proof into two distinct cases:

- Case 1: $\Pi^\star(A_1') \leq \Pi^\star(A_1 \cap \Omega)/2$ or $\Pi^\star(A_2') \leq \Pi^\star(A_2 \cap \Omega)/2$.
- Case 2: $\Pi^\star(A_1') > \Pi^\star(A_1 \cap \Omega)/2$ and $\Pi^\star(A_2') > \Pi^\star(A_2 \cap \Omega)/2$.

Note that these cases are mutually exclusive and exhaustive. We now consider these cases one by one.

Case 1: If we have $\Pi^\star(A_1') \leq \Pi^\star(A_1 \cap \Omega)/2$, then

$$\Pi^\star(A_1 \cap \Omega \setminus A_1') \geq \Pi^\star(A_1 \cap \Omega)/2. \tag{B.37}$$

We have

$$\begin{aligned}\int_{x \in A_1} \Theta(x, A_2) \pi^\star(x) dx \geq \int_{x \in A_1 \cap \Omega \setminus A_1'} \Theta(x, A_2) \pi^\star(x) dx &\overset{(i)}{\geq} \frac{\rho}{2} \int_{x \in A_1 \cap \Omega \setminus A_1'} \pi^\star(x) dx \\ &\overset{(ii)}{\geq} \frac{\rho}{4} \Pi^\star(A_1 \cap \Omega),\end{aligned}$$

where inequality (i) follows from the definition of the set A_1' in equation (B.36) and inequality (ii) follows from equation (B.37). For the case $\Pi^\star(A_2') \leq \Pi^\star(A_2 \cap \Omega)/2$, we use a similar argument with the role of A_1 and A_2 exchanged to obtain

$$\int_{x \in A_1} \Theta(x, A_2)\pi^\star(x)dx = \int_{x \in A_2} \Theta(x, A_1)\pi^\star(x)dx \geq \frac{\rho}{4}\Pi^\star(A_2 \cap \Omega).$$

Putting the pieces together for this case, we have established that

$$\int_{x \in A_1} \Theta(x, A_2)\pi^\star(x)dx \geq \frac{\rho}{4} \min\left\{\Pi^\star(A_1 \cap \Omega), \Pi^\star(A_2 \cap \Omega)\right\} = \frac{\rho}{4}\Pi^\star(A_1 \cap \Omega). \tag{B.38}$$

Case 2: We have $\Pi^\star(A_1') > \Pi^\star(A_1 \cap \Omega)/2$ and $\Pi^\star(A_2') > \Pi^\star(A_2 \cap \Omega)/2$. We first show that in this case the sets A_1' and A_2' are far away, and then we invoke the logarithmic isoperimetry inequality from Lemma C.6.

For any two vectors $u \in A_1'$ and $v \in A_2'$, we have

$$d_{\text{TV}}\big(\mathcal{T}_u, \mathcal{T}_v\big) \geq \Theta(u, A_1) - \Theta(v, A_1) = 1 - \Theta(u, A_2) - \Theta(v, A_1) > 1 - \rho.$$

Consequently, the assumption of the lemma implies that

$$d(A_1', A_2') \geq \Delta. \tag{B.39}$$

Using the fact that under the stationary distribution, the flow from A_1 to A_2 is equal to that from A_2 to A_1, we obtain

$$\begin{aligned}
\int_{x \in A_1} \Theta(x, A_2)\pi^\star(x)dx &= \frac{1}{2}\left(\int_{x \in A_1} \Theta(x, A_2)\pi^\star(x)dx + \int_{x \in A_2} \Theta(x, A_1)\pi^\star(x)dx\right) \\
&\geq \frac{1}{4}\left(\int_{x \in A_1 \cap \Omega \setminus A_1'} \Theta(x, A_2)\pi^\star(x)dx + \int_{x \in x \in A_2 \cap \Omega \setminus A_2'} \Theta(x, A_1)\pi^\star(x)dx\right) \\
&\geq \frac{\rho}{8}\Pi^\star(\Omega \setminus (A_1' \cup A_2')),
\end{aligned} \tag{B.40}$$

where the last inequality follows from the definition of the set A_1' in equation (B.36). Note that the sets A_1', A_2' and $\mathcal{X} \setminus (A_1' \cup A_2')$ partition $\mathcal{X}$. Using the condition (4.8) with the Ω-restricted distribution $\Pi_\Omega^\star$ with density $\pi_\Omega^\star$ defined as

$$\pi_\Omega^\star(x) = \frac{\pi^\star(x)\mathbf{1}_\Omega(x)}{\Pi^\star(\Omega)},$$

we obtain

$$\begin{aligned}
&\Pi^\star(\Omega \setminus (A_1' \cap A_2')) \\
&= \Pi^\star(\Omega) \cdot \Pi^\star_\Omega(\mathcal{X} \setminus (A_1' \cap A_2')) \\
&\overset{(i)}{\geq} \Pi^\star(\Omega) \cdot \frac{d(A_1', A_2')}{2\psi_{\mathfrak{e}}} \cdot \min\left\{\Pi^\star_\Omega(A_1'), \Pi^\star_\Omega(A_2')\right\} \cdot \log^{\mathfrak{e}}\left(1 + \frac{1}{\min\left\{\Pi^\star_\Omega(A_1'), \Pi^\star_\Omega\,(A_2')\right\}}\right) \\
&\overset{(ii)}{\geq} \Pi^\star(\Omega) \cdot \frac{\Delta}{4\psi_{\mathfrak{e}}} \min\left\{\Pi^\star(A_1 \cap \Omega), \Pi^\star(A_2 \cap \Omega)\right\} \cdot \log^{\mathfrak{e}}\left(1 + \frac{2}{\min\left\{\Pi^\star(A_1 \cap \Omega), \Pi^\star(A_2 \cap \Omega)\right\}}\right) \\
&\geq \frac{1}{2} \cdot \frac{\Delta}{4\psi_{\mathfrak{e}}} \cdot \Pi^\star(A_1 \cap \Omega) \cdot \log^{\mathfrak{e}}\left(1 + \frac{1}{\Pi^\star(A_1 \cap \Omega)}\right),
\end{aligned} \tag{B.41}$$

where step (i) follows from the assumption (4.8), step (ii) from the bound (B.39) and the facts that $\Pi^\star_\Omega(A_i') \geq \Pi^\star(A_i') \geq \frac{1}{2}\Pi^\star(A_i \cap \Omega)$ and that the map $x \mapsto x\log^{\mathfrak{e}}(1 + 1/x)$ is an increasing function for either $\mathfrak{e} = \frac{1}{2}$ or $\mathfrak{e} = 0$. Putting the pieces (B.40) and (B.41) together, we conclude that

$$\int_{x \in A_1} \Theta(x, A_2)\pi^\star(x)dx \geq \frac{\rho}{16} \cdot \frac{\Delta}{4\psi_{\mathfrak{e}}} \cdot \Pi^\star(A_1 \cap \Omega) \cdot \log^{\mathfrak{e}}\left(1 + \frac{1}{\Pi^\star(A_1 \cap \Omega)}\right). \tag{B.42}$$

Finally, the claim (B.35) follows from combining the two bounds (B.38) and (B.42) from the two separate cases.

Appendix C

Content Deferred from Chapter 5

We prove Lemma 5.1 and Corollary 5.1 in Appendices C.1 and C.2 respectively, and provide a detailed discussion about the trade-off between hyperparameter choices for HMC from our proofs in Appendix C.3.

C.1 Proof of Lemma 5.1

We first provide several convenient properties about the HMC proposal.

C.1.1 Properties of the HMC proposal

Recall the Hamiltonian Monte Carlo (HMC) with leapfrog integrator (5.3c). Using an induction argument, we find that the final states in one iteration of $\mathcal{K}$ steps of the HMC chain, denoted by $\mathbf{q}_{\mathcal{K}}$ and $\mathbf{p}_{\mathcal{K}}$ satisfy

$$\mathbf{p}_{\mathcal{K}} = \mathbf{p}_0 - \frac{\eta}{2}\nabla f(\mathbf{q}_0) - \sum_{j=1}^{\mathcal{K}-1}\nabla f(\mathbf{q}_j) - \frac{\eta}{2}\nabla f(\mathbf{q}_{\mathcal{K}}), \tag{C.1a}$$

$$\text{and} \quad \mathbf{q}_{\mathcal{K}} = \mathbf{q}_0 + \mathcal{K}\eta\mathbf{p}_0 - \frac{\mathcal{K}\eta^2}{2}\nabla f(\mathbf{q}_0) - \eta^2\sum_{j=1}^{\mathcal{K}-1}(\mathcal{K}-j)\,\nabla f(\mathbf{q}_j). \tag{C.1b}$$

It is easy to see that for $k \in [\mathcal{K}]$, $\mathbf{q}_k$ can be seen as a function of the initial state $\mathbf{q}_0$ and $\mathbf{p}_0$. We denote this function as the *forward mapping* F,

$$\mathbf{q}_k =: F_k(\mathbf{p}_0, \mathbf{q}_0) \quad \text{and} \quad \mathbf{q}_{\mathcal{K}} =: F_{\mathcal{K}}(\mathbf{p}_0, \mathbf{q}_0) =: F\left(\mathbf{p}_0, \mathbf{q}_0\right) \tag{C.1c}$$

where we introduced the simpler notation $F := F_{\mathcal{K}}$ for the final iterate. The forward mappings F_k and F are deterministic functions that only depends on the gradient ∇f, the number of leapfrog updates $\mathcal{K}$ and the step size η.

Denote $\mathbf{J}_x F$ as the Jacobian matrix of the forward mapping F with respect to the first variable. By definition, it satisfies

$$[\mathbf{J}_x F(x, q_0)]_{ij} = \frac{\partial}{\partial x_j} [F(x, q_0)]_i \,, \quad \text{for all} \quad i, j \in [d] \,. \tag{C.1d}$$

Similarly, denote $\mathbf{J}_y F$ as the Jacobian matrix of the forward mapping F with respect to the second variable. The following lemma characterizes the eigenvalues of the Jacobian $\mathbf{J}_x F$.

Lemma C.1. *Suppose the log density f is $\mathcal{L}$-smooth. For the number of leapfrog steps and step-size satisfying $\mathcal{K}^2\eta^2 \leq \frac{1}{4\mathcal{L}}$, we have*

$$\left\|\mathcal{K}\eta \mathbb{I}_d - \mathbf{J}_x F(x, y)\right\|_2 \leq \frac{1}{8}\mathcal{K}\eta, \quad \textit{for all} \quad x, y \in \mathcal{X} \ \textit{and}\ i \in [d] \,.$$

Also all eigenvalues of $\mathbf{J}_x F(x, y)$ have absolute value greater or equal to $\frac{7}{8}\mathcal{K}\eta$.

See Appendix C.1.3 for the proof.

Since the Jacobian is invertible for $\mathcal{K}^2\eta^2 \leq \frac{1}{4\mathcal{L}}$, we can define the inverse function of F with respect to the first variable as the backward mapping G. We have

$$F\left(G(x, y), y\right) = x, \quad \text{for all} \quad x, y \in \mathcal{X}. \tag{C.2}$$

Moreover as a direct consequence of Lemma C.1, we obtain that the magnitude of the eigenvalues of the Jacobian matrix $\mathbf{J}_x G(x, y)$ lies in the interval $\left[\frac{8}{9\mathcal{K}\eta}, \frac{8}{7\mathcal{K}\eta}\right]$. In the next lemma, we state another set of bounds on different Jacobian matrices:

Lemma C.2. *Suppose the log density f is $\mathcal{L}$-smooth. For the number of leapfrog steps and step-size satisfying $\mathcal{K}^2\eta^2 \leq \frac{1}{4\mathcal{L}}$, we have*

$$\left\|\mathbf{J}_y G(x, y)\right\|_2 \leq \frac{4}{3\mathcal{K}\eta}, \quad \textit{for all} \quad x, y \in \mathcal{X}, \quad \textit{and} \tag{C.3a}$$

$$\left\|\frac{\partial F_k(G(x, y), y)}{\partial y}\right\|_2 \leq 3, \qquad \textit{for all} \quad k \in [\mathcal{K}] \,. \tag{C.3b}$$

See Appendix C.1.4 for the proof.

Next, we would like to obtain a bound on the quantity $\frac{\partial \log \det \mathbf{J}_x G(x, \mathbf{q}_0)}{\partial y}$. Applying the chain rule, we find that

$$\frac{\partial \log \det \mathbf{J}_x G(x, \mathbf{q}_0)}{\partial y} = \begin{bmatrix} \operatorname{trace}\left([\mathbf{J}_x G(x, \mathbf{q}_0)]^{-1} \mathbf{J}_{xy_1} G(x, \mathbf{q}_0)\right) \\ \vdots \\ \operatorname{trace}\left([\mathbf{J}_x G(x, q_0)]^{-1} \mathbf{J}_{xy_d} G(x, \mathbf{q}_0)\right) \end{bmatrix}. \tag{C.4}$$

Here $\mathbf{J}_{xy}G(x, \mathbf{q}_0)$ is a third order tensor and we use $\mathbf{J}_{xy_l}G(x, \mathbf{q}_0)$ to denote the matrix corresponding to the l-th slice of the tensor which satisfies

$$[\mathbf{J}_{xy_l}G(x, \mathbf{q}_0)]_{ij} = \frac{\partial\partial}{\partial x_j y_l}\,[F(x, \mathbf{q}_0)]_i\,, \qquad \text{for all} \quad i, j, l \in [d]\,.$$

Lemma C.3. *Suppose the log density f is $\mathcal{L}$-smooth and $\mathcal{L}_H$-Hessian Lipschitz. For the number of leapfrog steps and step-size satisfying $\mathcal{K}^2\eta^2 \leq \frac{1}{4\mathcal{L}}$, we have*

$$\left\|\frac{\partial \log \det \mathbf{J}_x G(x, \mathbf{q}_0)}{\partial y}\right\|_2 = \left\|\begin{bmatrix} \text{trace}\left([\mathbf{J}_x G(x, \mathbf{q}_0)]^{-1}\mathbf{J}_{xy_1}G(x, \mathbf{q}_0)\right) \\ \vdots \\ \text{trace}\left([\mathbf{J}_x G(x, q_0)]^{-1}\mathbf{J}_{xy_d}G(x, \mathbf{q}_0)\right)\end{bmatrix}\right\|_2 \leq 2d\mathcal{K}^2\eta^2\mathcal{L}_H.$$

See Appendix C.1.5 for the proof.

As a direct consequence of the equation (C.1b) at k-th step of leapfrog updates, we obtain the following two bounds for the difference between successive F_k terms that come in handy later in our proofs.

Lemma C.4. *Suppose that the log density f is $\mathcal{L}$-smooth. For the number of leapfrog steps and step-size satisfying $\mathcal{K}^2\eta^2 \leq \frac{1}{4\mathcal{L}}$, we have*

$$\left\|F_k(\mathbf{p}_0, \mathbf{q}_0) - \mathbf{q}_0\right\|_2 \leq 2k\eta \left\|\mathbf{p}_0\right\|_2 + 2k^2\eta^2 \left\|\nabla f(\mathbf{q}_0)\right\|_2 \quad \textit{for } k \in [\mathcal{K}]\,, \ \textit{and} \tag{C.5a}$$

$$\left\|F_{k+1}(\mathbf{p}_0, \mathbf{q}_0) - F_k(\mathbf{p}_0, \mathbf{q}_0)\right\|_2 \leq 2\eta \left\|\mathbf{p}_0\right\|_2 + 2(k+1)\eta^2 \left\|\nabla f(\mathbf{q}_0)\right\|_2 \ \textit{for } k \in [\mathcal{K} - 1]\,. \tag{C.5b}$$

See Appendix C.1.6 for the proof.

We now turn to the proof the two claims in Lemma 5.1. Note that the claim (5.16a) states that the proposal distributions at two close points are close; the claim (5.16b) states that the proposal distribution and the transition distribution are close.

C.1.1.1 Proof of claim (5.16a) in Lemma 5.1

In order to bound the distance between proposal distributions of nearby points, we prove the following stronger claim: For a $\mathcal{L}$-smooth $\mathcal{L}_H$-Hessian-Lipschitz target distribution, the proposal distribution of the HMC algorithm with step size η and leapfrog steps $\mathcal{K}$ such that $\mathcal{K}\eta \leq \frac{1}{4\mathcal{L}}$ satisfies

$$d_{\text{TV}}\left(\mathcal{P}_{\mathbf{q}_0}, \mathcal{P}_{\widetilde{\mathbf{q}}_0}\right) \leq \left(\frac{2\left\|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0\right\|_2^2}{\mathcal{K}^2\eta^2} + 3\sqrt{d}\mathcal{K}\eta\mathcal{L}\left\|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0\right\|_2 + 4d\mathcal{K}^2\eta^2\mathcal{L}_H\left\|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0\right\|_2\right)^{1/2}, \tag{C.6}$$

for all $\mathbf{q}_0, \widetilde{\mathbf{q}}_0 \in \mathbb{R}^d$. Then for any two points $\mathbf{q}_0, \widetilde{\mathbf{q}}_0$ such that $\|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0\|_2 \leq \frac{1}{4}\mathcal{K}\eta$, under the condition (5.15a), i.e., $\mathcal{K}^2\eta^2 \leq \frac{1}{4\max\left\{d^{\frac{1}{2}}\mathcal{L}, d^{\frac{2}{3}}\mathcal{L}_H^{\frac{2}{3}}\right\}}$, we have

$$d_{\mathrm{TV}}\left(\mathcal{P}_{\mathbf{q}_0}, \mathcal{P}_{\widetilde{\mathbf{q}}_0}\right) \leq \left(\frac{1}{8} + \frac{3}{64} + \frac{1}{64}\right)^{1/2} \leq \frac{1}{2},$$

and the claim (5.16a) follows.

The proof of claim (C.6) involves the following steps: (1) we make use of the update rules (C.1b) and change of variable formula to obtain an expression for the density of $\mathbf{q}_\mathfrak{n}$ in terms of $\mathbf{q}_0$, (2) then we use Pinsker's inequality and derive expressions for the KL-divergence between the two proposal distributions, and (3) finally, we upper bound the KL-divergence between the two distributions using different properties of the forward mapping F from Appendix C.1.1.

According to the update rule (C.1b), the proposals from two initial points $\mathbf{q}_0$ and $\widetilde{\mathbf{q}}_0$ satisfy respectively

$$\mathbf{q}_{\mathcal{K}} = F(\mathbf{p}_0, \mathbf{q}_0), \quad \text{and} \quad \widetilde{\mathbf{q}}_{\mathcal{K}} = F(\widetilde{\mathbf{p}}_0, \widetilde{\mathbf{q}}_0),$$

where $\mathbf{p}_0$ and $\widetilde{\mathbf{p}}_0$ are independent random variable from Gaussian distribution $\mathcal{N}(0, \mathbb{I}_d)$.

Denote $p_{\mathbf{q}_0}$ as the density function of the proposal distribution $\mathcal{P}_{\mathbf{q}_0}$. For two different initial points $\mathbf{q}_0$ and $\widetilde{\mathbf{q}}_0$, the goal is to bound the total variation distance between the two proposal distribution, which is by definition

$$d_{\mathrm{TV}}\left(\mathcal{P}_{\mathbf{q}_0}, \mathcal{P}_{\widetilde{\mathbf{q}}_0}\right) = \frac{1}{2}\int_{x\in\mathcal{X}} \left|p_{\mathbf{q}_0}(x) - p_{\widetilde{\mathbf{q}}_0}(x)\right| dx. \tag{C.7}$$

Given $\mathbf{q}_0$ fixed, the random variable $\mathbf{q}_{\mathcal{K}}$ can be seen as a transformation of the Gaussian random variable $\mathbf{p}_0$ through the function $F(\cdot, \mathbf{q}_0)$. When F is invertible, we can use the change of variable formula to obtain an explicit expression of the density $p_{\mathbf{q}_0}$:

$$p_{\mathbf{q}_0}(x) = \varphi\left(G(x, \mathbf{q}_0)\right) \det\left(\mathbf{J}_x G(x, \mathbf{q}_0)\right), \tag{C.8}$$

where φ is the density of the standard Gaussian distribution $\mathcal{N}(0, \mathbb{I}_d)$. Note that even though explicit, directly bounding the total variation distance (C.7) using the complicated density expression (C.8) is difficult. We first use Pinsker's inequality [58] to give an upper bound of the total variance distance in terms of KL-divergence

$$d_{\mathrm{TV}}\left(\mathcal{P}_{\mathbf{q}_0}, \mathcal{P}_{\widetilde{\mathbf{q}}_0}\right) \leq \sqrt{2\,\mathrm{KL}(\mathcal{P}_{\mathbf{q}_0} \| \mathcal{P}_{\widetilde{\mathbf{q}}_0})}, \tag{C.9}$$

and then upper bound the KL-divergence. Plugging the density (C.8) into the KL-divergence formula, we obtain that

$$\begin{aligned}
\mathrm{KL}(\mathcal{P}_{\mathbf{q}_0} \| \mathcal{P}_{\widetilde{\mathbf{q}}_0}) &= \int_{\mathbb{R}^d} p_{\mathbf{q}_0}(x) \log\left(\frac{p_{\mathbf{q}_0}(x)}{p_{\widetilde{\mathbf{q}}_0}(x)}\right) dx \\
&= \int_{\mathbb{R}^d} p_{\mathbf{q}_0}(x) \left[\log\left(\frac{\varphi\left(G(x,\mathbf{q}_0)\right)}{\varphi\left(G(x,\widetilde{\mathbf{q}}_0)\right)}\right) + \log\det \mathbf{J}_x G(x,\mathbf{q}_0) - \log\det \mathbf{J}_x G(x,\widetilde{\mathbf{q}}_0)\right] dx \\
&= \underbrace{\int_{\mathbb{R}^d} p_{\mathbf{q}_0}(x) \left[\frac{1}{2}\left(-\|G(x,\mathbf{q}_0)\|_2^2 + \|G(x,\widetilde{\mathbf{q}}_0)\|_2^2\right)\right] dx}_{T_1} \\
&\qquad + \underbrace{\int_{\mathbb{R}^d} p_{\mathbf{q}_0}(x) \left[\log\det \mathbf{J}_x G(x,\mathbf{q}_0) - \log\det \mathbf{J}_x G(x,\widetilde{\mathbf{q}}_0)\right] dx}_{T_2} \qquad \text{(C.10)}
\end{aligned}$$

We claim the following bounds on the terms T_1 and T_2:

$$|T_1| \leq \frac{8}{9}\frac{\|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0\|_2^2}{\mathcal{K}^2\eta^2} + \frac{3}{2}\sqrt{d}\mathcal{K}\eta\mathcal{L}\,\|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0\|_2\,, \quad \text{and} \qquad \text{(C.11a)}$$

$$|T_2| \leq 2d\mathcal{K}^2\eta^2\mathcal{L}_H\,\|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0\|_2\,, \qquad \text{(C.11b)}$$

where the bound on T_2 follows readily from Lemma C.3:

$$\begin{aligned}
|T_2| &= \left|\int p_{\mathbf{q}_0}(x) \left[\log\det \mathbf{J}_x G(x,\mathbf{q}_0) - \log\det \mathbf{J}_x G(x,\widetilde{\mathbf{q}}_0)\right] dx\right| \\
&\leq \left\|\frac{\partial \log\det \mathbf{J}_x G(x,\mathbf{q}_0)}{\partial y}\right\|_2 \|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0\|_2 \\
&\leq 2d\mathcal{K}^2\eta^2\mathcal{L}_H\,\|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0\|_2\,. \qquad \text{(C.12)}
\end{aligned}$$

Putting together the inequalities (C.9), (C.10), (C.11a) and (C.11b) yields the claim (C.6).

It remains to prove the bound (C.11a) on T_1.

Proof of claim (C.11a): For the term T_1, we observe that

$$\frac{1}{2}\left(\|G(x,\widetilde{\mathbf{q}}_0)\|_2^2 - \|G(x,\mathbf{q}_0)\|_2^2\right) = \frac{1}{2}\|G(x,\mathbf{q}_0) - G(x,\widetilde{\mathbf{q}}_0)\|_2^2 - \left(G(x,\mathbf{q}_0) - G(x,\widetilde{\mathbf{q}}_0)\right)^\top G(x,\mathbf{q}_0).$$

The first term on the RHS can be bounded via the Jacobian of G with respect to the second variable. Applying the bound (C.3a) from Lemma C.2, we find that

$$\|G(x,\mathbf{q}_0) - G(x,\widetilde{\mathbf{q}}_0)\|_2 \leq \|\!|\mathbf{J}_y G(x,y)|\!\|_2\,\|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0)\|_2 \leq \frac{4}{3\mathcal{K}\eta}\|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0)\|_2\,. \qquad \text{(C.13)}$$

For the second part, we claim that there exists a deterministic function C of $\mathbf{q}_0$ and $\widetilde{\mathbf{q}}_0$ and independent of x, such that

$$\|G(x, \mathbf{q}_0) - G(x, \widetilde{\mathbf{q}}_0) - C(\mathbf{q}_0, \widetilde{\mathbf{q}}_0)\|_2 \leq \frac{3}{2}\mathcal{K}\eta\mathcal{L} \|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0\|_2 . \tag{C.14}$$

Assuming the claim (C.14) as given at the moment, we can further decompose the second part of T_1 into two parts:

$$(G(x, \mathbf{q}_0) - G(x, \widetilde{\mathbf{q}}_0))^\top G(x, \mathbf{q}_0) = (G(x, \mathbf{q}_0) - G(x, \widetilde{\mathbf{q}}_0) - C(\mathbf{q}_0, \widetilde{\mathbf{q}}_0))^\top G(x, \mathbf{q}_0) + C(\mathbf{q}_0, \widetilde{\mathbf{q}}_0)^\top G(x, \mathbf{q}_0) \tag{C.15}$$

Applying change of variables along with equation (C.8), we find that

$$\int p_{\mathbf{q}_0}(x) G(x, \mathbf{q}_0) dx = \int \varphi(x) x dx = 0.$$

Furthermore, we also have

$$\begin{aligned}\int_{x \in \mathcal{X}} p_{\mathbf{q}_0}(x) \|G(x, \mathbf{q}_0)\|_2 \, dx &= \int_{x \in \mathcal{X}} \varphi(x) \|x\|_2 \, dx \\ &\overset{(i)}{\leq} \left[\left(\int_{x \in \mathcal{X}} \varphi(x) \|x\|_2^2 \, dx\right)\left(\int_{x \in \mathcal{X}} \varphi(x) dx\right)\right]^{1/2} = \sqrt{d},\end{aligned}$$

where step (i) follows from Cauchy-Schwarz's inequality. Combining the inequalities (C.13), (C.14) and (C.15) together, we obtain the following bound on term T_1:

$$\begin{aligned}|T_1| &= \left|\int p_{\mathbf{q}_0}(x) \left[-\frac{1}{2} \|G(x, \mathbf{q}_0)\|_2^2 + \frac{1}{2} \|G(x, \widetilde{\mathbf{q}}_0)\|_2^2\right] dx\right| \\ &\leq \frac{1}{2}\left|\int p_{\mathbf{q}_0}(x) \|G(x, \mathbf{q}_0) - G(x, \widetilde{\mathbf{q}}_0)\|_2^2 \, dx\right| \\ &\qquad + \left|\int p_{\mathbf{q}_0}(x) \|G(x, \mathbf{q}_0) - G(x, \widetilde{\mathbf{q}}_0) - C(\mathbf{q}_0, \widetilde{\mathbf{q}}_0)\|_2 \|G(x, \mathbf{q}_0)\|_2 \, dx\right| \\ &\leq \frac{8}{9} \frac{\|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0\|_2^2}{\mathcal{K}^2\eta^2} + \frac{3}{2}\sqrt{d}\mathcal{K}\eta \|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0\|_2 ,\end{aligned} \tag{C.16}$$

which yields the claimed bound on T_1.

We now prove our earlier claim (C.14).

Proof of claim (C.14): For any pair of states $\mathbf{q}_0$ and $\widetilde{\mathbf{q}}_0$, invoking the definition (C.2) of the map $G(x, \cdot)$, we obtain the following implicit equations:

$$x = \mathbf{q}_0 + \mathcal{K}\eta G(x, \mathbf{q}_0) - \mathcal{K}\frac{\eta^2}{2}\nabla f(\mathbf{q}_0) - \eta^2 \sum_{j=1}^{\mathcal{K}-1}(\mathcal{K}-j)\nabla f(F_j(G(x, \mathbf{q}_0), \mathbf{q}_0)), \quad \text{and}$$

$$x = \widetilde{\mathbf{q}}_0 + \mathcal{K}\eta G(x, \widetilde{\mathbf{q}}_0) - \mathcal{K}\frac{\eta^2}{2}\nabla f(\widetilde{\mathbf{q}}_0) - \eta^2 \sum_{j=1}^{\mathcal{K}-1}(\mathcal{K}-j)\nabla f(F_j(G(x, \widetilde{\mathbf{q}}_0), \widetilde{\mathbf{q}}_0)).$$

Taking the difference between the two equations above, we obtain

$$\begin{aligned} G(x, \mathbf{q}_0) - G(x, \widetilde{\mathbf{q}}_0) - \frac{\mathbf{q}_0 - \widetilde{\mathbf{q}}_0}{\mathcal{K}\eta} - \frac{\eta}{2}\left(\nabla f(\mathbf{q}_0) - \nabla f(\widetilde{\mathbf{q}}_0)\right) & \\ = \frac{\eta^2}{\mathcal{K}\eta}\sum_{k=1}^{\mathcal{K}-1}(\mathcal{K}-j)\left(\nabla f(F_k(G(x, \mathbf{q}_0), \mathbf{q}_0)) - \nabla f(F_k(G(x, \widetilde{\mathbf{q}}_0), \widetilde{\mathbf{q}}_0))\right). & \end{aligned}$$

Applying $\mathcal{L}$-smoothness of f along with the bound (C.3b) from Lemma C.2, we find that

$$\begin{aligned} \left\|\nabla f(F_k(G(x, \mathbf{q}_0), \mathbf{q}_0)) - \nabla f(F_k(G(x, \widetilde{\mathbf{q}}_0), \widetilde{\mathbf{q}}_0))\right\|_2 &\leq \mathcal{L}\left|\!\left|\!\left|\frac{\partial F_k(G(x, y), y)}{\partial y}\right|\!\right|\!\right|_2 \left\|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0\right\|_2 \\ &\leq 3\mathcal{L}\left\|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0\right\|_2. \end{aligned}$$

Putting the pieces together, we find that

$$\left\|G(x, \mathbf{q}_0) - G(x, \widetilde{\mathbf{q}}_0) - \frac{\mathbf{q}_0 - \widetilde{\mathbf{q}}_0}{\mathcal{K}\eta} - \frac{1}{2}\left(\nabla f(\mathbf{q}_0) - \nabla f(\widetilde{\mathbf{q}}_0)\right)\right\|_2 \leq \frac{3\mathcal{K}\eta\mathcal{L}}{2}\left\|\mathbf{q}_0 - \widetilde{\mathbf{q}}_0\right\|_2,$$

which yields the claim (C.14).

C.1.1.2 Proof of claim (5.16b) in Lemma 5.1

We now bound the distance between the one-step proposal distribution $\mathcal{P}_x$ at point x and the one-step transition distribution $\mathcal{T}_x^{\text{before-lazy}}$ at x obtained after performing the accept-reject step (and no lazy step). Using equation (C.1a), we define the forward mapping E for the variable $\mathbf{p}_{\mathcal{K}}$ as follows

$$\mathbf{p}_{\mathcal{K}} = E(\mathbf{p}_0, \mathbf{q}_0) := \mathbf{p}_0 - \frac{\eta}{2}\nabla f(\mathbf{q}_0) - \eta\sum_{j=1}^{\mathcal{K}-1}\nabla f(\mathbf{q}_j) - \frac{\eta}{2}\nabla f(\mathbf{q}_{\mathcal{K}}).$$

Consequently, the probability of staying at x is given by

$$\mathcal{T}_x^{\text{before-lazy}}(\{x\}) = 1 - \int_{\mathcal{X}} \min\left\{1, \frac{\exp(-\mathcal{H}(E(z, x), F(z, x)))}{\exp(-\mathcal{H}(z, x))}\right\}\varphi_x(z)dz,$$

where the Hamiltonian $\mathcal{H}(q,p) = f(q) + \frac{1}{2}\|p\|_2^2$ was defined in equation (5.2). As a result, the TV-distance between the proposal and transition distribution is given by

$$\begin{aligned} d_{\mathrm{TV}}\big(\mathcal{P}_x, \mathcal{T}_x^{\text{before-lazy}}\big) &= 1 - \int_{\mathcal{X}} \min\left\{1, \frac{\exp(-\mathcal{H}(E(z,x), F(z,x)))}{\exp(-\mathcal{H}(z,x))}\right\}\varphi_x(z)dz \\ &= 1 - \mathbb{E}_{z\sim\mathcal{N}(0,\mathbb{I}_d)}\left[\min\left\{1, \frac{\exp(-\mathcal{H}(E(z,x), F(z,x)))}{\exp(-\mathcal{H}(z,x))}\right\}\right]. \qquad \text{(C.17)} \end{aligned}$$

An application of Markov's inequality yields that

$$\begin{aligned} &\mathbb{E}_{z\sim\mathcal{N}(0,\mathbb{I}_d)}\left[\min\left\{1, \frac{\exp(-\mathcal{H}(E(z,x), F(z,x)))}{\exp(-\mathcal{H}(z,x))}\right\}\right] \\ &\qquad \geq \alpha \mathbb{P}_{z\sim\mathcal{N}(0,\mathbb{I}_d)}\left[\frac{\exp(-\mathcal{H}(E(z,x), F(z,x)))}{\exp(-\mathcal{H}(z,x))} \geq \alpha\right], \qquad \text{(C.18)} \end{aligned}$$

for any $\alpha \in (0,1]$. Thus, to bound the distance $d_{\mathrm{TV}}\big(\mathcal{P}_x, \mathcal{T}_x^{\text{before-lazy}}\big)$, it suffices to derive a high probability lower bound on the ratio $\exp(-\mathcal{H}(E(z,x), F(z,x)))/\exp(-\mathcal{H}(z,x))$ when $z \sim \mathcal{N}(0,\mathbb{I}_d)$.

We now derive a lower bound on the following quantity:

$$\exp\left(-f(F(\mathbf{p}_0,\mathbf{q}_0)) + f(\mathbf{q}_0) - \frac{1}{2}\|E(\mathbf{p}_0,\mathbf{q}_0)\|_2^2 + \frac{1}{2}\|\mathbf{p}_0\|_2^2\right), \quad \text{when } \mathbf{p}_0 \sim \mathcal{N}(0,\mathbb{I}_d).$$

We derive the bounds on the two terms $-f(F(\mathbf{p}_0,\mathbf{q}_0)) + f(\mathbf{q}_0)$ and $\|E(\mathbf{p}_0,\mathbf{q}_0)\|_2^2$ separately.

Observe that

$$f(F(\mathbf{p}_0,\mathbf{q}_0)) - f(\mathbf{q}_0) = \sum_{j=0}^{\mathcal{K}-1}\left[f(F_{j+1}(\mathbf{p}_0,\mathbf{q}_0)) - f(F_j(\mathbf{p}_0,\mathbf{q}_0))\right].$$

The intuition is that it is better to apply Taylor expansion on closer points. Applying the third order Taylor expansion and using the smoothness assumptions (3.5a) and (5.5) for the function f, we obtain

$$f(x) - f(y) \leq \frac{(x-y)^\top}{2}(\nabla f(x) + \nabla f(y)) + \mathcal{L}_H\|x-y\|_2^3.$$

For the indices $j \in \{0,\ldots,\mathcal{K}-1\}$, using F_j as the shorthand for $F_j(\mathbf{p}_0,\mathbf{q}_0)$, we find that

$$\begin{aligned} f(F_{j+1}) - f(F_j) &\leq \frac{(F_{j+1}-F_j)^\top}{2}(\nabla f(F_{j+1}) + \nabla f(F_j)) + \mathcal{L}_H\|F_{j+1}-F_j\|_2^3 \\ &= \frac{1}{2}\eta\mathbf{p}_0^\top(\nabla f(F_{j+1}) + \nabla f(F_j)) \\ &\quad - \frac{\eta^2}{2}\left[\frac{1}{2}\nabla f(\mathbf{p}_0) + \sum_{k=1}^{j}\nabla f(F_k)\right]^\top(\nabla f(F_{j+1}) + \nabla f(F_j)) + \mathcal{L}_H\|F_{j+1}-F_j\|_2^3, \end{aligned}$$

(C.19)

where the last equality follows by definition (C.1c) of the operator F_j.

Now to bound the term $E(\mathbf{p}_0, \mathbf{q}_0)$, we observe that

$$\begin{aligned}\frac{\|E(\mathbf{p}_0, \mathbf{q}_0)\|_2^2}{2} &= \frac{\left\|\mathbf{p}_0 - \frac{\eta}{2}\nabla f(\mathbf{q}_0) - \eta\sum_{j=1}^{\mathcal{K}-1}\nabla f(F_j) - \frac{\eta}{2}\nabla f(F_{\mathcal{K}})\right\|_2^2}{2}\\ &= \frac{\|\mathbf{p}_0\|_2^2}{2} - \eta\mathbf{p}_0^\top\left(\frac{1}{2}\nabla f(\mathbf{q}_0) + \sum_{j=1}^{\mathcal{K}-1}\nabla f(F_j) + \frac{1}{2}\nabla f(F_{\mathcal{K}})\right)\\ &\qquad + \frac{\eta^2}{2}\Big\|\frac{1}{2}\nabla f(\mathbf{q}_0) + \sum_{j=1}^{\mathcal{K}-1}\nabla f(F_j) + \frac{1}{2}\nabla f(F_{\mathcal{K}})\Big\|_2^2. \end{aligned} \tag{C.20}$$

Putting the equations (C.19) and (C.20) together leads to cancellation of many gradient terms and we obtain

$$\begin{aligned}&- f(F(\mathbf{p}_0, \mathbf{q}_0)) + f(\mathbf{q}_0) - \frac{1}{2}\left\|E(\mathbf{p}_0, \mathbf{q}_0)\right\|_2^2 + \frac{1}{2}\left\|\mathbf{p}_0\right\|_2^2\\ &\geq \frac{\eta^2}{8}\left(\nabla f(\mathbf{q}_0) - \nabla f(F_{\mathcal{K}})\right)^\top\left(\nabla f(\mathbf{q}_0) + \nabla f(F_{\mathcal{K}})\right) - \mathcal{L}_H\sum_{j=0}^{\mathcal{K}-1}\|F_{j+1} - F_j\|_2^3\\ &\geq -\frac{\eta^2\mathcal{L}}{4}\left\|\mathbf{q}_0 - F(\mathbf{p}_0, \mathbf{q}_0)\right\|_2\left\|\nabla f(\mathbf{q}_0)\right\|_2 - \frac{\eta^2\mathcal{L}^2}{2}\left\|\mathbf{q}_0 - F(\mathbf{p}_0, \mathbf{q}_0)\right\|_2^2 - \mathcal{L}_H\sum_{j=0}^{\mathcal{K}-1}\|F_{j+1} - F_j\|_2^3\end{aligned} \tag{C.21}$$

The last inequality uses the smoothness condition (3.5a) for the function f. Plugging the bounds (C.5a) and (C.5b) in equation (C.21), we obtain a lower bound that only depends on $\|\mathbf{p}_0\|_2$ and $\|\nabla f(\mathbf{q}_0)\|_2$:

$$\begin{aligned}\text{RHS of (C.21)} \geq -2\mathcal{K}^2\eta^4\mathcal{L}^2\left\|\mathbf{p}_0\right\|_2^2 - 2\mathcal{K}\eta^3\mathcal{L}\left\|\mathbf{p}_0\right\|_2\left\|\nabla f(\mathbf{q}_0)\right\|_2 - 2\mathcal{K}^2\eta^4\mathcal{L}\left\|\nabla f(\mathbf{q}_0)\right\|_2^2\\ -\mathcal{L}_H\left(32\mathcal{K}\eta^3\left\|\mathbf{p}_0\right\|_2^3 + 8\mathcal{K}^4\eta^6\left\|\nabla f(\mathbf{q}_0)\right\|_2^3\right).\end{aligned} \tag{C.22}$$

According to assumption (5A), we have bounded gradient in the convex set Ω. For any $x \in \Omega$, we have $\|\nabla f(x)\|_2 \leq \mathcal{M}$. Standard Chi-squared tail bounds imply that

$$\mathbb{P}\left[\|\mathbf{p}_0\|_2^2 \leq d\alpha_1\right] \geq 1 - \frac{1}{16}, \quad \text{for } \alpha_1 = 1 + 2\sqrt{\log(16)} + 2\log(16). \tag{C.23}$$

Plugging the gradient bound and the bound (C.23) into equation (C.22), we conclude that there exists an absolute constant $c \leq 2000$ such that for η^2 satisfying equation (5.15b), namely

$$\eta^2 \leq \frac{1}{c\mathcal{L}}\min\left\{\frac{1}{\mathcal{K}^2}, \frac{1}{\mathcal{K}d^{\frac{1}{2}}}, \frac{1}{\mathcal{K}^{\frac{2}{3}}d^{\frac{1}{3}}\left(\frac{\mathcal{M}^2}{\mathcal{L}}\right)^{\frac{1}{3}}}, \frac{1}{\mathcal{K}\frac{\mathcal{M}}{\mathcal{L}^{\frac{1}{2}}}}, \frac{1}{\mathcal{K}^{\frac{2}{3}}d}\frac{\mathcal{L}}{\mathcal{L}_H^{\frac{2}{3}}}, \frac{1}{\mathcal{K}^{\frac{4}{3}}\frac{\mathcal{M}}{\mathcal{L}^{\frac{1}{2}}}}\left(\frac{\mathcal{L}}{\mathcal{L}_H^{\frac{2}{3}}}\right)^{\frac{1}{2}}\right\},$$

we have

$$\mathbb{P}\left[-f(F(\mathbf{p}_0, \mathbf{q}_0)) + f(\mathbf{q}_0) - \frac{1}{2}\left\|E(\mathbf{p}_0, \mathbf{q}_0)\right\|_2^2 + \frac{1}{2}\left\|\mathbf{p}_0\right\|_2^2 \geq -1/16\right] \geq 1 - \frac{1}{16}.$$

Plugging this bound in the inequality (C.18) yields that

$$\mathbb{E}_{z\sim\mathcal{N}(0,\mathbb{I}_d)}\left[\min\left\{1, \frac{\exp(-\mathcal{H}(E(z,x), F(z,x)))}{\exp(-\mathcal{H}(z,x))}\right\}\right] \geq 1 - \frac{1}{8},$$

which when plugged in equation (C.17) implies that $d_{\mathrm{TV}}\big(\mathcal{P}_x, \mathcal{T}_x^{\text{before-lazy}}\big) \leq 1/8$ for any $x \in \mathcal{R}_s$, as claimed. The proof is now complete.

C.1.2 Notation for proofs related to Lemma 5.1

We now collect some notation for the proofs of Lemmas C.1, C.2, C.3 and C.4. For simplicity, we adopt following the tensor notation.

Notations for tensor: Let $\mathcal{T} \in \mathbb{R}^{d\times d\times d}$ be a third order tensor. Let $U \in \mathbb{R}^{d\times d_1}$, $V \in \mathbb{R}^{d\times d_2}$, and $W \in \mathbb{R}^{d\times d_3}$ be three matrices. Then the multi-linear form applied on (U, V, W) is a tensor in $\mathbb{R}^{d_1\times d_2\times d_3}$:

$$\left[\mathcal{T}(U, V, W)\right]_{p,q,r} = \sum_{i,j,k\in[d]} \mathcal{T}_{ijk} U_{ip} V_{jq} W_{kr}.$$

In particular, for the vectors $u, v, w \in \mathbb{R}^d$, the quantity $\mathcal{T}(u, v, w)$ is a real number that depends linearly on u, v, w (tensor analogue of the quantity $u^\top Mv$ in the context of matrices and vector). Moreover, the term $\mathcal{T}(u, v, \mathbb{I}_d)$ denotes a vector in $\mathbb{R}^d$ (tensor analogue of the quantity Mv in the context of matrices and vector). Finally, the term $\mathcal{T}(u, \mathbb{I}_d, \mathbb{I}_d)$ represents a matrix in $\mathbb{R}^{d\times d}$.

C.1.3 Proof of Lemma C.1

We will prove an equivalent statement: for $\mathcal{K}^2\eta^2 \leq \frac{1}{4\mathcal{L}}$, there is a matrix $Q(x, y) \in \mathbb{R}^{d\times d}$ with $\|Q\|_2 \leq \frac{1}{8}$ such that

$$\mathbf{J}_x F(x, y) = \mathcal{K}\eta\left(\mathbb{I}_d - Q(x, y)\right), \quad \text{for all } x, y \in \mathcal{X}. \tag{C.24}$$

Recall from equation (C.1b) that the intermediate iterate $\mathbf{q}_k$ is defined recursively as

$$\mathbf{q}_k = F_k(\mathbf{p}_0, \mathbf{q}_0) = \mathbf{q}_0 + k\eta\mathbf{p}_0 - \frac{k\eta^2}{2}\nabla f(\mathbf{q}_0) - \eta^2\sum_{j=1}^{k-1}(k-j)\nabla f(\mathbf{q}_j) \quad \text{for} \quad 1 \leq k \leq \mathcal{K}.$$

Taking partial derivative with respective to the first variable, we obtain

$$\frac{\partial}{\partial \mathbf{p}_0}\mathbf{q}_k = \mathbf{J}_{\mathbf{p}_0}F_k(\mathbf{p}_0, \mathbf{q}_0) = k\eta\mathbb{I}_d - \eta^2 \sum_{j=1}^{k-1}(k-j)\nabla^2 f_{\mathbf{q}_j}\mathbf{J}_{\mathbf{p}_0}F_j(\mathbf{p}_0, \mathbf{q}_0), \tag{C.25}$$

where $\nabla^2 f_{\mathbf{q}_j}$ is the Hessian of f at $\mathbf{q}_j$. We claim that for $1 \le k \le \mathcal{K}$, there is a matrix $Q_k \in \mathbb{R}^{d\times d}$ with $\|Q_k\|_2 \le \frac{1}{8}$ such that

$$\mathbf{J}_{\mathbf{p}_0}F_k(\mathbf{p}_0, \mathbf{q}_0) = k\eta\left(\mathbb{I}_d - Q_k\right). \tag{C.26}$$

Note that substituting $k = \mathcal{K}$ in this claim yields the result of the lemma. We now prove the claim (C.26) using strong induction.

Base case ($k = 1, 2$): For the base case $k = 1, 2$, using equation (C.25), we have

$$\begin{aligned}
&\mathbf{J}_{\mathbf{p}_0}F_1(\mathbf{p}_0, \mathbf{q}_0) = \eta\mathbb{I}_d, \quad \text{and} \\
&\mathbf{J}_{\mathbf{p}_0}F_2(\mathbf{p}_0, \mathbf{q}_0) = 2\eta\mathbb{I}_d - \eta^2\nabla^2 f_{\mathbf{q}_1}\mathbf{J}_{\mathbf{p}_0}F_1(\mathbf{p}_0, \mathbf{q}_0) = 2\eta\left(\mathbb{I}_d - \frac{\eta^2}{2}\nabla^2 f_{\mathbf{q}_1}\right).
\end{aligned}$$

Combining the inequality $\|\nabla^2 f_{\mathbf{q}_1}\|_2 \le \mathcal{L}$ from smoothness assumption and the assumed stepsize bound $\eta^2 \le \frac{1}{4\mathcal{L}}$ yields

$$\|\frac{\eta^2}{2}\nabla^2 f_{\mathbf{q}_1}\|_2 \le \frac{1}{8}.$$

The statement in equation (C.26) is verified for $k = 1, 2$.

Inductive step: Assuming that the hypothesis holds for all iterations up to k, we now establish it for iteration $k + 1$. We have

$$\begin{aligned}
\mathbf{J}_{\mathbf{p}_0}F_{k+1}(\mathbf{p}_0, \mathbf{q}_0) &= (k+1)\eta\mathbb{I}_d - \eta^2\sum_{j=1}^{k}(k+1-j)\nabla^2 f_{\mathbf{q}_j}\mathbf{J}_{\mathbf{p}_0}F_j(\mathbf{p}_0, \mathbf{q}_0) \\
&\overset{(i)}{=} (k+1)\eta\mathbb{I}_d - \eta^2\sum_{j=1}^{k}(k+1-j)\nabla^2 f_{\mathbf{q}_j}\cdot j\eta\left(\mathbb{I}_d - Q_j\right) \\
&= (k+1)\eta(\mathbb{I}_d - Q_{k+1}),
\end{aligned}$$

where $Q_{k+1} = \frac{\eta^2}{k+1}\sum_{j=1}^{k}(k+1-j)j\nabla^2 f_{\mathbf{q}_j}(\mathbb{I}_d - Q_j)$. Equality (i) follows from the hypothesis of the induction. Finally, we verify that the spectral norm of Q_{k+1} is bounded by $\frac{1}{8}$,

$$\begin{aligned}
\|Q_{k+1}\|_2 &\le \frac{1}{k+1}\sum_{j=1}^{k}\|\eta^2(k+1-j)j\nabla^2 f_{\mathbf{q}_j}\|_2\|\mathbb{I}_d - Q_j\|_2\\
&\overset{(i)}{\le} \frac{1}{k+1}\sum_{j=1}^{k}\|\eta^2\frac{\mathcal{K}^2}{4}\nabla^2 f_{\mathbf{q}_j}\|_2\|\mathbb{I}_d - Q_j\|_2\\
&\overset{(ii)}{\le} \frac{1}{k+1}\sum_{j=1}^{k}\frac{1}{16}\left(1+\frac{1}{8}\right)\\
&\le \frac{1}{8}.
\end{aligned}$$

Inequality (i) follows from the inequality $(k+1-j)j \le \left(\frac{k+1-j+j}{2}\right)^2 \le \frac{\mathcal{K}^2}{4}$. Inequalilty (ii) follows from the assumption $\mathcal{K}^2\eta^2 \le \frac{1}{4\mathcal{L}}$ and the hypothesis $\|Q_j\|_2 \le \frac{1}{8}$. This completes the induction.

C.1.4 Proof of Lemma C.2

Recall that the backward mapping G is defined implicitly as

$$x = y + \mathcal{K}\eta G(x,y) - \frac{\mathcal{K}\eta^2}{2}\nabla f(y) - \eta^2\sum_{k=1}^{\mathcal{K}-1}(\mathcal{K}-k)\nabla f\left(F_k(G(x,y),y)\right). \tag{C.27}$$

First we check the derivatives of $F_k(G(x,y),y)$. Since $F_k(G(x,y),y)$ satisfies

$$F_k(G(x,y),y) = y + k\eta G(x,y) - \frac{k\eta^2}{2}\nabla f(y) - \eta^2\sum_{j=1}^{k-1}(k-j)\nabla f(F_j(G(x,y),y)),$$

taking derivative with respect to y, we obtain

$$\begin{aligned}
\frac{\partial}{\partial y}F_k(G(x,y),y) &= \mathbb{I}_d + k\eta\mathbf{J}_y G(x,y) - \frac{k\eta^2}{2}\nabla^2 f(y)\\
&\quad - \eta^2\sum_{j=1}^{k-1}(k-j)\nabla^2 f(F_j(G(x,y),y))\frac{\partial}{\partial y}F_j(G(x,y),y).
\end{aligned} \tag{C.28}$$

Using the same proof idea as in the previous lemma, we show by induction that for $1 \le k \le \mathcal{K}$, there exists matrices $A_k, B_k \in \mathbb{R}^{d\times d}$ with $\|A_k\|_2 \le \frac{1}{6}$ and $\|B_k\|_2 \le \frac{1}{8}$ such that

$$\frac{\partial}{\partial y}F_k(G(x,y),y) = (\mathbb{I}_d - A_k) + k\eta\left(\mathbb{I}_d - B_k\right)\mathbf{J}_y G(x,y). \tag{C.29}$$

Case $k = 1$: The case $k = 1$ can be easily checked according to equation (C.28), we have

$$\frac{\partial}{\partial y} F_1\left(G(x,y), y\right) = \mathbb{I}_d - \frac{\eta^2}{2}\nabla^2 f(y) + \eta \mathbf{J}_y G(x,y)$$

It is sufficient to set $A_1 = \frac{\eta^2}{2}\nabla^2 f(y)$ and $B_1 = 0$.

Case k to $k+1$: Assume the statement is verified until $k \geq 1$. For $k+1 \leq K$, according to equation (C.28), we have

$$\begin{aligned}
&\frac{\partial}{\partial y} F_{k+1}(G(x,y),y) \\
&= \mathbb{I}_d + (k+1)\eta \mathbf{J}_y G(x,y) - \frac{(k+1)\eta^2}{2}\nabla^2 f(y) - \eta^2 \sum_{j=1}^{k}(k+1-j)\nabla^2 f(F_j(G(x,y),y))\frac{\partial}{\partial y}F_j(G(x,y),y) \\
&= \mathbb{I}_d - \frac{(k+1)\eta^2}{2}\nabla^2 f(y) + (k+1)\eta \mathbf{J}_y G(x,y) \\
&\quad - \eta^2 \sum_{j=1}^{k}(k+1-j)\nabla^2 f(F_j(G(x,y),y))\left((\mathbb{I}_d - A_j) + j\eta\left(\mathbb{I}_d - B_j\right)\mathbf{J}_y G(x,y)\right) \\
&= \mathbb{I}_d - \frac{(k+1)\eta^2}{2}\nabla^2 f(y) - \eta^2 \sum_{j=1}^{k}(k+1-j)\nabla^2 f(F_j(G(x,y),y))(\mathbb{I}_d - A_j) \\
&\quad + (k+1)\eta \mathbf{J}_y G(x,y) - \eta^2 \sum_{j=1}^{k}(k+1-j)\nabla^2 f(F_j(G(x,y),y))\left(j\eta\left(\mathbb{I}_d - B_j\right)\mathbf{J}_y G(x,y)\right)
\end{aligned}$$

To conclude, it suffices to note the following values of A_{k+1} and B_{k+1}:

$$\begin{aligned}
A_{k+1} &= \frac{(k+1)\eta^2}{2}\nabla^2 f(y) + \eta^2 \sum_{j=1}^{k}(k+1-j)\nabla^2 f(F_j(G(x,y),y))(\mathbb{I}_d - A_j), \quad \text{and} \\
B_{k+1} &= \frac{1}{k+1}\eta^2 \sum_{j=1}^{k}(k+1-j)j\nabla^2 f(F_j(G(x,y),y))\left(\mathbb{I}_d - B_j\right).
\end{aligned}$$

We now have the following operator norm bounds:

$$\begin{aligned}
|\!|\!|A_{k+1}|\!|\!|_2 &\leq \frac{k+1}{2}\eta^2\mathcal{L} + \eta^2 \sum_{j=1}^{k}(k+1-j)\mathcal{L}(1+\frac{1}{6}) \leq \frac{7}{12}(k+1)^2\eta^2\mathcal{L} \leq \frac{1}{6}, \quad \text{and} \\
|\!|\!|B_{k+1}|\!|\!|_2 &\leq \frac{1}{k+1}\eta^2(1+\frac{1}{8})\mathcal{L}\sum_{j=1}^{k}(k+1-j)j = \frac{9}{8\cdot 6}k(k-1)\eta^2\mathcal{L} \leq \frac{1}{8}.
\end{aligned}$$

This concludes the proof of equation (C.29). As a particular case, for $k = \mathcal{K}$, we observe that

$$F_{\mathcal{K}}\left(G(x,y),y\right) = x.$$

Plugging it into equation (C.29), we obtain that

$$\mathbf{J}_y G(x,y) = \frac{1}{\mathcal{K}\eta}\left(\mathbb{I}_d - B_{\mathcal{K}}\right)^{-1}\left(\mathbb{I}_d - A_{\mathcal{K}}\right) \quad \Longrightarrow \quad \|\mathbf{J}_y G(x,y)\|_2 \le \frac{4}{3\mathcal{K}\eta}.$$

Plugging the bound on $\|\mathbf{J}_y G(x,y)\|_2$ back to equation (C.29) for other k, we obtain

$$\|\frac{\partial}{\partial y}F_k(G(x,y),y)\|_2 \le 3.$$

This concludes the proof of Lemma C.2.

C.1.5 Proof of Lemma C.3

Recall that the backward mapping G is defined implicitly as

$$x = y + \mathcal{K}\eta G(x,y) - \frac{\mathcal{K}\eta^2}{2}\nabla f(y) - \eta^2\sum_{k=1}^{\mathcal{K}-1}(\mathcal{K}-k)\nabla f\left(F_k(G(x,y),y)\right). \tag{C.30}$$

First we check the derivatives of $F_k(G(x,y),y)$. Since $F_k(G(x,y),y)$ satisfies

$$F_k(G(x,y),y) = y + k\eta G(x,y) - \frac{k\eta^2}{2}\nabla f(y) - \eta^2\sum_{j=1}^{k-1}(k-j)\nabla f(F_j(G(x,y),y)),$$

we have

$$\frac{\partial}{\partial x}F_k(G(x,y),y) = k\eta\mathbf{J}_x G(x,y) - \eta^2\sum_{j=1}^{k-1}(k-j)\nabla^2 f(F_j(G(x,y),y))\frac{\partial}{\partial x}F_j(G(x,y),y). \tag{C.31}$$

Similar to the proof of equation (C.26), we show by induction (proof omitted) that for $1 \le k \le \mathcal{K}$, there exists matrices $\widetilde{Q}_k \in \mathbb{R}^{d\times d}$ with $\|\widetilde{Q}_k\|_2 \le \frac{1}{2}$ such that

$$\frac{\partial}{\partial x}F_k(G(x,y),y) = k\eta\left(\mathbb{I}_d - \widetilde{Q}_k\right)\mathbf{J}_x G(x,y). \tag{C.32}$$

Then, by taking another derivative with respect to y_i in equation (C.31), we obtain

$$\begin{aligned}\frac{\partial\partial}{\partial x\partial y_i}F_k(G(x,y),y) &= k\eta\mathbf{J}_{xy_i}G(x,y)\\ &\quad - \eta^2\sum_{j=1}^{k-1}(k-j)\left\{\quad \nabla^3 f_{F_j(G(x,y),y)}\left(\frac{\partial F_j(G(x,y),y)}{\partial y_i},\mathbb{I}_d,\mathbb{I}_d\right)\frac{\partial}{\partial x}F_j(G(x,y),y)\right.\\ &\qquad\qquad \left. + \nabla^2 f_{F_j(G(x,y),y)}\frac{\partial\partial}{\partial x\partial y_i}F_j(G(x,y),y)\right\}\end{aligned} \tag{C.33}$$

Now we show by induction that for $1 \leq k \leq \mathcal{K}$, for any $\alpha \in \mathbb{R}^d$, we have

$$\left\| \sum_{i=1}^{d} \alpha_i \left(\frac{\partial \partial}{\partial x \partial y_i} F_k(G(x,y),y) \mathbf{J}_x G(x,y)^{-1} \right) \right\|_2 \leq 2k\eta \left\| \sum_{i=1}^{d} \alpha_i \left(\mathbf{J}_{xy_i} G(x,y) \mathbf{J}_x G(x,y)^{-1} \right) \right\|_2 + 2 \left\| \alpha \right\|_2 k^3 \eta^3 \mathcal{L}_H. \tag{C.34}$$

Case $k = 1$: We first examine the case $k = 1$. According to equation (C.33), we have

$$\sum_{i=1}^{d} \alpha_i \left(\frac{\partial \partial}{\partial x \partial y_i} F_1(G(x,y),y) \mathbf{J}_x G(x,y)^{-1} \right) = \eta \sum_{i=1}^{d} \alpha_i \left(\mathbf{J}_{xy_i} G(x,y) \mathbf{J}_x G(x,y)^{-1} \right).$$

The statement in equation (C.34) is easily verified for $k = 1$.

Case k to $k+1$: Assume the statement (C.34) is verified until k. For $k+1 \leq \mathcal{K}$, according to equation (C.33), we have

$$\begin{aligned}
&\sum_{i=1}^{d} \alpha_i \left(\frac{\partial \partial}{\partial x \partial y_i} F_{k+1}(G(x,y),y) \mathbf{J}_x G(x,y)^{-1} \right) \\
&= (k+1)\eta \sum_{i=1}^{d} \alpha_i \left(\mathbf{J}_{xy_i} G(x,y) \mathbf{J}_x G(x,y)^{-1} \right) \\
&\quad - \eta^2 \sum_{j=1}^{k} (k+1-j) \left\{ \nabla^3 f_{F_j(G(x,y),y)} \left(\sum_{i=1}^{d} \alpha_i \frac{\partial F_j(G(x,y),y)}{\partial y_i}, \mathbb{I}_d, \mathbb{I}_d \right) \frac{\partial}{\partial x} F_j(G(x,y),y) \mathbf{J}_x G(x,y)^{-1} \right\} \\
&\quad - \eta^2 \sum_{j=1}^{k} (k+1-j) \, \nabla^2 f_{F_j(G(x,y),y)} \sum_{i=1}^{d} \alpha_i \left(\frac{\partial \partial}{\partial x \partial y_i} F_j(G(x,y),y) \mathbf{J}_x G(x,y)^{-1} \right).
\end{aligned}$$

In the last equality, we have used the fact that $\nabla^3 f_{F_j(G(x,y),y)}$ is a multilinear form to enter the coefficients α_i in the tensor. Let

$$M_\alpha = \left\| \sum_{i=1}^{d} \alpha_i \left(\mathbf{J}_{xy_i} G(x,y) \mathbf{J}_x G(x,y)^{-1} \right) \right\|_2.$$

Applying the hypothesis of the induction, we obtain

$$\begin{aligned}
&\left\| \sum_{i=1}^{d} \alpha_i \left(\frac{\partial \partial}{\partial x \partial y_i} F_{k+1}(G(x,y),y) \mathbf{J}_x G(x,y)^{-1} \right) \right\|_2 \\
&\overset{(i)}{\leq} (k+1)\eta M_\alpha + \eta^2 \sum_{j=1}^{k} 4(k+1-j) j \mathcal{L}_H \left\| \alpha \right\|_2 + \eta^2 \sum_{j=1}^{k} (k+1-j) \mathcal{L} \left(2j\eta M + 2 \left\| \alpha \right\|_2 j^3 \eta^3 \mathcal{L}_H \right) \\
&\leq 2(k+1)\eta M_\alpha + 2 \left\| \alpha \right\|_2 (k+1)^3 \eta^3 \mathcal{L}_H.
\end{aligned}$$

The first inequality (i) used the second part of Lemma C.2 to bound $\|\frac{\partial}{\partial}F_k(G(x,y),y)\|_2$. This completes the induction. As a particular case for $k = \mathcal{K}$, we note that

$$F_{\mathcal{K}}(G(x,y),y) = F(G(x,y),y) = x,$$

and equation (C.33) for $k = \mathcal{K}$ gives

$$\begin{aligned} 0 &= \mathcal{K}\eta \mathbf{J}_{xy_i} G(x,y) \\ &- \eta^2 \sum_{j=1}^{\mathcal{K}-1} (\mathcal{K} - j) \Bigg\{ \nabla^3 f_{F_j(G(x,y),y)} \left(\frac{\partial F_j(G(x,y),y)}{\partial y_i}, \mathbb{I}_d, \mathbb{I}_d \right) \frac{\partial}{\partial x} F_j(G(x,y),y) \\ &+ \nabla^2 f_{F_j(G(x,y),y)} \frac{\partial\partial}{\partial x \partial y_i} F_j(G(x,y),y) \Bigg\}. \end{aligned}$$

Using the bound in equation (C.34), we have

$$\mathcal{K}\eta \left\| \sum_{i=1}^{d} \alpha_i \mathbf{J}_{xy_i} G(x,y) \mathbf{J}_x G(x,y)^{-1} \right\|_2 \leq \|\alpha\|_2 \mathcal{K}^3 \eta^3 \mathcal{L}_H + \frac{1}{2}\mathcal{K}\eta \left\| \sum_{i=1}^{d} \alpha_i \mathbf{J}_{xy_i} G(x,y) \mathbf{J}_x G(x,y)^{-1} \right\|_2 .$$

Hence, we obtain

$$\operatorname{trace}\left(\sum_{i=1}^{d} \alpha_i \mathbf{J}_{xy_i} G(x,y) \mathbf{J}_x G(x,y)^{-1} \right) \leq 2d \|\alpha\|_2 \mathcal{K}^2 \eta^2 \mathcal{L}_H.$$

This is valid for any $\alpha \in \mathbb{R}^d$, as a consequence, we have

$$\left\| \begin{bmatrix} \operatorname{trace}\left([\mathbf{J}_x G(x, \mathbf{q}_0)]^{-1} \mathbf{J}_{xy_1} G(x, \mathbf{q}_0) \right) \\ \vdots \\ \operatorname{trace}\left([\mathbf{J}_x G(x, q_0)]^{-1} \mathbf{J}_{xy_d} G(x, \mathbf{q}_0) \right) \end{bmatrix} \right\|_2 \leq 2d\mathcal{K}^2 \eta^2 \mathcal{L}_H.$$

This concludes the proof of Lemma C.3.

C.1.6 Proof of Lemma C.4

We first show equation (C.5b) by induction. Then equation (C.5a) is a direct consequence of equation (C.5b) by summing k terms together.

Case $k = 0$: We first examine the case $k = 0$. According to the definition of F_k in equation (C.1b), we have

$$F_1(\mathbf{p}_0, \mathbf{q}_0) = \mathbf{q}_0 + \eta \mathbf{p}_0 - \frac{\eta^2}{2} \nabla f(\mathbf{q}_0).$$

Then the case $k = 0$ is verified automatically via triangle inequality,

$$\|F_1(\mathbf{p}_0, \mathbf{q}_0) - \mathbf{q}_0\|_2 \leq \eta \|\mathbf{p}_0\|_2 + \frac{\eta^2}{2} \|\nabla f(\mathbf{q}_0)\|_2 .$$

Case k to $k+1$: Assume that the statement is verified until $k \geq 0$. For $k+1$, using F_j as the shorthand for $F_j(\mathbf{p}_0, \mathbf{q}_0)$, we obtain

$$\begin{aligned} & F_{k+2} - F_{k+1} \\ =& \eta \mathbf{p}_0 - \frac{\eta^2}{2} \nabla f(\mathbf{q}_0) - \eta^2 \sum_{j=1}^{k+1} \nabla f(F_j). \end{aligned}$$

Taking the norm, we have

$$\begin{aligned} \left\| F_{k+2} - F_{k+1} \right\|_2 \leq \eta \left\| \mathbf{p}_0 \right\|_2 + \frac{(2k+3)\eta^2}{2} \left\| \nabla f(\mathbf{q}_0) \right\|_2 + \eta^2 \sum_{j=1}^{k+1} \left\| \nabla f(F_j) - \nabla f(\mathbf{q}_0) \right\|_2 \\ \overset{(i)}{\leq} \eta \left\| \mathbf{p}_0 \right\|_2 + \frac{(2k+3)\eta^2}{2} \left\| \nabla f(\mathbf{q}_0) \right\|_2 + \eta^2 \sum_{j=1}^{k+1} \sum_{l=0}^{j} \left\| \nabla f(F_{l+1}) - \nabla f(F_l) \right\|_2 \\ \overset{(ii)}{\leq} \eta \left\| \mathbf{p}_0 \right\|_2 + \frac{(2k+3)\eta^2}{2} \left\| \nabla f(\mathbf{q}_0) \right\|_2 + \eta^2 \mathcal{L} \sum_{j=1}^{k+1} \sum_{l=0}^{j} \left\| F_{l+1} - F_l \right\|_2 \\ \overset{(iii)}{\leq} \eta \left\| \mathbf{p}_0 \right\|_2 + \frac{(2k+3)\eta^2}{2} \left\| \nabla f(\mathbf{q}_0) \right\|_2 + \eta^2 \mathcal{L} \sum_{j=1}^{k+1} \sum_{l=0}^{j} \left(2\eta \left\| \mathbf{p} \right\|_2 + 2(l+1)\eta^2 \left\| \nabla f(\mathbf{q}_0) \right\|_2 \right) \\ \overset{(iv)}{\leq} 2\eta \left\| \mathbf{p}_0 \right\|_2 + (2k+2)\eta^2 \left\| \nabla f(\mathbf{q}_0) \right\|_2 . \end{aligned}$$

Inequality (i) uses triangular inequality. Inequality (ii) uses $\mathcal{L}$-smoothness. Inequality (iii) applies the hypothesis of the induction and inequalities relies on the condition $\mathcal{K}^2\eta^2 \leq \frac{1}{4\mathcal{L}}$. This completes the induction.

C.2 Proof of Corollary 5.1

Before proving Corollary 5.1, we first state a more general corollary of Theorem 5.1 that does not specify the explicit choice of step size η and leapfrog steps $\mathcal{K}$. Then we specify two choices of the initial distribution μ_0 and hyper-parameters $(\mathcal{K}, \eta)$ to obtain part (a) and part (b) of Corollary 5.1.

Corollary C.1. *Consider an $(\mathcal{L}, \mathcal{L}_H, m)$-strongly log-concave target distribution $\Pi^\star$ (cf. Assumption (5B)). Fix $s = \frac{\delta^2}{2\beta}$. Then the $\frac{1}{2}$-lazy HMC algorithm with initial distribution $\mu_\star = \mathcal{N}(x^*, \frac{1}{\mathcal{L}}\mathbb{I}_d)$, step size η and leapfrog steps $\mathcal{K}$ chosen under the condition*

$$\begin{aligned} \eta^2 \leq \frac{1}{c\mathcal{L}} \min \Bigg\{ & \frac{1}{\mathcal{K}^2 d^{\frac{1}{2}}}, \frac{1}{\mathcal{K}^2 d^{\frac{2}{3}}} \frac{\mathcal{L}}{\mathcal{L}_H^{\frac{2}{3}}}, \frac{1}{\mathcal{K} d^{\frac{1}{2}}}, \frac{1}{\mathcal{K}^{\frac{2}{3}} d^{\frac{2}{3}} \kappa^{\frac{1}{3}} \mathfrak{a}(s)^{\frac{2}{3}}}, \\ & \frac{1}{\mathcal{K} d^{\frac{1}{2}} \kappa^{\frac{1}{2}} \mathfrak{a}(s)}, \frac{1}{\mathcal{K}^{\frac{2}{3}} d} \frac{\mathcal{L}}{\mathcal{L}_H^{\frac{2}{3}}}, \frac{1}{\mathcal{K}^{\frac{4}{3}} d^{\frac{1}{2}} \kappa^{\frac{1}{2}} \mathfrak{a}(s)} \left(\frac{\mathcal{L}}{\mathcal{L}_H^{\frac{2}{3}}} \right)^{\frac{1}{2}} \Bigg\} \end{aligned} \tag{C.35}$$

satisfies the mixing time bounds

$$\tau_2^{HMC}(\delta; \mu_0) \leq c \cdot \max\left\{\log \beta, \frac{1}{\mathcal{K}^2 \eta^2 m} \log\left(\frac{d \log \kappa}{\delta}\right)\right\}.$$

Proof of part (a) in Corollary 5.1: Taking the hyper-parameters $\mathcal{K} = d^{\frac{1}{4}}$ and $\eta = \eta_{\text{warm}}$ in equation (5.7b), we verify that η satisfies the condition (C.35). Given the warmness parameter $\beta = \mathcal{O}\left(\exp\left(d^{\frac{2}{3}}\kappa\right)\right)$, we have

$$\frac{1}{\mathcal{K}^2 \eta^2 m} \geq \log(\beta).$$

Plugging in the choice of $\mathcal{K}$ and η into Corollary C.1, we obtain the desired result.

Proof of part (b) in Corollary 5.1: We notice that the initial distribution $\mu_\star = \mathcal{N}(x^\star, \frac{1}{\mathcal{L}}\mathbb{I}_d)$ is $\kappa^{d/2}$-warm (see Corollary 1 in [78]). It is sufficient to plug in the hyper-parameters $\mathcal{K} = \kappa^{\frac{3}{4}}$ and $\eta = \eta_{\text{feasible}}$ into Corollary C.1 to obtain the desired result.

Now we turn back to prove Corollary C.1. In order to prove Corollary C.1, we require the the following lemma, which relates a $(\mathcal{L}, \mathcal{L}_H, m)$-strongly-logconcave target distribution to a regular target distribution.

Lemma C.5. *An $(\mathcal{L}, \mathcal{L}_H, m)$-strongly log-concave distribution is $(\mathcal{L}, \mathcal{L}_H, s, \psi_{\frac{1}{2}}, \mathcal{M})$-regular with high mass set $\Omega = \mathcal{R}_s$, log-isoperimetric constant $\psi_{\frac{1}{2}} = m^{-\frac{1}{2}}$ and $\mathcal{M} = \mathcal{L}\left(\frac{d}{m}\right)^{\frac{1}{2}} \mathfrak{a}(s)$, where the radius is defined in equation (5.7a) and the convex measurable set $\mathcal{R}_s$ defined in equation (5.18).*

Taking Lemma C.5 as given, Corollary C.1 is a direct consequence of Theorem 5.1 by plugging the specific values of $(\Omega, \psi_{\frac{1}{2}}, \mathcal{M})$ as a function of strong convexity parameter m. We now proceed to prove Lemma C.5.

C.2.1 Proof of Lemma C.5

First, we set Ω to $\mathcal{R}_s$ as defined in equation (5.18). Lemma Lemma 3.3 implies that this ball has probability under the target distribution lower bounded as $\Pi^\star(\mathcal{R}_s) \geq 1 - s$. Second, the gradient bound is a consequence of the bounded domain. For any $x \in \mathcal{R}_s$, we have

$$\|\nabla f(x)\|_2 = \|\nabla f(x) - \nabla f(x^\star)\|_2 \leq \mathcal{L}\|x - x^\star\|_2 \leq \mathcal{L}\left(\frac{d}{m}\right)^{\frac{1}{2}} \mathfrak{a}(s). \tag{C.36}$$

Third, we make use of a logarithmic isoperimetric inequality for log-concave distribution. We note that the logarithmic isoperimetric inequality has been introduced in Kannan et al. [136] for the uniform distribution on convex body and in Lee and Vempala [158] for log-concave distribution with a diameter. We extend this inequality to strongly log-concave distribution on $\mathbb{R}^d$ following a similar road-map and provide explicit constants.

Improved logarithmic isoperimetric inequality We now state the improved logarithmic isoperimetric inequality for strongly log-concave distributions.

Lemma C.6. *Let γ denote the density of the standard Gaussian distribution $\mathcal{N}\left(0, \sigma^2 \mathbb{I}_d\right)$, and let $\Pi^\star$ be a distribution with density $\pi^\star = q \cdot \gamma$, where q is a log-concave function. Then for any partition $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ of $\mathbb{R}^d$, we have*

$$\Pi^\star(\mathcal{S}_3) \geq \frac{d(\mathcal{S}_1, \mathcal{S}_2)}{2\sigma} \min\left\{\Pi^\star(\mathcal{S}_1), \Pi^\star(\mathcal{S}_2)\right\} \log^{\frac{1}{2}}\left(1 + \frac{1}{\min\left\{\Pi^\star(\mathcal{S}_1), \Pi^\star(\mathcal{S}_2)\right\}}\right). \tag{C.37}$$

See Appendix C.2.2 for the proof.

Taking Lemma C.6 as given for the moment, we turn to prove the logarithmic isoperimetric inequality for the Ω-restricted distribution $\Pi^\star_\Omega$ with density

$$\pi^\star_\Omega(x) = \frac{\pi^\star(x)\mathbf{1}_\Omega(x)}{\Pi^\star(\Omega)}.$$

Since f is m-strongly convex, the function $x \rightarrow f(x) - \frac{m}{2}\left\|x - x^\star\right\|_2^2$ is convex. Noting that the class of log-concave function is closed under multiplication and that the indicator function $\mathbf{1}_\Omega$ is log-concave, we conclude that the restricted density $\pi^\star_\Omega$ can be expressed as a product of a log-concave density and the density of the Gaussian distribution $\mathcal{N}(x^\star, \frac{1}{m}\mathbb{I}_d)$. Applying Lemma C.6 with $\sigma = \left(\frac{1}{m}\right)^{\frac{1}{2}}$, we obtain the desired logarithmic isoperimetric inequality with $\psi_{\frac{1}{2}} = \left(\frac{1}{m}\right)^{\frac{1}{2}}$, which concludes the proof of Lemma C.5.

C.2.2 Proof of Lemma C.6

The main tool for proving general isoperimetric inequalities is the localization lemma introduced by Lovász and Simonovits [167]. Similar result for the infinitesimal version of equation (C.37) have appeared as Theorem 1.1 in the paper [150] and Theorem 30 in the paper [158]. Intuitively, the localization lemma reduces a high-dimensional isoperimetric inequality to a one-dimensional inequality which is much easier to verify directly. In a few key steps, the proof follows a similar road map as the proof of logarithmic Cheeger inequality [136].

We first state an additional lemma that comes in handy for the proof.

Lemma C.7. *Let γ be the density of the one-dimensional Gaussian distribution $\mathcal{N}\left(\nu, \sigma^2\right)$ with mean ν and variance σ^2. Let ϱ be a one-dimensional distribution with density given by $\varrho = q \cdot \gamma$, where q is a log-concave function supported on $[0, 1]$. Let J_1, J_2, J_3 partition $[0, 1]$, then*

$$\varrho(J_3) \geq \frac{d(J_1, J_2)}{2\sigma} \min\left\{\varrho(J_1), \varrho(J_2)\right\} \log^{\frac{1}{2}}\left(1 + \frac{1}{\min\left\{\varrho(J_1), \varrho(J_2)\right\}}\right). \tag{C.38}$$

See Appendix C.2.3 for the proof.

We now turn to proving Lemma C.6 via contradiction: We assume that the claim (C.37) is not true for some partition, and then using well known localization techniques, we construct a one-dimensional distribution that violates Lemma C.7 resulting in a contradiction.

Suppose that there exists a partition $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ of $\mathbb{R}^d$, such that

$$\Pi^\star(\mathcal{S}_3) < \frac{d(\mathcal{S}_1, \mathcal{S}_2)}{2\sigma} \min\left\{\Pi^\star(\mathcal{S}_1), \Pi^\star(\mathcal{S}_2)\right\} \log^{\frac{1}{2}}\left(1 + \frac{1}{\min\left\{\Pi^\star(\mathcal{S}_1), \Pi^\star(\mathcal{S}_2)\right\}}\right). \tag{C.39}$$

Let $\nu > 0$ denote a sufficiently small number (to be specified exactly later), such that $\nu < \min\left\{\Pi^\star(\mathcal{S}_1), \Pi^\star(\mathcal{S}_2)\right\}$.

We now explain the construction of the one-dimensional density that is crucial for the rest of the argument. We define two functions $g : \mathcal{X} \to \mathbb{R}$ and $\mathcal{H} : \mathcal{X} \to \mathbb{R}$ as follows

$$g(x) = \frac{\pi^\star(x) \cdot \mathbf{1}_{\mathcal{S}_1}(x)}{\Pi^\star(\mathcal{S}_1) - \nu} - \pi^\star(x) \quad \text{and} \quad \mathcal{H}(x) = \frac{\pi^\star(x) \cdot \mathbf{1}_{\mathcal{S}_2}(x)}{\Pi^\star(\mathcal{S}_2) - \nu} - \pi^\star(x).$$

Clearly, we have

$$\int_{\mathcal{X}} g(x)dx > 0 \quad \text{and} \quad \int_{\mathcal{X}} \mathcal{H}(x)dx > 0.$$

By the localization lemma (Lemma 2.5 in the paper [167]; see the corrected form stated as Lemma 2.1 in the paper [137]), there exist two points $a \in \mathbb{R}^d, b \in \mathbb{R}^d$ and a linear function $l : [0, 1] \to \mathbb{R}_+$, such that

$$\int_0^1 l(t)^{d-1} g\left((1-t)a + tb\right) dt > 0 \quad \text{and} \quad \int_0^1 l(t)^{d-1} h\left((1-t)a + tb\right) dt > 0. \tag{C.40}$$

Define the one-dimensional density $\varrho : [0, 1] \to \mathbb{R}^+$ and the sets $J_i, i \in \{1, 2, 3\}$ as follows:

$$\varrho(t) = \frac{l(t)^{d-1}\pi^\star\left((1-t)a + tb\right)}{\int_0^1 l(u)^{d-1}\pi^\star\left((1-u)a + ub\right) du}, \quad \text{and} \tag{C.41}$$

$$J_i = \left\{t \in [0, 1] \mid (1-t)a + tb \in S_i\right\} \quad \text{for } i \in \{1, 2, 3\}. \tag{C.42}$$

We now show how the hypothesis (C.39) leads to a contradiction for the density ϱ. Plugging in the definiton of g and $\mathcal{H}$ into equation (C.40), we find that

$$\varrho(J_1) > \Pi^\star(\mathcal{S}_1) - \nu \quad \text{and} \quad \varrho(J_2) > \Pi^\star(\mathcal{S}_2) - \nu.$$

Since J_1, J_2, J_3 partition $[0, 1]$, it follows that

$$\varrho(J_3) < \Pi^\star(\mathcal{S}_3) + 2\nu.$$

Since the function $x \mapsto x \log^{\frac{1}{2}}(1 + 1/x)$ is monotonically increasing on $[0, 1]$, we have

$$\begin{aligned}
&\frac{d(\mathcal{S}_1, \mathcal{S}_2)}{2\sigma} \min\{\varrho(J_1), \varrho(J_2)\} \log^{\frac{1}{2}}\left(1 + \frac{1}{\min\{\varrho(J_1), \varrho(J_2)\}}\right) - \varrho(J_3) \\
&\quad \geq \frac{d(\mathcal{S}_1, \mathcal{S}_2)}{2\sigma} \min\{(\varrho(\mathcal{S}_1) - \nu), (\varrho(\mathcal{S}_2) - \nu)\} \cdot \\
&\qquad\qquad \log^{\frac{1}{2}}\left(1 + \frac{1}{\min\{(\varrho(\mathcal{S}_1) - \nu), (\varrho(\mathcal{S}_2) - \nu)\}}\right) - (\varrho(\mathcal{S}_3) + 2\nu)
\end{aligned}$$

The hypothesis (C.39) of the contradiction implies that we can find ν sufficiently small such that the RHS in the inequality above will be strictly positive. Consequently, we obtain

$$\frac{d(\mathcal{S}_1, \mathcal{S}_2)}{2\sigma} \min\{\varrho(J_1), \varrho(J_2)\} \log^{\frac{1}{2}}\left(1 + \frac{1}{\min\{\varrho(J_1), \varrho(J_2)\}}\right) > \varrho(J_3). \tag{C.43}$$

Additionally, for $t_1 \in J_1, t_2 \in J_2$, we have $(1 - t_1)a + t_1 b \in \mathcal{S}_1$ and $(1 - t_2)a + t_2 b \in \mathcal{S}_2$. As a result, we have

$$|t_1 - t_2| = \frac{1}{\|b - a\|_2} \|[(1 - t_1)a + t_1 b] - [(1 - t_2)a + t_2 b]\|_2 \geq \frac{1}{\|b - a\|_2} d(\mathcal{S}_1, \mathcal{S}_2),$$

which implies that

$$d(J_1, J_2) \geq \frac{1}{\|b - a\|_2} d(\mathcal{S}_1, \mathcal{S}_2). \tag{C.44}$$

Combining equations (C.43) and (C.44), we obtain that

$$\frac{\|b - a\|_2 \cdot d(J_1, J_2)}{2\sigma} \min\{\varrho(J_1), \varrho(J_2)\} \log^{\frac{1}{2}}\left(1 + \frac{1}{\min\{\varrho(J_1), \varrho(J_2)\}}\right) > \varrho(J_3), \tag{C.45}$$

which contradicts Lemma C.7. Indeed, this contradiction is immediate once we note that the new density ϱ can also be written as a product of log-concave density and a Gaussian density with variance $\frac{\sigma^2}{\|b-a\|_2^2}$.

C.2.3 Proof of Lemma C.7

We split the proof into three cases. Each one is more general than the previous one. First, we consider the case when q is a constant function on $[0, 1]$ and the sets J_1, J_2, J_3 are all intervals. In the second case, we consider a general log-concave q supported on $[0, 1]$ while we still assume that the sets J_1, J_2, J_3 are all intervals. Finally, in the most general case, we consider a general log-concave q supported on $[0, 1]$ and J_1, J_2, J_3 consist of an arbitrary partition of $[0, 1]$. The proof idea follows roughly that of Theorem 4.6 in Kannan et al. [136].

Our proof makes use of the Gaussian isoperimetric inequality which we now state (see e.g., equation (1.2) in [22]): Let Γ denote the standard univariate Gaussian distribution and let ψ_Γ and Ψ_Γ^{-1} denote its density and inverse cumulative distribution function respectively. Given a measurable set $\mathcal{S} \subset \mathbb{R}$, define its Γ-perimeter $\Gamma^+(\mathcal{S})$ as

$$\Gamma^+(\mathcal{S}) = \liminf_{h\to 0^+} \frac{\Gamma(\mathcal{S}+h) - \Gamma(\mathcal{S})}{h},$$

where $\mathcal{S}+h = \{t \in \mathbb{R} \mid \exists a \in \mathcal{S}, |t-a| < h\}$ denotes an h-neighborhood of $\mathcal{S}$. Then, we have

$$\Gamma^+(\mathcal{S}) \geq \psi_\Gamma(\Psi_\Gamma^{-1}(\Gamma(\mathcal{S}))), \tag{C.46}$$

where $\Gamma(\mathcal{S})$ denotes the Gaussian measure of the set $\mathcal{S}$. Furthermore, standard Gaussian tail bounds[1] estimate imply that

$$\psi_\Gamma(\Psi_\Gamma^{-1}(t)) \geq \frac{1}{2} t \log^{\frac{1}{2}}\left(1 + \frac{1}{t}\right), \qquad \text{for } t \in (0, \frac{1}{2}]. \tag{C.47}$$

Case 1: First, we consider the case when the function q is constant on $[0,1]$ and all of the sets J_1, J_2, J_3 are intervals. Without loss of generality, we can shift and scale the density function by changing the domain, and assume that the density ϱ is of the form $\varrho(t) \propto e^{-\frac{t^2}{2}}\mathbf{1}_{[a,d]}$. Additionally, we can assume that J_1, J_2, J_3 are of the form

$$J_1 = [a,b), \quad J_3 = [b,c], \quad \text{and} \quad J_2 = (c,d], \tag{C.48}$$

because the case when J_3 is not in the middle is a trivial case.

Applying the inequalities (C.46) and (C.47) with $A = J_2 = (c,d]$, we obtain that

$$\psi_\Gamma(c) = \Gamma^+(J_2) \geq \psi_\Gamma(\Psi_\Gamma^{-1}(\Gamma(J_2))) \geq \frac{\Gamma(J_2)}{2}\log^{\frac{1}{2}}\left(1 + \frac{1}{\Gamma(J_2)}\right), \tag{C.49}$$

Note that $\varrho(t) = \frac{\psi_\Gamma(t)}{\Psi_\Gamma(d)-\Psi_\Gamma(a)}\mathbf{1}_{[a,d]}(t)$ and $\varrho(J_2) = \frac{\Gamma(J_2)}{\Psi_\Gamma(d)-\Psi_\Gamma(a)}$. We have

$$\begin{aligned}
\varrho(J_3) = \int_b^c \varrho(t)dt \geq (c-b)\cdot\varrho(c) &= (c-b)\frac{\psi_\Gamma(c)}{\Psi_\Gamma(d)-\Psi_\Gamma(a)} \\
&\overset{(i)}{\geq} \frac{(c-b)}{2}\frac{\Gamma(J_2)}{\Psi_\Gamma(d)-\Psi_\Gamma(a)}\log^{\frac{1}{2}}\left(1+\frac{1}{\Gamma(J_2)}\right) \\
&\overset{(ii)}{\geq} \frac{c-b}{2}\varrho(J_2)\log^{\frac{1}{2}}\left(1+\frac{\Psi_\Gamma(d)-\Psi_\Gamma(a)}{\Gamma(J_2)}\right) \\
&\overset{(iii)}{=} \frac{c-b}{2}\varrho(J_2)\log^{\frac{1}{2}}\left(1+\frac{1}{\varrho(J_2)}\right) \\
&\overset{(iv)}{\geq} \frac{c-b}{2}\min\{\varrho(J_1),\varrho(J_2)\}\log^{\frac{1}{2}}\left(1+\frac{1}{\min\{\varrho(J_1),\varrho(J_2)\}}\right),
\end{aligned}$$

[1] E.g., see the discussion before equation 1 in the paper [13]. The constant 1/2 was estimated by plotting the continuous function on the left hand side via Mathematica.

where step (i) follows from the bound (C.49) and step (ii) follows from the relationship between ϱ and Γ and the facts that log is an increasing function and that $\Psi_\Gamma(d) - \Psi_\Gamma(a) \leq 1$. Step (iii) follows from the definition of ϱ and finally step (iv) follows from the increasing nature of the map $t \mapsto t\log^{1/2}\left(1 + \frac{1}{t}\right)$. This concludes the argument for Case 1.

Case 2: We now consider the case when q is a general log-concave function on $[0, 1]$ and J_1, J_2, J_3 are all intervals. Again we can assume that J_1, J_2, J_3 are of the form (C.48), i.e., they are given by $J_1 = [a, b)$, $J_3 = [b, c]$, and $J_2 = (c, d]$.

We consider a function $h(t) = \alpha e^{\beta t - \frac{t^2}{2\sigma^2}}$ such that $h(b) = q(b)$ and $h(c) = q(c)$.[2] Define $Q(t_1, t_2) = \int_{t_1}^{t_2} q(t)dt$ and $H(t_1, t_2) = \int_{t_1}^{t_2} h(t)dt$. Then since q has an extra log-concave component compared to h, we have

$$H(a, b) \geq Q(a, b), \quad H(c, d) \geq Q(c, d), \text{ but } H(b, c) \leq Q(b, c). \tag{C.50}$$

Using the individual bounds in equation (C.50), we have

$$\frac{H(a, b)}{H(b, c)} + \frac{H(c, d)}{H(b, c)} \geq \frac{Q(a, b)}{Q(b, c)} + \frac{Q(c, d)}{Q(b, c)}.$$

From the equation above and the fact that $H(a, b) + H(b, c) + H(c, d) = H(a, d)$, we obtain

$$\frac{H(b, c)}{H(a, d)} \leq \frac{Q(b, c)}{Q(a, d)}. \tag{C.51}$$

To prove the inequality in Case 2, here are two subcases depending on whether $H(a, d) \geq Q(a, d)$ or $H(a, d) < Q(a, d)$.

- If $H(a, d) \geq Q(a, d)$, then

$$\begin{aligned}
\frac{Q(b, c)}{Q(a, d)} &\overset{(i)}{\geq} \frac{H(b, c)}{Q(a, d)} \\
&\overset{(ii)}{\geq} \frac{c - b}{2} \cdot \frac{H(a, d)}{Q(a, d)} \cdot \frac{\min(H(a, b), H(c, d))}{H(a, d)} \cdot \log^{\frac{1}{2}}\left(1 + \frac{H(a, d)}{\min(H(a, b), H(c, d))}\right) \\
&\overset{(iii)}{\geq} \frac{c - b}{2} \cdot \frac{H(a, d)}{Q(a, d)} \cdot \frac{\min(Q(a, b), Q(c, d))}{H(a, d)} \cdot \log^{\frac{1}{2}}\left(1 + \frac{H(a, d)}{\min(Q(a, b), Q(c, d))}\right) \\
&\overset{(iv)}{\geq} \frac{c - b}{2} \cdot \frac{\min(Q(a, b), Q(c, d))}{Q(a, d)} \cdot \log^{\frac{1}{2}}\left(1 + \frac{Q(a, d)}{\min(Q(a, b), Q(c, d))}\right).
\end{aligned}$$

  Inequality (i) follows from equation (C.50); inequality (ii) follows from equation Case 1 because H is covered by Case 1; inequality (iii) uses the fact that the function $t \mapsto t\log^{\frac{1}{2}}\left(1 + \frac{1}{t}\right)$ is increasing; inequality (iv) follows from the assumption in this subcase $H(a, d) \geq Q(a, d)$.

[2] This idea of introducing exponential function appeared in Corollary 6.2 of Kannan et al. [136].

- Otherwise $H(a,d) < Q(a,d)$, then we have from equation (C.50)

$$\frac{H(a,b)}{H(a,d)} \geq \frac{Q(a,b)}{Q(a,d)}, \quad \frac{H(c,d)}{Q(a,d)} \geq \frac{Q(c,d)}{Q(a,d)}.$$

$$\begin{aligned}\frac{Q(b,c)}{Q(a,d)} &\overset{(i)}{\geq} \frac{H(b,c)}{H(a,d)}\\ &\overset{(ii)}{\geq} \frac{c-b}{2} \cdot \frac{\min(H(a,b),H(c,d))}{H(a,d)} \cdot \log^{\frac{1}{2}}\left(1 + \frac{H(a,d)}{\min(H(a,b),H(c,d))}\right)\\ &\overset{(iii)}{\geq} \frac{c-b}{2} \cdot \frac{\min(Q(a,b),Q(c,d))}{Q(a,d)} \cdot \log^{\frac{1}{2}}\left(1 + \frac{Q(a,d)}{\min(Q(a,b),Q(c,d))}\right).\end{aligned}$$

Inequality (i) follows from equation (C.51); inequality (ii) follows from equation Case 1; inequality (iii) uses the fact that the function $t \mapsto t \log^{\frac{1}{2}}\left(1 + \frac{1}{t}\right)$ is increasing.

In both subcases above, we conclude Case 2 using the results established in Case 1.

Case 3: Finally, we deal with the general case where J_1, J_2, J_3 each can be union of intervals and q is a general log-concave function on $[0,1]$. We show that this case can be reduced to the case of three intervals, namely, the previous case.

Let $\{(b_i, c_i)\}_{i \in \mathcal{I}}$ be all non-empty maximal intervals contained in J_3. Here the intervals can be either closed, open or half. That is, $(\cdot,\cdot)$ can be $[\cdot,\cdot]$, $]\cdot,\cdot[$, $[\cdot,\cdot[$ or $]\cdot,\cdot]$. For an interval (b_i, c_i), we define its left surround $LS((b_i,c_i))$ as

$$LS((b_i,c_i)) = \begin{cases} 2, & \text{if } \exists x_2 \in J_2, (x_2 \leq b_i) \text{ and } (\nexists x_1 \in J_1, x_2 < x_1 \leq b_i) \\ 1, & \text{if } \exists x_1 \in J_1, (x_1 \leq b_i) \text{ and } (\nexists x_2 \in J_2, x_1 < x_2 \leq b_i) \\ 0, & \text{otherwise}. \end{cases}$$

Similarly, we define $RS((b_i,c_i))$ as

$$RS((b_i,c_i)) = \begin{cases} 2, & \text{if } \exists x_2 \in J_2, (x_2 \geq c_i) \text{ and } (\nexists x_1 \in J_1, x_2 > x_1 \geq c_i) \\ 1, & \text{if } \exists x_1 \in J_1, (x_1 \geq c_i) \text{ and } (\nexists x_2 \in J_2, x_1 > x_2 \geq c_i) \\ 0, & \text{otherwise}. \end{cases}$$

We distinguish two types of intervals. Denote $G_2 \subset \mathcal{I}$ the set containing the indices of all intervals that are surrounded by either 1 or 2 but different.

$$G_2 \coloneqq \{i \in \mathcal{I} \mid (LS((b_i,c_i)), RS((b_i,c_i))) = (1,2) \text{ or } (2,1)\}.$$

Denote $G_1 \coloneqq \mathcal{I} \setminus G_2$ to be its complement. By the result settled in case 2, for $i \in G_2$, we have

$$\varrho([b_i,c_i]) \geq \frac{d(J_1,J_2)}{2\sigma} \varrho(I_i) \log^{\frac{1}{2}}\left(1 + \frac{1}{\varrho(I_i)}\right)$$

where I_i is either $[a, b_i]$ or $[c_i, d]$. Summing over all $i \in G_2$, we have

$$\begin{aligned} \varrho(J_3) \geq \sum_{i \in G_2} \varrho([b_i, c_i]) &\geq \frac{d(J_1, J_2)}{2\sigma} \sum_{i \in G_2} \varrho(I_i) \log^{\frac{1}{2}} \left(1 + \frac{1}{\varrho(I_i)}\right) \\ &\geq \frac{d(J_1, J_2)}{2\sigma} \varrho(\cup_{i \in G_2} I_i) \log^{\frac{1}{2}} \left(1 + \frac{1}{\varrho(\cup_{i \in G_2} I_i)}\right). \end{aligned} \tag{C.52}$$

The last inequality follows from the sub-additivity of the map: $x \mapsto x \log^{\frac{1}{2}}(1+x)$, i.e., for $x > 0$ and $y > 0$, we have

$$x \log^{\frac{1}{2}} \left(1 + \frac{1}{x}\right) + y \log^{\frac{1}{2}} \left(1 + \frac{1}{y}\right) \geq (x+y) \log^{\frac{1}{2}} \left(1 + \frac{1}{x+y}\right).$$

Indeed the sub-additivity follows immediately from the following observation:

$$\begin{aligned} &x \log^{\frac{1}{2}} \left(1 + \frac{1}{x}\right) + y \log^{\frac{1}{2}} \left(1 + \frac{1}{y}\right) - (x+y) \log^{\frac{1}{2}} \left(1 + \frac{1}{x+y}\right) \\ &\quad = x \left[\log^{\frac{1}{2}} \left(1 + \frac{1}{x}\right) - \log^{\frac{1}{2}} \left(1 + \frac{1}{x+y}\right)\right] + y \left[\log^{\frac{1}{2}} \left(1 + \frac{1}{y}\right) - \log^{\frac{1}{2}} \left(1 + \frac{1}{x+y}\right)\right] \\ &\quad \geq 0. \end{aligned}$$

Finally, we remark that either J_1 or J_2 is a subset of $\cup_{i \in G_2} I_i$. If not, there exists $u \in J_1 \setminus \cup_{i \in G_2} I_i$ and $v \in J_2 \setminus \cup_{i \in G_2} I_i$, such that u and v are separated by some inverval $(b_{i^*}, c_{i^*}) \subset J_3$ with $i^* \in G_2$. This is contradictory with the fact that either u or v must be included in I_{i^*}. Given equation (C.52), we use the fact that the function $x \mapsto x \log^{\frac{1}{2}} \left(1 + \frac{1}{x}\right)$ is monotonically increasing:

$$\varrho(J_3) \geq \frac{d(J_1, J_2)}{2\sigma} \min\left\{\varrho(J_1), \varrho(J_2)\right\} \log^{\frac{1}{2}} \left(1 + \frac{1}{\min\left\{\varrho(J_1), \varrho(J_2)\right\}}\right)$$

to conclude the proof.

C.3 Optimal choice for HMC hyper-parameters

In this section, we provide a detailed discussion about the optimal leapfrog steps choice for Metropolized HMC with strongly log-concave target distribution (Corollary 5.1). We also discuss a few improved convergence rates for Metropolized HMC under additional assumptions on the target distribution. Finally, we compare our results for Metropolized HMC with other versions of HMC namely unadjusted HMC and ODE-solved based HMC in Subsection C.3.2.

C.3.1 Optimal choices for Corollary C.1

Corollary C.1 provides an implicit condition that the step size η and leapfrog steps $\mathcal{K}$ should satisfy and provides a generic mixing time upper bound that depends on the choices made. We claim that the optimal choices of η and $\mathcal{K}$ according to Table C.1 lead to the following upper bound on number of gradient evaluations required by HMC to mix to δ-tolerance:

$$\mathcal{K} \cdot \tau_{\mathrm{TV}}^{\mathrm{HMC}}(\delta; \mu_0) \leq \mathcal{O}\left(\max\left\{d\kappa^{\frac{3}{4}}, d^{\frac{11}{12}}\kappa, d^{\frac{3}{4}}\kappa^{\frac{5}{4}}, d^{\frac{1}{2}}\kappa^{\frac{3}{2}}\right\} \cdot \log\frac{1}{\delta}\right). \tag{C.53}$$

This (upper) bound shows that HMC always requires fewer gradient evaluations when compared to MALA for mixing in total variation distance. However, such a bound requires a delicate choice of the leap frog steps $\mathcal{K}$ and η depending on the condition number κ and the dimension d, which might be difficult to implement in practice. We summarize these optimal choices in Table C.1.

Case	$\mathcal{K}$	η^2
$\kappa \in (0, d^{\frac{1}{3}})$	$\kappa^{\frac{3}{4}}$	$\frac{1}{c\mathcal{L}} \cdot d^{-1}\kappa^{-\frac{1}{2}}$
$\kappa \in [d^{\frac{1}{3}}, d^{\frac{2}{3}}]$	$d^{\frac{1}{4}}$	$\frac{1}{c\mathcal{L}} \cdot d^{-\frac{7}{6}}$
$\kappa \in (d^{\frac{2}{3}}, d]$	$d^{\frac{3}{4}}\kappa^{-\frac{3}{4}}$	$\frac{1}{c\mathcal{L}} \cdot d^{-\frac{3}{2}}\kappa^{\frac{1}{2}}$
$\kappa \in (d, \infty)$	1	$\frac{1}{c\mathcal{L}} \cdot d^{-\frac{1}{2}}\kappa^{-\frac{1}{2}}$

Table C.1: Optimal choices of leapfrog steps $\mathcal{K}$ and the step size η for the HMC algorithm for an $(m, \mathcal{L}, \mathcal{L}_H)$-regular target distribution such that $\mathcal{L}_H = O(\mathcal{L}^{\frac{3}{2}})$ used for the mixing time bounds in Corollary C.1. Here c denotes a universal constant.

Proof of claim (C.53): Recall that under the condition (C.35) (restated for reader's convenience)

$$\eta^2 \leq \frac{1}{c\mathcal{L}} \min\left\{\frac{1}{\mathcal{K}^2 d^{\frac{1}{2}}}, \frac{1}{\mathcal{K}^2 d^{\frac{2}{3}}}\frac{\mathcal{L}}{\mathcal{L}_H^{\frac{2}{3}}}, \frac{1}{\mathcal{K}d^{\frac{1}{2}}}, \frac{1}{\mathcal{K}^{\frac{2}{3}} d^{\frac{2}{3}}\kappa^{\frac{1}{3}}\mathfrak{a}(s)^{\frac{2}{3}}}, \frac{1}{\mathcal{K}d^{\frac{1}{2}}\kappa^{\frac{1}{2}}\mathfrak{a}(s)}, \frac{1}{\mathcal{K}^{\frac{2}{3}} d}\frac{\mathcal{L}}{\mathcal{L}_H^{\frac{2}{3}}}, \frac{1}{\mathcal{K}^{\frac{4}{3}} d^{\frac{1}{2}}\kappa^{\frac{1}{2}}\mathfrak{a}(s)}\left(\frac{\mathcal{L}}{\mathcal{L}_H^{\frac{2}{3}}}\right)^{\frac{1}{2}}\right\},$$

Corollary 5.1 guarantees that the HMC mixing time for the $\kappa^{\frac{d}{2}}$-warm initialization $\mu_\star = \mathcal{N}(x^\star, \mathcal{L}^{-1}\mathbb{I}_d)$, is

$$\tau_2^{\mathrm{HMC}}(\delta; \mu_0) = \mathcal{O}\left(d + \frac{\kappa}{\mathcal{K}^2\eta^2\mathcal{L}}\right),$$

where we have ignored logarithmic factors. In order to compare with MALA and other sampling methods, our goal is to optimize the number of gradient evaluations $\mathcal{G}_{\text{eval}}$ taken by HMC to mix:

$$\mathcal{G}_{\text{eval}} := \mathcal{K} \cdot \tau_{\text{TV}}^{\text{HMC}}(\delta; \mu_0) = \mathcal{O}\left(\mathcal{K}d + \frac{\kappa}{\mathcal{K}\eta^2\mathcal{L}}\right). \tag{C.54}$$

Plugging in the condition on η stated above, we obtain

$$\mathcal{G}_{\text{eval}} \leq \max\left\{\underbrace{\mathcal{K}d}_{=:T_1}, \quad \underbrace{\mathcal{K}\max\left(d^{\frac{1}{2}}\kappa, d^{\frac{2}{3}}\kappa\vartheta\right)}_{=:T_2}, \quad \underbrace{d^{\frac{1}{2}}\kappa^{\frac{3}{2}}}_{=:T_3}, \quad \underbrace{\mathcal{K}^{-\frac{1}{3}}d^{\frac{2}{3}}\kappa^{\frac{4}{3}}}_{=:T_4}, \quad \underbrace{\mathcal{K}^{-\frac{1}{3}}d\kappa\cdot\vartheta}_{=:T_5}, \quad \underbrace{\mathcal{K}^{\frac{1}{3}}d^{\frac{1}{2}}\kappa^{\frac{3}{2}}\cdot\vartheta^{\frac{1}{2}}}_{=:T_6}\right\} \tag{C.55}$$

where $\vartheta = \mathcal{L}_H^{\frac{2}{3}}/\mathcal{L}$. Note that this bound depends only on the relation between d, κ and the choice of $\mathcal{K}$. We now summarize the source of all of these terms in our proofs:

- T_1: This term is attributed to the warmness of the initial distribution. The distribution $\mu_\star$ is $\mathcal{O}(\kappa^d)$-warm. This term could be improved if we have a warmer initial distribution.
- T_2: This term appears in the proposal overlap bound from equation (5.16a) of Lemma 5.1 and more precisely, it comes from equation (C.6).
- T_3, T_4, T_5 and T_6: These terms pop-out from the accept-reject bound from equation (5.16b) of Lemma 5.1. More precisely, T_3 and T_4 are a consequence of the first three terms in equation (C.22), and T_5 and T_6 arise the last two terms in equation (C.22).

In Table C.2, we summarize how these six terms can be traded-off to derive the optimal parameter choices for Corollary C.1. The effective bound on $\mathcal{G}_{\text{eval}}$-the number of gradient evaluations required by HMC to mix, is given by the largest of the six terms.

κ versus d	optimal choice $\mathcal{K}$	T_1	T_2	T_3	T_4	T_5	T_6
		$\mathcal{K}d$	$\mathcal{K}d^{\frac{2}{3}}\kappa$	$d^{\frac{1}{2}}\kappa^{\frac{3}{2}}$	$\mathcal{K}^{-\frac{1}{3}}d^{\frac{2}{3}}\kappa^{\frac{4}{3}}$	$\mathcal{K}^{-\frac{1}{3}}d\kappa$	$\mathcal{K}^{\frac{1}{3}}d^{\frac{1}{2}}\kappa^{\frac{3}{2}}$
$\kappa \in [1, d^{\frac{1}{3}})$	$\mathcal{K} = \kappa^{\frac{3}{4}}$	$\boldsymbol{d\kappa^{\frac{3}{4}}}$	$d^{\frac{2}{3}}\kappa^{\frac{7}{4}}$	$d^{\frac{1}{2}}\kappa^{\frac{3}{2}}$	$d^{\frac{2}{3}}\kappa^{\frac{13}{12}}$	$\boldsymbol{d\kappa^{\frac{3}{4}}}$	$d^{\frac{1}{2}}\kappa^{\frac{7}{4}}$
$\kappa \in [d^{\frac{1}{3}}, d^{\frac{2}{3}}]$	$\mathcal{K} = d^{\frac{1}{4}}$	$d^{\frac{5}{4}}$	$\boldsymbol{d^{\frac{11}{12}}\kappa}$	$d^{\frac{1}{2}}\kappa^{\frac{3}{2}}$	$d^{\frac{7}{12}}\kappa^{\frac{4}{3}}$	$\boldsymbol{d^{\frac{11}{12}}\kappa}$	$d^{\frac{7}{12}}\kappa^{\frac{3}{2}}$
$\kappa \in (d^{\frac{2}{3}}, d]$	$\mathcal{K} = d^{\frac{3}{4}}\kappa^{-\frac{3}{4}}$	$d^{\frac{7}{4}}\kappa^{-\frac{3}{4}}$	$d^{\frac{19}{12}}\kappa^{\frac{1}{4}}$	$d^{\frac{1}{2}}\kappa^{\frac{3}{2}}$	$d^{\frac{5}{12}}\kappa^{\frac{19}{12}}$	$\boldsymbol{d^{\frac{3}{4}}\kappa^{\frac{5}{4}}}$	$\boldsymbol{d^{\frac{3}{4}}\kappa^{\frac{5}{4}}}$
$\kappa \in (d, \infty]$	$\mathcal{K} = 1$	d	$d^{\frac{2}{3}}\kappa$	$\boldsymbol{d^{\frac{1}{2}}\kappa^{\frac{3}{2}}}$	$d^{\frac{2}{3}}\kappa^{\frac{4}{3}}$	$d\kappa$	$\boldsymbol{d^{\frac{1}{2}}\kappa^{\frac{3}{2}}}$

Table C.2: Trade-off between the six terms $T_i, i = 1, \ldots 6$, from the bound (C.55) under the assumption $\vartheta = \mathcal{L}_H^{2/3}/\mathcal{L} \leq 1$. In the second column, we provide the optimal choice of $\mathcal{K}$ for the condition on κ stated in first column such that the maximum of the T_i's is smallest. For each row the dominant (maximum) term, and equivalently the effective bound on $\mathcal{G}_{\text{eval}}$ is displayed in bold (red).

C.3.1.1 Faster mixing time bounds

We now derive several mixing time bounds under additional assumptions: (a) when a warm start is available, and (b) the Hessian-Lipschitz constant is small.

Faster mixing time with warm start: When a better initialization with warmness $\beta \leq \mathcal{O}(e^{d^{\frac{2}{3}}\kappa})$ is available, and suppose that κ is much smaller than d. In such a case, the optimal choice turns out to be $\mathcal{K} = d^{\frac{1}{4}}$ (instead of $\kappa^{\frac{3}{4}}$) which implies a bound of $O\left(d^{\frac{11}{12}}\kappa \log\left(\frac{1}{\delta}\right)\right)$ on $\mathcal{G}_{\text{eval}}$ (this bound was also stated in Table 5.1).

Faster mixing time with small $\mathcal{L}_H$: Suppose in addition to warmness being not too large, $\beta \leq \mathcal{O}(e^{d^{\frac{2}{3}}\kappa})$, the Hessian-Lipschitz constant $\mathcal{L}_H$ is small enough $\mathcal{L}_H^{\frac{2}{3}} \ll \mathcal{L}$. In such a scenario, the terms T_5 and T_6 become negligible because of small $\mathcal{L}_H$ and T_1 is negligible because of small β. The terms T_3 and T_4 remain unchanged, and the term T_2 changes slightly. More precisely, for the case $\mathcal{L}_H^{\frac{2}{3}} \leq \frac{\mathcal{L}}{d^{\frac{1}{2}}\kappa^{\frac{1}{2}}}$ we obtain a slightly modified trade-off for the terms in the (C.55) for $\mathcal{G}_{\text{eval}}$ (summarized in Table C.3). If κ is small too, then we obtain a mixing time bound of order $d^{\frac{5}{8}}$. Via this artificially constructed example, we wanted to demonstrate two things. First, faster convergence rates are possible to derive under additional assumptions directly from our results. Suitable adaptation of our proof techniques might provide a faster rate of mixing for Metropolized HMC under additional assumptions like infinity semi-norm regularity condition made in other works [176] (but we leave a detailed derivation for future work). Second, it also demonstrates the looseness of our proof techniques since we were unable to recover an $\mathcal{O}(1)$ mixing time bound for sampling from a Gaussian target.

κ versus d	$\mathcal{K}$ optimal choice	T_1	T_2	T_3	T_4	T_5	T_6
		-	$\mathcal{K} d^{\frac{1}{2}} \kappa$	$d^{\frac{1}{2}} \kappa^{\frac{3}{2}}$	$\mathcal{K}^{-\frac{1}{3}} d^{\frac{2}{3}} \kappa^{\frac{4}{3}}$	-	-
$\kappa \in (0, d^{\frac{1}{2}})$	$\mathcal{K} = d^{\frac{1}{8}} \kappa^{\frac{1}{4}}$	-	$\mathbf{d^{\frac{5}{8}} \kappa^{\frac{5}{4}}}$	$d^{\frac{1}{2}} \kappa^{\frac{3}{2}}$	$\mathbf{d^{\frac{5}{8}} \kappa^{\frac{5}{4}}}$	-	-

Table C.3: Six terms in the HMC number of gradient evaluations bound under small hessian-Lipschitz constant and very warm start. The dominant term is highlighted in red.

Linearly transformed HMC (effect of mass function): In practice, it is often beneficial to apply linear transformations in HMC (cf. Section 4 [190]). At a high level, such a transformation can improve the conditioning of the problem and help HMC mix faster. For the target distribution $\Pi^\star$ with density proportional to e^{-f}, we can define a new distribution Π_h with density e^{-h} (up to normalization) such that $h(x) = f(M^{-\frac{1}{2}}x)$ where $M \in \mathbb{R}^{d \times d}$ is an invertible matrix. Then for a random sample $\widetilde{\mathbf{q}} \sim \Pi_h$, the distribution of $M^{\frac{1}{2}}\widetilde{\mathbf{q}}$ is $\Pi^\star$. When the new distribution h has a better condition number κ_h than the condition number κ of f, we can use HMC to draw approximate sample from Π_h and then transform the samples using the matrix M. Clearly the bound from Corollary C.1 guarantees that when κ_h is much smaller than κ, HMC on the new target Π_h would mix much faster than the HMC chain on $\Pi^\star$. This transformation is equivalent to the HMC algorithm with modified kinetic energy

$$\frac{d\mathbf{q}_t}{dt} = M^{-1}\mathbf{p}_t \quad \text{and} \quad \frac{d\mathbf{p}_t}{dt} = -\nabla f(\mathbf{q}_t),$$

which is easier to implement in practice. For a detailed discussion of this implementation, we refer the readers to the paper by Neal [190].

C.3.2 Comparison with guarantees for unadjusted versions of HMC

In this appendix, we compare our results with mixing time guarantees results on unadjusted and ODE solver based HMC chains. We summarize the number of gradient evaluations needed for Metropolized HMC to mix and those for other existing sampling results in Table C.4. Note that all the results summarized here are the best upper bounds in the literature for log-concave sampling. We present the results for a $(\mathcal{L}, \mathcal{L}_H, m)$-regular target distribution. We remark that all methods presented in Table C.4 requires the regularity assumptions (3.5a) and (3.5b), even though some do not require assumption (5.5).

Sampling algorithm	#Gradient evaluations
$^{\ddagger,\diamond}$Unadjusted HMC with leapfrog integrator [176]	$d^{\frac{1}{4}}\kappa^{\frac{11}{4}} \cdot \frac{1}{\delta^{1/2}}$
‡Underdamped Langevin [52]	$d^{\frac{1}{2}}\kappa^{2} \cdot \frac{1}{\delta}$
‡HMC with ODE solver, Thm. 1.6 in [153]	$d^{\frac{1}{2}}\kappa^{\frac{7}{4}} \cdot \frac{1}{\delta}$
*MALA [Thm. 3.1]	$\max\left\{d\kappa, d^{\frac{1}{2}}\kappa^{\frac{3}{2}}\right\} \cdot \log\frac{1}{\delta}$
*Metropolized HMC with leapfrog integrator [Cor. 5.1]	$\max\left\{d\kappa^{\frac{3}{4}}, d^{\frac{11}{12}}\kappa, d^{\frac{3}{4}}\kappa^{\frac{5}{4}}, d^{\frac{1}{2}}\kappa^{\frac{3}{2}}\right\} \cdot \log\frac{1}{\delta}$

Table C.4: Summary of the number of gradient evaluations needed for the sampling algorithms to converge to a $(m, \mathcal{L}, \mathcal{L}_H)$-regular target distribution with $\mathcal{L}_H = \mathcal{O}(\mathcal{L}^{\frac{3}{2}})$ within δ error from the target distribution (in total-variation distance* or 1-Wasserstein distance‡) (and $\diamond$ certain additional regularity conditions for the result by Mangoubi et al. [176]). Note that the unadjusted algorithms suffer from an exponentially worse dependency on δ when compared to the Metropolis adjusted chains.

Two remarks are in order. First, the error metric for the guarantees in the works [176, 52, 153] is 1-Wasserstein distance, while our results make use of $\mathcal{L}_2$ or TV distance. As a result, a direct comparison between these results is not possible although we provide an indirect comparison below. Second, the previous guarantees have a polynomial dependence on the inverse of error-tolerance $1/\delta$. In contrast, our results for MALA and Metropolized HMC have a logarithmic dependence $\log(1/\delta)$. For a well-conditioned target, i.e., when κ is a constant, all prior results have a better dependence on d when compared to our bounds.

Logarithmic vs polynomial dependence on $1/\delta$: We now provide an indirect comparison, between prior guarantees based on Wasserstein distance and our results based on TV-distance, for estimating expectations of Lipschitz-functions on bounded domains. MCMC algorithms are used to estimate expectations of certain functions of interest. Given an arbitrary function g and an MCMC algorithm, one of the ways to estimate $\Pi^\star(g) := \mathbb{E}_{X\sim\Pi^\star}[g(X)]$ is to use the k-th iterate from N independent runs of the chain. Let $X_i^{(k)}$ for $i = 1, \ldots, N$ denote the N i.i.d. samples at the k-th iteration of the chain and let μ_k denote the distribution of $X_i^{(k)}$, namely the distribution of the chain after k iterations. Then for the estimate

$\widehat{\Pi}_k(g) \coloneqq \frac{1}{N}\sum_{i=1}^N g(X_i^{(k)})$, the estimation error can be decomposed as

$$\begin{aligned}\Pi^\star(g) - \widehat{\Pi}_k(g) &= \int_{\mathbb{R}^d} g(x)\pi^\star(x)dx - \frac{1}{N}\sum_{i=1}^N g(X_i^{(k)}) \\ &= \underbrace{\int_{\mathbb{R}^d} g(x)\left[\pi^\star(x) - \mu_k(x)\right]dx}_{=:J_1\ \text{(Approximation bias)}} + \underbrace{\mathbb{E}_{\mu_k}\left[g(X)\right] - \frac{1}{N}\sum_{i=1}^N g(X_i^{(k)})}_{=:J_2\ \text{(Finite sample error)}}. \qquad \text{(C.56)}\end{aligned}$$

To compare different prior works, we assume that $\mathrm{Var}_{\mu_k}\left[g(X_1)\right]$ is bounded and thereby that the finite sample error J_2 is negligible for large enough N.[3] It remains to bound the error J_1 which can be done in two different ways depending on the error-metric used to provide mixing time guarantees for the Markov chain.

If the function g is ω-Lipschitz and k is chosen such that $\mathcal{W}_1(\Pi^\star, \mu_k) \leq \delta$, then we have $J_1 \leq \omega\delta =: J_{\text{Wass}}$. On the other hand, if the function g is bounded by B, and k is chosen such that $d_{\text{TV}}\big(\Pi^\star, \mu_k\big) \leq \delta$, then we obtain the bound $J_1 \leq B\delta =: J_{\text{TV}}$. We make use of these two facts to compare the number of gradient evaluations needed by unadjusted HMC or ODE solved based HMC and Metropolized HMC. Consider an ω-Lipschitz function g with support on a ball of radius R. Note that this function is uniformly bounded by $B = \omega R$. Now in order to to ensure that $J_1 \leq \delta$ (some user-specified small threshold), the choice of δ in the two cases (Wasserstein and TV distance) would be different leading to different number of gradient evaluations required by the two chains. More precisely, we have

$$\begin{aligned} J_1 \leq J_{\text{Wass}} = \omega\delta \leq \delta \quad &\Longrightarrow \quad \delta_{\text{wass}} = \frac{\delta}{\omega} \quad \text{and} \\ J_1 \leq J_{\text{TV}} = B\delta = \omega R\delta \leq \delta \quad &\Longrightarrow \quad \delta_{\text{TV}} = \frac{\delta}{\omega R}. \end{aligned}$$

To simplify the discussion, we consider well-conditioned (constant κ) strongly log-concave distributions such that most of the mass is concentrated on a ball of radius $\mathcal{O}(\sqrt{d})$ (cf. Appendix C.2.1) and consider $R = \sqrt{d}$. Then plugging the error-tolerances from the display above in Table C.4, we obtain that the number of gradient evaluations $\mathcal{G}_{\text{MC}}$ for different chains[4] would scale as

$$\mathcal{G}_{\text{unadj.-HMC}} \leq \mathcal{O}(\sqrt{\frac{d\omega}{\delta}}), \quad \mathcal{G}_{\text{ODE-HMC}} \leq \mathcal{O}(\frac{\omega\sqrt{d}}{\delta}), \text{and} \quad \mathcal{G}_{\text{Metro.-HMC}} \leq \mathcal{O}(d\log\frac{\omega\sqrt{d}}{\delta})$$

Clearly, depending on ω and the threshold δ, different chains would have better guarantees. When ω is large or δ is small, our results ensure the superiority of Metropolized-HMC

[3] Moreover, this error should be usually similar across different sampling algorithms since several algorithms are designed in a manner agnostic to a particular function g.

[4] The results for other HMCs often assume (different) additional conditions so that a direct comparison should be taken with a fine grain of salt.

over other versions. For example, higher-order moments can be functions of interest, i.e., $g(x) = \|x\|^{1+\nu}$ for which the Lipschitz-constant $\omega = O(d^{\nu})$ scales with d. For this function, we obtain the bounds:

$$\mathcal{G}_{\text{unadj.-HMC}} \leq \mathcal{O}(\frac{d^{\frac{1+\nu}{2}}}{\sqrt{\delta}}), \quad \mathcal{G}_{\text{ODE-HMC}} \leq \mathcal{O}(\frac{d^{\frac{1}{2}+\nu}}{\delta}), \quad \text{and} \quad \mathcal{G}_{\text{Metro.-HMC}} \leq \mathcal{O}(d(1+\nu)\log\frac{d}{\delta})$$

and thus Metropolized HMC takes fewer gradient evaluations than ODE-based HMC for $\nu > 1/2$ and unadjusted HMC for $\nu > 1$ (to ensure $J_1 \leq \delta$ (C.56)). We remark that the bounds for unadjusted-HMC require additional regularity conditions. From this informal comparison, we demonstrate that both the dimension dependency d and error dependency δ should be accounted for comparing unadjusted algorithms and Metropolized algorithms. Especially for estimating high-order moments, Metropolized algorithms with $\log(\frac{1}{\delta})$ dependency will be advantageous.

Appendix D

Content Deferred From Chapter 6 For The Vaidya Walk

In this appendix, we collect the technical content used earlier in the proof of mixing time bounds for the Vaidya walk. In particular, we prove Lemmas 6.1 and 6.2 in Appendices D.1 and D.2 respectively, with other auxiliary proofs in Appendix D.3. Furthermore, we provide the proof of Proposition 6.1 in Appendix D.4.

D.1 Proof of Lemma 6.1

In order to prove part (h), observe that for any $x \in \text{int}(\mathcal{K})$, the Hessian $\nabla^2 \mathcal{F}_x := \sum_{i=1}^n a_i a_i^\top / s_{x,i}^2$ is a sum of rank one positive semidefinite (PSD) matrices. Also, we can write $\nabla^2 \mathcal{F}_x = A_x^\top A_x$ where

$$A_x \coloneqq \begin{bmatrix} a_1^\top / s_{x,1} \\ \vdots \\ a_n^\top / s_{x,n} \end{bmatrix}.$$

Since $\text{rank}(A_x) = d$, we conclude that the matrix $\nabla^2 \mathcal{F}_x$ is invertible and thus, both the matrices $\nabla^2 \mathcal{F}_x$ and $\left(\nabla^2 \mathcal{F}_x\right)^{-1}$ are PSD. Since $\sigma_{x,i} = a_i^\top \left(\nabla^2 \mathcal{F}_x\right)^{-1} a_i / s_{x,i}^2$, we have $\sigma_{x,i} \geq 0$. Further, the fact that $a_i a_i^\top / s_{x,i}^2 \preceq \nabla^2 \mathcal{F}_x$ implies that $\sigma_{x,i} \leq 1$.

Turning to the proof of part (i), from the equality $\text{trace}(AB) = \text{trace}(BA)$, we obtain

$$\sum_{i=1}^n \sigma_{x,i} = \text{trace}\left(\sum_{i=1}^n \frac{a_i^\top \left(\nabla^2 \mathcal{F}_x\right)^{-1} a_i}{s_{x,i}^2}\right) = \text{trace}\left(\left(\nabla^2 \mathcal{F}_x\right)^{-1} \sum_{i=1}^n \frac{a_i a_i^\top}{s_{x,i}^2}\right) = \text{trace}(\mathbb{I}_d) = d.$$

Now we prove part (j). Using the fact that $\sigma_{x,i} \geq 0$, and an argument similar to part (h) we find that that the matrices V_x and V_x^{-1} are PSD. Since $\theta_{V_x,i} = a_i^\top V_x^{-1} a_i / s_{x,i}^2$, we have $\theta_{V_x,i} \geq 0$. It is straightforward to see that $\beta_{\mathrm{V}} \nabla^2 \mathcal{F}_x \preceq V_x$ which implies that $\theta_{V_x,i} \leq \sigma_{x,i}/\beta$.

Further, we also have $(\sigma_{x,i} + \beta_{\mathrm{V}}) \frac{a_i a_i^\top}{s_{x,i}^2} \preceq V_x$ and whence $\theta_{V_x,i} \leq 1/(\sigma_{x,i} + \beta_{\mathrm{V}})$. Combining the two inequalities yields the claim.

The other parts of the Lemma follow from Lemma 13, 14 and 15 in the paper [152].

D.2 Proof of Lemma 6.2

We prove the lemma for the following function

$$g_{\mathrm{V}}(\epsilon) := \min\left\{\sqrt{\frac{1}{20\left(1+\sqrt{2}\log(4/\epsilon)\right)}}, \frac{\epsilon}{\sqrt{18\log(2/\epsilon)}}, \sqrt{\frac{\epsilon}{86\sqrt{3}\gamma_2}}, \frac{\epsilon}{22\sqrt{5/3}\gamma_3}, \sqrt{\frac{\epsilon}{50\sqrt{105}\gamma_4}}\right\}, \tag{D.1}$$

where $\gamma_k = (2e/k \cdot \log(4/\epsilon))^{k/2}$ for $k = 2, 3$ and 4. A numerical calculation shows that $f(1/15) \geq 10^{-4}$.

D.2.1 Proof of claim (6.21a)

In order to bound the total variation distance $d_{\mathrm{TV}}\big(\mathcal{P}_x, \mathcal{P}_y\big)$, we apply Pinsker's inequality, which provides an upper bound on the TV-distance in terms of the KL divergence:

$$d_{\mathrm{TV}}\big(\mathcal{P}_x, \mathcal{P}_y\big) \leq \sqrt{2\,\mathrm{KL}(\mathcal{P}_x \| \mathcal{P}_y)}.$$

For Gaussian distributions, the KL divergence has a closed form expression. In particular, for two normal-distributions $\mathcal{G}_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{G}_2 = \mathcal{N}(\mu_2, \Sigma_2)$, the Kullback-Leibler divergence between the two is given by

$$\mathrm{KL}(\mathcal{G}_1 \| \mathcal{G}_2) = \frac{1}{2}\Big(\mathrm{trace}(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}) - d - \log\det(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}) + (\mu_1 - \mu_2)^\top \Sigma_1^{-1}(\mu_1 - \mu_2)\Big).$$

Substituting $\mathcal{G}_1 = \mathcal{P}_x$ and $\mathcal{G}_2 = \mathcal{P}_y$ into the above expression and applying Pinsker's inequality, we find that

$$\begin{aligned} d_{\mathrm{TV}}\big(\mathcal{P}_x, \mathcal{P}_y\big)^2 \leq 2\,\mathrm{KL}(\mathcal{P}_y \| \mathcal{P}_x) &= \mathrm{trace}(V_x^{-1/2} V_y V_x^{-1/2}) - d - \log\det(V_x^{-1/2} V_y V_x^{-1/2}) + \frac{\sqrt{nd}}{r^2}\left\|x - y\right\|_x^2 \\ &= \left\{\sum_{i=1}^d \left(\lambda_i - 1 + \log\frac{1}{\lambda_i}\right)\right\} + \frac{\sqrt{nd}}{r^2}\left\|x - y\right\|_x^2, \end{aligned} \tag{D.2}$$

where $\lambda_1, \ldots, \lambda_d > 0$ denote the eigenvalues of the matrix $V_x^{-1/2} V_y V_x^{-1/2}$, and we have used the facts that $\det(V_x^{-1/2} V_y V_x^{-1/2}) = \prod_{i=1}^d \lambda_i$ and $\mathrm{trace}(V_x^{-1/2} V_y V_x^{-1/2}) = \sum_{i=1}^d \lambda_i$. The following lemma is useful in bounding expression (D.2).

Lemma D.1. *For any scalar $t \in [0, 1/12]$ and any pair $x, y \in \operatorname{int}(\mathcal{K})$ such that $\|x - y\|_x \leq t/(nd)^{1/4}$, we have*

$$\left(1 - \frac{6t}{\sqrt{d}} + \frac{t^2}{d}\right) \mathbb{I}_d \preceq V_x^{-1/2} V_y V_x^{-1/2} \preceq \left(1 + \frac{6t}{\sqrt{d}} + \frac{t^2}{d}\right) \mathbb{I}_d,$$

where $\preceq$ denotes ordering in the PSD cone, and $\mathbb{I}_d$ is the d-dimensional identity matrix.

See Appendix D.3.3 for the proof of this claim.

For $\epsilon \in (0, 1/15]$ and $r = 10^{-4}$, we have $t = \epsilon r/2 \leq 1/12$, whence the eigenvalues $\{\lambda_i, i \in [d]\}$ can be sandwiched as

$$1 - \frac{3\epsilon r}{\sqrt{d}} + \frac{\epsilon^2 r^2}{4d} \leq \lambda_i \leq 1 + \frac{3\epsilon r}{\sqrt{d}} + \frac{\epsilon^2 r^2}{4d} \quad \text{for all } i \in d. \tag{D.3}$$

We are now ready to bound the TV distance between $\mathcal{P}_x$ and $\mathcal{P}_y$. Using the bound (D.2) and the inequality $\log \gamma \leq \gamma - 1$, valid for $\gamma > 0$, we obtain

$$d_{\mathrm{TV}}\big(\mathcal{P}_x, \mathcal{P}_y\big)^2 \leq \sum_{i=1}^{d} \left(\lambda_i - 2 + \frac{1}{\lambda_i}\right) + \frac{\sqrt{nd}}{r^2} \|x - y\|_x^2 .$$

Using the assumption that $\|x - y\|_x \leq \epsilon r / \left(2(nd)^{1/4}\right)$, and plugging in the bounds (D.3) for the eigenvalues $\{\lambda_i, i \in [d]\}$, we find that

$$\sum_{i=1}^{d} \left(\lambda_i - 2 + \frac{1}{\lambda_i}\right) + \frac{\sqrt{nd}}{r^2} \|x - y\|_x^2 \leq \frac{141\epsilon^2 r^2}{4} + \frac{\epsilon^2}{4}.$$

In asserting this inequality, we have used the facts that

$$\frac{1}{1 - 6\gamma + \gamma^2} \leq 1 + 6\gamma + 70\gamma^2, \quad \text{and} \quad \frac{1}{1 + 6\gamma + \gamma^2} \leq 1 - 6\gamma + 70\gamma^2 \quad \text{for all } \gamma \in \left[0, \tfrac{1}{12}\right].$$

Note that for any $r \in [0, 1/12]$ we have that $141r^2/4 \leq 1/2$. Putting the pieces together yields $d_{\mathrm{TV}}\big(\mathcal{P}_x, \mathcal{P}_y\big) \leq \epsilon$, as claimed.

D.2.2 Proof of claim (6.21b)

Note that

$$\mathcal{T}_x(\{x\}) = \mathcal{P}_x(\mathcal{K}^c) + 1 - \int_{\mathcal{K}} \min\left\{1, \frac{p_z(x)}{p_x(z)}\right\} p_x(z) dz, \tag{D.4}$$

where $\mathcal{K}^c$ denotes the complement of $\mathcal{K}$. Consequently, we find that

$$\begin{aligned} d_{\text{TV}}\left(\mathcal{P}_x, \mathcal{T}_x\right) &= \frac{1}{2}\left(\mathcal{T}_x(\{x\}) + \int_{\mathbb{R}^d} p_x(z)dz - \int_{\mathcal{K}} \min\left\{1, \frac{p_z(x)}{p_x(z)}\right\} p_x(z)dz\right) \\ &= \frac{1}{2}\left(\mathcal{P}_x(\mathcal{K}^c) + 2 - 2\int_{\mathbb{R}^d} \min\left\{1, \frac{p_z(x)}{p_x(z)}\right\} p_x(z)dz + 2\int_{\mathcal{K}^c} \min\left\{1, \frac{p_z(x)}{p_x(z)}\right\} p_x(z)dz\right) \\ &\leq \underbrace{\frac{3}{2}\mathcal{P}_x(\mathcal{K}^c)}_{=:\, S_1} + \underbrace{1 - \mathbb{E}_{z\sim\mathcal{P}_x}\left[\min\left\{1, \frac{p_z(x)}{p_x(z)}\right\}\right]}_{=:\, S_2}, \end{aligned} \tag{D.5}$$

Consequently, it suffices to show that both S_1 and S_2 are small, where the probability is taken over the randomness in the proposal z. In particular, we show that $S_1 \leq \epsilon$ and $S_2 \leq 4\epsilon$.

Bounding the term S_1: Since z is multivariate Gaussian with mean x and covariance $\frac{r^2}{\sqrt{nd}}V_x^{-1}$, we can write

$$z \overset{d}{=} x + \frac{r}{(nd)^{1/4}} V_x^{-1/2}\xi, \tag{D.6}$$

where $\xi \sim \mathcal{N}\left(0, \mathbb{I}_d\right)$ and $\overset{d}{=}$ denotes equality in distribution. Using equation (E.14) and definition (6.20b) of $\theta_{V_x,i}$, we obtain the bound

$$\frac{\left(a_i^\top\left(z-x\right)\right)^2}{s_{x,i}^2} = \frac{r^2}{(nd)^{\frac{1}{2}}}\left[\frac{a_i^\top V_x^{-1/2}\xi}{s_{x,i}}\right]^2 \overset{(i)}{\leq} \frac{r^2}{(nd)^{\frac{1}{2}}}\theta_{V_x,i}\left\|\xi\right\|_2^2 \overset{(ii)}{\leq} \frac{r^2}{d}\left\|\xi\right\|_2^2, \tag{D.7}$$

where step (i) follows from Cauchy-Schwarz inequality, and step (ii) from the bound on $\theta_{V_x,i}$ from Lemma 6.1(j). Define the events

$$\mathcal{E} := \left\{\frac{r^2}{d}\left\|\xi\right\|_2^2 < 1\right\} \quad \text{and} \quad \mathcal{E}' := \left\{z \in \text{int}\left(\mathcal{K}\right)\right\}.$$

Inequality (E.15) implies that $\mathcal{E} \subseteq \mathcal{E}'$ and hence $\mathbb{P}\left[\mathcal{E}'\right] \geq \mathbb{P}\left[\mathcal{E}\right]$. Using a standard Gaussian tail bound and noting that $r \leq \frac{1}{1+\sqrt{2/d\log(2/\epsilon)}}$, we obtain $\mathbb{P}\left[\mathcal{E}\right] \geq 1-\epsilon/2$ and whence $\mathbb{P}\left[\mathcal{E}'\right] \geq 1-\epsilon/2$. Thus, we have shown that $\mathbb{P}\left[z \notin \mathcal{K}\right] \leq \epsilon/2$ which implies that $S_1 \leq \epsilon$.

Bounding the term S_2: By Markov's inequality, we have

$$\mathbb{E}_{z\sim\mathcal{P}_x}\left[\min\left\{1, \frac{p_z(x)}{p_x(z)}\right\}\right] \geq \alpha\mathbb{P}\left[p_z(x) \geq \alpha p_x(z)\right] \quad \text{for all } \alpha \in (0,1]. \tag{D.8}$$

By definition (E.3) of p_x, we obtain

$$\frac{p_z(x)}{p_x(z)} = \exp\left(-\frac{\sqrt{nd}}{2r^2}\left(\left\|z-x\right\|_z^2 - \left\|z-x\right\|_x^2\right) + \frac{1}{2}\left(\log\det V_z - \log\det V_x\right)\right).$$

The following lemma provides us with useful bounds on the two terms in this expression, valid for any $x \in \text{int}(\mathcal{K})$.

Lemma D.2. *For any $\epsilon \in (0, 1/15]$ and $r \in (0, g_V(\epsilon)]$, we have*

$$\mathbb{P}_{z\sim\mathcal{P}_x}\left[\frac{1}{2}\log\det V_z - \frac{1}{2}\log\det V_x \geq -\epsilon\right] \geq 1-\epsilon, \quad \textit{and} \tag{D.9a}$$

$$\mathbb{P}_{z\sim\mathcal{P}_x}\left[\|z-x\|_z^2 - \|z-x\|_x^2 \leq 2\epsilon\frac{r^2}{\sqrt{nd}}\right] \geq 1-\epsilon. \tag{D.9b}$$

See Appendix D.3.4 for the proof of this claim.

Using Lemma D.2, we now complete the proof. For $r \leq g_{\text{V}}(\epsilon)$, we obtain

$$\frac{p_z(x)}{p_x(z)} \geq \exp(-2\epsilon) \geq 1-2\epsilon$$

with probability at least $1-2\epsilon$. Substituting $\alpha = 1-2\epsilon$ in inequality (E.16) yields that $S_2 \leq 4\epsilon$, as claimed.

D.3 Proofs of Lemmas D.1 and D.2

To prove Lemmas D.1 and D.2, we first state some additional notation, and additional technical results.

D.3.1 Notation

We begin with introducing some additional notation. Recall $A \in \mathbb{R}^{n\times d}$ is a matrix with $a_i^\top$ as its i-th row. For any positive integer p and any vector $v = (v_1, \ldots, v_p)^\top$, $\text{diag}(v) = \text{diag}(v_1, \ldots, v_p)$ denotes a $p \times p$ diagonal matrix with the i-th diagonal entry equal to v_i. Recall the definition of S_x:

$$S_x = \text{diag}(s_{x,1}, \ldots, s_{x,n}) \text{ where } s_{x,i} = b_i - a_i^\top x \text{ for each } i \in [n]. \tag{D.10}$$

Furthermore, define $A_x = S_x^{-1}A$ for all $x \in \text{int}(\mathcal{K})$, and let Υ_x denote the projection matrix for the column space of A_x, i.e.,

$$\Upsilon_x := A_x(A_x^\top A_x)^{-1}A_x^\top = A_x\nabla^2\mathcal{F}_x^{-1}A_x^\top. \tag{D.11}$$

Note that for the scores σ_x (6.7b), we have $\sigma_{x,i} = (\Upsilon_x)_{ii}$ for each $i \in [n]$. Let Σ_x be an $n \times n$ diagonal matrix defined as

$$\Sigma_x = \text{diag}(\sigma_{x,1}, \ldots, \sigma_{x,n}). \tag{D.12}$$

Let $\sigma_{x,i,j} \coloneqq (\Upsilon_x)_{ij}$, and let $\Upsilon_x^{(2)}$ denote the Hadamard product of Υ_x with itself, i.e.,

$$(\Upsilon_x^{(2)})_{ij} = \sigma_{x,i,j}^2 = \frac{\left(a_i^\top \nabla^2 \mathcal{F}_x^{-1} a_j\right)^2}{s_{x,i}^2 s_{x,j}^2} \quad \text{for all } i, j \in [n]. \tag{D.13}$$

Using the shorthand $\theta_x \coloneqq \theta_{V_x}$, we define

$$\Theta_x \coloneqq \operatorname{diag}\left(\theta_{x,1}, \ldots, \theta_{x,m}\right) \text{ where } \theta_{x,i} = \frac{a_i^\top V_x^{-1} a_i}{s_{x,i}^2} \quad \text{for } i \in [n], \text{ and}$$

$$\Xi_x \coloneqq (\theta_{x,i,j}^2) \text{ where } \theta_{x,i,j}^2 = \frac{\left(a_i^\top V_x^{-1} a_j\right)^2}{s_{x,i}^2 s_{x,j}^2} \quad \text{for } i, j \in [n].$$

In our new notation, we can re-write $V_x = A_x^\top \left(\Sigma_x + \beta_{\mathrm{V}} \mathbb{I}\right) A_x$.

D.3.2 Basic Properties

We begin by summarizing some key properties of various terms involved in our analysis.

Lemma D.3. *For any vector $x \in \operatorname{int}(\mathcal{K})$, the following properties hold:*

(a) $\sigma_{x,i} = \sum_{j=1}^n \sigma_{x,i,j}^2 = \sum_{j,k=1}^n \sigma_{x,i,j}\sigma_{x,j,k}\sigma_{x,k,i}$ for each $i \in [n]$,

(b) $\Sigma_x \succeq \Upsilon_x^{(2)}$,

(c) $\sum_{i=1}^n \theta_{x,i}\left(\sigma_{x,i} + \beta_V\right) = d$,

(d) $\forall i \in [n],\ \theta_{x,i} = \sum_{j=1}^n \left(\sigma_{x,j} + \beta_V\right)\theta_{x,i,j}^2$, for each $i \in [n]$,

(e) $\theta_x^\top \left(\Sigma_x + \beta_V \mathbb{I}\right)\theta_x = \sum_{i=1}^n \theta_{x,i}^2 \left(\sigma_{x,i} + \beta_V\right) \le \sqrt{nd}$, and

(f) $\beta_V \nabla^2 \mathcal{F}_x \preceq V_x \preceq \left(1 + \beta_V\right) \nabla^2 \mathcal{F}_x$.

Proof. We prove each property separately.

Part (a): Using $\mathbb{I}_d = \nabla^2 \mathcal{F}_x \left(\nabla^2 \mathcal{F}_x\right)^{-1}$, we find that

$$\sigma_{x,i} = \frac{a_i^\top \left(\nabla^2 \mathcal{F}_x\right)^{-1} \nabla^2 \mathcal{F}_x \left(\nabla^2 \mathcal{F}_x\right)^{-1} a_i}{s_{x,i}^2} = \frac{a_i^\top \left(\nabla^2 \mathcal{F}_x\right)^{-1} \nabla^2 \sum_{j=1}^n \frac{a_j^\top a_j}{s_{x,j}^2} \left(\nabla^2 \mathcal{F}_x\right)^{-1} a_i}{s_{x,i}^2} = \sum_{i,j=1}^n \sigma_{x,i,j}^2.$$

Applying a similar trick twice and performing some algebra, we obtain

$$\sigma_{x,i} = \frac{a_i^\top \left(\nabla^2 \mathcal{F}_x\right)^{-1} \nabla^2 \mathcal{F}_x \left(\nabla^2 \mathcal{F}_x\right)^{-1} \nabla^2 \mathcal{F}_x \left(\nabla^2 \mathcal{F}_x\right)^{-1} a_i}{s_{x,i}^2} = \sum_{i,j,k=1}^n \sigma_{x,i,j}\sigma_{x,j,k}\sigma_{x,k,i}.$$

Part (b): From part (a), we have that $\Sigma_x - \Upsilon_x^{(2)}$ is a symmetric and diagonally dominant matrix with non-negative entries on the diagonal. Applying Gershgorin's theorem [20, 117], we conclude that it is PSD.

Part (c): Since trace$(AB) =$ trace(BA), we have

$$\sum_{i=1}^{n} \theta_{x,i} \left(\sigma_{x,i} + \beta_{\mathrm{V}}\right) = \mathrm{trace}\left(V_x^{-1} \sum_{i=1}^{n} \left(\sigma_{x,i} + \beta_{\mathrm{V}}\right) \frac{a_i a_i^\top}{s_{x,i}^2} \right) = \mathrm{trace}\left(\mathbb{I}_d\right) = d.$$

Part (d): An argument similar to part (a) implies that

$$\theta_{x,i} = \frac{a_i^\top V_x^{-1} V_x V_x^{-1} a_i}{s_{x,i}^2} = \frac{a_i^\top V_x^{-1} \sum_{j=1}^{n} \left(\sigma_{x,i} + \beta_{\mathrm{V}}\right) \frac{a_j^\top a_j}{s_{x,j}^2} V_x^{-1} a_i}{s_{x,i}^2} = \sum_{i,j=1}^{n} \left(\sigma_{x,i} + \beta_{\mathrm{V}}\right) \theta_{x,i,j}^2.$$

Part (e): Using part (c) and Lemma 6.1(j) yields the claim.

Part (f): The left inequality is by the definition of V_x. The right inequality uses the fact that $\Sigma_x \preceq \mathbb{I}_d$. □

We now prove an important result that relates the *slackness* s_x and s_y at two points, in terms of $\|x - y\|_x$.

Lemma D.4. *For all* $x, y \in \mathrm{int}\left(\mathcal{K}\right)$, *we have*

$$\left| 1 - \frac{s_{y,i}}{s_{x,i}} \right| \leq \left(\frac{n}{d}\right)^{\frac{1}{4}} \left\|x - y\right\|_x \quad \textit{for each } i \in [n].$$

Proof. For any pair $x, y \in \mathrm{int}\left(\mathcal{K}\right)$ and index $i \in [n]$, we have

$$\begin{aligned}
\left(a_i^\top \left(x - y\right)\right)^2 = \left(\left(V_x^{-\frac{1}{2}} a_i\right)^\top V_x^{\frac{1}{2}} \left(x - y\right)\right)^2 &\overset{(i)}{\leq} \|V_x^{-\frac{1}{2}} a_i\|_2^2 \; \|V_x^{\frac{1}{2}} \left(x - y\right)\|_2^2 \\
&= a_i^T V_x^{-1} a_i \; \left\|x - y\right\|_x^2 \\
&= \theta_{x,i} s_{x,i}^2 \; \left\|x - y\right\|_x^2 \\
&\overset{(ii)}{\leq} \sqrt{\frac{n}{d}} s_{x,i}^2 \; \left\|x - y\right\|_x^2,
\end{aligned}$$

where step (i) follows from the Cauchy-Schwarz inequality, and step (ii) uses the bound $\theta_{x,i}$ from Lemma 6.1(j). Noting the fact that $a_i^\top (x - y) = s_{y,i} - s_{x,i}$, the claim follows after simple algebra. □

D.3.3 Proof of Lemma D.1

As a direct consequence of Lemma D.4, we find that

$$\left|1 - \frac{s_{y,i}}{s_{x,i}}\right| \leq \frac{t}{\sqrt{d}}, \quad \text{for any } x, y \in \operatorname{int}(\mathcal{K}) \text{ such that } \|x-y\|_x \leq \frac{t}{(nd)^{1/4}}.$$

The Hessian $\nabla^2 \mathcal{F}_y$ is thus sandwiched in terms of the Hessian $\nabla^2 \mathcal{F}_x$ as

$$\left(1 - \frac{t}{\sqrt{d}}\right)^2 \nabla^2 \mathcal{F}_x \preceq \nabla^2 \mathcal{F}_y \preceq \left(1 + \frac{t}{\sqrt{d}}\right)^2 \nabla^2 \mathcal{F}_x.$$

By the definition of $\sigma_{x,i}$ and $\sigma_{y,i}$, we have

$$\frac{\left(1 - \frac{t}{\sqrt{d}}\right)^2}{\left(1 + \frac{t}{\sqrt{d}}\right)^2} \sigma_{x,i} \leq \sigma_{y,i} \leq \frac{\left(1 + \frac{t}{\sqrt{d}}\right)^2}{\left(1 - \frac{t}{\sqrt{d}}\right)^2} \sigma_{x,i} \quad \text{for all } i \in [n]. \tag{D.14}$$

Consequently, we find that

$$\frac{\left(1 - \frac{t}{\sqrt{d}}\right)^2}{\left(1 + \frac{t}{\sqrt{d}}\right)^4} V_x \preceq V_y \preceq \frac{\left(1 + \frac{t}{\sqrt{d}}\right)^2}{\left(1 - \frac{t}{\sqrt{d}}\right)^4} V_x.$$

Note that

$$1 - 6\gamma + \gamma^2 \ \leq \ \frac{(1-\gamma)^2}{(1+\gamma)^4} \leq 1 + 6\gamma + \gamma^2 \quad \text{for any } \gamma \in \left[0, \tfrac{1}{12}\right].$$

Applying this sandwiching pair of inequalities with $\gamma = t/\sqrt{d}$ yields the claim.

D.3.4 Proof of Lemma D.2

We begin by defining

$$\varphi_{x,i} := \frac{\sigma_{x,i} + \beta_{\mathrm{V}}}{s_{x,i}^2} \text{ for } i \in [n], \quad \text{and} \quad \Psi_x := \frac{1}{2} \log\det V_x, \quad \text{for all } x \in \operatorname{int}(\mathcal{K}). \tag{D.15}$$

Further, for any two points x and z, let $\overline{xz}$ denote the set of points on the line segment joining x and z. The proof of Lemma D.2 is based on a Taylor series expansion, and so requires careful handling of σ, φ, Ψ and their derivatives. At a high level, the proof involves the following steps: (1) perform a Taylor series expansion around x and along the line segment $\overline{xz}$; (2) transfer the bounds of terms involving some point $y \in \overline{xz}$ to terms involving only x and z; and then (3) use concentration of Gaussian polynomials to obtain high probability bounds.

We now introduce some auxiliary results involved in these three steps. The following lemma provides expressions for gradients of σ, φ and Ψ and bounds for directional Hessian of φ and Ψ. Let $e_i \in \mathbb{R}^d$ denote a vector with 1 in the i-th position and 0 otherwise. For any $h \in \mathbb{R}^d$ and $x \in \text{int}(\mathcal{K})$, define $\eta_{x,h,i} = \eta_{x,i} := a_i^\top h / s_{x,i}$ for each $i \in [n]$.

Lemma D.5. *The following relations hold;*

(a) Gradient of σ: $\nabla \sigma_{x,i} = 2A_y^\top(\Sigma_x - P_x^{(2)})e_i$ for each $i \in [n]$.

(b) Gradient of φ: $\nabla \varphi_{x,i} = \dfrac{2}{s_{y,i}^2} A_x^\top \left[2\Sigma_x + \beta_V \mathbb{I} - P_x^{(2)}\right] e_i$ for each $i \in [n]$;

(c) Gradient of Ψ: $\nabla \Psi_x = A_x^\top \left(2\,\Sigma_x + \beta_V \mathbb{I} - P_x^{(2)}\right) \theta_x$;

(d) Bound on $\nabla^2 \varphi$: $s_{x,i}^2 \left|\frac{1}{2} h^\top \nabla^2 \varphi_{x,i} h\right| \leq 14\,(\sigma_{x,i} + \beta_V)\,\eta_{x,i}^2 + 11\,\sum_{j=1}^n \sigma_{x,i,j}^2 \eta_{x,j}^2$ for $i \in [n]$;

(e) Bound on $\nabla^2 \Psi$: $\left|\frac{1}{2} h^\top \left(\nabla^2 \Psi_x\right) h\right| \leq 13\,\sum_{i=1}^n (\sigma_{x,i} + \beta_V)\,\theta_{x,i}\eta_{x,i}^2 + \frac{17}{2}\,\sum_{i,j=1}^n \sigma_{x,i,j}^2 \theta_{x,i} \eta_{x,j}^2$.

See Section D.3.5 for the proof of this claim.

The following lemma that shows that for a random variable $z \sim \mathcal{P}_x$, the slackness $s_{z,i}$ is close to $s_{x,i}$ with high probability.

Lemma D.6. *For any $\epsilon \in (0, 1/4], r \in (0,1)$ and $x \in \text{int}(\mathcal{K})$, we have*

$$\mathbb{P}_{z \sim \mathcal{P}_x}\left[\forall i \in [n], \forall v \in \overline{xz},\ \frac{s_{x,i}}{s_{v,i}} \in \left(1 - \sqrt{1+\delta}\,r, 1 + \sqrt{1+\delta}\,r\right)\right] \geq 1 - \epsilon/4,$$

where $\delta = \sqrt{\frac{2}{d}\log\left(\frac{4}{\epsilon}\right)}$. Thus for any $d \geq 1$ and $r \leq 1/(20(1 + \sqrt{2}\log(4/\epsilon))^{1/2}$, we have

$$\mathbb{P}_{z \sim \mathcal{P}_x}\left[\forall i \in [n], \forall v \in \overline{xz},\ \frac{s_{x,i}}{s_{v,i}} \in (0.95, 1.05)\right] \geq 1 - \epsilon/4.$$

This result comes in handy for transferring bounds for different expressions in Taylor expansion involving an arbitrary y on $\overline{xz}$ to bounds on terms involving simply x. The proof follows from Lemma D.4 and a simple application of the standard Gaussian tail bounds and is thereby omitted. For brevity, we define the shorthand

$$\hat{a}_i = \frac{1}{s_{x,i}} V_x^{-1/2} a_i \quad \text{for each } i \in [n], \tag{D.16}$$

where we have omitted the dependence of $\hat{a}_i$ on x. In the following lemma, we state some tail bounds for particular Gaussian polynomials that arise in our analysis.

Lemma D.7. *For any $\epsilon \in (0, 1/15]$, define $\gamma_k = (2e/k \cdot \log(4/\epsilon))^{k/2}$ for $k = 2, 3$ and 4. Then for $\xi \sim \mathcal{N}(0, \mathbb{I}_d)$ and any $x \in \operatorname{int}(\mathcal{K})$ the following high probability bounds hold:*

$$\mathbb{P}\left[\sum_{i=1}^{n} (\sigma_{x,i} + \beta_V) \left(\hat{a}_i^\top \xi\right)^2 \leq \gamma_2 \sqrt{3} d\right] \geq 1 - \frac{\epsilon}{4}, \tag{D.17a}$$

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} (\sigma_{x,i} + \beta_V) \left(\hat{a}_i^\top \xi\right)^3\right| \leq \gamma_3 \sqrt{15}\, (nd)^{1/4}\right] \geq 1 - \frac{\epsilon}{4}, \tag{D.17b}$$

$$\mathbb{P}\left[\left|\sum_{i,j=1}^{n} \sigma_{x,i,j}^2 \left(\left(\frac{\hat{a}_i + \hat{a}_j}{2}\right)^\top \xi\right)^3\right| \leq \gamma_3 \sqrt{15}\, (nd)^{1/4}\right] \geq 1 - \frac{\epsilon}{4}, \tag{D.17c}$$

$$\mathbb{P}\left[\sum_{i=1}^{n} (\sigma_{x,i} + \beta_V) \left(\hat{a}_i^\top \xi\right)^4 \leq \gamma_4 \sqrt{105}\, (nd)^{1/2}\right] \geq 1 - \frac{\epsilon}{4}. \tag{D.17d}$$

See Section D.3.6 for the proof of these claims.

Now we summarize the final ingredients needed for our proofs. Recall that the Gaussian proposal z is related to the current state x via the equation

$$z \overset{d}{=} x + \frac{r}{(nd)^{1/4}} V_x^{-1/2} \xi, \tag{D.18}$$

where $\xi \sim \mathcal{N}(0, \mathbb{I}_d)$. We also use the following elementary inequalities:

$$\text{Cauchy-Schwarz inequality:} \qquad |u^\top v| \leq \|u\|_2 \|v\|_2 \tag{C-S}$$

$$\text{AM-GM inequality:} \qquad \nu\kappa \leq \frac{1}{2}(\nu^2 + \kappa^2). \tag{AM-GM}$$

$$\text{Sum of squares inequality:} \qquad \frac{1}{2}\|a + b\|_2^2 \leq \|a\|_2^2 + \|b\|_2^2, \tag{SSI}$$

Note that the sum-of-squares inequality is simply a vectorized version of the AM-GM inequality. With these tools, we turn to the proof of Lemma D.2. We split our analysis into parts.

D.3.4.1 Proof of claim (D.9a)

Using the second degree Taylor expansion, we have

$$\Psi_z - \Psi_x = (z - x)^\top \nabla \Psi_x + \frac{1}{2}(z - x)^\top \nabla^2 \Psi_y (z - x), \quad \text{for some } y \in \overline{xz}.$$

We claim that for $r \leq g_{\mathrm{V}}(\epsilon)$, we have

$$\mathbb{P}_z\left[(z - x)^\top \nabla \Psi_x \geq -\epsilon/2\right] \geq 1 - \epsilon/2, \quad \text{and} \tag{D.19a}$$

$$\mathbb{P}_z\left[\frac{1}{2}(z - x) \nabla^2 \Psi_y (z - x) \geq -\epsilon/2\right] \geq 1 - \epsilon/2. \tag{D.19b}$$

Note that the claim (D.9a) is a consequence of these two auxiliary claims, which we now prove.

Proof of bound (D.19a)**:** Equation (D.18) implies that $(z-x)^\top \nabla\Psi_x \sim \mathcal{N}\left(0, \frac{r^2}{\sqrt{nd}}\nabla\Psi_x^\top V_x^{-1}\nabla\Psi_x\right)$. We claim that

$$\nabla\Psi_x^\top V_x^{-1}\nabla\Psi_x \le 9\sqrt{nd} \quad \text{for all } x \in \text{int}\left(\mathcal{K}\right). \tag{D.20}$$

We prove this inequality at the end of this subsection. Taking it as given for now, let $\xi' \sim \mathcal{N}(0, 9r^2)$. Then using inequality (D.20) and a standard Gaussian tail bound, we find that

$$\mathbb{P}\left[(z-x)^\top \nabla\Psi_x \ge -\gamma\right] \ge \mathbb{P}\left[\xi' \ge -\gamma\right] \ge 1 - \exp(-\gamma^2/(18r^2)), \quad \text{valid for all } \gamma \ge 0.$$

Setting $\gamma = \epsilon/2$ and noting that $r \le \frac{\epsilon}{\sqrt{18\log(2/\epsilon)}}$ completes the claim.

Proof of bound (D.19b)**:** Let $\eta_{x,i} = \frac{a_i^\top(z-x)}{s_{x,i}} = \frac{r}{(mn)^{\frac{1}{4}}}\hat{a}_i^\top\xi$. Using Lemma D.5(e), we have

$$\begin{aligned}\left|\frac{1}{2}\left(z-x\right)\nabla^2\Psi_y\left(z-x\right)\right| &\le 13\sum_{i=1}^n \left(\sigma_{y,i}+\beta_{\mathrm{V}}\right)\theta_{y,i}\frac{s_{x,i}^2}{s_{y,i}^2}\eta_{x,i}^2 + \frac{17}{2}\sum_{i,j=1}^n \sigma_{y,i,j}^2\theta_{y,i}\frac{s_{x,j}^2}{s_{y,j}^2}\eta_{x,j}^2 \\ &\le \frac{43}{2}\sqrt{\frac{n}{d}}\sum_{i=1}^n \left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\frac{\left(\sigma_{y,i}+\beta_{\mathrm{V}}\right)s_{x,i}^2}{\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)s_{y,i}^2}\eta_{x,i}^2.\end{aligned} \tag{D.21}$$

Setting $\tau = 1.05$, we define the events $\mathcal{E}_1$ and $\mathcal{E}_2$ as follows:

$$\mathcal{E}_1 = \left\{\forall i \in [n],\ \frac{s_{x,i}}{s_{y,i}} \in [2-\tau, \tau]\right\}, \quad \text{and} \tag{D.22a}$$

$$\mathcal{E}_2 = \left\{\forall i \in [n],\ \frac{\sigma_{x,i}}{\sigma_{y,i}} \in \left[0, \frac{\tau^2}{(2-\tau)^2}\right]\right\}. \tag{D.22b}$$

It is straightforward to see that $\mathcal{E}_1 \subseteq \mathcal{E}_2$. Since $r \le \frac{1}{20\sqrt{1+\sqrt{2}\log(4/\epsilon)}}$, Lemma E.10 implies that $\mathbb{P}\left[\mathcal{E}_1\right] \ge 1-\epsilon/4$ whence $\mathbb{P}\left[\mathcal{E}_2\right] \ge 1-\epsilon/4$. Using these high probability bounds and the setting $\tau = 1.05$, we obtain that with probability at least $1-\epsilon/4$

$$\sqrt{\frac{n}{d}}\sum_{i=1}^n \left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\frac{\left(\sigma_{y,i}+\beta_{\mathrm{V}}\right)s_{x,i}^2}{\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)s_{y,i}^2}\eta_{x,i}^2 \le 2\sqrt{\frac{n}{d}}\sum_{i=1}^n \left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\eta_{x,i}^2 = \frac{2r^2}{d}\sum_{i=1}^n \left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\left(\hat{a}_i^\top\xi\right)^2. \tag{D.23}$$

Applying the high probability bound Lemma D.7 (D.17a) and the condition

$$r \le \sqrt{\frac{\epsilon}{86\sqrt{3}\gamma_2}}, \tag{D.24}$$

we obtain that with probability at least $1-\epsilon/2$,

$$\left|\frac{1}{2}\left(z-x\right)\nabla^2\Psi_y\left(z-x\right)\right| \ge -\epsilon/2,$$

as claimed.

Proof of bound (D.20): We now return to prove our earlier inequality (D.20). Using the expression for the gradient $\nabla \Psi_x$ from Lemma D.5(c), we have that for any vector $u \in \mathbb{R}^n$

$$\begin{aligned} u^\top \nabla \Psi_x \nabla \Psi_x^\top u &= \left\langle u, A_x^\top \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta_{\mathrm{V}} \mathbb{I}\right) \theta_x \right\rangle^2 \\ &= \left\langle A_x u, \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta_{\mathrm{V}} \mathbb{I}\right) \theta_x \right\rangle^2 \\ &= \left\langle \left(\Sigma_x + \beta_{\mathrm{V}} \mathbb{I}\right)^{\frac{1}{2}} A_x u, \left(\Sigma_x + \beta_{\mathrm{V}} \mathbb{I}\right)^{-1/2} \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta_{\mathrm{V}} \mathbb{I}\right) \theta_x \right\rangle^2 \\ &\leq u^\top V_x u \cdot \theta_x^\top \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta_{\mathrm{V}} \mathbb{I}\right) \left(\Sigma_x + \beta_{\mathrm{V}} \mathbb{I}\right)^{-1} \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta_{\mathrm{V}} \mathbb{I}\right) \theta_x \quad \text{(D.25)} \end{aligned}$$

where the last step follows from the Cauchy-Schwarz inequality. As a consequence of Lemma E.5(b), the matrix $\Sigma_x - \Upsilon_x^{(2)}$ is PSD. Thus, we have

$$0 \preceq 2\Sigma_x - \Upsilon_x^{(2)} + \beta_{\mathrm{V}} \mathbb{I} \preceq 3\left(\Sigma_x + \beta_{\mathrm{V}} \mathbb{I}\right).$$

Consequently, we find that

$$0 \preceq \underbrace{\left(3\Sigma_x + 3\beta_{\mathrm{V}} \mathbb{I}\right)^{-1/2} \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta_{\mathrm{V}} \mathbb{I}\right) \left(3\Sigma_x + 3\beta_{\mathrm{V}} \mathbb{I}\right)^{-1/2}}_{=:L} \preceq \mathbb{I}.$$

We deduce that all eigenvalues of the matrix L lie in the interval $[0, 1]$ and hence all the eigenvalues of the matrix L^2 belong to the interval $[0, 1]$. As a result, we have

$$\left(2\Sigma_x - \Upsilon_x^{(2)} + \beta_{\mathrm{V}} \mathbb{I}\right) \left(3\Sigma_x + 3\beta_{\mathrm{V}} \mathbb{I}\right)^{-1} \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta_{\mathrm{V}} \mathbb{I}\right) \preceq \left(3\Sigma_x + 3\beta_{\mathrm{V}} \mathbb{I}\right).$$

Thus, we obtain

$$\theta_x^\top \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta_{\mathrm{V}} \mathbb{I}\right) \left(\Sigma_x + \beta_{\mathrm{V}} \mathbb{I}\right)^{-1} \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta_{\mathrm{V}} \mathbb{I}\right) \theta_x \leq 9\theta_x^\top \left(\Sigma_x + \beta_{\mathrm{V}} \mathbb{I}\right) \theta_x. \quad \text{(D.26)}$$

Finally, applying Lemma E.5 and combining bounds (D.25) and (D.26) yields the claim.

D.3.4.2 Proof of claim (D.9b)

The quantity of interest can be written as

$$\|z - x\|_z^2 - \|z - x\|_x^2 = \sum_{i=1}^n \left(a_i^\top (z - x)\right)^2 \left(\varphi_{z,i} - \varphi_{x,i}\right).$$

We can write $z = x + \alpha u$, where α is a scalar and u is a unit vector in $\mathbb{R}^d$. Then we have

$$\|z - x\|_z^2 - \|z - x\|_x^2 = \alpha^2 \sum_{i=1}^n \left(a_i^\top u\right)^2 \left(\varphi_{z,i} - \varphi_{x,i}\right).$$

We apply a Taylor series expansion for $\sum_{i=1}^n \left(a_i^\top u\right)^2 (\varphi_{z,i} - \varphi_{x,i})$ around the point x, along the line u. There exists a point $y \in \overline{xz}$ such that

$$\sum_{i=1}^n \left(a_i^\top u\right)^2 (\varphi_{z,i} - \varphi_{x,i}) = \sum_{i=1}^n \left(a_i^\top u\right)^2 \left((z-x)^\top \nabla\varphi_{x,i} + \frac{1}{2}(z-x)^\top \nabla^2 \varphi_{y,i}(z-x) \right).$$

Multiplying both sides by α^2, and using the shorthand $\eta_{x,i} = \frac{a_i^\top (z-x)}{s_{x,i}}$, we obtain

$$\|z-x\|_z^2 - \|z-x\|_x^2 = \sum_{i=1}^n \eta_{x,i}^2 s_{x,i}^2 (z-x)^\top \nabla\varphi_{x,i} + \sum_{i=1}^n \eta_{x,i}^2 s_{x,i}^2 \frac{1}{2}(z-x)^\top \nabla^2\varphi_{y,i}(z-x). \tag{D.27}$$

Substituting the expression for $\nabla\varphi_{x,i}$ from Lemma D.5(b) in equation (D.27) and performing some algebra, the first term on the RHS of equation (D.27) can be written as

$$\sum_{i=1}^n \eta_{x,i}^2 s_{x,i}^2 (z-x)^\top \nabla\varphi_{x,i} = 2\sum_{i=1}^n \left(\frac{7}{3}\sigma_{x,i} + \beta_{\mathrm{V}}\right)\eta_{x,i}^3 - \frac{1}{3}\sum_{i,j=1}^n \sigma_{x,i,j}^2 (\eta_{x,i} + \eta_{x,j})^3. \tag{D.28}$$

On the other hand, using Lemma D.5 (d), we have

$$\frac{1}{2}s_{x,i}^2 \left|(z-x)^\top \nabla^2\varphi_{y,i}(z-x)\right| \le \frac{s_{x,i}^2}{s_{y,i}^2}\left[14(\sigma_{y,i}+\beta_{\mathrm{V}})\frac{s_{x,i}^2}{s_{y,i}^2}\eta_{x,i}^2 + 11\left(\sum_{j=1}^n \sigma_{y,i,j}^2 \eta_{x,j}^2 \frac{s_{x,j}^2}{s_{y,j}^2}\right)\right]. \tag{D.29}$$

Now, we use a fourth degree Gaussian polynomial to bound both the terms on the RHS of inequality (D.29). To do so, we use high probability bound for $s_{x,i}/s_{y,i}$. In particular, we use the high probability bounds for the events $\mathcal{E}_1$ and $\mathcal{E}_2$ defined in equations (D.22a) and (D.22b). Multiplying both sides of inequality (D.29) by $\eta_{x,i}^2$ and summing over the index i, we obtain that with probability at least $1-\epsilon/4$, we have

$$\begin{aligned}
\sum_{i=1}^n \eta_{x,i}^2 s_{x,i}^2 \left|\frac{1}{2}(z-x)^\top \nabla^2\varphi_{y,i}(z-x)\right| &\le \left[14\sum_{i=1}^n (\sigma_{y,i}+\beta_{\mathrm{V}})\frac{s_{x,i}^4}{s_{y,i}^4}\eta_{x,i}^4 + 11\sum_{i,j=1}^n \sigma_{y,i,j}^2\eta_{x,i}^2\eta_{x,j}^2\frac{s_{x,i}^2 s_{x,j}^2}{s_{y,i}^2 s_{y,j}^2}\right] \\
&\overset{(\text{hpb.(D.22a)})}{\le} \tau^4\left[14\sum_{i=1}^n (\sigma_{y,i}+\beta_{\mathrm{V}})\eta_{x,i}^4 + 11\sum_{i,j=1}^n \sigma_{y,i,j}^2\eta_{x,i}^2\eta_{x,j}^2\right] \\
&\overset{(\text{AM}-\text{GM})}{\le} \tau^4\left[14\sum_{i=1}^n (\sigma_{y,i}+\beta_{\mathrm{V}})\eta_{x,i}^4 + \frac{11}{2}\sum_{i,j=1}^n \sigma_{y,i,j}^2(\eta_{x,i}^4+\eta_{x,j}^4)\right] \\
&\overset{(\text{Lem. } E.5(a))}{\le} 25\tau^4\sum_{i=1}^n (\sigma_{y,i}+\beta_{\mathrm{V}})\eta_{x,i}^4 \\
&\overset{(\text{hpb.(D.22b)})}{\le} 50\sum_{i=1}^n (\sigma_{x,i}+\beta_{\mathrm{V}})\eta_{x,i}^4,
\end{aligned} \tag{D.30}$$

where "hpb" stands for high probability bound for events $\mathcal{E}_1$ and $\mathcal{E}_2$. In the last step, we have used the fact that $\tau^6/(2-\tau)^2 \leq 2$ for $\tau = 1.05$. Combining equations (D.27), (D.28) and (D.30) and noting that $\eta_{x,i} = r\hat{a}_i^\top \xi/(nd)^{1/4}$, we find that

$$\begin{aligned}\left|\|z-x\|_z^2 - \|z-x\|_x^2\right| &\leq \frac{14}{3}\left|\sum_{i=1}^n (\sigma_{x,i}+\beta_{\mathrm{V}})\,\eta_{x,i}^3\right| + \frac{8}{3}\left|\sum_{i,j=1}^n \sigma_{x,i,j}^2\left((\eta_{x,i}+\eta_{x,j})/2\right)^3\right| + 38\sum_{i=1}^n \sigma_{x,i}\eta_{x,i}^4 \\ &\leq \frac{14}{3}\frac{r^3}{(nd)^{3/4}}\left|\sum_{i=1}^n (\sigma_{x,i}+\beta_{\mathrm{V}})\left(\hat{a}_i^\top\xi\right)^3\right| + \frac{8}{3}\frac{r^3}{(nd)^{3/4}}\left|\sum_{i,j=1}^n \sigma_{x,i,j}^2\left(\frac{1}{2}(\hat{a}_i+\hat{a}_j)^\top\xi\right)^3\right| \\ &\quad + 50\,\frac{r^4}{nd}\sum_{i=1}^n (\sigma_{x,i}+\beta_{\mathrm{V}})\,(\hat{a}_i^\top\xi)^4, \end{aligned} \tag{D.31}$$

where the last step follows from the fact that $0 \leq \sigma_{x,i} \leq \sigma_{x,i} + \beta_{\mathrm{V}}$. In order to show that $\left|\|z-x\|_z^2 - \|z-x\|_x^2\right|$ is bounded as $\mathcal{O}\left(1/\sqrt{nd}\right)$ with high probability, it suffices to show that with high probability, the third and fourth degree polynomials of $\hat{a}_i^\top\xi$, that appear in bound (D.31), are bounded by $\mathcal{O}\left((nd)^{1/4}\right)$ and $\mathcal{O}\left(\sqrt{nd}\right)$ respectively.

Applying the bounds (D.17b), (D.17c) and (D.17d) from Lemma D.7, we have with probability at least $1-\epsilon$,

$$\|z-x\|_z^2 - \|z-x\|_x^2 \leq \frac{r^3}{\sqrt{nd}}\left(\frac{22\sqrt{15}\gamma_3}{3}\right) + \frac{r^4}{\sqrt{nd}}\left(50\sqrt{105}\gamma_4\right).$$

Using the condition

$$r \leq \min\left\{\frac{\epsilon}{22\sqrt{5/3}\gamma_3}, \sqrt{\frac{\epsilon}{50\sqrt{105}\gamma_4}}\right\}, \tag{D.32}$$

completes our proof of claim (D.9b).

D.3.5 Proof of Lemma D.5

We now derive the different expressions for derivatives and prove the bounds for Hessians of $x \mapsto \varphi_{x,i}$, $i \in [n]$ and $x \mapsto \Psi_x$. In this section we use the simpler notation $H_x := \nabla^2\mathcal{F}_x$.

D.3.5.1 Gradient of σ

Using $s_{x+h,i} = (b_i - a_i^\top(x+h)) = s_{x,i} - a_i^\top h$, we define the Hessian difference matrix

$$\Delta_{x,h}^H := H_{x+h} - H_x = \sum_{i=1}^n a_i a_i^\top\left(\frac{1}{(s_{x,i}-a_i^\top h)^2} - \frac{1}{s_{x,i}^2}\right). \tag{D.33}$$

Up to second order terms, we have

$$\frac{1}{s_{x+h,i}^2} = \frac{1}{s_{x,i}^2}\left[1+\frac{2a_i^\top h}{s_{x,i}}+\frac{3(a_i^\top h)^2}{s_{x,i}^2}\right]+\mathcal{O}\left(\|h\|_2^3\right), \tag{D.34a}$$

$$\Delta_{x,h}^H = \sum_{i=1}^n \frac{a_i a_i^\top}{s_{x,i}^2}\left[\frac{2a_i^\top h}{s_{x,i}}+\frac{3(a_i^\top h)^2}{s_{x,i}^2}\right]+\mathcal{O}\left(\|h\|_2^3\right), \tag{D.34b}$$

$$a_i^T H_{x+h}^{-1} a_i = a_i^\top H_x^{-1} a_i - a_i^\top H_x^{-1}\Delta_{x,h}^H H_x^{-1} a_i + a_i^\top H_x^{-1}\Delta_{x,h}^H H_x^{-1}\Delta_{x,h}^H H_x^{-1} a_i + \mathcal{O}\left(\|h\|_2^3\right). \tag{D.34c}$$

Collecting different first order terms in $\sigma_{x+h,i}-\sigma_{x,i}$, we obtain

$$\begin{aligned}\sigma_{x+h,i}-\sigma_{x,i} &= 2\,\frac{a_i^\top H_x^{-1} a_i}{s_{x,i}^2}\frac{a_i^\top h}{s_{x,i}} - 2\,\frac{a_i^\top H_x^{-1}\left(\sum_{j=1}^n \frac{a_j a_j^\top}{s_{x,j}^2}\frac{a_j^\top h}{s_{x,j}}\right)H_x^{-1} a_i}{s_{x,i}^2} + \mathcal{O}\left(\|h\|_2^2\right) \\ &= 2\left[\sigma_{x,i}\frac{a_i^\top h}{s_{x,i}} - \sum_{j=1}^n \sigma_{x,i,j}^2\,\frac{a_j^\top h}{s_{x,j}}\right] + \mathcal{O}\left(\|h\|_2^2\right) \\ &= 2\left[(\Sigma_x - \Upsilon_x^{(2)})S_x^{-1}A\right]_i h + \mathcal{O}\left(\|h\|_2^2\right).\end{aligned}$$

Dividing both sides by h and letting $h \to 0$ yields the claim.

D.3.5.2 Gradient of φ

Using the chain rule and the fact that $\nabla s_{x,i} = -a_i$, we find that

$$\begin{aligned}\nabla\varphi_{x,i} &= \frac{\nabla\sigma_{x,i}}{s_{x,i}^2} - 2\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\frac{\nabla s_{x,i}}{s_{x,i}^3} \\ &= \frac{2}{s_{x,i}^2}A^\top S_x^{-1}\left[2\Sigma_x + \beta_{\mathrm{V}}\,\mathbb{I} - \Upsilon_x^{(2)}\right]e_i,\end{aligned}$$

as claimed.

D.3.5.3 Gradient of Ψ

For convenience, let us restate equations (D.16) and (D.44):

$$\hat{a}_i = \frac{1}{s_{x,i}}V_x^{-1/2}a_i, \quad \text{and} \quad \sum_{i=1}^n \left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\hat{a}_i\hat{a}_i^\top = \mathbb{I}_d.$$

For a unit vector h, we have

$$h^\top\nabla\log\det V_x = \lim_{\delta\to 0}\frac{1}{\delta}\left[\operatorname{trace}\log\left(\sum_{i=1}^n \frac{\left(\sigma_{x+\delta h,i}+\beta_{\mathrm{V}}\right)}{\left(1-\delta a_i^\top h/s_{x,i}\right)^2}\hat{a}_i\hat{a}_i^\top\right) - \operatorname{trace}\log\left(\sum_{i=1}^n \left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\hat{a}_i\hat{a}_i^\top\right)\right]. \tag{D.35}$$

Let $\log L$ denote the logarithm of the matrix L. Keeping track of the first order terms on RHS of equation (D.35), we find that

$$\begin{aligned}
&\operatorname{trace}\left[\log\left(\sum_{i=1}^{n}\left(\sigma_{x+\delta h,i}+\beta_{\mathrm{V}}\right)\frac{\hat{a}_i\hat{a}_i^\top}{\left(1-\delta a_i^\top h/s_{x,i}\right)^2}\right)\right]-\operatorname{trace}\left[\log\left(\sum_{i=1}^{n}\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\hat{a}_i\hat{a}_i^\top\right)\right]\\
&=\operatorname{trace}\left[\log\left(\sum_{i=1}^{n}\left(\sigma_{x+\delta h,i}+\beta_{\mathrm{V}}+\delta h^\top\nabla\sigma_{x,i}\right)\left(1+2\delta\frac{a_i^\top h}{s_{x,i}^2}\right)\right)\right]-\operatorname{trace}\left[\log\left(\sum_{i=1}^{n}\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\hat{a}_i\hat{a}_i^\top\right)\right]+\mathcal{O}\,(\\
&=\operatorname{trace}\left[\sum_{i=1}^{n}\delta\left(2\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\frac{a_i^\top h}{s_{x,i}^2}+h^\top\nabla\sigma_{x,i}\right)\hat{a}_i\hat{a}_i^\top\right]+\mathcal{O}\left(\delta^2\right)\\
&=\delta\left(\sum_{i=1}^{n}\left(2\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\frac{a_i^\top h}{s_{x,i}^2}+h^\top\nabla\sigma_{x,i}\right)\theta_i\right)+\mathcal{O}\left(\delta^2\right),
\end{aligned}$$

where we have used the fact $\operatorname{trace}(\log\mathbb{I})=0$. Letting $\delta\to 0$ and substituting expression of $h^\top\nabla\sigma_x$ from part (a), we obtain

$$h^\top\nabla\log\det V_x=A_x^\top\left(4\Sigma_x+2\beta_{\mathrm{V}}\mathbb{I}-2\Upsilon_x^{(2)}\right)\Theta_x h.$$

D.3.5.4 Bound on Hessian $\nabla^2\varphi$

In terms of the shorthand $E_{ii}=e_ie_i^\top$, we claim that for any $h\in\mathbb{R}^d$,

$$\begin{aligned}
h^\top\nabla^2\varphi_{x,i}h=\frac{2}{s_{x,i}^2}h^\top A_x^\top\Big[&E_{ii}\left(3\left(\Sigma_x+\beta_{\mathrm{V}}\mathbb{I}\right)+7\Sigma_x-8\operatorname{diag}(\Upsilon_x^{(2)}e_i)\right)E_{ii}\\
&+\operatorname{diag}(\Upsilon_xe_i)(4\Upsilon_x-3\mathbb{I})\operatorname{diag}(\Upsilon_xe_i)\Big]A_xh.
\end{aligned}\tag{D.36}$$

Note that

$$\varphi_{x+h,i}-\varphi_{x,i}=\underbrace{\left(\frac{a_i^\top H_{x+h,i}^{-1}a_i}{s_{x+h,i}^4}-\frac{a_i^\top H_{x,i}^{-1}a_i}{s_{x,i}^4}\right)}_{=:A_1}+\underbrace{\beta_{\mathrm{V}}\left(\frac{1}{s_{x+h,i}^2}-\frac{1}{s_{x,i}^2}\right)}_{=:A_2}.\tag{D.37}$$

The second order Taylor expansion of $1/s_{x,i}^4$ is given by

$$\frac{1}{s_{x+h,i}^4}=\frac{1}{s_{x,i}^4}\left[1+\frac{4a_i^\top h}{s_{x,i}}+\frac{10(a_i^\top h)^2}{s_{x,i}^2}\right]+\mathcal{O}\left(\|h\|_2^3\right).$$

Let B_1 and B_2 denote the second order terms, i.e., the terms that are of order $\mathcal{O}\left(\|h\|_2^2\right)$, in Taylor expansion of A_1 and A_2 around x, respectively. Borrowing terms from equations (D.34a)-(D.34c) and simplifying we obtain

$$B_1 = 10\sigma_{x,i}\frac{(a_i^\top h)^2}{s_{x,i}^2} - 8\frac{a_i^\top h}{s_{x,i}}\sum_{j=1}^n \frac{\sigma_{x,i,j}^2}{s_{x,i}^2}\frac{a_j^\top h}{s_{x,j}} - 3\sum_{j=1}^n \frac{\sigma_{x,i,j}^2}{s_{x,i}^2}\frac{(a_j^\top h)^2}{s_{x,j}^2} + 4\sum_{j=1}^n\sum_{l=1}^n \frac{\sigma_{x,i,j}}{s_{x,i}}\sigma_{x,j,l}\frac{\sigma_{x,l,i}}{s_{x,i}}\frac{a_j^\top h}{s_{x,j}}\frac{a_l^\top h}{s_{x,l}},$$

and $B_2 = 3\beta_{\mathrm{V}}\dfrac{(a_i^\top h)^2}{s_{x,i}^2}$.

Observing that the second order term in the Taylor expansion of $\varphi_{x+h,i}$ around x, is exactly $\frac{1}{2}h^\top \nabla^2 \varphi_{x,i} h$ yields the claim (D.36). We now turn to prove the bound on the directional Hessian. Recall $\eta_{x,i} = a_i^\top h / s_{x,i}$. We have

$$\begin{aligned}
&s_{y,i}^2 \left|\frac{1}{2}h^\top \nabla^2 \varphi_{x,i} h\right| \\
&\quad = \left|3\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\eta_{x,i}^2 + 7\sigma_{x,i}\eta_{x,i}^2 - 8\sum_{j=1}^n \sigma_{x,i,j}^2\eta_{x,j}\eta_{x,i} - 3\sum_{j=1}^n \sigma_{x,i,j}^2\eta_{x,j}^2 + 4\sum_{j,k=1}^n \sigma_{x,i,j}\sigma_{x,j,k}\sigma_{x,k,i}\eta_{x,j}\eta_{x,k}\right| \\
&\quad \overset{(i)}{\leq} 10\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\eta_{x,i}^2 + 8\sum_{j=1}^n \sigma_{x,i,j}^2\left|\eta_{x,i}\eta_{x,j}\right| + 7\sum_{j=1}^n \sigma_{x,i,j}^2\eta_{x,j}^2 \\
&\quad \overset{(ii)}{\leq} 10\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\eta_{x,i}^2 + 4\sum_{j=1}^n \sigma_{x,i,j}^2\left(\eta_{x,i}^2+\eta_{x,j}^2\right) + 7\sum_{j=1}^n \sigma_{x,i,j}^2\eta_{x,j}^2 \\
&\quad \overset{(iii)}{\leq} 10\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\eta_{x,i}^2 + 4\sum_{j=1}^n \sigma_{x,i}\eta_{x,i}^2 + 4\sum_{j=1}^n \sigma_{x,i,j}^2\eta_{x,j}^2 + 7\sum_{j=1}^n \sigma_{x,i,j}^2\eta_{x,j}^2, \\
&\quad \overset{(iv)}{\leq} 14\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\eta_{x,i}^2 + 11\sum_{j=1}^n \sigma_{x,i,j}^2\eta_{x,j}^2,
\end{aligned}$$

where step (i) follows from the fact that $\operatorname{diag}(\Upsilon_y e_i)\Upsilon_y \operatorname{diag}(\Upsilon_y e_i) \preceq \operatorname{diag}(\Upsilon_y e_i)\operatorname{diag}(\Upsilon_y e_i)$ since Υ_y is an orthogonal projection matrix; step (ii) follows from AM-GM inequality; step (iii) follows from the symmetry of indices i and j and Lemma E.5(a), and step (iv) from the fact that $\sigma_{x,i} \leq \sigma_{x,i} + \beta_{\mathrm{V}}$.

D.3.5.5 Bound on Hessian $\nabla^2\Psi$

We have

$$\begin{aligned}\frac{1}{2}h^\top\left(\nabla^2\log\det V_x\right)h =& \frac{1}{2}\lim_{\delta\to 0}\frac{1}{\delta^2}\Bigg[\operatorname{trace}\log\left(\sum_{i=1}^n\frac{(\sigma_{x+\delta h,i}+\beta_{\mathrm{V}})}{\left(1-\delta a_i^\top h/s_{x,i}\right)^2}\hat{a}_i\hat{a}_i^\top\right)\\ &+\operatorname{trace}\log\left(\sum_{i=1}^n\frac{(\sigma_{x-\delta h,i}+\beta_{\mathrm{V}})}{\left(1+\delta a_i^\top h/s_{x,i}\right)^2}\hat{a}_i\hat{a}_i^\top\right)\\ &-2\operatorname{trace}\log\left(\sum_{i=1}^n\left(\sigma_x+\beta_{\mathrm{V}}\right)\hat{a}_i\hat{a}_i^\top\right)\Bigg].\end{aligned} \tag{D.38}$$

Up to second order terms, we havea

$$\begin{aligned}&\operatorname{trace}\left[\log\left(\sum_{i=1}^n\left(\sigma_{x+\delta h,i}+\beta_{\mathrm{V}}\right)\frac{\hat{a}_i\hat{a}_i^\top}{\left(1-\delta a_i^\top h/s_{x,i}\right)^2}\right)\right]\\ &=\operatorname{trace}\left[\log\left(\sum_{i=1}^n\left(\sigma_{x,i}+\beta_{\mathrm{V}}+\delta h^\top\nabla\sigma_{x,i}+\frac{1}{2}\delta^2h^\top\nabla^2\sigma_{x,i}h\right)\left(1+2\delta\frac{a_i^\top h}{s_{x,i}}+3\delta^2\left(\frac{a_i^\top h}{s_{x,i}}\right)^2\right)\hat{a}_i\hat{a}_i^\top\right)\right]\\ &=\operatorname{trace}\left[\sum_{i=1}^n\left(\sigma_{x,i}+\beta_{\mathrm{V}}+\delta h^\top\nabla\sigma_{x,i}+\frac{1}{2}\delta^2h^\top\nabla^2\sigma_{x,i}h\right)\left(1+2\delta\frac{a_i^\top h}{s_{x,i}}+3\delta^2\left(\frac{a_i^\top h}{s_{x,i}}\right)^2\right)\hat{a}_i\hat{a}_i^\top\right]\\ &\quad-\operatorname{trace}\left[\frac{1}{2}\left(\sum_{i=1}^n\left(\sigma_{x,i}+\beta_{\mathrm{V}}+\delta h^\top\nabla\sigma_{x,i}+\frac{1}{2}\delta^2h^\top\nabla^2\sigma_{x,i}h\right)\left(1+2\delta\frac{a_i^\top h}{s_{x,i}}+3\delta^2\left(\frac{a_i^\top h}{s_{x,i}}\right)^2\right)\hat{a}_i\hat{a}_i^\top\right)^2\right].\end{aligned}$$

We can similarly obtain the second order expansion of the term $\operatorname{trace}\log\left(\sum_{i=1}^n\frac{(\sigma_{x-\delta h,i}+\beta_{\mathrm{V}})}{\left(1+\delta a_i^\top h/s_{x,i}\right)^2}\hat{a}_i\hat{a}_i^\top\right)$. Recall $\eta_{x,i}=\frac{a_i^\top h}{s_{x,i}}$. Using part (a) to substitute $h^\top\nabla\sigma_{x,i}$, we obtain

$$\begin{aligned}\frac{1}{2}h^\top\left(\nabla^2\log\det V_x\right)h =& \sum_{i=1}^n\left(3\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\eta_{x,i}^2+4\left(\sigma_{x,i}\eta_{x,i}^2-\sum_{j=1}^n\sigma_{x,i,j}^2\eta_{x,i}\eta_{x,j}\right)+\frac{1}{2}h^\top\nabla^2\sigma_{x,i}h\right)\theta_i\\ &-2\Bigg[\sum_{i,j=1}^n\left(2\sigma_{x,i}+\beta_{\mathrm{V}}\right)\left(2\sigma_{x,j}+\beta_{\mathrm{V}}\right)\eta_{x,i}\eta_{x,j}\theta_{x,i,j}^2-2\sum_{i,j,k=1}^n\left(2\sigma_{x,i}+\beta_{\mathrm{V}}\right)\sigma_{x,j,k}^2\theta_{x,i,k}^2\eta_{x,i}\eta_{x,j}\\ &+\sum_{i,j,k,l=1}^n\sigma_{x,i,l}^2\sigma_{x,j,k}^2\theta_{x,k,l}^2\eta_{x,i}\eta_{x,j}\Bigg].\end{aligned} \tag{D.39}$$

We claim that the directional Hessian $h^\top\nabla^2\sigma_{x,i}h$ is given by

$$h^\top\nabla^2\sigma_{x,i}h = 2\,h^\top A_x^\top\left[E_{ii}(3\Sigma_x-4\operatorname{diag}(\Upsilon_x^{(2)}e_i))E_{ii}+\operatorname{diag}(\Upsilon_xe_i)(4\Upsilon_x-3\mathbb{I})\operatorname{diag}(\Upsilon_xe_i)\right]A_xh. \tag{D.40}$$

Assuming the claim at the moment we now bound $\left|h^\top \nabla^2 \Psi_x h\right|$. To shorten the notation, we drop the x-dependence of the terms $\sigma_{x,i}, \sigma_{x,i,j}, \theta_{x,i}$ and $\eta_{x,i}$. Since Υ_x is an orthogonal projection matrix, we have

$$\operatorname{diag}(\Upsilon_x e_i)\Upsilon_x \operatorname{diag}(\Upsilon_x e_i) \preceq \operatorname{diag}(\Upsilon_x e_i)\operatorname{diag}(\Upsilon_x e_i).$$

Using this fact and substituting the expression for $h^\top \nabla^2 \sigma_{x,i} h$ from equation (D.40) in equation (D.39), we obtain

$$\begin{aligned}
&\left|h^\top \nabla^2 \Psi_x h\right| \\
&\leq \sum_{i=1}^n \left[3\Big(\sigma_i + \beta_{\mathrm{V}}\Big)\eta_i^2 + 4\big(\sigma_i\eta_i^2 + \sum_{j=1}^n \sigma_{i,j}^2\eta_i\eta_j\big) + 3\sigma_i\eta_i^2 + 4\sum_{j=1}^n \sigma_{i,j}^2\eta_i\eta_j + 7\sum_{j=1}^n \sigma_{i,j}^2\eta_j^2\right]\theta_i \\
&\qquad + \left[8\sum_{i,j=1}^n (\sigma_i + \beta_{\mathrm{V}})(\sigma_j + \beta_{\mathrm{V}})\,\eta_i\eta_j\theta_{i,j}^2 + 8\sum_{i,j,k=1}^n (\sigma_i + \beta_{\mathrm{V}})\,\sigma_{j,k}^2\theta_{i,k}^2\eta_i\eta_j + 2\sum_{i,j,k,l=1}^n \sigma_{i,l}^2\sigma_{j,k}^2\theta_{k,l}^2\eta_i\eta_j\right].
\end{aligned}$$

Rearranging terms, we find that

$$\begin{aligned}
&\left|h^\top \nabla^2 \Psi_x h\right| \\
&\leq \sum_{i=1}^n \left[10\,(\sigma_i + \beta_{\mathrm{V}})\,\eta_i^2 + 8\sum_{j=1}^n \sigma_{i,j}^2\eta_i\eta_j + 7\sum_{j=1}^n \sigma_{i,j}^2\eta_j^2\right]\theta_i \\
&\qquad + \left[8\sum_{i,j=1}^n (\sigma_i + \beta_{\mathrm{V}})(\sigma_j + \beta_{\mathrm{V}})\,\eta_i\eta_j\theta_{i,j}^2 + 8\sum_{i,j,k=1}^n (\sigma_i + \beta_{\mathrm{V}})\,\sigma_{j,k}^2\theta_{i,k}^2\eta_i\eta_j + 2\sum_{i,j,k,l=1}^n \sigma_{i,l}^2\sigma_{j,k}^2\theta_{k,l}^2\eta_i\eta_j\right] \\
&\overset{(i)}{\leq} \sum_{i=1}^n \left[10\,(\sigma_i + \beta_{\mathrm{V}})\,\eta_i^2 + 4\sum_{j=1}^n \sigma_{i,j}^2\left(\eta_i^2 + \eta_j^2\right) + 7\sum_{j=1}^n \sigma_{i,j}^2\eta_j^2\right]\theta_i \\
&\qquad + \left[4\sum_{i,j=1}^n \big(\sigma_i + \beta_{\mathrm{V}}\big)\big(\sigma_j + \beta_{\mathrm{V}}\big)\theta_{i,j}^2(\eta_i^2 + \eta_j^2) + 4\sum_{i,j,k=1}^n \big(\sigma_i + \beta_{\mathrm{V}}\big)\sigma_{j,k}^2\theta_{i,k}^2(\eta_i^2 + \eta_j^2) + \sum_{i,j,k,l=1}^n \sigma_{i,l}^2\sigma_{j,k}^2\theta_{k,l}^2(\eta_i^2 + \eta_j^2)\right]
\end{aligned}$$

where in step (i) we have used the AM-GM inequality. Simplifying further, we obtain

$$\begin{aligned}
\left|h^\top \nabla^2 \Psi_y h\right| &\leq \sum_{i=1}^n \left[14\,(\sigma_i + \beta_{\mathrm{V}})\,\eta_i^2 + 11\sum_{j=1}^n \sigma_{i,j}^2\eta_j^2\right]\theta_i + \left[\sum_{i=1}^n 12\,(\sigma_i + \beta_{\mathrm{V}})\,\theta_i\eta_i^2 + \sum_{i,j=1}^n 6\sigma_{i,j}^2\theta_i\eta_j^2\right] \\
&= 26\sum_{i=1}^n (\sigma_i + \beta_{\mathrm{V}})\,\theta_i\eta_i^2 + 17\sum_{i,j=1}^n \sigma_{i,j}^2\theta_i\eta_j^2.
\end{aligned}$$

Dividing both sides by two completes the proof.

Proof of claim (D.40): In order to compute the directional Hessian of $x \mapsto \sigma_{x,i}$, we need to track the second order terms in equations (D.34a)-(D.34c). Collecting the second order terms (denoted by $\sigma_h^{(2)}$) in the expansion of $\sigma_{x+h,i} - \sigma_{x,i}$, we obtain

$$\begin{aligned}\sigma_h^{(2)} = 3\,\frac{a_i^\top H_x^{-1} a_i}{s_{x,i}^2}\,\frac{(a_i^\top h)^2}{s_{x,i}^2} &- 4\,\frac{a_i^\top H_x^{-1}\left(\sum_{j=1}^n \frac{a_j a_j^\top}{s_{x,j}^2}\frac{a_j^\top h}{s_{x,j}}\right) H_x^{-1} a_i}{s_{x,i}^2}\,\frac{a_i^\top h}{s_{x,i}} \\ &- 3\,\frac{a_i^\top H_x^{-1}\left(\sum_{j=1}^n \frac{a_j a_j^\top}{s_{x,j}^2}\frac{(a_j^\top h)^2}{s_{x,j}^2}\right) H_x^{-1} a_i}{s_{x,i}^2} \\ &+ 4\,\frac{a_i^\top H_x^{-1}\left(\sum_{j=1}^n \frac{a_j a_j^\top}{s_{x,j}^2}\frac{a_j^\top h}{s_{x,j}}\right) H_x^{-1}\left(\sum_{l=1}^n \frac{a_l a_l^\top}{s_{x,l}^2}\frac{a_l^\top h}{s_{x,l}}\right) a_i}{s_{x,i}^2}.\end{aligned}$$

We simply each term on the RHS one by one. Simplifying the first term, we obtain

$$3\,\frac{a_i^\top H_x^{-1} a_i}{s_{x,i}^2}\,\frac{(a_i^\top h)^2}{s_{x,i}^2} = 3\,\sigma_{x,i}\eta_{x,i}^2 = h^\top 3\,A_x^\top E_{ii}\Sigma_x E_{ii} A_x\,h.$$

For the second term, we have

$$\begin{aligned}4\,\frac{a_i^\top H_x^{-1}\left(\sum_{j=1}^n \frac{a_j a_j^\top}{s_{x,j}^2}\frac{a_j^\top h}{s_{x,j}}\right) H_x^{-1} a_i}{s_{x,i}^2}\,\frac{a_i^\top h}{s_{x,i}} &= 4\,\eta_{x,i}\sum_{j=1}^n \sigma_{x,i,j}^2\,\eta_{x,j} \\ &= 4\,h^\top A_x^\top E_{ii}\operatorname{diag}\left(\Upsilon_x^{(2)} e_i\right) E_{ii} A_x h.\end{aligned}$$

The third term can be simplified as follows:

$$\begin{aligned}3\,\frac{a_i^\top H_x^{-1}\left(\sum_{j=1}^n \frac{a_j a_j^\top}{s_{x,j}^2}\frac{(a_j^\top h)^2}{s_{x,j}^2}\right) H_x^{-1} a_i}{s_{x,i}^2} &= 3\sum_{j=1}^n \sigma_{x,i,j}^2\eta_{x,j}^2 \\ &= 3\,h^\top A_x^\top \operatorname{diag}\left(\Upsilon_x e_i\right)\operatorname{diag}\left(\Upsilon_x e_i\right) A_x h\end{aligned}$$

For the last term, we find that

$$\begin{aligned}4\,\frac{a_i^\top H_x^{-1}\left(\sum_{j=1}^n \frac{a_j a_j^\top}{s_{x,j}^2}\frac{a_j^\top h}{s_{x,j}}\right) H_x^{-1}\left(\sum_{l=1}^n \frac{a_l a_l^\top}{s_{x,l}^2}\frac{a_l^\top h}{s_{x,l}}\right) a_i}{s_{x,i}^2} &= 4\sum_{j,l=1}^n \sigma_{x,i,j}\,\sigma_{x,j,l}\,\sigma_{x,l,i}\,\eta_{x,j}\,\eta_{x,l} \\ &= 4\,h^\top A_x^\top \operatorname{diag}\left(\Upsilon_x e_i\right)\Upsilon_x\operatorname{diag}\left(\Upsilon_x e_i\right) A_x h.\end{aligned}$$

Putting together the pieces yields the expression (D.40).

D.3.6 Proof of Lemma D.7

The proof relies on the classical fact that the tails of a polynomial in Gaussian random variables decay exponentially independently of dimension. In particular, Theorem 6.7 of Janson [129] ensures that for any integers $d, k \geq 1$, any polynomial $f : \mathbb{R}^d \to \mathbb{R}$ of degree k, and any scalar $t \geq (2e)^{k/2}$, we have

$$\mathbb{P}\left[|f(\xi)| \geq t\left(\mathbb{E}f(\xi)^2\right)^{\frac{1}{2}}\right] \leq \exp\left(-\frac{k}{2e}t^{2/k}\right), \tag{D.41}$$

where $\xi \sim \mathcal{N}(0, \mathbb{I}_n)$ denotes a standard Gaussian vector in n dimensions.

Also, the following observations on the behavior of the vectors $\hat{a}_i$ from equation (D.16) are useful:

$$\|\hat{a}_i\|_2^2 = \theta_{x,i} \leq \sqrt{\frac{n}{d}} \quad \text{for all } i \in [n], \quad \text{and} \tag{D.42a}$$

$$(\hat{a}_i^\top \hat{a}_j)^2 = \theta_{x,i,j}^2 \quad \text{for all } i, j \in [n]. \tag{D.42b}$$

D.3.6.1 Proof of bound (D.17a)

We have

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^n (\sigma_{x,i} + \beta_{\mathrm{V}})\left(\hat{a}_i^\top \xi\right)^2\right)^2 &= \sum_{i,j=1}^n (\sigma_{x,i} + \beta_{\mathrm{V}})(\sigma_{x,j} + \beta_{\mathrm{V}})\,\mathbb{E}\left(\hat{a}_i^\top \xi\right)^2\left(\hat{a}_j^\top \xi\right)^2 \\
&= \sum_{i,j=1}^n (\sigma_{x,i} + \beta_{\mathrm{V}})(\sigma_{x,j} + \beta_{\mathrm{V}})\left(\|\hat{a}_i\|_2^2\,\|\hat{a}_j\|_2^2 + 2\left(\hat{a}_i^\top \hat{a}_j\right)^2\right) \\
&= \sum_{i,j=1}^n (\sigma_{x,i} + \beta_{\mathrm{V}})(\sigma_{x,j} + \beta_{\mathrm{V}})\left(\theta_{x,i}\theta_{x,j} + 2\theta_{x,i,j}^2\right) \\
&\overset{(i)}{=} d^2 + 2d \\
&\leq 3d^2,
\end{aligned}$$

where step (i) follows from properties (c) and (d) from Lemma E.5. Applying the bound (D.41) with $k = 2, t = e\log(\frac{4}{\epsilon})$ yields the claim. We verify that for $\epsilon \in (0, 1/15]$, $t \geq 2e$.

D.3.6.2 Proof of bound (D.17b)

Using Isserlis' theorem [128] for Gaussian moments, we obtain

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{n}\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\left(\hat{a}_i^\top\xi\right)^3\right)^2 &= \sum_{i,j=1}^{n}\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\mathbb{E}\left(\hat{a}_i^\top\xi\right)^3\left(\hat{a}_j^\top\xi\right)^3\\ &= 9\underbrace{\sum_{i,j=1}^{n}\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\left(\sigma_{x,j}+\beta_{\mathrm{V}}\right)\|\hat{a}_i\|_2^2\,\|\hat{a}_j\|_2^2\left(\hat{a}_i^\top\hat{a}_j\right)}_{=:N_1}\\ &\quad+6\underbrace{\sum_{i,j=1}^{n}\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\left(\sigma_{x,j}+\beta_{\mathrm{V}}\right)\left(\hat{a}_i^\top\hat{a}_j\right)^3}_{=:N_2}.\end{aligned}\tag{D.43}$$

We claim that the two terms in this sum are bounded as $N_1 \leq \sqrt{nd}$ and $N_2 \leq \sqrt{nd}$. Assuming the claims as given, we now complete the proof. Plugging in the bounds for N_1 and N_2 in equation (D.43) we find that $\mathbb{E}\big(\sum_{i=1}^n\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\left(\hat{a}_i^\top\xi\right)^3\big)^2 \leq 15\sqrt{nd}$. Applying the bound (D.41) with $k=3, t=\left(\frac{2e}{3}\log(4/\epsilon)\right)^{3/2}$ yields the claim. We also verify that for $\epsilon\in(0,1/15]$, $t\geq(2e)^{3/2}$. We now turn to proving the bounds on N_1 and N_2.

Bounding N_1: Let B be an $n\times d$ matrix with its i-th row given by $\sqrt{(\sigma_{x,i}+\beta_{\mathrm{V}})}\hat{a}_i^\top$. Observe that

$$\sum_{i=1}^{n}\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\hat{a}_i\hat{a}_i^\top = V_x^{-1/2}\left(\sum_{i=1}^{n}\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\frac{a_ia_i^\top}{s_{x,i}^2}\right)V_x^{-1/2} = V_x^{-1/2}V_xV_x^{-1/2} = \mathbb{I}_d. \tag{D.44}$$

Thus we have $B^\top B = \mathbb{I}_d$, which implies that $BB^\top$ is an orthogonal projection matrix. Letting $v\in\mathbb{R}^n$ be a vector such that $v_i = \sqrt{(\sigma_{x,i}+\beta_{\mathrm{V}})}\,\|\hat{a}_i\|_2^2$, we then have

$$\sum_{i,j=1}^{n}(\sigma_{x,i}+\beta_{\mathrm{V}})\|\hat{a}_i\|_2^2\,\hat{a}_i^\top\,(\sigma_{x,j}+\beta_{\mathrm{V}})\|\hat{a}_j\|_2^2\,\hat{a}_j = \left\|\sum_{i=1}^{n}(\sigma_{x,i}+\beta_{\mathrm{V}})\|\hat{a}_i\|_2^2\,\hat{a}_i\right\|_2^2 = \left\|B^\top v\right\|_2^2 \overset{(i)}{\leq} \|v\|_2^2,$$

where inequality (i) follows from the fact that $v^\top Pv \leq \|v\|_2^2$ for any orthogonal projection matrix P. Equation (D.42a) implies that $v_i^2 = \left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\theta_{x,i}^2$. Using Lemma E.5(e), we find that

$$\|v\|_2^2 = \sum_{i=1}^{n}\left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\theta_{x,i}^2 \leq \sqrt{nd}.$$

Bounding N_2: We see that

$$\begin{aligned}\sum_{i,j=1}^n \left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\left(\sigma_{x,j}+\beta_{\mathrm{V}}\right)\left(\hat{a}_i^\top \hat{a}_j\right)^3 &\overset{\text{(C−S)}}{\leq} \sum_{i,j=1}^n \left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\left(\sigma_{x,j}+\beta_{\mathrm{V}}\right)\left(\hat{a}_i^\top \hat{a}_j\right)^2 \left\|\hat{a}_i\right\|_2 \left\|\hat{a}_j\right\|_2 \\ &\overset{\text{(eqns.(D.42a),(D.42b))}}{\leq} \sum_{i,j=1}^n \left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\left(\sigma_{x,j}+\beta_{\mathrm{V}}\right)\theta_{x,i,j}^2 \sqrt{\theta_{x,i}\theta_{x,j}} \\ &\overset{\text{(Lem. 6.1}(j))}{\leq} \sqrt{\frac{n}{d}} \sum_{i,j=1}^n \left(\sigma_{x,i}+\beta_{\mathrm{V}}\right)\left(\sigma_{x,j}+\beta_{\mathrm{V}}\right)\theta_{x,i,j}^2 .\end{aligned}$$

We now apply Lemma E.5(d) followed by Lemma E.5(c) to obtain the claimed bound on N_2.

D.3.6.3 Proof of bound (D.17c)

Let $c_{i,j} = \dfrac{(\hat{a}_i + \hat{a}_j)}{2}$ for $i, j \in [n]$. Using Isserlis' theorem for Gaussian moments, we obtain

$$\begin{aligned}\mathbb{E}\left(\sum_{i,j=1}^n \sigma_{x,i,j}^2 \left(c_{i,j}^\top \xi\right)^3\right)^2 &= \sum_{i,j,k,l=1}^n \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \mathbb{E}\left(c_{i,j}^\top \xi\right)^3 \left(c_{k,l}^\top \xi\right)^3 \\ &= 9 \underbrace{\sum_{i,j,k,l=1}^n \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \|c_{i,j}\|_2^2 \left\|c_{k,l}\right\|_2^2 \left(c_{i,j}^\top c_{k,l}\right)}_{=:\, C_1} + 6 \underbrace{\sum_{i,j,k,l=1}^n \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \left(c_{i,j}^\top c_{k,l}\right)^3}_{=:\, C_2}\end{aligned}$$

We claim that $C_1 \leq \sqrt{nd}$ and $C_2 \leq \sqrt{nd}$. Assuming the claims as given, the result follows using similar arguments as in the previous part. We now bound $C_i, i = 1, 2$, using arguments similar to the ones used in Section D.3.6.2 to bound $N_i, i = 1, 2$, respectively. The following bounds on $\|c_{i,j}\|_2^2$ are used in the arguments that follow:

$$\|c_{i,j}\|_2^2 \overset{\text{SSI}}{\leq} \frac{1}{2}\left(\|\hat{a}_i\|_2^2 + \|\hat{a}_j\|_2^2\right) = \frac{1}{2}\left(\theta_{x,i} + \theta_{x,j}\right) \tag{D.45a}$$

$$\overset{\text{Lem. 6.1}(j)}{\leq} \sqrt{\frac{n}{d}}. \tag{D.45b}$$

Bounding C_1: Let B be the same $n \times d$ matrix as in the proof of previous part with its i-th row given by $\sqrt{(\sigma_{x,i} + \beta_{\mathrm{V}})}\hat{a}_i^\top$. Define the vector $u \in \mathbb{R}^d$ with entries given by

$u_i = \sum_{j=1}^n \sigma_{x,i,j}^2 \left\|c_{i,j}\right\|_2^2 / (\sigma_{x,i} + \beta_{\mathrm{V}})^{1/2}$. We have

$$\begin{aligned}
\sum_{i,j,k,l=1}^n \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \left\|c_{i,j}\right\|_2^2 \left\|c_{k,l}\right\|_2^2 \left(c_{i,j}^\top c_{k,l}\right) &\leq \left\| \sum_{i,j=1}^n \sigma_{x,i,j}^2 \left\|c_{i,j}\right\|_2^2 c_{i,j} \right\|_2^2 \\
&\overset{\text{(SSI)}}{\leq} \frac{1}{2}\left(\left\| \sum_{i,j=1}^n \sigma_{x,i,j}^2 \left\|c_{i,j}\right\|_2^2 \hat{a}_i \right\|_2^2 + \left\| \sum_{i,j=1}^n \sigma_{x,i,j}^2 \left\|c_{i,j}\right\|_2^2 \hat{a}_j \right\|_2^2 \right) \\
&= \left\|B^\top u\right\|_2^2 \\
&\overset{(i)}{\leq} \left\|u\right\|_2^2,
\end{aligned}$$

where inequality (i) follows from the fact that $v^\top P v \leq \|v\|_2^2$ for any orthogonal projection matrix P. It is left to bound the term u_i^2. We see that

$$\begin{aligned}
u_i^2 = \frac{1}{\sigma_{x,i} + \beta_{\mathrm{V}}} \sum_{j,k=1}^n \sigma_{x,i,j}^2 \sigma_{x,i,k}^2 \left\|c_{i,j}\right\|_2^2 \left\|c_{i,k}\right\|_2^2 &\overset{\text{(bnd. (D.45b))}}{\leq} \sqrt{\frac{n}{d}} \frac{1}{\sigma_{x,i} + \beta_{\mathrm{V}}} \sum_{j,k=1}^n \sigma_{x,i,j}^2 \sigma_{x,i,k}^2 \left\|c_{i,j}\right\|_2^2 \\
&\overset{\text{(Lem. }E.5(a))}{\leq} \sqrt{\frac{n}{d}} \frac{\sigma_{x,i}}{\sigma_{x,i} + \beta_{\mathrm{V}}} \sum_{j=1}^n \sigma_{x,i,j}^2 \left\|c_{i,j}\right\|_2^2 \\
&\overset{\text{(bnd. (D.45a))}}{\leq} \sqrt{\frac{n}{d}} \sum_{j=1}^n \sigma_{x,i,j}^2 \frac{\theta_{x,i} + \theta_{x,j}}{2}.
\end{aligned}$$

Now, summing over i and using symmetry of indices i, j, we find that

$$\|u\|_2^2 \leq \sqrt{\frac{n}{d}} \sum_{i=1}^n \sum_{j=1}^n \sigma_{x,i,j}^2 \theta_{x,i} \overset{\text{(Lem. }E.5(a))}{=} \sqrt{\frac{n}{d}} \sum_{i=1}^n \sigma_{x,i} \theta_{x,i} \overset{\text{(Lem. }E.5(c))}{\leq} \sqrt{nd},$$

thereby implying that $C_1 \leq \sqrt{nd}$.

Bounding C_2: Using the Cauchy-Schwarz inequality and the bound (D.45b), we find that

$$\begin{aligned}
\sum_{i,j,k,l=1}^n \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \left(c_{i,j}^\top c_{k,l}\right)^3 &\leq \sum_{i,j,k,l=1}^n \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \left(c_{i,j}^\top c_{k,l}\right)^2 \left\|c_{i,j}\right\|_2 \left\|c_{k,l}\right\|_2 \\
&\leq \sqrt{\frac{n}{d}} \sum_{i,j,k,l=1}^n \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \left(c_{i,j}^\top c_{k,l}\right)^2.
\end{aligned}$$

Using SSI and the symmetry of pairs of indices (i, j) and (k, l), we obtain

$$\sum_{i,j,k,l=1}^n \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \left(c_{i,j}^\top c_{k,l}\right)^2 \leq \sum_{i,j,k,l=1}^n \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \left(\hat{a}_i^\top \hat{a}_k\right)^2 = \sum_{i,k=1}^n \sigma_{x,i} \sigma_{x,k} \left(\hat{a}_i^\top \hat{a}_k\right)^2.$$

The resulting expression can be bounded as follows:

$$\sum_{i,k=1}^{n} \sigma_{x,i}\sigma_{x,k} \left(\hat{a}_i^\top \hat{a}_k\right)^2 \overset{\text{(eqn.(D.42b))}}{=} \sum_{i,k=1}^{n} \sigma_{x,i}\sigma_{x,k}\theta_{x,i,k}^2 \overset{\text{(Lem. } E.5(d))}{\leq} \sum_{i=1}^{n} \sigma_{x,i}\theta_{x,i} \overset{\text{(Lem. } E.5(c))}{\leq} n.$$

Putting the pieces together yields the claimed bound on C_2.

D.3.6.4 Proof of bound (D.17d)

Observe that $\hat{a}_i^\top \xi \sim \mathcal{N}(0, \theta_{x,i})$ and hence $\mathbb{E}\left(\hat{a}_i^\top \xi\right)^8 = 105\,\theta_{x,i}^4$. Thus we have

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n} \sigma_{x,i}\left(\hat{a}_i^\top \xi\right)^4\right)^2 &\overset{\text{C}-\text{S}}{\leq} \sum_{i,j=1}^{n} \sigma_{x,i}\sigma_{x,j}\left(\mathbb{E}\left(\hat{a}_i^\top \xi\right)^8\right)^{\frac{1}{2}}\left(\mathbb{E}\left(\hat{a}_j^\top \xi\right)^8\right)^{\frac{1}{2}} \\
&= 105\sum_{i,j=1}^{n} \sigma_{x,i}\sigma_{x,j}\theta_{x,i}^2\theta_{x,j}^2 \\
&= 105\left(\sum_{i=1}^{n} \sigma_{x,i}\theta_{x,i}^2\right)^2 \\
&\overset{\text{(Lem. } E.5(e))}{\leq} 105nd.
\end{aligned}$$

Applying the bound (D.41) with $k = 4, t = \left(\frac{e}{2}\log(4/\epsilon)\right)^2$ yields the result. We also verify that for $\epsilon \in (0, 1/15]$, we have $t \geq (2e)^2$

D.4 Proof of Lovász's Result: Proposition 6.1

We begin by formally defining the conductance (Φ) of a Markov chain on $(\mathcal{K}, \mathbb{B}(\mathcal{K}))$ with arbitrary transition operator $\mathcal{T}$ and stationary distribution $\Pi^\star$. We assume that the operator $\mathcal{T}$ is lazy and thereby the stationary distribution $\Pi^\star$ is unique. Let $\mathcal{T}_x = \mathcal{T}(\boldsymbol{\delta}_x)$ denote the transition distribution at point x, then the conductance Φ is defined as

$$\Phi := \inf_{\substack{\mathcal{S}\in\mathbb{B}(\mathcal{K}) \\ \Pi^\star(\mathcal{S})\in(0,1/2)}} \frac{\Phi(\mathcal{S})}{\Pi^\star(\mathcal{S})} \quad \text{where} \quad \Phi(\mathcal{S}) := \int_{\mathcal{S}} \mathcal{T}_u(\mathcal{K}\cap\mathcal{S}^c)d\Pi^\star(u) \quad \text{for any } \mathcal{S}\subseteq\mathcal{K}.$$

The conductance denotes the measure of the flow from a set to its complement relative to its own measure, when initialized in the stationary distribution. If the conductance is high, the following result shows that the Markov chain mixes fast.

Lemma D.8. *Theorem 1.4 [167] For any β-warm start μ_0, the mixing time of the Markov chain with conductance Φ is bounded as*

$$\left\|\mathcal{T}^k(\mu_0) - \Pi^\star\right\|_{TV} \leq \sqrt{\beta}\left(1 - \frac{\Phi^2}{2}\right)^k \leq \sqrt{\beta}\exp\left(-k\frac{\Phi^2}{2}\right).$$

Note that this result holds for a general distribution $\Pi^\star$ although we apply for uniform $\Pi^\star$. The result can be derived from Cheeger's inequality for continuous-space discrete-time Markov chain and elementary results in Calculus. See, e.g., Theorem 1.4 and Corollary 1.5 in the paper [167] for a proof. For ease in notation define $\mathcal{K}\backslash\mathcal{S} := \mathcal{K} \cap \mathcal{S}^c$. We now state a key isoperimetric inequality.

Lemma D.9 (Theorem 6 [163])**.** *For any measurable sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{K}$, we have*

$$\mathrm{vol}(\mathcal{K}\backslash\mathcal{S}_1\backslash\mathcal{S}_2) \cdot \mathrm{vol}(\mathcal{K}) \geq \mathfrak{d}_{\mathcal{K}}(\mathcal{S}_1, \mathcal{S}_2) \cdot \mathrm{vol}(\mathcal{S}_1) \cdot \mathrm{vol}(\mathcal{S}_2),$$

where $\mathfrak{d}_{\mathcal{K}}(\mathcal{S}_1, \mathcal{S}_2) := \inf_{x \in \mathcal{S}_1, y \in \mathcal{S}_2} \mathfrak{d}_{\mathcal{K}}(x, y)$*.*

Since $\Pi^\star$ is the uniform measure on $\mathcal{K}$, this lemma implies that

$$\Pi^\star(\mathcal{K}\backslash\mathcal{S}_1\backslash\mathcal{S}_2) \geq \mathfrak{d}_{\mathcal{K}}(\mathcal{S}_1, \mathcal{S}_2) \cdot \Pi^\star(\mathcal{S}_1) \cdot \Pi^\star(\mathcal{S}_2). \tag{D.46}$$

In fact, such an inequality holds for an arbitrary log-concave distribution [168]. In words, the inequality says that for a bounded convex set any two subsets which are far apart, can not have a large volume. Taking these lemmas as given, we now complete the proof.

Proof of Theorem 6.1: We first bound the conductance of the Markov chain using the assumptions of the lemma. From Lemma D.8, we see that the Markov chain mixes fast if all the sets $\mathcal{S}$ have a high conductance $\Phi(\mathcal{S})$. We claim that

$$\Phi \geq \frac{\rho\Delta}{64}, \tag{D.47}$$

from which the proof follows by applying Lemma D.8. We now prove the claim (D.47) along the lines of Theorem 11 in the paper [163]. In particular, we show that under the assumptions in the lemma, the sets with bad conductance are far apart and thereby have a small measure under $\Pi^\star$, whence the ratio $\Phi(\mathcal{S})/\Pi^\star(\mathcal{S})$ is not arbitrarily small. Consider a partition $\mathcal{S}_1, \mathcal{S}_2$ of the set $\mathcal{K}$ such that $\mathcal{S}_1$ and $\mathcal{S}_2$ are measurable. To prove claim (D.47), it suffices to show that

$$\frac{1}{\mathrm{vol}(\mathcal{K})} \int_{\mathcal{S}_1} \mathcal{T}_u(\mathcal{S}_2) du \geq \frac{\rho\Delta}{64} \cdot \min\left\{\Pi^\star(\mathcal{S}_1), \Pi^\star(\mathcal{S}_2)\right\}, \tag{D.48}$$

Define the sets

$$\mathcal{S}_1' := \left\{u \in \mathcal{S}_1 \middle| \widetilde{\mathcal{T}}_u(\mathcal{S}_2) < \frac{\rho}{2}\right\}, \quad \mathcal{S}_2' \quad := \left\{v \in \mathcal{S}_2 \middle| \widetilde{\mathcal{T}}_v(\mathcal{S}_1) < \frac{\rho}{2}\right\}, \quad \text{and} \quad \mathcal{S}_3' := \mathcal{K}\backslash\mathcal{S}_1'\backslash\mathcal{S}_2'. \tag{D.49}$$

Case 1: If we have $\text{vol}(\mathcal{S}_1') \leq \text{vol}(\mathcal{S}_1)/2$ and consequently $\text{vol}(\mathcal{K}\backslash\mathcal{S}_1') \geq \text{vol}(\mathcal{S}_1)/2$, then

$$\int_{\mathcal{S}_1} \mathcal{T}_u(\mathcal{S}_2)du \overset{(i)}{\geq} \frac{1}{2}\int_{\mathcal{S}_1\backslash\mathcal{S}_1'} \widetilde{\mathcal{T}}_u(\mathcal{S}_2)du \overset{(ii)}{\geq} \frac{\rho}{4}\text{vol}(\mathcal{S}_1) \overset{(iii)}{\geq} \frac{\rho\Delta}{4}\cdot\min\left\{\text{vol}(\mathcal{S}_1), \text{vol}(\mathcal{S}_2)\right\},$$

which implies the inequality (D.48) since $\Pi^\star$ is the uniform measure on $\mathcal{K}$. In the above sequence of inequalities, step (i) follows from the definition of the kernel $\mathcal{T}$, step (ii) follows from the definition of the set $\mathcal{S}_1'$ (D.49) and step (iii) from the fact that $\Delta < 1$. Dividing both sides by $\text{vol}(\mathcal{K})$ yields the inequality (D.48) and we are done.

Case 2: It remains to establish the inequality (D.48) for the case when $\text{vol}(\mathcal{S}_i') \geq \text{vol}(\mathcal{S}_i)/2$ for each $i \in \{1, 2\}$. Now for any $u \in \mathcal{S}_1'$ and $v \in \mathcal{S}_2'$ we have

$$\left\|\widetilde{\mathcal{T}}_u - \widetilde{\mathcal{T}}_v\right\|_{\text{TV}} \geq \widetilde{\mathcal{T}}_u(\mathcal{S}_1) - \widetilde{\mathcal{T}}_v(\mathcal{S}_1) = 1 - \widetilde{\mathcal{T}}_u(\mathcal{S}_2) - \widetilde{\mathcal{T}}_v(\mathcal{S}_1) > 1 - \rho,$$

and hence by assumption we have $\mathfrak{d}_{\mathcal{K}}(\mathcal{S}_1', \mathcal{S}_2') \geq \Delta$. Applying Lemma D.9 and the definition of $\mathcal{S}_3'$ (D.49) we find that

$$\text{vol}(\mathcal{S}_3')\cdot\text{vol}(\mathcal{K}) \geq \Delta\cdot\text{vol}(\mathcal{S}_1')\cdot\text{vol}(\mathcal{S}_2') \geq \frac{\Delta}{4}\cdot\text{vol}(\mathcal{S}_1)\cdot\text{vol}(\mathcal{S}_2). \tag{D.50}$$

Using this inequality and the fact that for any $x \in [0, 1]$ we have $x(1-x) \geq \min\left\{x, (1-x)\right\}/2$ we obtain that

$$\Pi^\star(\mathcal{S}_3') \geq \frac{\Delta}{4}\cdot\Pi^\star(\mathcal{S}_1)\cdot\Pi^\star(\mathcal{S}_2) \geq \frac{\Delta}{8}\min\left\{\Pi^\star(\mathcal{S}_1), \Pi^\star(\mathcal{S}_2)\right\}. \tag{D.51}$$

We claim that

$$\int_{\mathcal{S}_1} \mathcal{T}_u(\mathcal{S}_2)du = \int_{\mathcal{S}_2} \mathcal{T}_v(\mathcal{S}_1)dv. \tag{D.52}$$

Assuming the claim as given, we now complete the proof. Using the equation (D.52), we have

$$\begin{aligned}\frac{1}{\text{vol}(\mathcal{K})}\int_{\mathcal{S}_1} \mathcal{T}_u(\mathcal{S}_2)du &= \frac{1}{2\,\text{vol}(\mathcal{K})}\left(\int_{\mathcal{S}_1} \mathcal{T}_u(\mathcal{S}_2)du + \int_{\mathcal{S}_2} \mathcal{T}_v(\mathcal{S}_1)dv\right)\\ &\overset{(i)}{\geq} \frac{1}{2\,\text{vol}(\mathcal{K})}\left(\frac{1}{2}\int_{\mathcal{S}_1\backslash\mathcal{S}_1'} \widetilde{\mathcal{T}}_u(\mathcal{S}_2)du + \frac{1}{2}\int_{\mathcal{S}_2\backslash\mathcal{S}_2'} \widetilde{\mathcal{T}}_v(\mathcal{S}_2)dv\right)\\ &\overset{(ii)}{\geq} \frac{\rho}{8}\frac{\text{vol}(\mathcal{S}_3')}{\text{vol}(\mathcal{K})}\\ &\overset{(iii)}{\geq} \frac{\rho\Delta}{64}\min\left\{\Pi^\star(\mathcal{S}_1), \Pi^\star(\mathcal{S}_2)\right\},\end{aligned}$$

where step (i) follows from the definition of the kernel $\mathcal{T}$, step (ii) follows from the definition of the set $\mathcal{S}_3'$ (D.49) and step (iii) follows from the inequality (D.51). Putting together the pieces yields the claim (D.47).

It remains to prove the claim (D.52). We make use of the following result

$$\Phi(\mathcal{S}) = \Phi(\mathcal{K}\backslash\mathcal{S}) \quad \text{for any measurable } \mathcal{S} \subseteq \mathcal{K}. \tag{D.53}$$

Using equation (D.53) and noting that $\mathcal{S}_1 = \mathcal{K}\backslash\mathcal{S}_2$, we have

$$\frac{1}{\text{vol}(\mathcal{K})}\int_{\mathcal{S}_1} \mathcal{T}_u(\mathcal{S}_2)du = \int_{\mathcal{S}_1} \mathcal{T}_u(\mathcal{S}_2)\pi^\star(u)du = \Phi(\mathcal{S}_1) = \Phi(\mathcal{K}\backslash\mathcal{S}_1) = \frac{1}{\text{vol}(\mathcal{K})}\int_{\mathcal{S}_2} \mathcal{T}_v(\mathcal{S}_1)dv,$$

which yields equation (D.52).

Proof of result (D.53): Note that $\int_{\mathcal{K}} \mathcal{T}_u(\mathcal{S})d\Pi^\star(u) = \Pi^\star(\mathcal{S})$. Thus, we have

$$\Phi(\mathcal{K}\backslash\mathcal{S}) = \int_{\mathcal{K}\backslash\mathcal{S}} \mathcal{T}_u(\mathcal{S})d\Pi^\star(u) = \int_{\mathcal{K}} \mathcal{T}_u(\mathcal{S})d\Pi^\star(u) - \int_{\mathcal{S}} \mathcal{T}_u(\mathcal{S})d\Pi^\star(u) = \Pi^\star(\mathcal{S}) - \int_{\mathcal{S}} \mathcal{T}_u(\mathcal{S})d\Pi^\star(u).$$

Using the fact that $1 - \mathcal{T}_u(\mathcal{S}) = \mathcal{T}_u(\mathcal{K}\backslash\mathcal{S})$, we obtain

$$\Pi^\star(\mathcal{S}) - \int_{\mathcal{S}} \mathcal{T}_u(\mathcal{S})d\Pi^\star(u) = \int_{\mathcal{S}} d\Pi^\star(u) - \int_{\mathcal{S}} \mathcal{T}_u(\mathcal{S})d\Pi^\star(u) = \int_{\mathcal{S}} \mathcal{T}_u(\mathcal{K}\backslash\mathcal{S})d\Pi^\star(u) = \Phi(\mathcal{S}),$$

thereby yielding the claim (D.53).

Appendix E

Content Deferred From Chapter 6 For The John Walk

In this chapter, we provide the technical content related to John walk from the main text. We recap some ideas and notation in Appendix E.1, introduce some auxiliary results in Appendix E.2, and then prove Theorem 6.2 in Appendix E.3. We provide the proofs of the auxiliary results in Appendices E.4 and E.5.

E.1 Notation and recap

We recap the key ideas of the John walk for convenience. We have designed a new proposal distribution by making use of an *optimal set of weights* to define the new covariance structure for the Gaussian proposals, where optimality is defined with respect to a convex program (E.1). The optimality condition is closely related to the problem of finding the largest ellipsoid at any interior point of the polytope, such that the ellipsoid is contained within the polytope. This problem of finding the largest ellipsoid was first studied by John Fritz [133] in 1948 who showed that each convex body in $\mathbb{R}^d$ contains a unique ellipsoid of maximal volume. More recently, Lee and Sidford [152] make use of approximate John Ellipsoids to improve the convergence rate of interior point methods for linear programming. We refer the readers to their paper for more discussion about the use of John Ellipsoids for optimization problems. In this work, we make use of these ellipsoids for designing sampling algorithms with better theoretical bounds on the mixing times.

Let $\zeta_{x,i}/(b_i - a_i^\top x)^2$ denote the weight multiplied with $a_i a_i^\top$. Then the vector $\zeta_x = (\zeta_{x,1}, \ldots, \zeta_{x,n})^\top$ is computed by solving the following optimization problemma:

$$\zeta_x = \arg\min_{w \in \mathbb{R}^n} c_x(w) := \sum_{i=1}^{n} w_i - \frac{1}{\alpha_{\mathrm{J}}} \log\det\left(A^\top S_x^{-1} W^{\alpha_{\mathrm{J}}} S_x^{-1} A\right) - \beta_{\mathrm{J}} \sum_{i=1}^{n} \log w_i, \qquad \text{(E.1)}$$

where the parameters $\alpha_{\mathrm{J}}, \beta_{\mathrm{J}}$ are given by

$$\alpha_{\mathrm{J}} = 1 - \frac{1}{\log_2(2n/d)} \quad \text{and} \quad \beta_{\mathrm{J}} = \frac{d}{2n},$$

and W denotes an $n \times n$ diagonal matrix with $W_{ii} = w_i$ for each $i \in [n]$. In particular, for our proposal the inverse covariance matrix is proportional to J_x, where

$$J_x = \sum_{i=1}^{n} \zeta_{x,i} \frac{a_i a_i^\top}{(b_i - a_i^\top x)^2}. \tag{E.2}$$

where $\kappa := \kappa_{n,d} = \log_2(2n/d) = (1 - \alpha_{\mathrm{J}})^{-1}$.

Recall that for John walk with parameter $\frac{r}{d^{3/4}\kappa^2}$, the proposals at state x are drawn from the multivariate Gaussian distribution given by $\mathcal{N}\left(x, \frac{r^2}{d^{3/2}\kappa^4} J_x^{-1}\right)$, which we denote by $\mathcal{P}_x^{\mathrm{J}}$. In particular, the proposal density at point $x \in \operatorname{int}(\mathcal{K})$ is given by

$$p_x(z) := p(x, z) = \sqrt{\det J_x} \left(\frac{\kappa^4 d^{3/2}}{2\pi r^2}\right)^{d/2} \exp\left(-\frac{\kappa^4 d^{3/2}}{2r^2} \, (z - x)^\top J_x (z - x)\right). \tag{E.3}$$

E.2 Auxiliary results

We begin by proving basic properties of the weights ζ_x which are used throughout the remainder of the proofs. For $x \in \operatorname{int}(\mathcal{K}), w \in \mathbb{R}^n_{++}$, define the projection matrix $\Upsilon_{x,w}$ as follows

$$\Upsilon_{x,w} = W^{\alpha/2} A_x (A_x^\top W^\alpha A_x)^{-1} A_x^\top W^{\alpha/2}, \tag{E.4}$$

where $A_x = S_x^{-1} A$ and W is the $n \times n$ diagonal matrix with i-th diagonal entry given by w_i. Also, let

$$\sigma_{x,i} := \left(\Upsilon_{x,\zeta_x}\right)_{ii} \quad \text{for } x \in \operatorname{int}(\mathcal{K}) \text{ and } i \in [n]. \tag{E.5}$$

Define the *John slack sensitivity* θ_x^{J} as

$$\theta_x := \theta_x^{\mathrm{J}} := \left(\frac{a_1^\top J_x^{-1} a_1}{s_{x,1}^2}, \ldots, \frac{a_n^\top J_x^{-1} a_n}{s_{x,n}^2}\right)^\top \quad \text{for all } x \in \operatorname{int}(\mathcal{K}). \tag{E.6}$$

Further, for any $x \in \operatorname{int}(\mathcal{K})$, define the *John local norm at x* as

$$\left\|\cdot\right\|_{J_x} : v \mapsto \left\|J_x^{1/2} v\right\|_2 = \sqrt{\sum_{i=1}^{n} \zeta_{x,i} \frac{(a_i^\top v)^2}{s_{x,i}^2}}. \tag{E.7}$$

We now list some basic properties of the weights ζ_x and the local sensitivity θ_x.

Lemma E.1. *For any $x \in \mathrm{int}\,(\mathcal{K})$, the following properties are true:*

(a) (Implicit weight formula) $\zeta_{x,i} = \sigma_{x,i} + \beta_J$ *for all* $i \in [n]$,

(b) (Uniformity) $\zeta_{x,i} \in [\beta_J, 1 + \beta_J]$ *for all* $i \in [n]$,

(c) (Total size) $\sum_{i=1}^{n} \zeta_{x,i} = 3d/2$, *and*

(d) (Slack sensitivity) $\theta_{x,i} \in [0, 4]$ *for all* $i \in [n]$.

Lemma E.1 follows from Lemmas 14 and 15 by Lee and Sidford [152] and thereby we omit its proof.

Next, we state a key lemma that is crucial for proving the convergence rate of John walk. In this lemma, we provide bounds on difference in total variation norm between the proposal distributions of two nearby points.

Lemma E.2. *There exists a continuous non-decreasing function* $h : [0, 1/4] \to \mathbb{R}_+$ *with* $h(1/30) \geq 10^{-5}$, *such that for any* $\epsilon \in (0, 1/30]$, *the John walk with* $r \in [0, h(\epsilon)]$ *satisfies*

$$d_{\mathrm{TV}}\big(\mathcal{P}_x^J, \mathcal{P}_y^J\big) \leq \epsilon, \quad \textit{for all } x, y \in \mathrm{int}\,(\mathcal{K}) \textit{ such that } \|x - y\|_{J_x} \leq \frac{\epsilon r}{2\kappa^2 d^{3/4}}\ , \quad \textit{and} \tag{E.8a}$$

$$d_{\mathrm{TV}}\big(\mathcal{T}_{John}(\boldsymbol{\delta}_x), \mathcal{P}_x^J\big) \leq 5\epsilon, \ \textit{ for all } x \in \mathrm{int}\,(\mathcal{K}). \tag{E.8b}$$

See Section E.4 for its proof.

With these lemmas in hand, we are now ready to prove Theorem 6.2.

E.3 Proof of Theorem 6.2

The proof is similar to the proof of Theorem 1, and relies on the Lovász's Lemma. Here onwards, we use the following simplified notation

$$\mathcal{T}_x = \mathcal{T}_{\mathrm{John}}(\boldsymbol{\delta}_x), \mathcal{P}_x = \mathcal{P}_x^{\mathrm{J}} \text{ and } \|\cdot\|_x = \|\cdot\|_{J_x}.$$

In order to invoke Lovász's Lemma, we need to show that for any two points $x, y \in \mathrm{int}\,(\mathcal{K})$ with small cross-ratio $\mathfrak{d}_{\mathcal{K}}(x, y)$, the TV-distance $d_{\mathrm{TV}}\big(\mathcal{T}_x, \mathcal{T}_y\big)$ is also small.

We proceed with the proof in two steps: (A) first, we relate the cross-ratio $\mathfrak{d}_{\mathcal{K}}(x, y)$ to the John local norm of $x - y$ at x, and (B) we then use Lemma E.2 to show that if $x, y \in \mathrm{int}\,(\mathcal{K})$ are close in the John local-norm, then the transition kernels $\mathcal{T}_x$ and $\mathcal{T}_y$ are close in TV-distance.

Step (A): We claim that for all $x, y \in \text{int}\,(\mathcal{K})$, the cross-ratio can be lower bounded as

$$\mathfrak{b}_{\mathcal{K}}(x, y) \geq \frac{1}{\sqrt{3d/2}} \left\|x - y\right\|_x . \tag{E.9}$$

From the arguments in the proof of Theorem 1 (proof for the Vaidya Walk), we have

$$\mathfrak{b}_{\mathcal{K}}(x, y) \geq \max_{i \in [n]} \left| \frac{a_i^\top (x - y)}{s_{x,i}} \right| . \tag{E.10}$$

Using the fact that maximum of a set of non-negative numbers is greater than the weighted mean of the numbers and Lemma E.1, we find that

$$\mathfrak{b}_{\mathcal{K}}(x, y) \geq \sqrt{\frac{1}{\sum_{i=1}^n \zeta_{x,i}} \sum_{i=1}^n \zeta_{x,i} \frac{(a_i^\top (x - y))^2}{s_{x,i}^2}} = \frac{\left\|x - y\right\|_x}{\sqrt{3d/2}},$$

thereby proving the claim (E.9).

Step (B): By the triangle inequality, we have

$$d_{\text{TV}}\big(\mathcal{T}_x, \mathcal{T}_y\big) \leq d_{\text{TV}}\big(\mathcal{T}_x, \mathcal{P}_x\big) + d_{\text{TV}}\big(\mathcal{P}_x, \mathcal{P}_y\big) + d_{\text{TV}}\big(\mathcal{P}_y, \mathcal{T}_y\big).$$

Using Lemma E.2, we obtain that

$$d_{\text{TV}}\big(\mathcal{T}_x, \mathcal{T}_y\big) \leq 11\epsilon, \quad \forall x, y \in \text{int}\,(\mathcal{K}) \text{ such that } \left\|x - y\right\|_x \leq \frac{\epsilon r}{2\kappa^2 d^{3/4}}.$$

Consequently, the John walk satisfies the assumptions of Lovász's Lemma with

$$\Delta := \frac{1}{\sqrt{3d/2}} \cdot \frac{\epsilon r}{2\kappa^2 d^{3/4}} \quad \text{and} \quad \rho := 1 - 11\epsilon.$$

Plugging in $\epsilon = 1/30$, $r = 10^{-5}$, we obtain the claimed upper bound of $\mathcal{O}\left(\kappa^4 d^{5/2}\right)$ on the mixing time of the random walk.

E.4 Proof of Lemma E.2

We prove the lemma for the following function,

$$h(\epsilon) = \min \left\{ \frac{1}{25\sqrt{1 + \sqrt{2}\log(4/\epsilon)}}, \frac{\epsilon}{\left(2\sqrt{32}\gamma_{1,\epsilon}\right)}, \sqrt{\frac{\epsilon}{386\sqrt{24}\gamma_{2,\epsilon}}}, \frac{\epsilon}{5\sqrt{60}\gamma_{3,\epsilon}}, \right.$$
$$\left. \sqrt{\frac{\epsilon}{8\sqrt{1680}\gamma_{4,\epsilon}}}, \sqrt{\frac{\epsilon}{40\left(\gamma_{2,\epsilon}\gamma_{6,\epsilon}\sqrt{24}\sqrt{15120}\right)^{1/2}}}, \sqrt{\frac{\epsilon}{204800\gamma_{2,\epsilon}\sqrt{24}\log(32/\epsilon)}} \right\}.$$

where $\gamma_{1,\epsilon} = \log(2/\epsilon)$and $\gamma_{k,\epsilon} = \left(2e/k \cdot \log\left(16/\epsilon\right)\right)^{k/2}$ for $k = 2, 3, 4$ and 6. A numerical calculation shows that $h(1/30) \geq 10^{-5}$.

We now prove the two parts (E.8a) (E.8b) of the Lemma separately.

E.4.1 Proof of claim (E.8a)

Applying Pinsker's inequality, and plugging in the closed formed expression for the KL divergence between two Gaussian distributions we find that

$$d_{\mathrm{TV}}\big(\mathcal{P}_x, \mathcal{P}_y\big)^2 \le 2\,\mathrm{KL}(\mathcal{P}_y \| \mathcal{P}_x) = \mathrm{trace}(J_x^{-1/2} J_y J_x^{-1/2}) - d - \log\det(J_x^{-1/2} J_y J_x^{-1/2}) + \frac{\kappa^4 d^{3/2}}{r^2}\left\|x-y\right\|_x^2$$

$$= \sum_{i=1}^{d}\left(\lambda_i - 1 + \log\frac{1}{\lambda_i}\right) + \frac{\kappa^4 d^{3/2}}{r^2}\left\|x-y\right\|_x^2, \tag{E.11}$$

where $\lambda_1, \ldots, \lambda_d > 0$ denote the eigenvalues of the matrix $J_x^{-1/2} J_y J_x^{-1/2}$. To bound the expression (E.11), we make use of the following lemma:

Lemma E.3. *For any scalar $t \in [0, 1/64]$ and pair of points $x, y \in \mathrm{int}\,(\mathcal{K})$ such that $\left\|x-y\right\|_x \le t/\kappa^2$, we have*

$$\big(1 - 48t + 4t^2\big)\,\mathbb{I}_d \preceq J_x^{-1/2} J_y J_x^{-1/2} \preceq \big(1 + 48t + 4t^2\big)\,,$$

where $\preceq$ denotes ordering in the PSD cone and $\mathbb{I}_d$ denotes the d-dimensional identity matrix.

See Section E.5.3 for the proof of this lemma.

For $\epsilon \in (0, 1/30]$ and $r = 10^{-5}$, we have $t = \epsilon r/(2d^{3/4}) \le 1/64$, whence the eigenvalues $\{\lambda_i, i \in [d]\}$ can be sandwiched as

$$1 - \frac{24\epsilon r}{d^{3/4}} + \frac{\epsilon^2 r^2}{d^{3/2}} \le \lambda_i \le 1 + \frac{24\epsilon r}{d^{3/4}} + \frac{\epsilon^2 r^2}{d^{3/2}} \quad \text{for all } i \in d. \tag{E.12}$$

We are now ready to bound the TV distance between $\mathcal{P}_x$ and $\mathcal{P}_y$. Using the bound (E.11) and the inequality $\log\gamma \le \gamma - 1$, valid for $\gamma > 0$, we obtain

$$d_{\mathrm{TV}}\big(\mathcal{P}_x, \mathcal{P}_y\big)^2 \le \sum_{i=1}^{d}\left(\lambda_i - 2 + \frac{1}{\lambda_i}\right) + \frac{\kappa^4 d^{3/2}}{r^2}\left\|x-y\right\|_x^2.$$

Using the assumption that $\left\|x-y\right\|_x \le \epsilon r/\left(2\kappa^2 d^{3/4}\right)$, and plugging in the bounds (E.12) for the eigenvalues $\{\lambda_i, i \in [d]\}$, we find that

$$\sum_{i=1}^{d}\left(\lambda_i - 2 + \frac{1}{\lambda_i}\right) + \frac{\kappa^4 d^{3/2}}{r^2}\left\|x-y\right\|_x^2 \le \frac{2000\epsilon^2 r^2}{\sqrt{d}} + \frac{\epsilon^2}{4}.$$

In asserting this inequality, we have used the facts that

$$\frac{1}{1 - 24\gamma + \gamma^2} \le 1 + 24\gamma + 1000\gamma^2, \quad \text{and} \quad \frac{1}{1 + 24\gamma + \gamma^2} \le 1 - 24\gamma + 1000\gamma^2 \quad \text{for all } \gamma \in \big[0, \tfrac{1}{100}\big].$$

Note that for any $r \in [0, 1/100]$, we have that $2000r^2/\sqrt{d} \le 1/2$. Putting the pieces together yields $d_{\mathrm{TV}}\big(\mathcal{P}_x, \mathcal{P}_y\big) \le \epsilon$, as claimed.

E.4.2 Proof of claim (E.8b)

We have

$$d_{\mathrm{TV}}\left(\mathcal{P}_x, \mathcal{T}_x\right) \leq \underbrace{\frac{3}{2}\mathcal{P}_x(\mathcal{K}^c)}_{=:\, S_1} + \underbrace{1 - \mathbb{E}_{z\sim\mathcal{P}_x}\left[\min\left\{1, \frac{p_z(x)}{p_x(z)}\right\}\right]}_{=:\, S_2}, \tag{E.13}$$

where $\mathcal{K}^c$ denotes the complement of $\mathcal{K}$. We now show that $S_1 \leq \epsilon$ and $S_2 \leq 4\epsilon$, from which the claim follows.

Bounding the term S_1: Note that for $z \sim \mathcal{N}(x, \frac{r^2}{\kappa^2 d^{3/2}} J_x^{-1})$, we can write

$$z \overset{d}{=} x + \frac{r}{\kappa d^{3/4}} J_x^{-1/2}\xi, \tag{E.14}$$

where $\xi \sim \mathcal{N}(0, \mathbb{I}_d)$ and $\overset{d}{=}$ denotes equality in distribution. Using equation (E.14) and definition (E.6) of $\theta_{x,i}$, we obtain the bound

$$\frac{\left(a_i^\top(z-x)\right)^2}{s_{x,i}^2} = \frac{r^2}{\kappa^2 d^{3/2}}\left[\frac{a_i^\top J_x^{-1/2}\xi}{s_{x,i}}\right]^2 \overset{(i)}{\leq} \frac{r^2}{\kappa^2 d^{3/2}}\theta_{x,i}\left\|\xi\right\|_2^2 \overset{(ii)}{\leq} \frac{4r^2}{d}\left\|\xi\right\|_2^2, \tag{E.15}$$

where step (i) follows from Cauchy-Schwarz inequality, and step (ii) from part (d) of Lemma E.1. Define the events

$$\mathcal{E} := \left\{\frac{r^2}{d}\left\|\xi\right\|_2^2 < \frac{1}{4}\right\} \quad \text{and} \quad \mathcal{E}' := \left\{z \in \mathrm{int}\left(\mathcal{K}\right)\right\}.$$

Inequality (E.15) implies that $\mathcal{E} \subseteq \mathcal{E}'$ and hence $\mathbb{P}\left[\mathcal{E}'\right] \geq \mathbb{P}\left[\mathcal{E}\right]$. Using a standard Gaussian tail bound and noting that $r \leq \frac{1/2}{1+\sqrt{2/d\log(2/\epsilon)}}$, we obtain $\mathbb{P}\left[\mathcal{E}\right] \geq 1 - \epsilon/2$ and whence $\mathbb{P}\left[\mathcal{E}'\right] \geq 1 - \epsilon/2$. Thus, we have shown that $\mathbb{P}\left[z \notin \mathcal{K}\right] \leq \epsilon/2$ which implies that $S_1 \leq \epsilon$.

Bounding the term S_2: By Markov's inequality, we have

$$\mathbb{E}_{z\sim\mathcal{P}_x}\left[\min\left\{1, \frac{p_z(x)}{p_x(z)}\right\}\right] \geq \alpha\mathbb{P}\left[p_z(x) \geq \alpha p_x(z)\right] \quad \text{for all } \alpha \in (0, 1]. \tag{E.16}$$

By definition (E.3) of p_x, we obtain

$$\frac{p_z(x)}{p_x(z)} = \exp\left(-\frac{d^{3/2}\kappa^4}{2r^2}\left(\left\|z-x\right\|_z^2 - \left\|z-x\right\|_x^2\right) + \frac{1}{2}\left(\log\det J_z - \log\det J_x\right)\right).$$

The following lemma provides us with useful bounds on the two terms in this expression, valid for any $x \in \mathrm{int}\left(\mathcal{K}\right)$.

Lemma E.4. *For any $\epsilon \in (0, \frac{1}{4}]$ and $r \in (0, h(\epsilon)]$, we have*

$$\mathbb{P}_{z\sim\mathcal{P}_x}\left[\frac{1}{2}\log\det J_z - \frac{1}{2}\log\det J_x \geq -\epsilon\right] \geq 1-\epsilon, \quad \textit{and} \tag{E.17a}$$

$$\mathbb{P}_{z\sim\mathcal{P}_x}\left[\|z-x\|_z^2 - \|z-x\|_x^2 \leq 2\epsilon\frac{r^2}{\kappa^4 d^{3/2}}\right] \geq 1-\epsilon. \tag{E.17b}$$

We provide the of this lemma in Section E.5.4.

Using Lemma E.4, we now complete the proof of the Theorem 6.2. For $r \leq h(\epsilon)$, we obtain

$$\frac{p_z(x)}{p_x(z)} \geq \exp\left(-2\epsilon\right) \geq 1-2\epsilon$$

with probability at least $1-2\epsilon$. Substituting $\alpha = 1-2\epsilon$ in inequality (E.16) yields that $S_2 \leq 4\epsilon$, as claimed.

E.5 Proof of Lemmas E.3 and E.4

We first collect some additional notation, and state several technical lemmas in Appendices E.5.1 and E.5.2 that we then use to prove Lemmas E.3 and E.4 in Appendices E.5.3 and E.5.4 respectively.

E.5.1 Deterministic expressions and bounds

We begin by summarizing a few key properties of various terms involved in our analysis. Let $\Sigma_{x,w}$ be an $n\times n$ diagonal matrix defined as

$$\Sigma_{x,w} = \text{diag}\left(\sigma_{x,w,i},\ldots,\sigma_{x,w,n}\right) \text{ where } \sigma_{x,\zeta_x,w,i} = (\Upsilon_{x,w})_{ii}, i \in [n]. \tag{E.18a}$$

Let $\Upsilon_{x,w}^{(2)}$ denote the hadamard product of $\Upsilon_{x,w}$ with itself. Further define

$$\Lambda_{x,w} := \Sigma_{x,w} - \Upsilon_{x,w}^{(2)}. \tag{E.18b}$$

Lee and Sidford [152] proved that the weight vector ζ_x is the unique solution of the following fixed point equation:

$$w_i = \sigma_{x,w,i} + \beta_{\text{J}}, i \in [n]. \tag{E.19a}$$

To simplify notation, we use the following shorthands:

$$\sigma_x = \sigma_{x,\zeta_x}, \quad \Upsilon_x = \Upsilon_{x,\zeta_x}, \quad \Upsilon_x^{(2)} = \Upsilon_{x,\zeta_x}^{(2)}, \quad \Sigma_x = \Sigma_{x,\zeta_x}, \quad \Lambda_x = \Lambda_{x,\zeta_x}. \tag{E.19b}$$

Thus, we have the following relation:

$$\zeta_x = \sigma_{x,\zeta_x} + \beta_{\text{J}}\mathbf{1} = \sigma_x + \beta_{\text{J}}\mathbf{1}. \tag{E.19c}$$

Next, we collect some properties of various terms defined above.

Lemma E.5. *For any $x \in \operatorname{int}(\mathcal{K})$, the following properties hold:*

(a) $\sigma_{x,i} = \sum_{j=1}^{n} \sigma_{x,i,j}^2 = \sum_{j,k=1}^{n} \sigma_{x,i,j}\sigma_{x,j,k}\sigma_{x,k,i}$ *for each* $i \in [n]$,

(b) $\Sigma_x \succeq \Upsilon_x^{(2)}$,

(c) $\sum_{i=1}^{n} \zeta_{x,i}\theta_{x,i} = d$,

(d) $\theta_{x,i} = \sum_{j=1}^{n} \zeta_{x,i}\theta_{x,i,j}^2$, *for each* $i \in [n]$,

(e) $\theta_x^\top \Sigma_x \theta_x = \sum_{i=1}^{n} \theta_{x,i}^2 \zeta_{x,i} \leq 4d$, *and*

(f) $\beta_J \nabla^2 \mathcal{F}_x \preceq J_x \preceq (1+\beta_J)\nabla^2 \mathcal{F}_x$.

The proof is based on the ideas similar to Lemma 5 in the proof of the Vaidya walk and is thereby omitted.

The next lemma relates the change in *slackness* $s_{x,i} = b_i - a_i^\top x$ to the John-local norm at x.

Lemma E.6. *For all $x, y \in \operatorname{int}(\mathcal{K})$, we have*

$$\max_{i\in[n]} \left| 1 - \frac{s_{y,i}}{s_{x,i}} \right| \leq 2 \left\| x - y \right\|_x .$$

Proof. For any pair $x, y \in \operatorname{int}(\mathcal{K})$ and index $i \in [n]$, we have

$$\left(a_i^\top (x-y)\right)^2 \overset{(i)}{\leq} \| J_x^{-\frac{1}{2}} a_i \|_2^2 \; \| J_x^{\frac{1}{2}} (x-y) \|_2^2 = \theta_{x,i} s_{x,i}^2 \; \| x - y \|_x^2 \overset{(ii)}{\leq} 4 s_{x,i}^2 \; \| x - y \|_x^2 ,$$

where step (i) follows from the Cauchy-Schwarz inequality, and step (ii) uses the bound $\theta_{x,i}$ from Lemma E.1(d). Noting the fact that $a_i^\top (x-y) = s_{y,i} - s_{x,i}$, the claim follows after simple algebra. □

We now state various expressions and bounds for the first and second order derivatives of the different terms. To lighten notation, we introduce some shorthand notation. For any $y \in \operatorname{int}(\mathcal{K})$ and $h \in \mathbb{R}^d$, define the following terms:

$$d_{y,i} = \frac{a_i^\top h}{s_{y,i}}, \; i \in [n] \qquad D_y = \operatorname{diag}(d_{y,1}, \ldots, d_{y,n}), \tag{E.20a}$$

$$f_{y,i} = \frac{\nabla \zeta_{y,i}^\top h}{\zeta_{y,i}}, \; i \in [n] \qquad F_y = \operatorname{diag}(f_{y,1}, \ldots, f_{y,n}), \tag{E.20b}$$

$$\ell_{y,i} = \frac{1}{2} h^\top \nabla^2 \zeta_{y,i} h / \zeta_{y,i}, \; i \in [n] \qquad L_y = \operatorname{diag}(\ell_{y,1}, \ldots, \ell_{y,n}), \tag{E.20c}$$

$$\rho_y := (G_y - \alpha \Lambda_y) \begin{bmatrix} \ell_{y,1} \\ \vdots \\ \ell_{y,n} \end{bmatrix}, \tag{E.20d}$$

where for brevity in our notation we have omitted the dependence on h. The choice of h is specified as per the context. Further, we define for each $x \in \text{int}(\mathcal{K})$ and $i \in [n]$

$$\varphi_{x,i} := \frac{\zeta_{x,i}}{s_{x,i}^2}, \qquad \text{and} \qquad \Psi_x := \frac{1}{2}\log\det J_x, \tag{E.21}$$

$$\hat{a}_{x,i} := \frac{J_x^{-1/2}a_{x,i}}{s_{x,i}^2}, \qquad \text{and} \qquad \hat{b}_{x,i} := J_x^{-1/2}A_x\Lambda_x\left(G_x - \alpha\Lambda_x\right)^{-1}e_i. \tag{E.22}$$

Next, we state expressions for gradients of ζ, φ and Ψ and bounds for directional Hessian of σ, φ and Ψ which are used in various Taylor's series expansions and bounds in our proof.

Lemma E.7 (Calculus)**.** *For any $y \in \text{int}(\mathcal{K})$ and $h \in \mathbb{R}^n$, the following relations hold;*

(a) Gradient of ζ: $(f_{y,1}, \ldots, f_{y,n})^\top = 2\left(G_y - \alpha\Lambda_y\right)^{-1}\Lambda_y A_y h$;

(b) Hessian of ζ:

$$\left\|\rho_y\right\|_1 \leq 56\kappa^2 \sum_{i=1}^n \zeta_{y,i} d_{y,i}^2. \tag{E.23}$$

(c) Gradient of Ψ: $\nabla\Psi^\top h = \theta_y^\top G_y\left(\mathbb{I}_n + \left(G_y - \alpha\Lambda_y\right)^{-1}\Lambda_y\right)A_y h$.

(d) Gradient of φ: $\nabla\varphi_{y,i}^\top h = \varphi_{y,i}\left(2d_{y,i} + f_{y,i}\right)$.

(e) Bound on $\nabla^2\Psi$: $\frac{1}{2}\left|h^\top(\nabla^2\Psi)h\right| \leq \frac{1}{2}\left[\sum_{i=1}^n \zeta_{y,i}\,\theta_{y,i}\left[9\,d_{y,i}^2 + 4f_{y,i}^2\right] + \left|\sum_{i=1}^n \zeta_{y,i}\,\theta_{y,i}\ell_{y,i}\right|\right]$

(f) Bound on $\nabla^2\varphi$:

$$\left|\sum_{i=1}^n d_{y,i}^2 s_{y,i}^2 \frac{1}{2} h^\top\nabla^2\varphi_{y,i}h\right| \leq 3\sum_{i=1}^n \zeta_{y,i}d_{y,i}^4 + 2\left|\sum_{i=1}^n \zeta_{y,i}d_{y,i}^3 f_{y,i}\right| + \left|\sum_{i=1}^n \zeta_{y,i}d_{y,i}^2\ell_{y,i}\right|.$$

The proof is provided in Section E.5.5.1.

Next, we state some results that would be useful to provide explicit bounds for various terms like f_y, ℓ_y and ρ_y that appear in the statements of the previous lemma. Note that the following results do not have a corresponding analog in our analysis of the Vaidya walk.

Lemma E.8. *For any $c_1, c_2 \geq 0$, $y \in \text{int}(\mathcal{K})$, we have*

$$\left(c_1\mathbb{I}_n + c_2\Lambda_y\left(G_y - \alpha\Lambda_y\right)^{-1}\right)G_y\left(c_1\mathbb{I}_n + c_2\left(G_y - \alpha\Lambda_y\right)^{-1}\Lambda_y\right) \preceq \left(c_1 + c_2\right)^2\kappa^2 G_y,$$

where $\preceq$ denotes the ordering in the PSD cone.

Lemma E.9. *Let μ_y denote the $n \times n$ matrix $\left(G_y - \alpha\Lambda_y\right)^{-1} G_y$, and let $\mu_{y,i,j}$ denote its ij-th entry. Then for each $i \in [n]$ and $y \in \operatorname{int}(\mathcal{K})$, we have*

$$\mu_{y,i,i} \in [0, \kappa], \qquad \textit{and,} \tag{E.24a}$$

$$\sum_{j \neq i, j \in [n]} \frac{\mu_{y,i,j}^2}{\zeta_{y,j}} \leq \kappa^3. \tag{E.24b}$$

Corollary E.1. *Let $e_i \in \mathbb{R}^n$ denote the unit vector along i-th axis. Then for any $y \in \operatorname{int}(\mathcal{K})$, we have*

$$\left\| G_y \left(G_y - \alpha\Lambda_y\right)^{-1} e_i \right\|_1 \leq 3\sqrt{d}\kappa^{3/2}, \qquad \textit{for all } i \in [n]. \tag{E.25}$$

Consequently, we also have $\vert\kern-0.25ex\vert\kern-0.25ex\vert \left(G_y - \alpha\Lambda_y\right)^{-1} G_y \vert\kern-0.25ex\vert\kern-0.25ex\vert_\infty \leq 3\sqrt{d}\kappa^{3/2}$.

See Section E.5.5.2, E.5.5.3 and E.5.5.4 for the proofs of Lemma E.8, Lemma E.9 and Corollary E.1 respectively.

E.5.2 Tail Bounds

We now collect lemmas that provide us with useful tail bounds.

We start with a result that shows that for a random variable $z \sim \mathcal{P}_x$, the slackness $s_{z,i}$ is close to $s_{x,i}$ with high probability and consequently the weights $\zeta_{z,i}$ are also close to $\zeta_{x,i}$. This result comes in handy for transferring the remainder terms in Taylor expansions to the reference point (around which the series is being expanded).

Lemma E.10. *For any point $x \in \operatorname{int}(\mathcal{K})$ and $r \leq \frac{1}{25 \cdot \sqrt{1+\sqrt{2}\log(4/\epsilon)}}$, we have*

$$\mathbb{P}_{z \sim \mathcal{P}_x}\left[\forall i \in [n], \forall v \in \overline{xz},\ \frac{s_{x,i}}{s_{v,i}} \in [0.99, 1.01] \ \textit{and}\ \frac{\zeta_{x,i}}{\zeta_{v,i}} \in [0.96, 1.04]\right] \geq 1 - \epsilon/4 \tag{E.26a}$$

See Section E.5.6.1 for the proof of this lemma.

Next, we state high probability results for some Gaussian polynomials. These results are useful to bound various polynomials of the form $\sum_{i=1}^n \zeta_{x,i} d_{x,i}^k$, where $d_{x,i} = a_i^\top(z - x)/s_{x,i}$ and z is drawn from the transition distribution for the John walk at point x.

Lemma E.11 (Gaussian moment bounds)**.** *To simplify notations, all subscripts on x are omitted in the following statements. For any $\epsilon \in (0, 1/30]$, define $\gamma_{k,\epsilon} := \gamma_k = \left(2e/k \cdot \log\left(16/\epsilon\right)\right)^{k/2}$,* *for $k = 2, 3, 4$ and 6, then we have*

$$\mathbb{P}\left[\sum_{i=1}^n \zeta_i \left(\hat{a}_i^\top \xi\right)^2 \leq \gamma_2 \sqrt{24d}\right] \geq 1 - \frac{\epsilon}{16}, \tag{E.27a}$$

$$\mathbb{P}\left[\sum_{i=1}^n \zeta_i \left(\hat{a}_i^\top \xi\right)^3 \leq \gamma_3 \sqrt{60}d^{1/2}\right] \geq 1 - \frac{\epsilon}{16}, \tag{E.27b}$$

$$\mathbb{P}\left[\sum_{i=1}^n \zeta_i \left(\hat{a}_i^\top \xi\right)^2 \left(\hat{b}_i^\top \xi\right) \leq \gamma_3 \sqrt{240}\kappa d^{1/2}\right] \geq 1 - \frac{\epsilon}{16}, \tag{E.27c}$$

$$\mathbb{P}\left[\sum_{i=1}^n \zeta_i \left(\hat{a}_i^\top \xi\right)^4 \leq \gamma_4 \sqrt{1680d}\right] \geq 1 - \frac{\epsilon}{16}, \tag{E.27d}$$

$$\mathbb{P}\left[\sum_{i=1}^n \zeta_i \left(\hat{a}_i^\top \xi\right)^6 \leq \gamma_6 \sqrt{15120d}\right] \geq 1 - \frac{\epsilon}{16}. \tag{E.27e}$$

See Section E.5.6.2 for the proof.

E.5.3 Proof of Lemma E.3

As a direct consequence of Lemma E.6, for any $x, y \in \text{int}\,(\mathcal{K})$ such that $\|x - y\|_x \leq t/\kappa^2$, we have

$$\max_{i\in[n]} \left|1 - \frac{s_{y,i}}{s_{x,i}}\right| \leq \frac{2t}{\kappa^2}. \tag{E.28}$$

Bounding the terms in $\nabla^2 \mathcal{F}_x$ one by one, we obtain

$$\left(1 - \frac{2t}{\kappa^2}\right)^2 \nabla^2 \mathcal{F}_y \preceq \nabla^2 \mathcal{F}_x \preceq \left(1 + \frac{2t}{\kappa^2}\right)^2 \nabla^2 \mathcal{F}_y.$$

We claim that

$$\left\|\log \zeta_y - \log \zeta_x\right\|_\infty \leq 16t. \tag{E.29}$$

Assuming the claim as given at the moment, we now complete the proof. Putting the result (E.29) in matrix form, we obtain that $\exp\left(-16t\right)\mathbb{I}_n \preceq G_x^{-1} G_y \preceq \exp\left(16t\right)\mathbb{I}_n$, and hence

$$\exp\left(-16t\right)\zeta_{x,i} \leq \zeta_{y,i} \leq \exp\left(16t\right)\zeta_{x,i}. \tag{E.30}$$

Consequently, using the definition of J_x we have,

$$\underbrace{\left(1 - \frac{2t}{\kappa^2}\right)^2 \exp\left(-16t\right)}_{\gamma_\ell} J_x \leq J_y \leq \underbrace{\left(1 + \frac{2t}{\kappa^2}\right)^2 \exp\left(16t\right)}_{\gamma_u} J_y.$$

Letting $\gamma = 2t$, we obtain

$$\gamma_\ell \geq (1-\gamma)^2 \cdot \exp(-8\gamma) \overset{(i)}{\geq} 1 - 24\gamma + \gamma^2, \qquad \text{and} \qquad \gamma_u \leq (1+\gamma)^2 \cdot \exp(8\gamma) \overset{(ii)}{\leq} 1 + 24\gamma + \gamma^2,$$

where inequalities (i) and (ii) hold since $\gamma \leq 1/24$. Putting the pieces together, we find that

$$\left(1 - 48t + 4t^2\right) J_x \preceq J_y \preceq \left(1 - 48t + 4t^2\right) J_x$$

for $t \in [0, 1/48]$.

Now, we return to the proof of our earlier claim (E.29). We use an argument based on the continuity of the function $x \mapsto \log \zeta_x$. (Such an argument appeared in a similar scenario in [152].) For $\lambda \in [0,1]$, define $u_\lambda = \lambda y + (1-\lambda)x$. Let

$$\lambda^{\max} := \sup \left\{ \lambda \in [0,1] \middle| \left\| \log \zeta_{u_\lambda} - \log \zeta_x \right\|_\infty \leq 16t \right\}. \tag{E.31}$$

It suffices to establish that $\lambda^{\max} = 1$. Note that $\lambda = 0$ is feasible on the RHS of equation (E.31) and hence $\lambda^{\max}$ exists. Now for any $\lambda \in [0, \lambda^{\max}]$ and $i \in \{1, \ldots, n\}$, there exists v on the segment $\overline{u_\lambda x}$ such that

$$\left|\log \zeta_{u_\lambda, i} - \log \zeta_{x,i}\right| = \left| \left(\frac{\nabla \zeta_{v,i}}{\zeta_{v,i}} \right)^\top (u_\lambda - x) \right| \overset{(i)}{\leq} \left\| G_v^{-1} G_v' (y-x) \right\|_\infty = 2 \left\| (G_v - \alpha \Lambda_v)^{-1} \Lambda_v A_v (y-x) \right\|_\infty .$$

where in step (i) we have used the fact that $u_\lambda - x = \lambda(y-x)$ and $\lambda \in [0,1]$. We claim that

$$\left\| (G_v - \alpha\Lambda_v)^{-1} \Lambda_v u_1 \right\|_\infty \leq \kappa \left\| u_1 \right\|_\infty + 2\kappa^2 \left\| G_v^{1/2} u_1 \right\|_2 \quad \text{for any } u_1 \in \mathbb{R}^n. \tag{E.32}$$

We prove the claim at the end of this section. We now derive bounds for the two terms on the RHS of the equation (E.32) for $u_1 = A_v(y-x)$. Note that

$$\left\| A_v (y-x) \right\|_\infty = \max_i \left| \frac{s_{y,i} - s_{x,i}}{s_{v,i}} \right| = \max_i \left| \frac{s_{y,i} - s_{x,i}}{s_{x,i}} \right| \left| \frac{s_{x,i}}{s_{v,i}} \right| \overset{(i)}{\leq} \frac{2t}{\kappa^2 \left(1 - 2t/\kappa^2\right)} \overset{(ii)}{\leq} \frac{3t}{\kappa^2}.$$

Inequality (i) uses bound (E.28) and inequality (ii) follows by plugging in $t \leq 1/64$. Next, we have

$$\begin{aligned} \left\| G_v^{1/2} A_v (y-x) \right\|_2^2 = \sum_{i=1}^n \zeta_{x,i} \frac{\left(a_i^\top (y-x)\right)^2}{s_{x,i}^2} \frac{\zeta_{v,i}}{\zeta_{x,i}} \frac{s_{v,i}^2}{s_{x,i}^2} &\overset{(i)}{\leq} \left\| x - y \right\|_x^2 \max_{i \in [n]} \frac{\zeta_{v,i}}{\zeta_{x,i}} \frac{s_{v,i}^2}{s_{x,i}^2} \\ &\overset{(ii)}{\leq} \frac{t^2}{\kappa^4} \left(1 + (16t) + (16t)^2\right) \left(1 + \frac{2t}{\kappa^2}\right)^2 \\ &\overset{(iii)}{\leq} \frac{1.5t}{\kappa^4}, \end{aligned}$$

where step (i) follows from the definition of the local norm; step (ii) follows from bounds (E.28) and (E.31) and the fact that $e^x \leq 1 + x + x^2$ for all $x \in [0, 1/4]$; and inequality (iii) follows by plugging in $t \leq 1/64$. Putting the pieces together, we obtain

$$\left\|\log \zeta_{u_\lambda} - \log \zeta_x\right\|_\infty \leq 2(\kappa \cdot 3t/\kappa^2 + 2\kappa^2 \cdot 1.5t/\kappa^4) \leq 12t < 16t.$$

The strict inequality is valid for $\lambda = \lambda^{\max}$. Consequently, using the continuity of $x \mapsto \log \zeta_x$, we conclude that $\lambda^{\max} = 1$.

It is left to prove claim (E.32). Let $v := \left(G_v - \alpha\Lambda_v\right)^{-1} \Lambda_v u_1$. which implies $\left(G_v - \alpha\Lambda_v\right) v = \Lambda_v u_1$. Plugging the expression of G_v and Λ_v, we have

$$\left((1-\alpha)\Sigma_v + \beta_{\mathrm{J}}\mathbb{I}_n + \alpha\Upsilon_v^{(2)}\right) v = \left(\Sigma_v - \Upsilon_v^{(2)}\right) u_1.$$

Writing component wise, we find that for any $i \in [n]$, we have

$$\begin{aligned}
\left|\left((1-\alpha)\sigma_{v,i} + \beta_{\mathrm{J}}\right) v_i\right| &\leq \alpha \left|e_i^\top \Upsilon_v^{(2)} v\right| + \sigma_{v,i}\left|u_{1,i}\right| + \left|e_i^\top \Upsilon_v^{(2)} u_1\right| \\
&\overset{(i)}{\leq} \alpha\sigma_{v,i}\left\|\Sigma_v^{1/2} v\right\|_2 + \sigma_{v,i}\left\|u_1\right\|_\infty + \sigma_{v,i}\left\|\Sigma_v^{1/2} u_1\right\|_2 \\
&\overset{(ii)}{\leq} \alpha\sigma_{v,i}\left\|G_v^{1/2} v\right\|_2 + \sigma_{v,i}\left\|u_1\right\|_\infty + \sigma_{v,i}\left\|G_v^{1/2} u_1\right\|_2 \\
&\overset{(iii)}{\leq} \alpha\sigma_{v,i}\kappa\left\|G_v^{1/2} u_1\right\|_2 + \sigma_{v,i}\left\|u_1\right\|_\infty + \sigma_{v,i}\left\|G_v^{1/2} u_1\right\|_2,
\end{aligned} \tag{E.33}$$

where inequality (ii) from the fact that $\Sigma_y \preceq G_y$ and inequality (iii) from Lemma E.8 with $c_1 = 0, c_2 = 1$. To assert inequality (i), observe the following

$$\left|\sum_{j=1}^n \sigma_{y,i,j}^2 v_j\right| \leq \sum_{j=1}^n \sigma_{y,i,j}^2 \left|v_j\right| \overset{(a)}{\leq} \sigma_{y,i} \sum_{j=1}^n \sigma_{y,j}\left|v_j\right| \overset{(b)}{\leq} \sigma_{y,i} \sum_{j=1}^n \sqrt{\sigma_{y,j}}\left|v_j\right| = \sigma_{y,i}\left\|\Sigma_v^{1/2} v\right\|_2,$$

where step (a) follows from the fact that $\sigma_{y,i,j}^2 \leq \sigma_{y,i}\sigma_{y,j}$, and step (b) from the fac that $\sigma_{y,i} \in [0,1]$. Dividing both sides of inequality (E.33) by $\left((1-\alpha)\sigma_{v,i} + \beta_{\mathrm{J}}\right)$ and observing that $\sigma_{v,i}/\left((1-\alpha)\sigma_{v,i} + \beta_{\mathrm{J}}\right) \leq \kappa$, and $\alpha \in [0,1]$, yields the claim.

E.5.4 Proof of Lemma E.4

We prove Lemma E.4 in two parts: claim (E.17a) in Section E.5.4.1 and claim (E.17b) in Section E.5.4.2.

E.5.4.1 Proof of claim (E.17a)

Using the second order Taylor expansion, we have

$$\Psi_z - \Psi_x = (z-x)^\top \nabla\Psi_x + \frac{1}{2}(z-x)^\top \nabla^2\Psi_y (z-x), \quad \text{for some } y \in \overline{xz}.$$

We claim that for $r \leq h(\epsilon)$, we have

$$\mathbb{P}\left[(z-x)^\top \nabla \Psi_x \geq -\epsilon/2\right] \geq 1-\epsilon/2, \text{ and} \tag{E.34a}$$

$$\mathbb{P}\left[\frac{1}{2}(z-x)\nabla^2\Psi_y(z-x) \geq -\epsilon/2\right] \geq 1-\epsilon/2. \tag{E.34b}$$

Note that the claim (E.17a) follows from the above two claims.

Proof of bound (E.34a): We observe that

$$(z-x)^\top \nabla\Psi_x \sim \mathcal{N}\left(0, \frac{r^2}{\kappa^2 n}\nabla\Psi_x^\top J_x^{-1}\nabla\Psi_x\right).$$

Let $E_x = \mathbb{I}_n + (G_x - \alpha\Lambda_x)^{-1}\Lambda_x$. Substituting the expression of $\nabla\Psi_x$ from Lemma E.7 (c) and applying Cauchy-Schwarz inequality, we have that for any vector $u \in \mathbb{R}^d$

$$u^\top \nabla\Psi_x \nabla\Psi_x^\top u = (\theta_x^\top G_x E_x A_x u)^2 \leq \left(u^\top A_x^\top G_x A_x u\right)\cdot\left(\theta_x^\top G_x E_x G_x^{-1} E_x G_x \theta_x\right). \tag{E.35}$$

Observe that

$$G_x^{1/2} E_x G_x^{-1/2} = \mathbb{I}_n + (\mathbb{I}_n - \alpha G_x^{-1/2}\Lambda_x G_x^{-1/2})^{-1}(G_x^{-1/2}\Lambda_x G_x^{-1/2}).$$

Now, using the intermediate bound (E.58) from the proof of Lemma E.8, we obtain that

$$\mathbb{I}_n \preceq G_x^{1/2} E_x G_x^{-1/2} \preceq 2\kappa\mathbb{I}_n,$$

and hence $G_x \preceq G_x E_x G_x^{-1} E_x G_x \preceq 4\kappa^2 G_x$. Consequently, we have

$$\theta_x^\top G_x E_x G_x^{-1} E_x G_x \theta_x \leq 4\kappa^2 \theta_x^\top G_x \theta_x = 4\kappa^2 \sum_{i=1}^n \zeta_{x,i}\theta_{x,i}^2 \leq 16\kappa^2 d,$$

where the last step follows from Lemma E.5. Putting the pieces together into equation (E.35), we obtain $\nabla\Psi_x\nabla\Psi_x^\top \preceq 16\kappa^2 d J_x$ whence $J_x^{-1/2}\nabla\Psi_x\nabla\Psi_x^\top J_x^{-1/2} \preceq 16\kappa^2 d\mathbb{I}_d$. Noting that the matrix $J_x^{-1/2}\nabla\Psi_x\nabla\Psi_x^\top J_x^{-1/2}$ has rank one, we have

$$\nabla\Psi_x^\top J_x^{-1}\nabla\Psi_x = \text{trace}\left(J_x^{-1/2}\nabla\Psi_x\nabla\Psi_x^\top J_x^{-1/2}\right) \leq 16\kappa^2 d.$$

Using standard Gaussian tail bound, we have $\mathbb{P}\left((z-x)^\top\nabla\Psi_x \geq -\sqrt{32}\gamma_1 r\right) \geq 1 - \exp\left(-\gamma_1^2\right)$. Choosing $\gamma_1 = \log(2/\epsilon)$, and observing that

$$r \leq \frac{\epsilon}{(2\sqrt{32}\gamma_1)}, \tag{E.36}$$

yields the claim.

Proof of bound (E.34b): In the following proof, we use $h = z - x$ for definitions (E.20a)-(E.20d). According to Lemma E.7(e), we have

$$\left|\frac{1}{2}\left(z-x\right)^\top \nabla^2 \Psi_y \left(z-x\right)\right| \leq \sum_{i=1}^n \zeta_{y,i}\, \theta_{y,i} \left[\frac{9}{2}\, d_{y,i}^2 + 2 f_{y,i}^2\right] + \frac{1}{2}\left|\sum_{i=1}^n \zeta_{y,i}\, \theta_{y,i} \ell_{y,i}\right|$$

We claim that

$$\sum_{i=1}^n \zeta_{y,i}\, \theta_{y,i} \left[\frac{9}{2}\, d_{y,i}^2 + 2 f_{y,i}^2\right] + \frac{1}{2}\left|\sum_{i=1}^n \zeta_{y,i}\, \theta_{y,i} \ell_{y,i}\right| \leq 386\sqrt{d}\kappa^4 \sum_{i=1}^n \zeta_{y,i} d_{y,i}^2. \tag{E.37}$$

Assuming the claim as given at the moment, we now complete the proof. Note that y is some particular point on $\overline{xz}$ and its dependence on z is hard to characterize. Consequently, we transfer all the terms with dependence on y, to terms with dependence on x only. We have

$$\sum_{i=1}^n \zeta_{y,i} d_{y,i}^2 = \sum_{i=1}^n \zeta_{x,i} d_{x,i}^2 \underbrace{\frac{\zeta_{y,i}}{\zeta_{x,i}} \frac{s_{x,i}^2}{s_{y,i}^2}}_{\tau_{y,i}}.$$

We now invoke the following high probability bounds implied by Lemma E.10 and Lemma E.11 (E.27a) respectively

$$\mathbb{P}\left[\sup_{y\in\overline{xz}, i\in[n]} \tau_{y,i} \leq 1.1\right] \geq 1 - \epsilon/4, \quad \text{and,} \quad \mathbb{P}\left[\sum_{i=1}^n \zeta_{x,i} \left(\hat{a}_{x,i}^\top \xi\right)^2 \leq \gamma_2 \sqrt{24 d}\right] \geq 1 - \epsilon/16. \tag{E.38}$$

Since $h = z - x$, we have that $d_{x,i}^2 = \frac{r^2}{\kappa^2 d^{3/2}} \left(\hat{a}_{x,i}^\top \xi\right)^2$. Consequently, for

$$r \leq \sqrt{\frac{\epsilon}{386\sqrt{24}\gamma_2}}, \tag{E.39}$$

with probability at least $1 - \epsilon/2$, we have

$$\left|\frac{1}{2}\left(z-x\right)^\top \nabla^2 \Psi_y \left(z-x\right)\right| \overset{\text{eqn. (E.37)}}{\leq} 386\sqrt{d}\kappa^4 \sum_{i=1}^n \zeta_{y,i} d_{y,i}^2 \overset{\text{hpb (E.38)}}{\leq} \epsilon,$$

which completes the proof.

We now turn to the proof of claim (E.37). First we observe the following relationship between the terms $d_{y,i}$ and $f_{y,i}$:

$$\sum_{i=1}^n \zeta_{y,i} f_{y,i}^2 \overset{(i)}{=} 4 h^\top A_y^\top \Lambda_y \left(G_y - \alpha \Lambda_y\right)^{-1} G_y \left(G_y - \alpha\Lambda_y\right)^{-1} \Lambda_y A_y h \overset{(ii)}{\leq} 4\kappa^2 h^\top A_y^\top G_y A_y h = 4\kappa^2 \sum_{i=1}^n \zeta_{y,i} d_{y,i}^2, \tag{E.40}$$

where step (i) follows by plugging in the definition of $f_{y,i}$ (E.20b) and step (ii) by invoking Lemma E.8 with $c_1 = 0$ and $c_2 = 1$. Next, we relate the term on the LHS of equation (E.37) involving $\ell_{y,i}$ to a polynomial in $d_{y,i}$. Using Lemma E.7, we find that

$$\left|\sum_{i=1}^{n} \zeta_{y,i}\, \theta_{y,i}\ell_{y,i}\right| = \left|\left(\left(G_y - \alpha\Lambda_y\right)^{-1} G_y\theta_y\right)^\top \left(G_y - \alpha\Lambda_y\right)\ell_y\right| \leq \left\|\underbrace{\left(G_y - \alpha\Lambda_y\right)^{-1} G_y\theta_y}_{v_1}\right\|_\infty \left\|\underbrace{\left(G_y - \alpha\Lambda_y\right)\ell_y}_{\rho_y}\right\|_1,$$

where the last step follows from the Holder's inequality: for any two vectors $u, v \in \mathbb{R}^d$, we have that $u^\top v \leq \|u\|_\infty \|v\|_1$. Substituting the bound for the norm $\|v_1\|_\infty$ from Corollary E.1 and the bound on $\rho_{y,i}$ from Lemma E.7(b), we obtain that

$$\left|\sum_{i=1}^{n} \zeta_{y,i}\, \theta_{y,i}\ell_{y,i}\right| \leq 12\sqrt{n}\kappa^{3/2} \sum_{i=1}^{n} \left[7\zeta_{y,i}d_{y,i}^2 + 3\zeta_{y,i}f_{y,i}^2 + \sum_{j=1}^{n}\left(13d_{y,j}^2 + 6f_{y,j}^2\right)\Upsilon_{y,i,j}^2\right] \leq 672\sqrt{n}\kappa^4 \sum_{i=1}^{n} \zeta_{y,i}d_{y,i}^2,$$

where the last step follows from Lemma E.5(a) and the bound (E.40). The claim now follows.

E.5.4.2 Proof of claim (E.17b)

Writing $z = x + tu$, where t is a scalar and u is a unit vector in $\mathbb{R}^d$, we obtain

$$\|z - x\|_z^2 - \|z - x\|_x^2 = t^2 \sum_{i=1}^{n} \left(a_i^\top u\right)^2 \left(\varphi_{z,i} - \varphi_{x,i}\right).$$

Now, we use a Taylor's series expansion for $\sum_{i=1}^{n} \left(a_i^\top u\right)^2 \left(\varphi_{z,i} - \varphi_{x,i}\right)$ around the point x, along the line u. There exists a point $y \in \overline{xz}$ such that

$$\sum_{i=1}^{n} \left(a_i^\top u\right)^2 \left(\varphi_{z,i} - \varphi_{x,i}\right) = \sum_{i=1}^{n} \left(a_i^\top u\right)^2 \left(\left(z - x\right)^\top \nabla\varphi_{x,i} + \frac{1}{2}\left(z - x\right)^\top \nabla^2\varphi_{y,i}\left(z - x\right)\right).$$

Note that the point y in this discussion is not the same as the point y used in previous proofs, in particular in Section E.5.4.1. Multiplying both sides by t^2, and using the shorthand $d_{x,i} = \frac{a_i^\top(z-x)}{s_{x,i}}$, we obtain

$$\|z-x\|_z^2 - \|z-x\|_x^2 = \sum_{i=1}^{n} d_{x,i}^2 s_{x,i}^2 \left(z-x\right)^\top \nabla\varphi_{x,i} + \sum_{i=1}^{n} d_{x,i}^2 s_{x,i}^2 \frac{1}{2}\left(z-x\right)^\top \nabla^2\varphi_{y,i}\left(z-x\right). \tag{E.41}$$

We claim that for $r \leq h(\epsilon)$, we have

$$\mathbb{P}_{z\sim\mathcal{T}_x^{\text{J}}}\left[\sum_{i=1}^{n} d_{x,i}^2 s_{x,i}^2 \left(z-x\right)^\top \nabla\varphi_{x,i} \leq \epsilon\frac{r^2}{\kappa^4 d^{3/2}}\right] \geq 1 - \epsilon/2, \text{ and} \tag{E.42a}$$

$$\mathbb{P}_{z\sim\mathcal{T}_x^{\text{J}}}\left[\sup_{y\in\overline{xz}}\left(\sum_{i=1}^{n} d_{x,i}^2 s_{x,i}^2 \frac{1}{2}\left(z-x\right)^\top \nabla^2\varphi_{y,i}\left(z-x\right)\right) \leq \epsilon\frac{r^2}{\kappa^4 d^{3/2}}\right] \geq 1 - \epsilon/2. \tag{E.42b}$$

We now prove each claim separately.

Proof of bound (E.42a): Using Lemma E.7(d) and using $h = z - x$ where z is given by the relation (E.14), we find that

$$\begin{aligned}\sum_{i=1}^{n} d_{x,i}^2 s_{x,i}^2 \left(z-x\right)^\top \nabla \varphi_{x,i} &= \sum_{i=1}^{n} \zeta_{x,i} d_{x,i}^2 \left(2 d_{x,i} + f_{x,i}\right) \\ &= \frac{r^3}{d^{9/4}\kappa^6} \sum_{i=1}^{n} \zeta_{x,i} \left(\hat{a}_{x,i}^\top \xi\right)^3 + \frac{2r^3}{d^{9/4}\kappa^6} \sum_{i=1}^{n} \zeta_{x,i} \left(\hat{a}_{x,i}^\top \xi\right)^2 \left(\hat{b}_{x,i}^\top \xi\right)\end{aligned} \tag{E.43}$$

Using high probability bounds for the two terms in equation (E.43) from Lemma E.11, part (E.27b) and part (E.27c), we obtain that

$$\left|\sum_{i=1}^{n} d_{x,i}^2 s_{x,i}^2 \left(z-x\right)^\top \nabla \varphi_{x,i}\right| \leq \frac{5\sqrt{60}\gamma_3 r^3}{\kappa^5 d^{7/4}} \leq \epsilon \frac{r^2}{\kappa^4 d^{3/2}},$$

with probability at least $1 - \epsilon/2$. The last inequality uses the condition that

$$r \leq \frac{\epsilon}{5\sqrt{60}\gamma_3}. \tag{E.44}$$

The claim now follows.

Proof of bound (E.42b): Note that $d_{x,i}s_{x,i} = a_i^\top h = d_{y,i}s_{y,i}$ for any h. Using this equality for $h = z - x$, we find that

$$\begin{aligned}\left|\sum_{i=1}^{n} d_{x,i}^2 s_{x,i}^2 \frac{1}{2} h^\top \nabla^2 \varphi_{y,i} h\right| &= \left|\sum_{i=1}^{n} d_{y,i}^2 s_{y,i}^2 \frac{1}{2} h^\top \nabla^2 \varphi_{y,i} h\right| \\ &\overset{(i)}{\leq} \underbrace{3 \sum_{i=1}^{n} \zeta_{y,i} d_{y,i}^4}_{C_1} + \underbrace{2 \left|\sum_{i=1}^{n} \zeta_{y,i} d_{y,i}^3 f_{y,i}\right|}_{C_2} + \underbrace{\left|\sum_{i=1}^{n} \zeta_{y,i} d_{y,i}^2 \ell_{y,i}\right|}_{C_3},\end{aligned} \tag{E.45}$$

where step (i) follows from Lemma E.7(f). We can write C_1 as follows

$$\sum_{i=1}^{n} \zeta_{y,i} d_{y,i}^4 = \sum_{i=1}^{n} \zeta_{x,i} d_{x,i}^4 \frac{\zeta_{y,i}}{\zeta_{x,i}} \frac{d_{y,i}^4}{d_{x,i}^4} = \frac{r^4}{n^3\kappa^8} \sum_{i=1}^{n} \zeta_{x,i} \left(\hat{a}_{x,i}^\top \xi\right)^4 \frac{\zeta_{y,i}}{\zeta_{x,i}} \frac{d_{y,i}^4}{d_{x,i}^4}. \tag{E.46}$$

Now, we claim the following:

$$C_2 \leq 2 \frac{r^4}{n^3\kappa^7} \cdot \sqrt{\left[\sum_{i=1}^{n} \zeta_{x,i} \left(\hat{a}_{x,i}^\top \xi\right)^2 \frac{\zeta_{y,i}}{\zeta_{x,i}} \frac{d_{y,i}^2}{d_{x,i}^2}\right] \cdot \left[\sum_{i=1}^{n} \zeta_{x,i} \left(\hat{a}_{x,i}^\top \xi\right)^6 \frac{\zeta_{y,i}}{\zeta_{x,i}} \frac{d_{y,i}^6}{d_{x,i}^6}\right]}, \quad \text{and,} \tag{E.47a}$$

$$C_3 \leq 56 \frac{r^4}{n^3\kappa^{4.5}} \left(\sum_{i=1}^{n} \zeta_{x,i} \left(\hat{a}_{x,i}^\top \xi\right)^2 \frac{\zeta_{y,i}}{\zeta_{x,i}} \frac{d_{y,i}^2}{d_{x,i}^2}\right) \left(\max_i \left(\hat{a}_{x,i}^\top \xi\right)^2 \frac{d_{y,i}^2}{d_{x,i}^2} + \sqrt{\sum_{i=1}^{n} \zeta_{x,i} \left(\hat{a}_{x,i}^\top \xi\right)^4 \frac{\zeta_{y,i}}{\zeta_{x,i}} \frac{d_{y,i}^4}{d_{x,i}^4}}\right) \tag{E.47b}$$

Assuming the claims as given, we now complete the proof. Using Lemma E.10, we have

$$\mathbb{P}\left[\frac{\zeta_{y,i}}{\zeta_{x,i}}\frac{d_{y,i}^6}{d_{x,i}^6} \le 1.2\right] \ge 1 - \epsilon/4,$$

and consequently

$$\begin{aligned} 3C_1+2C_2+C_3 \le \frac{r^4}{d^3\kappa^{4.5}}\Bigg[&4\cdot\sum_{i=1}^n \zeta_{x,i}(\hat{a}_{x,i}^\top\xi)^4 + 10\cdot\left(\sum_{i=1}^n \zeta_{x,i}(\hat{a}_{x,i}^\top\xi)^2\cdot\sum_{i=1}^n \zeta_{x,i}(\hat{a}_{x,i}^\top\xi)^6\right)^{1/2} \\ &+ 100\cdot\sum_{i=1}^n \zeta_{x,i}\left(\hat{a}_{x,i}^\top\xi\right)^2\cdot\left(\max_i(\hat{a}_{x,i}^\top\xi)^2 + \Big(\sum_{i=1}^n \zeta_{x,i}(\hat{a}_{x,i}^\top\xi)^4\Big)^{1/2}\right)\Bigg], \end{aligned} \tag{E.48}$$

with probability at least $1-\epsilon/4$. Now, we observe that for all $i \in [n]$ and $x \in \operatorname{int}(\mathcal{K})$, we have

$$\left(\hat{a}_{x,i}^\top\xi\right) \sim \mathcal{N}(0, \theta_{x,i}) \quad \text{and} \quad \theta_{x,i} \le 4.$$

Invoking the standard tail bound for maximum of Gaussian random variables, we obtain

$$\mathbb{P}\left[\max_i \left|\left(\hat{a}_{x,i}^\top\xi\right)\right| \le 8\cdot\left(\sqrt{\log n} + \sqrt{\log(32/\epsilon)}\right)\right] \ge 1-\epsilon/16.$$

Using the fact that $2c_1c_2 \ge c_1 + c_2$ for all $c_1, c_2 \ge 1$, we obtain

$$\mathbb{P}\left[\max_i \left|\left(\hat{a}_{x,i}^\top\xi\right)\right| \le 16\cdot\sqrt{\log n}\cdot\sqrt{\log(32/\epsilon)}\right] \ge 1-\epsilon/16.$$

Combining this bound with the tail bounds for various Gaussian polynomials (E.27a), (E.27d), (E.27e) from Lemma E.11, and substituting in inequality (E.48), we obtain that

$$\begin{aligned} \left|\sum_{i=1}^n d_{x,i}^2 s_{x,i}^2 \frac{1}{2} h^\top \nabla^2 \varphi_{y,i} h\right| \le \frac{r^4}{\kappa^{6.5}d^3}\Bigg[&4\cdot\gamma_4\sqrt{1680}d + 10\left(\gamma_2\sqrt{24}d\cdot\gamma_6\sqrt{15120}d\right)^{1/2} \\ &+ 100\cdot\gamma_2\sqrt{24}d\cdot\left(256\cdot\log n\cdot\log(32/\epsilon) + \left(\gamma_4\sqrt{1680}d\right)^{1/2}\right)\Bigg] \end{aligned}$$

with probability at least $1-\epsilon/2$. In the above expression, the terms γ_i are a function of ϵ as defined in Lemma E.11. In particular, $\gamma_i := \gamma_{i,\epsilon} = (2e/i\cdot\log(16/\epsilon))^{i/2}$ for $i \in \{2,3,4,6\}$. Observing that $256\log(32/\epsilon) \ge \left(\gamma_4\sqrt{1680}\right)^{1/2}$, and that our choice of r satisfies

$$r^2 \le \min\left\{\frac{\epsilon}{8\sqrt{1680}\gamma_4}, \frac{\epsilon}{40\left(\gamma_2\gamma_6\sqrt{24}\sqrt{15120}\right)^{1/2}}, \frac{\epsilon}{204800\gamma_2\sqrt{24}\log(32/\epsilon)}\right\}, \tag{E.49}$$

we obtain

$$\left|\sum_{i=1}^n d_{x,i}^2 s_{x,i}^2 \frac{1}{2} h^\top \nabla^2 \varphi_{y,i} h\right| \leq \frac{r^2}{\kappa^4 d^{3/2}} \left[\frac{\epsilon}{2} + \frac{\epsilon}{4} + \frac{\epsilon}{8}\left(\frac{\log n}{\sqrt{d}} + 1\right)\right].$$

Asserting the additional condition $\sqrt{d} \geq \log n$, yields the claim.

It is now left to prove the bounds (E.47a) and (E.47b). We prove these bounds separately.

Bounding C_2: Applying Cauchy-Schwarz inequality, we have

$$\left|\sum_{i=1}^n \zeta_{y,i} d_{y,i}^3 f_{y,i}\right| \leq \left(\sum_{i=1}^n \zeta_{y,i} f_{y,i}^2 \cdot \sum_{i=1}^n \zeta_{y,i} d_{y,i}^6\right)^{1/2}$$

Using the bound (E.40), we obtain

$$\sum_{i=1}^n \zeta_{y,i} f_{y,i}^2 \leq 4\kappa^2 \sum_{i=1}^n \zeta_{y,i} d_{y,i}^2 = 4\kappa^2 \sum_{i=1}^n \zeta_{x,i} d_{x,i}^2 \frac{\zeta_{y,i}}{\zeta_{x,i}} \frac{d_{y,i}^2}{d_{x,i}^2}.$$

Substituting $h = z - x$ where z is given by relation (E.14), we obtain that $d_{x,i} = \frac{r}{d^{3/4}\kappa} \hat{a}_{x,i}^\top \xi$, and thereby

$$\sum_{i=1}^n \zeta_{y,i} f_{y,i}^2 \leq 4\kappa^2 \frac{r^2}{d^{3/2}\kappa^4} \sum_{i=1}^n \zeta_{x,i} (\hat{a}_{x,i}^\top \xi)^2 \frac{\zeta_{y,i}}{\zeta_{x,i}} \frac{d_{y,i}^2}{d_{x,i}^2}.$$

Doing similar algebra, we obtain $\sum_{i=1}^n \zeta_{y,i} d_{y,i}^6 = \frac{r^6}{d^{9/2}\kappa^{12}} \sum_{i=1}^n \zeta_{x,i} \left(\hat{a}_{x,i}^\top \xi\right)^6 \frac{\zeta_{y,i}}{\zeta_{x,i}} \frac{d_{y,i}^6}{d_{x,i}^6}$. Putting the pieces together yields the claim.

Bounding C_3: Recall that $\rho_y = (G_y - \alpha\Lambda_y)\ell_y$ (Lemma E.7) and $\mu_y = (G_y - \alpha\Lambda_y)^{-1} G_y$ (Lemma E.9). We have

$$\left|\sum_{i=1}^n \zeta_{y,i} d_{y,i}^2 \ell_{y,i}\right| = \mathbf{1} D_y^2 G_y \ell_y = \underbrace{\mathbf{1} D_y^2 G_y (G_y - \alpha\Lambda_y)^{-1}}_{=:u_y^\top} \underbrace{(G_y - \alpha\Lambda_y)\ell_y}_{\rho_y}.$$

Using the definition of u_y and μ_y, we obtain

$$u_{y,i} \coloneqq e_i^\top u_y = e_i^\top \left(G_y - \alpha\Lambda_y\right)^{-1} G_y D_y^2 \mathbf{1} = e_i^\top \mu_y D_y^2 \mathbf{1} = \mu_{y,i,i} d_{y,i}^2 + \sum_{j \in [n], j \neq i} \mu_{y,i,j} d_{y,j}^2.$$

Consequently, we have

$$\left|\sum_{i=1}^n u_{y,i} \rho_{y,i}\right| \leq \overbrace{\sum_{i=1}^n |\rho_{y,i}| \cdot \left|\mu_{y,i,i} d_{y,i}^2\right|}^{=:C_4} + \overbrace{\sum_{i=1}^n |\rho_{y,i}| \cdot \left(\sum_{j \in [n], j \neq i} \left|\mu_{y,i,j} d_{y,j}^2\right|\right)}^{=:C_5}$$

From Lemma E.9, we have that $\mu_{y,i,i} \in [0, \kappa]$. Hence, we have $C_4 \leq \left\|\rho_y\right\|_1 \cdot \kappa \cdot \max_{i\in[n]} d_{y,i}^2$. To bound C_5, we note that

$$\sum_{j\in[n],j\neq i} \left|\mu_{y,i,j} d_{y,j}^2\right| \overset{(i)}{\leq} \left(\sum_{j\in[n],j\neq i} \frac{\mu_{y,i,j}^2}{\zeta_{y,j}} \cdot \sum_{j=1}^n \zeta_{y,j} d_{y,j}^4\right)^{1/2} \overset{(ii)}{\leq} \left(\kappa^3 \cdot \sum_{j=1}^n \zeta_{x,j} d_{x,j}^4 \frac{\zeta_{y,j}}{\zeta_{x,j}} \frac{d_{y,j}^4}{d_{x,j}^4}\right)^{1/2},$$

where step (i) follows from Cauchy-Schwarz inequality and step (ii) from Lemma E.9. Putting the pieces together, we obtain that

$$\left|\sum_{i=1}^n \zeta_{y,i} d_{y,i}^2 \ell_{y,i}\right| \leq \left\|\rho_y\right\|_1 \cdot \left[\kappa \cdot \max_{i\in[n]} d_{y,i}^2 + \kappa^{3/2} \left(\sum_{j=1}^n \zeta_{x,j} d_{x,j}^4 \frac{\zeta_{y,j}}{\zeta_{x,j}} \frac{d_{y,j}^4}{d_{x,j}^4}\right)^{1/2}\right].$$

Using the bound on $\left\|\rho_y\right\|_1$ from Lemma E.7, we have

$$\left|\sum_{i=1}^n \zeta_{y,i} d_{y,i}^2 \ell_{y,i}\right| \leq \left(56\kappa^2 \sum_{i=1}^n \zeta_{y,i} d_{y,i}^2\right) \cdot \left[\kappa \cdot \max_{i\in[n]} d_{y,i}^2 + \kappa^{3/2} \left(\sum_{j=1}^n \zeta_{x,j} d_{x,j}^4 \frac{\zeta_{y,j}}{\zeta_{x,j}} \frac{d_{y,j}^4}{d_{x,j}^4}\right)^{1/2}\right].$$

Substituting the expression for $d_{x,i} = \frac{r}{\kappa^2 d^{3/4}} \left(\hat{a}_{x,i}^\top \xi\right)$ yields the claim.

E.5.5 Proofs of Lemmas from Section E.5.1

In this section we collect proofs of lemmas from Section E.5.1. Each lemma is proved in a different subsection.

E.5.5.1 Proof of Lemma E.7

Up to second order terms, we have

$$\frac{1}{s_{x+h,i}^2} = \frac{1}{s_{x,i}^2}\left[1 + \frac{2a_i^\top h}{s_{x,i}} + \frac{3(a_i^\top h)^2}{s_{x,i}^2}\right] + \mathcal{O}\left(\|h\|_2^3\right), \tag{E.50a}$$

$$\zeta_{y+h,i} = \zeta_{y,i} + h^\top \nabla\zeta_{y,i} + \frac{1}{2} h^\top \nabla^2 \zeta_{y,i} h + \mathcal{O}\left(\|h\|_2^3\right), \tag{E.50b}$$

$$\zeta_{y+h,i}^\alpha = \zeta_{y,i}^\alpha + \alpha\zeta_{y,i}^{\alpha-1}\left(h^\top \nabla\zeta_{y,i} + \frac{1}{2} h^\top \nabla^2 \zeta_{y,i} h\right) + \frac{\alpha(\alpha-1)}{2} \zeta_{y,i}^{\alpha-2}\left(h^\top \nabla\zeta_{y,i}\right)^2 + \mathcal{O}\left(\|h\|_2^3\right), \tag{E.50c}$$

Further, let

$$\widetilde{J}_y := A_y^\top G_y^\alpha A_y = \sum_{i=1}^n \zeta_{y,i}^\alpha \frac{a_i a_i^\top}{s_{y,i}^2}. \tag{E.50d}$$

Using equations (E.50a) and (E.50c), and substituting $d_{y,i} = a_i^\top h / s_{y,i}$, $f_{y,i} = h^\top \nabla \zeta_{y,i} / \zeta_{y,i}$ and $\ell_{y,i} = \frac{1}{2} h^\top \nabla^2 \zeta_{y,i} h / \zeta_{y,i}$, we find that

$$\widetilde{J}_{y+h} = \sum_{i=1}^n \left[1 + \alpha f_{y,i} + \alpha \ell_{y,i} + \frac{\alpha\,(\alpha-1)}{2} f_{y,i}^2\right] \left[1 + 2d_{y,i} + 3d_{y,i}^2\right] \zeta_{y,i}^\alpha \frac{a_i a_i^\top}{s_{y,i}^2} + \mathcal{O}\left(\|h\|_2^3\right).$$

Note that $d_{y,i}$ and $f_{y,i}$ are first order terms in $\|h\|_2$ and $\ell_{y,i}$ is a second order term in $\|h\|_2$. Thus we obtain

$$\widetilde{J}_{y+h} - \widetilde{J}_y = \underbrace{\sum_{i=1}^n \left(2d_{y,i} + \alpha f_{y,i}\right) \zeta_{y,i}^\alpha \frac{a_i a_i^\top}{s_{y,i}^2}}_{=:\Delta_{y,h}^{(1)}}$$
$$+ \underbrace{\sum_{i=1}^n \left[3d_{y,i}^2 + 2\alpha d_{y,i} f_{y,i} + \alpha \ell_{y,i} + \frac{\alpha(\alpha-1)}{2} f_{y,i}^2\right] \zeta_{y,i}^\alpha \frac{a_i a_i^\top}{s_{y,i}^2}}_{=:\Delta_{y,h}^{(2)}} + \mathcal{O}\left(\|h\|_2^3\right).$$

Let $\Delta_{y,h} \coloneqq \Delta_{y,h}^{(1)} + \Delta_{y,h}^{(2)}$. Note that $\Delta_{y,h}^{(i)}$ denotes the i-th order term in $\|h\|_2$. Finally, the following expansion also comes in handy for our derivations:

$$a_i^T \widetilde{J}_{y+h}^{-1} a_i = a_i^\top \widetilde{J}_y^{-1} a_i - a_i^\top \widetilde{J}_y^{-1} \Delta_{y,h} \widetilde{J}_y^{-1} a_i + a_i^\top \widetilde{J}_y^{-1} \Delta_{y,h} \widetilde{J}_y^{-1} \Delta_{y,h} \widetilde{J}_y^{-1} a_i + \mathcal{O}\left(\|h\|_2^3\right). \tag{E.50e}$$

Proof of part (a)—Gradient of weights: The expression for the gradient $\nabla \zeta_{y,i}$ is derived in Lemma 14 of the paper [152] and is thereby omitted.

Proof of part (b)—Hessian of weights: We claim that

$$\begin{aligned} \rho_y = \left(\mathbb{I} - \alpha \Lambda_y G_y^{-1}\right) \begin{bmatrix} \frac{1}{2} h^\top \nabla^2 \zeta_{y,1} h \\ \cdots \\ \frac{1}{2} h^\top \nabla^2 \zeta_{y,m} h \end{bmatrix} &= (2D_y + \alpha F_y) \Upsilon_y^{(2)} (2D_y + \alpha F_y) \mathbf{1} \\ &+ \left(\Sigma_y - \Upsilon_y^{(2)}\right) \left[2\alpha D_y F_y + 3D_y^2 + \tau_\alpha F_y^2\right] \mathbf{1} \\ &+ \operatorname{diag}\left(\Upsilon_y (2D_y + \alpha F_y) \Upsilon_y (2D_y + \alpha F_y) \Upsilon_y\right), \end{aligned} \tag{E.51}$$

where we have used diag(B) to denote the diagonal vector $(B_{1,1}, \ldots, B_{n,n})$ of the matrix B. Deferring the proof of this expression for the moment, we now derive a bound on the ℓ_1 norm of ρ_y. Expanding the i-th term of $\rho_{y,i}$ from equation (E.51), we obtain

$$\rho_{y,i} = (2d_{y,i} + \alpha f_{y,i}) \sum_{j=1}^n (2d_{y,j} + \alpha f_{y,j}) \Upsilon_{y,i,j}^2 + \left[2\alpha d_{y,i} f_{y,i} + 3d_{y,i}^2 + \tau_\alpha f_{y,i}^2\right] \sigma_{y,i}$$
$$- \sum_{j=1}^n \left[2\alpha d_{y,j} f_{y,j} + 3d_{y,j}^2 + \tau_\alpha f_{y,j}^2\right] \Upsilon_{y,i,j}^2 + \sum_{j,l=1}^n (2d_{y,j} + \alpha f_{y,j})(2d_{y,l} + \alpha f_{y,l}) \Upsilon_{y,i,j} \Upsilon_{y,j,l} \Upsilon_{y,l,i}.$$

Recall that $\alpha = 1 - 1/\log_2(2n/d)$. Since $n \geq d$ for polytopes, we have $\alpha \in [0, 1]$ and consequently $|\tau_\alpha| = |\alpha(\alpha-1)/2| \in [0, 1]$. Further note that Υ_x is an orthogonal projection matrix, and hence we have

$$\operatorname{diag}(\Upsilon_x e_i)\Upsilon_x \operatorname{diag}(\Upsilon_x e_i) \preceq \operatorname{diag}(\Upsilon_x e_i)\operatorname{diag}(\Upsilon_x e_i).$$

Combining these observations with the AM-GM inequality, we have

$$|\rho_{y,i}| \leq 7\sigma_{y,i}d_{y,i}^2 + 3\sigma_{y,i}f_{y,i}^2 + \sum_{j=1}^{n}\left(13d_{y,j}^2 + 6f_{y,j}^2\right)\Upsilon_{y,i,j}^2.$$

Summing both sides over the index i, we find that

$$\sum_{i=1}^{n}|\rho_{y,i}| \overset{(i)}{\leq} \sum_{i=1}^{n} 20\sigma_{y,i}d_{y,i}^2 + 9\sigma_{y,i}f_{y,i}^2 \overset{(ii)}{\leq} \sum_{i=1}^{n} 20\zeta_{y,i}d_{y,i}^2 + 9\zeta_{y,i}f_{y,i}^2 \overset{(iii)}{\leq} 56\kappa^2\sum_{i=1}^{n}\zeta_{y,i}d_{y,i}^2,$$

where step (i) follows from Lemma E.5 (a), step (ii) from Lemma E.1 (a) and step (iii) from the bound (E.40).

We now return to the proof of expression (E.51). Using equation (E.19c), we find that

$$\frac{1}{2}h^\top\nabla^2\zeta_{y,i}h = \frac{1}{2}h^\top\nabla^2\sigma_{y,i}h \quad \text{for all } i \in [n]. \tag{E.52}$$

Next, we derive the Taylor series expansion of $\sigma_{y,i}$. Using the definition of $\widetilde{J}_x$ (E.50d) in equation (E.4), we find that $\sigma_{y,i} = \zeta_{y,i}^\alpha \frac{a_i^\top \widetilde{J}_y^{-1}a_i}{s_{y,i}^2}$. To compute the difference $\sigma_{y+h,i} - \sigma_{y,i}$, we use the expansions (E.50a), (E.50c) and (E.50e). Letting $\tau_\alpha = \alpha(\alpha-1)/2$, we have

$$\begin{aligned}
\sigma_{y+h,i} &= \zeta_{y+h,i}^\alpha \frac{a_i^\top \widetilde{J}_{y+h}^{-1}a_i}{s_{y+h,i}^2} \\
&= \zeta_{y,i}^\alpha \frac{a_i^\top \widetilde{J}_{y+h}^{-1}a_i}{s_{y,i}^2}\left[1 + \alpha f_{y,i} + \alpha\ell_{y,i} + \tau_\alpha f_{y,i}^2\right]\left[1 + 2d_{y,i} + 3d_{y,i}^2\right] + \mathcal{O}\left(\|h\|_2^3\right) \\
&= \sigma_{y,i} + (2d_{y,i} + \alpha f_{y,i})\sigma_{y,i} - \sum_{j=1}^{n}(2d_{y,j} + \alpha f_{y,j})\Upsilon_{y,i,j}^2 + (2d_{y,i} + \alpha f_{y,i})\sum_{j=1}^{n}(2d_{y,j} + \alpha f_{y,j})\Upsilon_{y,i,j}^2 \\
&\quad + 2\alpha d_{y,i}f_{y,i}\sigma_{y,i} + \left[\alpha\ell_{y,i} + \tau_\alpha f_{y,i}^2 + 3d_{y,i}^2\right]\sigma_{y,i} - \sum_{j=1}^{n}\left[3d_{y,j}^2 + 2\alpha d_{y,j}f_{y,j} + \alpha\ell_{y,j} + \tau_\alpha f_{y,j}^2\right]\Upsilon_{y,i,j}^2 \\
&\quad + \sum_{j,l=1}^{n}(2d_{y,j} + \alpha f_{y,j})(2d_{y,l} + \alpha f_{y,l})\Upsilon_{y,i,j}\Upsilon_{y,j,l}\Upsilon_{y,l,i} + \mathcal{O}\left(\|h\|_2^3\right).
\end{aligned}$$

We identify the second order (in $\mathcal{O}\left(\|h\|_2^2\right)$) terms in the previous expression. Using the equation (E.52), these are indeed the terms that correspond to the terms $\frac{1}{2}h^\top\nabla^2\zeta_{y,i}h$, $i \in [n]$.

Substituting $\ell_{y,i} = \frac{1}{2}h^\top \nabla^2 \zeta_{y,i} h/\zeta_{y,i}$, we have

$$\begin{aligned}
&\frac{1}{2}h^\top \nabla^2 \zeta_{y,i} h \\
&= (2d_{y,i} + \alpha f_{y,i}) \sum_{j=1}^n (2d_{y,j} + \alpha f_{y,j}) \Upsilon_{y,i,j}^2 + 2\alpha d_{y,i} f_{y,i} \sigma_{y,i} + \left[\frac{\alpha}{2} \frac{h^\top \nabla^2 \zeta_{y,i} h}{\zeta_{y,i}} + \tau_\alpha f_{y,i}^2 + 3d_{y,i}^2 \right] \sigma_{y,i} \\
&- \sum_{j=1}^n \left[3d_{y,j}^2 + 2\alpha d_{y,j} f_{y,j} + \frac{\alpha}{2} \frac{h^\top \nabla^2 \zeta_{y,j} h}{\zeta_{y,j}} + \tau_\alpha f_{y,j}^2 \right] \Upsilon_{y,i,j}^2 + \sum_{j,l=1}^n (2d_{y,j} + \alpha f_{y,j})(2d_{y,l} + \alpha f_{y,l}) \Upsilon_{y,i,j} \Upsilon_{y,j,l} \Upsilon_{y,l,i}.
\end{aligned}$$

Collecting the different terms and doing some algebra yields the result (E.51).

Proof of part (c)—Gradient of $\log\det$**:** For a unit vector $h \in \mathbb{R}^d$, we have

$$h^\top \log\det J_y = \lim_{\delta \to 0} \frac{1}{\delta} (\log\det J_{y+\delta h} - \log\det J_y) = \lim_{\delta \to 0} \frac{1}{\delta} (\log\det J_y^{-1/2} J_{y+\delta h} J_y^{-1/2} - \log\det \mathbb{I}_d)$$

Let $\hat{a}_{y,i} := J_{y,i}^{-1/2} a_i / s_{y,i}$ for each $i \in [n]$. Using the property $\log\det B = \operatorname{trace}\log B$, where $\log B$ denotes the logarithm of the matrix and that $\log\det \mathbb{I}_d = 0$, we obtain

$$h^\top \log\det J_y = \lim_{\delta \to 0} \frac{1}{\delta} \left[\operatorname{trace}\log \left(\sum_{i=1}^n \frac{\zeta_{y+\delta h}}{(1 - \delta a_i^\top h/s_{y,i})} \hat{a}_{y,i} \hat{a}_{y,i}^\top \right) \right],$$

where we have substituted $s_{y+\delta h,i} = s_{y,i} - \delta a_i^\top h$. Keeping track of first order terms in δ, and noting that $\sum_{i=1}^n \zeta_{y,i} \hat{a}_{y,i} \hat{a}_{y,i}^\top = \mathbb{I}_d$, we find that

$$\begin{aligned}
\operatorname{trace}\log \left(\sum_{i=1}^n \frac{\zeta_{y+\delta h,i}}{(1 - \delta a_i^\top h/s_{y,i})} \hat{a}_{y,i} \hat{a}_{y,i}^\top \right) &= \operatorname{trace}\log \left[\sum_{i=1}^n (\zeta_{y,i} + \delta h^\top \nabla \zeta_{y,i}) \left(1 + \frac{2\delta a_i^\top h}{s_{y,i}} \right) \hat{a}_{y,i} \hat{a}_{y,i}^\top \right] + \mathcal{O}\left(\delta^2\right) \\
&= \operatorname{trace} \left[\sum_{i=1}^n \delta \left(\frac{2a_i^\top h}{s_{y,i}} + h^\top \nabla \zeta_{y,i} \right) \hat{a}_{y,i} \hat{a}_{y,i}^\top \right] + \mathcal{O}\left(\delta^2\right) \\
&= \sum_{i=1}^n \delta \left(\frac{2a_i^\top h}{s_{y,i}} + h^\top \nabla \zeta_{y,i} \right) \theta_{y,i} + \mathcal{O}\left(\delta^2\right)
\end{aligned}$$

where in the last step we have used the fact that $\operatorname{trace}(\hat{a}_{y,i} \hat{a}_{y,i}^\top) = \hat{a}_{y,i}^\top \hat{a}_{y,i} = \theta_{y,i}$ for each $i \in [n]$. Substituting the expression for $\nabla \zeta_y$ from part (a), and rearranging the terms yields the claimed expression in the limit $\delta \to 0$.

Proof of part (d)—Gradient of φ**:** Using the chain rule and the fact that $\nabla s_{y,i} = -a_i$, yields the result.

Proof of part (e): We claim that

$$\frac{1}{2}h^\top \nabla^2 \Psi_y h = \frac{1}{2}\left[\sum_{i=1}^n \zeta_{y,i}\theta_{y,i}(3d_{y,i}^2 + 2d_{y,i}f_{y,i} + \ell_{y,i}) - \frac{1}{2}\sum_{i,j=1}^n \zeta_{y,i}\zeta_{y,j}\theta_{y,i,j}^2\left(2d_{y,i}+f_{y,i}\right)\left(2d_{y,j}+f_{y,j}\right)\right].$$

The desired bound on $\left|h^\top \nabla^2 \Psi_y h\right|/2$ now follows from an application of AM-GM inequality with Lemma E.5(d).

We now derive the claimed expression for the directional Hessian of the function Ψ. We have

$$\frac{1}{2}h^\top\left(\nabla^2 \log\det J_y\right)h = \lim_{\delta\to 0}\frac{1}{2\delta^2}(\log\det J_y^{-1/2}J_{y+\delta h}J_y^{-1/2} + \log\det J_y^{-1/2}J_{y-\delta h}J_y^{-1/2} - 2\log\det \mathbb{I}_d)$$

$$= \frac{1}{2}\lim_{\delta\to 0}\frac{1}{\delta^2}\left[\operatorname{trace}\log\left(\sum_{i=1}^n \frac{\zeta_{y+\delta h}}{(1-\delta a_i^\top h/s_{y,i})}\hat{a}_{y,i}\hat{a}_{y,i}^\top\right) + \operatorname{trace}\log\left(\sum_{i=1}^n \frac{\zeta_{y-\delta h}}{(1+\delta a_i^\top h/s_{y,i})}\hat{a}_{y,i}\hat{a}_{y,i}^\top\right)\right].$$

Expanding the first term in the above expression, we find that

$$\operatorname{trace}\log\left(\sum_{i=1}^n \frac{\zeta_{y+\delta h,i}}{(1-\delta a_i^\top h/s_{y,i})}\hat{a}_{y,i}\hat{a}_{y,i}^\top\right)$$

$$= \operatorname{trace}\log\underbrace{\left[\sum_{i=1}^n\left(\zeta_{y,i} + \delta h^\top\nabla\zeta_{y,i} + \frac{\delta^2}{2}h^\top\nabla^2\zeta_{y,i}h\right)\left(1 + 2\delta\frac{a_i^\top h}{s_{y,i}} + 3\delta^2\frac{(a_i^\top h)^2}{s_{y,i}^2}\right)\hat{a}_{y,i}\hat{a}_{y,i}^\top\right]}_{=:\mathbb{I}_d + B} + \mathcal{O}\left(\delta^3\right).$$

Substituting the shorthand notation from equations (E.20a), (E.20b) and (E.20c), we have

$$B = \sum_{i=1}^n \zeta_{y,i}\left[\delta(2d_{y,i}+f_{y,i}) + \delta^2(3d_{y,i}^2 + 2d_{y,i}f_{y,i} + \ell_{y,i})\right]\hat{a}_{y,i}\hat{a}_{y,i}^\top + \mathcal{O}\left(\delta^3\right).$$

Now we make use of the following facts: (1) $\operatorname{trace}\log(\mathbb{I}_d + B) = \operatorname{trace}\left[B - \frac{B^2}{2} + \mathcal{O}\left(\|B\|^3\right)\right]$, (2) for each $i, j \in [n]$, we have $\operatorname{trace}(\hat{a}_{y,i}\hat{a}_j^\top) = \hat{a}_{y,i}^\top\hat{a}_j = \theta_{y,i,j}$, and (3) for each $i \in [n]$, we have $\theta_{y,i,i} = \theta_{y,i}$. Thus we obtain

$$\operatorname{trace}\log\left(\sum_{i=1}^n \frac{\zeta_{y+\delta h,i}}{(1-\delta a_i^\top h/s_{y,i})}\hat{a}_{y,i}\hat{a}_{y,i}^\top\right) = \sum_{i=1}^n \zeta_{y,i}\theta_{y,i}\left[\delta(2d_{y,i}+f_{y,i}) + \delta^2(3d_{y,i}^2 + 2d_{y,i}f_{y,i} + \ell_{y,i})\right]$$

$$- \frac{1}{2}\sum_{i,j=1}^n \zeta_{y,i}\zeta_{y,j}\theta_{y,i,j}^2\delta^2(2d_{y,i}+f_{y,i})(2d_{y,j}+f_{y,j}) + \mathcal{O}\left(\delta^3\right).$$

Similarly, we can obtain an expression for $\operatorname{trace}\log\left(\sum_{i=1}^n \frac{\zeta_{y-\delta h}}{(1+\delta a_i^\top h/s_{y,i})}\hat{a}_{y,i}\hat{a}_{y,i}^\top\right)$. Putting the pieces together, we obtain

$$\frac{1}{2}h^\top\left(\nabla^2\log\det J_y\right)h = \sum_{i=1}^n \zeta_{y,i}\theta_{y,i}(3d_{y,i}^2 + 2d_{y,i}f_{y,i} + \ell_{y,i}) - \frac{1}{2}\sum_{i,j=1}^n \zeta_{y,i}\zeta_{y,j}\theta_{y,i,j}^2(2d_{y,i}+f_{y,i})(2d_{y,j}+f_{y,j}). \tag{E.53}$$

Proof of part (f): We claim that

$$\frac{1}{2}h^\top \nabla^2 \varphi_{y,i} h = \varphi_{y,i}\left(2d_{y,i}f_{y,i} + 3d_{y,i}^2 + \ell_{y,i}\right). \tag{E.54}$$

The claim follows from a straightforward application of chain rule and substitution of the expressions for $\nabla\zeta_{y,i}$ and $\nabla^2\zeta_{y,i}$ in terms of the shorthand notation $d_{y,i}$, $f_{y,i}$ and $\ell_{y,i}$. Multiplying both sides of equation (E.54) with $d_{y,i}^2 s_{y,i}^2$ and summing over index i, we find that

$$\begin{aligned}\sum_{i=1}^n d_{y,i}^2 s_{y,i}^2 \frac{1}{2} h^\top \nabla \varphi_{y,i}^2 h = \sum_{i=1}^n d_{y,i}^2 s_{y,i}^2 \varphi_{y,i}\left[\ell_{y,i} + 2d_{y,i}f_{y,i} + 3d_{y,i}^2\right] &= \sum_{i=1}^n d_{y,i}^2 \zeta_{y,i}\left[\ell_{y,i} + 2d_{y,i}f_{y,i} + 3d_{y,i}^2\right] \\ &\leq \sum_{i=1}^n d_{y,i}^2 \zeta_{y,i}\left[\ell_{y,i} + f_{y,i}^2 + 4d_{y,i}^2\right],\end{aligned}$$

where in the last step we have used the AM-GM inequality. The claim follows.

E.5.5.2 Proof of Lemma E.8

We claim that

$$0 \preceq G_y^{-1/2}\left(c_1\mathbb{I}_n + c_2\Lambda_y\left(G_y - \alpha\Lambda_y\right)^{-1}\right)G_y^{1/2} \preceq (c_1 + c_2)\,\kappa\mathbb{I}_n. \tag{E.55}$$

The proof of the lemma is immediate from this claim, as for any PSD matrix $H \leq c\mathbb{I}_n$, we have $H^2 \leq c^2\mathbb{I}_n$.

We now prove claim (E.55). Note that

$$G_y^{-1/2}\Lambda_y\left(G_y - \alpha\Lambda_y\right)^{-1}G_y^{1/2} = \underbrace{G_y^{-1/2}\Lambda_y G_y^{-1/2}}_{:=B_y}(\mathbb{I}_n - \alpha_{\mathrm{J}} G_y^{-1/2}\Lambda_y G_y^{-1/2})^{-1}. \tag{E.56}$$

Note that the RHS is equal to the matrix $B_y(\mathbb{I}_n - \alpha_{\mathrm{J}} B_y)^{-1}$ which is symmetric. Observe the following ordering of the matrices in the PSD cone

$$\Sigma_y + \beta_{\mathrm{J}}\mathbb{I}_n = G_y \succeq \Sigma_y \succeq \Lambda_y = \Sigma_y - \Upsilon_y^{(2)} \succeq 0.$$

For the last step we have used the fact that $\Sigma_y - \Upsilon_y^{(2)}$ is a diagonally dominant matrix with non negative entries on the diagonal to conclude that it is a PSD matrix. Consequently, we have

$$B_y = G_y^{-1/2}\Lambda_y G_y^{-1/2} \preceq \mathbb{I}_n. \tag{E.57}$$

Further, recall that $\alpha_{\mathrm{J}} = (1 - 1/\kappa) \Leftrightarrow \kappa = (1 - \alpha_{\mathrm{J}})^{-1}$. As s result, we obtain

$$0 \preceq (\mathbb{I}_n - \alpha_{\mathrm{J}} G_y^{-1/2}\Lambda_y G_y^{-1/2})^{-1} \preceq \kappa\mathbb{I}_n.$$

Multiplying both sides by $B_y^{1/2}$ and using the relation (E.57), we obtain

$$0 \preceq B_y^{1/2}(\mathbb{I}_n - \alpha_{\mathrm{J}} G_y^{-1/2}\Lambda_y G_y^{-1/2})^{-1}B_y^{1/2} \preceq \kappa\mathbb{I}_n. \tag{E.58}$$

Using the fact that B_y commutes with $(\mathbb{I}_n - B_y)^{-1}$, we obtain $B_y(\mathbb{I}_n - \alpha_{\mathrm{J}} B_y)^{-1} \preceq \kappa\mathbb{I}_n$. Using observation (E.56) now completes the proof.

E.5.5.3 Proof of Lemma E.9

Without loss of generality, we can first prove the result for $i = 1$. Let $\nu := \mu_y^\top e_1$ denote the first row of the matrix μ_y. Observe that

$$e_1 = (G_y - \alpha\Lambda_y)\, G_y^{-1}\nu = \nu - \alpha\Sigma_y G_y^{-1}\nu + \alpha\Upsilon_y^{(2)} G_y^{-1}\nu \tag{E.59}$$

We now prove bounds (E.24a) and (E.24b) separately.

Proof of bound (E.24a): Multiplying the equation (E.59) on the left by $\nu^\top G_y^{-1}$, we obtain

$$\begin{aligned} g_1^{-1}\nu_1 &= \nu^\top G_y^{-1}\nu - \alpha\nu^\top G_y^{-1}\Sigma_y G_y^{-1}\nu + \alpha\nu^\top G_y^{-1}\Upsilon_y^{(2)} G_y^{-1}\nu \\ &\geq \nu^\top G_y^{-1}\nu - \alpha\nu^\top G_y^{-1}\Sigma_y G_y^{-1}\nu \\ &\geq \left(g_1^{-1} - \alpha\sigma_{y,1}/g_1^2\right)\nu_1^2. \end{aligned} \tag{E.60}$$

Rearranging terms, we obtain

$$0 \leq \nu_1 \leq \frac{\zeta_{y,1}}{\zeta_{y,1} - \alpha\sigma_{y,1}} \overset{(i)}{\leq} \kappa, \tag{E.61}$$

where inequality (i) follows from the facts that $\zeta_{y,j} \geq \sigma_{y,j}$ and $(1-\alpha) = \kappa$.

Proof of bound (E.24b): In our proof, we use the following improved lower bound for the term $\mu_{y,1,1} = \nu_1$.

$$\nu_1 \geq \frac{\zeta_{y,1}}{\zeta_{y,1} - \alpha\sigma_{y,1} + \alpha\sigma_{y,1}^2}, \tag{E.62}$$

Deferring the proof of this claim at the moment, we now complete the proof.

We begin by deriving a weighted ℓ_2-norm bound for the vector $\widetilde{\nu} = (\nu_2, \ldots, \nu_n)^\top$. Equation (E.60) implies

$$\zeta_{y,1}^{-1}\nu_1\left(1 - \nu_1 + \alpha\frac{\sigma_{y,1}}{\zeta_{y,1}}\nu_1\right) \geq \sum_{j=2}^{n}\nu_j^2\left(\zeta_{y,j}^{-1} - \alpha\zeta_{y,j}^{-2}\sigma_{y,j}\right) \overset{(i)}{\geq} (1-\alpha)\sum_{j=2}^{n}\frac{\nu_j^2}{\zeta_{y,j}},$$

where step (i) follows from the fact that $\zeta_{y,i} \geq \sigma_{y,i}$. Now, we upper bound the expression on the left hand side of the above inequality using the upper (E.61) and lower (E.62) bounds on ν_1:

$$\begin{aligned} \zeta_{y,1}^{-1}\nu_1\left(1 - \nu_1 + \alpha\frac{\sigma_{y,1}}{\zeta_{y,1}}\nu_1\right) &\leq \zeta_{y,1}^{-1}\frac{\zeta_{y,1}}{\zeta_{y,1} - \alpha\sigma_{y,1}}\left(1 - \left(1 - \alpha\frac{\sigma_{y,1}}{\zeta_{y,1}}\right)\frac{\zeta_{y,1}}{\zeta_{y,1} - \alpha\sigma_{y,1} + \alpha\sigma_{y,1}^2}\right) \\ &= \frac{\alpha\sigma_{y,1}^2}{\left(\zeta_{y,1} - \alpha\sigma_{y,1}\right)\left(\zeta_{y,1} - \alpha\sigma_{y,1} + \alpha\sigma_{y,1}^2\right)} \\ &\leq \kappa^2, \end{aligned}$$

where in the last step we have used the facts that $\zeta_{y,1} \geq \sigma_{y,1}$ and $(1-\alpha)^{-1} = \kappa$. Putting the pieces together, we obtain $\sum_{j=2}^n \nu_j^2 \zeta_{y,j}^{-1} \leq \kappa^3$, which is equivalent to our claim (E.24b) for $i = 1$. Since the choice of $i = 1$ was arbitrary, the claim (E.24b) follows.

It remains to prove our earlier claim (E.62). Writing equation (E.59) separately for the first coordinate and for the rest of the coordinates, we obtain

$$1 = \left(1 - \alpha\sigma_{y,1}\zeta_{y,1}^{-1} + \alpha\sigma_{y,1,1}^2\zeta_{y,j}^{-1}\right)\nu_1 + \alpha\sum_{j=2}^n \sigma_{y,1,j}^2\zeta_{y,j}^{-1}\nu_j, \quad \text{and} \tag{E.63a}$$

$$0 = \left(\mathbb{I}_{n-1} - \alpha\Sigma_y' G_y'^{-1}\right)\begin{pmatrix}\nu_2\\ \vdots \\ \nu_n\end{pmatrix} + \alpha\Upsilon_y'^{(2)}G_y'^{-1}\begin{pmatrix}\nu_2\\ \vdots \\ \nu_n\end{pmatrix} + \alpha\zeta_{y,1}^{-1}\nu_1\begin{pmatrix}\sigma_{y,1,2}^2\\ \vdots \\ \sigma_{y,1,n}^2\end{pmatrix}, \tag{E.63b}$$

where G_y' (respectively Σ_y', $\Upsilon_y'^{(2)}$) denotes the principal minor of G_y (respectively Σ_y, $\Upsilon_y^{(2)}$) obtained by excluding the first column and the first row. Multiplying both sides of the equation (E.63b) from the left by $\left(\nu_2, \cdots, \nu_n\right) G_y'^{-1}$, we obtain

$$0 = \sum_{j=2}^n \underbrace{\frac{1}{\zeta_{y,j}}\left(1 - \frac{\alpha\sigma_{y,j}}{\zeta_{y,j}}\right)}_{c_{y,j}}\nu_j^2 + \alpha\underbrace{\left(\nu_2, \cdots, \nu_n\right) G_y'^{-1}\Upsilon_y'^{(2)}G_y'^{-1}\begin{pmatrix}\nu_2\\ \vdots \\ \nu_n\end{pmatrix}}_{C_{y.2}} + \alpha\frac{\nu_1}{\zeta_{y,1}}\sum_{j=2}^n \frac{\sigma_{y,j}^2}{\zeta_{y,j}}\nu_j. \tag{E.64}$$

Observing that $\alpha \in [0,1]$ and $\zeta_{y,j} \geq \sigma_{y,j}$ for all $y \in \text{int}(\mathcal{K})$ and $j \in [n]$, we obtain $c_{y,j} \geq 0$. Further, note that $G_y'^{-1}\Upsilon_y'^{(2)}G_y'^{-1}$ is a PSD matrix and hence we have that $C_{y,2} \geq 0$. Putting the pieces together, we have

$$\alpha\frac{\nu_1}{\zeta_{y,1}}\sum_{j=2}^n \frac{\sigma_{y,j}^2}{\zeta_{y,j}}\nu_j \leq 0.$$

Combining this inequality with equation (E.63a) yields the claim.

E.5.5.4 Proof of Corollary E.1

Without loss of generality, we can prove the result for $i = 1$. Applying Cauchy-Schwarz inequality, we have

$$\left\|\nu\right\|_1 = \nu_1 + \sum_{j=2}^n |\nu_j| \leq \nu_1 + \sqrt{\sum_{j=2}^n \frac{\nu_j^2}{\zeta_{y,j}} \cdot \sum_{j=2}^n \zeta_{y,j}} \leq \kappa + \kappa^{3/2}\cdot\sqrt{1.5\ d} \leq 3\sqrt{d}\kappa^{3/2},$$

where to assert the last inequality we have used Lemma E.9 and Lemma E.1(c). The claim (E.25) follows. Further, noting that the infinity norm of a matrix is the ℓ_1-norm of its transpose, we obtain $\vert\vert\vert \left(G_y - \alpha\Lambda_y\right)^{-1} G_y \vert\vert\vert_\infty \leq 3\sqrt{d}\kappa^{3/2}$ as claimed.

E.5.6 Proofs of Lemmas from Section E.5.2

In this section, we collect proofs of auxiliary lemmas from Section E.5.2.

E.5.6.1 Proof of Lemma E.10

Using Lemma E.6, and the relation (E.14) we have

$$\left(1 - \frac{s_{z,i}}{s_{x,i}}\right)^2 \leq 4\frac{r^2}{\kappa^4 d^{3/2}}\xi^\top \xi, \tag{E.65}$$

where $\xi \sim \mathcal{N}(0, \mathbb{I}_d)$. Define

$$\Delta_s \coloneqq \max_{i\in[n],\ v\in\overline{xz}} \left|1 - \frac{s_{v,i}}{s_{x,i}}\right|. \tag{E.66}$$

Using the standard Gaussian tail bound, we observe that $\mathbb{P}_{\xi\sim\mathcal{N}(0,\mathbb{I}_n)}\left[\xi^\top\xi \geq d(1+\delta)\right] \leq 1 - \epsilon/4$ for $\delta = \sqrt{\frac{2}{d}}$. Plugging this bound in the inequality (E.65) and noting that for all $v \in \overline{xz}$ we have $\left\|v - x\right\|_{J_x} \leq \left\|z - x\right\|_{J_x}$, we obtain that

$$\mathbb{P}_{z\sim\mathcal{P}_x}\left[\Delta_s \leq \frac{2r^2(1+\sqrt{2/d}\log(4/\epsilon)}{\kappa^4\sqrt{d}}\right] \geq 1 - \epsilon/4.$$

Setting

$$r \leq 1/(25\sqrt{1+\sqrt{2}\log(4/\epsilon)}), \tag{E.67}$$

and noting that $\kappa^4\sqrt{d} \geq 1$ implies the claim (E.26a). Hence, we obtain that $\Delta_s < .005/\kappa^2$ and consequently $\max_{i\in[n],v\in\overline{xz}} s_{x,i}/s_{v,i} \in (0.99, 1.01)$ with probability at least $1 - \epsilon/4$.

We now claim that

$$\max_{i\in[n],v\in\overline{xz}} \frac{\zeta_{x,i}}{\zeta_{v,i}} \in \left[1 - 24\kappa^2\Delta_s, 1 + 24\kappa^2\Delta_s\right], \qquad \text{if } \Delta_s \leq \frac{1}{32\kappa^2}.$$

The result follows immediately from this claim. To prove the claim, note that equation (E.30) implies that if $\Delta_s \leq \frac{1}{32\kappa^2}$, then

$$\frac{\zeta_{v,i}}{\zeta_{x,i}} \in (e^{-8\kappa^2\Delta_s}, e^{8\kappa^2\Delta_s}) \quad \text{for all } i \in [n] \text{ and } v \in \overline{xz},$$

which implies that

$$\max_{i\in[n],v\in\overline{xz}} \frac{\zeta_{x,i}}{\zeta_{v,i}} \in (e^{-8\kappa^2\Delta_s}, e^{8\kappa^2\Delta_s}).$$

Asserting the facts that $e^x \leq 1 + 3x$ and $e^{-x} \geq 1 - 3x$, for all $x \in [0, 1]$ yields the claim.

E.5.6.2 Proof of Lemma E.11

The proof once again makes use of the classical tail bounds for polynomials in Gaussian random variables. We restate the classical result stated in equation (D.41) for convenience. For any $d \geq 1$, any polynomial $P : \mathbb{R}^d \to \mathbb{R}$ of degree k, and any $t \geq (2e)^{k/2}$, we have

$$\mathbb{P}\left[|P(\xi)| \geq t \left(\mathbb{E}P(\xi)^2\right)^{\frac{1}{2}}\right] \leq \exp\left(-\frac{k}{2e}t^{2/k}\right), \tag{E.68}$$

where $\xi \sim \mathcal{N}(0, \mathbb{I}_d)$ denotes the standard Gaussian vector in d dimensions.

Recall the notation from equation (E.22) and observe that

$$\|\hat{a}_{x,i}\|_2^2 = \theta_{x,i}, \quad \text{and} \quad \hat{a}_{x,i}^\top \hat{a}_{x,j} = \theta_{x,i,j}. \tag{E.69}$$

We also have

$$\sum_{i=1}^n \zeta_{x,i}\hat{a}_{x,i}\hat{a}_{x,i}^\top = J_x^{-1/2}\sum_{i=1}^n \zeta_{x,i}\frac{a_i a_i^\top}{s_{x,i}^2}J_x^{-1/2} = \mathbb{I}_d. \tag{E.70}$$

Further, using Lemma E.8 we obtain

$$\sum_{i=1}^n \zeta_{x,i}\hat{b}_{x,i}\hat{b}_{x,i}^\top = J_x^{-1/2}A_x\Lambda_x\left(G_x - \alpha\Lambda_x\right)^{-1}G_x\left(G_x - \alpha\Lambda_x\right)^{-1}\Lambda_x A_x^\top J_x^{-1/2} = 4\kappa^2\mathbb{I}_d. \tag{E.71}$$

Throughout this section, we consider a fixed point $x \in \text{int}\,(\mathcal{K})$. For brevity in our notation, we drop the dependence on x for terms like $\zeta_{x,i}, \theta_{x,i}, \hat{a}_{x,i}$ (etc.) and denote them simply by $\zeta_i, \theta_i, \hat{a}_i$ respectively.

We introduce some matrices and vectors that would come in handy for our proofs.

$$B = \begin{bmatrix}\sqrt{\zeta_1}\hat{a}_1^\top \\ \vdots \\ \sqrt{\zeta_n}\hat{a}_n^\top\end{bmatrix}, \quad B_b = \begin{bmatrix}\sqrt{\zeta_1}\hat{b}_1^\top \\ \vdots \\ \sqrt{\zeta_n}\hat{b}_n^\top\end{bmatrix}, \quad v = \begin{bmatrix}\sqrt{\zeta_1}\,\|\hat{a}_1\|_2^2 \\ \vdots \\ \sqrt{\zeta_n}\,\|\hat{a}_n\|_2^2\end{bmatrix}, \quad \text{and} \quad v^{ab} = \begin{bmatrix}\sqrt{\zeta_1}\hat{a}_1^\top\hat{b}_1 \\ \vdots \\ \sqrt{\zeta_n}\hat{a}_n^\top\hat{b}_n\end{bmatrix}. \tag{E.72}$$

We claim that

$$BB^\top \preceq \mathbb{I}_n, \quad \text{and} \quad B_bB_b^\top \preceq 4\kappa^2\mathbb{I}_n. \tag{E.73a}$$

To see these claims, note that equation (E.70) implies that $B^\top B = \mathbb{I}_d$ and consequently, $BB^\top$ is an orthogonal projection matrix and $BB^\top \preceq \mathbb{I}_n$. Next, note that from equation (E.71) we have that $B_b^\top B_b \preceq \kappa^2\mathbb{I}_d$, which implies that $B_bB_b^\top \preceq \kappa^2\mathbb{I}_n$. In asserting both these arguments, we have used the fact that for any matrix B, the matrices $BB^\top$ and $B^\top B$ are PSD and have same set of eigenvalues.

Next, we bound the ℓ_2 norm of the vectors v and v^{ab}:

$$\|v\|_2^2 = \sum_{i=1}^n \zeta_i \theta_i^2 \overset{\text{Lem. E.5 }(e)}{\leq} 4d, \quad \text{and} \tag{E.73b}$$

$$\left\|v^{ab}\right\|_2^2 = \sum_{i=1}^n \zeta_i \left(\hat{a}_i^\top \hat{b}_i\right)^2 \leq \sum_{i=1}^n \zeta_i \left\|\hat{a}_i\right\|_2^2 \left\|\hat{b}_i\right\|_2^2 \leq 4 \sum_{i=1}^n \zeta_i \left\|\hat{b}_i\right\|_2^2 = 4\,\text{trace}(B_b^\top B_b) \overset{\text{eqn. (E.73a)}}{\leq} 16\kappa^2 d. \tag{E.73c}$$

We now prove the five claims of the lemma separately.

Proof of bound (E.27a): Using Isserlis' theorem [128] for fourth order Gaussian moments, we have

$$\mathbb{E}\left(\sum_{i=1}^n \zeta_i \left(\hat{a}_i^\top \xi\right)^2\right)^2 = \sum_{i,j=1}^n \zeta_i \zeta_j \left(\|\hat{a}_i\|_2^2 \, \|\hat{a}_j\|_2^2 + 2\left(\hat{a}_i^\top \hat{a}_j\right)^2\right) = \sum_{i,j=1}^n \zeta_i \zeta_j \left(\theta_i \theta_j + 2\theta_{i,j}^2\right) \leq 24d^2,$$

where the last follows from Lemma E.5. Applying the bound (E.68) with $k = 2$ and $t = e\log(\frac{16}{\epsilon})$. Note that the bound is valid since $t \geq (2e)$ for all $\epsilon \in (0, 1/30]$.

Proof of bound (E.27b): Applying Isserlis' theorem for Gaussian moments, we obtain

$$\mathbb{E}\left(\sum_{i=1}^n \zeta_i \left(\hat{a}_i^\top \xi\right)^3\right)^2 = 9 \underbrace{\sum_{i,j=1}^n \zeta_i \zeta_j \, \|\hat{a}_i\|_2^2 \, \|\hat{a}_j\|_2^2 \left(\hat{a}_i^\top \hat{a}_j\right)}_{=:N_1} + 6 \underbrace{\sum_{i,j=1}^n \zeta_i \zeta_j \left(\hat{a}_i^\top \hat{a}_j\right)^3}_{=:N_2}.$$

We claim that $N_1 \leq 4d$ and $N_2 \leq 4d$. Assuming these claims as given at the moment, we now complete the proof. We have $\mathbb{E}\left(\sum_{i=1}^n \zeta_i \left(\hat{a}_i^\top \xi\right)^3\right)^2 \leq 60d$. Applying the bound (E.68) with $k = 3$ and $t = \left(\frac{2e}{3}\log\left(\frac{16}{\epsilon}\right)\right)^{3/2}$, and verifying that $t \geq (2e)^{3/2}$ for $\epsilon \in (0, 1/30]$ yields the claim.

We now turn to prove the bounds on N_1 and N_2. We have

$$N_1 = \sum_{i,j=1}^n \zeta_i \|\hat{a}_i\|_2^2 \, \hat{a}_i^\top \zeta_j \, \|\hat{a}_j\|_2^2 \, \hat{a}_j = \left\|\sum_{i=1}^n \zeta_i \, \|\hat{a}_i\|_2^2 \, \hat{a}_i\right\|_2^2 = \left\|B^\top v\right\|_2^2 \overset{\text{eqn. (E.73a)}}{\leq} \|v\|_2^2 \overset{\text{eqn. (E.73b)}}{\leq} 4d.$$

Next, applying Cauchy-Schwarz inequality and using equation (E.69), we obtain

$$N_2 = \sum_{i,j=1}^n \zeta_i \zeta_j \left(\hat{a}_i^\top \hat{a}_j\right)^3 \leq \sum_{i,j=1}^n \zeta_i \zeta_j \theta_{i,j}^2 \sqrt{\theta_i \theta_j} \overset{\text{(Lem. E.1 }(d))}{\leq} 4 \sum_{i,j=1}^n \zeta_i \zeta_j \theta_{i,j}^2 \overset{\text{(Lem. E.5 }(d))}{\leq} 4 \sum_{i=1}^n \zeta_i \theta_i = 4d.$$

Proof of bound (E.27c): Using Isserlis' theorem for Gaussian moments, we have

$$\mathbb{E}\left(\sum_{i=1}^{n}\zeta_i\left(\hat{a}_i^\top\xi\right)^2\left(\hat{b}_{x,i}^\top\xi\right)\right)^2 = \underbrace{\sum_{i,j=1}^{n}\zeta_i\zeta_j\left\|\hat{a}_i\right\|_2^2\left\|\hat{a}_j\right\|_2^2\left(\hat{b}_i^\top\hat{b}_j\right)}_{:=N_3} + 4\underbrace{\sum_{i,j=1}^{n}\zeta_i\zeta_j\left(\hat{a}_i^\top\hat{a}_j\right)\left(\hat{a}_i^\top\hat{b}_i\right)\left(\hat{a}_j^\top\hat{b}_j\right)}_{:=N_4}$$

$$+4\underbrace{\sum_{i,j=1}^{n}\zeta_i\zeta_j\left\|\hat{a}_i\right\|_2^2\left(\hat{b}_i^\top\hat{a}_j\right)\left(\hat{a}_j^\top\hat{b}_j\right)}_{:=N_5} + 2\underbrace{\sum_{i,j=1}^{n}\zeta_i\zeta_j\left(\hat{a}_i^\top\hat{a}_j\right)^2\left(\hat{b}_i^\top\hat{b}_j\right)}_{:=N_6} + 4\underbrace{\sum_{i,j=1}^{n}\zeta_i\zeta_j\left(\hat{a}_i^\top\hat{a}_j\right)\left(\hat{a}_i^\top\hat{b}_j\right)\left(\hat{b}_i^\top\hat{a}_j\right)}_{:=N_7}$$

We claim that all terms $N_k \leq 16\kappa^2 d, k \in \{3,4,5,6,7\}$. Putting the pieces together, we have

$$\mathbb{E}\left(\sum_{i=1}^{n}\zeta_i\left(\hat{a}_i^\top\xi\right)^2\left(\hat{b}_{x,i}^\top\xi\right)\right)^2 \leq 240\kappa^2 d.$$

Applying the bound (E.68) with $k = 3$ and $t = \left(\frac{2e}{3}\log\left(\frac{16}{\epsilon}\right)\right)^{3/2}$ yields the claim. Note that for the given definition of t, we have $t \geq (2e)^{3/2}$ for $\epsilon \in (0, 1/30]$ so that the bound (E.68) is valid.

It is now left to prove the bounds on N_k for $k \in \{3,4,5,6,7\}$. We have

$$N_3 = \sum_{i,j=1}^{n}\zeta_i\|\hat{a}_i\|_2^2\,\hat{b}_i^\top\zeta_j\left\|\hat{a}_j\right\|_2^2\hat{b}_j = \left\|\sum_{i=1}^{n}\zeta_i\left\|\hat{a}_i\right\|_2^2\hat{b}_i\right\|_2^2 = \left\|B_b^\top v\right\|_2^2 \overset{\text{eqn. (E.73a)}}{\leq} 4\kappa^2\left\|v\right\|_2^2 \overset{\text{eqn. (E.73b)}}{=\leq} 16\kappa^2 d,$$

$$N_4 = \sum_{i,j=1}^{n}\zeta_i\zeta_j\left(\hat{a}_i^\top\hat{a}_j\right)\left(\hat{a}_i^\top\hat{b}_i\right)\left(\hat{a}_j^\top\hat{b}_j\right) = \left\|B^\top v^{ab}\right\|_2^2 \overset{\text{eqn. (E.73a)}}{\leq} \left\|v^{ab}\right\|_2^2 \overset{\text{eqn. (E.73c)}}{\leq} 16\kappa^2 d, \quad \text{and}$$

$$N_5 = \sum_{i,j=1}^{n}\zeta_i\zeta_j\left\|\hat{a}_i\right\|_2^2\left(\hat{b}_i^\top\hat{a}_j\right)\left(\hat{a}_j^\top\hat{b}_j\right) = \left(B^\top v^{ab}\right)^\top\left(B_b^\top v\right) \overset{\text{C−S}}{\leq} \left\|B^\top v^{ab}\right\|_2\left\|B_b^\top v\right\|_2 \leq 16\kappa^2 d.$$

For the term N_6, we have

$$\begin{aligned}
N_6 = \sum_{i,j=1}^{n}\zeta_i\zeta_j\left(\hat{a}_i^\top\hat{a}_j\right)^2\left(\hat{b}_i^\top\hat{b}_j\right) \quad &\overset{\text{(C−S)}}{\leq} \quad \frac{1}{2}\sum_{i,j=1}^{n}\zeta_i\zeta_j\left(\hat{a}_i^\top\hat{a}_j\right)^2\left(\left\|\hat{b}_i\right\|_2^2 + \left\|\hat{b}_j\right\|_2^2\right) \\
&\overset{\text{(symm.in } i,j)}{=} \quad \sum_{i,j=1}^{n}\zeta_i\zeta_j\left(\hat{a}_i^\top\hat{a}_j\right)^2\left\|\hat{b}_i\right\|_2^2 \\
&\overset{\text{(eqn. (E.70))}}{\leq} \quad \sum_{i=1}^{n}\zeta_i\left\|\hat{a}_i\right\|_2^2\left\|\hat{b}_i\right\|_2^2 \\
&\overset{\text{(Lem. E.1}(d))}{\leq} \quad 4\sum_{i=1}^{n}\zeta_i\left\|\hat{b}_i\right\|_2^2 \\
&\overset{\text{(eqn. (E.73c))}}{\leq} \quad 16\kappa^2 d.
\end{aligned}$$

The bound on the term N_7 can be obtained in a similar fashion.

Proof of bound (E.27d): Observe that $\hat{a}_i^\top \xi \sim \mathcal{N}(0, \theta_i)$ and hence $\mathbb{E}\left(\hat{a}_i^\top \xi\right)^8 = 105\,\theta_i^4$. Thus, we have

$$\mathbb{E}\left(\sum_{i=1}^n \zeta_i \left(\hat{a}_i^\top \xi\right)^4\right)^2 \overset{\text{C−S}}{\leq} \sum_{i,j=1}^n \zeta_i \zeta_j \left(\mathbb{E}\left(\hat{a}_i^\top \xi\right)^8\right)^{\frac{1}{2}} \left(\mathbb{E}\left(\hat{a}_j^\top \xi\right)^8\right)^{\frac{1}{2}} = 105 \sum_{i,j=1}^n \zeta_i \zeta_j \theta_i^2 \theta_j^2 = 105 \left(\sum_{i=1}^n \zeta_i \theta_i^2\right)^2.$$

Now applying Lemma E.5, we obtain that $\mathbb{E}\left(\sum_{i=1}^n \zeta_i \left(\hat{a}_i^\top \xi\right)^4\right)^2 \leq 1680 d^2$. Consequently, applying the bound (E.68) with $k = 4$ and $t = \left(\frac{\epsilon}{2} \log\left(\frac{16}{\epsilon}\right)\right)^2$ and noting that $t \geq (2e)^2$ for $\epsilon \in (0, 1/30]$, yields the claim.

Proof of bound (E.27e): Using the fact that $\mathbb{E}\left(\hat{a}_i^\top \xi\right)^{12} = 945\,\theta_i^6$ and an argument similar to the previous part yields that $\mathbb{E}\left(\sum_{i=1}^n \zeta_i \left(\hat{a}_i^\top \xi\right)^6\right)^2 \leq 15120 d^2$. Finally, applying the bound (E.68) with $k = 6$ and $t = \left(\frac{\epsilon}{3} \log\left(\frac{16}{\epsilon}\right)\right)^3$, and verifying that $t \geq (2e)^3$ for $\epsilon \in (0, 1/30]$, yields the claim.

Appendix F

Content Deferred From Chapter 7

In this chapter, we present the proofs of our main results, namely Theorems 7.1 to 7.4 inAppendices F.1, F.2, F.3.2 and F.4. We prove Lemma 7.1 in Appendix F.5, and the other intermediate results used within the proofs in Appendix F.6. We also present additional results in Appendix F.7, visualization of log-likelihoods Appendix F.8, and extension of our theory for mixture of regression case in Appendix F.9.

F.1 Proof of Theorem 7.1

As alluded to earlier in Section 7.4.1—given Lemma 7.1—it suffices to prove the contraction property (7.10a) for the population operator M. Recall that $\theta^* = 0$ is a fixed point of the population EM operator (i.e., $M(0) = 0$). This fact, combined with the definition (7.8) of the M-update, yields

$$\|M(\theta)\|_2 = \|M(\theta) - M(\theta^*)\|_2 = \|\mathbb{E}\left[2(w_\theta(X) - w_0(X))X\right]\|_2 ,$$

where, in the unbalanced setting (7.3), the weight function w_θ (7.6) and the gradient $\nabla_\theta w_\theta$ take the form

$$w_\theta(X) = \frac{\pi}{\pi + (1-\pi)e^{-\frac{2\theta^\top X}{\sigma^2}}}, \text{ and } \nabla_\theta w_\theta(X) = \frac{\frac{2\pi(1-\pi)X}{\sigma^2}}{\left(\pi e^{-\frac{\theta^\top X}{\sigma^2}} + (1-\pi)e^{\frac{\theta^\top X}{\sigma^2}}\right)^2}.$$

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

$$\begin{aligned}
\|M(\theta)\|_2 &= \left\| \mathbb{E}\left[2X \int_0^1 h'(u) du \right] \right\|_2 \\
&= 4\pi(1-\pi) \left\| \int_0^1 \mathbb{E}\left[\frac{XX^\top}{\sigma^2 \left(\pi \exp\left(-\frac{\theta_u^\top X}{\sigma^2}\right) + (1-\pi)\exp\left(\frac{\theta_u^\top X}{\sigma^2}\right)\right)^2} \right] \theta du \right\|_2 \\
&\leq 4\pi(1-\pi)\|\theta\|_2 \max_{u\in[0,1]} \left\|\mathbb{E}\left[\Gamma_{\theta_u}(X)\right]\right\|_{\text{op}},
\end{aligned} \tag{F.1}$$

where in the last equation we have defined the matrix

$$\Gamma_{\theta_u}(X) := \frac{XX^\top}{\sigma^2 \left(\pi \exp\left(-\frac{\theta_u^\top X}{\sigma^2}\right) + (1-\pi)\exp\left(\frac{\theta_u^\top X}{\sigma^2}\right)\right)^2}. \tag{F.2}$$

Writing the mixture weight as $\pi = \frac{1}{2}(1-\rho)$, we claim that it suffices to show that

$$\max_{u\in[0,1]} \left\|\mathbb{E}\left[\Gamma_{\theta_u}(X)\right]\right\|_{\text{op}} \leq \frac{1-\rho^2/2}{1-\rho^2}. \tag{F.3}$$

Indeed, taking the last bound as given and substituting it into inequality (F.1), we find that

$$\|M(\theta)\|_2 \leq 4\pi(1-\pi)\frac{1-\rho^2/2}{1-\rho^2}\ \|\theta\|_2 \ = \ (1-\rho^2/2)\|\theta\|_2,$$

which yields the claim (7.10a) of Theorem 7.1.

F.1.1 Proof of claim (F.3):

We begin by making a convenient change of coordinates. Let $R \in \mathbb{R}^{d\times d}$ 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 (F.2) 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 $\left\|\mathbb{E}\left[\Gamma_{\theta_u}(X)\right]\right\|_{\text{op}} = \|B_{\theta_u}\|_{\text{op}}$, where

$$B_{\theta_u} := \mathbb{E}_V\left[\frac{VV^\top}{\left(\pi\exp\left(-\|\theta_u\|_2 V_1/\sigma\right) + (1-\pi)\exp\left(\|\theta_u\|_2 V_1/\sigma\right)\right)^2} \right].$$

Here $V_1 := Ve_1$ denotes the first coordinate of the random vector V. Note that the matrix B_{θ_u} is a diagonal matrix, with non-negative entries. Thus, in order to prove the bound (F.3), it suffices to show that

$$\max_{j\in[d]}[B_{\theta_u}]_{jj} \le \frac{1-\rho^2/2}{1-\rho^2}. \tag{F.4}$$

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

$$\begin{aligned}
&(\pi e^{-y} + (1-\pi)e^y) \in [\sqrt{(1-\rho^2)},\, 1], \qquad &&\text{if } e^y \in \left[1, \frac{1+\rho}{1-\rho}\right], \quad \text{and}\\
&(\pi e^{-y} + (1-\pi)e^y) > 1, &&\text{otherwise.}
\end{aligned} \tag{F.5}$$

Let $\mathcal{E}^c$ and $\mathbb{I}(\mathcal{E})$ respectively denote the complement and the indicator of any event $\mathcal{E}$. Define the event

$$\mathcal{E}_{\theta_u} := \left\{ e^{\|\theta_u\|_2 V_1/\sigma} \in \left[1, \frac{1+\rho}{1-\rho}\right]\right\}.$$

Using the observation (F.5) above and the fact that $V_1 \sim \mathcal{N}(0,1)$, we obtain

$$\begin{aligned}
[B_{\theta_u}]_{11} &= \mathbb{E}\left[\frac{V_1^2}{\left(\pi \exp\left(-\|\theta_u\|_2 V_1/\sigma\right) + (1-\pi)\exp\left(\|\theta_u\|_2 V_1/\sigma\right)\right)^2}\right]\\
&\le \frac{1}{(1-\rho^2)}\mathbb{E}\left[V_1^2\, \mathbb{I}(\mathcal{E}_{\theta_u})\right] + \mathbb{E}\left[V_1^2\, \mathbb{I}(\mathcal{E}_{\theta_u}^c)\right]\\
&= \frac{1-\rho^2 + \rho^2 \mathbb{E}\left[V_1^2\, \mathbb{I}(\mathcal{E}_{\theta_u})\right]}{(1-\rho^2)}.
\end{aligned} \tag{F.6}$$

Note that whenever $\theta_u \neq 0$, we have that $\mathcal{E}_{\theta_u} \subseteq \{V_1 \ge 0\}$ and consequently, we obtain that

$$\mathbb{E}\left[V_1^2\, \mathbb{I}(\mathcal{E}_{\theta_u})\right] \le \mathbb{E}\left[V_1^2\, \mathbb{I}(V_1 \ge 0)\right] = \frac{1}{2}. \tag{F.7}$$

Putting the inequalities (F.6) and (F.7) together, we conclude that
$[B_{\theta_u}]_{11} \le (1-\rho^2/2)/(1-\rho^2)$.

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

$$[B_{\theta_u}]_{jj} = \mathbb{E}\left[\frac{V_j^2}{\left(\pi \exp\left(-\left\|\theta_u\right\|_2 V_1/\sigma\right) + (1-\pi)\exp\left(\left\|\theta_u\right\|_2 V_1/\sigma\right)\right)^2}\right]$$

$$= \mathbb{E}\left[\frac{1}{\left(\pi \exp\left(-\left\|\theta_u\right\|_2 V_1/\sigma\right) + (1-\pi)\exp\left(\left\|\theta_u\right\|_2 V_1/\sigma\right)\right)^2}\right].$$

Invoking the definition of the event $\mathcal{E}_{\theta_u}$, we have

$$[B_{\theta_u}]_{jj} \leq \frac{1}{(1-\rho^2)}\mathbb{E}\left[\mathbb{I}(\mathcal{E}_{\theta_u})\right] + \mathbb{E}\left[\mathbb{I}(\mathcal{E}_{\theta_u}^c)\right] = \frac{1-\rho^2+\rho^2\mathbb{E}\left[\mathbb{I}(\mathcal{E}_{\theta_u})\right]}{(1-\rho^2)}.$$

Finally, 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.

F.2 Proof of Theorem 7.2

We split our proof into two parts, which correspond to the upper bound (7.13a) and the lower bound (7.13b) respectively.

F.2.1 Proof of the upper bound (7.13a)

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 7.1 from the unbalanced setting, we obtain that

$$\begin{aligned}\left\|M(\theta)\right\|_2 &= \left\|\mathbb{E}\left[2X\int_0^1 w'_{\theta_u}(X)^\top\theta_u du\right]\right\|_2 \\ &= 4\left\|\int_0^1 \mathbb{E}\left[\frac{XX^\top}{\sigma^2\left(e^{-\theta_u^\top X/\sigma^2}+e^{\theta_u^\top X/\sigma^2}\right)^2}\right]\theta du\right\|_2 \qquad \text{(F.8)} \\ &\leq 4\left\|\theta\right\|_2 \int_0^1 \left\|\mathbb{E}\left[\Gamma_{\theta_u}(X)\right]\right\|_{\text{op}} du,\end{aligned}$$

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 (7.13a), it suffices to show that

$$\int_0^1 \left\|\mathbb{E}\left[\Gamma_{\theta_u}(X)\right]\right\|_{\text{op}} du \leq \frac{1}{4}\left(p + \frac{1-p}{1+\left\|\theta\right\|_2^2/2\sigma^2}\right) = \frac{\gamma_{\text{up}}(\theta)}{4} \qquad \text{(F.9)}$$

where $p := (1 + \mathbb{P}_{Z\sim\mathcal{N}(0,1)}(|Z| \leq 1))/2 < 1$.

We now establish the claim (F.9). Like in proof of Theorem 7.1, we perform a change of coordinates using an orthogonal matrix 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 the matrix $X = \sigma R^\top V$ and $R\theta_u = \|\theta_u\|_2 e_1$ in the expression for Γ_{θ_u} and using the equality $\|R^\top BR\|_{\text{op}} = \|B\|_{\text{op}}$, valid for any matrix B and any orthogonal matrix R, we obtain that $\|\mathbb{E}[\Gamma_{\theta_u}(X)]\|_{\text{op}} = \|B_{\theta_u}\|_{\text{op}}$, where

$$B_{\theta_u} := \mathbb{E}_V\left[\frac{VV^\top}{\left(\exp\left(-\|\theta_u\|_2 V_1/\sigma\right) + \exp\left(\|\theta_u\|_2 V_1/\sigma\right)\right)^2}\right]. \tag{F.10}$$

Clearly, the matrix B_{θ_u} is a diagonal matrix with non-negative entries (note the abuse of notation: the definitions of the matrices Γ_{θ_u} and B_{θ_u} is different from the unbalanced case). Consequently, to obtain a bound for the operator norm of the matrix B_{θ_u}, it is sufficient to provide an upper bound on the diagonal entries of the matrix B_{θ_u}. In order to do so, we introduce an auxiliary claim:

Lemma F.1. *The ℓ_2-operator norm of the matrix B_{θ_u} defined in equation (F.10), is upper-bounded as*

$$\|B_{\theta_u}\|_{op} = \max_{j\in[d]}[B_{\theta_u}]_{jj} \leq \frac{p_2}{4} + \frac{(1-p_2)}{4}\frac{1}{(1 + \|\theta_u\|_2^2/(2\sigma^2))^2}, \tag{F.11}$$

where $p_2 = \mathbb{P}(|V_1| \leq 1) < 1$.

See Appendix F.6.1 for the proof of Lemma F.1.

Using Lemma F.1, we now complete the proof. Integrating both sides of the inequality (F.11) with respect to $u \in [0, 1]$, we find that

$$\begin{aligned}\int_0^1 \|B_{\theta_u}\|_{\text{op}} du &\leq \int_0^1 \frac{p_2}{4} du + \int_0^1 \frac{(1-p_2)}{4}\frac{1}{(1 + \|\theta_u\|_2^2/(2\sigma^2))^2} du \\ &= \frac{p_2}{4} + \frac{(1-p_2)}{4}\int_0^1 \frac{1}{(1 + u^2\|\theta\|_2^2/(2\sigma^2))^2} du.\end{aligned}$$

Direct computation of the second integral yields

$$\begin{aligned}\int_0^1 \frac{1}{(1 + u^2\|\theta\|_2^2/(2\sigma^2))^2} du &= \frac{1}{2}\left(\frac{1}{1 + \frac{\|\theta\|^2}{2\sigma^2}} + \frac{\tan^{-1}(\|\theta\|/(\sqrt{2}\sigma))}{\|\theta\|/(\sqrt{2}\sigma)}\right) \\ &\leq \frac{1}{2}\left(\frac{1}{1 + \frac{\|\theta\|^2}{2\sigma^2}} + 1\right),\end{aligned}$$

where the last inequality above follows since $\tan^{-1}(y) \leq y$, for all $y \geq 0$. Putting together the pieces yields

$$\int_0^1 \|\mathbb{E}\left[\Gamma_{\theta_u}(X)\right]\|_{\text{op}} du = \int_0^1 \|B_{\theta_u}\|_{\text{op}} du \leq \frac{(1+p_2)}{8} + \frac{(1-p_2)/8}{1+\|\theta\|_2^2/(2\sigma^2)},$$

which implies the claim (F.9) with $p = \frac{1+p_2}{2}$.

F.2.2 Proof of the lower bound (7.13b)

We now prove the lower bound (7.13b) of Theorem 7.2 on the population EM operator M. The argument involves Jensen's inequality and certain properties of the moment generating function (MGF) of the Gaussian distribution.

Recalling equation (F.8), we find that

$$\begin{aligned}\|M(\theta)\|_2 &= 4\Bigg\| \underbrace{\int_0^1 \mathbb{E}\left[\frac{XX^\top}{\sigma^2\left(\exp\left(-\theta_u^\top X/\sigma^2\right)+\exp\left(\theta_u^\top X/\sigma^2\right)\right)^2}\right] du}_{=:\Gamma_\theta}\ \theta\Bigg\|_2 \\ &\geq 4\lambda_{\min}\left(\Gamma_\theta\right)\|\theta\|_2\,, \end{aligned} \tag{F.12}$$

where $\lambda_{\min}(\Gamma_\theta)$ denotes the smallest eigenvalue of the square matrix Γ_θ. Following the change of variable $V := RX/\sigma$ used in the proof of upper bound (7.13a), we obtain that

$$\lambda_{\min}\left(\Gamma_\theta\right) = \lambda_{\min}\underbrace{\left(\mathbb{E}_V\left[\int_0^1 \frac{VV^\top}{\left(\exp\left(-\|\theta_u\|_2 V_1/\sigma\right)+\exp\left(\|\theta_u\|_2 V_1/\sigma\right)\right)^2} du\right]\right)}_{=:F_\theta}. \tag{F.13}$$

Clearly, the matrix F_θ is a diagonal matrix with non-negative diagonal entries and consequently, we have

$$\lambda_{\min}(F_\theta) = \min_{j\in[d]}[F_\theta]_{jj}. \tag{F.14}$$

In order to provide a lower bound on the diagonal entries of the matrix F_θ, we use the following auxiliary claim:

Lemma F.2. *For all vectors $\theta \in \mathbb{R}^d$ such that $\|\theta\|_2^2 \leq \frac{5\sigma^2}{8}$, the matrix F_θ defined in equation (F.13), satisfies the bounds*

$$[F_\theta]_{jj} \geq [F_\theta]_{11} \geq \frac{1}{4(1+2\|\theta\|_2^2/\sigma^2)} \quad \textit{for all } j \in [d]. \tag{F.15}$$

See Appendix F.6.2 for the proof of this claim.

Finally, combining the result of Lemma F.2 with equations (F.12) and (F.14), we conclude that

$$\frac{\|M(\theta)\|_2}{\|\theta\|_2} \geq 4\lambda_{\min}\left(\Gamma_\theta\right) \;=\; 4[F_\theta]_{11} \;\geq\; \frac{1}{(1+2\left\|\theta\right\|_2^2/\sigma^2)} = \gamma_{\text{low}}(\theta),$$

as claimed.

F.3 Proof of Theorem 7.3

The reader should recall the framework that was laid out in Section 7.5.1, especially Lemma 7.1 which was used to bound the deviation between the sample and population EM operators, as well the annulus-based localization argument (that breaks up the iterations of EM into different epochs) sketched out in Section 7.5.2. The proof of Theorem 7.3 is based on making this proof outline more precise.

F.3.1 Epochs and non-expansivity

Let us introduce the notation required to formalize the analysis that leads to the recursion (7.23). Recall that the recursion (7.23) generates the sequence $\{\alpha_\ell\}_{\ell\geq 0}$ given by

$$\alpha_0 = 0 \quad \text{and} \quad \alpha_{\ell+1} = \frac{\alpha_\ell}{3} + \frac{1}{6}, \quad \text{for } \ell = 0, 1, 2, \ldots. \tag{F.16a}$$

By inspection, this sequence is increasing and satisfies $\lim\limits_{\ell\to\infty} \alpha_\ell = 1/4$. Furthermore, we have $\alpha_\ell \leq 1/4 - \varepsilon$ for $\ell \geq \lceil \log(4/\varepsilon)/\log 3 \rceil$. For any given $\delta \in (0,1)$, define the following intermediate quantity

$$\omega = \sigma^2 \left(\frac{d + \log((2\ell_\varepsilon + 1)/\delta)}{n}\right) \text{ where } \ell_\varepsilon := \lceil \log(4/\varepsilon)/\log 3 \rceil + 1. \tag{F.16b}$$

Note that the lower bound on the sample size stated in the theorem ensures that $\omega \leq 1$. For the proof sketch provided in Section 7.5.2, we used the rough approximation $\omega \approx d/n$, which is adequate when tracking only the dependency on the pair (n, d).

For $\ell = 0, 1, 2, \ldots, \ell_\varepsilon - 1$, define the scalars t_ℓ and T_ℓ as

$$t_0 = \left\lceil \frac{2}{p} \log \frac{\|\theta_0\|_2}{\sqrt{2}\sigma\sqrt{\omega}} \right\rceil, \quad t_\ell = \left\lceil \frac{2}{p\omega^{2\alpha_{\ell+1}}} \log(1/\omega) \right\rceil, \quad \text{and } T_\ell = \sum_{j=0}^{\ell} t_j, \tag{F.16c}$$

where $\lceil y \rceil$ denotes the smallest integer greater than or equal to y, and the constant $p \in (0,1)$ is given by $p = \mathbb{P}(|X| \leq 1) + \frac{1}{2}\mathbb{P}(|X| > 1)$ 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 7.1, stated in the main text, gives us a bound on $\sup_{\|\theta\|_2 \le r} \|M_n(\theta) - M(\theta)\|_2$ for a given radius r. In the epoch-based argument, we have a sequence of such radii (corresponding to the outer radii of the annulus considered in each epoch), so that we need to control this same quantity uniformly over all radii r in the set $\mathcal{R}$ given by

$$\mathcal{R} = \left\{ \left\|\theta^0\right\|_2, \sqrt{2}\sigma\omega^{\alpha_0}, \ldots, \sqrt{2}\sigma\omega^{\alpha_{\ell_\varepsilon - 1}}, c'\sqrt{2}\sigma\omega^{\alpha_0}, \ldots, c'\sqrt{2}\sigma\omega^{\alpha_{\ell_\varepsilon - 1}} \right\}. \tag{F.17}$$

Here $c' = (2c_1\sigma/p + 1)$ denotes a constant independent of n, d, δ and ε where c_1 is the universal constant that appeared in the bound from Lemma 7.1. In order to do so, we apply a standard union bound with Lemma 7.1 and obtain that

$$\sup_{\|\theta\|_2 \le r} \|M_n(\theta) - M(\theta)\|_2 \le c_1 \sigma r \sqrt{\omega} \quad \text{for all } r \in \mathcal{R}, \tag{F.18}$$

with probability at least $1 - \delta$. Let $\mathcal{E}(n, d, \varepsilon, \delta)$ denote the event that the bound (F.18) holds.

With this notation in place, we start with our first claim. The sample-based EM operator is *non-expansive* in the following sense:

Lemma F.3. *Consider the sample-based EM iteration $\theta^{t+1} = M_n(\theta^t)$ with a sample size $n \ge (2c_1\sigma/p)^{1/(2\varepsilon)}\sigma^2(d + \log((2\ell_\varepsilon + 1)/\delta))$. Suppose that there exists an index $\ell \in \{0, 1, \ldots, \ell_\varepsilon - 1\}$ and an iteration number t such that $\|\theta^t\|_2 \le \sqrt{2}\sigma\omega^{\alpha_\ell}$. Then, conditional on the event $\mathcal{E}(n, d, \varepsilon, \delta)$ from equation (F.18), we have*

$$\|\theta^{t'}\|_2 \le \sqrt{2}\sigma\omega^{\alpha_\ell} \qquad \textit{for all } t' \ge t. \tag{F.19}$$

See Appendix F.6.3 for the proof of this claim.

F.3.2 Core of the argument

We now proceed to the core of the argument. Suppose that the sample size is lower bounded as

$$n \ge \max\left\{ c_2, \left(\frac{2c_1\sigma}{p} + 1\right)^{\frac{4}{\varepsilon}}, \left(\frac{\sqrt{2}c_1\left\|\theta^0\right\|_2}{p} + 1\right)^2 \right\} \cdot \sigma^2 (d + \log\left(\frac{3\log(4/\varepsilon)}{\delta}\right)), \tag{F.20}$$

where the constants c_1 and c_2 correspond to that from Lemma 7.1. Moreover, recall that the quantity ω and the time-steps T_ℓ were defined in equations (F.16b) and (F.16c) respectively. The core of the proof consists of the following:

F.3.2.1 Key claim

For all $\ell \in \{0, 1, \ldots, \ell_\varepsilon - 1\}$, we have

$$\left\|\theta^t\right\|_2 \leq \sqrt{2}\sigma\omega^{\alpha_\ell} \qquad \text{for all } t \geq T_\ell, \tag{F.21}$$

with probability at least $1 - \delta$.

Taking this claim as given, let us now show how the bounds in Theorem 7.3 hold for all $t \geq T_{\ell_\varepsilon - 1}$. Straightforward computations yield that

$$\begin{aligned} T_{\ell_\varepsilon-1} &\leq T_0 + (\ell_\varepsilon - 1)t_{\ell_\varepsilon-1} \\ &\leq \frac{4}{p}\left[\log\frac{\|\theta_0\|_2}{\sqrt{2}\sigma\sqrt{\omega}} + \log\frac{4}{\varepsilon}\cdot\omega^{1/2-2\varepsilon}\cdot\log\frac{n}{\sigma^2 d}\right] \\ &\leq \frac{8}{p}\left[\log\frac{\|\theta_0\|_2^2 n}{\sigma^2 d} + \left(\frac{n}{d}\right)^{\frac{1}{2}-2\varepsilon}\cdot\log\left(\frac{4}{\varepsilon}\right)\cdot\log\left(\frac{n}{\sigma^2 d}\right)\cdot\sigma^{4\varepsilon-1}\right]. \end{aligned} \tag{F.22}$$

In other words, equations (F.20) and (F.22) 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 $(d/n)^{1/4-\varepsilon}$ around the truth $\theta^* = 0$.

F.3.2.2 Proof of the key claim (F.21)

We prove this claim by an induction on the epoch index ℓ. All of the argument are performed conditioned on the event $\mathcal{E}(n, d, \varepsilon, \delta)$ defined in equation (F.18); note that this event occurs with probability at least $1 - \delta$. Moreover, we see that the sample size assumption (F.20) for Theorem 7.3 is larger than required in Lemma F.3 and hence we can invoke the non-expansiveness of the sample-based EM operator in our arguments to follow.

Proof of base case: ($\ell = 0$): We adopt the shorthand $\nu = \|\theta_0\|_2/\sqrt{2}\sigma$. The non-expansiveness property of the sample-based EM-operator (Lemma F.3) 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

$$\left\|\theta^{t+1}\right\|_2 \leq \left\|M_n(\theta^t) - M(\theta^t)\right\|_2 + \left\|M(\theta^t)\right\|_2 \tag{F.23a}$$

$$\overset{(i)}{\leq} c_1\sigma\cdot\nu\sqrt{2}\sigma\cdot\sqrt{\omega} + \gamma_{\text{up}}(\theta^t)\left\|\theta^t\right\|_2, \tag{F.23b}$$

where step (i) follows from using $r = \nu\sqrt{2}\sigma$ in the event (F.18), and applying Theorem 7.2 (for the two terms respectively). Noting that $\|\theta^t\|_2 \geq \sqrt{2}\sigma$, we also have that

$$\gamma_{\text{up}}(\theta^t) = 1 - p + \frac{p}{1 + \|\theta^t\|_2^2/\sigma^2} = 1 - \frac{p\left\|\theta^t\right\|_2^2}{\|\theta^t\|_2^2 + 2\sigma^2} \leq \underbrace{1 - \frac{p}{2}}_{\overline{\gamma}_0}.$$

Recursing the inequalities (F.23a) and (F.23b) from $t = 0$ up to $t = T_0$, and using the fact that $\gamma_{\text{up}}(\theta^t) \leq \overline{\gamma}_0$, we find that

$$\begin{aligned}\left\|\theta^{T_0}\right\|_2 &\leq c_1\sigma \cdot \nu\sqrt{2}\sigma \cdot \sqrt{\omega}(1 + \overline{\gamma}_0 + \ldots + \overline{\gamma}_0^{T_0 - 1}) + \overline{\gamma}_0^{T_0}\left\|\theta^0\right\|_2 \\ &\leq \frac{c_1\sigma \cdot \nu\sqrt{2}\sigma \cdot \sqrt{\omega}}{1 - \overline{\gamma}_0} + \overline{\gamma}_0^{T_0}\nu\sqrt{2}\sigma.\end{aligned}$$

Substituting the expressions $\overline{\gamma}_0 = 1 - p/2$ and $T_0 = \lceil (2/p)\log(\nu/\sqrt{\omega})\rceil$, we obtain that

$$\left\|\theta^{T_0}\right\|_2 \leq (2\nu c_1\sigma/p + 1)\sqrt{\omega}\sqrt{2}\sigma \leq \sqrt{2}\sigma,$$

where the last inequality follows from the fact that for the assumed bound (F.20) on n, we have $(2\nu c_1\sigma/p + 1)\sqrt{\omega} \leq 1$. The base-case now follows.

Proof of inductive step: Now we prove the inductive step. In particular, we assume that $\left\|\theta^{T_\ell}\right\|_2 \leq \sqrt{2}\sigma\omega^{\alpha_\ell}$ and show that $\left\|\theta^{T_{\ell+1}}\right\|_2 \leq \sqrt{2}\sigma\omega^{\alpha_{\ell+1}}$. Once again, Lemma F.3 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

$$\gamma_{\text{up}}(\theta^t) \leq 1 - \frac{p\omega^{2\alpha_{\ell+1}}}{1 + \omega^{2\alpha_{\ell+1}}} \leq \underbrace{1 - \frac{p\omega^{2\alpha_{\ell+1}}}{2}}_{=:\overline{\gamma}_\ell} \quad \text{for all } t \in \{T_\ell, \ldots, T_{\ell+1} - 1\}, \tag{F.24}$$

where the last step follows from the fact that $\omega \in [0, 1]$ and $\alpha_\ell \geq 0$. From our earlier definition (F.16c), we have $T_{\ell+1} = T_\ell + t_\ell$. We split the remainder of our proof in two parts, primarily to handle the constants. First, we show that

$$\left\|\theta^{T_\ell + \lceil t_{\ell+1}/2\rceil}\right\|_2 \leq c'\sqrt{2}\sigma\omega^{\alpha_{\ell+1}}, \tag{F.25a}$$

where $c' = (2c_1\sigma/p + 1)$ is a constant independent of n, d, δ and ε. Next we use this result to show that

$$\left\|\theta^{T_{\ell+1}}\right\|_2 = \left\|\theta^{T_\ell + t_{\ell+1}}\right\|_2 \leq \sqrt{2}\sigma\omega^{\alpha_{\ell+1}}, \tag{F.25b}$$

which completes the proof of the induction step. We now prove these two claims one by one.

Proof of claim (F.25a): Applying the triangle inequality yields

$$\begin{aligned}\left\|\theta^{T_\ell+1}\right\|_2 &\leq \left\|M_n(\theta^{T_\ell}) - M(\theta^{T_\ell})\right\|_2 + \left\|M(\theta^{T_\ell})\right\|_2 \\ &\overset{(i)}{\leq} c_1\sigma \cdot \sqrt{2}\sigma\omega^{\alpha_\ell} \cdot \sqrt{\omega} + \gamma_{\text{up}}(\theta^{T_\ell})\left\|\theta^{T_\ell}\right\|_2,\end{aligned} \tag{F.26}$$

where step (i) follows from using $r = \sqrt{2}\sigma\omega^{\alpha_\ell}$ in the event (F.18) and applying Theorem 7.2. Recursing the inequality (F.26) for $T \leq \lceil t_\ell/2 \rceil$ steps, and invoking the bound (F.24), i.e., $\gamma_{\text{up}}(\theta^t) \leq \overline{\gamma}_\ell$ for all $t \in \{T_\ell, \ldots, T_\ell + T\}$, we obtain that

$$\begin{aligned}
\left\|\theta^{T_\ell+T}\right\|_2 &\leq c_1\sigma \cdot \sqrt{2}\sigma\omega^{\alpha_\ell} \cdot \sqrt{\omega} \cdot (1 + \overline{\gamma}_\ell + \ldots + \overline{\gamma}_\ell^{T-1}) + \overline{\gamma}_\ell^T \left\|\theta^{T_\ell}\right\|_2 \\
&\leq \frac{c_1\sigma \cdot \sqrt{2}\sigma\omega^{\alpha_\ell} \cdot \sqrt{\omega}}{1-\overline{\gamma}_\ell} + \sqrt{2}\sigma\overline{\gamma}_\ell^T\omega^{\alpha_\ell} \\
&\overset{(i)}{\leq} c_1\sigma \cdot \sqrt{2}\sigma \cdot (2/p) \cdot \omega^{\alpha_\ell + 1/2 - 2\alpha_{\ell+1}} + e^{-Tp\omega^{2\alpha_{\ell+1}}/2} \cdot \sqrt{2}\sigma\omega^{\alpha_\ell} \\
&\overset{(ii)}{\leq} \sqrt{2}\sigma\omega^{\alpha_\ell+1/2-2\alpha_{\ell+1}} \cdot (2c_1\sigma/p + 1) \\
&\overset{(iii)}{=} c'\sqrt{2}\sigma\omega^{\alpha_{\ell+1}},
\end{aligned}$$

where step (i) follows from the inequality (F.24) and the consequent bound $\overline{\gamma}_\ell \leq e^{-p/(2\omega^{2\alpha_{\ell+1}})}$. Furthermore, in step (ii), we used the following bound

$$\overline{\gamma}_\ell^T \leq e^{-Tp\omega^{2\alpha_{\ell+1}}/2} \leq \omega^{1/2-2\alpha_{\ell+1}} \quad \text{for } T \geq \frac{(1-4\alpha_{\ell+1})}{p\omega^{2\alpha_{\ell+1}}}\log\frac{1}{\omega}, \tag{F.27}$$

and in step (iii) we invoked the relation (F.16a), 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 (F.27).

Proof of claim (F.25b)**:** The proof of this claim makes use of arguments similar to those used above in the proof of claim (F.25a). Starting at time $T_\ell + \lceil t_{\ell+1}/2 \rceil$, and applying the triangle inequality, we find that

$$\left\|\theta^{T_\ell+\lceil t_{\ell+1}/2\rceil+1}\right\|_2 \leq c_1\sigma \cdot c'\sqrt{2}\sigma\omega^{\alpha_{\ell+1}} \cdot \sqrt{\omega} + \overline{\gamma}_\ell \left\|\theta^{T_\ell+\lceil t_{\ell+1}/2\rceil}\right\|_2,$$

where we have used the bound (F.18) 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

$$\begin{aligned}
\left\|\theta^{T_\ell+\lceil t_{\ell+1}/2\rceil+T}\right\|_2 &\leq \sqrt{2}\sigma\omega^{\alpha_{\ell+1}+1/2-2\alpha_{\ell+1}} \cdot c' \cdot (2c_1\sigma/p + 1) \\
&= c'^2\omega^{1/2-2\alpha_{\ell+1}} \cdot \sqrt{2}\sigma\omega^{\alpha_{\ell+1}}.
\end{aligned}$$

Observe that $2\alpha_{\ell+1} - 1/2 \leq -2\varepsilon$ for all $\ell \leq \ell_\varepsilon - 1$ and that the sample size given by bound (F.20) satisfies $n \geq (c')^{4/\varepsilon}\sigma^2(d + \log(2\ell_\varepsilon/\delta))$; together, these facts imply that $c'^2\omega^{1/2-2\alpha_{\ell+1}} \leq 1$. The claim now follows.

F.4 Proof of Theorem 7.4

We now turn to the proof of the lower bound on the accuracy of EM fixed points, as stated in Theorem 7.4. Recalling the definition (7.9) of a sample-based EM operator M_n, the fixed point relation $M_n(\widehat{\theta}_n) = \widehat{\theta}_n$ can be re-written as

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tanh\left(\frac{\widehat{\theta}_n X_i}{\sigma^2}\right), \tag{F.28}$$

where $\widehat{\theta}_n$ denotes a fixed point solution. Our proof makes use of the following elementary bounds on the hyperbolic tangent function:

$$x \cdot \tanh(\alpha x) \geq \alpha x^2 - \frac{1}{3}\alpha^3 x^4, \quad \text{for } \alpha \geq 0, x \in \mathbb{R} \quad \text{and} \tag{F.29a}$$

$$x \cdot \tanh(\alpha x) \leq \alpha x^2 - \frac{1}{3}\alpha^3 x^4, \quad \text{for } \alpha < 0, x \in \mathbb{R}. \tag{F.29b}$$

In order to keep the proof self-contained, we prove these bounds at the end of this section. Now plugging in $\alpha = \widehat{\theta}_n/\sigma^2$ and using the bound (F.29a) for the case $\widehat{\theta}_n \geq 0$ and the bound (F.29b) 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 inequality above yields that

$$|\widehat{\theta}_n|^3 \geq \frac{3\sigma^2\left(\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - 1\right)|\widehat{\theta}_n|}{\frac{1}{n}\sum_{i=1}^{n} Y_i^4}. \tag{F.30}$$

Note that the random variables $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and thereby the quantity on the RHS above is a ratio of empirical moments of Gaussian random variables. In order to obtain a lower bound for $|\widehat{\theta}_n|$ from the inequality (F.30), we exploit a few standard probability bounds for the concentration of moments of standard Gaussian distribution (refer to Theorem 5.2 in Inglot [126] and Theorem 6.7 in Janson [129]). In particular, we have

$$\mathbb{P}\left[\frac{\sum_{i=1}^{n} Y_i^2}{n} - 1 \geq \frac{\log 17}{n} + \frac{\sqrt{\log 17}/4}{\sqrt{n}}\right] \geq \frac{1}{17}, \qquad \text{and} \tag{F.31a}$$

$$\mathbb{P}\left[\frac{\sum_{i=1}^{n} Y_i^4}{n} \leq c\right] \geq 1 - \frac{1}{34}, \tag{F.31b}$$

where $c = (e\log(34)/2)^2\sqrt{6}$. Plugging these bounds in the inequality (F.30), we find that

$$\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - 1}{\frac{1}{n}\sum_{i=1}^{n} Y_i^4} \geq \frac{\sqrt{\log 17}/4}{c} \cdot \frac{1}{\sqrt{n}}, \tag{F.32}$$

with probability at least $1/34$, where we have used the following elementary fact for two events $\mathcal{A}_1, \mathcal{A}_2$:

$$\mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2) = \mathbb{P}(\mathcal{A}_1) + \mathbb{P}(\mathcal{A}_2) - \mathbb{P}(\mathcal{A}_1 \cup \mathcal{A}_2) \geq \mathbb{P}(\mathcal{A}_1) + \mathbb{P}(\mathcal{A}_2) - 1.$$

Let $\mathcal{A}$ denote the event that "there are at least two non-zero fixed points $\widehat{\theta}_n$". We claim that $\mathcal{A}$ is contained within the event $\mathcal{B}$, defined as follows

$$\mathcal{A} \subseteq \underbrace{\left\{ \frac{1}{n} \sum_{i=1}^{n} Y_i^2 > 1 \right\}}_{=:\mathcal{B}}. \tag{F.33}$$

Deferring the proof of this claim to the end of this section, we now complete the proof of our original claim. Note that the event $\mathcal{B}$ is implied by the event in the bound (F.31a), 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 (F.30) by $|\widehat{\theta}_n|$ and using the bound (F.32), 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 in the theorem.

We now prove our earlier claims (F.29a)-(F.29b) and (F.33).

F.4.1 Proof of the bounds (F.29a) and (F.29b)

Note that it suffices to establish that

$$y \tanh(y) \geq y^2 - y^4/3, \qquad \text{for all } y \in \mathbb{R}. \tag{F.34}$$

Indeed, a change of variable $y = \alpha x$ and dividing both sides by α yield 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.

F.4.2 Proof of set-inclusion (F.33)

Consider the (random) function $g : \mathbb{R} \to \mathbb{R}$ such that $g(\theta) := M_n(\theta) - \theta$. Also introduce the shorthand $Z = \sum_{i=1}^n Y_i^2/n$, and note that $\mathcal{B} = \{Z > 1\}$. 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 $\mathcal{B}$, the function g has at least one strictly positive root. Since g is an odd function, we also have that under the same event, the function g has at least one strictly negative root. The claim now follows.

F.5 Proof of Lemma 7.1

The proof of this lemma is based on standard arguments to derive Rademacher complexity bounds [241, 246]. First, we reduce the supremum of random variables over an uncountable set to a finite maximum. We then symmetrize with Rademacher variables, and then apply the Ledoux-Talagrand contraction inequality. Finally, we exploit tail bounds on sub-Gaussian and sub-exponential random variables so as to obtain the desired claim.

Let $\mathbb{S}^d = \left\{u \in \mathbb{R}^d \mid \|u\|_2 = 1\right\}$ 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 $N \leq 17^d$. Using the usual discretization argument (see Chapter 6, [246]), we can show that

$$Z \leq \max_{j \in [N]} \frac{8 Z_{u^j}}{7}. \tag{F.35}$$

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 (7.8) and (7.9) of the EM operators M_n and

M, respectively, we find that

$$\begin{aligned}
Z_{u^k} &= \sup_{\theta \in \mathbb{B}(0,r)} \left\{ \frac{1}{n} \sum_{i=1}^{n} (2w_\theta(X_i) - 1) X_i^\top u^k - \mathbb{E}\left[(2w_\theta(X) - 1) X^\top u^k \right] \right\} \\
&= \left(\frac{1}{n} \sum_{i=1}^{n} X_i^\top u^k - \mathbb{E}[X^\top u^k] \right) \cdot (2\pi - 1) \\
&\quad + \sup_{\theta \in \mathbb{B}(0,r)} \left\{ \left(\frac{1}{n} \sum_{i=1}^{n} 2(w_\theta(X_i) - \pi) X_i^\top - \mathbb{E}\left[2(w_\theta(X) - \pi) \cdot X^\top \right] \right) u^k \right\} \\
&= A_{u^k} + B_{u^k},
\end{aligned}$$

and thereby that

$$Z \leq \frac{8}{7} \left\{ \max_{j \in [N]} A_{u^j} + \max_{j \in [N]} B_{u^j} \right\}. \tag{F.36}$$

Noting that $X_i^\top u^j \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ and that $N \leq 17^d$, standard concentration bounds yield that

$$\mathbb{P}\left[\max_{j \in [N]} A_{u^j} \leq |2\pi - 1|\, \sigma \sqrt{\frac{d \log 17 + \log(1/\delta)}{n}} \right] \geq 1 - \delta. \tag{F.37a}$$

On the other hand, for the random variables B_{u^j}, we claim the following bound

$$\mathbb{P}\left[\max_{j \in [N]} B_{u^j} \leq c' r \sigma^2 \sqrt{\frac{d + \log(1/\delta)}{n}} \right] \geq 1 - \delta. \tag{F.37b}$$

Putting the bounds (F.36) and (F.37) together yields the claim of the lemma.

F.5.0.1 Proof of the bound (F.37b)

Using a symmetrization bound [241, 246], we find that

$$\mathbb{E}[\exp(\lambda B_{u^k})] \leq \mathbb{E}\left[\exp\left(\sup_{\theta \in \mathbb{B}(0,r)} \frac{2\lambda}{n} \sum_{i=1}^{n} \varepsilon_i 2(w_\theta(X_i) - \pi) X_i^\top u^k \right) \right], \tag{F.38}$$

for any $\lambda > 0$ 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 [151]. For each fixed x, define the function $f_x(\theta) \coloneqq 2\left(w_\theta(x) - \pi\right)$. Since $w_0(x) = \pi$ for all x, we have $f_x(0) = 0$, so that this function is centered. Moreover, for any pair (θ, θ'), we have

$$|f_x(\theta) - f_x(\theta')| = |2w_\theta(x) - 2w_{\theta'}(x)| \leq 2\left|\theta^\top x - (\theta')^\top x\right|,$$

so that $f_x(\theta)$ is 2-Lipschitz in the quantity $\theta^\top x$. Consequently, applying the Ledoux-Talagrand contraction inequality for this map, we find that

$$\begin{aligned}\mathbb{E}\left[\exp\left(\sup_{\theta\in\mathbb{B}(0,r)} \frac{2\lambda}{n}\sum_{i=1}^n \varepsilon_i(2(w_\theta(X_i)-\pi)X_i^\top u^k)\right)\right]&\\ \leq \mathbb{E}\left[\exp\left(\sup_{\theta\in\mathbb{B}(0,r)} \frac{4\lambda}{n}\sum_{i=1}^n \varepsilon_i\theta^\top X_iX_i^\top u^k\right)\right].&\end{aligned}$$

Furthermore, using the fact that $\left\|u^k\right\|_2 = 1$ and the standard bound $u^\top Bv \leq \|u\|_2 \, |\!|\!| B |\!|\!|_{\text{op}} \, \|v\|_2$, we obtain that

$$\begin{aligned}\mathbb{E}\left[\exp\left(\sup_{\theta\in\mathbb{B}(0,r)} \frac{4\lambda}{n}\sum_{i=1}^n \varepsilon_i\theta^\top X_iX_i^\top u^k\right)\right]&\\ \leq \mathbb{E}\left[\exp\left(\sup_{\theta\in\mathbb{B}(0,r)} 4\lambda\|u^k\|_2\,\|\theta\|_2\,|\!|\!|\frac{1}{n}\sum_{i=1}^n \varepsilon_iX_iX_i^\top|\!|\!|_{\text{op}}\right)\right]&\\ \leq \mathbb{E}\left[\exp\left(4\lambda r\,|\!|\!|\frac{1}{n}\sum_{i=1}^n \varepsilon_iX_iX_i^\top|\!|\!|_{\text{op}}\right)\right].& \qquad \text{(F.39)}\end{aligned}$$

We now make two auxiliary claims:
(a) The operator norm of the matrix $\sum_{i=1}^n \varepsilon_iX_iX_i^\top/n$ can be bounded as follows:

$$|\!|\!|\frac{1}{n}\sum_{i=1}^n \varepsilon_iX_iX_i^\top|\!|\!|_{\text{op}} \leq 2\max_{j\in[N]}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(X_i^\top u^j)^2\right|. \qquad \text{(F.40a)}$$

(b) For all $(i,j)\in[n]\times[N]$, we have

$$\mathbb{E}\left[\exp(t\varepsilon_i(X_i^\top u^j)^2)\right] \leq \exp(17\cdot t^2\sigma^4) \quad \text{for all } |t|\leq \tfrac{1}{4\sigma^2}. \qquad \text{(F.40b)}$$

The claim (F.40a) follows by the same discretization argument that we used before (see Chapter 6 in the book [246]). We return to prove the claim (F.40b) at the end of this appendix.

Taking these claims as given for the moment, let us now complete the proof of the bound (F.37b). Putting together the pieces, we find that

$$\begin{aligned}
\mathbb{E}[\exp(\lambda B_{u^k})] &\overset{\text{(bnd. (F.38), (F.39))}}{\leq} \mathbb{E}\left[\exp\left(4\lambda r \,\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i X_i^\top\|_{\text{op}}\right)\right] \\
&\overset{\text{(eqn. (F.40a))}}{\leq} \mathbb{E}\left[\exp\left(\max_{j\in[N]}\frac{8\lambda r}{n}\left|\sum_{i=1}^{n}\varepsilon_i (X_i^\top u^j)^2\right|\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(\max_{j\in[N]}\frac{-8\lambda r}{n}\sum_{i=1}^{n}\varepsilon_i (X_i^\top u^j)^2\right)\right] \\
&\qquad + \mathbb{E}\left[\exp\left(\max_{j\in[N]}\frac{8\lambda r}{n}\sum_{i=1}^{n}\varepsilon_i (X_i^\top u^j)^2\right)\right] \\
&\overset{\text{(eqn. (F.40b))}}{\leq} 2N\cdot\prod_{i=1}^{n}\exp\left(17\cdot\frac{64\lambda^2 r^2}{n^2}\cdot\sigma^4\right)
\end{aligned}$$

for any $|\lambda| \leq n/(32r\sigma^2)$. Now invoking the inequality $2N \leq 34^d \leq e^{4d}$, we find that

$$\mathbb{E}[\exp(\lambda B_{u^k})] \leq \exp\left(c\cdot\lambda^2 r^2\sigma^4/n + 4d\right) \qquad \text{for any } k\in[N],$$

and sufficiently small λ. Now using the fact that $N \leq e^{3d}$, we obtain that

$$\mathbb{E}[\exp(\lambda\max_{j\in[N]} B_{u^j})] \leq N\exp(c\cdot 4\lambda^2 r^2\sigma^4/n + 4d) \leq \exp(c\cdot\lambda^2 r^2\sigma^4/n + 7d),$$

for some constant c. Using the standard approach for applying Chernoff bound, we have that

$$\max_{j\in[N]} B_{u^j} \leq cr\sigma^2\cdot\sqrt{\frac{d+\log(1/\delta)}{n}}, \qquad \text{with probability at least } 1-\delta,$$

as long as $n \geq c'(d+\log(1/\delta)$ for some suitable constants c and c'.

We now return to prove our earlier claim (F.40b).

Proof of claim (F.40b): Noting that $X_i^\top u^j \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$, and the fact that square of a sub-Gaussian random variable with parameter σ is a sub-exponential random variable with parameter $(4\sigma^2, 4\sigma^2)$, we obtain the following inequality [243]:

$$\mathbb{E}\left[\exp\left(t(X_i^\top u^j)^2 - t\mathbb{E}(X_i^\top u^j)^2\right)\right] \leq e^{16t^2\sigma^4} \quad \text{for all } |t| \leq \frac{1}{4\sigma^2}. \tag{F.41}$$

Noting that the random variable ε_i is independent of $X_i^\top v$, we find that

$$\begin{aligned}\mathbb{E}\left[\exp(t\varepsilon_i(X_i^\top u^j)^2)\right] &= \frac{1}{2}\mathbb{E}\left[\exp(t(X_i^\top u^j)^2)\right] + \frac{1}{2}\mathbb{E}\left[\exp(-t(X_i^\top u^j)^2)\right] \\ &\overset{(i)}{\leq} e^{16t^2\sigma^4}\cdot\frac{1}{2}\left[e^{t\sigma^2} + e^{-t\sigma^2}\right] \\ &\overset{(ii)}{\leq} e^{17t^2\sigma^4},\end{aligned}$$

for all $|t| \leq \frac{1}{4\sigma^2}$. In asserting the above sequence of steps, we have applied the inequality (F.41) along with the fact that $\mathbb{E}(X_i^\top u^j)^2 = \sigma^2$ to conclude step (i), and step (ii) follows from the inequality $e^x + e^{-x} \leq 2e^{x^2}$ for all $x \in \mathbb{R}$. The claim now follows.

F.6 Proofs of auxiliary lemmas

In this appendix, we present the proofs of the auxiliary lemmas used in the proofs of our main results.

F.6.1 Proof of Lemma F.1

We begin with the elementary inequality $\exp(y) + \exp(-y) \geq 2 + y^2$, valid for all $y \in \mathbb{R}$, to find that

$$\begin{aligned}[B_{\theta_u}]_{11} &= \mathbb{E}_V\left[\frac{V_1^2}{\left(\exp\left(-\|\theta_u\|_2 V_1/\sigma\right) + \exp\left(\|\theta_u\|_2 V_1/\sigma\right)\right)^2}\right] \\ &\leq \mathbb{E}_{V_1}\left[\frac{V_1^2}{\left(2 + V_1^2\|\theta_u\|_2^2/\sigma^2\right)^2}\right].\end{aligned} \tag{F.42}$$

Letting $\mathbb{I}_A$ 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

$$\begin{aligned}\mathbb{E}\left[\frac{V_1^2}{\left(2 + V_1^2\|\theta_u\|_2^2/\sigma^2\right)^2}\right] &= \mathbb{E}\left[\frac{V_1^2}{\left(2 + V_1^2\|\theta_u\|_2^2/\sigma^2\right)^2}\mathbb{I}_{\{|V_1|\leq 1\}}\right] \\ &\qquad + \mathbb{E}\left[\frac{V_1^2}{\left(2 + V_1^2\|\theta_u\|_2^2/\sigma^2\right)^2}\mathbb{I}_{\{|V_1|>1\}}\right] \\ &\leq \frac{1}{4}\mathbb{E}\left[V_1^2\mathbb{I}_{\{|V_1|\leq 1\}}\right] + \mathbb{E}\left[\frac{V_1^2}{\left(2 + \|\theta_u\|_2^2/\sigma^2\right)^2}\mathbb{I}_{\{|V_1|>1\}}\right].\end{aligned} \tag{F.43}$$

Here the final inequality is a consequence of the following observation:

$$\frac{V_1^2}{\left(2 + V_1^2\|\theta_u\|_2^2/\sigma^2\right)^2} \leq \begin{cases}\dfrac{V_1^2}{4} & \text{if } |V_1| \leq 1, \\ \dfrac{V_1^2}{(2 + \|\theta_u\|_2^2/\sigma^2)^2} & \text{if } |V_1| > 1.\end{cases} \tag{F.44}$$

Putting the inequalities (F.42) and (F.43) 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 can directly verify that $\mathbb{E}\left[V_1^2 \mathbb{I}_{\{|V_1|\geq 1\}}\right] = 1 - p_1$ and consequently obtain that

$$[B_{\theta_u}]_{11} \leq \frac{p_1}{4} + \frac{(1-p_1)}{4}\frac{1}{(1 + \|\theta_u\|_2^2/(2\sigma^2))^2}. \tag{F.45}$$

Now we bound the entries $[B_{\theta_u}]_{jj}$, $j \neq 1$. Using the standard inequality $\exp(y) + \exp(-y) \geq 2 + y^2$ once again and noting that $V_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, we find that

$$\begin{aligned}[B_{\theta_u}]_{jj} &= \mathbb{E}_V\left[\frac{V_j^2}{\left(\exp\left(-\|\theta_u\|_2 V_1/\sigma\right) + \exp\left(\|\theta_u\|_2 V_1/\sigma\right)\right)^2}\right] \\ &\leq \mathbb{E}_{V_1}\left[\frac{1}{\left(2 + V_1^2\|\theta_u\|_2^2/\sigma^2\right)^2}\right].\end{aligned} \tag{F.46}$$

Similar to observation (F.44), we also have that

$$[B_{\theta_u}]_{jj} \leq \frac{1}{4}\mathbb{E}\left[\mathbb{I}_{\{|V_1|\leq 1\}}\right] + \mathbb{E}\left[\frac{1}{\left(2 + \|\theta_u\|_2^2/\sigma^2\right)^2}\mathbb{I}_{\{|V_1|>1\}}\right]. \tag{F.47}$$

Define $p_2 := \mathbb{P}\left(|V_1| \leq 1\right)$. Putting together the inequalities (F.46) and (F.47), we obtain that

$$[B_{\theta_u}]_{jj} \leq \frac{p_2}{4} + \frac{(1-p_2)}{4}\frac{1}{(1 + \|\theta_u\|_2^2/(2\sigma^2))^2} \quad \text{for } j = 2, \ldots, d. \tag{F.48}$$

Note that

$$p_2 = \mathbb{P}\left(|V_1| \leq 1\right) = \mathbb{E}\left[\mathbb{I}_{\{|V_1|\leq 1\}}\right] > \mathbb{E}\left[V_1^2 \mathbb{I}_{\{|V_1|\leq 1\}}\right] = p_1,$$

and consequently, the bound on the RHS of inequality (F.48) is larger than the RHS of inequality (F.45). 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 $p_2 = \mathbb{P}\left(|V_1| \leq 1\right)$ and the claim (F.11) follows.

F.6.2 Proof of Lemma F.2

We now prove the claim (F.15) in two steps. First, we show that $[F_\theta]_{jj} \geq [F_\theta]_{11}$ for all $j \in [d]$. Then, we derive the claimed lower bound for $[F_\theta]_{11}$.

Proof of $[F_\theta]_{jj} \geq [F_\theta]_{11}$: For all $j \neq 1$, by changing the order of integration, we obtain that

$$\begin{aligned}
[F_\theta]_{jj} &= \int_0^1 \mathbb{E}_V \left[\frac{V_j^2}{(\exp(-\|\theta_u\|_2 V_1/\sigma) + \exp(\|\theta_u\|_2 V_1/\sigma))^2} \right] du \\
&\overset{(i)}{=} \int_0^1 \mathbb{E}_V \left[\frac{1}{(\exp(-\|\theta_u\|_2 V_1/\sigma) + \exp(\|\theta_u\|_2 V_1/\sigma))^2} \right] du \\
&\overset{(ii)}{\geq} \int_0^1 \mathbb{E}_V \left[\frac{V_1^2}{(\exp(-\|\theta_u\|_2 V_1/\sigma) + \exp(\|\theta_u\|_2 V_1/\sigma))^2} \right] du = [F_\theta]_{11} \,,
\end{aligned}$$

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.[1]

Lower bound on $[F_\theta]_{11}$: Substituting $\theta_u = u\theta$ in the expression for $[F_\theta]_{11}$, and noting that $\int_0^1 (e^{au} + e^{-au})^{-2} du = \tanh(a)/(4a)$, we obtain that

$$\begin{aligned}
[F_\theta]_{11} &= \mathbb{E}_{V_1} \left[\int_0^1 \frac{V_1^2}{(\exp(-u\|\theta\|_2 V_1/\sigma) + \exp(u\|\theta\|_2 V_1/\sigma))^2} du \right] \\
&= \mathbb{E}_{V_1} \left[\frac{\sigma V_1}{4\|\theta\|_2} \tanh \frac{\|\theta\|_2 V_1}{\sigma} \right] \\
&\overset{(i)}{\geq} \frac{1}{4} \mathbb{E}_{V_1} \left[\operatorname{sech}^2 \left(\frac{\|\theta\|_2 V_1}{\sigma} \right) \right] \\
&= \mathbb{E}_{V_1} \left[\frac{1}{(\exp(-\|\theta\|_2 V_1/\sigma) + \exp(\|\theta\|_2 V_1/\sigma))^2} \right],
\end{aligned}$$

[1] 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.

where step (i) follows from Stein's Lemma for standard Gaussian distribution[2]. Expanding the expression in the denominator, we obtain

$$\begin{aligned}[F_\theta]_{11} &= \mathbb{E}_{V_1}\left[\frac{1}{2+\exp\left(-2\|\theta\|_2 V_1/\sigma\right)+\exp\left(2\|\theta\|_2 V_1/\sigma\right)}\right] \\ &\geq \frac{1}{\mathbb{E}_{V_1}\left[2+\exp\left(-2\|\theta\|_2 V_1/\sigma\right)+\exp\left(2\|\theta\|_2 V_1/\sigma\right)\right]}, \end{aligned} \tag{F.49}$$

where the last inequality follows from Jensen's inequality applied with the convex function $y \mapsto \frac{1}{y}$ on $y \in (0, \infty)$. Noting that $V_1 \sim \mathcal{N}(0,1)$ and consequently that $\mathbb{E}_{V_1}\left(\exp\left(yV_1\right)\right) = e^{y^2/2}$ for all $y \in \mathbb{R}$, we obtain that

$$\begin{aligned}\mathbb{E}_{V_1}\Big[2+\exp\left(-2\|\theta\|_2 V_1/\sigma\right)+\exp\left(2\|\theta\|_2 V_1/\sigma\right)\Big] &= 2(1+e^{2\|\theta\|_2^2/\sigma^2}) \\ &\leq 4(1+2\left\|\theta\right\|_2^2/\sigma^2), \end{aligned} \tag{F.50}$$

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 (F.49) and (F.50) together yields the claimed lower bound for $[F_\theta]_{11}$.

F.6.3 Proof of Lemma F.3

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 $\|\theta^t\|_2 \geq \sqrt{2}\sigma\omega^{\alpha_{\ell+2}}$ and mimic the arguments that follow. Applying the triangle inequality, we find that

$$\begin{aligned}\left\|\theta^{t+1}\right\|_2 = \left\|M_n(\theta^t)\right\|_2 &\leq \left\|M_n(\theta^t) - M(\theta^t)\right\|_2 + \left\|M(\theta^t)\right\|_2 \\ &\overset{(i)}{\leq} c_1\sigma \cdot \sqrt{2}\sigma\omega^{\alpha_\ell} \cdot \sqrt{\omega} + \gamma_{\text{up}}(\theta^t)\left\|\theta^t\right\|_2 \\ &\overset{(ii)}{\leq} c_1\sigma \cdot \sqrt{2}\sigma\omega^{\alpha_\ell} \cdot \sqrt{\omega} + \left(1 - \frac{p\omega^{2\alpha_{\ell+1}}}{2}\right)\sqrt{2}\sigma\omega^{\alpha_\ell} \\ &= \left(1 - \frac{p\omega^{2\alpha_{\ell+1}}}{2} + c_1\sigma\sqrt{\omega}\right)\sqrt{2}\sigma\omega^{\alpha_\ell}, \end{aligned}$$

where step (i) follows from the bound (F.18) with $r = \|\theta^t\|_2 \leq \sqrt{2}\sigma\omega^{\alpha_\ell}$, and applying Theorem 7.2, and step (ii) follows from the condition that $\|\theta^t\|_2 \geq \sqrt{2}\sigma\omega^{\alpha_{\ell+1}}$ and consequently that $\gamma_{\text{up}}(\theta^t) \leq 1 - p\omega^{2\alpha_{\ell+1}}/2$. Note that $2\alpha_{\ell+1} - 1/2 \leq -2\varepsilon$ for all $\ell \leq \ell_\varepsilon - 1$ and $\omega \leq 1$. As a result, for $n \geq (2c_1\sigma/p)^{1/(2\varepsilon)}\sigma^2 d\log(2\ell_\varepsilon/\delta)$, we have that $\omega^{2\alpha_{\ell+1}-1/2} \geq \omega^{-2\varepsilon} \geq 2c_1\sigma/p$ and thereby that

$$\left(1 - \frac{p\omega^{2\alpha_{\ell+1}}}{2} + c_1\sigma\sqrt{\omega}\right) \leq 1.$$

[2] 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.

Putting all the pieces together yields the result.

F.7 Additional results

In this appendix, we provide additional results to support several claims from Chapter 7.

F.7.1 Effect of initial conditions

The next lemma shows that for the mixture models analyzed in Chapter 7, the population EM operator M maps any $\theta \in \mathbb{R}^d$ to a ball of radius $\sqrt{2/\pi}$ namely, the radius is independent of the dimension d. Given the uniform bounds provided in Lemma 7.1, loosely speaking, $\|M_n(\theta^0)\|_2$ is upper bounded by $\sqrt{2/\pi} + \|\theta^0\|_2 \sqrt{d/n}$ with high probability. Consequently, we make an implicit assumption while elaborating our results that $\|\theta^0\|_2$ is a constant and does not scale with dimension (provided that the sample size is large enough to keep the second term small).

Lemma F.4. *For both the unbalanced or balanced model fits (7.3), when the true model is standard Gaussian, we have*

$$\left\|M(\theta^0)\right\|_2 \leq \sqrt{\frac{2}{\pi}} \quad \textit{for any} \quad \theta^0 \in \mathbb{R}^d.$$

Proof. The proof of this lemma is a direct consequence of the change of basis ideas used in the proofs of Theorems 7.1 and 7.2 before. Using the definition of M and applying the transformation $V = RX/\sigma$ where $R \in \mathbb{R}^{d\times d}$ is an orthonormal matrix such that $R\theta = \|\theta\|_2 e_1$, and e_1 is the first canonical basis vector in $\mathbb{R}^d$, we find that

$$\begin{aligned}
\|M(\theta)\|_2 &= \left\|\mathbb{E}_X\left[\left(\frac{\pi e^{-\frac{\theta^\top X}{\sigma^2}} - (1-\pi)e^{\frac{\theta^\top X}{\sigma^2}}}{\pi e^{-\frac{\theta^\top X}{\sigma^2}} + (1-\pi)e^{\frac{\theta^\top X}{\sigma^2}}}\right) X\right]\right\|_2 \\
&= \left\|\mathbb{E}_V\left[\left(\frac{\pi e^{-\frac{\|\theta\|_2 V_1}{\sigma}} - (1-\pi)e^{\frac{\|\theta\|_2 V_1}{\sigma}}}{\pi e^{-\frac{\|\theta\|_2 V_1}{\sigma}} + (1-\pi)e^{\frac{\|\theta\|_2 V_1}{\sigma}}}\right) V\right]\right\|_2 \\
&= \left|\mathbb{E}_{V_1}\left[\left(\frac{\pi e^{-\frac{\|\theta\|_2 V_1}{\sigma}} - (1-\pi)e^{\frac{\|\theta\|_2 V_1}{\sigma}}}{\pi e^{-\frac{\|\theta\|_2 V_1}{\sigma}} + (1-\pi)e^{\frac{\|\theta\|_2 V_1}{\sigma}}}\right) V_1\right]\right| \\
&\leq \mathbb{E}_{V_1}\left[\left|\frac{\pi e^{-\frac{\|\theta\|_2 V_1}{\sigma}} - (1-\pi)e^{\frac{\|\theta\|_2 V_1}{\sigma}}}{\pi e^{-\frac{\|\theta\|_2 V_1}{\sigma}} + (1-\pi)e^{\frac{\|\theta\|_2 V_1}{\sigma}}}\right| |V_1|\right] \leq \mathbb{E}_{V_1}\left[|V_1|\right] = \sqrt{\frac{2}{\pi}}.
\end{aligned}$$

The claim now follows. □

F.7.2 Behavior of EM when the weight is unknown

We now discuss the case when the mixture weight $\pi \in (0, 1/2]$ in the model fit (7.3) is assumed to be unknown and is estimated jointly with the (single) location parameter μ using EM. (The scale parameter is still assumed to be known and fixed to the true value.) For our case, given a set of i.i.d. samples $\{X_i\}_{i=1}^n$, the sample EM operators $M_{1,n} : \mathbb{R}^d \times (0, 1/2] \mapsto (0, 1/2]$ and $M_{2,n} : \mathbb{R}^d \times (0, 1/2] \mapsto \mathbb{R}^d$ for the weight and location parameters respectively take the form

$$M_{1,n}(\theta, \pi) := \frac{1}{n}\sum_{i=1}^{n} w_{\theta,\pi}(X_i), \quad \text{and} \quad M_{2,n}(\theta, \pi) := \frac{1}{n}\sum_{i=1}^{n}(2w_{\theta,\pi}(X_i) - 1)X_i, \tag{F.51}$$

where the weight function $w_{\theta,\pi}$ is defined as

$$w_{\theta,\pi}(x) := \frac{\pi \exp\left(-\frac{\|\theta - x\|_2^2}{2\sigma^2}\right)}{\pi \exp\left(-\frac{\|\theta - x\|_2^2}{2\sigma^2}\right) + (1 - \pi) \exp\left(-\frac{\|\theta + x\|_2^2}{2\sigma^2}\right)}. \tag{F.52}$$

Taking the infinite sample limit, we can define the corresponding population EM operators M_1 and M_2 for the weight and location parameters as follows:

$$M_1(\theta, \pi) := \mathbb{E}_X \left[w_{\theta,\pi}(X)\right], \quad \text{and} \quad M_2(\theta, \pi) := \mathbb{E}_X \left[(2w_{\theta,\pi}(X) - 1)X\right], \tag{F.53}$$

where the expectation is over the true model $X \sim \mathcal{N}(0, I_d)$. The next results characterize the contraction properties of these population EM operators.

Lemma F.5. *For any $\mu \in \mathbb{R}^d$ and $\pi \in (0, 1/2]$, the population EM operators M_1 and M_2 satisfy*

$$|M_1(\theta, \pi) - \pi| \leq \frac{(1 - c\rho^2)\,\|\mu\|_2}{2}, \quad \textit{and} \quad \|M_2(\theta, \pi)\|_2 \leq \left(1 - \frac{\rho^2}{2}\right)\|\mu\|_2 \tag{F.54}$$

where $\rho := 1 - 2\pi \in (0, 1)$ and $c \in (1/2, 1)$ denotes a universal constant.

See the end of this appendix for the proof.

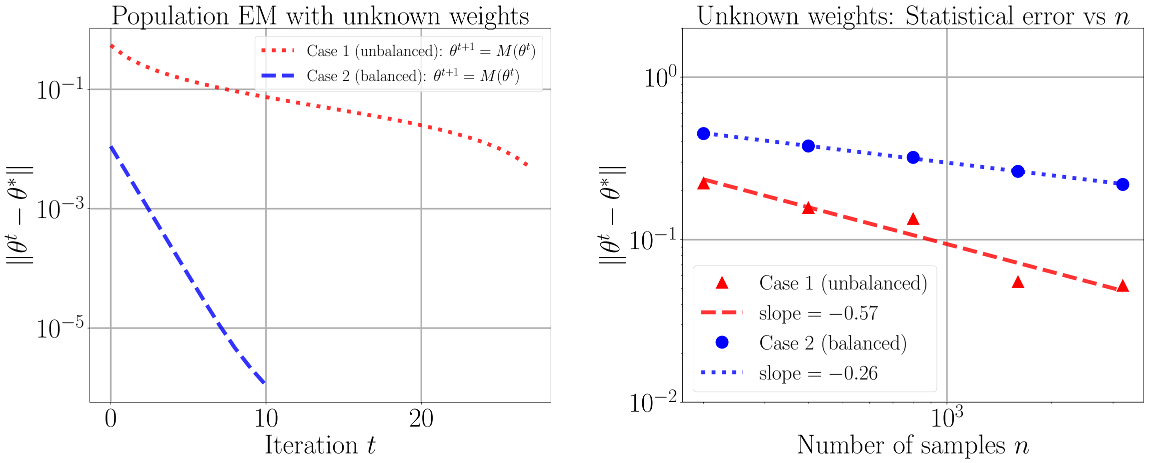
An immediate consequence of Lemma F.5 is the following. Let $\overline{\pi} < 1/2$ be any fixed constant. Consider the population EM sequence (π^t, μ^t) generated as $(\pi^{t+1}, \mu^{t+1}) = (M_1(\theta^t, \pi^t,), M_2(\mu^t, \pi^t))$ starting with an initialization $(\pi^0, \mu^0) \in (0, 1/2] \times \mathbb{R}^d$ such that

$$\pi^0 + \frac{\|\mu^0\|_2}{(1 - 2\overline{\pi})^2} \leq \overline{\pi}. \tag{F.55}$$

Then we have

$$\pi^t \leq \overline{\pi}, \qquad \left\|\mu^t\right\|_2 \leq \left(1 - \frac{(1 - 2\overline{\pi})^2}{2}\right)^{t+1} \left\|\theta^0\right\|_2.$$

In simple words, the weight sequence π^t remains bounded above by $\bar{\pi}$ and the sequence θ^t for the location parameter converges geometrically to $\theta^* = 0$. On the other hand, when the initialization does not satisfy the condition (F.55), the convergence of location parameter can become sub-linear, especially when $\pi^0 \approx 1/2$. In simple words, if the initial mixture is highly unbalanced, we would observe a geometric convergence and as we show in the next corollary sample EM estimates would have a statistical error of order $n^{-\frac{1}{2}}$. When the condition (F.55) is violated, loosely speaking the initial parameters are close to those of a balanced mixture and EM would depict the slower convergence on both algorithmic and (consequently) statistical fronts similar to the results stated in Theorem 7.2. However, a rigorous proof for the later case is beyond the scope of this work and we only provide some numerical evidence in Figure F.1.



(a) (b)

Figure F.1: Behavior of EM for the two-mixture over-specified fit (7.2) with unknown weights where the true model is $\mathcal{N}(0, I_2)$. We consider two different initializations. Case 1 (unbalanced): When the initialization condition (F.55) is met (in particular we set π much smaller than $\frac{1}{2}$). In Case 2 (balanced), we initialize the weight parameter very close to $\frac{1}{2}$. Panel (a) characterizes the population EM updates and panel (b) depicts the statistical error with sample size n for the two cases. We see that when the condition (F.55) is met, EM converges in few steps within error $n^{-\frac{1}{2}}$ error and on the other hand when the initial weight is near $\frac{1}{2}$ we observe a slow convergence of EM with a larger statistical error of order $n^{-\frac{1}{4}}$.

Corollary F.1. *Consider the sample EM sequences $\pi^{t+1} = M_{1,n}(\pi^t)$ and $\theta^{t+1} = M_{2,n}(\theta^t)$ with an initialization that satisfies the condition (F.55) for some $\overline{\pi} < 1/2$. Then for any* *fixed* $\delta \in (0,1)$ *and* $n \geq c\, d \log(1/\delta) \frac{\sigma^2}{\overline{\rho}^4}$, *we have*

$$\pi^t \leq \overline{\pi}, \qquad \left\|\theta^t\right\|_2 \leq \left\|\theta^0\right\|_2 \left[\left(1 - \overline{\rho}^2/2\right)^t + \frac{c'\sigma^2}{\overline{\rho}^2} \sqrt{\frac{d \log(1/\delta)}{n}}\;\right], \tag{F.56}$$

with probability at least $1 - \delta$, *where* c, c' *and* $\overline{\rho} = 1 - 2\overline{\pi} \in (0,1)$ *are universal constants.*

The proof is fairly straightforward given the proof of Theorem 7.1 and is thereby omitted. However, it remains to prove Lemma F.5.

Proof of Lemma F.5: The upper bound for $\|M_2(\mu, \pi)\|_2$ follows directly from the proof of Theorem 7.1. Turning to the other bound in equation (F.54), we see that

$$\begin{aligned} |M_1(\theta, \pi) - \pi| &= \pi(1-\pi) \left| \mathbb{E}_X \left[\frac{\exp\left(X^\top \mu\right) - \exp\left(-X^\top \mu\right)}{\pi \exp\left(X^\top \mu\right) + (1-\pi)\exp\left(-X^\top \mu\right)} \right] \right| \\ &\leq 2\pi(1-\pi) \left\|\mu\right\|_2 \max_{u \in [0,1]} \left\| \mathbb{E}\left[\bar{\Gamma}_{\theta_u}(X)\right] \right\|_{\text{op}}, \end{aligned}$$

where $\mu_u = u\mu$ for $u \in [0,1]$ and the matrix $\bar{\Gamma}_{\theta_u}(X)$ is defined as

$$\bar{\Gamma}_{\theta_u}(X) := \frac{X}{\sigma^2 \left(\pi \exp\left(-\frac{\theta_u^\top X}{\sigma^2}\right) + (1-\pi)\exp\left(\frac{\theta_u^\top X}{\sigma^2}\right)\right)^2}. \tag{F.57}$$

Invoking the transformation as that from the proof of Theorem 7.1 and mimicking the arguments presented there, we can verify that

$$\begin{aligned} \left|\mathbb{E}_X\left[\bar{\Gamma}_{\theta_u}(X)\right]\right| &= \left| \mathbb{E}_{V_1} \left[\frac{V_1}{\sigma \left(\pi \exp\left(-\left\|\theta_u\right\|_2 V_1/\sigma\right) + (1-\pi)\exp\left(\left\|\theta_u\right\|_2 V_1/\sigma\right)\right)^2} \right] \right| \\ &\leq \mathbb{E}_{V_1} \left[\frac{|V_1|}{\sigma \left(\pi \exp\left(-\left\|\theta_u\right\|_2 V_1/\sigma\right) + (1-\pi)\exp\left(\left\|\theta_u\right\|_2 V_1/\sigma\right)\right)^2} \right] \\ &\leq \frac{(1-\rho^2) + \rho^2 \mathbb{E}_{V_1}\left[|V_1|\, \mathbb{I}(V_1 \geq 0)\right]}{(1-\rho^2)} \\ &= \frac{1 - c\rho^2}{(1-\rho^2)} \end{aligned}$$

where $V_1 \sim \mathcal{N}(0,1)$ and $c = 1 - \mathbb{E}_{V_1}\left[|V_1|\, \mathbb{I}(V_1 \geq 0)\right] \in (1/2, 1)$. Putting the above results together yields the claimed bound.

F.8 Closer look at log-likelihood

In this appendix, we provide a further discussion on the difference between the unbalanced and balanced mixtures corresponding to the model (7.2) considered throughout our work. Recall that the expected (population) log-likelihood for the model fit (7.2) is given by

$$\mathcal{L}^\pi(\mu) = \mathbb{E}\left[\log\left(\pi\phi\left(X;\mu,\sigma^2 I_d\right) + (1-\pi)\phi\left(X;-\mu,\sigma^2 I_d\right)\right)\right],$$

where $\phi(\cdot;\theta,\sigma^2 I_d)$ denotes the probability density of the Gaussian distribution $\mathcal{N}(\theta,\sigma^2 I_d)$. Observe that

$$\arg\max_\theta \mathcal{L}^\pi(\mu) = \arg\min_\theta \mathrm{KL}(\mathcal{N}(0,\sigma^2 I_d) \| \pi\mathcal{N}(\theta,\sigma^2 I_d) + (1-\pi)\mathcal{N}(-\theta,\sigma^2 I_d)),$$

where $\mathrm{KL}(P\|Q)$ denotes the Kullback-Leibler divergence between the distributions P and Q. Since the true distribution belongs to the fitted class with $\theta^* = 0$, finding maximizer of the population log-likelihood would yield the true parameter θ^*. As alluded to in the main text, 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) given by equation (7.4).

We now use the nature of log-likelihood to justify the difference between unbalanced and balanced fits. Note that the Fisher information matrix $\mathcal{I}^\pi(\theta) := -\nabla_\theta^2 \mathcal{L}^\pi(\mu)$ for the fit (7.2) with mixture weights $(\pi, 1-\pi)$ is given by

$$[\mathcal{I}^\pi(\theta)]_{ii} = -4\pi(1-\pi)\mathbb{E}\left[\frac{Y_i^2}{(\pi\exp(\mu^\top Y) + (1-\pi)\exp(-\mu^\top Y))^2}\right] + 1$$

for $i \in [d]$ and

$$[\mathcal{I}^\pi(\theta)]_{ij} = -4\pi(1-\pi)\mathbb{E}\left[\frac{Y_i Y_j}{(\pi\exp(\mu^\top Y) + (1-\pi)\exp(-\mu^\top Y))^2}\right]$$

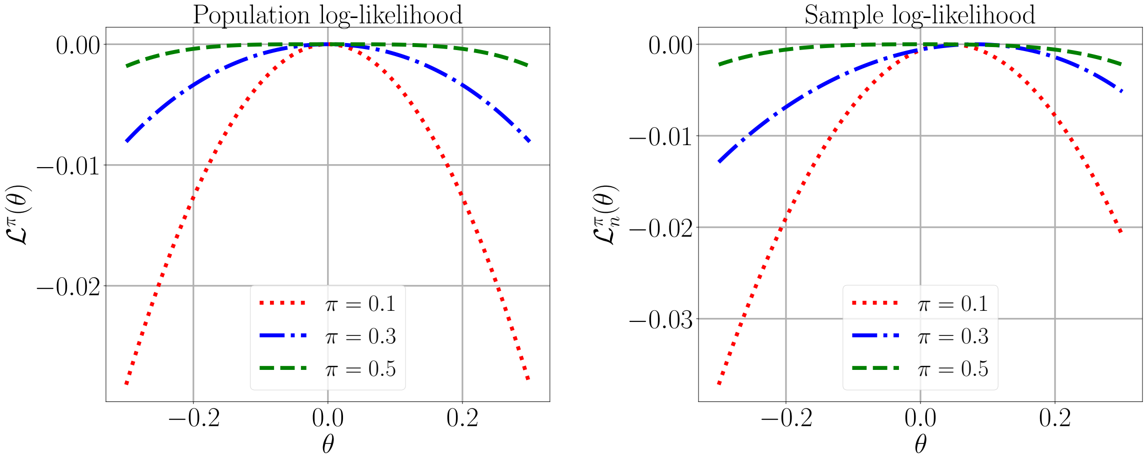
for $i, j \in [d]$ such that $i \neq j$. Here the expectations are taken under the true model $Y = (Y_1, \ldots, Y_d) \sim \mathcal{N}(0, I_d)$. Clearly, at $\theta = \theta^* = 0$, we have

$$\mathcal{I}^\pi(\theta^*) = \beta^\pi I_d, \quad \text{where} \quad \beta^\pi = -4\pi(1-\pi) + 1. \tag{F.58}$$

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 Fisher matrix is positive definite at θ^*, and, for the balanced fit with $\pi = 1/2$, it is singular at θ^*. Equivalently, the log-likelihood is strongly log-concave around θ^* for the unbalanced fit and weakly log-concave for the balanced fit.

We numerically computed the population log-likelihood and plotted it in Figure F.2(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

[3] Figure F.2(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.



(a) (b)

Figure F.2: 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 $\mathcal{L}^{\pi}$ (7.5) (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 $\mathcal{L}_n^{\pi}$ (7.4) 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.

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. As a result, we might expect why population EM may have fundamentally different rate of algorithmic convergence in the two model fits as observed in Figure 7.2.

Moreover, the singularity of Fisher matrix is known to lead to a slow down the statistical rate of MLE. It is well established [240] that when the Fisher matrix is invertible in a neighborhood of the true parameter, MLE has the parametric rate of $n^{-\frac{1}{2}}$, i.e., the MLE estimate is at a distance of order $n^{-\frac{1}{2}}$ from the true parameter θ^*. Moreover, as discussed in the introduction (**??**), several works [42, 194] have also shown than the singularity of the Fisher matrix may lead to a slower than $n^{-\frac{1}{2}}$ rate for the MLE. Since EM algorithm is designed to estimate MLE (and converges only to local maxima), we may loosely conclude that, for the singular case (balanced fit), a slower than parametric rate for the EM estimate is also expected.

F.9 Mixture of regression

In this appendix, we provide formal results for the slow convergence of EM for over-specified mixture of linear regression (as discussed in Section 7.6.2).

Given n samples from the mixture of regressions model (7.26), we use EM to fit the following model:

$$Y|X \sim \frac{1}{2}\mathcal{N}(\theta^\top X, 1) + \frac{1}{2}\mathcal{N}(-\theta^\top X, 1), \tag{F.59}$$

where we assume the knowledge of covariate design $X \sim \mathcal{N}(0, I_d)$. Given this joint model on (X, Y) the population log-likelihood for the model is given by

$$\mathcal{L}(\theta) = \mathbb{E}_{X,Y}\left[\log\left(\pi\phi\left(Y; \theta^\top X, \sigma^2 I_d\right) + (1-\pi)\phi\left(Y; -\theta^\top X, \sigma^2 I_d\right)\right)\right],$$

where $\phi(\cdot; \theta, \sigma^2 I_d)$ denotes the probability density of the Gaussian distribution $\mathcal{N}(\theta, \sigma^2 I_d)$. In Figure F.3, we plot this log-likelihood as a function of θ for two different values of θ^* and observe the following. When the mixture has strong signal ($\theta^* = 0.7$), the Hessian of log-likelihood is negative definite (strongly concave) but in the case of no signal $\theta^* = 0$ the Hessian degenerates at θ^* and the log-likelihood becomes weakly concave.

The behavior observed in Figure F.3 is reminiscent of the behavior of log-likelihood in the case of over-specified Gaussian mixtures considered in the main text (see Appendix F.8 and Figure F.2). We now show that such a similarity also implies a similar behavior for EM, which converges slowly on both algorithmic and statistical fronts (just like the over-specified Gaussian mixture case) for the fit (F.59).

Given this model, the sample EM operator $\overline{M}_n : \mathbb{R}^d \mapsto \mathbb{R}^d$ takes the form

$$\overline{M}_n(\theta) := \left(\sum_{i=1}^n X_i X_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n (2w_\theta(X_i, Y_i) - 1)X_i Y_i\right) \tag{F.60}$$

where we define

$$w_\theta(x, y) := \frac{\pi \exp\left(-\frac{(y-\mu^\top x)^2}{2}\right)}{\pi \exp\left(-\frac{(y-\mu^\top x)^2}{2}\right) + (1-\pi)\exp\left(-\frac{(y+\mu^\top x)^2}{2}\right)}. \tag{F.61}$$

Consequently, the population EM operator $\overline{M} : \mathbb{R}^d \mapsto \mathbb{R}^d$ is given by

$$\overline{M}(\theta) := \mathbb{E}_{(Y,X)}\left[(2w_\theta(X, Y) - 1)XY\right], \tag{F.62}$$

where the outer expectation is taken with respect to $X \sim \mathcal{N}(0, I_d)$ and $Y|X \sim \mathcal{N}((\theta^*)^\top X, 1)$ ($= \mathcal{N}(0, I_d)$ when $\theta^* = 0$). Given these notation, we now characterize the slow convergence of the population EM operator:

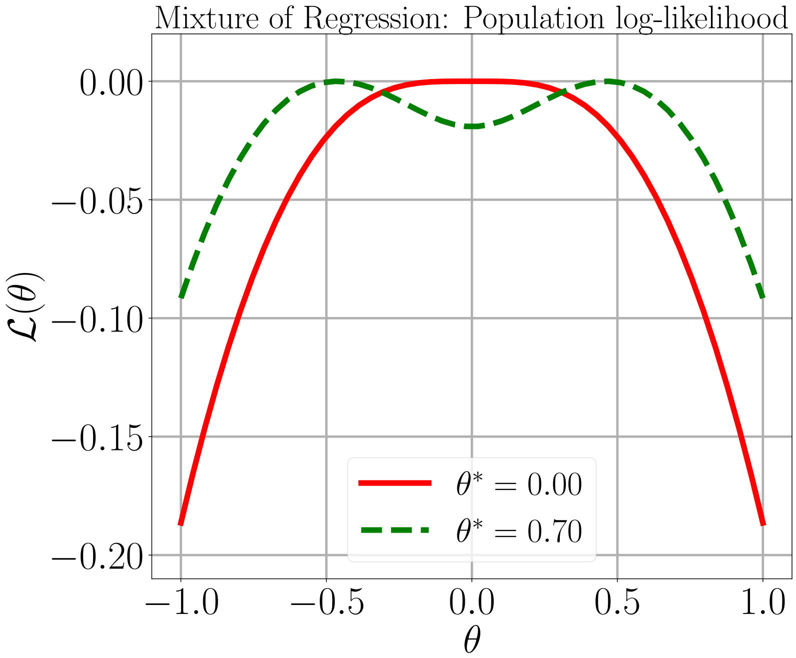


Figure F.3: Plots of the population log-likelihood for the mixture of regression model for $\theta^* \in \{0, 0.7\}$. We see that while the log-likelihood is clearly locally strongly concave around θ^* when $\theta^* = 0.7$, and it is rather flat (and weakly concave) for the case of no signal $\theta^* = 0$. This flatness in log-likelihood results in a slower rate of algorithmic and statistical convergence of EM in this setting thereby providing further evidence of the usefulness of our analysis of EM.

Lemma F.6. *Given the balanced model fit* (F.59) *to the true model* (7.26) *with $\theta^* = 0$, the population EM operator $\overline{M}$* (F.62) *satisfies the following bounds*

$$\|\mu\|_2 \left(1 - 3\|\mu\|_2^2\right) \leq \left\|\overline{M}(\mu)\right\|_2 \leq \|\mu\|_2 \left(1 - 2\|\mu\|_2^2\right) \tag{F.63}$$

for any $\mu \in \mathbb{R}^d$ such that $\|\mu\|_2 \leq 1/2$.

Proof is deferred to the end of this appendix.

We note that the assumption $\|\mu\|_2 \leq 1/2$ is a convenient technical assumption and is possibly loose in a similar manner as noted in Lemma F.4 for the Gaussian mixture case. Applying the localization argument in conjunction with the sub-geometric convergence of the population EM (Lemma F.6) yields the slow statistical convergence (of order $(d/n)^{\frac{1}{4}}$) of the sample EM:

Corollary F.2. *Consider the over-specified model fit (F.59) to the true model (7.26) with $\theta^* = 0$, and initialize the sample EM sequence $\mu^{t+1} = \overline{M}_n(\mu^t)$ with a θ^0 such that $\|\theta^0\|_2 \leq \frac{1}{2}$. Then, for any $\varepsilon \in (0, 1/4)$, $\delta \in (0,1)$, given a large sample size $n \geq c_1' d \log(\log(1/\varepsilon)/\delta)$, the sample EM updates satisfy*

$$\|\mu^t\|_2 \leq \left[\|\theta^0\|_2 \prod_{j=0}^{t-1}\left(1 - 2\|\mu^j\|_2^2\right)\right] + \sqrt{2}\left(\frac{(d + \log\frac{\log(4/\varepsilon)}{\delta})}{n}\right)^{\frac{1}{4}-\varepsilon},$$

for any iterate $t \geq c_2' \left(\frac{n}{d}\right)^{\frac{1}{2}-2\varepsilon} \log(n/d)\log(1/\varepsilon)$, with probability at least $1-\delta$. Here, c_1' and c_2' denote universal constants.

Given Lemma F.6, the proof of Corollary F.2 follows the same annulus-based localization road-map as of the proof of Theorem 7.3; and is thereby omitted. We now prove Lemma F.6.

Proof of Lemma F.6: We provide a proof sketch for the lemma based on an application of Taylor expansion. In particular, we define a transformation $V := RX$ where R is an orthonormal matrix such that $R\mu = \|\mu\|_2 e_1$ and e_1 denotes the first canonical basis vector in dimension d. After similar algebra as that of Theorem 7.2, we can verify that

$$\left\|\overline{M}(\mu)\right\|_2 = \mathbb{E}_{(Y,V_1)}\left[\tanh(V_1 Y \|\mu\|_2) V_1 Y\right],$$

where the outer expectation is taken with respect to $V_1, Y \sim \mathcal{N}(0,1)$ and V_1 and Y are independent. Using arguments similar to the bounds (F.29a) and (F.29b), we can derive that

$$x^2 - \frac{x^4}{3} \leq \tanh(x) \leq x^2 - \frac{x^4}{3} + \frac{2x^6}{15}$$

for all $x \in \mathbb{R}$. Given these bounds, we find that

$$\begin{aligned}
&\mathbb{E}_{(Y,V_1)}\left[\tanh(V_1 Y \|\mu\|_2) V_1 Y\right] \\
&\leq \mathbb{E}\left[(V_1Y)^2\right]\|\mu\|_2 - \frac{\mathbb{E}\left[(V_1Y)^4\right]\|\mu\|_2^3}{3} + \frac{2\mathbb{E}\left[(V_1Y)^6\right]\|\mu\|_2^5}{15} \\
&= \|\mu\|_2 - 3\|\mu\|_2^3 + 30\|\mu\|_2^5 \leq \|\mu\|_2\left(1 - 2\|\mu\|_2^2\right),
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{(Y,V_1)}\left[\tanh(V_1 Y \|\mu\|_2) V_1 Y\right] &\geq \mathbb{E}\left[(V_1Y)^2\right]\|\mu\|_2 - \frac{\mathbb{E}\left[(V_1Y)^4\right]\|\mu\|_2^3}{3} \\
&= \|\mu\|_2 - 3\|\mu\|_2^3 = \|\mu\|_2\left(1 - 3\|\mu\|_2^2\right)
\end{aligned}$$

for all $\mu \in \mathbb{R}^d$ such that $\|\mu\|_2 \leq 1/2$. Putting the above results together yields the lemma.

Appendix G

Content Deferred From Chapter 8

In this appendix, we state and prove a minimax bound in Appendix G.1 that establishes the sharpness of Theorem 8.1, provide the deferred proofs from main text—Theorem 8.1 and Lemmas 8.1 and 8.2 in Appendices G.2 to G.4 respectively. Furthermore, we state and prove a special property of the multivariate population EM operator (8.16) in Appendix G.6, and finally provide some additional details about the Wasserstein distance in Appendix G.7.

G.1 Minimax bound for Theorem 8.1

We now show that the error of order $n^{-\frac{1}{8}}$ in Theorem 8.1 (up to logarithmic factors) is, in fact, tight in the standard minimax sense. Given a compact set $\Omega \subset \mathbb{R} \times (0, \infty)$, and a set of true parameters $(\theta^*, \sigma^*) \in \Omega$, suppose that we draw n i.i.d. samples $\{X_i\}_{i=1}^n$ from a two-Gaussian mixture of the form $\frac{1}{2}\mathcal{N}(\theta^*, (\sigma^*)^2) + \frac{1}{2}\mathcal{N}(-\theta^*, (\sigma^*)^2)$. Let $(\widehat{\theta}_n, \widehat{\sigma}_n) \in \Omega$ denote any estimates—for the respective parameters—measurable with respect to the observed samples $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f_{\theta^*, \sigma^*}$ and let $\mathbb{E}_{(\theta^*, \sigma^*)}$ denote the corresponding expectation.

Proposition G.1. *There exists a universal constant $c_\Omega > 0$ (depending only on Ω), such that*

$$\inf_{(\widehat{\theta}_n, \widehat{\sigma}_n)} \sup_{(\theta^*, \sigma^*)} \mathbb{E}_{(\theta^*, \sigma^*)} \left[\left(|\widehat{\theta}_n| - |\theta^*| \right)^2 + \left| (\widehat{\sigma}_n)^2 - (\sigma^*)^2 \right| \right] \geq c_\Omega n^{-\frac{1}{4} - \delta} \qquad \textit{for any } \delta > 0.$$

See Appendix G.5.1 for the proof.

Based on the connection between location parameter θ_n^t and scale parameter σ_n^t in the EM updates (cf. Equation (8.3b)), the minimax lower bound in Proposition G.1 shows that the (non-squared) error of EM location updates $||\theta_n^t| - |\theta^*||$ is lower bounded by a term (arbitrarily close to) $n^{-\frac{1}{8}}$.

G.1.1 Proof of Proposition G.1

We now present the proof of the minimax bound. We introduce the shorthand $v := \sigma^2$ and $\eta := (\theta, v)$. First of all, we claim the following key upper bound of Hellinger distance between mixture densities f_{η_1}, f_{η_2} in terms of the distances among their corresponding parameters η_1 and η_2:

$$\inf_{\eta_1, \eta_2 \in \Omega} \frac{h\left(f_{\eta_1}, f_{\eta_2}\right)}{\left(\left(|\theta_1| - |\theta_2|\right)^2 + |v_1 - v_2|\right)^r} = 0 \qquad \text{for any } r \in (1, 4). \tag{G.1}$$

Moreover, for any two densities p and q, we denote the total variation distance between p and q by $V(p, q) := (1/2) \int |p(x) - q(x)|\, dx$. Similarly, the squared Hellinger distance between p and q is given as $h^2(p, q) = (1/2) \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx$.

Taking the claim (G.1) as given for the moment, let us complete the proof of Proposition G.1. Our proof relies on Le Cam's lemma for establishing minimax lower bounds. In particular, for any $r \in (1, 4)$ and for any $\epsilon > 0$ sufficiently small, according to the result in equation (G.1), there exist $\eta_1 = (\theta_1, v_1)$ and $\eta_2 = (\theta_2, v_2)$ such that $\left(|\theta_1| - |\theta_2|\right)^2 + |v_1 - v_2| = 2\epsilon$ and $h\left(f_{\eta_1}, f_{\eta_2}\right) \leq c\epsilon^r$ for some universal constant c. From Lemma 1 from Yu [257], we obtain that

$$\sup_{\eta \in \{\eta_1, \eta_2\}} \mathbb{E}_\eta \left[\left(\left|\widehat{\theta}_n\right| - |\theta|\right)^2 + \left|(\widehat{\sigma}_n)^2 - (\sigma)^2\right|\right] \gtrsim \epsilon \left(1 - V(f^n_{\eta_1}, f^n_{\eta_2})\right),$$

where f^n_η denotes the product of mixture densities f_η of the data $X_1, \ldots, X_n$. A standard relation between total variation distance and Hellinger distance leads to

$$V(f^n_{\eta_1}, f^n_{\eta_2}) \leq h(f^n_{\eta_1}, f^n_{\eta_2}) = \sqrt{1 - [1 - h^2(f_{\eta_1}, f_{\eta_2})]^n} \leq \sqrt{1 - [1 - c\epsilon^r]^n}.$$

By choosing $c\epsilon^r = 1/n$, we can verify that

$$\sup_{\eta \in \{\eta_1, \eta_2\}} \mathbb{E}_\eta \left[\left(\left|\widehat{\theta}_n\right| - |\theta|\right)^2 + \left|(\widehat{\sigma}_n)^2 - (\sigma)^2\right|\right] \gtrsim \epsilon \asymp n^{-1/r},$$

which establishes the claim of Proposition G.1.

G.1.1.1 Proof of claim (G.1)

In order to prove claim (G.1), it is sufficient to construct sequences $\eta_{1,n} = (\theta_{1,n}, v_{1,n})$ and $\eta_{2,n} = (\theta_{2,n}, v_{2,n})$ such that

$$h\left(f_{\eta_{1,n}}, f_{\eta_{2,n}}\right) / \left(\left(|\theta_{1,n}| - |\theta_{2,n}|\right)^2 + |v_{1,n} - v_{2,n}|\right)^r \to 0$$

as $n \to \infty$. Indeed, we construct these sequences as follows: $\theta_{2,n} = 2\theta_{1,n}$ and $v_{1,n} - v_{2,n} = 3\left(\theta_{1,n}\right)^2$ for all $n \geq 1$ while $\theta_{1,n} \to 0$ as $n \to \infty$. Direct computation leads to

$$f_{\eta_{1,n}}(x) - f_{\eta_{2,n}}(x) = \frac{1}{2} \underbrace{\left(\phi(x; -\theta_{1,n}, v_{1,n}) - \phi(x; -\theta_{2,n}, v_{2,n})\right)}_{T_{1,n}} + \frac{1}{2} \underbrace{\left(\phi(x; \theta_{1,n}, v_{1,n}) - \phi(x; \theta_{2,n}, v_{2,n})\right)}_{T_{2,n}}.$$

Invoking Taylor expansion up to the third order, we obtain that

$$T_{1,n} = \sum_{|\alpha|\le 3} \frac{(\theta_{2,n}-\theta_{1,n})^{\alpha_1}(v_{1,n}-v_{2,n})^{\alpha_2}}{\alpha_1!\alpha_2!} \frac{\partial^{|\alpha|}\phi}{\partial\theta^{\alpha_1}\partial v^{\alpha_2}}(x;-\theta_{2,n},v_{2,n}) + R_1(x),$$
$$T_{2,n} = \sum_{|\alpha|\le 3} \frac{(\theta_{1,n}-\theta_{2,n})^{\alpha_1}(v_{1,n}-v_{2,n})^{\alpha_2}}{\alpha_1!\alpha_2!} \frac{\partial^{|\alpha|}\phi}{\partial\theta^{\alpha_1}\partial v^{\alpha_2}}(x;\theta_{2,n},v_{2,n}) + R_2(x)$$

where $|\alpha| = \alpha_1+\alpha_2$ for $\alpha = (\alpha_1,\alpha_2)$. Here, $R_1(x)$ and $R_2(x)$ are Taylor remainders that have the following explicit representations

$$\begin{aligned} R_1(x) &:= 4\sum_{|\beta|=4} \frac{(\theta_{2,n}-\theta_{1,n})^{\beta_1}(v_{1,n}-v_{2,n})^{\beta_2}}{\beta_1!\beta_2!} \\ &\quad\times \int_0^1 (1-t)^3 \frac{\partial^4\phi}{\partial\theta^{\beta_1}\partial v^{\beta_2}}\left(x;-\theta_{2,n}+t(\theta_{2,n}-\theta_{1,n}), v_{2,n}+t(v_{1,n}-v_{2,n})\right) dt, \\ R_2(x) &:= 4\sum_{|\beta|=4} \frac{(\theta_{1,n}-\theta_{2,n})^{\beta_1}(v_{1,n}-v_{2,n})^{\beta_2}}{\beta_1!\beta_2!} \\ &\quad\times \int_0^1 (1-t)^3 \frac{\partial^4\phi}{\partial\theta^{\beta_1}\partial v^{\beta_2}}\left(x;\theta_{2,n}+t(\theta_{1,n}-\theta_{2,n}), v_{2,n}+t(v_{1,n}-v_{2,n})\right) dt. \end{aligned}$$

Recall from equation (8.2) that the univariate location-scale Gaussian distribution has the PDE structure of the following form

$$\frac{\partial^2\phi}{\partial\theta^2}(x;\mu,\sigma^2) = 2\frac{\partial\phi}{\partial\sigma^2}(x;\mu,\sigma^2).$$

Therefore, we can write the formulations of $T_{1,n}$ and $T_{2,n}$ as follows:

$$T_{1,n} = \sum_{|\alpha|\le 3} \frac{(\theta_{2,n}-\theta_{1,n})^{\alpha_1}(v_{1,n}-v_{2,n})^{\alpha_2}}{2^{\alpha_2}\alpha_1!\alpha_2!} \frac{\partial^{\alpha_1+2\alpha_2}\phi}{\partial\theta^{\alpha_1+2\alpha_2}}(x;-\theta_{2,n},v_{2,n}) + R_1(x),$$
$$T_{2,n} = \sum_{|\alpha|\le 3} \frac{(\theta_{1,n}-\theta_{2,n})^{\alpha_1}(v_{1,n}-v_{2,n})^{\alpha_2}}{2^{\alpha_2}\alpha_1!\alpha_2!} \frac{\partial^{\alpha_1+2\alpha_2}\phi}{\partial\theta^{\alpha_1+2\alpha_2}}(x;\theta_{2,n},v_{2,n}) + R_2(x).$$

Via a Taylor series expansion, we find that

$$\frac{\partial^{\alpha_1+2\alpha_2}\phi}{\partial\theta^{\alpha_1+2\alpha_2}}(x;\theta_{2,n},v_{2,n}) = \sum_{\tau=0}^{3-|\alpha|} \frac{(2\theta_{2,n})^{\tau}}{\tau!} \frac{\partial^{\alpha_1+2\alpha_2+\tau}\phi}{\partial\theta^{\alpha_1+2\alpha_2+\tau}}(x;-\theta_{2,n},v_{2,n}) + R_{2,\alpha}(x)$$

for any $\alpha = (\alpha_1, \alpha_2)$ such that $1 \leq |\alpha| \leq 3$. Here, $R_{2,\alpha}$ is Taylor remainder admitting the following representation

$$R_{2,\alpha}(x) = \sum_{\tau=4-|\alpha|} \frac{\tau\left(2\theta_{2,n}\right)^{\tau}}{\tau!} \int_0^1 (1-t)^{\tau-1} \frac{\partial^4 \phi}{\partial \theta^{\alpha_1+\tau} \partial v^{\alpha_2}} \left(x; -\theta_{2,n} + 2t\theta_{2,n}, v_{2,n}\right) dt.$$

Governed by the above results, we can rewrite $f_{\eta_{1,n}}(x) - f_{\eta_{2,n}}(x)$ as

$$f_{\eta_{1,n}}(x) - f_{\eta_{2,n}}(x) = \sum_{l=1}^{6} A_{l,n} \frac{\partial^l \phi}{\partial \theta^l}(x; -\theta_{2,n}, v_{2,n}) + R(x)$$

where the explicit formulations of $A_{l,n}$ and $R(x)$ are given by

$$\begin{aligned} A_{l,n} &:= \frac{1}{2} \sum_{\alpha_1,\alpha_2} \frac{1}{2^{\alpha_2}} \frac{(\theta_{2,n}-\theta_{1,n})^{\alpha_1}(v_{1,n}-v_{2,n})^{\alpha_2}}{\alpha_1!\alpha_2!} \\ &\quad + \frac{1}{2} \sum_{\alpha_1,\alpha_2,\tau} \frac{1}{2^{\alpha_2}} \frac{2^{\tau}(\theta_{2,n})^{\tau}(\theta_{1,n}-\theta_{2,n})^{\alpha_1}(v_{1,n}-v_{2,n})^{\alpha_2}}{\tau!\alpha_1!\alpha_2!}, \\ R(x) &:= \frac{1}{2}R_1(x) + \frac{1}{2}R_2(x) + \sum_{|\alpha|\leq 2} \frac{1}{2^{\alpha_2}} \frac{(\theta_{1,n}-\theta_{2,n})^{\alpha_1}(v_{1,n}-v_{2,n})^{\alpha_2}}{\alpha_1!\alpha_2!} R_{2,\alpha}(x) \end{aligned}$$

for any $l \in [6]$ and $x \in \mathbb{R}$. Here the ranges of α_1, α_2 in the first sum of $A_{l,n}$ satisfy $\alpha_1 + 2\alpha_2 = l$ and $1 \leq |\alpha| \leq 3$ while the ranges of α_1, α_2, τ in the second sum of $A_{l,n}$ satisfy $\alpha_1 + 2\alpha_2 + \tau = l$, $0 \leq \tau \leq 3 - |\alpha|$, and $1 \leq |\alpha| \leq 3$.

From the conditions that $\theta_{2,n} = 2\theta_{1,n}$ and $v_{1,n} - v_{2,n} = 3\left(\theta_{1,n}\right)^2$, we can check that $A_{l,n} = 0$ for all $1 \leq l \leq 3$. Additionally, we also have

$$\max\{|A_{4,n}|, |A_{5,n}|, |A_{6,n}|\} \precsim |\theta_{1,n}|^4.$$

Given the above results, we claim that

$$h\left(f_{\eta_{1,n}}, f_{\eta_{2,n}}\right) \precsim |\theta_{1,n}|^8. \tag{G.2}$$

Assume that the claim (G.2) is given. From the formulations of sequences $\eta_{1,n}$ and $\eta_{2,n}$, we can verify that

$$\left((|\theta_{1,n}| - |\theta_{2,n}|)^2 + |v_{1,n} - v_{2,n}|\right)^r \asymp |\theta_{1,n}|^{2r}.$$

Since $1 \leq r < 4$ and $\theta_{1,n} \to 0$ as $n \to \infty$, the above results lead to

$$h\left(f_{\eta_{1,n}}, f_{\eta_{2,n}}\right) / \left((|\theta_{1,n}| - |\theta_{2,n}|)^2 + |v_{1,n} - v_{2,n}|\right)^r \precsim |\theta_{1,n}|^{8-2r} \to 0.$$

As a consequence, we achieve the conclusion of the claim (G.1).

G.1.1.2 Proof of claim (G.2)

The definition of Hellinger distance leads to the following equations

$$
\begin{aligned}
2h^2\left(f_{\eta_{1,n}}, f_{\eta_{2,n}}\right) &= \int \frac{\left(f_{\eta_{1,n}}(x) - f_{\eta_{2,n}}(x)\right)^2}{\left(\sqrt{f_{\eta_{1,n}}(x)} + \sqrt{f_{\eta_{2,n}}(x)}\right)^2} dx \\
&= \int \frac{(\sum\limits_{l=4}^{6} A_{l,n} \dfrac{\partial^l \phi}{\partial \theta^l}(x; -\theta_{2,n}, v_{2,n}) + R(x))^2}{\left(\sqrt{f_{\eta_{1,n}}(x)} + \sqrt{f_{\eta_{2,n}}(x)}\right)^2} dx \\
&\lesssim \int \frac{\sum_{l=4}^{6} (A_{l,n})^2 \left(\dfrac{\partial^l \phi}{\partial \theta^l}(x; -\theta_{2,n}, v_{2,n})\right)^2 + R^2(x)}{\left(\sqrt{f_{\eta_{1,n}}(x)} + \sqrt{f_{\eta_{2,n}}(x)}\right)^2} dx, \qquad \text{(G.3)}
\end{aligned}
$$

where the last inequality is due to Cauchy-Schwarz's inequality. According to the structure of location-scale Gaussian density, the following inequalities hold

$$
\int \frac{\left(\dfrac{\partial^l \phi}{\partial \theta^l}(x; -\theta_{2,n}, v_{2,n})\right)^2}{\left(\sqrt{f_{\eta_{1,n}}(x)} + \sqrt{f_{\eta_{2,n}}(x)}\right)^2} dx \lesssim \int \frac{\left(\dfrac{\partial^l \phi}{\partial \theta^l}(x; -\theta_{2,n}, v_{2,n})\right)^2}{\phi(x; -\theta_{2,n}, v_{2,n})} dx < \infty \qquad \text{(G.4)}
$$

for $4 \leq l \leq 6$. Note that, for any $\beta = (\beta_1, \beta_2)$ such that $|\beta| = 4$, we have

$$
|\theta_{2,n} - \theta_{1,n}|^{\beta_1} |v_{1,n} - v_{2,n}|^{\beta_2} \asymp |\theta_{1,n}|^{4+\beta_2} \lesssim |\theta_{1,n}|^4 .
$$

With the above bounds, an application of Cauchy-Schwarz's inequality leads to

$$
\begin{aligned}
&\int \frac{R_1^2(x)}{\left(\sqrt{f_{\eta_{1,n}}(x)} + \sqrt{f_{\eta_{2,n}}(x)}\right)^2} dx \\
&\lesssim |\theta_{1,n}|^8 \sum_{|\beta|=4} \int \frac{\sup\limits_{t\in[0,1]} \left(\dfrac{\partial^4 \phi}{\partial \theta^{\beta_1} \partial v^{\beta_2}}(x; -\theta_{2,n} + t(\theta_{2,n} - \theta_{1,n}), v_{2,n} + t(v_{1,n} - v_{2,n}))\right)^2}{\phi(x; -\theta_{2,n}, v_{2,n})} dx \lesssim |\theta_{1,n}|^8 .
\end{aligned}
$$

With a similar argument, we also obtain that

$$
\int \frac{R_2^2(x)}{\left(\sqrt{f_{\eta_{1,n}}(x)} + \sqrt{f_{\eta_{2,n}}(x)}\right)^2} dx \lesssim |\theta_{1,n}|^8, \quad \max_{1 \leq |\alpha| \leq 4} \int \frac{R_{2,\alpha}^2(x)}{\left(\sqrt{f_{\eta_{1,n}}(x)} + \sqrt{f_{\eta_{2,n}}(x)}\right)^2} dx \lesssim |\theta_{1,n}|^8 .
$$

Governed by the above bounds, another application of Cauchy-Schwarz's inequality implies that

$$
\int \frac{R^2(x)}{\left(\sqrt{f_{\eta_{1,n}}(x)} + \sqrt{f_{\eta_{2,n}}(x)}\right)^2} dx \lesssim |\theta_{1,n}|^8 . \qquad \text{(G.5)}
$$

Combining the results from equations (G.3), (G.4), and (G.5), we achieve the conclusion of the claim (G.2).

G.2 Proof of Theorem 8.1

Our result makes use of the following corollary (proven in Appendix G.5.1):

Corollary G.1. *Given constants $\delta \in (0, 1)$ and $\varepsilon \in (0, 1/12]$, suppose that we generate the the sample-level EM sequence $\mu_n^{t+1} = M_{n,1}(\mu_n^t)$ starting from an initialization $|\mu_n^0| \in I'_\varepsilon$, and using a sample size n lower bounded as $n \gtrsim \log^{1/(12\varepsilon)}(\log(1/\varepsilon)/\delta)$. Then for all iterations $t \geq n^{1/2-6\varepsilon}\log(n)\log(1/\varepsilon)$, we have*

$$|\mu_n^t - \theta^*| \leq c_1 \left(\frac{1}{n}\log\frac{\log(1/\varepsilon)}{\delta}\right)^{\frac{1}{12}-\varepsilon}, \tag{G.6}$$

with probability at least $1 - \delta$.

Remark: We note that the sub-optimal bound (G.6) obtained from Corollary G.1 is not an artifact of the localization argument and arises due to the definition of the operator operator $\widetilde{M}_{n,1}$ (8.12a). As we have alluded to earlier, indeed a finer analysis with the population EM operator $\overline{M}_1$ is required to prove the rate of $n^{-1/8}$ stated in Theorem 8.1. However, a key assumption in the further derivation is that the sample EM iterates θ_n^t can converge to a ball of radius $r \precsim n^{-1/16}$ around θ^* in a finite number of steps, for which Corollary G.1 comes in handy.

We now begin with a sketch the two stage-argument, and then provide a rigorous proof for Theorem 8.1.

G.2.1 Proof sketch

As mentioned earlier, the pseudo-population operator $\widetilde{M}_{n,1}$ is not sufficient to achieve the sharp rate of EM iterates under the univariate symmetric Gaussian mixture fit. Therefore, we make use of corrected-population operator $\overline{M}_1$ to get a sharp statistical rate of EM. Our proof for the tight convergence rate of sample EM updates relies on a novel two-stage localization argument that we are going to sketch.

First stage argument: Plugging in $\varepsilon = 1/84$ in Corollary G.1, we obtain that for $t \gtrsim \sqrt{n}\log(n)$, with probability at least $1 - \delta$ we have that

$$\left|\theta_n^t - \theta^*\right| \leq cn^{-\frac{1}{14}}\log^{\frac{1}{14}}\frac{\log(1/\varepsilon)}{\delta} \leq n^{-\frac{1}{16}}, \tag{G.7}$$

where the second inequality follows from the large sample condition $n \geq c'\log^8\frac{\log 84}{\delta}$. All the following claims are made conditional on the event (G.7).

Second stage argument: In order to keep the presentation of the proof sketch simple, we do not track constant and logarithmic factors in the arguments to follow. In epoch ℓ, for any iteration t the EM iterates satisfy $\theta_n^t \in [n^{-a_{\ell+1}}, n^{-a_\ell}]$ where $a_{\ell+1} > a_\ell$ and $a_\ell \leq 1/16$. Applying Lemma 8.1 for such iterations, we find that with high probability

$$\left|\overline{M}_1(\theta_n^t)\right| \precsim \underbrace{(1 - n^{-6a_{\ell+1}})}_{=:\gamma_\ell} \left|\theta_n^t\right| \quad \text{and} \quad \left|M_{n,1}(\theta_n^t) - \overline{M}_1(\theta_n^t)\right| \precsim \frac{n^{-3a_\ell}}{\sqrt{n}},$$

where the first bound follows from the $1 - c\theta^6$ contraction bound (8.13b) and the second bound follows from the cubic-type Rademacher bound (8.13d). Invoking the basic triangle inequality T times, we obtain that

$$\left|\theta_n^{t+T}\right| \overset{(i)}{\precsim} e^{-Tn^{-6a_{\ell+1}}} n^{-a_\ell} + \frac{1}{1-\gamma_\ell} \cdot \frac{n^{-3a_\ell}}{\sqrt{n}} \overset{(ii)}{\precsim} \frac{1}{1-\gamma_\ell} \cdot \frac{n^{-3a_\ell}}{\sqrt{n}} = n^{6a_{\ell+1} - 3a_\ell - 1/2},$$

where in step (ii) we have used the fact that for large enough T, the first term is dominated by the second term in the RHS of step (i). To obtain a recursion for the sequence a_ℓ, we set the RHS equal to $n^{-a_{\ell+1}}$. Doing so yields the recursion

$$a_{\ell+1} = \frac{3a_\ell}{7} + \frac{1}{14}, \quad \text{where} \quad a_0 = 1/16. \tag{G.8a}$$

Solving for the limit $a_{\ell+1} = a_\ell = a_\star$, we find that $a_\star = 1/8$. Thus, we can conclude that sample EM iterates in the univariate setting converge to a ball of radius $n^{-1/8}$ as claimed in the theorem statement.

G.2.2 Formal proof of sample EM convergence rate

We now turn to providing a formal proof for the preceding arguments.

Notations: To make the proof comprehensible, some additional notations are necessary which we collect here. Let $\ell_\star = \lceil \log(8/\varepsilon)/\log(7/3) \rceil$ so that $a_{\ell_\star} \leq 1/8 - \varepsilon$. We define the following shorthand:

$$\omega := \frac{n}{c_{n,\delta}}, \quad \text{where} \quad c_{n,\delta} := \log^{10}(10n(\ell_\star + 1)/\delta). \tag{G.8b}$$

For $\ell = 0, \ldots, \ell_\star$, we define the time sequences t_ℓ and T_ℓ as follows:

$$t_0 = \sqrt{n}, \quad t_\ell = \left\lceil 10\omega^{6a_\ell} \log \omega \right\rceil, \quad \text{and} \quad T_\ell = \sum_{j=0}^{\ell} t_j. \tag{G.8c}$$

Direct computation leads to

$$T_{\ell_\star} \leq \sqrt{n} + \ell_\star t_{\ell_\star} \precsim \log\left(\frac{n \log\frac{1}{\varepsilon}}{c_{n,\delta}\delta}\right)\left(\frac{n}{c_{n,\delta}}\right)^{3/4-6\varepsilon} \precsim n^{3/4}. \tag{G.8d}$$

In order to facilitate the proof argument later, we define the following set

$$\mathcal{R} := \left\{\omega^{-a_1}, \ldots, \omega^{-a_{\ell_\star}}, c'\omega^{-a_1}, \ldots, c'\omega^{-a_{\ell_\star}}\right\}, \tag{G.8e}$$

where $c' := (5c_2 + 1)$. Here, c_2 is the universal constant from Lemma 8.1.

Formal argument: We show that with probability at least $1 - \delta$ the following holds:

$$\left|\theta_n^t\right| \leq \left(\frac{c_{n,\delta}}{n}\right)^{a_\ell} = \omega^{-a_\ell}, \qquad \text{for all } t \geq T_\ell, \text{ and } \ell \leq \ell_\star. \tag{G.9}$$

As a consequence of this claim and the definitions (G.8a)-(G.8d) of $a_{\ell_\star}$ and $T_{\ell_\star}$, we immediately obtain that

$$|\mu_n^t - \theta^*| \precsim \left(\frac{c_{n,\delta}}{n}\right)^{1/8-\varepsilon} \precsim \left(\frac{1}{n}\log^{10}\frac{10n\log(8/\varepsilon)}{\delta}\right)^{1/8-\varepsilon},$$

for all number of iterates $t \succsim n^{3/4-6\varepsilon}\log(n)\log(1/\varepsilon)$ with probability at least $1-\delta$ as claimed in Theorem 8.1.

We now define the high probability event that is crucial for our proof. For any $r \in \mathcal{R}$, define the event E_r as follows

$$E_r := \left\{\sup_{\theta\in\mathbb{B}(0,r)}\left|M_{n,1}(\theta) - \overline{M}_1(\theta)\right| \leq c_2 r^3\sqrt{\frac{\log^{10}(5n\left|\mathcal{R}\right|/\delta)}{n}}\right\}.$$

Then, for the event

$$\mathcal{E} := \bigcap_{r\in\mathcal{R}} E_r \cap \{\text{Event (G.7) holds }\}, \tag{G.10}$$

applying the union bound with Lemma 8.1 yields that $\mathbb{P}[\mathcal{E}] \geq 1 - \delta$. All the arguments that follow are conditional on the event $\mathcal{E}$ and hence hold with the claimed high probability.

In order to prove the claim (G.9), we make use of the following intermediate claim:

Lemma G.1. *Conditional on the event $\mathcal{E}$, if $|\theta| \leq \omega^{-a_\ell}$, then $|M_{n,1}(\theta)| \leq \omega^{-a_\ell}$ for any $\ell \leq \ell_\star$.*

Deferring the proof of Appendix G.5.2, we now establish the claim (G.9) conditional on the event $\mathcal{E}$ only for $t = T_\ell$ and when $|\theta_n^t| \in [\omega^{-a_{\ell+1}}, \omega^{-a_\ell}]$ in which we now prove using induction.

Proof of base case $\ell = 0$: Note that we have $a_0 = 1/16$ and that $n^{-1/16} \leq \omega^{1/16}$. Also, by the definition (G.10) we have that the event (G.7) $\subseteq \mathcal{E}$. Hence, under the event $\mathcal{E}$ we have that $|\theta_n^t| \leq n^{-1/16}$, for $t \gtrsim \sqrt{n}\log(n)$. Putting all the pieces together, we find that under the event $\mathcal{E}$, we have $|\theta_n^t| \leq n^{-1/16} \leq \omega^{1/16}$ and the base case follows.

Proof of inductive step: We now establish the inductive step. Note that Lemma G.1 implies that we need to show the following: if $|\theta_n^t| \leq \omega^{-a_\ell}$ for all $t \in \{T_\ell, T_\ell + 1, \ldots, T_{\ell+1} - 1\}$ for any given $\ell \leq \ell_\star$, then $|\theta_n^{T_{\ell+1}}| \leq \omega^{-a_{\ell+1}}$. We establish this claim in two steps:

$$\theta_n^{T_\ell + t_\ell/2} \leq c'\omega^{-a_{\ell+1}}, \quad \text{and,} \tag{G.11a}$$

$$\theta_n^{T_{\ell+1}} \leq \omega^{-a_{\ell+1}}, \tag{G.11b}$$

where $c' = (5c_2 + 1) \geq 1$ is a universal constant. Note that the inductive claim follows from the bound (G.11b). It remains to establish the two claims (G.11a) and (G.11b) which we now do one by one.

Proof of claim (G.11a): Let $\Theta_\ell = \{\theta : |\theta| \in [\omega^{-a_{\ell+1}}, \omega^{-a_\ell}]\}$. Now, conditional on the event $\mathcal{E}$, Lemma 8.1 implies that

$$\sup_{\theta \in \Theta_\ell} \left| M_{n,1}(\theta) - \overline{M}_1(\theta) \right| \leq c_2 \omega^{-3a_\ell - 1/2}, \quad \text{and} \quad \sup_{\theta \in \Theta_\ell} \left| \overline{M}_1(\theta)/\theta \right| \leq (1 - \omega^{-6a_{\ell+1}}/5) =: \gamma_\ell.$$

We can check that $\gamma_\ell \leq e^{-\omega^{6a_{\ell+1}}/5}$. Unfolding the basic triangle inequality $t_\ell/2$ times and noting that $\theta_n^t \in \Theta_\ell$ for all $t \in \{T_\ell, \ldots, T_\ell + t_\ell/2\}$, we obtain that

$$\begin{aligned}
\left|\theta_n^{T_\ell + t_\ell/2}\right| &\leq \gamma_\ell^{t_\ell/2} \left|\theta_n^{T_\ell}\right| + (1 + \gamma_\ell + \ldots + \gamma_\ell^{t_\ell/2 - 1}) c_2 \omega^{-3a_\ell - 1/2} \\
&\leq e^{-t_\ell \omega^{-6a_{\ell+1}}/10} \omega^{-a_\ell} + \frac{1}{1 - \gamma_\ell} c_2 \omega^{-3a_\ell - 1/2} \\
&\overset{(i)}{\leq} (1 + 5c_2) \omega^{6a_{\ell+1} - 3a_\ell - 1/2} \\
&\overset{(ii)}{=} (5c_2 + 1) \omega^{-a_{\ell+1}}
\end{aligned}$$

where step (i) follows from plugging in the value of γ_ℓ and invoking the definition (G.8c) of t_ℓ, which leads to

$$e^{-t_\ell \omega^{6a_{\ell+1}}/10} \omega^{-a_\ell} \leq \omega^{6a_{\ell+1} - 3a_\ell - 1/2}.$$

Moreover, step (ii) is a direct consequence of the definition (G.8a) of the sequence a_ℓ. Therefore, we achieve the conclusion of claim (G.11a).

Proof of claim (G.11b)**:** The proof of this step is very similar to the previous step, except that we now use the set $\Theta'_\ell = \{\theta : |\theta| \in [\omega^{-a_{\ell+1}}, c'\omega^{-a_{\ell+1}}]\}$ for our arguments. Applying Lemma 8.1, we have

$$\sup_{\theta\in\Theta'_\ell} \left|M_{n,1}(\theta) - \overline{M}_1(\theta)\right| \le c_2(c')^3\omega^{-3a_{\ell+1}-1/2}, \quad \text{and} \quad \sup_{\theta\in\Theta'_\ell} \left|\overline{M}_1(\theta)/\theta\right| \le \gamma_\ell.$$

Using the similar argument as that from the previous case, we find that

$$\begin{aligned}\left|\theta_n^{T_\ell+t_\ell/2+t_\ell/s2}\right| &\le e^{-t_\ell\omega^{6a_{\ell+1}}/10}c'\omega^{-a_{\ell+1}} + \frac{1}{1-\gamma_\ell}c_2(c')^3\omega^{-3a_{\ell+1}-1/2}\\ &\le (5c_2+1)(c')^3\omega^{4a_{\ell+1}-1/2}\cdot\omega^{-a_{\ell+1}}\\ &\overset{(i)}{\le} \omega^{-a_{\ell+1}}\end{aligned}$$

where step (i) follows from the inequality $e^{-t_\ell\omega^{6a_{\ell+1}}/10} \le \omega^{4a_{\ell+1}-1/2}$ and the inequality

$$\omega^{4a_{\ell+1}-1/2} \le \omega^{4a_{\ell_\star}-1/2} \le \omega^{-4\varepsilon} \le 1/(c')^4,$$

since $n \ge (c')^{1/\varepsilon}c_{n,\delta}$. The claim now follows.

G.3 Proof of Lemma 8.1

We now prove Lemma 8.1 which provides the basis for the two-staged proof of Theorem 8.1.

The proof for the contraction property (8.13b) of the corrected population operator $\overline{M}_1$ is similar to that of the property (8.13a) pseudo-population operator $\widetilde{M}_{n,1}$ (albeit with a few high probability arguments replaced by deterministic arguments). Hence, while we provide a complete proof of the bound (8.13a) (in Section G.3.1), we only provide a proof sketch for the bound (8.13b) at its end. Moreover the proofs of bounds (8.13c) and (8.13d) are provided in Sections G.3.2 and G.3.3 respectively.

G.3.1 Contraction bound for population operator $\widetilde{M}_{n,1}$

We begin by defining some notation. For $\varepsilon \in (0, 1/12]$ and $\alpha \ge 1/2 - 6\varepsilon$, we define the event $\mathcal{E}_\alpha$ and the interval $I_{\alpha,\varepsilon}$ as follows

$$\mathcal{E}_\alpha = \left\{\left|\sum_{j=1}^n X_j^2/n - 1\right| \le n^{-\alpha}\right\}, \quad \text{and,} \tag{G.12}$$

$$I_{\alpha,\varepsilon} = [3n^{-1/12+\varepsilon}, \sqrt{9/400 - n^{-\alpha}}], \tag{G.13}$$

where in the above notations we have omitted the dependence on n, as it is clear from the context. We also use the scalars a and b to denote the following:

$$a := 1 - n^{-\alpha} \quad \text{and} \quad b := 1 + n^{-\alpha}.$$

With the above notation in place, observe that standard chi-squared tail bounds yield that $\mathbb{P}[\mathcal{E}_\alpha] \geq 1 - e^{-n^{1-2\alpha}/8} \geq 1 - \delta$. Moreover, invoking the lower bound on n in Theorem 8.1, we have that $[3n^{-1/12+\varepsilon}, 1/10] \subseteq I_{\alpha,\varepsilon}$. Now conditional on the high probability event $\mathcal{E}_\alpha$, the population EM update $\widetilde{M}_{n,1}(\theta)$, in absolute value, can be upper and lower bounded as follows:

$$\left|\widetilde{M}_{n,1}(\theta)\right| \leq \mathbb{E}_Y\left[Y \tanh\left(\frac{Y\left|\theta\right|}{a-\theta^2}\right)\right] = \left|\theta\right| \underbrace{\mathbb{E}_Y\left[\frac{Y}{\left|\theta\right|}\tanh\left(\frac{\left|\theta\right| X}{a-\theta^2}\right)\right]}_{=:\overline{\gamma}(\theta)}, \quad \text{and,}$$
$$\left|\widetilde{M}_{n,1}(\theta)\right| \geq \mathbb{E}_Y\left[Y \tanh\left(\frac{X\left|\theta\right|}{b-\theta^2}\right)\right] = \left|\theta\right| \underbrace{\mathbb{E}_Y\left[\frac{Y}{\left|\theta\right|}\tanh\left(\frac{\left|\theta\right| Y}{b-\theta^2}\right)\right]}_{=:\underline{\gamma}(\theta)},$$

where the last two inequalities follows directly from the definition of $\widetilde{M}_{n,1}(\theta)$ in equation (8.12a), and from the fact that for any fixed $y, \theta \in \mathbb{R}$, the function $w \mapsto y\tanh(y\left|\theta\right|/(w-\theta^2))$ is non-increasing in w for $w > \theta^2$. Consequently, in order to complete the proof, it suffices to establish the following bounds:

$$1 - 3\theta^6/2 \leq \underline{\gamma}(\theta), \quad \text{and} \quad \overline{\gamma}(\theta) \leq (1 - \theta^6/5). \tag{G.14}$$

The following properties of the hyperbolic function $x \mapsto x\tanh(x)$ are useful for our proofs:

Lemma G.2. *For any $x \in \mathbb{R}$, the following holds*

$$\textit{(Lower bound):} \quad x\tanh(x) \geq x^2 - \frac{x^4}{3} + \frac{2x^6}{15} - \frac{17x^8}{315},$$
$$\textit{(Upper bound):} \quad x\tanh(x) \leq x^2 - \frac{x^4}{3} + \frac{2x^6}{15} - \frac{17x^8}{315} + \frac{62x^{10}}{2835}.$$

See Appendix G.5.3 for its proof.

Given the bounds in Lemma G.2, we derive the upper and lower bounds in the inequality (G.14) separately.

Upper bound for $\overline{\gamma}(\theta)$: Invoking the upper bound on $x\tanh(x)$ from Lemma G.2, we find that

$$\overline{\gamma}(\theta) \leq \frac{a-\theta^2}{\theta^2}\Bigg(\frac{\theta^2}{(a-\theta^2)^2}\mathbb{E}\left[Y^2\right] - \frac{\theta^4}{3(a-\theta^2)^4}\mathbb{E}\left[Y^4\right] + \frac{2\theta^6}{15(a-\theta^2)^6}\mathbb{E}\left[Y^6\right]$$
$$- \frac{17\theta^8}{315(a-\theta^2)^8}\mathbb{E}\left[Y^8\right] + \frac{62\theta^{10}}{2835(a-\theta^2)^{10}}\mathbb{E}\left[Y^{10}\right]\Bigg).$$

Recall that, for $Y \sim \mathcal{N}(0,1)$, we have $\mathbb{E}\left[Y^{2k}\right] = (2k-1)!!$ for all $k \geq 1$. Therefore, the last inequality can be simplified to

$$\overline{\gamma}(\theta) \leq \frac{1}{a-\theta^2} - \frac{\theta^2}{(a-\theta^2)^3} + \frac{2\theta^4}{(a-\theta^2)^5} - \frac{17\theta^6}{3(a-\theta^2)^7} + \frac{62\theta^8}{3(a-\theta^2)^9}. \tag{G.15}$$

When $n^{-\alpha} + \theta^2 \leq 9/400$, we can verify that the following inequalities hold:

$$\begin{aligned}
\frac{1}{1-n^{-\alpha}-\theta^2} &\leq 1 + (n^{-\alpha}+\theta^2) + (n^{-\alpha}+\theta^2)^2 + (n^{-\alpha}+\theta^2)^3 + 2(n^{-\alpha}+\theta^2)^4, \\
-\frac{\theta^2}{(1-n^{-\alpha}-\theta^2)^3} &\leq -\theta^2\left(1 + 3(n^{-\alpha}+\theta^2) + 6(n^{-\alpha}+\theta^2)^2 + 10(n^{-\alpha}+\theta^2)^3\right), \\
\frac{\theta^4}{(1-n^{-\alpha}-\theta^2)^5} &\leq \theta^4\left(1 + 5(n^{-\alpha}+\theta^2) + 16(n^{-\alpha}+\theta^2)^2\right), \\
-\frac{\theta^6}{(1-n^{-\alpha}-\theta^2)^7} &\leq -\theta^6\left(1 + 7(n^{-\alpha}+\theta^2)\right), \\
\frac{\theta^8}{(1-n^{-\alpha}-\theta^2)^9} &\leq 5\theta^8/4.
\end{aligned}$$

Substituting $a = 1 - n^{-\alpha}$ into the bound (G.15) and doing some algebra with the above inequalities and using the fact that $\max\{\theta, n^{-\alpha}\} \leq 1$ we have that

$$\overline{\gamma}(\theta) \leq 1 - \frac{2}{3}\theta^6 + \frac{61}{6}\theta^8 + 100n^{-\alpha} \leq 1 - \frac{2}{5}\theta^6 + 100n^{-\alpha} \leq 1 - \frac{1}{5}\theta^6.$$

The second last inequality above follows since $\theta \leq 3/20$, and the last inequality above utilizes the fact that if $\alpha \geq 1/2 - 6\varepsilon$, then $\theta^6/5 \geq 100n^{-\alpha}$ for all $\theta \geq 3n^{-1/12+\varepsilon}$. This completes the proof of the upper bound of $\overline{\gamma}(\theta)$.

Lower bound for $\underline{\gamma}(\theta)$: We start by utilizing the lower bound of $x\tanh(x)$ in the expression for $\underline{\gamma}(\theta)$, which yields:

$$\underline{\gamma}(\theta) \geq \frac{1}{b-\theta^2} - \frac{\theta^2}{(b-\theta^2)^3} + \frac{2\theta^4}{(b-\theta^2)^5} - \frac{17\theta^6}{3(b-\theta^2)^7}. \tag{G.16}$$

Since $|\theta| \in [3n^{-1/12+\varepsilon}, \sqrt{9/400 - n^{-\alpha}}]$ by assumption, we have the following lower bounds:

$$\begin{aligned}
\frac{1}{1+n^{-\alpha}-\theta^2} &\geq 1 + (\theta^2-n^{-\alpha}) + (\theta^2-n^{-\alpha})^2 + (\theta^2-n^{-\alpha})^3 + (\theta^2-n^{-\alpha})^4, \\
-\frac{\theta^2}{(1+n^{-\alpha}\theta^2)^3} &\geq -\theta^2 - \left(1 + 3(\theta^2-n^{-\alpha}) + 6(\theta^2-n^{--\alpha})^2 + 11(\theta^2-n^{-\alpha})^3\right), \\
\frac{\theta^4}{(1+n^{-\alpha}-\theta^2)^5} &\geq \theta^4\left(1 + 5(\theta^2-n^{-\alpha}) + 15(\theta^2-n^{-\alpha}\right), \\
-\frac{\theta^6}{(1+n^{-\alpha}-\theta^2)^7} &\geq -\theta^6\left(1 + 8(\theta^2-n^{-\alpha})\right).
\end{aligned}$$

Substituting $b = 1 + n^{-\alpha}$ into the bound (G.16) and doing some algebra with the above inequalities and using the fact that $\max\{\theta, n^{-\alpha}\} \leq 1$ we have that

$$\underline{\gamma}(\theta) \geq 1 - \frac{2}{3}\theta^6 - \frac{76}{3}\theta^8 - 100n^{-\alpha} \geq 1 - \frac{5}{4}\theta^6 - 100n^{-\alpha} \geq 1 - \frac{3}{2}\theta^6,$$

The second last inequality above follows since $\theta \leq 3/20$, and the last inequality above utilizes the fact that if $\alpha \geq 1/2 - 6\varepsilon$, then $\theta^6/4 \geq 100n^{-\alpha}$ for all $\theta \geq 3n^{-1/12+\varepsilon}$. This completes the proof of the lower bound of $\underline{\gamma}(\theta)$.

Proof of contraction bound for $\overline{M}_1$: Note that it suffices to repeat the arguments with $a = 1$ and $b = 1$ in the RHS of the inequalities (G.15) and (G.16) respectively. Given the other computations, the remaining steps are straightforward algebra and are thereby omitted.

G.3.2 Proof of perturbation bound for $\widetilde{M}_{n,1}$

We now prove the bound (8.13c) which is based on standard arguments to derive Rademacher complexity bounds. We first symmetrize with Rademacher variables, and apply the Ledoux-Talagrand contraction inequality. We then invoke results on sub-Gaussian and sub-exponential random variables, and finally perform the associated Chernoff-bound computations to obtain the desired result.

To ease the presentation, we denote $\alpha := 1/2 - 2\beta$ and $\mathcal{I} := [1 - n^{-\alpha} - 1/64, 1 - n^{-\alpha}]$. Next we fix $r \in [0, 1/8]$ and define $\widetilde{r} := \frac{r}{1-n^{-\alpha}-1/64}$. For sufficiently large n, we have $\widetilde{r} \leq 2r$. Recall the definition (G.12) of the event: $\mathcal{E}_\alpha = \{|\sum_{j=1}^n X_j^2/n - 1| \leq n^{-\alpha}\}$. Conditional on the event $\mathcal{E}_\alpha$, the following inequalities hold

$$\begin{aligned}\left|M_{n,1}(\theta) - \widetilde{M}_{n,1}(\theta)\right| &\leq \sup_{\theta \in \mathbb{B}(0,r), \sigma^2 \in \mathcal{I}} \left| \frac{1}{n}\sum_{i=1}^n X_i \tanh\left(\frac{X_i\theta}{\sigma^2}\right) - \mathbb{E}\left[Y \tanh\left(\frac{Y\theta}{\sigma^2}\right)\right]\right| \\ &\leq \sup_{\widetilde{\theta} \in \mathbb{B}(0,\widetilde{r})} \left|\widehat{M}_n(\widetilde{\theta}) - \widehat{M}(\widetilde{\theta})\right|,\end{aligned}$$

with all them valid for any $\theta \in \mathbb{B}(0, r)$. Here Y denotes a standard normal variate $\mathcal{N}(0, 1)$ whereas the operators $\widehat{M}$ and $\widehat{M}_n$ are defined as

$$\widehat{M}(\widetilde{\theta}) := \mathbb{E}[Y \tanh(Y\widetilde{\theta})] \quad \text{and} \quad \widehat{M}_n(\widetilde{\theta}) := \frac{1}{n}\sum_{i=1}^n X_i \tanh(X_i\widetilde{\theta}).$$

To facilitate the discussion later, we define the unconditional random variable

$$Z := \sup_{\widetilde{\theta} \in \mathbb{B}(0,\widetilde{r})} \left|\widehat{M}_n(\widetilde{\theta}) - \widehat{M}(\widetilde{\theta})\right|.$$

Employing standard symmetrization argument from empirical process theory [241], we find that

$$\mathbb{E}[\exp(\lambda Z)] \leq \mathbb{E}\left[\exp\left(\sup_{\widetilde{\theta}\in\mathbb{B}(0,\widetilde{r})} \frac{2\lambda}{n}\sum_{i=1}^{n}\varepsilon_i \tanh(X_i\widetilde{\theta})X_i\right)\right],$$

where $\varepsilon_i, i \in [n]$ are i.i.d. Rademacher random variables independent of $\{X_i, i \in [n]\}$. Noting that, the following inequality with hyperbolic function $\tanh(x)$ holds

$$\left|\tanh(x\widetilde{\theta}) - \tanh(x\widetilde{\theta}')\right| \leq \left|(\widetilde{\theta} - \widetilde{\theta}')x\right| \qquad \text{for all } x.$$

Consequently for any given x, the function $\widetilde{\theta} \mapsto \tanh(x\widetilde{\theta})$ is Lipschitz. Invoking the Ledoux-Talagrand contraction result for Lipschitz functions of Rademacher processes [151] and following the proof argument from Lemma 7.1, we obtain that

$$Z \leq c\widetilde{r}\sqrt{\frac{\log(1/\delta)}{n}}, \qquad \text{with probability } \geq 1-\delta,$$

for some universal constant c. Finally, using $\widetilde{r} \leq 2r$ for large n, we obtain that

$$\left|M_{n,1}(\theta) - \widetilde{M}_{n,1}(\theta)\right| \leq 2cr\sqrt{\frac{\log(1/\delta)}{n}}, \qquad \text{with probability } \geq 1-\delta-e^{-n^{1-2\alpha}/8},$$

where we have also used the fact that $\mathbb{P}[\mathcal{E}_\alpha] \geq 1 - e^{-n^{1-2\alpha}/8}$ from standard chi-squared tail bounds. The bound (8.13c) follows and we are done.

G.3.3 Proof of perturbation bound for $\overline{M}_1$

We now prove the bound (8.13d). Note that it suffices to establish the following point-wise result:

$$\left|\overline{M}_1(\theta) - M_{n,1}(\theta)\right| \precsim \frac{|\theta|^3\log^{10}(5n/\delta)}{\sqrt{n}} \quad \text{for all} \quad |\theta| \precsim n^{-1/16},$$

with probability at least $1-\delta$ for any given $\delta > 0$. For the reader's convenience, let us recall the definition of these operators

$$\overline{M}_1(\theta) = \mathbb{E}\Big[X\tanh(X\theta/(1-\theta^2))\Big], \tag{G.17a}$$

$$M_{n,1}(\theta) = \frac{1}{n}\sum_{i=1}^{n} X_i\tanh\Big(X_i\theta/(a_n-\theta^2)\Big), \tag{G.17b}$$

where $a_n := \sum_{i=1}^n X_i^2/n$. We further denote $\mu_k := \mathbb{E}_{X\sim\mathcal{N}(0,1)}[X^k]$, and $\widehat{\mu}_k := \frac{1}{n}\sum_{i=1}^n X_i^k$. From known results on Gaussian moments, we have $\mu_{2k} = (2k-1)!!$ for each integer $k = 1, 2, \ldots$.

For any given x and scalar b, consider the map $\theta \mapsto x\tanh(x\theta/(b-\theta^2))$. The 9-th order Taylor series for this function around $\theta = 0$ is given by

$$\begin{aligned} x\tanh(x\theta/(b-\theta^2)) = \frac{\theta x^2}{b} - \frac{\theta^3(x^4-3bx^2)}{3b^3} + \theta^5\left(\frac{2x^6}{15b^5} - \frac{x^4}{b^4} + \frac{x^2}{b^3}\right) \\ + \theta^7\left(-\frac{17x^8}{315b^7} + \frac{2x^6}{3b^6} - \frac{2x^4}{b^5} + \frac{x^2}{b^4}\right) \\ + \theta^9\left(\frac{62x^{10}}{2835b^9} - \frac{17x^8}{45b^8} + \frac{2x^6}{b^7} - \frac{10x^4}{3b^6} + \frac{x^2}{b^5}\right) + \varepsilon, \end{aligned} \tag{G.18}$$

where the remainder ε satisfies $\varepsilon \leq \mathcal{O}(\theta^{11})$. Plugging in this expansion with $b = 1$ on RHS of equation (G.17a) and taking expectation over $X \sim \mathcal{N}(0,1)$, we obtain

$$\overline{M}_1(\theta) = \theta + \theta^3\big(\sum_{k=1}^{2} c_{3,k}\mu_{2k}\big) + \theta^5\big(\sum_{k=1}^{3} c_{5,k}\mu_{2k}\big) + \theta^7\big(\sum_{k=1}^{4} c_{7,k}\mu_{2k}\big) + \theta^9\big(\sum_{k=1}^{5} c_{9,k}\mu_{2k}\big) + \varepsilon, \tag{G.19a}$$

where we have used the notation $\mu_k := \mathbb{E}_{X\sim\mathcal{N}(0,1)}[X^k]$ and $c_{j,k}$ denote universal constants. Furthermore, plugging in the same expansion (G.18) with $b = a_n$ on RHS of equation (G.17b), we obtain the following expansion for the sample EM operator

$$M_{n,1}(\theta) = \theta + \theta^3\big(\sum_{k=1}^{2} c_{3,k}\frac{\widehat{\mu}_{2k}}{a_n^{1+k}}\big) + \theta^5\big(\sum_{k=1}^{3} c_{5,k}\frac{\widehat{\mu}_{2k}}{a_n^{2+k}}\big) + \theta^7\big(\sum_{k=1}^{4} c_{7,k}\frac{\widehat{\mu}_{2k}}{a_n^{3+k}}\big) + \theta^9\big(\sum_{k=1}^{5} c_{9,k}\frac{\widehat{\mu}_{2k}}{a_n^{4+k}}\big) + \varepsilon_n, \tag{G.19b}$$

where $\widehat{\mu}_k$ denotes the sample mean of X^k, i.e., $\widehat{\mu}_k := \frac{1}{n}\sum_{i=1}^n X_i^k$. In order to lighten the notation, we introduce the following convenient shorthand:

$$\beta_j = \sum_{k=1}^{\frac{j+1}{2}} c_{j,k}\mu_{2k} \quad \text{and} \quad \widehat{\beta}_j = \sum_{k=1}^{\frac{j+1}{2}} c_{j,k}\frac{\widehat{\mu}_{2k}}{a_n^{\frac{j-1}{2}+k}} \quad \text{for } j \in \{3,5,7,9\} =: \mathcal{J}. \tag{G.20}$$

A careful inspection reveals that $\beta_3 = \beta_5 = 0$. With the above notations in place, we find that

$$\begin{aligned} \left|\overline{M}_1(\theta) - M_{n,1}(\theta)\right| &= \big|\sum_{j\in\mathcal{J}} \theta^j(\beta_j - \widehat{\beta}_j)\big| + \varepsilon \\ &=: M_1 + M_2. \end{aligned}$$

Therefore, it remains to establish that

$$M_1 \precsim \frac{|\theta|^3 \log^5(5n/\delta)}{\sqrt{n}} \quad \text{and} \quad M_2 \precsim \frac{|\theta|^3 \log^5(5n/\delta)}{\sqrt{n}}, \tag{G.21}$$

with probability at least $1-\delta$ for any given $\delta > 0$. Since the remainder term is of order θ^{11}, the assumption $|\theta| \precsim n^{-1/16}$ ensures that the remainder term is bounded by a term of order $\theta^3/\sqrt{n}$ and thus the bound (G.21) on the second term M_2 follows.

We now use concentration properties of Gaussian moments in order to prove the bound (G.21) on the first term M_1. Since $|\theta| \leq 1$, it suffices to show that

$$\sup_{j \in \mathcal{J}} \left|\beta_j - \widehat{\beta}_j\right| \precsim \frac{\log^5(5n/\delta)}{\sqrt{n}} \tag{G.22}$$

with probability at least $1-\delta$. Using the relation (G.20), we find that

$$\begin{aligned} \left|\beta_j - \widehat{\beta}_j\right| = |\sum_{k=1}^{\frac{j+1}{2}} \left(c_{j,k}\mu_{2k} - c_{j,k}\frac{\widehat{\mu}_{2k}}{a_n^{\frac{j-1}{2}+k}} \right) | &\leq \sum_{k=1}^{\frac{j+1}{2}} \frac{c_{j,k}}{a_n^{\frac{j-1}{2}+k}} \left|\mu_{2k} - \widehat{\mu}_{2k}\right| + c_{j,k}(1 - a_n^{-\frac{j-1}{2}-k})\mu_{2k} \\ &\leq C \sum_{k=1}^{\frac{j+1}{2}} \left(\left|\mu_{2k} - \widehat{\mu}_{2k}\right| + \frac{\mu_{2k}}{\sqrt{n}} \right), \end{aligned} \tag{G.23}$$

for any $j \in \mathcal{J}$. Here in the last step we have used the following bounds:

$$\max_{j \in \mathcal{J}, k \leq \frac{j+1}{2}} c_{j,k} \leq C \quad \text{and} \quad \max_{j \in \mathcal{J}, k \leq \frac{j+1}{2}} (1 - a_n^{-\frac{j-1}{2}-k}) \leq \frac{C}{\sqrt{n}}$$

for some universal constant C. Thus a lemma for the $1/\sqrt{n}$-concentration[1] of higher moments of Gaussian random variable is now useful:

Lemma G.3. *Let $X_1, \ldots, X_n$ are i.i.d. samples from $\mathcal{N}(0,1)$ and let $\mu_{2k} := \mathbb{E}_{X \sim \mathcal{N}(0,1)}[X^{2k}]$ and $\widehat{\mu}_{2k} := \frac{1}{n}\sum_{i=1}^n X_i^{2k}$. Then, we have*

$$\mathbb{P}\left(|\widehat{\mu}_{2k} - \mu_{2k}| \leq \frac{C_k \log^k(n/\delta)}{\sqrt{n}} \right) \geq 1 - \delta \quad \textit{for any} \quad k \geq 1,$$

where C_k denotes a universal constant depending only on k.

See the Appendix G.5.4 for the proof.

For any $\delta > 0$, consider the event

$$\mathcal{E} := \left\{ \left|\mu_{2k} - \widehat{\mu}_{2k}\right| \leq \frac{C_k \log^k(5n/\delta)}{\sqrt{n}} \quad \text{for all } k \in \{2, 4, \ldots, 10\} \right\}. \tag{G.24}$$

[1] The bound from Lemma G.3 is sub-optimal for $k = 1$ but is sharper than the standard tail bounds for Gaussian polynomials of degree $2k$ for $k \geq 2$. The $1/\sqrt{n}$ concentration of higher moments is necessary to derive the sharp rates stated in our results.

Straightforward application of union bound with Lemma G.3 yields that $\mathbb{P}[\mathcal{E}] \geq 1-\delta$. conditional on the event $\mathcal{E}$ inequality (G.22) implies that

$$
\begin{aligned}
\sup_{j\in\mathcal{J}} \left|\beta_j - \widehat{\beta}_j\right| &\leq C \sup_{j\in\mathcal{J}} \sum_{k=1}^{\frac{j+1}{2}} \left(\left|\mu_{2k} - \widehat{\mu}_{2k}\right| + \frac{\mu_{2k}}{\sqrt{n}}\right) \\
&\leq C \sup_{j\in\{3,5,7,9\}} \frac{j+1}{2} \left(\left|\mu_{j+1} - \widehat{\mu}_{j+1}\right| + \frac{(j+1)!!}{\sqrt{n}}\right) \\
&\overset{(i)}{\leq} C \sup_{j\in\{3,5,7,9\}} (j-1) \left(\left|C_{\frac{j+1}{2}} \frac{\log^{\frac{j+1}{2}}(5n/\delta)}{\sqrt{n}}\right| + \frac{(j+1)!!}{\sqrt{n}}\right) \\
&\overset{(ii)}{\leq} C\frac{\log^5(5n/\delta)}{\sqrt{n}},
\end{aligned}
\tag{G.25}
$$

where step (i) follows from the definition of the event (G.24) and in step (ii) using the fact that $j \leq 9$ is bounded we absorbed all the constants into a single constant. Since the event $\mathcal{E}$ has probability at least $1-\delta$, the claim (G.22) now follows.

G.3.4 Sharpness of bounds of Lemma 8.1

In Figure G.1, we numerically verify the linear and cubic scaling of the bounds stated in Lemma 8.1.

G.4 Proof of Lemma 8.2

The proof of the perturbation bound (8.17b) is a standard extension of $d = 1$ case presented above in Section G.3.2, and thereby is omitted.

We now present the proof of the contraction bound (8.17a), which has several similarities with the proofs of bounds (8.13a) and (8.13b) from Lemma 8.1. In order to simplify notation, we use the shorthand $Z_{n,d} := \frac{1}{nd}\sum_{j=1}^n \|X_j\|_2^2$. Recalling the definition (8.16) of operator $\widetilde{M}_{n,d}(\theta)$, we have

$$
\left\|\widetilde{M}_{n,d}(\theta)\right\|_2 = \left\|\mathbb{E}_{Y\sim\mathcal{N}(0,1)}\left[Y \tanh\left(\frac{Y^\top\theta}{Z_{n,d} - \|\theta\|_2^2/d}\right)\right]\right\|_2. \tag{G.26}
$$

We can find an orthonormal matrix R such that $R\theta = \|\theta\|_2 e_1$, where e_1 is the first canonical basis in $\mathbb{R}^d$. Define the random vector $V = RY$. Since $Y \sim \mathcal{N}(0, I_d)$, we have that $V \sim$

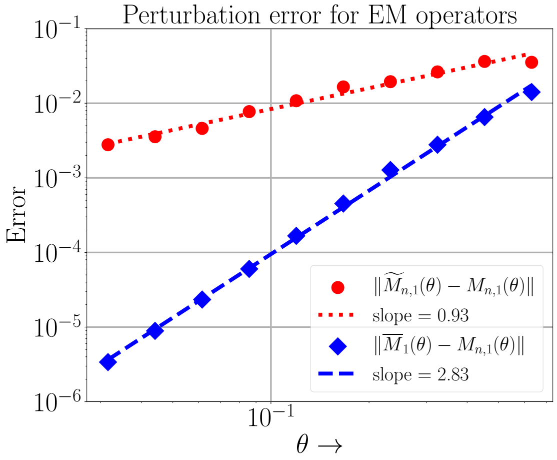


Figure G.1: Plots of the perturbation errors for the pseudo-population operator $\widetilde{M}_{n,1}$ (8.12a) and the corrected population operator $\overline{M}_1$ (8.12b) with respect to the sample EM operator $M_{n,1}$ (8.4), as a function of θ. From the least-squares fit on the log-log scale, we see that the error $\|\widetilde{M}_{n,1}(\theta) - M_{n,1}(\theta)\|$ scales linearly with θ, the error $\|\overline{M}_1(\theta) - M_{n,1}(\theta)\|$ has a cubic dependence on θ, in accordance with Lemma 8.1.

$\mathcal{N}(0, I_d)$. On performing the change of variables $Y = R^\top V$, we find that

$$\begin{aligned}\left\|\mathbb{E}_Y\left[Y \tanh\left(\frac{Y^\top\theta}{Z_{n,d} - \|\theta\|_2^2/d}\right)\right]\right\|_2 &= \left\|\mathbb{E}_V\left[R^\top V \tanh\left(\frac{\|\theta\|_2 V_1}{Z_{n,d} - \|\theta\|_2^2/d}\right)\right]\right\|_2 \\ &= \left|\mathbb{E}_{V_1}\left[V_1 \tanh\left(\frac{\|\theta\|_2 V_1}{Z_{n,d} - \|\theta\|_2^2/d}\right)\right]\right|\end{aligned}$$

where the final equality follows from the fact that

$$\mathbb{E}[R^\top V f(V_1)] = R^\top \mathbb{E}[V f(V_1)] = R^\top (\mathbb{E}[V_1 f(V_1)], 0, \ldots, 0)^\top.$$

Furthermore, the orthogonality of the matrix R implies that $\left\|\mathbb{E}[R^\top V f(V_1)]\right\|_2^2 = |\mathbb{E}[V_1 f(V_1)]|^2$.

In order to simplify the notation, we define the scalars a, b and the event $\mathcal{E}_{\alpha,d}$ as follows:

$$a := 1 - (nd)^{-\alpha}, \quad b := 1 + (nd)^{-\alpha}, \quad \text{and} \quad \mathcal{E}_{\alpha,d} = \left\{|Z_{n,d} - 1| \leq (nd)^{-\alpha}\right\}, \tag{G.27a}$$

where α is a suitable scalar to be specified later. Note that standard chi-squared tail bounds guarantee that

$$\mathbb{P}[\mathcal{E}_{\alpha,d}] \geq 1 - 2e^{-d^{2\alpha}n^{1-2\alpha}/8}. \tag{G.27b}$$

Now conditional on the event $\mathcal{E}_{\alpha,d}$, we have

$$\left\|\widetilde{M}_{n,d}(\theta)\right\|_2 \leq \left|\mathbb{E}_{V_1}\left[V_1 \tanh\left(\frac{\|\theta\|_2 V_1}{a-\|\theta\|_2^2/d}\right)\right]\right| = \|\theta\|_2 \underbrace{\mathbb{E}_{V_1}\left[\frac{V_1}{\|\theta\|_2}\tanh\left(\frac{\|\theta\|_2 V_1}{a-\|\theta\|_2^2/d}\right)\right]}_{=:\overline{\rho}(\theta)}, \quad \text{and,}$$

$$\left\|\widetilde{M}_{n,d}(\theta)\right\|_2 \geq \left|\mathbb{E}_{V_1}\left[V_1 \tanh\left(\frac{\|\theta\|_2 V_1}{b-\|\theta\|_2^2/d}\right)\right]\right| = \|\theta\|_2 \underbrace{\mathbb{E}_{V_1}\left[\frac{V_1}{\|\theta\|_2}\tanh\left(\frac{\|\theta\|_2 V_1}{b-\|\theta\|_2^2/d}\right)\right]}_{=:\underline{\rho}(\theta)},$$

where the above inequalities follow from the fact that for any fixed $y, \theta \in \mathbb{R}^d$, the function $w \mapsto y\tanh(y\|\theta\|_2/(w-\|\theta\|_2^2/d))$ is non-increasing in w for $w > \|\theta\|_2^2/d$.

Substituting $\alpha = 1/2 - 2\varepsilon$ in the bound (G.27b) and invoking the large sample size assumption in the theorem statement, we obtain that $\mathbb{P}[\mathcal{E}_{\alpha,d}] \geq 1-\delta$. Putting these observations together, it remains to prove that

$$\underline{\rho}(\theta) \geq \left(1 - \frac{3\|\theta\|_2^2}{4}\right)\|\theta\|_2^2, \qquad \text{and} \quad \overline{\rho}(\theta) \leq \left(1 - \left(1-\frac{1}{d}\right)\frac{\|\theta\|_2^2}{4}\right)\|\theta\|_2^2, \tag{G.28}$$

for all $5(d/n)^{-1/4+\varepsilon} \leq \|\theta\|_2^2 \leq (d-1)/(6d-1)$ conditional on the event $\mathcal{E}_{\alpha,d}$ for $\alpha = 1/2 - 6\varepsilon$ to obtain the conclusion of the theorem.

The proof of the claims in equation (G.28) relies on the following bounds on the hyperbolic function $\tanh(x)$. For any $x \in \mathbb{R}$, the following bounds hold:

$$\text{(Upper bound)} \quad x^2 - \frac{x^4}{3} + \frac{2x^6}{15} \;\geq\; x\tanh(x) \;\geq\; x^2 - \frac{x^4}{3} \quad \text{(Upper bound)}. \tag{G.29}$$

We omit the proof of these bounds, as it is very similar to that of similar results stated and proven later in Lemma G.2. We now turn to proving the bounds stated in equation (G.28) one-by-one.

Bounding $\overline{\rho}(\theta)$: Applying the upper bound (G.29) for $x\tanh(x)$, we obtain that

$$\overline{\rho}(\theta) \leq \frac{a-\|\theta\|_2^2/d}{\|\theta\|_2^2}\left(\frac{\|\theta\|_2^2}{(a-\|\theta\|_2^2/d)^2}\mathbb{E}\left[V_1^2\right] - \frac{\|\theta\|_2^4}{3(a-\|\theta\|_2^2/d)^4}\mathbb{E}\left[V_1^4\right] + \frac{2\|\theta\|_2^6}{15(a-\|\theta\|_2^2/d)^6}\mathbb{E}\left[V_1^6\right]\right).$$

Substituting $\mathbb{E}\left[V_1^{2k}\right] = (2k-1)!!$ for $k = 1, 2, 3$ in the RHS above, we find that

$$\overline{\rho}(\theta) \leq \frac{1}{a-\|\theta\|_2^2/d} - \frac{\|\theta\|_2^2}{(a-\|\theta\|_2^2/d)^3} + \frac{2\|\theta\|_2^4}{(a-\|\theta\|_2^2/d)^5}. \tag{G.30}$$

The condition $\|\theta\|_2^2 + (nd)^{-\alpha} \leq \frac{d-1}{6d-4} < 1/6$ implies the following bounds:

$$\begin{aligned}
\frac{1}{1-(nd)^{-\alpha} - \|\theta\|_2^2/d} &\leq 1 + \left((nd)^{-\alpha} + \|\theta\|_2^2/d\right) + 3/2 \cdot \left((nd)^{-\alpha} + \|\theta\|_2^2/d\right)^2, \\
\frac{1}{(1-(nd)^{-\alpha} - \|\theta\|_2^2/d)^3} &\geq 1 + 3\left((nd)^{-\alpha} + \|\theta\|_2^2/d\right), \\
\frac{1}{(1-(nd)^{-\alpha} - \|\theta\|_2^2/d)^5} &\leq 3/2.
\end{aligned}$$

Substituting the definitions (G.27a) of a and b and plugging the previous three bounds on the RHS of the inequality (G.30) yields that

$$\begin{aligned}
\overline{\rho}(\theta) &\leq 1 + \frac{\|\theta\|_2^2}{d} + \frac{3\|\theta\|_2^4}{2d^2} - \|\theta\|_2^2\left(1 + \frac{3\|\theta\|_2^2}{d}\right) + 3\|\theta\|_2^4 + \frac{11}{2}(nd)^{-\alpha} \\
&\leq 1 - \left(1 - \frac{1}{d}\right)\|\theta\|_2^2 + \left(3 - \frac{2}{d}\right)\|\theta\|_2^4 + \frac{11}{2}(nd)^{-\alpha} \\
&\leq 1 - \left(1 - \frac{1}{d}\right)\frac{\|\theta\|_2^2}{4}
\end{aligned}$$

where the last step follows from the following observations that

$$(3-2/d)\|\theta\|_2^4 \leq (1-1/d)\|\theta\|_2^2/2, \quad \text{for all } \|\theta\|_2 \leq (d-1)/(6d-4), \tag{G.31}$$

$$11(nd)^{-\alpha}/2 \leq (1-1/d)\|\theta\|_2^2/4, \quad \text{for all } \|\theta\|_2 \geq 5(d/n)^{-1/4+\varepsilon} \text{ when } \alpha = 1/2 - 2\varepsilon. \tag{G.32}$$

Therefore, the claim with an upper bound of $\overline{\rho}(\theta)$ now follows.

Bounding $\underline{\rho}(\theta)$: Using the lower bound (G.29) for $x\tanh(x)$, we find that

$$\underline{\rho}(\theta) \geq \frac{b - \|\theta\|_2^2/d}{\|\theta\|_2^2}\left(\frac{\|\theta\|_2^2}{(b-\|\theta\|_2^2/d)^2}\mathbb{E}\left[V_1^2\right] - \frac{\|\theta\|_2^4}{3(b-\|\theta\|_2^2/d)^4}\mathbb{E}\left[V_1^4\right]\right) \tag{G.33}$$

$$= \frac{1}{b - \|\theta\|_2^2/d} - \frac{\|\theta\|_2^2}{(b-\|\theta\|_2^2/d)^3}. \tag{G.34}$$

The condition $\|\theta\|_2 - (nd)^{-\alpha} \geq 0$ leads to

$$\begin{aligned}
\frac{1}{1+(nd)^{-\alpha} - \|\theta\|_2^2/d} &\geq 1 + \left(\|\theta\|_2^2/d - (nd)^{-\alpha}\right) + \left(\|\theta\|_2^2/d - (nd)^{-\alpha}\right)^2, \\
\frac{1}{(1+(nd)^{-\alpha} - \|\theta\|_2^2/d)^3} &\leq 1 + 4\left(\|\theta\|_2^2/d - (nd)^{-\alpha}\right).
\end{aligned}$$

Applying these inequalities to the bound (G.34), we obtain that

$$\begin{aligned}\underline{\rho}(\theta) &\geq 1 + \frac{\|\theta\|_2^2}{d} + \frac{\|\theta\|_2^4}{d^2} - \|\theta\|_2^2\left(1 + \frac{4\,\|\theta\|_2^2}{d}\right) - 2(nd)^{-\alpha} \\ &\overset{(i)}{\geq} 1 - \|\theta\|_2^2\left(1 - \frac{1}{d}\right) - \frac{\|\theta\|_2^2}{6}\left(\frac{4}{d} - \frac{1}{d^2}\right) - \frac{\|\theta\|_2^2\,(1 - 1/d)}{11} \\ &\geq 1 - \frac{3\,\|\theta\|_2^2}{4}\end{aligned}$$

where step (i) in the above inequalities follows from the observations (G.31)-(G.32) above. The lower bound (G.28) for $\underline{\rho}(\theta)$ now follows.

G.5 Proofs of auxiliary results

In this appendix, we collect the proofs of several auxiliary results used in the earlier proofs.

G.5.1 Proof of Corollary G.1

In order to ease the presentation, we only provide the proof sketch for the localization argument with this corollary. The detail proof argument for the corollary can be argued in similar fashion as that of Theorem 8.1. In particular, we consider the iterations t such that $\theta_n^t \in [n^{-a_\ell}, n^{-a_r}]$ where $a_\ell > a_r$. For all such iterations with θ_n^t, invoking Lemma 8.1, we find that

$$\left|\widetilde{M}_{n,1}(\theta_n^t)\right| \lesssim \underbrace{(1 - n^{-6a_\ell})}_{=:\gamma_{a_\ell}}\left|\theta_n^t\right| \quad \text{and} \quad \left|M_{n,1}(\theta_n^t) - \widetilde{M}_{n,1}(\theta_n^t)\right| \lesssim n^{-a_r}/\sqrt{n}.$$

Therefore, we obtain that

$$\left|\theta_n^{t+T}\right| \leq \left|\widetilde{M}_{n,1}(\theta_n^{t+T-1})\right| + \left|\widetilde{M}_{n,1}(\theta_n^{t+T-1}) - M_{n,1}(\theta_n^{t+T-1})\right| \leq \gamma_{a_\ell}\theta_n^{t+T-1} + n^{-a_r}/\sqrt{n}.$$

Unfolding the above inequality T times, we find that

$$\begin{aligned}\left|\theta_n^{t+T}\right| \leq \gamma_{a_\ell}^2(\theta_n^{t+T-2}) + n^{-a_r}/\sqrt{n}(1+\gamma_m) &\leq \gamma_{a_\ell}^T\theta_n^t + (1 + \gamma_{a_\ell} + \ldots + \gamma_{a_\ell}^{T-1})n^{-a_r}/\sqrt{n} \\ &\leq e^{-Tn^{-6a_\ell}}n^{-a_r} + \frac{1}{1-\gamma_{a_\ell}} \cdot n^{-a_r}/\sqrt{n}.\end{aligned}$$

As T is sufficiently large such that the second term is the dominant term, we find that that

$$\left|\theta_n^{t+T}\right| \lesssim \frac{1}{1-\gamma_{a_\ell}} \cdot n^{-a_r}/\sqrt{n} = n^{6a_\ell - a_r - 1/2}.$$

Setting the RHS equal to n^{-a_ℓ}, we obtain the recursion that

$$a_\ell = \frac{a_r}{7} + \frac{1}{14}. \tag{G.35}$$

Solving for the limit $a_\ell = a_r = a_\star$ yields that $a_\star = 1/12$. It suggests that we eventually have $\theta_n^t \to \mathbb{B}(0, n^{-\frac{1}{12}})$. As a consequence, we achieve the conclusion of the corollary.

G.5.2 Proof of Lemma G.1

Without loss of generality, we can assume that $|\theta| \in [\omega^{-a_{\ell+1}}, \omega^{-a_\ell}]$. Conditional on the event $\mathcal{E}$, we have that

$$\left|\overline{M}_1(\theta)\right| \leq (1 - \omega^{-6a_{\ell+1}}/5)\,|\theta| \quad \text{ and } \quad \left|M_{n,1}(\theta) - \overline{M}_1(\theta)\right| \leq c_2 \omega^{-3a_\ell}\omega^{-\frac{1}{2}}.$$

As a result, we have

$$\begin{aligned}
|M_{n,1}(\theta)| \leq \left|M_{n,1}(\theta) - \overline{M}_1(\theta)\right| + \left|\overline{M}_1(\theta)\right| &\leq (1 - \omega^{-6a_{\ell+1}}/5)\,|\theta| + c_2\omega^{-\frac{1}{2}}\omega^{-3a_\ell} \\
&\leq (1 - \omega^{-6a_{\ell+1}}/5 + c_2\omega^{-\frac{1}{2}}\omega^{-2a_\ell})\omega^{-a_\ell} \\
&\leq \omega^{-a_\ell}.
\end{aligned}$$

Here, to establish the last inequality, we have used the following observation: for $\omega = n/c_{n,\delta}$ and that $n \geq (c')^{1/\varepsilon} c_{n,\delta}$, we have

$$5c_2\omega^{6a_{\ell+1} - 2a_\ell - 1/2} \leq 5c_2\omega^{4a_\ell - 1/2} \leq c'\omega^{4a_{\ell_\star} - 1/2} \leq c'\omega^{-4\varepsilon} \leq 1/(c')^3 \leq 1,$$

which leads to $-\omega^{-6a_{\ell+1}}/5 + c_2\omega^{-\frac{1}{2}}\omega^{-2a_\ell} \leq 0$. As a consequence, we achieve the conclusion of the lemma.

G.5.3 Proof of Lemma G.2

The proof of this lemma relies on an evaluation of coefficients with x^{2k} as $k \geq 1$. In particular, we divide the proof of the lemma into two key parts:

Upper bound: From the definition of hyperbolic function $\tanh(x)$, it is sufficient to demonstrate that

$$x\left(\exp(x) - \exp(-x)\right) \leq \left(x^2 - \frac{x^4}{3} + \frac{2x^6}{15} - \frac{17x^8}{315} + \frac{62x^{10}}{2835}\right)\left(\exp(x) + \exp(-x)\right).$$

Invoking the Taylor series of $\exp(x)$ and $\exp(-x)$, the above inequality is equivalent to

$$\sum_{k=0}^{\infty} \frac{2x^{2k+2}}{(2k+1)!} \leq \left(x^2 - \frac{x^4}{3} + \frac{2x^6}{15} - \frac{17x^8}{315} + \frac{62x^{10}}{2835}\right)\left(\sum_{k=0}^{\infty} \frac{2x^{2k}}{(2k)!}\right).$$

Our approach to solve the above inequality is to show that the coefficients of x^{2k} in the LHS is smaller than that of x^{2k} in the RHS for all $k \geq 1$. In fact, when $1 \leq k \leq 3$, we can quickly check that the previous observation holds. For $k \geq 4$, it suffices to validate that

$$\frac{2}{(2k)!} - \frac{2}{3(2k-2)!} + \frac{4}{15(2k-4)!} - \frac{34}{315(2k-6)!} + \frac{124}{2835(2k-8)!} - \frac{2}{(2k+1)!} \geq 0.$$

Direct computation with the above inequality leads to

$$(k-1)(k-2)(k-3)(k-4)(496k^4 - 1736k^3 + 1430k^2 + 446k - 381) \geq 0$$

for all $k \geq 4$, which is always true. As a consequence, we achieve the conclusion with the upper bound of the lemma.

Lower bound: For the lower bound of the lemma, it is equivalent to prove that

$$\sum_{k=0}^{\infty} \frac{2x^{2k+2}}{(2k+1)!} \geq \left(x^2 - \frac{x^4}{3} + \frac{2x^6}{15} - \frac{17x^8}{315}\right)\left(\sum_{k=0}^{\infty} \frac{2x^{2k}}{(2k)!}\right).$$

Similar to the proof technique with the upper bound, we only need to verify that

$$\frac{2}{(2k)!} - \frac{2}{3(2k-2)!} + \frac{4}{15(2k-4)!} - \frac{34}{315(2k-6)!} - \frac{2}{(2k+1)!} \leq 0$$

for any $k \geq 3$. The above inequality is identical to

$$(k-1)(k-2)(k-3)(4352k^3 - 4352k^2 - 512k + 1472) \geq 0$$

for all $k \geq 3$, which always holds. Therefore, we obtain the conclusion with the lower bound of the lemma.

G.5.4 Proof of Lemma G.3

The proof of this lemma is based on appropriate truncation argument. More concretely, given any positive scalar τ, and the random variable $X \sim \mathcal{N}(0,1)$, consider the pair of truncated random variables (Y, Z) defined by:

$$Y := X^{2k}\mathbb{I}_{|X| \leq \tau} \quad \text{and} \quad Z := X^{2k}\mathbb{I}_{|X| \geq \tau}. \tag{G.36}$$

With the above notation in place, for n i.i.d. samples $X_1, \ldots, X_n$ from $\mathcal{N}(0,1)$, we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i^{2k} = \frac{1}{n}\sum_{i=1}^{n} Y_i + \frac{1}{n}\sum_{i=1}^{n} Z_i := S_{Y,n} + S_{Z,n}.$$

where $S_{Y,n}$ and $S_{Z,n}$, denote the averages of the random variables $Y_i's$ and $Z_i's$ respectively. Observe that $|Y_i| \leq \tau^{2k}$ for all $i \in [n]$; consequently, by standard sub-Gaussian concentration of bounded random variables, we have

$$\mathbb{P}\left(|S_{Y,n} - \mathbb{E}\left[Y\right]| \geq t_1\right) \leq 2\exp\left(-\frac{nt_1^2}{2\tau^{4k}}\right). \tag{G.37}$$

Next, applying Markov's inequality with the non-negative random variable $S_{Z,n}$, we find that

$$\mathbb{P}\left(S_{Z,n} \geq t_2\right) \leq \frac{\mathbb{E}\left[S_{Z,n}\right]}{t_2} = \frac{\mathbb{E}\left[Z\right]}{t_2}. \tag{G.38}$$

By definition of the truncated random variable Y, we have $\mathbb{E}[Y] \leq \mathbb{E}[X^{2k}]$; moreover, an application of Holder's inequality to $\mathbb{E}\left[Z\right]$ yields

$$\mathbb{E}\left[Z\right] = \mathbb{E}\left(X^{2k}\mathbb{I}_{|X|\geq\tau}\right) \leq \sqrt{\mathbb{E}\left[X^{4k}\right]}\sqrt{\mathbb{P}\left(|X| \geq \tau\right)} \leq \sqrt{2\mathbb{E}\left[X^{4k}\right]}\exp(-\tau^2/4).$$

Combining the bounds on $\mathbb{E}[Y]$ and $\mathbb{E}[Z]$ with the inequalities (G.37) and (G.38) we deduce that

$$\frac{\sum_{i=1}^n X_i^{2k}}{n} \leq \mathbb{E}\left[Y\right] + t_1 + t_2 \leq \mathbb{E}\left[X^{2k}\right] + t_1 + t_2, \quad \text{and,} \tag{G.39a}$$

$$\frac{\sum_{i=1}^n X_i^{2k}}{n} \geq \mathbb{E}\left[X^{2k}\right] - t_1 - t_2\sqrt{2\mathbb{E}\left[X^{4k}\right]}\exp(-\tau^2/4) \tag{G.39b}$$

with probability at least $1 - \exp\left(-\frac{nt_1^2}{2\tau^{4k}}\right) - \sqrt{2\mathbb{E}\left[X^{4k}\right]}\exp(-\tau^2/4)$. Finally, given any $\delta > 0$, choose the scalars τ, t_1, t_2 as follows:

$$\tau = 2\sqrt{\log\left(\frac{2\sqrt{2n\mathbb{E}\left[X^{4k}\right]}}{\delta}\right)}, \quad t_1 = \tau^2\sqrt{\frac{1}{n}\log\left(\frac{2}{\delta}\right)} \quad \text{and} \quad t_2 = \frac{1}{\sqrt{n}}.$$

Substituting the choice of t_1, t_2 and τ, in bounds (G.39a) and (G.39b) we conclude that with probability at least $1 - \delta$

$$\left|\frac{\sum_{i=1}^n X_i^{2k}}{n} - \mathbb{E}\left[X^{2k}\right]\right| \leq \frac{C_k \log^k(n/\delta)}{\sqrt{n}},$$

where C_k is a universal constant that depends only on k. This completes the proof of Lemma G.3.

G.6 Special contraction of population EM in one step

We now describe a special one-step contraction property of the population operator.

Lemma G.4. *For any vector θ^0 such that $\|\theta^0\| \leq \sqrt{d}$, we have $\|\widetilde{M}_{n,d}(\theta^0)\| \leq \sqrt{2/\pi}$ with probability at least $1-\delta$.*

The proof of this lemma is a straightforward application of the proof argument in Lemma 8.2 in Appendix G.4. In order to simplify notations, we use the shorthand $Z_{n,d} = \sum_{j=1}^{n} \|X_j\|_2^2 /(nd)$. Recalling the definition (8.16) of operator $\widetilde{M}_{n,d}$, we have

$$\left\|\widetilde{M}_{n,d}(\theta)\right\|_2 = \left\|\mathbb{E}_{Y\sim\mathcal{N}(0,1)}\left[Y\tanh\left(\frac{Y^\top\theta}{Z_{n,d} - \|\theta\|_2^2/d}\right)\right]\right\|_2.$$

As demonstrated in the proof of Theorem 8.2, we have the equivalence

$$\left\|\widetilde{M}_{n,d}(\theta)\right\|_2 = \mathbb{E}\left[V_1\tanh\left(\frac{\|\theta\|_2 V_1}{Z_{n,d} - \|\theta\|_2^2/d}\right)\right]$$

where $V_1 \sim \mathcal{N}(0,1)$. Since the function $x\tanh\left(\frac{\|\theta\|_2 x}{a-\|\theta\|_2^2/d}\right)$ is an even function in terms of x for any given a, we find that

$$\begin{aligned}\mathbb{E}\left[V_1\tanh\left(\frac{\|\theta\|_2 V_1}{Z_{n,d} - \|\theta\|_2^2/d}\right)\right] &= \mathbb{E}\left[|V_1|\tanh\left(\frac{\|\theta\|_2 |V_1|}{Z_{n,d} - \|\theta\|_2^2/d}\right)\right] \\ &\leq \mathbb{E}\left[|V_1|\right] = \sqrt{\frac{2}{\pi}}\end{aligned}$$

where the second inequality is due to the basic inequality $\tanh(x) \leq 1$ for all $x \in \mathbb{R}$. The inequality in the above display implies that regardless of the initialization θ^0, we always have $\left\|\widetilde{M}_{n,d}(\theta)\right\|_2 \leq \sqrt{2/\pi}$, as claimed.

G.7 Wasserstein Distance

In Figures 8.1 and 8.3, we use EM to estimate all the parameters of the fitted Gaussian mixture (e.g., the parameters $\{w_i, \mu_i, \Sigma_i, i \in [k]\}$ if the fitted mixture were $\mathcal{G} = \sum_{i=1}^{k} w_i\mathcal{N}(\mu_i, \Sigma_i)$) and use first-order Wasserstein distance between the fitted model and the true model to measure the quality of the estimate. Here we briefly summarize the definition of the first-order Wasserstein distance and refer the readers to the book [244] and the paper [113] for more details. Given two Gaussian mixture distributions of the form

$$\mathcal{G} = \sum_{i=1}^{k} w_i\mathcal{N}(\mu_i, \Sigma_i) \quad \text{and} \quad \mathcal{G}' = \sum_{j=1}^{k'} w_j\mathcal{N}(\mu_j', \Sigma_j),$$

the first-order Wasserstein distance between the two is given by

$$W_1(\mathcal{G}, \mathcal{G}') = \inf_{q \in \mathcal{Q}} \sum_{i=1}^{k} \sum_{j=1}^{k'} q_{ij} \left(\left\| \theta_i - \theta_j' \right\|_2 + \left\| \Sigma_i - \Sigma_j' \right\|_{\text{fro}} \right), \tag{G.40}$$

where $\left\| A \right\|_{\text{fro}}$ denotes the Frobenius norm of the matrix A (which in turn is defined as $\sqrt{\sum_{ij} A_{ij}^2}$). Moreover, $\mathcal{Q}$ denotes the set of all couplings on $[k] \times [k']$ such that

$$q_{ij} \in [0, 1], \qquad \sum_{i=1}^{k} q_{ij} = w_j' \quad \text{and} \quad \sum_{j=1}^{k'} q_{ij} = w_i \quad \text{for all} \quad i \in [k], j \in [k'].$$

We note that the optimization problem (G.40) is a linear program in the $k \times k'$ dimensional variable q and standard linear program solvers can be used for solving it. Also, we remark that here we have abused the notation slightly since the the definition of the Wasserstein distance above is typically used for the mixing measures which only depends on the parameters of the Gaussian mixture (and not the Gaussian density). Finally, applying definition (G.40), we can directly conclude that for the symmetric fit (8.1), we have

$$W_1 \left(\frac{1}{2}\mathcal{N}(\theta, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(-\theta, \sigma^2 I_d), \mathcal{N}(\theta_\star, \sigma_\star^2 I_d) \right) = \left\| \theta - \theta_\star \right\|_2 + \sqrt{d}\sqrt{|\sigma^2 - \sigma_\star^2|}, \tag{G.41}$$

where we have assumed that $\min \left\{ \left\| \theta - \theta_\star \right\|_2, \left\| -\theta - \theta_\star \right\|_2 \right\} = \left\| \theta - \theta_\star \right\|_2$.

Appendix H

Content Deferred From Chapter 9

In this chapter, we collect some tables deferred from the main text, derive the variance formula in Appendix H.1 that we used earlier to define the t-statistics (9.11a), and discuss several data cleaning details in Appendix H.2.

H.1 Derivation of variance formula in t-statistic

In this section, we derive a formula for the variance of $\widehat{\tau}_{\mathbf{G}} - \widehat{\tau}_{\text{ATE}}$, thereby justifying the formula for the plug-in estimator used in the definition of the t-statistic, which we repeat here for convenience.

$$\mathbb{T}_{\mathbf{G}} := \frac{\widehat{\tau}_{\mathbf{G}} - \widehat{\tau}_{\text{ATE}}}{\sqrt{\widehat{\text{Var}}(\widehat{\tau}_{\mathbf{G}} - \widehat{\tau}_{\text{ATE}})}}, \tag{H.1}$$

We first group terms to get

$$\begin{aligned}\widehat{\tau}_{\mathbf{G}} - \widehat{\tau}_{\text{ATE}} &= \left(\frac{1}{\text{abs}\,\mathbf{G}\cap\mathbf{T}}\sum_{i\in\mathbf{G}\cap\mathbf{T}} Y_i(1) - \frac{1}{\text{abs}\,\mathbf{G}\cap\mathbf{C}}\sum_{i\in\mathbf{G}\cap\mathbf{C}} Y_i(0)\right) - \left(\frac{1}{\text{abs}\,\mathbf{T}}\sum_{i\in\mathbf{T}} Y_i(1) - \frac{1}{\text{abs}\,\mathbf{C}}\sum_{i\in\mathbf{C}} Y_i(0)\right)\\ &= \alpha_1 \sum_{i\in\mathbf{G}\cap\mathbf{T}} Y_i(1) + \alpha_0 \sum_{i\in\mathbf{G}\cap\mathbf{C}} Y_i(0) + \beta_1 \sum_{i\in\mathbf{G}^c\cap\mathbf{T}} Y_i(1) + \beta_0 \sum_{i\in\mathbf{G}^c\cap\mathbf{C}} Y_i(0)\end{aligned}$$

where

$$\alpha_1 = \left(\frac{1}{\text{abs}\,\mathbf{G}\cap\mathbf{T}} - \frac{1}{\text{abs}\,\mathbf{T}}\right), \quad \alpha_0 = -\left(\frac{1}{\text{abs}\,\mathbf{G}\cap\mathbf{C}} - \frac{1}{\text{abs}\,\mathbf{C}}\right), \quad \beta_1 = -\frac{1}{\text{abs}\,\mathbf{T}}, \quad \text{and} \quad \beta_0 = \frac{1}{\text{abs}\,\mathbf{C}}.$$

Next, observe that even after we condition on $\mathcal{F}$, the collection of random variables $\{Y_i(1), Y_i(0) \colon 1 \leq i \leq N\}$ are fully independent, and furthermore, the terms within each sum are identically distributed. Applying the linearity of variance thus gives us

$$\begin{aligned}\text{Var}\left[\widehat{\tau}_{\mathbf{G}} - \widehat{\tau}_{\text{ATE}} \mid \mathcal{F}\right] = \alpha_1^2\, \text{abs}\,\mathbf{G}\cap\mathbf{T}\cdot \text{Var}\left[Y(1) \,\middle|\, \mathbf{G}\cap\mathbf{T}\right] + \alpha_0^2\, \text{abs}\,\mathbf{G}\cap\mathbf{C}\cdot \text{Var}\left[Y(0) \,\middle|\, \mathbf{G}\cap\mathbf{C}\right]\\ + \beta_1^2\, \text{abs}\,\mathbf{G}^c\cap\mathbf{T}\cdot \text{Var}\left[Y(1) \,\middle|\, \mathbf{G}^c\cap\mathbf{T}\right] + \beta_0^2\, \text{abs}\,\mathbf{G}^c\cap\mathbf{C}\cdot \text{Var}\left[Y(0) \,\middle|\, \mathbf{G}^c\cap\mathbf{C}\right]\end{aligned}$$

Estimator $\mathbf{M}$	$\overline{A}_{1,2}$	$\overline{A}_{2,3}$	$\overline{A}_{3,4}$	$\overline{A}_{4,5}$	$\overline{A}_{1,\min}$
t_logistic	**1.00**	0.67	**0.83**	0.25	**1.00**
causal_forest_2	**1.00**	0.50	**0.83**	0.17	**1.00**
x_lasso	**1.00**	0.50	0.58	0.67	**1.00**
x_rf	**1.00**	0.42	0.42	0.67	**1.00**
t_lasso	**1.00**	0.42	0.50	0.58	0.92
x_logistic	**1.00**	0.33	0.50	0.75	0.92
s_xgb	**1.00**	0.67	0.58	0.58	0.92
r_lassolasso	0.92	0.42	0.42	**0.92**	0.92
r_rfrf	0.92	0.50	0.42	0.50	0.92
r_lassorf	0.92	0.42	0.42	0.42	0.92
causal_forest_1	0.92	0.67	0.75	0.50	0.83
x_xgb	0.92	0.33	0.50	0.83	0.83
t_xgb	0.92	0.42	0.67	0.17	0.83
t_rf	0.92	**0.75**	0.50	0.33	0.83
causal_tree_2	0.92	**0.75**	0.25	0.42	0.75
s_rf	0.83	0.58	0.67	0.42	0.75
causal_tree_1	0.83	0.58	0.17	0.67	0.67

(a) GI Event

Estimator $\mathbf{M}$	$\overline{A}_{1,2}$	$\overline{A}_{2,3}$	$\overline{A}_{3,4}$	$\overline{A}_{4,5}$	$\overline{A}_{5,\max}$
t_lasso	0.33	0.42	0.42	**1.00**	**1.00**
x_xgb	0.33	0.50	0.58	0.92	0.92
x_logistic	0.50	0.50	0.42	0.92	0.92
r_rfrf	0.25	0.42	0.50	0.92	0.83
s_rf	0.42	0.42	0.42	0.92	0.83
x_lasso	0.50	0.33	0.50	0.83	0.75
t_rf	0.33	0.25	**0.67**	0.83	0.75
x_rf	0.50	0.33	0.58	0.83	0.75
t_logistic	0.33	0.25	0.58	0.83	0.75
r_lassorf	0.17	0.42	0.42	0.92	0.75
causal_forest_1	**0.67**	0.33	**0.67**	0.92	0.75
causal_forest_2	0.50	0.08	0.33	0.92	0.75
r_lassolasso	0.17	**0.75**	0.50	0.75	0.67
causal_tree_2	0.25	0.08	0.33	0.83	0.25
t_xgb	0.08	0.08	0.25	0.75	0.08

(b) CVT Event

Table H.1: Estimator-wise values of the mean scores $\overline{A}_{j,j+1}$ (9.8a) for $j = 1, 2, 3, 4$ for both GI and CVT events, $\overline{A}_{1,\min}$ (9.8b) for the GI event, and $\overline{A}_{5,\max}$ (9.8c) for the CVT event, where the mean was taken over the 12 validation folds, 4 each from the 3 random CV splits {cv_orig,cv_0,cv_1}. In each column the maximum score is highlighted in bold. The estimators are listed in the order sorted by the value in last column. Recall that each column was plotted earlier as a boxplot in Fig. 9.4(a).

where $\mathrm{Var}\left[Y(1) \mid \mathbf{G} \cap \mathbf{T}\right]$ denotes the variance of $Y(1)$ when conditioned on $X \in \mathbf{G}$ (recall our abuse of notation described in Section 9.3) and $T = 1$, with the other terms defined similarly. Simplifying this formula leads to

$$\begin{aligned}
&\mathrm{Var}\left[\widehat{\tau}_{\mathbf{G}} - \widehat{\tau}_{\mathrm{ATE}} \mid \mathcal{F}\right] \\
&= \left(1 - \frac{\mathrm{abs}\,\mathbf{G} \cap \mathbf{C}}{\mathrm{abs}\,\mathbf{C}}\right)^2 \cdot \frac{\mathrm{Var}\left[Y(0) \mid \mathbf{G} \cap \mathbf{C}\right]}{\mathrm{abs}\,\mathbf{G} \cap \mathbf{C}} + \left(1 - \frac{\mathrm{abs}\,\mathbf{G} \cap \mathbf{T}}{\mathrm{abs}\,\mathbf{T}}\right)^2 \cdot \frac{\mathrm{Var}\left[Y(1) \mid \mathbf{G} \cap \mathbf{T}\right]}{\mathrm{abs}\,\mathbf{G} \cap \mathbf{T}} \\
&+ \left(\frac{\mathrm{abs}\,\mathbf{G}^c \cap \mathbf{C}}{\mathrm{abs}\,\mathbf{C}}\right)^2 \cdot \frac{\mathrm{Var}\left[Y(0) \mid \mathbf{G}^c \cap \mathbf{C}\right]}{\mathrm{abs}\,\mathbf{G}^c \cap \mathbf{C}} + \left(\frac{\mathrm{abs}\,\mathbf{G}^c \cap \mathbf{T}}{\mathrm{abs}\,\mathbf{T}}\right)^2 \cdot \frac{\mathrm{Var}\left[Y(1) \mid \mathbf{G}^c \cap \mathbf{T}\right]}{\mathrm{abs}\,\mathbf{G}^c \cap \mathbf{T}} \\
&= \left(\frac{\mathrm{abs}\,\mathbf{G}^c \cap \mathbf{C}}{\mathrm{abs}\,\mathbf{C}}\right)^2 \cdot \left(\frac{\mathrm{Var}\left[Y(0) \mid \mathbf{G} \cap \mathbf{C}\right]}{\mathrm{abs}\,\mathbf{G} \cap \mathbf{C}} + \frac{\mathrm{Var}\left[Y(0) \mid \mathbf{G}^c \cap \mathbf{C}\right]}{\mathrm{abs}\,\mathbf{G}^c \cap \mathbf{C}}\right) \\
&+ \left(\frac{\mathrm{abs}\,\mathbf{G}^c \cap \mathbf{T}}{\mathrm{abs}\,\mathbf{T}}\right)^2 \cdot \left(\frac{\mathrm{Var}\left[Y(1) \mid \mathbf{G} \cap \mathbf{T}\right]}{\mathrm{abs}\,\mathbf{G} \cap \mathbf{T}} + \frac{\mathrm{Var}\left[Y(1) \mid \mathbf{G}^c \cap \mathbf{T}\right]}{\mathrm{abs}\,\mathbf{G}^c \cap \mathbf{T}}\right).
\end{aligned}$$

	$\texttt{Stab}(\mathbb{C})$-score in % with $\mathbf{G}_{\text{top}} = \widetilde{\mathbf{G}}_{\mathfrak{q}}$		
Cell $\mathbb{C}$ for GI event	$\mathfrak{q} = 0.2$	$\mathfrak{q} = 0.3$	Mean
{PPH=1}	**92**	**92**	**92**
{PSTRDS=1, HYPGRP=1}	**36**	**54**	**45**
{PSTRDS=1, ELDERLY=1}	**37**	**48**	**42**
{PNAPRXN=0, PSTRDS=1, ELDERLY=1}	23	18	21
{PNAPRXN=0, HYPGRP=1, PSTRDS=1}	25	8	17
{PSTRDS=1, PNSAIDS=0}	8	23	15
{WHITE=0, PSTRDS=1, ELDERLY=1}	18	3	11
{CHLGRP=1, HYPGRP=1}	17	2	10
{OBESE=1, WHITE=0, PSTRDS=1}	10	8	9
{PNAPRXN=0, ELDERLY=1}	0	18	9
{OBESE=1, WHITE=0}	0	17	8
{HYPGRP=1, PNSAIDS=0}	16	0	8
{WHITE=0, PNSAIDS=0}	14	0	7
{OBESE=1, WHITE=0, PNAPRXN=0}	3	10	7
{OBESE=1, PSTRDS=1, HYPGRP=1}	5	8	7
{PSTRDS=1, HYPGRP=1, ELDERLY=1}	12	0	6
{WHITE=0, PSTRDS=1, PNSAIDS=0}	10	2	6
{CHLGRP=1}	0	11	6
{PNAPRXN=0, HYPGRP=1}	0	10	5
{OBESE=1, PNSAIDS=0}	4	6	5

(a) GI Event

	$\texttt{Stab}(\mathbb{C})$-score in % with $\mathbf{G}_{\text{top}} = \widetilde{\mathbf{G}}^{c}_{\mathfrak{q}}$		
Cell $\mathbb{C}$ for CVT event	$\mathfrak{q} = 0.9$	$\mathfrak{q} = 0.8$	Mean
{ASPFDA=1}	**82**	50	**66**
{MALE=1, ELDERLY=1}	**70**	**57**	**64**
{ASCGRP=1}	**32**	**54**	**43**
{MALE=1}	0	**62**	31
{ELDERLY=1, SMOKE=1}	22	27	25
{MALE=1, ELDERLY=1, US=1}	30	0	15
{MALE=1, US=1}	0	26	13
{OBESE=1, ELDERLY=1}	0	21	10
{MALE=1, WHITE=1, ELDERLY=1}	20	0	10
{MALE=1, ASCGRP=1}	18	0	9
{WHITE=1, OBESE=1, ELDERLY=1}	0	15	8
{MALE=1, PPH=0, ELDERLY=1}	13	0	7
{MALE=1, WHITE=1}	0	12	6
{PPH=0, US=1, ASCGRP=1}	2	8	5
{WHITE=1, ELDERLY=1, SMOKE=1}	7	3	5
{ELDERLY=1, US=1, SMOKE=1}	7	3	5
{MALE=1, PPH=0}	0	9	4
{ELDERLY=1, US=1, CHLGRP=1}	0	8	4
{CHLGRP=1, ASCGRP=1}	8	0	4
{MALE=1, ELDERLY=1, SMOKE=1}	7	0	3

(b) CVT Event

Table H.2: $\texttt{Stab}(\mathbb{C})$-scores (in % rounded to nearest integer) for the top 20 cells $\mathbb{C}$ found by $\texttt{CellSearch}$-methodology for quantile-based top subgroups $\mathbf{G}_{\text{top}}$ of the ensemble CATE estimator. The cells are sorted by the "Mean" column of $\texttt{Stab}(\mathbb{C})$-scores, which in turn denote the average of the the scores in second and third columns. For each score column, cells corresponding to top-3 scores are displayed in bold. The choices $\mathfrak{q} = 0.2, 0.3$ for the GI event in panel **(a)**, and $\mathfrak{q} = 0.8, 0.9$ for the CVT event in panel **(b)** were made based on the results reported in Table 9.3 and the discussion around it.

H.2 Details on data cleaning with VIGOR and APPROVe

Here we collect additional details deferred from the main paper. First, we provide the details on how we identified the patients with prior history of GI event (PPH=1) for the VIGOR study. Although this subgroup was analyzed in the original study, the data files we had did not contain a membership indicator, not were there specific constructions on how to construct this subgroup. We applied a similar procedure to determine the patients with PPH=1 for the APPROVe study. Following that we describe the steps we followed to impute the GI ouctome as well as the features {ASPFDA, ASCGRP, HYPGRP, PSTRDS} for the APPROVe study. We also note that the other features, namely MALE and ELDERLY, reported in Table 9.6 could be readily identified from the demographics dataset for the APPROVe study, where the ELDERLY feature uses normalized age as detailed in Section 9.2.2.

Covariate (ABBRV)	Control No. (%)	Treatment No. (%)
Overall population	1300 (50.3)	1287 (49.7)
Demographics		
Whether *gender* is male (MALE=1)	805 (61.9)	804 (62.4)
Whether *adjusted age*† > 65 (ELDERLY=1)	338 (26.0)	329 (25.6)
Prior medical history		
of *GI PUB events** (PPH=1)	93 (7.2)	91 (7.1)
of *hypertension* (HYPGRP=1)	446 (34.3)	463 (36.0)
of *atherosclerotic cardiovascular disease* (ASCGRP=1)	121 (9.3)	129 (10.0)
indicating use of *aspirin* under FDA guidelines (ASPFDA=1)	70 (5.4)	81 (6.3)
Prior usage of drugs		
Whether used *glucocorticoids/steroids* (PSTRDS=1)	40 (3.1)	34 (2.6)
Outcomes		
Whether *GI event* occurred (GI=1)	6 (0.46)	27 (2.1)
Whether *CVT event* occurred (CVT=1)	32 (2.5)	57 (4.4)

Table H.3: Overview of the selected baseline covariates in the control and treatment arm of the APPROVe study. The treatment arm was given Vioxx, while the control arm was given placebo. †Adjusted age denotes age multiplied by the ratio of the life expectancy in the US to that in the individual's country of residence. *PUB stands for perforations, ulcers and bleeding.

PPH for both studies: To identify patients with a history of GI events, we identified a list of medical terms associated with such events, namely gastroduodenal perforation, obstruction, ulcer, or upper GI bleeding, from the medical history file. (We used REPTTERM field for this part as PREFTERM was not available in the medical history file for the VIGOR dataset.) Using this procedure, we identified 313 patients in the control arm, and 317 patients in the treatment arm who had a prior history of GI events (identified as PPH = 1). These number are off by 1 when compared to the 314 and 316 patients reported with PPH = 1 for the control and treatment arms respectively, by Bombardier et al in their paper on the VIGOR study [23].

To identify the patients with PPH = 1 for the APPROVe study, we used the medical terms identified above, with some adjustment for different spellings. Doing this gives us a subgroup of 184 patients. For this dataset, we used the PREFTERM in the medical history file for identification (since PREFTERM uses standardized terminology). Note that the paper on APPROVe study by Baron et al [12] does not report any information about the PPH feature.

GI outcome and other features for APPROVe study: On the VIGOR dataset, we identified all possible medical terms (PREFTERM field in the adverse event file) that were relevant and possibly associated with GI events during the treatment period. To be consistent with our procedure on VIGOR dataset, we excluded pre-treatment events, and included events that occurred during the treatment and post-study periods. A confirmed CVT event was a designated end point of the study, so these labels were directly provided to us in the study's data files. Such a process, i.e., using only a relevant list of medical terms in the adverse event file, correctly identified 166 out of 177 patients with GI events. Despite our best efforts, this procedure also falsely identified 12 out of the remaining 7899 patients who did not have a confirmed GI event. Next, we found that 33 patients in APPROVe had recorded adverse events with PREFTERM contained in the list of terms identified above (with some adjustment of different spellings) during the treatment or the post-study periods. We declared these 33 patients to have had a GI event. Because the APPROVe study did not aim to study GI toxicity, the paper on the study by Baron et al [12] does not report any information about the GI event as well as the risk factor features that we discuss next.

We followed a similar strategy to develop a mapping using the medical terms from the medical history file to the risk factor indicators for {ASPFDA, ASCGRP, HYPGRP} for the VIGOR study. Doing so, we correctly identified (i) 320/321 patients with ASPFDA = 1 (indication of aspirin usage by FDA due to their medical history), (ii) 453/454 patients with ASCGRP = 1 (history of atherosclerosis), and (iii) all 2385/2385 patients with HYPGRP = 1 (history of hypertension). For all three features ({ASPFD, ASCGRP, HYPGRP}), we did not have any false inclusion, i.e., using just the selected list of medical terms did not incorrectly impute a value of 1 for any patient. Finally to identify the patients with prior usage of glucortocoids (PSTRDS=1), we developed a mapping between the information from the concomitant therapy file and the PSTRDS indicator from the risk factor file. Our mapping correctly identified all 4479/4479 patients with PSTRDS = 1 but also falsely identified an additional 248 (out of the remaining 3597) patients.

The mappings described above were then used to impute the GI outcome and relevant missing features in the APPROVe study, thereby allowing us to report the "transfer" results for the subgroups found by StaDISC on the VIGOR study (Tables 9.4 and 9.5) to the APPROVe study (Table 9.6).