UCLA
UCLA Electronic Theses and Dissertations

Title
Distributed Stochastic Optimization in Non-Differentiable and Non-Convex Environments

Permalink
https://escholarship.org/uc/item/7pb746mg

Author
Vlaski, Stefan

Publication Date
2019

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Distributed Stochastic Optimization in Non-Differentiable and Non-Convex Environments

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical and Computer Engineering

by

Stefan Vlaski

2019

© Copyright by
Stefan Vlaski
2019

ABSTRACT OF THE DISSERTATION

Distributed Stochastic Optimization in
Non-Differentiable and Non-Convex Environments

by

Stefan Vlaski
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2019
Professor Ali H. Sayed, Chair

The first part of this dissertation considers distributed learning problems over networked agents. The general objective of distributed adaptation and learning is the solution of global, stochastic optimization problems through localized interactions and without information about the statistical properties of the data.

Regularization is a useful technique to encourage or enforce structural properties on the resulting solution, such as sparsity or constraints. A substantial number of regularizers are inherently non-smooth, while many cost functions are differentiable. We propose distributed and adaptive strategies that are able to minimize aggregate sums of objectives. In doing so, we exploit the structure of the individual objectives as sums of differentiable costs and non-differentiable regularizers. The resulting algorithms are adaptive in nature and able to continuously track drifts in the problem; their recursions, however, are subject to persistent perturbations arising from the stochastic nature of the gradient approximations and from disagreement across agents in the network. The presence of non-smooth, and potentially unbounded, regularizers enriches the dynamics of these recursions. We quantify the impact of this interplay and draw implications for steady-state performance as well as algorithm design and present applications in distributed machine learning and image reconstruction.

There has also been increasing interest in understanding the behavior of gradient-descent algorithms in non-convex environments. In this work, we consider stochastic cost functions,

where exact gradients are replaced by stochastic approximations and the resulting gradient noise persistently seeps into the dynamics of the algorithm. We establish that the diffusion learning algorithm continues to yield meaningful estimates in these more challenging, non-convex environments, in the sense that (a) despite the distributed implementation, individual agents cluster in a small region around the weighted network centroid in the mean-fourth sense, and (b) the network centroid inherits many properties of the centralized, stochastic gradient descent recursion, including the escape from strict saddle-points in time inversely proportional to the step-size and return of approximately second-order stationary points in a polynomial number of iterations.

In the second part of the dissertation, we consider centralized learning problems over networked feature spaces. Rapidly growing capabilities to observe, collect and process ever increasing quantities of information, necessitate methods for identifying and exploiting structure in high-dimensional feature spaces. Networks, frequently referred to as graphs in this context, have emerged as a useful tool for modeling interrelations among different parts of a data set. We consider graph signals that evolve dynamically according to a heat diffusion process and are subject to persistent perturbations. The model is not limited to heat diffusion but can be applied to modeling other processes such as the evolution of interest over social networks and the movement of people in cities. We develop an online algorithm that is able to learn the underlying graph structure from observations of the signal evolution and derive expressions for its performance. The algorithm is adaptive in nature and able to respond to changes in the graph structure and the perturbation statistics. Furthermore, in order to incorporate prior structural knowledge to improve classification performance, we propose a BRAIN strategy for learning, which enhances the performance of traditional algorithms, such as logistic regression and SVM learners, by incorporating a graphical layer that tracks and learns in real-time the underlying correlation structure among feature subspaces. In this way, the algorithm is able to identify salient subspaces and their correlations, while simultaneously dampening the effect of irrelevant features.

The dissertation of Stefan Vlaski is approved.

Lieven Vandenberghe

Suhas N. Diggavi

Abeer A. Alwan

Ali H. Sayed, Committee Chair

University of California, Los Angeles

2019

To my Family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, and foremost, I would like to express my deepest gratitude to Professor Ali H. Sayed, for his guidance as well as support throughout my studies and the opportunity to work on exciting and challenging projects. His passion and knowledge have inspired me to grow as a researcher; his patience and meticulous approach to science have laid the foundation to do so. Our interactions have offered invaluable lessons far beyond academic applications, and I am grateful to carry them with me for years to come.

I would like to thank Professor Alwan for her mentorship during my Masters studies at UCLA, as well as her participation on the doctoral committee for this dissertation. I would also like to thank Professor Christina Fragouli, Professor Suhas N. Diggavi and Professor Lieven Vandenberghe for their participation and feedback.

I am thankful to Deeona Columbia, Ryo Arreola, and Mandy Smith for their help during my time at UCLA. I am grateful to Professor Abdelhak Zoubir and Dr. Michael Muma at TU Darmstadt, whose mentorship led me to this path.

I am lucky to have met exceptional colleagues, collaborators and friends throughout my time at UCLA and while visiting EPFL – thank you for thoughtful discussions and occasional laughs: Chung-Kai Yu, Hawraa Salami, Bicheng Ying, Kun Yuan, Lucas Cassano, Sulaiman Alghunaim, Sina Basir-Kazeruni, Professor Dejan Marković, Zaid J. Towfic, Jianshu Chen, Xiaochuan Zhao, Steven Lee, Chengcheng Wang, Sara Al-Sayed, Saeed Ghazanfari Rad, Roula Nassif, Augusto Santos, Virginia Bordignon, Elsa Rizk, Guillermo Ortiz Jimenez, Hermina P. Maretić, Professor Pascal Frossard and Professor Ricardo Merched.

I am thankful to my family, Zaklina, Viktor, Slavka and Vasil, for their unconditional support, and to Cathrin, for being my partner along the way.

This dissertation is based upon work partially supported by the National Science Foundation under grants CCF-1011918, CCF-1524250 as well as ECCS-1407712 and DARPA project N66001-14-2-4029. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of

the National Science Foundation, the Department of Defense or the U.S. Government.

VITA

2013	B.Sc. in Electrical Engineering, Technical University Darmstadt, Germany.
2014	M.S. in Electrical Engineering, University of California, Los Angeles, CA, USA.
2014–2017	Research Assistant, Department of Electrical Engineering, University of California, Los Angeles, CA, USA.
2016	Intern, Apple Inc., Cupertino, CA, USA.
2017	Intern, Amazon Lab126, Sunnyvale, CA, USA.
2017–2019	Visiting Doctoral Assistant, École polytechnique fédérale de Lausanne, Switzerland.

PUBLICATIONS

Stefan Vlaski, Lieven Vandenberghe and Ali H. Sayed, "Regularized Diffusion Adaptation via Conjugate Smoothing," in preparation, September 2019.

Stefan Vlaski and Ali H. Sayed, "Second-Order Guarantees of Stochastic Gradient Descent in Non-Convex Optimization," submitted for publication, available as arXiv:1908.07023, August 2019.

Stefan Vlaski and Ali H. Sayed, "Distributed Learning in Non-Convex Environments – Part II: Polynomial Escape from Saddle-Points," submitted for publication, available as arXiv:1907.01849, July 2019.

Stefan Vlaski and Ali H. Sayed, "Distributed Learning in Non-Convex Environments – Part I: Agreement at a Linear Rate," submitted for publication, available as arXiv:1907.01848, July 2019.

Stefan Vlaski and Ali H. Sayed, "Diffusion Learning in Non-Convex Environments", in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5262–5266, Brighton, UK, May 2019.

Stefan Vlaski, Hermina P. Maretić, Roula Nassif, Pascal Frossard and Ali H. Sayed, "Online Graph Learning from Sequential Data", in Proceedings of the IEEE Data Science Workshop (DSW), pp. 190–194, Lausanne, Switzerland, June 2018.

Stefan Vlaski, Bicheng Ying and Ali H. Sayed, "The BRAIN Strategy for Online Learning", in Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 1285–1289, Washington, D.C., USA, December 2016.

Stefan Vlaski, Lieven Vandenberghe and Ali H. Sayed, "Diffusion Stochastic Optimization with Non-Smooth Regularizers", in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, pp. 4149–4153, March 2016.

Stefan Vlaski and Ali H. Sayed, "Proximal Diffusion for Stochastic Costs with Non-Differentiable Regularizers", in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3352–3356, Brisbane, Australia, April 2015.

CHAPTER 1

Introduction

1.1 Single-Agent Learning

Most learning problems can be formulated as stochastic optimization problems where the objective is to learn a parameter vector w that minimizes a risk $Q(w;\boldsymbol{x})$ over the distribution of the random data $\boldsymbol{x}$, i.e. [1]:

$$w^o \triangleq \arg\min_w \mathbb{E}_{\boldsymbol{x}}\, Q(w;\boldsymbol{x}) \triangleq \arg\min_w J(w) \tag{1.1}$$

where

$$J(w) \triangleq \mathbb{E}_{\boldsymbol{x}}\, Q(w;\boldsymbol{x}) \tag{1.2}$$

If $Q(w;\boldsymbol{x})$ is viewed as a penalty for the parameter set w given $\boldsymbol{x}$, then (1.1) can be viewed as the task of finding the parametrization w that gives the smallest *expected* penalty over the distribution of $\boldsymbol{x}$. The key challenge in learning by means of pursuing a solution to (1.1) is that in general, the distribution of the data $\boldsymbol{x}$ is unknown. Two main remedies exist for this challenge:

1.1.1 Empirical Risk Minimization

In empirical risk minimization, rather than solving (1.1) directly, S sample realizations of $\boldsymbol{x}$ are collected into a batch $\{x_s\}_{s=1}^S$ and the expectation is approximated by the sample mean [2]:

$$w^\star = \arg\min_w \frac{1}{S}\sum_{s=1}^{S} Q(w, x_s) \tag{1.3}$$

Note that, in general, the minimizer of the expected risk w^o will be different from the minimizer of the empirical risk $w^\star$. Under the assumption of ergodicity, and in light of the law of large numbers, we can nevertheless expect that $w^\star$ will be a reasonable estimate for w^o. This intuition can be formalized for a variety of data distributions and risk functions and is extensively studied [2–7].

It can be observed that (1.3), in contrast to (1.1), is now fully deterministic, and hence, $w^\star$ can be pursued by a variety of optimization algorithms. The most immediate solution is based on gradient descent:

$$w_i = w_{i-1} - \mu\nabla\left(\frac{1}{S}\sum_{s=1}^{S} Q(w, x_s)\right) = w_{i-1} - \frac{\mu}{S}\sum_{s=1}^{S}\nabla Q(w, x_s) \tag{1.4}$$

However, the approximation (1.3) has two main drawbacks. First, the formulation and solution of (1.3) requires the collection of a large number of samples $\{x_s\}_{s=1}^S$. This may not be feasible, particularly if (a) the sample size S is very large or (b) data is streaming in, requiring processing of samples on the fly. Second, guarantees on the accuracy of $w^\star$ relative to w^o are generally based on an ergodicity assumption, which is violated whenever data statistics drift over time.

1.1.2 Online Learning

Returning to (1.1), observe that if we had knowledge about the distribution of $\boldsymbol{x}$, we could simply iterate:

$$w_i = w_{i-1} - \mu\nabla J(w_{i-1}) = w_{i-1} - \mu\nabla\,\mathbb{E}_{\boldsymbol{x}}\,Q(w_{i-1};\boldsymbol{x}) \tag{1.5}$$

In order to derive the stochastic gradient algorithm, one can drop the expectation operation and replace the true gradient by an instantaneous approximation [1, 8]:

$$\widehat{\nabla J}(w) \triangleq \nabla Q(w;\boldsymbol{x}) \tag{1.6}$$

and instead iterate:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - \mu\widehat{\nabla J}(\boldsymbol{w}_{i-1}) = \boldsymbol{w}_{i-1} - \mu\nabla Q(\boldsymbol{w}_{i-1}; \boldsymbol{x}) \tag{1.7}$$

where $\boldsymbol{w}_i$ is denoted in boldface to emphasize the fact that it is now random. Observe that, rather than moving along the negative gradient direction, (approximate) descent occurs now relative to an *approximate* gradient direction. This approximation introduces noise into the evolution of the iterates $\boldsymbol{w}_i$. Indeed, if we denote:

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \triangleq \widehat{\nabla J}(\boldsymbol{w}_{i-1}) - \nabla J(\boldsymbol{w}_{i-1}) \tag{1.8}$$

we have for (1.7):

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - \mu\nabla J(\boldsymbol{w}_{i-1}) - \mu\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \tag{1.9}$$

Despite the presence of the gradient noise term $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$, it can be established that (1.7) will nevertheless approach a small region around the minimizer w^o under reasonable technical conditions on the cost functions and gradient noise term. Specifically, it holds in the mean-square sense that [1]:

$$\limsup_{i\to\infty} \mathbb{E}\,\|w^o - \boldsymbol{w}_i\|^2 = O(\mu) \tag{1.10}$$

1.1.3 Stochastic Gradient Algorithms for Empirical Risk Minimization

Observe from the gradient recursion to the deterministic, empirical risk (1.4), that every single gradient update from w_{i-1} to w_i requires the evaluation of S gradients, where S denotes the sample size. This can be prohibitively expensive, particularly for large data sizes. For this reason, a number of algorithms have been developed to alleviate the per-iteration cost of the gradient update in empirical risk minimization. The most basic algorithm is a variant of the online stochastic gradient algorithm (1.7), where instead of sampling from the true distribution of $\boldsymbol{x}$, at each iteration, data is sampled from the empirical distribution of $\boldsymbol{x}^{\text{emp}}$,

where:

$$\boldsymbol{x}^{\text{emp}} = \begin{cases} x_1, & \text{w.p. } \frac{1}{S}, \\ x_2, & \text{w.p. } \frac{1}{S}, \\ \vdots & \\ x_S, & \text{w.p. } \frac{1}{S}. \end{cases} \tag{1.11}$$

Then, the empirical risk minimization problem is equivalent to:

$$w^\star \triangleq \arg\min_w \frac{1}{S} \sum_{s=1}^{S} Q(w, x_s) = \arg\min_w \mathbb{E}_{\boldsymbol{x}^{\text{emp}}} Q(w; \boldsymbol{x}^{\text{emp}}) \tag{1.12}$$

This construction motivates the following stochastic gradient algorithm for empirical risk minimization:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - \mu \nabla Q(\boldsymbol{w}_{i-1}; \boldsymbol{x}^{\text{emp}}) \tag{1.13}$$

where the gradient now, in contrast to (1.4), is evaluated only at one sample per iteration. This construction reduces the computational complexity per iteration by a factor of S and can result in a significant improvement of the accuracy obtained after a limited number of gradient evaluations. It does, however, come at a cost. Since the true gradient, similarly to the online stochastic gradient iteration (1.7), is replaced by a stochastic gradient approximation, some gradient noise is introduced into the recursions, preventing the iterates $\boldsymbol{w}_i$ from converging to the minimizer $w^\star$ of (1.12). The work [9] has leveraged the analogy between the two problems (1.1) and (1.12) to obtain and accurate expression for the residual error in steady-state, namely:

$$\limsup_{i \to \infty} \mathbb{E} \left\| w^\star - \boldsymbol{w}_i \right\|^2 = O(\mu) \tag{1.14}$$

The observation that the residual error introduced by employing stochastic gradient approximations, rather than exact gradients, tends to be proportional to the variance introduced by the stochastic gradient noises, has sparked a line of work employing "variance-reduction" to reduce the variance of the stochastic gradient approximation over time [10]. These works generally rely on the assumption that the sample size S is finite, and are hence only applicable to empirical risk minimization (1.3).

Despite their apparent similarity, we draw in this work a clear distinction between the pursuit of w^o, the minimizer of the expected risk (1.1), and $w^\star$, the minimizer of the empirical risk (1.3). This distinction becomes particularly clear in the context of classification, where the empirical risk is generally referred to as the "training error", i.e., the performance of the parametrization w on the training data. The expected risk (1.1) on the other hand, denotes the expected performance of w on *unseen* data. While generalization theory loosely states that, as long as the classification surface is sufficiently simple, when compared to the sample size S, good training performance, i.e., a low empirical risk value, can guarantee small expected risk with high probability [2–4, 7], we emphasize that the fundamental objective of generalization ability is captured in the expected, rather than empirical risk. As such, in this dissertation, we will focus on developing learning solutions which are applicable to expected risk minimization. While these solutions will be applicable to empirical risk minimization by means of (1.11), we will not employ solutions which improve over (1.14) for the smaller class of empirical risk minimization problems.

1.2 Multi-Agent Learning

Rapid developments towards a networked and data-driven society have uncovered new challenges in the development of modern learning algorithms, where data driving modeling decisions is increasingly available at dispersed locations. Examples of such settings are social networks [11–13], power grids [14, 15], wireless sensor [16–18] and vehicular networks [19, 20] as well as cloud applications [21]. Limitations on communication, storage and computational resources as well as privacy and robustness concerns frequently prevent aggregation and processing of raw data at a central location [22].

Motivated by these considerations, the objective of distributed adaptation and learning is the solution of global, stochastic optimization problems across networks of agents through localized interactions and without information about the statistical properties of the data. The resulting algorithms are adaptive in nature and able to continuously track drifts in the problem. Extending the discussion from the single-agent problem (1.1), we now associate

with every agent, indexed by k a local cost function $J_k(w) : \mathbb{R}^M \to \mathbb{R}$ [1]:

$$J_k(w) \triangleq \mathbb{E}_{\boldsymbol{x}_k} Q_k(w; \boldsymbol{x}_k) \tag{1.15}$$

Observe that through the subscript k we emphasize different sources of heterogeneity across networks. Specifically, different agents may be observing data $\boldsymbol{x}_k$ from different distributions or may be interested in minimizing different risk functions $Q_k(\cdot;\cdot)$.

We consider a strongly-connected network consisting of N agents, depicted in Fig. 1.1. For any two agents k and ℓ, we attach a pair of non-negative coefficients $\{a_{\ell k}, a_{k\ell}\}$ to the edge linking them. The scalar $a_{\ell k}$ is used to scale data moving from agent ℓ to k; likewise, for $a_{k\ell}$. Strong-connectivity means that it is always possible to find a path, in either direction, with nonzero scaling weights linking any two agents (either directly if they are neighbors or indirectly through other agents). In addition, at least one agent k in the network possesses a self-loop with $a_{kk} > 0$. This condition ensures that at least one agent in the network has some confidence in its local information. Let $\mathcal{N}_k$ denote the set of neighbors of agent k. The coefficients $\{a_{\ell k}\}$ are convex combination weights that satisfy

$$a_{\ell k} \geq 0, \quad \sum_{\ell \in \mathcal{N}_k} a_{\ell k} = 1, \quad a_{\ell k} = 0 \text{ if } \ell \notin \mathcal{N}_k \tag{1.16}$$

If we introduce the combination matrix $A = [a_{\ell k}]$, it then follows from (1.16) and the strong-connectivity property that A is a left-stochastic primitive matrix. In view of the Perron-Frobenius Theorem [1,23,24], this ensures that A has a single eigenvalue at one while all other eigenvalues are inside the unit circle, so that $\rho(A) = 1$. Moreover, if we let p denote the right-eigenvector of A that is associated with the eigenvalue at one, and if we normalize the entries of p to add up to one, then it also holds that all entries of p are strictly positive, i.e.,

$$Ap = p, \quad \mathbb{1}^{\mathsf{T}} p = 1, \quad p_k > 0 \tag{1.17}$$

where the $\{p_k\}$ denote the individual entries of the Perron vector, p. One can then formulate

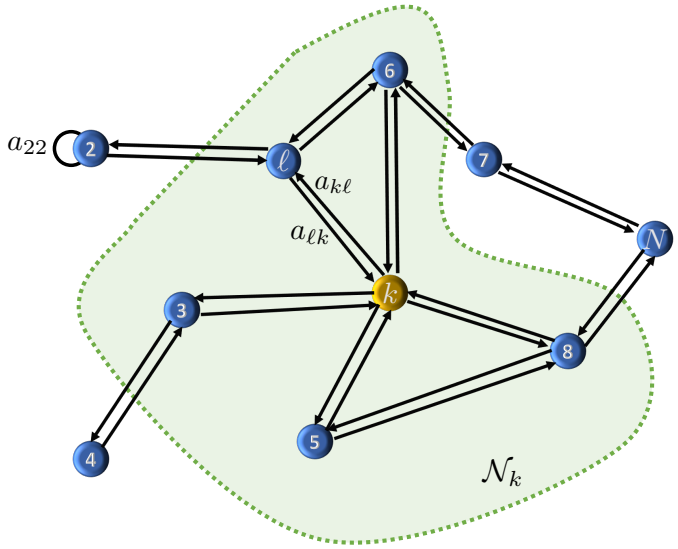

Figure 1.1: A network of N nodes with an emphasis on the neighborhood $\mathcal{N}_k$ of agent k.

the global learning problem [1, 25]:

$$w^o = \arg\min_{w} \sum_{k=1}^{N} p_k J_k(w) \tag{1.18}$$

The weights $\{p_k\}$ indicate that the resulting minimizer w^o can be interpreted as a Pareto solution for the collection of regularized risks $\{J_k(w)\}$ [1, 25]. The global optimization problem (1.18) can be approached through a variety of distributed algorithms, using both inexact [1, 26–28] and exact [29–31] gradients.

One approach for pursuing a solution of (1.18) in a distributed manner is the diffusion algorithm [1, 25].

Algorithm 1.1 Diffusion Strategy [1]

$$\boldsymbol{\phi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu \widehat{\nabla_w J}_k(\boldsymbol{w}_{k,i-1}) \tag{1.19}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell=1}^{N} a_{\ell k} \boldsymbol{\phi}_{\ell,i} \tag{1.20}$$

When the individual costs $J_k(w)$ are differentiable, and their weighted sum (1.18) is strongly-convex, the performance of this strategy has been studied in great detail. One of the key conclusions is that, despite the restriction of communication to localized interactions within neighborhoods, the iterates at every agent $\boldsymbol{w}_{k,i}$ cluster around the Pareto solution (1.18) in the mean-square sense, after a sufficient number of iterations [1, 25].

$$\limsup_{i\to\infty} \mathbb{E}\,\|w^o - \boldsymbol{w}_{k,i}\|^2 = O(\mu) \tag{1.21}$$

The effectiveness of the diffusion strategy for the pursuit of (1.18) has sparked a number of studies and extensions in recent years, including asynchronous [32], constrained [33], sub-gradient based [34] and multi-task [35–40] variations.

1.2.1 Regularized Learning

In many learning problems there exists a priori knowledge about the solution, such as sparsity or constraints. An effective method for encouraging the recovered solution to conform to this prior information is to add regularization $R_k(\cdot)$ to the data-dependent risk term $J_k(\cdot)$, i.e.,

$$w^o = \arg\min_{w} \sum_{k=1}^{N} p_k \left\{J_k(w) + R_k(w)\right\} \tag{1.22}$$

There are several useful works in the literature that study optimization problems with non-smooth regularizers primarily centered around sub-gradient [26, 41–44] and proximal [35, 45–48] constructions.

In this work, we will propose a modification of the diffusion strategy (1.19)–(1.20) based on the proximal operator. Recall that the proximal operator is defined as [49]:

$$\text{prox}_{\mu R_k}(x) \triangleq \arg\min_{u} \left(R_k(u) + \frac{1}{2\mu}\|x-u\|_2^2\right) \tag{1.23}$$

The proximal diffusion strategy then takes the form

Algorithm 1.2 Proximal Diffusion Strategy [50]

$$\boldsymbol{\phi}_{k,i} = \text{prox}_{\mu R_k}\left(\boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla_w J}_k(\boldsymbol{w}_{k,i-1})\right) \tag{1.24}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell=1}^{N} a_{\ell k}\boldsymbol{\phi}_{\ell,i} \tag{1.25}$$

Note that each agent k obtains an intermediate estimate $\boldsymbol{\phi}_{k,i}$ by a (stochastic) gradient step relative to $\widehat{\nabla_w J}_k(\boldsymbol{w}_{k,i-1})$ followed by a proximal step relative to the regularization term $R_k(\cdot)$. This corresponds to a (stochastic) proximal gradient update, an algorithm which is well studied in the centralized setting [49]. Following the proximal gradient update, agents then exchange their intermediate estimates $\boldsymbol{\phi}_{k,i}$ throughout their neighborhoods in (1.25) in the same manner as in the traditional diffusion algorithm.

In Chapter 2 we shall study the performance of the proximal diffusion strategy for the class of small regularizers. Small regularization weights are typically employed in an effort to reduce the noise present in the operation of the algorithm, without introducing significant bias relative to the unregularized solution. Such solutions encourage properties of the resulting estimate, without enforcing it. In particular, we will show in Chapter 2, that whenever the regularization strength is appropriately coupled with the step-size parameter, we have:

$$\limsup_{i\to\infty} \mathbb{E}\,\|w^o_{\text{unreg}} - \boldsymbol{w}_{k,i}\,\|^2 \leq O(\mu) + o(\mu) \tag{1.26}$$

where the term $O(\mu)$, similar to the centralized (1.10) and unregularized diffusion performances (1.21) arises from the stochastic gradient approximation, and is a higher-order term which corresponds to the bias introduced by regularizing the original problem.

For scenarios where the regularizers are general convex functions, we develop a more general strategy based on conjugate smoothing in Chapter 3, which involves a damped variation of the proximal diffusion strategy as a special case. In particular, we will replace each non-differentiable component, $R_k(w)$, by a differentiable approximation $R_k^{\delta}(w)$, parameterized by

$\delta > 0$. Subsequently, we can pursue

$$w_\delta^o = \arg\min_w \sum_{k=1}^{N} p_k \left\{ J_k(w) + R_k^\delta(w) \right\} \tag{1.27}$$

by means of the following, regularized diffusion strategy.

Algorithm 1.3 Regularized Diffusion Strategy [51]

$$\boldsymbol{\phi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu \widehat{\nabla_w J}_k(\boldsymbol{w}_{k,i-1}) \tag{1.28}$$

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{\phi}_{k,i} - \mu \nabla_w R_k^\delta(\boldsymbol{\phi}_{k,i}) \tag{1.29}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell=1}^{N} a_{\ell k} \boldsymbol{\psi}_{\ell,i} \tag{1.30}$$

We discover that performance guarantees under these general conditions require careful balancing of the step-size μ of the algorithm and a parameter δ used to construct the smooth approximation. We will describe a coupling relationship which ensures convergence for sufficiently small step-sizes, derive performance bounds, and show an application in group-Lasso regularized machine learning.

1.2.2 Non-Convex Learning

Driven by the need to solve increasingly complex optimization problems in signal processing and machine learning, there has been increasing interest in understanding the behavior of gradient-descent algorithms in non-convex environments. In contrast to (strongly) convex optimization problems, where a small gradient norm implies proximity to an optimal solution, non-convex loss surfaces contain many saddle-points, local minima and even maxima, where the gradient norm is small. Most available works on distributed non-convex optimization problems focus establishing convergence to first-order stationary points [52–58]. Recently, there has been growing interest in examining the ability of gradient descent implementations to escape from saddle points [59–61], since such points represent bottlenecks to the underlying learning problem [62].

In Chapters 5 and 6, we study the performance of the diffusion algorithm (1.19)–(1.20) for non-convex loss functions. We establish that the diffusion learning algorithm continues to yield meaningful estimates in these more challenging, non-convex environments, in the sense that (a) despite the distributed implementation, individual agents cluster in a small region around the network centroid in the mean-fourth sense, and (b) the network centroid inherits many properties of the centralized, stochastic gradient descent recursion, including escape from strict saddle points in $O(1/\mu)$ iterations and return of approximately second-order stationary points in a polynomial number of iterations.

1.3 Learning for Networked Feature Spaces

The first part of this dissertation focuses on the *design* of learning algorithms over networks, where the *network*, depicted in Fig. 1.1, acts as a *constraint* on exchanges of information between agents. In the second part of this dissertation, we take the alternative perspective of an *observer*, presented with data that has an internal, unknown, network structure. In data science applications, effective interpretation and processing of high-dimensional data is generally contingent on an understanding of the relationships that may exist between subsets of the data. This is particularly relevant for large-scale data sets. One useful way to capture interrelations among different parts of a data set is by means of a graph representation or model [63]. While data arising from some applications naturally lead to or suggest suitable graph representations for information flow, such as graphs representing networks or power grids, there are many instances where the underlying graph structure is not readily available and needs to be inferred from observations. Furthermore, even when the topology of the graph is known, the same may not hold for the weights on the edges of the graph, which describe the *strength* of the relationship. For example, in a social network, it may be less important to know whether two people are connected, than to know how much influence one person has on the other.

1.3.1 Online Graph Learning

In Chapter 8, we consider signals that evolve according to a heat diffusion process [64]. This process is related to a spatially sampled approximation of the second-order heat differential equation. The model is not limited to heat diffusion but can be applied to modeling other processes such as the evolution of interest over social networks [65] and the movement of people in cities [66]. We shall show that the problem of recovering the graph Laplacian, which parametrizes the heat diffusion process, from the time evolution of the observed signal, can be formulated as a strongly-convex and quadratic optimization problem. This in turn means that its minimizer can be sought efficiently by a variety of algorithms. We propose a (projected) stochastic gradient algorithm, which amounts to a Least-Mean-Squares (LMS)-type recursion and is adaptive in nature.

Algorithm 1.4 Laplacian LMS Strategy [67]

$$\overline{\boldsymbol{W}}_i = \overline{\boldsymbol{W}}_{i-1} + \mu \left(\overline{\boldsymbol{s}}_i - \overline{\boldsymbol{W}}_{i-1} \, \overline{\boldsymbol{s}}_{i-1} \right) \overline{\boldsymbol{s}}_{i-1}^{\mathsf{T}} \tag{1.31}$$

1.3.2 The BRAIN Strategy for Online Learning

In Chapter 9, rather than simply learn a graph from data, we leverage the learned graph to improve classification performance in a coupled and online fashion. Complexity is a double-edged sword for learning algorithms when the number of available samples for training in relation to the dimension of the feature space is small. This is because simple models do not sufficiently capture the nuances of the data set, while complex models overfit. While remedies such as regularization and dimensionality reduction exist, they can still suffer from overfitting or introduce bias. To address the issue of overfitting, the incorporation of prior structural knowledge is generally of paramount importance. In Chapter 9, we propose a BRAIN strategy for learning, which enhances the performance of traditional algorithms, such as logistic regression and SVM learners, by incorporating a graphical layer that tracks and learns in real-time the underlying correlation structure among feature subspaces. In this way, the algorithm is able to identify salient subspaces and their correlations, while

simultaneously dampening the effect of irrelevant features. This effect is particularly useful for high-dimensional feature spaces.

1.4 Organization

Chapters 2–6 focus on decentralized learning over networked agents, while Chapters 8 and 9 develop algorithms for centralized learning over networked feature spaces. Specifically, this dissertation is organized as follows:

- **Chapter 2**: We begin by introducing the proximal diffusion strategy for differentiable loss functions with non-differentiable regularizers. The performance of the strategy is quantified, and a coupling scheme for the regularization weight and the step-size is proposed, which leads to an asymptotically unbiased solution. The work in this chapter is based on material from reference [50].

- **Chapter 3**: In this chapter, we generalize the proximal diffusion strategy to allow for more general, and arbitrary convex regularization functions, by means of conjugate smoothing. We quantify the bias introduced by the smoothing procedure and establish the ability of the regularized diffusion strategy to approach the minimizer of the non-smooth cost with arbitrary accuracy. This chapter is based on the works [51, 68].

- **Chapter 4**: We show how the regularized diffusion strategy can be applied to matrix optimization and present an application in distributed image reconstruction.

- **Chapter 5 and 6**: We return to the study of smooth cost functions, but relax the convexity assumption commonly employed in the study of distributed algorithms. We establish that even in non-convex environments, iterates at individual continue to cluster around a network centroid, and proceed to study the dynamics of the representative centroid. We establish descent, even for (strict) saddle-points, and establish that the diffusion algorithm returns approximately second-order stationary points in a polynomial number of iterations. The material in these chapters is based on [69–71].

- **Chapter 7**: We focus on centralized learning problems and show how, relying primarily on mean-square arguments, second-order guarantees for stochastic gradient descent in non-convex environments can be obtained under conditions which are more general than typically assumed in the literature and applicable to a broader class of adaptation and learning problems. This chapter is based on material in [72].

- **Chapter 8**: This chapter considers data that arises from a heat diffusion process and presents the Laplacian LMS strategy for online graph learning. We study the performance of this strategy and derive mean-square error expressions. This chapter is based on the work [67].

- **Chapter 9**: When the objective of the learning problem is not to simply learn a graph describing relationships, but to leverage this information to improve performance in a classification task, the graph learning and classification problems can be coupled. Such procedure is proposed in Chapter 9, where a correlation layer is attached to traditional learning architectures such as logistic regression or SVM. We present an application in gene classification. The material in this chapter is based on [73].

- **Chapter 10**: The final chapter presents a summary of the contributions of this dissertation and a discussion of avenues for future research.

CHAPTER 2

Small Regularizers

In this chapter, we study the performance of the proximal diffusion strategy for small regularizers. The material is largely based on the work [50].

2.1 Motivation

Recall that our general problem of interest is:

$$w^o = \arg\min_{w} \sum_{k=1}^{N} p_k \left\{ J_k(w) + R_k(w) \right\} \tag{2.1}$$

This type of regularization can be motivated in one of two ways:

- The true objective is the minimizer

$$w^o_{\text{unreg}} \triangleq \arg\min_{w} \sum_{k=1}^{N} p_k J_k(w) \tag{2.2}$$

  However, there is prior information available about w^o_{unreg} (such as knowing that it is sparse, or that it is constrained to a certain region in space, or that it is close to some value). This knowledge is encoded through regularization $R_k(w)$, which is meant to mitigate the effect of noise on the algorithm. In these scenarios, the regularization is generally chosen small, so as to not bias the limiting point of the algorithm relative to w^o_{unreg}.

- The true objective is w^o. This is the case if properties encouraged by $R_k(w)$ are desired, albeit not necessarily present in w^o_{unreg}. Examples of such scenarios are constrained

optimization or sparsity-inducing regularizers meant to avoid overfitting in machine learning. These types of regularizers need not be small.

The more challenging case of arbitrary regularizers is treated in Chapter 3. In this chapter, we focus on *small* regularizers. To this end, let:

$$R_k(w) \triangleq \mu^\nu R_k^{\text{org}}(w) \tag{2.3}$$

where ν is a non-negative parameter and the regularization function $R_k^{\text{org}}(\cdot)$ does not need to be differentiable. Note that we allow for the regularization weight to depend on the step-size parameter of the algorithm. The motivation for this construction is the observation that the steady-state error of diffusion algorithms decreases linearly with μ [1], so that regularization becomes unnecessary as $\mu \to 0$ if the true objective is w^o_{unreg}.

2.2 Related Works

2.2.1 Differentiable Cost Functions

When the cost function at each agent k is differentiable, i.e., $R_k(w) = 0$ for all k, the minimizer of (2.1) can be sought through a variety of distributed strategies, such as consensus [26, 74–76] or diffusion [1, 22, 25]. For example, in the Adapt-then-Combine form of diffusion [1], each agent k runs the following recursion:

$$\boldsymbol{\phi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla_w J}_k(\boldsymbol{w}_{k,i-1}) \tag{2.4a}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell=1}^{N} a_{\ell k}\boldsymbol{\phi}_{\ell,i} \tag{2.4b}$$

In (2.4b), the symbol $\boldsymbol{w}_{k,i}$ denotes the iterate that is computed by agent k at iteration i, while $\boldsymbol{\psi}_{k,i}$ is an intermediate state resulting from the self-learning step (2.4a). It is shown in [1,22] that, under some reasonable technical conditions on the cost functions and gradient noise, the iterate $\boldsymbol{w}_{k,i}$ by each agent k converges in the mean-square sense to the unique

minimizer, w^o, of the following weighted aggregate cost:

$$w^o = \arg\min_{w} \sum_{k=1}^{N} p_k J_k(w) \tag{2.5}$$

within $O(\mu)$, namely,

$$\limsup_{i\to\infty} \mathbb{E}\|w^o - \boldsymbol{w}_{k,i}\|^2 \;=\; O(\mu) \tag{2.6}$$

so that all agents are able to approach the same global minimizer for a sufficiently small step-size.

2.2.2 Non-Differentiable Cost Functions

There are several useful works in the literature that study optimization problems with non-smooth regularizers. For example, the work [26] relies on the use of *sub-gradient* iterations but requires that the sub-gradients of the regularized risks, $J_k(w) + R_k(w)$, should be uniformly bounded. However, this condition is not satisfied in many important cases of interest, for example, even when $J_k(w)$ is simply quadratic in w (as happens in mean-square-error designs) or when the $R_k(w)$ are indicator functions used to encode constraints. Variations for specific choices of $J_k(\cdot)$ are examined in [41–44] where only the sub-gradients of $R_k(\cdot)$ are required to be bounded. For the case when the $R_k(w)$ are chosen as indicator functions in constrained problem formulations, a distributed diffusion strategy based on the use of suitable penalty functions is proposed and studied in [33].

Some other studies examine the performance of *inexact* proximal methods for particular sources of uncertainties in the gradient information. For example, in [77] regret bounds for stochastic proximal sub-gradient descent are derived under the assumption of Lipschitz continuous costs; the bounds there were limited to a single-agent implementation. The work in [45] considers inexact proximal gradient descent where the errors in the computation of the gradient and/or proximal operator are assumed to be deterministic and decay to zero. The work [46] builds on this analysis and develops a fast distributed implementation that enforces agreement among agents by embedding i communication steps between iterations

i and $i+1$ and letting $i \to \infty$. This construction can be reasonable in the deterministic context, where a given accuracy can be tolerated after finite time i, but is infeasible in the context of continuous adaptation and learning from streaming data since it will require the number of communication steps to grow unbounded. The authors of [30] remedy the need for increasing the number of communication steps between successive gradient updates by adding a correction term which ensures that the network converges to consensus for constant step-sizes and single communication exchanges as long as the cost functions are deterministic.

Distributed stochastic variations for mean-square error costs with bounded regularizer sub-gradients are proposed in [47, 48] for single-task problems and in [35] for multi-task environments.

Most of these prior works involve requirements that limit their application to important scenarios, whether in terms of requiring bounded sub-gradients, or focusing on quadratic costs. The purpose of this work is to propose a general distributed strategy and a line of analysis that is applicable to a wide class of stochastic costs and non-differentiable regularizers. For further review of the literature we refer the reader to Chapter 3.

2.3 Proximal Diffusion Strategy

To begin with, we recall that, in the purely deterministic context, the proximal operator relative to $R_k(\cdot)$ with step-size μ is defined by [49]:

$$\text{prox}_{\mu R_k}(x) \triangleq \arg\min_u \left(R_k(u) + \frac{1}{2\mu}\|x-u\|_2^2 \right) \tag{2.7}$$

Evaluating Eq. (2.7) at $x = w_{k,i-1} - \mu\nabla_w J_k(w_{k,i-1})$, which is the result of a gradient-descent step applied to $J_k(w)$, yields the proximal gradient descent iteration:

$$w_{k,i} = \text{prox}_{\mu R_k}\left\{w_{k,i-1} - \mu\nabla_w J_k(w_{k,i-1})\right\} \tag{2.8}$$

From the optimality condition for Eq. (2.7), namely that the sub-gradient set at the minimizer contains the zero-vector, it follows that [49, 78]:

$$w_{k,i} \in w_{k,i-1} - \mu \nabla_w J_k(w_{k,i-1}) - \mu \, \partial_w R_k \left(w_{k,i} \right) \tag{2.9}$$

where $\partial_w R_k \left(w_{k,i} \right)$ denotes the set of sub-gradients of $R_k(w)$ at $w_{k,i}$. The proximal operation (2.8) returns a particular sub-gradient vector, which we denote by $\widehat{\partial_w R_k}(w_{k,i})$. In this way, the resulting iterate can be written as

$$w_{k,i} = w_{k,i-1} - \mu \nabla_w J_k(w_{k,i-1}) - \mu \, \widehat{\partial_w R_k} \left(w_{k,i} \right) \tag{2.10}$$

Observe from (2.9) and (2.10) that $\nabla_w J_k(\cdot)$ is evaluated at $w_{k,i-1}$, whereas $\partial_w R_k(\cdot)$ is evaluated at $w_{k,i}$. This property sometimes motivates the alternative designation "forward-backward" operator for the proximal gradient step. Proximal gradient descent is of particular interest when (2.7) can be evaluated efficiently or even in closed form – see [79] for an overview of closed form solutions of (2.7) for particular $R_k(\cdot)$. In the case of the ℓ_1-norm, for example, the proximal operator reduces to soft-thresholding [80, 81].

Returning to (2.4a)–(2.4b), the above discussion motivates us to introduce the following proximal implementation of diffusion:

$$\boldsymbol{\phi}_{k,i} = \mathrm{prox}_{\mu R_k} \left\{ \boldsymbol{w}_{k,i-1} - \mu \widehat{\nabla_w J_k}(\boldsymbol{w}_{k,i-1}) \right\} \tag{2.11a}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell=1}^{N} a_{\ell k} \boldsymbol{\phi}_{\ell,i} \tag{2.11b}$$

where a proximal step has been added to (2.4a) as shown by (2.11a). This adjustment is meant to address the presence of the regularization term added in (2.3). Observe that (2.11a)–(2.11b) responds immediately to streaming data; it does not require repeated iterations between two successive time instants. We will further see that this implementation does also not require the gradient noise to be deterministic or to decay to zero.

The analysis in the subsequent sections will establish the following facts about the

stochastic implementation (2.11a)–(2.11b):

- In Section 2.5.1, it will be shown that, when the true gradient vectors are employed in (2.11a), then each agent in the diffusion strategy will converge to a unique fixed point, denoted by $w_{k,\infty}$.
- In Section 2.5.2, we will relate $w_{k,\infty}$ to the global minimizer $w^o_{\rm unreg}$ of (2.5) and show that $\|w^o_{\rm unreg} - w_{k,\infty}\|^2 \leq O(\mu^{2\nu}) + O(\mu^2)$.
- In Section 2.5.3, we will conclude that, for $\nu \geq 1/2$, recursion (2.11a)–(2.11b) *with* gradient noise converges to $w^o_{\rm unreg}$ within $O(\mu)$ in the mean-square-error sense.

The following two assumptions are needed in establishing the results — see [1] for explanations and motivation.

Assumption 2.1 (Lipschitz gradients)**.** *For any k, the gradient $\nabla_w J_k(\cdot)$ is Lipschitz*

$$0 < \lambda_{\min} I_N \leq H_k(w) \leq \lambda_{\max} I_N \tag{2.12}$$

Assumption 2.2 (Gradient Noise Process)**.** *For any k, the gradient noise process is defined as*

$$\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) = \widehat{\nabla_w J}_k(\boldsymbol{w}_{k,i-1}) - \nabla_w J_k(\boldsymbol{w}_{k,i-1}) \tag{2.13}$$

and satisfies

$$\mathbb{E}\,[\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})|\boldsymbol{\mathcal{F}}_{i-1}] = 0 \tag{2.14a}$$

$$\mathbb{E}\,\big[\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}\big] \leq \beta^2\|\boldsymbol{w}_{k,i-1}\|^2 + \sigma_s^2 \tag{2.14b}$$

for some non-negative constants $\{\beta^2, \sigma_s^2\}$, and where $\boldsymbol{\mathcal{F}}_{i-1}$ denotes the filtration generated by the random processes $\{\boldsymbol{w}_{\ell,j}\}$ for all $\ell = 1, 2, \ldots, N$ and $j \leq i-1$, i.e., $\boldsymbol{\mathcal{F}}_{i-1}$ represents the information that is available about the random processes $\{\boldsymbol{w}_{\ell,j}\}$ up to time $i-1$.

2.4 Operator Representation of Proximal Diffusion

We first show that the proximal diffusion strategy (2.11a)–(2.11b) can be represented as the concatenation of three operators, in a manner that extends the representation developed in [25] for the conventional diffusion iteration without proximal steps. We subsequently show that this mapping is contractive and invoke Banach's fixed-point theorem [82] to conclude that the proximal diffusion mapping has a unique fixed-point. We first introduce some notation and definitions. Thus, let

$$x = \operatorname{col}\{x_1, x_2, \ldots, x_N\} \in \mathbb{R}^{MN} \tag{2.15}$$

denote an $N \times 1$ block-column vector, where each x_k is $M \times 1$.

Definition 2.1. *(Combination Operator) The combination operator $T_A : \mathbb{R}^{MN} \to \mathbb{R}^{MN}$ is defined as the linear mapping:*

$$T_A(x) \triangleq (A^{\mathsf{T}} \otimes I_M)x = \operatorname{col}\left\{\sum_{\ell=1}^{N} a_{\ell k} x_\ell\right\} \tag{2.16}$$

where $A = [a_{\ell k}]$ is an $N \times N$ left-stochastic matrix and $\otimes$ denotes the Kronecker product operation. □

Definition 2.2. *(Block Gradient Descent Operator) The block gradient descent operator $T_G : \mathbb{R}^{MN} \to \mathbb{R}^{MN}$ is defined as the non-linear mapping:*

$$T_G(x) \triangleq \begin{bmatrix} x_1 - \mu\nabla_w J_1(x_1) \\ \vdots \\ x_N - \mu\nabla_w J_N(x_N) \end{bmatrix} \tag{2.17}$$

□

Definition 2.3. *(Stochastic Block Gradient Descent Operator) The stochastic block gradient descent operator* $\widehat{\boldsymbol{T}}_G : \mathbb{R}^{MN} \to \mathbb{R}^{MN}$ *is defined as the non-linear mapping:*

$$\widehat{\boldsymbol{T}}_G(\boldsymbol{x}) \triangleq \begin{bmatrix} \boldsymbol{x}_1 - \mu\widehat{\nabla_w J}_1(\boldsymbol{x}_1) \\ \vdots \\ \boldsymbol{x}_N - \mu\widehat{\nabla_w J}_N(\boldsymbol{x}_N) \end{bmatrix} = T_G(\boldsymbol{x}) + \mu\boldsymbol{s}(\boldsymbol{x}) \tag{2.18}$$

where

$$\boldsymbol{s}(\boldsymbol{x}) \triangleq \operatorname{col}\{\boldsymbol{s}_1(\boldsymbol{x}_1), \ldots, \boldsymbol{s}_N(\boldsymbol{x}_N)\} \tag{2.19}$$

is the (block) gradient noise vector. □

Definition 2.4. *(Block Proximal Operator) The block proximal operator* $T_P : \mathbb{R}^{MN} \to \mathbb{R}^{MN}$ *is defined as the non-linear mapping:*

$$T_P(x) \triangleq \begin{bmatrix} \operatorname{prox}_{\mu R_1}(x_1) \\ \vdots \\ \operatorname{prox}_{\mu R_N}(x_N) \end{bmatrix} \tag{2.20}$$

□

Using these operators, we can then rewrite the proximal diffusion algorithm (2.11a)–(2.11b) more compactly as the following concatenation of operators in terms of the network vector $\boldsymbol{w}_i = \operatorname{col}\{\boldsymbol{w}_{1,i}, \ldots, \boldsymbol{w}_{N,i}\}$:

$$\boldsymbol{w_i} = \widehat{\boldsymbol{T}}_{\mathrm{pd}}(\boldsymbol{w_{i-1}}) \triangleq T_A \circ T_P \circ \widehat{\boldsymbol{T}}_G(\boldsymbol{w_{i-1}}) \tag{2.21}$$

Without gradient noise, this relation reduces to:

$$w_i = T_{\mathrm{pd}}(w_{i-1}) \triangleq T_A \circ T_P \circ T_G(w_{i-1}) \tag{2.22}$$

Fig. 2.4 displays the stochastic proximal diffusion implementation as a cascade of operators. The following operator properties were derived in [25] for diffusion without proximal steps:

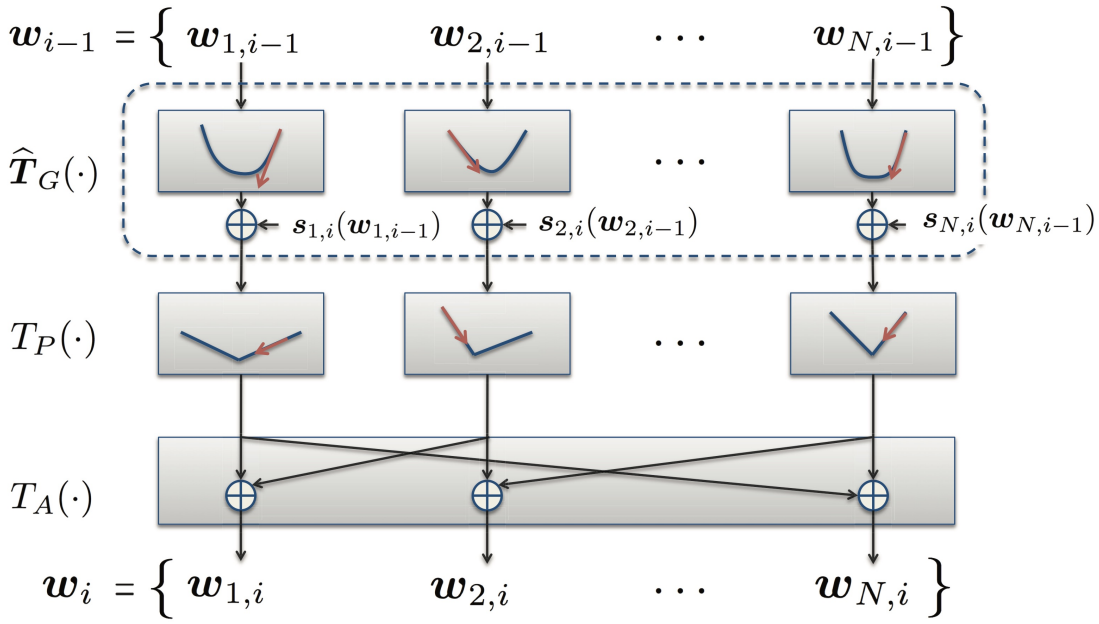

Figure 2.1: Proximal diffusion as a cascade of operators.

1. **(Linearity)**: $T_A(\cdot)$ is a linear operator.

2. **(Non-negativity)**: $P[x] \succeq 0$.

3. **(Scaling)**: For any $a \in \mathbb{R}$, $P[ax] = a^2 P[x]$.

4. **(Additivity)**: Suppose $\boldsymbol{x} = \text{col}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ and $\boldsymbol{y} = \text{col}\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ are random $N \times 1$ block vectors and furthermore $\mathbb{E}\boldsymbol{x}_k^{\mathsf{T}}\boldsymbol{y}_k = 0$. Then

$$\mathbb{E}P[\boldsymbol{x} + \boldsymbol{y}] = \mathbb{E}P[\boldsymbol{x}] + \mathbb{E}P[\boldsymbol{x}]. \tag{2.23}$$

5. **(Variance relations)**:

$$P[T_A(x)] \preceq A^{\mathsf{T}} P[x] \tag{2.24}$$

$$P[T_G(x) - T_G(y)] \preceq \gamma^2 P[x - y] \tag{2.25}$$

where

$$\gamma^2 \triangleq 1 - 2\mu\lambda_{\min} + \mu^2\lambda_{\max}^2 \tag{2.26}$$

6. **(Block Maximum Norm)**: The $\infty-$norm of $P[x]$ is the squared block maximum norm of x:

$$\|P[x]\|_\infty = \|x\|_{b,\infty}^2 \triangleq \max_{1\le k\le N} \|x_k\|^2 \tag{2.27}$$

7. **(Preservation of Inequality)**: Suppose vectors x, y and matrix F have non-negative entries, then $x \preceq y$ implies $Fx \preceq Fy$.

In order to incorporate the proximal operator into the analysis, we need an additional property:

Lemma 2.1 (Variance of Proximal Operator)**.** *Suppose each $R_k(\cdot)$ is a closed, convex function (i.e., its epigraph is a closed, convex set), then*

$$P\left[T_P(x) - T_P(y)\right] \preceq P[x-y]. \tag{2.28}$$

Proof. See Appendix 2.A. □

2.5 Main Results

2.5.1 Fixed-Point of Deterministic Recursion

Lemma 2.2 (Contractive Mapping)**.** *The deterministic proximal diffusion operator $T_{\rm pd}(\cdot)$ defined in (2.22) satisfies*

$$\|T_{\rm pd}(x) - T_{\rm pd}(y)\|_{b,\infty} \le \gamma \cdot \|x-y\|_{b,\infty} \tag{2.29}$$

with $\gamma^2 \triangleq 1 - 2\mu\lambda_{\min} + \mu^2\lambda_{\max}^2$, and where $\|\cdot\|_{b,\infty}$ denotes the block maximum norm [1]. The condition on μ to guarantee $\gamma^2 < 1$ is:

$$0 < \mu < \frac{2\lambda_{\min}}{\lambda_{\max}^2} \tag{2.30}$$

It then follows from Banach's fixed point theorem [82, 83] that $w_i = T_{\rm pd}(w_{i-1})$ converges to a unique fixed-point, w_∞, geometrically.

Proof. See Appendix 2.B. □

2.5.2 Bias Analysis

Now we analyze how far this fixed point w_∞ is from the desired global solution, $w^o_{\rm unreg}$. In steady-state, the deterministic fixed-point equation (2.22) can be unfolded as follows:

$$\phi_{k,\infty} = \mathrm{prox}_{\mu R_k}\left\{w_{k,\infty} - \mu\nabla_w J_k(w_{k,\infty})\right\} \tag{2.31a}$$

$$w_{k,\infty} = \sum_{\ell=1}^{N} a_{\ell k}\phi_{\ell,\infty} \tag{2.31b}$$

To proceed, we introduce an assumption of bounded sub-gradients, which is common in the sub-gradient [26,77] and distributed proximal gradient [46] literature, namely, that for every agent k, the set of sub-differentials $\partial_w R_k^{\rm org}(w)$ is uniformly bounded, i.e. for all w:

$$\|\partial_w R_k^{\rm org}(w)\| \leq \eta_k^{\rm org} \tag{2.32}$$

for some non-negative constant $\eta_k^{\rm org}$. For convex functions, the statement is equivalent to requiring $R_k^{\rm org}(w)$ to be Lipschitz continuous with constant $\eta_k^{\rm org}$. For the scaled costs $R_k(w) \triangleq \mu^\nu R_k^{\rm org}(w)$, condition (2.32) translates to:

$$\|\partial_w R_k(w)\| \leq \mu^\nu \eta_k^{\rm org} \triangleq \eta_k = O(\mu^\nu) \tag{2.33}$$

Now we subtract Eqs. (2.31a) and (2.31b) from $w^o_{\rm unreg}$ and define the error variables $\widetilde{w}_{k,\infty} = w^o_{\rm unreg} - w_{k,\infty}$. This leads to the error recursion:

$$\widetilde{\phi}_{k,\infty} = \widetilde{w}_{k,\infty} + \mu\nabla_w J_k(w_{k,\infty}) + \mu\widehat{\partial_w R}_k(\phi_{k,\infty}) \tag{2.34a}$$

$$\widetilde{w}_{k,\infty} = \sum_{\ell=1}^{N} a_{\ell k}\widetilde{\phi}_{\ell,\infty} \tag{2.34b}$$

Using the mean-value theorem [1, 84], we can write:

$$\nabla_w J_k(w_{k,\infty}) = \nabla_w J_k(w^o_{\text{unreg}}) - H_{k,\infty}\widetilde{w}_{k,\infty} \tag{2.35}$$

where $H_{k,\infty}$ denotes the Hessian of $J_k(w)$ at $w_{k,\infty}$. We get

$$\widetilde{\phi}_{k,\infty} = (I_M - \mu H_{k,\infty})\,\widetilde{w}_{k,\infty} + \mu\nabla_w J_k(w^o_{\text{unreg}}) + \mu\widehat{\partial_w R}_k(\phi_{k,\infty}) \tag{31a}$$

$$\widetilde{w}_{k,\infty} = \sum_{\ell=1}^{N} a_{\ell k}\widetilde{\phi}_{\ell,\infty} \tag{31b}$$

We next introduce the following extended vectors and matrices:

$$\widetilde{w}_\infty \triangleq \text{col}\,\{\widetilde{w}_{1,\infty}, \ldots, \widetilde{w}_{N,\infty}\} \tag{2.37}$$

$$\mathcal{A} \triangleq A \otimes I_M \tag{2.38}$$

$$\mathcal{H}_\infty \triangleq \text{diag}\,\{H_{1,\infty}, \ldots, H_{N,\infty}\} \tag{2.39}$$

$$\mathcal{B}_\infty \triangleq \mathcal{A}^{\mathsf{T}}(I_{MN} - \mu\mathcal{H}_\infty) \tag{2.40}$$

$$g^o \triangleq \text{col}\left\{\nabla_w J_1(w^o_{\text{unreg}}), \ldots, \nabla_w J_N(w^o_{\text{unreg}})\right\} \tag{2.41}$$

$$r_\infty \triangleq \text{col}\left\{\widehat{\partial_w R}_1(\phi_{1,\infty}), \ldots, \widehat{\partial_w R}_N(\phi_{N,\infty})\right\} \tag{2.42}$$

With these quantities, relations (31a)–(31b) lead to:

$$\widetilde{w}_\infty = \mathcal{B}_\infty\widetilde{w}_\infty + \mu\mathcal{A}^{\mathsf{T}}\,(g^o + r_\infty)\,. \tag{2.43}$$

Because A is a left-stochastic and primitive matrix, it admits a Jordan decomposition of the form $A = V_\epsilon J V_\epsilon^{-1}$

$$V_\epsilon = \left[\begin{array}{c|c} p & V_R \end{array}\right], \quad J = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & J_\epsilon \end{array}\right], \quad V_\epsilon^{-1} = \left[\begin{array}{c} \mathbb{1}^{\mathsf{T}} \\ \hline V_L^{\mathsf{T}} \end{array}\right] \tag{2.44}$$

where all diagonal entries of J_ϵ are inside the unit circle and J_ϵ consists of Jordan blocks with the value ϵ on the first lower diagonal instead of ones [1, 23]. Pre-multiplying both sides

of (2.43) by $\mathcal{V}_\epsilon^{\mathsf{T}} = V_\epsilon^{\mathsf{T}} \otimes I_M$ gives:

$$\overline{w}_\infty = \overline{\mathcal{B}}_\infty \overline{w}_\infty + \mu \mathcal{V}_\epsilon^{\mathsf{T}} \mathcal{A}^{\mathsf{T}} \left(g^o + r_\infty\right) \tag{2.45}$$

where $\overline{w}_\infty = \mathcal{V}_\epsilon^{\mathsf{T}} \widetilde{w}_\infty$ and $\overline{\mathcal{B}}_\infty = \mathcal{V}_\epsilon^{\mathsf{T}} \mathcal{B}_\infty (\mathcal{V}_\epsilon^{-1})^{\mathsf{T}}$. It follows that

$$\overline{w}_\infty = \mu \left(I_{MN} - \overline{\mathcal{B}}_\infty\right)^{-1} \mathcal{V}_\epsilon^{\mathsf{T}} \mathcal{A}^{\mathsf{T}} \left(g^o + r_\infty\right). \tag{2.46}$$

It was shown in [1, p. 541, Lemma 9.4] that, for sufficiently small step-sizes, it holds that

$$\left(I_{MN} - \overline{\mathcal{B}}_\infty\right)^{-1} = \left[\begin{array}{c|c} O(1/\mu) & O(1) \\ \hline O(1) & O(1) \end{array}\right] \tag{2.47}$$

where the leading $(1,1)$ block has dimensions $M \times M$. It can further be verified from the decomposition of V_ϵ in (2.44), that

$$\mathcal{V}_\epsilon^{\mathsf{T}} \mathcal{A}^{\mathsf{T}} \left(g^o + r_\infty\right) = \left[\begin{array}{c} \sum_{\ell=1}^N p_\ell \widehat{\partial_w R}_\ell(\phi_{\ell,\infty}) \\ \hline O(1) + \mathcal{V}_R^{\mathsf{T}} \mathcal{A}^{\mathsf{T}} r_\infty \end{array}\right] \tag{2.48}$$

Theorem 2.1. *Under assumption (2.32) and for small μ, the steady-state bias of the deterministic proximal diffusion recursion is bounded as:*

$$\|w^o_{\text{unreg}} - w_{k,\infty}\|^2 \leq O\left(\mu^{2\nu}\right) + O(\mu^2) \tag{2.49}$$

Proof. The result follows from (2.33) and (2.47)–(2.48). □

2.5.3 Evolution of Stochastic Recursion

We now examine how close the stochastic recursion $\boldsymbol{w}_i = \widehat{\boldsymbol{T}}_{\text{pd}}(\boldsymbol{w}_{i-1})$ approaches w^o_{unreg}. For this purpose, we introduce the mean-square perturbation vector at time i relative to w_∞:

$$\text{MSP}_i \triangleq \text{col}\left\{\mathbb{E}\|\boldsymbol{w}_{k,i} - w_{k,\infty}\|^2\right\} \in \mathbb{R}^N \tag{2.50}$$

Lemma 2.3. *The* MSP *at time i can be recursively bounded as:*

$$\mathrm{MSP}_i \preceq \left(\gamma^2 + 2\mu^2\beta^2\right) A^{\mathsf{T}}\mathrm{MSP}_{i-1} + \mu^2 d \tag{2.51}$$

where $d = O(1)$. A sufficient condition on μ for stability of (2.51) is:

$$0 < \mu < \frac{2\lambda_{\min}}{\lambda_{\max}^2 + 2\beta^2} \tag{2.52}$$

It follows that

$$\limsup_{i\to\infty} \|\mathrm{MSP}_i\|_\infty = O(\mu). \tag{2.53}$$

Proof. See Appendix 2.C. □

The following theorem ties all results together.

Theorem 2.2. *For sufficiently small step-sizes and $\nu \geq 1/2$, the steady-state MSD* of the proximal diffusion algorithm (2.11a)–(2.11b) is

$$\limsup_{i\to\infty} \mathbb{E}\|w_{\mathrm{unreg}}^o - \boldsymbol{w}_{k,i}\|^2 = O(\mu) \tag{2.54}$$

Proof. The result follows from (2.49) and (2.53). □

2.6 Numerical Results

Consider a network of $N = 10$ agents and $M = 20$. The network topology is shown in Fig. 2.2. Observations $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}\}$ for each agent k are generated according to the linear regression model $\boldsymbol{d}_k = \boldsymbol{u}_k w_{\mathrm{unreg}}^o + \boldsymbol{v}_k$, where $\boldsymbol{u}_{k,i}$ and $\boldsymbol{v}_k(i)$ are zero-mean Gaussian random variables with power shown in Fig. 2.3. The true w_{unreg}^o is sparse with only one non-zero element. For the special case with $J_k(w) = \mathbb{E}\|\boldsymbol{d}_k - \boldsymbol{u}_k w\|^2$ and $R_k^{\mathrm{org}}(w) = \|w\|_1$, we compare the performance of the regularized proximal diffusion implementation (2.11a)–(2.11b) and the unregularized diffusion implementation (2.4a)–(2.4b). Fig. 2.6 displays the steady-state MSD for different choices of the step-size parameter.

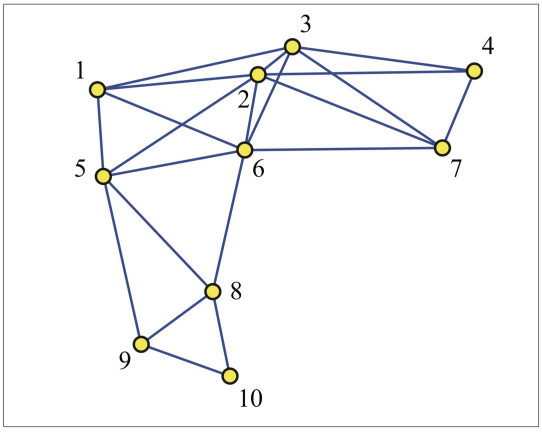

Figure 2.2: Network topology.

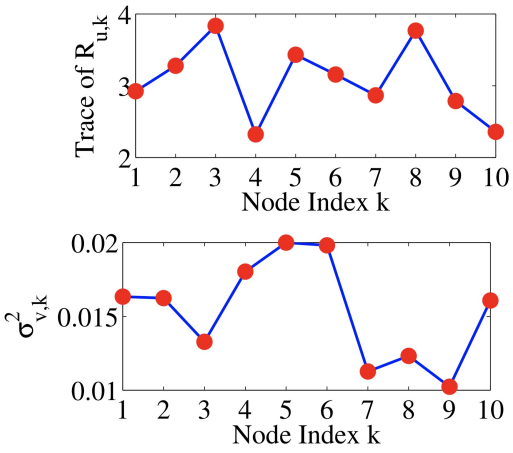

Figure 2.3: Data statistics.

2.A Proof of Lemma 2.1

We first note that for some generic regularization term that closed and convex, the solution of the proximal operator exists, is unique, and satisfies the following non-expansiveness property [85]:

$$\|\text{prox}_{\mu R}(x) - \text{prox}_{\mu R}(y)\| \leq \|x - y\|. \tag{2.55}$$

Now, from the definitions of $T_P(\cdot)$ and $P[\cdot]$ in:

$$\begin{aligned} P[T_P(x) - T_P(y)] &= \begin{bmatrix} \|\text{prox}_{\mu R_1}(x_1) - \text{prox}_{\mu R_1}(y_1)\|^2 \\ \vdots \\ \|\text{prox}_{\mu R_N}(x_N) - \text{prox}_{\mu R_N}(y_N)\|^2 \end{bmatrix} \\ &\preceq \begin{bmatrix} \|x_1 - y_1\|^2 \\ \vdots \\ \|x_N - y_N\|^2 \end{bmatrix} \\ &\preceq T_P[x - y] \end{aligned} \tag{2.56}$$

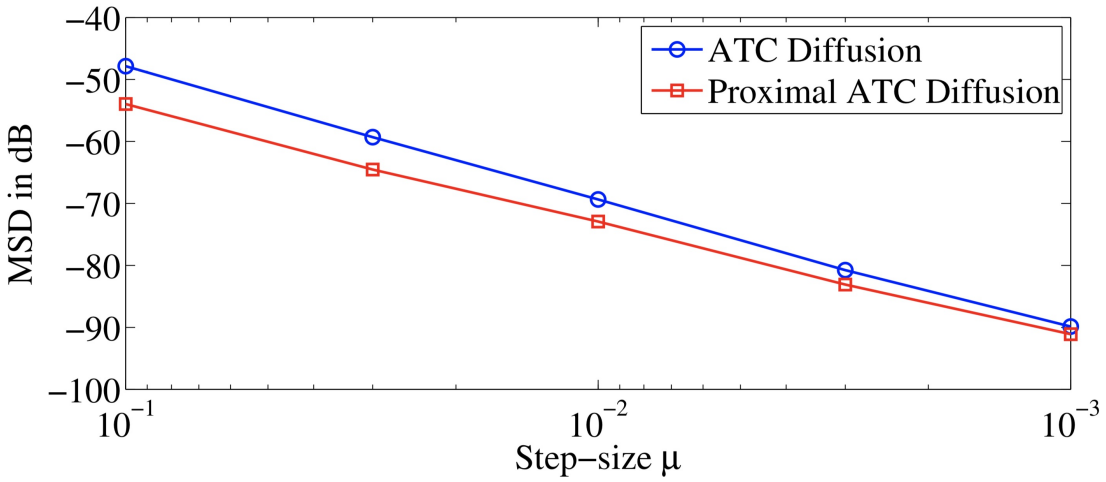

Figure 2.4: Performance comparison for $\nu = 1$.

2.B Proof of Lemma 2.2

Apply the operator properties from the previous section

$$\begin{aligned} P[T_{\text{pd}}(x) - T_{\text{pd}}(y)] =& P\left[T_A \circ T_P \circ T_G(x) - T_A \circ T_P \circ T_G(y)\right] \\ \overset{(a)}{\leq}& A^{\mathsf{T}} P\left[T_P \circ T_G(x) - T_P \circ T_G(y)\right] \\ \overset{(b)}{\leq}& A^{\mathsf{T}} P\left[T_G(x) - T_G(y)\right] \\ \overset{(c)}{\leq}& A^{\mathsf{T}} \gamma^2 P\left[x - y\right] \end{aligned} \tag{2.57}$$

where (a) and (c) are due to the variance relations of operators and (b) is due to Lemma 2.1. Now,

$$\begin{aligned} \|P[T_{\text{pd}}(x) - T_{\text{pd}}(y)]\|_\infty &\leq \|A^{\mathsf{T}} \gamma^2 P\left[x - y\right]\|_\infty \\ &\leq \|A^{\mathsf{T}} \gamma^2\|_\infty \cdot \|P\left[x - y\right]\|_\infty \\ &= \gamma^2 \cdot \|A^{\mathsf{T}}\|_\infty \cdot \|P\left[x - y\right]\|_\infty \\ &= \gamma^2 \cdot \|P\left[x - y\right]\|_\infty \end{aligned} \tag{2.58}$$

Inequality (2.29) follows after applying the block maximum norm property (2.27). The condition on μ follows from the expression for γ^2 (2.26).

2.C Proof of Lemma 2.3

The entries of this perturbation vector satisfy the following inequality recursion:

$$\begin{aligned}
\mathrm{MSP}_i &\triangleq \mathbb{E}P[\boldsymbol{w}_i - w_\infty] \\
&= \mathbb{E}P\left[T_A \circ T_P \circ \widehat{\boldsymbol{T}}_G(\boldsymbol{w}_{i-1}) - T_A \circ T_P \circ T_G(w_\infty)\right] \\
&= \mathbb{E}P\left[T_A\left(T_P \circ \widehat{\boldsymbol{T}}_G(\boldsymbol{w}_{i-1}) - T_P \circ T_G(w_\infty)\right)\right] \\
&\overset{(a)}{\preceq} A^{\mathsf{T}}\mathbb{E}P\left[T_P \circ \widehat{\boldsymbol{T}}_G(\boldsymbol{w}_{i-1}) - T_P \circ T_G(w_\infty)\right] \\
&\overset{(b)}{\preceq} A^{\mathsf{T}}\mathbb{E}P\left[\widehat{\boldsymbol{T}}_G(\boldsymbol{w}_{i-1}) - T_G(w_\infty)\right] \\
&\overset{(c)}{=} A^{\mathsf{T}}\mathbb{E}P\left[T_G(\boldsymbol{w}_{i-1}) + \mu\boldsymbol{s}_i\left(\boldsymbol{w}_{i-1}\right) - T_G(w_\infty)\right] \\
&\overset{(d)}{=} A^{\mathsf{T}}\mathbb{E}P\left[T_G(\boldsymbol{w}_{i-1}) - T_G(w_\infty)\right] + \mu^2 A^{\mathsf{T}}\mathbb{E}P\left[\boldsymbol{s}_i\left(\boldsymbol{w}_{i-1}\right)\right] \\
&\overset{(e)}{\preceq} \gamma^2 A^{\mathsf{T}}\mathbb{E}P\left[\boldsymbol{w}_{i-1} - w_\infty\right] + \mu^2 A^{\mathsf{T}}\mathbb{E}P\left[\boldsymbol{s}_i\left(\boldsymbol{w}_{i-1}\right)\right] \\
&= \gamma^2 A^{\mathsf{T}} \cdot \mathrm{MSP}_{i-1} + \mu^2 A^{\mathsf{T}}\mathbb{E}P\left[\boldsymbol{s}_i\left(\boldsymbol{w}_{i-1}\right)\right] \qquad (2.59)
\end{aligned}$$

where (a) and (e) are due to the variance properties, (b) is due to Lemma 2.1, (c) is due to the definition of $\widehat{\boldsymbol{T}}_G(\cdot)$, and (d) is due the additivity property. Computing the $\infty-$norm of both sides of (2.51) on both sides yields:

$$\begin{aligned}
\|\mathrm{MSP}_i\|_\infty &\leq \|\left(\gamma^2 + 2\mu^2\beta^2\right) A^{\mathsf{T}}\mathrm{MSP}_{i-1} + \mu^2 d\|_\infty \\
&\leq \left(\gamma^2 + 2\mu^2\beta^2\right)\|\mathrm{MSP}_{i-1}\|_\infty + \mu^2\|d\|_\infty
\end{aligned}$$

so that

$$\limsup_{i\to\infty}\ \|\mathrm{MSP}_i\|_\infty \leq \frac{\mu\|d\|_\infty}{2\lambda_{\min} - \mu\left(\lambda_{\max}^2 + \beta_{\max}^2\right)} = O(\mu) \qquad (2.60)$$

CHAPTER 3

General Regularizers

We now consider general convex regularizers $R_k(w)$, which are no longer required to be small or have bounded sub-gradients. We will further relax assumptions on the differentiable parts of the cost function $J_k(w)$ and consider a broader class of updates with respect to $R_k(w)$ than the proximal step. The material in this chapter is largely based on the works [51, 68].

The purpose of this chapter is to develop and study a distributed strategy for Pareto optimization of an aggregate cost consisting of regularized risks. Each risk is modeled as the expectation of some loss function with unknown probability distribution while the regularizers are assumed deterministic, but are not required to be differentiable or even continuous. The individual, regularized, cost functions are distributed across a strongly-connected network of agents and the Pareto optimal solution is sought by appealing to a multi-agent diffusion strategy. To this end, the regularizers are smoothed by means of infimal convolution and it is shown that the Pareto solution of the approximate, smooth problem can be made arbitrarily close to the solution of the original, non-smooth problem. Performance bounds are established under conditions that are weaker than assumed before in the literature, and hence applicable to a broader class of adaptation and learning problems.

3.1 Introduction

The objective of distributed learning is the solution of global, stochastic optimization problems across networks of agents through localized interactions and without information about the statistical properties of the data. Using streaming data, the resulting strategies are adaptive in nature and able to track drifts in the location of the minimizers due to variations in

the statistical properties of the data. Regularization is one useful technique to encourage or enforce structural properties on the sought after minimizer, such as sparsity or constraints. A substantial number of regularizers are inherently non-smooth, while many cost functions are differentiable. These article proposes a *fully-decentralized* and *adaptive* strategy that is able to minimize an aggregate sum of regularized costs. To do so, we fully exploit the structure of the individual objectives as sums of differentiable costs and non-differentiable regularizers.

Notation: Throughout the manuscript, random quantities are denoted in boldface. Matrices are denoted in capital letters while vectors and scalars are denoted in small-case letters. The symbol $\leq$ denotes a regular inequality, while $\preceq$ denotes an *element-wise* inequality.

3.1.1 Problem Formulation

We consider a strongly-connected network consisting of N agents. For any two agents k and ℓ, we attach a pair of non-negative coefficients $\{a_{\ell k}, a_{k\ell}\}$ to the edge linking them. The scalar $a_{\ell k}$ is used to scale data moving from agent ℓ to k; likewise, for $a_{k\ell}$. Strong-connectivity means that it is always possible to find a path, in either direction, with nonzero scaling weights linking any two agents (either directly if they are neighbors or indirectly through other agents). In addition, at least one agent k in the network possesses a self-loop with $a_{kk} > 0$. This condition ensures that at least one agent in the network has some confidence in its local information. Let $\mathcal{N}_k$ denote the set of neighbors of agent k. The coefficients $\{a_{\ell k}\}$ are convex combination weights that satisfy

$$a_{\ell k} \geq 0, \quad \sum_{\ell \in \mathcal{N}_k} a_{\ell k} = 1, \quad a_{\ell k} = 0 \text{ if } \ell \notin \mathcal{N}_k \tag{3.1}$$

If we introduce the combination matrix $A = [a_{\ell k}]$, it then follows from (3.1) and the strong-connectivity property that A is a left-stochastic primitive matrix. In view of the Perron-Frobenius Theorem [1,23,24], this ensures that A has a single eigenvalue at one while all other eigenvalues are inside the unit circle, so its spectral radius is given by $\rho(A) = 1$. Moreover, if we let p denote the right-eigenvector of A that is associated with the eigenvalue at one,

and if we normalize the entries of p to add up to one, then it also holds that all entries of p are strictly positive, i.e.,

$$Ap = p, \quad \mathbb{1}^{\mathsf{T}} p = 1, \quad p_k > 0 \tag{3.2}$$

where the $\{p_k\}$ denote the individual entries of the Perron vector, p [1].

We associate with each agent k a risk function $J_k(w) : \mathbb{R}^M \to \mathbb{R}$, assumed differentiable. In most adaptation and learning problems, risk functions are expressed as the expectation of loss functions. Hence, we assume that each risk function is of the form $J_k(w) = \mathbb{E}\, Q(w; \boldsymbol{x})$, where $Q(\cdot)$ is the loss function and $\boldsymbol{x}$ denotes random data. The expectation is computed over the distribution of this data (note that, in our notation, we use boldface letters for random quantities and normal letters for deterministic quantities or data realizations). We also associate with agent k a regularization term, $R_k(w) : \mathbb{R}^M \to \mathbb{R}$, which is a known deterministic function although possibly non-differentiable. Regularization factors of this form can, for example, help induce sparsity properties (such as using ℓ_1 or elastic-net regularizers) [86–88].

The objective we are interested in is to devise a fully distributed strategy to seek the minimizer of the following weighted aggregate cost, denoted by w^o:

$$w^o = \arg\min_{w \in \mathbb{R}^M} \sum_{k=1}^{N} p_k \left\{ J_k(w) + R_k(w) \right\} \tag{3.3}$$

The weights $\{p_k\}$ indicate that the resulting minimizer w^o can be interpreted as a Pareto solution for the collection of regularized risks $\{J_k(w) + R_k(w)\}$ [1,25] and will depend on the entries of the Perron eigenvector in a manner specified further below. We are particularly interested in determining this Pareto solution in the *stochastic* setting when the distribution of the data $\boldsymbol{x}$ is unknown. This means that the risks $J_k(w)$, or their gradient vectors, are also unknown. As such, *approximate* gradient vectors will need to be employed. A common construction in stochastic approximation theory is to employ the following choice at each iteration i [1,8]:

$$\widehat{\nabla J_k}(w) \;=\; \nabla Q_k(w; \boldsymbol{x}_i) \tag{3.4}$$

where $\boldsymbol{x}_i$ represents the data that is available (observed) at time i. The difference between

the true gradient vector and its approximation is called gradient noise. This noise will seep into the operation of the distributed algorithm and one main challenge is to show that, despite its presence, the proposed solution is able to approach w^o asymptotically. A second challenge we face in constructing an effective distributed solution is the non-smoothness (non-differentiability) of the regularizers. Motivated by a technique proposed in [89] in the context of *single* agent optimization, we will address this difficulty in the multi-agent case by introducing a smoothed version of the regularizers and then showing that the solution w^o can still be recovered under this substitution as the size of the smoothing parameter is reduced. We adopt a general formulation that will be shown to include proximal iterations as a special case.

3.1.2 Related Works in the Literature

The literature on distributed optimization is extensive. Some early strategies include incremental [90], consensus or decentralized gradient descent [26, 29, 91, 92], and the diffusion algorithm [1,22,25,27,93]. When exact gradients are employed, these strategies converge to a small area around the minimizer of the aggregate cost at a linear rate [25, 29]. Exact convergence requires diminishing step-sizes, resulting in sublinear rates of convergence. A number of more recent works focusing primarily on deterministic optimization, have proposed variations yielding linear rates of convergence pursued either by employing corrections in the primal domain [28, 30, 94–102] or primal-dual strategies [103–112] where [28, 101, 104, 111] allow for stochastic gradient approximations and [97, 107] consider empirical risk minimization problems. In many applications, the choice $p_k = \frac{1}{N}$ for all k is desirable, corresponding to an equally weighted Pareto solution. A number of works develop algorithms for $p_k = \frac{1}{N}$ over directed graphs [113–115].

One common method for handling non-differentiable cost functions is the utilization of sub-gradient recursions, where the ordinary gradient is replaced by sub-gradients [26, 91, 92, 104, 105, 111]. Most often, these works assume the sub-gradients are *bounded*. This condition is not satisfied in many important cases of interest, for example, even when $J_k(w)$

is simply quadratic in w (as happens in mean-square-error designs) or when the $R_k(w)$ are indicator functions used to encode constraints. Variations for specific choices of costs functions are examined in [41–44] where only the subgradients of $R_k(\cdot)$ are required to be bounded. The work [34] generalized these conditions to allow for (sub-)gradients that are "affine-Lipschitz", which holds for many, but not all costs and regularizers of interest, such as indicator functions. For the case when the $R_k(w)$ are chosen as indicator functions in constrained problem formulations, as an alternative to projection based schemes [91, 92, 104, 105], a distributed diffusion strategy based on the use of suitable penalty functions was proposed and studied in [33].

Some other studies pursue distributed solutions by relying instead on the use of proximal iterations (as opposed to sub-gradient iterations); an accessible survey on the proximal operator and its properties appears in [49]. For example, for purely deterministic costs, distributed proximal strategies are developed in [30, 46, 95, 96, 98]. Stochastic variations for mean-square error costs with bounded regularizer subgradients are proposed in [47, 48] for single-task problems and in [35] for multi-task environments. A strategy for general stochastic costs with *small*, Lipschitz continuous regularizers is studied in [50].

3.1.3 Contributions

The purpose of this chapter is to propose a general distributed strategy and a line of analysis that is applicable to a wide class of stochastic costs and non-differentiable regularizers. The first step in the solution will involve replacing each non-differentiable component, $R_k(w)$, by a differentiable approximation $R_k^\delta(w)$, parameterized by $\delta > 0$, such that

$$\|w^o - w_\delta^o\|^2 \leq O(\delta) \tag{3.5}$$

The accuracy of the approximation is controlled through the smoothing parameter δ. Subsequently, we will solve for the minimizer:

$$w_\delta^o = \arg\min_w \sum_{k=1}^{N} p_k \left\{ J_k(w) + R_k^\delta(w) \right\} \tag{3.6}$$

Smoothing non-differentiable costs via infimal convolution [89,116,117] is a popular technique in the deterministic optimization literature, and it can be used to motivate some known algorithms, such as the proximal point algorithm [49]. The technique has been mainly developed for deterministic optimization by *single* stand-alone agents. In this chapter, we pursue an extension in two non-trivial directions. First, we consider networked agents (rather than a single agent) working together to solve the aggregate optimization problem (3.3) (or (3.6)) and, second, the risk functions involved are a combination of stochastic costs defined as the expectations of certain loss functions and deterministic regularizers. Moreover, the probability distribution of the data is assumed unknown and, therefore, the aggregate risks themselves are not known but can only be approximated. The challenge is to devise a distributed strategy that is able to converge to the desired Pareto solution despite these difficulties.

We note that an alternative smoothing procedure by means of adding small stochastic perturbations is considered in [118] and extended to decentralized stochastic optimization in [119], requiring bounded subgradients. In contrast, our focus is on smooth stochastic risks regularized by non-smooth, deterministic risks. Splitting the smooth stochastic part from the non-differentiable deterministic risk, and smoothing only the deterministic risk via a deterministic procedure will allow us to only require looser bounds on both components.

In the next sections we will explain how to construct the smooth approximation, $R_k^\delta(w)$, by appealing to conjugate functions and will show that the distance $\|w^o - w_\delta^o\|$ can be made arbitrarily small for $\delta \to 0$. We then present an algorithm to solve for the minimizer of (3.6) in a distributed manner and derive bounds on its performance. The analysis in future sections will rely on the following common assumptions [1, 22, 27]:

Assumption 3.1 (Lipschitz gradients)**.** *For each k, the gradient $\nabla J_k(\cdot)$ is Lipschitz, namely,*

there exists $\lambda_U \geq 0$ *such that for any* $x, y \in \mathbb{R}^M$*:*

$$\|\nabla J_k(x) - \nabla J_k(y)\| \leq \lambda_U \|x - y\| \tag{3.7}$$

□

Assumption 3.2 (Strong Convexity)**.** *The weighted aggregate of the differentiable risks is strongly convex, namely, there exists* $\lambda_L \geq 0$ *such that for any* $x, y \in \mathbb{R}^M$*:*

$$(x - y)^{\mathsf{T}} \cdot \sum_{k=1}^{N} p_k \left(\nabla_w J_k(x) - \nabla_w J_k(y)\right) \geq \lambda_L \|x - y\|^2 \tag{3.8}$$

□

Assumption 3.3 (Regularizers)**.** *For each* k*,* $R_k(\cdot)$ *is closed convex.* □

3.2 Algorithm Formulation

3.2.1 Construction of Smooth Approximation

To begin with, following the works [89, 116], we explain how smoothing of the regularizers is performed. Thus, recall that the conjugate function, denoted by $R_k^\star(w)$, of a regularizer $R_k(w)$ is defined as

$$R_k^\star(w) \triangleq \sup_{u \in \operatorname{dom} R_k} \left\{ w^{\mathsf{T}} u - R_k(u) \right\}. \tag{3.9}$$

A useful property of conjugate functions is that $R_k^\star(w)$ is always closed convex regardless of whether $R_k(w)$ is convex or not.

Definition 3.1 (Proximity function [89])**.** *A proximity function* $d(\cdot)$ *for a closed convex set* C *is a continuous, strongly-convex function with* $C \subseteq \operatorname{dom} d(\cdot)$*. We center and normalize the function so that*

$$\min_{w \in C} d(w) = 0 \tag{3.10}$$

and

$$\arg\min_{w \in C} \; d(w) = 0 \tag{3.11}$$

which exists and is unique, since $d(w)$ is strongly-convex. Furthermore, the proximity function is scaled to satisfy the following normalization (which means that its strong-convexity constant is set to one):

$$d(w) \geq \frac{1}{2}\|w\|^2. \tag{3.12}$$

□

Definition 3.2 (Smooth approximation [89])**.** *We choose a proximity function over $C = \operatorname{dom} R_k^\star(w)$ and define the smooth approximation of $R_k(\cdot)$ as:*

$$\begin{aligned} R_k^\delta(w) \triangleq & \max_{u \in \operatorname{dom} R_k^\star} \left\{ w^\mathsf{T} u - R_k^\star(u) - \delta \cdot d(u) \right\} \\ =& (R_k^\star + \delta \cdot d)^\star (w) \end{aligned} \tag{3.13}$$

□

The maximum in (3.13) is attained for all w since $R_k^\star(u) + \delta \cdot d(u)$ is strongly convex. Thus, observe that the smooth approximation for $R_k(w)$, which we are denoting by $R_k^\delta(w)$, is obtained by first perturbing the conjugate function $R_k^\star(u)$ by $\delta \cdot d(u)$ and then conjugating the result again. The perturbation makes the sum $R_k^\star(u) + \delta \cdot d(u)$ a strongly-convex function. The motivation behind this construction is the fact that the conjugate of a strongly-convex function is *differentiable* everywhere and, therefore, $R_k^\delta(w)$ is differentiable everywhere. This intuition is formalized in the following known theorem [89], preceded by an elementary lemma [120].

Lemma 3.1 (Conjugate subgradients [120])**.** *If $G(\cdot)$ is some closed and convex function, the subgradients of $G(\cdot)$ and its conjugate $G^\star(\cdot)$ are related as:*

$$v \in \partial G(w) \longleftrightarrow w \in \partial G^\star(v) \tag{3.14}$$

Proof. The theorem is from [120]. For reference, the proof is repeated in Appendix 3.A. □

Theorem 3.1 (Gradient of smooth approximation [89]). *Any $R_k^\delta(w)$ constructed according to* (3.13) *is differentiable with gradient vector*

$$\nabla R_k^\delta(w) = \underset{u \in \operatorname{dom} R_k^\star}{\arg\max} \left\{ w^\mathsf{T} u - R_k^\star(u) - \delta \cdot d(u) \right\}. \tag{3.15}$$

Furthermore, the gradient is co-coercive, i.e., it satisfies:

$$(x-y)^\mathsf{T} \left(\nabla R_k^\delta(x) - \nabla R_k^\delta(y) \right) \geq \delta \|\nabla R_k^\delta(x) - \nabla R_k^\delta(y)\|^2 \tag{3.16}$$

By Cauchy-Schwarz, this implies Lipschitz continuity, i.e.,

$$\|\nabla R_k^\delta(x) - \nabla R_k^\delta(y)\| \leq \frac{1}{\delta} \|x-y\|. \tag{3.17}$$

Proof. The theorem is from [89]. For reference, the proof is repeated in Appendix 3.B. □

The feasibility of stochastic-gradient algorithms for the minimization of (3.6) hinges on the assumption that (3.15) can be evaluated in closed form or at least easily. Fortunately, this is the case for a large class of regularizers of interest — see [79] for an overview of closed form solutions in the special case $d(\cdot) = \frac{1}{2}\|\cdot\|^2$ and [89, 116] for other distance choices. For example, for every function where the proximal operator [49]:

$$R_k^\delta(w) = \min_u \left(R_k(w) + \frac{1}{2\delta} \|w-u\|^2 \right) \tag{3.18}$$

can be evaluated in closed form, we can let $d(\cdot) \triangleq \frac{1}{2}\|\cdot\|^2$ and obtain [49]:

$$\nabla R_k^\delta(w) = \frac{1}{\delta} \left(w - \operatorname{prox}_{\delta R_k}(w) \right). \tag{3.19}$$

Depending on the regularizers $R_k(\cdot)$, other proximity functions may be more appropriate [89]. We point out that the smooth approximation (3.13) can equivalently be written as [116]:

$$R_k^\delta(w) = \min_{u \in \operatorname{dom} R_k} \left\{ R_k(u) + \delta \cdot d^\star \left(\frac{w-u}{\delta} \right) \right\} \tag{3.20}$$

To verify this, observe that

$$\begin{aligned}
R_k^{\delta}(w) &= \min_{u\in\mathrm{dom}\,R_k} \left\{ R_k(u) + \delta\cdot\sup_z \left\{ z^{\mathsf{T}}\left(\frac{w-u}{\delta}\right) - d\,(z)\right\}\right\} \\
&= \min_{u\in\mathrm{dom}\,R_k} \left\{ R_k(u) + \sup_z \left\{ z^{\mathsf{T}}\,(w-u) - \delta\cdot d\,(z)\right\}\right\} \\
&= \sup_z \left\{ \inf_u \left\{ -z^{\mathsf{T}}u + R_k(u)\right\} + z^{\mathsf{T}}w - \delta\cdot d\,(z)\right\} \\
&= \sup_z \left\{ -\sup_u \left\{ R_k(u) - z^{\mathsf{T}}u\right\} + z^{\mathsf{T}}w - \delta\cdot d\,(z)\right\} \\
&= \max_z \left\{ z^{\mathsf{T}}w - R_k^{\star}(z) - \delta\cdot d\,(z)\right\} \qquad (3.21)
\end{aligned}$$

Expression (3.20) is known as the infimal convolution.

3.2.2 Accuracy of the Smooth Approximation

Replacing the original optimization problem (3.3) by the smoothed cost (3.6) naturally results in a bias, since the new minimizer w_δ^o will generally be different from the original minimizer w^o. This bias, when not properly controlled, can degrade the performance of the algorithm. For this reason, a number of works have examined the smoothing bias introduced through conjugate smoothing under various conditions on the cost functions. In the centralized setting, when $N = 1$, it has been established that $R_k^{\delta}(w) \to R_k(w)$ both pointwise and epigraphically, which implies $w_\delta^o \to w^o$ as $\delta \to 0$ [121], while [122] showed a sum of costs $\sum_{k=1}^N p_k R_k(w)$, when smoothed individually, will continue to converge epigraphically. While encouraging, these results do not guarantee a rate at which $w_\delta^o \to w^o$, complicating the choice of the smoothing parameter δ. Pointwise convergence has been strengthened to uniform convergence, i.e., $\left|R_k(w) - R_k^{\delta}(w)\right| \leq O(\delta)$ for costs with bounded subgradients for $N = 1$ [89, 116] and for a collection of costs, each with bounded subgradients in [117].

We present here a variation of these results by restricting ourselves to *strongly*-convex costs, but allowing for regularizers with *unbounded* sub-gradients and establishing $\|w^o - w_\delta^o\|^2 \leq O(\delta)$ rather than simply $w_\delta^o \to w^o$.

Theorem 3.2 (Accuracy of smooth approximation)**.** *The bias introduced by smoothing the*

original problem diminishes linearly with δ, i.e.,

$$\|w^o - w_\delta^o\|^2 \leq \frac{2}{\lambda_L} \sum_{k=1}^{N} p_k \delta d\left(r_k^o\right) = O(\delta) \tag{3.22}$$

where $r_k^o \in \partial R_k(w^o)$ such that

$$\sum_{k=1}^{N} p_k \left\{\nabla J_k(w^o) + r_k^o\right\} = 0 \tag{3.23}$$

This collection of $\{r_k^o\}$ is guaranteed to exist, since $w^o \triangleq \arg\min \sum_{k=1}^{N} p_k \left\{J_k(w) + R_k(w)\right\}$.

Proof. Appendix 3.C. □

3.2.3 Regularized Diffusion Strategy

Now that we have established a method for constructing a differentiable approximation for each regularizer, we can solve for the minimizer of (3.6) by resorting to the following (adapt-then-combine form of the) diffusion strategy [1, 22, 27]:

$$\boldsymbol{\phi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) - \mu\nabla R_k^\delta(\boldsymbol{w}_{k,i-1}) \tag{3.24}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell=1}^{N} a_{\ell k}\boldsymbol{\phi}_{\ell,i} \tag{3.25}$$

where $\mu > 0$ is a small step-size parameter. In this implementation, each agent k first performs the stochastic-gradient update (3.24), starting from its existing iterate value $\boldsymbol{w}_{k,i-1}$, and obtains an intermediate iterate $\boldsymbol{\phi}_{k,i}$. Subsequently, agent k consults with its neighbors and combines their intermediate iterates into $\boldsymbol{w}_{k,i}$ according to (3.25). Motivated by the construction in [33], we can refine (3.24)–(3.25) further as follows. We first introduce an

auxiliary variable $\boldsymbol{\psi}_{k,i}$ and rewrite (3.24) in the equivalent form:

$$\boldsymbol{\phi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \tag{3.26}$$

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{\phi}_{k,i} - \mu\nabla R_k^{\delta}(\boldsymbol{w}_{k,i-1}) \tag{3.27}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell=1}^{N} a_{\ell k}\boldsymbol{\psi}_{\ell,i} \tag{3.28}$$

We can now appeal to an incremental-type argument [90, 123] by noting that it is reasonable to expect $\boldsymbol{\phi}_{k,i}$ to be an improved estimate for w_δ^o compared to $\boldsymbol{w}_{k,i-1}$. Therefore, we replace $\boldsymbol{w}_{k,i-1}$ in (3.27) by $\boldsymbol{\phi}_{k,i}$ and arrive at the following regularized diffusion implementation.

Algorithm 3.1 Regularized Diffusion Strategy

$$\boldsymbol{\phi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \tag{3.29}$$

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{\phi}_{k,i} - \mu\nabla R_k^{\delta}(\boldsymbol{\phi}_{k,i}) \tag{3.30}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell=1}^{N} a_{\ell k}\boldsymbol{\psi}_{\ell,i} \tag{3.31}$$

Example 3.1 (Proximal Diffusion Learning)**.** *Choosing* $d(w) = \frac{1}{2}\|w\|^2$ *turns the smooth approximation* (3.13) *into*

$$R_k^{\delta}(w) = \left(R_k^{\star}(w) + \frac{\delta}{2}\|w\|^2\right)^{\star} \tag{3.32}$$

which is the well-known Moreau envelope [49]. It can be rewritten equivalently as

$$R_k^{\delta}(w) = \min_u \left(R_k(w) + \frac{1}{2\delta}\|w-u\|^2\right) \tag{3.33}$$

where the minimizing argument is identified as the proximal operator:

$$\mathrm{prox}_{\delta R_k}(w) = \arg\min_u \left(R_k(w) + \frac{1}{2\delta}\|w-u\|^2\right) \tag{3.34}$$

For many costs $R_k(w)$*, the proximal operator can be evaluated in closed form. The gradient*

of the Moreau envelope can also be written as

$$\nabla R_k^{\delta}(w) = \frac{1}{\delta}\left(w - \text{prox}_{\delta R_k}(w)\right). \tag{3.35}$$

This allows us to rewrite iterations (3.29)–(3.31) *as*

$$\boldsymbol{\phi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \tag{3.36}$$

$$\boldsymbol{\psi}_{k,i} = \left(1 - \frac{\mu}{\delta}\right)\boldsymbol{\phi}_{k,i} + \frac{\mu}{\delta}\text{prox}_{\delta R_k}(\boldsymbol{\phi}_{k,i}) \tag{3.37}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell=1}^{N} a_{\ell k}\boldsymbol{\phi}_{\ell,i} \tag{3.38}$$

which is a damped variation of the proximal diffusion algorithm studied in [50] under the stronger assumption of small Lipschitz continuous regularizers. □

3.3 Convergence Analysis

3.3.1 Centralized Recursion

We now examine the convergence properties of the diffusion strategy (3.29)–(3.31). To do so, and motivated by the approach introduced in [27], it is useful to introduce the following centralized recursion to serve as a frame of reference:

$$w_i = w_{i-1} - \mu\sum_{k=1}^{N} p_k\nabla J_k(w_{i-1}) - \mu\sum_{k=1}^{N} p_k\nabla R_k^{\delta}(w_{i-1}) \tag{3.39}$$

This recursion amounts to a gradient-descent iteration applied to the smoothed aggregate cost in (3.6) under the assumption that the risk functions (and therefore their gradients) are known. For convenience of presentation, we introduce the central operator $T_c(x) : \mathbb{R}^M \to \mathbb{R}^M$ defined as follows:

$$T_c(x) \triangleq x - \mu\sum_{k=1}^{N} p_k\nabla J_k(x) - \mu\sum_{k=1}^{N} p_k\nabla R_k^{\delta}(x) \tag{3.40}$$

so that the reference recursion (3.39) becomes $w_i = T_c(w_{i-1})$.

Lemma 3.2 (Contraction mapping)**.** *Assume $\mu \leq 2\delta$. Then, the centralized recursion* (3.39) *satisfies*

$$\|T_c(x) - T_c(y)\| \leq \gamma_c \|x - y\| \tag{3.41}$$

where $\gamma_c > 0$ can be made strictly less than one by selecting sufficiently small μ and is given by:

$$\gamma_c = 1 - \mu\lambda_L + \mu^2 \left(\frac{\lambda_U^2}{2 - \frac{\mu}{\delta}} \right). \tag{3.42}$$

From Banach's fixed point theorem [82] and (3.40)*, we conclude that for sufficiently small μ, $w_i = T_c(w_{i-1})$ converges exponentially to the unique fixed-point w_δ^o, the minimizer of* (3.6)*.*

Proof. Appendix 3.D. □

3.3.2 Network Basis Transformation

We are now ready to examine the behavior of the diffusion strategy (3.29)–(3.31), which employs stochastic gradients. Structurally, our argument follows those in [27] in the absence of regularizers. We begin by introducing the following extended vectors and matrices, which collect quantities of interest from across all agents in the network:

$$\boldsymbol{\mathcal{w}}_i \triangleq \operatorname{col}\{\boldsymbol{w}_{1,i}, \ldots, \boldsymbol{w}_{N,i}\} \tag{3.43}$$

$$\mathcal{A} \triangleq A \otimes I_M \tag{3.44}$$

$$g(\boldsymbol{\mathcal{w}}_i) \triangleq \operatorname{col}\{\nabla_w J_1(\boldsymbol{w}_{1,i}), \ldots, \nabla_w J_N(\boldsymbol{w}_{N,i})\} \tag{3.45}$$

$$\widehat{g}(\boldsymbol{\mathcal{w}}_i) \triangleq \operatorname{col}\left\{\widehat{\nabla_w J_1}(\boldsymbol{w}_{1,i}), \ldots, \widehat{\nabla_w J_N}(\boldsymbol{w}_{N,i})\right\} \tag{3.46}$$

$$r(\boldsymbol{\mathcal{w}}_i) \triangleq \operatorname{col}\left\{\nabla_w R_1^\delta(\boldsymbol{w}_{1,i}), \ldots, \nabla_w R_N^\delta(\boldsymbol{w}_{N,i})\right\} \tag{3.47}$$

$$q(\boldsymbol{\mathcal{w}}_i) \triangleq r(\boldsymbol{\mathcal{w}}_i - \mu g(\boldsymbol{\mathcal{w}}_i)) \tag{3.48}$$

$$\widehat{q}(\boldsymbol{\mathcal{w}}_i) \triangleq r(\boldsymbol{\mathcal{w}}_i - \mu \widehat{g}(\boldsymbol{\mathcal{w}}_i)) \tag{3.49}$$

Using these definitions, iterations (3.29)–(3.31) show that the network vector $\boldsymbol{\mathcal{w}}_i$ evolves according to the following dynamics:

$$\boldsymbol{\mathcal{w}}_i = \mathcal{A}^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_{i-1} - \mu\mathcal{A}^{\mathsf{T}}\left(\widehat{g}(\boldsymbol{\mathcal{w}}_{i-1}) + \widehat{q}(\boldsymbol{\mathcal{w}}_{i-1})\right) \tag{3.50}$$

By construction, the combination matrix A is left-stochastic and primitive and hence admits a Jordan decomposition of the form $A = V_\epsilon J V_\epsilon^{-1}$ with [1, 27]:

$$V_\epsilon = \begin{bmatrix} p & V_R \end{bmatrix}, \quad J = \begin{bmatrix} 1 & 0 \\ 0 & J_\epsilon \end{bmatrix}, \quad V_\epsilon^{-1} = \begin{bmatrix} \mathbb{1}^{\mathsf{T}} \\ V_L^{\mathsf{T}} \end{bmatrix} \tag{3.51}$$

where J_ϵ is a block Jordan matrix with the eigenvalues $\lambda_2(A)$ through $\lambda_N(A)$ on the diagonal and ϵ on the first lower sub-diagonal. The extended matrix $\mathcal{A}$ then satisfies $\mathcal{A} = \mathcal{V}_\epsilon \mathcal{J} \mathcal{V}_\epsilon^{-1}$ with $\mathcal{V}_\epsilon = V_\epsilon \otimes I_N$, $\mathcal{J} = J \otimes I_N$, $\mathcal{V}_\epsilon^{-1} = V_\epsilon^{-1} \otimes I_N$. Multiplying both sides of (3.50) by $\mathcal{V}_\epsilon^{\mathsf{T}}$ and introducing the transformed iterate vector $\boldsymbol{\mathcal{w}}_i' \triangleq \mathcal{V}_\epsilon^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_i$, we obtain

$$\boldsymbol{\mathcal{w}}_i' = \mathcal{J}^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_{i-1}' - \mu\mathcal{J}^{\mathsf{T}}\mathcal{V}_\epsilon^{\mathsf{T}}\left(\widehat{g}(\boldsymbol{\mathcal{w}}_{i-1}) + \widehat{q}(\boldsymbol{\mathcal{w}}_{i-1})\right) \tag{3.52}$$

Following [1, 27], we can exploit the structure of the decomposition (3.51) to provide further insight into this transformed recursion. Let $\boldsymbol{\mathcal{w}}_i = \mathrm{col}\left\{\boldsymbol{w}_{c,i}, \boldsymbol{\mathcal{w}}_{e,i}\right\}$, where $\boldsymbol{w}_{c,i} \in \mathbb{R}^{N\times 1}$ and $\boldsymbol{\mathcal{w}}_{e,i} \in \mathbb{R}^{(N-1)M\times 1}$. Then, recursion (3.52) can be decomposed as

$$\boldsymbol{w}_{c,i} = \boldsymbol{w}_{c,i-1} - \mu\left(p^{\mathsf{T}} \otimes I_N\right)\left(\widehat{g}(\boldsymbol{\mathcal{w}}_{i-1}) + \widehat{q}(\boldsymbol{\mathcal{w}}_{i-1})\right) \tag{3.53}$$

$$\boldsymbol{\mathcal{w}}_{e,i} = \mathcal{J}_\epsilon^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_{e,i-1} - \mu\mathcal{J}_\epsilon^{\mathsf{T}}\mathcal{V}_R^{\mathsf{T}}\left(\widehat{g}(\boldsymbol{\mathcal{w}}_{i-1}) + \widehat{q}(\boldsymbol{\mathcal{w}}_{i-1})\right) \tag{3.54}$$

Note from $\boldsymbol{\mathcal{w}}_i' = \mathcal{V}_\epsilon^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_i$, that [27]:

$$\boldsymbol{w}_{c,i} = \left(p^{\mathsf{T}} \otimes I_M\right)\boldsymbol{\mathcal{w}}_i = \sum_{k=1}^{N} p_k\,\boldsymbol{w}_{k,i} \tag{3.55}$$

Hence, $\boldsymbol{w}_{c,i}$ is the weighted centroid vector of all iterates $\boldsymbol{w}_{k,i}$ across the network. From $\boldsymbol{\mathcal{W}}_i = (\mathcal{V}_\epsilon^{-1})^{\mathsf{T}} \boldsymbol{\mathcal{W}}_i'$ on the other hand, one obtains [27]:

$$\boldsymbol{\mathcal{W}}_i = \mathbb{1} \otimes \boldsymbol{w}_{c,i} + \mathcal{V}_L \, \boldsymbol{\mathcal{W}}_{e,i} \tag{3.56}$$

so that $\boldsymbol{\mathcal{W}}_{e,i}$ can be interpreted as the deviation of individual estimates from the weighted centroid vector $\boldsymbol{w}_{c,i}$ across the network.

We examine the centroid recursion (3.53) in greater detail. Thus, note that

$$\begin{aligned}
\boldsymbol{w}_{c,i} =& \; \boldsymbol{w}_{c,i-1} - \mu \left(p^{\mathsf{T}} \otimes I_M\right) \left(\widehat{g}(\boldsymbol{\mathcal{W}}_{i-1}) + \widehat{q}(\boldsymbol{\mathcal{W}}_{i-1})\right) \\
=& \; \boldsymbol{w}_{c,i-1} - \mu \left(p^{\mathsf{T}} \otimes I_M\right) \left(g(\mathbb{1} \otimes \boldsymbol{w}_{c,i-1}) + r(\mathbb{1} \otimes \boldsymbol{w}_{c,i-1})\right) \\
& - \mu \left(p^{\mathsf{T}} \otimes I_M\right) \left(g(\boldsymbol{\mathcal{W}}_{i-1}) + q(\boldsymbol{\mathcal{W}}_{i-1}) - g(\mathbb{1} \otimes \boldsymbol{w}_{c,i-1}) - q(\mathbb{1} \otimes \boldsymbol{w}_{c,i-1})\right) \\
& - \mu \left(p^{\mathsf{T}} \otimes I_M\right) \left(\widehat{g}(\boldsymbol{\mathcal{W}}_{i-1}) + \widehat{q}(\boldsymbol{\mathcal{W}}_{i-1}) - g(\boldsymbol{\mathcal{W}}_{i-1}) - q(\boldsymbol{\mathcal{W}}_{i-1})\right) \\
& - \mu \left(p^{\mathsf{T}} \otimes I_M\right) \left(q(\boldsymbol{\mathcal{W}}_{i-1}) - r(\boldsymbol{\mathcal{W}}_{i-1})\right) \\
=& \; T_c(\boldsymbol{w}_{c,i-1}) - \mu \left(p^{\mathsf{T}} \otimes I_M\right) \left(\boldsymbol{t}_{i-1} + \boldsymbol{s}_i + \boldsymbol{u}_{i-1}\right)
\end{aligned} \tag{3.57}$$

where we replaced

$$\begin{aligned}
& \boldsymbol{w}_{c,i-1} - \mu \left(p^{\mathsf{T}} \otimes I_M\right) \left(g(\mathbb{1} \otimes \boldsymbol{w}_{c,i-1}) + r(\mathbb{1} \otimes \boldsymbol{w}_{c,i-1})\right) \\
=& \; \boldsymbol{w}_{c,i-1} - \mu \sum_{k=1}^{N} p_k \nabla J_k(\boldsymbol{w}_{c,i-1}) - \mu \sum_{k=1}^{N} p_k \nabla R_k^{\delta}(\boldsymbol{w}_{c,i-1}) \\
\overset{(3.40)}{=}& \; T_c(\boldsymbol{w}_{c,i-1})
\end{aligned} \tag{3.58}$$

and introduced the perturbation terms:

$$\boldsymbol{t}_{i-1} = g(\boldsymbol{\mathcal{W}}_{i-1}) + q(\boldsymbol{\mathcal{W}}_{i-1}) - g(\mathbb{1} \otimes \boldsymbol{w}_{c,i-1}) - q(\mathbb{1} \otimes \boldsymbol{w}_{c,i-1}) \tag{3.59}$$

$$\boldsymbol{s}_i = \widehat{g}(\boldsymbol{\mathcal{W}}_{i-1}) + \widehat{q}(\boldsymbol{\mathcal{W}}_{i-1}) - g(\boldsymbol{\mathcal{W}}_{i-1}) - q(\boldsymbol{\mathcal{W}}_{i-1}) \tag{3.60}$$

$$\boldsymbol{u}_{i-1} = q(\boldsymbol{\mathcal{W}}_{i-1}) - r(\boldsymbol{\mathcal{W}}_{i-1}) \tag{3.61}$$

It follows from (3.57) that the centroid recursion is a perturbed version of the central recursion introduced earlier in (3.40). The perturbation arising from disagreement across agents in the network is captured in $\boldsymbol{t}_{i-1}$, while stochastic perturbations due to instantaneous gradient approximations is captured in $\boldsymbol{s}_i$. The incremental implementation causes $\boldsymbol{u}_{i-1}$. It is therefore reasonable to expect that $\boldsymbol{w}_{c,i}$ will evolve close to the central variable w_i from (3.39), which was already shown to converge to w_δ^o in Lemma 3.2. This intuition was formalized for the classical diffusion algorithm without regularizers in [27]. Motivated by that work, we define $\widetilde{\boldsymbol{w}}_{c,i-1} = w_\delta^o - \boldsymbol{w}_{c,i-1}$. Since w_δ^o is a fixed-point of $T_c(\cdot)$, i.e., $w_\delta^o = T_c(w_\delta^o)$, the error $\widetilde{\boldsymbol{w}}_{c,i-1}$ satisfies the recursion

$$\widetilde{\boldsymbol{w}}_{c,i-1} = T_c(w_\delta^o) - T_c(\boldsymbol{w}_{c,i-1}) + \mu \left(p^{\mathsf{T}} \otimes I_M\right) \left(\boldsymbol{t}_{i-1} + \boldsymbol{s}_i + \boldsymbol{u}_{i-1}\right) \tag{3.62}$$

With the same perturbation terms, expression (3.54) turns into

$$\boldsymbol{\mathcal{W}}_{e,i} = \mathcal{J}_\epsilon^{\mathsf{T}} \boldsymbol{\mathcal{W}}_{e,i-1} - \mu \mathcal{J}_\epsilon^{\mathsf{T}} \mathcal{V}_R^{\mathsf{T}} \big(\boldsymbol{t}_{i-1} + \boldsymbol{s}_i + \boldsymbol{u}_{i-1} - g(\mathbb{1} \otimes \boldsymbol{w}_{c,i-1}) - r(\mathbb{1} \otimes \boldsymbol{w}_{c,i-1})\big) \tag{3.63}$$

We employ the following common assumption on the perturbations caused by the gradient noise [1, 22, 27].

Assumption 3.4 (Gradient noise process)**.** *For each k, the gradient noise process is defined as*

$$\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) = \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) - \nabla J_k(\boldsymbol{w}_{k,i-1}) \tag{3.64}$$

and satisfies

$$\mathbb{E}\left[\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) | \boldsymbol{\mathcal{F}}_{i-1}\right] = 0 \tag{3.65a}$$

$$\mathbb{E}\left[\| \boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2 | \boldsymbol{\mathcal{F}}_{i-1}\right] \leq \beta^2 \| \boldsymbol{w}_{k,i-1} \|^2 + \sigma^2 \tag{3.65b}$$

for some non-negative constants $\{\beta^2, \sigma^2\}$, and where $\boldsymbol{\mathcal{F}}_{i-1}$ denotes the filtration generated by the random processes $\{\boldsymbol{w}_{\ell,j}\}$ for all $\ell = 1, 2, \ldots, N$ and $j \leq i-1$, i.e., $\boldsymbol{\mathcal{F}}_{i-1}$ represents the information that is available about the random processes $\{\boldsymbol{w}_{\ell,j}\}$ up to time $i-1$. □

For a block-vector $\boldsymbol{x} \in \mathbb{R}^{MN\times 1}$ consisting of N blocks of size $M \times 1$, let [27]:

$$P[\boldsymbol{x}] = \operatorname{col}\left\{\mathbb{E}\,\|\boldsymbol{x}_1\|^2, \ldots, \mathbb{E}\,\|\boldsymbol{x}_N\|^2\right\} \in \mathbb{R}^{N\times 1} \tag{3.66}$$

Note that $\mathbb{1}^{\mathsf{T}} P[\boldsymbol{x}] = \mathbb{E}\,\|\boldsymbol{x}\|^2$. Furthermore, let $v_{L,k}$ denote the k-th row of V_L and let $\nu = \max_k \|v_{L,k} \otimes I_M\|$, which is independent of μ and δ.

Lemma 3.3 (Bounds on perturbation terms)**.** *The perturbation terms in* (3.62) *satisfy the following bounds:*

$$P[\boldsymbol{t}_{i-1}] \preceq \left(2\lambda_U^2 + 4\frac{1+\mu^2}{\delta^2}\right)\nu^2 \mathbb{1}\mathbb{1}^{\mathsf{T}} P[\boldsymbol{\mathcal{W}}_{e,i-1}] \tag{3.67}$$

$$P[\boldsymbol{u}_{i-1}] \preceq \frac{\mu^2}{\delta^2}\big(3\lambda_U^2\nu^2\mathbb{1}\mathbb{1}^{\mathsf{T}} P[\boldsymbol{\mathcal{W}}_{e,i-1}] + 3\lambda_U^2 P[\mathbb{1}\otimes\widetilde{\boldsymbol{w}}_{c,i-1}] + 3P[g(\mathbb{1}\otimes w_\delta^o)]\big) \tag{3.68}$$

$$P[\boldsymbol{s}_i - \mathbb{E}\,\boldsymbol{s}_i] \preceq 3\beta^2 P[\mathbb{1}\otimes\widetilde{\boldsymbol{w}}_{c,i-1}] + 3\beta^2\nu^2\mathbb{1}\mathbb{1}^{\mathsf{T}} P[\boldsymbol{\mathcal{W}}_{e,i-1}] + 3\beta^2 P[\mathbb{1}\otimes w_\delta^o] + \sigma^2\mathbb{1} \tag{3.69}$$

$$P[\mathbb{E}\,\boldsymbol{s}_i] \preceq 3\beta^2\frac{\mu^2}{\delta^2}P[\mathbb{1}\otimes\widetilde{\boldsymbol{w}}_{c,i-1}] + 3\beta^2\frac{\mu^2}{\delta^2}\nu^2\mathbb{1}\mathbb{1}^{\mathsf{T}} P[\boldsymbol{\mathcal{W}}_{e,i-1}] + 3\beta^2\frac{\mu^2}{\delta^2}P[\mathbb{1}\otimes w_\delta^o] + \frac{\mu^2}{\delta^2}\sigma^2\mathbb{1} \tag{3.70}$$

$$P[g(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})] \preceq 2\lambda_U^2 P[\mathbb{1}\otimes\widetilde{\boldsymbol{w}}_{c,i-1}] + 2P[g(\mathbb{1}\otimes w_\delta^o)] \tag{3.71}$$

$$P[r(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})] \preceq \frac{2}{\delta^2}P[\mathbb{1}\otimes\widetilde{\boldsymbol{w}}_{c,i-1}] + 2P[r(\mathbb{1}\otimes w_\delta^o)] \tag{3.72}$$

Proof. Appendix 3.E. □

3.3.3 Mean-Square-Error Bounds

Using the bounds on the perturbation terms obtained in Lemma 3.3, we can formulate a recursive bound on the mean-square error.

Lemma 3.4 (Mean-Square-Error Recursion)**.** *The variances of $\widetilde{\boldsymbol{w}}_{c,i}$ and $\boldsymbol{\mathcal{W}}_{e,i}$ are coupled and recursively bounded as*

$$\begin{bmatrix}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i}\|^2\\ \mathbb{E}\,\|\,\boldsymbol{\mathcal{W}}_{e,i}\,\|^2\end{bmatrix} \preceq \Gamma\begin{bmatrix}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2\\ \mathbb{E}\,\|\,\boldsymbol{\mathcal{W}}_{e,i-1}\,\|^2\end{bmatrix} + \begin{bmatrix}\frac{\mu^3}{\delta^2}b_1 + \frac{\mu^3}{\delta^2}b_2 + \mu^2 b_3\\ \frac{\mu^2}{\delta^2}b_4 + \mu^2 b_5 + \frac{\mu^4}{\delta^2}b_6\end{bmatrix} \tag{3.73}$$

where

$$\Gamma = \begin{bmatrix} \gamma_c + \frac{\mu^3}{\delta^2}h_1 + \mu^2 h_2 & \frac{\mu}{\delta^2}h_3 + \mu h_4 + \frac{\mu^3}{\delta^2}h_5 + \mu^2 h_6 \\ \frac{\mu^2}{\delta^2}h_7 + \mu^2 h_8 + \frac{\mu^4}{\delta^2}h_9 & \|\mathcal{J}_\epsilon\| + \frac{\mu^2}{\delta^2}h_{10} + \mu^2 h_{11} + \frac{\mu^4}{\delta^2}h_{12} \end{bmatrix} \tag{3.74}$$

$$\gamma_c \triangleq 1 - \mu\lambda_L + \mu^2 \left(\frac{\lambda_U^2}{2 - \frac{\mu}{\delta}} \right) \tag{3.75}$$

$$a_1 \triangleq \frac{1}{\lambda_L - \mu \frac{\lambda_U^2}{2 - \frac{\mu}{\delta}}} = O(1) \tag{3.76}$$

$$a_2 \triangleq \frac{25N\|\mathcal{J}_\epsilon\|^2 \|\mathcal{V}_R\|^2}{1 - \|J_\epsilon\|} = O(1) \tag{3.77}$$

$$h_1 \triangleq 9(\beta^2 + \lambda_U^2)a_1 = O(1) \tag{3.78}$$

$$h_2 \triangleq 3\beta^2 = O(1) \tag{3.79}$$

$$h_3 \triangleq 3\nu^2 a_1 = O(1) \tag{3.80}$$

$$h_4 \triangleq 6\nu^2 \lambda_U^2 a_1 = O(1) \tag{3.81}$$

$$h_5 \triangleq 9\nu^2(\lambda_U^2 + \beta^2)a_1 = O(1) \tag{3.82}$$

$$h_6 \triangleq 3\nu^2\beta^2 = O(1) \tag{3.83}$$

$$h_7 \triangleq 2a_2 = O(1) \tag{3.84}$$

$$h_8 \triangleq \left(2\lambda_U^2 + \frac{1 - \|\mathcal{J}_\epsilon\|}{25} 3\beta^2 \right) a_2 = O(1) \tag{3.85}$$

$$h_9 \triangleq 3\left(\lambda_U^2 + \beta^2\right) a_2 = O(1) \tag{3.86}$$

$$h_{10} \triangleq \nu^2 a_2 = O(1) \tag{3.87}$$

$$h_{11} \triangleq \nu^2 \left(2\lambda_U^2 + \frac{1 - \|\mathcal{J}_\epsilon\|}{25} 3\beta^2 \right) a_2 = O(1) \tag{3.88}$$

$$h_{12} \triangleq \nu^2 \left(1 + 3\lambda_U^2 + 3\beta^2\right) a_2 = O(1) \tag{3.89}$$

$$b_1 \triangleq 9a_1 \|g(w_\delta^o)\|^2 = O(1) \tag{3.90}$$

$$b_2 \triangleq 3a_1(3\beta^2\|w_\delta^o\|^2 + \sigma^2) = O(1) \tag{3.91}$$

$$b_3 \triangleq 3\beta^2\|w_\delta^o\|^2 + \sigma^2 = O(1) \tag{3.92}$$

$$b_4 \triangleq 2a_2 \left(\delta^2 \|r(\mathbb{1} \otimes w_\delta^o)\|^2\right) = O(1) \tag{3.93}$$

$$b_5 \triangleq 2a_2\|g(\mathbb{1} \otimes w_\delta^o)\|^2 + \|\mathcal{J}_\epsilon\|^2 \|\mathcal{V}_R\|^2 \left(3\beta^2 N\|w_\delta^o\|^2 + N\sigma^2\right) = O(1) \tag{3.94}$$

$$b_6 \triangleq a_2\left(3\|g(\mathbb{1}\otimes w_\delta^o)\|^2 + 3\beta^2 N\|w_\delta^o\|^2 + N\sigma^2\right) = O(1) \tag{3.95}$$

Proof. Appendix 3.F. □

It is evident from expression (3.74) that the stability of the driving matrix Γ depends critically on the fraction between the step-size μ and the smoothing parameter δ. Motivated by this observation, let us set, for a small $\kappa > 0$:

$$\delta = \mu^{\frac{1}{2}-\kappa} \tag{3.96}$$

so that

$$\frac{\mu}{\delta^2} = \mu^{2\kappa} \to 0 \text{ as } \mu \to 0 \tag{3.97}$$

Under this construction, the driving matrix satisfies

$$\Gamma = \begin{bmatrix} \gamma_c + O(\mu^2) & O(\mu^{2\kappa}) \\ O(\mu^{1+2\kappa}) & \|\mathcal{J}_\epsilon\| + O(\mu^{1+2\kappa}) \end{bmatrix} \tag{3.98}$$

which ensures that the off-diagonal coupling terms diminish as $\mu, \delta \to 0$.

Lemma 3.5. *Let $\delta = \mu^{\frac{1}{2}-\kappa}$, $\frac{1}{2} > \kappa > 0$. Then there exists a small enough μ, such that $\rho(\Gamma) < 1$. Furthermore,*

$$\limsup_{i\to\infty} \begin{bmatrix} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 \\ \mathbb{E}\,\|\,\boldsymbol{\mathcal{W}}_{e,i}\,\|^2 \end{bmatrix} \preceq \begin{bmatrix} O(\mu) + O(\mu^{4\kappa}) \\ O(\mu^{1+2\kappa}) \end{bmatrix} \tag{3.99}$$

Proof. See Appendix 3.G. □

Theorem 3.3. *Let $\delta = \mu^{\frac{1}{2}-\kappa}$, $\frac{1}{2} > \kappa > \frac{1}{4}$. Then it holds that for sufficiently small μ,*

$$\limsup_{i\to\infty}\ \mathbb{E}\,\|w_\delta^o - \boldsymbol{w}_{k,i}\,\|^2 = O(\mu) \tag{3.100}$$

Proof. We have

$$\begin{aligned}\mathbb{E}\,\|w_\delta^o - \boldsymbol{w}_{k,i}\,\|^2 &= \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i} + (v_{L,k}\otimes I_M)\,\boldsymbol{\mathcal{W}}_{e,i}\,\|^2 \\ &\leq 2\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 + 2\nu^2\,\mathbb{E}\,\|\,\boldsymbol{\mathcal{W}}_{e,i}\,\|^2 \end{aligned} \tag{3.101}$$

so that the theorem follows after taking the limit and applying Lemma 3.5. □

3.4 Application: Division of Labor in Machine Learning

We illustrate the performance of the algorithm in an online machine learning problem over a heterogeneous network. Given random binary class variables $\boldsymbol{\gamma} = \pm 1$ and feature vectors $\boldsymbol{h} \in \mathbb{R}^M$, the general objective in single-agent machine learning is to find a classifier $c^\star(\boldsymbol{h})$, such that

$$c^\star \triangleq \arg\min_c \text{Prob}\left\{c(\boldsymbol{h}) \neq \boldsymbol{\gamma}\right\}. \tag{3.102}$$

We restrict the class of permissible classifiers to linear classifiers of the form $c(\boldsymbol{h}) = \boldsymbol{h}^{\mathsf{T}} w$ with $w \in \mathbb{R}^M$ and approximate (3.102) by the logistic cost to obtain

$$w^o \triangleq \arg\min_w \;\mathbb{E}\;\ln\left[1 + e^{-\boldsymbol{\gamma}\boldsymbol{h}^{\mathsf{T}} w}\right] \tag{3.103}$$

3.4.1 Group Lasso

Regularization is an effective technique to incorporate prior structural knowledge about the classifier into the optimization problem as a means to avoiding overfitting and improving generalization ability. For example, when the linear classifier is known to be sparse, regularization through the ℓ_1-norm, also known as Lasso-regularization, has been shown to encourage sparse solutions [88]. When there is further knowledge about the structure of the sparsity, the group-Lasso has been proposed [124, 125]. It takes the form

$$R(w) = \sum_k \lambda_k \|D_k w\|_1 = \sum_k \lambda_k \|w_g^k\|_1 \tag{3.104}$$

where

$$w_g^k \triangleq D_k w \tag{3.105}$$

and D_k denotes a diagonal selection matrix with entries 0 or 1 where 1's appear for entries of w belonging to a group. The resulting regularizer then encourages all elements of a group to either be active or equal to zero jointly [124, 125]. Note that (3.104) is in the form of a sum-of-costs and hence immediately decomposable.

3.4.2 Network Structure

We consider a network consisting of 3 types of agents: fully-informed ($\mathcal{F}$), data-informed ($\mathcal{D}$), and structure-informed ($\mathcal{S}$) agents. Fully-informed agents have access to streaming realizations $\{\boldsymbol{\gamma}_k(i), \boldsymbol{h}_{k,i}\}$ as well as knowledge about a subset of covariates of w which are likely to be sparse, collected in w_g^k. These agents are equipped with the regularized cost $J_k(w) + R_k(w)$, where

$$J_k(w) = \mathbb{E}\ \ln\left[1 + e^{-\boldsymbol{\gamma}_k \boldsymbol{h}_k^{\mathsf{T}} w}\right] + \rho_2 \|w\|_2^2 \tag{3.106}$$

$$R_k(w) = \rho_1 \|w_g^k\|_1 \tag{3.107}$$

for $k \in \mathcal{F}$. Data-informed agents have access to streaming realizations $\{\boldsymbol{\gamma}_k(i), \boldsymbol{h}_{k,i}\}$, but no knowledge about the structure of sparsity in w. They are equipped with

$$J_k(w) = \mathbb{E}\ \ln\left[1 + e^{-\boldsymbol{\gamma}_k \boldsymbol{h}_k^{\mathsf{T}} w}\right] + \rho_2 \|w\|_2^2 \tag{3.108}$$

$$R_k(w) = 0 \tag{3.109}$$

for $k \in \mathcal{D}$. Structure-informed agents have information about the sparsity of w, but no access to realizations of feature vectors. They are equipped with

$$J_k(w) = 0 \tag{3.110}$$

$$R_k(w) = \rho_1 \|w_g^k\|_1 \tag{3.111}$$

for $k \in \mathcal{S}$. Similar to ordinary ℓ_1-norm regularization, the proximal operator of $\rho_1\|w_g^k\|_1$ is available in closed form as a variation of soft-thresholding. Note that $\|w_g^k\|_1 = \|D_k w\|_1$, where D_k is a diagonal matrix with $D_{(ii)} = 1$, if the $i-th$ element of w is likely to be sparse and 0 otherwise. We then obtain

$$\text{prox}_{\delta\rho_1\|w_g^k\|_1}(w) = D_k \text{prox}_{\delta\rho_1\|w\|_1}(w). \tag{3.112}$$

It is hence possible for each agent k to run (3.29)–(3.31). As long as at least one agent in the network is either fully-informed or data-informed, the weighted sum of costs across the network is strongly convex and assumptions 3.1 through 3.3 are satisfied. We conclude from Theorem 3.3 that all agents in the network will converge to the neighborhood of:

$$\begin{aligned} w^o = \arg\min_w \sum_{k\in\mathcal{F}\cup\mathcal{D}} p_k \left\{\mathbb{E}\ \ln\left[1+e^{-\gamma_k \boldsymbol{h}_k^{\mathsf{T}} w}\right]\right\} \\ +\rho_2 \cdot \text{card}(\mathcal{F}\cup\mathcal{D})\|w\|_2^2 + \sum_{k\in\mathcal{F}\cup\mathcal{S}} p_k\|w_g^k\|_1 \end{aligned} \tag{3.113}$$

where $\text{card}(\mathcal{F}\cup\mathcal{D})$ denotes the cardinality of the set $\mathcal{F}\cup\mathcal{D}$, i.e. the number of agents who are either fully or data-informed. This classifier minimizes the weighted average logistic cost across the network, hence incorporating data from all agents, regularized by the ℓ_2-norm and weighted group Lasso. Through local interactions, both data and structural information is diffused across the entire network, allowing all agents, irrespective of their type and available information, to arrive at an accurate classification decision.

3.4.3 Numerical Results

Performance is illustrated on the network depicted in Fig. 3.1, consisting of a total of $N = 40$ agents, 20 of which are data-informed and 10 each of which are fully and structure informed respectively. The network is heterogeneous in both the types of available information and the noise profile of feature realizations, when data is available. Features are generated as

$$\boldsymbol{h}_{k,i} = \boldsymbol{\gamma}(i)\begin{pmatrix}1 & 1 & \cdots & 0 & 0\end{pmatrix}^{\mathsf{T}} + \boldsymbol{v}_k(i) \tag{3.114}$$

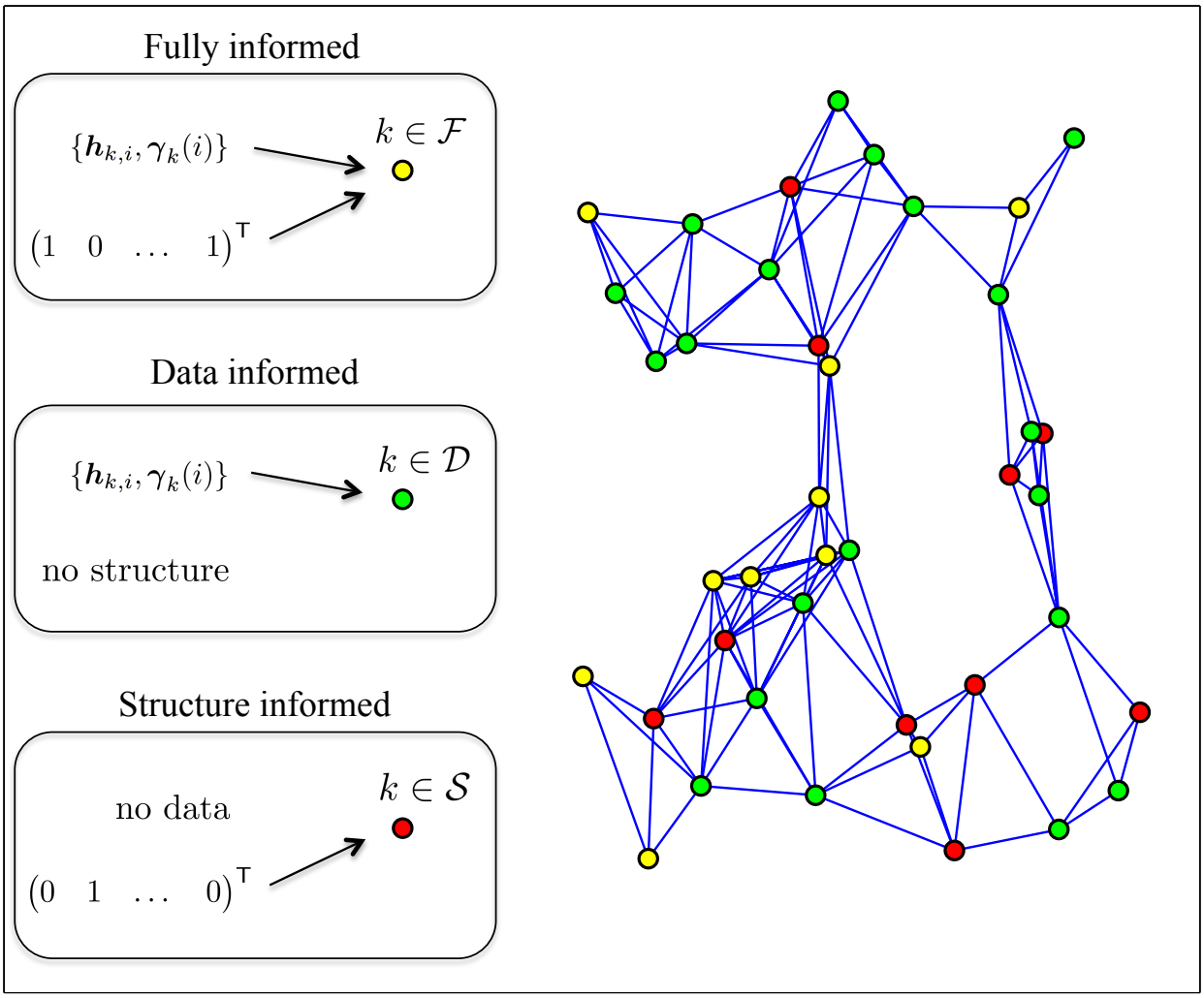

Figure 3.1: Sample network consisting of $N = 40$ agents, $\text{card}(\mathcal{F}) = 10$, $\text{card}(\mathcal{D}) = 20$, $\text{card}(\mathcal{S}) = 10$. Fully-informed agents have access to data as well as partial structural information. Data-informed agents observe realizations of the feature vector along with class-labels, but have no information on the structure of the classifier. Structure-informed agents do not have access to data, but do have partial information on sparse elements.

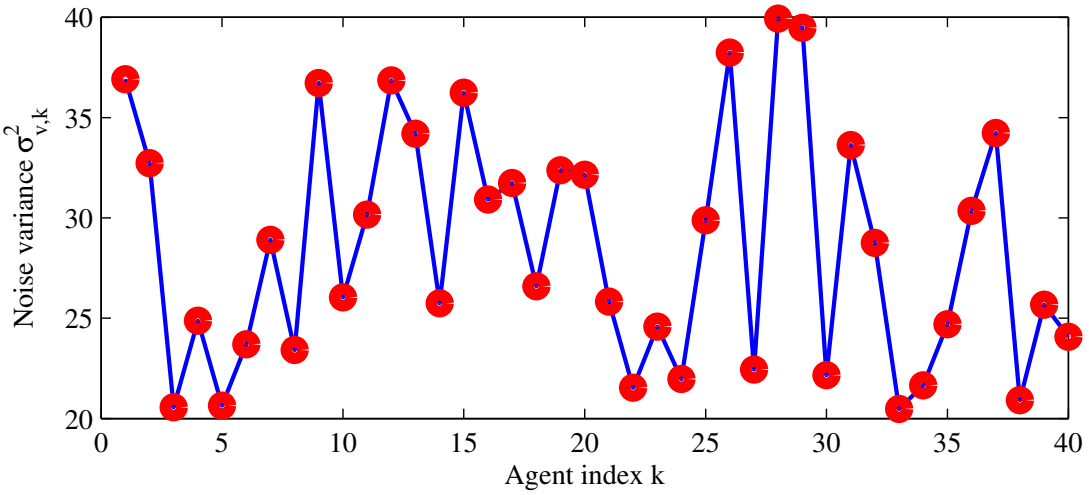

Figure 3.2: Noise profile across the network for training (if $k \in \mathcal{F} \cup \mathcal{D}$) and testing.

where $\boldsymbol{v}_k(i) \sim \mathcal{N}(0, \sigma^2_{v,k})$ and $\begin{pmatrix} 1 & 1 & \cdots & 0 & 0 \end{pmatrix}^{\mathsf{T}}$ consists of 50 leading 1's followed by 50 trailing 0's. It is evident, that all class information is contained in the first half of the feature vector. This information is dispersed across the network as follows. The noise profile across the network is depicted in Fig. 3.4.3.

Each agent with $k \in \mathcal{F} \cup \mathcal{S}$, i.e., fully and data-informed agents, are supplied with 5 indices, chosen uniformly at random, of irrelevant feature covariates. They use this information to augment their cost by an appropriate regularization as in (3.107) and (3.111).

3.A Proof of Lemma 3.1

Let $v \in \partial G(w)$. From the definition of the conjugate:

$$G^\star(v) = \sup_u \ \big(v^{\mathsf{T}} u - G(u)\big) \tag{3.115}$$

The optimality condition of the above supremum dictates that

$$0 \in v - \partial G(w) \iff w = \arg\max_u \ \big(v^{\mathsf{T}} u - G(u)\big) \tag{3.116}$$

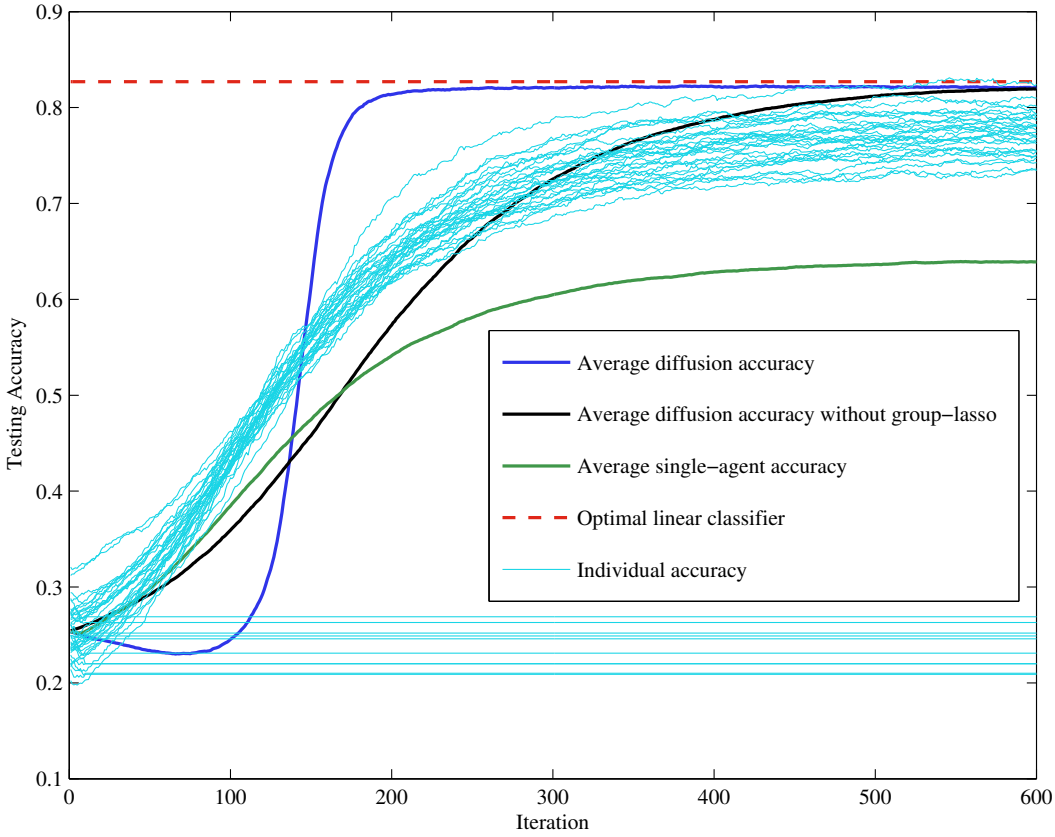

Figure 3.3: Classifier performance on separate testing set.

so for $v \in \partial G(w)$, the supremum (3.115) is attained at w. Then

$$G^\star(v) = v^\mathsf{T} w - G(w). \tag{3.117}$$

Now for any x (where the supremum might in general not be attained):

$$\begin{aligned} G^\star(x) &= \sup_u \left(x^\mathsf{T} u - G(u)\right) \\ &\geq x^\mathsf{T} w - G(w) \\ &= v^\mathsf{T} w - G(w) + w^\mathsf{T}(x - v) \\ &= G^\star(v) + w^\mathsf{T}(x - v) \end{aligned} \tag{3.118}$$

By definition, any vector that satisfies $G^\star(x) - G^\star(v) \geq w^\mathsf{T}(x - v)$ for all x is a subgradient of $G^\star(\cdot)$ at v, i.e., $w \in \partial G^\star(v)$. The other direction follows analogously, after noting that for closed, convex functions $(G^\star(\cdot))^\star = G(\cdot)$.

3.B Proof of Theorem 3.1

Let $u^o \in \partial(R_k^\star + \delta \cdot d)^\star(w) = \partial R_k^\delta(w)$. From Lemma 3.1, this is equivalent to

$$w \in \partial R_k^\star(u^o) + \delta \cdot d(u^o) \tag{3.119}$$

which due to optimality conditions is equivalent to

$$u^o = \operatorname*{arg\,max}_{u \in \operatorname{dom} R_k^\star} \left\{ w^\mathsf{T} u - R_k^\star(u) - \delta \cdot d(u) \right\}. \tag{3.120}$$

Since $R_k^\star(w) + \delta \cdot d(w)$ is strongly-convex, the minimizer u^o is unique and the above holds for any $u^o \in \partial R_k^\delta(w)$. We conclude that the set $\partial R_k^\delta(w)$ and hence

$$\left\{ \partial R_k^\delta(w) \right\} = \nabla R_k^\delta(w) = u^o. \tag{3.121}$$

To prove the bound on the gradient of the smooth approximation, let $u_1^o = \nabla R_k^\delta(w_1)$ and $u_2^o = \nabla R_k^\delta(w_2)$ for any w_1, w_2. From Lemma 3.1, this implies $w_1 \in \partial R_k^\star(u_1^o) + \delta \cdot \partial\, d(u_1^o)$ and $w_2 \in \partial R_k^\star(u_2^o) + \delta \cdot \partial\, d(u_2^o)$. From the strong-convexity of $\delta \cdot d(\cdot)$, we have:

$$(R_k^\star(u_1^o) + \delta \cdot \partial\, d(u_1^o) - \partial R_k^\star(u_2^o) + \delta \cdot \partial\, d(u_2^o))^{\mathsf{T}}(u_1^o - u_2^o) \geq \delta \|u_1^o - u_2^o\|^2 \tag{3.122}$$

Plugging in $w_1 \in \partial R_k^\star(u_1^o) + \delta \cdot \partial\, d(u_1^o)$ and $w_2 \in \partial R_k^\star(u_2^o) + \delta \cdot \partial\, d(u_2^o)$ as well as $u_1^o = \nabla R_k^\delta(w_1)$ and $u_2^o = \nabla R_k^\delta(w_2)$ yields

$$(w_1 - w_2)^{\mathsf{T}}(\nabla R_k^\delta(w_1) - \nabla R_k^\delta(w_2)) \geq \delta \|\nabla R_k^\delta(w_1) - \nabla R_k^\delta(w_2)\|^2 \tag{3.123}$$

which is the co-coercitivity property (3.16).

3.C Proof of Theorem 3.2

For ease of exposition, let us introduce

$$F(w) \triangleq \sum_{k=1}^{N} p_k \left\{ J_k(w) + R_k(w) \right\} \tag{3.124}$$

$$F^\delta(w) \triangleq \sum_{k=1}^{N} p_k \left\{ J_k(w) + R_k^\delta(w) \right\} \tag{3.125}$$

We establish a string of inequalities around the difference in function values $F(w^o) - F^\delta(w_\delta^o)$. On one hand, we have

$$
\begin{aligned}
& F(w^o) - F^\delta(w_\delta^o) \\
&= \sum_{k=1}^N p_k \left\{ J_k(w^o) + R_k(w^o) \right\} - \sum_{k=1}^N p_k \left\{ J_k(w_\delta^o) + R_k^\delta(w_\delta^o) \right\} \\
&= \sum_{k=1}^N p_k \left\{ J_k(w^o) - J_k(w_\delta^o) \right\} + \sum_{k=1}^N p_k \left\{ R_k(w^o) - R_k^\delta(w_\delta^o) \right\} \\
&\overset{(a)}{=} \sum_{k=1}^N p_k \left\{ J_k(w^o) - J_k(w_\delta^o) \right\} + \sum_{k=1}^N p_k \left\{ R_k(w^o) - \max_u \left(u^{\mathsf{T}} w_\delta^o - R_k^\star(u) - \delta d(u) \right) \right\} \\
&\overset{(b)}{=} \sum_{k=1}^N p_k \left\{ J_k(w^o) - J_k(w_\delta^o) \right\} \\
&\quad + \sum_{k=1}^N p_k \left\{ R_k(w^o) - \left(\nabla R_k^\delta(w_\delta^o) \right)^{\mathsf{T}} w_\delta^o + R_k^\star(\nabla R_k^\delta(w_\delta^o)) + \delta d \left(\nabla R_k^\delta(w_\delta^o) \right) \right\} \\
&\overset{(c)}{\geq} \sum_{k=1}^N p_k \left\{ J_k(w^o) - J_k(w_\delta^o) \right\} + \sum_{k=1}^N p_k \left\{ \nabla R_k^\delta(w_\delta^o)^{\mathsf{T}} w^o - \left(\nabla R_k^\delta(w_\delta^o) \right)^{\mathsf{T}} w_\delta^o + \delta d \left(\nabla R_k^\delta(w_\delta^o) \right) \right\} \\
&\overset{(d)}{\geq} \sum_{k=1}^N p_k \nabla J_k(w_\delta^o)^{\mathsf{T}} (w^o - w_\delta^o) + \frac{\lambda_L}{2} \|w^o - w_\delta^o\|^2 \\
&\quad + \sum_{k=1}^N p_k \left\{ \nabla R_k^\delta(w_\delta^o)^{\mathsf{T}} w^o - \left(\nabla R_k^\delta(w_\delta^o) \right)^{\mathsf{T}} w_\delta^o + \delta d \left(\nabla R_k^\delta(w_\delta^o) \right) \right\} \\
&= \sum_{k=1}^N p_k \big(\nabla J_k(w_\delta^o) + \nabla R_k^\delta(w_\delta^o) \big)^{\mathsf{T}} (w^o - w_\delta^o) + \frac{\lambda_L}{2} \|w^o - w_\delta^o\|^2 + \sum_{k=1}^N p_k \delta d \left(\nabla R_k^\delta(w_\delta^o) \right) \\
&\overset{(e)}{=} \frac{\lambda_L}{2} \|w^o - w_\delta^o\|^2 + \sum_{k=1}^N p_k \delta d \left(\nabla R_k^\delta(w_\delta^o) \right)
\end{aligned}
\tag{3.126}
$$

Here, (a) follows from the definition of the smooth approximation (3.13), (b) follows from the expression for the gradient of the smooth approximation (3.15), (c) follows from the property $R^\star(x) \triangleq \sup_u \left(u^{\mathsf{T}} x - R(u) \right) \geq y^{\mathsf{T}} x - R(y) \ \forall \ x, y$, (d) follows from the aggregate strong convexity (3.8) and (e) follows from the definition of w_δ^o and the minimizer of the smoothed aggregate cost.

To prove the upper bound, we bound the bias for each agent individually. To begin with,

note that convexity of $J_k(\cdot)$ and $R_k(\cdot)$ yields for all $r_k(w^o) \in \partial R_k(w^o)$:

$$J_k(w_\delta^o) - J_k(w^o) \geq (\nabla J_k(w^o))^{\mathsf{T}} (w_\delta^o - w^o) \iff J_k(w^o) - J_k(w_\delta^o) \leq (\nabla J_k(w^o))^{\mathsf{T}} (w^o - w_\delta^o) \tag{3.127}$$

$$R_k(u) - R_k(w^o) \geq (r_k(w^o))^{\mathsf{T}} (u - w^o) \tag{3.128}$$

Then,

$$\begin{aligned}
&J_k(w^o) + R_k(w^o) - J_k(w_\delta^o) - R_k^\delta(w_\delta^o) \\
&= J_k(w^o) + R_k(w^o) - J_k(w_\delta^o) - \min_u \left\{ R_k(u) + \delta d^\star \left(\frac{w_\delta^o - u}{\delta} \right) \right\} \\
&= J_k(w^o) - J_k(w_\delta^o) - \min_u \left\{ R_k(u) - R_k(w^o) + \delta d^\star \left(\frac{w_\delta^o - u}{\delta} \right) \right\} \\
&\leq (\nabla J_k(w^o))^{\mathsf{T}} (w^o - w_\delta^o) - \min_u \left\{ (r_k(w^o))^{\mathsf{T}} (u_k - w^o) + \delta d^\star \left(\frac{w_\delta^o - u}{\delta} \right) \right\} \\
&= (\nabla J_k(w^o) + r_k(w^o))^{\mathsf{T}} (w^o - w_\delta^o) - \min_u \left\{ (r_k(w^o))^{\mathsf{T}} (u - w_\delta^o) + \delta d^\star \left(\frac{w_\delta^o - u}{\delta} \right) \right\} \\
&\overset{(a)}{=} (\nabla J_k(w^o) + r_k(w^o))^{\mathsf{T}} (w^o - w_\delta^o) - \delta \min_v \left\{ -(r_k(w^o))^{\mathsf{T}} v + d^\star(v) \right\} \\
&= (\nabla J_k(w^o) + r_k(w^o))^{\mathsf{T}} (w^o - w_\delta^o) + \delta \max_v \left\{ (r_k(w^o))^{\mathsf{T}} v - d^\star(v) \right\} \\
&\overset{(b)}{=} (\nabla J_k(w^o) + r_k(w^o))^{\mathsf{T}} (w^o - w_\delta^o) + \delta d(r_k(w^o))
\end{aligned} \tag{3.129}$$

where (a) follows after a change of variables $v \triangleq \frac{w_\delta^o - u}{\delta}$ and (b) is a result of the definition of the conjugate function. Returning to the aggregate cost, we then have

$$\begin{aligned}
&\sum_{k=1}^N p_k \{ J_k(w^o) + R_k(w^o) \} - \sum_{k=1}^N p_k \left\{ J_k(w_\delta^o) + R_k^\delta(w_\delta^o) \right\} \\
&= \sum_{k=1}^N p_k \left\{ J_k(w^o) + R_k(w^o) - J_k(w_\delta^o) + R_k^\delta(w_\delta^o) \right\} \\
&\leq \sum_{k=1}^N p_k \left\{ (\nabla J_k(w^o) + r_k(w^o))^{\mathsf{T}} (w^o - w_\delta^o) \right\} + \sum_{k=1}^N p_k \delta d(r_k(w^o)) \\
&= \left\{ \sum_{k=1}^N p_k (\nabla J_k(w^o) + r_k(w^o)) \right\}^{\mathsf{T}} (w^o - w_\delta^o) + \sum_{k=1}^N p_k \delta d(r_k(w^o))
\end{aligned} \tag{3.130}$$

By definition, w^o is the minimizer of $\sum_{k=1}^{N} p_k \{J_k(w^o) + R_k(w^o)\}$, so there exist subgradients $r_k^o \in \partial R_k(w^o)$, such that

$$\sum_{k=1}^{N} p_k \left(\nabla J_k(w^o) + r_k^o\right) = 0 \tag{3.131}$$

Then,

$$\sum_{k=1}^{N} p_k \{J_k(w^o) + R_k(w^o)\} - \sum_{k=1}^{N} p_k \left\{J_k(w_\delta^o) + R_k^\delta(w_\delta^o)\right\} \leq \sum_{k=1}^{N} p_k \delta d\left(r_k^o\right) = O(\delta) \tag{3.132}$$

We conclude from (3.126):

$$\frac{\lambda_L}{2}\|w^o - w_\delta^o\|^2 + \sum_{k=1}^{N} p_k \delta d\left(\nabla R_k^\delta(w_\delta^o)\right) \leq F(w^o) - F^\delta(w_\delta^o) \leq \sum_{k=1}^{N} p_k \delta d\left(r_k^o\right) \tag{3.133}$$

The result follows after rearranging.

3.D Proof of Lemma 3.2

Let α be an arbitrary real number such that $0 < \alpha < 1$. Then

$$
\begin{aligned}
&\|T_c(x) - T_c(y)\|^2 \\
=&\Big\|x - y - \mu \sum_{k=1}^{N} p_k \Big\{\nabla J_k(x) - \nabla J_k(y) + \nabla R_k^\delta(x) - \nabla R_k^\delta(y)\Big\}\Big\|^2 \\
=&\|x-y\|^2 + \mu^2 \Big\| \sum_{k=1}^{N} p_k \Big\{\nabla J_k(x) - \nabla J_k(y) + \nabla R_k^\delta(x) - \nabla R_k^\delta(y)\Big\}\Big\|^2 \\
&- 2\mu \sum_{k=1}^{N} p_k (x-y)^{\mathsf{T}} \left(\nabla J_k(x) - \nabla J_k(y)\right) - 2\mu \sum_{k=1}^{N} p_k (x-y)^{\mathsf{T}} \left(\nabla R_k^\delta(x) - \nabla R_k^\delta(y)\right) \\
\overset{(a)}{\leq}&\|x-y\|^2 + \mu^2 \sum_{k=1}^{N} p_k \Big\|\nabla J_k(x) - \nabla J_k(y) + \nabla R_k^\delta(x) - \nabla R_k^\delta(y)\Big\|^2 \\
&- 2\mu\lambda_L \|x-y\|^2 - 2\mu\delta \sum_{k=1}^{N} p_k \left\|\nabla R_k^\delta(x) - \nabla R_k^\delta(y)\right\|^2 \\
\overset{(b)}{\leq}&\|x-y\|^2 + \mu^2 \sum_{k=1}^{N} p_k \frac{1}{\alpha} \Big\|\nabla J_k(x) - \nabla J_k(y)\Big\|^2 + \mu^2 \sum_{k=1}^{N} p_k \frac{1}{1-\alpha} \Big\|\nabla R_k^\delta(x) - \nabla R_k^\delta(y)\Big\|^2 \\
&- 2\mu\lambda_L \|x-y\|^2 - 2\mu\delta \sum_{k=1}^{N} p_k \left\|\nabla R_k^\delta(x) - \nabla R_k^\delta(y)\right\|^2
\end{aligned} \tag{3.134}
$$

where (a) follows from Jensen's inequality, strong convexity (3.8), and co-coercitivity (3.16), and (b) from $\|a+b\|^2 \leq \frac{1}{\alpha}\|a\|^2 + \frac{1}{1-\alpha}\|b\|^2$ for any $a, b \in \mathbb{R}^M$. Since, by assumption, $\mu < 2\delta$, we select $\alpha = 1 - \frac{\mu}{2\delta}$. This results in $\frac{\mu^2}{1-\alpha} = 2\mu\delta$ and allows us to cancel all terms involving

$\nabla_w R_k^\delta(\cdot)$ in the above inequality. Hence,

$$\begin{aligned}
&\|T_c(x) - T_c(y)\|^2 \\
\leq &\|x-y\|^2 + \mu^2 \sum_{k=1}^N p_k \frac{1}{1-\frac{\mu}{2\delta}} \left\|\nabla J_k(x) - \nabla J_k(y)\right\|^2 - 2\mu\lambda_L \|x-y\|^2 \\
\overset{(a)}{\leq} &\|x-y\|^2 + \frac{\mu^2\lambda_U^2}{1-\frac{\mu}{2\delta}}\|x-y\|^2 - 2\mu\lambda_L \|x-y\|^2 \\
= &\left(1 - 2\mu\lambda_L + \mu^2 \frac{\lambda_U^2}{1-\frac{\mu}{2\delta}}\right)\|x-y\|^2 \\
\overset{(b)}{\leq} &\left(1 - \mu\lambda_L + \mu^2 \frac{\lambda_U^2}{2-\frac{\mu}{\delta}}\right)^2 \|x-y\|^2
\end{aligned} \tag{3.135}$$

where (a) is due to the Lipschitz property (3.7) and (b) is due to $1-a \leq (1-\frac{1}{2}a)^2$ for all $a \in \mathbb{R}$. From Banach's fixed-point theorem, we know that as long as $\gamma_c < 1$, $w_i = T_c(w_{i-1})$ converges exponentially to a unique fixed point, which satisfies $w_\infty = T_c(w_\infty)$. From (3.40), we conclude that

$$\sum_{k=1}^N p_k \nabla J_k(w_\infty) + \sum_{k=1}^N p_k \nabla R_k^\delta(w_\infty) = 0 \tag{3.136}$$

so that from (3.6), $w_\infty = w_\delta^o$.

3.E Proof of Lemma 3.3

The proof of the first three inequalities relies on the Lipschitz properties of the gradients and the decomposition (3.53)–(3.54). First, we bound the terms arising from the disagreement across the network. Denote the k-th element of $P[\cdot]$ by $P_{(k)}[\cdot]$. Then

$$\begin{aligned}
&P_{(k)}[\boldsymbol{t}_{i-1}] \\
= &\ \mathbb{E}\,\|\nabla J_k(\boldsymbol{w}_{c,i-1}) - \nabla J_k(\boldsymbol{w}_{k,i-1}) \\
&\quad + \nabla R_k^\delta(\boldsymbol{w}_{c,i-1} - \mu\nabla J_k(\boldsymbol{w}_{c,i-1})) - \nabla R_k^\delta(\boldsymbol{w}_{k,i-1} - \mu\nabla J_k(\boldsymbol{w}_{k,i-1}))\|^2 \\
\overset{(a)}{\leq} &\ 2\,\mathbb{E}\,\|\nabla J_k(\boldsymbol{w}_{c,i-1}) - \nabla J_k(\boldsymbol{w}_{k,i-1})\|^2 \\
&\quad + 2\,\mathbb{E}\,\|\nabla R_k^\delta(\boldsymbol{w}_{c,i-1} - \mu\nabla J_k(\boldsymbol{w}_{c,i-1})) - \nabla R_k^\delta(\boldsymbol{w}_{k,i-1} - \mu\nabla J_k(\boldsymbol{w}_{k,i-1}))\|^2
\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{\leq} 2\lambda_U^2\,\mathbb{E}\,\|\,\boldsymbol{w}_{c,i-1}-\boldsymbol{w}_{k,i-1}\,\|^2+\frac{2}{\delta^2}\,\mathbb{E}\,\|\,\boldsymbol{w}_{c,i-1}-\mu\nabla J_k(\boldsymbol{w}_{c,i-1})-\boldsymbol{w}_{k,i-1}+\mu\nabla J_k(\boldsymbol{w}_{k,i-1})\|^2\\
&\overset{(c)}{\leq}\left(2\lambda_U^2+4\frac{1+\mu^2}{\delta^2}\right)\mathbb{E}\,\|\,\boldsymbol{w}_{c,i-1}-\boldsymbol{w}_{k,i-1}\,\|^2\\
&\overset{(d)}{=}\left(2\lambda_U^2+4\frac{1+\mu^2}{\delta^2}\right)\mathbb{E}\,\|\,(v_{L,k}\otimes I_M)\,\boldsymbol{\mathcal{W}}_{e,i-1}\,\|^2\\
&\leq\left(2\lambda_U^2+4\frac{1+\mu^2}{\delta^2}\right)\nu^2\,\mathbb{E}\,\|\,\boldsymbol{\mathcal{W}}_{e,i-1}\,\|^2\\
&=\left(2\lambda_U^2+4\frac{1+\mu^2}{\delta^2}\right)\nu^2\mathbb{1}^{\mathsf{T}}P[\boldsymbol{\mathcal{W}}_{e,i-1}]
\end{aligned}
\tag{3.137}
$$

where (a) is due Jensen's inequality, (b) and (c) are due to Lipschitz continuity of the gradients and (d) is due to $\boldsymbol{\mathcal{W}}_i=\mathbb{1}\otimes\boldsymbol{w}_{c,i}+\mathcal{V}_L\,\boldsymbol{\mathcal{W}}_{e,i}$. Stacking both sides of the above inequality yields (3.67).

Now consider $\boldsymbol{u}_{i-1}$, which arises from the incremental implementation:

$$
\begin{aligned}
&P_{(k)}[\boldsymbol{u}_{i-1}]\\
&=\mathbb{E}\,\|\nabla R_k^{\delta}(\boldsymbol{w}_{k,i-1})-\nabla R_k^{\delta}(\boldsymbol{w}_{k,i-1}-\mu\nabla J_k(\boldsymbol{w}_{k,i-1}))\|^2\\
&\overset{(a)}{\leq}\frac{\mu^2}{\delta^2}\,\mathbb{E}\,\|\nabla J_k(\boldsymbol{w}_{k,i-1})\|^2\\
&=\frac{\mu^2}{\delta^2}\,\mathbb{E}\,\|\nabla J_k(\boldsymbol{w}_{k,i-1})-\nabla J_k(\boldsymbol{w}_{c,i-1})+\nabla J_k(\boldsymbol{w}_{c,i-1})-\nabla J_k(w_\delta^o)+\nabla J_k(w_\delta^o)\|^2\\
&\overset{(b)}{\leq}\frac{\mu^2}{\delta^2}\big(3\lambda_U^2\nu^2\mathbb{1}^{\mathsf{T}}P[\boldsymbol{\mathcal{W}}_{e,i-1}]+3\lambda_U^2\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2+3\|\nabla J_k(w_\delta^o)\|^2\big)
\end{aligned}
\tag{3.138}
$$

where (a) is due to Lipschitz continuity of $\nabla R_k^{\delta}(w)$ and (b) is due to Jensen's inequality and Lipschitz continuity of $\nabla J_k(w)$. Upon stacking we obtain (3.68).

Next, we bound the perturbations caused by the gradient noise $\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i})=\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1})-\nabla J_k(\boldsymbol{w}_{k,i-1})$. While a loose upper bound can be obtained immediately from Jensen's inequality, it turns out that the incremental implementation (3.30) along with the co-coercivity (3.16) of $\nabla R_k^{\delta}(w)$ have a variance reducing effect on the recursion:

$$
\begin{aligned}
&P_{(k)}[\boldsymbol{s}_i^g+\boldsymbol{s}_i^p-\mathbb{E}\,\boldsymbol{s}_i^p]\\
&\overset{(a)}{\leq}P_{(k)}[\boldsymbol{s}_i^g+\boldsymbol{s}_i^p]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\left\|\nabla J_k(\boldsymbol{w}_{k,i-1}) - \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1})\right\|^2 \\
&\quad + \mathbb{E}\left\|\nabla R_k^{\delta}(\boldsymbol{w}_{k,i-1} - \mu\nabla J_k(\boldsymbol{w}_{k,i-1})) - \nabla R_k^{\delta}(\boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}))\right\|^2 \\
&\quad + 2\,\mathbb{E}\left(\nabla J_k(\boldsymbol{w}_{k,i-1}) - \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1})\right)^{\mathsf{T}} \\
&\qquad \times \left(\nabla R_k^{\delta}(\boldsymbol{w}_{k,i-1} - \mu\nabla J_k(\boldsymbol{w}_{k,i-1})) - \nabla R_k^{\delta}(\boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}))\right) \\
&= \mathbb{E}\left\|\nabla J_k(\boldsymbol{w}_{k,i-1}) - \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1})\right\|^2 \\
&\quad + \mathbb{E}\left\|R_k^{\delta}(\boldsymbol{w}_{k,i-1} - \mu\nabla J_k(\boldsymbol{w}_{k,i-1})) - \nabla R_k^{\delta}(\boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}))\right\|^2 \\
&\quad - \frac{2}{\mu}\,\mathbb{E}\left(\boldsymbol{w}_{k,i-1} - \mu\nabla J_k(\boldsymbol{w}_{k,i-1}) - \left(\boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1})\right)\right)^{\mathsf{T}} \\
&\quad \times \left(\nabla R_k^{\delta}(\boldsymbol{w}_{k,i-1} - \mu\nabla J_k(\boldsymbol{w}_{k,i-1})) - \nabla R_k^{\delta}(\boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}))\right) \\
&\overset{(b)}{\leq} \mathbb{E}\left\|\nabla J_k(\boldsymbol{w}_{k,i-1}) - \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1})\right\|^2 \\
&\quad + \mathbb{E}\left\|R_k^{\delta}(\boldsymbol{w}_{k,i-1} - \mu\nabla J_k(\boldsymbol{w}_{k,i-1})) - \nabla R_k^{\delta}(\boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}))\right\|^2 \\
&\quad - \frac{2\delta}{\mu}\,\mathbb{E}\left\|\nabla R_k^{\delta}(\boldsymbol{w}_{k,i-1} - \mu\nabla J_k(\boldsymbol{w}_{k,i-1})) - \nabla R_k^{\delta}(\boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}))\right\|^2 \\
&= \mathbb{E}\left\|\nabla J_k(\boldsymbol{w}_{k,i-1}) - \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1})\right\|^2 \\
&\quad - \left(\frac{2\delta}{\mu} - 1\right)\mathbb{E}\left\|R_k^{\delta}(\boldsymbol{w}_{k,i-1} - \mu\nabla J_k(\boldsymbol{w}_{k,i-1})) - \nabla R_k^{\delta}(\boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}))\right\|^2 \\
&\overset{(c)}{\leq} \mathbb{E}\left\|\nabla J_k(\boldsymbol{w}_{k,i-1}) - \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1})\right\|^2 \\
&= \mathbb{E}\,\|\,\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2 \\
&\overset{(d)}{\leq} \beta^2\,\mathbb{E}\,\|\,\boldsymbol{w}_{k,i-1}\,\|^2 + \sigma^2
\end{aligned}
\tag{3.139}
$$

where (a) follows from $\mathbb{E}\,\|\boldsymbol{x} - \mathbb{E}\,\boldsymbol{x}\|^2 \leq \mathbb{E}\,\|\boldsymbol{x}\|^2$ for any $\boldsymbol{x}$, (b) follows from co-coercitivity (3.16), (c) follows from $\mu < 2\delta$ and (d) is due to (3.65b). Now from $\boldsymbol{w}_{k,i-1} = \boldsymbol{w}_{c,i-1} + (v_{L,k} \otimes I)\,\boldsymbol{\mathcal{W}}_{e,i-1}$, we can bound

$$
\begin{aligned}
\|\,\boldsymbol{w}_{k,i-1}\,\|^2 &= \|\,\boldsymbol{w}_{c,i-1} + (v_{L,k} \otimes I)\,\boldsymbol{\mathcal{W}}_{e,i-1}\,\|^2 \\
&= \|\,\boldsymbol{w}_{c,i-1} - w_{\delta}^{o} + (v_{L,k} \otimes I)\,\boldsymbol{\mathcal{W}}_{e,i-1} + w_{\delta}^{o}\|^2 \\
&\leq 3\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + 3\nu^2\mathbb{1}^{\mathsf{T}}P[\boldsymbol{\mathcal{W}}_{e,i-1}] + 3\|w_{\delta}^{o}\|^2.
\end{aligned}
\tag{3.140}
$$

where we appealed to Jensen's inequality again. Eq. (3.69) follows after stacking. Next, note that because $\|\mathbb{E}\,\boldsymbol{x}\|^2 \leq \mathbb{E}\,\|\boldsymbol{x}\|^2$

$$P[\mathbb{E}\,\boldsymbol{s}_i^p] \preceq P[\boldsymbol{s}_i^p]. \tag{3.141}$$

Subsequently,

$$\begin{aligned} P_{(k)}[\boldsymbol{s}_i^p] &= \mathbb{E}\,\|\nabla R_k^\delta(\boldsymbol{w}_{k,i-1} - \mu\nabla J_k(\boldsymbol{w}_{k,i-1})) - \nabla R_k^\delta(\boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}))\|^2 \\ &\overset{(a)}{\leq} \frac{\mu^2}{\delta^2}\,\mathbb{E}\,\|\nabla J_k(\boldsymbol{w}_{k,i-1}) - \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1})\|^2 \\ &= \frac{\mu^2}{\delta^2}\|\,\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2 \end{aligned} \tag{3.142}$$

where (a) is due to (3.17), so that similarly to the above

$$\begin{aligned} P[\mathbb{E}\,\boldsymbol{s}_i^p] \preceq\; & 3\beta^2\frac{\mu^2}{\delta^2}P[\mathbb{1}\otimes\widetilde{\boldsymbol{w}}_{c,i-1}] + 3\beta^2\frac{\mu^2}{\delta^2}\nu^2\mathbb{1}\mathbb{1}^{\mathsf{T}}P[\boldsymbol{\mathcal{W}}_{e,i-1}] \\ & + 3\beta^2\frac{\mu^2}{\delta^2}P[\mathbb{1}\otimes w_\delta^o] + \frac{\mu^2}{\delta^2}\sigma^2\mathbb{1} \end{aligned} \tag{3.143}$$

which is (3.70). Next,

$$\begin{aligned} P_{(k)}[g(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})] &= \mathbb{E}\,\|\nabla J_k(\boldsymbol{w}_{c,i-1})\|^2 \\ &= \mathbb{E}\,\|\nabla J_k(\boldsymbol{w}_{c,i-1}) - \nabla J_k(w_\delta^o) + \nabla J_k(w_\delta^o)\|^2 \\ &\leq 2\lambda_U^2\,\mathbb{E}\,\|\,\boldsymbol{w}_{c,i-1} - w_\delta^o\|^2 + 2\|\nabla J_k(w_\delta^o)\|^2 \end{aligned} \tag{3.144}$$

which implies (3.71) after stacking. Eq. (3.72) follows analogously.

3.F Proof of Lemma 3.4

We make use of Jensen's inequality $\|x+y\|^2 \leq \frac{1}{\alpha}\|x\|^2 + \frac{1}{1-\alpha}\|y\|^2$ for all x, y and $0 < \alpha < 1$:

$$\begin{aligned} &\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i}\|^2 \\ =\;& \mathbb{E}\left\|T_c(\boldsymbol{w}_{c,i-1}) - T_c(w_\delta^o) + \mu\left(p^{\mathsf{T}}\otimes I_M\right)\left(\boldsymbol{t}_{i-1} + \boldsymbol{u}_{i-1} + \boldsymbol{s}_i - \mathbb{E}\,\boldsymbol{s}_i + \mathbb{E}\,\boldsymbol{s}_i\right)\right\|^2 \end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{=} \mathbb{E}\left\|T_c(\boldsymbol{w}_{c,i-1}) - T_c(w_\delta^o) + \mu\left(p^\mathsf{T}\otimes I_M\right)\left(\boldsymbol{t}_{i-1}+\boldsymbol{u}_{i-1}+\mathbb{E}\,\boldsymbol{s}_i\right)\right\|^2 \\
&\quad + \mu^2\,\mathbb{E}\left\|\left(p^\mathsf{T}\otimes I_M\right)\left(\boldsymbol{s}_i-\mathbb{E}\,\boldsymbol{s}_i\right)\right\|^2 \\
&\overset{(b)}{\le} \frac{1}{\gamma_c}\,\mathbb{E}\left\|T_c(\boldsymbol{w}_{c,i-1}) - T_c(w_\delta^o)\right\|^2 + \frac{\mu^2}{1-\gamma_c}\,\mathbb{E}\left\|\left(p^\mathsf{T}\otimes I_M\right)\left(\boldsymbol{t}_{i-1}+\boldsymbol{u}_{i-1}+\mathbb{E}\,\boldsymbol{s}_i\right)\right\|^2 \\
&\quad + \mu^2\,\mathbb{E}\left\|\left(p^\mathsf{T}\otimes I_M\right)\left(\boldsymbol{s}_i-\mathbb{E}\,\boldsymbol{s}_i\right)\right\|^2 \\
&\overset{(c)}{\le} \gamma_c\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + \frac{\mu^2}{1-\gamma_c}\,\mathbb{E}\left\|\left(p^\mathsf{T}\otimes I_M\right)\left(\boldsymbol{t}_{i-1}+\boldsymbol{u}_{i-1}+\mathbb{E}\,\boldsymbol{s}_i\right)\right\|^2 \\
&\quad + \mu^2\,\mathbb{E}\left\|\left(p^\mathsf{T}\otimes I_M\right)\left(\boldsymbol{s}_i-\mathbb{E}\,\boldsymbol{s}_i\right)\right\|^2 \\
&\overset{(d)}{\le} \gamma_c\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + \frac{\mu^2}{1-\gamma_c}p^\mathsf{T}P[\boldsymbol{t}_{i-1}+\boldsymbol{u}_{i-1}+\mathbb{E}\,\boldsymbol{s}_i] + \mu^2 p^\mathsf{T}P[\boldsymbol{s}_i-\mathbb{E}\,\boldsymbol{s}_i] \\
&\overset{(e)}{\le} \gamma_c\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + \frac{3\mu^2}{1-\gamma_c}p^\mathsf{T}\Big(P[\boldsymbol{t}_{i-1}] + P[\boldsymbol{u}_{i-1}] + P[\mathbb{E}\,\boldsymbol{s}_i]\Big) + \mu^2 p^\mathsf{T}P[\boldsymbol{s}_i-\mathbb{E}\,\boldsymbol{s}_i] \\
&\overset{(f)}{\le} \gamma_c\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + \frac{3\mu^2}{1-\gamma_c}p^\mathsf{T}\Big(\Big(2\lambda_U^2+\frac{1+\mu^2}{\delta^2}\Big)\nu^2\mathbb{1}\mathbb{1}^\mathsf{T}P[\boldsymbol{\mathcal{W}}_{e,i-1}] \\
&\quad + \frac{\mu^2}{\delta^2}\big(3\lambda_U^2\nu^2\mathbb{1}\mathbb{1}^\mathsf{T}P[\boldsymbol{\mathcal{W}}_{e,i-1}] + 3\lambda_U^2 P[\mathbb{1}\otimes\widetilde{\boldsymbol{w}}_{c,i-1}] + 3P[g(\mathbb{1}\otimes w_\delta^o)]\big) \\
&\quad + 3\beta^2\frac{\mu^2}{\delta^2}P[\mathbb{1}\otimes\widetilde{\boldsymbol{w}}_{c,i-1}] + 3\beta^2\frac{\mu^2}{\delta^2}\nu^2\mathbb{1}\mathbb{1}^\mathsf{T}P[\boldsymbol{\mathcal{W}}_{e,i-1}] + 3\beta^2\frac{\mu^2}{\delta^2}P[\mathbb{1}\otimes w_\delta^o] + \frac{\mu^2}{\delta^2}\sigma^2\mathbb{1}\Big) \\
&\quad + \mu^2 p^\mathsf{T}\Big(3\beta^2 P[\mathbb{1}\otimes\widetilde{\boldsymbol{w}}_{c,i-1}] + 3\beta^2\nu^2\mathbb{1}\mathbb{1}^\mathsf{T}P[\boldsymbol{\mathcal{W}}_{e,i-1}] + 3\beta^2 P[\mathbb{1}\otimes w_\delta^o] + \sigma^2\mathbb{1}\Big) \\
&\overset{(g)}{=} \gamma_c\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + \frac{3\mu^2}{1-\gamma_c}\Big(\Big(2\lambda_U^2+\frac{1+\mu^2}{\delta^2}\Big)\nu^2\,\mathbb{E}\,\|\,\boldsymbol{\mathcal{W}}_{e,i-1}\,\|^2 \\
&\quad + \frac{\mu^2}{\delta^2}\big(3\lambda_U^2\nu^2\,\mathbb{E}\,\|\,\boldsymbol{\mathcal{W}}_{e,i-1}\,\|^2 + 3\lambda_U^2\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + 3\|g(w_\delta^o)\|^2\big) \\
&\quad + 3\beta^2\frac{\mu^2}{\delta^2}\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + 3\beta^2\frac{\mu^2}{\delta^2}\nu^2\,\mathbb{E}\,\|\,\boldsymbol{\mathcal{W}}_{e,i-1}\,\|^2 + 3\beta^2\frac{\mu^2}{\delta^2}\|w_\delta^o\|^2 + \frac{\mu^2}{\delta^2}\sigma^2\Big) \\
&\quad + \mu^2\Big(3\beta^2\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + 3\beta^2\nu^2\,\mathbb{E}\,\|\,\boldsymbol{\mathcal{W}}_{e,i-1}\,\|^2 + 3\beta^2\|w_\delta^o\|^2 + \sigma^2\Big) \\
&\overset{(h)}{=} \left(\gamma_c + \frac{9\mu^4(\beta^2+\lambda_U^2)}{(1-\gamma_c)\delta^2} + 3\mu^2\beta^2\right)\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 \\
&\quad + \left(\frac{3\mu^2\nu^2}{1-\gamma_c}\left(2\lambda_U^2 + \frac{1+\mu^2+3\mu^2\lambda_U^2+3\mu^2\beta^2}{\delta^2}\right) + 3\mu^2\beta^2\nu^2\right)\mathbb{E}\,\|\,\boldsymbol{\mathcal{W}}_{e,i-1}\,\|^2 \\
&\quad + \frac{3\mu^4}{(1-\gamma_c)\delta^2}\left(3\|g(w_\delta^o)\|^2 + 3\beta^2\|w_\delta^o\|^2 + \sigma^2\right) + \mu^2\left(3\beta^2\|w_\delta^o\|^2 + \sigma^2\right) \\
&\overset{(i)}{=} \left(\gamma_c + \frac{\mu^3}{\delta^2}\frac{9(\beta^2+\lambda_U^2)}{\lambda_L - \mu\frac{\lambda_U^2}{2-\frac{\mu}{\delta}}} + 3\mu^2\beta^2\right)\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2
\end{aligned}
$$

$$+\left(\frac{\mu}{\delta^2}\frac{3\nu^2}{\lambda_L-\mu\frac{\lambda_U^2}{2-\frac{\mu}{\delta}}}+\mu\frac{6\nu^2}{\lambda_L-\mu\frac{\lambda_U^2}{2-\frac{\mu}{\delta}}}\lambda_U^2+\frac{\mu^3}{\delta^2}\frac{9\nu^2}{\lambda_L-\mu\frac{\lambda_U^2}{2-\frac{\mu}{\delta}}}(\lambda_U^2+\beta^2)+3\mu^2\beta^2\nu^2\right)\mathbb{E}\,\|\,\boldsymbol{\mathcal{w}}_{e,i-1}\,\|^2$$
$$+\frac{\mu^3}{\delta^2}\frac{9}{\lambda_L-\mu\frac{\lambda_U^2}{2-\frac{\mu}{\delta}}}\|g(w_\delta^o)\|^2+\frac{\mu^3}{\delta^2}\frac{3}{\lambda_L-\mu\frac{\lambda_U^2}{2-\frac{\mu}{\delta}}}\left(3\beta^2\|w_\delta^o\|^2+\sigma^2\right)+\mu^2\left(3\beta^2\|w_\delta^o\|^2+\sigma^2\right) \tag{3.145}$$

In step (a), cross-terms are eliminated because $\mathbb{E}\,\{\boldsymbol{s}_i-\mathbb{E}\,\boldsymbol{s}_i\}=0$. Step (b) is due to $\gamma_c<1$ and Jensen's inequality, (c) is due to Lemma 3.2, (d) and (e) follow from Jensen's inequality. The bounds from Lemma 3.3 are used in (f) and (g) is due to $\mathbb{1}^\mathsf{T}P[\boldsymbol{x}]=\mathbb{E}\,\|\boldsymbol{x}\|^2$ for $\boldsymbol{x}\in\mathbb{R}^{MN}$ and $p^\mathsf{T}P[\mathbb{1}\otimes\boldsymbol{y}]=\mathbb{E}\,\|\boldsymbol{y}\|^2$ for $\boldsymbol{y}\in\mathbb{R}^M$. In (i), the terms are rearranged to expose the dependence on μ and δ more clearly.

Now let us turn to the mean-square recursion of $\boldsymbol{\mathcal{w}}_{e,i}$. First note that $\rho(\mathcal{J}_\epsilon)=\lambda_2(A)<1$. Since $\mathcal{J}_\epsilon$ has a Jordan structure, this means that we can chose ϵ small enough, such that $\|\mathcal{J}_\epsilon\|_2=\rho(\mathcal{J}_\epsilon^\mathsf{T}\mathcal{J}_\epsilon)\le\|\mathcal{J}_\epsilon^\mathsf{T}\mathcal{J}_\epsilon\|_\infty<1$. Then,

$$\begin{aligned}
&\mathbb{E}\,\|\,\boldsymbol{\mathcal{w}}_{e,i}\,\|^2\\
&=\mathbb{E}\left\|\mathcal{J}_\epsilon^\mathsf{T}\,\boldsymbol{\mathcal{w}}_{e,i-1}+\mu\mathcal{J}_\epsilon^\mathsf{T}\mathcal{V}_R^\mathsf{T}\big(\boldsymbol{t}_{i-1}+\boldsymbol{u}_{i-1}+\boldsymbol{s}_i-\mathbb{E}\,\boldsymbol{s}_i+\mathbb{E}\,\boldsymbol{s}_i-g(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})-r(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})\big)\right\|^2\\
&\overset{(a)}{=}\mathbb{E}\left\|\mathcal{J}_\epsilon^\mathsf{T}\,\boldsymbol{\mathcal{w}}_{e,i-1}+\mu\mathcal{J}_\epsilon^\mathsf{T}\mathcal{V}_R^\mathsf{T}\big(\boldsymbol{t}_{i-1}+\boldsymbol{u}_{i-1}+\mathbb{E}\,\boldsymbol{s}_i-g(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})-r(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})\big)\right\|^2\\
&+\mu^2\,\mathbb{E}\left\|\mathcal{J}_\epsilon^\mathsf{T}\mathcal{V}_R^\mathsf{T}\,(\boldsymbol{s}_i-\mathbb{E}\,\boldsymbol{s}_i)\right\|^2\\
&\overset{(b)}{\le}\frac{1}{\|\mathcal{J}_\epsilon\|}\;\mathbb{E}\left\|\mathcal{J}_\epsilon^\mathsf{T}\,\boldsymbol{\mathcal{w}}_{e,i-1}\right\|^2+\frac{\mu^2}{1-\|\mathcal{J}_\epsilon\|}\,\mathbb{E}\left\|\mathcal{J}_\epsilon^\mathsf{T}\mathcal{V}_R^\mathsf{T}\big(\boldsymbol{t}_{i-1}+\boldsymbol{u}_{i-1}+\mathbb{E}\,\boldsymbol{s}_i\right.\\
&\left.-g(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})-r(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})\big)\right\|^2+\mu^2\,\mathbb{E}\left\|\mathcal{J}_\epsilon^\mathsf{T}\mathcal{V}_R^\mathsf{T}\,(\boldsymbol{s}_i-\mathbb{E}\,\boldsymbol{s}_i)\right\|^2\\
&\overset{(c)}{\le}\|\mathcal{J}_\epsilon\|\;\mathbb{E}\,\|\boldsymbol{\mathcal{w}}_{e,i-1}\|^2+\frac{\mu^2\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2}{1-\|\mathcal{J}_\epsilon\|}\,\mathbb{E}\left\|\boldsymbol{t}_{i-1}+\boldsymbol{u}_{i-1}+\mathbb{E}\,\boldsymbol{s}_i\right.\\
&\left.-g(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})-r(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})\right\|^2+\mu^2\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2\,\mathbb{E}\,\|\boldsymbol{s}_i-\mathbb{E}\,\boldsymbol{s}_i\|^2\\
&\overset{(d)}{\le}\|\mathcal{J}_\epsilon\|\;\mathbb{E}\,\|\boldsymbol{\mathcal{w}}_{e,i-1}\|^2+\frac{25\mu^2\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2}{1-\|\mathcal{J}_\epsilon\|}\Big(\,\mathbb{E}\,\|\boldsymbol{t}_{i-1}\|^2+\mathbb{E}\,\|\boldsymbol{u}_{i-1}\|^2\\
&+\mathbb{E}\,\|\mathbb{E}\,\boldsymbol{s}_i^p\|^2+\mathbb{E}\,\|g(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})\|^2+\mathbb{E}\,\|r(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})\|^2\Big)+\mu^2\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2\,\mathbb{E}\,\|\boldsymbol{s}_i-\mathbb{E}\,\boldsymbol{s}_i\|^2\\
&\overset{(e)}{=}\|\mathcal{J}_\epsilon\|\;\mathbb{E}\,\|\boldsymbol{\mathcal{w}}_{e,i-1}\|^2+\frac{25\mu^2\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2}{1-\|\mathcal{J}_\epsilon\|}\mathbb{1}^\mathsf{T}\Big(P[\boldsymbol{t}_{i-1}]+P[\boldsymbol{u}_{i-1}]+P[\mathbb{E}\,\boldsymbol{s}_i]
\end{aligned}$$

$$
\begin{aligned}
&\quad + P[g(\mathbb{1}\otimes \boldsymbol{w}_{c,i-1})] + P[r(\mathbb{1}\otimes \boldsymbol{w}_{c,i-1})]\Big) + \mu^2\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2\mathbb{1}^\mathsf{T}P[\boldsymbol{s}_i - \mathbb{E}\,\boldsymbol{s}_i]\\
&\overset{(f)}{\leq} \|\mathcal{J}_\epsilon\|\ \mathbb{E}\,\|\boldsymbol{\mathcal{w}}_{e,i-1}\|^2 + \frac{25\mu^2\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2}{1-\|\mathcal{J}_\epsilon\|}\mathbb{1}^\mathsf{T}\Big(\Big(2\lambda_U^2 + \frac{1+\mu^2}{\delta^2}\Big)\nu^2\mathbb{1}\mathbb{1}^\mathsf{T}P[\boldsymbol{\mathcal{w}}_{e,i-1}]\\
&+\frac{\mu^2}{\delta^2}\big(3\lambda_U^2\nu^2\mathbb{1}\mathbb{1}^\mathsf{T}P[\boldsymbol{\mathcal{w}}_{e,i-1}] + 3\lambda_U^2 P[\mathbb{1}\otimes\widetilde{\boldsymbol{w}}_{c,i-1}] + 3P[g(\mathbb{1}\otimes w_\delta^o)]\big)\\
&+3\beta^2\frac{\mu^2}{\delta^2}P[\mathbb{1}\otimes\widetilde{\boldsymbol{w}}_{c,i-1}] + 3\beta^2\frac{\mu^2}{\delta^2}\nu^2\mathbb{1}\mathbb{1}^\mathsf{T}P[\boldsymbol{\mathcal{w}}_{e,i-1}] + 3\beta^2\frac{\mu^2}{\delta^2}P[\mathbb{1}\otimes w_\delta^o] + \frac{\mu^2}{\delta^2}\sigma^2\mathbb{1}\\
&+2\lambda_U^2 P[\mathbb{1}\otimes\widetilde{\boldsymbol{w}}_{c,i-1}] + 2P[g(\mathbb{1}\otimes w_\delta^o)] + \frac{2}{\delta^2}P[\mathbb{1}\otimes\widetilde{\boldsymbol{w}}_{c,i-1}] + 2P[r(\mathbb{1}\otimes w_\delta^o)]\Big)\\
&+\mu^2\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2\mathbb{1}^\mathsf{T}\Big(3\beta^2 P[\mathbb{1}\otimes\widetilde{\boldsymbol{w}}_{c,i-1}] + 3\beta^2\nu^2\mathbb{1}\mathbb{1}^\mathsf{T}P[\boldsymbol{\mathcal{w}}_{e,i-1}] + 3\beta^2 P[\mathbb{1}\otimes w_\delta^o] + \sigma^2\mathbb{1}\Big)\\
&\overset{(g)}{=} \|\mathcal{J}_\epsilon\|\ \mathbb{E}\,\|\boldsymbol{\mathcal{w}}_{e,i-1}\|^2 + \frac{25\mu^2\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2}{1-\|\mathcal{J}_\epsilon\|}\Big(\Big(2\lambda_U^2 + \frac{1+\mu^2}{\delta^2}\Big)\nu^2 N\,\mathbb{E}\,\|\,\boldsymbol{\mathcal{w}}_{e,i-1}\,\|^2\\
&+\frac{\mu^2}{\delta^2}\big(3\lambda_U^2\nu^2 N\,\mathbb{E}\,\|\,\boldsymbol{\mathcal{w}}_{e,i-1}\,\|^2 + 3\lambda_U^2 N\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + 3\|g(\mathbb{1}\otimes w_\delta^o)\|^2\big)\\
&+3\beta^2\frac{\mu^2}{\delta^2}N\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + 3\beta^2\frac{\mu^2}{\delta^2}\nu^2 N\,\mathbb{E}\,\|\,\boldsymbol{\mathcal{w}}_{e,i-1}\,\|^2 + 3\beta^2\frac{\mu^2}{\delta^2}N\|w_\delta^o\|^2 + \frac{\mu^2}{\delta^2}N\sigma^2\\
&+2\lambda_U^2 N\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + 2\|g(\mathbb{1}\otimes w_\delta^o)\|^2 + \frac{2}{\delta^2}N\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + 2\|r(\mathbb{1}\otimes w_\delta^o)\|^2\Big)\\
&+\mu^2\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2\Big(3\beta^2 N\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2 + 3\beta^2\nu^2 N\,\mathbb{E}\,\|\,\boldsymbol{\mathcal{w}}_{e,i-1}\,\|^2 + 3\beta^2 N\|w_\delta^o\|^2 + N\sigma^2\Big)\\
&= \left(\|\mathcal{J}_\epsilon\| + \mu^2\nu^2 N\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2\Big(\frac{25}{1-\|\mathcal{J}_\epsilon\|}\left(2\lambda_U^2 + \frac{1+\mu^2}{\delta^2} + 3\frac{\mu^2}{\delta^2}(\lambda_U^2+\beta^2)\right) + 3\beta^2\Big)\right)\\
&\qquad\times \mathbb{E}\,\|\boldsymbol{\mathcal{w}}_{e,i-1}\|^2\\
&+\mu^2 N\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2\left(\frac{25}{1-\|\mathcal{J}_\epsilon\|}\left(3\lambda_U^2\frac{\mu^2}{\delta^2} + 3\beta^2\frac{\mu^2}{\delta^2} + 2\lambda_U^2 + \frac{2}{\delta^2}\right) + 3\beta^2\right)\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2\\
&+\frac{25\mu^2\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2}{1-\|\mathcal{J}_\epsilon\|}\left(\left(2+3\frac{\mu^2}{\delta^2}\right)\|g(\mathbb{1}\otimes w_\delta^o)\|^2 + 3\beta^2\frac{\mu^2}{\delta^2}N\|w_\delta^o\|^2 + \frac{\mu^2}{\delta^2}N\sigma^2 + 2\|r(\mathbb{1}\otimes w_\delta^o)\|^2\right)\\
&+\mu^2\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2\Big(3\beta^2 N\|w_\delta^o\|^2 + N\sigma^2\Big)\\
&= \left(\|\mathcal{J}_\epsilon\| + \frac{\mu^2}{\delta^2}\frac{25\nu^2 N\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2}{1-\|\mathcal{J}_\epsilon\|} + \mu^2\frac{25\nu^2 N\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2}{1-\|\mathcal{J}_\epsilon\|}\left(2\lambda_U^2 + \frac{1-\|\mathcal{J}_\epsilon\|}{25}3\beta^2\right)\right.\\
&\qquad\left.+\frac{\mu^4}{\delta^2}\frac{25\nu^2 N\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2}{1-\|\mathcal{J}_\epsilon\|}\left(1+3\beta^2+3\lambda_U^2\right)\right)\mathbb{E}\,\|\boldsymbol{\mathcal{w}}_{e,i-1}\|^2\\
&+\left(\frac{\mu^2}{\delta^2}\frac{50N\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2}{1-\|\mathcal{J}_\epsilon\|} + \mu^2\frac{25N\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2}{1-\|\mathcal{J}_\epsilon\|}\left(2\lambda_U^2 + \frac{1-\|\mathcal{J}_\epsilon\|}{25}3\beta^2\right)\right.\\
&\qquad\left.+\frac{\mu^4}{\delta^2}\frac{75N\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2}{1-\|\mathcal{J}_\epsilon\|}\big(\lambda_U^2+\beta^2\big)\right)\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i-1}\|^2
\end{aligned}
$$

$$+\frac{25\mu^2\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2}{1-\|\mathcal{J}_\epsilon\|}\left(\left(2+3\frac{\mu^2}{\delta^2}\right)\|g(\mathbb{1}\otimes w_\delta^o)\|^2+3\beta^2\frac{\mu^2}{\delta^2}N\|w_\delta^o\|^2+\frac{\mu^2}{\delta^2}N\sigma^2+2\|r(\mathbb{1}\otimes w_\delta^o)\|^2\right)$$
$$+\mu^2\|\mathcal{J}_\epsilon\|^2\|\mathcal{V}_R\|^2\Big(3\beta^2N\|w_\delta^o\|^2+N\sigma^2\Big) \tag{3.146}$$

In step (a), cross-terms are eliminated because $\mathbb{E}\,\{\boldsymbol{s}_i-\mathbb{E}\,\boldsymbol{s}_i\}=0$. Step (b) is due to $\|\mathcal{J}_\epsilon\|<1$ and Jensen's inequality, (c) is due to the sub-multiplicative property of norms, (d) follows from Jensen's inequality, and (e) is due to $\mathbb{1}^{\mathsf{T}}P[\boldsymbol{x}]=\mathbb{E}\,\|x\|^2$. The bounds from Lemma 3.3 are used in (f) and (g) is due to $\mathbb{1}^{\mathsf{T}}P[\boldsymbol{x}]=\mathbb{E}\,\|\boldsymbol{x}\|^2$ for $\boldsymbol{x}\in\mathbb{R}^{MN}$ and $\mathbb{1}^{\mathsf{T}}P[\mathbb{1}\otimes\boldsymbol{y}]=N\cdot\mathbb{E}\,\|\boldsymbol{y}\|^2$ for $\boldsymbol{y}\in\mathbb{R}^M$.

3.G Proof of Lemma 3.5

For $\delta=\mu^{\frac{1}{2}-\kappa}$ and small step-sizes μ,

$$\Gamma=\begin{bmatrix}1-\mu\lambda_L+O(\mu^2) & O(\mu^{2\kappa})\\ O(\mu^{1+2\kappa}) & \|\mathcal{J}_\epsilon\|+O(\mu^{1+2\kappa})\end{bmatrix} \tag{3.147}$$

so that

$$\|\Gamma\|_1=\max\left\{1-\mu\lambda_L+O(\mu^{1+2\kappa}),\|\mathcal{J}_\epsilon\|+O(\mu^{2\kappa})\right\}<1 \tag{3.148}$$

for small enough μ. Since $\rho(\Gamma)\leq\|\Gamma\|_1<1$, Γ is stable. It is also invertible and we obtain

$$\limsup_{i\to\infty}\begin{bmatrix}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{c,i}\|^2\\ \mathbb{E}\,\|\,\boldsymbol{w}_{e,i}\,\|^2\end{bmatrix}\preceq(I-\Gamma)^{-1}\begin{bmatrix}O(\mu^2)\\ O(\mu^{1+2\kappa})\end{bmatrix} \tag{3.149}$$

Using the matrix inversion lemma, we have

$$\begin{aligned}(I-\Gamma)^{-1}&=\begin{bmatrix}\mu\lambda_L-O(\mu^2) & -O(\mu^{2\kappa})\\ -O(\mu^{1+2\kappa}) & 1-\|\mathcal{J}_\epsilon\|-O(\mu^{1+2\kappa})\end{bmatrix}^{-1}\\&=\begin{bmatrix}O(\mu) & -O(\mu^{2\kappa})\\ -O(\mu^{1+2\kappa}) & O(1)\end{bmatrix}^{-1}\end{aligned}$$

$$= \begin{bmatrix} O(\mu^{-1}) & O(\mu^{-1+2\kappa}) \\ O(\mu^{2\kappa}) & O(1) \end{bmatrix} \tag{3.150}$$

The result follows after multiplication and cancellation.

CHAPTER 4

Extension to Matrix Variables

4.1 Problem and Algorithm Formulation

Up to this point, we have restricted the optimization variable w and the iterates $\boldsymbol{w}_{k,i}$ to be vector-valued. We now broaden our scope to consider problems of the form

$$W^o = \arg\min_{W} \sum_{k=1}^{N} p_k \left\{ J_k(W) + R_k(W) \right\} \tag{4.1}$$

where $W \in \mathbb{R}^{N_1 \times N_2}$ and $J_k(\cdot) \,:\, \mathbb{R}^{N_1 \times N_2} \to \mathbb{R},\ R_k(\cdot) \,:\, \mathbb{R}^{N_1 \times N_2} \to \mathbb{R}$. Problems of this form frequently appear, for example, in image processing, when the structure of W is important. We will illustrate an application from image reconstruction further below in section 4.3. Our discussion in Chapter 3, motivates the following algorithm.

Algorithm 4.1 Regularized Diffusion Strategy for Matrix Variables

$$\boldsymbol{\Phi}_{k,i} = \boldsymbol{W}_{k,i-1} - \mu \widehat{\nabla_W J}_k(\boldsymbol{W}_{k,i-1}) \tag{4.2}$$

$$\boldsymbol{\Psi}_{k,i} = \boldsymbol{\Phi}_{k,i} - \mu \nabla_w R_k^{\delta}(\boldsymbol{\Phi}_{k,i}) \tag{4.3}$$

$$\boldsymbol{W}_{k,i} = \sum_{\ell=1}^{N} a_{\ell k} \boldsymbol{\Psi}_{\ell,i} \tag{4.4}$$

4.2 Analogy to Vector Optimization

There is no need to repeat the analysis from Chapter 3, since any problem of the form (4.1) can be mapped to an auxiliary problem, where the optimization variable is vector valued.

To this end, define for $W = (w_1, w_2, \cdots, w_{N_2}) \in \mathbb{R}^{N_1 \times N_2}$ with columns $w_1, w_2, \ldots, w_{N_2}$:

$$\text{vec}(W) \triangleq \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{N_2} \end{bmatrix} \in \mathbb{R}^{N_1 \times N_2} \tag{4.5}$$

and the corresponding inverse operation

$$\text{unvec}\left(\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{N_2} \end{bmatrix} \right) \triangleq W. \tag{4.6}$$

We can then define auxiliary functions

$$J_k^{\text{aux}}(w) = J_k(\text{unvec}(w)) \ : \ \mathbb{R}^{N_1 N_2 \times 1} \to \mathbb{R} \tag{4.7}$$

$$R_k^{\text{aux}}(w) = R_k(\text{unvec}(w)) \ : \ \mathbb{R}^{N_1 N_2 \times 1} \to \mathbb{R} \tag{4.8}$$

and an auxiliary problem

$$w^o = \arg\min_{w} \sum_{k=1}^{N} p_k \left\{ J_k^{\text{aux}}(w) + R_k^{\text{aux}}(w) \right\} \tag{4.9}$$

Running the original regularized diffusion strategy (4.2)–(4.4) on (4.9) is then equivalent to running the regularized diffusion strategy for matrix variables (4.2)–(4.4) on (4.1). As such, our conclusions and performance guarantees continue to hold.

4.3 Distributed Image Reconstruction

We consider a scenario, where an image $A \in \mathbb{R}^{N_1} \times \mathbb{R}^{N_2}$ is observed partially by a collection of M agents labeled 1 through M. To formalize this, define a sampling set $\mathcal{S}$, consisting of

index pairs (r,c) of A and a sampling operator $S_{\mathcal{S}}[\cdot] \;:\; \mathbb{R}^{N_1\times N_2} \to \mathbb{R}^{N_1\times N_2}$. The (r,c)-th element of $S[A]$ is given by:

$$S_{\mathcal{S}}[A]_{(r,c)} \triangleq \begin{cases} A_{(r,c)} & \text{if } (r,c)\in\mathcal{S} \\ 0 & \text{otherwise} \end{cases}. \tag{4.10}$$

The problem of reconstructing A from $S[A]$, when $S[A]$ is available at a centralized location is well-studied. It is known matrix completion in the general setting, or image reconstruction in image processing. Under the assumption that the image is smooth in the sense that A is low-rank, the image can be reconstructed using [126]:

$$W^o = \arg\min_W \frac{1}{2}\|S_{\mathcal{S}}[A-W]\|_F^2 + \rho_1\|W\|_* + \frac{\rho_2}{2}\|W\|_F^2 \tag{4.11}$$

Here, $\|W\|_*$ denotes the nuclear norm, i.e. the sum of the singular values of W:

$$\|W\|_* = \sum_{i=1}^{\text{rank}(W)} \sigma_i(W) \tag{4.12}$$

The nuclear norm of W corresponds to the ℓ_1-norm on the singular values of W, which encourages sparse (i.e. low-rank) solutions.

If instead of A directly, we only have access to noisy perturbations of A through

$$\boldsymbol{A}_i = A + \boldsymbol{V}_i \tag{4.13}$$

we can formulate a stochastic variation of (4.11):

$$W^o = \arg\min_W \frac{1}{2}\,\mathbb{E}\,\|S_{\mathcal{S}}[\boldsymbol{A}_i - W]\|_F^2 + \rho_1\|W\|_* + \frac{\rho_2}{2}\|W\|_F^2 \tag{4.14}$$

To distribute (4.14), define additionally for every agent k, an agent specific sampling set $\mathcal{S}_k$ and assume $\mathcal{S}_k \cap \mathcal{S}_\ell = \emptyset$ for all k,ℓ and $\cup_{k=1}^M \mathcal{S}_k = \mathcal{S}$. In other words, every agent observes a different part of the image, while the network as a whole observes all of $\mathcal{S}$. At each time

instance i, agent k then observes

$$\boldsymbol{A}_{k,i} = S_{\mathcal{S}\cap\mathcal{S}_k}[A] + \boldsymbol{V}_i \tag{4.15}$$

where $\boldsymbol{V}_i$ is a zero-mean noise term. The centralized problem (4.11) can then be decomposed as

$$W^o = \arg\min_{W} \sum_{k=1}^{N} p_k \left\{ \frac{1}{2}\,\mathbb{E}\,\|S_{\mathcal{S}\cap\mathcal{S}_k}[\boldsymbol{A} - W]\|_F^2 + \frac{\rho_2}{2}\|S_{\mathcal{S}_k}[W]\|_F^2 + \rho_1\|W\|_* \right\} \tag{4.16}$$

If we let

$$J_k(W) \triangleq \frac{1}{2}\,\mathbb{E}\,\|S_{\mathcal{S}\cap\mathcal{S}_k}[\boldsymbol{A} - W]\|_F^2 + \frac{\rho_2}{2}\|S_{\mathcal{S}_k}[W]\|_F^2 \tag{4.17}$$

$$R_k(W) \triangleq \rho_1\|W\|_* \tag{4.18}$$

then (4.16) is of the form of (4.1), and hence lends itself to a distributed solution using (4.2)–(4.4). If we let $d(W) = \frac{1}{2}\|W\|_F^2$ in the construction of the smooth approximation, we have

$$\widehat{\nabla J}_k(W) = -S_{\mathcal{S}\cap\mathcal{S}_k}[\boldsymbol{A}_{k,i} - W] + \rho_2 S_{\mathcal{S}_k}[W] \tag{4.19}$$

$$\nabla R_k^{\delta}(W) = \frac{1}{\delta}\left(W - \mathrm{prox}_{\delta\|\cdot\|_*}(W)\right) \tag{4.20}$$

The proximal operator of $\|\cdot\|_*$ is available in closed form and corresponds to soft-thresholding on the singular values of W [127]. We arrive at the algorithm.

Algorithm 4.2 Regularized Diffusion Strategy for Matrix Completion

$$\boldsymbol{\Phi}_{k,i} = \boldsymbol{W}_{k,i-1} + \mu S_{\mathcal{S}\cap\mathcal{S}_k}[\boldsymbol{A} - W] - \mu\rho_2 S_{\mathcal{S}_k}[W] \tag{4.21}$$

$$\boldsymbol{\Psi}_{k,i} = \left(1 - \frac{\mu}{\delta}\right)\boldsymbol{\Phi}_{k,i} + \frac{\mu}{\delta}\,\mathrm{prox}_{\delta\|\cdot\|_*}(\boldsymbol{\Phi}_{k,i}) \tag{4.22}$$

$$\boldsymbol{W}_{k,i} = \sum_{\ell=1}^{N} a_{\ell k}\boldsymbol{\Psi}_{\ell,i} \tag{4.23}$$

4.3.1 Numerical Results

We present the algorithm with the image A, shown in Fig. 4.1, which has been corrupted by passing it through a sampling operator $S_{\mathcal{S}}[\cdot]$, where each index pair (r, c) is either 0 or 1 with probabilities 0.2 and 0.8 respectively. The corrupted image $S_{\mathcal{S}}[A]$ is shown in Fig. 4.2. In addition to the fact that the image is globally corrupted through $S_{\mathcal{S}}[\cdot]$, each agent only

Figure 4.1: Original image A.

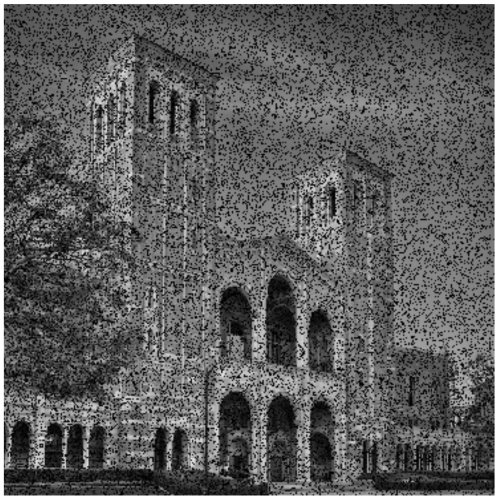

Figure 4.2: Corrupted image $S_{\mathcal{S}}[A]$.

observes a noisy subset of $S_{\mathcal{S}}[A]$. To illustrate the flow of information, we chose the number of agents to be $M = 12$ and decompose the image into 3 rows and 4 columns of blocks of equal size. This is illustrated in Fig. 4.3.

Each agent observes then at iteration i:

$$\boldsymbol{A}_{k,i} = S_{\mathcal{S}\cap\mathcal{S}_k}[A] + \boldsymbol{V}_i \tag{4.24}$$

where $\boldsymbol{V}_i$ consists of i.i.d. elements drawn from a normal distribution with zero mean and unit variance. Algorithm parameters are set to $\mu = 0.9$, $\delta = 2$, $\rho_1 = 1$, $rho_2 = 100$ and agents give equal weight to data received from each of their neighbors. The evolution of the algorithm is shown in Figs. 4.4–4.7.

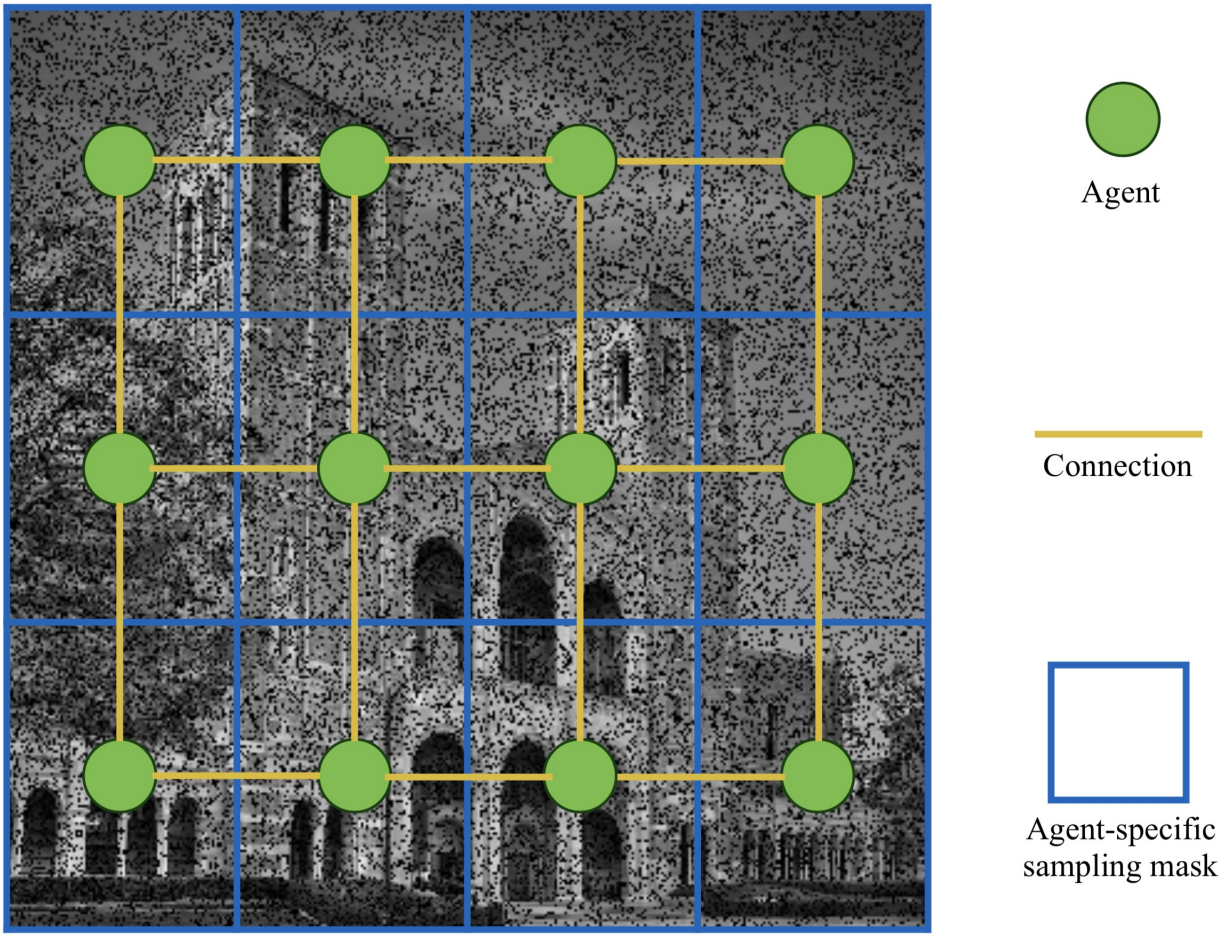

Figure 4.3: The sampled image is decomposed into 12 blocks of size 150×200. Each agent only has access to the block it has been assigned. For example, the top-left agent only sees the top-left block of the sampled image. Agents are allowed to exchange estimates, if their respective blocks share an edge.

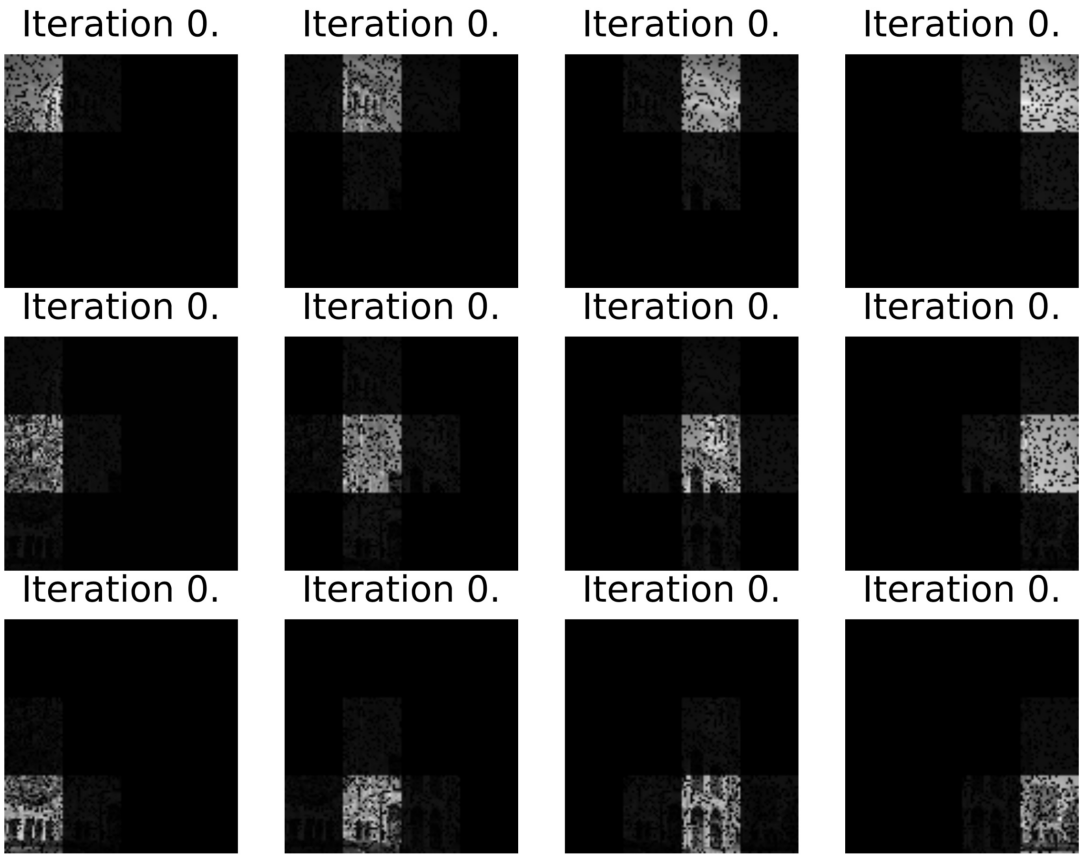

Figure 4.4: Each agent's estimate of the full image after a single iteration.

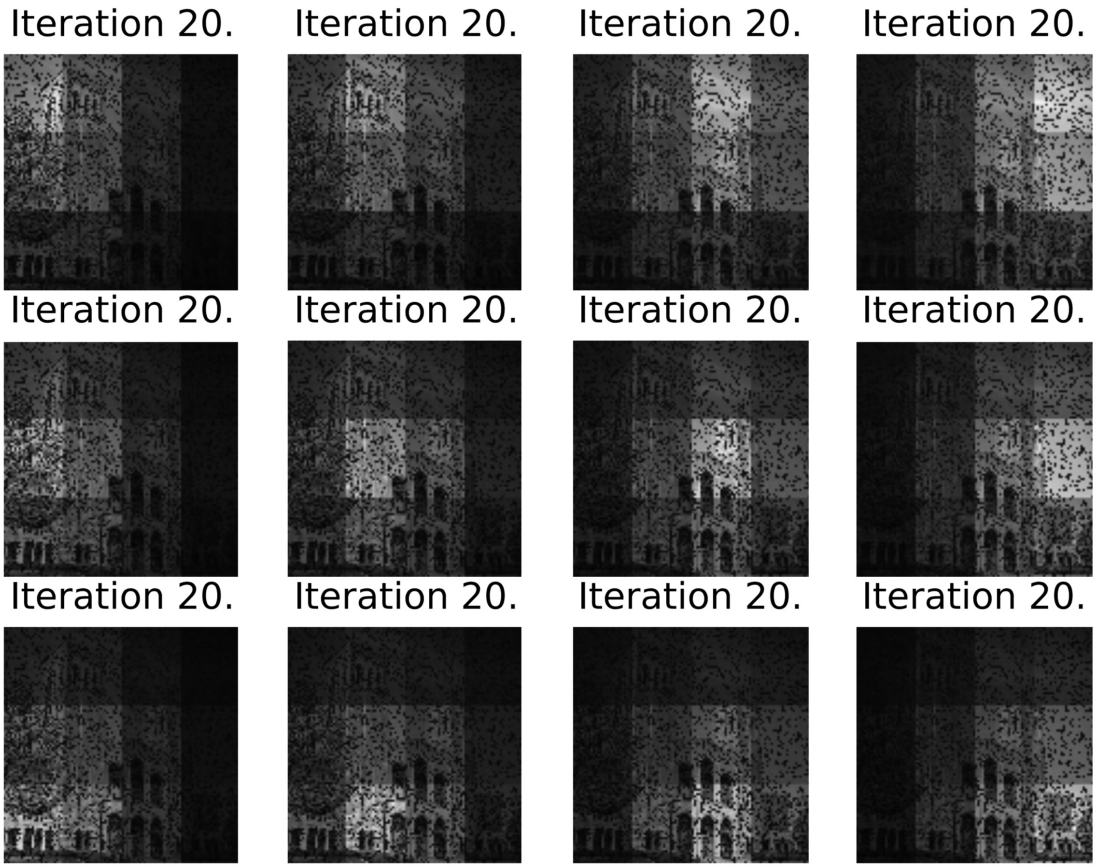

Figure 4.5: After 20 iterations, it can be observed how the information from each agent is radiated into its neighborhood.

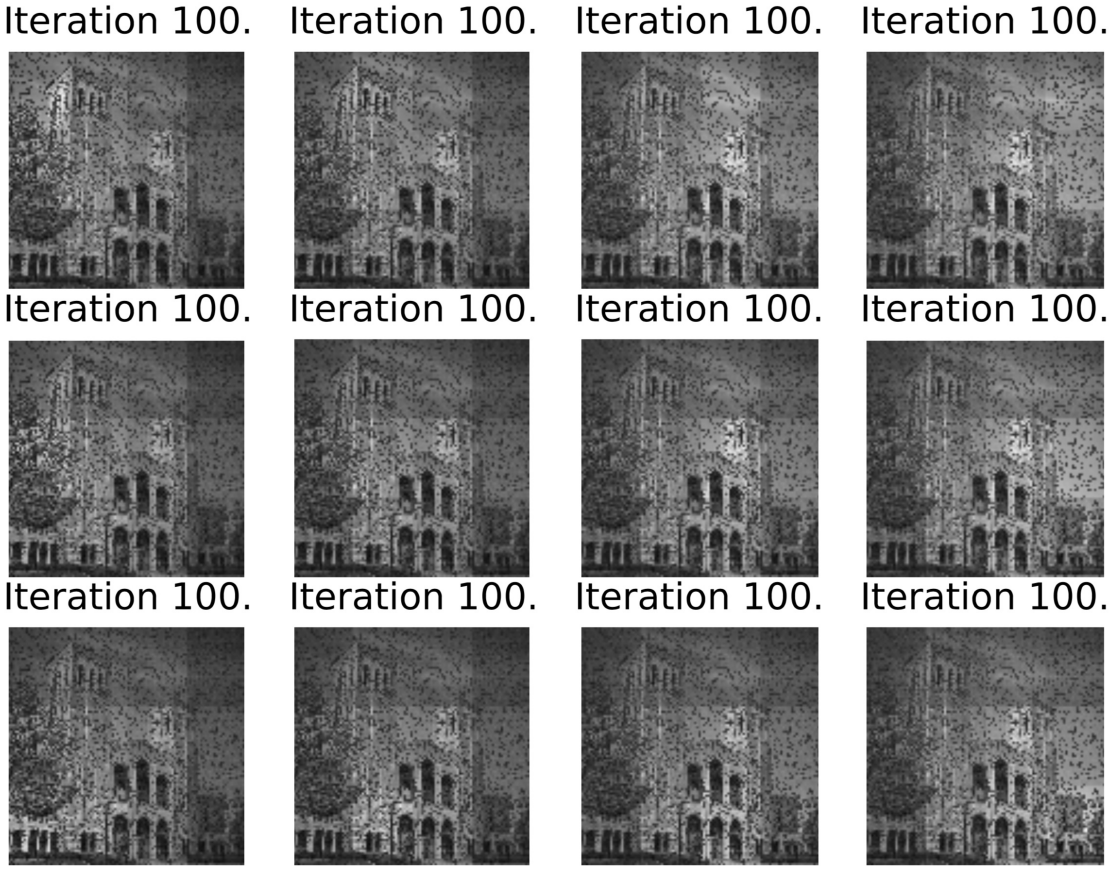

Figure 4.6: After 100 iterations, the agents have almost reached consensus and continue to refine their solution to move closer to the global minimizer.

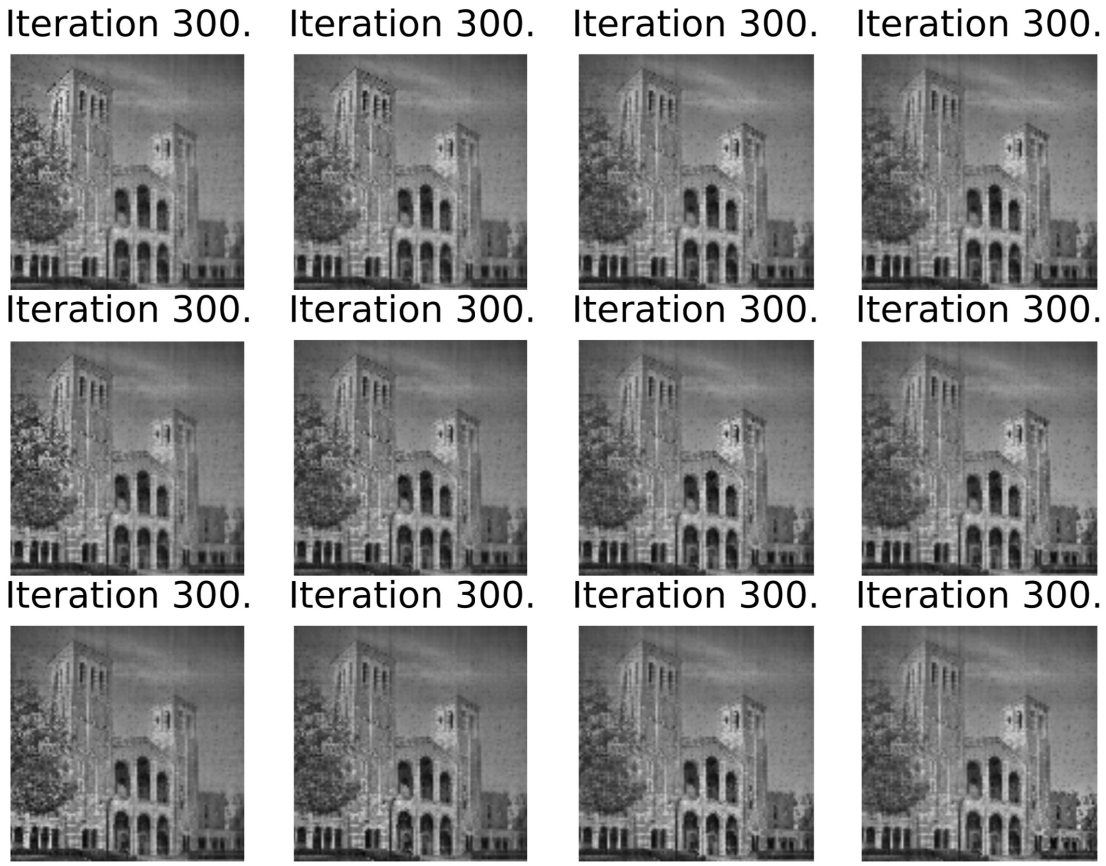

Figure 4.7: After 300 iterations, the full image has been recovered at every agent.

CHAPTER 5

Decentralized Non-Convex Learning — Short-Term Model

Driven by the need to solve increasingly complex optimization problems in signal processing and machine learning, there has been increasing interest in understanding the behavior of gradient-descent algorithms in non-convex environments. In this and the following Chapter 6, we consider stochastic cost functions, where exact gradients are replaced by stochastic approximations and the resulting gradient noise persistently seeps into the dynamics of the algorithm. We establish that the diffusion learning strategy continues to yield meaningful estimates non-convex scenarios in the sense that the iterates by the individual agents will cluster in a small region around the network centroid in the mean-fourth sense. We use this insight to motivate a short-term model for network evolution over a finite-horizon. In Chapter 6, we leverage this model to establish descent of the diffusion strategy through saddle points in $O(1/\mu)$ steps and the return of approximately second-order stationary points in a polynomial number of iterations. The materials in this chapter are based on the works [69,70].

5.1 Introduction

The broad objective of distributed adaptation and learning is the solution of global, stochastic optimization problems by networked agents through localized interactions and in the absence of information about the statistical properties of the data. When constant, rather than diminishing, step-sizes are employed, the resulting algorithms are adaptive in nature and are able to adapt to drifts in the data statistics. In this chapter, we consider a collection of N agents, where each agent k is equipped with a stochastic risk of the form $J_k(w) =$

$\mathbb{E}_x\, Q_k(w;\boldsymbol{x}_k)$ with $Q_k(w;\boldsymbol{x}_k)$ referring to the loss function, $w \in \mathbb{R}^M$ denoting a parameter vector, and $\boldsymbol{x}_k$ referring to the stochastic data. The expectation is over the probability distribution of the data. The objective of the network is to seek the Pareto solution:

$$\min_{w} J(w), \qquad \text{where } J(w) \triangleq \sum_{k=1}^{N} p_k J_k(w) \tag{5.1}$$

where the p_k are positive weights that are normalized to add up to one and will be specified further below; in particular, in the special case when the $\{p_k\}$ are identical, they can be removed from (5.1). Algorithms for the solution of (5.1) have been studied extensively over recent years both with inexact [1, 26–28] and exact [29–31] gradients. Here, we focus on the following diffusion strategy, which has been shown in previous works to provide enhanced performance and stability guarantees under constant step-size learning and adaptive scenarios [1, 22]:

$$\boldsymbol{\phi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \tag{5.2a}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell=1}^{N} a_{\ell k}\boldsymbol{\phi}_{\ell,i} \tag{5.2b}$$

where $\widehat{\nabla J}_k(\cdot)$ denotes a stochastic approximation for the true local gradient $\nabla J_k(\cdot)$. The intermediate estimate $\boldsymbol{\phi}_{k,i}$ is obtained at agent k by taking a stochastic gradient update relative to the local cost $J_k(\cdot)$. The intermediate estimates are then fused across local neighborhoods where $a_{\ell k}$ are convex combination weights satisfying:

$$a_{\ell k} \geq 0, \quad \sum_{\ell\in\mathcal{N}_k} a_{\ell k} = 1, \quad a_{\ell k} = 0 \text{ if } \ell \notin \mathcal{N}_k \tag{5.3}$$

The symbol $\mathcal{N}_k$ denotes the set of neighbors of agent k.

Assumption 5.1 (Strongly-connected graph). *We shall assume that the graph described by the weighted combination matrix $A = [a_{\ell k}]$ is strongly-connected [1]. This means that there exists a path with nonzero weights between any two agents in the network and, moreover, at least one agent has a nontrivial self-loop, $a_{kk} > 0$.* □

It then follows from the Perron-Frobenius theorem [1,23,24] that A has a single eigenvalue at one while all other eigenvalues are strictly inside the unit circle, so that $\rho(A) = 1$. Moreover, if we let p denote the right-eigenvector of A that is associated with the eigenvalue at one, and if we normalize the entries of p to add up to one, then it also holds that all entries of p are strictly positive, i.e.,

$$Ap = p, \quad \mathbb{1}^{\mathsf{T}} p = 1, \quad p_k > 0 \tag{5.4}$$

where the $\{p_k\}$ denote the individual entries of the Perron vector, p.

5.1.1 Related Works

The performance of the diffusion algorithm (5.2a)–(5.2b) has been studied extensively in differentiable settings [22,27], with extensions to multi-task [128], constrained [33], and non-differentiable [34] environments. A common assumption in these works, along with others studying the behavior of distributed optimization algorithms in general, is that of *convexity* (or strong-convexity) of the aggregate risk $J(w)$. While many problems of interest such as least-squares estimation [1], logistic regression [1], and support vector machines [129] are convex, there has been increased interest in the optimization of *non-convex* cost functions. Such problems appear frequently in the design of robust estimators [130] and the training of more complex machine learning architectures such as those involving dictionary learning [131] and artificial neural networks [62].

Motivated by these applications, recent works have pursued the study of optimization algorithms for non-convex problems, both in the centralized and distributed settings [52–54, 54, 55, 57–61, 132–143]. While some works focus on establishing convergence to a stationary point [52–57], there has been growing interest in examining the ability of gradient descent implementations to escape from saddle points, since such points represent bottlenecks to the underlying learning problem [62]. We defer a detailed discussion on the plethora of related works on second-order guarantees [59–61, 132–140, 144] to Chapter 6, where we will be establishing the ability of the diffusion strategy (5.2a)–(5.2b) to escape strict-saddle points efficiently. For ease of reference, the modeling conditions and results from this and related

works are summarized in Table 5.1.

The key contributions of Chapters 5 and 6 are three-fold. To the best of our knowledge, we present the first analysis establishing *efficient* (i.e., polynomial) escape from strict-saddle points in the *distributed* setting. Second, we establish that the gradient noise process is sufficient to ensure efficient escape without the need to alter it by adding artificial forms of perturbations, interlacing steps with small and large step-sizes, or imposing a dispersive noise assumption as long as a gradient noise component is present in the descent direction. Third, relative to the existing literature on *centralized* non-convex optimization, where the focus is mostly on deterministic or *finite-sum* optimization, our modeling conditions are specifically tailored to the scenario of learning from stochastic *streaming* data. In particular, we only impose bounds on the gradient noise variance in expectation, rather than assume a bound with probability one [134, 138] or a sub-Gaussian distribution [139]. Furthermore, we assume that any Lipschitz conditions only hold on the *expected* stochastic gradient approximation, rather than for every realization, with probability one [135–137].

	Modeling conditions					Results	
	Gradient	Hessian	Initialization	Perturbations	Step-size	Stationary	Saddle
Centralized							
[132]	Lipschitz	—	—	SGD + Annealing	diminishing	✓	asymptotic†
[59]	Lipschitz & bounded⋆	Lipschitz	—	i.i.d. and bounded w.p. 1	constant	✓	polynomial
[133]	Lipschitz	—	Random	—	constant	✓	asymptotic
[60]	Lipschitz	Lipschitz	—	Selective & bounded w.p. 1	constant	✓	polynomial
[134]	Lipschitz	Lipschitz	—	SGD, bounded w.p. 1	alternating	✓	polynomial
[135]	Lipschitz	Lipschitz	—	Bounded variance, Lipschitz w.p. 1	constant	✓	polynomial
[136]	Lipschitz	Lipschitz	—	Bounded variance, Lipschitz w.p. 1	constant	✓	polynomial
[137]	Lipschitz	Lipschitz	—	Bounded variance, Lipschitz w.p. 1	constant	✓	polynomial
[138]	Lipschitz	Lipschitz	—	SGD, bounded w.p. 1	constant	✓	polynomial
[139]	Lipschitz	Lipschitz	—	SGD + Gaussian	constant	✓	polynomial
Decentralized							
[52]	Cont. differentiable	—	—	SGD	diminishing	✓	—
[53]	Lipschitz & bounded	—	—	—	constant	✓	—
[54]	Bounded disagreement	bounded moments	—	SGD	constant	✓	—
[56]	Lipschitz	bounded moments	—	SGD	constant	✓	—
[57]	Lipschitz	—	—	—	constant	✓	—
[55]	Lipschitz + prox	—	—	—	constant	✓	—
[58]	Lipschitz & bounded	—	—	i.i.d.	diminishing	✓	—
[61]	Lipschitz	Exists	Random	—	constant	✓	asymptotic
[140]	Bounded disagreement	—	—	SGD + Annealing	diminishing	✓	asymptotic†
This work	**Bounded disagreement**	**Lipschitz**	**—**	**Bounded moments**	**constant**	✓	**polynomial**

Table 5.1: Comparison of modeling assumptions and results for gradient-based methods. Statements marked with ⋆ are not explicitly stated but are implied by other conditions. The works marked with † establish global (asymptotic) convergence, which of course implies escape from saddle-points.

5.1.2 Preview of Results

We first establish that in non-convex environments, as was already shown earlier in [27] for convex environments, the evolution of the individual iterates $\boldsymbol{w}_{k,i}$ at the agents continues to be well-described by the evolution of the weighted centroid vector $\sum_{k=1}^{N} p_k \boldsymbol{w}_{k,i}$ in the sense that the iterates from across the network will cluster around this centroid after sufficient iterations in the mean-fourth sense. We subsequently consider two cases separately and establish descent in both of them. The first case corresponds to the region where the gradient at the network centroid is large and establish that descent can occur in one iteration. The second and more challenging case occurs when the gradient norm is small, but there is a sufficiently negative eigenvalue in the Hessian matrix. We establish in Chapter 6 that the recursion will continue to descend along the aggregate cost at a rate of $O(\mu)$ per $O(1/\mu)$ iterations. Combined with the first result, this descent relation allows us to provide guarantees about the second-order optimality of the returned iterates.

The flow of the argument is summarized in Fig. 5.1. We decompose $\mathbb{R}^M$ into the set of approximate first-order stationary points, i.e., those with $\|\nabla J(w)\|^2 \leq O(\mu)$ and the complement, i.e., the large-gradient regime. For the large-gradient regime, descent is established in Theorem 5.2. Motivated by prior works establishing second-order guarantees [59, 60, 144], we proceed to further decompose the set of approximate first-order stationary points into those that are τ-strict-saddle, i.e., those that have a Hessian with significant negative eigenvalue $\lambda_{\min}\left(\nabla^2 J(w)\right) \leq -\tau$, and the complement, which are approximately second-order stationary points. For τ-strict-saddle points we establish descent in Theorem 6.1. Finally, in Theorem 6.2, we conclude that the centroid will reach an approximately second-order stationary point in a *polynomial* number of iterations.

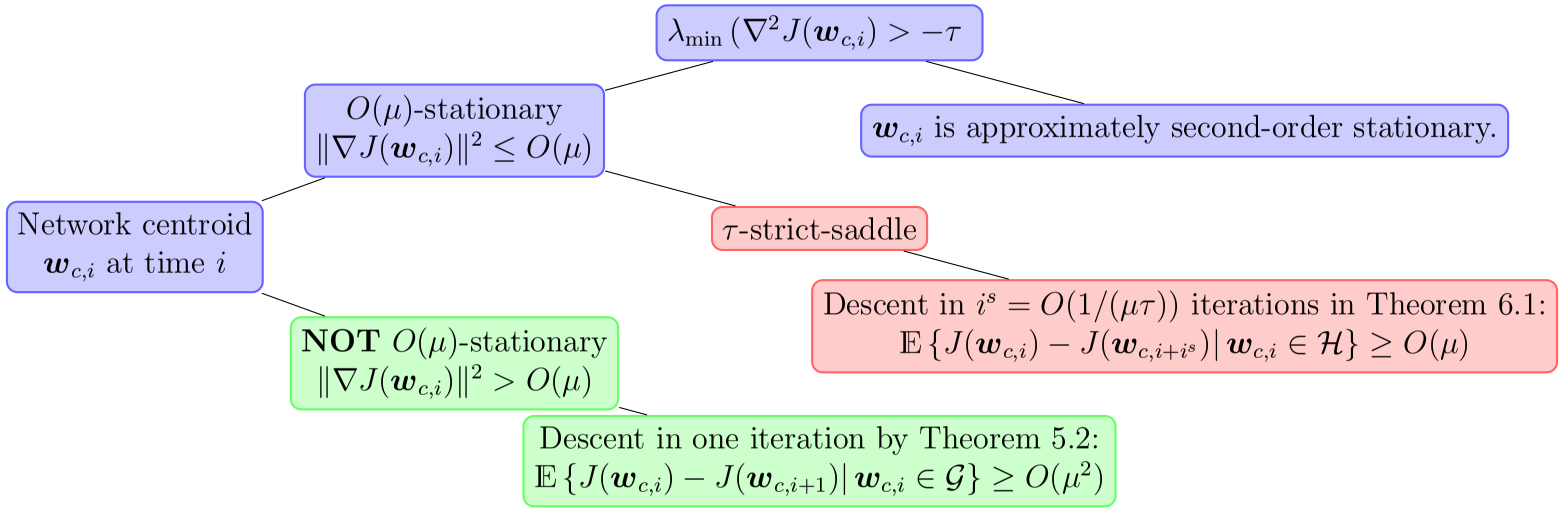

Figure 5.1: Classification of approximately stationary points. Theorem 5.2 in this chapter establishes descent in the green branch. The red branch is treated in Chapter 5. The two results are combined in Theorem 6.2 to establish the return of a second-order stationary point with high probability.

5.2 Evolution Analysis

We shall perform the analysis under the following common assumptions on the gradients and their approximations.

Assumption 5.2 (**Lipschitz gradients**). *For each k, the gradient $\nabla J_k(\cdot)$ is Lipschitz, namely, for any $x, y \in \mathbb{R}^M$:*

$$\|\nabla J_k(x) - \nabla J_k(y)\| \leq \delta \|x - y\| \tag{5.5}$$

In light of (5.1) *and Jensen's inequality, this implies for the aggregate cost:*

$$\|\nabla J(x) - \nabla J(y)\| \leq \delta \|x - y\| \tag{5.6}$$

□

The Lipschitz gradient conditions (5.5) and (5.6) imply bounds on the both the function value and the Hessian matrix (when it exists), which will be used regularly throughout the derivations. In particular, we have for the function values:

$$J(y) \leq J(x) + \nabla J(x)^{\mathsf{T}} (y - x) + \frac{\delta}{2} \|x - y\|^2 \tag{5.7}$$

For the Hessian matrix we have [1]:

$$-\delta I \leq \nabla^2 J(x) \leq \delta I \tag{5.8}$$

Assumption 5.3 (**Bounded gradient disagreement**). *For each pair of agents k and ℓ, the gradient disagreement is bounded, namely, for any $x \in \mathbb{R}^M$:*

$$\|\nabla J_k(x) - \nabla J_\ell(x)\| \leq G \tag{5.9}$$

□

This assumption is similar to the one used in [54] to establish first-order stationarity under constant step-size selection and [140] for global optimality under a diminishing step-size with annealing. Note that condition (5.9) is weaker than the more common assumption of bounded gradients. Condition (5.9) is automatically satisfied in cases where the expected risks $J_k(\cdot)$ are common (though agents still may see different realizations of data), or in the case of centralized stochastic gradient descent where the number of agents is one. This condition is also satisfied whenever agent-specific risks with bounded gradients are regularized by common regularizers with potentially unbounded gradients, as is common in many machine learning applications. Observe that (5.9) implies a similar condition on the deviation from the centralized gradient via Jensen's inequality:

$$\begin{aligned} \|\nabla J_k(x) - \nabla J(x)\| &= \left\| \sum_{\ell=1}^{N} p_\ell \left(\nabla J_k(x) - \nabla J_\ell(x) \right) \right\| \\ &\leq \sum_{\ell=1}^{N} p_\ell \left\| \nabla J_k(x) - \nabla J_\ell(x)) \right\| \leq G \end{aligned} \tag{5.10}$$

Definition 5.1 (**Filtration**). *We denote by $\boldsymbol{\mathcal{F}}_i$ the filtration generated by the random processes $\boldsymbol{w}_{k,j}$ for all k and $j \leq i$:*

$$\boldsymbol{\mathcal{F}}_i \triangleq \{\boldsymbol{\mathcal{W}}_0, \boldsymbol{\mathcal{W}}_1, \ldots, \boldsymbol{\mathcal{W}}_i\} \tag{5.11}$$

where $\boldsymbol{\mathcal{W}}_j \triangleq \operatorname{col}\{\boldsymbol{w}_{1,j}, \ldots, \boldsymbol{w}_{k,j}\}$ contains the iterates across the network at time j. Informally, $\boldsymbol{\mathcal{F}}_i$ captures all information that is available about the stochastic processes $\boldsymbol{w}_{k,j}$ across the network up to time i.

□

Throughout the following derivations, we will frequently rely on appropriate conditionings to make the analysis tractable. A frequent theme will be the exchange of conditioning on filtrations by conditioning on events. To this end, the following lemma will be used repeatedly.

Lemma 5.1 (**Conditioning**). *Suppose $\boldsymbol{w} \in \mathbb{R}^M$ is a random variable measurable by $\boldsymbol{\mathcal{F}}$. In*

other words, $\boldsymbol{w}$ is deterministic conditioned on $\boldsymbol{\mathcal{F}}$ and

$$\mathbb{E}\left\{\boldsymbol{w}\,|\boldsymbol{\mathcal{F}}\right\} = \boldsymbol{w} \tag{5.12}$$

Then,

$$\mathbb{E}\left\{\mathbb{E}\left\{\boldsymbol{x}\,|\boldsymbol{\mathcal{F}}\right\}|\,\boldsymbol{w}\in\mathcal{S}\right\} = \mathbb{E}\left\{\boldsymbol{x}\,|\,\boldsymbol{w}\in\mathcal{S}\right\} \tag{5.13}$$

for any deterministic set $\mathcal{S}\subseteq\mathbb{R}^M$ and random $\boldsymbol{x}\in\mathbb{R}^M$.

Proof. Denote by $\mathbb{I}_{\mathcal{S}}(\boldsymbol{w})$ the random indicator function:

$$\mathbb{I}_{\mathcal{S}}(\boldsymbol{w}) = \begin{cases} 1, \text{ if } \boldsymbol{w}\in\mathcal{S} \\ 0, \text{ otherwise.} \end{cases} \tag{5.14}$$

Since $\boldsymbol{w}$ is measurable by $\boldsymbol{\mathcal{F}}$, then $\mathbb{I}_{\mathcal{S}}(\boldsymbol{w})$ is measurable by $\boldsymbol{\mathcal{F}}$ as well. In other words, the event $\boldsymbol{w}\in\mathcal{S}$ is deterministic conditioned on $\boldsymbol{\mathcal{F}}$. Furthermore, for the random variable $\boldsymbol{x}\,\mathbb{I}_{\mathcal{S}}(\boldsymbol{w})$, we have:

$$\begin{aligned}\mathbb{E}\left\{\boldsymbol{x}\,\mathbb{I}_{\mathcal{S}}(\boldsymbol{w})\right\} &= \mathbb{E}\left\{\boldsymbol{x}\,\mathbb{I}_{\mathcal{S}}(\boldsymbol{w})|\,\boldsymbol{w}\in\mathcal{S}\right\}\cdot\Pr\left\{\boldsymbol{w}\in\mathcal{S}\right\} \\ &\quad+\mathbb{E}\left\{\boldsymbol{x}\,\mathbb{I}_{\mathcal{S}}(\boldsymbol{w})|\,\boldsymbol{w}\notin\mathcal{S}\right\}\cdot\Pr\left\{\boldsymbol{w}\notin\mathcal{S}\right\} \\ &= \mathbb{E}\left\{\boldsymbol{x}\,|\,\boldsymbol{w}\in\mathcal{S}\right\}\cdot\Pr\left\{\boldsymbol{w}\in\mathcal{S}\right\}\end{aligned} \tag{5.15}$$

Rearranging yields:

$$\mathbb{E}\left\{\boldsymbol{x}\,|\,\boldsymbol{w}\in\mathcal{S}\right\} = \frac{\mathbb{E}\left\{\boldsymbol{x}\,\mathbb{I}_{\mathcal{S}}(\boldsymbol{w})\right\}}{\Pr\left\{\boldsymbol{w}\in\mathcal{S}\right\}} \tag{5.16}$$

Similarly, for the random variable $\mathbb{E}\left\{\boldsymbol{x}\,|\boldsymbol{\mathcal{F}}\right\}\mathbb{I}_{\mathcal{S}}(\boldsymbol{w})$, we have:

$$\begin{aligned}&\mathbb{E}\left\{\mathbb{E}\left\{\boldsymbol{x}\,|\boldsymbol{\mathcal{F}}\right\}\mathbb{I}_{\mathcal{S}}(\boldsymbol{w})\right\} \\ =\;&\mathbb{E}\left\{\mathbb{E}\left\{\boldsymbol{x}\,|\boldsymbol{\mathcal{F}}\right\}\mathbb{I}_{\mathcal{S}}(\boldsymbol{w})|\,\boldsymbol{w}\in\mathcal{S}\right\}\cdot\Pr\left\{\boldsymbol{w}\in\mathcal{S}\right\} \\ &+\mathbb{E}\left\{\mathbb{E}\left\{\boldsymbol{x}\,|\boldsymbol{\mathcal{F}}\right\}\mathbb{I}_{\mathcal{S}}(\boldsymbol{w})|\,\boldsymbol{w}\notin\mathcal{S}\right\}\cdot\Pr\left\{\boldsymbol{w}\notin\mathcal{S}\right\} \\ =\;&\mathbb{E}\left\{\mathbb{E}\left\{\boldsymbol{x}\,|\boldsymbol{\mathcal{F}}\right\}|\,\boldsymbol{w}\in\mathcal{S}\right\}\cdot\Pr\left\{\boldsymbol{w}\in\mathcal{S}\right\}\end{aligned} \tag{5.17}$$

It then follows that:

$$\begin{aligned}\mathbb{E}\left\{\mathbb{E}\left\{\boldsymbol{x}\,|\boldsymbol{\mathcal{F}}\right\}|\,\boldsymbol{w}\in\mathcal{S}\right\} &\overset{(5.17)}{=} \frac{\mathbb{E}\left\{\mathbb{E}\left\{\boldsymbol{x}\,|\boldsymbol{\mathcal{F}}\right\}\mathbb{I}_{\mathcal{S}}(\boldsymbol{w})\right\}}{\Pr\left\{\boldsymbol{w}\in\mathcal{S}\right\}}\\ &\overset{(a)}{=} \frac{\mathbb{E}\left\{\mathbb{E}\left\{\boldsymbol{x}\,\mathbb{I}_{\mathcal{S}}(\boldsymbol{w})|\boldsymbol{\mathcal{F}}\right\}\right\}}{\Pr\left\{\boldsymbol{w}\in\mathcal{S}\right\}} \overset{(b)}{=} \frac{\mathbb{E}\left\{\boldsymbol{x}\,\mathbb{I}_{\mathcal{S}}(\boldsymbol{w})\right\}}{\Pr\left\{\boldsymbol{w}\in\mathcal{S}\right\}}\\ &\overset{(5.16)}{=} \mathbb{E}\left\{\boldsymbol{x}\,|\,\boldsymbol{w}\in\mathcal{S}\right\} \end{aligned} \tag{5.18}$$

where in step (a) we pulled $\mathbb{I}_{\mathcal{S}}(\boldsymbol{w})$ into the inner expectation, since it is deterministic conditioned on $\boldsymbol{\mathcal{F}}$ and (b) follows from the law of total expectation. □

Assumption 5.4 (**Gradient noise process**). *For each k, the gradient noise process is defined as*

$$\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) = \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) - \nabla J_k(\boldsymbol{w}_{k,i-1}) \tag{5.19}$$

and satisfies

$$\mathbb{E}\left\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})|\boldsymbol{\mathcal{F}}_{i-1}\right\} = 0 \tag{5.20a}$$

$$\mathbb{E}\left\{\|\,\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^4|\boldsymbol{\mathcal{F}}_{i-1}\right\} \leq \sigma^4 \tag{5.20b}$$

for some non-negative constant σ^4. We also assume that the gradient noise processes are pairwise uncorrelated over the space conditioned on $\boldsymbol{\mathcal{F}}_{i-1}$, i.e.:

$$\mathbb{E}\left\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\,\boldsymbol{s}_{\ell,i}(\boldsymbol{w}_{\ell,i-1})^{\mathsf{T}}|\boldsymbol{\mathcal{F}}_{i-1}\right\} = 0 \tag{5.21}$$

□

Property (5.20a) means that the gradient noise construction is unbiased on average. Property (5.20b) means that the fourth-moment of the gradient noise is bounded. These properties are automatically satisfied for several costs of interest [1, 22]. Note, that the bound on

the fourth-order moment, in light of Jensen's inequality, immediately implies:

$$\begin{aligned}\mathbb{E}\left\{\|\,\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}\right\} &= \mathbb{E}\left\{\sqrt{\|\,\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^4}|\boldsymbol{\mathcal{F}}_{i-1}\right\} \\ &\leq \sqrt{\mathbb{E}\left\{\|\,\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^4|\boldsymbol{\mathcal{F}}_{i-1}\right\}} \overset{(5.20b)}{\leq} \sigma^2 \end{aligned} \tag{5.22}$$

While our primary interest is in the development of algorithms that allow for learning from *streaming* data, we remark briefly that the results obtained in this work are equally applicable to empirical risk minimization via stochastic gradient descent, by assuming that the streaming data is selected according to a particular distribution.

Example 5.1 (Empirical Risk Minimization)**.** Suppose the costs $J_k(\cdot)$ are empirical based on locally collected data $\{x_{k,s}\}_{s=1}^{S}$ and take the form:

$$J_k(w) = \frac{1}{S}\sum_{s=1}^{S} Q(w, x_{k,s}) \tag{5.23}$$

In empirical risk minimization (ERM) problems, we are interested in finding a vector w^o that minimizes the following empirical risk over the data across the *entire* network:

$$w^o \triangleq \arg\min_{w} \frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{S}\sum_{s=1}^{S} Q(w, x_{k,s})\right) \tag{5.24}$$

If we introduce the uniformly-distributed random variable $\boldsymbol{x}_k = x_{k,s}$ with probability $\frac{1}{S}$ for all s, then the cost (5.24) is equivalent to solving:

$$w^o = \arg\min_{w} \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}_{\boldsymbol{x}_k}\, Q(w, \boldsymbol{x}_k) \tag{5.25}$$

which is of the same form as (5.1) with $p_k = \frac{1}{N}$. The resulting gradient noise process satisfies the assumptions imposed in this chapter under appropriate conditions on the risk $Q(\cdot,\cdot)$. This observation has been leveraged to accurately quantify the performance of stochastic gradient descent, as well as mini-batch and importance sampling generalizations, for empirical minimization of convex risks in [9]. □

5.2.1 Network basis transformation

In analyzing the dynamics of the distributed algorithm (5.2a)–(5.2b), it is useful to introduce the following extended quantities by collecting variables from across the network:

$$\boldsymbol{\mathcal{w}}_i \triangleq \operatorname{col}\{\boldsymbol{w}_{1,i}, \ldots, \boldsymbol{w}_{N,i}\} \tag{5.26}$$

$$\mathcal{A} \triangleq A \otimes I_M \tag{5.27}$$

$$\widehat{\boldsymbol{g}}(\boldsymbol{\mathcal{w}}_i) \triangleq \operatorname{col}\left\{\widehat{\nabla J_1}(\boldsymbol{w}_{1,i}), \ldots, \widehat{\nabla J}_N(\boldsymbol{w}_{N,i})\right\} \tag{5.28}$$

where $\otimes$ denotes the Kronecker product operation. We can then write the diffusion recursion (5.2a)–(5.2b) compactly as

$$\boldsymbol{\mathcal{w}}_i = \mathcal{A}^{\mathsf{T}}\left(\boldsymbol{\mathcal{w}}_{i-1} - \mu\widehat{\boldsymbol{g}}(\boldsymbol{\mathcal{w}}_{i-1})\right) \tag{5.29}$$

By construction, the combination matrix A is left-stochastic and primitive and hence admits a Jordan decomposition of the form $A = V_\epsilon J V_\epsilon^{-1}$ with [1, 27]:

$$V_\epsilon = \begin{bmatrix} p & V_R \end{bmatrix}, \quad J = \begin{bmatrix} 1 & 0 \\ 0 & J_\epsilon \end{bmatrix}, \quad V_\epsilon^{-1} = \begin{bmatrix} \mathbb{1}^{\mathsf{T}} \\ V_L^{\mathsf{T}} \end{bmatrix} \tag{5.30}$$

where J_ϵ is a block Jordan matrix with the eigenvalues $\lambda_2(A)$ through $\lambda_N(A)$ on the diagonal and ϵ on the first lower sub-diagonal. The extended matrix $\mathcal{A}$ then satisfies $\mathcal{A} = \mathcal{V}_\epsilon \mathcal{J} \mathcal{V}_\epsilon^{-1}$ with $\mathcal{V}_\epsilon = V_\epsilon \otimes I_N$, $\mathcal{J} = J \otimes I_N$, $\mathcal{V}_\epsilon^{-1} = V_\epsilon^{-1} \otimes I_N$. The spectral properties of A and its corresponding eigendecomposition have been exploited extensively in the study of the diffusion learning strategy in the *convex* setting [1, 27], and will continue to be useful in *non-convex* scenarios.

Multiplying both sides of (5.29) by $\left(p^{\mathsf{T}} \otimes I\right)$ from the left, we obtain in light of (5.4):

$$\left(p^{\mathsf{T}} \otimes I\right)\boldsymbol{\mathcal{w}}_i = \left(p^{\mathsf{T}} \otimes I\right)\boldsymbol{\mathcal{w}}_{i-1} - \mu\left(p^{\mathsf{T}} \otimes I\right)\widehat{\boldsymbol{g}}(\boldsymbol{\mathcal{w}}_{i-1}) \tag{5.31}$$

Letting $\boldsymbol{w}_{c,i} \triangleq \sum_{k=1}^{K} p_k \boldsymbol{w}_{k,i} = \left(p^{\mathsf{T}} \otimes I\right)\boldsymbol{\mathcal{w}}_i$ and exploiting the block-structure of the gradient

term, we find:

$$\boldsymbol{w}_{c,i} = \boldsymbol{w}_{c,i-1} - \mu \sum_{k=1}^{N} p_k \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \tag{5.32}$$

Note that $\boldsymbol{w}_{c,i}$ is a convex combination of iterates across the network and can be viewed as a weighted centroid. The recursion for $\boldsymbol{w}_{c,i}$ is reminiscent of a stochastic gradient step associated with the aggregate cost $\sum_{k=1}^{N} p_k J_k(w)$ with the exact gradients $\nabla J_k(\cdot)$ replaced by stochastic approximations $\widehat{\nabla J}_k(\cdot)$ *and* with the stochastic gradients evaluated at $\boldsymbol{w}_{k,i-1}$, rather than $\boldsymbol{w}_{c,i-1}$. In fact, we can write:

$$\boldsymbol{w}_{c,i} = \boldsymbol{w}_{c,i-1} - \mu \sum_{k=1}^{N} p_k \nabla J_k(\boldsymbol{w}_{c,i-1}) - \mu \boldsymbol{d}_{i-1} - \mu\, \boldsymbol{s}_i \tag{5.33}$$

where we defined the perturbation terms:

$$\boldsymbol{d}_{i-1} \triangleq \sum_{k=1}^{N} p_k \left(\nabla J_k(\boldsymbol{w}_{k,i-1}) - \nabla J_k(\boldsymbol{w}_{c,i-1}) \right) \tag{5.34}$$

$$\boldsymbol{s}_i \triangleq \sum_{k=1}^{N} p_k \left(\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) - \nabla J_k(\boldsymbol{w}_{k,i-1}) \right) \tag{5.35}$$

We use the subscript $i-1$ for $\boldsymbol{d}_{i-1}$ to emphasize that it depends on data up to time $i-1$, in contrast to $\boldsymbol{s}_i$ which is also dependent on the most recent data from time i. Observe that $\boldsymbol{d}_{i-1}$ arises from the disagreement within the network, and in particular that if each $\boldsymbol{w}_{k,i-1}$ remains close to the network centroid $\boldsymbol{w}_{c,i-1}$, this perturbation will be small in light of the Lipschitz condition (5.5) on the gradients. The second perturbation term $\boldsymbol{s}_i$ arises from the noise introduced by stochastic gradient approximations at each agent. We now establish that recursion (5.33) will continue to exhibit some of the desired properties of (centralized) gradient descent, despite the presence of persistent and coupled perturbation terms.

5.2.2 Network disagreement

To begin with, we study more closely the evolution of the individual estimates $\boldsymbol{w}_{k,i}$ relative to the network centroid $\boldsymbol{w}_{c,i}$. Multiplying (5.29) by $\mathcal{V}_R^{\mathsf{T}} \triangleq \left(V_R^{\mathsf{T}} \otimes I\right)$ from the left yields in

light of (5.30):

$$\begin{aligned}\mathcal{V}_R^{\mathsf{T}}\boldsymbol{\mathcal{w}}_i &= \mathcal{V}_R^{\mathsf{T}}\mathcal{A}^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_{i-1} - \mu\mathcal{V}_R^{\mathsf{T}}\mathcal{A}^{\mathsf{T}}\widehat{\boldsymbol{\mathcal{g}}}(\boldsymbol{\mathcal{w}}_{i-1}) \\ &= \mathcal{V}_R^{\mathsf{T}}\mathcal{A}^{\mathsf{T}}\mathcal{V}_L\mathcal{V}_R^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_{i-1} - \mu\mathcal{V}_R^{\mathsf{T}}\mathcal{A}^{\mathsf{T}}\mathcal{V}_L\mathcal{V}_R^{\mathsf{T}}\widehat{\boldsymbol{\mathcal{g}}}(\boldsymbol{\mathcal{w}}_{i-1}) \\ &= J_\epsilon^{\mathsf{T}}\mathcal{V}_R^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_{i-1} - \mu J_\epsilon^{\mathsf{T}}\mathcal{V}_R^{\mathsf{T}}\widehat{\boldsymbol{\mathcal{g}}}(\boldsymbol{\mathcal{w}}_{i-1})\end{aligned} \tag{5.36}$$

Then, for the deviation from the network centroid:

$$\begin{aligned}\boldsymbol{\mathcal{w}}_i - \boldsymbol{\mathcal{w}}_{c,i} &= \boldsymbol{\mathcal{w}}_i - \left(\mathbb{1}p^{\mathsf{T}}\otimes I\right)\boldsymbol{\mathcal{w}}_i \\ &= \left(I - \left(\mathbb{1}p^{\mathsf{T}}\otimes I\right)\right)\boldsymbol{\mathcal{w}}_i \\ &= \left(\left(V_\epsilon^{-1}\otimes I\right)^{\mathsf{T}}(V_\epsilon\otimes I)^{\mathsf{T}} - \left(\mathbb{1}p^{\mathsf{T}}\otimes I\right)\right)\boldsymbol{\mathcal{w}}_i \\ &\overset{(5.30)}{=} \mathcal{V}_L\mathcal{V}_R^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_i\end{aligned} \tag{5.37}$$

so that the deviation from the centroid can be easily recovered from $\mathcal{V}_R^{\mathsf{T}}\boldsymbol{\mathcal{w}}_i$ in (5.36). Proceeding with (5.36), we find:

$$\begin{aligned}\left\|\mathcal{V}_R^{\mathsf{T}}\boldsymbol{\mathcal{w}}_i\right\|^4 &= \left\|J_\epsilon^{\mathsf{T}}\mathcal{V}_R^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_{i-1} - \mu J_\epsilon^{\mathsf{T}}\mathcal{V}_R^{\mathsf{T}}\widehat{\boldsymbol{\mathcal{g}}}(\boldsymbol{\mathcal{w}}_{i-1})\right\|^4 \\ &\overset{(a)}{\leq} \left\|J_\epsilon^{\mathsf{T}}\right\|^4\left\|\mathcal{V}_R^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_{i-1} - \mu\mathcal{V}_R^{\mathsf{T}}\widehat{\boldsymbol{\mathcal{g}}}(\boldsymbol{\mathcal{w}}_{i-1})\right\|^4 \\ &\overset{(b)}{\leq} \left\|J_\epsilon^{\mathsf{T}}\right\|\left\|\mathcal{V}_R^{\mathsf{T}}\,\boldsymbol{\mathcal{w}}_{i-1}\right\|^4 + \mu^4\frac{\left\|J_\epsilon^{\mathsf{T}}\right\|^4}{\left(1-\left\|J_\epsilon^{\mathsf{T}}\right\|\right)^3}\left\|\mathcal{V}_R^{\mathsf{T}}\widehat{\boldsymbol{\mathcal{g}}}(\boldsymbol{\mathcal{w}}_{i-1})\right\|^4\end{aligned} \tag{5.38}$$

where (a) follows from the sub-multiplicative property of norms, and (b) follows from Jensen's inequality $\|a+b\|^4 \leq \frac{1}{\alpha^3}\|a\|^4 + \frac{1}{(1-\alpha)^3}\|b\|^4$ with

$$\alpha = \left\|J_\epsilon^{\mathsf{T}}\right\| \triangleq \sqrt{\rho\left(J_\epsilon J_\epsilon^{\mathsf{T}}\right)} \leq \sqrt{\left\|J_\epsilon J_\epsilon^{\mathsf{T}}\right\|_1} \leq \sqrt{\lambda_2^2+\epsilon^2} < 1 \tag{5.39}$$

for sufficiently small ϵ due to Assumption 5.1, where $\lambda_2 \triangleq \rho\left(A - \mathbb{1}p^{\mathsf{T}}\right)$. We observe that the term $\left\|\mathcal{V}_R^{\mathsf{T}}\boldsymbol{\mathcal{w}}_i\right\|^4$ contracts at an exponential rate given by $\left\|J_\epsilon^{\mathsf{T}}\right\| \approx \lambda_2$ for small ϵ, also known as the mixing rate of the graph. Iterating this relation and applying Assumptions 5.1–5.4, we obtain the following result. We note that similar results have been obtained before in the

literature, see for example [27] for strongly convex costs and extended later in [54] for the non-convex setting.

Theorem 5.1 (**Network disagreement (4th order)**). *Under assumptions 5.1–5.4, the network disagreement is bounded after sufficient iterations $i \geq i_o$ by:*

$$\begin{aligned}&\mathbb{E}\left\|\boldsymbol{\mathcal{w}}_i - \left(\mathbb{1}p^{\mathsf{T}} \otimes I\right)\boldsymbol{\mathcal{w}}_i\right\|^4 \\ &\leq \mu^4\|\mathcal{V}_L\|^4 \frac{\left\|J_\epsilon^{\mathsf{T}}\right\|^4}{\left(1-\|J_\epsilon^{\mathsf{T}}\|\right)^4}\|\mathcal{V}_R^{\mathsf{T}}\|^4 N^2\left(G^4+\sigma^4\right) + o(\mu^4)\end{aligned} \tag{5.40}$$

where

$$i_o = \frac{\log\left(o(\mu^4)\right)}{\log\left(\|J_\epsilon^{\mathsf{T}}\|\right)} \tag{5.41}$$

and $o(\mu^4)$ denotes a term that is higher in order than μ^4.

Proof. Appendix 5.A. □

Note again, that Jensen's inequality immediately implies for the second-order moment:

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{\mathcal{w}}_i - \left(\mathbb{1}p^{\mathsf{T}} \otimes I\right)\boldsymbol{\mathcal{w}}_i\right\|^2 &= \mathbb{E}\sqrt{\|\boldsymbol{\mathcal{w}}_i - \left(\mathbb{1}p^{\mathsf{T}} \otimes I\right)\boldsymbol{\mathcal{w}}_i\|^4} \\ &\leq \sqrt{\mathbb{E}\,\|\boldsymbol{\mathcal{w}}_i - \left(\mathbb{1}p^{\mathsf{T}} \otimes I\right)\boldsymbol{\mathcal{w}}_i\|^4} \\ &\overset{(a)}{\leq} \mu^2\|\mathcal{V}_L\|^2 \frac{\left\|J_\epsilon^{\mathsf{T}}\right\|^2}{\left(1-\|J_\epsilon^{\mathsf{T}}\|\right)^2}\|\mathcal{V}_R^{\mathsf{T}}\|^2 N\left(G^2+\sigma^2\right) + o(\mu^2)\end{aligned} \tag{5.42}$$

where (a) follows from (5.40) and sub-additivity of the square root, i.e. $\sqrt{x+y} \leq \sqrt{x}+\sqrt{y}$. This result establishes that, for every agent k, we have after sufficient iterations $i \geq i_o$:

$$\mathbb{E}\,\|\,\boldsymbol{w}_{k,i} - \boldsymbol{w}_{c,i}\,\|^2 \leq O(\mu^2) \tag{5.43}$$

or, by Markov's inequality [145]:

$$\Pr\left\{\|\,\boldsymbol{w}_{k,i} - \boldsymbol{w}_{c,i}\,\|^2 \geq O(\mu)\right\} \leq O(\mu) \tag{5.44}$$

and hence $\boldsymbol{w}_{k,i}$ will be arbitrarily close to $\boldsymbol{w}_{c,i}$ with arbitrarily high probability for all agents.

This result has two implications. First, it allows us to use the network centroid $\boldsymbol{w}_{c,i}$ as a proxy for all iterates $\boldsymbol{w}_{k,i}$ in the network, since all agents will cluster around the network centroid after sufficient iterations. Second, it allows us to bound the perturbation terms encountered in (5.33).

Lemma 5.2 (**Perturbation bounds (2nd and 4th order)**). *Under assumptions 5.1–5.4 and for sufficiently small step-sizes μ, the perturbation terms are bounded as:*

$$\left(\mathbb{E}\,\|\boldsymbol{d}_{i-1}\|^2\right)^2 \leq \mathbb{E}\,\|\boldsymbol{d}_{i-1}\|^4 \leq O(\mu^4) \tag{5.45}$$

$$\left(\mathbb{E}\left\{\|\boldsymbol{s}_i\|^2|\boldsymbol{\mathcal{F}}_{i-1}\right\}\right)^2 \leq \mathbb{E}\left\{\|\boldsymbol{s}_i\|^4|\boldsymbol{\mathcal{F}}_{i-1}\right\} \leq \sigma^4 \tag{5.46}$$

after sufficient iterations $i \geq i_0$.

Proof. Appendix 5.B. □

Definition 5.2 (Sets). *To simplify the notation in the sequel, we introduce following sets:*

$$\mathcal{G} \triangleq \left\{w : \|\nabla J(w)\|^2 \geq \mu\frac{c_2}{c_1}\left(1+\frac{1}{\pi}\right)\right\} \tag{5.47}$$

$$\mathcal{G}^C \triangleq \left\{w : \|\nabla J(w)\|^2 < \mu\frac{c_2}{c_1}\left(1+\frac{1}{\pi}\right)\right\} \tag{5.48}$$

$$\mathcal{H} \triangleq \left\{w : w \in \mathcal{G}^C, \lambda_{\min}\left(\nabla^2 J(w)\right) \leq -\tau\right\} \tag{5.49}$$

$$\mathcal{M} \triangleq \left\{w : w \in \mathcal{G}^C, \lambda_{\min}\left(\nabla^2 J(w)\right) > -\tau\right\} \tag{5.50}$$

where τ is a small positive parameter, c_1 and c_2 are constants:

$$c_1 \triangleq \frac{1}{2}\left(1-2\mu\delta\right) = O(1) \tag{5.51}$$

$$c_2 \triangleq \delta\sigma^2/2 = O(1) \tag{5.52}$$

and $0 < \pi < 1$ is a parameter to be chosen. Note that $\mathcal{G}^C = \mathcal{H} \cup \mathcal{M}$. We also define the probabilities $\pi_i^{\mathcal{G}} \triangleq \Pr\{\boldsymbol{w}_{c,i} \in \mathcal{G}\}$, $\pi_i^{\mathcal{H}} \triangleq \Pr\{\boldsymbol{w}_{c,i} \in \mathcal{H}\}$ and $\pi_i^{\mathcal{M}} \triangleq \Pr\{\boldsymbol{w}_{c,i} \in \mathcal{M}\}$. Then for all i, we have $\pi_i^{\mathcal{G}} + \pi_i^{\mathcal{H}} + \pi_i^{\mathcal{M}} = 1$. □

The definitions (5.47)–(5.50) decompose the parameter-space $\mathbb{R}^M$ into two disjoint sets $\mathcal{G}$ and $\mathcal{G}^C$, and further sub-divides $\mathcal{G}^C$ into $\mathcal{H}$ and $\mathcal{M}$. The set $\mathcal{G}$ denotes the set all points w where the norm of the gradient is large, while $\mathcal{G}^C = \mathcal{H} \cup \mathcal{M}$ denotes the set of all points where the norm of the gradient is small, i.e., approximately first-order stationary points. In a manner similar to related works on the escape from strict-saddle points, we further decompose the set $\mathcal{G}^C$ of approximate first-order stationary points into those points $w \in \mathcal{H}$ that do have a significant negative eigenvalue, and those in $\mathcal{M}$ that do not [59, 60, 144]. Points in the parameter space that have a small gradient norm and *no* significant negative eigenvalue are referred to as *second-order* stationary points, while points in $\mathcal{H}$ are known as *strict* saddle-points due to the presence of a strictly negative eigenvalue in the Hessian matrix. In the sequel, we will establish descent for centroids in $\mathcal{G}$ in Theorem 5.2 and centroids in $\mathcal{H}$ in Theorem 6.1, and hence the approach of a point in $\mathcal{M}$ with high probability after a polynomial number of iterations in Theorem 6.2. Second-order stationary points are generally more likely to be "good" minimizers than first-order stationary points, which could even correspond to local maxima. Furthermore, for a certain class of cost functions, known as "strict-saddle" functions, second-order stationary points always correspond to local minima for sufficiently small τ [59].

5.2.3 Evolution of the network centroid

Having established in (5.42), that after sufficient iterations, all agents in the network will have contracted around the centroid in a small cluster for small step-sizes, we can now leverage $\boldsymbol{w}_{c,i}$ as a proxy for all $\boldsymbol{w}_{k,i}$. From Assumption 5.2 and (5.7), we have the following bound:

$$J(\boldsymbol{w}_{c,i}) \leq J(\boldsymbol{w}_{c,i-1}) + \nabla J(\boldsymbol{w}_{c,i-1})^{\mathsf{T}} (\boldsymbol{w}_{c,i} - \boldsymbol{w}_{c,i-1}) + \frac{\delta}{2} \|\boldsymbol{w}_{c,i} - \boldsymbol{w}_{c,i-1}\|^2 \tag{5.53}$$

From (5.33), we then obtain:

$$\begin{aligned} J(\boldsymbol{w}_{c,i}) \leq J(\boldsymbol{w}_{c,i-1}) &- \mu\|\nabla J(\boldsymbol{w}_{c,i-1})\|^2 \\ &- \mu\nabla J(\boldsymbol{w}_{c,i-1})^{\mathsf{T}}\left(\boldsymbol{d}_{i-1}+\boldsymbol{s}_i\right) \\ &+ \mu^2\frac{\delta}{2}\|\nabla J(\boldsymbol{w}_{c,i-1})+\boldsymbol{d}_{i-1}+\boldsymbol{s}_i\|^2 \end{aligned} \tag{5.54}$$

This relation, along with (5.33) and the results from Lemma 5.2, allow us to establish the following theorem.

Theorem 5.2 (**Descent relation**). *Beginning at $\boldsymbol{w}_{c,i-1}$ in the large gradient regime $\mathcal{G}$, we can bound:*

$$\mathbb{E}\left\{J(\boldsymbol{w}_{c,i})\middle|\,\boldsymbol{w}_{c,i-1}\in\mathcal{G}\right\} \leq \mathbb{E}\left\{J(\boldsymbol{w}_{c,i-1})\middle|\,\boldsymbol{w}_{c,i-1}\in\mathcal{G}\right\} - \mu^2\frac{c_2}{\pi} + \frac{O(\mu^3)}{\pi_{i-1}^{\mathcal{G}}} \tag{5.55}$$

as long as $\pi_{i-1}^{\mathcal{G}} = \Pr\{\boldsymbol{w}_{c,i-1}\in\mathcal{G}\} \neq 0$ where the relevant constants are listed in definition 5.2. On the other hand, beginning at $\boldsymbol{w}_{c,i-1}\in\mathcal{M}$, we can bound:

$$\mathbb{E}\left\{J(\boldsymbol{w}_{c,i})\middle|\,\boldsymbol{w}_{c,i-1}\in\mathcal{M}\right\} \leq \mathbb{E}\left\{J(\boldsymbol{w}_{c,i-1})\middle|\,\boldsymbol{w}_{c,i-1}\in\mathcal{M}\right\} + \mu^2 c_2 + \frac{O(\mu^3)}{\pi_{i-1}^{\mathcal{M}}} \tag{5.56}$$

as long as $\pi_{i-1}^{\mathcal{M}} = \Pr\{\boldsymbol{w}_{c,i-1}\in\mathcal{M}\} \neq 0$.

Proof. Appendix 5.D. □

Relation (5.55) guarantees a lower bound on the expected improvement when the gradient norm at the current iterate is sufficiently large, i.e. $\boldsymbol{w}_{c,i-1}\in\mathcal{G}$ is not an approximately first-order stationary point. On the other hand, when $\boldsymbol{w}_{c,i-1}\in\mathcal{M}$, inequality (5.56) it establishes an upper bound on the expected ascent. The respective bounds can be balanced by appropriately choosing π, which will be leveraged in Chapter 6. We are left to treat the third possibility, namely $\boldsymbol{w}_{c,i-1}\in\mathcal{H}$. In this case, since the norm of the gradient is small, it is no longer possible to guarantee descent in a single iteration. We shall study the dynamics in more detail in the sequel.

5.2.4 Behavior around stationary points

In the vicinity of saddle-points, the norm of the gradient is not sufficiently large to guarantee descent at every iteration as indicated by (5.55). Instead, we will study the cumulative effect of the gradient, as well as perturbations, over several iterations. For this purpose, we introduce the following second-order condition on the cost functions, which is common in the literature [1, 59, 60].

Assumption 5.5 (Lipschitz Hessians). *Each $J_k(\cdot)$ is twice-differentiable with Hessian $\nabla^2 J_k(\cdot)$ and, there exists $\rho \geq 0$ such that:*

$$\|\nabla^2 J_k(x) - \nabla^2 J_k(y)\| \leq \rho \|x - y\| \tag{5.57}$$

By Jensen's inequality, this implies that $J(\cdot) = \sum_{k=1}^{N} p_k J_k(\cdot)$ also satisfies:

$$\|\nabla^2 J(x) - \nabla^2 J(y)\| \leq \rho \|x - y\| \tag{5.58}$$

□

Let $i^\star$ denote an arbitrary point in time. We use $i^\star$ in order to emphasize approximately first-order stationary points, where the norm of the gradient is small. Such first-order stationary points $\boldsymbol{w}_{c,i^\star} \in \mathcal{G}^C$ could either be in the set of second-order stationary points $\mathcal{M}$ or in the set of strict-saddle points $\mathcal{H}$. Our objective is to show that when $\boldsymbol{w}_{c,i^\star} \in \mathcal{H}$, we can guarantee descent after several iterations. To this end, starting at $i^\star$, we have for $i \geq 0$:

$$\boldsymbol{w}_{c,i^\star+i+1} = \boldsymbol{w}_{c,i^\star+i} - \mu \nabla J(\boldsymbol{w}_{c,i^\star+i}) - \mu \boldsymbol{d}_{i^\star+i} - \mu\, \boldsymbol{s}_{i^\star+i+1} \tag{5.59}$$

Subsequent analysis will rely on an auxiliary model, referred to as a short-term model. It will be seen that this model is more tractable and evolves "close" to the true recursion under the second-order smoothness condition on the Hessian matrix (5.58) and as long as the iterates remain close to a stationary point. A similar approach has been introduced and used to great advantage in the form of a "long-term model" to derive accurate mean-square

deviation performance expressions for strongly-convex costs in [1, 22, 32, 146]. The approach was also used to provide a "quadratic approximation" to establish the ability of stochastic gradient based algorithms to escape from strict saddle-points in the single-agent case under i.i.d. perturbations in [59].

For the driving gradient term in (5.59), we have from the mean-value theorem [1]:

$$\nabla J(\boldsymbol{w}_{c,i^\star+i}) - \nabla J(\boldsymbol{w}_{c,i^\star}) = \boldsymbol{H}_{i^\star+i}\,(\boldsymbol{w}_{c,i^\star+i} - \boldsymbol{w}_{c,i^\star}) \tag{5.60}$$

where

$$\boldsymbol{H}_{i^\star+i} \triangleq \int_0^1 \nabla^2 J\left((1-t)\,\boldsymbol{w}_{c,i^\star+i} + t\,\boldsymbol{w}_{c,i^\star}\right) dt \tag{5.61}$$

Subtracting (5.59) from $\boldsymbol{w}_{c,i^\star}$, we obtain:

$$\begin{aligned}\boldsymbol{w}_{c,i^\star} - \boldsymbol{w}_{c,i^\star+i+1} &= \boldsymbol{w}_{c,i^\star} - \boldsymbol{w}_{c,i^\star+i} + \mu\nabla J(\boldsymbol{w}_{c,i^\star+i}) + \mu\boldsymbol{d}_{i^\star+i} + \mu\,\boldsymbol{s}_{i^\star+i+1} \\ &= (I - \mu\boldsymbol{H}_{i^\star+i})\,(\boldsymbol{w}_{c,i^\star} - \boldsymbol{w}_{c,i^\star+i}) + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) \\ &\quad + \mu\boldsymbol{d}_{i^\star+i} + \mu\,\boldsymbol{s}_{i^\star+i+1}\end{aligned} \tag{5.62}$$

We introduce short-hand notation for the deviation:

$$\widetilde{\boldsymbol{w}}_i^{\,i^\star} \triangleq \boldsymbol{w}_{c,i^\star} - \boldsymbol{w}_{c,i^\star+i} \tag{5.63}$$

Note that $\widetilde{\boldsymbol{w}}_i^{\,i^\star}$ denotes the deviation of the network centroid $\boldsymbol{w}_{c,i^\star+i}$ at time $i^\star + i$ from the initial, approximately first-order stationary point $\boldsymbol{w}_{c,i^\star}$. Establishing escape from saddle-points is equivalent to establishing the growth of $\widetilde{\boldsymbol{w}}_i^{\,i^\star}$ whenever $\boldsymbol{w}_{c,i^\star} \in \mathcal{H}$. We hence expect the deviation to grow over time, but would like to establish that $\boldsymbol{w}_{c,i^\star+i}$ moves away from $\boldsymbol{w}_{c,i^\star}$ in a direction of descent. We can then write more compactly:

$$\begin{aligned}\widetilde{\boldsymbol{w}}_{i+1}^{\,i^\star} = &\ (I - \mu\boldsymbol{H}_{i^\star+i})\,\widetilde{\boldsymbol{w}}_i^{\,i^\star} + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) \\ &+ \mu\boldsymbol{d}_{i^\star+i} + \mu\,\boldsymbol{s}_{i^\star+i+1}\end{aligned} \tag{5.64}$$

The time-varying nature of $\boldsymbol{H}_{i^\star+i}$ makes this recursion difficult to study. We hence introduce the following auxiliary recursion, initialized at $\boldsymbol{w}'_{c,i^\star} = \boldsymbol{w}_{c,i^\star}$, where $\boldsymbol{H}_{i^\star+i}$ is replaced by $\nabla^2 J(\boldsymbol{w}_{c,i^\star})$ and the perturbation term $\mu \boldsymbol{d}_{i^\star+i}$ is omitted:

$$\begin{aligned}\boldsymbol{w}_{c,i^\star} - \boldsymbol{w}'_{c,i^\star+i+1} = &\left(I - \mu\nabla^2 J(\boldsymbol{w}_{c,i^\star})\right)\left(\boldsymbol{w}_{c,i^\star} - \boldsymbol{w}'_{c,i^\star+i}\right) \\ &+ \mu\nabla J(\boldsymbol{w}_{c,i^\star}) + \mu\,\boldsymbol{s}_{i^\star+i+1}\end{aligned} \tag{5.65}$$

or, more compactly, with $\widetilde{\boldsymbol{w}}_i^{\prime i^\star} \triangleq \boldsymbol{w}_{c,i^\star} - \boldsymbol{w}'_{c,i^\star+i}$

$$\widetilde{\boldsymbol{w}}_{i+1}^{\prime i^\star} = \left(I - \mu\nabla^2 J(\boldsymbol{w}_{c,i^\star})\right)\widetilde{\boldsymbol{w}}_i^{\prime i^\star} + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) + \mu\,\boldsymbol{s}_{i^\star+i+1} \tag{5.66}$$

Of course, this second model is only useful in studying the behavior of the original recursion (5.59) if the iterates generated by both models remain close to each other, which we shall prove to be true. Specifically, if we write:

$$\boldsymbol{w}'_{i^\star+i+1} = \boldsymbol{w}_{i^\star+i+1} + \boldsymbol{u}_{i^\star+i+1} \tag{5.67}$$

then $\boldsymbol{u}_{i^\star+i+1}$ will be shown to be negligible in some sense. Results along this line have been established in the centralized and distributed contexts for strongly-convex costs [1, 22] and in the centralized setting for strict saddle points [59]. We show here that this conclusion holds more generally in the vicinity of $O(\mu)$-first-order stationary points. Before establishing deviation bounds, we establish a short lemma which will be used repeatedly.

Lemma 5.3 (A limiting result)**.** *For $T, \mu, \delta > 0$ and $k \in \mathbb{Z}_+$ with $\mu < \frac{1}{\delta}$, we have:*

$$\lim_{\mu\to 0}\left(\frac{(1+\mu\delta)^k}{(1-\mu\delta)^{k-1}}\right)^{\frac{T}{\mu}} = e^{-T\delta+2kT\delta} = O(1) \tag{5.68}$$

Proof. Appendix 5.C. □

Lemma 5.4 (Deviation bounds)**.** *Suppose* $\Pr\{\boldsymbol{w}_{c,i^\star} \in \mathcal{H}\} \neq 0$. *Then, the following quanti-*

ties are conditionally bounded:

$$\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \leq O(\mu) + \frac{O(\mu^2)}{\pi_{i^\star}^{\mathcal{H}}} \tag{5.69}$$

$$\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|^3 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \leq O(\mu^{3/2}) + \frac{O(\mu^3)}{\pi_{i^\star}^{\mathcal{H}}} \tag{5.70}$$

$$\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|^4 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \leq O(\mu^2) + \frac{O(\mu^4)}{\pi_{i^\star}^{\mathcal{H}}} \tag{5.71}$$

$$\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i^{i^\star} - \widetilde{\boldsymbol{w}}_i'^{i^\star}\right\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \leq O(\mu^2) + \frac{O(\mu^2)}{\pi_{i^\star}^{\mathcal{H}}} \tag{5.72}$$

$$\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i'^{i^\star}\right\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \leq O(\mu) + \frac{O(\mu^2)}{\pi_{i^\star}^{\mathcal{H}}} \tag{5.73}$$

for $i \leq \frac{T}{\mu}$, where T denotes an arbitrary constant that is independent of the step-size μ.

Proof. Appendix 5.E. □

These deviation bounds establish that, beginning at a strict-saddle point $\boldsymbol{w}_{c,i^\star}$ at time $i^\star$ the iterates will remain close to $\boldsymbol{w}_{c,i^\star}$ for the next $O(1/\mu)$ iterations. Consequently, the short-term model will be sufficiently accurate for the next $O(1/\mu)$ iterations. We will establish formally in Chapter 6 that the small-deviation bounds in Lemma 5.4 ensure descent of the true recursion can be inferred by studying only the evolution of the short-term model, which is significantly more tractable.

5.3 Application: Robust Regression

Consider a scenario where each agent k in the network observes streaming realizations $\{\boldsymbol{\gamma}(k,i), \boldsymbol{h}_{k,i}\}$ from the linear model $\boldsymbol{\gamma}(k) = \boldsymbol{h}_k^{\mathsf{T}} w^o + \boldsymbol{v}(k)$ where $\boldsymbol{\gamma}(k)$ denotes scalar observations and $\boldsymbol{v}(k)$ denotes measurement noise. One common approach for estimating w^o in a distributed setting is via least-mean-square error estimation, resulting in the local cost functions:

$$J_k^{\mathrm{MSE}}(w) = \mathbb{E}\left\|\boldsymbol{\gamma}(k) - \boldsymbol{h}_k^{\mathsf{T}} w\right\|^2 \tag{5.74}$$

The resulting problem is convex and has been studied extensively in the literature. While effective under the assumption of Gaussian noise, and similar well-behaved noise conditions, this approach is susceptible to outliers caused by heavy-tailed distributions for $\boldsymbol{v}(k)$ [130]. This is caused by the fact that the quadratic risk penalizes errors proportionally to their squared norm, and as such has a tendency to over-correct outliers, even if they are rare. Several alternative robust cost functions have been suggested in the literature. We consider two in particular in order to illustrate the advantages of allowing for non-convex costs in the context of robust estimation, namely the Huber loss $Q_k^{\mathrm{H}}(w;\boldsymbol{x}_k)$ and Tukey's biweight loss $Q_k^{\mathrm{B}}(w;\boldsymbol{x}_k)$ [130]. For ease of notation, let $\boldsymbol{e}(w) \triangleq \boldsymbol{\gamma}(k) - \boldsymbol{h}_k^{\mathsf{T}} w$. Then:

$$Q_k^{\mathrm{H}}(w;\boldsymbol{x}_k) = \begin{cases} \frac{1}{2}|\boldsymbol{e}(w)|^2, & \text{for } |\boldsymbol{e}(w)| \leq c_H \\ c_H|\boldsymbol{e}(w)| - \frac{1}{2}c_H^2, & \text{for } |\boldsymbol{e}(w)| > c_H. \end{cases} \tag{5.75}$$

$$Q_k^{\mathrm{B}}(w;\boldsymbol{x}_k) = \begin{cases} \frac{c_B^2}{6}\left(1 - \left(1 - \frac{|\boldsymbol{e}(w)|^2}{c_B^2}\right)^3\right), & \text{for } |\boldsymbol{e}(w)| \leq c_B \\ \frac{c_B^2}{6} & \text{otherwise} \end{cases} \tag{5.76}$$

where c_H, c_B are tuning constants. The Huber cost is merely convex (and not strongly-convex), while the Tukey loss is non-convex. Both losses satisfy assumptions 5.1–5.4 imposed in this chapter. In particular, since the Huber risk $J_k^{\mathrm{H}}(w)$ has a unique, local minimum, which also happens to be locally strongly-convex, we can conclude that despite the absence of strong-convexity, the algorithm will converge to within $O(\mu)$ of the global minimum. The Tukey loss on the other hand, is non-convex, and is therefore a more challenging problem. The setting for the simulation results is shown in Figures 5.2–5.3.

Performance is illustrated in Fig. 5.4–5.5. We first show the performance of each cost in the nominal scenario, where $\boldsymbol{v}(k) \sim \mathcal{N}(0, \sigma_v^2)$. We observe that the distributed strategies outperform the non-cooperative ones, and that despite differences in the rate of convergence, there is negligible difference in the performance of the mean-square-error, Huber and Tukey variations. In the presence of outliers, modeled as a bimodal distribution with $\boldsymbol{v}(k) \sim (1-\epsilon)\mathcal{N}(0, \sigma_v^2) + \epsilon\mathcal{N}(10, \sigma_v^2)$ and $\epsilon = 0.1$, the performance of the mean-square-error solution

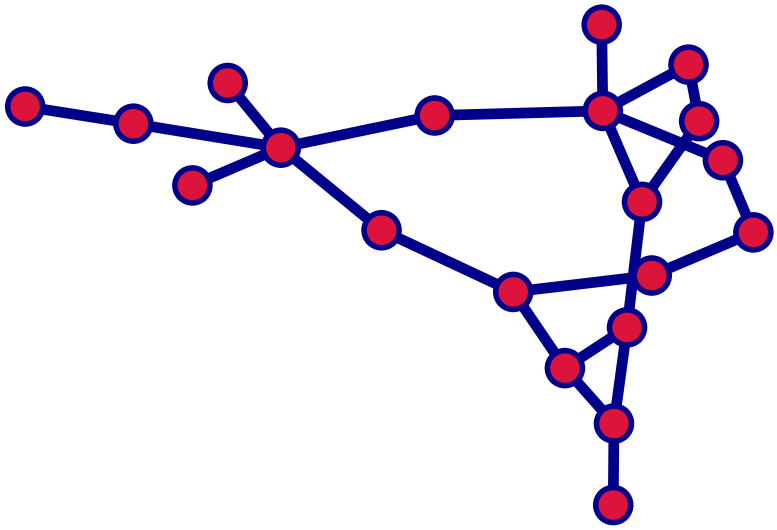

Figure 5.2: Graph with $N = 20$ nodes.

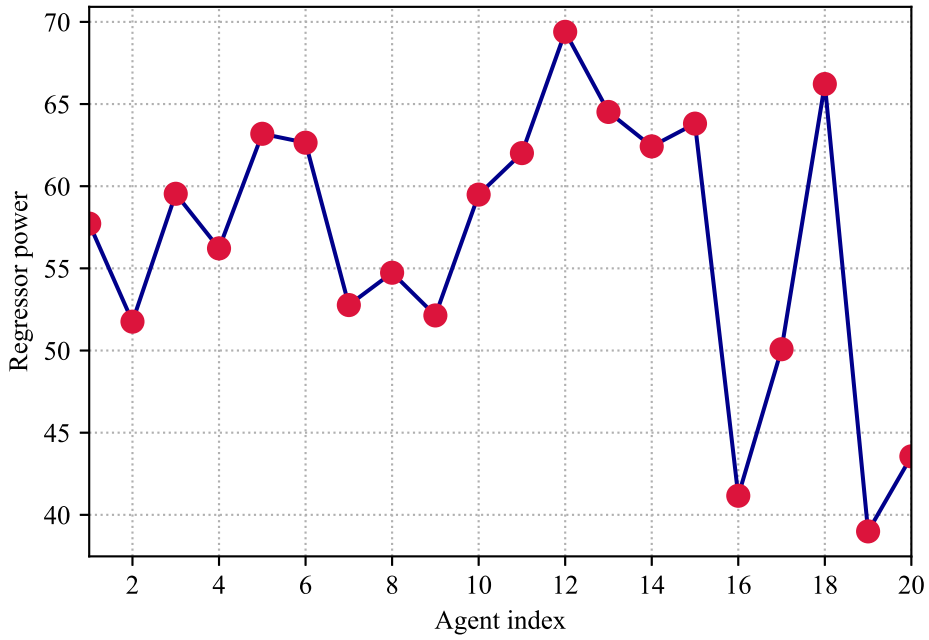

Figure 5.3: Regressor power $\mathrm{Tr}\,(R_{h,k})$ at each agent.

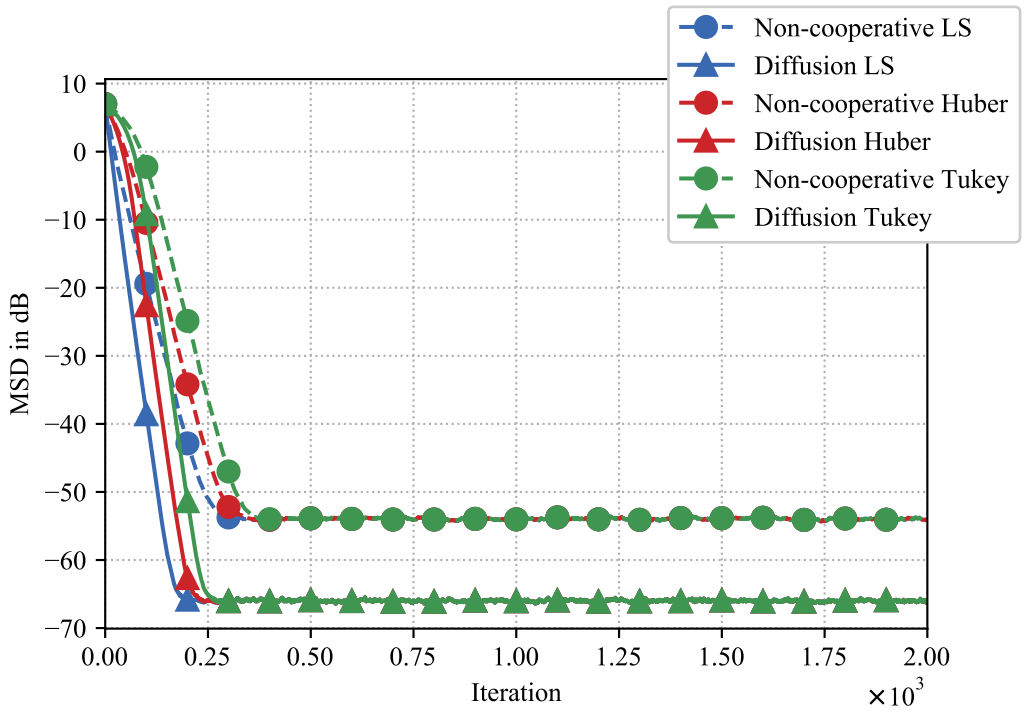

Figure 5.4: Performance in the nominal case.

dramatically deteriorates, as is to be expected in the presence of deviations from the nominal model.

5.A Proof of Lemma 5.1

Starting from (5.36), taking norms of both sides and computing the fourth power, we find:

$$\begin{aligned}\left\|\mathcal{V}_R^{\mathsf{T}}\boldsymbol{w}_i\right\|^4 &= \left\|J_\epsilon^{\mathsf{T}}\mathcal{V}_R^{\mathsf{T}}\boldsymbol{w}_{i-1} + \mu J_\epsilon^{\mathsf{T}}\mathcal{V}_R^{\mathsf{T}}\widehat{\boldsymbol{g}}(\boldsymbol{w}_{i-1})\right\|^4 \\ &\leq \left\|J_\epsilon^{\mathsf{T}}\right\|^4\left\|\mathcal{V}_R^{\mathsf{T}}\boldsymbol{w}_{i-1} + \mu\mathcal{V}_R^{\mathsf{T}}\widehat{\boldsymbol{g}}(\boldsymbol{w}_{i-1})\right\|^4 \\ &\overset{(a)}{\leq} \left\|J_\epsilon^{\mathsf{T}}\right\|\left\|\mathcal{V}_R^{\mathsf{T}}\boldsymbol{w}_{i-1}\right\|^4 + \mu^4\frac{\left\|J_\epsilon^{\mathsf{T}}\right\|^4}{\left(1-\left\|J_\epsilon^{\mathsf{T}}\right\|\right)^3}\left\|\mathcal{V}_R^{\mathsf{T}}\widehat{\boldsymbol{g}}(\boldsymbol{w}_{i-1})\right\|^4 \end{aligned} \tag{5.77}$$

where step (a) follows from convexity of $\|\cdot\|^4$ and Jensen's inequality, i.e. $\|a+b\|^4 = \frac{1}{\alpha^3}\|a\|^4 + \frac{1}{(1-\alpha)^3}\|b\|^4$. To begin with, we study the stochastic gradient term in some greater

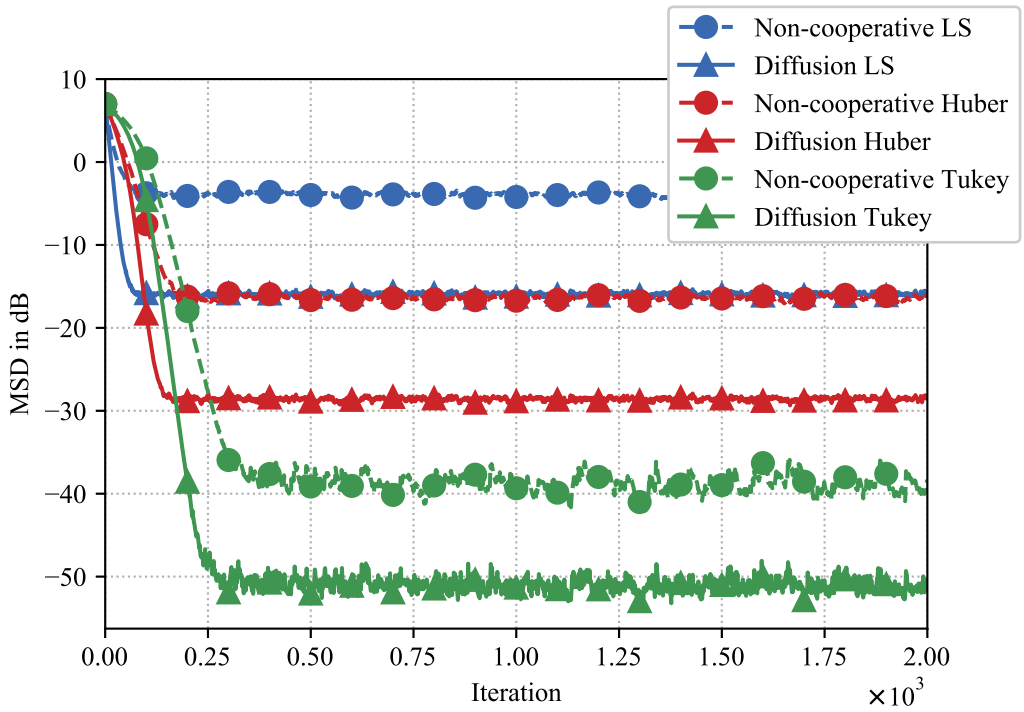

Figure 5.5: Performance in the corrupted case.

detail. We have:

$$\begin{aligned}\left\|\mathcal{V}_R^{\mathsf{T}}\widehat{\boldsymbol{g}}(\boldsymbol{\mathcal{W}}_{i-1})\right\|^4 &= \left\|\mathcal{V}_R^{\mathsf{T}}g(\boldsymbol{\mathcal{W}}_{i-1}) + \mathcal{V}_R^{\mathsf{T}}\mathrm{col}\left\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right\}\right\|^4 \\ &\leq 8\left\|\mathcal{V}_R^{\mathsf{T}}g(\boldsymbol{\mathcal{W}}_{i-1})\right\|^4 + 8\left\|\mathcal{V}_R^{\mathsf{T}}\mathrm{col}\left\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right\}\right\|^4 \end{aligned} \tag{5.78}$$

For the first term we have:

$$\begin{aligned}8\left\|\mathcal{V}_R^{\mathsf{T}}g(\boldsymbol{\mathcal{W}}_{i-1})\right\|^4 &\overset{(a)}{=} 8\left\|\mathcal{V}_R^{\mathsf{T}}g(\boldsymbol{\mathcal{W}}_{i-1}) - \left(\mathbb{1}p^{\mathsf{T}}\otimes I\right)g(\boldsymbol{\mathcal{W}}_{i-1})\right\|^4 \\ &\overset{(b)}{\leq} 8\left\|\mathcal{V}_R^{\mathsf{T}}\right\|^4\left\|g(\boldsymbol{\mathcal{W}}_{i-1}) - \left(\mathbb{1}p^{\mathsf{T}}\otimes I\right)g(\boldsymbol{\mathcal{W}}_{i-1})\right\|^4 \\ &\overset{(c)}{=} 8\left\|\mathcal{V}_R^{\mathsf{T}}\right\|^4\left(\sum_{k=1}^{N}\left\|\nabla J_k(\boldsymbol{w}_{k,i-1}) - \nabla J(\boldsymbol{w}_{k,i-1})\right\|^2\right)^2 \\ &\overset{(5.9)}{\leq} 8\left\|\mathcal{V}_R^{\mathsf{T}}\right\|^4\left(\sum_{k=1}^{N}G^2\right)^2 \leq 8\left\|\mathcal{V}_R^{\mathsf{T}}\right\|^4N^2G^4 \end{aligned} \tag{5.79}$$

where (a) follows from the fact that (5.30) implies $V_R^{\mathsf{T}}\mathbb{1} = 0$, (b) follows from the sub-multiplicity of norms and (c) expands $\|\cdot\|^2$. For the gradient noise term we find under

expectation:

$$
\begin{aligned}
8\,\mathbb{E}\left\|\mathcal{V}_R^{\mathsf{T}}\mathrm{col}\left\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right\}\right\|^4 &= 8\left\|\mathcal{V}_R^{\mathsf{T}}\right\|^4\mathbb{E}\left\|\mathrm{col}\left\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right\}\right\|^4 \\
&= 8\left\|\mathcal{V}_R^{\mathsf{T}}\right\|^4\mathbb{E}\left(\sum_{k=1}^{N}\left\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right\|^2\right)^2 \\
&\overset{(a)}{\leq} 8\left\|\mathcal{V}_R^{\mathsf{T}}\right\|^4 N\sum_{k=1}^{N}\mathbb{E}\left\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right\|^4 \\
&\overset{(5.20b)}{\leq} 8\left\|\mathcal{V}_R^{\mathsf{T}}\right\|^4 N\sum_{k=1}^{N}\sigma^4 = 8\left\|\mathcal{V}_R^{\mathsf{T}}\right\|^4 N^2\sigma^4
\end{aligned}
\tag{5.80}
$$

where (a) follows from Cauchy-Schwarz, which implies $\left(\sum_{k=1}^{N}x_k\right)^2 \leq N\sum_{k=1}^{N}x_k^2$. Plugging these relations back into (5.77), we obtain:

$$
\mathbb{E}\left\|\mathcal{V}_R^{\mathsf{T}}\boldsymbol{\mathcal{W}}_i\right\|^4 \leq \left\|J_\epsilon^{\mathsf{T}}\right\|\mathbb{E}\left\|\mathcal{V}_R^{\mathsf{T}}\boldsymbol{\mathcal{W}}_{i-1}\right\|^4 + \mu^4\frac{8\left\|J_\epsilon^{\mathsf{T}}\right\|^4}{\left(1-\|J_\epsilon^{\mathsf{T}}\|\right)^3}\|\mathcal{V}_R^{\mathsf{T}}\|^4 N^2\left(G^4+\sigma^4\right) \tag{5.81}
$$

We can iterate, starting from $i=0$, to obtain:

$$
\begin{aligned}
\mathbb{E}\left\|\mathcal{V}_R^{\mathsf{T}}\boldsymbol{\mathcal{W}}_i\right\|^4 &\leq \left\|J_\epsilon^{\mathsf{T}}\right\|^i\mathbb{E}\left\|\mathcal{V}_R^{\mathsf{T}}\mathcal{W}_0\right\|^4 + \mu^4\frac{8\left\|J_\epsilon^{\mathsf{T}}\right\|^4}{\left(1-\|J_\epsilon^{\mathsf{T}}\|\right)^3}\|\mathcal{V}_R^{\mathsf{T}}\|^4 N^2\left(G^4+\sigma^4\right)\sum_{n=1}^{i}\|J_\epsilon^{\mathsf{T}}\|^{n-1} \\
&\overset{(a)}{\leq} \left\|J_\epsilon^{\mathsf{T}}\right\|^i\mathbb{E}\left\|\mathcal{V}_R^{\mathsf{T}}\mathcal{W}_0\right\|^4 + \mu^4\frac{8\left\|J_\epsilon^{\mathsf{T}}\right\|^4}{\left(1-\|J_\epsilon^{\mathsf{T}}\|\right)^4}\|\mathcal{V}_R^{\mathsf{T}}\|^4 N^2\left(G^4+\sigma^4\right) \\
&\overset{(b)}{\leq} o(\mu^4) + \mu^4\frac{8\left\|J_\epsilon^{\mathsf{T}}\right\|^4}{\left(1-\|J_\epsilon^{\mathsf{T}}\|\right)^4}\|\mathcal{V}_R^{\mathsf{T}}\|^4 N^2\left(G^4+\sigma^4\right)
\end{aligned}
\tag{5.82}
$$

where (a) follows from $\sum_{n=1}^{i}\|J_\epsilon^{\mathsf{T}}\|^{n-1} \leq \sum_{n=1}^{\infty}\|J_\epsilon^{\mathsf{T}}\|^{n-1} = \left(1-\|J_\epsilon^{\mathsf{T}}\|\right)^{-1}$, and (b) holds whenever:

$$
\begin{aligned}
&\left\|J_\epsilon^{\mathsf{T}}\right\|^i\mathbb{E}\left\|\mathcal{V}_R^{\mathsf{T}}\mathcal{W}_0\right\|^4 \leq o(\mu^4) \iff \left\|J_\epsilon^{\mathsf{T}}\right\|^i \leq o(\mu^4) \\
\iff & i\log\left(\left\|J_\epsilon^{\mathsf{T}}\right\|\right) \leq \log\left(o(\mu^4)\right) \iff i \geq \frac{\log\left(o(\mu^4)\right)}{\log\left(\|J_\epsilon^{\mathsf{T}}\|\right)}
\end{aligned}
\tag{5.83}
$$

Finally, we have from (5.37) under (5.83):

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{\mathcal{W}}_i-\left(\mathbb{1}p^{\mathsf{T}}\otimes I\right)\boldsymbol{\mathcal{W}}_i\right\|^4 &= \mathbb{E}\left\|\mathcal{V}_L\mathcal{V}_R^{\mathsf{T}}\boldsymbol{\mathcal{W}}_i\right\|^4\\ &\overset{(a)}{\leq} \|\mathcal{V}_L\|^4\,\mathbb{E}\left\|\mathcal{V}_R^{\mathsf{T}}\boldsymbol{\mathcal{W}}_i\right\|^4\\ &\overset{(5.82)}{\leq} \mu^4\|\mathcal{V}_L\|^4\frac{\left\|J_\epsilon^{\mathsf{T}}\right\|^4}{\left(1-\|J_\epsilon^{\mathsf{T}}\|\right)^4}\|\mathcal{V}_R^{\mathsf{T}}\|^4N^2\left(G^4+\sigma^4\right)+o(\mu^4)\end{aligned}\tag{5.84}$$

where (a) follows from the sub-multiplicative property of norms.We conclude that all agents in the network will contract around the centroid vector $\left(\mathbb{1}p^{\mathsf{T}}\otimes I\right)\boldsymbol{\mathcal{W}}_i$ after sufficient iterations.

5.B Proof of Lemma 5.2

We begin by studying the perturbation term $\boldsymbol{s}_i$. We have:

$$\begin{aligned}\mathbb{E}\left\{\|\boldsymbol{s}_i\|^4|\boldsymbol{\mathcal{F}}_{i-1}\right\} &= \mathbb{E}\left\{\left\|\sum_{k=1}^N p_k\left(\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1})-\nabla J_k(\boldsymbol{w}_{k,i-1})\right)\right\|^4|\boldsymbol{\mathcal{F}}_{i-1}\right\}\\ &\overset{(a)}{\leq} \sum_{k=1}^N p_k\,\mathbb{E}\left\{\left\|\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1})-\nabla J_k(\boldsymbol{w}_{k,i-1})\right\|^4|\boldsymbol{\mathcal{F}}_{i-1}\right\}\\ &\overset{(b)}{\leq} \sum_{k=1}^N p_k\sigma^4 = \sigma^4\end{aligned}\tag{5.85}$$

where (a) follows from $\sum_{k=1}^{N} p_k = 1$ and Jensen's inequality and (b) follows from the fourth-order moment condition in Assumption 5.4. For the second perturbation term, we have

$$\begin{aligned}
\|\boldsymbol{d}_{i-1}\|^4 &= \left\| \sum_{k=1}^{N} p_k \left(\nabla J_k(\boldsymbol{w}_{k,i-1}) - \nabla J_k(\boldsymbol{w}_{c,i-1}) \right) \right\|^4 \\
&\overset{(a)}{\leq} \sum_{k=1}^{N} p_k \|\nabla J_k(\boldsymbol{w}_{k,i-1}) - \nabla J_k(\boldsymbol{w}_{c,i-1})\|^4 \\
&\overset{(b)}{\leq} \delta^4 \sum_{k=1}^{N} p_k \|\boldsymbol{w}_{k,i-1} - \boldsymbol{w}_{c,i-1}\|^4 \\
&\leq \delta^4 p_{\max} \sum_{k=1}^{N} \|\boldsymbol{w}_{k,i-1} - \boldsymbol{w}_{c,i-1}\|^4 \\
&\leq \delta^4 p_{\max} \left(\sum_{k=1}^{N} \|\boldsymbol{w}_{k,i-1} - \boldsymbol{w}_{c,i-1}\|^2 \right)^2 \\
&= \delta^4 p_{\max} \|\boldsymbol{\mathcal{w}}_{i-1} - \boldsymbol{\mathcal{w}}_{c,i-1}\|^4
\end{aligned} \tag{5.86}$$

where (a) again follows from Jensen's inequality, (b) follows from the Lipschitz gradient condition in Assumption 5.2, and we introduced $\boldsymbol{\mathcal{w}}_{c,i-1} \triangleq \mathbb{1} \otimes \boldsymbol{w}_{c,i-1}$. Result (5.45) follows by applying (5.84) to (5.86).

5.C Proof of Lemma 5.3

For the natural logarithm of the expression, we have:

$$\begin{aligned}
&\log \left(\frac{(1+\mu\delta)^k}{(1-\mu\delta)^{k-1}} \right)^{\frac{T}{\mu}} \\
&= \frac{T}{\mu} \left(k \log(1+\mu\delta) - (k-1) \log(1-\mu\delta) \right)
\end{aligned} \tag{5.87}$$

Since the logarithm is continuous over $\mathbb{R}_+$, we have:

$$\begin{aligned}
&\log\left(\lim_{\mu\to 0}\left(\frac{(1+\mu\delta)^k}{(1-\mu\delta)^{k-1}}\right)^{\frac{T}{\mu}}\right)\\
&=\lim_{\mu\to 0}\log\left(\left(\frac{(1+\mu\delta)^k}{(1-\mu\delta)^{k-1}}\right)^{\frac{T}{\mu}}\right)\\
&=\lim_{\mu\to 0}\frac{T}{\mu}\left(k\log\left(1+\mu\delta\right)-(k-1)\log\left(1-\mu\delta\right)\right)\\
&=kT\lim_{\mu\to 0}\frac{\log\left(1+\mu\delta\right)}{\mu}-(k-1)T\lim_{\mu\to 0}\frac{\log\left(1-\mu\delta\right)}{\mu}
\end{aligned}\tag{5.88}$$

We examine the fraction inside the limit more closely. Since both the numerator and denominator of the fraction approach zero as $\mu\to 0$, we apply L'Hôpital's rule:

$$\lim_{\mu\to 0}\frac{\log\left(1\pm\mu\delta\right)}{\mu}=\lim_{\mu\to 0}\frac{\pm\delta}{1\pm\mu\delta}=\pm\delta\tag{5.89}$$

Hence, we find:

$$\lim_{\mu\to 0}\left(\frac{(1+\mu\delta)^k}{(1-\mu\delta)^{k-1}}\right)^{\frac{T}{\mu}}=e^{kT\delta+(k-1)T\delta}=e^{-T\delta+2kT\delta}\tag{5.90}$$

5.D Proof of Theorem 5.2

We begin with (5.54) and take expectations conditioned on $\boldsymbol{\mathcal{W}}_{i-1}$ to obtain:

$$
\begin{aligned}
&\mathbb{E}\left\{J(\boldsymbol{w}_{c,i})\middle|\,\boldsymbol{\mathcal{W}}_{i-1}\right\} \\
&\overset{(a)}{\leq} J(\boldsymbol{w}_{c,i-1}) - \mu\|\nabla J(\boldsymbol{w}_{c,i-1})\|^2 - \mu\nabla J(\boldsymbol{w}_{c,i-1})^{\mathsf{T}}\boldsymbol{d}_{i-1} \\
&\quad + \mu^2\frac{\delta}{2}\|\nabla J(\boldsymbol{w}_{c,i-1}) + \boldsymbol{d}_{i-1}\|^2 + \mu^2\frac{\delta}{2}\,\mathbb{E}\left\{\|\boldsymbol{s}_i\|^2\middle|\,\boldsymbol{\mathcal{W}}_{i-1}\right\} \\
&\overset{(b)}{\leq} J(\boldsymbol{w}_{c,i-1}) - \mu\|\nabla J(\boldsymbol{w}_{c,i-1})\|^2 + \frac{\mu}{2}\|\nabla J(\boldsymbol{w}_{c,i-1})\|^2 \\
&\quad + \frac{\mu}{2}\|\boldsymbol{d}_{i-1}\|^2 + \mu^2\delta\|\nabla J(\boldsymbol{w}_{c,i-1})\|^2 + \mu^2\delta\|\boldsymbol{d}_{i-1}\|^2 \\
&\quad + \mu^2\frac{\delta}{2}\,\mathbb{E}\left\{\|\boldsymbol{s}_i\|^2\middle|\,\boldsymbol{\mathcal{W}}_{i-1}\right\} \\
&\overset{(c)}{\leq} J(\boldsymbol{w}_{c,i-1}) - \frac{\mu}{2}\left(1-2\mu\delta\right)\|\nabla J(\boldsymbol{w}_{c,i-1})\|^2 \\
&\quad + \frac{\mu}{2}\left(1+2\mu\delta\right)\|\boldsymbol{d}_{i-1}\|^2 + \mu^2\frac{\delta}{2}\sigma^2
\end{aligned}
\tag{5.91}
$$

where cross-terms were removed in (a) due to the conditional zero-mean condition (5.20a), (b) follows from $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and from $-2a^{\mathsf{T}}b \leq \|a\|^2 + \|b\|^2$ and (c) is a result of grouping terms and Lemma 5.2.

Note that (5.91) continues to be random due to the conditioning on $\boldsymbol{\mathcal{W}}_{i-1}$, but that it holds for every choice of $\boldsymbol{\mathcal{W}}_{i-1}$ with probability 1. Furthermore, since $\boldsymbol{w}_{c,i-1} = \sum_{k=1}^{N} p_k\,\boldsymbol{w}_{k,i-1}$, the centroid $\boldsymbol{w}_{c,i-1}$ is deterministic conditioned on $\boldsymbol{\mathcal{W}}_{i-1}$. As such, the event $\boldsymbol{w}_{c,i-1} \in \mathcal{G}$ is deterministic conditioned on $\boldsymbol{\mathcal{W}}_{i-1}$, and (5.91) holds for every $\boldsymbol{w}_{c,i-1} \in \mathcal{G}$. We can hence take

expectations over $\boldsymbol{w}_{c,i-1} \in \mathcal{G}$ and apply Lemma 5.1 to find:

$$\begin{aligned}
\mathbb{E}\left\{J(\boldsymbol{w}_{c,i})|\,\boldsymbol{w}_{c,i-1}\in\mathcal{G}\right\} \leq\; & \mathbb{E}\left\{J(\boldsymbol{w}_{c,i-1})|\,\boldsymbol{w}_{c,i-1}\in\mathcal{G}\right\} \\
& -\frac{\mu}{2}\left(1-2\mu\delta\right)\mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_{c,i-1})\|^2|\,\boldsymbol{w}_{c,i-1}\in\mathcal{G}\right\} \\
& +\frac{\mu}{2}\left(1+2\mu\delta\right)\mathbb{E}\left\{\|\boldsymbol{d}_{i-1}\|^2|\,\boldsymbol{w}_{c,i-1}\in\mathcal{G}\right\}+\mu^2\frac{\delta}{2}\sigma^2 \\
\overset{(a)}{\leq}\; & \mathbb{E}\left\{J(\boldsymbol{w}_{c,i-1})|\,\boldsymbol{w}_{c,i-1}\in\mathcal{G}\right\}-\mu^2 c_1\frac{c_2}{c_1}\left(1+\frac{1}{\pi}\right) \\
& +O(\mu)\,\mathbb{E}\left\{\|\boldsymbol{d}_{i-1}\|^2|\,\boldsymbol{w}_{c,i-1}\in\mathcal{G}\right\}+\mu^2 c_2 \\
\overset{(b)}{\leq}\; & \mathbb{E}\left\{J(\boldsymbol{w}_{c,i-1})|\,\boldsymbol{w}_{c,i-1}\in\mathcal{G}\right\}-\mu^2\frac{c_2}{\pi} \\
& +\frac{\mu}{2}\left(1+2\mu\delta\right)\mathbb{E}\left\{\|\boldsymbol{d}_{i-1}\|^2|\,\boldsymbol{w}_{c,i-1}\in\mathcal{G}\right\}
\end{aligned} \tag{5.92}$$

In step (a) we applied definition 5.2, and in particular, that from (5.47) $\|\nabla J(\boldsymbol{w}_{c,i-1})\|^2 \geq \mu\frac{c_2}{c_1}\left(1+\frac{1}{\pi}\right)$ whenever $\boldsymbol{w}_{c,i-1}\in\mathcal{G}$, which implies:

$$\mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_{c,i-1})\|^2|\,\boldsymbol{w}_{c,i-1}\in\mathcal{G}\right\}\geq\mu\frac{c_2}{c_1}\left(1+\frac{1}{\pi}\right) \tag{5.93}$$

We also collected constants into c_1 and c_2 defined in (5.51)–(5.52) for brevity. Step (b) is obtained by grouping terms. Note that from lemma 5.2, we have a bound on $\mathbb{E}\,\|\boldsymbol{d}_{i-1}\|^2$, but not on the partial expectation conditioned over $\boldsymbol{w}_{c,i-1}\in\mathcal{G}$. We can decompose the full expectation:

$$\begin{aligned}
\mathbb{E}\left\{\|\boldsymbol{d}_{i-1}\|^2\right\} =\; & \mathbb{E}\left\{\|\boldsymbol{d}_{i-1}\|^2|\,\boldsymbol{w}_{c,i-1}\in\mathcal{G}\right\}\cdot\pi_{i-1}^{\mathcal{G}} \\
& +\mathbb{E}\left\{\|\boldsymbol{d}_{i-1}\|^2|\,\boldsymbol{w}_{c,i-1}\in\mathcal{G}^C\right\}\cdot\pi_{i-1}^{\mathcal{G}^C}\overset{(5.45)}{\leq}O(\mu^2)
\end{aligned} \tag{5.94}$$

which implies

$$\mathbb{E}\left\{\|\boldsymbol{d}_{i-1}\|^2|\,\boldsymbol{w}_{c,i-1}\in\mathcal{G}\right\}\leq\frac{O(\mu^2)}{\pi_{i-1}^{\mathcal{G}}} \tag{5.95}$$

so that we obtain for (5.92):

$$\mathbb{E}\left\{J(\boldsymbol{w}_{c,i})|\,\boldsymbol{w}_{c,i-1}\in\mathcal{G}\right\}\leq\mathbb{E}\left\{J(\boldsymbol{w}_{c,i-1})|\,\boldsymbol{w}_{c,i-1}\in\mathcal{G}\right\}-\mu^2\frac{c_2}{\pi}+\frac{O(\mu^3)}{\pi_{i-1}^{\mathcal{G}}}\tag{5.96}$$

Similarly:

$$\begin{aligned}\mathbb{E}\left\{J(\boldsymbol{w}_{c,i})|\,\boldsymbol{w}_{c,i-1}\in\mathcal{M}\right\}\leq&\;\mathbb{E}\left\{J(\boldsymbol{w}_{c,i-1})|\,\boldsymbol{w}_{c,i-1}\in\mathcal{M}\right\}\\&-\frac{\mu}{2}\left(1-2\mu\delta\right)\mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_{c,i-1})\|^2|\,\boldsymbol{w}_{c,i-1}\in\mathcal{M}\right\}\\&+\frac{\mu}{2}\left(1+2\mu\delta\right)\mathbb{E}\left\{\|\boldsymbol{d}_{i-1}\|^2|\,\boldsymbol{w}_{c,i-1}\in\mathcal{M}\right\}+\mu^2\frac{\delta}{2}\sigma^2\\\overset{(a)}{\leq}&\;\mathbb{E}\left\{J(\boldsymbol{w}_{c,i-1})|\,\boldsymbol{w}_{c,i-1}\in\mathcal{M}\right\}+\mu^2c_2\\&+\frac{\mu}{2}\left(1+2\mu\delta\right)\mathbb{E}\left\{\|\boldsymbol{d}_{i-1}\|^2|\,\boldsymbol{w}_{c,i-1}\in\mathcal{M}\right\}\\\overset{(b)}{\leq}&\;\mathbb{E}\left\{J(\boldsymbol{w}_{c,i-1})|\,\boldsymbol{w}_{c,i-1}\in\mathcal{M}\right\}+\mu^2c_2+\frac{O(\mu^3)}{\pi_{i-1}^{\mathcal{M}}}\end{aligned}\tag{5.97}$$

where (a) follows from the fact that $\|\nabla J(\boldsymbol{w}_{c,i-1})\|^2\geq 0$ with probability 1 and (b) made use of the same argument that led to (5.96).

5.E Proof of Lemma 5.4

We refer to (5.64). Suppose $i \leq \frac{T}{\mu}$, where T is an arbitrary constant independent of μ. We then have for $i \geq 0$:

$$\begin{aligned}
\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_{i+1}^{i^\star}\right\|^2|\boldsymbol{\mathcal{F}}_{i^\star+i}\right\} &\overset{(5.64)}{=} \mathbb{E}\left\{\left\|\left(I-\mu\boldsymbol{H}_{i^\star+i}\right)\widetilde{\boldsymbol{w}}_i^{i^\star}+\mu\nabla J(\boldsymbol{w}_{c,i^\star})\right.\right. \\
&\qquad\left.\left.+\,\mu\boldsymbol{d}_{i^\star+i}+\mu\,\boldsymbol{s}_{i^\star+i+1}\right\|^2|\boldsymbol{\mathcal{F}}_{i^\star+i}\right\} \\
&\overset{(a)}{=} \left\|\left(I-\mu\boldsymbol{H}_{i^\star+i}\right)\widetilde{\boldsymbol{w}}_i^{i^\star}+\mu\nabla J(\boldsymbol{w}_{c,i^\star})+\mu\boldsymbol{d}_{i^\star+i}\right\|^2 \\
&\quad+\mu^2\,\mathbb{E}\left\{\|\boldsymbol{s}_{i^\star+i+1}\|^2|\boldsymbol{\mathcal{F}}_{i^\star+i}\right\} \\
&\overset{(b)}{=} \frac{1}{1-\mu\delta}\left\|\left(I-\mu\boldsymbol{H}_{i^\star+i}\right)\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|^2+\frac{\mu}{\delta}\|\nabla J(\boldsymbol{w}_{c,i^\star})+\boldsymbol{d}_{i^\star+i}\|^2 \\
&\quad+\mu^2\,\mathbb{E}\left\{\|\boldsymbol{s}_{i^\star+i+1}\|^2|\boldsymbol{\mathcal{F}}_{i^\star+i}\right\} \\
&\overset{(c)}{=} \frac{1}{1-\mu\delta}\left\|\left(I-\mu\boldsymbol{H}_{i^\star+i}\right)\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|^2+2\frac{\mu}{\delta}\|\nabla J(\boldsymbol{w}_{c,i^\star})\|^2 \\
&\quad+2\frac{\mu}{\delta}\|\boldsymbol{d}_{i^\star+i}\|^2+\mu^2\,\mathbb{E}\left\{\|\boldsymbol{s}_{i^\star+i+1}\|^2|\boldsymbol{\mathcal{F}}_{i^\star+i}\right\} \\
&\overset{(d)}{\leq} \frac{(1+\mu\delta)^2}{1-\mu\delta}\left\|\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|^2+2\frac{\mu}{\delta}\|\nabla J(\boldsymbol{w}_{c,i^\star})\|^2 \\
&\quad+2\frac{\mu}{\delta}\|\boldsymbol{d}_{i^\star+i}\|^2+\mu^2\,\mathbb{E}\left\{\|\boldsymbol{s}_{i^\star+i+1}\|^2|\boldsymbol{\mathcal{F}}_{i^\star+i}\right\} \qquad (5.98)
\end{aligned}$$

where (a) follows from the conditional zero-mean property of the gradient noise term in Assumption 5.4, (b) follows from Jensen's inequality

$$\|a+b\|^2 \leq \frac{1}{\alpha}\|a\|^2+\frac{1}{1-\alpha}\|b\|^2 \qquad (5.99)$$

with $\alpha=\mu\delta<1$ and (c) follows from the same inequality with $\alpha=\frac{1}{2}$. Step (d) follows from the sub-multiplicative property of norms along with $-\delta I \leq \nabla^2 J(\boldsymbol{w}_{c,i^\star}) \leq \delta I$, which follows from the Lipschitz gradient condition in Assumption 5.2. Since $\boldsymbol{w}_{c,i^\star}$ is deterministic

conditioned on $\boldsymbol{\mathcal{F}}_{i^\star+i}$ we can now take expectations over $\boldsymbol{w}_{c,i^\star} \in \mathcal{H}$ to obtain:

$$
\begin{aligned}
&\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_{i+1}^{i^\star}\right\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\leq \frac{(1+\mu\delta)^2}{1-\mu\delta}\,\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_{i}^{i^\star}\right\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\quad + 2\frac{\mu}{\delta}\,\mathbb{E}\left\{\|\boldsymbol{d}_{i^\star+i}\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\quad + 2\frac{\mu}{\delta}\,\mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_{c,i^\star})\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\quad + \mu^2\,\mathbb{E}\left\{\|\boldsymbol{s}_{i^\star+i+1}\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\overset{(a)}{\leq} \frac{(1+\mu\delta)^2}{1-\mu\delta}\,\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_{i}^{i^\star}\right\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} + 2\frac{\mu}{\delta}\cdot\frac{O(\mu^2)}{\pi_{i^\star}^{\mathcal{H}}} \\
&\quad + 2\frac{\mu}{\delta}\cdot O(\mu) + O(\mu^2) \\
&\leq \frac{(1+\mu\delta)^2}{1-\mu\delta}\,\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_{i}^{i^\star}\right\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} + O(\mu^2) + \frac{O(\mu^3)}{\pi_{i^\star}^{\mathcal{H}}}
\end{aligned}
\tag{5.100}
$$

where (a) follows from the perturbation bounds in Lemma 5.2 and the starting assumption that $\boldsymbol{w}_{c,i^\star}$ is an $O(\mu)$-square stationary point. Note that, at time $i = 0$, we have:

$$
\widetilde{\boldsymbol{w}}_{0}^{i^\star} = \boldsymbol{w}_{c,i^\star} - \boldsymbol{w}_{c,i^\star+0} = 0 \tag{5.101}
$$

and hence the initial deviation is zero, by definition. Iterating, starting at $i = 0$ yields:

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\leq \left(\sum_{n=0}^{i-1}\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\right)^n\right)\left(O(\mu^2)+\frac{O(\mu^3)}{\pi_{i^\star}^{\mathcal{H}}}\right) \\
&= \frac{1-\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\right)^i}{1-\frac{(1+\mu\delta)^2}{1-\mu\delta}}\left(O(\mu^2)+\frac{O(\mu^3)}{\pi_{i^\star}^{\mathcal{H}}}\right) \\
&= \frac{\left(\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\right)^i-1\right)(1-\mu\delta)}{1+2\mu\delta+\mu^2\delta^2-1+\mu\delta}\left(O(\mu^2)+\frac{O(\mu^3)}{\pi_{i^\star}^{\mathcal{H}}}\right) \\
&= \frac{\left(\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\right)^i-1\right)(1-\mu\delta)}{3\delta+\mu\delta^2}\left(O(\mu)+\frac{O(\mu^2)}{\pi_{i^\star}^{\mathcal{H}}}\right) \\
&\leq \frac{\left(\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\right)^{\frac{T}{\mu}}-1\right)(1-\mu\delta)}{3\delta+\mu\delta^2}\left(O(\mu)+\frac{O(\mu^2)}{\pi_{i^\star}^{\mathcal{H}}}\right) \\
&= O(\mu)+\frac{O(\mu^2)}{\pi_{i^\star}^{\mathcal{H}}}
\end{aligned} \tag{5.102}$$

where the last line follows from Lemma 5.3 after noting that:

$$\begin{aligned}
&\frac{\left(\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\right)^{\frac{T}{\mu}}-1\right)(1-\mu\delta)}{3\delta+\mu\delta^2} \\
&\leq \frac{\left(\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\right)^{\frac{T}{\mu}}-1\right)(1-\mu\delta)}{3\delta} \\
&\leq \frac{\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\right)^{\frac{T}{\mu}}-\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\right)^{\frac{T}{\mu}}\mu\delta-1+\mu\delta}{3\delta} \\
&\leq \frac{\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\right)^{\frac{T}{\mu}}-1}{3\delta}
\end{aligned} \tag{5.103}$$

This establishes (5.69). We proceed to establish a bound on the fourth-order moment. Using the inequality [1]:

$$\|a+b\|^4 \leq \|a\|^4 + 3\|b\|^4 + 8\|a\|^2\|b\|^2 + 4\|a\|^2\left(a^{\mathsf{T}}b\right) \tag{5.104}$$

we have:

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_{i+1}^{i^\star}\right\|^4 |\boldsymbol{\mathcal{F}}_{i^\star+i}\right\} \\
&\leq \left\|(I-\mu\boldsymbol{H}_{i^\star+i})\,\widetilde{\boldsymbol{w}}_i^{i^\star} + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) + \mu\boldsymbol{d}_{i^\star+i}\right\|^4 \\
&\quad + 3\mu^4\,\mathbb{E}\left\{\|\boldsymbol{s}_{i^\star+i+1}\|^4\,|\boldsymbol{\mathcal{F}}_{i^\star+i}\right\} \\
&\quad + 8\mu^2\left\|(I-\mu\boldsymbol{H}_{i^\star+i})\,\widetilde{\boldsymbol{w}}_i^{i^\star} + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) + \mu\boldsymbol{d}_{i^\star+i}\right\|^2 \\
&\qquad \times \mathbb{E}\left\{\|\boldsymbol{s}_{i^\star+i+1}\|^2\,|\boldsymbol{\mathcal{F}}_{i^\star+i}\right\} \\
&\quad + 4\mu\left\|(I-\mu\boldsymbol{H}_{i^\star+i})\,\widetilde{\boldsymbol{w}}_i^{i^\star} + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) + \mu\boldsymbol{d}_{i^\star+i}\right\|^2 \\
&\qquad \times \left((I-\mu\boldsymbol{H}_{i^\star+i})\,\widetilde{\boldsymbol{w}}_i^{i^\star} + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) + \mu\boldsymbol{d}_{i^\star+i}\right)^{\mathsf{T}} \\
&\qquad\quad \times \left(\mathbb{E}\left\{\boldsymbol{s}_{i^\star+i+1}\,|\boldsymbol{\mathcal{F}}_{i^\star+1}\right\}\right) \\
&\overset{(a)}{=} \left\|(I-\mu\boldsymbol{H}_{i^\star+i})\,\widetilde{\boldsymbol{w}}_i^{i^\star} + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) + \mu\boldsymbol{d}_{i^\star+i}\right\|^4 \\
&\quad + 3\mu^4\,\mathbb{E}\left\{\|\boldsymbol{s}_{i^\star+i+1}\|^4\,|\boldsymbol{\mathcal{F}}_{i^\star+i}\right\} \\
&\quad + 8\mu^2\left\|(I-\mu\boldsymbol{H}_{i^\star+i})\,\widetilde{\boldsymbol{w}}_i^{i^\star} + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) + \mu\boldsymbol{d}_{i^\star+i}\right\|^2 \\
&\qquad \times \mathbb{E}\left\{\|\boldsymbol{s}_{i^\star+i+1}\|^2\,|\boldsymbol{\mathcal{F}}_{i^\star+i}\right\} \\
&\overset{(b)}{=} \left\|(I-\mu\boldsymbol{H}_{i^\star+i})\,\widetilde{\boldsymbol{w}}_i^{i^\star} + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) + \mu\boldsymbol{d}_{i^\star+i}\right\|^4 + O(\mu^4) \\
&\quad + \left\|(I-\mu\boldsymbol{H}_{i^\star+i})\,\widetilde{\boldsymbol{w}}_i^{i^\star} + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) + \mu\boldsymbol{d}_{i^\star+i}\right\|^2 O(\mu^2) \\
&\overset{(c)}{=} \left\|(I-\mu\boldsymbol{H}_{i^\star+i})\,\widetilde{\boldsymbol{w}}_i^{i^\star} + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) + \mu\boldsymbol{d}_{i^\star+i}\right\|^4 + O(\mu^4) \\
&\quad + \Big(\left\|(I-\mu\boldsymbol{H}_{i^\star+i})\,\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|^2 \\
&\qquad + \mu^2\|\nabla J(\boldsymbol{w}_{c,i^\star})\|^2 + \mu^2\|\boldsymbol{d}_{i^\star+i}\|^2\Big)O(\mu^2)
\end{aligned} \tag{5.105}$$

where in step (a) we dropped cross-terms due to the conditional zero-mean property of the gradient noise in Assumption 5.4, step (b) follows from the fourth-order conditions on the gradient noise in Assumption 5.4 along with the perturbation bounds in Lemma 5.2, and (c) follows from Jensen's inequality, i.e. $\|a+b+c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$. Taking expectations over $\boldsymbol{w}_{c,i^\star} \in \mathcal{H}$ on both sides and collecting constant factors along with μ in appropriate $O(\cdot)$ terms:

$$
\begin{aligned}
&\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_{i+1}^{i^\star}\right\|^4 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\leq \mathbb{E}\left\{\left\| \left(I-\mu\boldsymbol{H}_{i^\star+i}\right)\widetilde{\boldsymbol{w}}_{i}^{i^\star} + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) \right.\right. \\
&\qquad \left.\left. + \mu\boldsymbol{d}_{i^\star+i}\right\|^4 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} + O(\mu^4) \\
&\quad + \Big(\mathbb{E}\left\{\left\|\left(I-\mu\boldsymbol{H}_{i^\star+i}\right)\widetilde{\boldsymbol{w}}_{i}^{i^\star}\right\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\qquad + \mu^2\,\mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_{c,i^\star})\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\qquad + \mu^2\,\mathbb{E}\left\{\|\boldsymbol{d}_{i^\star+i}\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\}\Big) O(\mu^2) \\
&\leq \mathbb{E}\left\{\left\| \left(I-\mu\boldsymbol{H}_{i^\star+i}\right)\widetilde{\boldsymbol{w}}_{i}^{i^\star} + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) \right.\right. \\
&\qquad \left.\left. + \mu\boldsymbol{d}_{i^\star+i}\right\|^4 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} + O(\mu^4) \\
&\quad + \Big((1+\mu\delta)^2\,\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_{i^\star}^{i}\right\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\qquad + \mu^2\,\mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_{c,i^\star})\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\qquad + \mu^2\,\mathbb{E}\left\{\|\boldsymbol{d}_{i^\star+i}\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\}\Big) O(\mu^2) \\
&\leq \mathbb{E}\left\{\left\| \left(I-\mu\boldsymbol{H}_{i^\star+i}\right)\widetilde{\boldsymbol{w}}_{i}^{i^\star} + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) \right.\right. \\
&\qquad \left.\left. + \mu\boldsymbol{d}_{i^\star+i}\right\|^4 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} + O(\mu^4) \\
&\quad + \left((1+\mu\delta)^2 O(\mu) + \mu^2 O(\mu) + \mu^2 \frac{O(\mu^2)}{\pi_{i^\star}^{\mathcal{H}}} \right) O(\mu^2) \\
&= \mathbb{E}\left\{\left\| \left(I-\mu\boldsymbol{H}_{i^\star+i}\right)\widetilde{\boldsymbol{w}}_{i}^{i^\star} + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) \right.\right. \\
&\qquad \left.\left. + \mu\boldsymbol{d}_{i^\star+i}\right\|^4 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} + O(\mu^3) + \frac{O(\mu^6)}{\pi_{i^\star}^{\mathcal{H}}}
\end{aligned}
\tag{5.106}
$$

Finally, from Jensen's inequality, we find for $0 < \alpha < 1$:

$$\|a+b\|^4 = \frac{1}{\alpha^3}\|a\|^4 + \frac{1}{(1-\alpha)^3}\|b\|^4 \tag{5.107}$$

and hence for $\alpha = 1-\mu\delta$ and $0 < \mu < \frac{1}{\delta}$:

$$\begin{aligned}
&\mathbb{E}\left\{\Big\| (I-\mu\boldsymbol{H}_{i^\star+i})\,\widetilde{\boldsymbol{w}}_i^{i^\star} + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) \right. \\
&\qquad \left. + \mu\boldsymbol{d}_{i^\star+i}\Big\|^4 |\, \boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\} \\
&\overset{(5.107)}{\leq} \frac{(1+\mu\delta)^4}{(1-\mu\delta)^3}\,\mathbb{E}\left\{\Big\|\widetilde{\boldsymbol{w}}_i^{i^\star}\Big\|^4 |\, \boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\} \\
&\qquad + \frac{\mu^4}{\mu^3\delta^3}\,\mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_{c,i^\star}) + \boldsymbol{d}_{i^\star+i}\|^4 |\, \boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\} \\
&\overset{(5.107)}{\leq} \frac{(1+\mu\delta)^4}{(1-\mu\delta)^3}\,\mathbb{E}\left\{\Big\|\widetilde{\boldsymbol{w}}_i^{i^\star}\Big\|^4 |\, \boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\} \\
&\qquad + 8\frac{\mu}{\delta^3}\Big(\,\mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_{c,i^\star})\|^4 |\, \boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\} \\
&\qquad\qquad + \mathbb{E}\left\{\|\boldsymbol{d}_{i^\star+i}\|^4 |\, \boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\}\Big) \\
&\leq \frac{(1+\mu\delta)^4}{(1-\mu\delta)^3}\,\mathbb{E}\left\{\Big\|\widetilde{\boldsymbol{w}}_i^{i^\star}\Big\|^4 |\, \boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\} \\
&\qquad + 8\frac{\mu}{\delta^3}\left(O(\mu^2) + \frac{O(\mu^4)}{\pi_i^{\mathcal{H}}}\right) \\
&\leq \frac{(1+\mu\delta)^4}{(1-\mu\delta)^3}\,\mathbb{E}\left\{\Big\|\widetilde{\boldsymbol{w}}_i^{i^\star}\Big\|^4 |\, \boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\} + O(\mu^3) + \frac{O(\mu^5)}{\pi_i^{\mathcal{H}}}
\end{aligned} \tag{5.108}$$

Hence,

$$\begin{aligned}
&\mathbb{E}\left\{\Big\|\widetilde{\boldsymbol{w}}_{i+1}^{i^\star}\Big\|^4 |\, \boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\} \\
&\leq \frac{(1+\mu\delta)^4}{(1-\mu\delta)^3}\,\mathbb{E}\left\{\Big\|\widetilde{\boldsymbol{w}}_i^{i^\star}\Big\|^4 |\, \boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\} + O(\mu^3) + \frac{O(\mu^5)}{\pi_i^{\mathcal{H}}}
\end{aligned} \tag{5.109}$$

Recall again that $\widetilde{\boldsymbol{w}}_0^{i^\star} = 0$ and therefore iterating yields:

$$
\begin{aligned}
&\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|^4 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\leq \left(\sum_{n=0}^{i-1}\left(\frac{(1+\mu\delta)^4}{(1-\mu\delta)^3}\right)^n\right)\left(O(\mu^3)+\frac{O(\mu^5)}{\pi_i^{\mathcal{H}}}\right) \\
&= \frac{1-\left(\frac{(1+\mu\delta)^4}{(1-\mu\delta)^3}\right)^i}{1-\frac{(1+\mu\delta)^4}{(1-\mu\delta)^3}}\left(O(\mu^3)+\frac{O(\mu^5)}{\pi_i^{\mathcal{H}}}\right) \\
&= \frac{\left(\left(\frac{(1+\mu\delta)^4}{(1-\mu\delta)^3}\right)^i-1\right)(1-\mu\delta)^3}{(1+\mu\delta)^4-(1-\mu\delta)^3}\left(O(\mu^3)+\frac{O(\mu^5)}{\pi_i^{\mathcal{H}}}\right) \\
&\leq \frac{\left(\frac{(1+\mu\delta)^4}{(1-\mu\delta)^3}\right)^i-1}{(1+\mu\delta)^4-(1-\mu\delta)^3}\left(O(\mu^3)+\frac{O(\mu^5)}{\pi_i^{\mathcal{H}}}\right) \\
&\overset{(a)}{\leq} \frac{\left(\frac{(1+\mu\delta)^4}{(1-\mu\delta)^3}\right)^i-1}{O(\mu)}\left(O(\mu^3)+\frac{O(\mu^5)}{\pi_i^{\mathcal{H}}}\right) \\
&= \left(\left(\frac{(1+\mu\delta)^4}{(1-\mu\delta)^3}\right)^i-1\right)O(\mu^2) \\
&\leq \left(\left(\frac{(1+\mu\delta)^4}{(1-\mu\delta)^3}\right)^{\frac{T}{\mu}}-1\right)\left(O(\mu^2)+\frac{O(\mu^4)}{\pi_i^{\mathcal{H}}}\right) \\
&\leq O(\mu^2)+\frac{O(\mu^4)}{\pi_i^{\mathcal{H}}}
\end{aligned}
\tag{5.110}
$$

where in (a) we expanded:

$$
\begin{aligned}
&(1+\mu\delta)^4-(1-\mu\delta)^3 \\
&= 1+4\mu\delta+O(\mu^2)-1+3\mu\delta-O(\mu^2) = O(\mu)
\end{aligned}
\tag{5.111}
$$

and the last step follows from Lemma 5.3. This establishes (5.71). Eq. (5.70) then follows from Jensen's inequality via:

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|^3 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\leq \left(\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|^4 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\}\right)^{3/4} \\
&\leq \left(O(\mu^2) + \frac{O(\mu^4)}{\pi_{i^\star}^{\mathcal{H}}}\right)^{3/4} \\
&= O(\mu^{3/2}) + \frac{O(\mu^3)}{(\pi_{i^\star}^{\mathcal{H}})^{4/3}} \\
&\leq O(\mu^{3/2}) + \frac{O(\mu^3)}{\pi_{i^\star}^{\mathcal{H}}}
\end{aligned} \tag{5.112}$$

We now study the difference between the short-term model (5.66) and the true recursion (5.64). We have:

$$\begin{aligned}
&\boldsymbol{w}_{c,i^\star+i+1} - \boldsymbol{w}'_{c,i^\star+i+1} \\
&= -\widetilde{\boldsymbol{w}}_{i+1}^{i^\star} + \widetilde{\boldsymbol{w}}'^{i^\star}_{i+1} \\
&= -(I - \mu \boldsymbol{H}_{i^\star+i})\,\widetilde{\boldsymbol{w}}_i^{i^\star} - \mu \nabla J(\boldsymbol{w}_{c,i^\star}) - \mu \boldsymbol{d}_{i^\star+i} - \mu\, \boldsymbol{s}_{i^\star+i+1} \\
&\quad + \left(I - \mu \nabla^2 J(\boldsymbol{w}_{c,i^\star})\right) \widetilde{\boldsymbol{w}}'^{i^\star}_i + \mu \nabla J(\boldsymbol{w}_{c,i^\star}) + \mu\, \boldsymbol{s}_{i^\star+i+1} \\
&= -(I - \mu \boldsymbol{H}_{i^\star+i})\,\widetilde{\boldsymbol{w}}_i^{i^\star} - \mu \boldsymbol{d}_{i^\star+i} + \left(I - \mu \nabla^2 J(\boldsymbol{w}_{c,i^\star})\right) \widetilde{\boldsymbol{w}}'^{i^\star}_i \\
&= \left(I - \mu \nabla^2 J(\boldsymbol{w}_{c,i^\star})\right)\left(\boldsymbol{w}_{c,i^\star+i} - \boldsymbol{w}'_{c,i^\star+i}\right) - \mu \boldsymbol{d}_{i^\star+i} \\
&\quad + \mu \left(\boldsymbol{H}_{i^\star+i} - \nabla^2 J(\boldsymbol{w}_{c,i^\star})\right) \widetilde{\boldsymbol{w}}_i^{i^\star}
\end{aligned} \tag{5.113}$$

Before proceeding, note that the difference between the Hessians in the driving term can be bounded as:

$$
\begin{aligned}
&\left\|\nabla^2 J(\boldsymbol{w}_{c,i^\star}) - \boldsymbol{H}_{i^\star+i}\right\| \\
&= \left\|\nabla^2 J(\boldsymbol{w}_{c,i^\star}) - \int_0^1 \nabla^2 J\left((1-t)\,\boldsymbol{w}_{c,i^\star+i} + t\,\boldsymbol{w}_{c,i^\star}\right) dt\right\| \\
&= \left\|\int_0^1 \left(\nabla^2 J(\boldsymbol{w}_{c,i^\star}) - \nabla^2 J\left((1-t)\,\boldsymbol{w}_{c,i^\star+i} + t\,\boldsymbol{w}_{c,i^\star}\right)\right) dt\right\| \\
&\overset{(a)}{\leq} \int_0^1 \left\|\nabla^2 J(\boldsymbol{w}_{c,i^\star}) - \nabla^2 J\left((1-t)\,\boldsymbol{w}_{c,i^\star+i} + t\,\boldsymbol{w}_{c,i^\star}\right)\right\| dt \\
&\overset{(b)}{\leq} \rho \int_0^1 \left\|(1-t)\,\boldsymbol{w}_{c,i^\star} - (1-t)\,\boldsymbol{w}_{c,i^\star+i}\right\| dt \\
&= \rho \left\|\widetilde{\boldsymbol{w}}_i^{i^\star}\right\| \int_0^1 (1-t)dt = \frac{\rho}{2}\left\|\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|
\end{aligned}
\tag{5.114}
$$

where (a) follows Jensen's inequality and (b) follows form the Lipschitz Hessian assumption 5.5. Returning to (5.113) and taking norms yields:

$$
\begin{aligned}
&\left\|\,\boldsymbol{w}_{c,i^\star+i+1} - \boldsymbol{w}'_{c,i^\star+i+1}\,\right\|^2 \\
&= \Big\| \left(I - \mu\nabla^2 J(\boldsymbol{w}_{c,i^\star})\right)\left(\boldsymbol{w}_{c,i^\star+i} - \boldsymbol{w}'_{c,i^\star+i}\right) \\
&\quad - \mu\boldsymbol{d}_{i^\star+i} + \mu\left(\boldsymbol{H}_{i^\star+i} - \nabla^2 J(\boldsymbol{w}_{c,i^\star})\right)\widetilde{\boldsymbol{w}}_i^{i^\star}\Big\|^2 \\
&\overset{(a)}{\leq} \frac{1}{1-\mu\delta}\left\|\left(I - \mu\nabla^2 J(\boldsymbol{w}_{c,i^\star})\right)\left(\boldsymbol{w}_{c,i^\star+i} - \boldsymbol{w}'_{c,i^\star+i}\right)\right\|^2 \\
&\quad + \frac{\mu^2}{\mu\delta}\left\|\boldsymbol{d}_{i^\star+i} + \left(\boldsymbol{H}_{i^\star+i} - \nabla^2 J(\boldsymbol{w}_{c,i^\star})\right)\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|^2 \\
&\overset{(b)}{\leq} \frac{1}{1-\mu\delta}\left\|\left(I - \mu\nabla^2 J(\boldsymbol{w}_{c,i^\star})\right)\left(\boldsymbol{w}_{c,i^\star+i} - \boldsymbol{w}'_{c,i^\star+i}\right)\right\|^2 \\
&\quad + 2\frac{\mu}{\delta}\left(\left\|\boldsymbol{d}_{i^\star+i}\right\|^2 + \left\|\left(\boldsymbol{H}_{i^\star+i} - \nabla^2 J(\boldsymbol{w}_{c,i^\star})\right)\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|^2\right) \\
&\overset{(5.114)}{\leq} \frac{(1+\mu\delta)^2}{1-\mu\delta}\left\|\boldsymbol{w}_{c,i^\star+i} - \boldsymbol{w}'_{c,i^\star+i}\right\|^2 \\
&\quad + 2\frac{\mu}{\delta}\left(\left\|\boldsymbol{d}_{i^\star+i}\right\|^2 + \frac{\rho}{2}\left\|\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|^4\right)
\end{aligned}
\tag{5.115}
$$

where (a) again follows from Jensen's inequality (5.99) with $\alpha = 1-\mu\delta$ and (b) follows from the same inequality with $\alpha = \frac{1}{2}$. Taking expectations over $\boldsymbol{w}_{c,i^\star} \in \mathcal{H}$ yields:

$$\begin{aligned}
&\mathbb{E}\left\{\| \boldsymbol{w}_{c,i^\star+i+1} - \boldsymbol{w}'_{c,i^\star+i+1} \|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\leq \frac{(1+\mu\delta)^2}{1-\mu\delta}\, \mathbb{E}\left\{\left\|\boldsymbol{w}_{c,i^\star+i} - \boldsymbol{w}'_{c,i^\star+i}\right\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\quad + 2\frac{\mu}{\delta}\, \mathbb{E}\left\{\|\boldsymbol{d}_{i^\star+i}\|^2 \,|\, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\quad + \frac{\rho\mu}{\delta}\, \mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|^4 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\overset{(a)}{\leq} \frac{(1+\mu\delta)^2}{1-\mu\delta}\, \mathbb{E}\left\|\boldsymbol{w}_{c,i^\star+i} - \boldsymbol{w}'_{c,i^\star+i}\right\|^2 + O(\mu^3) + \frac{O(\mu^3)}{\pi_{i^\star}^{\mathcal{H}}}
\end{aligned} \tag{5.116}$$

where (a) follows from the bound on the network disagreement in Lemma 5.4.

Since both the true and the short-term model are initialized at $\boldsymbol{w}_{c,i^\star}$, we have $\boldsymbol{w}_{c,i^\star+0} - \boldsymbol{w}'_{c,i^\star+0} = 0$. Iterating and applying the same argument as above leads to:

$$\mathbb{E}\,\| \boldsymbol{w}_{c,i^\star+i+1} - \boldsymbol{w}'_{c,i^\star+i+1} \|^2 \leq O(\mu^2) + \frac{O(\mu^2)}{\pi_{i^\star}^{\mathcal{H}}} \tag{5.117}$$

which is (5.72).

CHAPTER 6

Decentralized Non-Convex Learning — Escape from Saddle-Points

The diffusion strategy for distributed learning from streaming data employs local stochastic gradient updates along with exchange of iterates over neighborhoods. In Chapter 5 we established that agents cluster around a network centroid and proceeded to study the dynamics of this point. We established expected descent in non-convex environments in the large-gradient regime and introduced a short-term model to examine the dynamics over finite-time horizons. Using this model, we establish in this chapter that the diffusion strategy is able to escape from strict saddle-points in $O(1/\mu)$ iterations; it is also able to return approximately second-order stationary points in a polynomial number of iterations. Relative to prior works on the polynomial escape from saddle-points, most of which focus on *centralized* perturbed or stochastic gradient descent, our approach requires less restrictive conditions on the gradient noise process. The materials in this chapter are based on [71].

6.1 Introduction

We consider a network of N agents. Each agent k is equipped with a local, stochastic cost of the form $J_k(w) = \mathbb{E}_x\, Q_k(w; \boldsymbol{x}_k)$, where $w \in \mathbb{R}^M$ denotes a parameter vector and $\boldsymbol{x}_k$ denotes random data. In Chapter 5, we consider a global optimization problem of the form:

$$\min_{w} J(w), \qquad \text{where } J(w) \triangleq \sum_{k=1}^{N} p_k J_k(w) \tag{6.1}$$

where the weights p_k are a function of the combination weights $a_{\ell k}$ and will be specified further below in (6.4).

Solutions to such problems via distributed strategies can be pursued through a variety of algorithms, including those of the consensus and diffusion type [1, 26–31]. In Chapter 5, we studied the diffusion strategy strategy due to its proven enhanced performance in adaptive environments in response to streaming data and drifting conditions [1, 147]. The strategy takes the form:

$$\boldsymbol{\phi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \tag{6.2a}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell=1}^{N} a_{\ell k}\boldsymbol{\phi}_{\ell,i} \tag{6.2b}$$

Note that the gradient step (6.2a) employs a *stochastic* gradient approximation $\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1})$, rather than the true gradient $\nabla J_k(\boldsymbol{w}_{k,i-1})$. The random approximation of the true gradient based on sampled data introduces persistent gradient noise, which seeps into the evolution of the algorithm. A commonly employed construction is $\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) = \nabla Q_k(\boldsymbol{w}_{k,i-1}; \boldsymbol{x}_k)$; nevertheless, we consider general stochastic gradient approximations $\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1})$ under suitable conditions on the induced gradient noise process (Assumptions 6.4 and 6.7 further ahead). Prior works have studied the dynamics of the diffusion strategy (6.2a)–(6.2b) and examined the implications of the gradient noise term in the *strongly-convex* setting [1, 27, 146]. In particular, it has been shown that despite the presence of gradient noise, the iterates $\boldsymbol{w}_{k,i}$ will approach the global solution $w^\star \triangleq \arg\min_w J(w)$ to the problem (6.1) in the mean-square-error sense, namely it will hold that $\limsup_{i\to\infty} \mathbb{E}\,\|w^\star - \boldsymbol{w}_{k,i}\|^2 = O(\mu)$.

In Chapter 5 we showed that many of the desirable properties of the diffusion algorithm continue to hold in the more challenging non-convex setting. We established that all agents will cluster around a common network centroid after sufficient iterations and established expected descent of the network centroid in the large-gradient regime. In this part of the work we establish that the diffusion strategy is able to escape strict-saddle points and return second-order stationary points in polynomial time.

6.1.1 Related Works

A general discussion on decentralized algorithms for optimization and learning can be found in Chapter 5. In this section, we focus on works studying the ability of algorithms to escape strict saddle-points and reach second-order stationary points, which is the focus of this part. The desire to obtain guarantees for the escape from saddle-points is motivated by the observation that in many problems of interest, such as neural networks, saddle-points can correspond to bottlenecks of the optimization problem. As such, guarantees of convergence to first-order stationary points, i.e., points where the norm of the gradient is small, need not be sufficient to establish good performance. For this reason, there has been interest in the guarantee of convergence to second-order stationary points. Approximate second-order stationary points, like first-order stationary points, are required to have a small gradient norm, but are also restricted in terms of the smallest eigenvalues of their Hessian matrices.

Works that study the ability of gradient descent algorithms to escape strict saddle-points can broadly be classified into two approaches. The first class is based on the fact that there is at least one direction of descent at every saddle-point and leverage either second-order information [144] or first-order strategies for identifying a negative-curvature direction [135–137] to identify the descent direction. Our work falls into a second class of strategies, which exploit the fact that *strict* saddle-points (defined later) are unstable in the sense that small perturbations allow for the iterates to escape from the saddle point almost surely. Along these lines, it has been shown in [133] that under an appropriately chosen random initialization scheme, the gradient descent algorithm converges to minimizers almost surely. The work [61] further leveraged this fact to establish that distributed gradient descent with appropriately chosen initialization escapes saddle points. When subjected to persistent, but diminishing perturbations, known as annealing, *asymptotic* almost sure convergence to global minimizers of gradient descent-type algorithms has also been established in the centralized [132] and more recently in the distributed setting [140]. All these useful results, while powerful in theory, still do not provide a guarantee that the procedures are efficient in the sense that they would return accurate solutions after a finite number of iterations. Actually, despite the fact

that gradient descent with random initialization escapes saddle-points almost surely [133], it has been established that this process can take exponentially long [148], rendering the procedure impractical.

These observations have sparked interest in the design of methods that have the ability to escape saddle-points *efficiently*, where efficiency is loosely defined as yielding success in polynomial, rather than exponential time. The authors in [59] add persistent, i.i.d. perturbations to the exact gradient descent algorithm and establish polynomial escape from saddle-points, while the work [60] adds perturbations only when the presence of a saddle-point is detected. It is important to note that in most of these works, perturbations or random initializations are selected and introduced with the explicit purpose of allowing the algorithm to escape from unstable stationary points. For example, random initialization is followed by exact gradient updates in the works [61, 133], while the perturbations in [60] are applied only when a saddle-point is detected via the norm of the gradient. All of these techniques still require knowledge of the *exact gradient*. While the authors of [59] consider persistent gradient perturbations, these are nevertheless assumed to be independently and identically distributed.

Motivated by these considerations, in this chapter, we focus on implementations that employ *stochastic* gradient approximations and *constant* step-sizes. This is driven by the fact that computation of the exact gradients $\nabla J_k(\cdot)$ is generally infeasible in practice because (a) data may be streaming in, making it impossible to compute $\nabla \mathbb{E}_{x_k} Q_k(\cdot; \boldsymbol{x}_k)$ in the absence of knowledge about the distribution of the data or (b) the data set, while available as a batch, may be so large that efficient computation of the full gradient is infeasible. As such, the exact gradient will need to be replaced by an approximate *stochastic* gradient, which ends up introducing in a natural manner some form of *gradient noise* into the operation of the algorithm; this noise is the difference between the true gradient and its approximation. The gradient noise seeps into the operation of the algorithm continually and becomes coupled with the evolution of the iterates, resulting in perturbations that are neither identically nor independently distributed over time. For instance, the presence of the gradient noise process complicates the dynamics of the iterate evolution relative to the centralized recursions

considered in [59].

There have been some recent works that study *stochastic* gradient scenarios as well. However, these methods alter the gradient updates in specific ways or require the gradient noise to satisfy particular conditions. For example, the work [139] proposes the addition of Gaussian noise to the naturally occurring gradient noise, while the authors of [134] leverage alternating step-sizes. The works [135–137] introduce an intermediate negative-curvature-search step. All of these works alter the traditional stochastic gradient algorithm in order to ensure efficient escape from saddle-points. The work [138] studies the traditional stochastic gradient algorithm under a dispersive noise assumption.

The key contributions of this work are three-fold. To the best of our knowledge, we present the first analysis establishing *efficient* (i.e., polynomial) escape from strict-saddle points in the *distributed* setting. Second, we establish that the gradient noise process is sufficient to ensure efficient escape without the need to alter it by adding artificial forms of perturbations, interlacing steps with small and large step-sizes or imposing a dispersive noise assumption, as long as there is a gradient noise component present in some descent direction for every strict saddle-point. Third, relative to the existing literature on *centralized* non-convex optimization, where the focus is mostly on deterministic or *finite-sum* optimization, our modeling conditions are specifically tailored to the scenario of learning from stochastic *streaming* data. In particular, we only impose bounds on the gradient noise variance in expectation, rather than assume a bound with probability 1 [134, 138] or a sub-Gaussian distribution [139]. Furthermore, we assume that any Lipschitz conditions only hold on the *expected* stochastic gradient approximation, rather than for every realization, with probability 1 [135–137].

For reference, we refer the reader back to Table 5.1 in Chapter 5 for a summary of related works and modeling conditions.

6.2 Review of Chapter 5

6.2.1 Modeling Conditions

In this section, we briefly list the modeling conditions employed in Chapter 5 for reference.

Assumption 6.1 (**Strongly-connected graph**). *The combination weights in* (6.2b) *are convex combination weights satisfying:*

$$a_{\ell k} \geq 0, \quad \sum_{\ell \in \mathcal{N}_k} a_{\ell k} = 1, \quad a_{\ell k} = 0 \text{ if } \ell \notin \mathcal{N}_k \tag{6.3}$$

The symbol $\mathcal{N}_k$ denotes the set of neighbors of agent k. We shall assume that the graph described by the weighted combination matrix $A = [a_{\ell k}]$ is strongly-connected [1]. This means that there exists a path with nonzero weights between any two agents in the network and, moreover, at least one agent has a nontrivial self-loop, $a_{kk} > 0$. □

The Perron-Frobenius theorem [1, 23, 24] then implies that A has a spectral radius of one and a single eigenvalue at one. The corresponding eigenvector can be normalized to satisfy:

$$Ap = p, \quad \mathbb{1}^{\mathsf{T}} p = 1, \quad p_k > 0 \tag{6.4}$$

where the $\{p_k\}$ denote the individual entries of the Perron vector, p.

Assumption 6.2 (**Lipschitz gradients**). *For each k, the gradient $\nabla J_k(\cdot)$ is Lipschitz, namely, for any $x, y \in \mathbb{R}^M$:*

$$\|\nabla J_k(x) - \nabla J_k(y)\| \leq \delta \|x - y\| \tag{6.5}$$

In light of (6.1) *and Jensen's inequality, this implies for the aggregate cost:*

$$\|\nabla J(x) - \nabla J(y)\| \leq \delta \|x - y\| \tag{6.6}$$

□

The Lipschitz gradient conditions (6.5) and (6.6) imply

$$J(y) \leq J(x) + \nabla J(x)^{\mathsf{T}}(y-x) + \frac{\delta}{2}\|x-y\|^2 \tag{6.7}$$

For the Hessian matrix we have [1]:

$$-\delta I \leq \nabla^2 J(x) \leq \delta I \tag{6.8}$$

Assumption 6.3 (**Bounded gradient disagreement**). *For each pair of agents k and ℓ, the gradient disagreement is bounded, namely, for any $x \in \mathbb{R}^M$:*

$$\|\nabla J_k(x) - \nabla J_\ell(x)\| \leq G \tag{6.9}$$

□

Definition 6.1 (**Filtration**). *We denote by $\boldsymbol{\mathcal{F}}_i$ the filtration generated by the random processes $\boldsymbol{w}_{k,j}$ for all k and $j \leq i$:*

$$\boldsymbol{\mathcal{F}}_i \triangleq \{\boldsymbol{\mathcal{W}}_0, \boldsymbol{\mathcal{W}}_1, \ldots, \boldsymbol{\mathcal{W}}_i\} \tag{6.10}$$

where $\boldsymbol{\mathcal{W}}_j \triangleq \operatorname{col}\{\boldsymbol{w}_{1,j}, \ldots, \boldsymbol{w}_{k,j}\}$ contains the iterates across the network at time j. Informally, $\boldsymbol{\mathcal{F}}_i$ captures all information that is available about the stochastic processes $\boldsymbol{w}_{k,j}$ across the network up to time i. □

Assumption 6.4 (**Gradient noise process**). *For each k, the gradient noise process is defined as*

$$\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) = \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) - \nabla J_k(\boldsymbol{w}_{k,i-1}) \tag{6.11}$$

and satisfies

$$\mathbb{E}\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})|\boldsymbol{\mathcal{F}}_{i-1}\} = 0 \tag{6.12a}$$

$$\mathbb{E}\{\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^4|\boldsymbol{\mathcal{F}}_{i-1}\} \leq \sigma^4 \tag{6.12b}$$

for some non-negative constant σ^4. We also assume that the gradient noise processes are pairwise uncorrelated over the space conditioned on $\boldsymbol{\mathcal{F}}_{i-1}$, i.e.:

$$\mathbb{E}\left\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\,\boldsymbol{s}_{\ell,i}(\boldsymbol{w}_{\ell,i-1})^{\mathsf{T}}|\boldsymbol{\mathcal{F}}_{i-1}\right\}=0 \tag{6.13}$$

□

The fourth-order condition also implies via Jensen's inequality:

$$\mathbb{E}\left\{\|\,\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}\right\}\leq\sigma^2 \tag{6.14}$$

Definition 6.2 (Sets)**.** *To simplify the notation in the sequel, we introduce following sets:*

$$\mathcal{G}\triangleq\left\{w:\|\nabla J(w)\|^2\geq\mu\frac{c_2}{c_1}\left(1+\frac{1}{\pi}\right)\right\} \tag{6.15}$$

$$\mathcal{G}^C\triangleq\left\{w:\|\nabla J(w)\|^2<\mu\frac{c_2}{c_1}\left(1+\frac{1}{\pi}\right)\right\} \tag{6.16}$$

$$\mathcal{H}\triangleq\left\{w:w\in\mathcal{G}^C,\lambda_{\min}\left(\nabla^2J(w)\right)\leq-\tau\right\} \tag{6.17}$$

$$\mathcal{M}\triangleq\left\{w:w\in\mathcal{G}^C,\lambda_{\min}\left(\nabla^2J(w)\right)>-\tau\right\} \tag{6.18}$$

where τ is a small positive parameter, c_1 and c_2 are constants:

$$c_1\triangleq\frac{1}{2}\left(1-2\mu\delta\right)=O(1) \tag{6.19}$$

$$c_2\triangleq\delta\sigma^2/2=O(1) \tag{6.20}$$

and $0<\pi<1$ is a parameter to be chosen. Note that $\mathcal{G}^C=\mathcal{H}\cup\mathcal{M}$. We also define the probabilities $\pi_i^{\mathcal{G}}\triangleq\Pr\{\boldsymbol{w}_{c,i}\in\mathcal{G}\}$, $\pi_i^{\mathcal{H}}\triangleq\Pr\{\boldsymbol{w}_{c,i}\in\mathcal{H}\}$ and $\pi_i^{\mathcal{M}}\triangleq\Pr\{\boldsymbol{w}_{c,i}\in\mathcal{M}\}$. Then for all i, we have $\pi_i^{\mathcal{G}}+\pi_i^{\mathcal{H}}+\pi_i^{\mathcal{M}}=1$. □

Assumption 6.5 (**Lipschitz Hessians**)**.** *Each $J_k(\cdot)$ is twice-differentiable with Hessian $\nabla^2J_k(\cdot)$ and, there exists $\rho\geq0$ such that:*

$$\|\nabla^2J_k(x)-\nabla^2J_k(y)\|\leq\rho\|x-y\| \tag{6.21}$$

By Jensen's inequality, this implies that $J(\cdot) = \sum_{k=1}^{N} p_k J_k(\cdot)$ *also satisfies:*

$$\|\nabla^2 J(x) - \nabla^2 J(y)\| \leq \rho \|x - y\| \tag{6.22}$$

□

Similarly to the quadratic upper bound that follows from the Lipschitz condition on the first-derivative (6.7), this new Lipschitz condition on the second-derivative implies a cubic upper bound on the function values [144]:

$$J(y) \leq J(x) + \nabla J(x)^{\mathsf{T}}(y - x) + \frac{1}{2}(y - x)^{\mathsf{T}} \nabla^2 J(x)(y - x) + \frac{\rho}{6}\|y - x\|^3 \tag{6.23}$$

6.2.2 Review of Results

An important quantity in the network dynamics of (6.2a)–(6.2b) is the weighted network centroid:

$$\boldsymbol{w}_{c,i} \triangleq \sum_{k=1}^{N} p_k \, \boldsymbol{w}_{k,i} \tag{6.24}$$

where the weights p_k are elements of the Perron vector, defined in (6.4), which in turn is a function of the graph topology and weights. The network centroid can be shown to evolve according to a perturbed, centralized, exact gradient descent recursion [27]:

$$\boldsymbol{w}_{c,i} = \boldsymbol{w}_{c,i-1} - \mu \sum_{k=1}^{N} p_k \nabla J_k(\boldsymbol{w}_{c,i-1}) - \mu \boldsymbol{d}_{i-1} - \mu \, \boldsymbol{s}_i \tag{6.25}$$

where we defined the perturbation terms:

$$\boldsymbol{d}_{i-1} \triangleq \sum_{k=1}^{N} p_k \left(\nabla J_k(\boldsymbol{w}_{k,i-1}) - \nabla J_k(\boldsymbol{w}_{c,i-1}) \right) \tag{6.26}$$

$$\boldsymbol{s}_i \triangleq \sum_{k=1}^{N} p_k \left(\widehat{\nabla J_k}(\boldsymbol{w}_{k,i-1}) - \nabla J_k(\boldsymbol{w}_{k,i-1}) \right) \tag{6.27}$$

In Chapter 5 we established that, under assumptions 6.1–6.4, all agents will cluster around the network centroid in the mean-fourth sense:

$$\mathbb{E}\left\|\boldsymbol{\mathcal{W}}_i - \left(\mathbb{1}p^{\mathsf{T}} \otimes I\right)\boldsymbol{\mathcal{W}}_i\right\|^4 \leq \mu^4\|\mathcal{V}_L\|^4 \frac{\left\|J_\epsilon^{\mathsf{T}}\right\|^4}{\left(1-\|J_\epsilon^{\mathsf{T}}\|\right)^4}\|\mathcal{V}_R^{\mathsf{T}}\|^4 N^2\left(G^4+\sigma^4\right) + o(\mu^4) \tag{6.28}$$

for $i \geq i_o$ where $i_o \triangleq \log\left(o(\mu^4)\right)/\log\left(\left\|J_\epsilon^{\mathsf{T}}\right\|\right)$. This result has two implications. First, it establishes that, despite the fact that agents may be descending along different cost functions, and despite the fact that they may have been initialized close to different local minima, the entire network will eventually agree on a common iterate in the mean-fourth sense (and via Markov's inequality with high probability). Furthermore, it allows us to bound the perturbation terms appearing in (6.25) as:

$$\left(\mathbb{E}\,\|\boldsymbol{d}_{i-1}\|^2\right)^2 \leq \mathbb{E}\,\|\boldsymbol{d}_{i-1}\|^4 \leq O(\mu^4) \tag{6.29}$$

$$\left(\mathbb{E}\left\{\|\boldsymbol{s}_i\|^2|\boldsymbol{\mathcal{F}}_{i-1}\right\}\right)^2 \leq \mathbb{E}\left\{\|\boldsymbol{s}_i\|^4|\boldsymbol{\mathcal{F}}_{i-1}\right\} \leq \sigma^4 \tag{6.30}$$

after sufficient iterations $i \geq i_0$. We conclude that all iterates, after sufficient iterations, approximately track the network centroid $\boldsymbol{w}_{c,i}$, which in turn follows a perturbed gradient descent recursion, where the perturbation terms can be appropriately bounded.

We then proceeded to study the evolution of the network centroid and establish expected descent in the large gradient regime, i.e.:

$$\mathbb{E}\left\{J(\boldsymbol{w}_{c,i})|\,\boldsymbol{w}_{c,i-1} \in \mathcal{G}\right\} \leq \mathbb{E}\left\{J(\boldsymbol{w}_{c,i-1})|\,\boldsymbol{w}_{c,i-1} \in \mathcal{G}\right\} - \mu^2\frac{c_2}{\pi} + \frac{O(\mu^3)}{\pi_{i-1}^{\mathcal{G}}} \tag{6.31}$$

where the set $\mathcal{G}$ introduced in Definition 6.2 denotes the set of points with sufficiently large gradients $\|\nabla J(w)\|^2 \geq O(\mu)$.

While this argument could have been continued to establish the return of approximately first-order stationary points in the complement $\mathcal{G}^C = \mathcal{M}\cup\mathcal{H}$, our objective here is to establish the return of second-order stationary points in $\mathcal{M}$, which is a subset of $\mathcal{G}^C$. This requires the escape from strict-saddle points in $\mathcal{H}$. In the vicinity of first-order stationary points, a single gradient step is no longer sufficient to guarantee descent, and as such it is necessary to

study the cumulative effect of the gradient, as well as perturbations, over several iterations. We laid the ground work for this in Chapter 5 by introducing a short-term model, which is more tractable and sufficiently accurate for a limited number of iterations. This approach has been used successfully to accurately quantify the performance of adaptive networks in convex environments [1] and establish the ability of centralized perturbed gradient descent to escape saddle-points [59]. Around a first-order stationary points $\boldsymbol{w}_{c,i^\star}$ at time $i^\star$, the short-term model is obtained by first applying the mean-value theorem to (6.25) and obtain:

$$\widetilde{\boldsymbol{w}}_{i+1}^{i^\star} = (I - \mu \boldsymbol{H}_{i^\star+i})\, \widetilde{\boldsymbol{w}}_i^{i^\star} + \mu \nabla J(\boldsymbol{w}_{c,i^\star}) + \mu \boldsymbol{d}_{i^\star+i} + \mu\, \boldsymbol{s}_{i^\star+i+1} \tag{6.32}$$

where $\widetilde{\boldsymbol{w}}_i^{i^\star}$ denotes the deviation from the initial point $\boldsymbol{w}_{c,i^\star}$, i.e. $\widetilde{\boldsymbol{w}}_i^{i^\star} = \boldsymbol{w}_{c,i^\star} - \boldsymbol{w}_{c,i^\star+i}$ and

$$\boldsymbol{H}_{i^\star+i} \triangleq \int_0^1 \nabla^2 J\left((1-t)\, \boldsymbol{w}_{c,i^\star+i} + t\, \boldsymbol{w}_{c,i^\star}\right) dt \tag{6.33}$$

The short-term model is then obtained by replacing $\boldsymbol{H}_{i^\star+i}$ by $\nabla^2 J(\boldsymbol{w}_{c,i^\star})$ and dropping the driving term $\mu \boldsymbol{d}_{i^\star+i}$:

$$\widetilde{\boldsymbol{w}}'^{i^\star}_{i+1} = \left(I - \mu \nabla^2 J(\boldsymbol{w}_{c,i^\star})\right) \widetilde{\boldsymbol{w}}'^{i^\star}_i + \mu \nabla J(\boldsymbol{w}_{c,i^\star}) + \mu\, \boldsymbol{s}_{i^\star+i+1} \tag{6.34}$$

where again $\widetilde{\boldsymbol{w}}'^{i^\star}_i$ denotes the deviation from the initialization $\widetilde{\boldsymbol{w}}'^{i^\star}_i = \boldsymbol{w}_{c,i^\star} - \boldsymbol{w}'_{c,i^\star+i}$. In [70, Lemma 4], we established that the short-term model (6.34) is a meaningful approximation of (6.32) in the sense that for a limited number of iterations $i \leq \frac{T}{\mu}$, we have the following

bounds:

$$\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i^{i\star}\right\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \leq O(\mu) + \frac{O(\mu^2)}{\pi_{i^\star}^{\mathcal{H}}} \tag{6.35}$$

$$\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i^{i\star}\right\|^3 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \leq O(\mu^{3/2}) + \frac{O(\mu^3)}{\pi_{i^\star}^{\mathcal{H}}} \tag{6.36}$$

$$\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i^{i\star}\right\|^4 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \leq O(\mu^2) + \frac{O(\mu^4)}{\pi_{i^\star}^{\mathcal{H}}} \tag{6.37}$$

$$\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i^{i\star} - \widetilde{\boldsymbol{w}}_i^{\prime i\star}\right\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \leq O(\mu^2) + \frac{O(\mu^2)}{\pi_{i^\star}^{\mathcal{H}}} \tag{6.38}$$

$$\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i^{\prime i\star}\right\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \leq O(\mu) + \frac{O(\mu^2)}{\pi_{i^\star}^{\mathcal{H}}} \tag{6.39}$$

We will now proceed to argue that these deviation bounds allow us to establish decent of (6.32) by means of studying descent of (6.34) and leverage this fact to show that the diffusion strategy will continue to descend through strict-saddle points in Theorem 6.1. This result, along with the descent for large gradients established in Theorem 5.2 will allow us to guarantee the return of an approximately second-order stationary points in Theorem 6.2. The argument is summarized in Fig. 6.1.

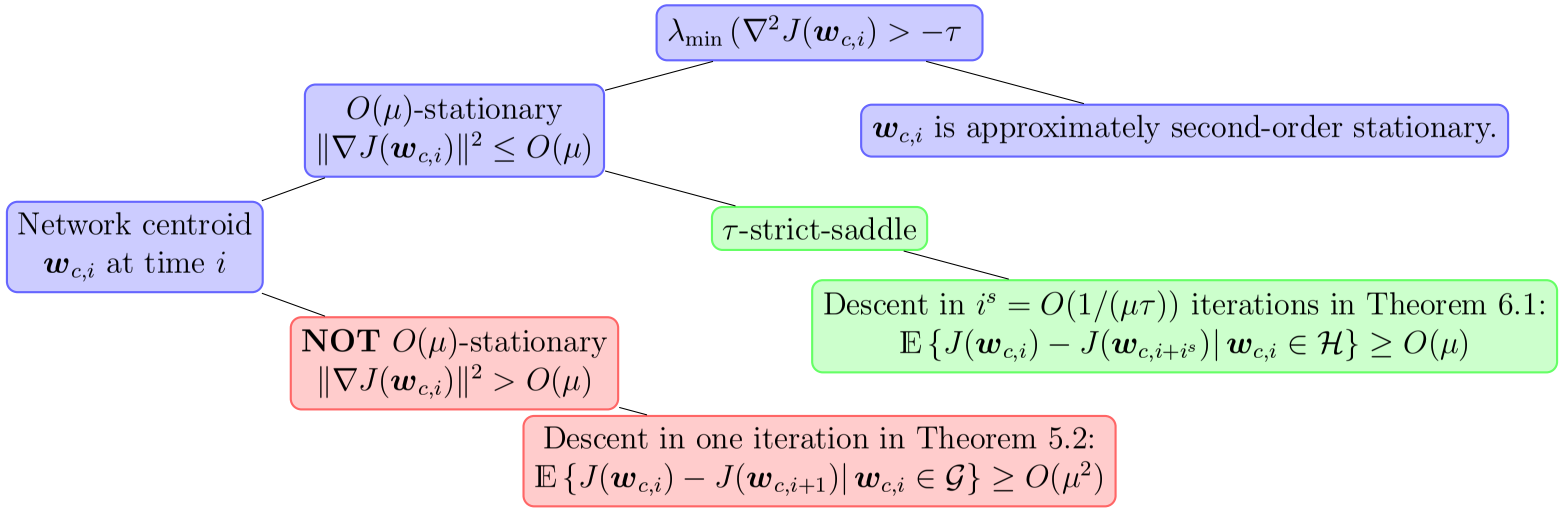

Figure 6.1: Classification of approximately stationary points. Theorem 6.1 in this chapter establishes descent in the green branch. The red branch is treated in Chapter 5. The two results are combined in Theorem 6.2 to establish the return of a second-order stationary point with high probability.

6.3 Escape from Saddle-Points

The deviation bounds (6.35)–(6.39) establish that, for the first $O(1/\mu)$ iterations following a first-order stationary points $\boldsymbol{w}_{c,i^\star}$, the trajectories of the true recursion (6.32) the short-term model (6.34) will remain close. As a consequence, we are able to guarantee descent of $J(\boldsymbol{w}_{c,i^\star+i})$ by studying $J(\boldsymbol{w}'_{c,i^\star+i})$. Note from (6.7) that

$$J(\boldsymbol{w}_{c,i^\star+i}) \leq J(\boldsymbol{w}'_{c,i^\star+i}) + \nabla J\left(\boldsymbol{w}'_{c,i^\star+i}\right)^{\mathsf{T}}\left(\boldsymbol{w}_{c,i^\star+i} - \boldsymbol{w}'_{c,i^\star+i}\right) + \frac{\delta}{2}\left\|\boldsymbol{w}_{c,i^\star+i} - \boldsymbol{w}'_{c,i^\star+i}\right\|^2 \quad (6.40)$$

Taking conditional expectation yields:

$$\begin{aligned}
\mathbb{E}\left\{J(\boldsymbol{w}_{c,i^\star+i})\,|\,\boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \leq\ & \mathbb{E}\left\{J(\boldsymbol{w}'_{c,i^\star+i})|\,\boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
& + \mathbb{E}\left\{\nabla J\left(\boldsymbol{w}'_{c,i^\star+i}\right)^{\mathsf{T}}\left(\boldsymbol{w}_{c,i^\star+i} - \boldsymbol{w}'_{c,i^\star+i}\right)\,|\,\boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
& + \frac{\delta}{2}\,\mathbb{E}\left\{\left\|\boldsymbol{w}_{c,i^\star+i} - \boldsymbol{w}'_{c,i^\star+i}\right\|^2|\,\boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \qquad (6.41)
\end{aligned}$$

The two terms appearing on the right-hand side can be bounded as:

$$\begin{aligned}
& \mathbb{E}\left\{\nabla J\left(\boldsymbol{w}'_{c,i^\star+i}\right)^{\mathsf{T}}\left(\boldsymbol{w}_{c,i^\star+i} - \boldsymbol{w}'_{c,i^\star+i}\right)\,|\,\boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\overset{(a)}{\leq} \sqrt{\mathbb{E}\left\{\left\|\nabla J\left(\boldsymbol{w}'_{c,i^\star+i}\right)\right\|^2|\,\boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\}} \\
&\quad \times \sqrt{\mathbb{E}\left\{\left\|\boldsymbol{w}_{c,i^\star+i} - \boldsymbol{w}'_{c,i^\star+i}\right\|^2|\,\boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\}} \\
&\overset{(6.38)}{\leq} \sqrt{O(\mu)}\sqrt{O(\mu^2) + \frac{O(\mu^2)}{\pi_{i^\star}^{\mathcal{H}}}} \\
&= O\left(\mu^{3/2}\right) + \frac{O(\mu^{3/2})}{\sqrt{\pi_{i^\star}^{\mathcal{H}}}} \\
&\overset{(b)}{\leq} O\left(\mu^{3/2}\right) + \frac{O(\mu^{3/2})}{\pi_{i^\star}^{\mathcal{H}}} \qquad (6.42)
\end{aligned}$$

where (a) follows from Cauchy-Schwarz, (b) follows from $\sqrt{\pi_{i^\star}^{\mathcal{H}}} \geq \pi_{i^\star}^{\mathcal{H}}$ since $\pi_{i^\star}^{\mathcal{H}} \leq 1$ so that:

$$\begin{aligned}
&\mathbb{E}\left\{J(\boldsymbol{w}_{c,i^\star+i})|\,\boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\leq \mathbb{E}\left\{J(\boldsymbol{w}'_{c,i^\star+i})|\,\boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} + O\left(\mu^{3/2}\right) + \frac{O(\mu^{3/2})}{\pi_{i^\star}^{\mathcal{H}}}
\end{aligned} \tag{6.43}$$

We conclude that the function value at $\boldsymbol{w}_{c,i^\star+i}$ after i iterations is upper-bounded by the function evaluated at the short-term model $\boldsymbol{w}'_{c,i^\star+i}$ with an additional approximation error that is bounded. We conclude that it is sufficient to study the dynamics of the short-term model, which is more tractable. Specifically, in light of the bound (6.23) following from the Lipschitz-Hessian Assumption 6.5, we have:

$$\begin{aligned}
J(\boldsymbol{w}'_{c,i^\star+i}) \leq\, & J(\boldsymbol{w}_{c,i^\star}) - \nabla J(\boldsymbol{w}_{c,i^\star})^{\mathsf{T}} \widetilde{\boldsymbol{w}}_i'^{i^\star} \\
& + \frac{1}{2}\|\widetilde{\boldsymbol{w}}_i'^{i^\star}\|^2_{\nabla^2 J(\boldsymbol{w}_{c,i^\star})} + \frac{\rho}{6}\|\widetilde{\boldsymbol{w}}_i'^{i^\star}\|^3
\end{aligned} \tag{6.44}$$

In order to establish escape from saddle-points, we need to carefully bound each term appearing on the right-hand side of (6.44), and to this end, we will need study the effect to the gradient noise term over several iterations. For this purpose, we introduce the following smoothness condition on the gradient noise covariance [1]:

Assumption 6.6 (**Lipschitz covariances**). *The gradient noise process has a Lipschitz covariance matrix, i.e.,*

$$R_{s,k}(\boldsymbol{w}_{k,i-1}) \triangleq \mathbb{E}\left\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})^{\mathsf{T}}|\boldsymbol{\mathcal{F}}_{i-1}\right\} \tag{6.45}$$

satisfies

$$\|R_{s,k}(x) - R_{s,k}(y)\| \leq \beta_R \|x-y\|^\gamma \tag{6.46}$$

for some β_R and $0 < \gamma \leq 4$. □

Definition 6.3. *We define the aggregate gradient noise covariance as:*

$$\mathcal{R}_{s,i}\left(\boldsymbol{\mathcal{W}}_{i-1}\right) = \mathbb{E}\left\{\boldsymbol{s}_i\,\boldsymbol{s}_i^{\mathsf{T}}\,|\boldsymbol{\mathcal{F}}_{i-1}\right\} \tag{6.47}$$

where $\boldsymbol{s}_i \triangleq \sum_{k=1}^{N} p_k \, \boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})$ *denotes the aggregate gradient noise term introduced earlier in* (6.27). □

Note that in light of this definition and the assumption that the gradient noise process is conditionally uncorrelated over space as in (6.13), we have:

$$\begin{aligned}
\mathcal{R}_{s,i}(\boldsymbol{\mathcal{w}}_{i-1}) &= \mathbb{E}\left\{\boldsymbol{s}_i \, \boldsymbol{s}_i^{\mathsf{T}} \,|\boldsymbol{\mathcal{F}}_{i-1}\right\} \\
&= \mathbb{E}\left\{\left(\sum_{k=1}^{N} p_k \, \boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right)\left(\sum_{k=1}^{N} p_k \, \boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right)^{\mathsf{T}} |\boldsymbol{\mathcal{F}}_{i-1}\right\} \\
&= \mathbb{E}\left\{\sum_{k=1}^{N} p_k^2 \, \boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) \, \boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})^{\mathsf{T}} |\boldsymbol{\mathcal{F}}_{i-1}\right\} \\
&= \sum_{k=1}^{N} p_k^2 \, \mathbb{E}\left\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) \, \boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})^{\mathsf{T}} |\boldsymbol{\mathcal{F}}_{i-1}\right\} \\
&= \sum_{k=1}^{N} p_k^2 R_{s,k}(\boldsymbol{w}_{k,i-1})
\end{aligned} \tag{6.48}$$

so that the aggregate gradient noise covariance is a weighted combination of the individual gradient noise covariances, albeit evaluated at different iterates. In light of the smoothness assumption 6.6, we are nevertheless able to approximate the aggregate noise covariance by one that is evaluated at the centroid.

Lemma 6.1 (**Noise covariance at centroid**). *Under assumptions 6.1–6.6 and for sufficiently small step-sizes* μ*, we have for all* i *and* $w \in \mathbb{R}^M$*:*

$$\|\mathcal{R}_{s,i}(\mathbb{1} \otimes \boldsymbol{w}_{c,i-1}) - \mathcal{R}_{s,i}(\mathbb{1} \otimes w)\| \leq p_{\max}\beta_R \, \|\boldsymbol{w}_{c,i-1} - w\|^{\gamma} \tag{6.49}$$

$$\|\mathcal{R}_{s,i}(\boldsymbol{\mathcal{w}}_{c,i-1}) - \mathcal{R}_{s,i}(\boldsymbol{\mathcal{w}}_{i-1})\| \leq p_{\max}\beta_R \, \|\boldsymbol{\mathcal{w}}_{c,i-1} - \boldsymbol{\mathcal{w}}_{i-1}\|^{\gamma} \tag{6.50}$$

Proof. Appendix 6.A. □

Note that from the bound on the aggregate gradient noise variance (6.14), we can upper

bound the gradient noise covariance:

$$\|\mathcal{R}_{s,i}(\mathcal{w})\| = \left\|\mathbb{E}\,\boldsymbol{s}_i\,\boldsymbol{s}_i^{\mathsf{T}}\right\| \overset{(a)}{\leq} \mathbb{E}\left\|\boldsymbol{s}_i\,\boldsymbol{s}_i^{\mathsf{T}}\right\| = \mathbb{E}\,\|\boldsymbol{s}_i\|^2 \overset{(6.14)}{\leq} \sigma^2 \tag{6.51}$$

where (a) follows from Jensen's inequality. In order to ensure escape from saddle-points, we introduce a similar, lower-bound condition.

Assumption 6.7 (**Gradient noise in strict saddle-points**). *Suppose w is an approximate strict-saddle points, i.e., $w \in \mathcal{H}$ and denote the eigendecomposition of the Hessian as $\nabla^2 J(w) = V\Lambda V^{\mathsf{T}}$. We introduce the decomposition:*

$$V = \begin{bmatrix} V^{\geq 0} & V^{<0} \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \Lambda^{\geq 0} & 0 \\ 0 & \Lambda^{<0} \end{bmatrix} \tag{6.52}$$

where $\Lambda^{\geq 0} \geq 0$ and $\Lambda^{<0} < 0$. Then, we assume that:

$$\lambda_{\min}\left(\left(V^{<0}\right)^{\mathsf{T}}\mathcal{R}_s\left(\mathbb{1}\otimes w\right)V^{<0}\right) \geq \sigma_\ell^2 \tag{6.53}$$

for some $\sigma_\ell^2 > 0$ and all $w \in \mathcal{H}$. □

Assumption 6.7 is similar to the condition in [134], where alternating step-sizes are employed, and essentially states than for every strict-saddle point in the set $\mathcal{H}$, there is gradient noise present along some descent direction, spanned by the eigenvectors corresponding to the negative eigenvalues of the Hessian $\nabla^2 J(\cdot)$.

Theorem 6.1 (Descent through strict saddle-points). *Suppose* $\Pr\{\boldsymbol{w}_{c,i^\star} \in \mathcal{H}\} \neq 0$, *i.e., $\boldsymbol{w}_{c,i^\star}$ is approximately stationary with significant negative eigenvalue. Then, iterating for i^s iterations after $i^\star$ with*

$$i^s = \frac{\log\left(2M\frac{\sigma^2}{\sigma_\ell^2}+1\right)}{\log(1+2\mu\tau)} \leq O\left(\frac{1}{\mu\tau}\right) \tag{6.54}$$

guarantees

$$\begin{aligned}&\mathbb{E}\left\{J(\boldsymbol{w}_{c,i^\star+i^s})\middle|\,\boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\}\\ &\leq \mathbb{E}\left\{J(\boldsymbol{w}_{c,i^\star})\middle|\,\boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\}-\frac{\mu}{2}M\sigma^2+o(\mu)+\frac{o(\mu)}{\pi_{i^\star}^{\mathcal{H}}}\end{aligned}\tag{6.55}$$

Proof. Appendix 6.B. □

This result establishes that, even if $\boldsymbol{w}_{c,i^\star}$ is an $O(\mu)$-square-stationary point and Theorem 5.2 can no longer guarantee sufficient descent, the expected function value at the network centroid will continue to decrease, as long as the Hessian matrix has a sufficiently negative eigenvalue.

6.4 Main Result

In Theorem 5.2, we established a descent condition for points with large gradient norm $\boldsymbol{w}_{c,i}\in\mathcal{G}$, while Theorem 6.1 guarantees descent in i^s iterations for strict-saddle points $\boldsymbol{w}_{c,i}\in\mathcal{H}$. Together, they establish descent whenever $\boldsymbol{w}_{c,i}\in\mathcal{G}\cup\mathcal{H}=\mathcal{M}^C$. Hence, we conclude that, as long as the cost is bounded from below, the algorithm must necessarily reach a point in $\mathcal{M}$ after a finite amount of iterations. This intuition is formalized in the following theorem.

Theorem 6.2. *For sufficiently small step-sizes μ, we have with probability $1-\pi$, that $\boldsymbol{w}_{c,i^o}\in\mathcal{M}$, i.e., $\|\nabla J(\boldsymbol{w}_{c,i^o})\|^2\leq O(\mu)$ and $\lambda_{\min}\left(\nabla^2 J(\boldsymbol{w}_{c,i^o})\right)\geq-\tau$ in at most i^o iterations, where*

$$i^o\leq\frac{(J(w_{c,0})-J^o)}{\mu^2 c_2\pi}i^s\tag{6.56}$$

and i^s denotes the escape time from Theorem 6.1, i.e.,

$$i^s=\frac{\log\left(2M\frac{\sigma^2}{\sigma_\ell^2}+1\right)}{\log(1+2\mu\tau)}\leq O\left(\frac{1}{\mu\tau}\right)\tag{6.57}$$

Proof. Appendix 6.C. □

This final result states that with probability $1-\pi$, where we are free to choose the desired confidence level, the diffusion strategy (6.2a)–(6.2b) will have visited an approximately second-order stationary point after at most i^o iterations.

6.5 Simulation Results

In this section, we consider an example that will allow us to visualize the ability of the diffusion strategy to escape saddle-points. Given a binary class label $\boldsymbol{\gamma} \in \{0, 1\}$ and feature vector $\boldsymbol{h} \in \mathbb{R}^M$, we consider a neural network with a single, linear hidden layer and a logistic activation function leading into the output layer:

$$\widehat{\boldsymbol{\gamma}}\,(\boldsymbol{h}) \triangleq \frac{1}{1+e^{-w_1^{\mathsf{T}} W_2\,\boldsymbol{h}}} \tag{6.58}$$

with weights $w_1 \in \mathbb{R}^L, W_2 \in \mathbb{R}^{L\times M}$ of appropriate dimensions. A popular risk function for training is the cross-entropy loss:

$$Q(w_1, W_2; \boldsymbol{\gamma}, \boldsymbol{h}) \triangleq -\boldsymbol{\gamma}\log(\widehat{\boldsymbol{\gamma}}) - (1-\boldsymbol{\gamma})\log(1-\widehat{\boldsymbol{\gamma}}) \tag{6.59}$$

Note that, the first term is non-zero, while the second term is zero if, and only if, $\boldsymbol{\gamma} = 1$, in which case we have:

$$-\boldsymbol{\gamma}\log(\widehat{\boldsymbol{\gamma}}) = \log\left(1+e^{-w_1^{\mathsf{T}} W_2\,\boldsymbol{h}}\right) \tag{6.60}$$

Similarly, the second term is non-zero while the first term is zero if, and only if, $\boldsymbol{\gamma} = 0$, which implies:

$$\begin{aligned}
-(1-\boldsymbol{\gamma})\log(1-\widehat{\boldsymbol{\gamma}}) &= -\log\left(1 - \frac{1}{1+e^{-w_1^{\mathsf{T}} W_2\,\boldsymbol{h}}}\right) \\
&= -\log\left(\frac{e^{-w_1^{\mathsf{T}} W_2\,\boldsymbol{h}}}{1+e^{-w_1^{\mathsf{T}} W_2\,\boldsymbol{h}}}\right) \\
&= -\log\left(\frac{1}{1+e^{w_1^{\mathsf{T}} W_2\,\boldsymbol{h}}}\right) \\
&= \log\left(1+e^{w_1^{\mathsf{T}} W_2\,\boldsymbol{h}}\right) \qquad (6.61)
\end{aligned}$$

Letting $\boldsymbol{\gamma}' \in \{-1, 1\}$ such that:

$$\boldsymbol{\gamma}' \triangleq \begin{cases} -1, & \text{if } \boldsymbol{\gamma} = 0 \\ 1, & \text{if } \boldsymbol{\gamma} = 1. \end{cases} \qquad (6.62)$$

we can hence simplify (6.59) to an equivalent logistic loss:

$$Q(w_1, W_2; \boldsymbol{\gamma}', \boldsymbol{h}) = \log\left(1+e^{-\boldsymbol{\gamma}' w_1^{\mathsf{T}} W_2\,\boldsymbol{h}}\right) \qquad (6.63)$$

The regularized learning problem can then be formulated as:

$$J(w_1, W_2) = \mathbb{E}\, Q(w_1, W_2; \boldsymbol{\gamma}', \boldsymbol{h}) + \frac{\rho}{2}\|w_1\|^2 + \frac{\rho}{2}\|W_2\|_F^2 \qquad (6.64)$$

which fits into the framework (6.1) treated in this chapter. In order to be able to visualize and enumerate all stationary points of (6.64), we assume in the sequel that $M = L = 1$ so that all involved quantities are scalar variables. We can then find:

$$\nabla J(w_1, W_2) = \mathbb{E}\begin{pmatrix} \rho w_1 - \frac{\boldsymbol{\gamma}' W_2\,\boldsymbol{h}}{e^{\boldsymbol{\gamma}' w_1 W_2\,\boldsymbol{h}}} \\ \rho W_2 - \frac{\boldsymbol{\gamma}' w_1\,\boldsymbol{h}}{e^{\boldsymbol{\gamma}' w_1 W_2\,\boldsymbol{h}}} \end{pmatrix} \qquad (6.65)$$

The cost surface is depicted in Fig. 6.2. It can be observed from the figure, and analytically verified, that $J(\cdot)$ has two local minima in the positive and negative quadrants, respectively,

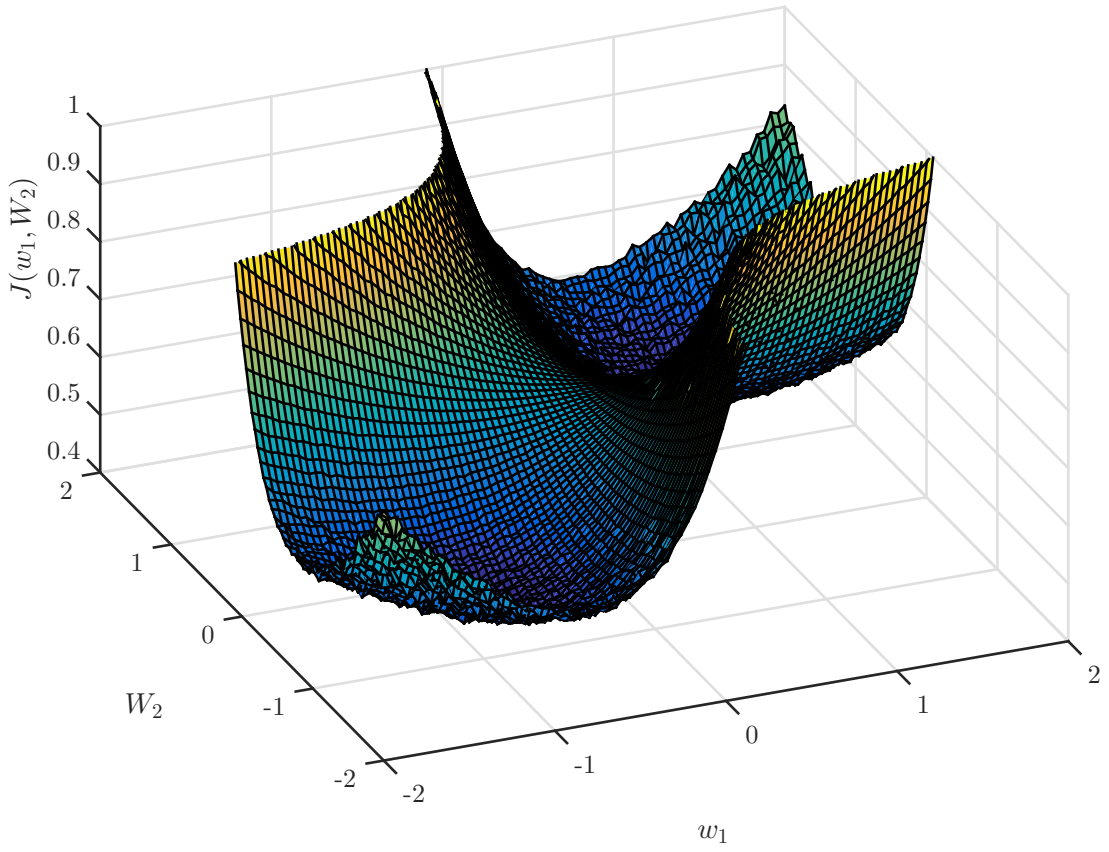

Figure 6.2: Cost surface of a simple neural network with $\rho = 0.1$.

and a single saddle-point at $w_1 = W_2 = 0$. The Hessian matrix of $J(\cdot)$ at $w_1 = W_2 = 0$ evaluates to:

$$\nabla^2 J(0,0) = \begin{pmatrix} \rho & -\mathbb{E}\,\frac{\boldsymbol{\gamma}'\boldsymbol{h}}{2} \\ -\mathbb{E}\,\frac{\boldsymbol{\gamma}'\boldsymbol{h}}{2} & \rho \end{pmatrix} \tag{6.66}$$

For this example, we let $\Pr\{\boldsymbol{\gamma}' = -1\} = \Pr\{\boldsymbol{\gamma}' = 1\} = \frac{1}{2}$ and $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{\gamma}', 1)$. Then, we obtain $\mathbb{E}\,\boldsymbol{\gamma}'\boldsymbol{h} = 1$. We also let $\rho = 0.1$, so that:

$$\nabla^2 J(0,0) = \begin{pmatrix} 0.1 & -0.5 \\ -0.5 & 0.1 \end{pmatrix} \tag{6.67}$$

which has an eigenvalue at -0.4 with corresponding eigenvector $\operatorname{col}\{1,1\}$. This implies that $w_1 = W_2 = 0$ is a strict saddle-point with local descent direction $\operatorname{col}\{1,1\}$. It turns out,

however, that the gradient noise induced by the immediate stochastic gradient approximation $\widehat{\nabla J}(\cdot) = \nabla Q(\cdot;\boldsymbol{\gamma}',\boldsymbol{h})$ does not have a gradient noise component in the descent direction $\mathrm{col}\{1,1\}$ at the strict saddle-point $w_1 = W_2 = 0$. Indeed, note that with probability one we have $\nabla Q(0,0;\boldsymbol{\gamma}',\boldsymbol{h}) = \mathrm{col}\{0,0\} = \nabla J(0,0)$ so that the gradient noise vanishes at $w_1 = W_2 = 0$. Hence, initializing all agents at $w_1 = W_2 = 0$ and iterating (6.2a)–(6.2b) would cause them to remain there with probability 1. This suggests that assumption 6.7 is not merely a technical condition but indeed necessary. To satisfy the assumption we construct the stochastic gradient approximation as:

$$\widehat{\nabla J}(w_1, W_2) \triangleq \nabla Q(w_1, W_2;\boldsymbol{\gamma}',\boldsymbol{h}) + \boldsymbol{v}\cdot\mathrm{col}\{1,1\} \tag{6.68}$$

where $\boldsymbol{v} \sim \mathcal{N}(0,1)$ acts only in the direction $\mathrm{col}\{1,1\}$ and ensures that gradient noise is present in the descent direction around the strict saddle-point at $w_1 = W_2 = 0$. Two realizations of the evolution are shown in Figures 6.3–6.4.

6.A Proof of Lemma 6.1

Recall that

$$\boldsymbol{s}_i \triangleq \sum_{k=1}^{N} p_k\,\boldsymbol{s}_{k,i}\left(\boldsymbol{w}_{k,i-1}\right) \tag{6.69}$$

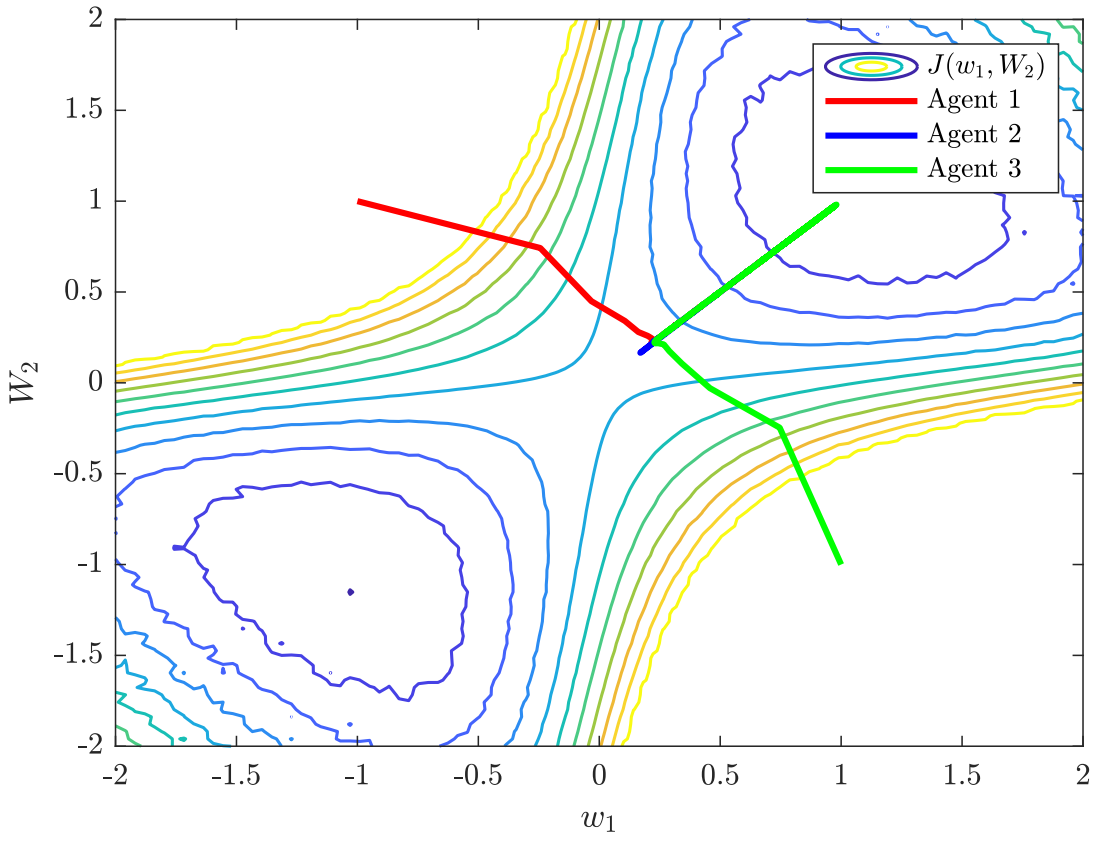

Figure 6.3: Agents are initialized at different points in space, but nevertheless quickly cluster. They then jointly travel away from the strict saddle-point and towards one of the local minimizers.

and hence (6.48) holds. Using the smoothness assumption on the gradient noise term (6.46), we can write:

$$\begin{aligned}
&\mathbb{E}\left\{\boldsymbol{s}_i\,\boldsymbol{s}_i^{\mathsf{T}}\,|\boldsymbol{\mathcal{F}}_{i-1}\right\} \\
&= \sum_{k=1}^{N} p_k^2 R_{s,k}\left(\boldsymbol{w}_{k,i-1}\right) \\
&= \sum_{k=1}^{N} p_k^2 R_{s,k}\left(\boldsymbol{w}_{c,i-1}\right) \\
&\quad + \left(\sum_{k=1}^{N} p_k^2 R_{s,k}\left(\boldsymbol{w}_{k,i-1}\right) - \sum_{k=1}^{N} p_k^2 R_{s,k}\left(\boldsymbol{w}_{c,i-1}\right)\right)
\end{aligned} \tag{6.70}$$

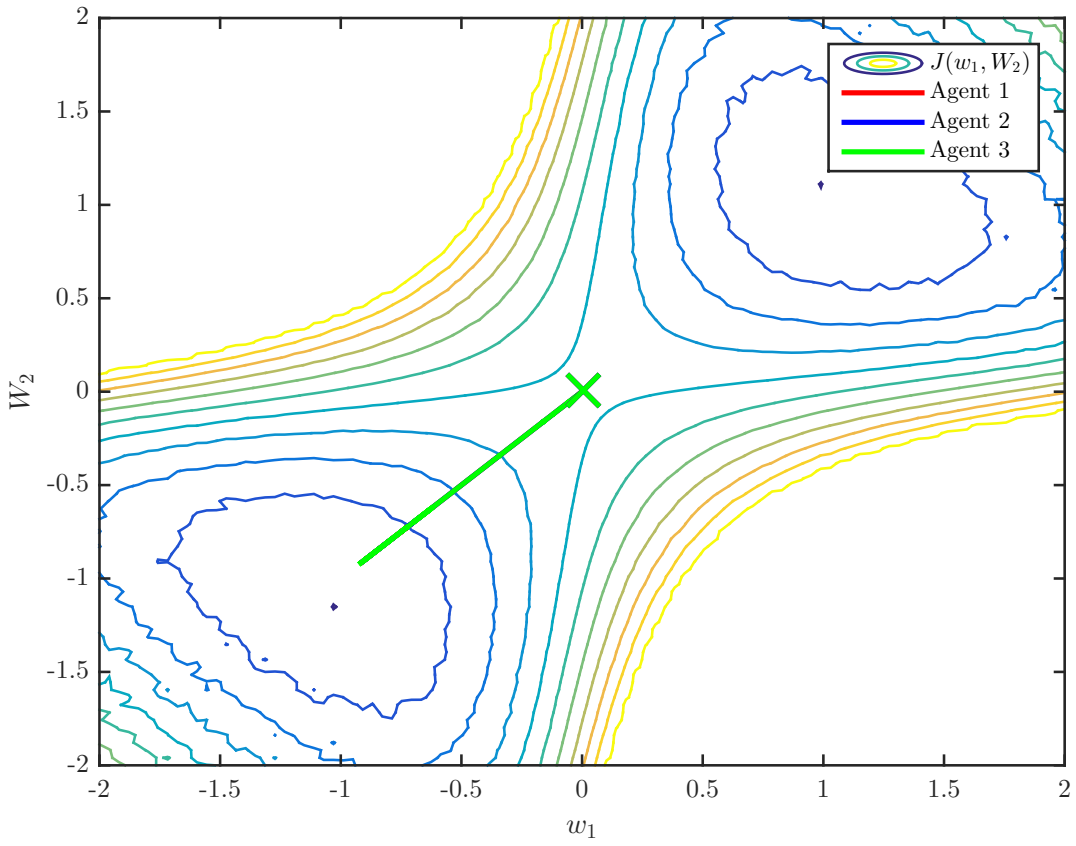

Figure 6.4: Agents are initialized together precisely in the strict saddle-point. The presence of the gradient perturbation allows them to jointly escape the saddle-point.

so that:

$$
\begin{aligned}
&\|\mathcal{R}_s\left(\mathbb{1}\otimes \boldsymbol{w}_{c,i-1}\right) - \mathcal{R}_s\left(\mathbb{1}\otimes w\right)\| \\
&= \left\|\sum_{k=1}^{N} p_k^2 R_{s,k}\left(\boldsymbol{w}_{c,i-1}\right) - \sum_{k=1}^{N} p_k^2 R_{s,k}\left(w\right)\right\| \\
&= \left\|\sum_{k=1}^{N} p_k^2 \left(R_{s,k}\left(\boldsymbol{w}_{c,i-1}\right) - R_{s,k}\left(w\right)\right)\right\| \\
&\overset{(a)}{\leq} \sum_{k=1}^{N} p_k \left\|p_k \left(R_{s,k}\left(\boldsymbol{w}_{c,i-1}\right) - R_{s,k}\left(w\right)\right)\right\| \\
&\leq p_{\max} \sum_{k=1}^{N} p_k \left\|R_{s,k}\left(\boldsymbol{w}_{c,i-1}\right) - R_{s,k}\left(w\right)\right\|
\end{aligned}
$$

$$\overset{(b)}{\leq} p_{\max}\beta_R \left\|\boldsymbol{w}_{c,i-1} - w\right\|^{\gamma} \tag{6.71}$$

where (a) follows from Jensen's inequality and (b) follows from the Lipschitz condition on the gradient noise covariance (6.46) and $\sum_{k=1}^{N} p_k = 1$. Similarly:

$$\begin{aligned}
&\left\|R_s\left(\boldsymbol{\mathcal{W}}_{i-1}\right) - R_s\left(\boldsymbol{\mathcal{W}}_{c,i-1}\right)\right\| \\
&= \left\|\sum_{k=1}^{N} p_k^2 R_{s,k}\left(\boldsymbol{w}_{k,i-1}\right) - \sum_{k=1}^{N} p_k^2 R_{s,k}\left(\boldsymbol{w}_{c,i-1}\right)\right\| \\
&= \left\|\sum_{k=1}^{N} p_k^2 \left(R_{s,k}\left(\boldsymbol{w}_{k,i-1}\right) - R_{s,k}\left(\boldsymbol{w}_{c,i-1}\right)\right)\right\| \\
&\leq \sum_{k=1}^{N} p_k \left\|p_k \left(R_{s,k}\left(\boldsymbol{w}_{k,i-1}\right) - R_{s,k}\left(\boldsymbol{w}_{c,i-1}\right)\right)\right\| \\
&\leq p_{\max}\beta_R \sum_{k=1}^{N} p_k \left\|\boldsymbol{w}_{k,i-1} - \boldsymbol{w}_{c,i-1}\right\|^{\gamma} \\
&\overset{(a)}{\leq} p_{\max}\beta_R \sum_{k=1}^{N} p_k \left\|\boldsymbol{\mathcal{W}}_{i-1} - \boldsymbol{\mathcal{W}}_{c,i-1}\right\|^{\gamma} \\
&= p_{\max}\beta_R \left\|\boldsymbol{\mathcal{W}}_{i-1} - \boldsymbol{\mathcal{W}}_{c,i-1}\right\|^{\gamma}
\end{aligned} \tag{6.72}$$

where (a) follows from the fact that x^{γ} is monotonically increasing in γ for $x, \gamma > 0$ and:

$$\begin{aligned}
\left\|\boldsymbol{\mathcal{W}}_{i-1} - \boldsymbol{\mathcal{W}}_{c,i-1}\right\|^2 &= \sum_{k=1}^{N} \left\|\boldsymbol{w}_{k,i-1} - \boldsymbol{w}_{c,i-1}\right\|^2 \\
&\geq \left\|\boldsymbol{w}_{\ell,i-1} - \boldsymbol{w}_{c,i-1}\right\|^2, \quad \forall\ \ell
\end{aligned} \tag{6.73}$$

6.B Proof of Theorem 6.1

We shall carefully bound each of the terms appearing on the right-hand side of (6.44), which we repeat here again for reference:

$$\begin{aligned} J(\boldsymbol{w}'_{c,i^\star+i}) \leq J(\boldsymbol{w}_{c,i^\star}) - \nabla J(\boldsymbol{w}_{c,i^\star})^{\mathsf{T}} \widetilde{\boldsymbol{w}}'^{i^\star}_i \\ + \frac{1}{2}\big\|\widetilde{\boldsymbol{w}}'^{i^\star}_i\big\|^2_{\nabla^2 J(\boldsymbol{w}_{c,i^\star})} + \frac{\rho}{6}\big\|\widetilde{\boldsymbol{w}}'^{i^\star}_i\big\|^3 \end{aligned} \tag{6.74}$$

We begin by establishing a bound on the linear term in (6.74). Iterating the recursive relation for the short-term model (6.34) and taking expectations conditioned on $\boldsymbol{\mathcal{F}}_{i^\star+i}$ yields:

$$\begin{aligned} \mathbb{E}\left\{\widetilde{\boldsymbol{w}}'^{i^\star}_{i+1}|\boldsymbol{\mathcal{F}}_{i^\star+i}\right\} &= \big(I - \mu\nabla^2 J(\boldsymbol{w}_{c,i^\star})\big)\,\widetilde{\boldsymbol{w}}'^{i^\star}_i \\ &\quad + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) + \mu\,\mathbb{E}\left\{\boldsymbol{s}_{i^\star+i+1}\,|\boldsymbol{\mathcal{F}}_{i^\star+i}\right\} \\ &= \big(I - \mu\nabla^2 J(\boldsymbol{w}_{c,i^\star})\big)\,\widetilde{\boldsymbol{w}}'^{i^\star}_i + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) \end{aligned} \tag{6.75}$$

where the gradient-noise term disappeared in light of

$$\mathbb{E}\left\{\boldsymbol{s}_{i^\star+i+1}\,|\boldsymbol{\mathcal{F}}_{i^\star+i}\right\} = 0 \tag{6.76}$$

by Assumption 6.4. Note that $\boldsymbol{\mathcal{F}}_{i^\star+i}$ denotes the information captured in $\boldsymbol{w}_{k,j}$ up to time $i^\star+i$, while $\boldsymbol{\mathcal{F}}_{i^\star}$ denotes the information available up to time $i^\star$. Hence:

$$\boldsymbol{\mathcal{F}}_{i^\star+i} = \boldsymbol{\mathcal{F}}_{i^\star} \cup \text{filtration}\left\{\boldsymbol{w}_{k,i^\star+1},\ldots,\boldsymbol{w}_{k,i^\star+i}\right\} \tag{6.77}$$

Hence, taking expectation of (6.75) conditioned on $\boldsymbol{\mathcal{F}}_{i^\star}$ removes the elements not contained in $\boldsymbol{\mathcal{F}}_{i^\star}$ and yields:

$$\mathbb{E}\left\{\widetilde{\boldsymbol{w}}'^{i^\star}_{i+1}|\boldsymbol{\mathcal{F}}_{i^\star}\right\} = \big(I - \mu\nabla^2 J(\boldsymbol{w}_{c,i^\star})\big)\,\mathbb{E}\left\{\widetilde{\boldsymbol{w}}'^{i^\star}_i\,|\boldsymbol{\mathcal{F}}_{i^\star}\right\} + \mu\nabla J(\boldsymbol{w}_{c,i^\star}) \tag{6.78}$$

Since $\widetilde{\boldsymbol{w}}'^{i^\star}_0 = 0$, iterating starting at $i = 0$ yields:

$$\mathbb{E}\left\{\widetilde{\boldsymbol{w}}'^{i^\star}_i | \boldsymbol{\mathcal{F}}_{i^\star}\right\} = \mu\left(\sum_{k=1}^{i}\left(I - \mu\nabla^2 J(\boldsymbol{w}_{c,i^\star})\right)^{k-1}\right)\nabla J(\boldsymbol{w}_{c,i^\star}) \tag{6.79}$$

This allows us to bound the linear term appearing in (6.74) as:

$$\begin{aligned}
&-\mathbb{E}\left\{\nabla J(\boldsymbol{w}_{c,i^\star})^{\mathsf{T}}\widetilde{\boldsymbol{w}}'^{i^\star}_i | \boldsymbol{\mathcal{F}}_{i^\star}\right\} \\
=\ &-\nabla J(\boldsymbol{w}_{c,i^\star})^{\mathsf{T}}\,\mathbb{E}\left\{\widetilde{\boldsymbol{w}}'^{i^\star}_i | \boldsymbol{\mathcal{F}}_{i^\star}\right\} \\
\overset{(6.79)}{=}\ &-\mu\nabla J(\boldsymbol{w}_{c,i^\star})^{\mathsf{T}}\left(\sum_{k=1}^{i}\left(I - \mu\nabla^2 J(\boldsymbol{w}_{c,i^\star})\right)^{k-1}\right)\nabla J(\boldsymbol{w}_{c,i^\star}) \\
=\ &-\mu\|\nabla J(\boldsymbol{w}_{c,i^\star})\|^2_{\sum_{k=1}^{i}\left(I-\mu\nabla^2 J(\boldsymbol{w}_{c,i^\star})\right)^{k-1}}
\end{aligned} \tag{6.80}$$

We now examine the quadratic term in (6.74). To this end, we introduce the eigenvalue decomposition of the Hessian around the iterate at time $i^\star$:

$$\nabla^2 J(\boldsymbol{w}_{c,i^\star}) \triangleq \boldsymbol{V}_{i^\star}\boldsymbol{\Lambda}_{i^\star}\boldsymbol{V}_{i^\star}^{\mathsf{T}} \tag{6.81}$$

Note that both $\boldsymbol{V}_{i^\star}$ and $\boldsymbol{\Lambda}_{i^\star}$ inherit their randomness from $\boldsymbol{w}_{c,i^\star}$. As such, they are random but become deterministic when conditioning on $\boldsymbol{\mathcal{F}}_{i^\star}$. This fact will be exploited further below. To begin with, note that:

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{w}}'^{i^\star}_{i+1}\right\|^2_{\nabla^2 J(\boldsymbol{w}_{c,i^\star})} &= \left\|\widetilde{\boldsymbol{w}}'^{i^\star}_{i+1}\right\|^2_{\boldsymbol{V}_{i^\star}\boldsymbol{\Lambda}_{i^\star}\boldsymbol{V}_{i^\star}^{\mathsf{T}}} \\
&= \left\|\boldsymbol{V}_{i^\star}^{\mathsf{T}}\boldsymbol{w}_{c,i^\star} - \boldsymbol{V}_{i^\star}^{\mathsf{T}}\boldsymbol{w}'_{c,i^\star+i+1}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}} \\
&= \left\|\overline{\boldsymbol{w}}'^{i^\star}_{i+1}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}}
\end{aligned} \tag{6.82}$$

where we introduced:

$$\overline{\boldsymbol{w}}'^{i^\star}_{i+1} \triangleq \boldsymbol{V}_{i^\star}^{\mathsf{T}}\widetilde{\boldsymbol{w}}'^{i^\star}_{i+1} \tag{6.83}$$

Under this transformation, recursion (6.34) is also diagonalized, yielding:

$$\begin{aligned}\overline{\boldsymbol{w}}'^{i^\star}_{i+1} &\triangleq \boldsymbol{V}^{\mathsf{T}}_{i^\star}\widetilde{\boldsymbol{w}}'^{i^\star}_{i+1} \\ &= \boldsymbol{V}^{\mathsf{T}}_{i^\star}\left(I-\mu\nabla^2 J(\boldsymbol{w}_{c,i^\star})\right)\boldsymbol{V}_{i^\star}\boldsymbol{V}^{\mathsf{T}}_{i^\star}\widetilde{\boldsymbol{w}}'^{i^\star}_{i} \\ &\quad + \mu\boldsymbol{V}^{\mathsf{T}}_{i^\star}\nabla J(\boldsymbol{w}_{c,i^\star}) + \mu\boldsymbol{V}^{\mathsf{T}}_{i^\star}\,\boldsymbol{s}_{i^\star+i+1} \\ &= (I-\mu\boldsymbol{\Lambda}_{i^\star})\,\overline{\boldsymbol{w}}'^{i^\star}_{i} + \mu\overline{\nabla} J(\boldsymbol{w}_{c,i^\star}) + \mu\overline{\boldsymbol{s}}_{i^\star+i+1}\end{aligned} \tag{6.84}$$

with $\overline{\nabla} J(\boldsymbol{w}_{c,i^\star}) \triangleq \boldsymbol{V}^{\mathsf{T}}_{i^\star}\nabla J(\boldsymbol{w}_{c,i^\star})$ and $\overline{\boldsymbol{s}}_{i^\star+i+1} \triangleq \boldsymbol{V}^{\mathsf{T}}_{i^\star}\,\boldsymbol{s}_{i^\star+i+1}$. The presence of the gradient term, which is deterministic conditioned on $\boldsymbol{\mathcal{F}}_{i^\star}$ complicates the analysis of the evolution. It can be removed by (conditionally) centering the random variable. Specifically, applying the same transformation to the conditional mean recursion (6.78), and subtracting the transformed conditional mean on both sides of (6.84), we find:

$$\overline{\boldsymbol{w}}'^{i^\star}_{i+1} - \mathbb{E}\left\{\overline{\boldsymbol{w}}'^{i^\star}_{i+1}|\boldsymbol{\mathcal{F}}_{i^\star}\right\} = (I-\mu\boldsymbol{\Lambda}_{i^\star})\left(\overline{\boldsymbol{w}}'^{i^\star}_{i} - \mathbb{E}\left\{\overline{\boldsymbol{w}}'^{i^\star}_{i}|\boldsymbol{\mathcal{F}}_{i^\star}\right\}\right) + \mu\overline{\boldsymbol{s}}_{i^\star+i+1} \tag{6.85}$$

which allows us to cancel the driving term involving the gradient. For brevity, define the (conditionally) centered random variable:

$$\check{\boldsymbol{w}}'^{i^\star}_{i+1} = \overline{\boldsymbol{w}}'^{i^\star}_{i+1} - \mathbb{E}\left\{\overline{\boldsymbol{w}}'^{i^\star}_{i+1}|\boldsymbol{\mathcal{F}}_{i^\star}\right\} \tag{6.86}$$

so that:

$$\check{\boldsymbol{w}}'^{i^\star}_{i+1} = (I-\mu\boldsymbol{\Lambda}_{i^\star})\,\check{\boldsymbol{w}}'^{i^\star}_{i} + \mu\overline{\boldsymbol{s}}_{i^\star+i+1} \tag{6.87}$$

Before proceeding, note that we can express:

$$\begin{aligned}\mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^{i^\star}_{i}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}}|\boldsymbol{\mathcal{F}}_{i^\star}\right\} &= \mathbb{E}\left\{\left\|\overline{\boldsymbol{w}}'^{i^\star}_{i} - \mathbb{E}\left\{\overline{\boldsymbol{w}}'^{i^\star}_{i}|\boldsymbol{\mathcal{F}}_{i^\star}\right\}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}}|\boldsymbol{\mathcal{F}}_{i^\star}\right\} \\ &= \mathbb{E}\left\{\left\|\overline{\boldsymbol{w}}'^{i^\star}_{i}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}}|\boldsymbol{\mathcal{F}}_{i^\star}\right\} - \left\|\mathbb{E}\left\{\overline{\boldsymbol{w}}'^{i^\star}_{i}|\boldsymbol{\mathcal{F}}_{i^\star}\right\}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}}\end{aligned} \tag{6.88}$$

Hence, we have:

$$\begin{aligned}\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i^{\prime i^\star}\right\|^2_{\nabla^2 J(\boldsymbol{w}_{c,i^\star})}|\boldsymbol{\mathcal{F}}_{i^\star}\right\} &= \mathbb{E}\left\{\left\|\overline{\boldsymbol{w}}_i^{\prime i^\star}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}}|\boldsymbol{\mathcal{F}}_{i^\star}\right\} \\ &= \mathbb{E}\left\{\left\|\check{\boldsymbol{w}}_i^{\prime i^\star}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}}|\boldsymbol{\mathcal{F}}_{i^\star}\right\} + \left\|\mathbb{E}\left\{\overline{\boldsymbol{w}}_i^{\prime i^\star}|\boldsymbol{\mathcal{F}}_{i^\star}\right\}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}} \end{aligned} \tag{6.89}$$

In order to make claims about $\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i^{\prime i^\star}\right\|^2_{\nabla^2 J(\boldsymbol{w}_{c,i^\star})}|\boldsymbol{\mathcal{F}}_{i^\star}\right\}$ by studying $\mathbb{E}\left\{\left\|\check{\boldsymbol{w}}_i^{\prime i^\star}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}}|\boldsymbol{\mathcal{F}}_{i^\star}\right\}$, we need to establish a bound on $\left\|\mathbb{E}\left\{\overline{\boldsymbol{w}}_i^{\prime i^\star}|\boldsymbol{\mathcal{F}}_{i^\star}\right\}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}}$. We have:

$$\begin{aligned}&\left\|\mathbb{E}\left\{\overline{\boldsymbol{w}}_i^{\prime i^\star}|\boldsymbol{\mathcal{F}}_{i^\star}\right\}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}} \\ &= \left\|\mathbb{E}\left\{\boldsymbol{V}_{i^\star}^{\mathsf{T}}\widetilde{\boldsymbol{w}}_i^{\prime i^\star}|\boldsymbol{\mathcal{F}}_{i^\star}\right\}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}} \\ &\overset{(6.79)}{=} \mu^2\left\|\boldsymbol{V}_{i^\star}^{\mathsf{T}}\left(\sum_{k=1}^{i}\left(I-\mu\nabla^2 J(\boldsymbol{w}_{c,i^\star})\right)^{k-1}\right)\nabla J(\boldsymbol{w}_{c,i^\star})\right\|^2_{\boldsymbol{\Lambda}_{i^\star}} \\ &= \mu^2\left\|\left(\sum_{k=1}^{i}\left(I-\mu\boldsymbol{\Lambda}_{i^\star}\right)^{k-1}\right)\overline{\nabla} J(\boldsymbol{w}_{c,i^\star})\right\|^2_{\boldsymbol{\Lambda}_{i^\star}} \\ &= \mu^2\overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\mathsf{T}}\left(\sum_{k=1}^{i}\left(I-\mu\boldsymbol{\Lambda}_{i^\star}\right)^{k-1}\right)\boldsymbol{\Lambda}_{i^\star}\left(\sum_{k=1}^{i}\left(I-\mu\boldsymbol{\Lambda}_{i^\star}\right)^{k-1}\right)\overline{\nabla} J(\boldsymbol{w}_{c,i^\star}) \end{aligned} \tag{6.90}$$

We shall order the eigenvalues of $\nabla^2 J(\boldsymbol{w}_{c,i^\star})$, such that its eigendecomposition has a block structure:

$$\boldsymbol{V}_{i^\star} = \begin{bmatrix} \boldsymbol{V}_{i^\star}^{\geq 0} & \boldsymbol{V}_{i^\star}^{<0} \end{bmatrix}, \quad \boldsymbol{\Lambda}_{i^\star} = \begin{bmatrix} \boldsymbol{\Lambda}_{i^\star}^{\geq 0} & 0 \\ 0 & \boldsymbol{\Lambda}_{i^\star}^{<0} \end{bmatrix} \tag{6.91}$$

with $\delta I \geq \boldsymbol{\Lambda}_{i^\star}^{\geq 0} \geq 0$ and $\boldsymbol{\Lambda}_{i^\star}^{<0} < 0$. Note that since $\nabla^2 J(\boldsymbol{w}_{c,i^\star})$ is random, the decomposition itself is random as well. Nevertheless, it exists with probability one. We also decompose the transformed gradient vector with appropriate dimensions:

$$\overline{\nabla} J(\boldsymbol{w}_{c,i^\star}) = \operatorname{col}\left\{\overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\geq 0}, \overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{<0}\right\} \tag{6.92}$$

We can then decompose (6.90):

$$\left\|\mathbb{E}\left\{\overline{\boldsymbol{w}}_i^{\prime i^\star}|\boldsymbol{\mathcal{F}}_{i^\star}\right\}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}}$$

$$
\begin{aligned}
&= \mu^2 \overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\mathsf{T}} \left(\sum_{k=1}^{i} (I - \mu \boldsymbol{\Lambda}_{i^\star})^{k-1} \right) \boldsymbol{\Lambda}_{i^\star} \\
&\qquad \times \left(\sum_{k=1}^{i} (I - \mu \boldsymbol{\Lambda}_{i^\star})^{k-1} \right) \overline{\nabla} J(\boldsymbol{w}_{c,i^\star}) \\
&= \mu^2 \Big(\overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\succeq 0} \Big)^{\mathsf{T}} \left(\sum_{k=1}^{i} \big(I - \mu \boldsymbol{\Lambda}_{i^\star}^{\succeq 0}\big)^{k-1} \right) \boldsymbol{\Lambda}_{i^\star}^{\succeq 0} \\
&\qquad \times \left(\sum_{k=1}^{i} \big(I - \mu \boldsymbol{\Lambda}_{i^\star}^{\succeq 0}\big)^{k-1} \right) \overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\succeq 0} \\
&\quad + \mu^2 \Big(\overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\prec 0} \Big)^{\mathsf{T}} \left(\sum_{k=1}^{i} \big(I - \mu \boldsymbol{\Lambda}_{i^\star}^{\prec 0}\big)^{k-1} \right) \boldsymbol{\Lambda}_{i^\star}^{\prec 0} \\
&\qquad \times \left(\sum_{k=1}^{i} \big(I - \mu \boldsymbol{\Lambda}_{i^\star}^{\prec 0}\big)^{k-1} \right) \overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\prec 0} \\
&\overset{(a)}{\leq} \mu^2 \Big(\overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\succeq 0} \Big)^{\mathsf{T}} \left(\sum_{k=1}^{i} \big(I - \mu \boldsymbol{\Lambda}_{i^\star}^{\succeq 0}\big)^{k-1} \right) \boldsymbol{\Lambda}_{i^\star}^{\succeq 0} \\
&\qquad \times \left(\sum_{k=1}^{i} \big(I - \mu \boldsymbol{\Lambda}_{i^\star}^{\succeq 0}\big)^{k-1} \right) \overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\succeq 0} \\
&\overset{(b)}{\leq} \mu^2 \Big(\overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\succeq 0} \Big)^{\mathsf{T}} \left(\sum_{k=1}^{\infty} \big(I - \mu \boldsymbol{\Lambda}_{i^\star}^{\succeq 0}\big)^{k-1} \right) \boldsymbol{\Lambda}_{i^\star}^{\succeq 0} \\
&\qquad \times \left(\sum_{k=1}^{i} \big(I - \mu \boldsymbol{\Lambda}_{i^\star}^{\succeq 0}\big)^{k-1} \right) \overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\succeq 0} \\
&\overset{(c)}{=} \mu^2 \Big(\overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\succeq 0} \Big)^{\mathsf{T}} \big(\mu \boldsymbol{\Lambda}_{i^\star}^{\succeq 0}\big)^{-1} \boldsymbol{\Lambda}_{i^\star}^{\succeq 0} \\
&\qquad \times \left(\sum_{k=1}^{i} \big(I - \mu \boldsymbol{\Lambda}_{i^\star}^{\succeq 0}\big)^{k-1} \right) \overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\succeq 0} \\
&= \mu \Big(\overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\succeq 0} \Big)^{\mathsf{T}} \left(\sum_{k=1}^{i} \big(I - \mu \boldsymbol{\Lambda}_{i^\star}^{\succeq 0}\big)^{k-1} \right) \overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\succeq 0} \\
&\overset{(d)}{\leq} \mu \Big(\overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\succeq 0} \Big)^{\mathsf{T}} \left(\sum_{k=1}^{i} \big(I - \mu \boldsymbol{\Lambda}_{i^\star}^{\succeq 0}\big)^{k-1} \right) \overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\succeq 0} \\
&\quad + \mu \Big(\overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\prec 0} \Big)^{\mathsf{T}} \left(\sum_{k=1}^{i} \big(I - \mu \boldsymbol{\Lambda}_{i^\star}^{\prec 0}\big)^{k-1} \right) \overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\prec 0} \\
&\leq \mu \overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\mathsf{T}} \left(\sum_{k=1}^{i} (I - \mu \boldsymbol{\Lambda}_{i^\star})^{k-1} \right) \overline{\nabla} J(\boldsymbol{w}_{c,i^\star})
\end{aligned}
$$

$$= \mu \big\| \overline{\nabla} J(\boldsymbol{w}_{c,i^\star}) \big\|^2_{\sum_{k=1}^{i} (I - \mu \boldsymbol{\Lambda}_{i^\star})^{k-1}} \tag{6.93}$$

where (a) follows from $\boldsymbol{\Lambda}_{i^\star}^{<0} < 0$, (b) follows from:

$$\sum_{k=1}^{k} \left(I - \mu \boldsymbol{\Lambda}_{i^\star}^{\geq 0} \right)^{k-1} \leq \sum_{k=1}^{\infty} \left(I - \mu \boldsymbol{\Lambda}_{i^\star}^{\geq 0} \right)^{k-1} \tag{6.94}$$

for $\mu < \frac{1}{\delta}$. Step (c) follows from the formula for the geometric matrix series, and (d) follows from:

$$\mu \Big(\overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\geq 0} \Big)^{\mathsf{T}} \left(\sum_{k=1}^{i} \left(I - \mu \boldsymbol{\Lambda}_{i^\star}^{\geq 0} \right)^{k-1} \right) \overline{\nabla} J(\boldsymbol{w}_{c,i^\star})^{\geq 0} \geq 0 \tag{6.95}$$

Comparing (6.93) to (6.80), we find that we can bound:

$$- \mathbb{E} \left\{ \nabla J(\boldsymbol{w}_{c,i^\star})^{\mathsf{T}} \widetilde{\boldsymbol{w}}_i'^{i^\star} | \boldsymbol{\mathcal{F}}_{i^\star} \right\} + \big\| \mathbb{E} \left\{ \overline{\boldsymbol{w}}_i'^{i^\star} | \boldsymbol{\mathcal{F}}_{i^\star} \right\} \big\|^2_{\boldsymbol{\Lambda}_{i^\star}} \leq 0 \tag{6.96}$$

To recap, we can simplify (6.74) as:

$$\mathbb{E} \left\{ J(\boldsymbol{w}'_{c,i^\star+i}) | \boldsymbol{\mathcal{F}}_{i^\star} \right\} \leq J(\boldsymbol{w}_{c,i^\star}) + \frac{1}{2} \mathbb{E} \left\{ \big\| \check{\boldsymbol{w}}'^{i^\star}_i \big\|^2_{\boldsymbol{\Lambda}_{i^\star}} | \boldsymbol{\mathcal{F}}_{i^\star} \right\} + \frac{\rho}{6} \mathbb{E} \left\{ \big\| \widetilde{\boldsymbol{w}}'^{i^\star}_i \big\|^3 | \boldsymbol{\mathcal{F}}_{i^\star} \right\} \tag{6.97}$$

We proceed with the now simplified quadratic term. Motivated by a technique employed for the analysis of adaptive filters and stochastic gradient algorithms in *convex* environments [1, 149], we square both sides of (6.87) under an arbitrary diagonal weighting matrix $\boldsymbol{\Sigma}_i$, deterministic conditioned on $\boldsymbol{w}_{c,i^\star}$ and $\boldsymbol{w}_{c,i^\star+i}$, to obtain:

$$\begin{aligned} \big\| \check{\boldsymbol{w}}'^{i^\star}_{i+1} \big\|^2_{\boldsymbol{\Sigma}_i} &= \big\| (I - \mu \boldsymbol{\Lambda}_{i^\star}) \, \check{\boldsymbol{w}}'^{i^\star}_i + \mu \overline{\boldsymbol{s}}_{i^\star+i+1} \big\|^2_{\boldsymbol{\Sigma}_i} \\ &= \big\| (I - \mu \boldsymbol{\Lambda}_{i^\star}) \, \check{\boldsymbol{w}}'^{i^\star}_i \big\|^2_{\boldsymbol{\Sigma}_i} + \mu^2 \| \overline{\boldsymbol{s}}_{i^\star+i+1} \|^2_{\boldsymbol{\Sigma}_i} + 2\mu \check{\boldsymbol{w}}'^{i^\star \mathsf{T}}_i (I - \mu \boldsymbol{\Lambda}_{i^\star}) \, \boldsymbol{\Sigma}_i \overline{\boldsymbol{s}}_{i^\star+i+1} \end{aligned} \tag{6.98}$$

Note that upon conditioning on $\boldsymbol{\mathcal{F}}_{i^\star+i}$, all elements of the cross-term, aside from $\overline{\boldsymbol{s}}_{i^\star+i+1}$, become deterministic, and as such the term disappears when taking expectations. We obtain:

$$\begin{aligned}
\mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^{i^\star}_{i+1}\right\|^2_{\boldsymbol{\Sigma}_i}|\boldsymbol{\mathcal{F}}_{i^\star+i}\right\} &= \left\|(I-\mu\boldsymbol{\Lambda}_{i^\star})\,\check{\boldsymbol{w}}'^{i^\star}_{i}\right\|^2_{\boldsymbol{\Sigma}_i} + \mu^2\,\mathbb{E}\left\{\|\overline{\boldsymbol{s}}_{i^\star+i+1}\|^2_{\boldsymbol{\Sigma}_i}|\boldsymbol{\mathcal{F}}_{i^\star+i}\right\} \\
&= \left\|\check{\boldsymbol{w}}'^{i^\star}_{i}\right\|^2_{\boldsymbol{\Sigma}_i-2\mu\boldsymbol{\Lambda}_{i^\star}\boldsymbol{\Sigma}_i+\mu^2\boldsymbol{\Lambda}_{i^\star}\boldsymbol{\Sigma}_i\boldsymbol{\Lambda}_{i^\star}} + \mu^2\mathrm{Tr}\left(\boldsymbol{V}_{i^\star}\boldsymbol{\Sigma}_i\boldsymbol{V}_{i^\star}^{\mathsf{T}}\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{i^\star+i}\right)\right) \\
&= \left\|\check{\boldsymbol{w}}'^{i^\star}_{i}\right\|^2_{\boldsymbol{\Sigma}_i-2\mu\boldsymbol{\Lambda}_{i^\star}\boldsymbol{\Sigma}_i} + \mu^2\mathrm{Tr}\left(\boldsymbol{V}_{i^\star}\boldsymbol{\Sigma}_i\boldsymbol{V}_{i^\star}^{\mathsf{T}}\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star}\right)\right) \\
&\quad + \mu^2\mathrm{Tr}\left(\boldsymbol{V}_{i^\star}\boldsymbol{\Sigma}_i\boldsymbol{V}_{i^\star}^{\mathsf{T}}\left(\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{i^\star+i}\right)-\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star}\right)\right)\right) \\
&\quad + \mu^2\left\|\check{\boldsymbol{w}}'^{i^\star}_{i}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}\boldsymbol{\Sigma}_i\boldsymbol{\Lambda}_{i^\star}}
\end{aligned} \tag{6.99}$$

We proceed to bound the last two terms. First, we have:

$$\begin{aligned}
&\mathrm{Tr}\left(\boldsymbol{V}_{i^\star}\boldsymbol{\Sigma}_i\boldsymbol{V}_{i^\star}^{\mathsf{T}}\left(\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{i^\star+i}\right)-\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star}\right)\right)\right) \\
&\overset{(a)}{\leq} \left\|\boldsymbol{V}_{i^\star}\boldsymbol{\Sigma}_i\boldsymbol{V}_{i^\star}^{\mathsf{T}}\right\|\left\|\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{i^\star+i}\right)-\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star}\right)\right\| \\
&\leq \left\|\boldsymbol{V}_{i^\star}\boldsymbol{\Sigma}_i\boldsymbol{V}_{i^\star}^{\mathsf{T}}\right\|\left\|\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{i^\star+i}\right)-\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star+i}\right)+\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star+i}\right)-\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star}\right)\right\| \\
&\leq \left\|\boldsymbol{V}_{i^\star}\boldsymbol{\Sigma}_i\boldsymbol{V}_{i^\star}^{\mathsf{T}}\right\|\left\|\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{i^\star+i}\right)-\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star+i}\right)\right\| \\
&\quad + \left\|\boldsymbol{V}_{i^\star}\boldsymbol{\Sigma}_i\boldsymbol{V}_{i^\star}^{\mathsf{T}}\right\|\left\|\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star+i}\right)-\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star}\right)\right\| \\
&\overset{(b)}{\leq} \rho\left(\boldsymbol{\Sigma}_i\right)\beta_R p_{\max}\left(\left\|\boldsymbol{w}_{c,i^\star+i}-\boldsymbol{w}_{c,i^\star}\right\|^\gamma+\left\|\boldsymbol{\mathcal{W}}_{c,i^\star+i}-\boldsymbol{\mathcal{W}}_{i^\star+i}\right\|^\gamma\right) \\
&= \rho\left(\boldsymbol{\Sigma}_i\right)\beta_R p_{\max}\left(\left\|\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|^\gamma+\left\|\boldsymbol{\mathcal{W}}_{c,i^\star+i}-\boldsymbol{\mathcal{W}}_{i^\star+i}\right\|^\gamma\right)
\end{aligned} \tag{6.100}$$

where (a) follows from Cauchy-Schwarz, since $\mathrm{Tr}(A^{\mathsf{T}}B)$ is an inner product over the space of symmetric matrices, and hence, $|\mathrm{Tr}(A^{\mathsf{T}}B)| \leq \|A\|\|B\|$, and (b) follows from Lemma (6.1). For the second term, we have:

$$\begin{aligned}
\left\|\check{\boldsymbol{w}}'^{i^\star}_{i}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}\boldsymbol{\Sigma}_i\boldsymbol{\Lambda}_{i^\star}} &\leq \rho\left(\boldsymbol{\Lambda}_{i^\star}\boldsymbol{\Sigma}_i\boldsymbol{\Lambda}_{i^\star}\right)\left\|\check{\boldsymbol{w}}'^{i^\star}_{i}\right\|^2 \\
&\leq \delta^2\rho\left(\boldsymbol{\Sigma}_i\right)\left\|\check{\boldsymbol{w}}'^{i^\star}_{i}\right\|^2
\end{aligned} \tag{6.101}$$

We conclude that

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^{i^\star}_{i+1}\right\|^2_{\boldsymbol{\Sigma}_i}|\boldsymbol{\mathcal{F}}_{i^\star}\right\} \\
&= \mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^{i^\star}_{i}\right\|^2_{\boldsymbol{\Sigma}_i-2\mu\boldsymbol{\Lambda}_{i^\star}\boldsymbol{\Sigma}_i}|\boldsymbol{\mathcal{F}}_{i^\star}\right\} \\
&\quad + \mu^2\mathrm{Tr}\left(\boldsymbol{V}_{i^\star}\boldsymbol{\Sigma}_i\boldsymbol{V}^{\mathsf{T}}_{i^\star}\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star}\right)\right) + \mu^2\rho\left(\boldsymbol{\Sigma}_i\right)\mathbb{E}\left\{\boldsymbol{q}_{i^\star+i}|\boldsymbol{\mathcal{F}}_{i^\star}\right\}
\end{aligned} \tag{6.102}$$

where

$$\boldsymbol{q}_{i^\star+i} \triangleq \beta_R p_{\max}\left(\left\|\widetilde{\boldsymbol{w}}^{i^\star}_i\right\|^\gamma + \left\|\boldsymbol{\mathcal{W}}_{c,i^\star+i} - \boldsymbol{\mathcal{W}}_{i^\star+i}\right\|^\gamma\right) + \delta^2\left\|\check{\boldsymbol{w}}'^{i^\star}_i\right\|^2 \tag{6.103}$$

For brevity, we define

$$\boldsymbol{D} \triangleq I - 2\mu\boldsymbol{\Lambda}_{i^\star} \tag{6.104}$$

$$\boldsymbol{Y} \triangleq \boldsymbol{V}^{\mathsf{T}}_{i^\star}\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star}\right)\boldsymbol{V}_{i^\star} \tag{6.105}$$

With these substitutions we obtain:

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^{i^\star}_{i+1}\right\|^2_{\boldsymbol{\Sigma}_i}|\boldsymbol{\mathcal{F}}_{i^\star}\right\} \\
&= \mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^{i^\star}_{i}\right\|^2_{\boldsymbol{D}\boldsymbol{\Sigma}_i}|\boldsymbol{\mathcal{F}}_{i^\star}\right\} + \mu^2\mathrm{Tr}\left(\boldsymbol{\Sigma}_i\boldsymbol{Y}\right) + \mu^2\rho\left(\boldsymbol{\Sigma}_i\right)\mathbb{E}\left\{\boldsymbol{q}_{i^\star+i}|\boldsymbol{\mathcal{F}}_{i^\star}\right\}
\end{aligned} \tag{6.106}$$

At $i = 0$, we have:

$$\check{\boldsymbol{w}}'^{i^\star}_0 = \overline{\boldsymbol{w}}'^{i^\star}_0 - \mathbb{E}\left\{\overline{\boldsymbol{w}}'^{i^\star}_0|\boldsymbol{\mathcal{F}}_{i^\star}\right\} = 0 - 0 = 0 \tag{6.107}$$

Letting $\boldsymbol{\Sigma}_i = \boldsymbol{\Lambda}_{i^\star}\boldsymbol{D}^i$, we can iterate to obtain:

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^{i^\star}_{i+1}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}}|\boldsymbol{\mathcal{F}}_{i^\star}\right\} \\
&= \mu^2\sum_{n=0}^{i}\mathrm{Tr}\left(\boldsymbol{\Lambda}_{i^\star}\boldsymbol{D}^n\boldsymbol{Y}\right) + \mu^2\sum_{n=0}^{i}\rho\left(\boldsymbol{\Lambda}_{i^\star}\boldsymbol{D}^n\right)\cdot\mathbb{E}\left\{\boldsymbol{q}_{i^\star+n}|\boldsymbol{\mathcal{F}}_{i^\star}\right\} \\
&= \mu^2\mathrm{Tr}\left(\boldsymbol{\Lambda}_{i^\star}\left(\sum_{n=0}^{i}\boldsymbol{D}^n\right)\boldsymbol{Y}\right) + \mu^2\sum_{n=0}^{i}\rho\left(\boldsymbol{\Lambda}_{i^\star}\boldsymbol{D}^n\right)\cdot\mathbb{E}\left\{\boldsymbol{q}_{i^\star+n}|\boldsymbol{\mathcal{F}}_{i^\star}\right\}
\end{aligned} \tag{6.108}$$

since $\overline{\boldsymbol{w}}'_{c,i^\star+i+1} = \overline{\boldsymbol{w}}_{c,i^\star}$ at $i = 0$. Our objective is to show that the first term on the right-hand side yields sufficient descent (i.e., will be sufficiently negative), while the second term is small enough to be negligible. To this end, we again make use of the structured eigendecomposition (6.91). We have:

$$
\begin{aligned}
&\mu^2 \mathrm{Tr}\left(\boldsymbol{\Lambda}_{i^\star}\left(\sum_{n=0}^{i}\boldsymbol{D}^n\right)\boldsymbol{V}_{i^\star}^{\mathsf{T}}\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star}\right)\boldsymbol{V}_{i^\star}\right)\\
&\overset{(a)}{=}\mu^2\mathrm{Tr}\Bigg(\boldsymbol{\Lambda}_{i^\star}^{\geq 0}\left(\sum_{n=0}^{i}\left(I-2\mu\boldsymbol{\Lambda}_{i^\star}^{\geq 0}\right)^n\right)\\
&\qquad\qquad\times\left(\boldsymbol{V}_{i^\star}^{\geq 0}\right)^{\mathsf{T}}\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star}\right)\boldsymbol{V}_{i^\star}^{\geq 0}\Bigg)\\
&\quad+\mu^2\mathrm{Tr}\Bigg(\boldsymbol{\Lambda}_{i^\star}^{<0}\left(\sum_{n=0}^{i}\left(I-2\mu\boldsymbol{\Lambda}_{i^\star}^{<0}\right)^n\right)\\
&\qquad\qquad\times\left(\boldsymbol{V}_{i^\star}^{<0}\right)^{\mathsf{T}}\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star}\right)\boldsymbol{V}_{i^\star}^{<0}\Bigg)\\
&\overset{(b)}{=}\mu^2\mathrm{Tr}\Bigg(\boldsymbol{\Lambda}_{i^\star}^{\geq 0}\left(\sum_{n=0}^{i}\left(I-2\mu\boldsymbol{\Lambda}_{i^\star}^{\geq 0}\right)^n\right)\\
&\qquad\qquad\times\left(\boldsymbol{V}_{i^\star}^{\geq 0}\right)^{\mathsf{T}}\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star}\right)\boldsymbol{V}_{i^\star}^{\geq 0}\Bigg)\\
&\quad-\mu^2\mathrm{Tr}\Bigg(\left(-\boldsymbol{\Lambda}_{i^\star}^{<0}\right)\left(\sum_{n=0}^{i}\left(I-2\mu\boldsymbol{\Lambda}_{i^\star}^{<0}\right)^n\right)\\
&\qquad\qquad\times\left(\boldsymbol{V}_{i^\star}^{<0}\right)^{\mathsf{T}}\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star}\right)\boldsymbol{V}_{i^\star}^{<0}\Bigg)\\
&\overset{(c)}{\leq}\mu^2\mathrm{Tr}\left(\boldsymbol{\Lambda}_{i^\star}^{\geq 0}\left(\sum_{n=0}^{i}\left(I-2\mu\boldsymbol{\Lambda}_{i^\star}^{\geq 0}\right)^n\right)\right)\\
&\quad\times\lambda_{\max}\left(\left(\boldsymbol{V}_{i^\star}^{\geq 0}\right)^{\mathsf{T}}\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star}\right)\boldsymbol{V}_{i^\star}^{\geq 0}\right)\\
&\quad-\mu^2\mathrm{Tr}\left(\left(-\boldsymbol{\Lambda}_{i^\star}^{<0}\right)\left(\sum_{n=0}^{i}\left(I-2\mu\boldsymbol{\Lambda}_{i^\star}^{<0}\right)^n\right)\right)\\
&\quad\times\lambda_{\min}\left(\left(\boldsymbol{V}_{i^\star}^{<0}\right)^{\mathsf{T}}\mathcal{R}_s\left(\boldsymbol{\mathcal{W}}_{c,i^\star}\right)\boldsymbol{V}_{i^\star}^{<0}\right)\\
&\overset{(d)}{\leq}\mu^2\mathrm{Tr}\left(\boldsymbol{\Lambda}_{i^\star}^{\geq 0}\left(\sum_{n=0}^{i}\left(I-2\mu\boldsymbol{\Lambda}_{i^\star}^{\geq 0}\right)^n\right)\right)\sigma^2
\end{aligned}
$$

$$- \mu^2 \mathrm{Tr}\left(\left(-\boldsymbol{\Lambda}_{i^\star}^{<0}\right) \left(\sum_{n=0}^{i} \left(I - 2\mu \boldsymbol{\Lambda}_{i^\star}^{<0}\right)^n \right) \right) \sigma_\ell^2 \tag{6.109}$$

where in (a) we decomposed the trace since $\boldsymbol{\Lambda}_{i^\star} \left(\sum_{n=0}^{i} \boldsymbol{D}^n \right)$ is a diagonal matrix, (b) applies $-\left(-\boldsymbol{\Lambda}_{i^\star}^{<0}\right) = \boldsymbol{\Lambda}_{i^\star}^{<0}$. Step (b) follows from $\mathrm{Tr}(A)\lambda_{\min}(B) \leq \mathrm{Tr}(AB) \leq \mathrm{Tr}(A)\lambda_{\max}(B)$ which holds for $A = A^\mathsf{T}, B = B^\mathsf{T} \geq 0$, and (c) follows from the bounded covariance property (6.51) and Assumption 6.7. For the positive term, we have:

$$\begin{aligned}
&\mu^2 \mathrm{Tr}\left(\boldsymbol{\Lambda}_{i^\star}^{\geq 0} \left(\sum_{n=0}^{i} \left(I - 2\mu \boldsymbol{\Lambda}_{i^\star}^{\geq 0}\right)^n \right) \right) \sigma^2 \\
&\overset{(a)}{\leq} \mu^2 \mathrm{Tr}\left(\boldsymbol{\Lambda}_{i^\star}^{\geq 0} \left(\sum_{n=0}^{\infty} \left(I - 2\mu \boldsymbol{\Lambda}_{i^\star}^{\geq 0}\right)^n \right) \right) \sigma^2 \\
&\overset{(b)}{\leq} \mu^2 \mathrm{Tr}\left(\boldsymbol{\Lambda}_{i^\star}^{\geq 0} \left(2\mu \boldsymbol{\Lambda}_{i^\star}^{\geq 0}\right)^{-1} \right) \sigma^2 \overset{(c)}{\leq} \frac{\mu}{2} M \sigma^2
\end{aligned} \tag{6.110}$$

where (a) follows since $I - 2\mu\boldsymbol{\Lambda}_{i^\star}^{\geq 0}$ is elementwise non-negative for $\mu \leq \frac{2}{\delta}$, (b) follows from $\sum_{n=0}^{\infty} A^n = (I - A)^{-1}$ and (c) follows since $\nabla^2 J(\boldsymbol{w}_{c,i^\star})$ is of dimension M.

For the negative term, we have under expectation conditioned on $\boldsymbol{w}_{c,i^\star} \in \mathcal{H}$:

$$\begin{aligned}
&\mathbb{E}\left\{ \mathrm{Tr}\left(\left(-\boldsymbol{\Lambda}_{i^\star}^{<0}\right) \left(\sum_{n=0}^{i} \left(I - 2\mu \boldsymbol{\Lambda}_{i^\star}^{<0}\right)^n \right) \right) \sigma_\ell^2 \,\middle|\, \boldsymbol{w}_{c,i^\star} \in \mathcal{H} \right\} \\
&\overset{(a)}{\geq} \mathbb{E}\left\{ \tau \left(\sum_{n=0}^{i} (1 + 2\mu\tau)^n \right) \sigma_\ell^2 \,\middle|\, \boldsymbol{w}_{c,i^\star} \in \mathcal{H} \right\} \\
&\overset{(b)}{=} \tau \left(\sum_{n=0}^{i} (1 + 2\mu\tau)^n \right) \sigma_\ell^2 \overset{(c)}{=} \tau \frac{1 - (1 + 2\mu\tau)^{i+1}}{1 - (1 + 2\mu\tau)} \sigma_\ell^2 \\
&= \frac{1}{2\mu} \left((1 + 2\mu\tau)^{i+1} - 1 \right) \sigma_\ell^2
\end{aligned} \tag{6.111}$$

Step (a) makes use of the fact that $\left(-\boldsymbol{\Lambda}_{i^\star}^{<0}\right) \left(\sum_{n=0}^{i} \left(I - 2\mu \boldsymbol{\Lambda}_{i^\star}^{<0}\right)^n \right)$ is a diagonal matrix, where all elements are non-negative. Hence, its trace can be bounded by any of its diagonal

elements:

$$\begin{aligned}
&\operatorname{Tr}\left(\left(-\boldsymbol{\Lambda}_{i^\star}^{<0}\right)\left(\sum_{n=0}^{i}\left(I-2\mu\boldsymbol{\Lambda}_{i^\star}^{<0}\right)^n\right)\right)\\
&\overset{(6.17)}{\geq}\ \tau\left(\sum_{n=0}^{i}\left(1+2\mu\tau\right)^n\right)
\end{aligned}\tag{6.112}$$

In (b) we dropped the expectation since the expression is no longer random, and (c) is the result of a geometric series. We return to the full expression (6.109) and find:

$$\begin{aligned}
&\mu^2\,\mathbb{E}\left\{\operatorname{Tr}\left(\boldsymbol{\Lambda}_{i^\star}\left(\sum_{n=0}^{i}\boldsymbol{D}^n\right)\right.\right.\\
&\qquad\left.\left.\times\,\boldsymbol{V}_{i^\star}^{\mathsf{T}}\mathcal{R}_s\left(\boldsymbol{w}_{c,i^\star}\right)\boldsymbol{V}_{i^\star}\right)\,|\,\boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\}\\
&\leq\frac{\mu}{2}M\sigma^2-\frac{\mu}{2}\left(\left(1+2\mu\tau\right)^{i+1}-1\right)\sigma_\ell^2\overset{(a)}{\leq}\ -\frac{\mu}{2}M\sigma^2
\end{aligned}\tag{6.113}$$

where (a) holds if, and only if,

$$\begin{aligned}
&\frac{\mu}{2}M\sigma^2-\frac{\mu}{2}\left(\left(1+2\mu\tau\right)^{i+1}-1\right)\sigma_\ell^2\leq-\frac{\mu}{2}M\sigma^2\\
\iff&2M\frac{\sigma^2}{\sigma_\ell^2}+1\leq\left(1+2\mu\tau\right)^{i+1}\\
\iff&\log\left(2M\frac{\sigma^2}{\sigma_\ell^2}+1\right)\leq(i+1)\log\left(1+2\mu\tau\right)\\
\iff&\frac{\log\left(2M\frac{\sigma^2}{\sigma_\ell^2}+1\right)}{\log\left(1+2\mu\tau\right)}\leq i+1\\
\iff&\frac{\log\left(2M\frac{\sigma^2}{\sigma_\ell^2}+1\right)}{O(\mu\tau)}\leq i+1
\end{aligned}\tag{6.114}$$

where the last line follows from $\lim_{x\to 0}1/x\log(1+x)=1$. We conclude that there exists a bounded i^s such that:

$$\mu^2\,\mathbb{E}\left\{\operatorname{Tr}\left(\boldsymbol{\Lambda}_{i^\star}\left(\sum_{n=0}^{i^s}\boldsymbol{D}^n\right)\boldsymbol{V}_{i^\star}^{\mathsf{T}}\mathcal{R}_s\left(\boldsymbol{w}_{c,i^\star}\right)\boldsymbol{V}_{i^\star}\right)\right\}\leq-\frac{\mu}{2}M\sigma^2\tag{6.115}$$

Applying this relation to (6.108) and taking expectations over $\boldsymbol{w}_{c,i^\star} \in \mathcal{H}$, we obtain:

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^{i^\star}_{i^s+1}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}} | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\leq \mu^2 \sum_{n=0}^{i^s} \mathbb{E}\left\{\left(\mathrm{Tr}\left(\boldsymbol{\Lambda}_{i^\star}\boldsymbol{D}^n\right) \cdot \mathbb{E}\left\{\boldsymbol{q}_{i^\star+n} | \boldsymbol{\mathcal{F}}_{i^\star}\right\}\right) | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} - \frac{\mu}{2} M\sigma^2
\end{aligned} \tag{6.116}$$

We now bound the perturbation term:

$$\begin{aligned}
&\mu^2 \sum_{n=0}^{i^s} \mathbb{E}\left\{\left(\rho\left(\boldsymbol{\Lambda}_{i^\star}\boldsymbol{D}^n\right) \cdot \mathbb{E}\left\{\boldsymbol{q}_{i^\star+n} | \boldsymbol{\mathcal{F}}_{i^\star}\right\}\right) | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\leq \mu^2 \sum_{n=0}^{i^s} \mathbb{E}\left\{\left(\rho\left(\delta I(I+2\mu\delta I)^n\right) \cdot \mathbb{E}\left\{\boldsymbol{q}_{i^\star+n} | \boldsymbol{\mathcal{F}}_{i^\star}\right\}\right) | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&= \mu^2 \sum_{n=0}^{i^s} \left(\delta(1+2\mu\delta)^n \cdot \mathbb{E}\left\{\boldsymbol{q}_{i^\star+n} | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\}\right) \\
&\overset{(6.103)}{=} \mu^2 \sum_{n=0}^{i^s} \delta(1+2\mu\delta)^n \cdot \Bigg(\beta_R p_{\max}\Big(\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_i^{i^\star}\right\|^\gamma | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \\
&\qquad + \mathbb{E}\left\{\|\boldsymbol{\mathcal{W}}_{c,i^\star+i} - \boldsymbol{\mathcal{W}}_{i^\star+i}\|^\gamma | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\}\Big) \\
&\qquad + \delta^2 \mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^{i^\star}_i\right\|^2 | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\}\Bigg) \\
&\leq \mu^2 \sum_{n=0}^{i^s} \delta(1+2\mu\delta)^n \cdot \left(O(\mu^\gamma) + \frac{O(\mu^\gamma)}{\pi^{\mathcal{H}}_{i^\star}} + O(\mu^2)\right) \\
&\leq \delta\left(\sum_{n=0}^{i^s}(1+2\mu\delta)^n\right)\left(O(\mu^{2+\gamma}) + \frac{O(\mu^{2+\gamma})}{\pi^{\mathcal{H}}_{i^\star}}\right) \\
&\overset{(a)}{\leq} O(\mu^{1+\gamma}) + \frac{O(\mu^{1+\gamma})}{\pi^{\mathcal{H}}_{i^\star}} = \; o(\mu) + \frac{o(\mu)}{\pi^{\mathcal{H}}_{i^\star}}
\end{aligned} \tag{6.117}$$

where (a) follows from Lemma [70, Lemma 3]. We conclude:

$$\mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^{i^\star}_{i^s+1}\right\|^2_{\boldsymbol{\Lambda}_{i^\star}} | \, \boldsymbol{w}_{c,i^\star} \in \mathcal{H}\right\} \leq -\frac{\mu}{2} M\sigma^2 + o(\mu) + \frac{o(\mu)}{\pi^{\mathcal{H}}_{i^\star}} \tag{6.118}$$

Returning to (6.97), we find:

$$\begin{aligned}
&\mathbb{E}\left\{J(\boldsymbol{w}'_{c,i^\star+i})|\,\boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\}\\
&\leq \mathbb{E}\left\{J(\boldsymbol{w}_{c,i^\star})|\,\boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\}+\frac{1}{2}\,\mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^{i^\star}_i\right\|^2_{\boldsymbol{\Lambda}_{i^\star}}|\,\boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\}\\
&\quad+\frac{\rho}{6}\,\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}'^{i^\star}_i\right\|^3|\,\boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\}\\
&\leq \mathbb{E}\left\{J(\boldsymbol{w}_{c,i^\star})|\,\boldsymbol{w}_{c,i^\star}\in\mathcal{H}\right\}-\frac{\mu}{2}M\sigma^2+o(\mu)+\frac{o(\mu)}{\pi^{\mathcal{H}}_{i^\star}}
\end{aligned}\tag{6.119}$$

6.C Proof of Theorem 6.2

The proof follows by constructing a particular telescoping sum and subsequently applying [70, Theorem 2] and 6.1. In a manner similar to [59], we define the stochastic process:

$$\mathbf{t}(k+1)=\begin{cases}\mathbf{t}(k)+1, & \text{if } \boldsymbol{w}_{c,\mathbf{t}(k)}\in\mathcal{G},\\ \mathbf{t}(k)+1, & \text{if } \boldsymbol{w}_{c,\mathbf{t}(k)}\in\mathcal{M},\\ \mathbf{t}(k)+i_s, & \text{if } \boldsymbol{w}_{c,\mathbf{t}(k)}\in\mathcal{H}.\end{cases}\tag{6.120}$$

where $\mathbf{t}(0)=0$. We then have:

$$\begin{aligned}
&\mathbb{E}\left\{J(\boldsymbol{w}_{c,\mathbf{t}(k)})-J(\boldsymbol{w}_{c,\mathbf{t}(k+1)})|\,\boldsymbol{w}_{c,\mathbf{t}(k)}\in\mathcal{G}\right\}\\
&=\mathbb{E}\left\{J(\boldsymbol{w}_{c,\mathbf{t}(k)})-J(\boldsymbol{w}_{c,\mathbf{t}(k)+1})|\,\boldsymbol{w}_{c,\mathbf{t}(k)}\in\mathcal{G}\right\}\\
&\geq \mu^2\frac{c_2}{\pi}-O(\mu^3)-\frac{O(\mu^3)}{\pi^{\mathcal{G}}_i}
\end{aligned}\tag{6.121}$$

and

$$\begin{aligned}
&\mathbb{E}\left\{J(\boldsymbol{w}_{c,\mathbf{t}(k)})-J(\boldsymbol{w}_{c,\mathbf{t}(k+1)})|\,\boldsymbol{w}_{c,\mathbf{t}(k)}\in\mathcal{H}\right\}\\
&=\mathbb{E}\left\{J(\boldsymbol{w}_{c,\mathbf{t}(k)})-J(\boldsymbol{w}_{c,\mathbf{t}(k)+i^s})|\,\boldsymbol{w}_{c,\mathbf{t}(k)}\in\mathcal{H}\right\}\\
&\geq \frac{\mu}{2}M\sigma^2-o(\mu)-\frac{o(\mu)}{\pi^{\mathcal{H}}_i}
\end{aligned}\tag{6.122}$$

Finally, we have:

$$\begin{aligned}
&\mathbb{E}\left\{J(\boldsymbol{w}_{c,\mathbf{t}(k)}) - J(\boldsymbol{w}_{c,\mathbf{t}(k+1)})\middle|\, \boldsymbol{w}_{c,\mathbf{t}(k)} \in \mathcal{M}\right\} \\
&= \mathbb{E}\left\{J(\boldsymbol{w}_{c,\mathbf{t}(k)}) - J(\boldsymbol{w}_{c,\mathbf{t}(k)+1})\middle|\, \boldsymbol{w}_{c,\mathbf{t}(k)} \in \mathcal{M}\right\} \\
&\geq -\mu^2 c_2 - O(\mu^3) - \frac{O(\mu^3)}{\pi_i^{\mathcal{M}}} \qquad (6.123)
\end{aligned}$$

where (a) follows since $\mathbf{t}(k+1)-\mathbf{t}(k) = 1$ when $\boldsymbol{w}_{c,\mathbf{t}(k)} \in \mathcal{M}$. We can combine these relations to obtain:

$$\begin{aligned}
&\mathbb{E}\left\{J(\boldsymbol{w}_{c,\mathbf{t}(k)}) - \mathbb{E}\, J(\boldsymbol{w}_{c,\mathbf{t}(k+1)})\right\} \\
&= \mathbb{E}\left\{J(\boldsymbol{w}_{c,\mathbf{t}(k)}) - \mathbb{E}\, J(\boldsymbol{w}_{c,\mathbf{t}(k+1)})\middle|\, \boldsymbol{w}_{c,\mathbf{t}(k)} \in \mathcal{G}\right\} \cdot \pi_{\mathbf{t}(k)}^{\mathcal{G}} \\
&\quad + \mathbb{E}\left\{J(\boldsymbol{w}_{c,\mathbf{t}(k)}) - \mathbb{E}\, J(\boldsymbol{w}_{c,\mathbf{t}(k+1)})\middle|\, \boldsymbol{w}_{c,\mathbf{t}(k)} \in \mathcal{H}\right\} \cdot \pi_{\mathbf{t}(k)}^{\mathcal{H}} \\
&\quad + \mathbb{E}\left\{J(\boldsymbol{w}_{c,\mathbf{t}(k)}) - \mathbb{E}\, J(\boldsymbol{w}_{c,\mathbf{t}(k+1)})\middle|\, \boldsymbol{w}_{c,\mathbf{t}(k)} \in \mathcal{M}\right\} \cdot \pi_{\mathbf{t}(k)}^{\mathcal{M}} \\
&= \left(\mu^2 \frac{c_2}{\pi} - O(\mu^3) - \frac{O(\mu^3)}{\pi_i^{\mathcal{G}}}\right) \cdot \pi_{\mathbf{t}(k)}^{\mathcal{G}} \\
&\quad + \left(\frac{\mu}{2} M\sigma^2 - o(\mu) - \frac{o(\mu)}{\pi_i^{\mathcal{H}}}\right) \cdot \pi_{\mathbf{t}(k)}^{\mathcal{H}} \\
&\quad + \left(-\mu^2 c_2 - O(\mu^3) - \frac{O(\mu^3)}{\pi_i^{\mathcal{M}}}\right) \cdot \pi_{\mathbf{t}(k)}^{\mathcal{M}} \\
&= \mu^2 \frac{c_2}{\pi} \cdot \pi_{\mathbf{t}(k)}^{\mathcal{G}} + \left(\frac{\mu}{2} M\sigma^2 - o(\mu)\right) \cdot \pi_{\mathbf{t}(k)}^{\mathcal{H}} \\
&\quad - \mu^2 c_2 \cdot \pi_{\mathbf{t}(k)}^{\mathcal{M}} - o(\mu^2) \qquad (6.124)
\end{aligned}$$

Suppose $\pi_{\mathbf{t}(k)}^{\mathcal{M}} \leq 1 - \pi$ for all i. Then $\pi_{\mathbf{t}(k)}^{\mathcal{G}} + \pi_{\mathbf{t}(k)}^{\mathcal{H}} \geq \pi$ for all i, and

$$\begin{aligned}
&\mathbb{E}\left\{J(\boldsymbol{w}_{c,\mathbf{t}(k)}) - \mathbb{E}\, J(\boldsymbol{w}_{c,\mathbf{t}(k+1)})\right\} \\
&\geq \mu^2 \frac{c_2}{\pi} \cdot \left(\pi - \pi_{\mathbf{t}(k)}^{\mathcal{H}}\right) + \left(\frac{\mu}{2} M\sigma^2 - o(\mu)\right) \cdot \pi_{\mathbf{t}(k)}^{\mathcal{H}} \\
&\quad - \mu^2 c_2 \cdot (1-\pi) - o(\mu^2) \\
&= \mu^2 c_2 \pi + \left(\frac{\mu}{2} M\sigma^2 - \mu^2 \frac{c_2}{\pi} - o(\mu)\right) \pi_{\mathbf{t}(k)}^{\mathcal{H}} - o(\mu^2) \\
&\overset{(a)}{\geq} \mu^2 c_2 \pi - o(\mu^2) \qquad (6.125)
\end{aligned}$$

where (a) holds whenever $\frac{\mu}{2}M\sigma^2 - \mu^2\frac{c_2}{\pi} - o(\mu) \geq 0$, which holds whenever μ is sufficiently small. We hence have by telescoping:

$$\begin{aligned}
&J(w_{c,0}) - J^o \\
\geq\ &\mathbb{E}\, J(w_{c,\mathbf{t}(0)}) - \mathbb{E}\, J(\boldsymbol{w}_{c,\mathbf{t}(k)}) \\
=\ &\mathbb{E}\, J(w_{c,\mathbf{t}(0)}) - \mathbb{E}\, J(\boldsymbol{w}_{c,\mathbf{t}(1)}) \\
&+ \mathbb{E}\, J(\boldsymbol{w}_{c,\mathbf{t}(1)}) - \mathbb{E}\, J(\boldsymbol{w}_{c,\mathbf{t}(2)}) \\
&+ \cdots \\
&+ \mathbb{E}\, J(\boldsymbol{w}_{c,\mathbf{t}(k-1)}) - \mathbb{E}\, J(\boldsymbol{w}_{c,\mathbf{t}(k)}) \\
\geq\ &\mu^2 c_2 \pi k
\end{aligned} \tag{6.126}$$

Rearranging yields:

$$k \leq \frac{J(w_{c,0}) - J^o}{\mu^2 c_2 \pi} \tag{6.127}$$

We conclude by definition of the stochastic process $\boldsymbol{t}_k$:

$$i = \boldsymbol{t}(k) \leq k \cdot i^s \leq \frac{(J(w_{c,0}) - J^o)}{\mu^2 c_2 \pi} i^s \tag{6.128}$$

CHAPTER 7

Centralized Non-Convex Optimization

Recent years have seen increased interest in performance guarantees of gradient descent algorithms for non-convex optimization. A number of works have uncovered that gradient noise plays a critical role in the ability of gradient descent recursions to efficiently escape saddle-points and reach second-order stationary points. Most available works limit the gradient noise component to be bounded with probability one or sub-Gaussian and leverage concentration inequalities to arrive at high-probability results. We present an alternate approach, relying primarily on mean-square arguments and show that a more relaxed relative bound on the gradient noise variance is sufficient to ensure efficient escape from saddle-points without the need to inject additional noise, employ alternating step-sizes or rely on a global dispersive noise assumption, as long as a gradient noise component is present in a descent direction for every saddle-point. The material in this chapter are based on [72].

In this chapter, we consider optimization problems of the form:

$$w^o \triangleq \arg\min_{w\in\mathbb{R}^M} J(w) \tag{7.1}$$

where $J(w)$ is a risk function defined as the expectation of a loss function, i.e.,

$$J(w) \triangleq \mathbb{E}_{\boldsymbol{x}}\, Q(w;\boldsymbol{x}) \tag{7.2}$$

where the expectation is over the distribution of the data variable $\boldsymbol{x}$. We wish to study first-order methods for pursuing solutions of (7.1), i.e., recursions of the form:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - \mu\widehat{\nabla J}\,(\boldsymbol{w}_{i-1}) \tag{7.3}$$

where $\widehat{\nabla J}(\boldsymbol{w}_{i-1})$ denotes some suitable update direction. When the gradient of $J(\cdot)$ can be evaluated, which in general requires the distribution of $\boldsymbol{x}$ to be known, then one popular and effective construction is to employ the actual gradient vector:

$$\widehat{\nabla J}^{\,\mathrm{G}}(\boldsymbol{w}_{i-1}) \triangleq \nabla J(\boldsymbol{w}_{i-1}) \tag{7.4}$$

When the distribution of $\boldsymbol{x}$ is unknown, we can instead rely on the stochastic gradient approximation [8]:

$$\widehat{\nabla J}^{\,\mathrm{SG}}(\boldsymbol{w}_{i-1}) \triangleq \nabla Q(\boldsymbol{w}_{i-1}, \boldsymbol{x}_i) \tag{7.5}$$

where $\nabla Q(\boldsymbol{w}_{i-1}, \boldsymbol{x}_i)$ denotes an instantaneous approximation of $\nabla J(\boldsymbol{w}_{i-1})$ based on the realization $\boldsymbol{x}_i$ observed at time i. For strongly *convex* cost functions $J(\cdot)$, both gradient (7.4) and stochastic gradient (7.5) implementations of (7.3) are very well behaved and well studied in the literature – see, e.g., [22,84] and the references therein. One particular conclusion is that, under suitable conditions on the loss function and data distribution, descent along the true gradient $\nabla J(\boldsymbol{w}_{i-1})$ results in linear convergence to the minimizer w^o, while stochastic "descent" along the instantaneous gradient approximation (7.5) results in a small performance degradation in steady-state for small step-sizes, i.e., $\limsup_{i\to\infty} \mathbb{E}\,\|w^o - \boldsymbol{w}_i\|^2 \leq O(\mu)$ [1].

One surprising fact that arises when considering non-convex cost functions is that employing stochastic or perturbed gradient directions is generally beneficial and can in fact improve the ability of an algorithm to escape saddle-points. For example, recursion (7.3) with true gradients (7.4) can take exponentially long to escape from saddle-points [148]. However, by simply perturbing the gradient by adding i.i.d. noise will allow the algorithm to escape strict saddle-points in polynomial time [59]. More formally, perturbed gradient descent takes the form [59]:

$$\widehat{\nabla J}^{\,\mathrm{PG}}(\boldsymbol{w}_{i-1}) \triangleq \nabla J(\boldsymbol{w}_{i-1}) + \boldsymbol{v}_i \tag{7.6}$$

where $\boldsymbol{v}_i$ is some i.i.d. perturbation term with positive definite covariance matrix. When the true gradient $\nabla J(\boldsymbol{w}_{i-1})$ is unavailable, the perturbation can be added instead to the

instantaneous gradient approximation [139]:

$$\widehat{\nabla J}^{\,\mathrm{PSG}}(\boldsymbol{w}_{i-1}) \triangleq \nabla Q(\boldsymbol{w}_{i-1}, \boldsymbol{x}_i) + \boldsymbol{v}_i \tag{7.7}$$

In this chapter, we will study a generic update direction $\widehat{\nabla J}(\boldsymbol{w}_{i-1})$ and examine the dynamics of (7.3) in non-convex environments under conditions that are more relaxed than typically assumed in the recent literature. To this end, we introduce the gradient noise process:

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \triangleq \nabla J(\boldsymbol{w}_{i-1}) - \widehat{\nabla J}(\boldsymbol{w}_{i-1}) \tag{7.8}$$

and write (7.3) as:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - \mu \nabla J(\boldsymbol{w}_{i-1}) - \mu\, \boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \tag{7.9}$$

Any particular choice for the gradient estimate $\widehat{\nabla J}(\boldsymbol{w}_{i-1})$ will induce a different gradient noise process (7.8) with varying properties. For example, while employing construction (7.6) results in i.i.d. gradient noise, a general construction of the form (7.5) will generally result in a gradient noise process that is no longer i.i.d.

7.1 Related Works

The results and proof techniques presented in this chapter are related to Chapters 5 and 6, which considered instead *distributed* optimization problems under an *absolute* variance bound on the gradient noise. The contribution of this current work in relation to these earlier studies is two-fold. First, we focus here solely on the case of single-agent optimization, i.e., on *centralized* as opposed to *decentralized* implementations. Second, and more importantly, by limiting our analysis to the single-agent setting, we are able to relax the *absolute* variance condition employed in [70, 71] to a mixed variance bound consisting of a mixture of *relative and absolute* components, thus leading to new performance guarantees in the centralized case.

There have of course been several other useful works on non-convex optimization using

first-order methods in the literature. The primary focus in these earlier works has been establishing convergence to first-order stationary points, i.e., points where the gradient vanishes so that $\nabla J(\boldsymbol{w}_{i-1}) = 0$ as $i \to \infty$ [142, 150–152]. First-order stationarity by itself however, is generally not a sufficient guarantee of a desirable solution since the set of first-order stationary points includes saddle-points and even local maxima. For this reason, in more recent years, there has been growing interest in convergence guarantees that exclude such undesirable first-order stationary points. To do so, one also examines second-order conditions. In particular, recall that second-order stationary points are those where not only the gradient vector is zero, but there are also restrictions on the smallest eigenvalue of the Hessian matrix at their locations [141]. These restrictions, when chosen to exclude local maxima and strict saddle-points can help ensure convergence towards local minima. Actually, under such restrictions, the stationary points can be shown to *always* correspond to local minima for some functions of interest [59, 62, 153–155].

One approach for ensuring convergence to these desirable second-order stationary points is by incorporating second-order information via the Hessian matrix into the update relation [144, 156]. Such a construction helps ensure that a descent direction can be identified even when the gradient vanishes and no longer carries directional information. For many, especially large-scale problems, evaluating the Hessian matrix at every iteration can be prohibitively costly. This fact has spawned a number of works that continue to employ first-order schemes for identifying a descent direction around saddle-points for both deterministic and stochastic optimization [135–137].

A second class of methods for the escape from saddle-points exploits the fact that strict saddle-points (defined later) are unstable, in the sense that small perturbations, either induced during initialization [61, 133] or added to the true gradient direction [59, 60, 157], will cause iterates to approach second-order stationary points almost surely. These algorithms require knowledge of the true gradient $\nabla J(\boldsymbol{w}_{i-1})$, which generally requires information about the distribution of $\boldsymbol{x}$. Strategies for *stochastic* optimization, where instantaneous approximations $\nabla Q(\boldsymbol{w}_{i-1}, \boldsymbol{x}_i)$ are employed in place of the true gradient $\nabla J(\boldsymbol{w}_{i-1})$ have also been studied recently. The works [132,140] and [139] consider perturbed stochastic gradients (7.7)

with diminishing and constant step-sizes, respectively, while [134] employs (7.5) by interlacing small and large step-sizes and the works [70, 71, 138] descend along (7.5) with constant step-sizes. This chapter is most related to these latter references — we shall make a detailed distinction when discussing the modeling conditions below. We also note that a number of recent works consider variance reduced strategies for the setting where $J(\cdot)$ corresponds to an empirical risk based on a finite number of samples [137, 142, 143]. In contrast, our focus is on the *streaming* data setting, where the sample size tends to infinity and traditional variance reduction techniques are inapplicable.

7.2 Modeling Conditions

7.2.1 Smoothness Conditions

We employ the following smoothness assumptions.

Assumption 7.1 (**Lipschitz gradients**). *The gradient $\nabla J(\cdot)$ is Lipschitz, namely, there exists $\delta > 0$ such that for any x, y:*

$$\|\nabla J(x) - \nabla J(y)\| \leq \delta \|x - y\| \tag{7.10}$$

□

Assumption 7.2 (**Lipschitz Hessians**). *The cost $J(\cdot)$ is twice-differentiable and there exists $\rho \geq 0$ such that:*

$$\|\nabla^2 J(x) - \nabla^2 J(y)\| \leq \rho \|x - y\| \tag{7.11}$$

□

Assumption 7.1 is common in the study of gradient algorithms, even for the minimization of convex function [1] and first-order stationarity in non-convex environments [150, 151]. It

implies a quadratic upper bound on the cost:

$$J(y) \leq J(x) + \nabla J(x)^{\mathsf{T}}(y-x) + \frac{\delta}{2}\|x-y\|^2 \tag{7.12}$$

and uniform lower and upper bounds on the Hessian matrix:

$$-\delta I \leq \nabla^2 J(x) \leq \delta I \tag{7.13}$$

The stronger Assumption 7.2 is not necessary to establish convergence to first-order stationary points [150]. It is frequently employed to characterize more granularly the dynamics of (stochastic) gradient algorithms around first-order stationary points, both to establish the ability of various gradient algorithms to escape saddle-points [59, 133, 137, 139] or to study the mean-square deviation of stochastic gradient implementations from minimizers in the strongly-convex setting [1]. It implies a tighter upper bound than (7.12) [144]:

$$J(y) \leq J(x) + \nabla J(x)^{\mathsf{T}}(y-x) + \frac{1}{2}(y-x)^{\mathsf{T}}\nabla^2 J(x)(y-x) + \frac{\rho}{6}\|y-x\|^3 \tag{7.14}$$

7.2.2 Gradient Noise Conditions

We shall employ the following conditions on the gradient noise process (7.8).

Definition 7.1 (Filtration). *We denote by $\boldsymbol{\mathcal{F}}_i$ the filtration generated by the random processes $\boldsymbol{w}_j$ for all $j \leq i$:*

$$\boldsymbol{\mathcal{F}}_i \triangleq \{\boldsymbol{w}_0, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_i\} \tag{7.15}$$

Informally, $\boldsymbol{\mathcal{F}}_i$ captures all information that is available about the stochastic processes $\boldsymbol{w}_j$ up to time i. □

Assumption 7.3 (Gradient noise process). *The gradient noise process* (7.8) *satisfies:*

$$\mathbb{E}\,\{\boldsymbol{s}_i(\boldsymbol{w}_{i-1})|\boldsymbol{\mathcal{F}}_{i-1}\} = 0 \tag{7.16}$$

$$\mathbb{E}\,\big\{\|\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4|\boldsymbol{\mathcal{F}}_{i-1}\big\} \leq \beta^4\|\nabla J(\boldsymbol{w}_{i-1})\|^4 + \sigma^4 \tag{7.17}$$

for some non-negative constants β^4, σ^4. $\square$

The fourth-order condition (7.17) also implies a bound on the second-order moment via Jensen's inequality:

$$\begin{aligned}\mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}\right\} &\leq \sqrt{\beta^4\|\nabla J(\boldsymbol{w}_{i-1})\|^4+\sigma^4} \\ &\overset{(a)}{\leq} \beta^2\|\nabla J(\boldsymbol{w}_{i-1})\|^2+\sigma^2 \end{aligned} \tag{7.18}$$

where (a) follows from the sub-additivity of the square root. Condition (7.18) is the same as the one employed in [151] to study first-order stationarity under a diminishing step-size rule and corresponds to a mixture of the absolute and relative noise components appearing in [84]. It is weaker than the condition assumed in works on second-order stationarity. For example, the works [59, 138] require the gradient noise process to be uniformly bounded for all $\boldsymbol{w}_i$ with probability one. This condition is relaxed in [139] by requiring the difference $\nabla J(\boldsymbol{w}_{i-1})-\nabla Q\left(\boldsymbol{w}_{i-1}, \boldsymbol{x}_i\right)$ to be sub-Gaussian and further in [70,71] by allowing for a uniform bound on the fourth-order moment. Works that employ bounded or sub-Gaussian gradient perturbation generally rely on concentration relations, which explicitly exploit the bounded or sub-Gaussian nature of the gradient noise process [139].

In this chapter, we take a different approach by anchoring our analysis around mean-square arguments. This allows us to track the evolution of the iterates $\boldsymbol{w}_i$ in the mean-square sense, rather than with high probability and avoid the need for restrictive probability bounds on the gradient noise process. Observe that condition (7.17) is weaker than a uniform bound on the fourth moment of the gradient noise process, since we allow for a relative component in the form of $\beta^4\|\nabla J(\boldsymbol{w}_{i-1})\|^4$. This condition allows for the gradient noise variance to grow away from first-order stationary points and in particular does not enforce a uniform bound on the gradient noise variance as seen from (7.18). In place of stronger bounds on the gradient noise variance, we employ a smoothness condition on the gradient noise covariance, previously employed for characterizing the mean-square deviation of stochastic gradient algorithms around the minimizer in strongly convex optimization [1].

Assumption 7.4 (**Lipschitz covariances**). *The gradient noise process has a Lipschitz covariance matrix, i.e.,*

$$R_s(\boldsymbol{w}_{i-1}) \triangleq \mathbb{E}\left\{\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\boldsymbol{s}_i(\boldsymbol{w}_{i-1})^{\mathsf{T}}|\boldsymbol{\mathcal{F}}_{i-1}\right\} \tag{7.19}$$

satisfies

$$\|R_s(x) - R_s(y)\| \leq \beta_R \|x - y\|^{\gamma} \tag{7.20}$$

for some β_R *and* $0 < \gamma \leq 4$. □

This condition essentially ensures that the second-order moment of the gradient noise process is approximately invariant so long as the iterates $\boldsymbol{w}_{i-1}$ remain sufficiently close. From the bound on the aggregate gradient noise variance (7.18), we can upper bound the gradient noise covariance as follows:

$$\begin{aligned}
&\|R_s\left(\boldsymbol{w}_{i-1}\right)\| \\
&\leq \mathbb{E}\left\{\left\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\,\boldsymbol{s}_i(\boldsymbol{w}_{i-1})^{\mathsf{T}}\right\| |\boldsymbol{\mathcal{F}}_i\right\} \\
&= \mathbb{E}\left\{\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 |\boldsymbol{\mathcal{F}}_i\right\} \\
&\overset{(7.18)}{\leq} \beta^2\|\nabla J(\boldsymbol{w}_{i-1})\|^2 + \sigma^2
\end{aligned} \tag{7.21}$$

Before introducing the final assumption, we formally define first and second-order stationary points, similar to prior works on second-order stationary guarantees [59, 70, 71, 144]. We decompose the space $w \in \mathbb{R}^M$ into four sets.

Definition 7.2 (Sets). *To simplify the notation in the sequel, we introduce following sets:*

$$\mathcal{G} \triangleq \left\{w : \|\nabla J(w)\|^2 \geq \mu\frac{c_2}{c_1}\left(1 + \frac{1}{\pi}\right)\right\} \tag{7.22}$$

$$\mathcal{G}^C \triangleq \left\{w : \|\nabla J(w)\|^2 < \mu\frac{c_2}{c_1}\left(1 + \frac{1}{\pi}\right)\right\} \tag{7.23}$$

$$\mathcal{H} \triangleq \left\{w : w \in \mathcal{G}^C, \lambda_{\min}\left(\nabla^2 J(w)\right) \leq -\tau\right\} \tag{7.24}$$

$$\mathcal{M} \triangleq \left\{w : w \in \mathcal{G}^C, \lambda_{\min}\left(\nabla^2 J(w)\right) > -\tau\right\} \tag{7.25}$$

where τ is a small positive parameter, c_1 and c_2 are constants:

$$c_1 \triangleq 1 - \mu\frac{\delta}{2}\left(1+\beta^2\right) = O(1) \tag{7.26}$$

$$c_2 \triangleq \frac{\delta}{2}\sigma^2 = O(1) \tag{7.27}$$

and $0 < \pi < 1$ is a parameter to be chosen. Note that $\mathcal{G}^C = \mathcal{H} \cup \mathcal{M}$. We also define the probabilities $\pi_i^{\mathcal{G}} \triangleq \Pr\{\boldsymbol{w}_i \in \mathcal{G}\}$, $\pi_i^{\mathcal{H}} \triangleq \Pr\{\boldsymbol{w}_i \in \mathcal{H}\}$ and $\pi_i^{\mathcal{M}} \triangleq \Pr\{\boldsymbol{w}_i \in \mathcal{M}\}$. Then, for all i, we have $\pi_i^{\mathcal{G}} + \pi_i^{\mathcal{H}} + \pi_i^{\mathcal{M}} = 1$. □

As explained in [70, 71], the above definition first decomposes the space $\mathbb{R}^M$ into the set $\mathcal{G}$, where the squared norm of the gradient is larger than $O(\mu)$ and its complement $\mathcal{G}^C$. Since the squared norm of the gradient in $\mathcal{G}^C$ is not precisely equal to zero, but nevertheless small for small step-sizes μ, we refer to these points as approximately first-order stationary. The set of approximate first-order stationary points is further decomposed into those where the Hessian matrix has a strictly negative eigenvalue $\mathcal{H}$, and those who do not $\mathcal{M}$. The set of points $\mathcal{H}$ correspond to approximate *strict* saddle-points, and are points where a descent direction could be identified from the Hessian matrix. Points in $\mathcal{M}$ are referred to as approximately second-order stationary, since they are indistinguishable from minima based on first and second-order information.

Assumption 7.5 (Gradient noise in strict saddle-points). *Suppose w is an approximate strict-saddle point, i.e., $w \in \mathcal{H}$. Introduce the eigendecomposition of the Hessian matrix as $\nabla^2 J(w) = V\Lambda V^{\mathsf{T}}$ and let the decomposition:*

$$V = \begin{bmatrix} V^{\geq 0} & V^{<0} \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \Lambda^{\geq 0} & 0 \\ 0 & \Lambda^{<0} \end{bmatrix} \tag{7.28}$$

where $\Lambda^{\geq 0} \geq 0$ and $\Lambda^{<0} < 0$. Then, we assume that:

$$\lambda_{\min}\left(\left(V^{<0}\right)^{\mathsf{T}} R_s(w)\, V^{<0}\right) \geq \sigma_\ell^2 \tag{7.29}$$

for some $\sigma_\ell^2 > 0$ and all $w \in \mathcal{H}$. □

As explained in [70,71], assumption 7.5 is similar to the condition in [134], where alternating step-sizes are employed, and ensures that at every strict saddle-point there is a gradient noise component in a descent direction with non-zero probability. It will be leveraged to establish the ability of recursion (7.3) to escape strict saddle-points. Note that, in contrast to the global dispersive noise assumption [138], condition (7.29) is only required to hold locally in the vicinity of strict saddle-points. When there is no prior information, condition (7.29) can always be guaranteed by choosing the update direction to be the perturbed stochastic gradient direction (7.7) with $\boldsymbol{v}_i \sim \mathcal{N}(0, \sigma_\ell^2 I)$, as is done in [139]. Under this construction, the additional perturbation $\boldsymbol{v}_i$ plays a similar role to ridge regularization, which is frequently added to convex optimization problems to ensure strong convexity and hence improved convergence behavior in the absence of a priori strong convexity guarantees. An alternative construction is to add perturbations selectively, when a saddle-point is detected by calculating the gradient norm, resulting in an algorithm similar to [60].

Remark #1: In order to make the notation more compact, and whenever it is clear from context, we shall omit the argument $\boldsymbol{w}_{i-1}$ from the gradient noise term and write instead $\boldsymbol{s}_i \triangleq \boldsymbol{s}_i(\boldsymbol{w}_{i-1})$ with the understanding that the gradient noise at time i is a function of the iterate $\boldsymbol{w}_{i-1}$ at time $i-1$ in addition to the data $\boldsymbol{x}_i$ at time i.

Remark #2: The proof technique used to establish the main theorems in the next section are motivated by the arguments used in the works [70,71] for distributed optimization in non-convex environments. The main difference is that the arguments need to be adjusted to accommodate the more relaxed relative variance bound (7.17) in the single-agent case.

7.3 Performance Analysis

7.3.1 Preliminary Lemmas

Before proceeding with the analysis, we list some preliminary lemmas, which will be used repeatedly throughout.

Lemma 7.1 (Conditioning [70]). *Suppose $\boldsymbol{w} \in \mathbb{R}^M$ is a random variable measurable by*

$\boldsymbol{\mathcal{F}}$. *In other words,* $\boldsymbol{w}$ *is deterministic conditioned on* $\boldsymbol{\mathcal{F}}$ *and* $\mathbb{E}\{\boldsymbol{w}\,|\boldsymbol{\mathcal{F}}\} = \boldsymbol{w}$. *Then,*

$$\mathbb{E}\Big\{\,\mathbb{E}\{\boldsymbol{x}\,|\boldsymbol{\mathcal{F}}\}\,|\,\boldsymbol{w}\in\mathcal{S}\Big\} = \mathbb{E}\{\boldsymbol{x}\,|\,\boldsymbol{w}\in\mathcal{S}\} \tag{7.30}$$

for any deterministic set $\mathcal{S}\subseteq\mathbb{R}^M$ *and random* $\boldsymbol{x}\in\mathbb{R}^M$. □

Lemma 7.2 (A limiting result). *For* $T,\mu,\delta>0$ *and* $k\in\mathbb{Z}_+$ *with* $\mu<\frac{1}{\delta}$, *we have:*

$$\lim_{\mu\to 0}\left(\frac{(1+\mu\delta)^k+O(\mu^2)}{(1-\mu\delta)^{k-1}}\right)^{\frac{T}{\mu}} = e^{-T\delta+2kT\delta} = O(1) \tag{7.31}$$

Proof. This lemma is a minor variation of the result in [70]. The adjusted proof is listed in Appendix 7.A. □

7.3.2 Large-Gradient Regime

Theorem 7.1. *For sufficiently small step-sizes:*

$$\mu \le \frac{2}{\delta\,(1+\beta^2)} \tag{7.32}$$

and when the gradient at $\boldsymbol{w}_i$ *is sufficiently large, i.e.,* $\boldsymbol{w}_i\in\mathcal{G}$, *the stochastic gradient recursion* (7.3) *yields descent in expectation in one iteration, namely,*

$$\mathbb{E}\{J(\boldsymbol{w}_{i+1})|\,\boldsymbol{w}_i\in\mathcal{G}\} \le \mathbb{E}\{J(\boldsymbol{w}_i)|\,\boldsymbol{w}_i\in\mathcal{G}\} - \mu^2\frac{c_2}{\pi} \tag{7.33}$$

On the other hand, when $\boldsymbol{w}_i\in\mathcal{M}$, *we can bound the expected ascent:*

$$\mathbb{E}\{J(\boldsymbol{w}_{i+1})|\,\boldsymbol{w}_i\in\mathcal{M}\} \le \mathbb{E}\{J(\boldsymbol{w}_i)|\,\boldsymbol{w}_i\in\mathcal{M}\} + \mu^2 c_2 \tag{7.34}$$

Proof. Appendix 7.B. □

Theorem 7.1 ensures that, whenever $\boldsymbol{w}_i\in\mathcal{G}$, i.e., whenever the gradient is sufficiently large, one can expect descent in one iteration. This descent relation is similar to those used to es-

tablish convergence to first-order stationary points [151]. In fact, repeatedly applying (7.33) would allow us to conclude that $\boldsymbol{w}_i$ must eventually reach $\mathcal{G}^C$ with high probability, as long as $J(\cdot)$ is bounded from below. In contrast to strongly convex optimization however, where a small gradient norm always implies vicinity to the global minimizer, first-order stationary points can be arbitrarily far from a local minimum in non-convex surfaces. For this reason, we will proceed to study the behavior around strict-saddle points in the sequel.

7.3.3 Escape from Saddle-Points

Beginning at a strict saddle-point $\boldsymbol{w}_i \in \mathcal{H}$ and for any $j \geq 0$, we have from (7.3):

$$\boldsymbol{w}_{i+j+1} = \boldsymbol{w}_{i+j} - \mu \nabla J(\boldsymbol{w}_{i+j}) - \mu\, \boldsymbol{s}_{i+j+1}(\boldsymbol{w}_{i+j}) \tag{7.35}$$

Subtracting this relation from $\boldsymbol{w}_i$, we find:

$$\boldsymbol{w}_i - \boldsymbol{w}_{i+j+1} = \boldsymbol{w}_i - \boldsymbol{w}_{i+j} + \mu \nabla J(\boldsymbol{w}_{i+j}) + \mu\, \boldsymbol{s}_{i+j+1}(\boldsymbol{w}_{i+j}) \tag{7.36}$$

We shall study the evolution of the deviation $\boldsymbol{w}_i - \boldsymbol{w}_{i+j+1}$ over several iterations $j \geq 0$. For brevity, we define:

$$\widetilde{\boldsymbol{w}}_{j+1}^{i} \triangleq \boldsymbol{w}_i - \boldsymbol{w}_{i+j+1} \tag{7.37}$$

so that (7.36) becomes:

$$\widetilde{\boldsymbol{w}}_{j+1}^{i} = \widetilde{\boldsymbol{w}}_{j}^{i} + \mu \nabla J(\boldsymbol{w}_{i+j}) + \mu\, \boldsymbol{s}_{i+j+1}(\boldsymbol{w}_{i+j}) \tag{7.38}$$

From the mean-value theorem we find [1]:

$$\nabla J(\boldsymbol{w}_{i+j}) - \nabla J(\boldsymbol{w}_i) = \boldsymbol{H}_{i+j}\,(\boldsymbol{w}_{i+j} - \boldsymbol{w}_i) \stackrel{(7.37)}{=} -\boldsymbol{H}_{i+j}\widetilde{\boldsymbol{w}}_{j}^{i} \tag{7.39}$$

where

$$\boldsymbol{H}_{i+j} \triangleq \int_0^1 \nabla^2 J\left((1-t)\,\boldsymbol{w}_{i+j} + t\,\boldsymbol{w}_i\right) dt \tag{7.40}$$

so that (7.38) can be reformulated to:

$$\widetilde{\boldsymbol{w}}_{j+1}^{i} = (I - \mu \boldsymbol{H}_{i+j})\, \widetilde{\boldsymbol{w}}_{j}^{i} + \mu \nabla J(\boldsymbol{w}_i) + \mu\, \boldsymbol{s}_{i+j+1}(\boldsymbol{w}_{i+j}) \tag{7.41}$$

In a manner similar to [1, 25, 59], we replace the random and time-varying matrix $\boldsymbol{H}_{i+j}$ by the Hessian matrix $\nabla^2 J(\boldsymbol{w}_i)$ evaluated at the starting point i. This substitution obviously leads to an approximate recursion in place of (7.41); we shall denote its state vector by $\widetilde{\boldsymbol{w}}_{j+1}^{\prime i}$ instead of $\widetilde{\boldsymbol{w}}_{j+1}^{i}$, as seen below in (7.42). The point is that while the Hessian $\nabla^2 J(\boldsymbol{w}_i)$ is random and depends on the time instance i, it becomes deterministic and constant when conditioning on $\boldsymbol{\mathcal{F}}_i$ and iterating over $j \geq 0$. We thus arrive at the following recursion, which we shall refer to as the *short-term* model:

$$\widetilde{\boldsymbol{w}}_{j+1}^{\prime i} = \big(I - \mu \nabla^2 J(\boldsymbol{w}_i)\big)\, \widetilde{\boldsymbol{w}}_{j}^{\prime i} + \mu \nabla J(\boldsymbol{w}_i) + \mu\, \boldsymbol{s}_{i+j+1}(\boldsymbol{w}_{i+j}) \tag{7.42}$$

where

$$\widetilde{\boldsymbol{w}}_{j+1}^{\prime i} \triangleq \boldsymbol{w}_i - \boldsymbol{w}_{i+j+1}^{\prime} \tag{7.43}$$

The fact that the driving matrix $I - \mu \nabla^2 J(\boldsymbol{w}_i)$ is constant for all $j \geq 0$ ensures that (7.42) is a more tractable recursion than (7.41). In order for this model to be useful, however, we need to ensure that the function $J(\boldsymbol{w}_{i+j}^{\prime})$ evaluated at the iterate of the short-term model carries sufficient information about the actual recursion of interest, i.e., $J(\boldsymbol{w}_{i+j})$. We begin by establishing a set of deviation bounds over a finite time horizon. These ensure that the iterates $\boldsymbol{w}_{i+j}^{\prime}$ and $\boldsymbol{w}_{i+j}$ remain close for a bounded number of iterations, which will allow us to relate $J(\boldsymbol{w}_{i+j}^{\prime})$ and $J(\boldsymbol{w}_{i+j})$ further below.

Lemma 7.3 (**Deviation bounds**). *The following quantities are conditionally bounded:*

$$\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_j^i\right\|^2 | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \leq O(\mu) \tag{7.44}$$

$$\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_j^i\right\|^3 | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \leq O(\mu^{3/2}) \tag{7.45}$$

$$\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_j^i\right\|^4 | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \leq O(\mu^2) \tag{7.46}$$

$$\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_j^i - \widetilde{\boldsymbol{w}}'^i_j\right\|^2 | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \leq O(\mu^2) \tag{7.47}$$

$$\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}'^i_j\right\|^2 | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \leq O(\mu) \tag{7.48}$$

for $j \leq \frac{T}{\mu}$, where T denotes an arbitrary constant that is independent of the step-size μ.

Proof. Appendix 7.C. □

These deviation bounds, along with the smoothness conditions on $J(\cdot)$ allow us to establish the following corollary.

Corollary 7.1 (**Short-term model accuracy**). *Beginning at $\boldsymbol{w}_i \in \mathcal{H}$, the short term model is accurate over a finite horizon $j \leq \frac{T}{\mu}$, i.e.,*

$$\mathbb{E}\left\{J(\boldsymbol{w}_{i+j}) | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \leq \mathbb{E}\left\{J(\boldsymbol{w}'_{i+j}) | \, \boldsymbol{w}_i \in \mathcal{H}\right\} + O(\mu^{3/2}) \tag{7.49}$$

for $j \leq \frac{T}{\mu}$, where T denotes an arbitrary constant that is independent of the step-size μ.

Proof. Appendix 7.D. □

We conclude that $J(\cdot)$ evaluated at the true iterate $\boldsymbol{w}_{i+j}$ is upper bounded by $J(\cdot)$ evaluated at the short-term model $\boldsymbol{w}'_{i+j}$ (up to an approximation error $O(\mu^{3/2})$ that will turn out to be negligible for small step-sizes), so long as both recursions are initialized at strict-saddle points $\boldsymbol{w}_i \in \mathcal{H}$.

Theorem 7.2 (**Descent through strict saddle-points**). *Beginning at a strict saddle-*

point $\boldsymbol{w}_i \in \mathcal{H}$ *and iterating for* i^s *iterations after* i *with*

$$i^s = \frac{\log\left(2M\frac{\sigma^2}{\sigma_\ell^2} + 1 + O(\mu)\right)}{\log(1+2\mu\tau)} \leq O\left(\frac{1}{\mu\tau}\right) \tag{7.50}$$

guarantees

$$\begin{aligned} &\mathbb{E}\left\{J(\boldsymbol{w}_{i+i^s})\middle|\, \boldsymbol{w}_i \in \mathcal{H}\right\} \\ &\leq \mathbb{E}\left\{J(\boldsymbol{w}_i)\middle|\, \boldsymbol{w}_i \in \mathcal{H}\right\} - \frac{\mu}{2}M\sigma^2 + o(\mu) \end{aligned} \tag{7.51}$$

Proof. Appendix 7.E. □

We conclude that when $\boldsymbol{w}_i$ reaches an approximately strict-saddle points in $\mathcal{H}$, where the gradient norm alone is no longer sufficient to guarantee descent in a single iteration, we can nevertheless guarantee descent after $O(1/\mu)$ iterations. Recall that Theorem 7.1 guarantees descent for points in $\mathcal{G}$. As such, Theorems 7.1 and 7.2 together guarantee (expected) descent whenever $\boldsymbol{w}_i \notin \mathcal{M}$ and, as long as $J(\cdot)$ is bounded from below, they ensure that $\boldsymbol{w}_i$ must eventually reach a point in $\mathcal{M}$. This argument is formalized in the final theorem.

Theorem 7.3. *Suppose* $J(w) \geq J^o$*. Then, for sufficiently small step-sizes* μ*, we have with probability* $1-\pi$*, that* $\boldsymbol{w}_{i^o} \in \mathcal{M}$*, i.e.,* $\|\nabla J(\boldsymbol{w}_{i^o})\|^2 \leq O(\mu)$ *and* $\lambda_{\min}\left(\nabla^2 J(\boldsymbol{w}_{i^o})\right) \geq -\tau$ *in at most* i^o *iterations, where*

$$i^o \leq \frac{(J(w_0) - J^o)}{\mu^2 c_2 \pi} i^s \tag{7.52}$$

and i^s *denotes the escape time from Theorem 7.2.*

Proof. Appendix 7.F. □

7.4 Simulation Results

In this section, we consider a simple example, arising from a single-hidden-layer neural network with a linear hidden layer and a logistic activation function leading into the output

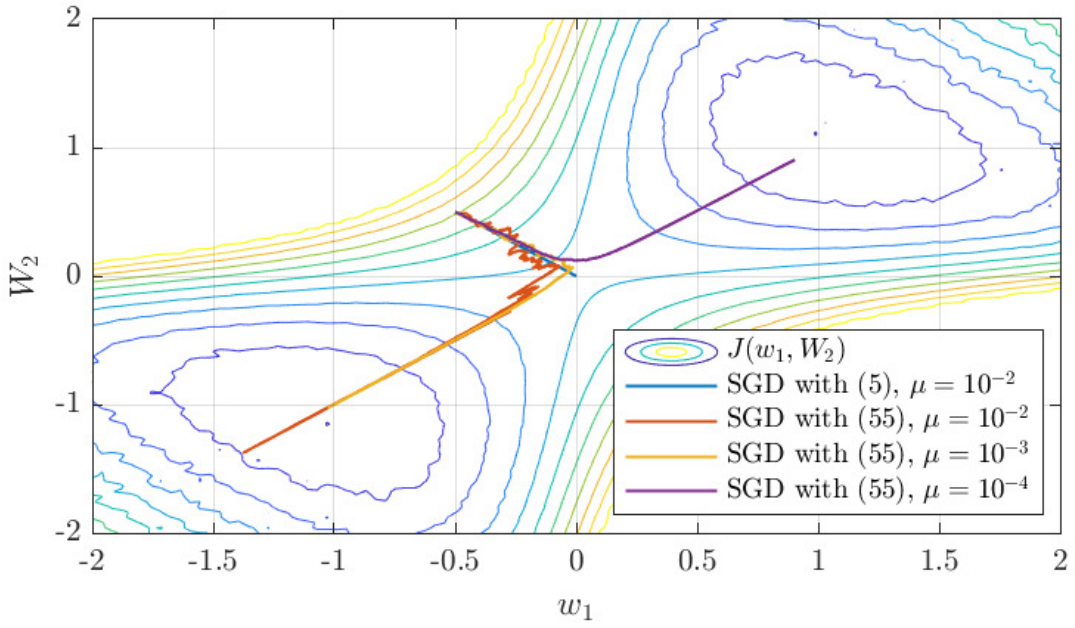

Figure 7.1: Cost surface of a simple neural network with $\rho = 0.1$ and sample trajectories. The symmetric nature of the loss and initialization result in an equal probability of escaping towards the local minimum in the positive or negative quadrant.

layer. The cross-entropy loss for such a structure can be simplified to an equivalent logistic loss [71]:

$$Q(w_1, W_2; \boldsymbol{\gamma}, \boldsymbol{h}) = \log\left(1 + e^{-\boldsymbol{\gamma} w_1^{\mathsf{T}} W_2 \boldsymbol{h}}\right) \tag{7.53}$$

The regularized learning problem can then be formulated as:

$$J(w_1, W_2) = \mathbb{E}\, Q(w_1, W_2; \boldsymbol{\gamma}, \boldsymbol{h}) + \frac{\rho}{2}\|w_1\|^2 + \frac{\rho}{2}\|W_2\|_F^2 \tag{7.54}$$

The cost surface is depicted in Fig. 7.1. The cost $J(\cdot)$ has two local minima in the positive and negative quadrants, respectively, and a single strict saddle-point at $w_1 = W_2 = 0$. We initialize $w_0 = \mathrm{col}\{-0.5, 0.5\}$ and compare the direct stochastic gradient descent implementation (7.5) with:

$$\widehat{\nabla J}(w_1, W_2) \triangleq \nabla Q(w_1, W_2; \boldsymbol{\gamma}, \boldsymbol{h}) + \boldsymbol{s} \cdot \mathrm{col}\{1, 1\} \tag{7.55}$$

where $\boldsymbol{s} \sim \mathcal{N}(0, 1)$ and the direction $\mathrm{col}\{1, 1\}$ corresponds to the local descent direction at the strict saddle-point $w_1 = W_2 = 0$. The particular choice of the direction is informed by the

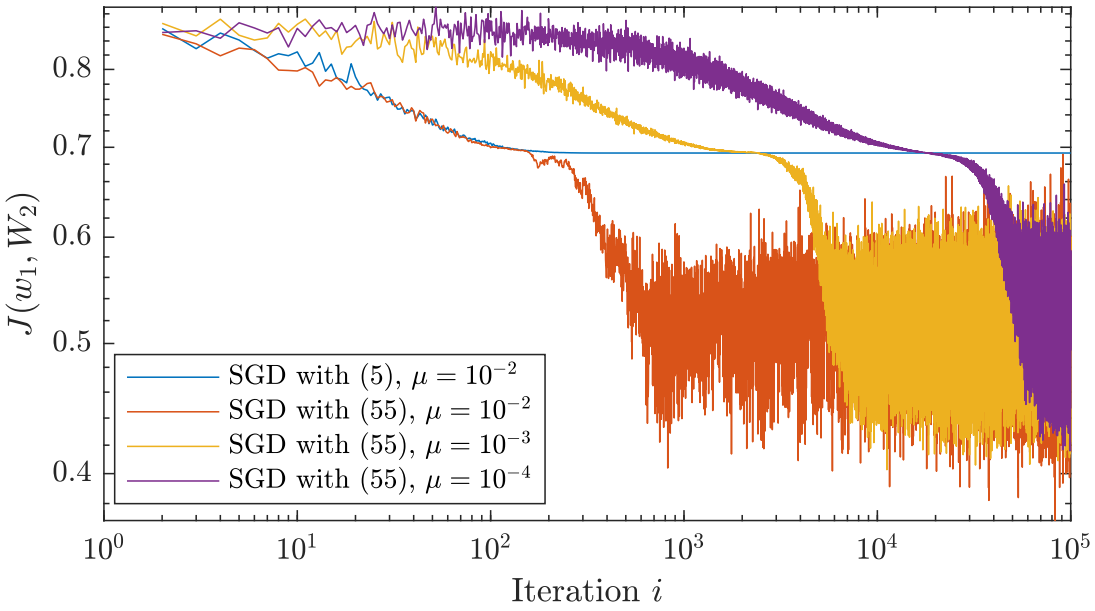

Figure 7.2: Evolution of the function value.

analysis and Assumption 7.5 and will allow us to verify whether condition (7.29) is indeed necessary. A realization of the learning curve is depicted in Fig. 7.2. It can be observed that the stochastic gradient recursion is outperformed by (7.55), since Assumption 7.5 is not satisfied for (7.5). Furthermore, it is evident that the escape time increases at a rate of $O(1/\mu)$ as μ decreases, suggesting the tightness of the escape time (7.50).

7.A Proof of Lemma 7.2

The proof techniques in these appendices are generally similar to the ones used in our works [70, 71] albeit after some necessary adjustments to account for the relative variance bound (7.17) and the adjusted relations in Definition 7.2.

To begin with, for the natural logarithm of the expression, we have:

$$\begin{aligned} &\log\left(\frac{(1+\mu\delta)^k + O(\mu^2)}{(1-\mu\delta)^{k-1}}\right)^{\frac{T}{\mu}} \\ &= \frac{T}{\mu}\left(\log\Big((1+\mu\delta)^k + O(\mu^2)\Big) - (k-1)\log(1-\mu\delta)\right) \end{aligned} \tag{7.56}$$

Since the logarithm is continuous over $\mathbb{R}_+$, we have:

$$\begin{aligned}
&\log\left(\lim_{\mu\to 0}\left(\frac{(1+\mu\delta)^k+O(\mu^2)}{(1-\mu\delta)^{k-1}}\right)^{\frac{T}{\mu}}\right)\\
&=\lim_{\mu\to 0}\log\left(\left(\frac{(1+\mu\delta)^k+O(\mu^2)}{(1-\mu\delta)^{k-1}}\right)^{\frac{T}{\mu}}\right)\\
&=\lim_{\mu\to 0}\frac{T}{\mu}\left(\log\left((1+\mu\delta)^k+O(\mu^2)\right)-(k-1)\log\left(1-\mu\delta\right)\right)\\
&=\lim_{\mu\to 0}\frac{T}{\mu}\left(\log\left((1+\mu\delta)^k\right)-(k-1)\log\left(1-\mu\delta\right)\right)\\
&=\lim_{\mu\to 0}\frac{T}{\mu}\left(k\log\left((1+\mu\delta)\right)-(k-1)\log\left(1-\mu\delta\right)\right)\\
&=kT\lim_{\mu\to 0}\frac{\log\left(1+\mu\delta\right)}{\mu}-(k-1)T\lim_{\mu\to 0}\frac{\log\left(1-\mu\delta\right)}{\mu}
\end{aligned}\tag{7.57}$$

We examine the fraction inside the limit more closely. Since both the numerator and denominator of the fraction approach zero as $\mu\to 0$, we apply L'Hôpital's rule:

$$\lim_{\mu\to 0}\frac{\log\left(1\pm\mu\delta\right)}{\mu}=\lim_{\mu\to 0}\frac{\pm\delta}{1\pm\mu\delta}=\pm\delta \tag{7.58}$$

Hence, we find:

$$\lim_{\mu\to 0}\left(\frac{(1+\mu\delta)^k+O(\mu^2)}{(1-\mu\delta)^{k-1}}\right)^{\frac{T}{\mu}}=e^{kT\delta+(k-1)T\delta}=e^{-T\delta+2kT\delta} \tag{7.59}$$

7.B Proof of Lemma 7.1

Since $J(\cdot)$ has δ-Lipschitz gradients:

$$J(\boldsymbol{w}_{i+1})\leq J(\boldsymbol{w}_i)+\nabla J(\boldsymbol{w}_i)^{\mathsf{T}}\left(\boldsymbol{w}_{i+1}-\boldsymbol{w}_i\right)+\frac{\delta}{2}\|\boldsymbol{w}_{i+1}-\boldsymbol{w}_i\|^2 \tag{7.60}$$

From (7.3), we find:

$$\begin{aligned}
&J(\boldsymbol{w}_{i+1}) \\
&\leq J(\boldsymbol{w}_i) + \nabla J(\boldsymbol{w}_i)^{\mathsf{T}} \left(-\widehat{\nabla J}(\boldsymbol{w}_i)\right) + \frac{\delta}{2}\left\|-\mu\widehat{\nabla J}(\boldsymbol{w}_i)\right\|^2 \\
&\leq J(\boldsymbol{w}_i) - \mu\nabla J(\boldsymbol{w}_i)^{\mathsf{T}}\nabla J(\boldsymbol{w}_i) - \mu\nabla J(\boldsymbol{w}_i)^{\mathsf{T}}\,\boldsymbol{s}_{i+1}(\boldsymbol{w}_i) \\
&\quad + \mu^2\frac{\delta}{2}\|\nabla J(\boldsymbol{w}_i) + \boldsymbol{s}_{i+1}(\boldsymbol{w}_i)\|^2
\end{aligned} \tag{7.61}$$

Under conditional expectation, we have:

$$\begin{aligned}
&\mathbb{E}\,\{J(\boldsymbol{w}_{i+1})|\boldsymbol{\mathcal{F}}_i\} \\
&\leq J(\boldsymbol{w}_i) - \mu\|\nabla J(\boldsymbol{w}_i)\|^2 - \mu\nabla J(\boldsymbol{w}_i)^{\mathsf{T}}\,\mathbb{E}\,\{\boldsymbol{s}_{i+1}(\boldsymbol{w}_i)|\boldsymbol{\mathcal{F}}_i\} \\
&\quad + \mu^2\frac{\delta}{2}\,\mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_i) + \boldsymbol{s}_{i+1}(\boldsymbol{w}_i)\|^2|\boldsymbol{\mathcal{F}}_i\right\} \\
&= J(\boldsymbol{w}_i) - \mu\left(1 - \mu\frac{\delta}{2}\right)\|\nabla J(\boldsymbol{w}_i)\|^2 \\
&\quad + \mu^2\frac{\delta}{2}\,\mathbb{E}\left\{\|\boldsymbol{s}_{i+1}(\boldsymbol{w}_i)\|^2|\boldsymbol{\mathcal{F}}_i\right\} \\
&\leq J(\boldsymbol{w}_i) - \mu\left(1 - \mu\frac{\delta}{2}\left(1+\beta^2\right)\right)\|\nabla J(\boldsymbol{w}_i)\|^2 + \mu^2\frac{\delta}{2}\sigma^2 \\
&\overset{(a)}{=} J(\boldsymbol{w}_i) - \mu c_1\|\nabla J(\boldsymbol{w}_i)\|^2 + \mu^2 c_2
\end{aligned} \tag{7.62}$$

where (a) follows from (7.26)–(7.27). Taking expectations conditioned on $\boldsymbol{w}_i \in \mathcal{G}$, we find:

$$\begin{aligned}
&\mathbb{E}\,\{J(\boldsymbol{w}_{i+1})|\,\boldsymbol{w}_i \in \mathcal{G}\} \\
&\leq\ \mathbb{E}\,\{J(\boldsymbol{w}_i)|\,\boldsymbol{w}_i \in \mathcal{G}\} - \mu c_1\,\mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_i)\|^2|\,\boldsymbol{w}_i \in \mathcal{G}\right\} + \mu^2 c_2 \\
&\leq\ \mathbb{E}\,\{J(\boldsymbol{w}_i)|\,\boldsymbol{w}_i \in \mathcal{G}\} - \mu c_1 \cdot \mu\frac{c_2}{c_1}\left(1 + \frac{1}{\pi}\right) + \mu^2 c_2 \\
&=\ \mathbb{E}\,\{J(\boldsymbol{w}_i)|\,\boldsymbol{w}_i \in \mathcal{G}\} - \mu^2\frac{c_2}{\pi}
\end{aligned} \tag{7.63}$$

On the other hand, starting from (7.62) and taking expectations conditioned on $\boldsymbol{w}_i \in \mathcal{M}$, we have:

$$\begin{aligned}
&\mathbb{E}\left\{J(\boldsymbol{w}_{i+1})|\,\boldsymbol{w}_i \in \mathcal{M}\right\} \\
&\leq \mathbb{E}\left\{J(\boldsymbol{w}_i)|\,\boldsymbol{w}_i \in \mathcal{M}\right\} - \mu c_1\,\mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_i)\|^2|\,\boldsymbol{w}_i \in \mathcal{M}\right\} + \mu^2 c_2 \\
&\overset{(a)}{\leq} \mathbb{E}\left\{J(\boldsymbol{w}_i)|\,\boldsymbol{w}_i \in \mathcal{M}\right\} + \mu^2 c_2
\end{aligned} \tag{7.64}$$

where (a) follows since $c_1 = 1 - \mu\frac{\delta}{2}\,(1+\beta^2) \geq 0$ whenever $\mu \leq \frac{2}{\delta(1+\beta^2)}$.

7.C Proof of Lemma 7.3

We refer to (7.41). Suppose $j \leq \frac{T}{\mu}$, where T is an arbitrary constant independent of μ. We then have for $j \geq 0$:

$$
\begin{aligned}
&\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_{j+1}^{i}\right\|^{2}|\boldsymbol{\mathcal{F}}_{i+j}\right\} \\
&\overset{(7.41)}{=} \mathbb{E}\left\{\left\|\left(I-\mu\boldsymbol{H}_{i+j}\right)\widetilde{\boldsymbol{w}}_{j}^{i}+\mu\nabla J(\boldsymbol{w}_{i})+\mu\,\boldsymbol{s}_{i+j+1}\right\|^{2}|\boldsymbol{\mathcal{F}}_{i+j}\right\} \\
&\overset{(a)}{=} \left\|\left(I-\mu\boldsymbol{H}_{i+j}\right)\widetilde{\boldsymbol{w}}_{j}^{i}+\mu\nabla J(\boldsymbol{w}_{i})\right\|^{2} \\
&\quad+\mu^{2}\,\mathbb{E}\left\{\|\boldsymbol{s}_{i+j+1}\|^{2}|\boldsymbol{\mathcal{F}}_{i+j}\right\} \\
&\overset{(b)}{=} \frac{1}{1-\mu\delta}\left\|\left(I-\mu\boldsymbol{H}_{i+j}\right)\widetilde{\boldsymbol{w}}_{j}^{i}\right\|^{2}+\frac{\mu}{\delta}\|\nabla J(\boldsymbol{w}_{i})\|^{2} \\
&\quad+\mu^{2}\,\mathbb{E}\left\{\|\boldsymbol{s}_{i+j+1}\|^{2}|\boldsymbol{\mathcal{F}}_{i+j}\right\} \\
&\overset{(c)}{\leq} \frac{(1+\mu\delta)^{2}}{1-\mu\delta}\left\|\widetilde{\boldsymbol{w}}_{j}^{i}\right\|^{2}+\frac{\mu}{\delta}\|\nabla J(\boldsymbol{w}_{i})\|^{2} \\
&\quad+\mu^{2}\,\mathbb{E}\left\{\|\boldsymbol{s}_{i+j+1}\|^{2}|\boldsymbol{\mathcal{F}}_{i+j}\right\} \\
&\overset{(d)}{\leq} \frac{(1+\mu\delta)^{2}}{1-\mu\delta}\left\|\widetilde{\boldsymbol{w}}_{j}^{i}\right\|^{2}+\frac{\mu}{\delta}\|\nabla J(\boldsymbol{w}_{i})\|^{2} \\
&\quad+\mu^{2}\beta^{2}\|\nabla J(\boldsymbol{w}_{i+j})\|^{2}+\mu^{2}\sigma^{2} \\
&= \frac{(1+\mu\delta)^{2}}{1-\mu\delta}\left\|\widetilde{\boldsymbol{w}}_{j}^{i}\right\|^{2}+\frac{\mu}{\delta}\|\nabla J(\boldsymbol{w}_{i})\|^{2} \\
&\quad+\mu^{2}\beta^{2}\|\nabla J(\boldsymbol{w}_{i})+\nabla J(\boldsymbol{w}_{i+j})-\nabla J(\boldsymbol{w}_{i})\|^{2}+\mu^{2}\sigma^{2} \\
&\overset{(e)}{\leq} \frac{(1+\mu\delta)^{2}}{1-\mu\delta}\left\|\widetilde{\boldsymbol{w}}_{j}^{i}\right\|^{2}+\frac{\mu}{\delta}\|\nabla J(\boldsymbol{w}_{i})\|^{2}+2\mu^{2}\beta^{2}\|\nabla J(\boldsymbol{w}_{i})\|^{2} \\
&\quad+2\mu^{2}\beta^{2}\|\nabla J(\boldsymbol{w}_{i+j})-\nabla J(\boldsymbol{w}_{i})\|^{2}+\mu^{2}\sigma^{2} \\
&\overset{(f)}{\leq} \frac{(1+\mu\delta)^{2}+(1-\mu\delta)2\mu^{2}\beta^{2}\delta^{2}}{1-\mu\delta}\left\|\widetilde{\boldsymbol{w}}_{j}^{i}\right\|^{2} \\
&\quad+\mu\left(\frac{1}{\delta}+2\mu\beta^{2}\right)\|\nabla J(\boldsymbol{w}_{i})\|^{2}+\mu^{2}\sigma^{2} \\
&\overset{(g)}{\leq} \frac{(1+\mu\delta)^{2}+O(\mu^{2})}{1-\mu\delta}\left\|\widetilde{\boldsymbol{w}}_{j}^{i}\right\|^{2}+O(\mu)\|\nabla J(\boldsymbol{w}_{i})\|^{2}+\mu^{2}\sigma^{2} \qquad (7.65)
\end{aligned}
$$

where (a) follows from the conditional zero-mean property of the gradient noise term in Assumption 7.3, (b) follows from Jensen's inequality

$$\|a+b\|^2 \leq \frac{1}{\alpha}\|a\|^2 + \frac{1}{1-\alpha}\|b\|^2 \tag{7.66}$$

with $\alpha = \mu\delta < 1$ and (c) follows from the sub-multiplicative property of norms along with $-\delta I \leq \nabla^2 J(\boldsymbol{w}_i) \leq \delta I$, which follows from the Lipschitz gradient condition in Assumption 7.1. We can now take expectations over $\boldsymbol{w}_i \in \mathcal{H}$ to obtain:

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_{j+1}^{i}\right\|^2 | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \\
&\leq \frac{(1+\mu\delta)^2 + O(\mu^2)}{1-\mu\delta}\,\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_{j}^{i}\right\|^2 | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \\
&\quad + O(\mu)\,\mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_i)\|^2 | \, \boldsymbol{w}_i \in \mathcal{H}\right\} + O(\mu^2) \\
&\overset{(a)}{\leq} \frac{(1+\mu\delta)^2 + O(\mu^2)}{1-\mu\delta}\,\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_{j}^{i}\right\|^2 | \, \boldsymbol{w}_i \in \mathcal{H}\right\} + O(\mu^2)
\end{aligned} \tag{7.67}$$

where (a) follows from the definition of the set $\mathcal{H}$ (7.24). Note that, at time $i = 0$, we have:

$$\widetilde{\boldsymbol{w}}_0^i = \boldsymbol{w}_i - \boldsymbol{w}_{i+0} = 0 \tag{7.68}$$

and hence the initial deviation is zero, by definition. Iterating, starting at $j = 0$ yields:

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_j^i\right\|^2 | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \\
&\leq \left(\sum_{n=0}^{j-1}\left(\frac{(1+\mu\delta)^2 + O(\mu^2)}{1-\mu\delta}\right)^n\right) O(\mu^2) \\
&= \frac{1 - \left(\frac{(1+\mu\delta)^2+O(\mu^2)}{1-\mu\delta}\right)^j}{1 - \frac{(1+\mu\delta)^2+O(\mu^2)}{1-\mu\delta}} O(\mu^2) \\
&= \frac{\left(\left(\frac{(1+\mu\delta)^2+O(\mu^2)}{1-\mu\delta}\right)^j - 1\right)(1-\mu\delta)}{1 + 2\mu\delta + \mu^2\delta^2 - 1 + \mu\delta} O(\mu^2) \\
&= \frac{\left(\left(\frac{(1+\mu\delta)^2+O(\mu^2)}{1-\mu\delta}\right)^j - 1\right)(1-\mu\delta)}{3\delta + \mu\delta^2} O(\mu) \\
&\leq \frac{\left(\left(\frac{(1+\mu\delta)^2+O(\mu^2)}{1-\mu\delta}\right)^{\frac{T}{\mu}} - 1\right)(1-\mu\delta)}{3\delta + \mu\delta^2} O(\mu) \\
&= O(\mu)
\end{aligned} \tag{7.69}$$

where the last line follows from Lemma 7.2 after noting that:

$$\begin{aligned}
&\frac{\left(\left(\frac{(1+\mu\delta)^2+O(\mu^2)}{1-\mu\delta}\right)^{\frac{T}{\mu}} - 1\right)(1-\mu\delta)}{3\delta + \mu\delta^2} \\
&\leq \frac{\left(\left(\frac{(1+\mu\delta)^2+O(\mu^2)}{1-\mu\delta}\right)^{\frac{T}{\mu}} - 1\right)(1-\mu\delta)}{3\delta} \\
&\leq \frac{\left(\frac{(1+\mu\delta)^2+O(\mu^2)}{1-\mu\delta}\right)^{\frac{T}{\mu}}}{3\delta}
\end{aligned} \tag{7.70}$$

This establishes (7.44). We proceed to establish a bound on the fourth-order moment. Using the inequality [1]:

$$\|a+b\|^4 \leq \|a\|^4 + 3\|b\|^4 + 8\|a\|^2\|b\|^2 + 4\|a\|^2\left(a^{\mathsf{T}}b\right) \tag{7.71}$$

we have:

$$
\begin{aligned}
&\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_{j+1}^{i}\right\|^{4}|\boldsymbol{\mathcal{F}}_{i+j}\right\} \\
&\leq\left\|\left(I-\mu \boldsymbol{H}_{i+j}\right) \widetilde{\boldsymbol{w}}_{j}^{i}+\mu \nabla J\left(\boldsymbol{w}_{i}\right)\right\|^{4} \\
&\quad+3 \mu^{4} \mathbb{E}\left\{\left\|\boldsymbol{s}_{i+j+1}\right\|^{4}\left|\boldsymbol{\mathcal{F}}_{i+j}\right.\right\} \\
&\quad+8 \mu^{2}\left\|\left(I-\mu \boldsymbol{H}_{i+j}\right) \widetilde{\boldsymbol{w}}_{j}^{i}+\mu \nabla J\left(\boldsymbol{w}_{i}\right)\right\|^{2} \\
&\qquad \times \mathbb{E}\left\{\left\|\boldsymbol{s}_{i+j+1}\right\|^{2}\left|\boldsymbol{\mathcal{F}}_{i+j}\right.\right\} \\
&\quad+4 \mu\left\|\left(I-\mu \boldsymbol{H}_{i+j}\right) \widetilde{\boldsymbol{w}}_{j}^{i}+\mu \nabla J\left(\boldsymbol{w}_{i}\right)\right\|^{2} \\
&\qquad \times\left(\left(I-\mu \boldsymbol{H}_{i+j}\right) \widetilde{\boldsymbol{w}}_{j}^{i}+\mu \nabla J\left(\boldsymbol{w}_{i}\right)\right)^{\mathsf{T}} \\
&\qquad\quad \times\left(\mathbb{E}\left\{\boldsymbol{s}_{i+j+1}\left|\boldsymbol{\mathcal{F}}_{i+1}\right.\right\}\right) \\
&\overset{(a)}{=}\left\|\left(I-\mu \boldsymbol{H}_{i+j}\right) \widetilde{\boldsymbol{w}}_{j}^{i}+\mu \nabla J\left(\boldsymbol{w}_{i}\right)\right\|^{4} \\
&\quad+3 \mu^{4} \mathbb{E}\left\{\left\|\boldsymbol{s}_{i+j+1}\right\|^{4}\left|\boldsymbol{\mathcal{F}}_{i+j}\right.\right\} \\
&\quad+8 \mu^{2}\left\|\left(I-\mu \boldsymbol{H}_{i+j}\right) \widetilde{\boldsymbol{w}}_{j}^{i}+\mu \nabla J\left(\boldsymbol{w}_{i}\right)\right\|^{2} \\
&\qquad \times \mathbb{E}\left\{\left\|\boldsymbol{s}_{i+j+1}\right\|^{2}\left|\boldsymbol{\mathcal{F}}_{i+j}\right.\right\} \\
&\overset{(b)}{\leq}\left\|\left(I-\mu \boldsymbol{H}_{i+j}\right) \widetilde{\boldsymbol{w}}_{j}^{i}+\mu \nabla J\left(\boldsymbol{w}_{i}\right)\right\|^{4} \\
&\quad+3 \mu^{4}\left(\left\|\nabla J\left(\boldsymbol{w}_{i+j}\right)\right\|^{4}+\sigma^{4}\right) \\
&\quad+8 \mu^{2}\left\|\left(I-\mu \boldsymbol{H}_{i+j}\right) \widetilde{\boldsymbol{w}}_{j}^{i}+\mu \nabla J\left(\boldsymbol{w}_{i}\right)\right\|^{2} \\
&\qquad \times\left(\left\|\nabla J\left(\boldsymbol{w}_{i+j}\right)\right\|^{2}+\sigma^{2}\right)
\end{aligned}
\tag{7.72}
$$

where in step (a) we dropped cross-terms due to the conditional zero-mean property of the gradient noise in Assumption 7.3, step (b) follows from the fourth-order conditions on the gradient noise in Assumption 7.3. We shall bound each term one by one. From Jensen's inequality, we find for $0<\alpha<1$:

$$
\|a+b\|^{4}=\frac{1}{\alpha^{3}}\|a\|^{4}+\frac{1}{(1-\alpha)^{3}}\|b\|^{4} \tag{7.73}
$$

and hence for the first term on the right-hand side of (7.72) with $\alpha = 1-\mu\delta$ and $0 < \mu < \frac{1}{\delta}$:

$$\begin{aligned}
&\left\|(I-\mu\boldsymbol{H}_{i+j})\,\widetilde{\boldsymbol{w}}_j^i + \mu\nabla J(\boldsymbol{w}_i)\right\|^4 \\
&\leq \frac{(1+\mu\delta)^4}{(1-\mu\delta)^3}\left\|\widetilde{\boldsymbol{w}}_j^i\right\|^4 + \frac{\mu^4}{\mu^3\delta^3}\|\nabla J(\boldsymbol{w}_i)\|^4 \\
&= \frac{(1+\mu\delta)^4}{(1-\mu\delta)^3}\left\|\widetilde{\boldsymbol{w}}_j^i\right\|^4 + O(\mu)\|\nabla J(\boldsymbol{w}_i)\|^4
\end{aligned} \tag{7.74}$$

After taking expectations conditioned on $\boldsymbol{w}_i \in \mathcal{H}$, we find:

$$\begin{aligned}
&\mathbb{E}\left\{\left\|(I-\mu\boldsymbol{H}_{i+j})\,\widetilde{\boldsymbol{w}}_j^i + \mu\nabla J(\boldsymbol{w}_i)\right\|^4 |\,\boldsymbol{w}_i \in \mathcal{H}\right\} \\
&\leq \frac{(1+\mu\delta)^4}{(1-\mu\delta)^3}\,\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_j^i\right\|^4 |\,\boldsymbol{w}_i \in \mathcal{H}\right\} \\
&\quad + O(\mu)\,\mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_i)\|^4 |\,\boldsymbol{w}_i \in \mathcal{H}\right\} \\
&\overset{(7.24)}{\leq} \frac{(1+\mu\delta)^4}{(1-\mu\delta)^3}\,\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_j^i\right\|^4 |\,\boldsymbol{w}_i \in \mathcal{H}\right\} + O(\mu^3)
\end{aligned} \tag{7.75}$$

For the second term we have, again from (7.73) with $\alpha = \frac{1}{2}$:

$$\begin{aligned}
&3\mu^4\left(\|\nabla J(\boldsymbol{w}_{i+j})\|^4 + \sigma^4\right) \\
&= 3\mu^4\left(\|\nabla J(\boldsymbol{w}_i) + \nabla J(\boldsymbol{w}_{i+j}) - \nabla J(\boldsymbol{w}_i)\|^4 + \sigma^4\right) \\
&\overset{(7.73)}{\leq} 3\mu^4\left(8\|\nabla J(\boldsymbol{w}_i)\|^4 + 8\|\nabla J(\boldsymbol{w}_{i+j}) - \nabla J(\boldsymbol{w}_i)\|^4 + \sigma^4\right) \\
&\overset{(7.73)}{\leq} 3\mu^4\left(8\|\nabla J(\boldsymbol{w}_i)\|^4 + 8\delta^4\left\|\widetilde{\boldsymbol{w}}_j^i\right\|^4 + \sigma^4\right) \\
&= O(\mu^4)\|\nabla J(\boldsymbol{w}_i)\|^4 + O(\mu^4)\left\|\widetilde{\boldsymbol{w}}_j^i\right\|^4 + O(\mu^4)
\end{aligned} \tag{7.76}$$

After taking expectations over $\boldsymbol{w}_i \in \mathcal{H}$ we have:

$$\begin{aligned}
&\mathbb{E}\left\{3\mu^4\left(\|\nabla J(\boldsymbol{w}_{i+j})\|^4+\sigma^4\right)\,|\,\boldsymbol{w}_i\in\mathcal{H}\right\}\\
&\leq O(\mu^4)\,\mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_i)\|^4|\,\boldsymbol{w}_i\in\mathcal{H}\right\}\\
&\quad+O(\mu^4)\,\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_j^i\right\|^4|\,\boldsymbol{w}_i\in\mathcal{H}\right\}+O(\mu^4)\\
&\leq O(\mu^4)\,\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_j^i\right\|^4|\,\boldsymbol{w}_i\in\mathcal{H}\right\}+O(\mu^4)
\end{aligned}\tag{7.77}$$

For the last term, we have:

$$
\begin{aligned}
&8\mu^2\big\|(I-\mu\boldsymbol{H}_{i+j})\,\widetilde{\boldsymbol{w}}_j^i+\mu\nabla J(\boldsymbol{w}_i)\big\|^2\left(\|\nabla J(\boldsymbol{w}_{i+j})\|^2+\sigma^2\right)\\
&=8\mu^2\big\|(I-\mu\boldsymbol{H}_{i+j})\,\widetilde{\boldsymbol{w}}_j^i+\mu\nabla J(\boldsymbol{w}_i)\big\|^2\|\nabla J(\boldsymbol{w}_{i+j})\|^2\\
&\quad+8\mu^2\big\|(I-\mu\boldsymbol{H}_{i+j})\,\widetilde{\boldsymbol{w}}_j^i+\mu\nabla J(\boldsymbol{w}_i)\big\|^2\sigma^2\\
&\overset{(7.66)}{\leq}8\mu^2\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^2+\frac{\mu}{\delta}\|\nabla J(\boldsymbol{w}_i)\|^2\right)\|\nabla J(\boldsymbol{w}_{i+j})\|^2\\
&\quad+8\mu^2\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^2+\frac{\mu}{\delta}\|\nabla J(\boldsymbol{w}_i)\|^2\right)\sigma^2\\
&=8\mu^2\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^2+\frac{\mu}{\delta}\|\nabla J(\boldsymbol{w}_i)\|^2\right)\\
&\quad\times\|\nabla J(\boldsymbol{w}_i)+\nabla J(\boldsymbol{w}_{i+j})-\nabla J(\boldsymbol{w}_i)\|^2\\
&\quad+8\mu^2\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^2+\frac{\mu}{\delta}\|\nabla J(\boldsymbol{w}_i)\|^2\right)\sigma^2\\
&\overset{(7.66)}{\leq}8\mu^2\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^2+\frac{\mu}{\delta}\|\nabla J(\boldsymbol{w}_i)\|^2\right)\\
&\quad\times\left(2\|\nabla J(\boldsymbol{w}_i)\|^2+2\|\nabla J(\boldsymbol{w}_{i+j})-\nabla J(\boldsymbol{w}_i)\|^2\right)\\
&\quad+8\mu^2\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^2+\frac{\mu}{\delta}\|\nabla J(\boldsymbol{w}_i)\|^2\right)\sigma^2\\
&\overset{(7.66)}{\leq}8\mu^2\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^2+\frac{\mu}{\delta}\|\nabla J(\boldsymbol{w}_i)\|^2\right)\\
&\quad\times\left(2\|\nabla J(\boldsymbol{w}_i)\|^2+2\delta^2\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^2\right)\\
&\quad+8\mu^2\left(\frac{(1+\mu\delta)^2}{1-\mu\delta}\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^2+\frac{\mu}{\delta}\|\nabla J(\boldsymbol{w}_i)\|^2\right)\sigma^2\\
&=O(\mu^2)\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^4+O(\mu^3)\|\nabla J(\boldsymbol{w}_i)\|^4\\
&\quad+O(\mu^2)\|\nabla J(\boldsymbol{w}_i)\|^2\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^2+O(\mu^2)\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^2\\
&\quad+O(\mu^3)\|\nabla J(\boldsymbol{w}_i)\|^2
\end{aligned}
\tag{7.78}
$$

After taking conditional expectations:

$$
\begin{aligned}
&\mathbb{E}\left\{8\mu^2\big\|(I-\mu\boldsymbol{H}_{i+j})\,\widetilde{\boldsymbol{w}}_j^i+\mu\nabla J(\boldsymbol{w}_i)\big\|^2\right.\\
&\qquad\left.\times\left(\|\nabla J(\boldsymbol{w}_{i+j})\|^2+\sigma^2\right)\,|\,\boldsymbol{w}_i\in\mathcal{H}\right\}\\
&\le O(\mu^2)\,\mathbb{E}\left\{\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^4|\,\boldsymbol{w}_i\in\mathcal{H}\right\}\\
&\quad+O(\mu^3)\,\mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_i)\|^4|\,\boldsymbol{w}_i\in\mathcal{H}\right\}\\
&\quad+O(\mu^2)\,\mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_i)\|^2\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^2|\,\boldsymbol{w}_i\in\mathcal{H}\right\}\\
&\quad+O(\mu^2)\,\mathbb{E}\left\{\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^2|\,\boldsymbol{w}_i\in\mathcal{H}\right\}\\
&\quad+O(\mu^3)\,\mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_i)\|^2|\,\boldsymbol{w}_i\in\mathcal{H}\right\}\\
&\le O(\mu^2)\,\mathbb{E}\left\{\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^4|\,\boldsymbol{w}_i\in\mathcal{H}\right\}+O(\mu^3)\cdot O(\mu^2)\\
&\quad+O(\mu^2)\,\mathbb{E}\left\{O(\mu)\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^2|\,\boldsymbol{w}_i\in\mathcal{H}\right\}+O(\mu^2)\cdot O(\mu)\\
&\quad+O(\mu^3)\cdot O(\mu)\\
&\le O(\mu^2)\,\mathbb{E}\left\{\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^4|\,\boldsymbol{w}_i\in\mathcal{H}\right\}+O(\mu^3)
\end{aligned}
\tag{7.79}
$$

Returning to (7.72), after taking expectations over $\boldsymbol{w}_i\in\mathcal{H}$ on both sides and grouping terms we find:

$$
\begin{aligned}
&\mathbb{E}\left\{\big\|\widetilde{\boldsymbol{w}}_{j+1}^i\big\|^4|\,\boldsymbol{w}_i\in\mathcal{H}\right\}\\
&\le\frac{(1+\mu\delta)^4+O(\mu^2)}{(1-\mu\delta)^3}\,\mathbb{E}\left\{\big\|\widetilde{\boldsymbol{w}}_j^i\big\|^4|\,\boldsymbol{w}_i\in\mathcal{H}\right\}+O(\mu^3)
\end{aligned}
\tag{7.80}
$$

Recall again that $\widetilde{\boldsymbol{w}}_0^i = 0$ and therefore iterating yields:

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_j^i\right\|^4 | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \\
&\leq \left(\sum_{n=0}^{j-1}\left(\frac{(1+\mu\delta)^4 + O(\mu^2)}{(1-\mu\delta)^3}\right)^n\right) O(\mu^3) \\
&= \frac{1 - \left(\frac{(1+\mu\delta)^4+O(\mu^2)}{(1-\mu\delta)^3}\right)^j}{1 - \frac{(1+\mu\delta)^4+O(\mu^2)}{(1-\mu\delta)^3}} O(\mu^3) \\
&= \frac{\left(\left(\frac{(1+\mu\delta)^4+O(\mu^2)}{(1-\mu\delta)^3}\right)^j - 1\right)(1-\mu\delta)^3}{(1+\mu\delta)^4 + O(\mu^2) - (1-\mu\delta)^3} O(\mu^3) \\
&\leq \frac{\left(\frac{(1+\mu\delta)^4+O(\mu^2)}{(1-\mu\delta)^3}\right)^j - 1}{(1+\mu\delta)^4 + O(\mu^2) - (1-\mu\delta)^3} O(\mu^3) \\
&\leq \frac{\left(\frac{(1+\mu\delta)^4+O(\mu^2)}{(1-\mu\delta)^3}\right)^j}{(1+\mu\delta)^4 + O(\mu^2) - (1-\mu\delta)^3} O(\mu^3) \\
&\overset{(a)}{\leq} \frac{\left(\frac{(1+\mu\delta)^4+O(\mu^2)}{(1-\mu\delta)^3}\right)^j}{O(\mu)} O(\mu^3) \\
&= \left(\frac{(1+\mu\delta)^4 + O(\mu^2)}{(1-\mu\delta)^3}\right)^j O(\mu^2) \\
&\leq \left(\frac{(1+\mu\delta)^4 + O(\mu^2)}{(1-\mu\delta)^3}\right)^{\frac{T}{\mu}} O(\mu^2) \\
&\leq O(\mu^2)
\end{aligned} \tag{7.81}$$

where in (a) we expanded:

$$\begin{aligned}
&(1+\mu\delta)^4 + O(\mu^2) - (1-\mu\delta)^3 \\
&= 1 + 4\mu\delta + O(\mu^2) - 1 + 3\mu\delta - O(\mu^2) = O(\mu)
\end{aligned} \tag{7.82}$$

and the last step follows from Lemma 7.2. This establishes (7.46). Eq. (7.45) then follows from Jensen's inequality via:

$$\begin{aligned}&\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_j^i\right\|^3 | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \leq \left(\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_j^i\right\|^4 | \, \boldsymbol{w}_i \in \mathcal{H}\right\}\right)^{3/4} \\ &\leq \big(O(\mu^2)\big)^{3/4} = O(\mu^{3/2}) \end{aligned} \tag{7.83}$$

We now study the difference between the short-term model (7.42) and the true recursion (7.41). We have:

$$\begin{aligned}&\boldsymbol{w}_{i+j+1} - \boldsymbol{w}'_{i+j+1} \\ &= -\widetilde{\boldsymbol{w}}_{i+1}^i + \widetilde{\boldsymbol{w}}'^i_{i+1} \\ &= -(I - \mu \boldsymbol{H}_{i+i})\,\widetilde{\boldsymbol{w}}_j^i - \mu \nabla J(\boldsymbol{w}_i) - \mu\, \boldsymbol{s}_{i+j+1} \\ &\quad + \big(I - \mu \nabla^2 J(\boldsymbol{w}_i)\big)\,\widetilde{\boldsymbol{w}}'^i_i + \mu \nabla J(\boldsymbol{w}_i) + \mu\, \boldsymbol{s}_{i+j+1} \\ &= -(I - \mu \boldsymbol{H}_{i+i})\,\widetilde{\boldsymbol{w}}_j^i + \big(I - \mu \nabla^2 J(\boldsymbol{w}_i)\big)\,\widetilde{\boldsymbol{w}}'^i_i \\ &= \big(I - \mu \nabla^2 J(\boldsymbol{w}_i)\big)\,\big(\boldsymbol{w}_{i+j} - \boldsymbol{w}'_{i+j}\big) \\ &\quad + \mu\,\big(\boldsymbol{H}_{i+j} - \nabla^2 J(\boldsymbol{w}_i)\big)\,\widetilde{\boldsymbol{w}}_j^i \end{aligned} \tag{7.84}$$

Before proceeding, note that the difference between the Hessians in the driving term can be bounded as:

$$\begin{aligned}&\big\|\nabla^2 J(\boldsymbol{w}_i) - \boldsymbol{H}_{i+i}\big\| \\ &= \left\|\nabla^2 J(\boldsymbol{w}_i) - \int_0^1 \nabla^2 J\left((1-t)\,\boldsymbol{w}_{i+j} + t\,\boldsymbol{w}_i\right) dt\right\| \\ &= \left\|\int_0^1 \left(\nabla^2 J(\boldsymbol{w}_i) - \nabla^2 J\left((1-t)\,\boldsymbol{w}_{i+j} + t\,\boldsymbol{w}_i\right)\right) dt\right\| \\ &\overset{(a)}{\leq} \int_0^1 \big\|\nabla^2 J(\boldsymbol{w}_i) - \nabla^2 J\left((1-t)\,\boldsymbol{w}_{i+j} + t\,\boldsymbol{w}_i\right)\big\|\, dt \\ &\overset{(b)}{\leq} \rho \int_0^1 \|(1-t)\,\boldsymbol{w}_i - (1-t)\,\boldsymbol{w}_{i+j}\|\, dt \\ &= \rho\,\big\|\widetilde{\boldsymbol{w}}_j^i\big\| \int_0^1 (1-t) dt = \frac{\rho}{2}\,\big\|\widetilde{\boldsymbol{w}}_j^i\big\| \end{aligned} \tag{7.85}$$

where (a) follows Jensen's inequality and (b) follows form the Lipschitz Hessian assumption 7.2. Returning to (7.84) and taking norms yields:

$$\begin{aligned}
&\| \boldsymbol{w}_{i+j+1} - \boldsymbol{w}'_{i+j+1} \|^2 \\
&= \Big\| \big(I - \mu\nabla^2 J(\boldsymbol{w}_i)\big) \big(\boldsymbol{w}_{i+j} - \boldsymbol{w}'_{i+j}\big) \\
&\quad + \mu \big(\boldsymbol{H}_{i+j} - \nabla^2 J(\boldsymbol{w}_i)\big) \widetilde{\boldsymbol{w}}^i_j \Big\|^2 \\
&\overset{(a)}{\leq} \frac{1}{1-\mu\delta} \big\| \big(I - \mu\nabla^2 J(\boldsymbol{w}_i)\big) \big(\boldsymbol{w}_{i+j} - \boldsymbol{w}'_{i+j}\big) \big\|^2 \\
&\quad + \frac{\mu^2}{\mu\delta} \big\| \big(\boldsymbol{H}_{i+j} - \nabla^2 J(\boldsymbol{w}_i)\big) \widetilde{\boldsymbol{w}}^i_j \big\|^2 \\
&\overset{(b)}{\leq} \frac{1}{1-\mu\delta} \big\| \big(I - \mu\nabla^2 J(\boldsymbol{w}_i)\big) \big(\boldsymbol{w}_{i+j} - \boldsymbol{w}'_{i+j}\big) \big\|^2 \\
&\quad + \frac{\mu}{\delta} \big\| \big(\boldsymbol{H}_{i+j} - \nabla^2 J(\boldsymbol{w}_i)\big) \widetilde{\boldsymbol{w}}^i_j \big\|^2 \\
&\overset{(7.85)}{\leq} \frac{(1+\mu\delta)^2}{1-\mu\delta} \big\| \boldsymbol{w}_{i+j} - \boldsymbol{w}'_{i+j} \big\|^2 + \frac{\mu}{\delta}\frac{\rho}{2} \big\| \widetilde{\boldsymbol{w}}^i_j \big\|^4
\end{aligned} \tag{7.86}$$

where (a) again follows from Jensen's inequality (7.66) with $\alpha = 1 - \mu\delta$ and (b) follows from the same inequality with $\alpha = \frac{1}{2}$. Taking expectations over $\boldsymbol{w}_i \in \mathcal{H}$ yields:

$$\begin{aligned}
&\mathbb{E} \Big\{ \| \boldsymbol{w}_{i+j+1} - \boldsymbol{w}'_{i+j+1} \|^2 \big| \, \boldsymbol{w}_i \in \mathcal{H} \Big\} \\
&\leq \frac{(1+\mu\delta)^2}{1-\mu\delta} \, \mathbb{E} \Big\{ \big\| \boldsymbol{w}_{i+j} - \boldsymbol{w}'_{i+j} \big\|^2 \big| \, \boldsymbol{w}_i \in \mathcal{H} \Big\} \\
&\quad + \frac{\mu}{\delta}\frac{\rho}{2} \, \mathbb{E} \Big\{ \big\| \widetilde{\boldsymbol{w}}^i_j \big\|^4 \big| \, \boldsymbol{w}_i \in \mathcal{H} \Big\} \\
&\overset{(7.81)}{\leq} \frac{(1+\mu\delta)^2}{1-\mu\delta} \, \mathbb{E} \big\| \boldsymbol{w}_{i+j} - \boldsymbol{w}'_{i+j} \big\|^2 + O(\mu^3)
\end{aligned} \tag{7.87}$$

Since both the true and the short-term model are initialized at $\boldsymbol{w}_i$, we have $\boldsymbol{w}_{i+0} - \boldsymbol{w}'_{i+0} = 0$. Iterating and applying the same argument as above leads to:

$$\mathbb{E} \, \| \boldsymbol{w}_{i+j+1} - \boldsymbol{w}'_{i+j+1} \|^2 \leq O(\mu^2) \tag{7.88}$$

which is (7.47).

7.D Proof of Lemma 7.1

Recall that $J(\cdot)$ has δ-Lipschitz gradients, which implies:

$$J(\boldsymbol{w}_{i+j}) \leq J(\boldsymbol{w}'_{i+j}) + \nabla J\left(\boldsymbol{w}'_{i+j}\right)^{\mathsf{T}}\left(\boldsymbol{w}_{i+j} - \boldsymbol{w}'_{i+j}\right) + \frac{\delta}{2}\left\|\boldsymbol{w}_{i+j} - \boldsymbol{w}'_{i+j}\right\|^2 \tag{7.89}$$

In the vicinity of saddle-points, we can refine the upper bound (7.89) by taking expectations conditioned on $\boldsymbol{w}_i \in \mathcal{H}$:

$$\begin{aligned}
&\mathbb{E}\left\{J(\boldsymbol{w}_{i+j})|\,\boldsymbol{w}_i \in \mathcal{H}\right\} \\
\leq\ &\mathbb{E}\left\{J(\boldsymbol{w}'_{i+j})|\,\boldsymbol{w}_i \in \mathcal{H}\right\} \\
&+ \mathbb{E}\left\{\nabla J\left(\boldsymbol{w}'_{i+j}\right)^{\mathsf{T}}\left(\boldsymbol{w}_{i+j} - \boldsymbol{w}'_{i+j}\right)|\,\boldsymbol{w}_i \in \mathcal{H}\right\} \\
&+ \frac{\delta}{2}\,\mathbb{E}\left\{\left\|\boldsymbol{w}_{i+j} - \boldsymbol{w}'_{i+j}\right\|^2|\,\boldsymbol{w}_i \in \mathcal{H}\right\} \\
\overset{(a)}{\leq}\ &\mathbb{E}\left\{J(\boldsymbol{w}'_{i+j})|\,\boldsymbol{w}_i \in \mathcal{H}\right\} \\
&+ \sqrt{\mathbb{E}\left\{\left\|\nabla J\left(\boldsymbol{w}'_{i+j}\right)\right\|^2|\,\boldsymbol{w}_i \in \mathcal{H}\right\}} \\
&\quad\times\sqrt{\mathbb{E}\left\{\left\|\boldsymbol{w}_{i+j} - \boldsymbol{w}'_{i+j}\right\|^2|\,\boldsymbol{w}_i \in \mathcal{H}\right\}} \\
&+ \frac{\delta}{2}\,\mathbb{E}\left\{\left\|\boldsymbol{w}_{i+j} - \boldsymbol{w}'_{i+j}\right\|^2|\,\boldsymbol{w}_i \in \mathcal{H}\right\} \\
\overset{(a)}{\leq}\ &\mathbb{E}\left\{J(\boldsymbol{w}'_{i+j})|\,\boldsymbol{w}_i \in \mathcal{H}\right\} \\
&+ \sqrt{\mathbb{E}\left\{2\|\nabla J\left(\boldsymbol{w}_i\right)\|^2 + 2\delta^2\left\|\widetilde{\boldsymbol{w}}'^{i}_{j}\right\|^2|\,\boldsymbol{w}_i \in \mathcal{H}\right\}} \\
&\quad\times\sqrt{\mathbb{E}\left\{\left\|\boldsymbol{w}_{i+j} - \boldsymbol{w}'_{i+j}\right\|^2|\,\boldsymbol{w}_i \in \mathcal{H}\right\}} \\
&+ \frac{\delta}{2}\,\mathbb{E}\left\{\left\|\boldsymbol{w}_{i+j} - \boldsymbol{w}'_{i+j}\right\|^2|\,\boldsymbol{w}_i \in \mathcal{H}\right\} \\
\overset{(b)}{\leq}\ &\mathbb{E}\left\{J(\boldsymbol{w}'_{i+j})|\,\boldsymbol{w}_i \in \mathcal{H}\right\} \\
&+ O\left(\mu^{1/2}\right)\sqrt{\mathbb{E}\left\{\left\|\boldsymbol{w}_{i+j} - \boldsymbol{w}'_{i+j}\right\|^2|\,\boldsymbol{w}_i \in \mathcal{H}\right\}} \\
&+ \frac{\delta}{2}\,\mathbb{E}\left\{\left\|\boldsymbol{w}_{i+j} - \boldsymbol{w}'_{i+j}\right\|^2|\,\boldsymbol{w}_i \in \mathcal{H}\right\} \\
\overset{(c)}{\leq}\ &\mathbb{E}\left\{J(\boldsymbol{w}'_{i+j})|\,\boldsymbol{w}_i \in \mathcal{H}\right\} + O(\mu^{3/2})
\end{aligned} \tag{7.90}$$

where (a) follows from:

$$\begin{aligned}
&\left\|\nabla J\left(\boldsymbol{w}_{i+j}'\right)\right\|^2 \\
&= \left\|\nabla J\left(\boldsymbol{w}_i\right) + \nabla J\left(\boldsymbol{w}_{i+j}'\right) - \nabla J\left(\boldsymbol{w}_i\right)\right\|^2 \\
&\leq 2\|\nabla J\left(\boldsymbol{w}_i\right)\|^2 + 2\left\|\nabla J\left(\boldsymbol{w}_{i+j}'\right) - \nabla J\left(\boldsymbol{w}_i\right)\right\|^2 \\
&\leq 2\|\nabla J\left(\boldsymbol{w}_i\right)\|^2 + 2\delta^2\left\|\boldsymbol{w}_{i+j}' - \boldsymbol{w}_i\right\|^2
\end{aligned} \tag{7.91}$$

Step (b) follows from Cauchy-Schwarz inequality and (c) is a result of the definition of $\mathcal{H}$ as approximately strict-saddle points (7.24) and (7.48) and (c) is a result of (7.47).

7.E Proof of Theorem 7.2

The argument generally mirrors the proof to [71, Theorem 1] after accounting for the relative variance bound (7.17) by noting that, around first-order stationary points, the relative component $\beta^4\|\nabla J(\boldsymbol{w}_i)\|^4$ will necessarily be small.

From Corollary 7.1, we have:

$$\mathbb{E}\left\{J(\boldsymbol{w}_{i+j})|\,\boldsymbol{w}_i \in \mathcal{H}\right\} \leq \mathbb{E}\left\{J(\boldsymbol{w}_{i+j}')|\,\boldsymbol{w}_i \in \mathcal{H}\right\} + O(\mu^{3/2}) \tag{7.92}$$

so long as $j \leq \frac{T}{\mu}$. We can hence proceed by studying $\mathbb{E}\left\{J(\boldsymbol{w}_{i+j}')|\mathcal{H}\right\}$ and will add the approximation error $O(\mu^{3/2})$ to the end result. From (7.14) we find:

$$\begin{aligned}
J(\boldsymbol{w}_{i+j}') \leq J(\boldsymbol{w}_i) - \nabla J(\boldsymbol{w}_i)^{\mathsf{T}}\widetilde{\boldsymbol{w}}'^i_j + \frac{1}{2}\left\|\widetilde{\boldsymbol{w}}'^i_j\right\|^2_{\nabla^2 J(\boldsymbol{w}_i)} \\
+ \frac{\rho}{6}\left\|\widetilde{\boldsymbol{w}}'^i_j\right\|^3
\end{aligned} \tag{7.93}$$

We will bound each term appearing on the right-hand side. From (7.42) we find after

conditioning on $\boldsymbol{\mathcal{F}}_{i+j}$:

$$\begin{aligned}
&\mathbb{E}\left\{\widetilde{\boldsymbol{w}}'^{i}_{j+1}|\boldsymbol{\mathcal{F}}_{i+j}\right\} \\
=\;&\left(I-\mu\nabla^2 J(\boldsymbol{w}_i)\right)\widetilde{\boldsymbol{w}}'^{i}_{j}+\mu\nabla J(\boldsymbol{w}_i)+\mu\,\mathbb{E}\left\{\boldsymbol{s}_{i+j+1}\,|\boldsymbol{\mathcal{F}}_{i+j}\right\} \\
\overset{(7.16)}{=}\;&\left(I-\mu\nabla^2 J(\boldsymbol{w}_i)\right)\widetilde{\boldsymbol{w}}'^{i}_{j}+\mu\nabla J(\boldsymbol{w}_i)
\end{aligned} \tag{7.94}$$

Note that $\boldsymbol{\mathcal{F}}_{i+j}$ denotes the information captured in $\boldsymbol{w}_{k,j}$ up to time $i+j$, while $\boldsymbol{\mathcal{F}}_i$ denotes the information available up to time i. Hence:

$$\boldsymbol{\mathcal{F}}_{i+j}=\boldsymbol{\mathcal{F}}_i\cup\text{filtration}\left\{\boldsymbol{w}_{k,i+1},\ldots,\boldsymbol{w}_{k,i+j}\right\} \tag{7.95}$$

Hence, taking expectation of (7.94) conditioned on $\boldsymbol{\mathcal{F}}_i$ removes the elements not contained in $\boldsymbol{\mathcal{F}}_i$ and yields:

$$\begin{aligned}
\mathbb{E}\left\{\widetilde{\boldsymbol{w}}'^{i}_{j+1}|\boldsymbol{\mathcal{F}}_i\right\}=\;&\left(I-\mu\nabla^2 J(\boldsymbol{w}_i)\right)\mathbb{E}\left\{\widetilde{\boldsymbol{w}}'^{i}_{j}|\boldsymbol{\mathcal{F}}_i\right\} \\
&+\mu\nabla J(\boldsymbol{w}_i)
\end{aligned} \tag{7.96}$$

Since $\widetilde{\boldsymbol{w}}'^{i}_{0}=0$, iterating starting at $j=0$ yields:

$$\mathbb{E}\left\{\widetilde{\boldsymbol{w}}'^{i}_{j}|\boldsymbol{\mathcal{F}}_i\right\}=\mu\left(\sum_{k=1}^{j}\left(I-\mu\nabla^2 J(\boldsymbol{w}_i)\right)^{k-1}\right)\nabla J(\boldsymbol{w}_i) \tag{7.97}$$

This allows us to bound the linear term appearing in (7.93) as:

$$\begin{aligned}
&-\mathbb{E}\left\{\nabla J(\boldsymbol{w}_i)^{\mathsf{T}}\widetilde{\boldsymbol{w}}'^{i}_{j}|\boldsymbol{\mathcal{F}}_i\right\} \\
=\;&-\nabla J(\boldsymbol{w}_i)^{\mathsf{T}}\,\mathbb{E}\left\{\widetilde{\boldsymbol{w}}'^{i}_{j}|\boldsymbol{\mathcal{F}}_i\right\} \\
\overset{(7.97)}{=}\;&-\mu\nabla J(\boldsymbol{w}_i)^{\mathsf{T}}\left(\sum_{k=1}^{j}\left(I-\mu\nabla^2 J(\boldsymbol{w}_i)\right)^{k-1}\right)\nabla J(\boldsymbol{w}_i) \\
=\;&-\mu\|\nabla J(\boldsymbol{w}_i)\|^2_{\sum_{k=1}^{j}\left(I-\mu\nabla^2 J(\boldsymbol{w}_i)\right)^{k-1}}
\end{aligned} \tag{7.98}$$

To study the quadratic term in (7.93), we introduce the eigenvalue decomposition of the

Hessian around the iterate at time i:

$$\nabla^2 J(\boldsymbol{w}_i) \triangleq \boldsymbol{V}_i \boldsymbol{\Lambda}_i \boldsymbol{V}_i^{\mathsf{T}} \tag{7.99}$$

which motivates the transformation:

$$\begin{aligned} \big\|\widetilde{\boldsymbol{w}}'^{i}_{j+1}\big\|^2_{\nabla^2 J(\boldsymbol{w}_i)} &= \big\|\widetilde{\boldsymbol{w}}'^{i}_{j+1}\big\|^2_{\boldsymbol{V}_i \boldsymbol{\Lambda}_i \boldsymbol{V}_i^{\mathsf{T}}} \\ &= \big\|\boldsymbol{V}_i^{\mathsf{T}} \boldsymbol{w}_i - \boldsymbol{V}_i^{\mathsf{T}} \boldsymbol{w}'_{i+j+1}\big\|^2_{\boldsymbol{\Lambda}_i} \\ &= \big\|\overline{\boldsymbol{w}}'^{i}_{j+1}\big\|^2_{\boldsymbol{\Lambda}_i} \end{aligned} \tag{7.100}$$

where we introduced:

$$\overline{\boldsymbol{w}}'^{i}_{j+1} \triangleq \boldsymbol{V}_i^{\mathsf{T}} \widetilde{\boldsymbol{w}}'^{i}_{j+1} \tag{7.101}$$

Under this transformation, recursion (7.42) is also diagonalized, yielding:

$$\begin{aligned} &\overline{\boldsymbol{w}}'^{i}_{j+1} \\ &\triangleq \boldsymbol{V}_i^{\mathsf{T}} \widetilde{\boldsymbol{w}}'^{i}_{j+1} \\ &= \boldsymbol{V}_i^{\mathsf{T}} \big(I - \mu \nabla^2 J(\boldsymbol{w}_i)\big) \boldsymbol{V}_i \boldsymbol{V}_i^{\mathsf{T}} \widetilde{\boldsymbol{w}}'^{i}_{j} \\ &\quad + \mu \boldsymbol{V}_i^{\mathsf{T}} \nabla J(\boldsymbol{w}_i) + \mu \boldsymbol{V}_i^{\mathsf{T}} \boldsymbol{s}_{i+j+1} \\ &= (I - \mu \boldsymbol{\Lambda}_i)\, \overline{\boldsymbol{w}}'^{i}_{j} + \mu \overline{\nabla} J(\boldsymbol{w}_i) + \mu \overline{\boldsymbol{s}}_{i+j+1} \end{aligned} \tag{7.102}$$

with $\overline{\nabla} J(\boldsymbol{w}_i) \triangleq \boldsymbol{V}_i^{\mathsf{T}} \nabla J(\boldsymbol{w}_i)$ and $\overline{\boldsymbol{s}}_{i+j+1} \triangleq \boldsymbol{V}_i^{\mathsf{T}} \boldsymbol{s}_{i+j+1}$. Applying the same transformation to the conditional mean recursion (7.96), and subtracting the transformed conditional mean on both sides of (7.102), we find:

$$\begin{aligned} &\overline{\boldsymbol{w}}'^{i}_{j+1} - \mathbb{E}\,\big\{\overline{\boldsymbol{w}}'^{i}_{j+1} | \boldsymbol{\mathcal{F}}_i\big\} \\ &= (I - \mu \boldsymbol{\Lambda}_i)\,\big(\overline{\boldsymbol{w}}'^{i}_{j} - \mathbb{E}\,\big\{\overline{\boldsymbol{w}}'^{i}_{j} | \boldsymbol{\mathcal{F}}_i\big\}\big) + \mu \overline{\boldsymbol{s}}_{i+j+1} \end{aligned} \tag{7.103}$$

which allows us to cancel the driving term involving the gradient. For brevity, define the

(conditionally) centered random variable:

$$\check{\boldsymbol{w}}'^{i}_{j+1} = \overline{\boldsymbol{w}}'^{i}_{j+1} - \mathbb{E}\left\{\overline{\boldsymbol{w}}'^{i}_{j+1}|\boldsymbol{\mathcal{F}}_i\right\} \tag{7.104}$$

so that:

$$\check{\boldsymbol{w}}'^{i}_{j+1} = (I - \mu\boldsymbol{\Lambda}_i)\,\check{\boldsymbol{w}}'^{i}_{j} + \mu\overline{\boldsymbol{s}}_{i+j+1} \tag{7.105}$$

Before proceeding, note that we can express:

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^{i}_{j}\right\|^2_{\boldsymbol{\Lambda}_i}|\boldsymbol{\mathcal{F}}_i\right\} \\
&= \mathbb{E}\left\{\left\|\overline{\boldsymbol{w}}'^{i}_{j} - \mathbb{E}\left\{\overline{\boldsymbol{w}}'^{i}_{j}|\boldsymbol{\mathcal{F}}_i\right\}\right\|^2_{\boldsymbol{\Lambda}_i}|\boldsymbol{\mathcal{F}}_i\right\} \\
&= \mathbb{E}\left\{\left\|\overline{\boldsymbol{w}}'^{i}_{j}\right\|^2_{\boldsymbol{\Lambda}_i}|\boldsymbol{\mathcal{F}}_i\right\} - \left\|\mathbb{E}\left\{\overline{\boldsymbol{w}}'^{i}_{j}|\boldsymbol{\mathcal{F}}_i\right\}\right\|^2_{\boldsymbol{\Lambda}_i}
\end{aligned} \tag{7.106}$$

Hence, we have:

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}'^{i}_{j}\right\|^2_{\nabla^2 J(\boldsymbol{w}_i)}|\boldsymbol{\mathcal{F}}_i\right\} \\
&= \mathbb{E}\left\{\left\|\overline{\boldsymbol{w}}'^{i}_{j}\right\|^2_{\boldsymbol{\Lambda}_i}|\boldsymbol{\mathcal{F}}_i\right\} \\
&= \mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^{i}_{j}\right\|^2_{\boldsymbol{\Lambda}_i}|\boldsymbol{\mathcal{F}}_i\right\} + \left\|\mathbb{E}\left\{\overline{\boldsymbol{w}}'^{i}_{j}|\boldsymbol{\mathcal{F}}_i\right\}\right\|^2_{\boldsymbol{\Lambda}_i}
\end{aligned} \tag{7.107}$$

In order to make claims about $\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}'^{i}_{j}\right\|^2_{\nabla^2 J(\boldsymbol{w}_i)}|\boldsymbol{\mathcal{F}}_i\right\}$ by studying $\mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^{i}_{j}\right\|^2_{\boldsymbol{\Lambda}_i}|\boldsymbol{\mathcal{F}}_i\right\}$, we need

to establish a bound on $\left\|\mathbb{E}\left\{\overline{\boldsymbol{w}}_j'^i|\boldsymbol{\mathcal{F}}_i\right\}\right\|_{\boldsymbol{\Lambda}_i}^2$. We have:

$$\begin{aligned}
&\left\|\mathbb{E}\left\{\overline{\boldsymbol{w}}_j'^i|\boldsymbol{\mathcal{F}}_i\right\}\right\|_{\boldsymbol{\Lambda}_i}^2 \\
&= \left\|\mathbb{E}\left\{\boldsymbol{V}_i^{\mathsf{T}}\widetilde{\boldsymbol{w}}_j'^i|\boldsymbol{\mathcal{F}}_i\right\}\right\|_{\boldsymbol{\Lambda}_i}^2 \\
&\overset{(7.97)}{=} \mu^2\left\|\boldsymbol{V}_i^{\mathsf{T}}\left(\sum_{k=1}^{j}\left(I-\mu\nabla^2 J(\boldsymbol{w}_i)\right)^{k-1}\right)\nabla J(\boldsymbol{w}_i)\right\|_{\boldsymbol{\Lambda}_i}^2 \\
&= \mu^2\left\|\left(\sum_{k=1}^{j}\left(I-\mu\boldsymbol{\Lambda}_i\right)^{k-1}\right)\overline{\nabla} J(\boldsymbol{w}_i)\right\|_{\boldsymbol{\Lambda}_i}^2 \\
&= \mu^2\overline{\nabla} J(\boldsymbol{w}_i)^{\mathsf{T}}\left(\sum_{k=1}^{j}\left(I-\mu\boldsymbol{\Lambda}_i\right)^{k-1}\right)\boldsymbol{\Lambda}_i \\
&\qquad\times\left(\sum_{k=1}^{j}\left(I-\mu\boldsymbol{\Lambda}_i\right)^{k-1}\right)\overline{\nabla} J(\boldsymbol{w}_i)
\end{aligned} \tag{7.108}$$

We shall order the eigenvalues of $\nabla^2 J(\boldsymbol{w}_i)$, such that its eigendecomposition has a block structure:

$$\boldsymbol{V}_i = \begin{bmatrix} \boldsymbol{V}_i^{\geq 0} & \boldsymbol{V}_i^{<0} \end{bmatrix}, \quad \boldsymbol{\Lambda}_i = \begin{bmatrix} \boldsymbol{\Lambda}_i^{\geq 0} & 0 \\ 0 & \boldsymbol{\Lambda}_i^{<0} \end{bmatrix} \tag{7.109}$$

with $\delta I \geq \boldsymbol{\Lambda}_i^{\geq 0} \geq 0$ and $\boldsymbol{\Lambda}_i^{<0} < 0$. Note that since $\nabla^2 J(\boldsymbol{w}_i)$ is random, the decomposition itself is random as well. Nevertheless, it exists with probability one. We also decompose the transformed gradient vector with appropriate dimensions:

$$\overline{\nabla} J(\boldsymbol{w}_i) = \operatorname{col}\left\{\overline{\nabla} J(\boldsymbol{w}_i)^{\geq 0}, \overline{\nabla} J(\boldsymbol{w}_i)^{<0}\right\} \tag{7.110}$$

We can then decompose (7.108):

$$\begin{aligned}
&\left\|\mathbb{E}\left\{\overline{\boldsymbol{w}}_j'^i|\boldsymbol{\mathcal{F}}_i\right\}\right\|_{\boldsymbol{\Lambda}_i}^2 \\
&= \mu^2\overline{\nabla} J(\boldsymbol{w}_i)^{\mathsf{T}}\left(\sum_{k=1}^{j}\left(I-\mu\boldsymbol{\Lambda}_i\right)^{k-1}\right)\boldsymbol{\Lambda}_i \\
&\qquad\times\left(\sum_{k=1}^{j}\left(I-\mu\boldsymbol{\Lambda}_i\right)^{k-1}\right)\overline{\nabla} J(\boldsymbol{w}_i)
\end{aligned}$$

$$
\begin{aligned}
&= \mu^2 \Big(\overline{\nabla} J(\boldsymbol{w}_i)^{\geq 0}\Big)^{\mathsf{T}} \left(\sum_{k=1}^{j} \left(I - \mu \boldsymbol{\Lambda}_i^{\geq 0}\right)^{k-1}\right) \boldsymbol{\Lambda}_i^{\geq 0} \\
&\qquad \times \left(\sum_{k=1}^{j} \left(I - \mu \boldsymbol{\Lambda}_i^{\geq 0}\right)^{k-1}\right) \overline{\nabla} J(\boldsymbol{w}_i)^{\geq 0} \\
&\quad + \mu^2 \Big(\overline{\nabla} J(\boldsymbol{w}_i)^{<0}\Big)^{\mathsf{T}} \left(\sum_{k=1}^{j} \left(I - \mu \boldsymbol{\Lambda}_i^{<0}\right)^{k-1}\right) \boldsymbol{\Lambda}_i^{<0} \\
&\qquad \times \left(\sum_{k=1}^{j} \left(I - \mu \boldsymbol{\Lambda}_i^{<0}\right)^{k-1}\right) \overline{\nabla} J(\boldsymbol{w}_i)^{<0} \\
&\overset{(a)}{\leq} \mu^2 \Big(\overline{\nabla} J(\boldsymbol{w}_i)^{\geq 0}\Big)^{\mathsf{T}} \left(\sum_{k=1}^{j} \left(I - \mu \boldsymbol{\Lambda}_i^{\geq 0}\right)^{k-1}\right) \boldsymbol{\Lambda}_i^{\geq 0} \\
&\qquad \times \left(\sum_{k=1}^{j} \left(I - \mu \boldsymbol{\Lambda}_i^{\geq 0}\right)^{k-1}\right) \overline{\nabla} J(\boldsymbol{w}_i)^{\geq 0} \\
&\overset{(b)}{\leq} \mu^2 \Big(\overline{\nabla} J(\boldsymbol{w}_i)^{\geq 0}\Big)^{\mathsf{T}} \left(\sum_{k=1}^{\infty} \left(I - \mu \boldsymbol{\Lambda}_i^{\geq 0}\right)^{k-1}\right) \boldsymbol{\Lambda}_i^{\geq 0} \\
&\qquad \times \left(\sum_{k=1}^{j} \left(I - \mu \boldsymbol{\Lambda}_i^{\geq 0}\right)^{k-1}\right) \overline{\nabla} J(\boldsymbol{w}_i)^{\geq 0} \\
&\overset{(c)}{=} \mu^2 \Big(\overline{\nabla} J(\boldsymbol{w}_i)^{\geq 0}\Big)^{\mathsf{T}} \left(\mu \boldsymbol{\Lambda}_i^{\geq 0}\right)^{-1} \boldsymbol{\Lambda}_i^{\geq 0} \\
&\qquad \times \left(\sum_{k=1}^{j} \left(I - \mu \boldsymbol{\Lambda}_i^{\geq 0}\right)^{k-1}\right) \overline{\nabla} J(\boldsymbol{w}_i)^{\geq 0} \\
&= \mu \Big(\overline{\nabla} J(\boldsymbol{w}_i)^{\geq 0}\Big)^{\mathsf{T}} \left(\sum_{k=1}^{j} \left(I - \mu \boldsymbol{\Lambda}_i^{\geq 0}\right)^{k-1}\right) \overline{\nabla} J(\boldsymbol{w}_i)^{\geq 0} \\
&\overset{(d)}{\leq} \mu \Big(\overline{\nabla} J(\boldsymbol{w}_i)^{\geq 0}\Big)^{\mathsf{T}} \left(\sum_{k=1}^{j} \left(I - \mu \boldsymbol{\Lambda}_i^{\geq 0}\right)^{k-1}\right) \overline{\nabla} J(\boldsymbol{w}_i)^{\geq 0} \\
&\quad + \mu \Big(\overline{\nabla} J(\boldsymbol{w}_i)^{<0}\Big)^{\mathsf{T}} \left(\sum_{k=1}^{j} \left(I - \mu \boldsymbol{\Lambda}_i^{<0}\right)^{k-1}\right) \overline{\nabla} J(\boldsymbol{w}_i)^{<0} \\
&\leq \mu \overline{\nabla} J(\boldsymbol{w}_i)^{\mathsf{T}} \left(\sum_{k=1}^{j} \left(I - \mu \boldsymbol{\Lambda}_i\right)^{k-1}\right) \overline{\nabla} J(\boldsymbol{w}_i) \\
&= \mu \big\| \overline{\nabla} J(\boldsymbol{w}_i) \big\|^2_{\sum_{k=1}^{j} (I - \mu \boldsymbol{\Lambda}_i)^{k-1}}
\end{aligned} \tag{7.111}
$$

where (a) follows from $\boldsymbol{\Lambda}_i^{\preceq 0} < 0$, (b) follows from:

$$\sum_{k=1}^{j} \left(I - \mu \boldsymbol{\Lambda}_i^{\succeq 0}\right)^{k-1} \leq \sum_{k=1}^{\infty} \left(I - \mu \boldsymbol{\Lambda}_i^{\succeq 0}\right)^{k-1} \tag{7.112}$$

for $\mu < \frac{1}{\delta}$. Step (c) follows from the formula for the geometric matrix series, and (d) follows from:

$$\mu \Big(\overline{\nabla} J(\boldsymbol{w}_i)^{\succeq 0}\Big)^{\mathsf{T}} \left(\sum_{k=1}^{j} \left(I - \mu \boldsymbol{\Lambda}_i^{\succeq 0}\right)^{k-1}\right) \overline{\nabla} J(\boldsymbol{w}_i)^{\succeq 0} \geq 0 \tag{7.113}$$

Comparing (7.111) to (7.98), we find that we can bound:

$$-\mathbb{E}\left\{\nabla J(\boldsymbol{w}_i)^{\mathsf{T}} \widetilde{\boldsymbol{w}}'^i_j | \boldsymbol{\mathcal{F}}_i\right\} + \left\|\mathbb{E}\left\{\overline{\boldsymbol{w}}'^i_j | \boldsymbol{\mathcal{F}}_i\right\}\right\|^2_{\boldsymbol{\Lambda}_i} \leq 0 \tag{7.114}$$

To recap, we can simplify (7.93) as:

$$\begin{aligned} &\mathbb{E}\left\{J(\boldsymbol{w}'_{i+j}) | \boldsymbol{\mathcal{F}}_i\right\} \\ &\leq J(\boldsymbol{w}_i) + \frac{1}{2}\,\mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^i_j\right\|^2_{\boldsymbol{\Lambda}_i} | \boldsymbol{\mathcal{F}}_i\right\} + \frac{\rho}{6}\,\mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}'^i_j\right\|^3 | \boldsymbol{\mathcal{F}}_i\right\} \end{aligned} \tag{7.115}$$

We proceed with the now simplified quadratic term. Motivated by a technique employed for the analysis of adaptive filters and stochastic gradient algorithms in *convex* environments [1, 149], we square both sides of (7.105) under an arbitrary diagonal weighting matrix $\boldsymbol{\Sigma}_i$, deterministic conditioned on $\boldsymbol{w}_i$ and $\boldsymbol{w}_{i+j}$, to obtain:

$$\begin{aligned} &\left\|\check{\boldsymbol{w}}'^i_{j+1}\right\|^2_{\boldsymbol{\Sigma}_i} \\ &= \left\|\left(I - \mu \boldsymbol{\Lambda}_i\right) \check{\boldsymbol{w}}'^i_j + \mu \overline{\boldsymbol{s}}_{i+j+1}\right\|^2_{\boldsymbol{\Sigma}_i} \\ &= \left\|\left(I - \mu \boldsymbol{\Lambda}_i\right) \check{\boldsymbol{w}}'^i_j\right\|^2_{\boldsymbol{\Sigma}_i} + \mu^2 \|\overline{\boldsymbol{s}}_{i+j+1}\|^2_{\boldsymbol{\Sigma}_i} \\ &\quad + 2\mu \check{\boldsymbol{w}}'^{i\,\mathsf{T}}_j \left(I - \mu \boldsymbol{\Lambda}_i\right) \boldsymbol{\Sigma}_i \overline{\boldsymbol{s}}_{i+j+1} \end{aligned} \tag{7.116}$$

Note that upon conditioning on $\boldsymbol{\mathcal{F}}_{i+j}$, all elements of the cross-term, aside from $\overline{\boldsymbol{s}}_{i+j+1}$,

become deterministic, and as such the term disappears when taking expectations. We obtain:

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^{i}_{j+1}\right\|^2_{\boldsymbol{\Sigma}_i}|\boldsymbol{\mathcal{F}}_{i+j}\right\} \\
&= \left\|(I-\mu\boldsymbol{\Lambda}_i)\,\check{\boldsymbol{w}}'^{i}_{j}\right\|^2_{\boldsymbol{\Sigma}_i} + \mu^2\,\mathbb{E}\left\{\|\overline{\boldsymbol{s}}_{i+j+1}\|^2_{\boldsymbol{\Sigma}_i}|\boldsymbol{\mathcal{F}}_{i+j}\right\} \\
&= \left\|\check{\boldsymbol{w}}'^{i}_{j}\right\|^2_{\boldsymbol{\Sigma}_i-2\mu\boldsymbol{\Lambda}_i\boldsymbol{\Sigma}_i+\mu^2\boldsymbol{\Lambda}_i\boldsymbol{\Sigma}_i\boldsymbol{\Lambda}_i} \\
&\quad + \mu^2\mathrm{Tr}\left(\boldsymbol{V}_i\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{\mathsf{T}}R_s\left(\boldsymbol{w}_{i+j}\right)\right) \\
&= \left\|\check{\boldsymbol{w}}'^{i}_{j}\right\|^2_{\boldsymbol{\Sigma}_i-2\mu\boldsymbol{\Lambda}_i\boldsymbol{\Sigma}_i} + \mu^2\mathrm{Tr}\left(\boldsymbol{V}_i\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{\mathsf{T}}R_s\left(\boldsymbol{w}_{i}\right)\right) \\
&\quad + \mu^2\mathrm{Tr}\left(\boldsymbol{V}_i\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{\mathsf{T}}\left(R_s\left(\boldsymbol{w}_{i+j}\right)-R_s\left(\boldsymbol{w}_{i}\right)\right)\right) \\
&\quad + \mu^2\left\|\check{\boldsymbol{w}}'^{i}_{j}\right\|^2_{\boldsymbol{\Lambda}_i\boldsymbol{\Sigma}_i\boldsymbol{\Lambda}_i}
\end{aligned} \tag{7.117}$$

We proceed to bound the last two terms. First, we have:

$$\begin{aligned}
&\mathrm{Tr}\left(\boldsymbol{V}_i\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{\mathsf{T}}\left(R_s\left(\boldsymbol{w}_{i+j}\right)-R_s\left(\boldsymbol{w}_{i}\right)\right)\right) \\
&\overset{(a)}{\leq} \left\|\boldsymbol{V}_i\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{\mathsf{T}}\right\|\left\|R_s\left(\boldsymbol{w}_{i+j}\right)-R_s\left(\boldsymbol{w}_{i}\right)\right\| \\
&\overset{(b)}{\leq} \rho\left(\boldsymbol{\Sigma}_i\right)\beta_R\left\|\widetilde{\boldsymbol{w}}^{i}_{j}\right\|^{\gamma}
\end{aligned} \tag{7.118}$$

where (a) follows from Cauchy-Schwarz, since $\mathrm{Tr}(A^{\mathsf{T}}B)$ is an inner product over the space of symmetric matrices, and hence, $|\mathrm{Tr}(A^{\mathsf{T}}B)| \leq \|A\|\|B\|$, and (b) follows from Assumption 7.4. For the second term, we have:

$$\begin{aligned}
\left\|\check{\boldsymbol{w}}'^{i}_{j}\right\|^2_{\boldsymbol{\Lambda}_i\boldsymbol{\Sigma}_i\boldsymbol{\Lambda}_i} &\leq \rho\left(\boldsymbol{\Lambda}_i\boldsymbol{\Sigma}_i\boldsymbol{\Lambda}_i\right)\left\|\check{\boldsymbol{w}}'^{i}_{j}\right\|^2 \\
&\leq \delta^2\rho\left(\boldsymbol{\Sigma}_i\right)\left\|\check{\boldsymbol{w}}'^{i}_{j}\right\|^2
\end{aligned} \tag{7.119}$$

We conclude that

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^{i}_{j+1}\right\|^2_{\boldsymbol{\Sigma}_i}|\boldsymbol{\mathcal{F}}_{i}\right\} \\
&= \mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^{i}_{j}\right\|^2_{\boldsymbol{\Sigma}_i-2\mu\boldsymbol{\Lambda}_i\boldsymbol{\Sigma}_i}|\boldsymbol{\mathcal{F}}_{i}\right\} + \mu^2\mathrm{Tr}\left(\boldsymbol{V}_i\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{\mathsf{T}}R_s\left(\boldsymbol{w}_{i}\right)\right) \\
&\quad + \mu^2\rho\left(\boldsymbol{\Sigma}_i\right)\mathbb{E}\left\{\boldsymbol{q}_{i+j}|\boldsymbol{\mathcal{F}}_{i}\right\}
\end{aligned} \tag{7.120}$$

where

$$\boldsymbol{q}_{i+j} \triangleq \beta_R \left\| \widetilde{\boldsymbol{w}}_j^i \right\|^{\gamma} + \delta^2 \left\| \check{\boldsymbol{w}}'^i_j \right\|^2 \tag{7.121}$$

For brevity, we define

$$\boldsymbol{D} \triangleq I - 2\mu\boldsymbol{\Lambda}_i \tag{7.122}$$

$$\boldsymbol{Y} \triangleq \boldsymbol{V}_i^{\mathsf{T}} R_s\left(\boldsymbol{w}_i\right) \boldsymbol{V}_i \tag{7.123}$$

With these substitutions we obtain:

$$\begin{aligned} &\mathbb{E}\left\{ \left\| \check{\boldsymbol{w}}'^i_{j+1} \right\|^2_{\boldsymbol{\Sigma}_i} | \boldsymbol{\mathcal{F}}_i \right\} \\ &= \mathbb{E}\left\{ \left\| \check{\boldsymbol{w}}'^i_{j} \right\|^2_{\boldsymbol{D}\boldsymbol{\Sigma}_i} | \boldsymbol{\mathcal{F}}_i \right\} + \mu^2 \mathrm{Tr}\left(\boldsymbol{\Sigma}_i \boldsymbol{Y}\right) + \mu^2 \rho\left(\boldsymbol{\Sigma}_i\right) \mathbb{E}\left\{ \boldsymbol{q}_{i+j} | \boldsymbol{\mathcal{F}}_i \right\} \end{aligned} \tag{7.124}$$

At $j = 0$, we have $\check{\boldsymbol{w}}'^i_0 = 0$. Letting $\boldsymbol{\Sigma}_j = \boldsymbol{\Lambda}_i \boldsymbol{D}^j$, we can iterate to obtain:

$$\begin{aligned} &\mathbb{E}\left\{ \left\| \check{\boldsymbol{w}}'^i_{j+1} \right\|^2_{\boldsymbol{\Lambda}_i} | \boldsymbol{\mathcal{F}}_i \right\} \\ &= \mu^2 \sum_{n=0}^{j} \mathrm{Tr}\left(\boldsymbol{\Lambda}_i \boldsymbol{D}^n \boldsymbol{Y}\right) \\ &\quad + \mu^2 \sum_{n=0}^{j} \rho\left(\boldsymbol{\Lambda}_i \boldsymbol{D}^n\right) \cdot \mathbb{E}\left\{ \boldsymbol{q}_{i+n} | \boldsymbol{\mathcal{F}}_i \right\} \\ &= \mu^2 \mathrm{Tr}\left(\boldsymbol{\Lambda}_i \left(\sum_{n=0}^{j} \boldsymbol{D}^n \right) \boldsymbol{Y} \right) \\ &\quad + \mu^2 \sum_{n=0}^{j} \rho\left(\boldsymbol{\Lambda}_i \boldsymbol{D}^n\right) \cdot \mathbb{E}\left\{ \boldsymbol{q}_{i+n} | \boldsymbol{\mathcal{F}}_i \right\} \end{aligned} \tag{7.125}$$

since $\overline{\boldsymbol{w}}'_{i+j+1} = \overline{\boldsymbol{w}}_i$ at $j = 0$. Our objective is to show that the first term on the right-hand side yields sufficient descent (i.e., will be sufficiently negative), while the second term is small enough to be negligible. To this end, we again make use of the structured eigendecomposi-

tion (7.109). We have:

$$
\begin{aligned}
&\mu^2 \mathrm{Tr}\left(\boldsymbol{\Lambda}_i \left(\sum_{n=0}^{j} \boldsymbol{D}^n\right) \boldsymbol{V}_i^{\mathsf{T}} R_s\left(\boldsymbol{w}_i\right) \boldsymbol{V}_i\right) \\
&\overset{(a)}{=} \mu^2 \mathrm{Tr}\Bigg(\boldsymbol{\Lambda}_i^{\geq 0} \left(\sum_{n=0}^{j} \left(I-2\mu \boldsymbol{\Lambda}_i^{\geq 0}\right)^n\right) \\
&\qquad \times \left(\boldsymbol{V}_i^{\geq 0}\right)^{\mathsf{T}} R_s\left(\boldsymbol{w}_i\right) \boldsymbol{V}_i^{\geq 0}\Bigg) \\
&\quad + \mu^2 \mathrm{Tr}\Bigg(\boldsymbol{\Lambda}_i^{<0} \left(\sum_{n=0}^{j} \left(I-2\mu \boldsymbol{\Lambda}_i^{<0}\right)^n\right) \\
&\qquad \times \left(\boldsymbol{V}_i^{<0}\right)^{\mathsf{T}} R_s\left(\boldsymbol{w}_i\right) \boldsymbol{V}_i^{<0}\Bigg) \\
&\overset{(b)}{=} \mu^2 \mathrm{Tr}\Bigg(\boldsymbol{\Lambda}_i^{\geq 0} \left(\sum_{n=0}^{j} \left(I-2\mu \boldsymbol{\Lambda}_i^{\geq 0}\right)^n\right) \\
&\qquad \times \left(\boldsymbol{V}_i^{\geq 0}\right)^{\mathsf{T}} R_s\left(\boldsymbol{w}_i\right) \boldsymbol{V}_i^{\geq 0}\Bigg) \\
&\quad - \mu^2 \mathrm{Tr}\Bigg(\left(-\boldsymbol{\Lambda}_i^{<0}\right) \left(\sum_{n=0}^{j} \left(I-2\mu \boldsymbol{\Lambda}_i^{<0}\right)^n\right) \\
&\qquad \times \left(\boldsymbol{V}_i^{<0}\right)^{\mathsf{T}} R_s\left(\boldsymbol{w}_i\right) \boldsymbol{V}_i^{<0}\Bigg) \\
&\overset{(c)}{\leq} \mu^2 \mathrm{Tr}\left(\boldsymbol{\Lambda}_i^{\geq 0} \left(\sum_{n=0}^{j} \left(I-2\mu \boldsymbol{\Lambda}_i^{\geq 0}\right)^n\right)\right) \\
&\quad \times \lambda_{\max}\left(\left(\boldsymbol{V}_i^{\geq 0}\right)^{\mathsf{T}} R_s\left(\boldsymbol{w}_i\right) \boldsymbol{V}_i^{\geq 0}\right) \\
&\quad - \mu^2 \mathrm{Tr}\left(\left(-\boldsymbol{\Lambda}_i^{<0}\right) \left(\sum_{n=0}^{j} \left(I-2\mu \boldsymbol{\Lambda}_i^{<0}\right)^n\right)\right) \\
&\quad \times \lambda_{\min}\left(\left(\boldsymbol{V}_i^{<0}\right)^{\mathsf{T}} R_s\left(\boldsymbol{w}_i\right) \boldsymbol{V}_i^{<0}\right) \\
&\overset{(d)}{\leq} \mu^2 \mathrm{Tr}\left(\boldsymbol{\Lambda}_i^{\geq 0} \left(\sum_{n=0}^{j} \left(I-2\mu \boldsymbol{\Lambda}_i^{\geq 0}\right)^n\right)\right)\left(\beta^2\|\nabla J(\boldsymbol{w}_i)\|^2+\sigma^2\right) \\
&\quad - \mu^2 \mathrm{Tr}\left(\left(-\boldsymbol{\Lambda}_i^{<0}\right) \left(\sum_{n=0}^{j} \left(I-2\mu \boldsymbol{\Lambda}_i^{<0}\right)^n\right)\right) \sigma_\ell^2
\end{aligned}
\tag{7.126}
$$

where in (a) we decomposed the trace since $\boldsymbol{\Lambda}_i \left(\sum_{n=0}^{j} \boldsymbol{D}^n\right)$ is a diagonal matrix, (b) applies

$-\left(-\boldsymbol{\Lambda}_i^{<0}\right) = \boldsymbol{\Lambda}_i^{<0}$. where in (a) we decomposed the trace since $\boldsymbol{\Lambda}_i \left(\sum_{n=0}^{j} \boldsymbol{D}^n\right)$ is a diagonal matrix and applied $-\left(-\boldsymbol{\Lambda}_i^{<0}\right) = \boldsymbol{\Lambda}_i^{<0}$. Step (b) follows from $\mathrm{Tr}(A)\lambda_{\min}(B) \leq \mathrm{Tr}(AB) \leq \mathrm{Tr}(A)\lambda_{\max}(B)$ which holds for $A = A^{\mathsf{T}}, B = B^{\mathsf{T}} \geq 0$, and (c) follows from the bounded covariance property (7.21) and Assumption 7.5.For the positive term, we have:

$$\begin{aligned}
&\mu^2 \mathrm{Tr}\left(\boldsymbol{\Lambda}_i^{\geq 0}\left(\sum_{n=0}^{j}\left(I - 2\mu\boldsymbol{\Lambda}_i^{\geq 0}\right)^n\right)\right)\left(\beta^2\|\nabla J(\boldsymbol{w}_i)\|^2 + \sigma^2\right) \\
&\overset{(a)}{\leq} \mu^2 \mathrm{Tr}\left(\boldsymbol{\Lambda}_i^{\geq 0}\left(\sum_{n=0}^{\infty}\left(I - 2\mu\boldsymbol{\Lambda}_i^{\geq 0}\right)^n\right)\right)\left(\beta^2\|\nabla J(\boldsymbol{w}_i)\|^2 + \sigma^2\right) \\
&\overset{(b)}{\leq} \mu^2 \mathrm{Tr}\left(\boldsymbol{\Lambda}_i^{\geq 0}\left(2\mu\boldsymbol{\Lambda}_i^{\geq 0}\right)^{-1}\right)\left(\beta^2\|\nabla J(\boldsymbol{w}_i)\|^2 + \sigma^2\right) \\
&\overset{(c)}{\leq} \frac{\mu}{2} M\left(\beta^2\|\nabla J(\boldsymbol{w}_i)\|^2 + \sigma^2\right)
\end{aligned} \tag{7.127}$$

where (a) follows since $I - 2\mu\boldsymbol{\Lambda}_i^{\geq 0}$ is elementwise non-negative for $\mu \leq \frac{2}{\delta}$, (b) follows from $\sum_{n=0}^{\infty} A^n = (I - A)^{-1}$ and (c) follows since $\nabla^2 J(\boldsymbol{w}_i)$ is of dimension M. Hence, under expectation:

$$\begin{aligned}
&\mu^2\, \mathbb{E}\left\{\mathrm{Tr}\left(\boldsymbol{\Lambda}_i^{\geq 0}\left(\sum_{n=0}^{j}\left(I - 2\mu\boldsymbol{\Lambda}_i^{\geq 0}\right)^n\right)\right)\right. \\
&\qquad\qquad \left.\times\left(\beta^2\|\nabla J(\boldsymbol{w}_i)\|^2 + \sigma^2\right) \,|\, \boldsymbol{w}_i \in \mathcal{H}\right\} \\
&\leq \frac{\mu}{2} M\left(\beta^2\, \mathbb{E}\left\{\|\nabla J(\boldsymbol{w}_i)\|^2 |\, \boldsymbol{w}_i \in \mathcal{H}\right\} + \sigma^2\right) \\
&\overset{(7.24)}{\leq} \frac{\mu}{2} M\left(\beta^2 \cdot O(\mu) + \sigma^2\right) = \frac{\mu}{2} M\sigma^2 + O(\mu^2)
\end{aligned} \tag{7.128}$$

For the negative term, we have under expectation conditioned on $\boldsymbol{w}_i \in \mathcal{H}$:

$$\begin{aligned}
&\mathbb{E}\left\{\mathrm{Tr}\left(\left(-\boldsymbol{\Lambda}_i^{<0}\right)\left(\sum_{n=0}^{j}\left(I-2\mu\boldsymbol{\Lambda}_i^{<0}\right)^n\right)\right)\sigma_\ell^2\,\middle|\,\boldsymbol{w}_i\in\mathcal{H}\right\}\\
&\overset{(a)}{\geq}\mathbb{E}\left\{\tau\left(\sum_{n=0}^{j}(1+2\mu\tau)^n\right)\sigma_\ell^2\,\middle|\,\boldsymbol{w}_i\in\mathcal{H}\right\}\\
&\overset{(b)}{=}\tau\left(\sum_{n=0}^{j}(1+2\mu\tau)^n\right)\sigma_\ell^2\overset{(c)}{=}\tau\frac{1-(1+2\mu\tau)^{i+1}}{1-(1+2\mu\tau)}\sigma_\ell^2\\
&=\frac{1}{2\mu}\left((1+2\mu\tau)^{j+1}-1\right)\sigma_\ell^2
\end{aligned}\tag{7.129}$$

Step (a) makes use of the fact that $\left(-\boldsymbol{\Lambda}_i^{<0}\right)\left(\sum_{n=0}^{j}\left(I-2\mu\boldsymbol{\Lambda}_i^{<0}\right)^n\right)$ is a diagonal matrix, where all elements are non-negative. Hence, its trace can be bounded by any of its diagonal elements:

$$\begin{aligned}
&\mathrm{Tr}\left(\left(-\boldsymbol{\Lambda}_i^{<0}\right)\left(\sum_{n=0}^{j}\left(I-2\mu\boldsymbol{\Lambda}_i^{<0}\right)^n\right)\right)\\
&\overset{(7.24)}{\geq}\tau\left(\sum_{n=0}^{j}(1+2\mu\tau)^n\right)
\end{aligned}\tag{7.130}$$

In (b) we dropped the expectation since the expression is no longer random, and (c) is the result of a geometric series. We return to the full expression (7.126) and find:

$$\begin{aligned}
&\mu^2\,\mathbb{E}\left\{\mathrm{Tr}\left(\boldsymbol{\Lambda}_i\left(\sum_{n=0}^{j}\boldsymbol{D}^n\right)\boldsymbol{V}_i^{\mathsf{T}}R_s\left(\boldsymbol{w}_i\right)\boldsymbol{V}_i\right)|\,\boldsymbol{w}_i\in\mathcal{H}\right\}\\
&\leq\frac{\mu}{2}M\sigma^2+O(\mu^2)-\frac{\mu}{2}\left((1+2\mu\tau)^{j+1}-1\right)\sigma_\ell^2\\
&\overset{(a)}{\leq}-\frac{\mu}{2}M\sigma^2
\end{aligned}\tag{7.131}$$

where (a) holds if, and only if,

$$\begin{aligned}
&\frac{\mu}{2}M\sigma^2+O(\mu^2)-\frac{\mu}{2}\left((1+2\mu\tau)^{j+1}-1\right)\sigma_\ell^2\leq-\frac{\mu}{2}M\sigma^2\\
\iff&2M\frac{\sigma^2}{\sigma_\ell^2}+O(\mu)+1\leq(1+2\mu\tau)^{j+1}\\
\iff&\log\left(2M\frac{\sigma^2}{\sigma_\ell^2}+1+O(\mu)\right)\leq(j+1)\log\left(1+2\mu\tau\right)\\
\iff&\frac{\log\left(2M\frac{\sigma^2}{\sigma_\ell^2}+1+O(\mu)\right)}{\log\left(1+2\mu\tau\right)}\leq j+1\\
\iff&\frac{\log\left(2M\frac{\sigma^2}{\sigma_\ell^2}+1+O(\mu)\right)}{O(\mu\tau)}\leq j+1
\end{aligned}\tag{7.132}$$

where the last line follows from $\lim_{x\to0}1/x\log(1+x)=1$. We conclude that there exists a bounded i^s such that:

$$\begin{aligned}
&\mu^2\,\mathbb{E}\left\{\mathrm{Tr}\left(\boldsymbol{\Lambda}_i\left(\sum_{n=0}^{i^s}\boldsymbol{D}^n\right)\boldsymbol{V}_i^{\mathsf{T}}R_s\left(\boldsymbol{w}_i\right)\boldsymbol{V}_i\right)\right\}\\
\leq&-\frac{\mu}{2}M\sigma^2
\end{aligned}\tag{7.133}$$

Applying this relation to (7.125) and taking expectations over $\boldsymbol{w}_i\in\mathcal{H}$, we obtain:

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^i_{i^s+1}\right\|^2_{\boldsymbol{\Lambda}_i}|\,\boldsymbol{w}_i\in\mathcal{H}\right\}\\
\leq&\,\mu^2\sum_{n=0}^{i^s}\mathbb{E}\left\{\left(\mathrm{Tr}\left(\boldsymbol{\Lambda}_i\boldsymbol{D}^n\right)\cdot\mathbb{E}\left\{\boldsymbol{q}_{i+n}|\boldsymbol{\mathcal{F}}_i\right\}\right)|\,\boldsymbol{w}_i\in\mathcal{H}\right\}\\
&-\frac{\mu}{2}M\sigma^2
\end{aligned}\tag{7.134}$$

We now bound the perturbation term:

$$
\begin{aligned}
&\mu^2 \sum_{n=0}^{i^s} \mathbb{E}\left\{\left(\rho\left(\boldsymbol{\Lambda}_i \boldsymbol{D}^n\right) \cdot \mathbb{E}\left\{\boldsymbol{q}_{i+n} | \boldsymbol{\mathcal{F}}_i\right\}\right) | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \\
&\leq \mu^2 \sum_{n=0}^{i^s} \mathbb{E}\left\{\left(\rho\left(\delta I (I + 2\mu\delta I)^n\right) \cdot \mathbb{E}\left\{\boldsymbol{q}_{i+n} | \boldsymbol{\mathcal{F}}_i\right\}\right) | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \\
&= \mu^2 \sum_{n=0}^{i^s} \left(\delta(1+2\mu\delta)^n \cdot \mathbb{E}\left\{\boldsymbol{q}_{i+n} | \, \boldsymbol{w}_i \in \mathcal{H}\right\}\right) \\
&\overset{(7.121)}{=} \mu^2 \sum_{n=0}^{i^s} \delta(1+2\mu\delta)^n \cdot \Bigg(\beta_R \, \mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}_j^i\right\|^\gamma | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \\
&\qquad\qquad + \delta^2 \, \mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^i_j\right\|^2 | \, \boldsymbol{w}_i \in \mathcal{H}\right\}\Bigg) \\
&\leq \mu^2 \sum_{n=0}^{i^s} \delta(1+2\mu\delta)^n \cdot \left(O(\mu^\gamma) + O(\mu^2)\right) \\
&\leq \delta \left(\sum_{n=0}^{i^s} (1+2\mu\delta)^n\right) O(\mu^{2+\gamma}) \\
&\overset{(a)}{\leq} O(\mu^{1+\gamma}) = \; o(\mu)
\end{aligned} \tag{7.135}
$$

where (a) follows from Lemma 7.2. We conclude:

$$
\mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^i_{i^s+1}\right\|^2_{\boldsymbol{\Lambda}_i} | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \leq -\frac{\mu}{2} M\sigma^2 + o(\mu) \tag{7.136}
$$

Returning to (7.115), we find:

$$
\begin{aligned}
&\mathbb{E}\left\{J(\boldsymbol{w}'_{i+j}) | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \\
&\leq \mathbb{E}\left\{J(\boldsymbol{w}_i) | \, \boldsymbol{w}_i \in \mathcal{H}\right\} + \frac{1}{2} \mathbb{E}\left\{\left\|\check{\boldsymbol{w}}'^i_j\right\|^2_{\boldsymbol{\Lambda}_i} | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \\
&\quad + \frac{\rho}{6} \mathbb{E}\left\{\left\|\widetilde{\boldsymbol{w}}'^i_j\right\|^3 | \, \boldsymbol{w}_i \in \mathcal{H}\right\} \\
&\leq \mathbb{E}\left\{J(\boldsymbol{w}_i) | \, \boldsymbol{w}_i \in \mathcal{H}\right\} - \frac{\mu}{2} M\sigma^2 + o(\mu)
\end{aligned} \tag{7.137}
$$

and with (7.92) we prove the result.

7.F Proof of Theorem 7.3

In a manner similar to [59], we define the stochastic process:

$$\mathbf{t}(k+1) = \begin{cases} \mathbf{t}(k)+1, & \text{if } \boldsymbol{w}_{\mathbf{t}(k)} \in \mathcal{G}, \\ \mathbf{t}(k)+1, & \text{if } \boldsymbol{w}_{\mathbf{t}(k)} \in \mathcal{M}, \\ \mathbf{t}(k)+i_s, & \text{if } \boldsymbol{w}_{\mathbf{t}(k)} \in \mathcal{H}. \end{cases} \tag{7.138}$$

where $\mathbf{t}(0) = 0$. From Theorem 7.1, we have:

$$\begin{aligned} &\mathbb{E}\left\{J(\boldsymbol{w}_{\mathbf{t}(k)}) - J(\boldsymbol{w}_{\mathbf{t}(k+1)})|\, \boldsymbol{w}_{\mathbf{t}(k)} \in \mathcal{G}\right\} \\ &= \mathbb{E}\left\{J(\boldsymbol{w}_{\mathbf{t}(k)}) - J(\boldsymbol{w}_{\mathbf{t}(k)+1})|\, \boldsymbol{w}_{\mathbf{t}(k)} \in \mathcal{G}\right\} \\ &\geq \mu^2 \frac{c_2}{\pi} \end{aligned} \tag{7.139}$$

and

$$\begin{aligned} &\mathbb{E}\left\{J(\boldsymbol{w}_{\mathbf{t}(k)}) - J(\boldsymbol{w}_{\mathbf{t}(k+1)})|\, \boldsymbol{w}_{\mathbf{t}(k)} \in \mathcal{M}\right\} \\ &= \mathbb{E}\left\{J(\boldsymbol{w}_{\mathbf{t}(k)}) - J(\boldsymbol{w}_{\mathbf{t}(k)+1})|\, \boldsymbol{w}_{\mathbf{t}(k)} \in \mathcal{M}\right\} \\ &\geq -\mu^2 c_2 \end{aligned} \tag{7.140}$$

while Theorem 7.2 ensures:

$$\begin{aligned} &\mathbb{E}\left\{J(\boldsymbol{w}_{\mathbf{t}(k)}) - J(\boldsymbol{w}_{\mathbf{t}(k+1)})|\, \boldsymbol{w}_{\mathbf{t}(k)} \in \mathcal{H}\right\} \\ &= \mathbb{E}\left\{J(\boldsymbol{w}_{\mathbf{t}(k)}) - J(\boldsymbol{w}_{\mathbf{t}(k)+i^s})|\, \boldsymbol{w}_{\mathbf{t}(k)} \in \mathcal{H}\right\} \\ &\geq \frac{\mu}{2} M\sigma^2 - o(\mu) \end{aligned} \tag{7.141}$$

Together, they yield:

$$\begin{aligned}
&\mathbb{E}\left\{J(\boldsymbol{w}_{\mathbf{t}(k)}) - \mathbb{E}\, J(\boldsymbol{w}_{\mathbf{t}(k+1)})\right\} \\
&= \mathbb{E}\left\{J(\boldsymbol{w}_{\mathbf{t}(k)}) - \mathbb{E}\, J(\boldsymbol{w}_{\mathbf{t}(k+1)}) | \, \boldsymbol{w}_{\mathbf{t}(k)} \in \mathcal{G}\right\} \cdot \pi_{\mathbf{t}(k)}^{\mathcal{G}} \\
&\quad + \mathbb{E}\left\{J(\boldsymbol{w}_{\mathbf{t}(k)}) - \mathbb{E}\, J(\boldsymbol{w}_{\mathbf{t}(k+1)}) | \, \boldsymbol{w}_{\mathbf{t}(k)} \in \mathcal{H}\right\} \cdot \pi_{\mathbf{t}(k)}^{\mathcal{H}} \\
&\quad + \mathbb{E}\left\{J(\boldsymbol{w}_{\mathbf{t}(k)}) - \mathbb{E}\, J(\boldsymbol{w}_{\mathbf{t}(k+1)}) | \, \boldsymbol{w}_{\mathbf{t}(k)} \in \mathcal{M}\right\} \cdot \pi_{\mathbf{t}(k)}^{\mathcal{M}} \\
&\geq \mu^2 \frac{c_2}{\pi} \cdot \pi_{\mathbf{t}(k)}^{\mathcal{G}} + \left(\frac{\mu}{2} M\sigma^2 - o(\mu)\right) \cdot \pi_{\mathbf{t}(k)}^{\mathcal{H}} - \mu^2 c_2 \cdot \pi_{\mathbf{t}(k)}^{\mathcal{M}}
\end{aligned} \tag{7.142}$$

Suppose $\pi_{\mathbf{t}(k)}^{\mathcal{M}} \leq 1 - \pi$ for all i. Then $\pi_{\mathbf{t}(k)}^{\mathcal{G}} + \pi_{\mathbf{t}(k)}^{\mathcal{H}} \geq \pi$ and

$$\begin{aligned}
&\mathbb{E}\left\{J(\boldsymbol{w}_{\mathbf{t}(k)}) - \mathbb{E}\, J(\boldsymbol{w}_{\mathbf{t}(k+1)})\right\} \\
&\geq \mu^2 \frac{c_2}{\pi} \cdot \left(\pi - \pi_{\mathbf{t}(k)}^{\mathcal{H}}\right) + \left(\frac{\mu}{2} M\sigma^2 - o(\mu)\right) \cdot \pi_{\mathbf{t}(k)}^{\mathcal{H}} \\
&\quad - \mu^2 c_2 \cdot (1 - \pi) \\
&= \mu^2 c_2 \pi + \left(\frac{\mu}{2} M\sigma^2 - \mu^2 \frac{c_2}{\pi} - o(\mu)\right) \pi_{\mathbf{t}(k)}^{\mathcal{H}} \\
&\overset{(a)}{\geq} \mu^2 c_2 \pi
\end{aligned} \tag{7.143}$$

where (a) holds whenever $\frac{\mu}{2} M\sigma^2 - \mu^2 \frac{c_2}{\pi} - o(\mu) \geq 0$, which holds whenever μ is sufficiently small. We hence have by telescoping:

$$\begin{aligned}
J(w_0) - J^o &\geq \mathbb{E}\, J(w_{\mathbf{t}(0)}) - \mathbb{E}\, J(\boldsymbol{w}_{\mathbf{t}(k)}) \\
&= \mathbb{E}\, J(w_{\mathbf{t}(0)}) - \mathbb{E}\, J(\boldsymbol{w}_{\mathbf{t}(1)}) \\
&\quad + \mathbb{E}\, J(\boldsymbol{w}_{\mathbf{t}(1)}) - \mathbb{E}\, J(\boldsymbol{w}_{\mathbf{t}(2)}) \\
&\quad + \cdots \\
&\quad + \mathbb{E}\, J(\boldsymbol{w}_{\mathbf{t}(k-1)}) - \mathbb{E}\, J(\boldsymbol{w}_{\mathbf{t}(k)}) \\
&\geq \mu^2 c_2 \pi k
\end{aligned} \tag{7.144}$$

Rearranging yields:

$$k \leq \frac{J(w_0) - J^o}{\mu^2 c_2 \pi} \tag{7.145}$$

We conclude by definition of the stochastic process $\boldsymbol{t}(k)$:

$$i = \boldsymbol{t}(k) \leq k \cdot i^s \leq \frac{(J(w_0) - J^o)}{\mu^2 c_2 \pi} i^s \tag{7.146}$$

CHAPTER 8

Graph Learning from Streaming Data

Graphs provide a powerful framework to represent high-dimensional but structured data, and to make inferences about relationships between subsets of the data. In this chapter we consider graph signals that evolve dynamically according to a heat diffusion process and are subject to persistent perturbations. We develop an online algorithm that is able to learn the underlying graph structure from observations of the signal evolution. The algorithm is adaptive in nature and in particular able to respond to changes in the graph structure and the perturbation statistics. The material in this chapter appeared in [67].

8.1 Related Works

The earliest works related to graph learning are based on *sparse* estimation of precision matrices, i.e., inverse covariance matrices [158, 159]. The work in [160] introduced structural constraints to ensure that the learned (regularized Laplacian) matrix describes a valid graph. A string of subsequent works [161–163] leverage the concept of a "smooth signal over a graph". The drawback of these approaches is that the smoothness assumption may not be satisfied in some important applications, particularly if the graph signal is dynamic or perturbed by events on the graph.

The interpretation of graph-shifts as a generalization of the traditional shift operation in digital signal processing has motivated a number of generalizations of DSP concepts to the graph domain. Autoregressive graph filters in terms of polynomials of the adjacency matrix are used in [164] to model the signal evolution over the graph and infer the adjacency matrix. The heat diffusion model is considered in [66], where an algorithm is proposed to leverage a

collection of *independent* samples which are modeled as the superposition of a small number of perturbations that diffuse over the graph.

Both of these recent works allow for dynamic signals that evolve according to some graph topology that is subsequently learned. This is achieved by collecting all available samples and solving an optimization problem based on a batch of data. As such, even though the model allows for dynamic signals, the algorithms themselves are not dynamic; the underlying assumption is that the model parameters are fixed. In contrast, in this work, we develop a truly adaptive solution that responds to streaming data and has the potential to track drifts in both the graph and data statistics under the heat diffusion model. Dynamic algorithms for the estimation of edge probabilities in social interactions are developed in [165, 166] and for autoregressive graph processes in [167].

8.2 Framework

8.2.1 Graph Model

We consider weighted, undirected graphs without self-loops. Every pair of vertices i and j is assigned a weight $a_{ij} = a_{ji}$, which quantifies their relative influence, in a manner made precise in the signal model further below. We collect these weights into an adjacency matrix $A = [a_{ij}]$ that satisfies the following properties:

$$\text{Symmetry: } A = A^{\mathsf{T}} \tag{8.1}$$

$$\text{Non-negativity: } a_{ij} \geq 0, \ \forall \ i, j \tag{8.2}$$

$$\text{No self-loops: } a_{ii} = 0, \ \forall \ i \tag{8.3}$$

A common and useful matrix to describe and study graphs is the Laplacian matrix, defined as:

$$L \triangleq \operatorname{diag}(A\mathbb{1}) - A \tag{8.4}$$

Under conditions (8.1)–(8.3) on the adjacency matrix, the graph Laplacian L satisfies the following properties [168]:

$$\text{Symmetry: } L = L^{\mathsf{T}} \tag{8.5}$$

$$\text{Non-positive off-diagonal elements: } \ell_{ij} \leq 0, \ \forall \ i \neq j \tag{8.6}$$

$$\text{Positive definite: } L \succeq 0 \tag{8.7}$$

$$\text{Nullspace: } L\frac{1}{\sqrt{N}}\mathbb{1} = 0 \tag{8.8}$$

8.2.2 Signal Model

We shall assume that we observe discrete samples of a continuous time graph process $s(t) \in \mathbb{R}^N$, which evolves according to the differential equation [64]:

$$s'(t) = -L^{\star}s(t) + p(t) \tag{8.9}$$

where $L^{\star} \in \mathbb{R}^{N \times N}$ denotes the Laplacian matrix of the underlying graph linking the entries of $s(t)$, and $p(t) \in \mathbb{R}^N$ describes a process that drives the signal dynamics. The variable $p(t)$ can either be viewed as an outside force, which influences the evolution of the signal, or some internal events that subsequently diffuse over the graph.

The homogeneous solution to

$$s_h'(t) = -L^{\star}s_h(t) \tag{8.10}$$

is given by

$$s_h(t) = e^{-tL^{\star}}s(0) \tag{8.11}$$

and the particular solution amounts to

$$s_p(t) = \int_0^t e^{-(t-u)L^{\star}}p(u)du \tag{8.12}$$

The solution to the differential equation (8.9) has the form:

$$s(t) = e^{-tL^\star} s(0) + \int_0^t e^{-(t-u)L^\star} p(u) du \tag{8.13}$$

Example 8.1 (Heat Diffusion with a Single Event)**.** *Assume that the system is initially at rest* ($s(0) = 0$) *and* $p(t) = p_1 \delta(t - t_1)$*. Then for* $t \geq t_1$*:*

$$s(t) = \int_0^t e^{-(t-u)L^\star} p_1 \delta(u - t_1) du = e^{-(t-t_1)L^\star} p_1 \tag{8.14}$$

Example 8.2 (Heat Diffusion with Multiple Events)**.** *Assume that the system is initially at rest* ($s(0) = 0$) *and* $p(t) = \sum_{k=1}^{K} p_k \delta(t - t_k)$*. Then for* $t \geq \max_k t_k$*:*

$$s(t) = \sum_{k=1}^{K} e^{-(t-t_k)L^\star} p_k \tag{8.15}$$

studied in [66].

We have access to the evolution of the graph signal beginning at some time t_0 and subsequently at times $t_i = t_0 + iT, i > 0$, where $i \in \mathbb{N}$ denotes the i-th sample and $T \in \mathbb{R}_{>0}$ denotes the sampling period. We observe a recursive relationship between adjacent samples, that is critical for this work, namely the fact that:

$$s(t_i) = e^{-TL^\star} s(t_{i-1}) + \int_{t_i - T}^{t_i} e^{-(t_i - u)L^\star} p(u) du \tag{8.16}$$

Note that the relationship between $s(t_i)$ and $s(t_{i-1})$ only depends on $L^\star$ and on the perturbations $p(t)$ between t_i and t_{i-1}. We move into the discrete domain by letting $s_i \triangleq s(t_0 + iT)$ and $p_i \triangleq \int_{t_i - T}^{t_i} e^{-(t_i - u)L^\star} p(u) du$ so that (8.16) becomes:

$$s_i = e^{-TL^\star} s_{i-1} + p_i \tag{8.17}$$

Since we are generally not provided with the perturbations that drive the system, we shall

model the driving term p_i as a stochastic random variable, so that:

$$\boldsymbol{s}_i = e^{-TL^\star}\,\boldsymbol{s}_{i-1} + \boldsymbol{p}_i \tag{8.18}$$

where we are using boldface notation to refer to random variables.

The objective of this work is to develop a solution that allows for the estimation of $L^\star$ from streaming realizations $\boldsymbol{s}_i$. These types of algorithms generally operate by evaluating the prediction error of the current estimate on the incoming observation and adjusting the estimate based on this error. Under the non-linear model (8.18), every such iteration requires the evaluation of a matrix exponential and is computationally expensive. This is particularly critical in scenarios where the graph size is large.

8.2.3 An Equivalent Linear Model

On the face of it, it is straightforward to define

$$W^\star \triangleq e^{-TL^\star} \tag{8.19}$$

so that the relation becomes

$$\boldsymbol{s}_i = W^\star\,\boldsymbol{s}_{i-1} + \boldsymbol{p}_i \tag{8.20}$$

However, it is important to remember that $L^\star$ is a Laplacian matrix and hence required to satisfy properties (8.5)–(8.8). It turns out that an equivalent set of properties can be imposed on $W^\star$ to ensure that $L^\star = \frac{-1}{T}\ln(W^\star)$ describes a valid Laplacian matrix and hence a valid graph. To begin with, we introduce the eigendecomposition of the Laplacian matrix:

$$L^\star = V\Lambda_L V^{\mathsf{T}} \tag{8.21}$$

Expanding the matrix exponential as an infinite sum and recalling that $VV^{\mathsf{T}} = I$, we obtain:

$$\begin{aligned} W^{\star} = e^{-TL^{\star}} &= \sum_{k=0}^{\infty} \frac{(-T)^k}{k!} (L^{\star})^k = \sum_{k=0}^{\infty} \frac{(-T)^k}{k!} \left(VL^{\star}\Lambda^{\mathsf{T}}\right)^k \\ &= V \left(\sum_{k=0}^{\infty} \frac{(-T)^k}{k!} \Lambda^k \right) V^{\mathsf{T}} = V e^{-T\Lambda} V^{\mathsf{T}} \end{aligned} \tag{8.22}$$

where $e^{-T\Lambda} = \operatorname{diag}\left\{e^{-T\lambda_k(L^{\star})}\right\}$ since Λ is diagonal. This means that the matrix exponential preserves the set of eigenvectors of $L^{\star}$ and there is a simple relationship between the eigenvalues of $W^{\star}$ and $L^{\star}$. This relation also provides a method for calculating the matrix logarithm. Given the eigendecomposition $W^{\star} = V\Lambda_W V^{\mathsf{T}}$, the logarithm is given by $\ln(W^{\star}) = V \ln(\Lambda_W) V^{\mathsf{T}}$, where $\ln(\Lambda_W) = \operatorname{diag}\{\ln(\lambda_k(W^{\star}))\}$. This allows us to establish the following conditions on $W^{\star}$ to ensure that $L^{\star}$ describes a valid graph.

Lemma 8.1 (Conditions on $W^{\star}$)**.** *Let $W \in \mathbb{R}^{N\times N}$ and $L = \frac{-1}{T}\ln(W)$. Then, L is a valid Graph Laplacian* if, and only if, *W satisfies the following properties:*

$$\textit{Symmetry: } W = W^{\mathsf{T}} \tag{8.23}$$

$$\textit{Non-negative elements: } w_{ij} \geq 0, \ \forall\ i,j \tag{8.24}$$

$$\textit{Spectral bound: } I \succeq W \succ 0 \tag{8.25}$$

$$\textit{Stochastic: } W\mathbb{1} = \mathbb{1} \tag{8.26}$$

Proof. Appendix 8.A. □

8.2.4 Graph Signal Evolution

Observe that since $\rho(W^{\star}) = 1$, the recursion described by (8.20) is not mean-square stable. This means that, while the recursion will converge in the mean as long as $\mathbb{E}\,\boldsymbol{p}_i = 0$, the same does not hold for covariance matrix of $\boldsymbol{s}_i$. It turn out, however, that the centered signal across the graph is mean-square stable as long as the graph is connected. We make this statement precise in the following.

Assumption 8.1 (Connected graph)**.** *The graph described by A and L is connected. In*

other words, there is a path of non-zero weights from any vertex to any other vertex in the graph. □

It then follows that the eigenvalue at zero has multiplicity one with unique (normalized) eigenvector $\frac{1}{\sqrt{N}}\mathbb{1}$ [168]. A direct consequence of this property is that the graph Laplacian has a particular eigenstructure $L^\star = V\Lambda_L V^\mathsf{T}$ where:

$$V = \begin{bmatrix} \frac{1}{\sqrt{N}}\mathbb{1} & \overline{V} \end{bmatrix}, \quad \Lambda_L = \begin{bmatrix} 0 & 0 \\ 0 & \overline{\Lambda}_L \end{bmatrix} \tag{8.27}$$

and critically $\overline{\Lambda}_L$ is *strictly* positive definite:

$$\overline{\Lambda}_L \succ 0 \tag{8.28}$$

The driving matrix $W^\star$ inherits a similar structure from $L^\star$ via (8.22). In particular, we have $W^\star = V\Lambda_W V^\mathsf{T}$, where:

$$V = \begin{bmatrix} \frac{1}{\sqrt{N}}\mathbb{1} & \overline{V} \end{bmatrix}, \quad \Lambda_W = \begin{bmatrix} 1 & 0 \\ 0 & \overline{\Lambda}_W \end{bmatrix} \tag{8.29}$$

and $\overline{\Lambda}_W = e^{-T\overline{\Lambda}_L}$, which due to (8.28) implies that

$$0 \prec \overline{\Lambda}_W \prec I \tag{8.30}$$

so that $\rho\left(\overline{\Lambda}_W\right) < 1$. The mean across the graph of the signal at time i is given by $\boldsymbol{s}_{c,i} = \frac{1}{N}\mathbb{1}^\mathsf{T}\boldsymbol{s}_i$. Subtracting this mean yields the centered graph signal $\overline{\boldsymbol{s}}_i$:

$$\overline{\boldsymbol{s}}_i \triangleq \boldsymbol{s}_i - \boldsymbol{s}_{c,i} = \left(I - \frac{1}{N}\mathbb{1}\mathbb{1}^\mathsf{T}\right)\boldsymbol{s}_i \tag{8.31}$$

It is important to recognize that the mean contains no information about the graph. This

is because for any doubly stochastic W:

$$W\,\boldsymbol{s}_i = W\,(\overline{\boldsymbol{s}}_i + \mathbb{1}\otimes \boldsymbol{s}_{c,i}) = W\overline{\boldsymbol{s}}_i + \mathbb{1}\otimes \boldsymbol{s}_{c,i} \tag{8.32}$$

In other words, the mean is passed through independently of W. For the evolution of the centered signal, we can now write:

$$\overline{\boldsymbol{s}}_i = \overline{W}^{\star}\overline{\boldsymbol{s}}_{i-1} + \overline{\boldsymbol{p}}_i \tag{8.33}$$

where we defined:

$$\overline{W}^{\star} \triangleq W^{\star} - \frac{1}{N}\mathbb{1}\mathbb{1}^{\mathsf{T}}, \quad \overline{\boldsymbol{p}}_i \triangleq \left(I - \frac{1}{N}\mathbb{1}\mathbb{1}^{\mathsf{T}}\right)\boldsymbol{p}_i \tag{8.34}$$

The eigendecomposition of $\overline{W}^{\star} = V\Lambda_{\overline{W}}V^{\mathsf{T}}$ is related to the decomposition of $W^{\star}$ via

$$V = \left[\begin{array}{cc} \frac{1}{\sqrt{N}}\mathbb{1} & \overline{V} \end{array}\right], \quad \Lambda_{\overline{W}} = \left[\begin{array}{cc} 0 & 0 \\ 0 & \overline{\Lambda}_W \end{array}\right] \tag{8.35}$$

so that the only change is the replacement of the eigenvalue at 1 by 0 and critically now $\rho\left(\overline{W}^{\star}\right) < 1$. We can examine in detail the evolution of the first and second-order statistics of $\overline{\boldsymbol{s}}_i$.

Assumption 8.2 (Statistics of the Perturbation Terms)**.** *The statistics of the centered perturbations $\overline{\boldsymbol{p}}_i = \left(I - \frac{1}{N}\mathbb{1}\mathbb{1}^{\mathsf{T}}\right)\boldsymbol{p}_i$ satisfy the following two conditions for all i:*

$$\mathbb{E}\,\overline{\boldsymbol{p}}_i = 0 \tag{8.36}$$

$$\mathbb{E}\,\overline{\boldsymbol{p}}_i\overline{\boldsymbol{p}}_i^{\mathsf{T}} = R_{\overline{\boldsymbol{p}}} < \infty \tag{8.37}$$

Furthermore, the perturbation $\overline{\boldsymbol{p}}_i$ at time i is independent of $\overline{\boldsymbol{p}}_{i-k}$ for $k > 0$. □

Lemma 8.2 (Signal evolution)**.** *Suppose the network is initially at rest, i.e., $s_0 = 0$ and denote $\mathbb{E}\,\overline{\boldsymbol{p}}_i\overline{\boldsymbol{p}}_i^{\mathsf{T}} = R_{\overline{\boldsymbol{p}}}$. Then, the first and second-order statistics of the graph process described*

by (8.18) *evolve according to:*

$$\mathbb{E}\,\overline{\boldsymbol{s}}_i = 0 \tag{8.38}$$

$$\mathbb{E}\,\overline{\boldsymbol{s}}_i\overline{\boldsymbol{s}}_i^{\mathsf{T}} \triangleq R_{\overline{\boldsymbol{s}}_i} = \sum_{k=0}^{i-1}\left(\overline{W}^{\star}\right)^{i-k} R_{\overline{\boldsymbol{p}}}\left(\overline{W}^{\star}\right)^{i-k} \tag{8.39}$$

Furthermore, the second-order moment converges and we have:

$$\lim_{i\to\infty} R_{\overline{\boldsymbol{s}}_i} \triangleq R_{\infty} \tag{8.40}$$

where R_{∞} is the solution to the discrete Lyapunov equation:

$$R_{\overline{\boldsymbol{p}}} = R_{\infty} - \overline{W}^{\star} R_{\infty} \overline{W}^{\star} \tag{8.41}$$

Proof. Appendix 8.B. □

To strengthen our intuition of this result, let us briefly consider the simplified case where $R_{\overline{\boldsymbol{p}}} = \sigma_p^2 I$. Then

$$R_{\infty} = \sigma_p^2 \sum_{k=0}^{\infty}\left(\overline{W}^{\star}\right)^{2k} = \sigma_p^2\left(I - \left(\overline{W}^{\star}\right)^2\right)^{-1} \tag{8.42}$$

If we consider the trace of the covariance matrix as notion of variation, we have

$$\begin{aligned}
\mathrm{Tr}\left(R_{\infty}\right) &= \sum_{k=1}^{N}\lambda_k\left(R_{\infty}\right) = \sum_{k=1}^{N}\frac{\sigma_p^2}{1-\lambda_k^2\left(\overline{W}^{\star}\right)} \\
&= \sigma_p^2 + \sum_{k=2}^{N}\frac{\sigma_p^2}{1-\lambda_k^2\left(\overline{W}^{\star}\right)} = \sigma_p^2 + \sum_{k=2}^{N}\frac{\sigma_p^2}{1-\lambda_k^2\left(W^{\star}\right)} \\
&= \sigma_p^2 + \sum_{k=2}^{N}\frac{\sigma_p^2}{1-e^{-2T\lambda_k(L^{\star})}}
\end{aligned} \tag{8.43}$$

Recall that T is the sampling clock of the system and note that that $\mathrm{Tr}\left(R_{\infty}\right)$ decreases to $N\sigma_p^2 = \mathrm{Tr}\left(R_{\overline{\boldsymbol{p}}}\right)$ as $T\lambda_k\left(L^{\star}\right) \to \infty\ \forall\ k$. This means that the variation in the system in

steady-state is determined by the product of the sampling clock and the eigenvalues of the Laplacian matrix. The non-zero eigenvalues of the Laplacian are a measure for how fast the graph mixes. In other words, to preserve variation in steady-state, a quickly mixing graph requires a small sampling period, while slowly mixing graphs allow for less frequent sampling.

8.3 Graph Learning

We now formulate the following optimization problem for learning $\overline{W}^{\star}$:

$$\overline{W}^{\star} = \arg\min_{\overline{W}\in\mathcal{C}} \frac{1}{2}\,\mathbb{E}\,\|\overline{\boldsymbol{s}}_i - \overline{W}\overline{\boldsymbol{s}}_{i-1}\|^2 \triangleq \arg\min_{\overline{W}\in\mathcal{C}} \mathbb{E}\,J_i\left(\overline{W}\right) \tag{8.44}$$

where $\mathcal{C}$ is a constraint-set. The cost $J_i(\cdot)$ depends on i because the statistics of $\boldsymbol{s}_{i-1}$ evolve as described in the previous lemma. A natural construction is to choose $\mathcal{C}$ to be the set of matrices that result in a valid Laplacian matrix. It turns out, however, that this is not necessary since $J_i(\overline{W})$ is strongly-convex.

Lemma 8.3 (Properties of the cost)**.** *The cost specified in* (8.44) *is Lipschitz continuous and strongly-convex. Specifically, for all $\overline{W} \in \mathbb{R}^{N\times N}$, we have:*

$$J_i(\overline{W}) \geq \frac{\nu_i}{2}\|\overline{W}^{\star} - \overline{W}\|^2 + \frac{1}{2}\mathrm{Tr}\left(R_{\overline{\boldsymbol{p}}}\right) \tag{8.45}$$

$$J_i(\overline{W}) \leq \frac{\delta_i}{2}\|\overline{W}^{\star} - \overline{W}\|^2 + \frac{1}{2}\mathrm{Tr}\left(R_{\overline{\boldsymbol{p}}}\right) \tag{8.46}$$

where

$$\delta_i = \lambda_{\max}\left(R_{\overline{\boldsymbol{s}}_{i-1}}\right),\ \nu_i = \lambda_{\min}\left(R_{\overline{\boldsymbol{s}}_{i-1}}\right) \tag{8.47}$$

Moreover, $\overline{W}^{\star}$ defined in (8.34) *is the unique minimizer of $J_i(\overline{W})$ for all i.*

Proof. Appendix 8.C. □

It follows from this property that the enforcement of properties of $\overline{W}^{\star}$ is in fact not necessary when designing algorithms for the solution of (8.44), since any algorithm that converges to

a minimizer of (8.44) will converge to its unique minimizer, $\overline{W}^{\star}$, which by definition already satisfies all properties that lead to a valid graph Laplacian. Of course, it is reasonable to believe that the addition of constraints and regularization may lead to an increased rate of convergence and/or improved performance in steady-state at the cost of increased computational cost per iteration.

To begin with, we shall pursue the minimizer of (8.44) in the absence of constraints by means of a stochastic gradient descent algorithm.

Algorithm 8.1 Laplacian LMS Strategy

$$\overline{\boldsymbol{W}}_i = \overline{\boldsymbol{W}}_{i-1} + \mu \left(\overline{\boldsymbol{s}}_i - \overline{\boldsymbol{W}}_{i-1} \, \overline{\boldsymbol{s}}_{i-1} \right) \overline{\boldsymbol{s}}_{i-1}^{\mathsf{T}} \tag{8.48}$$

It is essentially a matrix valued variation of the least-mean squares (LMS) algorithm. To derive approximate expressions for its performance, we shall adopt an assumption on the step-size μ, which is common in the literature [149].

Assumption 8.3 (Small step-size and independence)**.** *Assume the step-size μ is sufficiently small, so that in the limit, $\|\overline{W}^{\star} - \overline{\boldsymbol{W}}_i\|^2$ reaches a steady-state distribution and $\overline{W}^{\star} - \overline{\boldsymbol{W}}_i$ is independent of $\overline{\boldsymbol{s}}_i$.*

Theorem 8.1 (Performance for small step-sizes)**.** *Under Assumption 8.3, the mean-square deviation of the estimate from the true minimizer $\overline{W}^{\star}$ is given by:*

$$\lim_{i \to \infty} \mathbb{E} \left\| \overline{W}^{\star} - \overline{\boldsymbol{W}}_i \right\|^2 \approx \mu \frac{N \mathrm{Tr}\left(R_{\overline{\boldsymbol{p}}}\right)}{2} \tag{8.49}$$

Proof. The proof is essentially the same as the one for the traditional LMS algorithm [149]. □

Performance of the algorithm can be improved by including projections in the update relation. Recall that W is obtained from $\overline{W}$ via $W = \overline{W} + \frac{1}{N} \mathbb{1}\mathbb{1}^{\mathsf{T}}$. This means that a necessary

condition for the properties from Lemma 8.1 to be satisfied is:

$$\begin{aligned}
&\overline{\boldsymbol{W}}_i \in \mathcal{C}_{\text{ele}} \cup \mathcal{C}_{\text{sym}} \cup \mathcal{C}_{\text{null}} \cup \mathcal{C}_{\text{spec}} \qquad (8.50)\\
&\mathcal{C}_{\text{ele}} \triangleq \left\{ \overline{W} \middle| \overline{w}_{ij} \geq -\frac{1}{N} \right\} \qquad \mathcal{C}_{\text{sym}} \triangleq \left\{ \overline{W} \middle| \overline{W} = \overline{W}^{\mathsf{T}} \right\}\\
&\mathcal{C}_{\text{null}} \triangleq \left\{ \overline{W} \middle| \overline{W}\mathbb{1} = 0 \right\} \qquad \mathcal{C}_{\text{spec}} \triangleq \left\{ \overline{W} \middle| 0 \preceq \overline{W} \preceq I \right\}
\end{aligned}$$

Projections onto each of these sets can be evaluated in closed form:

$$\left[\text{Proj}_{\mathcal{C}_{\text{ele}}}\left(\overline{W}\right)\right]_{ij} = \begin{cases} \overline{w}_{ij}, & \text{if } \overline{w}_{ij} \geq -\frac{1}{N} \\ -\frac{1}{N}, & \text{otherwise} \end{cases} \qquad (8.51)$$

$$\text{Proj}_{\mathcal{C}_{\text{sym}}}\left(\overline{W}\right) = \frac{1}{2}\left(\overline{W} + \overline{W}^{\mathsf{T}}\right) \qquad (8.52)$$

$$\text{Proj}_{\mathcal{C}_{\text{null}}}\left(\overline{W}\right) = \overline{W} - \frac{1}{N}\overline{W}\mathbb{1}\mathbb{1}^{\mathsf{T}} \qquad (8.53)$$

$$\text{Proj}_{\mathcal{C}_{\text{spec}}}\left(\overline{W}\right) = V\Lambda_t V^{\mathsf{T}} \qquad (8.54)$$

where the last projection is given in terms of the eigendecomposition of the argument $\overline{W} = V\Lambda V^{\mathsf{T}}$ by thresholding the eigenvalues:

$$[\Lambda_t]_{ii} = \begin{cases} 0, & \text{if } [\Lambda]_{ii} < 0 \\ [\Lambda]_{ii}, & \text{if } 0 \leq [\Lambda]_{ii} \leq 1 \\ 1, & \text{otherwise} \end{cases} \qquad (8.55)$$

We can now interlace these projections with the stochastic gradient update to obtain two algorithms, which explicitly incorporate the structural constraints. Note that the first three projections (8.51)–(8.53) are simple in the sense that they require $O(N^2)$ operations where N is the size of the graph, whereas (8.54) requires a full eigenvalue decomposition. Hence, we can formulate two projected variants of the algorithm. The Type I implementation only enforces simple projections, while Type II enforces all properties.

Algorithm 8.2 Projected Laplacian LMS Strategy I and II

$$\overline{\boldsymbol{W}}_i' = \overline{\boldsymbol{W}}_{i-1} + \mu \left(\overline{\boldsymbol{s}}_i - \overline{\boldsymbol{W}}_{i-1} \, \overline{\boldsymbol{s}}_{i-1} \right) \overline{\boldsymbol{s}}_{i-1}^{\mathsf{T}} \tag{8.56}$$

$$\overline{\boldsymbol{W}}_i'' = \mathrm{Proj}_{\mathcal{C}_{\mathrm{sym}}} \left(\mathrm{Proj}_{\mathcal{C}_{\mathrm{null}}} \left(\mathrm{Proj}_{\mathcal{C}_{\mathrm{ele}}} \left(\overline{\boldsymbol{W}}_i' \right) \right) \right) \tag{8.57}$$

$$\overline{\boldsymbol{W}}_i = \begin{cases} \overline{\boldsymbol{W}}_i'', & \text{for Type I} \\ \mathrm{Proj}_{\mathcal{C}_{\mathrm{spec}}} \left(\overline{\boldsymbol{W}}_i'' \right), & \text{for Type II.} \end{cases} \tag{8.58}$$

Whenever an estimate of the graph Laplacian is required, it is obtained via:

$$\widehat{L} = \frac{-1}{T} \ln \left(\overline{\boldsymbol{W}}_i \right) \tag{8.59}$$

8.4 Simulation Results

We illustrate the performance of the algorithm in recovering $\overline{W}^{\star}$ as well as the graph structure on a network with $N = 30$ nodes. The perturbation terms are modeled as following a normal distribution with $\boldsymbol{p}_i \sim \mathcal{N}(0, I)$ and the sampling period is $T = 1$. The observations $\boldsymbol{s}_i$ are generated according to (8.18) and processed according to the algorithms developed in this work. The true graphs is generated using the Barabasi-Albert model [169], upon which random weights between 0.1 and 1.0 are attached to each non-zero edge. After $500,000$ iterations, there is a sudden change in the network topology, to illustrate the ability of the algorithm to adapt. The second graph and its adjacency matrix are depicted in Fig. 8.1–8.2. In the graph representation, small weights are depicted as thin and light lines, while strong weights are dark and thick. Bright colors in the adjacency matrix correspond to large weights.

The recovered graph and adjacency matrix at the final iteration using Algorithm 2 Type I are depicted in Fig. 8.3–8.4. Color and weight maps are the same as in the representation of the true graph. Key connections along with their weights and the general structure of the graph are accurately recovered. Note that no weights are truly set to zero, resulting in a number of low-weight connections. This is due to the fact that no sparsity prior was

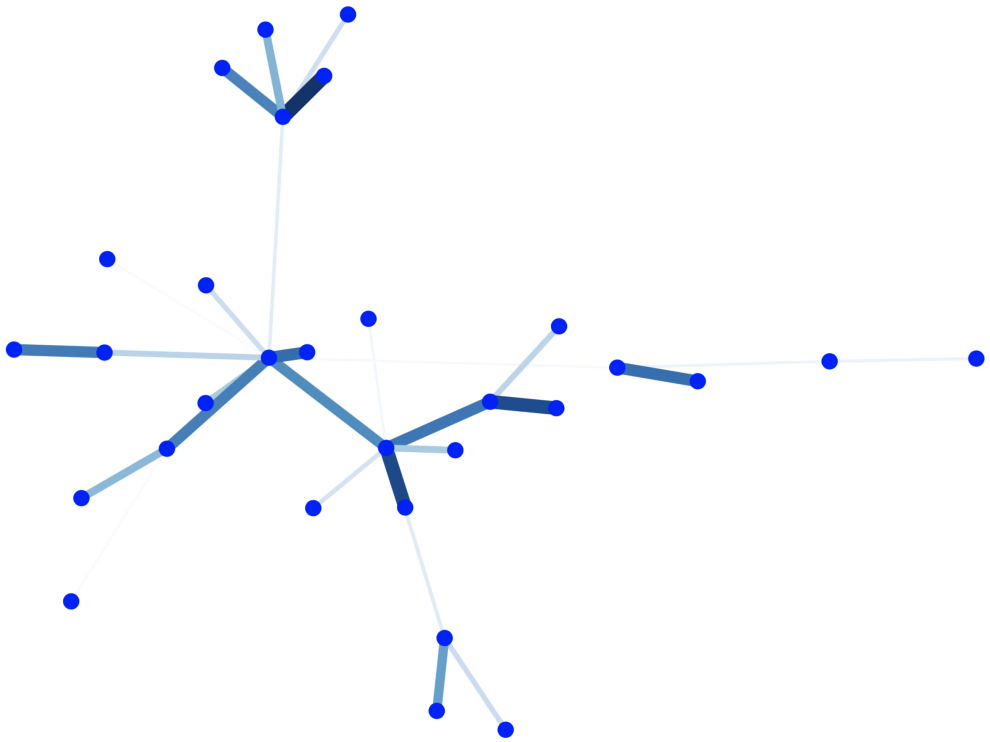

Figure 8.1: True graph.

imposed on the weight matrix. If desired, they can be removed during post-processing via simple thresholding.

The mean-square deviation of $\overline{\boldsymbol{W}}_i$ from $\overline{W}^\star$ is depicted in Fig. 8.5. All methods converge in the mean-square sense to a region around the true minimizer. The theoretical expression (8.49) accurately predicts the performance of the projection-free algorithm, while adding projections improves performance. Observe that notably, in this scenario, the addition of the spectral constraint to the simple constraints yields a negligible improvement, as both learning curves overlap.

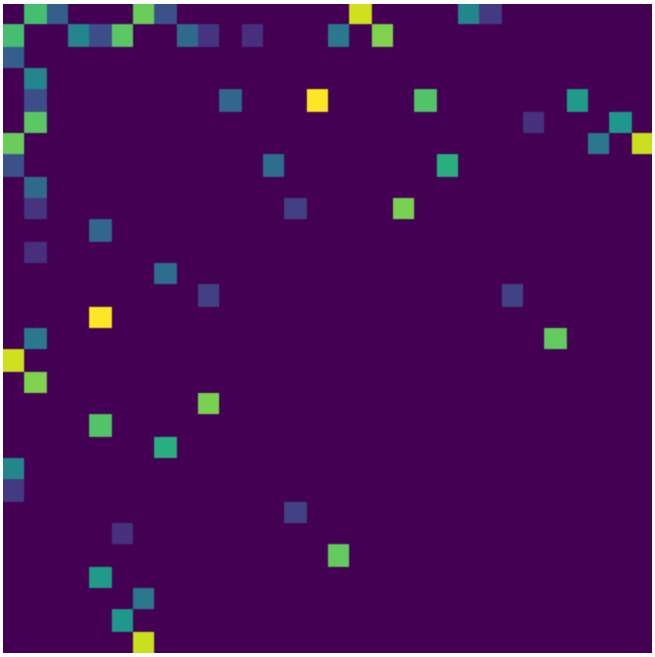

Figure 8.2: True adjacency matrix.

8.A Proof of Lemma 8.1

Equivalence between the symmetry relations (8.5) and (8.23) follows immediately from (8.22). The same goes for the spectral bounds (8.7) and (8.25), after noting that

$$\lambda_k(L) \geq 0 \iff 1 \geq e^{-T\lambda_k(L)} > 0, \ \forall \ k \tag{8.60}$$

Equivalence between the nullspace condition (8.8) and stochasticity (8.26) follows again from (8.22). Specifically, we know from (8.22) that the matrix exponential preserves the eigenvectors and maps the eigenvalues according to:

$$\lambda_k\left(W\right) = e^{-T\lambda_k(L)} \tag{8.61}$$

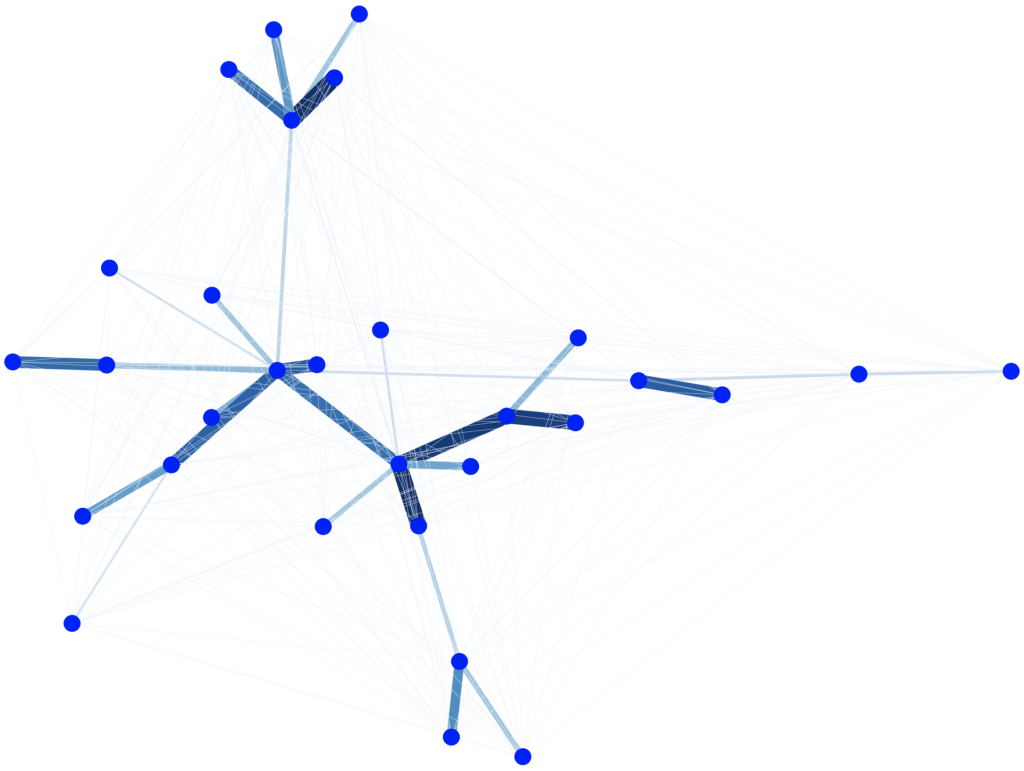

Figure 8.3: Graph recovered using the Projected Laplacian LMS Strategy I.

The nullspace condition (8.8) states that $\frac{1}{\sqrt{N}}\mathbb{1}$ is an eigenvector for L with eigenvalue 0, which is equivalent to the statement that $\frac{1}{\sqrt{N}}\mathbb{1}$ is an eigenvector for W with eigenvalue $e^{-T\cdot 0} = 1$.

To establish the equivalence between the non-positivity constraint (8.6) and the non-negativity constraint (8.24), note that condition (8.6) ensures that L has non-positive off-diagonal elements, which implies that $-TL$ has non-negative off-diagonal elements. Such matrices, known as "Metzler" matrices, and the corresponding matrix exponentials appear frequently in the study of positive linear systems [170]. In particular, it has been shown that $e^{T(-L)}$ has positive elements, if and only if $(-L)$ is a Metzler matrix [171, Example 1.4.b], which is our desired equivalence.

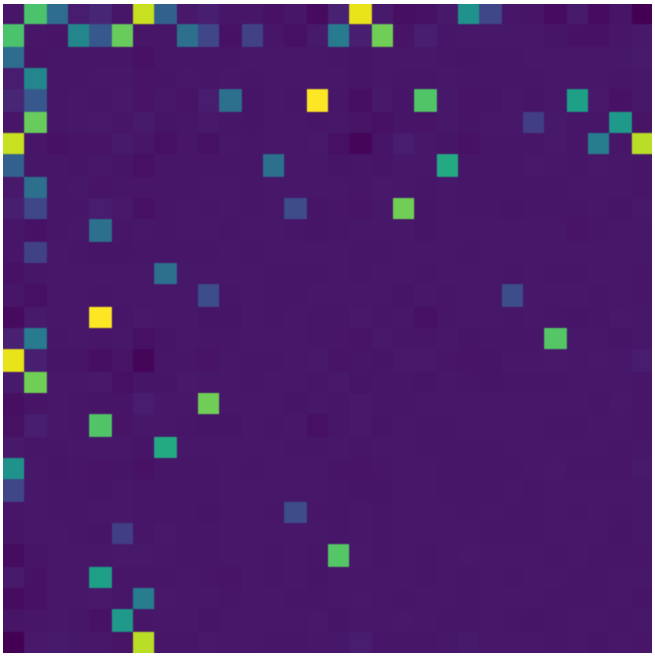

Figure 8.4: Adjacency matrix recovered using the Projected Laplacian LMS Strategy I.

8.B Proof of Lemma 8.2

Iterating (8.33), we have

$$\overline{\boldsymbol{s}_i} = \left(\overline{W}^{\star}\right)^{i} \overline{s}_0 + \sum_{k=1}^{i} \left(\overline{W}^{\star}\right)^{i-k} \overline{\boldsymbol{p}}_k \tag{8.62}$$

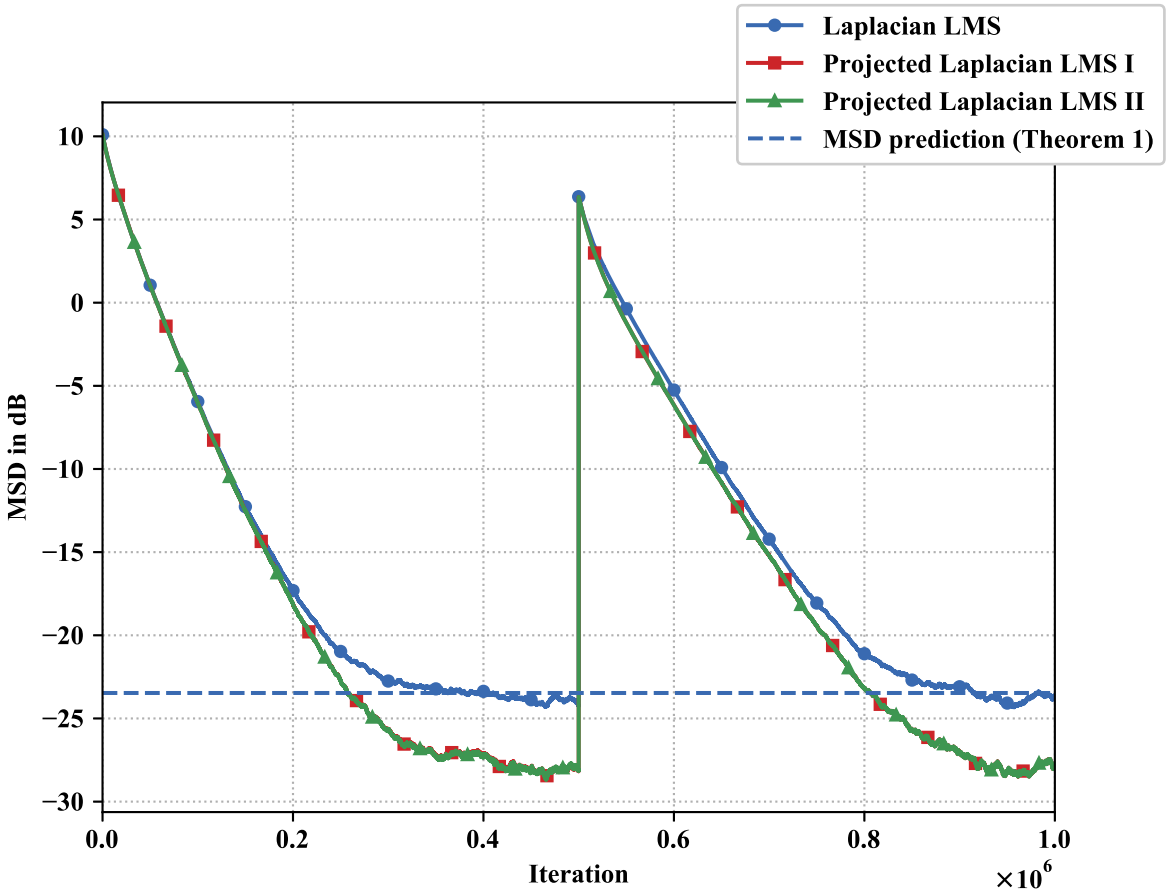

Figure 8.5: Mean-Square Deviation.

Taking expectation and noting that $\overline{s}_0 = 0$ and $\mathbb{E}\,\overline{p}_i = 0$, we obtain the zero-mean relation. The second-order relation follows from:

$$\begin{aligned}
\mathbb{E}\,\overline{\boldsymbol{s}_i \boldsymbol{s}_i}^{\mathsf{T}} &= \mathbb{E}\left(\sum_{k=1}^{i}\left(\overline{W}^{\star}\right)^{i-k}\overline{\boldsymbol{p}}_k\right)\left(\sum_{k=1}^{i}\left(\overline{W}^{\star}\right)^{i-k}\overline{\boldsymbol{p}}_k\right)^{\mathsf{T}} \\
&= \mathbb{E}\sum_{k=1}^{i}\left(\overline{W}^{\star}\right)^{i-k}\overline{\boldsymbol{p}}_k\overline{\boldsymbol{p}}_k^{\mathsf{T}}\left(\overline{W}^{\star}\right)^{i-k} \\
&= \sum_{k=1}^{i}\left(\overline{W}^{\star}\right)^{i-k}R_{\overline{\boldsymbol{p}}}\left(\overline{W}^{\star}\right)^{i-k} \\
&= \sum_{k=0}^{i-1}\left(\overline{W}^{\star}\right)^{k}R_{\overline{\boldsymbol{p}}}\left(\overline{W}^{\star}\right)^{k} \qquad (8.63)
\end{aligned}$$

This sum appears frequently in control theory. It converges if $\rho\left(\overline{W}^{\star}\right) < 1$, in which case [172]

$$\lim_{i\to\infty} \mathbb{E}\,\overline{\boldsymbol{s}}_i\overline{\boldsymbol{s}}_i^{\mathsf{T}} = \sum_{k=0}^{\infty}\left(\overline{W}^{\star}\right)^k R_{\overline{\boldsymbol{p}}}\left(\overline{W}^{\star}\right)^k \triangleq R_{\infty} \tag{8.64}$$

where R_∞ is the solution to the Lyapunov equation

$$\overline{W}^{\star} R_{\infty} \overline{W}^{\star} = R_{\infty} - R_{\overline{\boldsymbol{p}}} \tag{8.65}$$

8.C Proof of Lemma 8.3

We first establish that $\overline{W}^{\star}$ is in fact a minimizer of $J_i\left(\overline{W}\right)$. Its gradient relative to $\overline{W}$ is given by:

$$\nabla J_i(\overline{W}) = -\,\mathbb{E}\left(\overline{\boldsymbol{s}}_i - \overline{W}\overline{\boldsymbol{s}}_{i-1}\right)\overline{\boldsymbol{s}}_{i-1}^{\mathsf{T}} \tag{8.66}$$

Evaluated at $\overline{W}^{\star}$, we have:

$$\begin{aligned}\nabla J_i(\overline{W}^{\star}) &= -\,\mathbb{E}\left(\overline{\boldsymbol{s}}_i - \overline{W}^{\star}\overline{\boldsymbol{s}}_{i-1}\right)\overline{\boldsymbol{s}}_{i-1}^{\mathsf{T}} \\ &= -\,\mathbb{E}\left(\overline{\boldsymbol{s}}_i - \overline{\boldsymbol{s}}_i - \overline{\boldsymbol{p}}_i\right)\overline{\boldsymbol{s}}_{i-1}^{\mathsf{T}} \\ &= 0\end{aligned} \tag{8.67}$$

so that $\overline{W}^{\star}$ is indeed a minimizer of $J_i(\overline{W})$ for all i.

Now, we can write:

$$
\begin{aligned}
J_i(W) &= \frac{1}{2}\,\mathbb{E}\,\|\,\boldsymbol{s}_i - W\,\boldsymbol{s}_{i-1}\,\|^2 \\
&= \frac{1}{2}\,\mathbb{E}\,\|W^\star\,\boldsymbol{s}_{i-1} + \boldsymbol{p}_i - W\,\boldsymbol{s}_{i-1}\,\|^2 \\
&= \frac{1}{2}\,\mathbb{E}\,\|\,(W^\star - W)\,\boldsymbol{s}_{i-1}\,\|^2 + \frac{1}{2}\,\mathbb{E}\,\|\,\boldsymbol{p}_i\,\|^2 \\
&= \frac{1}{2}\,\mathbb{E}\,\mathrm{Tr}\left(\boldsymbol{s}_{i-1}^{\mathsf{T}}\,\widetilde{W}^{\mathsf{T}}\widetilde{W}\,\boldsymbol{s}_{i-1}\right) + \frac{1}{2}\mathrm{Tr}\,(R_p) \\
&= \frac{1}{2}\,\mathbb{E}\,\mathrm{Tr}\left(\widetilde{W}^{\mathsf{T}}\widetilde{W}\,\boldsymbol{s}_{i-1}\,\boldsymbol{s}_{i-1}^{\mathsf{T}}\right) + \frac{1}{2}\mathrm{Tr}\,(R_p) \\
&= \frac{1}{2}\mathrm{Tr}\left(\widetilde{W}R_{\boldsymbol{s},i-1}\widetilde{W}^{\mathsf{T}}\right) + \frac{1}{2}\mathrm{Tr}\,(R_p) \\
&= \frac{1}{2}\mathrm{Tr}\left(\widetilde{W}V_s\Lambda_s V_s^{\mathsf{T}}\widetilde{W}^{\mathsf{T}}\right) + \frac{1}{2}\mathrm{Tr}\,(R_p) \\
&= \frac{1}{2}\mathrm{Tr}\left(\overline{W}^{\mathsf{T}}\Lambda_s\overline{W}\right) + \frac{1}{2}\mathrm{Tr}\,(R_p) \\
&\geq \frac{1}{2}\lambda_{\min}\left(R_{\boldsymbol{s}_{i-1}}\right)\mathrm{Tr}\left(\overline{W}^{\mathsf{T}}\overline{W}\right) + \frac{1}{2}\mathrm{Tr}\,(R_p) \\
&= \frac{1}{2}\lambda_{\min}\left(R_{\boldsymbol{s}_{i-1}}\right)\left\|\overline{W}\right\|^2 + \frac{1}{2}\mathrm{Tr}\,(R_p) \\
&= \frac{1}{2}\lambda_{\min}\left(R_{\boldsymbol{s}_{i-1}}\right)\left\|\widetilde{W}\right\|^2 + \frac{1}{2}\mathrm{Tr}\,(R_p)
\end{aligned}
\tag{8.68}
$$

The upper bound yielding the Lipschitz constant follows analogously.

CHAPTER 9

Interpretative Learning via the BRAIN strategy

The material in this chapter appeared in [73].

9.1 Introduction

Given feature vectors $\boldsymbol{h} \in \mathbb{R}^M$ and binary class labels $\boldsymbol{\gamma} \in \{\pm 1\}$, the broad objective of learning solutions is to seek classifiers $c(\boldsymbol{h})$ from the set $\mathcal{C}$ that solve [1, 86, 173, 174]:

$$c^{\star}(\boldsymbol{h}) = \arg\min_{c(\cdot)\in\mathcal{C}} \mathbb{P}\{c(\boldsymbol{h}) \neq \boldsymbol{\gamma}\} \tag{9.1}$$

The exact solution of (9.1) is generally intractable, mainly because it requires knowledge of the joint probability distribution of the feature and class variables. It is customary to replace the cost function by some regularized convex risk function and to seek instead the classifier, $c^{o}(\boldsymbol{h})$, that minimizes:

$$c^{o}(\boldsymbol{h}) = \arg\min_{c(\cdot)\in\mathcal{C}} \mathbb{E}\, Q\big(c(\cdot)\,;\, \boldsymbol{h}, \boldsymbol{\gamma}\big) + R\big(c(\cdot)\big) \tag{9.2}$$

In this formulation, the term $R(c)$ denotes a regularizer intended to endow $c^{o}(\boldsymbol{h})$ with useful properties (such as sparsity), and $Q(\cdot)$ is a loss function. Under the assumption that the stochastic process generating realizations $\{h_n, \gamma(n)\}_{n=0}^{N-1}$ is ergodic, the mean in (9.2) can be approximated by its sample average, resulting in an empirical risk minimization problem directly in terms of the training data:

$$c^{o}(\boldsymbol{h}) \triangleq \arg\min_{c(\cdot)\in\mathcal{C}} \frac{1}{N}\sum_{n=0}^{N-1} Q\big(c(\cdot); h_n, \gamma(n)\big) + R\big(c(\cdot)\big) \tag{9.3}$$

Various machine learning algorithms are derived from this perspective. Examples include logistic regression [86], support-vector machines [175, 176], as well as neural networks [177, 178] and deep neural networks [179, 180]. When the number of available samples N is much larger than the dimension of the feature space M and the VC dimension of the classifier set $\mathcal{C}$, it follows from the Vapnik-Chervonenkis theory [175] that the solution of (9.3) will result in a classifier with good generalization ability. This property refers to the fact that, although the classifier has been trained on a finite number of training data, it will still perform reasonably well on unseen data.

On the other hand, it is well known that when the number of samples N available for training is limited, appropriate prevention of overfitting becomes necessary. This scenario is common in cases where data collection is expensive, for example in biomedical applications, or when there is lack of information about the nature of the features, resulting in the collection of high dimensional feature data. Among the most commonly used remedies for overfitting are dimensionality reduction, feature selection, and regularization, all of which effectively reduce the complexity of the classifier set. However, several useful methods for dimensionality reduction, such as principal component analysis [173, 181] or Fisher discriminant analysis [182, 183], can still suffer from overfitting for small sample sizes.

In this work, we propose a framework for classification that involves an adaptive "soft" feature selection mechanism involving a graph topology that is also learned and tuned online during the same training process. The proposed framework is motivated by the observation that many feature spaces in practice include an implicit structure that may be learned and exploited for enhanced classification performance. The graph topology is used for this purpose; its role is to learn and track correlations among feature subspaces over time, and this information is fed into the learning algorithm in real-time. By doing so, the resulting learning mechanism reduces the complexity of the classifier and combats overfitting. Once trained, one prominent feature of the proposed solution is that it provides an "x-ray" view into the correlation structure of the feature space, offering an opportunity for iterative refinement of the features.

Figure 9.1 provides a high level overview of the proposed architecture; it includes elements

that are meant to mimic processing in the brain. The block with dictionary learning agents plays the role of a local memory that learns and stores foundational atoms (or basis) for the representation of feature subspaces. The block with the graph topology plays the role of interconnections that are also learned from correlations among the feature subspaces. Thus, while traditional learning algorithms focus on learning a mapping from the feature space to the class label, the proposed learning strategy slices the feature space into subspaces and incorporates learned memory and correlation graphs. We refer to the architecture in the figure as the BRAIN strategy, where the acronym stands for **B**lock-**R**educed **A**daptation and **I**nference from **N**etworked subspaces. Due to space limitations, in this article, however, we do not study the BRAIN structure in its generality. As a proof of concept, we shall ignore the dictionary blocks (i.e., we let the basis be the feature vectors themselves) and illustrate the enhancement that already results from exploiting the graphical correlation information alone.

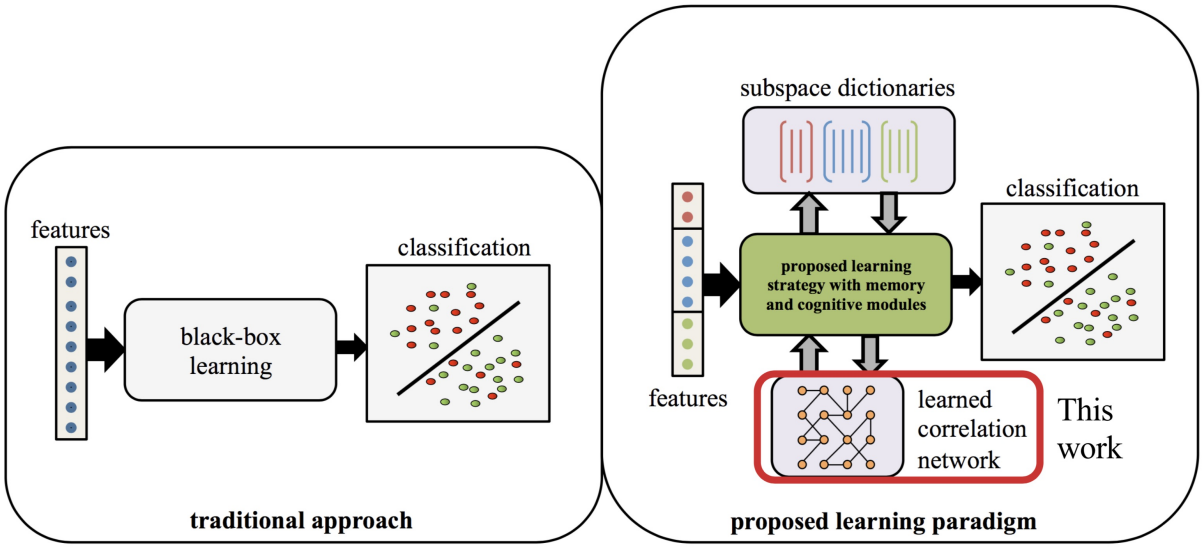

Figure 9.1: (*left*) Traditional learning paradigm. (*right*) The BRAIN strategy with dictionary and correlation networks.

9.1.1 Relation to other works

Graphs have been used before as a useful tool to encode dependency among random variables, as happens, for example, in Bayesian and Markov networks [184, 185]. These structures are appropriate when there is a fundamental understanding of how the variables relate to each other. For the case when this information is not available, algorithms with latent variables, such as expectation-maximization algorithms [173,186], restricted Boltzmann machines [187], or deep belief networks [188] are generally employed. One drawback of such architectures is that, while powerful when trained with sufficient amounts of training data, the populated hidden layers are not always interpretable. In contrast, given only the structure of the feature space and no information about correlation, our solution attaches a single correlation layer to the shallow learner. Unlike deep strategies, this layer does not play a role in the actual classification decision, but rather learns and tracks low-variability representations of the feature space. Moreover, this layer does not operate directly on the feature data but rather on scalar score variables defined in (9.7). These steps enable the algorithm to more accurately learn the subset of informative features and after convergence provides an insight into the correlation structure that resides in the feature space with respect to the classification decision.

9.2 Algorithm Formulation

Consider a large feature vector $\boldsymbol{h} \in \mathbb{R}^M$, which can be divided into a collection of sub-vectors $\boldsymbol{h}_k \in \mathbb{R}^{M_k}, k = 1, \ldots, K$, so that $\boldsymbol{h} = \text{col}\{\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_K\}$. These collections are application specific and can, for example, correspond to different bands in a densely sampled spectrogram, different performance metrics for a sector of the economy, or different regions of the human genome. In this work, we consider linear classifiers of the form:

$$c(\boldsymbol{h}) \triangleq \text{sign}(w^{\mathsf{T}} \boldsymbol{h}) \tag{9.4}$$

which can be decomposed under the assumed structure for the feature space into

$$c(\boldsymbol{h}) = \text{sign}\left(\sum_{k=1}^{K} w_k^{\mathsf{T}}\, \boldsymbol{h}_k\right) \tag{9.5}$$

where $w_k \in \mathbb{R}^{M_k}$ is the sub-vector of the linear classifier associated with $\boldsymbol{h}_k$, i.e.,

$$w \triangleq \text{col}\left\{w_1, w_2, \ldots, w_K\right\}, \quad \boldsymbol{h} \triangleq \text{col}\left\{\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_K\right\} \tag{9.6}$$

Traditional methods for dimensionality reduction operate directly on $\boldsymbol{h} \in \mathbb{R}^M$. They include projections techniques, such as PCA [173, 181] or FDA [182, 183] and selection techniques based on various measures of information — see for example [189, 190]. These methods rely on the computation of statistics of the feature vectors; accurate estimation of these statistics is challenging in high-dimensional spaces. Furthermore, projection based transformations are agnostic to the underlying structure of $\boldsymbol{h} = \text{col}\{\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_K\}$. The resulting classifier, based on a scrambled feature vector, can become difficult to interpret. In contrast, we propose to operate on the classifier soft sub-scores, defined as:

$$\boldsymbol{s}_w = \text{col}\{w_1^{\mathsf{T}}\, \boldsymbol{h}_1, w_2^{\mathsf{T}}\, \boldsymbol{h}_2, \ldots, w_K^{\mathsf{T}}\, \boldsymbol{h}_K\} \in \mathbb{R}^K, \tag{9.7}$$

This vector is of dimension $K \ll M$. Working with this reduced dimension has several advantages. First, overfitting is less likely to occur, as the dimension under consideration is significantly smaller than the underlying dimension of the feature space for appropriately chosen sub-vectors $\boldsymbol{h}_k$. Second, the structure of the feature space is preserved, allowing for more interpretable results. Third, we exploit the information gathered from statistics of $\boldsymbol{s}_w$ in real-time by feeding it back into the computation of w. This additional information results in more accurate estimate of w, which in turn stabilizes the statistics of $\boldsymbol{s}_w$ by reducing the weight of noisy features. This closed loop results in more accurate identification of relevant features.

To motivate the mechanism proposed in the sequel, recall that the general objective of feature selection in the context of classification is the identification of subsets of the feature

vector $\boldsymbol{h}$, that are highly correlated with the class label $\boldsymbol{\gamma}$. Here we propose to obtain a measure of this correlation by analyzing the statistics of $\{w_k^{\mathsf{T}}\,\boldsymbol{h}_k\}$ directly. To this end, we recall the definition of the Pearson correlation coefficient of two scalar random variables $\boldsymbol{x}, \boldsymbol{y}$ with means $\mu_{\boldsymbol{x}}, \mu_{\boldsymbol{y}}$ and standard deviations $\sigma_{\boldsymbol{x}}, \sigma_{\boldsymbol{y}}$:

$$\rho_{\boldsymbol{x},\boldsymbol{y}} = \frac{\mathbb{E}\left[(\boldsymbol{x}-\mu_{\boldsymbol{x}})(\boldsymbol{y}-\mu_{\boldsymbol{y}})\right]}{\sigma_{\boldsymbol{x}}\sigma_{\boldsymbol{y}}} \tag{9.8}$$

We collect the absolute values of the pairwise correlation coefficients for the individual predictions, $\rho_{w_k^{\mathsf{T}}\boldsymbol{h}_k, w_\ell^{\mathsf{T}}\boldsymbol{h}_\ell}$, into a symmetric matrix $A \in [0,1]^{K\times K}$, so that the element in the ℓ-th row and k-th column is defined as:

$$A^{(\ell k)} \triangleq a_{\ell k} \triangleq |\rho_{w_\ell^{\mathsf{T}}\boldsymbol{h}_\ell, w_k^{\mathsf{T}}h_k}| \tag{9.9}$$

This matrix is a measure of the linear correlations among predictions based on subsets of feature vector $\boldsymbol{h}$. A value $a_{\ell k}$ close to zero indicates that it is difficult to linearly predict the classification score based on the k-th sub-vector from the ℓ-th sub-vector and vice-versa. This implies that at least one of the sub-vectors contributes little information to the classification decision. Motivated by this observation, we proceed to interpret the A matrix as an adjacency matrix to a K-node graph, where each node k represents a sub-vector $\boldsymbol{h}_k$ of the feature vector $\boldsymbol{h}$. This is illustrated in Fig. 9.2, which corresponds to one particular realization of the BRAIN structure; in future works we will examine more elaborate structures, involving, in addition, local memory and dictionaries evolving over time.

The strength of the link between nodes k and ℓ is given by the absolute value of the Pearson correlation coefficient $\rho_{w_k^{\mathsf{T}}\boldsymbol{h}_k, w_\ell^{\mathsf{T}}\boldsymbol{h}_\ell}$. Nodes with strong connections have a tendency to agree in their predictions of the class variable. These predictions are in turn based on the vectors $\boldsymbol{h}_k$ and $\boldsymbol{h}_\ell$, respectively. In the sequel, we will show how to incorporate the learned correlation information into the online update of the learning algorithm. The objective is to devise an algorithm, where opinions of nodes in a strongly connected cluster are reinforced. This behavior mimics the fact that specific neural connections in the brain are reinforced as a result of learning [191].

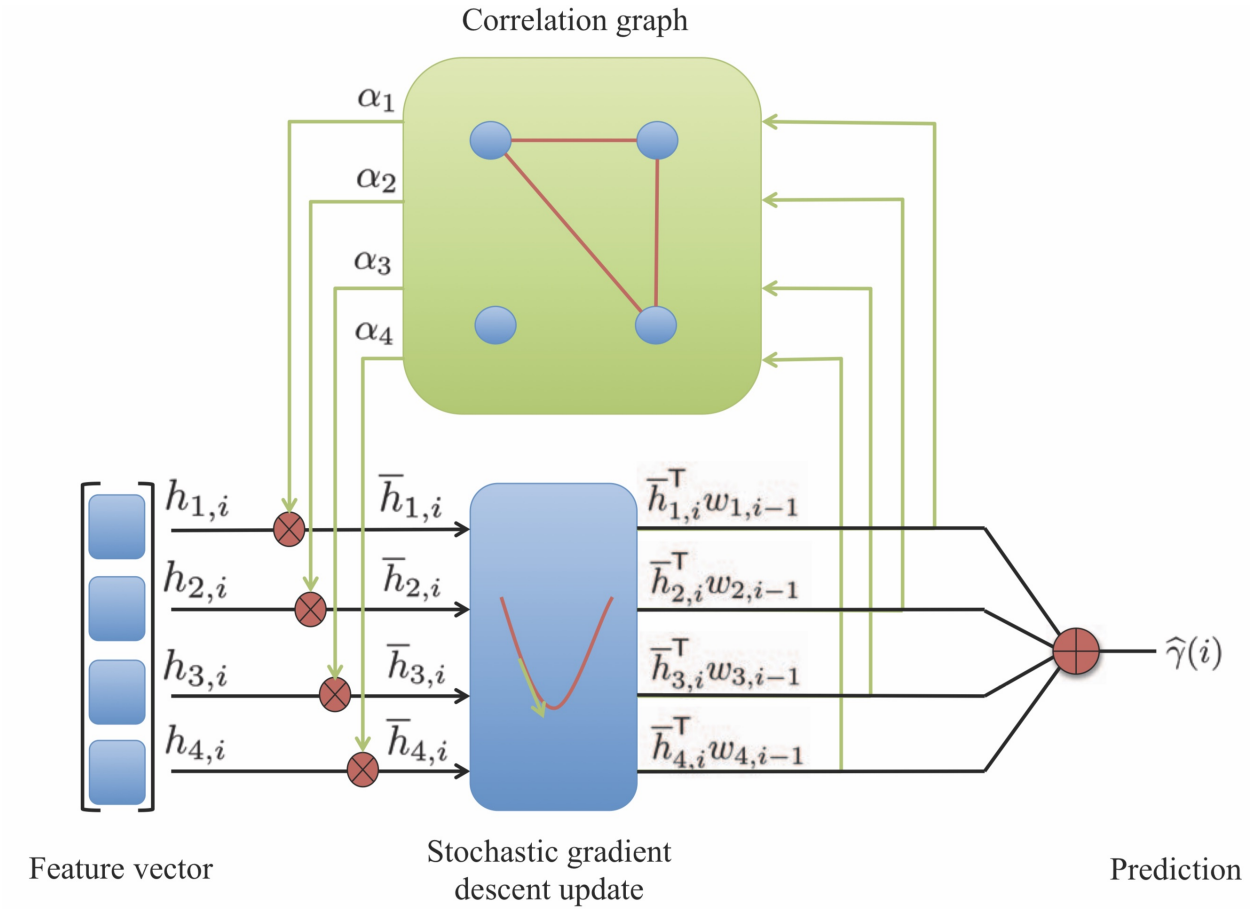

Figure 9.2: Illustration of a correlation layer placed on top of an online learning algorithm.

9.3 Correlation-Aware Online Update

We associate with each node k a scalar weight α_k, which is obtained from the adjacency matrix A of the graph according to

$$\alpha_k = \sum_{\ell=1}^{K} a_{\ell k} = \sum_{\ell=1}^{K} a_{k\ell} \tag{9.10}$$

These weights can be interpreted as a measure of trust placed in node k by its neighbors. This trust, loosely speaking, is the result of agreeing on classification decisions during past realizations of the feature vectors. We use this measure to scale incoming sub-vectors of the feature vector. The full algorithm, applied to a dataset of observations, $\{h_n, \gamma(n)\}_{n=0}^{N-1}$, generated form random variables $\{\boldsymbol{h}, \boldsymbol{\gamma}\}$, is summarized below where the notation $\partial Q(\cdot)$

refers to the gradient vector of $Q(\cdot)$ when it is differentiable or to a sub-gradient vector when it is non-differentiable. Likewise, for $\partial R(\cdot)$.

Algorithm 9.1 Online BRAIN strategy

Parameters: ν, N

Initialize: w_0, m_0, Σ_0

Run:

for $i < N$ **do**

Statistics:

$$s_{i-1} = \left(w_{1,i-1}^{\mathsf{T}} h_{1,i}, w_{2,i-1}^{\mathsf{T}} h_{2,i}, \ldots, w_{K,i-1}^{\mathsf{T}} h_{K,i}\right)^{\mathsf{T}}$$

$$m_i = (1-\nu) m_{i-1} + \nu s_{i-1}$$

$$\Sigma_i = (1-\nu)\Sigma_{i-1} + \nu\,(s_{i-1} - m_i)\,(s_{i-1} - m_i)^{\mathsf{T}}$$

Weights:

$$a_{\ell k}(i) = \frac{\Sigma_i^{(\ell k)}}{\sqrt{\Sigma_i^{(\ell\ell)} \Sigma_i^{(kk)}}},\ \forall\ \ell, k$$

$$\alpha_k(i) = \sum_{\ell=1}^{K} a_{\ell k}(i),\ \forall\ k$$

Learning:

$$\overline{h}_i = \mathrm{col}\{\alpha_1(i) h_{1,i}, \alpha_2(i) h_{2,i}, \ldots, \alpha_K(i) h_{K,i}\}$$

$$w_i = w_{i-1} - \mu \cdot \partial Q(w_{i-1}; \overline{h}_i, \gamma(i)) - \mu \cdot \partial R(w_{i-1})$$

end for

Return: w_N, Σ_N, A_N

Observe that the above algorithm is fully online, which is particularly useful when the feature vector is high-dimensional. The statistics information of $w_{k,i}^{\mathsf{T}} h_{k,i}$ is estimated adaptively, where the parameter ν controls the trade-off between accuracy of estimation and speed of convergence. The update of the weight vector w_i for classification is performed through a stochastic gradient step. For example, for online logistic regression, where

$$Q(w) = \ln\left(1 + e^{-\boldsymbol{\gamma}\,\boldsymbol{h}^{\mathsf{T}} w}\right), \quad R(w) = \delta \|w\|^2, \tag{9.11}$$

the algorithm would take the form

$$w_i = (1 - \mu\delta) w_{i-1} - \mu \gamma(i) \overline{h}_i \left(1 + e^{-\gamma(i)\overline{h}_i}\right)^{-1}. \tag{9.12}$$

For support vector machines with ℓ_2 regularization, where

$$Q(w) = \max\left(1 - \boldsymbol{\gamma}\,\boldsymbol{h}^{\mathsf{T}} w, 0\right), \quad R(w) = \delta \|w\|^2, \tag{9.13}$$

the algorithm becomes

$$w_i = (1 - \mu\delta)w_{i-1} - \mu\gamma(i)\overline{h}_i \cdot \mathbb{I}[\,\gamma(i)\overline{h}_i^{\mathsf{T}} w_{i-1} < 1\,] \tag{9.14}$$

where μ is the step size. The notation $\mathbb{I}[\cdot]$ represents the 0-1 indicator function, which is equal to 1 when the statement is true and 0 when it is false.

9.4 Simulation Results

9.4.1 Artificial Data

We begin by illustrating performance on synthetic data. The dataset is generated using the `make_classification` function[1] from the `sklearn.datasets` Python module [192], which is adapted from one of the datasets in the 2003 NIPS feature selection challenge [193]. The method allows for the specification of the number of informative and non-informative features. We generate $N = 3000$ feature vectors of dimension $M = 400$, where only the first 40 indices contain class information. The remaining 360 indices contain noise.

To begin with, we confirm that the statistics of classifier scores indeed allow the classifier to learn the subset of informative features. In Fig. 9.3 we show the evolution of the correlation network, which controls the weighting of the incoming feature blocks. We represent the weight α_k of node k through the size of its dot, and the correlation between a pair of sub-vectors of h through the thickness of the connecting link. The correlation matrix Σ is initialized as the identity matrix, resulting in a set of K unconnected nodes and $\alpha_k = \frac{1}{K}$, depicted in the leftmost plot. The second plot depicts the state of the correlation network after convergence, resulting in a fully connected network, albeit with two dominant nodes, namely nodes 1 and 2, which correspond to the first 40 elements of the feature vector, which is the informative subset. This dominance becomes more clear in the rightmost plot, where weak links were removed by simple thresholding. It is evident that there is a strong link

[1] `http://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_classification.html`

between nodes 1 and 2. This means that decisions formed from the first 20 elements and those formed from the second set of 20 features have a strong tendency to agree. These are in fact the informative subset, as constructed. In contrast, none of the decisions formed based on the remaining 18 subsets show any meaningful correlation.

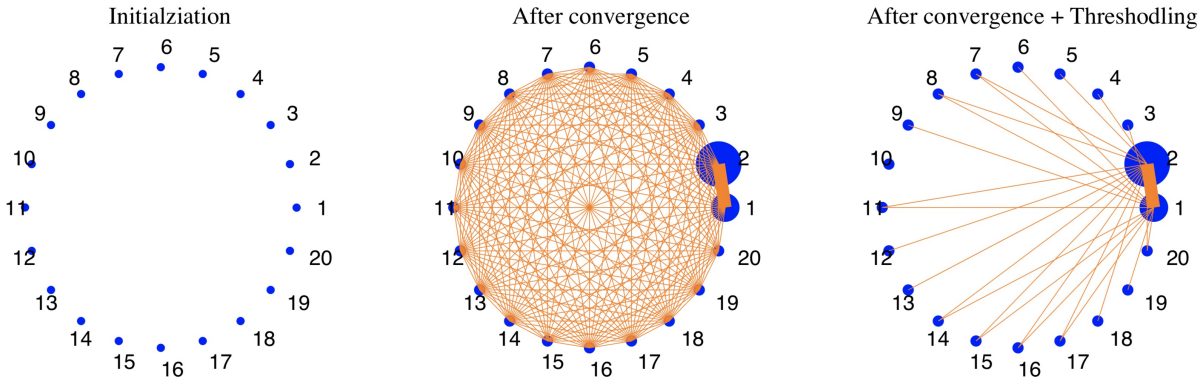

Figure 9.3: Evolution of correlation network of classifier sub-scores.

Figure 9.4 shows the evolution of the classification accuracy for ordinary online logistic regression compared to BRAIN online logistic regression. For this particular example, we observe fastest convergence and highest performance for a damping factor $\nu = 0.01$ and $K = 20$ equally spaced divisions of the feature vector. We show two additional accuracy evolutions to assess the sensitivity of the algorithm performance for varying design choices.

9.4.2 p53 Mutants Dataset

Here, we test the performance of the algorithm on real data, compiled in the University of California, Irvine (UCI) Machine Learning Repository[2], discussed in [194]. The dataset contains biophysical features pertaining to the p53 protein, which is also known as a tumor suppressor protein. When active, p53 guards the genome against cancer. The objective is to predict the state of p53 (active or inactive) from $M = 5408$ features. One challenge in this dataset is that the classes are highly unbalanced, with a majority of p53 realizations being active (healthy). Of the 16772 available instances, only 286 are inactive. To remedy this

[2] http://archive.ics.uci.edu/ml/datasets/p53+Mutants

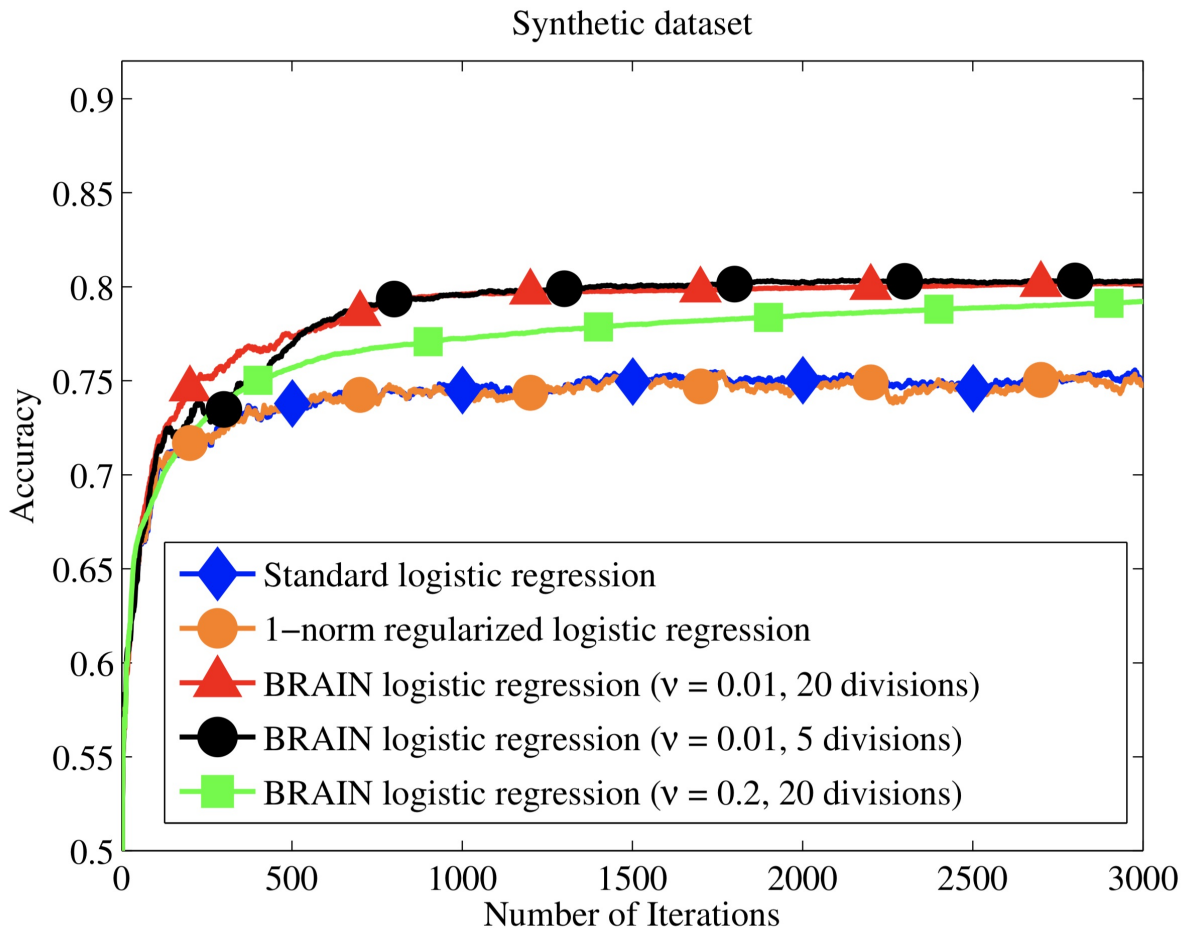

Figure 9.4: Learning curves for logistic regression with and without the correlation layer on synthetic data.

imbalance, we randomly select 286 active instances. After leaving 72 samples for testing, we are left with a training set of size $N = 500$.

Here we endow a support vector machine with the correlation layer and compare performance against ordinary SVM in Fig. 9.5. Since we have no prior information on the structure of the feature space, we divide the feature vector into $K = 50$ divisions of equal size. To allow the algorithms to converge, we run multiple passes over the small training set.

We show the learned correlation network in Fig. 9.6. We observe that the nodes in the bottom left form a cluster (nodes 25–43) and contribute most strongly to the classification decision.

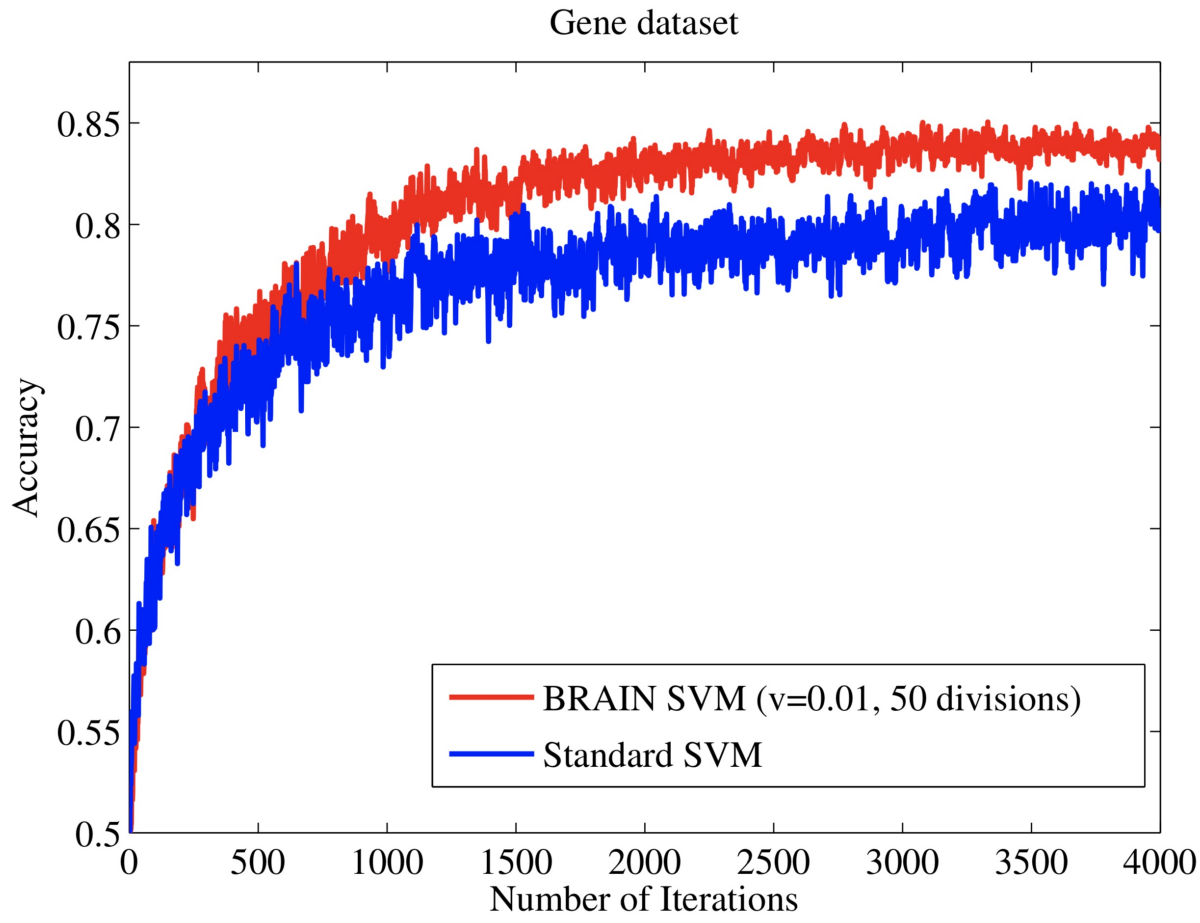

Figure 9.5: Learning curves for Support-Vector-Machine with and without correlation layer on gene data, $\mu = 0.01$, $\nu = 0.01$, and $\rho = 0.01$.

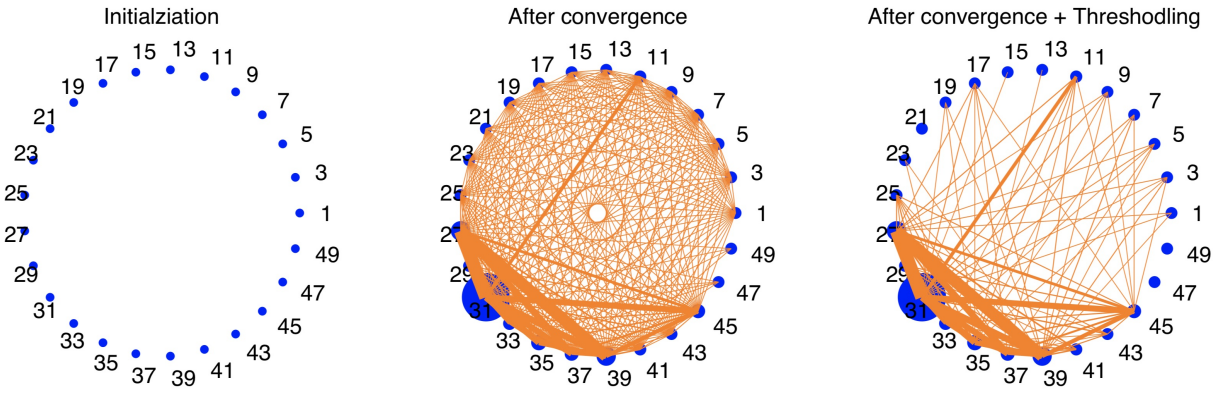

Figure 9.6: Correlation network evolution on p53 mutants.

CHAPTER 10

Conclusions and Future Issues

In this dissertation, we developed distributed strategies for continuous adaptation and learning in the presence of non-smooth regularizers. For the case when regularizers are chosen small, we studied the performance of the proximal diffusion recursion and showed that despite the lack of smoothness and persistent gradient noise, the algorithm is able to converge to the minimizer of the aggregate cost within $O(\mu)$ in the mean-square sense, assuming that the step-size and regularization parameter are appropriately coupled (Theorem 2.1). We proceeded to extend the strategy to allow for arbitrary convex regularizers by constructing a smooth approximation based on conjugate smoothing. We examined the relationship between the step-size, smoothing-parameter and stability-conditions and determined an expression for the coupling between step-size and smoothing parameter, which ensures that the limiting point of the algorithm converges to the minimizer of the original problem as $\mu \to 0$ (Theorem 3.3). We illustrated the algorithms through applications in machine learning and image reconstruction. Avenues for future research are the examination of the effect of the proximity function in the construction of the smooth approximation on the algorithm as well as perturbations caused by persistent errors in the evaluation of the proximal operators.

A second contribution of this dissertation is the establishment of second-order guarantees for the diffusion strategy in smooth, but non-convex environments. In particular, we established a descent relation for the network centroid around strict saddle-points under the condition that a noise component is present in at least one descent direction (Theorem 6.1). This relation, along with the more commonly established descent in the large-gradient regime, allowed us to provide a second-order stationarity guarantee in polynomial time (Theorem 6.2). Open questions for future consideration include the examination of second-order guarantees

in the presence of non-smooth terms and an analysis of the various randomization schemes employed in practice, such as for example dropout [195], under the gradient noise framework.

In the second part of the dissertation, focusing on learning from data exhibiting an internal network structure, we proposed the Laplacian LMS Strategy and variants involving projections for learning the graph characterizing a heat diffusion model. The resulting algorithm takes the form of an adaptive filter, and is able to learn the true, underlying graph with arbitrarily high accuracy for sufficiently small step-sizes (Theorem 8.1). Detailed exploration of the tracking performance of the adaptive algorithm as a function of the observed graph and its evolution is left for future research.

We also proposed a BRAIN strategy to enhance the performance of online classifiers for high-dimensional feature spaces. We illustrated results and performance on both artificially generated and real data examples and observed experimentally that **(a)** the correlation layer is able to identify the subset of informative features; **(b)** this information seeps into the final weight vector, and **(c)** this results in improved performance when compared to regular versions of the respective online learners. This work opens avenues for further research. Recall that one of the key features of the correlation layer is that it weights features based on classifier sub-scores. These can be interpreted as single-dimensional projections of the sub-feature vectors with reduced variance. More elaborate and possibly higher-dimensional, albeit still variance-reduced, representations can be considered by means of dictionaries, which are updated in an online manner, similar to [196]. A second opportunity for improvement is the automatic and iterative refinement of feature vector divisions in the absence of exact prior knowledge. In this work, we were able to show performance improvement with evenly spaced divisions, but do not make a claim of optimality. On the other hand, correlation networks after convergence contain information on the amount of information contained in a given feature subset. This information can be used to inform a restructuring of the feature vector subsets, before running the algorithm again with the previous weight vector as a starting point. In this manner, the information provided by the correlation graph can be more fully exploited. Finally, distributed implementations can be pursued, along the lines of [1, 22].

REFERENCES

[1] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, July 2014.

[2] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, Sep. 1999.

[3] V. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability & Its Applications*, vol. 16, no. 2, pp. 264–280, 1971.

[4] M. J. Kearns and U. V. Vazirani, *An Introduction to Computational Learning Theory*, MIT Press, Cambridge, MA, USA, 1994.

[5] B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, 1996.

[6] A. Blum, *On-Line Algorithms in Machine Learning*, pp. 306–325, Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[7] O. Bousquet, S. Boucheron, and G. Lugosi, *Introduction to Statistical Learning Theory*, pp. 169–207, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[8] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 09 1951.

[9] K. Yuan, B. Ying, S. Vlaski, and A. H. Sayed, "Stochastic gradient descent with finite samples sizes," in *Proc. of IEEE MLSP*, Vietri sul Mare, Italy, Sep. 2016, pp. 1–6.

[10] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Proc. of NIPS*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., pp. 315–323. Curran Associates, Inc., 2013.

[11] E. Otte and R. Rousseau, "Social network analysis: a powerful strategy, also for the information sciences," *Journal of Information Science*, vol. 28, no. 6, pp. 441–453, 2002.

[12] D. Acemoglu, M. A. Dahleh, I. Lobel, and A. Ozdaglar, "Bayesian learning in social networks," *The Review of Economic Studies*, vol. 78, no. 4, pp. 1201–1236, 2011.

[13] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi, "Non-bayesian social learning," *Games and Economic Behavior*, vol. 76, no. 1, pp. 210 – 225, 2012.

[14] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart grid — the new and improved power grid: A survey," *IEEE Communications Surveys Tutorials*, vol. 14, no. 4, pp. 944–980, Fourth 2012.

[15] P. McDaniel and S. McLaughlin, "Security and privacy challenges in the smart grid," *IEEE Security Privacy*, vol. 7, no. 3, pp. 75–77, May 2009.

[16] L. Krishnamachari, D. Estrin, and S. Wicker, "The impact of data aggregation in wireless sensor networks," in *Proceedings 22nd International Conference on Distributed Computing Systems Workshops*, July 2002, pp. 575–578.

[17] A. H. Sayed, A. Tarighat, and N. Khajehnouri, "Network-based wireless location: challenges faced in developing techniques for accurate wireless location information," *IEEE Signal Processing Magazine*, vol. 22, no. 4, pp. 24–40, July 2005.

[18] R. V. Kulkarni, A. Forster, and G. K. Venayagamoorthy, "Computational intelligence in wireless sensor networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 13, no. 1, pp. 68–96, First 2011.

[19] G. Karagiannis, O. Altintas, E. Ekici, G. Heijenk, B. Jarupan, K. Lin, and T. Weil, "Vehicular networking: A survey and tutorial on requirements, architectures, challenges, standards and solutions," *IEEE Communications Surveys Tutorials*, vol. 13, no. 4, pp. 584–616, Fourth 2011.

[20] E. Ahmed and H. Gharavi, "Cooperative vehicular networking: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 996–1014, March 2018.

[21] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

[22] A. H. Sayed, "Adaptive networks," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 460–497, April 2014.

[23] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 2003.

[24] S. U. Pillai, T. Suel, and S. Cha, "The Perron-Frobenius theorem: Some of its applications," *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 62–75, March 2005.

[25] J. Chen and A. H. Sayed, "Distributed Pareto optimization via diffusion strategies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 2, pp. 205–220, April 2013.

[26] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Automatic Control*, vol. 54, no. 1, pp. 48–61, Jan 2009.

[27] J. Chen and A. H. Sayed, "On the learning behavior of adaptive networks - Part I: Transient analysis," *IEEE Transactions on Information Theory*, vol. 61, no. 6, pp. 3487–3517, June 2015.

[28] R. Xin, A. K. Sahu, U. A. Khan, and S. Kar, "Distributed stochastic optimization with gradient tracking over strongly-connected networks," *available as arXiv:1903.07266*, March 2019.

[29] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1835–1854, 2016.

[30] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[31] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, "Exact diffusion for distributed optimization and learning – Part II: Convergence analysis," *IEEE Transactions on Signal Processing*, vol. 67, no. 3, pp. 724–739, Feb 2019.

[32] X. Zhao and A. H. Sayed, "Asynchronous adaptation and learning over networks – Part II: Performance analysis," *IEEE Transactions on Signal Processing*, vol. 63, no. 4, pp. 827–842, Feb 2015.

[33] Z. J. Towfic and A. H. Sayed, "Adaptive penalty-based distributed stochastic convex optimization," *IEEE Trans. on Signal Process.*, vol. 62, no. 15, pp. 3924–3938, Aug. 2014.

[34] B. Ying and A. H. Sayed, "Performance limits of stochastic sub-gradient learning, Part II: Multi-agent case," *Signal Processing*, vol. 144, pp. 253 – 264, 2018.

[35] R. Nassif, C. Richard, A. Ferrari, and A. H. Sayed, "Proximal multitask learning over networks with sparsity-inducing coregularization," *Trans. Sig. Proc.*, vol. 64, no. 23, pp. 6329–6344, Dec. 2016.

[36] R. Nassif, S. Vlaski, and A. H. Sayed, "Learning over multitask graphs - Part I: Stability analysis," *available as arXiv:1805.08535*, May 2018.

[37] R. Nassif, S. Vlaski, and A. H. Sayed, "Learning over multitask graphs - Part II: Performance analysis," *available as arXiv:1805.08547*, May 2018.

[38] R. Nassif, S. Vlaski, C. Richard, and A. H. Sayed, "A regularization framework for learning over multitask graphs," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 297–301, Feb 2019.

[39] R. Nassif, S. Vlaski, and A. H. Sayed, "Adaptation and learning over networks under subspace constraints – Part I: Stability analysis," *available as arXiv:1905.08750*, June 2019.

[40] R. Nassif, S. Vlaski, and A. H. Sayed, "Adaptation and learning over networks under subspace constraints – Part II: Performance analysis," *available as arXiv:1906.12250*, June 2019.

[41] P. Di Lorenzo, S. Barbarossa, and A. H. Sayed, "Sparse diffusion LMS for distributed adaptive estimation," in *Proc. IEEE ICASSP*, Kyoto, Japan, March 2012, pp. 3281–3284.

[42] P. Di Lorenzo and A. H. Sayed, "Sparse distributed learning based on diffusion adaptation," *IEEE Trans. Signal Process.*, vol. 61, no. 6, pp. 1419–1433, March 2013.

[43] S. Chouvardas, K. Slavakis, Y. Kopsinis, and S. Theodoridis, "A sparsity promoting adaptive algorithm for distributed learning," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5412–5425, Oct. 2012.

[44] Y. Liu, C. Li, and Z. Zhang, "Diffusion sparse least-mean squares over networks," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4480–4485, Aug 2012.

[45] M. Schmidt, N. L. Roux, and F. R. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Proc. Advances in Neural Information Processing Systems 24*, Granada, Spain, 2011, pp. 1458–1466.

[46] A. I. Chen and A. Ozdaglar, "A fast distributed proximal-gradient method," in *Proc. Annual Allerton Conference on Communication, Control, and Computing*, Allerton, USA, Oct. 2012, pp. 601–608.

[47] W. M. Wee and I. Yamada, "A proximal splitting approach to regularized distributed adaptive estimation in diffusion networks," in *Proc. IEEE ICASSP*, Vancouver, Canada, May 2013, pp. 5420–5424.

[48] P. Di Lorenzo, "Diffusion adaptation strategies for distributed estimation over Gaussian Markov random fields," *IEEE Transactions on Signal Process.*, vol. 62, no. 21, pp. 5748–5760, Nov. 2014.

[49] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2013.

[50] S. Vlaski and A. H. Sayed, "Proximal diffusion for stochastic costs with non-differentiable regularizers," in *Proc. IEEE ICASSP*, Brisbane, Australia, April 2015, pp. 3352–3356.

[51] S. Vlaski, L. Vandenberghe, and A. H. Sayed, "Diffusion stochastic optimization with non-smooth regularizers," in *Proc. of IEEE ICASSP*, Shanghai, China, March 2016, pp. 4149–4153.

[52] P. Bianchi and J. Jakubowicz, "Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization," *IEEE Transactions on Automatic Control*, vol. 58, no. 2, pp. 391–405, Feb 2013.

[53] P. Di Lorenzo and G. Scutari, "NEXT: in-network nonconvex optimization," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 2, pp. 120–136, June 2016.

[54] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent," in *Advances in Neural Information Processing Systems 30*, pp. 5330–5340. Curran Associates, Inc., 2017.

[55] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," *IEEE Transactions on Signal Processing*, vol. 66, no. 11, pp. 2834–2848, June 2018.

[56] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, "d^2: Decentralized training over decentralized data," in *Proceedings of the 35th International Conference on Machine Learning*, 10–15 Jul 2018, vol. 80, pp. 4848–4856.

[57] Y. Wang, W. Yin, and J. Zeng, "Global convergence of ADMM in nonconvex nonsmooth optimization," *Journal of Scientific Computing*, vol. 78, no. 1, pp. 29–63, Jan. 2019.

[58] T. Tatarenko and B. Touri, "Non-convex distributed optimization," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3744–3757, Aug. 2017.

[59] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—online stochastic gradient for tensor decomposition," in *Proc. of Conference on Learning Theory*, Paris, France, 2015, pp. 797–842.

[60] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *Proc. of ICML*, Sydney, Australia, Aug. 2017, pp. 1724–1732.

[61] A. Daneshmand, G. Scutari and V. Kungurtsev, "Second-order guarantees of distributed gradient algorithms," *available as arXiv:1809.08694*, Sep. 2018.

[62] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, "The Loss Surfaces of Multilayer Networks," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, San Diego, May 2015, pp. 192–204.

[63] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, May 2013.

[64] F. Chung, "The heat kernel as the pagerank of a graph," *Proceedings of the National Academy of Sciences*, vol. 104, no. 50, pp. 19735–19740, 2007.

[65] H. Ma, H. Yang, M. R. Lyu, and I. King, "Mining social networks using heat diffusion processes for marketing candidates selection," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, New York, NY, 2008, pp. 233–242.

[66] D. Thanou, X. Dong, D. Kressner, and P. Frossard, "Learning heat diffusion graphs," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 3, pp. 484–499, Sep. 2017.

[67] S. Vlaski, H. P. Maretić, R. Nassif, P. Frossard, and A. H. Sayed, "Online graph learning from sequential data," in *Proc. of IEEE Data Science Workshop (DSW)*, Lausanne, Switzerland, June 2018, pp. 190–194.

[68] S. Vlaski, L. Vandenberghe, and A. H. Sayed, "Regularized Diffusion Adaptation via Conjugate Smoothing," *in preparation*, September 2019.

[69] S. Vlaski and A. H. Sayed, "Diffusion learning in non-convex environments," in *Proc. of IEEE ICASSP*, Brighton, UK, May 2019, pp. 5262–5266.

[70] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments – Part I: Agreement at a Linear rate," *submitted for publication, available as arXiv:1907.01848*, July 2019.

[71] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments – Part II: Polynomial escape from saddle-points," *submitted for publication, available as arXiv:1907.01849*, July 2019.

[72] S. Vlaski and A. H. Sayed, "Second-order guarantees of stochastic gradient descent in non-convex optimization," *submitted for publication, available as arXiv:1908.07023*, August 2019.

[73] S. Vlaski, B. Ying, and A. H. Sayed, "The brain strategy for online learning," in *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Washington, DC, USA, Dec 2016, pp. 1285–1289.

[74] J. N. Tsitsiklis and M. Athans, "Convergence and asymptotic agreement in distributed decision problems," *IEEE Trans. Automatic Control*, vol. 29, no. 1, pp. 42–50, Jan 1984.

[75] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, Jan 2007.

[76] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1847–1864, Nov. 2010.

[77] J. Duchi and Y. Singer, "Efficient online and batch learning using forward backward splitting," *Journal of Machine Learning Research*, vol. 10, pp. 2899–2934, 2009.

[78] Y. Murakami, M. Yamagishi, M. Yukawa, and I. Yamada, "A sparse adaptive filtering using time-varying soft-thresholding techniques," in *Proc. IEEE ICASSP*, Dallas, USA, March 2010, pp. 3734–3737.

[79] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," *Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer Optimization and Its Applications*, pp. 185–221, Springer, NY, 2011.

[80] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[81] J. Chen, Z. J. Towfic, and A. H. Sayed, "Dictionary learning over distributed models," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 1001–1016, February 2015.

[82] E. Kreyszig, *Introductory Functional Analysis with Applications*, John Wiley & Sons, 1989.

[83] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, 1997.

[84] B. T. Polyak, *Introduction to Optimization*, Optimization Software, 1997.

[85] L. Vandenberghe, "Optimization Methods for Large-Scale Systems," *UCLA EE236C Lecture Notes*, 2014.

[86] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.

[87] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. on Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

[88] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[89] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical Programming*, vol. 103, no. 1, pp. 127–152, 2005.

[90] D. P. Bertsekas, "A new class of incremental gradient methods for least squares problems," *SIAM J. Optim.*, vol. 7, no. 4, pp. 913–926, April 1997.

[91] S. Sundhar Ram, A. Nedic, and V. V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *Journal of Optimization Theory and Applications*, vol. 147, no. 3, pp. 516–545, Dec 2010.

[92] A. Nedic, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, April 2010.

[93] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *Trans. Sig. Proc.*, vol. 56, no. 7, pp. 3122–3136, July 2008.

[94] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[95] P. Di Lorenzo and G. Scutari, "Next: In-network nonconvex optimization," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 2, pp. 120–136, June 2016.

[96] Y. Sun, G. Scutari, and D. Palomar, "Distributed nonconvex multiagent optimization over time-varying networks," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, Nov 2016, pp. 788–794.

[97] A. Mokhtari and A. Ribeiro, "Dsa: Decentralized double stochastic averaging gradient algorithm," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2165–2199, Jan. 2016.

[98] Y. Sun and G. Scutari, "Distributed nonconvex optimization for sparse representation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 4044–4048.

[99] A. Nedich, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2597–2633, 1 2017.

[100] R. Xin and U. A. Khan, "A linear algorithm for optimization over directed graphs with geometric convergence," *IEEE Control Systems Letters*, vol. 2, no. 3, pp. 315–320, July 2018.

[101] S. Pu and A. Nedić, "A distributed stochastic gradient tracking method," *available as arXiv:1803.07741*, March 2018.

[102] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, "Exact diffusion for distributed optimization and learning—part i: Algorithm development," *IEEE Transactions on Signal Processing*, vol. 67, no. 3, pp. 708–723, Feb 2019.

[103] D. Jakovetic, J. Xavier, and J. M. F. Moura, "Cooperative convex optimization in networked systems: Augmented lagrangian algorithms with directed gossip communication," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3889–3902, Aug 2011.

[104] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Transactions on Automatic Control*, vol. 57, no. 3, pp. 592–606, March 2012.

[105] K. I. Tsianos and M. G. Rabbat, "Distributed dual averaging for convex optimization under communication delays," in *Proc. American Control Conference (ACC)*, Montreal, Canada, June 2012, pp. 1067–1072.

[106] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the admm in decentralized consensus optimization," *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1750–1761, April 2014.

[107] M. Jaggi, V. Smith, M. Takáč, J. Terhorst, S. Krishnan, T. Hofmann, and M. I. Jordan, "Communication-efficient distributed dual coordinate ascent," in *Proc. International Conference on Neural Information Processing Systems*, Montreal, Canada, 2014, pp. 3068–3076.

[108] Q. Ling, W. Shi, G. Wu, and A. Ribeiro, "Dlm: Decentralized linearized alternating direction method of multipliers," *IEEE Transactions on Signal Processing*, vol. 63, no. 15, pp. 4051–4064, Aug 2015.

[109] D. Jakovetić, J. M. F. Moura, and J. Xavier, "Linear convergence rate of a class of distributed augmented lagrangian algorithms," *IEEE Transactions on Automatic Control*, vol. 60, no. 4, pp. 922–936, April 2015.

[110] K. Seaman, F. Bach, S. Bubeck, Yin T. Lee, and L. Massoulié, "Optimal algorithms for smooth and strongly convex distributed optimization in networks," in *Proc. International Conference on Machine Learning*, Sydney, Australia, 2017, pp. 3027–3036.

[111] G. Lan, S. Lee, and Y. Zhou, "Communication-efficient algorithms for decentralized and stochastic optimization," *Mathematical Programming*, Dec 2018.

[112] D. Jakovetić, "A unification and generalization of exact distributed first-order methods," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 5, no. 1, pp. 31–46, March 2019.

[113] A. Nedić, A. Olshevsky, and M. G. Rabbat, "Network topology and communication-computation tradeoffs in decentralized optimization," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 953–976, May 2018.

[114] A. Nedić and A. Olshevsky, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 3936–3947, Dec 2016.

[115] S. Pu, W. Shi, J. Xu and A. Nedić, "A push-pull gradient method for distributed optimization in networks," *available as arXiv:1803.07588*, March 2018.

[116] A. Beck and M. Teboulle, "Smoothing and first order methods: A unified framework," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 557–580, 2012.

[117] Y.-L. Yu, "Better approximation and faster algorithm using the proximal average," in *Advances in Neural Information Processing Systems 26*, pp. 458–466. 2013.

[118] J. Duchi, P. Bartlett, and M. Wainwright, "Randomized smoothing for stochastic optimization," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 674–701, 2012.

[119] K. Scaman, F. Bach, S. Bubeck, Y. T. Lee, and L. Massoulié, "Optimal algorithms for non-smooth distributed optimization in networks," in *Proc. International Conference on Neural Information Processing Systems*, Montreal, Canada, 2018, pp. 2745–2754.

[120] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[121] C. Planiden and X. Wang, "Strongly convex functions, moreau envelopes, and the generic nature of convex functions with strong minimizers," *SIAM Journal on Optimization*, vol. 26, no. 2, pp. 1341–1364, 2016.

[122] H. Bauschke, R. Goebel, Y. Lucet, and X. Wang, "The proximal average: Basic theory," *SIAM Journal on Optimization*, vol. 19, no. 2, pp. 766–785, 2008.

[123] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, 1997.

[124] Y. Kim, J. Kim, and Y. Kim, "Blockwise sparse regression," *Statistica Sinica*, vol. 16, no. 2, pp. 375–390, 2006.

[125] L. Meier, S. van de Geer, and P. Buhlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society Series B*, vol. 70, pp. 53–71, 02 2008.

[126] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, June 2010.

[127] J. Cai, E. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[128] R. Nassif, C. Richard, A. Ferrari, and A. H. Sayed, "Proximal multitask learning over networks with sparsity-inducing coregularization," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6329–6344, Dec 2016.

[129] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, July 1998.

[130] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma, *Robust Statistics for Signal Processing*, Cambridge University Press, 2018.

[131] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, March 2011.

[132] S. Gelfand and S. Mitter, "Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$," *SIAM Journal on Control and Optimization*, vol. 29, no. 5, pp. 999–1018, 1991.

[133] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *29th Annual Conference on Learning Theory*, New York, 2016, pp. 1246–1257.

[134] H. Daneshmand, J. Kohler, A. Lucchi and T. Hofmann, "Escaping saddles with stochastic gradients," *available as arXiv:1803.05999*, March 2018.

[135] C. Fang, C. J. Li, Z. Lin, and T. Zhang, "SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator," in *Proc. of NIPS*, pp. 689–699. Montreal, Canada, 2018.

[136] Z. Allen-Zhu, "Natasha 2: Faster non-convex optimization than SGD," in *Proc. of NIPS*, pp. 2675–2686. Montreal, Canada, Dec. 2018.

[137] Z. Allen-Zhu and Y. Li, "NEON2: Finding local minima via first-order oracles," in *Proc. of NIPS*, pp. 3716–3726. Montreal, Canada, Dec. 2018.

[138] C. Fang, Z. Lin and T. Zhang, "Sharp analysis for nonconvex sgd escaping from saddle points," *available as arXiv:1902.00247*, Feb. 2019.

[139] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade and M. I. Jordan, "Stochastic gradient descent escapes saddle points efficiently," *available as arXiv:1902.04811*, Feb. 2019.

[140] B. Swenson, S. Kar, H. V. Poor and J. M. F. Moura, "Annealing for distributed global optimization," *available as arXiv:1903.07258*, March 2019.

[141] P. Jain and P. Kar, "Non-convex optimization for machine learning," *Foundations and Trends in Machine Learning*, vol. 10, no. 3-4, pp. 142–336, 2017.

[142] S. J. Reddi, A. Hefny, S. Sra, B. Póczós, and A. Smola, "Stochastic variance reduction for nonconvex optimization," in *Proc. of ICML*, New York, NY, USA, 2016, pp. 314–323.

[143] R. Ge, Z. Li, W. Wang and X. Wang, "Stabilized SVRG: Simple variance reduction for nonconvex optimization," *available as arXiv:1905.00529*, May 2019.

[144] Y. Nesterov and B.T. Polyak, "Cubic regularization of newton method and its global performance," *Mathematical Programming*, vol. 108, no. 1, pp. 177–205, Aug 2006.

[145] A. Klenke, *Probability Theory: A Comprehensive Course*, Springer, 2013.

[146] J. Chen and A. H. Sayed, "On the learning behavior of adaptive networks – Part II: Performance analysis," *IEEE Transactions on Information Theory*, vol. 61, no. 6, pp. 3518–3548, June 2015.

[147] S.-Y. Tu and A. H. Sayed, "Diffusion strategies outperform consensus strategies for distributed estimation over adaptive networks," *Trans. Sig. Proc.*, vol. 60, no. 12, pp. 6217–6234, Dec. 2012.

[148] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, B. Poczos and A. Singh, "Gradient descent can take exponential time to escape saddle points," *available as arXiv:1705.10412*, May 2017.

[149] A. H. Sayed, *Adaptive Filters*, John Wiley & Sons, Inc., 2008.

[150] Y. Nesterov, *Introductory Lectures on Convex Programming Volume I: Basic Course*, Springer, 1998.

[151] D. Bertsekas and J. Tsitsiklis, "Gradient convergence in gradient methods with errors," *SIAM Journal on Optimization*, vol. 10, no. 3, pp. 627–642, 2000.

[152] F. Facchinei, V. Kungurtsev, L. Lampariello, G. Scutari, "Ghost Penalties in Nonconvex Constrained Optimization: Diminishing Stepsizes and Iteration Complexity," *available as arXiv:1709.03384*, Sep. 2017.

[153] K. Kawaguchi, "Deep learning without poor local minima," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., pp. 586–594. Curran Associates, Inc., 2016.

[154] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., pp. 2973–2981. Curran Associates, Inc., 2016.

[155] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., pp. 3873–3881. Curran Associates, Inc., 2016.

[156] F. E. Curtis, D. P. Robinson, and M. Samadi, "A trust region algorithm with a worst-case iteration complexity of $o(\epsilon^{-3/2})$ for nonconvex optimization," *Mathematical Programming*, vol. 162, pp. 1–32, 2017.

[157] C. Jin, P. Netrapalli, and M. I. Jordan, "Accelerated gradient descent escapes saddle points faster than gradient descent," in *Proceedings of the 31st Conference On Learning Theory*, Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, Eds. 06–09 Jul 2018, vol. 75 of *Proceedings of Machine Learning Research*, pp. 1042–1085, PMLR.

[158] A. P. Dempster, "Covariance selection," *Biometrics*, vol. 28, no. 1, pp. 157–175, 1972.

[159] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical Lasso," vol. 9, pp. 432–41, Aug. 2008.

[160] B. Lake and J. Tenenbaum, "Discovering structure by learning sparse graphs," in *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, Jan. 2010, pp. 778–783.

[161] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst, "Learning laplacian matrix in smooth graph signal representations," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6160–6173, 2016.

[162] V. Kalofolias, "How to learn a graph from smooth signals," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, Cadiz, Spain, May 2016, vol. 51, pp. 920–929.

[163] E. Pavez and A. Ortega, "Generalized laplacian precision matrix estimation for graph signal processing," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 6350–6354.

[164] J. Mei and J. M. F. Moura, "Signal processing on graphs: Causal modeling of unstructured data," *IEEE Transactions on Signal Processing*, vol. 65, no. 8, pp. 2077 – 2092, Apr. 2017.

[165] D. Durante and D. B. Dunson, "Locally adaptive dynamic networks," *Ann. Appl. Stat.*, vol. 10, no. 4, pp. 2203–2232, 12 2016.

[166] K. S. Xu and A. O. Hero, "Dynamic stochastic blockmodels for time-evolving social networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 4, pp. 552–562, Aug 2014.

[167] B. Zaman, L. M. Lopez-Ramos, D. Romero, and B. Beferull-Lozano, "Online topology estimation for vector autoregressive processes in data networks," in *IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Dec 2017, pp. 1–5.

[168] F. R. K. Chung, *Spectral Graph Theory*, American Mathematical Society, 1997.

[169] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, pp. 47–97, Jan. 2002.

[170] L. Farina and S. Rinaldi, *Positive linear systems. Theory and applications*, John Wiley & Sons, Inc., 06 2000.

[171] W. Arendt, *Characterization of positive semigroups on Co(X)*, pp. 122–162, Springer Berlin Heidelberg, 1986.

[172] C.-T. Chen, *Linear System Theory and Design*, Oxford University Press, Inc., New York, NY, USA, 3rd edition, 1998.

[173] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[174] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*, Academic Press, 2015.

[175] V. Vapnik, *Statistical Learning Theory*, Wiley NY, 1998.

[176] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[177] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[178] S. Haykin, *Neural Networks and Learning Machines (3rd Edition)*, Prentice Hall, 2009.

[179] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[180] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–1105. Lake Tahoe, USA, 2012.

[181] I. Jolliffe, *Principal Component Analysis*, Wiley NY, 2002.

[182] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 2013.

[183] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Mullers, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX*, pp. 41–48. Madison, USA, Aug 1999.

[184] C. M. Bishop, "Model-based machine learning," *Phil. Trans. Royal Society of London A*, vol. 371, no. 1984, pp. 1–17, 2012.

[185] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, 2009.

[186] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, University College London, London, United Kingdom, May 2003.

[187] G. E. Hinton, "A practical guide to training restricted boltzmann machines," in *Neural Networks: Tricks of the Trade: Second Edition*, pp. 599–619. Springer, 2012.

[188] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, July 2006.

[189] M. A. Hall, *Correlation-based Feature Selection for Machine Learning*, Ph.D. thesis, University of Waikato, Hamilton, New Zealand, April 1999.

[190] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *J. Mach. Learn. Res.*, vol. 13, pp. 27–66, Jan. 2012.

[191] Y. Zhang, R. H. Cudmore, D.-T. Lin, D. J. Linden, and R. L. Huganir, "Visualization of NMDA receptor-dependent AMPA receptor synaptic plasticity in vivo," *Nat Neurosci*, vol. 18, no. 3, pp. 402–407, 03 2015.

[192] 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*, vol. 12, pp. 2825–2830, 2011.

[193] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror, "Result Analysis of the NIPS 2003 Feature Selection Challenge," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 545–552. Montreal, Canada, 2005.

[194] S. A. Danziger, S. J. Swamidass, Jue Zeng, L. R. Dearth, Qiang Lu, J. H. Chen, J. Cheng, V. P. Hoang, H. Saigo, R. Luo, P. Baldi, R. K. Brachmann, and R. H. Lathrop, "Functional census of mutation sequence spaces: the example of p53 cancer rescue mutants," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 3, no. 2, pp. 114–125, April 2006.

[195] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[196] J. Chen, Z. J. Towfic, and A. H. Sayed, "Dictionary learning over distributed models," *IEEE Transactions on Signal Processing*, vol. 63, no. 4, pp. 1001–1016, Feb 2015.