Mini-Batch Gradient-Based Algorithms on High-Dimensional Quadratic Problems: Exact Dynamics and Consequences

Andrew Cheng

Department of Mathematics and Statistics

McGill University, Montreal

April, 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

©Andrew Cheng, 2023

Abstract

As learning models become ever complex (as shown in Figure 1.1), the importance of efficient optimization algorithms for training these models becomes ever more apparent (illustrated by Figure 1.2). A popular class of these algorithms are the mini-batch gradient-based methods (see Definition 2) which includes stochastic gradient descent [50], Nesterov acceleration [40], and Polyak momentum [49]. Despite the popularity and practicality of these algorithms, the theoretical understanding of these algorithms are limited even in the simplest of cases. This thesis aims to bridge the gap between empirical observations and our theoretical understanding of the behaviour of mini-batch gradient-based methods under the setting of quadratic models (See Assumption 6)

To motivate further analysis of these algorithms, we re-examine the conventional worst-case complexity bounds of gradient-based methods. In short, these bounds only consider the maximum and minimum eigenvalues of the problem, and fail to capture typical runtime behaviour [48]. To overcome this limitation, we adopt a physics-inspired approach from random matrix theory that leverages the *universality* phenomenon. By placing a distribution on the data, we show that we can incorporate information from the eigenspectrum of the Hessian of the quadratic problem to provide a sharper and more robust analysis.

In the high-dimensional setting, we demonstrate through concentration of measure results that the two source of randomness - the sampling mechanism of the algorithm and the data distribution - averages out. Furthermore, we illustrate the deep connection between polynomials and mini-batch gradient-based algorithms - that is, we can write

iterates of the algorithm as a sum of polynomials that depends on the eigenspectrum of $\mathbf{A}\mathbf{A}^T$ for a given design matrix $\mathbf{A}$. These results allow us to state the main theorem of this thesis: Under general quadratic functions, the dynamics of every mini-batch gradient-based algorithm can be exactly captured by a deterministic equation Ψ which solves a discrete Volterra equation.

In turn, exact dynamics allows us to quantify the behaviour of interesting quadratic statistical models, including training and generalization errors as well as the random features model [6, 37]. Moreover, we are able to study Ψ in order to analyze properties of the celebrated Polyak momentum algorithm under the least-squares setting. Our analysis provides sufficient conditions for convergence, the relationship between convergence rate and batch-size for fixed hyperparmeters, and limiting behaviours.

Abrégé

Les modèles d'apprentissage deviennent de plus en plus complexes (comme le montre la figure 1.1), donc l'importance d'algorithmes d'optimisation efficaces pour la formation de ces modèles devient de plus en plus évidente (illustrée par la figure 1.2). Une classe populaire de ces algorithmes est celle des méthodes basées sur le mini-batchgradient (voir définition 2), qui comprend la descente de gradient stochastique et le momentum de Polyak [47]). Malgré la popularité et le caractère pratique de ces algorithmes, leur compréhension théorique est limitée, même dans les cas les plus simples. Cette thèse vise à combler le fossé entre les observations empiriques et notre compréhension théorique du comportement des méthodes basées sur le gradient par mini-lots.

Pour motiver une analyse de ces algorithmes, nous réexaminons les limites conventionnelles de complexité dans le pire des cas des méthodes basées sur le gradient.

En bref, ces limites ne prennent en compte que les valeurs propres maximales et minimales du problème et ne parviennent pas à rendre compte du comportement typique de l'exécution [46].

Pour surmonter cette limitation, nous adoptons une approche inspirée par la théorie des matrices aléatoires, basée sur la physique, qui tire parti du phénomène d'universalité. En plaçant une distribution sur les données, nous montrons que nous pouvons incorporer des informations provenant de l'ensemble du spectre des valeurs propres du problème, ce qui permet d'obtenir une analyse plus précise et plus robuste.

Dans le cadre de la haute dimension, nous démontrons par des résultats de concentration de mesures que les deux sources d'erreur aléatoire, le mécanisme d'échantillonnage

de l'algorithme et la distribution des données - s'équilibrent. En outre, en utilisant le lien profond entre les polynômes et les algorithmes basés sur le gradient en mini-lots, nous pouvons énoncer le principal théorème de cette thèse : Sous des fonctions quadratiques générales, la dynamique de chaque algorithme à base de gradient par mini-lots peut être exactement capturée par une équation déterministe Ψ qui résout une équation de Volterra discrète.

À son tour, les dynamiques exactes nous permettent de quantifier le comportement des modèles statistiques quadratiques intéressants, incluant les erreurs d'apprentissage et de généralisation ainsi que le modèle des caractéristiques aléatoires [6, 35].

En outre, nous sommes en mesure d'étudier Ψ afin d'analyser les propriétés du célèbre algorithme de momentum de Polyak dans le cadre des moindres carrés. Notre analyse fournit des conditions suffisantes pour la convergence, la relation entre le taux de convergence et la taille du lot pour des hyperparmètres fixes, et les comportements limitants

Acknowledgements

I would like to express my profound gratitude to my supervisors Courtney and Elliot Paquette as they have been a driving force behind one of the best academic decisions I have made thus far - pursuing my master's at McGill University. Courtney and Elliot have been instrumental in revealing my passion for mathematical research and imparting the knowledge that I have gained thus far. Courtney and Elliot, thank you for always being available to meet, for patiently answering my questions, encouraging my natural voice in mathematics, and setting an excellent example as mentors and role models in the research community. Needless to say, it is impossible to show my full appreciation to my supervisors in an acknowledgment section.

Also, I would like to express my thanks to my undergraduate supervisors Russell Steele and Archer Yang. They exposed me to statistical research which eventually led me to pursuing research in machine learning. More importantly, they have provided continual support and guidance over the years in research and beyond.

My friend and collaborator, Kiwon Lee, epitomizes the PhD student I aspire to be - caring, patient, and incredibly knowledgeable. He made my master's degree journey very enjoyable, and I consider myself fortunate to be able to call him a friend. In addition, I have met many brilliant and kindhearted friends during my time at McGill. They have brought me so much and joy and made my life so much richer.

I would also like to extend my appreciation to Courtney Paquette, McGill University, Mitacs, Fonds de recherche du Québec, and SURA for generously funding my studies and research.

Last but not least, I could not have progressed this far without my family's unwavering support. Thank you to my mom and dad for always supporting my dream of pursuing research and providing emotional support. I also extend my thanks to my brother, Timothy, who serves as an embodiment of kindness that I strive to emulate.

Contributions

This thesis focuses on results that were established by a joint effort involving Kiwon Lee (McGill PhD student) and co-advisors of the author, Courtney and Elliot Paquette. Some of the results can be found in [32] and another project with the same collaborators that is in progress. We delineate the author's original work by chapter.

Chapter 3 focuses on the main theorem of exact dynamics. These results were established by the author and his supervisors Courtney and Elliot Paquette. These results follow from similar proof techniques established by Kiwon Lee in [32]. This will be the most technically demanding chapter. First, we provide the formal problem setup. Second, we provide assumptions on the data and the models under consideration. Third, we present the main proposition of the chapter which shows that the iterates of mini-batch gradient-based methods can be written as a linear combination of polynomials. Fourth, we prove the main theorem which illustrates that the dynamics of every mini-batch-gradient-based algorithm under quadratic loss functions can be captured by exact dynamics.

Chapter 4 shows the main theorem in practice by providing an exact dynamic analysis of mini-batch Polyak momentum. The main theorem which enables such an analysis was established by Kiwon Lee. With this result, the author provided experiments illustrating the main theorem and its implications as well as established results concerning the selection of optimal hyperparameters and conditions for convergence. We show that the exact dynamics is captured by a deterministic function ψ that solves the Volterra integral equation (2.15) and is completely determined by the eigenspectrum of the problem. This enables us to address several open fundamental questions regarding Polyak momentum:

1. What are the conditions of convergence?

2. How does the rate of convergenece depend on batch size?

3. How does one select optimal hyperparameters?

Table of Contents

Abstract i

Abrégé iii

Acknowledgements v

Contributions vii

List of Figures xvii

List of Tables xviii

1 Introduction **1**

1.1 First order algorithms and risk minimization 2

1.1.1 Problem statement 3

1.1.2 First order methods 8

1.1.3 Examples of mini-batch gradient-based methods. 9

1.2 Motivation for random matrix theory 11

2 Preliminaries **14**

2.1 Strong convexity and implications 14

2.1.1 First order upper bound 16

2.1.2 First order lower bound 16

2.1.3 The eigenvalue connection 17

2.2 Worst-case analysis of gradient-based methods 17

2.2.1 First example: gradient descent's suboptimal rate 18

2.2.2 Second example: stochastic gradient descent 20

2.2.3 Shortcomings of worst-case analysis . . . 27
2.3 Momentum machinery on quadratics: rates and batchsizes . . . 28
2.3.1 Full-batch rates . . . 29
2.4 Resolvents . . . 30
2.5 Volterra equation . . . 33
2.6 Exact dynamics: motivating examples . . . 34
2.7 Concentration of measure . . . 36
2.8 Overview . . . 38

3 Exact Dynamic Results **40**
3.1 Formal problem set-up . . . 40
3.2 From optimization algorithms to polynomials . . . 42
3.2.1 Examples of polynomials . . . 48
3.3 Resolvents and Statistics . . . 54
3.4 Main Theorem . . . 55
3.5 Motivating applications . . . 65
3.5.1 Training loss . . . 66
3.5.2 Generalization . . . 67
3.5.3 Random features . . . 69
3.6 Martingale Bounds . . . 71
3.7 Martingale error terms are small . . . 78
3.7.1 Error from the gradient term $\mathcal{E}_t^{\nabla}$ . . . 80
3.7.2 Error from the on diagonal term $\mathcal{E}_t^{\nabla^2\text{-Diag}}$ . . . 83
3.7.3 Error from the off-diagonal term . . . 101
3.8 Summary . . . 110

4 Exact dynamics: Polyak heavy-ball **112**
4.1 Setting and assumptions . . . 112
4.2 Deterministic Equivalent of Polyak heavy-ball . . . 114

4.3 Convolution Volterra analysis . . . 117
4.3.1 The Malthusian exponent and complexity . . . 119
4.3.2 Two regimes for the Malthusian exponent . . . 119
4.4 Performance of SGD+M: implicit conditioning ratio (ICR) . . . 122
4.5 Proofs of results . . . 126
4.5.1 Derivation of the dynamics of Polyak momentum (SGD+M) . . . 127
4.5.2 Change of basis . . . 128
4.5.3 Evolution of f . . . 129
4.6 Estimates based on concentration of measure on the high–dimensional orthogonal group . . . 138
4.6.1 Control of the errors from the Initial Conditions . . . 139
4.6.2 Control of the beta errors . . . 141
4.6.3 Control of the Key lemma errors . . . 142
4.6.4 Control of the Martingale error . . . 143
4.6.5 Proof of Theorem 9 . . . 153
4.7 Proof of Main Results . . . 154
4.7.1 Learning rate assumption and kernel bound . . . 154
4.7.2 Malthusian exponent and convergence rate . . . 158
4.7.3 Choice of optimal learning rate and momentum . . . 160
4.8 Numerical Simulations . . . 166

List of Figures

1.1 State of the art models exhibit a positive correlation between number of training samples and complexity of the model over time. The blue line denotes the line $y = x$ which alludes to a linear relationship between the complexity of the models and time. ChatGPT3 [1] and AlphaGo Zero [52] are shown as black dots. This figure is generated using the *Parameter, Compute and Data Trends in Machine Learning* dataset found on *Lesswrong.com* . . 3

1.2 State of the art models exhibit quadratic scaling of compute expenses (# of flops) with respect to the complexity of the model (number of parameters). From Figure 1.1, we observed a linear trend between time and complexity of models. These two figures, in conjunction, illustrate the necessity of efficient optimization methods as we continue to build more complex models. This figure is generated using the *Parameter, Compute and Data Trends in Machine Learning* dataset found on *Lesswrong.com* 4

2.1 Worst-case analysis captures algorithmic performance on pathological instances but do not reflect typical behaviour of algorithms. Figure from [48]. 28

2.2 Probability density function of the Marchenko-Pastur law with varying dimensionality ration $r = d/n$. 32

2.3 Empirical distribution of eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}/n$ where $\boldsymbol{A}$ is a matrix with iid standard Gaussian entries. 32

2.4 Concentration of SGD+M on a Gaussian random least squares problem. . . 34

3.1 Following the setup in Section 3.5.1 we can predict the training loss under the iterates of Nesterov acceleration under the strongly convex setting. We specified the data matrix $\boldsymbol{A}$ as iid standard Gaussian such that the rows are standardized accordingly. 65
3.2 Following the setup in Section 3.5.2 we can predict the test loss under the iterates of Nesterov acceleration under the strongly convex setting. We specified the data matrix $\boldsymbol{A}$ as iid standard Gaussian such that the rows are standardized accordingly. 67

4.1 **Different convergence rate regions: problem constrained regime versus algorithmically constrained regime** for Gaussian random least squares problem with $(n = 2000 \times d = 1000)$. Plots are functions of momentum (x-axis) and learning rate (y-axis). Analytic expression for $\lambda_{2,\max}$ (see (4.6), (4.38)) – convergence rate of forcing term $F(t)$ – given in (top row, column 1) represents the problem constrained region. (top row, column 2) plots 1/(Malthusian exponent) ((4.13), for details see Subsection 4.8); black region is where the Malthusian exponent Ξ does not exist. This represents the algorithmically constrained region. Finally, (top row, column 3 and bottom row) plots convergence rate of SGD+M $= \max\{\lambda_{2,\max}, \Xi^{-1}\}$, (see (4.14)), for various batch fractions. When the Malthusian exponent does not exist (black), $\lambda_{2,\max}$ takes over the convergence rate of SGD+M; otherwise the noise in the algorithm (i.e. Malthusian exponent Ξ) dominates. Optimal parameters that maximize $\lambda_{2,\max}$ denoted by Polyak parameters (orange circle, (4.15)) and the optimal parameters for SGD+M (orange dot); below red line is the problem constrained region; otherwise the algorithmic constrained region. When batch fractions $\zeta = 0.85$ and $\zeta = 0.7$ (top row and bottom row, column 1) (i.e., large batch), the SGD+M convergence rate is the deterministic momentum rate of $1/\sqrt{\kappa}$. As the batch fraction decreases ($\zeta = 0.25$), the convergence rate becomes that of SGD and the optimal parameters of SGD+M and Polyak parameters are quite far from each other. The Malthusian exponent (algorithmically constrained region) starts to control the SGD+M rate as batch fraction $\to 0$. 121

4.2 For each value of the batch fraction, ζ, we run SGD+M for 50 iterations on (normalized) MNIST data set using a random features set-up with Gaussian weight matrix $\boldsymbol{W} \in \mathbb{R}^{784\times d}$ (see App. 4.8 for details) and targets odd/even. We record the function value of the last iterate. The momentum and learning rate parameters are set to be near-optimal (4.20). Gray dot is the computed ICR value. At the predicted $\zeta =$ ICR (gray dot), there is a change in the behavior of the last iterate. For $\zeta \leq$ ICR, the value of the last iterate monotonically decreases until it hits the ICR. For $\zeta \geq$ ICR, we see no improvement in the value of the last iterate. This agrees with the theory that the convergence rate does not change. 123

4.3 **SGD+M vs. Theory on even/odd MNIST.** MNIST ($60,000 \times 28 \times 28$ images) [31] is reshaped into a single matrix of dimension $60,000 \times 784$ (preconditioned to have centered rows of norm-1), representing 60,000 samples of 10 digits. The target $\boldsymbol{b}$ satisfies $b_i = 0.5$ if the i^{th} sample is an odd digit and $b_i = -0.5$ otherwise. SGD+M was run 10 times with $(\Delta = 0.8, \gamma = 0.001, \zeta = 0.5)$ and the empirical Volterra was run once with $(R = 11,000$, $\tilde{R} = 5300)$. The 10^{th} to 90^{th} percentile interval is displayed for the loss values of 10 runs of SGD+M. While MNIST data set does not satisfy our eigenvalue assumption on the data matrix, the solution to the Volterra equation on MNIST data set captures the dynamics of SGD+M. See Subsection 4.8 for more details. 126

4.4 **SGD+M vs. Theory on even/odd MNIST.** MNIST ($60,000 \times 28 \times 28$ images) [31] is reshaped into a single matrix of dimension $60,000 \times 784$ (preconditioned to have centered rows of norm-1), representing 60,000 samples of 10 digits. The target $\boldsymbol{b}$ satisfies $\boldsymbol{b}_i = 0.5$ if the i^{th} sample is an odd digit and $\boldsymbol{b}_i = -0.5$ otherwise. SGD+M was run 10 times with $\Delta = 0.8$, various values of ζ, and learning rates $\gamma = 0.005, 0.001, 0.0005$ (left to right, top to bottom) and empirical Volterra was run once with ($R = 11,000$, $\tilde{R} = 5300$). The R and $\tilde{R}$ values were found by running a grid-search. The 10^{th} to 90^{th} percentile interval is displayed for the loss values of 10 runs of SGD+M. Volterra predicts the convergent behavior of SGD+M in this setting. 168

4.5 **Different convergence rate regions for MNIST dataset.** Plots are functions of momentum (x-axis) and learning rate (y-axis). Optimal parameters that maximize $\lambda_{2,\max}$ denoted by Polyak parameters (orange circle, (4.15)) and the optimal parameters for SGD+M (orange dot); below red line is the problem constrained region; otherwise the algorithmic constrained region. The MNIST data set is standardized. As the batch fraction decreases (left $\zeta = 0.7$ to right $\zeta = 0.25$), the optimal parameters of SGD+M and Polyak parameters are quite far from each other. The Malthusian exponent (algorithmically constrained region) starts to control the SGD+M rate as batch fraction $\to 0$. 170

4.6 **Convergence behavior near the ICR.** For each value of batch fraction ζ, run SGD+M for 20 (left) and 50 (right) iterations (colored lines – blue, green, and purple) and record the function value of the last iterate. The momentum and learning rate parameters are set to be near-optimal (see (4.20)). Gray dot is the computed ICR (4.22), ζ value. Data matrix $\boldsymbol{A} \in \mathbb{R}^{d,n}$ Gaussian entries, $\tilde{\boldsymbol{x}} \sim N(\mathbf{0}, 1/n\mathbf{I}_d)$, $\boldsymbol{x}_0 = \mathbf{0}$ (R = 1.0), and $\boldsymbol{\eta} \sim N(\mathbf{0}, 0.0001/n\mathbf{I}_n)$ ($\tilde{R} = 0.0001$). Different colored lines (blue, green, purple) correspond to running SGD+M with different values of the ratio d/n. At the predicted $\zeta =$ ICR (gray dot), there is a noticeable change in the behavior of the last iterate. For ζ values less than the ICR, the value of the last iterate gets smaller as ζ increases. Then the batch fraction ζ hits the ICR and we see little to no improvement in the value of the last iterate. This agrees *exactly* with our theory for batch fraction saturation (Prop 15 and Prop. 16). For $\zeta \geq$ ICR, the convergence rate does not change; thus the values of the last iterates are approximately all equal in this regime. For $\zeta <$ ICR, our theory predicts the convergence rate improves as $\zeta \to$ ICR, $\mathcal{O}(\zeta/\bar{\kappa})$. Hence the value of the last iterate decreases here. Moreover (left), SGD+M dynamics match the predicted last value given by the Volterra equation (red) (see Thm 7). 171

4.7 **Convergence rate regions for Gaussian random least squares.** Same setup as in Figure 4.1 but for a wider range of batch fractions. 172

List of Tables

2.1 Convergence rates of various algorithms with momentum under the quadratic loss function. 29

Chapter 1

Introduction

In the age of data-driven endeavours, optimization for algorithms plays a pivotal role in many recent scientific ventures. This is prevalent in many recent discoveries. AlphaGo [51], the first computer program to defeat a Go world champion, was (in part) trained on a dataset of 30 million positions each drawn from a unique game of self-play of Go. AlphaFold [29] which solved the 50-year-old protein folding problem [7] was trained on a large dataset of approximately 29,000 proteins. ChatGPT [1], a large language model trained on hundreds of billions of words, provides, arguably, the most sophisticated AI text interactions. The unifying theme amongst these three impressive models: They were all trained on large-dimensional datasets driven by mini-batch gradient-based algorithms (see Definition 2).

The growing necessity for efficient machine learning optimization methods is clear, but our theoretical understanding of even the simplest of these methods are rather limited. This is especially true for industry-standard algorithms such as stochastic gradient descent (SGD) and its variants which are known for their practical use in large-scale data settings (e.g. AlphaGo used SGD [51]). Conventional analysis of these algorithms [12, 13, 34] result in worst-case complexity bounds which do not capture the empirical behaviour of these algorithms [48]. In Section 2.2.3 we explain why the worst-case bounds are limited.

In response, this thesis intends to contribute to a better understanding of the fundamental properties of these algorithms through the perspective of *exact dynamics* (See Subsections 1.2 and 2.6). We take a physics-inspired approach by leveraging the *universality* phenomenon of random matrix theory [17] which allows us to address several fundamental problems concerning the celebrated Polyak momentum algorithm [49]: How does the convergence rate depend on batch-size? How should one select the optimal hyperparameters (stepsize, momentum, etc.)? What is the limiting behaviour of the algorithm?

In this chapter, we briefly introduce the main themes of the paper: Empirical risk minimization, gradient-based methods and its variants, random matrix theory, concentration of measure, and exact dynamics. In Chapter 2, we refine our understanding of the main themes by presenting previous lines of work to motivate the contributions of this thesis, relevant mathematical tools, and making certain notions like exact dynamics more precise.

1.1 First order algorithms and risk minimization

In 1951, Robbins and Monro constructed an iterative algorithm to solve a root-finding problem where the function (not directly observable) is represented by an expected value of an observable random variable [50]. Their work pioneered the field of stochastic approximation methods: A family of model-free algorithms used to seek the extrema of functions that are unobservable, but where only noisy observations are available. Our focus is on a subclass of these methods known as mini-batch gradient descent (see Definition 2). These methods have recently found extensive applications in the field of machine learning and statistics, particularly in large-scale data settings. These methods have far-reaching consequences, from enabling algorithms to be differentially private [53, 56] to driving the learning process behind transformer architectures [19] which form the backbone of many novel artificial intelligence applications (GPT4, Alphafold, etc.).

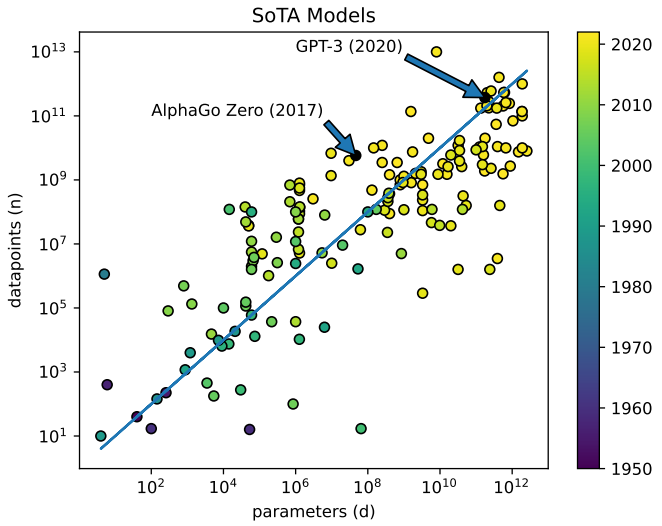

Figure 1.1: State of the art models exhibit a positive correlation between number of training samples and complexity of the model over time. The blue line denotes the line $y = x$ which alludes to a linear relationship between the complexity of the models and time. ChatGPT3 [1] and AlphaGo Zero [52] are shown as black dots. This figure is generated using the *Parameter, Compute and Data Trends in Machine Learning* dataset found on *Lesswrong.com*

In this section, we provide the canonical machine learning optimization problem statement, the definition of mini-batch methods, and conclude with examples of popular algorithms.

1.1.1 Problem statement

Optimization algorithms are the workhorses that drive the learning process of machine learning and statistical models. We formalize this notion in the *supervised machine learning* context which is the setting of this thesis.

Supervised learning and parametric risk minimization Supervised machine learning seeks a prediction model $h : \mathcal{A} \to \mathcal{B}$ so that given $\boldsymbol{a} \in \mathcal{A}$, the model's output $h(\boldsymbol{a})$ is reflective of the truth $b \in \mathcal{B}$. Conventionally, we assume there exists a joint distribution

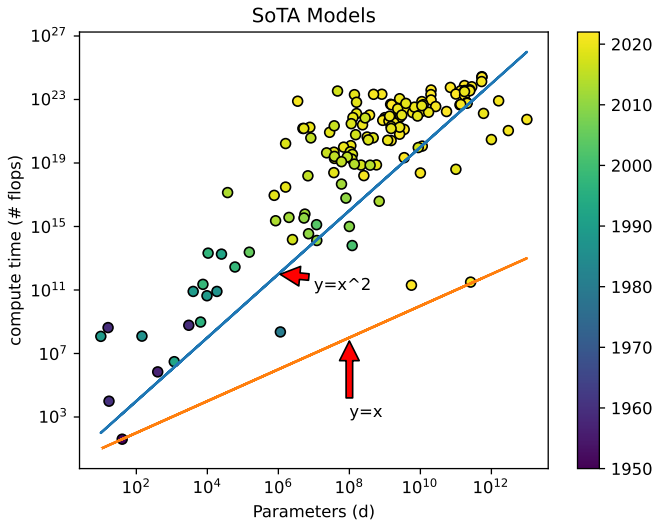

Figure 1.2: State of the art models exhibit quadratic scaling of compute expenses (# of flops) with respect to the complexity of the model (number of parameters). From Figure 1.1, we observed a linear trend between time and complexity of models. These two figures, in conjunction, illustrate the necessity of efficient optimization methods as we continue to build more complex models. This figure is generated using the *Parameter, Compute and Data Trends in Machine Learning* dataset found on *Lesswrong.com*

$P(\boldsymbol{a}, b)$ that captures the true relationship between $\boldsymbol{a}$ and b. That is, we assume the datum we observe $(\boldsymbol{a}, b)$ is drawn from $P(\boldsymbol{a}, b)$. In order to quantify the accuracy of a model $h \in \mathcal{H}$, one designs a *loss function* $\ell : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$ that compares the prediction of the model $h(a)$ against the truth b and outputs an error value. The expected error value of $h(\boldsymbol{a})$ against b over the distribution $P(\boldsymbol{a}, b)$ is known as the *expected risk*. In symbols, the expected risk of a model $h \in \mathcal{H}$ is

$$R(h) = \mathbb{E}_{(a,b)\sim P(a,b)} \left[\ell \left(h(a), b \right) \right] \qquad \text{(Expected Risk)}. \tag{1.1}$$

Needless to say, the goal is to find $h \in \mathcal{H}$ to minimize (1.1).

Often, the distribution $P(a, b)$ is too complex to compute (1.1). Instead, we draw n samples from $P(a, b)$ which forms our dataset $\{(a_i, b_i)\}_{i=1}^{n}$ and estimate (1.1) by computing the *empirical risk* of $h \in \mathcal{H}$

$$R_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(h(a_i, b_i)\right) \qquad \text{(Empirical Risk)}. \tag{1.2}$$

A popular class of models in statistics and machine learning, and of interest to us, are the parametric models. That is, each model h is parameterized by a real vector $\boldsymbol{x} \in \mathbb{R}^d$ and optimization of h is done over $\boldsymbol{x}$. For simplicity, assume $h : \mathbb{R}^{d_{\boldsymbol{a}}} \times \mathbb{R}^d \to \mathbb{R}$ and $y \in \mathbb{R}$. The parametric class $\mathcal{H}$ satisfies the structure

$$\mathcal{H} \stackrel{\text{def}}{=} \{h(\cdot, ; \boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^d\}.$$

Then (1.1) and (1.2) are translated, respectively, to

$$\begin{aligned} \text{Expected Risk} = R(\boldsymbol{x}) &= \mathbb{E}\left[\ell\left(h(\boldsymbol{a}_i; \boldsymbol{x}), b_i\right)\right] \\ \text{Empirical Risk} = R_n(\boldsymbol{x}) &= \frac{1}{n}\sum_{i=1}^{n} \ell\left(h(\boldsymbol{a}_i; \boldsymbol{x}), b_i\right) \end{aligned} \tag{1.3}$$

for a given model $h \in \mathcal{H}$. We assume that $P(\boldsymbol{a}, b)$ is too complex to compute the expected risk. Hence, we use the empirical risk R_n to guide us to a desirable model $h \in \mathcal{H}$.

Discrete-sum loss. A natural question is: How do we use the empirical risk to find a desirable parametric model $h(\cdot; \cdot) \in \mathcal{H}$? Before answering this question, we simplify our notation. Given a parametric model $h \in \mathcal{H}$, a fixed parameter $\boldsymbol{x} \in \mathbb{R}^d$ and a dataset $\{(\boldsymbol{a}_i, b_i)\}_{i=1}^n$ consisting of n independent samples drawn from $P(\boldsymbol{a}, b)$ we denote

$$f_i(\boldsymbol{x}) = \ell\left(h(\boldsymbol{a}_i, ; \boldsymbol{x}), b_i\right) \tag{1.4}$$

as the loss incurred under the ith sample. Then the empirical risk (1.3), under the realized dataset $\{(\boldsymbol{a}_i, b_i)\}_{i=1}^n$, can be written as

$$R_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}). \tag{1.5}$$

To answer the original question, we find a desirable model $h \in \mathcal{H}$ by optimizing (1.5) as

$$\arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}) \tag{1.6}$$

for a given data set $\{(\boldsymbol{a}_i, b_i)\}_{i=1}^n$ drawn independently from $P(\boldsymbol{a}, b)$. The form of (1.6) is known as the *discrete-sum loss* and is a standard optimization problem in machine learning. Solving (1.6) is the essence of the learning process of machine learning models in the supervised setting. In order to develop algorithms to find a solution to (1.6) under a reasonable number of computations, we impose some structure on f and turn to the theory of convex optimization.

Convex optimization. Suppose the loss $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable and strongly convex where the function $f(\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n f_i(\boldsymbol{x})$. Then (1.6) can be phrased as

$$\arg\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}), \qquad \text{where } f \in C^2 \text{ and strongly convex.} \tag{1.7}$$

The structure of (1.7) is known as the *unconstrained (strongly) convex optimization problem*. This problem is well-studied and the convex optimization literature [13,15] presents many efficient algorithms to solve (1.7). The conditions imply there exists an optimal value $\boldsymbol{x}^*$ so that $\inf_{\boldsymbol{x}} f(\boldsymbol{x}) = \boldsymbol{x}^*$ and $\nabla f(\boldsymbol{x}^*) = 0$. In some cases, one may solve (1.7) analytically by solving a set of d equations in d variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d$. However, even if one were able to do so, if d were large then the computational costs become quickly infeasible. Instead, we aim to solve (1.7) using iterative algorithms. That is, algorithms that return a sequence of iterates $\{\boldsymbol{x}_k : k \geq 0\} \subseteq \mathbb{R}^d$ with the hope that for small $\epsilon > 0$ there exists a moderate $N \in \mathbb{N}$ so that for every $k \geq N$ we have $d\left(f(\boldsymbol{x}_k), f(\boldsymbol{x}^*)\right) \leq \epsilon$ with respect to some metric $d : \mathbb{R}^2 \to \mathbb{R}$. This thesis focuses on a subclass of iterative methods known as the mini-batch gradient-based methods. As these are probabilistic algorithms it is necessary to convert the notion of deterministic convergence to its probabilistic analog (See Subsection 2.6).

Before introducing the methods that solve (1.7) we provide two standard examples of (1.7).

Examples of loss functions. We introduce two popular models that are extensively used throughout the statistics literature: the least squares model, and the logistic regression model which motivate the need for iterative methods.

Least-squares. The least squares problem has been extensively studied from a statistical, numerical, and optimization perspective [5, 13, 23]. Formally, the *least squares problem* is defined as:

$$\arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ f(\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2 \right\} \tag{1.8}$$

for a given the data matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ and labels $\boldsymbol{b} \in \mathbb{R}^n$. In the case that $\boldsymbol{A}$ is full rank then one can solve (1.8) analytically as

$$\boldsymbol{x}^* = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{b} \tag{1.9}$$

where $\boldsymbol{x}^*$ satisfies $f(\boldsymbol{x}^*) = 0$. The computational bottleneck of (1.9) is matrix multiplication which is on the order of $\mathcal{O}(n^2 \cdot d)$ operations. If the linear system is inconsistent then there is no solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, i.e., one cannot drive the loss function in (1.8) to zero. In this case, one can solve (1.8) by applying the Penrose-inverse of the data matrix $\boldsymbol{A}^\dagger$ onto $\boldsymbol{x}$, i.e.,

$$\boldsymbol{x}^* = \boldsymbol{A}^\dagger\boldsymbol{x} \qquad \text{solves (1.8) when } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \text{ is inconsistent.}$$

Finding $\boldsymbol{A}^\dagger$ requires using the singular value decomposition (SVD) which has computational cost on the order of $\mathcal{O}(n \cdot d^2)$. Hence, these analytic methods become infeasible when n and d become large. One solution is to turn to first order methods such as gradient descent which has computation cost on the order of $\mathcal{O}(N \cdot n \cdot d)$ where N is the number of iterates returned by the algorithm. We will explore such methods in the next section.

Binary logistic regression. There are situations where the loss function has no analytic solution, as for example we are trying to solve a binary prediction problem. That is, our data consists of the input $\boldsymbol{a}_i \in \mathbb{R}^d$ and we seek to predict the true labels $y_i \in \{0, 1\}$. One natural model to approach solve this task is logistic regression with the corresponding loss

$$\arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ f(\boldsymbol{x}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} y_i \log\left(\frac{1}{1+e^{-\boldsymbol{x}^T\boldsymbol{a}_i}}\right) + (1-y_i)\log\left(1-\frac{1}{1+e^{-\boldsymbol{x}^T\boldsymbol{a}_i}}\right)\right\}. \tag{1.10}$$

Therefore, one solves (1.10) using iterative methods.

In the next section, we introduce a special class of iterative methods known as the *first order iterative methods.* In Section 2.2, we show that first order iterative methods can actually solve (1.7).

1.1.2 First order methods

First order iterative methods (also known as gradient-based methods) form the standard toolkit used to solve (1.7). In essence, they are algorithms which can be written as a linear combination of the previous gradients and the initial iterate.

Definition 1 (Gradient-based method)**.** A deterministic optimization algorithm is called a *gradient-based method* if each update of the algorithm can be written as a linear combination of the initial iterate and previous gradients. In other words, if every update is of the form

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_0 + \sum_{j=0}^{k} c_{k,j} \nabla f(\boldsymbol{x}_j), \tag{1.11}$$

for some scalar values $c_{k,j}$.

Intuitively, (1.11) tells us that gradient-based methods leverage information about the initial iterate $\boldsymbol{x}_0$ and the past trajectory of the first order information about f in order to form the next iterate.

Every gradient-based method has a natural stochastic correspondence. Specifically, we can approximate the full gradient $\nabla f(\boldsymbol{x}_i)$ with a stochastic version, that is, we sample a batch $B_j \subset \{1, 2 \ldots, n\}$ of cardinality β uniformly at random without replacement, and approximate $\nabla f(\boldsymbol{x}_j)$ with $\sum_{i \in B_j} \nabla f_i(\boldsymbol{x}_j)$. This allows us to define a *mini-batch gradient-based method*.

Definition 2 (Mini-batch gradient-based method)**.** Given a gradient-based algorithm, we define a stochastic optimization algorithm, called a *mini-batch gradient-based method*, if at each update one generated uniformly at random a batch $B_i \subset \{1, 2, \ldots, n\}$ and the update satisfies,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_0 + \sum_{j=0}^{k} c_{k,j} \sum_{i \in B_j} \nabla f_i(\boldsymbol{x}_j) \tag{1.12}$$

for some scalars $c_{k,i}$. In the special case that the batch size $|B_i| = 1$ for all $i \geq 0$, (1.12) is known as a *stochastic gradient-based method*.

Indeed, the definition of mini-batch gradient-based methods generalize the notion of gradient-based methods. However, a special class of mini-batch gradient-based methods that are important to us are those that incorporate the momentum machinery. We provide examples in the next section.

1.1.3 Examples of mini-batch gradient-based methods.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and continuously differentiable function and denote $f_i(\boldsymbol{x})$ as the loss function incurred from the ith sample. To make Definition 2 more concrete, we provide a few examples of popular mini-batch gradient-based methods.

Stochastic gradient descent (SGD). Stochastic gradient descent [50] is the most basic stochastic gradient-based method. Its iterates evolve as

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} - \gamma \sum_{i \in B_k} \nabla f_i(\boldsymbol{x}_{k-1})$$

where $\gamma > 0$ is the stepsize (or learning rate) parameter.

Stochastic heavy-ball. The iterates of stochastic heavy-ball [49] are

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} - \gamma \sum_{i \in B_k} \nabla f_i(\boldsymbol{x}_{k-1}) + \Delta(\boldsymbol{x}_{k-1} - \boldsymbol{x}_{k-2}) \tag{1.13}$$

with momentum parameter $\Delta \geq 0$, stepsize $\gamma > 0$, and some fixed $\boldsymbol{x}_0 \in \mathbb{R}$. The momentum parameter influences the extent to which the direction of the previous update influences the next one. An alternative view of the heavy ball method is obtained by expanding (1.13):

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} - \gamma \sum_{j=1}^{k-1} \Delta^{k-j-1} \sum_{i \in B_k} \nabla f_i(\boldsymbol{x}_{k-1}).$$

This gives the interpretation that each step is an exponentially decaying average of past gradients. Roughly speaking, this smoothes out the trajectory taken by $\{\boldsymbol{x}_k : k \geq 0\}$ via an exponential smoothing effect.

Stochastic Nesterov acceleration Nesterov's accelerated method [40] generates iterates satisfying the recurrence for $k \geq 1$

$$\boldsymbol{x}_{k+1} = \big[1 + \Delta_{k-1}\big]\big[\boldsymbol{x}_k - \gamma \sum_{i \in B_{k+1}} \nabla f_i(\boldsymbol{x}_k)\big] - \Delta_{k-1}\big[\boldsymbol{x}_{k-1} - \gamma \sum_{i \in B_k} \nabla f_i(\boldsymbol{x}_{k-1})\big] \tag{1.14}$$

with the initial iteration such that $\boldsymbol{x}_0 \in \mathbb{R}^d$ and one gradient step for $\boldsymbol{x}_1$, that is, $\boldsymbol{x}_1 = \boldsymbol{x}_0 - \gamma \nabla f(\boldsymbol{x}_0)$. The learning rate $\gamma > 0$ and momentum parameter $\Delta(k) \geq 0$ satisfy

$$\Delta_k = \begin{cases} \frac{k}{k+3} & \text{if } f \text{ is convex} \\ \Delta & \text{if } f \text{ is strongly convex.} \end{cases} \tag{1.15}$$

Each of these algorithms have been extensively studied through the lens of stochastic and ordinary differential equations [34]. Though useful for bounding the performance of algorithms, worst-case analysis results often have large discrepancies with empirical behaviour of algorithms [48]. One explanation of the discrepancy is that worst-case analysis often neglects the information encoded in high dimensional datasets and often only fixates on the extremum of the eigenspectrum of the problem. This thesis aims to analyze these algorithms by incorporating information about the entire spectrum of the problem. In order to do so, we leverage tools from random matrix theory.

1.2 Motivation for random matrix theory

In Section 1.1.1 we mentioned that given a class of parametric function $\mathcal{H} \stackrel{\text{def}}{=} \{h(\cdot; \boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^d\}$, a dataset $\{(\boldsymbol{a}_i, y_i)\}_{i=1}^n$ realized from independently sampling from $P(\boldsymbol{a}, y)$, and a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we would like to minimize the empirical risk (1.6). One way to investigate the structure of the problem (1.6) is to study the properties of the distribution of the data $P(\boldsymbol{a}, y)$. In order to do so, we can write the dataset $\{(\boldsymbol{a}_i, y_i)\}_{i=1}^n$ in the form of a matrix:

$$\boldsymbol{A} \stackrel{\text{def}}{=} \left(\begin{array}{ccc|c} — & \boldsymbol{a}_1 & — & y_1 \\ \cdots & \vdots & \cdots & \vdots \\ — & \boldsymbol{a}_n & — & y_n \end{array}\right) \in \mathbb{R}^{n \times (d+1)}$$

where the ith sample $(\boldsymbol{a}_i, y_i) \in \mathbb{R}^{d+1}$ forms the ith row of $\boldsymbol{A}$. Then the data matrix $\boldsymbol{A}$ consists of realizations of random variables distributed according to $P(\boldsymbol{a}, y)$. From a probabilistic perspective, i.e., treating each entry of $\boldsymbol{A}$ as a random variable, the matrix $\boldsymbol{A}$ forms a random matrix. That is, a random matrix $\boldsymbol{A}$ is just a matrix with random variables as entries. Then we can leverage results from random matrix theory to study the structure of $P(\boldsymbol{a}, y)$ through the random matrix $\boldsymbol{A}$.

Random matrix theory (RMT) originates from the study of the eigenspectrum of $\boldsymbol{A}\boldsymbol{A}^T$ with $\boldsymbol{A}_{ij}$ distributed as standard normal random variables [55]. The field rapidly pro-

gressed when Eugene Wigner determined the energy level spacings of nuclei can be analyzed by studying the eigenvalues of symmetric matrices $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with independent Bernoulli entries. Wigner opted for the asymptotic analysis in his study of the eigenvalue distribution. Thus began the branch of *large dimensional random matrix theory*, the study of the asymptotic behaviour of eigenspectrums of large dimensional random matrices. We make use of this line of approach to provide asymptotically exact approximations of the training performance of mini-batch gradient-based methods.

Large dimensional RMT provides asymptotically exact approximations. Large dimensional RMT has found many modern applications, from wireless communications, financial statistics, and the analysis of algorithms [17]. One can analyze complex systems by using a random matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ to model the system. By constraining the dimensional growth of a random matrix so that $d/n \to r \in (0, \infty)$, one can leverage large dimensional RMT theory which provides exact asymptotics of the eigenspectrum of $\boldsymbol{A}\boldsymbol{A}^T$. This allows one to precisely quantify a complex system's (e.g. training loss under stochastic algorithms) behaviour in terms of the eigenspectrum of $\boldsymbol{A}$.

In the context of optimization algorithms, we will show that by leveraging RMT, the training dynamics of many popular stochastic gradient-based methods can be captured exactly by a deterministic quantity depending only on the eigenspectrum of the data matrix $\boldsymbol{A}$ which has many practical applications.

Exact Dynamics. As previously mentioned, random matrix theory provides an (asymptotically) exact description of many complex systems. One such example (and the focus of this thesis) is determining the exact training dynamics of stochastic first order algorithms. By this, we mean that our aim is to construct a deterministic function $\psi(k) : \mathbb{R} \to \mathbb{R}$ that captures the behaviour of the loss function f. More precisely, if $\{x_k : k \geq 0\}$ are the iterates returned by running a stochastic algorithm on the realization of a random data

matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ then the approximation

$$\psi(k) \approx f(x_k) \qquad \text{for all } k \geq 0 \tag{1.16}$$

becomes equality with high probability in the high-dimensional limit of $\boldsymbol{A}$. This has many practical consequences such as determining the population risk of linear regression models and analyzing the behaviour of the random features model [3]. These notions will be made precise in Subsection 2.6.

Chapter 2

Preliminaries

The central theme of this thesis is analyzing mini-batch-gradient-based methods from the perspective of random matrix theory. As such, we will review some standard approaches of analyzing these methods (worst-case analysis) and why they may be lacking. Also, we introduce some important mathematical tools from random matrix theory and concentration of measure results that are relevant to our discussion. Finally, we make notions such as stochastic approximation and exact dynamics precise.

2.1 Strong convexity and implications

We will review the standard approaches of analyzing gradient-based and mini-batch-gradient-based methods under the context of minimizing a strongly convex function (see (1.7)). Bouttou et al. [12] lists three reasons why the analysis of strongly convex is a good starting point: First, it is particularly relevant for practical machine learning, as convex functions are often regularized by a strongly convex function. Second, often times the objective function is not globally (strongly) convex, but is so in the neighborhood of local minimizers, thus the analysis holds for such local regions of the search space. Lastly, the analysis of non-strongly convex models can be seen as extensions of the results of the

strongly-convex case, making the strongly-convex case a good starting point for deriving a more general theory.

The goal of this section is to gain some insight into the standard worse-case analysis of deterministic and stochastic first-order algorithms on strongly convex problems. In particular, to observe that standard worst-case analysis hinges on bounding the performance in terms of the smallest and largest eigenvalues of the Hessian.

We constrain our study to the case where the loss function $f : \mathbb{R}^d \to \mathbb{R}$ is strongly convex and twice continuously differentiable. The implications of these assumptions result in some of the conventional worst-case analysis of first-order algorithms. First we give the definition of strong convexity. Then we present two characterizations of strong convexity under different smoothness assumptions.

Definition 3. *The loss function $f : \mathbb{R}^d \to \mathbb{R}$ is strongly convex if there exists a constant $m > 0$ such that the function*

$$g(x) = f(x) - \frac{m}{2}\|x\|^2 \qquad \textit{is convex}, \qquad \forall x \in \mathbb{R}^d. \tag{2.1}$$

Definition 4. *We say that a twice continuously differentiable loss function $f : \mathbb{R}^d \to \mathbb{R}$ is m-strongly convex on $\mathbb{R}^d$ if there exists a constant $m > 0$ such that*

$$y^T \nabla^2 f(x) y \geq m y^T y$$

for all $y \in \mathbb{R}^d$. In particular, this implies the smallest eigenvalue of $\nabla^2 f$ is lower bounded by m.

Definition 5. *We say that a differentiable loss function $f : \mathbb{R}^d \to \mathbb{R}$ is m-strongly convex on $\mathbb{R}^d$ if there exists $m > 0$ such that*

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2}\|y - x\|^2. \tag{2.2}$$

This is clearly a stronger lower bound than convexity (which we recover if $m = 0$). Intuitively, strong convexity implies a quadratic lower bound on the growth of f.

As f is strongly convex, there exists an optimum x^* so that $\inf_x f(x) = f(x^*)$. Now we introduce a notion of convergence typically used in worst-case analysis.

Definition 6 (ϵ-suboptimal)**.** *The ϵ-suboptimal set $S_\epsilon \subseteq \mathbb{R}^d$ for a loss function $f : \mathbb{R}^d \to \mathbb{R}$ is the set of feasible points so that*

$$f(x) - f(x^*) \leq \epsilon, \qquad \forall x \in S_\epsilon.$$

Intuitively, given a strongly-convex optimization problem f, a user defines ϵ to be the error tolerance and S_ϵ to be a set of feasible solutions to the corresponding optimization problem.

2.1.1 First order upper bound

Strong convexity tells us that knowing the gradient and the strong convexity constant m suffices in indicating whether ϵ-suboptimality is satisfied or not. To see this, observe that the right-hand side of (2.2) is minimized when $y = x - (1/m)\nabla f(x)$ for fixed x. Plugging this into (2.2) yields

$$f(y) \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|^2, \qquad \text{for all } y \in \mathbb{R}^d.$$

Setting $y = x^*$ gives us the condition

$$f(x) - f(x^*) \leq \frac{1}{2m}\|\nabla f(x)\|^2. \tag{2.3}$$

Then ϵ-suboptimality is satisfied whenever the norm of the gradient $\|\nabla f(x)\|$ is sufficiently small. Conventionally, this is known as the first-order stopping criterion.

2.1.2 First order lower bound

The definition of C^2-smooth implies that on every compact set there exists $M > 0$ so that the Hessian $\nabla^2 f$ is upper bounded by M. Let m be the strong convexity constant for f as

seen in Definition 4. Then $m \leq M$ and for all $x, y \in \mathbb{R}^d$ the following holds:

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} \|y - x\|^2. \tag{2.4}$$

Fixing x and minimizing y on both sides gives us

$$f(x) - f(x^*) \geq \frac{1}{2M} \|\nabla f(x)\|^2.$$

which provides a lower bound given by the norm of the gradient.

2.1.3 The eigenvalue connection

From f strongly convex on a compact set $\mathbb{R}^d$ and C^2-smooth we have the inequality

$$m\|y\|^2 \leq \nabla^2 f(x) \leq M\|y\|^2 \qquad \forall x, y \in \mathbb{R}^d \tag{2.5}$$

for some $m, M > 0$. The constants M (m) are actually upper (lower) bounds on the largest (smallest) eigenvalue of $\nabla^2 f(x)$ for every $x \in \mathbb{R}^d$. Hence the ratio $\tilde{M} = M/m$ provides an upper bound on the condition number of $\nabla^2 f(x)$. In particular, if we set y to be an eigenvalue of $\nabla^2 f(x)$, then the tightest bound for (2.5) is replacing m and M with the smallest and largest eigenvalue, respectively, of $\nabla^2 f(x)$. This fact provides some insight into how worst-case analysis only incorporates information pertaining to the extremum of the eigenspectrum of $\nabla^2 f(x)$. In the next section, we make use of all these notions by starting with the worst-case analysis of gradient descent.

2.2 Worst-case analysis of gradient-based methods

Using the notions in the previous section, we show standard worst-case analysis results and provide their interpretations.

2.2.1 First example: gradient descent's suboptimal rate

The worst-case analysis of gradient descent will leverage the strongly-convex and smoothness constants $M, m > 0$. To begin, let $f : \mathbb{R}^d \to \mathbb{R}$ be a strongly convex and twice continuously differentiable function. We recall the gradient descent algorithm as

$$x_k = x_{k-1} - \gamma \nabla f(x_k) \tag{2.6}$$

where the fixed constant $\gamma > 0$ is called the step size. For ease of analysis, we assume we have an *optimal step size oracle* that tells us for every step k the optimal step size γ^* so that

$$\gamma^* \stackrel{\text{def}}{=} \arg\min_{\gamma \geq 0} \{f(x_k - \gamma \nabla f(x_k))\}$$

for any chosen initial iterate $x_0 \in \mathbb{R}^d$.

Theorem 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a strongly convex and twice continiously differentiable function with the aforementioned parameters m, M. Let $\{x_k : k \geq 0\}$ be the iterates generated by* (2.6)*. For any $\epsilon > 0$ we have ϵ-suboptimality, i.e. $f(x_k) - f(x^*) \leq \epsilon$ is satisfied whenever*

$$k \geq \frac{\log\left(\frac{f(x_0) - f(x^*)}{\epsilon}\right)}{\log\left(\frac{1}{1 - m/M}\right)}. \tag{2.7}$$

Proof. Since f is strongly convex on $\mathbb{R}^d$ we have from (2.4),

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2}\|y - x\|^2 \qquad \forall x, y \in \mathbb{R}^d.$$

Setting $y = x_k - \gamma \nabla f(x_k)$ and $x = x_k$ gives

$$\begin{aligned} f\left(x_k - \gamma \nabla f(x_k)\right) &\leq f(x_k) + \nabla f(x_k)^T \left(-\gamma \nabla f(x_k)\right) + \frac{M}{2}\| - \gamma \nabla f(x_k)\|^2 \\ &= f(x_k) - \gamma \|\nabla f(x_k)\|^2 + \frac{M\gamma^2}{2}\|\nabla f(x_k)\|^2. \end{aligned}$$

Choosing stepsize $\gamma = \frac{1}{M}$ results in

$$f\left(x_k - \gamma \nabla f(x_k)\right) \leq f(x_k) - \frac{1}{2M}\|\nabla f(x_k)\|^2.$$

By the minimality of γ^*,

$$f\left(x_k - \gamma^* \nabla f(x_k)\right) \leq f(x_k) - \frac{1}{2M}\|\nabla f(x_k)\|^2.$$

Under the *optimal step size oracle* assumption,

$$f\left(x_{k+1}\right) \leq f(x_k) - \frac{1}{2M}\|\nabla f(x_k)\|^2.$$

Subtracting $f(x^*)$ on both sides and applying the bound $\nabla^2 f(x) \geq 2m\left(f(x) - f(x^*)\right)$ (see (2.3)) yields

$$\begin{aligned} f(x_{k+1}) - f(x^*) &\leq f(x_k) - f(x^*) - \frac{1}{2M}\|\nabla f(x_k)\|^2 \\ &\leq f(x_k) - f(x^*) - \frac{m}{M}\left(f(x_k) - f(x^*)\right) \\ &= \left(1 - \frac{m}{M}\right)\left(f(x_k) - f(x^*)\right). \end{aligned}$$

Unravelling the recurrence gives us the relation

$$f(x_k) - f(x^*) \leq \left(1 - \frac{m}{M}\right)^k \left(f(x_0) - f(x^*)\right)$$

which implies the desired ϵ-suboptimality condition. □

Equation (2.7) tells us the number of iterations to ϵ-suboptimality relies on the initialization of the algorithm $x_0 \in \mathbb{R}^d$, the upper bound on the condition number m/M of the matrix $\nabla^2 f(x)$, and the user-defined error tolerance ϵ.

2.2.2 Second example: stochastic gradient descent

This section provides a standard worst-case analysis of stochastic gradient descent with fixed step sizes. The proof shown here can be readily generalized to a larger class of stochastic iterative methods as seen in [12].

Let $f : \mathbb{R}^d \to \mathbb{R}$ be strongly convex and twice continuously differentiable. Following the setup introduced in Section 1.1.1, assume we have a data set $\{(a_i, b_i)\}_{i=1}^n$ sampled independently from a distribution $P(a, b)$, a class of parametric models

$$\mathcal{H} = \{h(\cdot, x) : x \in \mathbb{R}^d\},$$

and a loss function ℓ. We define the loss function incurred under the ith sample as

$$f_i(x) \stackrel{\text{def}}{=} \ell\left(h(\boldsymbol{a}_i, ; x), b_i\right).$$

Stochastic gradient descent (SGD) has iterates that evolve by selecting uniformly at random an index $i_k \in \{1, 2, \ldots, n\}$ and makes the update

$$x_{k+1} = x_k - \gamma \nabla f_{i_k}(x_k). \tag{2.8}$$

with some fixed initialization $x \in \mathbb{R}^d$. Instead of taking the full gradient (as in gradient descent), SGD only computes the gradient with respect to a single datum. Intuitively, SGD sacrifices the accuracy of the full gradient update in order for cheap computational costs as the cost is independent of the size of the dataset. On the other hand, gradient descent requires the computation of the full gradient which scales with the size of the data.

SGD is random because of its inherent sampling mechanism. Therefore to analyze its worst-case behaviour, we require a different set of assumptions than the deterministic (i.e. gradient descent) case. However, the strongly-convex and smoothness constants m, M will play a similar role as in the gradient descent case.

Digression to Probability Theory. We begin by briefly introducing some notions from probability theory. More details can be found in standard measure-theoretic probability textbooks [10, 20].

Definition 7 (σ-algebra)**.** *Let Ω be some set and let $P(\Omega)$ be its power set. A set $\mathcal{F} \subseteq P(\Omega)$ is a σ-algebra over Ω if and only if it satisfies the three properties:*

1. *$\Omega \in \mathcal{F}$.*
2. *If $A \in \mathcal{F}$ then $A^C \in \mathcal{F}$.*
3. *If $\{A_i : i \in I\} \subseteq \mathcal{F}$ forms a countable collection of sets in $\mathcal{F}$ then $\bigcup_{i \in I} A_i \in \mathcal{F}$.*

The set Ω is often referred to as the *sample space*. The notion of sample space allows us to think of outcomes of experiments, albeit in an abstract manner. One can think of Ω as all possible outcomes of an experiment. Then one can construct $\mathcal{F}$, a σ-algebra over Ω, and interpret a subset $A \in \mathcal{F}$ as an *event* of an experiment. The pair $(\Omega, \mathcal{F})$ is referred to as a *measurable space*.

Remark 1 (Generated σ-algebra)**.** *Given a random variable $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$, i.e. a mapping between two measurable spaces, the σ-algebra generated by X denoted $\sigma(X)$ is defined as*

$$\sigma(X) = X^{-1}(\mathcal{F}').$$

More generally, if for some index set I

$$X_i : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$$

then we denote by

$$\sigma(X_i : i \in I) = \sigma\left(\bigcup_{i \in I} \sigma\left(X_i\right)\right)$$

the smallest σ-algebra containing all $\sigma(X_i)$.

Definition 8 (Borel σ-algebra of $\mathbb{R}$)**.** *The Borel σ-algebra of $\mathbb{R}$, written $\mathcal{B}$, is the σ-algebra generated by the open sets. That is, if $\mathcal{O}$ denotes the collection of all open subsets of $\mathbb{R}$, then $\mathcal{B} = \sigma(\mathcal{O})$.*

One is often interested in quantifying the likelihood of an event occurring when running an experiment. To make this possible we endow the measurable space with a *probability measure*.

Definition 9 (Probability measure)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space. A probability measure* Pr *on the sample space $(\Omega, \mathcal{F})$ is a real-valued function defined on the collection of sets $\mathcal{F}$ satisfying*

1. $\Pr(A) \geq 0$ *for all* $A \in \mathcal{F}$.
2. $\Pr(\Omega) = 1$.
3. *If* $\{A_i : i \in I\} \subseteq \mathcal{F}$ *forms a countable and pairwise disjoint collection of sets then*

$$\Pr\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} \Pr(A_i).$$

Definition 10 (Probability space)**.** *A probability space is a measure space $(\Omega, \mathcal{F}, \Pr)$ where $\Pr(\Omega) = 1$ where*

1. *The set Ω is called the sample space.*
2. *The set of events $\mathcal{F}$ denotes the σ-algebra over Ω.*
3. *The measure* Pr *for the measurable space $(\Omega, \mathcal{F})$ denotes the probability measure.*

The probability space allows us to describe and assign probability measure to events of an experiment. However, this notion is static, in the sense that the assignment of probability measure to events are invariant with respect to time. However, in many cases, our intuition tells us that the probability of some events will vary as we gain more information as time elapses. The concept of *filtration* allows us to encode this intuition.

Definition 11 (Filtration)**.** *Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space. For every $k \in \mathbb{N}$ let $\mathcal{F}_k$ be a sub-σ-algebra of $\mathcal{F}$, that is, $\mathcal{F}_k$ is a subset of $\mathcal{F}$ and is a σ-algebra of Ω. Then $\{\mathcal{F}_k : k \geq 0\}$ is called a filtration if $\mathcal{F}_k \subseteq \mathcal{F}_\ell$ for all $k \leq \ell$.*

All stochastic quantities defined hereafter live on a probability space denoted by $(\Omega, \mathcal{F}, \Pr)$ with probability measure $\Pr$ and the σ-algebra $\mathcal{F}$ containing subsets of Ω. Let $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ be a random variable mapping into the Borel σ-algebra. We use the standard shorthand, $\sigma(X)$ for the σ-algebra generated by the random variable X.

Let $\{\mathcal{F}_k : k \geq 0\}$ be a filtration where $\mathcal{F}_k = \sigma(\boldsymbol{A}, \boldsymbol{b}, x_0, x_1, \ldots, x_k)$ be the σ-algebra generated by the SGD iterates up to time k and the data matrix $\boldsymbol{A}$ and label vector $\boldsymbol{b}$. We use the notation $\mathbb{E}[\cdot|\mathcal{F}_k]$ to denote the expectation taken with respect to the random variable i_k which samples uniformly at random from $\{1, \ldots, n\}$ conditioning on information provided by $\mathcal{F}_k$. Occasionally, we may use the overloaded notation $\mathbb{E}_{i_k}[\cdot|\mathcal{F}_k]$. The following lemma quantifies the Markovian behaviour of the expected change in the loss at time $k+1$.

Lemma 1. *Suppose f is strongly convex and twice continuously then the iterates of SGD satisfy the following inequality*

$$\mathbb{E}_{i_k}[f(x_{k+1})|\mathcal{F}_k] - f(x_k) \leq -\gamma \nabla f(x_k)^T \mathbb{E}_{i_k}[\nabla f_{i_k}(x_k)] + \frac{\gamma^2}{2} \mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(x_k)\|^2 | \mathcal{F}_k\right]$$

where the σ-algebra $\mathcal{F}_k \overset{def}{=} \sigma(\boldsymbol{A}, \boldsymbol{b}, x_0, x_1, \ldots, x_k)$ contains information about the trajectory of the algorithm up to time k.

Proof. Recall by strong convexity we have the bound

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} \|y - x\|^2.$$

Fixing $y = x_{k+1}$ and $x = x_k$ and using the SGD iterative relation (2.8) we obtain

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{M}{2} \|x_{k+1} - x_k\|^2 \\ &= -\gamma \nabla f(x_k)^T \nabla f_{i_k}(x_k) + \frac{M}{2} \| - \gamma \nabla f_{i_k}(x_k)\|^2. \end{aligned}$$

Taking conditional expectations,

$$\mathbb{E}_{i_k}\left[f(x_{k+1})|\mathcal{F}_k\right] - f(x_k) \leq -\gamma \nabla f(x_k)^T \mathbb{E}_{i_k}[\nabla f_{i_k}(x_k)|\mathcal{F}_k] + \frac{M\gamma^2}{2}\mathbb{E}\left[\|\nabla f_{i_k}(x_k)\|^2|\mathcal{F}_k\right]$$

where we used the fact that $f(x_k)$ and $\nabla f(x_k)$ are measurable with respect to $\mathcal{F}_k$ and do not depend on the random quantity i_k. □

Lemma 1 indicates the Markovian dyanmics of SGD: The expected change of the loss in the $(k+1)$th step only depends on the most recent past (i.e. x_k), the uniformly sampling i_k, stepsize γ, and the second moment of the stochastic gradient $\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2|\mathcal{F}_k]$ and ignores the path to x_k. Moreover, Lemma 1 gives a deterministic upper bound to this expected change.

Corollary 1. *Since ∇f_{i_k} is an unbiased estimator of the full-gradient $\nabla f(x_k)$ we have from Lemma 1*

$$\mathbb{E}_{i_k}[f(x_{k+1})|\mathcal{F}_k] - f(x_k) \leq -\gamma\|\nabla f(x_k)\|^2 + \frac{\gamma^2}{2}\mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(x_k)\|^2|\mathcal{F}_k\right] \tag{2.9}$$

Proof. To see the unbiasedness of the estimator, observe

$$E_{i_k}[\nabla f_{i_k}(x_k)|\mathcal{F}_k] = \sum_{i=1}^{n}\frac{1}{n}\nabla f_i(x_k) = \frac{1}{n}\nabla\sum_{i=1}^{n} f_i(x) = \nabla f(x_k).$$

and the result directly follows from Lemma 1. □

We obtain good behaviour of SGD if the right-hand side of (2.9) is small for each k. This motivates the need for assumptions on the second moment of $\nabla f_{i_k}(x_k)$. It turns out, convergence of SGD is guaranteed (up to a neighborhood of the optimum) if we have the following assumption.

Assumption 1. *Assume there exists constants $C_1 > 0$ and $C_2 \geq 1$ so that the following holds for all k,*

$$\mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(x_k)\|^2\middle|\mathcal{F}_k\right] \leq C_1 + C_2\|\nabla f(x_k)\|^2. \tag{2.10}$$

Assumption 1 says that at every time k the second moment of the stochastic gradient can be bounded by a deterministic quantity depending on the full gradient. That is, the stochastic gradient, in expectation, can be comparable to that of the full gradient. With this assumption, we can refine our deterministic upper bound to be of the following.

Lemma 2. *Under Assumption 1 and f twice continuously differentiable and strongly convex, the iterates of SGD satisfies the following for all $k \in \mathbb{N}$:*

$$\mathbb{E}_{i_k}\left[f(x_{k+1})|\mathcal{F}_k\right] - f(x_k) \leq -\left(1 - \frac{\gamma M C_2}{2}\right)\gamma\|\nabla f(x_k)\|^2 + \frac{\gamma^2 M C_1}{2}. \tag{2.11}$$

Proof. Applying Lemma 1 and Assumption 1 gives

$$\begin{aligned}
\mathbb{E}_{i_k}[f(x_{k+1})|\mathcal{F}_k] - f(x_k) &\leq \gamma\|\nabla f(x_k)\|^2 + \frac{M\gamma^2}{2}\mathbb{E}\left[\|\nabla f_{i_k}(x_k)\|\big|\mathcal{F}_k\right] \\
&\leq -\gamma\|\nabla f(x_k)\|^2 + \frac{M\gamma^2}{2}\left(C_1 + C_2\|\nabla f(x_k)\|^2\right) \\
&= -\left(1 - \frac{\gamma M C_2}{2}\right)\gamma\|\nabla f(x_k)\|^2 + \frac{\gamma^2 M C_1}{2}.
\end{aligned}$$

□

For sufficiently small stepsizes γ the righthand side of (2.11) is negative and scales with $\|\nabla f(x_k)\|^2$. However, the second term could be sufficiently large and may indicate that the loss value increases. Moreover, the second term alludes to the fact that for fixed stepsizes $\gamma > 0$, SGD does not converge to the optimum, but rather to a neighborhood of the optimum.

Worst-case analysis of SGD. We use $\mathbb{E}[\cdot]$ to denote the expected value taken with respect to the joint distribution of all random variables. That is, since x_k is completely determined by the realizations of the independent random variables $x_1, x_2, \cdots x_{k-1}$, the expectation of $f(x_k)$ for $k \in \mathbb{N}$ will be denoted as

$$E[f(x_k)] = \mathbb{E}_{i_1}\mathbb{E}_{i_2}\cdots\mathbb{E}_{i_{k-1}}[f(x_k)].$$

Then we have the following worst-case analysis which shows that under the SGD iterates, the loss function converges, in expectation, in a neighborhood of the optimal value.

Theorem 2 (SGD worst-case analysis, fixed stepsize.)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be strongly convex and twice continuously differentiable and suppose Assumption 1 holds. Then the iterates under SGD with fixed stepsize γ satisfying*

$$0 < \gamma \leq \frac{1}{C_2 M}$$

satisfies the following inequality for all $k \in \mathbb{N}$:

$$\begin{aligned}\mathbb{E}\left[f(x_k) - f(x^*)\right] &\leq \frac{\gamma M C_1}{2m} + (1-\gamma m)^k \left(f(x_0) - f(x^*) - \frac{\gamma M C_1}{2m}\right) \\ &\xrightarrow{k\to\infty} \frac{\gamma M C_1}{2m}.\end{aligned}$$

Proof. By Lemma 2 and the stepsize condition

$$\begin{aligned}\mathbb{E}_{i_k}[f(x_{k+1})|\mathcal{F}_k] - f(x_k) &\leq \gamma\|\nabla f(x_k)\|^2 + \frac{M\gamma^2}{2}\mathbb{E}\left[\|\nabla f_{i_k}(x_k)\|\big|\mathcal{F}_k\right] \\ &\leq -\frac{\gamma}{2}\|\nabla f(x_k)\|^2 + \frac{\gamma^2 M C_1}{2} \\ &\leq -\gamma m\left(f(x_k) - f(x^*)\right) + \frac{\gamma^2 M C_1}{2}. \qquad \text{(Apply (2.3))}\end{aligned}$$

Subtracting the optimal value $f(x^*)$ on both sides gives

$$\mathbb{E}_{i_k}[f(x_{k+1})|\mathcal{F}_k] - f(x^*) \leq (1-\gamma m)\left(f(x_k) - f(x^*)\right) + \frac{\gamma^2 M C_1}{2}.$$

Taking the full expectation $\mathbb{E}[\cdot]$ on both sides give

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] \leq (1-\gamma m)\mathbb{E}\left[f(x_k) - f(x^*)\right] + \frac{\gamma^2 M C_1}{2}$$

In order to put it into a recurrent form, subtract $\frac{\gamma M C_1}{2m}$ on both sides

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] - \frac{\gamma C_1 M}{2m} \leq (1 - \gamma m)\left(\mathbb{E}\left[f(x_k) - f(x^*)\right] - \frac{\gamma M C_1}{2m}\right). \tag{2.12}$$

Applying the recurrence bound (2.12) k-times achieves the desired result. Moreover, (2.12) is a contraction inequality as $C_1 \geq 1$ and the stepsize condition gives

$$0 < \gamma m \leq \frac{m}{C_1 M} \leq \frac{m}{M} \leq 1.$$

□

2.2.3 Shortcomings of worst-case analysis

Assumptions of worst-case analysis. The analysis of SGD and gradient descent's performance hinges on the bounds

$$2m\left(f(x_k) - f(x^*)\right) \leq \|\nabla f(x_k)\|^2$$

and

$$f(x_{k+1}) - f(x^*) \leq \gamma \|\nabla f(x_k)\|^2 + \frac{M\gamma^2}{2}\|\nabla f(x_k)\|^2.$$

Moreover, the analysis assumes rather limiting stepsize conditions ($\gamma = \frac{1}{M}$ and $0 < \gamma < \frac{1}{MC_2}$ for gradient descent and SGD, respectively). Therefore the analysis is only as strong as how accurate these inequalities are in capturing the true relation between the iterates and the optimum on a specific problem instance $(\boldsymbol{A}, \boldsymbol{b})$ constrained to these stepsize values. In reality, these analysis captures the behaviour of algorithms on pathological data but poorly reflect the typical behaviour of algorithms [48].

Worst-case analysis dimension-invariant. The analysis for both gradient descent and SGD holds for $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ regardless of the size of $n, d > 0$. In other words, if n, d are large, the analysis does *not* incorporate information embedded in the structure of high-

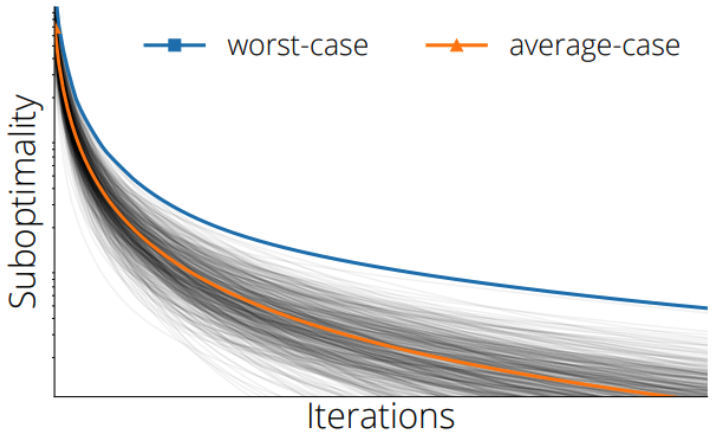

Figure 2.1: Worst-case analysis captures algorithmic performance on pathological instances but do not reflect typical behaviour of algorithms. Figure from [48].

dimensional datasets which often drastically differs from the low-dimensional settings. In fact, the only properties of the problem $\boldsymbol{A}$ the analysis incorporates in its bounds are M and m which form proxies of the largest and smallest eigenvalues of the Hessian $\nabla^2 f(x)$), respectively. This thesis aims to incorporate the information of high-dimensional structures by incorporating the entire spectrum of the problem instance into our analysis.

2.3 Momentum machinery on quadratics: rates and batch-sizes

In this thesis, we analyze two popular mini-batch methods that incorporate the momentum machinery: mini-batch Polyak momentum and Nesterov acceleration with momentum (see subsection 1.1.3 for their definitions) on quadratic functions. A quadratic function $f(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$ satisfies the form

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T \boldsymbol{S}\boldsymbol{x} + \boldsymbol{h}^T \boldsymbol{x} + u$$

Algorithm	Convergence Rate
Gradient Descent	$\lvert f(\boldsymbol{x}^*) - f(\boldsymbol{x}_k)\rvert \leq C_1 \left(\frac{\kappa-1}{\kappa+1}\right)^k$
Polyak Heavy-ball	$\lvert f(\boldsymbol{x}^*) - f(\boldsymbol{x}_k)\rvert \leq C_2 \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k$
Nesterov Acceleration	$\lvert f(\boldsymbol{x}^*) - f(\boldsymbol{x}_k)\rvert \leq C_3 \left(1 - \frac{2}{\sqrt{3k+1}}\right)^k$

Table 2.1: Convergence rates of various algorithms with momentum under the quadratic loss function.

for some positive definite matrix $\boldsymbol{S} \in \mathbb{R}^{d\times d}$, vector $\boldsymbol{h} \in \mathbb{R}^d$ and scalar $u \in \mathbb{R}$. Although both algorithms are extensively used to train complex models (e.g. neural networks), relatively little is known about their fundamental properties. In particular, as far as we know, there is no analysis that quantifies the convergence rate of these algorithms based on batchsize. However, convergence rate of these algorithms *without* momentum in terms of batchsize has been explored [11, 35]. In this subsection, we provide context to the problem of quantifying rates in terms of batchsize under the quadratic loss.

2.3.1 Full-batch rates

Under the quadratic loss, it is known that for *full-batch* gradient descent, Polyak heavy-ball, and Nesterov acceleration have contraction constants respectively given by

$$\frac{\kappa-1}{\kappa+1}, \qquad \frac{\sqrt{k}-1}{\sqrt{k}+1} \qquad \text{and} \qquad 1-\frac{2}{\sqrt{3k+1}} \qquad \text{where } \kappa \overset{\text{def}}{=} \frac{\lambda_{\max}(\boldsymbol{S})}{\lambda_{\min}(\boldsymbol{S})}.$$

Since $\kappa \geq 1$ we have that heavy-ball and Nesterov acceleration has superior convergence rates. See [11,41] for proof of these results. However, in the small-batch regime, that is when batch-size $\beta \approx 1$ stochastic heavy-ball momentum has the same rate as SGD when step sizes are adjusted correctly [44]. Moreover when $\beta = 1$, [30] showed examples of linear regression where Polyak momentum and Nesterov's accelerated momentum provably fail to achieve faster rates than vanilla SGD. Also, [27, 57] illustrate instances where incorporating momentum to SGD does not provide any performance gain. In short, un-

der the *small batch regime*, i.e., $\beta \approx 1$, SGD, Polyak momentum, and Nesterov acceleration with momentum, have the same rates. In the *large batch regime*, i.e., $\beta \approx n$, we have Polyak momentum and Nesterov acceleration with momentum provides superior rates. The question becomes: *How do the rates of these algorithms depend on β?*

The discussion in the previous paragraph suggests that in small-batch regimes, incorporating momentum to SGD is unnecessary. However, in large-batch regimes, the momentum machinery provides a speed up of a factor of $\sqrt{k}$. In this thesis, we give an exact analysis of how the contraction constant of Polyak heavy-ball depend on batch-size which illustrates a phase transition phenomena depending on a provably optimal batch-size β^*. In turn, we rigorously define the *small batch* and *large batch regimes* based off this phenomena.

2.4 Resolvents

Many of the assumptions in this thesis will be expressed in terms of a mathematical object known as the resolvent.

Definition 12. *The resolvent of a random matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}$ is*

$$R(z;\boldsymbol{X}) = (z\boldsymbol{I}_d - \boldsymbol{X})^{-1} \qquad \textit{for } z \in \mathbb{C}.$$

In this thesis, we restrict to the setting where $\boldsymbol{X}$ is almost-surely symmetric. In this case, the resolvent is a complex analytic matrix-valued function that encodes the spectral properties of $\boldsymbol{X}$. To see this, we may apply the eigenvalue decomposition of $\boldsymbol{X}$ to obtain

$$R(z;\boldsymbol{X}) = \boldsymbol{U}\text{diag}\left(z-\lambda_1,\ldots,z-\lambda_2\right)^{-1}\boldsymbol{U}^T.$$

Moreover, if Ω is a simple contour enclosing the eigenspectrum of $\boldsymbol{X}$ and φ is analytic then

$$\varphi(\boldsymbol{X}) = \frac{1}{2\pi i} \oint_{\Omega} \varphi(z) R(z; \boldsymbol{X}) dz.$$

In essence, the resolvent gives access to the functionals of random matrices by contour integration in the complex plane. Another useful application of the resolvent is that it gives acccess to the eigenvalue distribution of random matrices through the application of the inverse Stieltjes transform. That is, define

$$\mu_{\boldsymbol{X}} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i}(\boldsymbol{X})$$

where $\lambda(\boldsymbol{X}) = \{\lambda_1(\boldsymbol{X}), \ldots, \lambda_d(\boldsymbol{X})\}$ is the eigenspectrum of $\boldsymbol{X}$ and $\delta_{\lambda_i}(\boldsymbol{X})$ is the Dirac-delta function at the i-th eigenvalue of $\boldsymbol{X}$. Then for all $a, b \notin \lambda(\boldsymbol{X})$ the Stieltjes transform relation gives us

$$\int_a^b \mu_{\boldsymbol{X}}(d\lambda) = \lim_{\epsilon \downarrow 0} \int_a^b \frac{1}{\pi} \Im[m_{\boldsymbol{X}}(x + i\epsilon)] dx$$

where $\Im(\cdot)$ denotes the imaginary part of a complex number where

$$m_{\boldsymbol{X}}(z) \stackrel{\text{def}}{=} \int \frac{\mu \boldsymbol{X}(d\lambda)}{\lambda - z} = \frac{1}{n} \mathsf{tr}\left(R(z; \boldsymbol{X})\right).$$

One approach in determining the eigenvalue distribution of a random matrix $\boldsymbol{X}$ is to construct a deterministic matrix $\tilde{R}(z; \boldsymbol{X})$ so that

$$\tilde{R}(z; \boldsymbol{X}) - R(z : \boldsymbol{X}) \xrightarrow{a.s.} 0$$

For more details on random matrix identities especially in a machine learning context see [33].

Marchenkor-Pastur law. For example, one setting that we consider in this thesis is the Marchenko-Pastur law (see [36]) where we consider the random matrix $\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T$ and $\boldsymbol{A} \in$

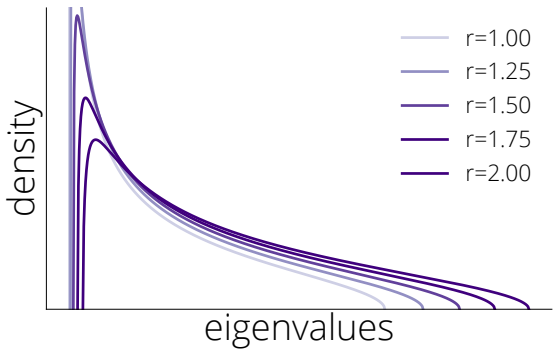

Figure 2.2: Probability density function of the Marchenko-Pastur law with varying dimensionality ration $r = d/n$.

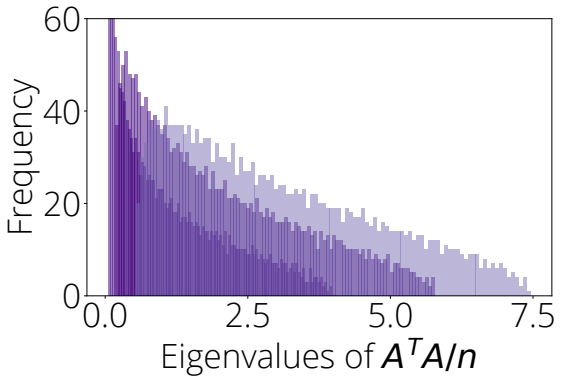

Figure 2.3: Empirical distribution of eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}/n$ where $\boldsymbol{A}$ is a matrix with iid standard Gaussian entries.

$\mathbb{R}^{n\times d}$ has i.i.d. mean zero and variance $\sigma^2 < \infty$. Then the resolvent

$$R(z; \boldsymbol{A}\boldsymbol{A}^T) = \left(\frac{1}{n}\boldsymbol{A}\boldsymbol{A}^T - z\mathbf{I}_d\right)^{-1} \tag{2.13}$$

has the deterministic correspondence

$$\tilde{R}(z; \boldsymbol{A}\boldsymbol{A}^T) = \int \frac{\mu_{MP}(d\lambda)}{\lambda - z}\mathbf{I}_d$$

where the Marchenko-Pastur measure μ_{MP} satisfies

$$\begin{gathered} d\mu_{MP}(\lambda) \overset{\text{def}}{=} \delta_0(\lambda)\max\{1-r, 0\} + \frac{r\sqrt{(\lambda-\lambda^-)(\lambda^+-\lambda)}}{2\pi\lambda}1_{[\lambda^-,\lambda^+]}, \\ \text{where} \quad \lambda^- \overset{\text{def}}{=} (1-\sqrt{\frac{1}{r}})^2 \quad \text{and} \quad \lambda^+ \overset{\text{def}}{=} (1+\sqrt{\frac{1}{r}})^2 \end{gathered} \tag{2.14}$$

and where $r = d/n$ is the dimensionality ratio. See Figures 2.2 and 2.3 for an illustration of the Marchenko-Pastur law. In the coming sections, we use the resolvent to study the eigenspectrum of the Hessian of the least squares problem $\boldsymbol{A}^T\boldsymbol{A}$. In short, the resolvent is a rich mathematical tool from random matrix theory that extensively draws from linear algebra, complex analysis, and probability theory with far reaching applications. As such,

we limit our discussion of this tool to the bare minimum required to state and prove the majority of the assumptions in this thesis.

2.5 Volterra equation

The Volterra equations are a subclass of integral equations. In this thesis, we will make use of the linear Volterra equation of the second kind. This equation has found applications in population dynamics [26] and actuarial science in the form of the renewal equation [14].

Definition 13. *A linear Volterra equation of the second kind is of the form*

$$\psi(t) = f(t) + \int_a^t \psi(s)W(t-s)\,ds. \tag{2.15}$$

The function $f(t)$ denotes the forcing term which depends only on time t. The other term on the right-hand side is a convolution between a kernel term $W(t,s)$ with the previous values of ψ. The Volterra equation has a discrete analogue given by the following definition.

Definition 14. *The discrete Volterra equation of the second kind is of the form*

$$\psi(t+1) = F(t+1) + \sum_{k=0}^{t} \psi(k)K(t-k). \tag{2.16}$$

We will show that, under a class of quadratic functions, every mini-batch gradient-based method has a corresponding equation of the form (2.16) where its solution $\psi(\cdot)$ captures the exact training dynamics. In doing so, we can revert to the theory of integral equations to express interesting properties of the dynamics such as limiting behaviour and selection of optimal hyperparameters. Given a mini-batch gradient-based method, we will explore the intuition behind how the forcing term $f(t)$ captures its deterministic behaviour and how the convolution term captures its inherit randomness. More gener-

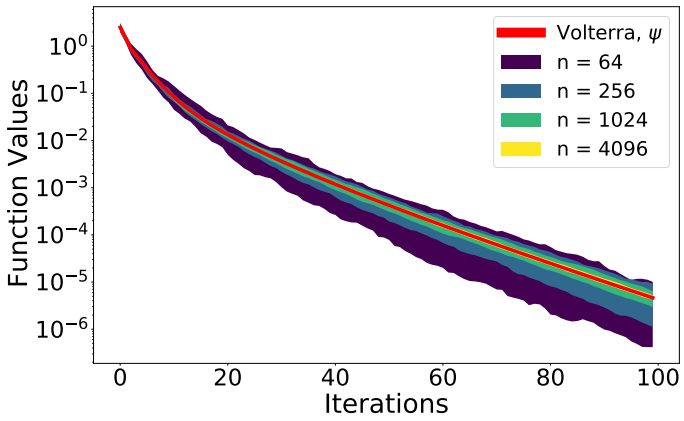

Figure 2.4: Concentration of SGD+M on a Gaussian random least squares problem.

ally, one can draw analogies between the population risk of models and the population dynamics in a demographic context under the lens of the Volterra integral equation.

Recall the shortcomings of worst-case analysis mentioned in subsection 2.2.3. This motivates the need for exact dynamics to capture the behaviour of optimization algorithms. To capture the exact dynamics of a (stochastic) optimization algorithm is to construct a deterministic function $\psi(t) : \mathbb{R} \to \mathbb{R}$ that exactly reflects the behaviour of the algorithm. Formally, given some finite time horizon $T > 0$, loss function f, and iterates $\{x_k : 0 \leq k \leq T\}$ returned by an optimization algorithm, $\psi : \mathbb{R} \to \mathbb{R}$ captures the exact dynamics of the algorithm if

$$\sup_{0 \leq k \leq T} |\psi(k) - f(k)| = 0. \tag{2.17}$$

2.6 Exact dynamics: motivating examples

There are several lines of work of exact dynamics which can be split into two perspectives: Population exact dynamics [32, 37, 43] and the finite-sample exact dynamics [16]. Here, "finite-sample" means that one is concerned with explicit estimates that are either dimension-free or that precisely capture the dependence of the problem on relevant dimensional parameters. Unlike many classical probabilistic results that focus on limit

theorems, nonasymptotic methods provide a precise expression of the interrelation between different parameters in high-dimensional problems. However, the finite-sample requires stronger assumptions (typically some form of Gaussianity) on the data. In our discussion, we limit ourselves to the analysis of asymptotics (i.e. population dynamics) which require taking all these parameters to the limit in a fixed relation to each other (e.g. $n/d \to r \in (0, \infty)$). Nevertheless, the population dynamics have numerous implications, and we provide three examples in our discussion.

Since the algorithms of focus (e.g. mini-batch gradient-based methods) are stochastic then the iterates $\{x_k : k \geq 0\}$ form a stochastic process and the notion of exact dynamics (2.17) requires a probabilistic statement that strengthens as dimension scales up.

Overwhelming probability. We say an event B holds *with overwhelming probability* (w.o.p.) if, for every fixed $D > 0$, $\Pr(B) \geq 1 - C_D n^{-D}$ for some C_D independent of d. In this thesis, we will show that there exists a solution to the Volterra equation ψ and constants $C, \tilde{C}, c > 0$ such that the event

$$\sup_{0 \leq k \leq T} |f(x_t) - \psi(t)| > C n^{-\tilde{C}}$$

holds with overwhelming probability. In symbols,

$$\Pr\left(\sup_{0 \leq k \leq T} |f(x_t) - \psi(t)| > C n^{-\tilde{C}} \right) \leq D n^{-c}. \tag{2.18}$$

for some $D > 0$ and least-squares loss function f. Equation (2.18) says the probability of the event that $f(\boldsymbol{x}_t)$ uniformly converges to $\psi(t)$ tends to one quicker than any polynomial rate as $n, d \to \infty$. This is what we mean by *capturing exact dynamics* in the asymptotics case for stochastic algorithms. Equation (2.18) is illustrated in Figure 2.4 on the least squares problems under the iterates of mini-batch gradient descent with momentum.

Analysis of algorithms. In asymptotic analysis, $\psi(k)$ in (2.17) captures the training loss of the algorithm when $n, d \to \infty$ (see Figure 2.4). This allows us to use ψ to study the behaviour and properties of algorithms under high-dimensional data. In this thesis, we analyze ψ in order to establish interesting properties of Polyak momentum and Nesterov acceleration such as conditions of convergence, optimal selection of hyperparameters, and limiting behaviour.

Generalization. We can also use the exact dynamics to capture the expected risk ((1.3)) of certain models. For example, in linear regression, one can write the expected risk $R(x)$ as a deterministic function depending on the exact dynamics ψ. Therefore, in such settings one can compute exact generalization errors under large dimensional settings. See Subsection 3.5.2 for more details.

Random features. The random features setting was introduced in [6] for scaling kernel machines. Random features models provide a rich and tractable class of models to understand generalization phenomena. As random features models satisfy our assumptions, we can (asymptotically) characterize all the rich phenomena illustrated by random features like *double-descent* [39] often seen in extremely complex models, for e.g. neural networks. In fact, characterizing the double-descent phenomena via random features model has already been established [38] which shows that the double-descent phenomena is not a unique characteristic of neural networks. See Subsection 3.5.3 for more details.

2.7 Concentration of measure

As our analysis is centered on stochastic optimization algorithms in high-dimensional settings, we employ tools from the field of high-dimensional probability to establish our results. In particular, we use these tools to establish (2.18). Generally speaking, we will explicitly construct ψ and express the difference between $\psi(t)$ and $f(x_t)$ in terms of error

terms $\mathcal{E}_k(t)$. In symbols,

$$\psi(t) - f(x_t) = \mathcal{E}_1(t) + \cdots \mathcal{E}_m(t), \qquad \forall t \in \{0, 1, \ldots, T\}.$$

Then we establish that each error term $\mathcal{E}_k(t) \to 0$ in probability as dimensions $n, d \to \infty$. The tools we use reflect the notion that many high-dimensional objects concentrate around their mean. First, recall the basic tail bound:

Proposition 1 (Chebyshev's inequality)**.** *Let X be a random variable with mean μ and variance σ^2. Then, for any $t > 0$ we have*

$$\Pr\left(|X - \mu| \geq t\right) \leq \frac{\sigma^2}{t^2}.$$

Chebyshev's inequality is a general inequality in the sense that it can be applied to all random variables (provided its second moment exists). However, this inequality is weak in the sense that the bound does not get stronger as dimension grows. By placing stronger assumptions on X we can achieve tighter bounds that explicitly depend on dimensions. For example, if we assume X is almost-surely bounded in an interval $[a, b]$ with $b > a$ then we have Hoeffding's inequality [54].

Proposition 2 (Hoeffding's inequality.)**.** *Let $\boldsymbol{X} \in \mathbb{R}^n$ where $\boldsymbol{X} = (X_1, \cdots X_n)$ and X_i's are independent random variables so that* $\Pr\left(X_i \in [a, b]\right) = 1$ *for some $a < b$, and let $\epsilon > 0$. Then*

$$\Pr(\frac{1}{n}\sum_{j=1}^{n} X_i - \mathbb{E}[X_1] \geq \epsilon) \leq \exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

It turns out that the high-dimensional probability tools we use are just variants of Hoeffding's inequality. For example, since the mini-batch gradient-based methods (recall Definition 2) are inherently random due to the uniform random batch sampling $B_j \subseteq \{1, \ldots, n\}$, we will require concentration of measure results for sampling from a finite population of n points.

Proposition 3 (Proposition 1.4, [9]). *Let $\mathcal{X} = (x_1, \cdots, x_n)$ be a finite population of n points and $X_1, \cdots, X_\beta$ be a random sample drawn without replacement from $\mathcal{X}$. Let*

$$a = \min_{1 \le i \le n} x_i \text{ and } b = \max_{1 \le i \le n} x_i.$$

Also let

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

be the mean and variance of $\mathcal{X}$, respectively. Then for all $\epsilon > 0$,

$$\mathbb{P}\left(\frac{1}{\beta}\sum_{i=1}^{\beta} X_i - \mu \ge \epsilon\right) \le \exp\left(-\frac{\beta\epsilon^2}{2\sigma^2 + (2/3)(b-a)\epsilon}\right).$$

We will use this result to show some of the error terms $\mathcal{E}_k(t)$ tend to zero as $n, d \to \infty$ (see Section 3.7.1).

2.8 Overview

This thesis focuses on results that were established by a joint effort involving Kiwon Lee (McGill PhD) and Courtney and Elliot Paquette (McGill professors). Some of the results can be found in our publication [32]. We delineate the author's original work by chapter.

Chapter 3 focuses on the main theorem of exact dynamics. These results were established by the author and his supervisors Courtney and Elliot Paquette. These results follow from similar proof techniques established by Kiwon Lee in [32]. This will be the most technically demanding chapter. First, we provide the formal problem setup. Second, we provide assumptions on the data and the models under consideration. Third, we present the main proposition of the chapter which shows that the iterates of mini-batch gradient-based methods can be written as a linear combination of polynomials. Fourth, we prove the main theorem which illustrates that the dynamics of every mini-batch-gradient-based algorithm under quadratic loss functions can be captured by exact dynamics.

Chapter 4 shows the main theorem in practice by providing an exact dynamic analysis of mini-batch Polyak momentum. The main theorem which enables such an analysis was established by Kiwon Lee. With this result, the author provided experiments illustrating the main theorem and its implications as well as established results concerning the selection of optimal hyperparameters and conditions for convergence. We show that the exact dynamics is captured by a deterministic function ψ that solves the Volterra integral equation (2.15) and is completely determined by the eigenspectrum of the problem. This enables us to address several open fundamental questions regarding Polyak momentum:

1. What are the conditions of convergence?

2. How does the rate of convergenece depend on batch size?

3. How does one select optimal hyperparameters?

Chapter 6 gives an analysis of mini-batch Nesterov acceleration. This work was established by the author. By an exact dynamics approach, we prove sufficient conditions of convergence for mini-batch Nesterov acceleration in the strongly convex case. In the convex case, we show one can represent the evolution of the iterates by Chebyshev polynomials.

Chapter 3

Exact Dynamic Results

This chapter establishes the main theorem of this thesis. That is, we will show that under a general class of quadratic functions, we can provide exact dynamics of any algorithm satisfying Definition 2. To this end, we will provide the formal setup, necessary assumptions, connections between optimization and polynomials and technical bounds that are required to state and prove the main theorem. Some assumptions, particularly those related to the data distribution, may appear contrived, so we will show several practical settings that satisfy these assumptions.

3.1 Formal problem set-up

To formalize the analysis, we define the *ℓ^2-regularized least squares problem*:

$$\arg\min_{x\in\mathbb{R}^d}\left\{f(\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2+\frac{\delta}{2}\|\boldsymbol{x}\|^2 = \sum_{i=1}^{n}\underbrace{\frac{1}{2}\left((\boldsymbol{a}_i\boldsymbol{x}-b_i)^2+\frac{\delta}{n}\|\boldsymbol{x}\|^2\right)}_{\stackrel{\text{def}}{=}f_i(\boldsymbol{x})}\right\}. \tag{3.1}$$

The fixed parameter $\delta > 0$ controls the regularization strength and it is independent of n and d. We focus on setups where the parameter choices n and d are large, but we do not require that they are proportional. Instead we need the following:

Assumption 2 (Polynomially related)**.** *There is an $\rho \in (0, 1)$ so that*

$$d^{\rho} \leq n \leq d^{1/\rho}.$$

Moreover our results only gain power when one (and hence both) of these parameters are large.

The data matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ and the labels $\boldsymbol{b}$ may be deterministic or random; we formulate our theorems for deterministic $\boldsymbol{A}$ and $\boldsymbol{b}$ in (3.1) satisfying various assumptions, and in the the applications of our theorems to statistical settings, we shall show that the random $\boldsymbol{A}$ and $\boldsymbol{b}$ (3.1) satisfy these assumptions. These assumptions are motivated by the case where the augmented matrix $[\boldsymbol{A} \,|\, \boldsymbol{b}]$ has rows that are independent and sampled from some common distribution. We also note that the problem (3.1) is homogeneous, in that if we divide all of $\boldsymbol{A}$, $\boldsymbol{b}$, and $\sqrt{\delta}$ by any desired scalar, we produce an equivalent optimization problem. As such, we adopt the following convention without loss of generality.

Assumption 3 (Data–target normalization)**.** *There is a constant $C > 0$ independent of d and n such that the spectral norm of $\boldsymbol{A}$ is bounded by C and the target vector $\boldsymbol{b} \in \mathbb{R}^n$ is normalized so that $\|\boldsymbol{b}\|^2 \leq C$.*

Assumption 3 and homogeneity allows us to conduct an analysis that is unaffected by the dimensionality of our problem. More importantly, we also assume that the data and targets resemble typical unstructured high-dimensional random matrices. One of the principal qualitative properties of high-dimensional random matrices is the *delocalization of their eigenvectors*, which refers to the statistical similarity of the eigenvectors to uniform random elements from the Euclidean sphere. The precise mathematical description of this assumption is most easily given in terms of resolvent bounds. The resolvent $R(z; \boldsymbol{M})$ of a matrix $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ is

$$R(z; \boldsymbol{M}) = (\boldsymbol{M} - z\mathbf{I}_d)^{-1} \quad \text{for } z \in \mathbb{C}.$$

In terms of the resolvent, we suppose the following:

Assumption 4. *Suppose Ω is a bounded contour enclosing $[0, 1+\|\boldsymbol{A}\|^2]$ at distance $1/2$. Suppose there is an $\alpha_0 \in (0, \frac{1}{4})$ for which*

1. $\max_{z\in\Omega} \max_{1\le i\le n} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{b}| \le n^{\alpha_0 - 1/2}$.
2. $\max_{z\in\Omega} \max_{1\le i\ne j\le n} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_j^T| \le n^{\alpha_0 - 1/2}$.
3. $\max_{z\in\Omega} \max_{1\le i\le n} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_i - \frac{1}{n} tr R(z; \boldsymbol{A}\boldsymbol{A}^T)| \le n^{\alpha_0 - 1/2}$.

We use the notation $|\Omega| = \max_{z\in\Omega} |\Omega(z)|$ which we assume to be bounded independent of n and d.

Only the resolvent of $\boldsymbol{A}\boldsymbol{A}^T$ appears in these assumptions, and so in effect we are only assuming statistical properties on the left singular-vectors of $\boldsymbol{A}$. This assumption reflects the common formulation (3.1) in which the rows of $\boldsymbol{A}$ are independent, and so the left singular-vectors of $\boldsymbol{A}$ are expected to be delocalized (under some mildness assumptions on the distributions of the rows). The first condition, which involves the interaction between $\boldsymbol{A}\boldsymbol{A}^T$ and $\boldsymbol{b}$, can be understood as requiring that $\boldsymbol{b}$ is not too strongly aligned with the left singular-vectors of $\boldsymbol{A}$. The other two conditions can be viewed as corollaries of delocalization of the left singular-vectors.

3.2 From optimization algorithms to polynomials

In this section we look at the classical connection between optimization algorithms, iterative methods, and polynomials [18,21,22,28,46]. While the idea of analyzing optimization algorithms from the perspective of polynomials is well-established, analyzing stochastic versions of these algorithms from this perspective is a new approach (see *e.g.,* [42]) that differs from the common analysis seen in [12]. This connection between polynomials and (stochastic) optimization algorithms will be crucial in proving the generalization result (3.49) in Theorem 3. Before our discussion of this connection, we recall the notion of mini-batch gradient-based methods seen in Definition 2.

Definition 15 (Mini-batch gradient-based method)**.** Given a gradient-based algorithm, we define a stochastic optimization algorithm, called a *mini-batch gradient-based method*, if at each update one generated uniformly at random a batch $B_i \subset \{1, 2, \ldots, n\}$ and the update satisfies,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_0 + \sum_{j=0}^{k} c_{k,j} \sum_{i \in B_j} \nabla f_i(\boldsymbol{x}_j) \tag{3.2}$$

for some scalars $c_{k,i}$. In the special case that the batch size $|B_i| = 1$ for all $i \geq 0$, (3.2) is known as a *stochastic gradient-based method*.

We denote the proportion of the number of samples n relative to the size of the batch B_i taken by the mini-batch gradient-based method as the *batch fraction*,

$$\zeta \overset{\text{def}}{=} \frac{|B_i|}{n} = \frac{\beta}{n}. \tag{3.3}$$

As for the initialization $\boldsymbol{x}_0$, we need to suppose that it does not interact too strongly with the *right* singular-vectors of $\boldsymbol{A}$. In the spirit of Assumption 4, it suffices to assume the following:

Assumption 5. *Let Ω be the same contour as in Assumption 4 and let $\alpha_0 \in (0, \frac{1}{4})$. Then*

$$\max_{z \in \Omega} \max_{1 \leq i \leq d} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T) \boldsymbol{A}\boldsymbol{x}_0| \leq n^{\alpha_0 - 1/2}.$$

For solving the ℓ^2-regularized least squares problem (3.1), the updates in (3.2) precisely are given by

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{x}_0 + \sum_{j=0}^{k} c_{k,j} \sum_{i \in B_j} \left(\boldsymbol{a}_i \boldsymbol{x}_j - b_i + \frac{\delta}{n} \boldsymbol{x}_j \right) \\ &= \boldsymbol{x}_0 + \sum_{j=0}^{k} c_{k,j} \left(\boldsymbol{A}^T \boldsymbol{P}_{j+1} (\boldsymbol{A}\boldsymbol{x}_j - \boldsymbol{b}) + \delta \frac{|B_j|}{n} \boldsymbol{x}_j \right) \\ &= \boldsymbol{x}_0 + \sum_{j=0}^{k} c_{k,j} \left(\boldsymbol{A}^T \boldsymbol{P}_{j+1} (\boldsymbol{A}\boldsymbol{x}_j - \boldsymbol{b}) + \delta \zeta \boldsymbol{x}_j \right), \end{aligned} \tag{3.4}$$

where $\boldsymbol{P}_{j+1} = \sum_{i \in B_j} \boldsymbol{e}_i \boldsymbol{e}_i^T$ for $\boldsymbol{e}_i \in \mathbb{R}^n$ with 1 in the i-th coordinate and 0 otherwise. Next we define the *martingale increment* to be, for all $j \geq 1$,

$$\begin{aligned} \mathring{\boldsymbol{M}}_j &\stackrel{\text{def}}{=} \mathbb{E}\left[\boldsymbol{A}^T \boldsymbol{P}_j \left(\boldsymbol{A}\boldsymbol{x}_{j-1} - \boldsymbol{b}\right) \middle| \mathcal{F}_{j-1}\right] - \boldsymbol{A}^T \boldsymbol{P}_j \left(\boldsymbol{A}\boldsymbol{x}_{j-1} - \boldsymbol{b}\right) \\ &= \zeta \boldsymbol{A}^T (\boldsymbol{A}\boldsymbol{x}_{j-1} - \boldsymbol{b}) - \boldsymbol{A}^T \boldsymbol{P}_j (\boldsymbol{A}\boldsymbol{x}_{j-1} - \boldsymbol{b}). \end{aligned} \tag{3.5}$$

The filtration $\{\mathcal{F}_j\}$ which contains the past iterate information, that is, $\mathcal{F}_0 = \sigma(\boldsymbol{x}_0, \boldsymbol{A}, \boldsymbol{b})$ and $\mathcal{F}_j = \sigma(\boldsymbol{x}_j, \boldsymbol{x}_{j-1}, \ldots, \boldsymbol{x}_0, \boldsymbol{A}, \boldsymbol{b})$. A simple computation shows

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{x}_0 + \sum_{j=0}^{k} c_{k,j} \left(\boldsymbol{A}^T \boldsymbol{P}_{j+1} (\boldsymbol{A}\boldsymbol{x}_j - \boldsymbol{b}) + \delta\zeta \boldsymbol{x}_j \right) \\ &= \boldsymbol{x}_0 + \sum_{j=0}^{k} \zeta c_{k,j} \left[\boldsymbol{A}^T (\boldsymbol{A}\boldsymbol{x}_j - \boldsymbol{b}) + \delta \boldsymbol{x}_j \right] - \sum_{j=0}^{k} c_{k,j} \mathring{\boldsymbol{M}}_{j+1} \\ &= \boldsymbol{x}_0 + \sum_{j=0}^{k} \zeta c_{k,j} (\delta \mathbf{I} + \boldsymbol{A}^T \boldsymbol{A}) \boldsymbol{x}_j - \sum_{j=0}^{k} \zeta c_{k,j} \boldsymbol{A}^T \boldsymbol{b} - \sum_{j=0}^{k} c_{k,j} \mathring{\boldsymbol{M}}_{j+1}. \end{aligned} \tag{3.6}$$

Given any mini-batch gradient-based method, we can associate to the method a *residual polynomial* P_k, *iteration polynomial* Q_k, and *noise polynomial* $N_{k,j}$ which are polynomials of degree k, precisely as followed.

Proposition 4 (Polynomials and mini-batch gradient-based methods)**.** *Consider a mini-batch gradient-based method with coefficients $c_{k,j}$. Define the sequence of polynomials $\{P_k, Q_k\}_{k=0}^{\infty}$*

and $\{N_{k,j}\}_{k\in[0,\infty),j\leq k}$ *recursively by*

$$P_0(\boldsymbol{H}) = \boldsymbol{I} \quad \textit{and} \quad P_{k+1}(\boldsymbol{H}) = \boldsymbol{I} + \sum_{j=0}^{k} \zeta c_{k,j} \boldsymbol{H} P_j(\boldsymbol{H}),$$

$$Q_0(\boldsymbol{H}) = \boldsymbol{0} \quad \textit{and} \quad Q_{k+1}(\boldsymbol{H}) = \sum_{j=0}^{k} c_{k,j}(\zeta \boldsymbol{H} Q_j(\boldsymbol{H}) - \boldsymbol{I})$$

$$N_{0,0}(\boldsymbol{H}) = \boldsymbol{0} \quad \textit{and} \quad N_{k+1,j}(\boldsymbol{H}) = \begin{cases} \sum_{i=j}^{k} \zeta c_{k,i} \boldsymbol{H} N_{i,j}(\boldsymbol{H}) - c_{k,j-1}, & \textit{if } j = 1, \ldots, k+1 \\ \sum_{i=0}^{k} \zeta c_{k,i} \boldsymbol{H} N_{i,0}(\boldsymbol{H}), & \textit{if } j = 0 \\ -c_{k,k}, & \textit{if } j = k+1 \end{cases}, \tag{3.7}$$

where the matrix $\boldsymbol{H} \overset{\textit{def}}{=} \delta \boldsymbol{I} + \boldsymbol{A}^T \boldsymbol{A}$*. Then we can express the iterates* $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ *generated by the mini-batch gradient-based method in terms of these polynomials:*

$$\boldsymbol{x}_k = P_k(\boldsymbol{H})\boldsymbol{x}_0 + Q_k(\boldsymbol{H})\zeta \boldsymbol{A}^T \boldsymbol{b} + \sum_{j=0}^{k} N_{k,j}(\boldsymbol{H}) \mathring{\boldsymbol{M}}_j. \tag{3.8}$$

Proof. We show this by induction. For $k = 0$, the result clearly holds. We suppose that

$$\boldsymbol{x}_k = P_k(\boldsymbol{H})\boldsymbol{x}_0 + Q_k(\boldsymbol{H})\zeta \boldsymbol{A}^T \boldsymbol{b} + \sum_{j=0}^{k} N_{k,j}(\boldsymbol{H}) \mathring{\boldsymbol{M}}_j. \tag{3.9}$$

We will show that this holds for $\boldsymbol{x}_{k+1}$. By (3.6), the definitions of the polynomials P_k, Q_k, and $N_{k,j}$, and the induction hypothesis, we have that

$$\begin{aligned}
\boldsymbol{x}_{k+1} &= \boldsymbol{x}_0 + \sum_{j=0}^{k} \zeta c_{k,j} \boldsymbol{H}\boldsymbol{x}_j - \sum_{j=0}^{k} \zeta c_{k,j} \boldsymbol{A}^T\boldsymbol{b} - \sum_{j=0}^{k} c_{k,j} \mathring{\boldsymbol{M}}_{j+1} \\
&= \boldsymbol{x}_0 + \sum_{j=0}^{k} \zeta c_{k,j} \boldsymbol{H}\left[P_j(\boldsymbol{H})\boldsymbol{x}_0 + Q_j(\boldsymbol{H})\zeta\boldsymbol{A}^T\boldsymbol{b} + \sum_{i=0}^{j} N_{j,i}(\boldsymbol{H})\mathring{\boldsymbol{M}}_i \right] \\
&\quad - \sum_{j=0}^{k} c_{k,j}\zeta\boldsymbol{A}^T\boldsymbol{b} - \sum_{j=0}^{k} c_{k,j}\mathring{\boldsymbol{M}}_{j+1} \\
&= \left[\mathbf{I} + \sum_{j=0}^{k} \zeta c_{k,j}\boldsymbol{H}P_j(\boldsymbol{H})\right]\boldsymbol{x}_0 + \sum_{j=0}^{k} c_{k,j}[\zeta\boldsymbol{H}Q_j(\boldsymbol{H}) - \mathbf{I}]\zeta\boldsymbol{A}^T\boldsymbol{b} \\
&\quad + \sum_{j=0}^{k}\sum_{i=0}^{j} \zeta c_{k,j}\boldsymbol{H}N_{j,i}(\boldsymbol{H})\mathring{\boldsymbol{M}}_i - \sum_{j=0}^{k} c_{k,j}\mathring{\boldsymbol{M}}_{j+1} \\
&= P_{k+1}(\boldsymbol{H})\boldsymbol{x}_0 + Q_{k+1}(\boldsymbol{H})\zeta\boldsymbol{A}^T\boldsymbol{b} + \sum_{j=0}^{k}\sum_{i=j}^{k} \zeta c_{k,i}\boldsymbol{H}N_{i,j}(\boldsymbol{H})\mathring{\boldsymbol{M}}_j - \sum_{j=0}^{k} c_{k,j}\mathring{\boldsymbol{M}}_{j+1} \\
&= P_{k+1}(\boldsymbol{H})\boldsymbol{x}_0 + Q_{k+1}(\boldsymbol{H})\zeta\boldsymbol{A}^T\boldsymbol{b} + \sum_{j=1}^{k}\left[\sum_{i=j}^{k} \zeta c_{k,i}\boldsymbol{H}N_{i,j}(\boldsymbol{H}) - c_{k,(j-1)}\right]\mathring{\boldsymbol{M}}_j \\
&\quad + \sum_{i=0}^{k} \zeta c_{k,i}\boldsymbol{H}N_{i,0}(\boldsymbol{H})\mathring{\boldsymbol{M}}_0 - c_{k,k}\mathring{\boldsymbol{M}}_{k+1} \\
&= P_{k+1}(\boldsymbol{H})\boldsymbol{x}_0 + Q_{k+1}(\boldsymbol{H})\zeta\boldsymbol{A}^T\boldsymbol{b} + \sum_{j=0}^{k+1} N_{k+1,j}(\boldsymbol{H})\mathring{\boldsymbol{M}}_j.
\end{aligned}$$

We note that $P_{k+1}(\boldsymbol{H}), Q_{k+1}(\boldsymbol{H})$, and $N_{k+1,j}(\boldsymbol{H})$ are $k+1$-degree polynomials. The result immediately follows. □

Remark 2. *In this thesis we will define* $N_{k,j} = \mathbf{0}$ *for* $j > k$.

It is clear that P_k and Q_k are the residual and iterative polynomials, respectively, as in the full batch (*i.e.*, $\zeta = 1$) setting, but with coefficients c_{kj} scaled by the batch fraction, ζ. In this regard, one can view

$$y_k \stackrel{\text{def}}{=} P_k(\boldsymbol{H})\boldsymbol{x}_0 + Q_k(\boldsymbol{H})\zeta\boldsymbol{A}^T\boldsymbol{b} \tag{3.10}$$

as the k-th iterate of the full-batch (batch fraction, $\zeta = 1$) gradient-based method but with coefficients c_{kj} scaled by the batch fraction ζ. That is, $\{y_k\}$ with $\boldsymbol{y}_0 \overset{\text{def}}{=} \boldsymbol{x}_0$ are iterates generated by the gradient-based method

$$\boldsymbol{y}_{k+1} = \boldsymbol{x}_0 + \sum_{j=0}^{k} \zeta c_{k,j} \nabla f(\boldsymbol{y}_k). \tag{3.11}$$

In light of this and Proposition 4, we can express the iterates generated by the mini-batch gradient-based algorithm as

$$\begin{gathered} \boldsymbol{x}_k = \boldsymbol{y}_k + \sum_{j=0}^{k} N_{k,j}(\boldsymbol{H}) \mathring{\boldsymbol{M}}_j \\ \text{where the iterates } \boldsymbol{y}_k \text{ satisfy} \quad \boldsymbol{y}_{k+1} = \boldsymbol{x}_0 + \sum_{j=0}^{k} \zeta c_{kj} \nabla f(\boldsymbol{y}_k). \end{gathered} \tag{3.12}$$

Corollary 2 (Property of noise polynomials)**.** *Consider a mini-batch gradient-based method with coefficients $c_{k,j}$ and its corresponding noise polynomials $\{N_{k,j}\}_{k\in[0,\infty),j\leq k}$ as defined in Proposition 4. The noise polynomials satisfy*

$$N_{k,0}(\boldsymbol{H}) = \boldsymbol{0} \qquad k \geq 0$$

where the matrix $\boldsymbol{H} = \delta \boldsymbol{I} + \boldsymbol{A}^T \boldsymbol{A}$.

Proof. Consider the base case where $k = 0$ then $N_{0,0} = \boldsymbol{0}$ by definition. Assuming that $N_{k,0}(\boldsymbol{H}) = \boldsymbol{0}$ for $0 \leq i \leq k$, observe that

$$N_{k+1,0}(\boldsymbol{H}) = \sum_{i=1}^{k} \zeta c_{k,i} \boldsymbol{H} N_{i,0}(\boldsymbol{H}) = \boldsymbol{0}.$$

By strong induction the result holds. □

3.2.1 Examples of polynomials

Motivated by the identity linking the iterates to polynomials in Proposition 4, we derive these polynomials for some well-known stochastic optimization methods. For examples of the residual and iterative polynomials for popular algorithms in the full-batch setting (*i.e.*, batch fraction $\zeta = 1$), see [46]. We will exclude the proof of stochastic heavy-ball as the idea is similiar to that of the proof of Nesterov acceleration.

Stochastic gradient descent (SGD). Due to its simplicity in the recurrence of iterates for stochastic gradient descent [50], the polynomials P_k, Q_k, and $N_{k,j}$ are explicit. For learning rate $\gamma > 0$, the iterates of stochastic gradient are

$$\begin{aligned} \boldsymbol{x}_k &= \boldsymbol{x}_{k-1} - \gamma \sum_{i \in B_{k-1}} \nabla f_i(\boldsymbol{x}_{k-1}) \\ &= \boldsymbol{x}_{k-1} - \gamma \big[\boldsymbol{A}^T \boldsymbol{P}_k(\boldsymbol{A}\boldsymbol{x}_{k-1} - \boldsymbol{b}) + \zeta\delta \boldsymbol{x}_{k-1} \big] \\ &= \boldsymbol{x}_{k-1} - \gamma\zeta \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x}_{k-1} - \boldsymbol{b}) - \gamma\zeta\delta\boldsymbol{x}_{k-1} + \gamma \mathring{\boldsymbol{M}}_k, \end{aligned} \tag{3.13}$$

where $\boldsymbol{x}_0 \in \mathbb{R}^d$ is some initial vector. By unraveling the recursion, we deduce that

$$\boldsymbol{x}_k = \left(\mathbf{I} - \gamma\zeta \boldsymbol{H}_\delta\right)^k \boldsymbol{x}_0 + \sum_{t=1}^{k} \left(\mathbf{I} - \gamma\zeta \boldsymbol{H}_\delta\right)^{k-t} \gamma\zeta \boldsymbol{A}^T \boldsymbol{b} + \sum_{t=1}^{k} \gamma \left(\mathbf{I} - \gamma\zeta \boldsymbol{H}_\delta\right)^{k-t} \mathring{\boldsymbol{M}}_t, \qquad k \geq 0. \tag{3.14}$$

Here we defined $\boldsymbol{H}_\delta \overset{\text{def}}{=} \delta \mathbf{I}_d + \boldsymbol{A}^T\boldsymbol{A}$.

Stochastic heavy-ball. For learning rate $\gamma > 0$, the iterates of stochastic heavy-ball are

$$\begin{aligned} \boldsymbol{x}_k &= \boldsymbol{x}_{k-1} - \gamma \sum_{i \in B_k} \nabla f_i(\boldsymbol{x}_{k-1}) + \Delta(\boldsymbol{x}_{k-1} - \boldsymbol{x}_{k-2}) \\ &= \boldsymbol{x}_{k-1} - \gamma \big[\boldsymbol{A}^T \boldsymbol{P}_k(\boldsymbol{A}\boldsymbol{x}_{k-1} - \boldsymbol{b}) + \zeta\delta \boldsymbol{x}_{k-1} \big] + \Delta(\boldsymbol{x}_{k-1} - \boldsymbol{x}_{k-2}) \\ &= \boldsymbol{x}_{k-1} - \gamma\zeta \boldsymbol{A}^T \left(\boldsymbol{A}\boldsymbol{x}_{k-1} - \boldsymbol{b}\right) - \gamma\zeta\delta\boldsymbol{x}_{k-1} + \Delta(\boldsymbol{x}_{k-1} - \boldsymbol{x}_{k-2}) + \gamma \mathring{\boldsymbol{M}}_k \\ &= \left(\mathbf{I} - \gamma\zeta \left(\boldsymbol{A}^T\boldsymbol{A} + \delta\mathbf{I}\right) + \Delta\mathbf{I}\right) \boldsymbol{x}_{k-1} - \Delta\boldsymbol{x}_{k-2} + \gamma \left(\zeta \boldsymbol{A}^T\boldsymbol{b} + \mathring{\boldsymbol{M}}_k\right) \end{aligned} \tag{3.15}$$

where $\boldsymbol{x}_0 \in \mathbb{R}^d$ is some initial vector. By unraveling the recursion and applying the theory of generating functions, we deduce that

$$\boldsymbol{x}_{k+1} = \Delta^{k/2} U_k(\boldsymbol{C})\boldsymbol{x}_1 - \Delta^{\frac{k+1}{2}} \mathbf{1}_{\{k\geq 1\}} U_k(\boldsymbol{C})\boldsymbol{x}_0 + \gamma \Delta^{k/2} \sum_{t=1}^{k} U_k(\boldsymbol{C}) \left(\mathring{\boldsymbol{M}}_t + \zeta \boldsymbol{A}^T \boldsymbol{b} \right), \qquad k \geq 0$$

where $\{U_k : k \geq 0\}$ are the Chebyshev polynomials of the second kind and

$$\boldsymbol{C} \stackrel{\text{def}}{=} \mathbf{I} - \gamma\zeta \left(\boldsymbol{A}^T \boldsymbol{A} + \delta \mathbf{I} \right) + \Delta \mathbf{I} \tag{3.16}$$

Stochastic Nesterov acceleration (convex). Nesterov's celebrated accelerated method [41] generates iterates satisfying the recurrence for $k \geq 1$

$$\boldsymbol{x}_{k+1} = \big[1 + \Delta_{k-1}\big]\big[\boldsymbol{x}_k - \gamma \nabla f(\boldsymbol{x}_k)\big] - \Delta_{k-1}\big[\boldsymbol{x}_{k-1} - \gamma \nabla f(\boldsymbol{x}_{k-1})\big] \tag{3.17}$$

with the initial iteration such that $\boldsymbol{x}_0 \in \mathbb{R}^d$ and one gradient step for $\boldsymbol{x}_1$, that is, $\boldsymbol{x}_1 = \boldsymbol{x}_0 - \gamma \nabla f(\boldsymbol{x}_0)$. The learning rate $\gamma > 0$ and momentum parameter $\Delta(k) \geq 0$ satisfy

$$\Delta_k = \begin{cases} \frac{k}{k+3} & \text{if } \sigma_{\min}(\boldsymbol{A}) = 0 \\ \Delta & \text{if } \sigma_{\min}(\boldsymbol{A}) > 0. \end{cases} \tag{3.18}$$

The stochastic version of (3.17) replaces $\nabla f(\boldsymbol{x}_k)$ with a mini-batch $\sum_{i \in B_k} \nabla f_i(\boldsymbol{x}_k)$. The result is the following when applied to the ℓ^2-regularized least squares problem in the

strongly convex case, for $k \geq 1$,

$$
\begin{aligned}
\boldsymbol{x}_{k+1} &= \big[1+\Delta_{k-1}\big]\big[\boldsymbol{x}_k - \gamma\big(\boldsymbol{A}^T\boldsymbol{P}_{k+1}(\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}) + \delta\zeta\boldsymbol{x}_k\big)\big] \\
&\quad - \Delta_{k-1}\big[\boldsymbol{x}_{k-1} - \gamma\big(\boldsymbol{A}^T\boldsymbol{P}_{k+1}(\boldsymbol{A}\boldsymbol{x}_{k-1} - \boldsymbol{b}) + \delta\zeta\boldsymbol{x}_{k-1}\big)\big] \\
&= \big(1+\Delta_{k-1}\big)\big(\mathbf{I} - \gamma\zeta\boldsymbol{H}_\delta\big)\boldsymbol{x}_k - \Delta_{k-1}(\mathbf{I} - \gamma\zeta\boldsymbol{H}_\delta)\boldsymbol{x}_{k-1} \\
&\quad + \mathring{\boldsymbol{M}}_{k+1} - \gamma\zeta\boldsymbol{A}^T\boldsymbol{b} \\
&= P_k(\boldsymbol{H})\left[(\mathbf{I} - \gamma\zeta\boldsymbol{H})\boldsymbol{x}_0 + \gamma\mathring{\boldsymbol{M}}_1 + \gamma\zeta\boldsymbol{A}^T\boldsymbol{b}\right] + Q_k(\boldsymbol{H})\boldsymbol{x}_0 - \sum_{t=1}^{k}\gamma P_{k-t}(\boldsymbol{H})\zeta\boldsymbol{A}^T\boldsymbol{b} \\
&+ \sum_{t=1}^{k}\gamma P_{k-t}(\boldsymbol{H})\mathring{\boldsymbol{M}}_{t+1} \\
&= \left[P_k(\boldsymbol{H})(\mathbf{I} - \gamma\zeta\boldsymbol{H}) + Q_k(\boldsymbol{H})\right]\boldsymbol{x}_0 + \left[\gamma P_k(\boldsymbol{H}) - \gamma\sum_{t=1}^{k}P_{k-t}(\boldsymbol{H})\right]\zeta\boldsymbol{A}^T\boldsymbol{b} \\
&+ \left[\gamma P_k(\boldsymbol{H})\mathring{\boldsymbol{M}}_1 + \sum_{t=2}^{k+1}\gamma P_{k-t+1}(\boldsymbol{H})\mathring{\boldsymbol{M}}_t\right]
\end{aligned}
\tag{3.19}
$$

where

$$
\begin{aligned}
P_k(\lambda) &\stackrel{\text{def}}{=} \left[\Delta(1-\gamma\zeta\lambda)\right]^{k/2} U_k\left(\frac{(1+\Delta)\sqrt{1-\gamma\zeta\lambda}}{2\sqrt{\Delta}}\right), \qquad k \geq 0 \\
Q_k(\lambda) &\stackrel{\text{def}}{=} \begin{cases} -\left[\Delta(1-\gamma\zeta\lambda)\right]^{\frac{k+2}{2}} U_{k-1}\left(\frac{(1+\Delta)\sqrt{1-\gamma\zeta\lambda}}{2\sqrt{\Delta}}\right) & \text{if} \quad k \geq 1 \\ 0 & \text{if} \quad k = 0. \end{cases}
\end{aligned}
\tag{3.20}
$$

Denoting the residual polynomials in Proposition 4 as $P_k^{(\text{res})}, Q_k^{(\text{res})}, N_{k,j}^{(\text{res})}$ and by comparing (3.19) to Proposition 4 we obtain

$$
\begin{aligned}
P_{k+1}^{(res)}(\boldsymbol{H}) &= P_k(\boldsymbol{H})(\mathbf{I} - \gamma\zeta\boldsymbol{H}) + Q_k(\boldsymbol{H}) \qquad \forall k \geq 0 \\
Q_{k+1}^{(\text{res})}(\boldsymbol{H}) &= \gamma P_k(\boldsymbol{H}) - \gamma\sum_{t=1}^{k}P_{k-t}(\boldsymbol{H}) \qquad \forall k \geq 0 \\
N_{k+1,0}^{(\text{res})}(\boldsymbol{H}) &= \mathbf{0}, \qquad N_{k+1,j}^{(\text{res})}(\boldsymbol{H}) = \gamma P_{k+1-j}(\boldsymbol{H}) \qquad \forall k \geq 0, \forall 1 \leq j \leq k+1
\end{aligned}
\tag{3.21}
$$

Proof. We consider the case $\sigma_{min}(\boldsymbol{A}) \geq 0$. We can rewrite (3.19) for $k \geq 1$ as

$$\begin{aligned}
\begin{pmatrix} \boldsymbol{x}_{k+1} \\ \boldsymbol{x}_k \end{pmatrix} &= \begin{pmatrix} (1+\Delta)(\mathbf{I}-\gamma\zeta\boldsymbol{H}) & -\Delta(\mathbf{I}-\gamma\zeta\boldsymbol{H}) \\ \mathbf{I} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_k \\ \boldsymbol{x}_{k-1} \end{pmatrix} + \begin{pmatrix} \gamma\mathring{\boldsymbol{M}}_{k+1} - \gamma\zeta\boldsymbol{A}^T\boldsymbol{b} \\ \mathbf{0} \end{pmatrix} \\
&= \boldsymbol{G}^k \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_0 \end{pmatrix} + \sum_{t=1}^{k} \boldsymbol{G}^{k-t} \begin{pmatrix} \gamma\mathring{\boldsymbol{M}}_{t+1} - \gamma\zeta\boldsymbol{A}^T\boldsymbol{b} \\ \mathbf{0} \end{pmatrix} \qquad (3.22) \\
\text{where} \qquad \boldsymbol{G} &\stackrel{\text{def}}{=} \begin{pmatrix} (1+\Delta)(\mathbf{I}-\gamma\zeta\boldsymbol{H}) & -\Delta(\mathbf{I}-\gamma\zeta\boldsymbol{H}) \\ \mathbf{I} & \mathbf{0} \end{pmatrix}.
\end{aligned}$$

To construct the residual polynomials it suffices to find the upper left and upper right block matrices of the matrix $\boldsymbol{G}^k$ for $k \geq 0$. To this end, we observe that the upper left entry of $\boldsymbol{G}^k$ evolves as the polynomial P_k for $k \geq 0$ given by the following three-term recurrence,

$$\begin{gathered}
P_{k+1}(\lambda) = (1+\Delta)(1-\gamma\zeta\lambda)P_k(\lambda) - \Delta(1-\gamma\zeta\lambda)P_{k-1}(\lambda), \qquad k \geq 1, \\
\text{with} \qquad P_0(\lambda) = 1, \qquad P_1(\lambda) = (1+\Delta)(1-\gamma\zeta\lambda)
\end{gathered} \qquad (3.23)$$

We generate an explicit representation for the polynomial by constructing the generating function for the polynomials P_k, namely

$$\mathfrak{G}(\lambda, t) \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} t^k P_k(\lambda)$$

and identifying it as a scaled Chebyshev polynomial of the first kind. That is, using (3.23),

$$\begin{aligned}
\sum_{k=0}^{\infty} t^k P_{k+2}(\lambda) &= (1+\Delta)(1-\gamma\zeta\lambda)\sum_{k=0}^{\infty} t^k P_{k+1}(\lambda) - \Delta(1-\gamma\zeta\lambda)\sum_{k=0}^{\infty} t^k P_k(\lambda) \\
\text{(initial conditions)} \qquad & \frac{1}{t^2}\left[\mathfrak{G}(\lambda,t) - 1 - t(1+\Delta)(1-\gamma\zeta\lambda)\right] \\
&= \frac{(1+\Delta)(1-\gamma\zeta\lambda)}{t}\left[\mathfrak{G}(\lambda,t) - 1\right] - \Delta(1-\gamma\zeta\lambda)\mathfrak{G}(\lambda,t)
\end{aligned}$$

We solve this expression for $\mathfrak{G}$, which gives us

$$\mathfrak{G}(\lambda, t) = \frac{1}{1 - (1+\Delta)(1-\gamma\zeta\lambda)t + \Delta(1-\zeta\gamma\lambda)t^2}. \tag{3.24}$$

We note that (3.24) resembles the generating function for the Chebyshev polynomial of the second kind, namely,

$$\sum_{k=0}^{\infty} U_k(x)t^k = \frac{1}{1 - 2tx + t^2}. \tag{3.25}$$

To link (3.24) and (3.25), we perform the substitution $t \to \frac{\tilde{t}}{\sqrt{\Delta(1-\gamma\zeta\lambda)}}$ on (3.24) which gives us

$$\mathfrak{G}(\lambda, \tilde{t}) = \frac{1}{1 - 2x\tilde{t} + \tilde{t}^2} \tag{3.26}$$

where $x = \frac{(1+\Delta)\sqrt{1-\gamma\zeta\lambda}}{2\sqrt{\Delta}}$. We compare (3.25) and (3.26) to derive an expression for the polynomials P_k

$$P_k(\lambda) = [\Delta(1-\gamma\zeta\lambda)]^{k/2} U_k\left(\frac{(1+\Delta)\sqrt{1-\gamma\zeta\lambda}}{2\sqrt{\Delta}}\right), \qquad k \geq 0 \tag{3.27}$$

In light of (3.22), we find the polynomials Q_k by finding the upper right block matrix of $\boldsymbol{G}^k$. We do this by exploiting the structure of $\boldsymbol{G}$ and that $P_k(\boldsymbol{H})$ is the upper left block matrix of $\boldsymbol{G}^k$. Observe that for $k \geq 1$

$$\boldsymbol{G}^k = \boldsymbol{G}^{k-1}\boldsymbol{G}$$

$$\begin{pmatrix} * & Q_k(\boldsymbol{H}) \\ * & * \end{pmatrix} = \begin{pmatrix} P_{k-1}(\boldsymbol{H}) & * \\ * & 0 \end{pmatrix} \begin{pmatrix} (1+\Delta)(1-\gamma\zeta\boldsymbol{H}) & -\Delta(1-\gamma\zeta\boldsymbol{H}) \\ \mathbf{I} & 0 \end{pmatrix}$$

and so the upper right block matrix (i.e. Q_k) evolves as

$$\begin{aligned} Q_k(\boldsymbol{H}) &= -\Delta(1-\gamma\zeta\boldsymbol{H})P_{k-1}(\boldsymbol{H}) \\ &= \begin{cases} -[\Delta(1-\gamma\zeta\boldsymbol{H})]^{\frac{k+2}{2}} U_{k-1}\left(\frac{(1+\Delta)\sqrt{1-\gamma\zeta\boldsymbol{H}}}{2\sqrt{\Delta}}\right) & \text{if} \quad k \geq 1 \\ \mathbf{0} & \text{if} \quad k = 0. \end{cases} \end{aligned} \tag{3.28}$$

Using (3.22) and the initialization $\boldsymbol{x}_1 = \boldsymbol{x}_0 - \gamma \nabla f(\boldsymbol{x}_0)$ we can now write

$$\begin{aligned}\boldsymbol{x}_{k+1} &= P_k(\boldsymbol{H})\boldsymbol{x}_1 + Q_k(\boldsymbol{H})\boldsymbol{x}_0 - \sum_{t=1}^{k} \gamma P_{k-t}(\boldsymbol{H})\zeta \boldsymbol{A}^T\boldsymbol{b} + \sum_{t=1}^{k} P_{k-t}(\boldsymbol{H})\gamma \mathring{\boldsymbol{M}}_{t+1} \\ &= \left[P_k(\boldsymbol{H}) + Q_k(\boldsymbol{H})\right]\boldsymbol{x}_0 - \gamma P_k(\boldsymbol{H})\nabla f(\boldsymbol{x}_0) - \sum_{t=1}^{k} \gamma P_{k-t}(\boldsymbol{H})\zeta \boldsymbol{A}^T\boldsymbol{b} + \sum_{t=1}^{k} \gamma P_{k-t}(\boldsymbol{H})\mathring{\boldsymbol{M}}_{t+1}\end{aligned} \tag{3.29}$$

To write this equation in the form of Proposition 4 we can write

$$\begin{aligned}\nabla f(\boldsymbol{x}_0) &= \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{b}) + \delta \boldsymbol{x}_0 = \boldsymbol{H}\boldsymbol{x}_0 - \boldsymbol{A}^T\boldsymbol{b} \\ -\gamma P_k(\boldsymbol{H})\nabla f(\boldsymbol{x}_0) &= -\gamma P_k(\boldsymbol{H})\boldsymbol{H}\boldsymbol{x}_0 + \frac{\gamma}{\zeta} P_k(\boldsymbol{H})\zeta \boldsymbol{A}^T\boldsymbol{b}\end{aligned}$$

and so

$$\begin{aligned}\boldsymbol{x}_{k+1} = \left[P_k(\boldsymbol{H})\left(\mathbf{I} - \gamma \boldsymbol{H}\right) + Q_k(\boldsymbol{H})\right]\boldsymbol{x}_0 + \left[\frac{\gamma}{\zeta} P_k(\boldsymbol{H}) - \sum_{t=1}^{k} \gamma P_{k-t}(\boldsymbol{H})\right]\zeta \boldsymbol{A}^T\boldsymbol{b} \\ + \sum_{t=1}^{k} \gamma P_{k-t}(\boldsymbol{H})\mathring{\boldsymbol{M}}_{t+1}\end{aligned}$$

This seems a bit unnatural to me, so instead of taking a full-gradient on the first step, let's take a batch-stochastic one. To this end let $\boldsymbol{x}_1 = \boldsymbol{x}_0 - \gamma \sum_{i \in B_0} \nabla f(\boldsymbol{x}_0)$. Then

$$\sum_{i \in B_0} \nabla f(\boldsymbol{x}_0) = \boldsymbol{A}^T \boldsymbol{P}_1(\boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{b}) + \delta \frac{|B_0|}{n}\boldsymbol{x}_0 \tag{3.30}$$

and

$$\begin{aligned}\boldsymbol{x}_1 &= \boldsymbol{x}_0 - \gamma\left(\boldsymbol{A}^T\boldsymbol{P}_1(\boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{b}) + \delta\frac{|B_0|}{n}\boldsymbol{x}_0\right) \\ &= \boldsymbol{x}_0 + \gamma\left[\zeta \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{b}) - \boldsymbol{A}^T P_1(\boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{b})\right] - \gamma\zeta\left[\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{b}) + \delta\boldsymbol{x}_0\right] \\ &= (\mathbf{I} - \gamma\zeta\boldsymbol{H})\boldsymbol{x}_0 + \gamma\mathring{\boldsymbol{M}}_1 + \gamma\zeta\boldsymbol{A}^T\boldsymbol{b}\end{aligned} \tag{3.31}$$

then plugging (3.31) into (3.29) we obtain, in the form of Proposition 4

$$
\begin{aligned}
\boldsymbol{x}_{k+1} &= P_k(\boldsymbol{H})\left[(\mathbf{I}-\gamma\zeta\boldsymbol{H})\boldsymbol{x}_0+\gamma\mathring{\boldsymbol{M}}_1+\gamma\zeta\boldsymbol{A}^T\boldsymbol{b}\right]+Q_k(\boldsymbol{H})\boldsymbol{x}_0-\sum_{t=1}^{k}\gamma P_{k-t}(\boldsymbol{H})\zeta\boldsymbol{A}^T\boldsymbol{b} \\
&\quad+\sum_{t=1}^{k}\gamma P_{k-t}(\boldsymbol{H})\mathring{\boldsymbol{M}}_{t+1} \\
&= \left[P_k(\boldsymbol{H})(\mathbf{I}-\gamma\zeta\boldsymbol{H})+Q_k(\boldsymbol{H})\right]\boldsymbol{x}_0+\left[\gamma P_k(\boldsymbol{H})-\gamma\sum_{t=1}^{k}P_{k-t}(\boldsymbol{H})\right]\zeta\boldsymbol{A}^T\boldsymbol{b} \\
&\quad+\left[\gamma P_k(\boldsymbol{H})\mathring{\boldsymbol{M}}_1+\sum_{t=2}^{k+1}\gamma P_{k-t+1}(\boldsymbol{H})\mathring{\boldsymbol{M}}_t\right]
\end{aligned}
\tag{3.32}
$$

□

Remark 3. *One result of writing the iterates of mini-batch gradient-based methods in terms of residual polynomials as in Proposition 4 is that one can leverage the rich theory of polynomials to perform analysis on algorithms. For example, for Nesterov acceleration, we wrote the iterates in terms of sums of Chebyshev polynomials of the second kind, which is a well-studied polynomial. In fact, in future work, we leverage the theory of Chebyshev polynomials of the second kind in order to perform the type of analysis seen in Chapter 4. Also, one can redo the entire analysis of Chapter 4 from the perspective of polynomials.*

3.3 Resolvents and Statistics

The statistics we consider are all of the following form:

Definition 16. *A function $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$ is quadratic if it is a degree-2 polynomial or equivalently if it can be represented by*

$$\mathcal{R}(x) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{S}\boldsymbol{x}+\boldsymbol{h}^T\boldsymbol{x}+u. \tag{3.33}$$

We assume that $\boldsymbol{S} \in \mathbb{R}^{d\times d}$ is symmetric (not necessarily positive semidefinite), $\boldsymbol{h} \in \mathbb{R}^d$ is a vector, and $u \in \mathbb{R}$ is a constant. For any quadratic, define the H^2-norm:

$$||\mathcal{R}||_{H^2} \overset{def}{=} |u|+||\nabla\mathcal{R}(0)||_2+||\nabla^2\mathcal{R}||_2 = \|\boldsymbol{S}\|+\|\boldsymbol{h}\|+|u|.$$

To execute our exact dynamic of statistics, we require an additional assumption on the quadratic in the same spirit as Assumption 4:

Assumption 6 (Quadratic statistics)**.** *Suppose $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$ is quadratic, i.e., there is a symmetric matrix $\boldsymbol{S} \in \mathbb{R}^{d\times d}$, a vector $\boldsymbol{h} \in \mathbb{R}^d$, and a constant $u \in \mathbb{R}$ so that*

$$\mathcal{R}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{S}\boldsymbol{x} + \boldsymbol{h}^T\boldsymbol{x} + u. \tag{3.34}$$

We assume that $\mathcal{R}$ satisfies $\|\mathcal{R}\|_{H^2} \le C$ for some C independent of n and d. Moreover we assume the following (for the same Ω and α_0) as in Assumption 4:

$$\max_{z,y\in\Omega}\max_{1\le i\le n} |\boldsymbol{e}_i^T\boldsymbol{A}\boldsymbol{T}\boldsymbol{A}^T\boldsymbol{e}_i - \tfrac{1}{n}tr\big(\boldsymbol{A}\boldsymbol{T}\boldsymbol{A}^T\big)| \le \|\boldsymbol{S}\|n^{\alpha_0-1/2} \quad \textit{where} \quad \boldsymbol{T} \overset{\textit{def}}{=} R(z;\boldsymbol{A}^T\boldsymbol{A})\cdot\boldsymbol{S}\cdot R(y;\boldsymbol{A}^T\boldsymbol{A}). \tag{3.35}$$

This assumption ensures that the quadratic $\mathcal{R}$ has a Hessian which is not too correlated with any of the left singular vectors of $\boldsymbol{A}$. By incorporating the resolvent, we will be able to show that (3.35) holds for all polynomials in $\boldsymbol{A}^T\boldsymbol{A}$. Establishing Assumption 6 can be non-trivial in the cases when the quadratic statistic $\mathcal{R}$ has complicated dependence on $\boldsymbol{A}$. In simple cases, (especially for the case of the empirical risk and the norm) it follows automatically from Assumption 4.

3.4 Main Theorem

In this section we prove Theorem 3 below after we have introduced some definitions. The proof of Theorem 3 consists of constructing two deterministic quantities Ψ_t and Ω_t such that the difference of these quantities against the loss and any quadratic function (satisfying Assumption 6) under mini-batch gradient-based iterates, $f(\boldsymbol{x}_t)$ and $\mathcal{R}(\boldsymbol{x}_t)$, respectively, can be expressed as a summation of error terms that vanish in the high-dimensional limit. The proofs that these terms are small will be deferred to the next section.

To begin, recall that the dynamics of the iterates of any mini-batch gradient-based method under any quadratic function $\mathcal{R}(\cdot)$ satisfying Assumption 6 above, is expressed as

$$\mathcal{R}(\boldsymbol{x}_k) = \frac{1}{2}\boldsymbol{x}_k^T \boldsymbol{S}\boldsymbol{x}_k + \boldsymbol{h}^T\boldsymbol{x}_k + \boldsymbol{u}, \tag{3.36}$$

where the matrix $\boldsymbol{S} \in \mathbb{R}^{d\times d}$ is symmetric, the vector $\boldsymbol{h} \in \mathbb{R}^d$, $u \in \mathbb{R}$ is a constant, and the iterates $\boldsymbol{x}_k$ have updates that satisfy (3.2). We recall the mini-batch gradient-based iterates by (3.12)

$$\boldsymbol{x}_k = \boldsymbol{y}_k + \sum_{j=0}^{k} N_{k,j}(\boldsymbol{H})\mathring{\boldsymbol{M}}_j \tag{3.37}$$

where the iterates $\boldsymbol{y}_k$ satisfy $\boldsymbol{y}_{k+1} = \boldsymbol{x}_0 + \sum_{j=0}^{k}\zeta c_{kj}\nabla f(\boldsymbol{y}_k)$. Since everything is evaluated at $\boldsymbol{H} = \delta\mathbf{I} + \boldsymbol{A}^T\boldsymbol{A}$, we will often suppress the $\boldsymbol{H}$, that is, $\boldsymbol{N}_{k,j} = N_{k,j}(\boldsymbol{H})$.

Using (3.37), we can express the quadratic statistic $\mathcal{R}$, satisfying Assumption 6, applied to any mini-batch gradient-based method as follows:

$$\begin{aligned}
\mathcal{R}(\boldsymbol{x}_t) &= \mathcal{R}(\boldsymbol{y}_t) + \nabla\mathcal{R}(\boldsymbol{y}_t)^T\left(\sum_{k=1}^{t}\boldsymbol{N}_{t,k}\mathring{\boldsymbol{M}}_k\right) + \frac{1}{2}\left(\sum_{k=1}^{t}\boldsymbol{N}_{t,k}\mathring{\boldsymbol{M}}_k\right)^T(\nabla^2\mathcal{R})\left(\sum_{k=1}^{t}\boldsymbol{N}_{t,k}\mathring{\boldsymbol{M}}_k\right)\\
&= \mathcal{R}(\boldsymbol{y}_t) + \nabla\mathcal{R}(\boldsymbol{y}_t)^T\left(\sum_{k=1}^{t}\boldsymbol{N}_{t,k}\mathring{\boldsymbol{M}}_k\right)\\
&+ \frac{1}{2}\sum_{k=1}^{t}\mathring{\boldsymbol{M}}_k^T\boldsymbol{N}_{t,k}(\nabla^2\mathcal{R})\boldsymbol{N}_{t,k}\mathring{\boldsymbol{M}}_k + \sum_{k_1>k_2}\mathring{\boldsymbol{M}}_{k_1}^T\boldsymbol{N}_{t,k_1}(\nabla^2\mathcal{R})\boldsymbol{N}_{t,k_2}\mathring{\boldsymbol{M}}_{k_2}\\
&= \mathcal{R}(\boldsymbol{y}_t) + \frac{1}{2}\sum_{k=1}^{t}\sum_{\ell=1}^{n}\frac{(\zeta-\zeta^2)}{n}\mathrm{tr}\left(\boldsymbol{A}\tilde{\boldsymbol{N}}_{t,k}(\nabla^2 R;\boldsymbol{H})\boldsymbol{A}^T\right)\boldsymbol{w}_{k-1,\ell}^2 + \mathcal{E}_t^{\nabla}(\mathcal{R}) + \mathcal{E}_t^{\nabla^2\text{-Diag}}(\mathcal{R})\\
&+ \mathcal{E}_t^{\nabla^2-\text{Off}}(\mathcal{R})
\end{aligned} \tag{3.38}$$

where we define $\tilde{\boldsymbol{N}}_{t,k}(\boldsymbol{M};\boldsymbol{H}) \stackrel{\text{def}}{=} N_{t,k}(\boldsymbol{H})\boldsymbol{M}N_{t,k}(\boldsymbol{H})$ and the error terms to be

$$\mathcal{E}_t^{\nabla}(\mathcal{R}) \stackrel{\text{def}}{=} \nabla\mathcal{R}(\boldsymbol{y}_t)^T\left(\sum_{k=1}^{t}\boldsymbol{N}_{t,k}\mathring{\boldsymbol{M}}_k\right) \tag{3.39}$$

$$\mathcal{E}_t^{\nabla^2\text{-Diag}}(\mathcal{R}) \stackrel{\text{def}}{=} \frac{1}{2}\sum_{k=1}^{t}\mathring{\boldsymbol{M}}_k^T\boldsymbol{N}_{t,k}(\nabla^2\mathcal{R})\boldsymbol{N}_{t,k}\mathring{\boldsymbol{M}}_k \tag{3.40}$$

$$-\frac{1}{2}\sum_{k=1}^{t}\sum_{\ell=1}^{n}\frac{(\zeta-\zeta^2)}{n}\text{tr}\left(\boldsymbol{A}\tilde{\boldsymbol{N}}_{(t,k)}(\nabla^2 R;\boldsymbol{H})\boldsymbol{A}^T\right)\boldsymbol{w}_{k-1,\ell}^2 \tag{3.41}$$

$$\mathcal{E}_t^{\nabla^2\text{-Off}}(\mathcal{R}) \stackrel{\text{def}}{=} \sum_{k_1>k_2}^{t}\mathring{\boldsymbol{M}}_{k_1}^T\boldsymbol{N}_{t,k_1}(\nabla^2\mathcal{R})\boldsymbol{N}_{t,k_2}\mathring{\boldsymbol{M}}_{k_2}. \tag{3.42}$$

The terms, $\mathcal{E}_t^{\nabla}(\mathcal{R})$, $\mathcal{E}_t^{\nabla^2\text{-Diag}}(\mathcal{R})$, and $\mathcal{E}_t^{\nabla^2-\text{Off}}(\mathcal{R})$, are error terms that vanish when n or d is sufficiently large. We often will drop the $\mathcal{R}$ in the definition of the error terms if it is clear. In order to prove Theorem 3, we require a discrete version of Gronwall's inequality [24] used in the proof of

Lemma 3 (Discrete Gronwall Inequality)**.** *Let $\mathcal{E}(k)$ and $\tilde{\mathcal{K}}(k)$ be non-negative, non-decreasing sequences and $F(k)$ be a non-negative sequence defined for $k = 0, 1, 2, \ldots$. For any $T > 0$, suppose the sequences satisfy*

$$F(T) \leq \mathcal{E}(T) + \tilde{\mathcal{K}}(T)\sum_{0\leq k<T}F(k), \tag{3.43}$$

and $F(0) \leq \mathcal{E}(0)$. Then for any $T \geq 0$,

$$F(T) \leq \mathcal{E}(T) + T\cdot\mathcal{E}(T)\tilde{\mathcal{K}}(T)\exp\left(T\cdot\tilde{\mathcal{K}}(T)\right). \tag{3.44}$$

Proof. First, we will show by induction that the following holds

$$F(T) \leq \mathcal{E}(T) + \sum_{0\leq k<T}\tilde{\mathcal{K}}(T)\mathcal{E}(k)\prod_{k<j<T}\left(1+\tilde{\mathcal{K}}(j)\right) \tag{3.45}$$

For the base case, $T = 0$, the result holds by assumption that $F(0) \leq \mathcal{E}(0)$. Let $T > 0$ be arbitrary and assume that (3.45) holds for every $0 \leq \ell < T$. By (3.43) and the induction

hypothesis,

$$
\begin{aligned}
F(T) &\le \mathcal{E}(T) + \tilde{\mathcal{K}}(T) \sum_{0\le k<T} F(k) \\
&\le \mathcal{E}(T) + \tilde{\mathcal{K}}(T) \sum_{0\le k<T} \left(\mathcal{E}(k) + \tilde{\mathcal{K}}(k) \sum_{0\le j<k} \mathcal{E}(j) \prod_{j<i<k} \left(1+\tilde{\mathcal{K}}(j)\right) \right) \\
&= \mathcal{E}(T) + \sum_{0\le j<T} \mathcal{E}(j)\tilde{\mathcal{K}}(T) + \sum_{0\le j<T} \mathcal{E}(j)\tilde{\mathcal{K}}(T) \sum_{j<k<T} \tilde{\mathcal{K}}(k) \prod_{j<i<k} \left(1+\tilde{\mathcal{K}}(i)\right) \qquad (3.46) \\
&= \mathcal{E}(T) + \sum_{0\le j<T} \mathcal{E}(j)\tilde{\mathcal{K}}(T) \left(1 + \sum_{j<k<T} \tilde{\mathcal{K}}(k) \prod_{j<i<k} \left(1+\tilde{\mathcal{K}}(i)\right) \right) \\
&= \mathcal{E}(T) + \sum_{0\le j<T} \mathcal{E}(j)\tilde{\mathcal{K}}(T) \prod_{j<i<T} \left(1+\tilde{\mathcal{K}}(i)\right)
\end{aligned}
$$

where in the last equality we used the implicit induction

$$
\begin{aligned}
1 + \sum_{j<k<T} g_k \prod_{j<i<k} (1+g_i) &= 1 + g_{j+1}\cdot 1 + g_{j+2}(1+g_{j+1}) + \cdots + g_{T-1}(1+g_{j+1})\cdots(1+g_{T-2}) \\
&= (1+g_{j+1})(1 + g_{j+2} + \cdots + g_{T-1}(1+g_{j+2})\cdots(1+g_{T-2})) \\
&= \cdots \\
&= \prod_{j<i<T} (1+g_i)
\end{aligned}
$$

for any sequence $\{g_n\}$. By induction, we have shown that (3.45) holds for all $T \ge 0$.

Now we show that (3.44) holds. Since $1 + x \le \exp(x)$ for $x \in \mathbb{R}$, $\mathcal{E}(j) \le \mathcal{E}(T)$, and $\tilde{\mathcal{K}}(j) \le \tilde{\mathcal{K}}(T)$ for all $j \le T$, we obtain by (3.45)

$$
\begin{aligned}
F(T) &\le \mathcal{E}(T) + \sum_{0\le j<T} \mathcal{E}(j)\tilde{\mathcal{K}}(T) \prod_{j<i<T} \exp\left(\tilde{\mathcal{K}}(i)\right) \\
&\le \mathcal{E}(T) + \mathcal{E}(T)\tilde{\mathcal{K}}(T) \sum_{0\le j<T} \prod_{j<i<T} \exp\left(\tilde{\mathcal{K}}(T)\right) \qquad (3.47) \\
&\le \mathcal{E}(T) + T\cdot\mathcal{E}(T)\tilde{\mathcal{K}}(T) \exp\left(T\cdot\tilde{\mathcal{K}}(T)\right),
\end{aligned}
$$

and we have shown that (3.44) holds. □

Recall the vector $\boldsymbol{w}_t = \boldsymbol{A}\boldsymbol{x}_t - \boldsymbol{b}$ and empirical loss function

$$\mathcal{L}(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2.$$

It will be convenient to work with a stopped process based on the stopping time ϑ:

$$\vartheta \stackrel{\text{def}}{=} \inf\{t \geq 0 : ||\boldsymbol{w}_t|| > n^\theta\} \tag{3.48}$$

for some $\theta > 0$ to be determined. We define the stopped process of the iterates and the residual vector to be

$$\boldsymbol{x}_t^\vartheta \stackrel{\text{def}}{=} \boldsymbol{x}_{t\wedge\vartheta} \quad \text{and} \quad \boldsymbol{w}_t^\vartheta \stackrel{\text{def}}{=} \boldsymbol{w}_{t\wedge\vartheta} \quad t \geq 0.$$

Then we will remove the stopping time ϑ and get the result for the entire sequence $\boldsymbol{x}_t$. We are now ready to state our main result, the dynamics of any mini-batch gradient-based method. We will first prove that Theorem 3 holds for the stopped process $\boldsymbol{x}_t^\vartheta$.

Theorem 3 (Dynamics of Mini-Batch Gradient-Based Methods)**.** *Suppose Assumptions 2, 3 4, and 5 hold on the data matrix $\boldsymbol{A}$, targets $\boldsymbol{b}$, and initialization $\boldsymbol{x}_0$ with $\alpha_0 \in (0, 1/4)$. Let $\mathcal{R}(\cdot)$ be a general quadratic function satisfying Assumption 6 and the batch size satisfies $\beta/n = \zeta$ for some $\zeta > 0$. For $T > 0$, there exists $\tilde{C} > 0$ such that for any $c > 0$, there exists $D > 0$ satisfying*

$$\Pr\left[\sup_{0\leq t\leq T}\left\|\begin{pmatrix}\mathcal{L}(\boldsymbol{x}_t)\\ \mathcal{R}(\boldsymbol{x}_t)\end{pmatrix} - \begin{pmatrix}\Psi(t)\\ \Omega(t)\end{pmatrix}\right\| > C(T)n^{-\tilde{C}}\right] \leq Dn^{-c}, \tag{3.49}$$

where the $\boldsymbol{x}_t$ are any iterates with updates that satisfy (3.2) *(and therefore* (3.12)*) and $C(T)$ is a constant depending on T, independent of n and d. The functions Ψ and Ω are given as*

$$\begin{aligned}\Psi(t) &= \mathcal{L}(\boldsymbol{y}_t) + \sum_{k=1}^{t}(\zeta - \zeta^2)\mathcal{K}(t, k; \nabla^2\mathcal{L})\Psi(k-1)\\ \Omega(t) &= \mathcal{R}(\boldsymbol{y}_t) + \sum_{k=1}^{t}(\zeta - \zeta^2)\mathcal{K}(t, k; \nabla^2\mathcal{R})\Psi(k-1)\end{aligned} \tag{3.50}$$

where we define the quantities

$$\boldsymbol{y}_{t+1} = \boldsymbol{x}_0 + \sum_{k=0}^{t} \zeta c_{tk} \nabla f(\boldsymbol{y}_k) \qquad t \geq 0 \tag{3.51}$$

$$\tilde{\boldsymbol{N}}_{t,k}(\boldsymbol{M}; \boldsymbol{H}) = N_{t,k}(\boldsymbol{H}) \boldsymbol{M} N_{t,k}(\boldsymbol{H}) \qquad t \geq 0 \qquad \textit{and} \qquad 0 \leq k \leq t \tag{3.52}$$

$$\mathcal{K}(t, k; \boldsymbol{M}) = \frac{1}{n} tr\left(\tilde{\boldsymbol{N}}_{t,k}(\boldsymbol{M}; \boldsymbol{H}) \boldsymbol{A}^T \boldsymbol{A} \right) \qquad t \geq 0 \qquad \textit{and} \qquad 0 \leq k \leq t. \tag{3.53}$$

Proof. Let $\{N_{t,k}\}$ be the noise polynomial that corresponds to the mini-batch gradient-based iterates $\boldsymbol{x}_t$ (see Definition 2). As before, we will suppress the $\boldsymbol{H} = \boldsymbol{A}\boldsymbol{A}^T + \delta \mathbf{I}$ and simply use $\boldsymbol{N}_{t,k} = N_{t,k}(\boldsymbol{H})$ throughout this proof. First, we apply (3.38) replacing $\mathcal{R}$ with the least-squares loss function (3.1), f, so that

$$\mathcal{L}(\boldsymbol{x}_t) = \mathcal{L}(\boldsymbol{y}_t) + \nabla \mathcal{L}(\boldsymbol{y}_t)^T \left(\sum_{k=1}^{t} \boldsymbol{N}_{t,k} \mathring{\boldsymbol{M}}_k \right) + \frac{1}{2} \left(\sum_{k=1}^{t} \boldsymbol{N}_{t,k} \mathring{\boldsymbol{M}}_k \right)^T (\nabla^2 \mathcal{L}) \left(\sum_{k=1}^{t} \boldsymbol{N}_{t,k} \mathring{\boldsymbol{M}}_k \right)$$

Proceeding in the same way as (3.38) with $\mathcal{L}$ in place of $\mathcal{R}$ gives us

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}_t) = \mathcal{L}(\boldsymbol{y}_t) + \frac{1}{2} \sum_{k=1}^{t} \sum_{\ell=1}^{n} \frac{(\zeta - \zeta^2)}{n} \text{tr}\left(\boldsymbol{A} \tilde{\boldsymbol{N}}_{t,k}(\nabla^2 \mathcal{L}; \boldsymbol{H}) \boldsymbol{A}^T \right) \boldsymbol{w}_{k-1,\ell}^2 \\ + \mathcal{E}_t^{\nabla}(\mathcal{L}) + \mathcal{E}_t^{\nabla^2\text{-Diag}}(\mathcal{L}) + \mathcal{E}_t^{\nabla^2 - \text{Off}}(\mathcal{L}). \end{aligned} \tag{3.54}$$

From now on we suppress the argument of $\tilde{\boldsymbol{N}}_{t,k}$, that is, we write $\tilde{\boldsymbol{N}}_{t,k} = \tilde{\boldsymbol{N}}_{t,k}(\nabla^2\mathcal{L}; \boldsymbol{H})$. Hence for every $0 \le t \le T$ we get

$$
\begin{aligned}
|\mathcal{L}(\boldsymbol{x}_{t\wedge\vartheta}) - \Psi(t\wedge\vartheta)| \le \Bigg| & \sum_{k=1}^{t\wedge\vartheta} (\zeta-\zeta^2)\mathcal{K}\left(t\wedge\vartheta, k; \nabla^2\mathcal{L}\right)\Psi(k-1) \\
& \qquad - \frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\sum_{\ell=1}^{n}\frac{(\zeta-\zeta^2)}{n}\mathrm{tr}\left(\boldsymbol{A}\tilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\right)(\boldsymbol{w}^{\vartheta}_{k-1,\ell})^2\Bigg| \\
& + \left|\mathcal{E}^{\nabla}_{t\wedge\vartheta}(\mathcal{L})\right| + \left|\mathcal{E}^{\nabla^2\text{-Diag}}_{t\wedge\vartheta}(\mathcal{L})\right| + \left|\mathcal{E}^{\nabla^2\text{-Off}}_{t\wedge\vartheta}(\mathcal{L})\right| \\
& \le (\zeta-\zeta^2)\sum_{k=1}^{t\wedge\vartheta}\left|\mathcal{K}(t\wedge\vartheta, k; \nabla^2\mathcal{L})\right|\cdot\left|\mathcal{L}(\boldsymbol{x}^{\vartheta}_{k-1}) - \Psi\left((k-1)\wedge\vartheta\right)\right| \\
& + \left|\mathcal{E}^{\nabla}_{t\wedge\vartheta}(\mathcal{L})\right| + \left|\mathcal{E}^{\nabla^2\text{-Diag}}_{t\wedge\vartheta}(\mathcal{L})\right| + \left|\mathcal{E}^{\nabla^2\text{-Off}}_{t\wedge\vartheta}(\mathcal{L})\right|
\end{aligned}
\tag{3.55}
$$

where in the last inequality, we used the fact that $\mathcal{L}(\boldsymbol{x}_k) = \frac{1}{2}\|\boldsymbol{w}_k\|^2$ and $\mathcal{K}(t\wedge\vartheta, k; \boldsymbol{M}) = \frac{1}{n}\mathrm{tr}\left(\boldsymbol{A}\tilde{\boldsymbol{N}}_{t\wedge\vartheta,k}(\boldsymbol{M}; \boldsymbol{H})\boldsymbol{A}^T\right)$. In particular, using $t\wedge\vartheta \le T$, we obtain from (3.55),

$$
\begin{aligned}
\max_{0\le t\le T}|\mathcal{L}(\boldsymbol{x}_{t\wedge\vartheta}) - \Psi(t\wedge\vartheta)| \le & \max_{0\le t\le T}\left(\left|\mathcal{E}^{\nabla}_{t\wedge\vartheta}(\mathcal{L})\right| + \left|\mathcal{E}^{\nabla^2\text{-Diag}}_{t\wedge\vartheta}(\mathcal{L})\right| + \left|\mathcal{E}^{\nabla^2\text{-Off}}_{t\wedge\vartheta}(\mathcal{L})\right|\right) \\
& + \sum_{k=1}^{T}\left((\zeta-\zeta^2)\max_{0\le s\le T, 0\le\ell\le s}\left|\mathcal{K}(s\wedge\vartheta, \ell\wedge\vartheta; \nabla^2\mathcal{L})\right|\right)\left(\max_{0\le s\le k}\left|\mathcal{L}(\boldsymbol{x}^{\vartheta}_s) - \Psi(s\wedge\vartheta)\right|\right)
\end{aligned}
\tag{3.56}
$$

.

We define the following terms in order to rewrite (3.56) in a recursive form

$$
\begin{aligned}
F(T) &\stackrel{\text{def}}{=} \max_{0\le t\le T}\left|\mathcal{L}(\boldsymbol{x}^{\vartheta}_t) - \Psi(t\wedge\vartheta)\right| \\
\mathcal{E}(T;\mathcal{L}) &\stackrel{\text{def}}{=} \max_{0\le t\le T}\left|\mathcal{E}^{\nabla}_{t\wedge\vartheta}(\mathcal{L})\right| + \left|\mathcal{E}^{\nabla^2\text{-Diag}}_{t\wedge\vartheta}(\mathcal{L})\right| + \left|\mathcal{E}^{\nabla^2\text{-Off}}_{t\wedge\vartheta}(\mathcal{L})\right|, \\
\text{and} \quad \tilde{\mathcal{K}}(T;\mathcal{L}) &\stackrel{\text{def}}{=} (\zeta-\zeta^2)\max_{0\le s\le T, 0\le\ell\le s}\left|\mathcal{K}(s\wedge\vartheta, \ell\wedge\vartheta; \nabla^2\mathcal{L})\right|.
\end{aligned}
$$

Rewriting (3.56) in terms of these quantities gives the recursive form

$$F(T) \leq \mathcal{E}(T;\mathcal{L}) + \sum_{0\leq k<T} \tilde{\mathcal{K}}(T;\mathcal{L})F(k) \qquad \text{for } T \geq 0. \tag{3.57}$$

Note that $\{F(T) : T \in \mathbb{N}\}, \{\mathcal{E}(T;\mathcal{L}) : T \in \mathbb{N}\}$, and $\{\tilde{\mathcal{K}}(T;\mathcal{L}) : T \in \mathbb{N}\}$ are non-negative sequences. Moreover $\mathcal{E}(i)$ and $\tilde{\mathcal{K}}(i)$ are non-decreasing, and

$$F(0) = |\mathcal{L}(\boldsymbol{x}_0) - \Psi(0)| = 0 \leq \mathcal{E}(0).$$

Thus, the assumptions of Lemma 3 hold and we have that

$$\begin{aligned} F(T) &\leq \mathcal{E}(T;\mathcal{L}) + T \cdot \mathcal{E}(T;\mathcal{L})\tilde{\mathcal{K}}(T) \exp\left(T \cdot \tilde{\mathcal{K}}(T;\mathcal{L})\right) \\ &\leq C(T;\mathcal{L})\mathcal{E}(T;\mathcal{L}), \end{aligned} \tag{3.58}$$

where $C(T;\mathcal{L})$ is a constant independent of n and d. Moreover, observe that for $0 \leq t \leq T$, we have

$$\begin{aligned} |\Omega(t\wedge\vartheta) - \mathcal{R}(\boldsymbol{x}_{t\wedge\vartheta})| \leq \Bigg| &\sum_{k=1}^{t\wedge\vartheta} (\zeta-\zeta^2)\mathcal{K}\left((t\wedge\vartheta)-k;\nabla^2\mathcal{R}\right)\Psi(k-1) \\ &\quad - \frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\sum_{\ell=1}^{n} \frac{(\zeta-\zeta^2)}{n}\text{tr}\left(\boldsymbol{A}\tilde{\boldsymbol{N}}_{t\wedge\vartheta,k}\boldsymbol{A}^T\right)(\boldsymbol{w}^{\vartheta}_{k-1,\ell})^2\Bigg| \\ &+ \left|\mathcal{E}^{\nabla}_{t\wedge\vartheta}(\mathcal{R})\right| + \left|\mathcal{E}^{\nabla^2\text{-Diag}}_{t\wedge\vartheta}(\mathcal{R})\right| + \left|\mathcal{E}^{\nabla^2\text{-Off}}_{t\wedge\vartheta}(\mathcal{R})\right| \\ &\leq (\zeta-\zeta^2)\sum_{k=1}^{t\wedge\vartheta}\left|\frac{1}{n}\text{tr}\left(\tilde{\boldsymbol{N}}_{t\wedge\vartheta,k}\boldsymbol{A}^T\boldsymbol{A}\right)\right| \cdot \left|\Psi((k-1)\wedge\vartheta) - \mathcal{L}(\boldsymbol{x}^{\vartheta}_{k-1})\right| \\ &\quad + \left|\mathcal{E}^{\nabla}_{t\wedge\vartheta}(\mathcal{R})\right| + \left|\mathcal{E}^{\nabla^2\text{-Diag}}_{t\wedge\vartheta}(\mathcal{R})\right| + \left|\mathcal{E}^{\nabla^2\text{-Off}}_{t\wedge\vartheta}(\mathcal{R})\right| \end{aligned} \tag{3.59}$$

where we used the fact that $\mathcal{L}(\boldsymbol{x}_k) = \frac{1}{2}\|\boldsymbol{w}_k\|^2$ and $\tilde{\boldsymbol{N}}_{t,k} \overset{\text{def}}{=} \tilde{\boldsymbol{N}}_{t,k}(\nabla^2\mathcal{R}; \boldsymbol{H})$. Taking the supremum on both sides and applying (3.58) yields

$$\begin{aligned}
\max_{0\le t\le T} |\Omega(t\wedge\vartheta) - \mathcal{R}(\boldsymbol{x}_{t\wedge\vartheta})| &\le \max_{0\le t\le T} \left|\Psi(t\wedge\vartheta) - \mathcal{L}(\boldsymbol{x}_t^{\vartheta})\right| \sum_{k=1}^{t\wedge\vartheta} \left|\frac{1}{n}\text{tr}\left(\tilde{\boldsymbol{N}}_{t\wedge\vartheta,k}\boldsymbol{A}^T\boldsymbol{A}\right)\right| \\
&\quad + \max_{0\le t\le T} \left|\mathcal{E}^{\nabla}_{t\wedge\vartheta}(\mathcal{R})\right| + \max_{0\le t\le T} \left|\mathcal{E}^{\nabla^2\text{-Diag}}_{t\wedge\vartheta}(\mathcal{R})\right| + \max_{0\le t\le T} \left|\mathcal{E}^{\nabla^2\text{-Off}}_{t\wedge\vartheta}(\mathcal{R})\right| \\
&\le F(T) \max_{0\le t\le T} \left\{ \sum_{k=1}^{t\wedge\vartheta} \left|\frac{1}{n}\text{tr}\left(\tilde{\boldsymbol{N}}_{t\wedge\vartheta,k}\boldsymbol{A}^T\boldsymbol{A}\right)\right| \right\} + \mathcal{E}(T;\mathcal{R}) \\
&\le C(T;\mathcal{L})\mathcal{E}(T;\mathcal{L}) \max_{0\le t\le T} \left\{ \sum_{k=1}^{t\wedge\vartheta} \left|\frac{1}{n}\text{tr}\left(\tilde{\boldsymbol{N}}_{t\wedge\vartheta,k}\boldsymbol{A}^T\boldsymbol{A}\right)\right| \right\} + \mathcal{E}(T;\mathcal{R}) \\
&\le C(T;\mathcal{L};\mathcal{R}) \max\{\mathcal{E}(T;\mathcal{L}), \mathcal{E}(T;\mathcal{R})\}.
\end{aligned} \tag{3.60}$$

Here $C(T;\mathcal{L};\mathcal{R})$ is a constant independent of n and d. Given some $\alpha_0 \in (0, 1/4)$ we assign values of $\alpha, \alpha', \theta, \eta$, and δ so that each of the error terms in $\mathcal{E}(T;\mathcal{L})$ and $\mathcal{E}(T;\mathcal{R})$ becomes vanishingly small as n, d tend to infinity. To this end we assign the following values

$$\begin{aligned}
&\alpha := \frac{\alpha_0}{2} + \frac{1}{8} \qquad \alpha' := \frac{\alpha_0}{4} + \frac{3}{16} \qquad \delta := \frac{\alpha_0}{8} + \frac{7}{32} \\
&\theta = \frac{1}{4}\left(-\frac{\alpha_0}{4} + \frac{1}{16}\right) \qquad \eta := \frac{\delta - \alpha}{2}.
\end{aligned} \tag{3.61}$$

Using these values for Proposition 8, Proposition 9, and Proposition 10 yields

$$\begin{aligned}
\mathcal{E}(T;\mathcal{R}) &\overset{\text{def}}{=} \max_{0\le t\le T} \left|\mathcal{E}^{\nabla}_{t\wedge\vartheta}\right| + \left|\mathcal{E}^{\nabla^2\text{-Diag}}_{t\wedge\vartheta}\right| + \left|\mathcal{E}^{\nabla^2\text{-Off}}_{t\wedge\vartheta}\right| \\
&\le \max_{0\le t\le T} \left|\mathcal{E}^{\nabla}_{t\wedge\vartheta}\right| + \max_{0\le t\le T} \left|\mathcal{E}^{\nabla^2\text{-Diag}}_{t\wedge\vartheta}\right| + \max_{0\le t\le T} \left|\mathcal{E}^{\nabla^2\text{-Off}}_{t\wedge\vartheta}\right| \\
&\le C(T;\mathcal{R}) n^{-c} \qquad \text{w.o.p.}
\end{aligned} \tag{3.62}$$

for some $c > 0$ and similarly,

$$\mathcal{E}(T;\mathcal{L}) \le C(T;\mathcal{L}) n^{-c} \qquad \text{w.o.p.}$$

for some $c > 0$. Hence we have shown that the following holds with overwhelming probability

$$\max_{0\le t\le T} |\Omega(t\wedge\vartheta) - \mathcal{R}(\boldsymbol{x}_{t\wedge\vartheta})| \le C(T;\mathcal{L};\mathcal{R})n^{-c} \qquad \text{and} \tag{3.63}$$

$$\max_{0\le t\le T} \left|\mathcal{L}(\boldsymbol{x}_t^{\vartheta}) - \Psi(t\wedge\vartheta)\right| \le C(T;\mathcal{L})n^{-c} \tag{3.64}$$

for some constants $c > 0$ and $C(T;\mathcal{L}), C(T;\mathcal{L};\mathcal{R})$ which are independent of n and d and only depends on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta, T, |\Omega|$. To finish off the proof, we now show that the stopping time satisfies $\vartheta > T$ with overwhelming probability. For sufficiently large n, by the definition of the stopping time, the fact that $\mathcal{L}(\boldsymbol{x}_t) = \frac{1}{2}\|\boldsymbol{w}_t\|_2^2$, and using the constant $c > 0$ in (3.63) yields

$$\begin{aligned}
\Pr(\vartheta > T) &\ge \Pr\left(\max_{0\le t\le T} \left|\mathcal{L}(\boldsymbol{x}_t^{\vartheta}) - \Psi(t\wedge\vartheta)\right| \le \frac{1}{2}n^{2\theta} - \max_{0\le t\le T}\Psi(t)\right) \\
&\ge 1 - \Pr\left(\max_{0\le t\le T} \left|\mathcal{L}(\boldsymbol{x}_t^{\vartheta}) - \Psi(t\wedge\vartheta)\right| \ge \frac{1}{2}n^{2\theta} - \max_{0\le t\le T}\Psi(t)\right) \\
&\ge 1 - \Pr\left(\max_{0\le t\le T} \left|\mathcal{L}(\boldsymbol{x}_t^{\vartheta}) - \Psi(t\wedge\vartheta)\right| \ge n^{-c}\right) \\
&\ge 1 - Dn^{-\tilde{c}}
\end{aligned} \tag{3.65}$$

for some constant $\tilde{c} > 0$ where the last line follows from the result of (3.63). We note that $\max_{0\le t\le T}\Psi(t)$ is independent of n and d and the maximum is taken over a finite set so the maximum is finite. Using (3.63) and (3.65) we get

$$\begin{aligned}
&\Pr\left(\max_{0\le t\le T} |\mathcal{L}(\boldsymbol{x}_t) - \Psi(t)| > n^{-c}\right) = \Pr\left(\max_{0\le t\le T} |\mathcal{L}(\boldsymbol{x}_t) - \Psi(t)| \left(\mathbf{1}_{\{\vartheta>T\}} + \mathbf{1}_{\{\vartheta\le T\}}\right) > n^{-c}\right) \\
&\le \Pr\left(\max_{0\le t\le T} |\mathcal{L}(\boldsymbol{x}_t) - \Psi(t)| \mathbf{1}_{\{\vartheta\le T\}} > n^{-c}\right) + \Pr\left(\max_{0\le t\le T} \left|\mathcal{L}(\boldsymbol{x}_t^{\vartheta}) - \Psi(t\wedge\vartheta)\right| > n^{-c}\right) \\
&\le Dn^{-\tilde{c}}
\end{aligned} \tag{3.66}$$

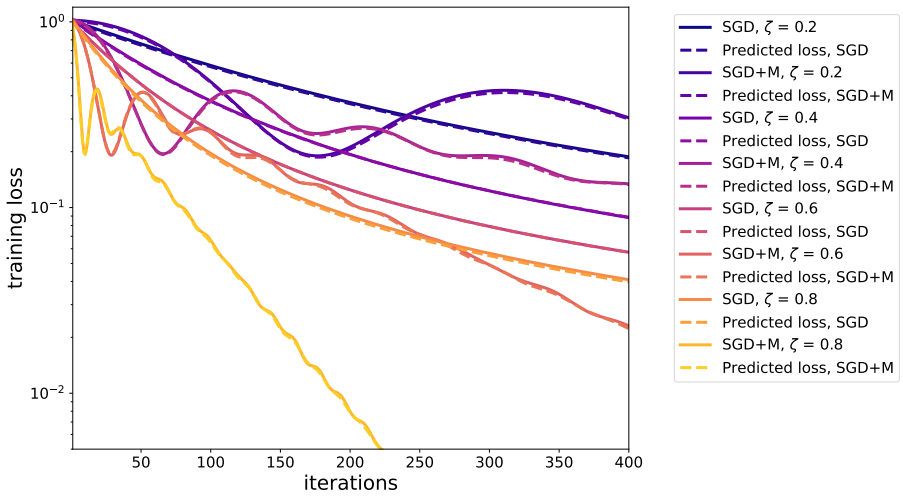

Figure 3.1: Following the setup in Section 3.5.1 we can predict the training loss under the iterates of Nesterov acceleration under the strongly convex setting. We specified the data matrix $\boldsymbol{A}$ as iid standard Gaussian such that the rows are standardized accordingly.

for some constants D and $c > 0$ that are independent of n and d. This shows

$$\max_{0 \le t \le T} |\Psi(t) - \mathcal{L}(\boldsymbol{x}_t)| \le n^{-c}$$

holds with overwhelming probability. Similarly, we can show

$$\max_{0 \le t \le T} |\Omega(t) - \mathcal{R}(\boldsymbol{x}_t)| \le n^{-c}$$

holds with overwhelming probability which gives us what we need. □

3.5 Motivating applications

We present some settings which satisfy our setup. In particular, our setup enables any quadratic test error. The following results can be found in [45].

3.5.1 Training loss

One important quadratic statistic, which allows for analysis of the optimization aspects of SGD in high dimensions, is the ℓ^2-regularized loss function f in (3.1). Then provided $\boldsymbol{A}, \boldsymbol{b}$ satisfy Assumptions 4, 5, and 6, $\boldsymbol{x}_0$ is iid subgaussian, then Theorem 3 shows that $f(\boldsymbol{x}_k)$ concentrates around the solution of a Volterra integral equation. A natural setup which satisfies Assumption 4 and 5 is the following:

Assumption 7. *Suppose $M > 0$ is a constant. Suppose that $\boldsymbol{\Sigma}$ is a positive semi-definite $d \times d$ matrix with tr $\boldsymbol{\Sigma}$ and $\|\boldsymbol{\Sigma}\| \leq M\sqrt{d} < \infty$. Suppose that $\boldsymbol{A}$ is a random matrix $\boldsymbol{A} = \boldsymbol{Z}\sqrt{\boldsymbol{\Sigma}}$ where $\boldsymbol{Z}$ is an $n \times d$ matrix of independent, mean 0, variance 1 entries with subgausssian norm at most $M < \infty$, and suppose $n \leq Md$. Finally suppose that $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{\xi}$ for $\boldsymbol{\beta}, \boldsymbol{\xi}$ iid centered subgaussian satisfying $\|\boldsymbol{\beta}\|^2 = R$ and $\|\boldsymbol{\xi}\|^2 = \tilde{R}\frac{n}{d}$ for some signal and noise constant $R, \tilde{R} > 0$, respectively.*

Under these assumptions we conclude:

Theorem 4. *Suppose $(\boldsymbol{A}, \boldsymbol{b})$ satisfy Assumption 7, $\delta > 0$ and $\boldsymbol{x}_0$ is iid centered subgaussian with $\mathbb{E}\|\boldsymbol{x}_0\|^2 = \hat{R}$. Suppose that $\gamma > 0$ then for some $\epsilon > 0$, for all $T > 0$, and for all $D > 0$ there is a $C > 0$ such that*

$$\Pr\left(\sup_{0\leq t\leq T}\left\|\begin{pmatrix} f(\boldsymbol{x}_t) \\ \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{\beta}\|^2 \end{pmatrix} - \begin{pmatrix} \Psi(t) \\ \Omega(t) \end{pmatrix}\right\|^2 > d^{-\epsilon}\right) \leq Cd^{-D},$$

where $\Psi(t)$ and $\Omega(t)$ solves (3.49) *with $\mathcal{R} = \frac{1}{2}\|\cdot - \boldsymbol{\beta}\|^2$.*

Remark. Theorem 4 states that we can simultaneously capture the dynamics of the training loss in terms of the loss values $f(\boldsymbol{x}_t)$ and the distance from the optimal iterate $\frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{\beta}\|$.

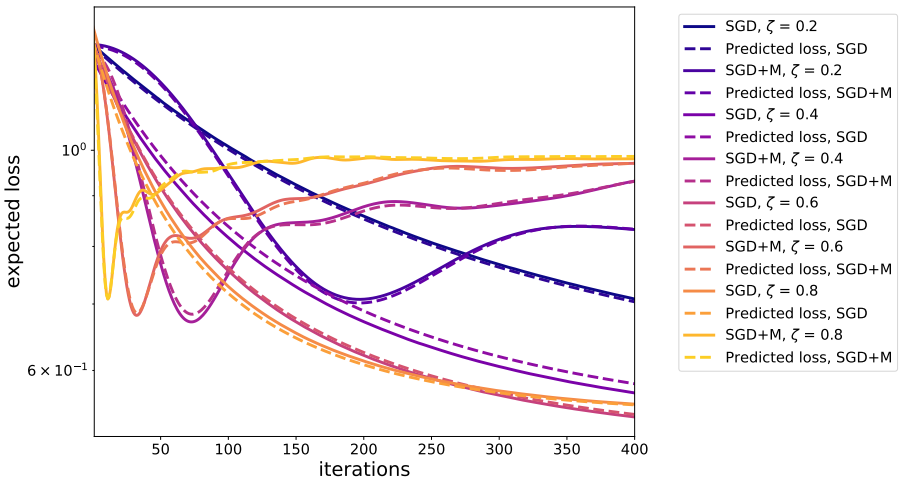

Figure 3.2: Following the setup in Section 3.5.2 we can predict the test loss under the iterates of Nesterov acceleration under the strongly convex setting. We specified the data matrix $\boldsymbol{A}$ as iid standard Gaussian such that the rows are standardized accordingly.

3.5.2 Generalization

In the standard linear regression setup, we suppose that $\boldsymbol{A}$ is generated by taking n independent d-dimensional samples from a centered distribution $\mathcal{D}_f$ which we assume to be standardized (mean 0 and expected sample-norm-squared 1). We let the matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ be the feature covariance of $\mathcal{D}_f$, that is

$$\boldsymbol{\Sigma}_f \stackrel{\text{def}}{=} \mathbb{E}[\boldsymbol{a}\boldsymbol{a}^T], \qquad \text{where} \qquad \boldsymbol{a} \sim \mathcal{D}_f. \tag{3.67}$$

Suppose there is as linear ground truth function $\beta : \mathbb{R}^d \to \mathbb{R}$, which for simplicity we suppose to have $\beta(0) = 0$. In this case, we identify β with a vector using the representation $\boldsymbol{a} \to \boldsymbol{\beta}^T \boldsymbol{a}$. We suppose that our data is drawn from a distribution $\mathcal{D}$ on $\mathbb{R}^d \times \mathbb{R}$, with the property that

$$\mathbb{E}[b|\boldsymbol{a}] = \boldsymbol{\beta}^T \boldsymbol{a}, \qquad \text{where} \qquad (\boldsymbol{a}, b) \sim \mathcal{D},$$

and the data $\boldsymbol{a} \sim \mathcal{D}_f$.

Hence we suppose that $[\boldsymbol{A}|\boldsymbol{b}]$ is a $\mathbb{R}^{n\times d} \times \mathbb{R}^{n\times 1}$ matrix on independent samples from $\mathcal{D}$. The vector $\boldsymbol{x}_t$ represents an estimate of $\boldsymbol{\beta}$, and the expected (population) is

$$\mathcal{R}(\boldsymbol{x}_t) \stackrel{\text{def}}{=} \frac{1}{2}\mathbb{E}\left[(b - \boldsymbol{x}_t^T\boldsymbol{a})^2|\boldsymbol{x}_t\right]$$

where expectation is taken over $(\boldsymbol{a}, b) \sim \mathcal{D}$, and where $(\boldsymbol{a}, b)$ is a sample independent of $\boldsymbol{x}_t$. We can rewrite the expected risk in as a bias-variance decomposition in terms of the feature covariance matrix $\boldsymbol{\Sigma}_f$ and the noise $\eta^2 \stackrel{\text{def}}{=} \mathbb{E}\left[(b - \boldsymbol{\beta}^T\boldsymbol{a})^2\right]$ to give

$$\mathcal{R}(\boldsymbol{x}_t) = \frac{1}{2}\eta^2 + \frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{x}_t\right)^T \boldsymbol{\Sigma}_f \left(\boldsymbol{\beta} - \boldsymbol{x}_t\right). \tag{3.68}$$

It is important to recall that the sequence $\{\boldsymbol{x}_t\}_{t\geq 0}$ is generated from iterates of any mini-batch gradient-based method applied to ℓ^2-regularized least-squares problem (3.1).

In the case that $(\boldsymbol{a}, b)$ is jointly Gaussian, it follows that we may represent

$$\boldsymbol{a} = \boldsymbol{\Sigma}_f^{1/2}\boldsymbol{z}, \qquad b = \boldsymbol{\beta}^T\boldsymbol{a} + \eta w, \qquad \text{where} \qquad (\boldsymbol{z}, w) \sim N(0, \mathbf{I}_d \oplus 1).$$

Therefore, it follows that the iterates $\boldsymbol{x}_t$ are generated from a mini-batch gradient-based method applied to the problem:

$$\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \frac{\delta}{2}\|\boldsymbol{x}\|^2 \qquad \text{where} \qquad \boldsymbol{b} = \boldsymbol{A}\boldsymbol{\beta} + \eta\boldsymbol{w},$$

and the vector $\boldsymbol{w}$ is iid $N(0, 1)$ random variables, independent of $\boldsymbol{A}$. This is also known as the generative model with noise.

Moreover, if D satisfies Assumption 7 (with $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_f$) then the population risk $\mathcal{R}(\boldsymbol{x}_t)$ is well approximated by Ω:

Theorem 5. *Suppose $(\boldsymbol{A}, \boldsymbol{b})$ saitsfy Assumption 7, $\delta > 0$ and $\boldsymbol{x}_0$ is iid centered subgaussian with $\mathbb{E}\|\boldsymbol{x}_0\|^2 = \hat{R}$. Suppose $\gamma > 0$ then for some $\epsilon > 0$, for all $T > 0$, and for all $D > 0$ there is a $C > 0$*

such that

$$\Pr\left[\sup_{0\leq t\leq T}\left\|\begin{pmatrix} f(\boldsymbol{x}_t) \\ \mathcal{R}(\boldsymbol{x}_t)\end{pmatrix} - \begin{pmatrix}\Psi(t) \\ \Omega(t)\end{pmatrix}\right\| > d^{-\epsilon}\right] \leq Cd^{-D},$$

where Ψ_t *and* Ω_t *solves* (3.49) *with* $\mathcal{R}$ *given by* (3.68).

Remark. Under Assumption 7 (and in-distribution) that $\eta^2 = \frac{\tilde{R}}{d}$. In the case of out-of-distribution regression, we have that $\eta^2 \neq \frac{\tilde{R}}{d}$ as the η represents the population noise.

3.5.3 Random features

We follow a setup based upon [4, 38]. As before, we suppose that the data matrix $\boldsymbol{X}$ is generated by taking n independent n_0-dimensional samples from a centered distribution $\mathcal{D}_f$ with feature covariance

$$\boldsymbol{\Sigma}_f \stackrel{\text{def}}{=} \mathbb{E}[\boldsymbol{X}_i^T \boldsymbol{X}_i], \qquad \text{where} \boldsymbol{X}_i \in \mathbb{R}^{1\times n_0} \text{ and } \boldsymbol{X}_i \sim \mathcal{D}_f.$$

We suppose for simplicity that $\boldsymbol{X}$ is a data matrix having dimension $n \times n_0$ whose iid rows are drawn from a multivariate Gaussian with covariance $\boldsymbol{\Sigma}_f$ and nice covariance structure:

Assumption 8. *The distribution* $\mathcal{D}_f$ *is multivariate normal and the covariance matrix* $\boldsymbol{\Sigma}_f$ *of the random features data satisfies for some* $C > 0$

$$\frac{1}{n_0} tr\left(\boldsymbol{\Sigma}_f\right) = 1 \quad \textit{and} \qquad \|\boldsymbol{\Sigma}_f\| \leq C.$$

This allows $\boldsymbol{X}$ to be represented equivalently as $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}/\sqrt{n_0}$ for a iid standard Gaussian matrix $\boldsymbol{Z}$. We suppose that the random features projection matrix $\boldsymbol{W} \in \mathbb{R}^{n_0\times d}$ is an iid matrix having standard Gaussian entries and independent of $\boldsymbol{Z}$ so that $\boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}\boldsymbol{W}/\sqrt{n_0}$ is element-wise standardized.

We let σ be an activation function satisfying:

Assumption 9. *The activation function satisfies for some* $C_0, C_1 \geq 0$

$$|\sigma'(x)| \leq C_0 e^{C_1|x|}, \qquad \textit{for all } x \in \mathbb{R}, \qquad \textit{and for standard normal } Z, \qquad \mathbb{E}\sigma(Z) = 0.$$

For example, the rectified linear unit (ReLU) with an appropriate shift satisfies Assumption 9.

We note from the outset, the growth rate of the derivative of the activation function implies a similar bound on the growth rate of the underlying activation function σ. As before, we suppose the data $[\boldsymbol{X}|\boldsymbol{b}]$ is arranged in the matrix $\mathbb{R}^n \times (\mathbb{R}^{n_0} \times \mathbb{R})$ where each row is an independent sample from $\mathcal{D}$. We now transform the data $\boldsymbol{X} \in \mathbb{R}^{n \times n_0}$ by

$$\boldsymbol{A} = \sigma\left(\boldsymbol{X}\boldsymbol{W}/\sqrt{n_0}\right) \in \mathbb{R}^{n \times d},$$

where $\boldsymbol{W} \in \mathbb{R}^{n_0 \times d}$ is a matrix independent of $[\boldsymbol{X}|\boldsymbol{b}]$ of independent standard normals. The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is applied element-wise.

We introduce the following notation

$$\begin{aligned} \boldsymbol{\Sigma}_\sigma(\boldsymbol{W}) &\stackrel{\text{def}}{=} \mathbb{E}\left[\sigma(\boldsymbol{X}_i\boldsymbol{W}/\sqrt{n_o})^T\sigma(\boldsymbol{X}_i\boldsymbol{W}/\sqrt{n_0})|\boldsymbol{W}\right] \qquad \text{and} \\ \hat{\sigma}(\boldsymbol{W}) &\stackrel{\text{def}}{=} \mathbb{E}[\boldsymbol{X}_i^T\sigma(\boldsymbol{X}_i\boldsymbol{W}/\sqrt{n_0})|\boldsymbol{W}]. \end{aligned} \tag{3.69}$$

Then we can write the expected risk as

$$\begin{aligned} \mathcal{R}(\boldsymbol{x}_t) &\stackrel{\text{def}}{=} \mathbb{E}[(b - \boldsymbol{x}_t^T\sigma(\boldsymbol{X}_i\boldsymbol{W}/\sqrt{n_0}))^2|\boldsymbol{x}_t, \boldsymbol{W}] \\ &= \eta^2 + \mathbb{E}\left[(\boldsymbol{X}_i\boldsymbol{\beta} - \sigma(\boldsymbol{X}_i\boldsymbol{W}/\sqrt{n_0})\boldsymbol{x}_t)^2|\boldsymbol{x}_t, \boldsymbol{W}\right] \\ &= \eta^2 + \boldsymbol{\beta}^T\boldsymbol{\Sigma}_f\boldsymbol{\beta} + \boldsymbol{x}_t^T\boldsymbol{\Sigma}_\sigma(\boldsymbol{W})\boldsymbol{x}_t - 2\boldsymbol{\beta}^T\hat{\sigma}(\boldsymbol{W})\boldsymbol{x}_t, \\ &\qquad \text{where } (\boldsymbol{X}_i, b) \sim D \text{ and } \mathbb{E}[b|\boldsymbol{X}_i] = \boldsymbol{X}_i\boldsymbol{\beta}. \end{aligned} \tag{3.70}$$

The ℓ^2-regularized least-squares problem is now

$$\min_x \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \frac{\delta}{2}\|\boldsymbol{x}\|^2 \qquad \text{where} \qquad \boldsymbol{b} = \boldsymbol{X}\boldsymbol{\beta} + \eta\boldsymbol{w}$$

which forms the random features regression problem. One can see this as a two-layer neural network model, in which the hidden layer has d nodes. However, the hidden layer weights are simply generated randomly in advanced and left untrained. The optimization is only performed on the final layers's weights $(\boldsymbol{x})$.

Theorem 6. *Suppose that n, d, n_0 are proportinally related. Suppose that the data matrix $\boldsymbol{X}$ stiasfies Assumption 8 and the random features $\boldsymbol{W}$ are iid standard normal. Suppose $\boldsymbol{b} = \boldsymbol{X}\boldsymbol{\beta} + \eta\boldsymbol{w}$ with $\boldsymbol{\beta}, \boldsymbol{w}$ independent isotropic subgaussian vectors with $\mathbb{E}\|\boldsymbol{\beta}\|^2 = 1/n_0$ and $\mathbb{E}\|\boldsymbol{w}\|^2 = 1$ and η bounded independent of n. Suppose the activation function satisfies Assumption 9 and $\gamma > 0$. Suppose the initialization $\boldsymbol{x}_0$ is iid centered subgaussian with $\mathbb{E}\|\boldsymbol{x}_0\|^2 = \hat{R}$. Then for some $\epsilon > 0$, for all $T > 0$, and for all $D > 0$ there is a $C > 0$ such that*

$$\Pr\left[\sup_{0\le t\le T}\left\|\begin{pmatrix} f(\boldsymbol{x}_t) \\ \mathcal{R}(\boldsymbol{x}_t)\end{pmatrix} - \begin{pmatrix}\Psi(t) \\ \Omega(t)\end{pmatrix}\right\| > d^{-\epsilon}\right] \le Cd^{-D},$$

where Ψ_t and Ω_t solves (3.49) *with $\mathcal{R}$ given by* (3.70).

3.6 Martingale Bounds

In this section, without loss of generality, we normalize our data matrix $\boldsymbol{A}$ so that it has row sum equals 1, namely

$$||\boldsymbol{A}_i||_2 = 1 \quad i \in [n] \tag{3.71}$$

Proposition 5 (Resolvent and Bounded Entries)**.** *Fix a constant $T > 0$. Suppose Assumptions 4 hold for the closed, bounded contour Ω. For any $\alpha > \alpha_0 + \theta$,*

$$\begin{aligned} \max_{0\le t\le T}\max_{z\in\Omega}\max_{1\le i\le n}|e_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_t^\vartheta| &\le C(|\Omega|, \|\boldsymbol{A}^T\boldsymbol{A}\|, T)n^{\alpha-1/2} \\ \textit{and} \quad \max_{0\le t\le T}\max_{1\le i\le n}|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_t^\vartheta| &\le C(|\Omega|, \|\boldsymbol{A}^T\boldsymbol{A}\|, T)n^{\alpha-1/2} \end{aligned} \tag{3.72}$$

with overwhelming probability (conditioned on $\mathcal{F}_0$), where C is constant depending on $\|\Omega\|$, $\|\boldsymbol{A}^T\boldsymbol{A}\|$, and time T, but independent of n and d.

We immediately get a consequence which yields a bound on the individual entries of $\boldsymbol{w}_t^{\vartheta}$.

Lemma 4 (Coordinates are Small in n)**.** *Suppose the assumptions of Proposition 5 hold with some $T \geq 0$. For all $\alpha > \alpha_0 + \theta$,*

$$\begin{aligned} \max_{0\leq t\leq T}\max_{1\leq i\leq n} |\boldsymbol{e}_i^T \boldsymbol{w}_t^{\vartheta}| &\leq C(|\Omega|, \|\boldsymbol{A}^T\boldsymbol{A}\|, T) n^{\alpha-1/2} \\ \textit{and} \quad \max_{0\leq t\leq T}\max_{1\leq i\leq n} |\boldsymbol{e}_i^T \boldsymbol{A}\boldsymbol{x}_t^{\vartheta}| &\leq C(|\Omega|, \|\boldsymbol{A}^T\boldsymbol{A}\|, T) n^{\alpha-1/2} \end{aligned} \tag{3.73}$$

with overwhelming probability (conditioned on $\mathcal{F}_0$).

Proof of Lemma 4. . From Cauchy's integral formula, we express the i-th entry of $\boldsymbol{A}\boldsymbol{x}_t^{\vartheta}$ and $\boldsymbol{w}_t^{\vartheta}$, respectively, as

$$\begin{aligned} |\boldsymbol{e}_i^T \boldsymbol{A}\boldsymbol{x}_t^{\vartheta}| &= \left|\frac{-1}{2\pi i}\oint_{\Omega} \boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_t^{\vartheta}\,\mathrm{d}z\right| \leq \frac{|\Omega|}{2\pi}\max_{1\leq i\leq n}\max_{z\in\Omega}\left|\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_t^{\vartheta}\right| \\ |\boldsymbol{e}_i^T \boldsymbol{w}_t^{\vartheta}| &= \left|\frac{-1}{2\pi i}\oint_{\Omega} \boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_t^{\vartheta}\,\mathrm{d}z\right| \leq \frac{|\Omega|}{2\pi}\max_{1\leq i\leq n}\max_{z\in\Omega}\left|\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_t^{\vartheta}\right| \end{aligned} \tag{3.74}$$

The result then immediately follows after applying Proposition 5. □

Before showing the proof of Proposition 5, we state a *Bernstein-type concentration result* for sampling without replacement. This result says the randomness from mini-batch sampling does not deviate too much from the "expectation". It will be used to show Proposition 5.

Proposition 6. *Let $\mathcal{X} = (x_1, \cdots, x_n)$ be a finite population of n points and $X_1, \ldots, X_\beta$ be a random sample drawn without replacement from $\mathcal{X}$. Let*

$$a = \min_{1\leq i\leq n} x_i \quad \textit{and} \quad b = \max_{1\leq i\leq n} x_i.$$

Also let

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \textit{and} \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

be the mean and variance of $\mathcal{X}$, respectively. Then for all $\epsilon > 0$,

$$\Pr(\frac{1}{\beta}\sum_{i=1}^{\beta} X_i - \mu \geq \epsilon) \leq \exp\left(-\frac{\beta\epsilon^2}{2\sigma^2 + (2/3)(b-a)\epsilon}\right).$$

We now prove the general bound on the entries of $\boldsymbol{w}_t^{\vartheta}$.

Proof of Proposition 5. For any fixed $\alpha > \alpha_0 + \theta$, define an increasing sequence $\alpha(t)$ where $t = 0, 1, \ldots, T$ such that $\alpha_0 + \theta < \alpha(0) < \alpha(1) < \ldots < \alpha(T) \leq \alpha$. We will show for any $0 \leq t \leq T$

$$\begin{aligned} \max_{z\in\Omega}\max_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_t^{\vartheta}| &\leq C(|\Omega|, \|\boldsymbol{A}^T\boldsymbol{A}\|, t)n^{\alpha(t)-1/2} \\ \text{and} \quad \max_{z\in\Omega}\max_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_t^{\vartheta}| &\leq C(|\Omega|, \|\boldsymbol{A}^T\boldsymbol{A}\|, t)n^{\alpha(t)-1/2}, \end{aligned} \tag{3.75}$$

where C is a constant depending on $|\Omega|$, $\|\boldsymbol{A}^T\boldsymbol{A}\|$, and t but it is independent of n and d.

For $t = 0$, a simple computations yield

$$\begin{aligned} \max_{z\in\Omega}\max_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_0^{\vartheta}| &= \max_{z\in\Omega}\max_{1\leq i\leq n} \left|\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)(\boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{b})\right| \\ &\leq \max_{z\in\Omega}\max_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}^T\boldsymbol{A})\boldsymbol{A}\boldsymbol{x}_0| + \max_{z\in\Omega}\max_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{b}|. \end{aligned}$$

Therefore, for the $\boldsymbol{w}_0^{\vartheta}$ term, the inequality (3.75) follows after applying Assumption 4 and Assumption 5. By setting $\boldsymbol{b} = 0$ in the inequality above, we get that (3.75) holds for $\boldsymbol{A}\boldsymbol{x}_0^{\vartheta}$.

For $1 \leq t \leq \vartheta$, we prove by induction. We already showed that

$$\begin{aligned} \max_{z\in\Omega}\max_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_0^{\vartheta}| &\leq C(|\Omega|, \|\boldsymbol{A}^T\boldsymbol{A}\|, 0)(n^{\alpha(0)-1/2}), \\ \text{and} \quad \max_{z\in\Omega}\max_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_0^{\vartheta}| &\leq C(|\Omega|, \|\boldsymbol{A}^T\boldsymbol{A}\|, 0)(n^{\alpha(0)-1/2}). \end{aligned} \tag{3.76}$$

where C is some function depending on $|\Omega|$ and $\|\boldsymbol{A}^T\boldsymbol{A}\|$. From (3.2), the mini-batch gradient method update gives, for $t \geq 1$,

$$\begin{aligned}
\boldsymbol{w}_t^{\vartheta} &= \boldsymbol{w}_0^{\vartheta} + \sum_{k=0}^{t-1} c_{(t-1),k} \boldsymbol{A}\boldsymbol{A}^T \boldsymbol{P}_{k+1} \boldsymbol{w}_k^{\vartheta} + \delta\zeta \sum_{k=0}^{t-1} c_{(t-1),k} \boldsymbol{A}\boldsymbol{x}_k^{\vartheta} \\
\text{and} \quad \boldsymbol{A}\boldsymbol{x}_t^{\vartheta} &= \boldsymbol{A}\boldsymbol{x}_0^{\vartheta} + \sum_{k=0}^{t-1} c_{(t-1),k} \boldsymbol{A}\boldsymbol{A}^T \boldsymbol{P}_{k+1} \boldsymbol{w}_k^{\vartheta} + \delta\zeta \sum_{k=0}^{t-1} c_{(t-1),k} \boldsymbol{A}\boldsymbol{x}_k^{\vartheta},
\end{aligned}$$

so by multiplying $\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)$ on both sides, we get

$$\begin{aligned}
\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T) \boldsymbol{w}_t^{\vartheta} &= \boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T) \boldsymbol{w}_0^{\vartheta} + \sum_{k=0}^{t-1} c_{(t-1),k} \boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T) \boldsymbol{A}\boldsymbol{A}^T \boldsymbol{P}_{k+1} \boldsymbol{w}_k^{\vartheta} \\
&\quad + \delta\zeta \sum_{k=0}^{t-1} c_{(t-1),k} \boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T) \boldsymbol{A}\boldsymbol{x}_k^{\vartheta}, \\
\text{and} \quad \boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T) \boldsymbol{A}\boldsymbol{x}_t^{\vartheta} &= \boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T) \boldsymbol{A}\boldsymbol{x}_0^{\vartheta} + \sum_{k=0}^{t-1} c_{(t-1),k} \boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T) \boldsymbol{A}\boldsymbol{A}^T \boldsymbol{P}_{k+1} \boldsymbol{w}_k^{\vartheta} \\
&\quad + \delta\zeta \sum_{k=0}^{t-1} c_{(t-1),k} \boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T) \boldsymbol{A}\boldsymbol{x}_k^{\vartheta}.
\end{aligned} \tag{3.77}$$

Now assume the induction hypothesis, that is, for each $0 \leq t-1 < \vartheta$

$$\begin{aligned}
&\max_{z\in\Omega} \max_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T) \boldsymbol{w}_{t-1}^{\vartheta}| \leq C(|\Omega|, \|\boldsymbol{A}^T\boldsymbol{A}\|, \gamma, t-1) \cdot n^{\alpha(t-1)-1/2}, \\
\text{and} \quad &\max_{z\in\Omega} \max_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T) \boldsymbol{A}\boldsymbol{x}_{t-1}^{\vartheta}| \leq C(|\Omega|, \|\boldsymbol{A}^T\boldsymbol{A}\|, \gamma, t-1) \cdot n^{\alpha(t-1)-1/2},
\end{aligned} \tag{3.78}$$

with overwhelming probability. From (3.77), let $Y_{i,k} \overset{\text{def}}{=} \boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T) \boldsymbol{A}\boldsymbol{A}^T \boldsymbol{P}_{k+1} (\boldsymbol{A}\boldsymbol{x}_k^{\vartheta} - \boldsymbol{b})$. We will use Bernstein-type inequality. Note that $Y_{i,k} = \sum_{\ell\in B_k} X_{\ell,i,k}$ where $X_{\ell,i,k} \overset{\text{def}}{=} (\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T) \boldsymbol{A}\boldsymbol{A}^T)_\ell (\boldsymbol{A}\boldsymbol{x}_k^{\vartheta} - \boldsymbol{b})_\ell$. Observe,

$$\mu_k \overset{\text{def}}{=} \frac{1}{n} \sum_{\ell=1}^{n} X_{\ell,i,k} = \frac{1}{n} \boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T) \boldsymbol{A}\boldsymbol{A}^T \boldsymbol{w}_k^{\vartheta},$$

and

$$\sigma_k^2 \stackrel{\text{def}}{=} \frac{1}{n}\sum_{\ell=1}^{n}(\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{A}^T)_\ell^2(\boldsymbol{A}\boldsymbol{x}_k^\vartheta - \boldsymbol{b})_\ell^2 - \mu^2.$$

As for bounding μ, observe

$$\begin{aligned} R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{A}^T &= R(z;\boldsymbol{A}\boldsymbol{A}^T)(\boldsymbol{A}\boldsymbol{A}^T - z\mathbf{I} + z\mathbf{I}) \\ &= \mathbf{I} + zR(z;\boldsymbol{A}\boldsymbol{A}^T). \end{aligned} \tag{3.79}$$

By left multiplying $\boldsymbol{e}_i^T$ and right multiplying $\boldsymbol{w}_k^\vartheta$, we get that

$$\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_k^\vartheta = \boldsymbol{e}_i^T\boldsymbol{w}_k^\vartheta + z\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_k^\vartheta$$

so that by using the induction hypothesis, with $z \in \Omega$, we have

$$\mu_k \le \frac{1}{n}|\Omega|C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,k)n^{\alpha(k)-1/2} \le C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,t)n^{\alpha(k)-3/2}, \quad \text{w.o.p.}$$

For the constant, we abuse notation so that $C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,t) = |\Omega| \max\limits_{0\le k\le t-1} C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,k)$. We will use this sense for all constants below. As for σ^2, we have

$$\begin{aligned} \sigma_k^2 &\le \frac{1}{n}\max_{\ell\neq i}\{(\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{A}^T)_\ell^2\}\|\boldsymbol{A}\boldsymbol{x}_k^\vartheta - \boldsymbol{b}\|^2 + \frac{1}{n}(\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{A}^T)_i^2|(\boldsymbol{A}\boldsymbol{x}_k^\vartheta - \boldsymbol{b})_i|^2. \\ &= \frac{1}{n}\max_{\ell\neq i}\{(z\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_\ell)^2\}\|\boldsymbol{A}\boldsymbol{x}_k^\vartheta - \boldsymbol{b}\|^2 + \frac{1}{n}(1 + z\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_i)^2|(\boldsymbol{A}\boldsymbol{x}_k^\vartheta - \boldsymbol{b})_i|^2. \end{aligned} \tag{3.80}$$

The last equality follows from (3.79). By applying Assumption 4,

$$\begin{aligned} \max_{z\in\Omega}\max_{i\neq j}\left|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_j\right| &< n^{\alpha_0-1/2} \quad \text{and} \\ \max_{z\in\Omega}\max_{i}\left|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_i - \frac{1}{n}\text{tr}(\boldsymbol{A}\boldsymbol{A}^T)\right| &< n^{\alpha_0-1/2}, \end{aligned} \tag{3.81}$$

together with the induction hypothesis and the definition of the stopping time ϑ, we have

$$\sigma_k^2 \le C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,\gamma,t)\left[\frac{1}{n}(n^{(2\alpha_0-1)+2\theta}) + \frac{1}{n}(n^{2\alpha(k)-1})\right] = C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,\gamma,t)n^{2\alpha(k)-2}, \quad \text{w.o.p.}$$

Here we used that $\alpha_0+\theta < \alpha(k)$. Let $b_k \stackrel{\text{def}}{=} \max_{1\le\ell\le n} X_{l,i,k} = \max_{1\le\ell\le n}\{(e_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{A}^T)_\ell(\boldsymbol{A}\boldsymbol{x}_k^\vartheta - \boldsymbol{b})_\ell\}$. Again using (3.79), one equates b_k, as follows,

$$b_k = \max_{1\le\ell\le n}\{\big(\boldsymbol{e}_i^T(\mathbf{I}+zR(z;\boldsymbol{A}\boldsymbol{A}^T))\boldsymbol{e}_\ell\big)\boldsymbol{e}_\ell^T\boldsymbol{w}_k^\vartheta\}. \tag{3.82}$$

We will bound $\boldsymbol{e}_\ell^T\boldsymbol{w}_k^\vartheta$ using the induction hypothesis (3.78). For the other term, we look at two cases $\ell = i$ and $\ell \neq i$ and use Assumption 4:

$$\begin{aligned}
\ell \neq i \qquad & |\boldsymbol{e}_i^T(\mathbf{I}+zR(z;\boldsymbol{A}\boldsymbol{A}^T))\boldsymbol{e}_\ell| \le |\Omega|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_\ell \le |\Omega| n^{\alpha_0-1/2}\\
\ell = i \qquad & |\boldsymbol{e}_i^T(\mathbf{I}+zR(z;\boldsymbol{A}\boldsymbol{A}^T))\boldsymbol{e}_i| \le 1+|\Omega||\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_i|\\
& \qquad\qquad 1+|\Omega|\big(|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_i - \tfrac{1}{n}\mathrm{tr}(\boldsymbol{A}\boldsymbol{A}^T)| + \tfrac{1}{n}\mathrm{tr}(\boldsymbol{A}\boldsymbol{A}^T)\big)\\
& \qquad\qquad \le 1+|\Omega|(n^{\alpha_0-1/2}+\|\boldsymbol{A}^T\boldsymbol{A}\|).
\end{aligned}$$

Using this with the induction hypothesis (3.78) on $\boldsymbol{e}_\ell^T\boldsymbol{w}_k^\vartheta$, we get a bound on b_k:

$$\begin{aligned}
b_k &= \max_{1\le\ell\le n}\{\big(\boldsymbol{e}_i^T(\mathbf{I}+zR(z;\boldsymbol{A}\boldsymbol{A}^T))\boldsymbol{e}_\ell\big)\boldsymbol{e}_\ell^T\boldsymbol{w}_k^\vartheta\}\\
&\le \big[|\Omega|n^{\alpha_0-1/2}+1+|\Omega|(n^{\alpha_0-1/2}+\|\boldsymbol{A}^T\boldsymbol{A}\|)\big]C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,k)n^{\alpha(k)-1/2}\\
&\le C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,\gamma,t)n^{\alpha(k)-1/2}, \quad \text{w.o.p.}
\end{aligned} \tag{3.83}$$

Note in the last inequality we used that $\alpha_0 < 1/2$ to conclude that $\alpha_0+\alpha(k)-1 < \alpha(k)-1/2$.

Recall Proposition 6, that is,

$$\mathbb{P}\left(\frac{1}{\beta}\sum_{\ell\in B_k} X_{\ell,i,k} - \mu_k \ge \epsilon\right) \le \exp\left(-\frac{\beta\epsilon^2}{2\sigma_k^2+(2/3)(b_k-a)\epsilon}\right). \tag{3.84}$$

Let $\epsilon = n^{-3/2+\alpha(t)}$ in the above. Therefore after noting that β/n is the constant ζ, we deduce that the expression on the right-hand side inside the exponential is

$$-\frac{\beta\epsilon^2}{2\sigma^2+(2/3)(b-a)\epsilon} \lesssim \frac{-n\cdot n^{2(-3/2+\alpha(t))}}{n^{2\alpha(k)-2}+n^{\alpha(k)+\alpha(t)-2}} \lesssim -n^{\alpha(t)-\alpha(k)}. \tag{3.85}$$

Here we use $a \lesssim b$ to mean that there is a constant $C > 0$ depending on $t, |\Omega|, \|\boldsymbol{A}^T\boldsymbol{A}\|$, but independent of n, such that $a \leq Cb$. Since $\alpha(t) > \alpha(k)$ for all $k < t$, the probability in (3.84) goes down faster than any polynomial in n. Therefore, from (3.77) and the probability (3.84) with (3.85), we have that

$$\begin{aligned}
&\max_{z\in\Omega}\max_{1\leq i\leq n}\left|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_t^\vartheta\right| \\
&\qquad \leq \max_{z\in\Omega}\max_{1\leq i\leq n}|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_0^\vartheta| + \beta\sum_{k=0}^{t-1} c_{(t-1),k}[\mu_k+\epsilon] + \delta\zeta\sum_{k=0}^{t-1} c_{(t-1),k}\boldsymbol{A}\boldsymbol{x}_k^\vartheta \\
&\qquad \leq C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,0)n^{\alpha(0)-1/2} + C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,t)\sum_{k=0}^{t-1} c_{(t-1),k}(n^{\alpha(k)-1/2}+n^{\alpha(t)-1/2}) \\
&\qquad\quad + \delta\zeta\sum_{k=0}^{t-1} c_{(t-1),k}C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,k)n^{\alpha(k)-1/2} \\
&\qquad \leq C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,t)n^{\alpha(t)-1/2},
\end{aligned} \tag{3.86}$$

with overwhelming probability. Here we used that $\alpha(k) < \alpha(t)$ for all $k < t$. Similarly, from (3.77) and the probability (3.84) with (3.85), one deduces, with overwhelming probability,

$$\begin{aligned}
&\max_{z\in\Omega}\max_{1\leq i\leq n}\left|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_t^\vartheta\right| \\
&\qquad \leq \max_{z\in\Omega}\max_{1\leq i\leq n}|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_0^\vartheta| + \beta\sum_{k=0}^{t-1} c_{(t-1),k}[\mu_k+\epsilon] + \delta\zeta\sum_{k=0}^{t-1} c_{(t-1),k}\boldsymbol{A}\boldsymbol{x}_k^\vartheta \\
&\qquad \leq C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,0)n^{\alpha(0)-1/2} + C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,t)\sum_{k=0}^{t-1} c_{(t-1),k}(n^{\alpha(k)-1/2}+n^{\alpha(t)-1/2}) \\
&\qquad\quad + \delta\zeta\sum_{k=0}^{t-1} c_{(t-1),k}C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,k)n^{\alpha(k)-1/2} \\
&\qquad \leq C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,t)n^{\alpha(t)-1/2}.
\end{aligned} \tag{3.87}$$

The induction step is now shown and the result (3.75) follows for $0 \leq t \leq \vartheta$ with $t \leq T$.

For $T \geq t > \vartheta$, we have $\boldsymbol{w}_t^\vartheta = \boldsymbol{w}_{t+1}^\vartheta = \boldsymbol{w}_\vartheta^\vartheta$. We already showed that the result (3.75) holds for $(\boldsymbol{w}_0^\vartheta, \boldsymbol{A}\boldsymbol{x}_0^\vartheta)$ and $(\boldsymbol{w}_t^\vartheta, \boldsymbol{A}\boldsymbol{x}_t^\vartheta)$ where $t \leq \vartheta$ and in particular when $t = \vartheta$. Thus,

we immediately get that the result (3.75) holds for $T \geq t > \vartheta$. From (3.75), the desired proposition follows after noting that $\alpha(0) < \alpha(1) < \ldots < \alpha(T) \leq \alpha$ and defining $C(|\Omega|, \|\boldsymbol{A}^T\boldsymbol{A}\|, T) = \max_{0 \leq t \leq T} C(|\Omega|, \|\boldsymbol{A}^T\boldsymbol{A}\|, t)$. □

3.7 Martingale error terms are small

We begin this section by introducing some definitions, propositions, and lemmas that are required to bound the martingale error terms. The martingale errors $\mathcal{E}_t^{\nabla}$, $\mathcal{E}_t^{\nabla^2\text{-Diag}}$, and $\mathcal{E}_t^{\nabla^2\text{-Off}}$, (3.39), (3.41), and (3.42), respectively, arise from randomly sampling a mini-batch at every iteration. The following Bernstein-type concentration result for sampling without replacement tells us that this randomness does not deviate too much from the "expectation".

Proposition 7 (Proposition 1.4, [9]). *Let $\mathcal{X} = (x_1, \cdots, x_n)$ be a finite population of n points and $X_1, \cdots, X_\beta$ be a random sample drawn without replacement from $\mathcal{X}$. Let*

$$a = \min_{1 \leq i \leq n} x_i \text{ and } b = \max_{1 \leq i \leq n} x_i.$$

Also let

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

be the mean and variance of $\mathcal{X}$, respectively. Then for all $\epsilon > 0$,

$$\mathbb{P}\left(\frac{1}{\beta}\sum_{i=1}^{\beta} X_i - \mu \geq \epsilon\right) \leq \exp\left(-\frac{\beta\epsilon^2}{2\sigma^2 + (2/3)(b-a)\epsilon}\right).$$

Recall the vector $\boldsymbol{w}_t = \boldsymbol{A}\boldsymbol{x}_t - \boldsymbol{b}$. We define the quantity ϑ to be

$$\vartheta \stackrel{\text{def}}{=} \inf\{t \geq 0 : ||\boldsymbol{w}_t||_2 > n^\theta\} \tag{3.88}$$

for some $\theta > 0$ to be determined. We define the stopped process of the iterates and the residual vector to be

$$\boldsymbol{x}_t^{\vartheta} \stackrel{\text{def}}{=} \boldsymbol{x}_{t\wedge\vartheta} \quad \text{and} \quad \boldsymbol{w}_t^{\vartheta} \stackrel{\text{def}}{=} \boldsymbol{w}_{t\wedge\vartheta} \quad t \geq 0.$$

Lemma 5 (Key Lemma)**.** *Fix $T > 0$ and suppose Assumption 6 holds. Let $p_k : \mathbb{C} \to \mathbb{C}$ be a k-degree polynomial with $k \leq T$ and coefficients which are independent of n and d. Then for some C depending on T and $|\Omega|$, the following holds with overwhelming probability*

$$\max_{0\leq k\leq T} \max_{1\leq \ell\leq n} \left| \boldsymbol{e}_\ell^T \boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T\boldsymbol{e}_\ell - \tfrac{1}{n} tr\left(\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T\right) \right| \leq n^{\alpha_0 - 1/2} C(T, |\Omega|) \|\mathcal{R}\|_{H^2}$$
$$\textit{where} \quad \boldsymbol{W} \stackrel{\textit{def}}{=} p_k(\boldsymbol{A}^T\boldsymbol{A}) \cdot (\nabla^2 \mathcal{R}) \cdot p_k(\boldsymbol{A}^T\boldsymbol{A}).$$

Proof. Define the matrix $\boldsymbol{K}(z, y) \stackrel{\text{def}}{=} \boldsymbol{A}R(z; \boldsymbol{A}^T\boldsymbol{A})(\nabla^2\mathcal{R})R(y; \boldsymbol{A}^T\boldsymbol{A})\boldsymbol{A}^T$. Let Ω by the contour and $\alpha_0 \in (0, \frac{1}{2})$ in Assumption 6. By Cauchy's integral formula, we can write

$$\begin{aligned} \boldsymbol{e}_\ell^T \boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T\boldsymbol{e}_\ell &= \frac{-1}{(2\pi i)^2} \oint_\Omega \boldsymbol{e}_\ell^T p_k(z) p_k(y) \boldsymbol{K}(z, y) \boldsymbol{e}_\ell \, \mathrm{d}z \, \mathrm{d}y \\ \tfrac{1}{n}\mathrm{tr}(\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T) &= \frac{-1}{(2\pi i)^2} \oint_\Omega p_k(z) p_k(y) \tfrac{1}{n}\mathrm{tr}\big(\boldsymbol{K}(z, y)\big) \, \mathrm{d}z \, \mathrm{d}y. \end{aligned} \tag{3.89}$$

Using Assumption 6, the following holds with overwhelming probability

$$\begin{aligned} |\boldsymbol{e}_\ell^T \boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T\boldsymbol{e}_\ell - \tfrac{1}{n}\mathrm{tr}(\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T)| &\leq \left| \frac{-1}{(2\pi i)^2} \oint_\Omega p_k(z) p_k(y) \big[\boldsymbol{e}_\ell^T \boldsymbol{K}(z, y) \boldsymbol{e}_\ell - \tfrac{1}{n}\mathrm{tr}(\boldsymbol{K}(z, y))\big] \, \mathrm{d}z \, \mathrm{d}y \right| \\ &\leq n^{\alpha_0 - 1/2} \frac{\|\mathcal{R}\|_{H^2}}{4\pi^2} \oint_\Omega |p_k(z) p_k(y)| |\, \mathrm{d}z| \, |\mathrm{d}y| \\ &\leq n^{\alpha_0 - 1/2} \frac{|\Omega|^2}{4\pi^2} \max_{z\in\Omega} |p_k(z)|^2 \|\mathcal{R}\|_{H^2}, \end{aligned} \tag{3.90}$$

Because the contour Ω is bounded, $\max_{0\leq k\leq T} \max_{z\in\Omega} |p_k(z)|$ is bounded independent of n and d, but it is dependent on T. Thus we get

$$\begin{aligned} \max_{0\leq k\leq T} \max_{1\leq \ell\leq n} |\boldsymbol{e}_\ell^T \boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T\boldsymbol{e}_\ell - \tfrac{1}{n}\mathrm{tr}(\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T)| &\leq \max_{0\leq k\leq T} n^{\alpha_0 - 1/2} \frac{|\Omega|^2}{4\pi^2} \max_{z\in\Omega} |p_k(z)|^2 \|\mathcal{R}\|_{H^2} \\ &\leq n^{\alpha_0 - 1/2} C(T, |\Omega|) \|\mathcal{R}\|_{H^2}, \end{aligned} \tag{3.91}$$

where the constant C depends on T and $|\Omega|$. The result immediately follows. $\square$

In the subsections below we provide the proofs in bounding the errors $\mathcal{E}_t^{\nabla}$, $\mathcal{E}_t^{\nabla^2\text{-Diag}}$, and $\mathcal{E}_t^{\nabla^2\text{-Off}}$, (3.39), (3.41), and (3.42). We recall that for any mini-batch gradient-based method there exists a corresponding k-degree noise polynomial, $N_{t,k}(\boldsymbol{H}) : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ where $\boldsymbol{H} = \boldsymbol{A}\boldsymbol{A}^T + \delta\mathbf{I}$, as defined in Proposition 4. Since everything is evaluated at $\boldsymbol{H}$, we will often suppress the $\boldsymbol{H}$, that is, $\boldsymbol{N}_{t,k} = \boldsymbol{N}_{t,k}(\boldsymbol{H})$.

In this section, we will show that the error terms in $\mathcal{E}_t^{\nabla}, \mathcal{E}_t^{\nabla^2\text{-Diag}}$, and $\mathcal{E}_t^{\nabla^2\text{-Off}}$ are small. For reference, we recall these terms here:

$$\mathcal{E}_t^{\nabla} \stackrel{\text{def}}{=} \nabla\mathcal{R}(\boldsymbol{y}_t)^T \left(\sum_{k=1}^{t} \boldsymbol{N}_{t,k}\mathring{\boldsymbol{M}}_k\right)$$

$$\mathcal{E}_t^{\nabla^2\text{-Diag}} \stackrel{\text{def}}{=} \frac{1}{2}\sum_{k=1}^{t} \mathring{\boldsymbol{M}}_k^T \boldsymbol{N}_{t,k}(\nabla^2\mathcal{R})\boldsymbol{N}_{t,k}\mathring{\boldsymbol{M}}_k - \frac{1}{2}\sum_{k=1}^{t}\sum_{\ell=1}^{n} \frac{(\zeta-\zeta^2)}{n}\text{tr}\left(\boldsymbol{A}\tilde{\boldsymbol{N}}_{t,k}(\nabla^2\mathcal{R};\boldsymbol{H})\boldsymbol{A}^T\right)\boldsymbol{w}_{k-1,\ell}^2$$

$$\mathcal{E}_t^{\nabla^2\text{-Off}} \stackrel{\text{def}}{=} \gamma^2 \sum_{k_1<k_2} \mathring{\boldsymbol{M}}_{k_1}^T \boldsymbol{N}_{t,k_1}(\nabla^2\mathcal{R})\boldsymbol{N}_{t,k_2}\mathring{\boldsymbol{M}}_{k_2}$$

where as before, we write $\widetilde{\boldsymbol{N}}_{t,k}(\boldsymbol{M};\boldsymbol{H}) \stackrel{\text{def}}{=} N_{t,k}(\boldsymbol{H})\boldsymbol{M}N_{t,k}(\boldsymbol{H})$. We are interested in making these errors small when conditioned on the stopping time ϑ.

3.7.1 Error from the gradient term $\mathcal{E}_t^{\nabla}$

In this section, we will show that $\mathcal{E}_t^{\nabla}$ is small.

Proposition 8. *For all $\alpha' > \alpha_0 + \theta$, the following holds with overwhelming probability*

$$\max_{0\le t\le T} |\mathcal{E}_{t\wedge\vartheta}^{\nabla}| \le n^{\alpha'-1/2}. \tag{3.92}$$

Proof of Proposition 8. Recall for $t > 0$, we have

$$\mathcal{E}_{t\wedge\vartheta}^{\nabla} = \sum_{k=1}^{t\wedge\vartheta} (\nabla\mathcal{R}(\boldsymbol{y}_t^{\vartheta})^T \boldsymbol{N}_{(t\wedge\vartheta,k)} \left[\zeta\boldsymbol{A}^T\left(\boldsymbol{A}\boldsymbol{x}_{k-1}^{\vartheta} - b\right) - \boldsymbol{A}^T\boldsymbol{P}_k\left(\boldsymbol{A}\boldsymbol{x}_{k-1}^{\vartheta} - \boldsymbol{b}\right)\right] \tag{3.93}$$

We make use of Proposition 21, (proof is similar to Proposition 10 in [32]) by defining the following quantities:

$$X_I^{(t,k)} \stackrel{\text{def}}{=} -\nabla\mathcal{R}(\boldsymbol{y}_t^{\vartheta})^T \boldsymbol{N}_{(t\wedge\vartheta,k)} \boldsymbol{A}^T \boldsymbol{e}_I \boldsymbol{e}_I^T \left(\boldsymbol{A}\boldsymbol{x}_{k-1}^{\vartheta} - \boldsymbol{b}\right), \tag{3.94}$$

where $\boldsymbol{e}_I$ denotes the I-th elementary basis that is included in the random batch B at the k-th iteration. Moreover, we define

$$\mu_{(t,k)} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i\in[n]} \boldsymbol{X}_i^{(t,k)} = -\frac{1}{n}\sum_{i=1}^{n} \nabla\mathcal{R}(\boldsymbol{y}_t^{\vartheta})^T \boldsymbol{N}_{(t\wedge\vartheta,k)} \boldsymbol{A}^T \boldsymbol{e}_i \boldsymbol{e}_i^T \left(\boldsymbol{A}\boldsymbol{x}_{k-1}^{\vartheta} - \boldsymbol{b}\right) \tag{3.95}$$

and hence

$$\mathcal{E}_{t\wedge\vartheta}^{\nabla} = \sum_{k=1}^{t\wedge\vartheta} \left(\sum_{i\in B_k} X_i^{(t,k)} - \beta\mu_{(t,k)} \right) \tag{3.96}$$

Since $k \le t \le T$, we apply Lemma 4 with an α chosen so that $\alpha' > \alpha > \alpha_0 + \theta$ to conclude that

$$\max_{1\le j\le n} \left|\left(\boldsymbol{A}\boldsymbol{x}_{k-1}^{\vartheta} - \boldsymbol{b}\right)_j\right|^2 \le Cn^{2\alpha-1}, \quad \text{w.o.p.} \tag{3.97}$$

where the constant C depends on $\|\boldsymbol{A}^T\boldsymbol{A}\|, |\Omega|, T$, but it is independent of n and d. For each $k = 0, 1, 2 \ldots, t$, we can impose an upper bound on the empirical variance as follows:

$$\begin{aligned}
\sigma_{(t,k)}^2 &\stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \left(\boldsymbol{X}_i^{(t,k)}\right)^2 - \left(\mu_{(t,k)}\right)^2 \\
&\le \frac{1}{n}\sum_{i=1}^{n} \left(\boldsymbol{X}_i^{(t,k)}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(\nabla\mathcal{R}(\boldsymbol{y}_t^{\vartheta})^T \boldsymbol{N}_{(t\wedge\vartheta,k)} \boldsymbol{A}^T \boldsymbol{e}_i\right)^2 \left(\boldsymbol{e}_i^T \left(\boldsymbol{A}\boldsymbol{x}_{k-1}^{\vartheta} - \boldsymbol{b}\right)\right)^2 \\
&\le \frac{1}{n} \max_{1\le j\le n} \left|\left(\boldsymbol{A}\boldsymbol{x}_{k-1}^{\vartheta} - \boldsymbol{b}\right)_j\right|^2 \cdot ||\nabla\mathcal{R}(\boldsymbol{y}_t^{\vartheta})||^2 \cdot ||\boldsymbol{N}_{(t\wedge\vartheta,k)}||^2 \cdot ||\boldsymbol{A}||^2 \\
&\le C(t) n^{2\alpha-2} \quad \text{w.o.p.},
\end{aligned}$$

where the constant C is dependent on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta, T, |\Omega|$, and time t (here take $\max$ of the constants over $1 \le k \le t$), but it is independent of n or d. All constants

going forward will also have this property and we note that the constant will change from line to line. Similarly, using the same α as in (3.97), we can bound the following quantity:

$$\begin{aligned} b_{(t,k)} = \max_{1\le i\le n} \boldsymbol{X}_i^{(t,k)} &\le \max_{1\le i\le n} \left|\nabla\mathcal{R}(\boldsymbol{y}_t^{\vartheta})^T \boldsymbol{N}_{(t\wedge\vartheta,k)} \boldsymbol{A}^T \boldsymbol{e}_i \boldsymbol{e}_i^T \left(\boldsymbol{A}\boldsymbol{x}_{k-1}^{\vartheta} - \boldsymbol{b}\right)\right| \\ &\le \max_{1\le i\le n} |\left(A\boldsymbol{x}_{k-1}^{\vartheta} - \boldsymbol{b}\right)_i| \cdot ||\nabla\mathcal{R}(\boldsymbol{y}_t^{\vartheta})||_2 \cdot ||\boldsymbol{N}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T|| \\ &\le C(t) n^{\alpha-1/2} \quad \text{w.o.p.} \end{aligned} \tag{3.98}$$

Applying Proposition 21 gives

$$\begin{aligned} \Pr\left(\sum_{i=1}^{\beta} \boldsymbol{X}_i^{(t,k)} - \beta\mu_{(t,k)} \ge \tilde{\varepsilon}\right) &= \Pr\left(\frac{1}{\beta}\sum_{i=1}^{\beta} \boldsymbol{X}_i^{(t,k)} - \mu_{(t,k)} \ge \frac{\tilde{\varepsilon}}{\beta}\right) \\ &\le \exp\left(-\frac{\beta(\frac{\tilde{\varepsilon}}{\beta})^2}{2\sigma_{(t,k)}^2 + \frac{2}{3}(b_{(t,k)} - a)(\frac{\tilde{\varepsilon}}{\beta})}\right) \end{aligned} \tag{3.99}$$

We first note that $\beta/n = \zeta$. Set $\tilde{\varepsilon} = \varepsilon/T$ with $\varepsilon = n^{\alpha'-1/2}$ and $a = 0$. Using the upper bounds on $b_{(t,k)}$ and $\sigma_{(t,k)}^2$,

$$\begin{aligned} \frac{\beta(\frac{\tilde{\varepsilon}}{\beta})^2}{2\sigma_{(t,k)}^2 + \frac{2}{3}(b_{(t,k)} - a)(\frac{\tilde{\varepsilon}}{\beta})} &\ge C(t) \cdot \frac{T^2 n^{-1}\varepsilon^2}{n^{2\alpha-2} + Tn^{\alpha-1/2}n^{-1}\varepsilon} \ge C(t) \cdot \frac{T^2}{n^{2(\alpha-\alpha')} + Tn^{\alpha-\alpha'}} \\ &\ge C(T) \cdot \frac{T^2}{n^{2(\alpha-\alpha')} + Tn^{\alpha-\alpha'}} \underset{n\to\infty}{\to} \infty. \end{aligned} \tag{3.100}$$

Here again $C(t)$ is a positive constant independent of n and k. We note that we chose α so that $\alpha' > \alpha > \alpha_0 + \theta$. We can then lower bound $C(t)$ with $C(T)$ simply by letting $C(T) = \min_{1\le t\le T} C(t) > 0$. Hence

$$\Pr\left(\sum_{i=1}^{\beta} \boldsymbol{X}_i^{(t,k)} - \beta\mu_{(t,k)} \ge \tilde{\varepsilon}\right) \le \exp\left(-C(T)n^c\right) \quad \text{when } \tilde{\varepsilon} = n^{\alpha'-1/2}/T \tag{3.101}$$

for some $c > 0$. Note that the constants c and $C(T)$ are independent of t and k, only depends on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, |\Omega|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta$ and T. Similarly, by taking $-\boldsymbol{X}_i^{(t,k)}$ and using

the same bounds on $b_{(t,k)}$ and $\sigma^2_{(t,k)}$, we get that

$$\Pr\left(-\big[\sum_{i=1}^{\beta} \boldsymbol{X}_i^{(t,k)} - \beta\mu_{(t,k)}\big] \geq \tilde{\varepsilon}\right) \leq \exp\left(-C(T)n^c\right) \quad \text{when } \tilde{\varepsilon} = n^{\alpha'-1/2}/T. \tag{3.102}$$

Therefore, it follows that

$$\Pr\left(\left|\sum_{i=1}^{\beta} \boldsymbol{X}_i^{(t,k)} - \beta\mu_{(t,k)}\right| \geq \tilde{\varepsilon}\right) \leq 2\exp\left(-C(T)n^c\right) \quad \text{when } \tilde{\varepsilon} = n^{\alpha'-1/2}/T, \tag{3.103}$$

for some $c > 0$ and $C(T) > 0$ where the constants do not depend on t, but do depend on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, |\Omega|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta$ and T. We set $\varepsilon = n^{\alpha'-1/2}$. Applying the union bound twice and using (3.103), we get

$$\begin{aligned}
\Pr\left(\max_{0\leq t\leq T} |\mathcal{E}^{\nabla}_{t\wedge\vartheta}| \geq \varepsilon\right) &\leq \sum_{t=0}^{T} \Pr\left(|\mathcal{E}^{\nabla}_{t\wedge\vartheta}| \geq \varepsilon\right) \\
&\leq \sum_{t=0}^{T} \Pr\left(\sum_{k=1}^{t\wedge\vartheta} \left|\left(\sum_{i\in B_k} X_i^{(t,k)} - \beta\mu_{(t,k)}\right)\right| \geq \varepsilon\right) \\
&\leq \sum_{t=0}^{T}\sum_{k=1}^{t\wedge\vartheta} \Pr\left(\left|\sum_{i=1}^{\beta} X_i^{(t,k)} - \beta\mu_{(t,k)}\right| \geq \varepsilon/(t\wedge\vartheta)\right) \\
&\leq \sum_{t=0}^{T}\sum_{k=1}^{t\wedge\vartheta} \Pr\left(\left|\sum_{i=1}^{\beta} X_i^{(t,k)} - \beta\mu_{(t,k)}\right| \geq \varepsilon/T\right) \\
&\leq T^2 \cdot \exp(-C(T)n^c)
\end{aligned}$$

for some $c, C(T) > 0$. In the penultimate inequality, we used that $t \wedge \vartheta \leq T$. For the last inequality, we note that $\varepsilon/T = \tilde{\varepsilon}$ in (3.103) that the constants, $c, C(T) > 0$ in (3.103) hold for all $1 \leq t \leq T$. The result immediately follows. □

3.7.2 Error from the on diagonal term $\mathcal{E}_t^{\nabla^2\text{-Diag}}$

In this subsection, we define $\widetilde{\boldsymbol{N}}_{t,k} \stackrel{\text{def}}{=} \boldsymbol{N}_{t,k}(\boldsymbol{H})(\nabla^2\mathcal{R})\boldsymbol{N}_{t,k}(\boldsymbol{H})$ and we express the ij-th entry of $\widetilde{\boldsymbol{N}}_{t,k}$ as $\widetilde{N}_{ij}^{(t,k)}$. Recall $\boldsymbol{w}_k \stackrel{\text{def}}{=} \boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}$. In the following proposition, we decompose

the error term from the diagonal components of the Hessian, $\mathcal{E}_t^{\nabla^2\text{-Diag}}$, into six terms. We will show each of these terms are small. We will show the following proposition:

Proposition 9. *Let $\alpha_0 \in (0, 1/4)$ as specified in Assumption 5. For any α, α', and θ satisfying $0 < \theta < \alpha' - \alpha$ and $\alpha_0 + \theta < \alpha < \alpha'$ we have*

$$\max_{0\le t\le T} |\mathcal{E}_{t\wedge\vartheta}^{\nabla^2-\text{Diag}}| \le C(T) n^{2\alpha'-1/2} \quad \textit{w.o.p.}$$

for some constant $C(T)$ that is independent of n and d and only depends on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|$, $\|\boldsymbol{x}_0\|$, $\|\mathcal{R}\|_{H^2}$, γ, ζ, T, and $|\Omega|$.

Before proving Proposition 9, we provide a few results.

Lemma 6 (Lemma B.1 [42])**.** *Suppose that $\boldsymbol{u}$ and $\boldsymbol{v}$ are fixed vectors in $\mathbb{R}^n$. Then*

$$\mathbb{E}\left[\Big(\sum_{i\in B} u_i v_i\Big)^2\right] = \frac{\beta}{n}\frac{\beta-1}{n-1}(\boldsymbol{u}^T\boldsymbol{v})^2 + \left(\frac{\beta}{n} - \frac{\beta}{n}\frac{\beta-1}{n-1}\right)\sum_{i=1}^n (u_i v_i)^2.$$

We now decompose the error $\mathcal{E}_t^{\nabla^2-\text{Diag}}$ into six other error terms.

Lemma 7. *We can decompose the on-diagonal error term $\mathcal{E}_t^{\nabla^2\text{-Diag}}$ as follows*

$$\begin{aligned}\mathcal{E}_t^{\nabla^2\text{-Diag}} &\overset{\text{def}}{=} \frac{1}{2}\sum_{k=1}^t \mathring{\boldsymbol{M}}_k^T \widetilde{\boldsymbol{N}}_{t,k} \mathring{\boldsymbol{M}}_k - \frac{\gamma^2}{2}\sum_{k=1}^t\sum_{\ell=1}^n \frac{(\zeta-\zeta^2)}{n} tr\left(\boldsymbol{A}\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T\right)\boldsymbol{w}_{k-1,\ell}^2 \\ &= \mathcal{E}_t^{(\nabla^2:\text{KL})} + \mathcal{E}_t^{(\nabla^2:\text{Z.1})} + \mathcal{E}_t^{(\nabla^2:\text{Z.2})} + \mathcal{E}_t^{(\nabla^2:\text{B.1})} + \mathcal{E}_t^{(\nabla^2:\text{B.2})} + \mathcal{E}_t^{(\nabla^2:\text{HW})}\end{aligned} \tag{3.104}$$

where the six error terms are

$$\mathcal{E}_t^{(\nabla^2:\,\mathrm{KL})} = \frac{1}{2}(\zeta-\zeta^2)\sum_{k=1}^{t}\sum_{ij}\sum_{\ell=1}^{n}\left(A_{\ell j}A_{\ell i}w_{k-1,\ell}^2\widetilde{N}_{ij}^{(t,k)}\right) - \frac{1}{2}\sum_{k=1}^{t}\sum_{\ell=1}^{n}\frac{(\zeta-\zeta^2)}{n}tr\left(\boldsymbol{A}\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T\right)w_{k-1,\ell}^2$$

$$\mathcal{E}_t^{(\nabla^2:\mathrm{Z.1})} = \frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\left(\frac{\beta(\beta-1)}{n(n-1)}-\zeta^2\right)\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\widetilde{N}_{ij}^{(t,k)}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}$$

$$\mathcal{E}_t^{(\nabla^2:\mathrm{Z.2})} = \frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\sum_{\ell=1}^{n}\left(\zeta^2-\frac{\beta(\beta-1)}{n(n-1)}\right)A_{\ell j}A_{\ell i}\widetilde{N}_{ij}^{(t,k)}w_{k-1,\ell}^2$$

$$\mathcal{E}_t^{(\nabla^2:\mathrm{B.1})} = \frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\left[\zeta^2\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\widetilde{N}_{ij}^{(t,k)}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}-\zeta\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\widetilde{N}_{ij}^{(t,k)}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right]$$

$$\mathcal{E}_t^{(\nabla^2:\mathrm{B.2})} = \frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\left[\zeta^2\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\widetilde{N}_{ij}^{(t,k)}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}-\zeta\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\widetilde{N}_{ij}^{(t,k)}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right]$$

$$\mathcal{E}_t^{(\nabla^2:\mathrm{HW})} = \frac{1}{2}\sum_{k=1}^{t}\sum_{ij}\widetilde{N}_{ij}^{(t,k)}\left[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}-\mathbb{E}[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_t\boldsymbol{w}_{k-1}|\mathcal{F}_{k-1}]\right]. \tag{3.105}$$

We indexed the error terms by the acronyms inspired by the results applied to bound them: Key Lemma (KL), Zeta (Z), Bardanet (B), and Hanson-Wright (HW).

Proof. Observe that

$$\frac{1}{2}\sum_{k=1}^{t}\mathring{\boldsymbol{M}}_k^T\widetilde{\boldsymbol{N}}_{t,k}\mathring{\boldsymbol{M}}_k = \frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\mathring{\boldsymbol{M}}_{k,i}\widetilde{N}_{ij}^{(t,k)}\mathring{\boldsymbol{M}}_{k,j} \tag{3.106}$$

and that the product of the entries of the Martingale increments defined in (3.5) gives us

$$\begin{aligned}
\mathring{\boldsymbol{M}}_{k,i}\mathring{\boldsymbol{M}}_{k,j} &= \left[\zeta \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} - \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1}\right]\left[\zeta \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} - \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1}\right] \\
&= \zeta^2 \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} - \zeta \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \\
&\quad - \zeta \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} + \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \\
&= \zeta^2 \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} - 2\zeta^2 \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} \\
&\quad + \left(\zeta^2 \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} - \zeta \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{w}_{k-1}\right) \\
&\quad + \left(\zeta^2 \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} - \zeta \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1}\right) \\
&\quad + \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \\
&= -\zeta^2 \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} + \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \\
&+ \mathcal{E}_k^{(\nabla^2,1)}(i,j) + \mathcal{E}_k^{(\nabla^2,2)}(i,j) + \boldsymbol{e}_i^T
\end{aligned} \tag{3.107}$$

where we introduce two error terms

$$\begin{aligned}
\mathcal{E}_k^{(\nabla^2,1)}(i,j) &\stackrel{\text{def}}{=} \zeta^2 \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} - \zeta \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} \\
\mathcal{E}_k^{(\nabla^2,2)}(i,j) &\stackrel{\text{def}}{=} \zeta^2 \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} - \zeta \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1}.
\end{aligned} \tag{3.108}$$

Observe the conditional expectation of $\zeta \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1}$ is $\zeta^2 \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{w}_{k-1}$ (same for the other term). We will show, in fact, that $\zeta \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1}$ will concentrate around its mean.

We now consider the term $\boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1}$ in (3.107) and we perform a Doob's decomposition on this term, that is,

$$\begin{aligned}
&\boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} = \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \\
&- \mathbb{E}\left[\boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} | \mathcal{F}_{k-1}\right] + \mathbb{E}\left[\boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} | \mathcal{F}_{k-1}\right].
\end{aligned} \tag{3.109}$$

The first term in the inequality will be small as $\boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1}$ will concentrate around its mean. It remains to simplify the conditional expectation term

$$\mathbb{E}\left[\boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} | \mathcal{F}_{k-1}\right]$$

which we do so below. Noting that $\boldsymbol{P}_k = \sum_{m \in B_{k-1}} \boldsymbol{e}_m \boldsymbol{e}_m^T$,

$$\mathbb{E}\left[\boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} | \mathcal{F}_{k-1}\right] = \mathbb{E}\left[\left(\sum_{m \in B_k} A_{mi} w_{k-1,m}\right)\left(\sum_{\ell \in B_k} A_{\ell j} w_{k-1,\ell}\right) | \mathcal{F}_{k-1}\right].$$

We want to apply Lemma 6 to the above, but it is not in the form that Lemma 6. To get it into the correct form, we use polarization:

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{m \in B_{k-1}} A_{mi} w_{k-1,m}\right)\left(\sum_{\ell \in B_{k-1}} A_{\ell j} w_{k-1,\ell}\right) | \mathcal{F}_{k-1}\right] \\
&= \frac{1}{4}\mathbb{E}\left[\left(\sum_{m \in B_{k-1}} A_{mj} w_{k-1,m} + A_{mi} w_{k-1,m}\right)^2 | \mathcal{F}_{k-1}\right] \\
&- \frac{1}{4}\mathbb{E}\left[\left(\sum_{m \in B_{k-1}} A_{mj} w_{k-1,m} - A_{mi} w_{k-1,m}\right)^2 | \mathcal{F}_{k-1}\right].
\end{aligned} \tag{3.110}$$

Now we apply Lemma 6 to each term in (3.110) and thus, after simplifying,

$$
\begin{aligned}
\mathbb{E}\big[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}|\mathcal{F}_{k-1}\big] &= \frac{1}{4}\mathbb{E}\bigg[\bigg(\sum_{m\in B_{k-1}} A_{mj}w_{k-1,m}+A_{mi}w_{k-1,m}\bigg)^2|\mathcal{F}_{k-1}\bigg]\\
&\quad-\frac{1}{4}\mathbb{E}\bigg[\bigg(\sum_{m\in B_{k-1}} A_{mj}w_{k-1,m}-A_{mi}w_{k-1,m}\bigg)^2|\mathcal{F}_{k-1}\bigg]\\
&=\frac{1}{4}\left[\frac{\beta(\beta-1)}{n(n-1)}\left[(\boldsymbol{A}\boldsymbol{e}_j+\boldsymbol{A}\boldsymbol{e}_i)^T\boldsymbol{w}_{k-1}\right]^2+\left(\frac{\beta}{n}-\frac{\beta(\beta-1)}{n(n-1)}\right)\sum_{\ell=1}^n\left(A_{\ell j}+A_{\ell i}\right)^2w_{k-1,\ell}^2\right]\\
&-\frac{1}{4}\left[\frac{\beta(\beta-1)}{n(n-1)}\left[(\boldsymbol{A}\boldsymbol{e}_j-\boldsymbol{A}\boldsymbol{e}_i)^T\boldsymbol{w}_{k-1}\right]^2+\left(\frac{\beta}{n}-\frac{\beta(\beta-1)}{n(n-1)}\right)\sum_{\ell=1}^n\left(A_{\ell j}-A_{\ell i}\right)^2w_{k-1,\ell}^2\right]\\
&=\left(\frac{\beta(\beta-1)}{n(n-1)}\right)\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}+\left(\frac{\beta}{n}-\frac{\beta(\beta-1)}{n(n-1)}\right)\sum_{\ell=1}^n A_{\ell j}A_{\ell i}w_{k-1,\ell}^2\\
&=\zeta^2\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}+(\zeta-\zeta^2)\sum_{\ell=1}^n A_{\ell j}A_{\ell i}w_{k-1,\ell}^2\\
&+\left(\frac{\beta(\beta-1)}{n(n-1)}-\zeta^2\right)\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\\
&+\left(\left[\frac{\beta}{n}-\frac{\beta(\beta-1)}{n(n-1)}\right]-(\zeta-\zeta^2)\right)\sum_{\ell=1}^n A_{\ell j}A_{\ell i}w_{k-1,\ell}^2.
\end{aligned}
\tag{3.111}
$$

In the last equality, we added and subtracted terms corresponding to $\zeta\approx\frac{\beta-1}{n-1}$ when n is large. Plugging in (3.111) into (3.107) we obtain

$$
\begin{aligned}
\mathring{\boldsymbol{M}}_{k,i}\mathring{\boldsymbol{M}}_{k,j} &= (\zeta-\zeta^2)\sum_{\ell=1}^n A_{\ell j}A_{\ell i}w_{k-1,\ell}^2+\left(\frac{\beta(\beta-1)}{n(n-1)}-\zeta^2\right)\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\\
&\quad+\left(\left[\frac{\beta}{n}-\frac{\beta(\beta-1)}{n(n-1)}\right]-(\zeta-\zeta^2)\right)\sum_{\ell=1}^n A_{\ell j}A_{\ell i}w_{k-1,\ell}^2+\mathcal{E}_k^{(\nabla^2,1)}(i,j)+\mathcal{E}_k^{(\nabla^2,2)}(i,j)\\
&\quad+\left(\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}-\mathbb{E}\left[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}|\mathcal{F}_{k-1}\right]\right),
\end{aligned}
\tag{3.112}
$$

where $\mathcal{E}_k^{(\nabla^2,1)}(i,j)$ and $\mathcal{E}_k^{(\nabla^2,2)}(i,j)$ are defined in (3.108). Returning to (3.106) using the martingale increment computation (3.112),

$$
\begin{aligned}
&\frac{1}{2}\sum_{k=1}^{t}\mathring{\boldsymbol{M}}_k^T\widetilde{\boldsymbol{N}}_{t,k}\mathring{\boldsymbol{M}}_k - \frac{1}{2}\sum_{k=1}^{t}\sum_{\ell=1}^{n}\frac{(\zeta-\zeta^2)}{n}\text{tr}\left(\boldsymbol{A}\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T\right)w_{k-1,\ell}^2\\
&=\frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\mathring{\boldsymbol{M}}_{k,i}\widetilde{N}_{ij}^{(t,k)}\mathring{\boldsymbol{M}}_{k,j} - \frac{1}{2}\sum_{k=1}^{t}\sum_{\ell=1}^{n}\frac{(\zeta-\zeta^2)}{n}\text{tr}\left(\boldsymbol{A}\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T\right)w_{k-1,\ell}^2\\
&=\frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}(\zeta-\zeta^2)\sum_{\ell=1}^{n}A_{\ell j}A_{\ell i}w_{k-1,\ell}^2\widetilde{N}_{ij}^{(t,k)} - \frac{1}{2}\sum_{k=1}^{t}\sum_{\ell=1}^{n}\frac{(\zeta-\zeta^2)}{n}\text{tr}\left(\boldsymbol{A}\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T\right)w_{k-1,\ell}^2\\
&+\frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\left(\frac{\beta(\beta-1)}{n(n-1)}-\zeta^2\right)\widetilde{N}_{ij}^{(t,k)}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\\
&+\frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\left(\zeta^2-\frac{\beta(\beta-1)}{n(n-1)}\right)\widetilde{N}_{ij}^{(t,k)}\sum_{\ell=1}^{n}A_{\ell j}A_{\ell i}w_{k-1,\ell}^2\\
&+\frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\widetilde{N}_{ij}^{(t,k)}\left[\mathcal{E}_k^{(\nabla^2,1)}(i,j)+\mathcal{E}_k^{(\nabla^2,2)}(i,j)\right]\\
&+\frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\widetilde{N}_{ij}^{(t,k)}\left[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}-\mathbb{E}\left[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}|\mathcal{F}_{k-1}\right]\right].
\end{aligned}
\tag{3.113}
$$

The result follows by matching up terms in (3.113) with (3.105) and noting that $\mathcal{E}_t^{(\nabla^2:\text{B.1})} = \frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\widetilde{N}_{ij}^{(t,k)}\mathcal{E}_k^{(\nabla^2,1)}(i,j)$ and $\mathcal{E}_t^{(\nabla^2:\text{B.2})} = \frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\widetilde{N}_{ij}^{(t,k)}\mathcal{E}_k^{(\nabla^2,2)}(i,j)$ and the batch fraction, $\zeta = \frac{\beta}{n}$. □

We now show that each of the terms in (3.113) are small with overwhelming probability.

Lemma 8. *Fix $T > 0$. For any $\alpha > \alpha_0 + \theta$,*

$$
\max_{0\le t\le T}\left|\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:KL)}\right| \le C(T)n^{\alpha_0-1/2+2\theta} \tag{3.114}
$$

holds with overwhelming probability.

Proof. By applying Lemma 5 (Key Lemma) and Lemma 4 we get

$$\begin{aligned}
&\max_{0\le t\le T}\left|\mathcal{E}_t^{(\nabla^2:\mathrm{KL})}\right| \\
&= \frac{1}{2}(\zeta-\zeta^2)\max_{0\le t\le T}\left|\sum_{k=1}^{t\wedge\vartheta}\sum_{\ell=1}^{n}\left(\boldsymbol{e}_\ell^T\boldsymbol{A}\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\boldsymbol{e}_\ell w_{k-1,\ell}^2\right)-\sum_{k=1}^{t\wedge\vartheta}\sum_{\ell=1}^{n}\frac{1}{n}\mathrm{tr}\left(\boldsymbol{A}\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\right)w_{k-1,\ell}^2\right| \\
&\le \frac{1}{2}(\zeta-\zeta^2)\cdot T\cdot\max_{0\le t\le T}\max_{1\le\ell\le n}\left|\boldsymbol{e}_\ell^T\boldsymbol{A}\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\boldsymbol{e}_\ell-\frac{1}{n}\mathrm{tr}\left(\boldsymbol{A}\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\right)\right|\cdot\|\boldsymbol{w}_{k-1}^{\vartheta}\|^2 \\
&\le C(T)n^{\alpha_0-1/2}\cdot n^{2\theta}
\end{aligned} \tag{3.115}$$

where $C(T)$ depends on $T, |\Omega|$ and γ^2 but independent of n and d. □

Lemma 9. *Fix $T>0$. Let $\alpha'>\alpha_0+\theta$ and $\alpha>0$ such that $\alpha'>\alpha>\alpha_0+\theta$. The following holds with overwhelming probability*

$$\max_{0\le t\le T}|\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\mathrm{B.1})}|\le n^{\alpha'-1/2}, \tag{3.116}$$

provided $\theta<\alpha'-\alpha$.

Proof. Recall for $t>0$, we have

$$\begin{aligned}
\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\mathrm{B.1})} &= \frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\sum_{i,j}\left[\zeta^2\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\widetilde{N}_{ij}^{(t\wedge\vartheta,k)}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}-\zeta\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{t-1}\widetilde{N}_{ij}^{(t\wedge\vartheta,k)}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right] \\
&= \frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\left[\zeta^2\boldsymbol{w}_{k-1}^T\boldsymbol{A}\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\boldsymbol{w}_{k-1}-\zeta\boldsymbol{w}_{k-1}^T\boldsymbol{P}_k\boldsymbol{A}\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right].
\end{aligned}$$

Proceeding similarly to Proposition 8 for $\mathcal{E}_t^{\nabla}$, we define

$$X_I^{(t,k)}\overset{\text{def}}{=}-\frac{\zeta}{2}(\boldsymbol{w}_{k-1}^{\vartheta})^T\boldsymbol{e}_I\boldsymbol{e}_I^T\boldsymbol{A}\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\boldsymbol{w}_{k-1}^{\vartheta}\quad\text{and}\quad\mu^{(t,k)}\overset{\text{def}}{=}\frac{1}{n}\sum_{i=1}^{n}X_i^{(t,k)}, \tag{3.117}$$

and we observe that

$$\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\mathrm{B.1})}=\sum_{k=1}^{t\wedge\vartheta}\left(\sum_{i\in B_{k-1}}X_i^{(t,k)}-\beta\mu^{(t,k)}\right).$$

Before proceeding, we note that $\|\widetilde{\boldsymbol{N}}_{t,k}\|$ for $t \geq k$ is bounded by a constant depending on $\|\boldsymbol{A}^T\boldsymbol{A}\|$ and t, but independent of n and d. We apply Lemma 4 with α satisfying $\alpha' > \alpha > \alpha_0 + \theta$ and get that $\max_{1\leq j\leq n} |w_{k-1,j}| \leq n^{\alpha-1/2}$. Using the definition of ϑ, we have the bounds

$$\begin{aligned}
\sigma^2_{(t,k)} &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i^{(t,k)}\right)^2 - \left(\mu^{(t,k)}\right)^2 \\
&\leq \frac{\zeta^2}{4n}\sum_{i=1}^{n}\left(\boldsymbol{e}_i^T\boldsymbol{w}_{k-1}^{\vartheta}\right)^2\left(\boldsymbol{e}_i^T\boldsymbol{A}\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\boldsymbol{w}_{k-1}^{\vartheta}\right)^2 \\
&\leq \frac{\zeta^2}{4n}\max_{1\leq j\leq n}|w_{k-1,j}^{\vartheta}|^2\cdot||\boldsymbol{A}\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\boldsymbol{w}_{k-1}^{\vartheta}||^2 \\
&\leq \frac{\zeta^2}{4n}\max_{1\leq j\leq n}|w_{k-1,j}^{\vartheta}|^2||\boldsymbol{w}_{k-1}^{\vartheta}||^2||\boldsymbol{A}||^4\cdot||\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}||^2 \\
&\leq C(t)n^{2(\alpha+\theta-1)} \qquad \text{w.o.p,}
\end{aligned}$$

where the constant C is dependent on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta, T, |\Omega|$, and time t (here take $\max$ of the constants over $1 \leq k \leq t$), but it is independent of n or d. All constants going forward will also have this property and we note that the constant will change from line to line.

Similarly, applying Lemma 4 with α, we can bound the following quantity:

$$\begin{aligned}
b &= \max_{1\leq i\leq n} X_i^{(t,k)} \\
&\leq \max_{1\leq i\leq n}|\frac{\zeta}{2}(\boldsymbol{w}_{k-1}^{\vartheta})^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{A}\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\boldsymbol{w}_{k-1}^{\vartheta}| \\
&\leq \frac{\zeta}{2}\max_{1\leq i\leq n}|w_{k-1,i}^{\vartheta}|\cdot||\boldsymbol{e}_i^T\boldsymbol{A}\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}||\cdot||\boldsymbol{A}^T\boldsymbol{w}_{k-1}^{\vartheta}|| \qquad (3.118) \\
&\leq \frac{\zeta}{2}\left(\max_{1\leq i\leq n}|w_{k-1,i}|\right)\cdot||\boldsymbol{w}_{k-1}^{\vartheta}||\cdot\|\boldsymbol{A}\|^2\cdot\|\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}\| \\
&\leq C(t)n^{\alpha-\frac{1}{2}+\theta} \qquad \text{w.o.p.}
\end{aligned}$$

Applying Proposition 21 gives

$$\begin{aligned}\Pr\left(\sum_{i=1}^{\beta} X_i^{(t,k)} - \beta\mu^{(t,k)} \geq \tilde{\varepsilon}\right) &= \Pr\left(\frac{1}{\beta}\sum_{i=1}^{\beta} X_i^{(t,k)} - \mu^{(t,k)} \geq \frac{\tilde{\varepsilon}}{\beta}\right) \\ &\leq \exp\left(-\frac{\beta(\frac{\tilde{\varepsilon}}{\beta})^2}{2\sigma_{(t,k)}^2 + \frac{2}{3}(b-a)(\frac{\tilde{\varepsilon}}{\beta})}\right).\end{aligned} \tag{3.119}$$

We note that $\beta/n = \zeta$. Set $\tilde{\varepsilon} = \varepsilon/T$ with $\varepsilon = n^{\alpha'-1/2}$ where $\alpha' > \alpha$ and a = 0. Using the upper bounds on b and $\sigma_{(t,k)}^2$,

$$\frac{\beta(\frac{\tilde{\varepsilon}}{\beta})^2}{2\sigma_{(t,k)}^2 + \frac{2}{3}(b-a)(\frac{\tilde{\varepsilon}}{\beta})} \geq C(t)\cdot\frac{T^{-2}n^{-1}\varepsilon^2}{n^{2(\alpha+\theta-1)} + T^{-1}n^{\alpha-1/2+\theta}n^{-1}\varepsilon} \geq C(t)\cdot\frac{T^{-2}}{n^{2(\alpha-\alpha'+\theta)} + T^{-1}n^{\alpha-\alpha'+\theta}}.$$

Here again $C(t)$ is a positive constant independent of n and d. By the choice of $\alpha' > \alpha$ and $\alpha' - \alpha > \theta$, we have that the right-hand-side goes to infinity. We can then lower bound $C(t)$ with $C(T)$ simply by letting $C(T) = \min_{1\leq t\leq T} C(t) > 0$. Hence

$$\Pr\left(\sum_{i=1}^{\beta} \boldsymbol{X}_i^{(t,k)} - \beta\mu^{(t,k)} \geq \tilde{\varepsilon}\right) \leq \exp\left(-C(T)n^c\right) \quad \text{when } \tilde{\varepsilon} = n^{\alpha'-1/2}/T \tag{3.120}$$

for some $c > 0$. Note that the constants c and $C(T)$ are independent of t and k, only depends on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, |\Omega|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta$ and T. Similarly, by taking $-\boldsymbol{X}_i^{(t,k)}$ and using the same bounds on b and $\sigma_{(t,k)}^2$, we get that

$$\Pr\left(-\big[\sum_{i=1}^{\beta} \boldsymbol{X}_i^{(t,k)} - \beta\mu^{(t,k)}\big] \geq \tilde{\varepsilon}\right) \leq \exp\left(-C(T)n^c\right) \quad \text{when } \tilde{\varepsilon} = n^{\alpha'-1/2}/T. \tag{3.121}$$

Therefore, it follows that

$$\Pr\left(\left|\sum_{i=1}^{\beta} \boldsymbol{X}_i^{(t,k)} - \beta\mu^{(t,k)}\right| \geq \tilde{\varepsilon}\right) \leq 2\exp\left(-C(T)n^c\right) \quad \text{when } \tilde{\varepsilon} = n^{\alpha'-1/2}/T, \tag{3.122}$$

for some $c > 0$ and $C(T) > 0$ where the constants do not depend on t and n and d. The constants do depend on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, |\Omega|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta$ and T. We set $\varepsilon = n^{\alpha'-1/2}$.

Applying the union bound twice and using (3.122), we get

$$\begin{aligned}
\Pr\left(\max_{0\le t\le T}|\mathcal{E}^{(\nabla^2:\text{B.1})}_{t\wedge\vartheta}|\ge\varepsilon\right) &\le \sum_{t=0}^{T}\Pr\left(|\mathcal{E}^{(\nabla^2:\text{B.1})}_{t\wedge\vartheta}|\ge\varepsilon\right)\\
&\le \sum_{t=0}^{T}\Pr\left(\sum_{k=1}^{t\wedge\vartheta}\left|\left(\sum_{i\in B_k}X_i^{(t,k)}-\beta\mu^{(t,k)}\right)\right|\ge\varepsilon\right)\\
&\le \sum_{t=0}^{T}\sum_{k=1}^{t\wedge\vartheta}\Pr\left(\left|\sum_{i=1}^{\beta}X_i^{(t,k)}-\beta\mu^{(t,k)}\right|\ge\varepsilon/(t\wedge\vartheta)\right)\\
&\le \sum_{t=0}^{T}\sum_{k=1}^{t\wedge\vartheta}\Pr\left(\left|\sum_{i=1}^{\beta}X_i^{(t,k)}-\beta\mu^{(t,k)}\right|\ge\varepsilon/T\right)\\
&\le T^2\cdot\exp(-C(T)n^c)
\end{aligned}$$

for some $c, C(T) > 0$. In the penultimate inequality, we used that $t \wedge \vartheta \le T$. For the last inequality, we note that $\varepsilon/T = \tilde{\varepsilon}$ in (3.122) and that the constants, $c, C(T) > 0$ in (3.122) hold for all $1 \le t \le T$. The result immediately follows. □

Corollary 3. *Fix $T > 0$. Let $\alpha' > \alpha_0 + \theta$ and $\alpha > 0$ such that $\alpha' > \alpha > \alpha_0 + \theta$. The following holds with overwhelming probability*

$$\max_{0\le t\le T\wedge\vartheta}|\mathcal{E}^{(\nabla^2:\text{B.2})}_{t\wedge\vartheta}| = C(T)n^{\alpha'-1/2}, \tag{3.123}$$

provided $\alpha' - \alpha > \theta$.

Proof. We observe that for $t \geq 0$ we have

$$\begin{aligned}
\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{B.2})} &= \frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\sum_{i,j}\left[\zeta^2 \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} \widetilde{N}_{ij}^{(t,k)} \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k-1} - \zeta \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \widetilde{N}_{ij}^{(t,k)} \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k-1}\right] \\
&= \frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\left[\zeta^2 \left(\boldsymbol{A}^T \boldsymbol{w}_{k-1}\right)_i \widetilde{N}_{ij}^{(t,k)} \left(\boldsymbol{A}^T \boldsymbol{w}_{k-1}\right)_j - \zeta \left(\boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1}\right)_j \widetilde{N}_{ij}^{(t,k)} \left(\boldsymbol{A}^T \boldsymbol{w}_{k-1}\right)_i\right] \\
&= \frac{1}{2}\sum_{k=1}^{t}\left[\zeta^2 \left(\boldsymbol{A}^T \boldsymbol{w}_{k-1}\right)^T \widetilde{\boldsymbol{N}}_{t,k} \left(\boldsymbol{A}^T \boldsymbol{w}_{k-1}\right) - \zeta \left(\boldsymbol{A}^T \boldsymbol{w}_{k-1}\right)^T \widetilde{\boldsymbol{N}}_{t,k} \left(\boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1}\right)\right] \\
&= \frac{1}{2}\sum_{k=1}^{t}\left[\zeta^2 \left(\boldsymbol{A}^T \boldsymbol{w}_{k-1}\right)^T \widetilde{\boldsymbol{N}}_{t,k} \left(\boldsymbol{A}^T \boldsymbol{w}_{k-1}\right) - \zeta \left(\boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1}\right)^T \widetilde{\boldsymbol{N}}_{t,k} \left(\boldsymbol{A}^T \boldsymbol{w}_{k-1}\right)\right] \\
&= \mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{B.1})}
\end{aligned}$$

where in the penultimate step we used the fact that $\widetilde{\boldsymbol{N}}_{(t,k)}$ is a symmetric matrix for each $t \in [T \wedge \vartheta]$. By applying Proposition 9 we achieve our desired result. □

Lemma 10. *The following holds*

$$\max_{0\leq t\leq T} |\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{Z.1})}| \leq n^{2\theta-1} \quad \textit{and} \quad \max_{0\leq t\leq T} |\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{ Z.2})}| \leq C(T) n^{2\theta-1}. \tag{3.124}$$

Proof. A simple computation shows that

$$\frac{\beta(\beta-1)}{n(n-1)} - \zeta^2 = \frac{n\beta(\beta-1) - \beta^2(n-1)}{n^2(n-1)} = \frac{\zeta(\zeta-1)}{n-1} \leq C(\zeta) n^{-1}. \tag{3.125}$$

First we show the result for $\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\mathrm{Z}.1)}$ (see (3.105)). By applying the definition of the stopping time ϑ, we deduce that

$$\begin{aligned}
|\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\mathrm{Z}.1)}| &\leq \frac{\zeta(\zeta-1)}{n-1}\cdot\frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}|\sum_{i,j}\left(\boldsymbol{A}^T\boldsymbol{w}_{k-1}^{\vartheta}\right)_i\widetilde{N}_{ij}^{(t,k)}\left(\boldsymbol{A}^T\boldsymbol{w}_{k-1}^{\vartheta}\right)_j| \\
&= \frac{\zeta(\zeta-1)}{n-1}\cdot\frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}|\left(\boldsymbol{A}^T\boldsymbol{w}_{t-1}^{\vartheta}\right)^T\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T\boldsymbol{w}_{k-1}^{\vartheta}| \\
&\leq \frac{\zeta(\zeta-1)}{n-1}\cdot\frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}||\boldsymbol{A}||^2\cdot||\widetilde{\boldsymbol{N}}_{t,k}||\cdot||\boldsymbol{w}_{k-1}^{\vartheta}||^2 \\
&\leq C(T)n^{2\theta-1},
\end{aligned}$$

where the constant C is dependent on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta, T$, and $|\Omega|$, (here take $\max$ of the constants over $1 \leq k \leq T$), but it is independent of n or d. Now taking the maximum over $0 \leq t \leq T$ proves the result.

Next we show the result holds for $\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{Z.2})}$ (see (3.105)). Applying Lemma (5) with $\boldsymbol{W} = \widetilde{\boldsymbol{N}}_{t,k}$, we deduce that

$$
\begin{aligned}
|\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{Z.2})}| &\leq \frac{\zeta(\zeta-1)}{n-1}\cdot\frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\left|\sum_{i,j}\sum_{\ell=1}^{n}\boldsymbol{A}_{\ell j}\boldsymbol{A}_{\ell i}\widetilde{N}_{ij}^{(k,t)}(\boldsymbol{w}_{k-1,\ell}^{\vartheta})^2\right| \\
&= \frac{\zeta(\zeta-1)}{n-1}\cdot\frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\left|\sum_{\ell=1}^{n}(\boldsymbol{A}\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T)_{\ell\ell}(\boldsymbol{w}_{k-1,\ell}^{\vartheta})^2\right| \\
&= \frac{\zeta(\zeta-1)}{n-1}\cdot\frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\left|\sum_{\ell=1}^{n}\left[(\boldsymbol{A}\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T)_{\ell\ell}-\frac{1}{n}\text{tr}\left(\boldsymbol{A}\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T\right)+\frac{1}{n}\text{tr}(\boldsymbol{A}\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T)\right](\boldsymbol{w}_{k-1,\ell}^{\vartheta})^2\right| \\
&\leq \frac{\zeta(\zeta-1)}{n-1}\cdot\frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\left|\sum_{\ell=1}^{n}\left[(\boldsymbol{A}\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T)_{\ell\ell}-\frac{1}{n}\text{tr}\left(\boldsymbol{A}\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T\right)\right](\boldsymbol{w}_{k-1,\ell}^{\vartheta})^2\right| \\
&\quad+\frac{\zeta(\zeta-1)}{n-1}\cdot\frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\left|\sum_{\ell=1}^{n}\frac{1}{n}\text{tr}(\boldsymbol{A}\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T)(\boldsymbol{w}_{k-1,\ell}^{\vartheta})^2\right| \\
&\leq \frac{\zeta(\zeta-1)}{n-1}\cdot\frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\max_{1\leq\ell\leq n}|(\boldsymbol{A}\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T)_{\ell\ell}-\frac{1}{n}\text{tr}\left(\boldsymbol{A}\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T\right)|\,||\boldsymbol{w}_{k-1}^{\vartheta}||_2^2 \\
&\quad+\frac{\zeta(\zeta-1)}{n-1}\cdot\frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\frac{1}{n}\text{tr}\left(\boldsymbol{A}\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T\right)||\boldsymbol{w}_{k-1}^{\vartheta}||_2^2 \\
&\leq C'(t)\cdot n^{\alpha_0-1/2}\cdot n^{2\theta-1}+C(t)n^{2\theta-1},
\end{aligned}
$$

where the constants C and C' are dependent on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta, T, |\Omega|$, and time t (here take $\max$ of the constants over $1 \leq k \leq t$), but it is independent of n or d. Since $0 \leq t \leq T$ and t is finite, we can take the maximum over both sides. □

Hanson-Wright Error Terms, $\mathcal{E}_t^{(\nabla^2:\text{HW})}$. In this section, we provide the results required to control the errror

$$
\begin{aligned}
\mathcal{E}_t^{(\nabla^2:\text{HW})} &= \frac{1}{2}\sum_{k=1}^{t}\sum_{ij}\widetilde{N}_{ij}^{(t,k)}\left[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}-\mathbb{E}[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_t\boldsymbol{w}_{k-1}|\mathcal{F}_{k-1}]\right] \\
&= \frac{1}{2}\sum_{k=1}^{t}\left[(\boldsymbol{P}_k\boldsymbol{w}_{k-1})^T\boldsymbol{A}\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T(\boldsymbol{P}_k\boldsymbol{w}_{k-1})-\mathbb{E}\left[(\boldsymbol{P}_k\boldsymbol{w}_{k-1})^T\boldsymbol{A}\widetilde{\boldsymbol{N}}_{t,k}\boldsymbol{A}^T(\boldsymbol{P}_k\boldsymbol{w}_{k-1})\,|\mathcal{F}_{k-1}\right]\right]
\end{aligned}
\tag{3.126}
$$

The following definition, lemma, and remarks will help us control $\mathcal{E}_t^{(\nabla^2:\mathrm{HW})}$.

Definition 17 (Convex concentration property, [2]). *Let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^n$. We will say that $\boldsymbol{X}$ has the convex concentration property with constant K if for every $1-$Lipschitz convex function $\varphi : \mathbb{R}^n \to \mathbb{R}$. we have $\mathbb{E}[\varphi(\boldsymbol{X})] < \infty$ and for every $t > 0$,*

$$\Pr(|\varphi(\boldsymbol{X}) - \mathbb{E}\varphi(\boldsymbol{X})| \geq t) \leq 2\exp(-t^2/K^2).$$

Remark 4. Let x_i be an entry of $\boldsymbol{X}$. By a simple scaling, the previous remark can extend to $x_1, \cdots, x_n \in [a, b]$, in which case K in the definition above will be replaced by $K(b-a)$.

What is interesting for us is that vectors obtained via sampling without replacement follow the convex concentration property ([2, Remark 2.3]). More precisely, if $x_1, \cdots, x_n \in [0, 1]$ and the random vector $\boldsymbol{X} = (X_1, \cdots, X_m)$ with $m \leq n$ is obtained by sampling without replacement m numbers from the set $\{x_1, \cdots, x_n\}$, then $\boldsymbol{X}$ satisfies the convex concentration property with an absolute constant K. In this sense, the following lemma ([2, Theorem 2.5]) will be useful to us.

Lemma 11 (Hanson-Wright concentration for sampling without replacement, Theorem 2.5 [2]). *Let $\boldsymbol{X}$ be a mean zero random vector in $\mathbb{R}^n$. If $\boldsymbol{X}$ has the convex concentration property with constant K, then for any $n \times n$ matrix $\boldsymbol{A}$ and every $t > 0$,*

$$\Pr(|\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} - \mathbb{E}\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}| \geq t) \leq 2\exp\left(-\frac{1}{C}\min\left(\frac{t^2}{2K^4\|\boldsymbol{A}\|_{HS}^2}, \frac{t}{K^2\|\boldsymbol{A}\|}\right)\right),$$

for some universal constant C.

Remark 5. The assumption that $\boldsymbol{X}$ is centered is introduced just to simplify the statement of the theorem. Note that if $\boldsymbol{X}$ has the convex concentration property with constant K, then so does $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]$. Moreover, observe,

$$\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} = (\tilde{\boldsymbol{X}} + \mathbb{E}[\boldsymbol{X}])^T\boldsymbol{A}(\tilde{\boldsymbol{X}} + \mathbb{E}[\boldsymbol{X}]) = \tilde{\boldsymbol{X}}^T\boldsymbol{A}\tilde{\boldsymbol{X}} + 2\tilde{\boldsymbol{X}}^T\boldsymbol{A}\mathbb{E}[\boldsymbol{X}] + \mathbb{E}[\boldsymbol{X}]^T\boldsymbol{A}\mathbb{E}[\boldsymbol{X}]$$

and $\quad \mathbb{E}\left[\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\right] = \mathbb{E}\left[\tilde{\boldsymbol{X}}^T\boldsymbol{A}\tilde{\boldsymbol{X}}\right] + \mathbb{E}\left[\boldsymbol{X}\right]^T\boldsymbol{A}\mathbb{E}\left[\boldsymbol{X}\right],$

as $\mathbb{E}\left[\tilde{\boldsymbol{X}}^T \boldsymbol{A} \mathbb{E}\left[\boldsymbol{X}\right]\right] = 0$. This implies by Lemma 18 and convex concentration property for linear functions,

$$\begin{aligned}
&\Pr(|\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}]| \geq t) \\
&\quad \leq \Pr(|\tilde{\boldsymbol{X}}^T\boldsymbol{A}\tilde{\boldsymbol{X}} - \mathbb{E}[\tilde{\boldsymbol{X}}^T\boldsymbol{A}\tilde{\boldsymbol{X}}]| \geq t/3) + 2\Pr(|\tilde{\boldsymbol{X}}^T\boldsymbol{A}\mathbb{E}[\boldsymbol{X}] - \mathbb{E}\left[\tilde{\boldsymbol{X}}^T\boldsymbol{A}\mathbb{E}[\boldsymbol{X}]\right]| \geq t/3) \\
&\quad \leq 2\exp\left(-\frac{1}{C}\min\left(\frac{t^2}{2\cdot 9K^4\|\boldsymbol{A}\|_{HS}^2}, \frac{t}{3K^2\|\boldsymbol{A}\|}\right)\right) + 2\cdot 2\exp\left(-\frac{t^2}{9K^2\|\boldsymbol{A}\mathbb{E}[\boldsymbol{X}]\|_2^2}\right).
\end{aligned}$$

Using Lemma 18 and Remark 5 we can show $\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\mathrm{HW})}$ is small for large n and d.

Lemma 12. *For all α' and α such that $\alpha' > \alpha > \alpha_0 + \theta$ and θ, we have*

$$\max_{1\leq t\leq T} |\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:HW)}| \leq C(T)n^{2\alpha'-1/2} \quad \textit{w.o.p.},$$

provided $0 < \theta < 2\alpha' - \alpha$.

Proof. Let $\boldsymbol{X}_k \overset{\text{def}}{=} \boldsymbol{P}_k\boldsymbol{w}_{k-1}^{\vartheta}$ and $\boldsymbol{D}^{(t-k)} \overset{\text{def}}{=} \boldsymbol{A}\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T$. Then

$$\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\mathrm{HW})} = \frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\left[\boldsymbol{X}_k^T\boldsymbol{D}^{(t-k)}\boldsymbol{X}_k - \mathbb{E}\left[\boldsymbol{X}_k^T\boldsymbol{D}^{(t-k)}\boldsymbol{X}_k|\mathcal{F}_{k-1}\right]\right]. \tag{3.127}$$

In view of union bounds, it suffices to impose bounds on each summand of (3.127) since $k = 1, \cdots, t\wedge\vartheta$ and $t \leq T$. We first specify K as the absolute constant in Definition 17. In light of Remark 5 we replace K with $K\cdot M_k$ where $M_k \overset{\text{def}}{=} \max_{i\in[n]}|(\boldsymbol{P}_k\boldsymbol{w}_{k-1}^{\vartheta})_i|$. By our choice of α such that $\alpha > \alpha_0 + \theta$ and $\alpha < \alpha'$, then we get from Lemma 4 with α,

$$\max_{1\leq k\leq T}\max_{1\leq i\leq n}|\boldsymbol{e}_i^T\boldsymbol{w}_{k-1}^{\vartheta}| \leq C(T)n^{\alpha-1/2}$$

and we obtain

$$M_k = \max_{i\in[n]}|(\boldsymbol{X}_k)_i| = \max_{i\in[n]}|(\boldsymbol{P}_k\boldsymbol{w}_{k-1}^{\vartheta})_i| \leq \max_{1\leq k\leq T}\max_{1\leq i\leq n}|\boldsymbol{e}_i^T\boldsymbol{w}_{k-1}^{\vartheta}| \leq C(T)n^{\alpha-1/2}. \tag{3.128}$$

In order to apply Lemma 18, we need to compute $\|\boldsymbol{D}^{(t-k)}\|_{HS}$ and $\|\boldsymbol{D}^{(t-k)}\|$. Now observe that

$$\|\boldsymbol{D}^{(t-k)}\|_{HS}^2 \leq \|\boldsymbol{A}\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}\|^2\|\boldsymbol{A}\|_{HS}^2 \leq \|\boldsymbol{A}\|^4\|\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}\|^2 \cdot n \leq C(T)n \tag{3.129}$$

and

$$\|\boldsymbol{D}^{(t-k)}\|_2 = \|\boldsymbol{A}\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\| \leq \|\boldsymbol{A}\|^2\|\widetilde{\boldsymbol{N}}_{(t\wedge\vartheta,k)}\| \leq C(T). \tag{3.130}$$

We note that $t \leq T$ and $\|\widetilde{Q}^{(t\wedge k)}\|$ can be upper bounded by constants independent of n and d. Lastly, we bound $||\boldsymbol{D}^{(t-k)}\mathbb{E}[\boldsymbol{X}_k|\mathcal{F}_{k-1}]||$ where $\mathbb{E}[\boldsymbol{X}_k \,|\, \mathcal{F}_{k-1}] = (\mu_1, \ldots, \mu_n)$ and $\mu_\ell = \zeta \boldsymbol{w}_{k-1,\ell}^{\vartheta}$. Using the definition of the stopping time ϑ and (3.130) we obtain

$$\|\boldsymbol{D}^{(t-k)}\mathbb{E}[\boldsymbol{X}_k|\mathcal{F}_{k-1}]\| \leq \|\boldsymbol{D}^{(t-k)}\| \cdot \|\mathbb{E}[\boldsymbol{X}_k \,|\, \mathcal{F}_{k-1}]\| = \zeta\|\boldsymbol{D}^{(t-k)}\|\|\boldsymbol{w}_{k-1}^{\vartheta}\| \leq C(T)n^{\theta} \tag{3.131}$$

Using Lemma 18 (Hanson-Wright) and the remark following it, we have

$$\begin{aligned}
&\Pr(|\boldsymbol{X}_k^T\boldsymbol{D}^{(t-k)}\boldsymbol{X}_k - \mathbb{E}\boldsymbol{X}_k^T\boldsymbol{D}^{(t-k)}\boldsymbol{X}_k| \geq \tilde{\epsilon} \,|\, \mathcal{F}_{k-1}) \\
&\qquad \leq 2\exp\left(-\frac{1}{C}\min\left(\frac{\tilde{\epsilon}^2}{2\cdot 9M_k^4K^4\|\boldsymbol{D}^{(t-k)}\|_{HS}^2}, \frac{\tilde{\epsilon}}{3M_k^2K^2\|\boldsymbol{D}^{(t-k)}\|}\right)\right) \\
&\qquad\quad + 2\cdot 2\exp\left(-\frac{\tilde{\epsilon}^2}{9M_k^2K^2\|\boldsymbol{D}^{(t-k)}(\mathbb{E}[\boldsymbol{X}_k \,|\, \mathcal{F}_{k-1}])\|^2}\right).
\end{aligned}$$

Let $\tilde{\epsilon} = n^{2\alpha'-1/2}\cdot 2(T)^{-1}$. By using (3.128), (3.129), (3.130), and (3.131),

$$\frac{\tilde{\epsilon}^2}{2\cdot 9M_k^4K^4\|\boldsymbol{D}^{(t-k)}\|_{HS}^2} \geq C(T)\cdot\frac{n^{4\alpha'-1}}{n^{4\alpha-2}n} = C(T)\cdot\frac{1}{n^{4(\alpha-\alpha')}}, \tag{3.132}$$

$$\frac{\tilde{\epsilon}}{3M_k^2K^2\|\boldsymbol{D}^{(t-k)}\|} \geq C(T)\cdot\frac{n^{2\alpha'-1/2}}{n^{2\alpha-1}} = C(T)\cdot\frac{1}{n^{2(\alpha-\alpha')-1/2}}, \tag{3.133}$$

$$\text{and}\quad \frac{\tilde{\epsilon}^2}{M_k^2K^2\|\boldsymbol{D}^{(t-k)}(\mathbb{E}[\boldsymbol{X}_k \,|\, \mathcal{F}_{k-1}])\|_2^2} \geq C(T)\cdot\frac{n^{4\alpha'-1}}{n^{2\alpha-1}n^{2\theta}} = C(T)\cdot\frac{1}{n^{2\alpha-4\alpha'+2\theta}}, \tag{3.134}$$

where $C(T)$ is independent of n and k, and only depends on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, |\Omega|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta$ and T. Therefore by our choice of $\alpha' > \alpha$ and $0 < \theta < 2\alpha' - \alpha$, we get

$$\boldsymbol{P}(|\boldsymbol{X}_k^T\boldsymbol{D}\boldsymbol{X}_k - \mathbb{E}\boldsymbol{X}_k^T\boldsymbol{D}\boldsymbol{X}_k| \geq \tilde{\epsilon}|\mathcal{F}_{k-1}) \leq 2\exp\left(-C(T)n^c\right) \qquad \text{when } \tilde{\epsilon} = \frac{2n^{2\alpha'-1/2}}{T} \tag{3.135}$$

for some $c > 0$ and $T > 0$. We set $\epsilon = n^{2\alpha' - 1/2}$. Applying two union bounds and (3.135), we get

$$\begin{aligned}
&\Pr\left(\max_{0\le t\le T} |\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\mathrm{HW})}| \ge \epsilon\right) \le \sum_{t=0}^{T} \Pr\left(|\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\mathrm{HW})}| \ge n^{2\alpha'-1/2}\right) \\
&\le \sum_{t=0}^{T} \Pr\left(\frac{1}{2}\sum_{k=1}^{t\wedge\vartheta} \left|\boldsymbol{X}_k^T \boldsymbol{D}^{(t-k)} \boldsymbol{X}_k - \mathbb{E}[\boldsymbol{X}_k^T \boldsymbol{D}^{(t-k)} \boldsymbol{X}_k | \mathcal{F}_{k-1}]\right| \ge \epsilon\right) \\
&\le \sum_{t=0}^{T} \sum_{k=1}^{t\wedge\vartheta} \Pr\left(\left|\boldsymbol{X}_k^T \boldsymbol{D}^{(t-k)} \boldsymbol{X}_k - \mathbb{E}[\boldsymbol{X}_k^T \boldsymbol{D}^{(t-k)} \boldsymbol{X}_k | \mathcal{F}_{k-1}]\right| \ge \frac{2n^{2\alpha'-1/2}}{T}\right) \\
&\le 2T^2 \exp(-C(T) n^c)
\end{aligned} \tag{3.136}$$

for some $c, C(T) > 0$. In the penultimate inequality, we used that $t \wedge \vartheta \le T$. The result immediately follows. □

Now we return back to Proposition 9 as we have shown all the components are small.

Proof of Proposition 9. Let $\alpha_0 \in (0, 1/4)$ as specified in Assumption 5. Let α, α', and θ satisfying $0 < \theta < \alpha' - \alpha$ and $\alpha_0 + \theta < \alpha < \alpha'$. Then applying the decomposition in Lemma 7 and the lemmas showing the decomposed error terms are small when n and d are large (i.e. Lemma 8, 9, 10, 23, and Corollary 3) we get

$$\begin{aligned}
\max_{0\le t\le T} |\mathcal{E}_t^{\nabla^2\text{-Diag}}| &\le \max_{0\le t\le T} |\mathcal{E}_t^{(\nabla^2:\mathrm{KL})}| + \max_{0\le t\le T} |\mathcal{E}_t^{(\nabla^2:\mathrm{B.1})}| + \max_{0\le t\le T} |\mathcal{E}_t^{(\nabla^2:\mathrm{B.2})}| \\
&\quad + \max_{0\le t\le T} |\mathcal{E}_t^{(\nabla^2:\mathrm{Z.1})}| + \max_{0\le t\le T} |\mathcal{E}_t^{(\nabla^2:\mathrm{Z.2})}| + \max_{0\le t\le T} |\mathcal{E}_t^{(\nabla^2:\mathrm{HW})}| \\
&\le C(T) n^{\alpha_0 - 1/2 + 2\theta} + 2C(T) n^{\alpha' - 1/2} + 2C(T) n^{2\theta - 1} + C(T) n^{2\alpha' - 1/2} \\
&\le C(T) n^{2\alpha' - 1/2}
\end{aligned} \tag{3.137}$$

holds with overwhelming probability. □

3.7.3 Error from the off-diagonal term

In this section, we control the term

$$\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Off}} = \sum_{1\le k_1<k_2}^{t\wedge\vartheta} \mathring{\boldsymbol{M}}_{k_1}^T \boldsymbol{N}_{(t\wedge\vartheta,k_1)}(\nabla^2\mathcal{R})\boldsymbol{N}_{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2} \tag{3.138}$$

where we recall the martingale increment (3.5) and the matrix $\boldsymbol{N}_{t,k}$ from Proposition 4, respectively,

$$\mathring{\boldsymbol{M}}_k = \zeta \boldsymbol{A}^T \boldsymbol{w}_{k-1} - \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \quad \text{and} \quad \boldsymbol{N}_{t,k} = \boldsymbol{N}_{t,k}(\boldsymbol{H}). \tag{3.139}$$

We will show that this error term is small, the main result of this section.

Proposition 10. *For all α and δ such that $\alpha_0 + \theta < \alpha < \delta < 1/4$ and $0 < \theta < \frac{1}{4} - \delta$ and for any η satisfying $0 < \eta < \delta - \alpha$ we have that the quadratic-off-diagonal term satisfies*

$$\max_{0\le t\le T} \left|\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-}Off}\right| = C(T)\cdot n^{\alpha-\delta+\eta} \qquad w.o.p., \tag{3.140}$$

for some constant $C(T) > 0$ that is independent of n and d and only depends on $||\Omega||$, $\|\boldsymbol{A}^T\boldsymbol{A}\|$, γ and T.

To prove this result, we first fix the k_2 terms and show the resulting summation is small using Bardenet [9, Proposition 1.4] result (Proposition 21). To do so, we need to show certain terms are themselves small, which depend on the martingale increment $\mathring{\boldsymbol{M}}_{k_2}$. This will require us to use Hanson-Wright, Lemma 18. Combining both Proposition 21 (Bardenet) and Lemma 18 (Hanson-Wright), Proposition 10 will follow

To this end, for convenience, we define

$$\boldsymbol{V}^{(t\wedge\vartheta,k_2)} \overset{\text{def}}{=} \sum_{k_1=1}^{k_2-1} \mathring{\boldsymbol{M}}_{k_1}^T \boldsymbol{N}_{(t\wedge\vartheta,k_1)}(\nabla^2\mathcal{R})\boldsymbol{N}_{(t\wedge\vartheta,k_2)} \qquad k_2 = 1,2,\cdots t\wedge\vartheta \tag{3.141}$$

$$\boldsymbol{Y}^{(t\wedge\vartheta,k_2)} \overset{\text{def}}{=} \sum_{j\in B_{k_2-1}} \left(\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\right)_j (\boldsymbol{A}\boldsymbol{x}_{k_2-1} - \boldsymbol{b})_j \qquad k_2 = 1,2,\cdots t\wedge\vartheta \tag{3.142}$$

$$\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)} \stackrel{\text{def}}{=} \left(\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\right)_j \left(\boldsymbol{A}\boldsymbol{x}_{k_2-1}-\boldsymbol{b}\right)_j \qquad\qquad k_2 = 1,2,\cdots t\wedge\vartheta. \tag{3.143}$$

The following lemma and its corollary will be useful in showing that the off-diagonal term is small.

Lemma 13. *Let $\delta > 0$ such that $\alpha_0 + \theta < \delta < 1/4$ and $0 < \theta < \frac{1}{4} - \delta$. Then the following holds with overwhelming probability*

$$\max_{1\le t\le T}\max_{1\le k_2\le t\wedge\vartheta} \|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\|^2 \le C(T)\cdot n^{1-2\delta}. \tag{3.144}$$

Proof. We have

$$\begin{aligned}
&\max_{0\le t\le T}\max_{1\le k_2\le t\wedge\vartheta} \|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\| = \max_{0\le t\le T}\max_{1\le k_2\le t\wedge\vartheta} \left\|\left(\sum_{k_1=1}^{k_2-1} \mathring{\boldsymbol{M}}_{k_1}^T \boldsymbol{N}_{(t\wedge\vartheta,k_1)}(\nabla^2\mathcal{R})\boldsymbol{N}_{(t\wedge\vartheta,k_2)}\right)\boldsymbol{A}^T\right\| \\
&\le T\max_{0\le t\le T}\max_{1\le k_2\le t\wedge\vartheta}\max_{1\le k_1\le k_2-1} \left\|\left(\mathring{\boldsymbol{M}}_{k_1}^T \boldsymbol{N}_{(t\wedge\vartheta,k_1)}\left(\nabla^2\mathcal{R}\right)\boldsymbol{N}_{(t\wedge\vartheta,k_2)}\right)\boldsymbol{A}^T\right\| \\
&\le T\max_{0\le t\le T}\max_{1\le k_2\le t\wedge\vartheta}\max_{1\le k_1\le k_2-1} \|\mathring{\boldsymbol{M}}_{k_1}\|\cdot\|\boldsymbol{N}_{(t\wedge\vartheta,k_1)}\|\cdot\|\boldsymbol{N}_{(t\wedge\vartheta,k_2)}\|\cdot\|\boldsymbol{A}\|\cdot\|\nabla^2\mathcal{R}\| \\
&\le T\left(\max_{1\le k_1\le (T\wedge\vartheta)-1}\|\mathring{\boldsymbol{M}}_{k_1}\|\right)\cdot\left(\max_{0\le t\le T}\max_{1\le k_2\le t\wedge\vartheta}\|\boldsymbol{N}_{(t\wedge\vartheta,k_2)}\|^2\cdot\|\boldsymbol{A}\|\cdot\|\nabla^2\mathcal{R}\|\right) \\
&\le C(T)\cdot\max_{1\le k_1\le (T\wedge\vartheta)-1}\|\mathring{\boldsymbol{M}}_{k_1}\|
\end{aligned} \tag{3.145}$$

which implies

$$\max_{0\le t\le T}\max_{1\le k_2\le t\wedge\vartheta} \|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\|^2 \le C(T)\cdot\max_{1\le k_1\le (T\wedge\vartheta)-1}\|\mathring{\boldsymbol{M}}_{k_1}\|^2 \tag{3.146}$$

where the constant C is dependent on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta, |\Omega|$, and time T, but it is independent of n or d.

Fix $k_1 \leq (T \wedge \vartheta) - 1$. By adding and subtracting the mean of the quadratic martingale term and applying the triangle inequality we get

$$\|\mathring{\boldsymbol{M}}_{k_1}\|^2 \leq \left| \mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} - \mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} \mid \mathcal{F}_{k_1-1} \right] \right| + \left| \mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} \mid \mathcal{F}_{k_1-1} \right] \right|. \tag{3.147}$$

We address the first term in the sum by applying Lemma 18 (Hanson-Wright). We specify K as the absolute constant in Definition 17. In light of remark 5, we replace K with $K \cdot M_{k_1}$ where $M_{k_1} \overset{\text{def}}{=} \max_{1 \leq i \leq n} |\mathring{\boldsymbol{M}}_{k_1,i}|$. By the definition of the stopping time ϑ and the fact that $k_1 \leq (T \wedge \vartheta) - 1$ we get

$$\begin{aligned}
M_{k_1} &= \max_{1 \leq i \leq n} |\mathring{\boldsymbol{M}}_{k_1,i}| \\
&= \max_{1 \leq i \leq n} |(\zeta \boldsymbol{A}^T \boldsymbol{w}_{k_1-1})_i - (\boldsymbol{A}^T \boldsymbol{P}_{k_1} \boldsymbol{w}_{k_1-1})_i| \\
&\leq \max_{1 \leq i \leq n} \zeta |\boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k_1-1}| + \max_{1 \leq i \leq n} |\boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_{k_1} \boldsymbol{w}_{k_1-1}| \\
&\leq \zeta \cdot \|\boldsymbol{A}\| \cdot \|\boldsymbol{w}_{k_1-1}\| + \|\boldsymbol{A}\| \cdot \|\boldsymbol{w}_{k_1-1}\| \\
&\leq C(\zeta, \|\boldsymbol{A}\|) \cdot n^{\theta}
\end{aligned} \tag{3.148}$$

where $C(\zeta, \|\boldsymbol{A}\|)$ is independent of n or d. Choosing $\epsilon = n^{1-2\delta}$ and applying Lemma 18 we get

$$\begin{aligned}
&\Pr\left(\left| \mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} - \mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} \mid \mathcal{F}_{k_1-1} \right] \right| \geq \epsilon \right) \\
&\qquad \leq 2 \exp\left(-\frac{1}{D} \min\left(\frac{\epsilon^2}{2 M_{k_1}^4 K^4 \|\mathbf{I}\|_{HS}^2}, \frac{\epsilon}{M_{k_1}^2 K^2 \|\mathbf{I}\|} \right) \right).
\end{aligned} \tag{3.149}$$

for some universal constant $D > 0$. Moreover we have that

$$\begin{aligned}
&\frac{\epsilon^2}{2 M_{k_1}^4 K^4 \|\mathbf{I}\|_{HS}^2} \geq C \cdot \frac{n^{2-4\delta}}{n^{4\theta} n} = C \cdot \frac{1}{n^{4(\delta+\theta)-1}} \\
\text{and} \qquad &\frac{\epsilon}{M_{k_1}^2 K^2 \|\mathbf{I}\|} \geq C \cdot \frac{n^{1-2\delta}}{n^{2\theta}} = C \cdot \frac{1}{n^{2(\delta+\theta)-1}}
\end{aligned}$$

where $C > 0$ is independent of n and d. By our assumptions on δ and θ, namely $\delta+\theta < 1/4$, we have

$$\Pr\left(\left|\mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} - \mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} \mid \mathcal{F}_{k_1-1}\right]\right| \geq n^{1-2\delta}\right) \leq \exp(-Cn^c) \tag{3.150}$$

for some $c > 0$ and constant C independent of n and d.

Now we bound $\mathbb{E}\left[\|\mathring{\boldsymbol{M}}_{k_1}\|^2 \mid \mathcal{F}_{k_1-1}\right]$ in (3.147). From (3.112), we know that

$$\begin{aligned}
\mathring{\boldsymbol{M}}_{k_1,i}\mathring{\boldsymbol{M}}_{k_1,i} &= (\zeta-\zeta^2)\sum_{\ell=1}^n A_{\ell i}A_{\ell i}w_{k_1-1,\ell}^2 + \left(\frac{\beta(\beta-1)}{n(n-1)} - \zeta^2\right)\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k_1-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k_1-1} \\
&\quad + \left(\left[\frac{\beta}{n} - \frac{\beta(\beta-1)}{n(n-1)}\right] - (\zeta-\zeta^2)\right)\sum_{\ell=1}^n A_{\ell i}A_{\ell i}w_{k_1-1,\ell}^2 + \mathcal{E}_{k_1}^{(\nabla^2,1)}(i,i) + \mathcal{E}_{k_1}^{(\nabla^2,2)}(i,i) \\
&\quad + \left(\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_{k_1}\boldsymbol{w}_{k_1-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_{k_1}\boldsymbol{w}_{k-1} - \mathbb{E}\left[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_{k_1}\boldsymbol{w}_{k-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_{k_1}\boldsymbol{w}_{k_1-1}|\mathcal{F}_{k_1-1}\right]\right),
\end{aligned} \tag{3.151}$$

where $\mathcal{E}_{k_1}^{(\nabla^2,1)}(i,i)$ and $\mathcal{E}_{k_1}^{(\nabla^2,2)}(i,i)$ are defined in (3.108). When we take conditional expectation, we see that $\mathbb{E}[\mathcal{E}_{k_1}^{(\nabla^2,1)}(i,i) \mid \mathcal{F}_{k_1-1}], \mathbb{E}[\mathcal{E}_{k_1}^{(\nabla^2,2)}(i,i) \mid \mathcal{F}_{k_1-1}] = 0$. Therefore, we deduce that

$$\begin{aligned}
\mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1,i}\mathring{\boldsymbol{M}}_{k_1,i}|\mathcal{F}_{k_1-1}\right] &= (\zeta-\zeta^2)\sum_{\ell=1}^n A_{\ell i}A_{\ell i}w_{k_1-1,\ell}^2 + \left(\frac{\beta(\beta-1)}{n(n-1)} - \zeta^2\right)\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k_1-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k_1-1} \\
&\quad + \left(\left[\frac{\beta}{n} - \frac{\beta(\beta-1)}{n(n-1)}\right] - (\zeta-\zeta^2)\right)\sum_{\ell=1}^n A_{\ell i}A_{\ell i}w_{k_1-1,\ell}^2 \\
&= (\zeta-\zeta^2)\sum_{\ell=1}^n A_{\ell i}A_{\ell i}w_{k_1-1,\ell}^2 + \frac{\zeta(\zeta-1)}{n-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k_1-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k_1-1} \\
&\quad - \frac{\zeta(\zeta-1)}{n-1}\sum_{\ell=1}^n A_{\ell i}A_{\ell i}w_{k_1-1,\ell}^2.
\end{aligned} \tag{3.152}$$

where in the last line we used the fact $\frac{\beta(\beta-1)}{n(n-1)} - \zeta^2 = \frac{\zeta(\zeta-1)}{n-1}$. By summing over *i* and applying the definition of the stopping time, Lemma 5, the fact that $\zeta \in (0,1)$, and that $k_1 \leq (T \wedge$

$\vartheta) - 1$ we get

$$
\begin{aligned}
\left|\mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} \,|\, \mathcal{F}_{k_1-1}\right]\right| &\leq \left|\frac{\zeta(\zeta-1)}{n-1}\right| \sum_{i=1}^{d} \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k_1-1} \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k_1-1} \\
&+ \left|\frac{\zeta(\zeta-1)}{n-1}\right| \sum_{i=1}^{d}\sum_{\ell=1}^{n} A_{\ell i} A_{\ell i} w_{k_1-1,\ell}^2 + (\zeta-\zeta^2) \sum_{i=1}^{d}\sum_{\ell=1}^{n} \boldsymbol{A}_{\ell i}^2 (\boldsymbol{w}_{k_1-1,\ell})^2 \\
&= \left|\frac{\zeta(\zeta-1)}{n-1}\right| \sum_{i=1}^{d} \left(\boldsymbol{A}^T \boldsymbol{w}_{k_1-1}\right)_i^2 + \left(\left|\frac{\zeta(\zeta-1)}{n-1}\right| + (\zeta-\zeta^2)\right) \sum_{\ell=1}^{n} \boldsymbol{w}_{k_1-1,\ell}^2 \left(\sum_{i=1}^{d} \boldsymbol{A}_{\ell i}^2\right) \\
&= \left|\frac{\zeta(\zeta-1)}{n-1}\right| \|\boldsymbol{A}^T \boldsymbol{w}_{k_1-1}\|^2 + \left(\left|\frac{\zeta(\zeta-1)}{n-1}\right| + (\zeta-\zeta^2)\right) \sum_{\ell=1}^{n} \boldsymbol{w}_{k_1-1,\ell}^2 \left(\boldsymbol{A}\boldsymbol{A}^T\right)_{\ell\ell} \\
&\leq \left|\frac{\zeta(\zeta-1)}{n-1}\right| \|\boldsymbol{A}\|^2 \|\boldsymbol{w}_{k_1-1}\|^2 + \left(\left|\frac{\zeta(\zeta-1)}{n-1}\right| + (\zeta-\zeta^2)\right) \max_{1\leq j\leq n} \left(\boldsymbol{A}\boldsymbol{A}^T\right)_{jj} \|\boldsymbol{w}_{k_1-1}\|^2 \\
&\leq C(T) n^{2\theta-1} \\
&+ C(T) n^{2\theta} \left(\left|\frac{\zeta(\zeta-1)}{n-1}\right| + (\zeta-\zeta^2)\right) \left|\max_{j\in[n]} \left(\boldsymbol{A}\boldsymbol{A}^T\right)_{jj} + \frac{1}{n}\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{A}^T\right) - \frac{1}{n}\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{A}^T\right)\right| \\
&\leq C(T) n^{2\theta-1} \\
&+ C(T) n^{2\theta} \left(\left|\frac{\zeta(\zeta-1)}{n-1}\right| + (\zeta-\zeta^2)\right) \left(\left|\max_{j\in[n]} \left(\boldsymbol{A}\boldsymbol{A}^T\right)_{jj} - \frac{1}{n}\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{A}^T\right)\right| + \left|\frac{1}{n}\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{A}^T\right)\right|\right) \\
&\leq C(T) n^{2\theta-1} + C(T) n^{2\theta} \left(\left|\frac{\zeta(\zeta-1)}{n-1}\right| + (\zeta-\zeta^2)\right) \left(C(T) n^{\alpha_0 - 1/2} + \left|\frac{1}{n}\mathrm{tr}\left(\boldsymbol{A}\boldsymbol{A}^T\right)\right|\right) \\
&\leq C(T) n^{2\theta-1} + C(T) n^{\alpha_0+2\theta-1/2} + C(T) n^{2\theta}.
\end{aligned}
\tag{3.153}
$$

In the last inequality, we used that $\frac{1}{n}\mathrm{tr}(\boldsymbol{A}\boldsymbol{A}^T)$ is independent of n and d by our assumptions on the data matrix $\boldsymbol{A}$. By the assumptions on θ and δ, we have that

$$
\left|\mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} \,|\, \mathcal{F}_{k_1-1}\right]\right| \leq C(T) n^{1-2\delta}, \tag{3.154}
$$

where $C(T)$ is independent of n and d, and only depends on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, |\Omega|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta$ and T. Putting everything together, we have from (3.145), (3.150), and (3.154) that

$$\begin{aligned}\max_{0\le t\le T}\max_{1\le k_2\le t\wedge\vartheta}\|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\| &\le C(T)\max_{1\le k_1\le (T\wedge\vartheta)-1}\|\mathring{\boldsymbol{M}}_{k_1}\| \\ &\le C(T)\Bigg(\max_{1\le k_1\le t\wedge\vartheta}\left|\mathring{\boldsymbol{M}}_{k_1}^T\mathbf{I}\mathring{\boldsymbol{M}}_{k_1}-\mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1}^T\mathbf{I}\mathring{\boldsymbol{M}}_{k_1}\mid\mathcal{F}_{k-1}\right]\right|+\max_{1\le k_1\le t\wedge\vartheta}\left|\mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1}^T\mathbf{I}\mathring{\boldsymbol{M}}_{k_1}\mid\mathcal{F}_{k-1}\right]\right|\Bigg) \\ &\le C(T)n^{1-2\delta}\qquad\text{w.o.p.}\end{aligned} \tag{3.155}$$

The result is now shown. □

We now show that $X_j^{(t\wedge\vartheta,k_2)}$ (3.143) is small.

Lemma 14. *For all α and δ such that $\alpha_0+\theta<\alpha<\delta<1/4$ and $0<\theta<\frac{1}{4}-\delta$ we have*

$$\max_{0\le t\le T}\frac{1}{n}\sum_{j=1}^n|\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)}|^2\le C(T)n^{2(\alpha-\delta)-1}\quad\textit{and}\quad\max_{0\le t\le T}\max_{j\in[n]}|\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)}|\le C(T)n^{\alpha-\delta}, \tag{3.156}$$

for $0\le k_2\le t\wedge\vartheta$.

Proof. Let $0\le t\le T$ and $0\le k_2\le t\wedge\vartheta$. We achieve the first result in (3.156) as follows

$$\begin{aligned}\sum_{j=1}^n|\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)}|^2&=\sum_{j=1}^n\left|\left(\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\right)_j\left(\boldsymbol{A}\boldsymbol{x}_{k_2-1}-\boldsymbol{b}\right)_j\right|^2 \\ &\le\max_{1\le j\le n}|\boldsymbol{w}_{k_2,j}^{\vartheta}|^2\sum_{j=1}^n\left|\left(\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\right)_j\right|^2=\max_j|\boldsymbol{w}_{k_2,j}^{\vartheta}|^2\cdot||\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T||^2 \\ &\le C(t)\cdot n^{2\alpha-1}\cdot n^{1-2\delta}=C(t)n^{2(\alpha-\delta)}\end{aligned} \tag{3.157}$$

where we applied Lemma 4 and Lemma 13 in the penultimate step. This immediately gives us

$$\max_{0\le t\le T}\frac{1}{n}\sum_{j=1}^n|\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)}|^2\le C(T)\cdot n^{2(\alpha-\delta)-1} \tag{3.158}$$

where $C(T)=\max_{0\le t\le T}C(t)$. This gives the first inequality in (3.156).

For the second inequality in (3.156), using (3.157), we get

$$\max_{1\leq j\leq n}|\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)}|\leq\sqrt{\sum_{j=1}^{n}|\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)}|^2}\leq C(t)n^{\alpha-\delta}$$

which immediately yields

$$\max_{0\leq t\leq T}\max_{1\leq j\leq n}|\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)}|\leq C(T)n^{\alpha-\delta} \tag{3.159}$$

where again $C(t)$ is independent of n and d and $C(T)=\max\limits_{0\leq t\leq T}C(t)$ and the result is shown. □

Now we have all the results in order to prove the main proposition, Proposition 10, of this section.

Proof of Proposition 10. Let $0\leq t\leq T$. First, we rewrite $\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Off}}$ so that we can apply Lemma 14. To this end, we have

$$\begin{aligned}\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Off}}&=\sum_{k_1<k_2}^{t\wedge\vartheta}\mathring{\boldsymbol{M}}_{k_1}^T\boldsymbol{N}_{(t\wedge\vartheta,k_1)}(\nabla^2R)\boldsymbol{N}_{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2}\\&=\sum_{k_2=1}^{t\wedge\vartheta}\left(\sum_{k_1=1}^{k_2-1}\mathring{\boldsymbol{M}}_{k_1}^T\boldsymbol{N}_{(t\wedge\vartheta,k_1)}\left(\nabla^2\mathcal{R}\right)\boldsymbol{N}_{(t\wedge\vartheta,k_2)}\right)\mathring{\boldsymbol{M}}_{k_2}\\&=\sum_{k_2=1}^{t\wedge\vartheta}\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2},\end{aligned} \tag{3.160}$$

where $\boldsymbol{V}^{(t\wedge\vartheta,k_2)}$ is defined in (3.141). For each $1\leq k_2\leq t\wedge\vartheta$, we get

$$\begin{aligned}|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2}|&=\left|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\boldsymbol{P}_{k_2}\boldsymbol{w}_{k_2-1}^{\vartheta}-\mathbb{E}\left[\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\boldsymbol{P}_{k_2}\boldsymbol{w}_{k_2-1}^{\vartheta}\middle|\mathcal{F}_{k_2-1}\right]\right|\\&=\left|\sum_{j\in B_{k_2-1}}\left(\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\right)_j\boldsymbol{w}_{k_2-1,j}^{\vartheta}-\mathbb{E}\left[\sum_{j\in B_{k_2-1}}\left(\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\right)_j\boldsymbol{w}_{k_2-1,j}^{\vartheta}|\mathcal{F}_{k_2-1}\right]\right|\\&=\left|\boldsymbol{Y}^{(t\wedge\vartheta,k_2)}-\mathbb{E}\left[\boldsymbol{Y}^{(t\wedge\vartheta,k_2)}|\mathcal{F}_{k_2-1}\right]\right|,\end{aligned} \tag{3.161}$$

where $\boldsymbol{Y}^{(t\wedge\vartheta,k_2)}$ is defined in (3.142). Fix $\tilde{\epsilon}>0$. Using (3.161), we have

$$\begin{aligned}\Pr\left(|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_t|>\tilde{\epsilon}\right)&=\mathbb{E}\left[\mathbf{1}_{|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2}|>\tilde{\epsilon}}\right]\\&=\mathbb{E}\left[\mathbb{E}\big[\mathbf{1}_{|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2}|>\tilde{\epsilon}}\,|\,\mathcal{F}_{k_2-1}\big]\right]\\&=\mathbb{E}\left[\Pr\left(|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2}|>\tilde{\epsilon}\,|\,\mathcal{F}_{k_2-1}\right)\right]\\&=\mathbb{E}\left[\Pr\left(|\boldsymbol{Y}^{(t\wedge\vartheta,k_2)}-\mathbb{E}\left[\boldsymbol{Y}^{(t\wedge\vartheta,k_2)}\,|\,\mathcal{F}_{k_2-1}\right]|>\tilde{\epsilon}\right)\right].\end{aligned}\tag{3.162}$$

This means we can work with the quantity $\Pr\left(|\boldsymbol{Y}^{(t\wedge\vartheta,k_2)}-\mathbb{E}\left[\boldsymbol{Y}^{(t\wedge\vartheta,k_2)}\,|\,\mathcal{F}_{k_2-1}\right]|>\tilde{\epsilon}\right)$ which allows us to apply Proposition 21 [9]. In light of the syntax of Proposition 21 and for $1\leq k_2\leq t\wedge\vartheta$, we use $\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)}=\left(\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\right)_j\left(\boldsymbol{A}\boldsymbol{x}_{k_2-1}-\boldsymbol{b}\right)_j$ defined in (3.143), and define

$$b_{t\wedge\vartheta,k_2}\overset{\text{def}}{=}\max_{1\leq j\leq n}\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)},\quad \mu_{(t\wedge\vartheta,k_2)}\overset{\text{def}}{=}\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)},$$
$$\text{and}\qquad \sigma^2_{(t\wedge\vartheta,k_2)}\overset{\text{def}}{=}\frac{1}{n}\sum_{j=1}^{n}\left(\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)}-\mu_{(t\wedge\vartheta,k_2)}\right)^2.$$

By Lemma 14, we have the following upper bounds that hold with overwhelming probability

$$\left|b_{(t\wedge\vartheta,k_2)}\right|\leq C(T)\cdot n^{\alpha-\delta}\qquad\text{and}\qquad\sigma^2_{(t\wedge\vartheta,k_2)}\leq C(T)\cdot n^{2(\alpha-\delta)-1}\tag{3.163}$$

for some constant $C(T)$ depending on $||\Omega||, \|\boldsymbol{A}^T\boldsymbol{A}\|$, and T but independent of n and d. Applying Proposition 21 yields

$$\begin{aligned}\Pr\left(\left|\boldsymbol{Y}^{(t\wedge\vartheta,k_2)} - \mathbb{E}\left[\boldsymbol{Y}^{(t\wedge\vartheta,k_2)}|\mathcal{F}_{k_2-1}\right]\right| > \tilde{\epsilon}\right) &= \Pr\left(\left|\sum_{j\in B_{k_2-1}} \boldsymbol{X}_j^{(t\wedge\vartheta,k_2)} - \beta\mu_{t\wedge\vartheta,k_2}\right| \geq \tilde{\epsilon}\right) \\ &= \Pr\left(\left|\frac{1}{\beta}\sum_{j\in B_{k_2-1}} \boldsymbol{X}_j^{(t\wedge\vartheta,k_2)} - \mu_{t\wedge\vartheta,k_2}\right| \geq \tilde{\epsilon}/\beta\right) \\ &\leq 2\exp\left(-\frac{\beta(\frac{\tilde{\epsilon}}{\beta})^2}{2\sigma^2_{(t\wedge\vartheta,k_2)} + \frac{2}{3}(b_{(t\wedge\vartheta,k_2)} - a)(\frac{\tilde{\epsilon}}{\beta})}\right)\end{aligned} \tag{3.164}$$

Let $\tilde{\epsilon} := n^{\alpha-\delta+\eta}\cdot T^{-1}$ Using (3.163), we obtain

$$\begin{aligned}\frac{\beta(\frac{\tilde{\epsilon}}{\beta})^2}{2\sigma^2_{(t\wedge\vartheta,k_2)} + \frac{2}{3}(b_{(t\wedge\vartheta,k_2)} - a)(\frac{\tilde{\epsilon}}{\beta})} &\geq C(T)\frac{T^{-2}\cdot n^{-1}n^{2(\alpha-\delta+\eta)}}{n^{2(\alpha-\delta)-1} + T^{-1}\cdot n^{\alpha-\delta}n^{-1}n^{\alpha-\delta+\eta}} \\ &\geq C(T)\frac{1}{n^{-2\eta}+n^{-\eta}} \to \infty.\end{aligned}$$

Therefore it follows that

$$\Pr\left(|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2}| > \tilde{\epsilon}\right) \leq 2\exp(-C(T)n^c) \qquad \text{when } \tilde{\epsilon} = T^{-1}\cdot n^{\alpha-\delta+\eta} \tag{3.165}$$

for some constant $c > 0$ and $C(T)$ where the constants are independent of n and d and only depend on $||\Omega||, \|\boldsymbol{A}^T\boldsymbol{A}\|, \gamma$ and T. Here again $C(T)$ is a positive constant independent of n and d. Let $\epsilon = n^{\alpha-\delta+\eta}$. Using (3.160) and (3.165), and observing that $\boldsymbol{V}^{(t\wedge\vartheta,k_2)} = \boldsymbol{0}$ for

$k_2 > t \wedge \vartheta$ we get

$$\begin{aligned}
\Pr\left(\max_{0\le t\le T}\left|\mathcal{E}^{\nabla^2\text{-Off}}_{t\wedge\vartheta}\right| \ge \epsilon\right) &= \sum_{t=0}^{T}\Pr\left(\sum_{k_2=1}^{t\wedge\vartheta}|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2}| \ge \epsilon\right) \\
&\le \sum_{t=0}^{T}\sum_{k_2=1}^{T}\Pr\left(|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2}| \ge \frac{\epsilon}{T}\right) \\
&\le 2\sum_{t=0}^{T}\sum_{k_2=1}^{T}\exp\left(-C(T)n^c\right) = 2T^2\cdot\exp\left(-C(T)n^c\right).
\end{aligned} \tag{3.166}$$

This gets used our desired result.

3.8 Summary

In the first section, we provided the connection between mini-batch gradient-based methods and polynomials via Proposition 4. Namely, each mini-batch gradient-based method has an associated residual polynomial P_k, iteration polynomial Q_k, and noise polynomials $N_{k,j}$ which are defined recursively. Then we provided explicit polynomials for SGD, Polyak heavy-ball, and stochastic Nesterov acceleration (in the convex case) and their proofs. Then we provided Proposition 5 of the resolvent $R(z : \boldsymbol{A}\boldsymbol{A}^T)$ which allows us to bound the entries of the stopped loss value process $(\omega_k^{\vartheta} : k \ge 0)$ in our analysis of the main theorem.

In Section 3.4 we state and prove the main theorem under the assumption of small martingale errors $\mathcal{E}$. We show that the ℓ^2-regularized loss 3.1 under the iterates $\boldsymbol{x}_k$ of any mini-batch gradient-based method $f(\boldsymbol{x}_t)$ can be captured by a deterministic function $\Psi(t)$. Moreover, any general quadratic $\mathcal{R}(\cdot)$ satisfying Assumption 6, under mini-batch gradient-based iterates $\boldsymbol{x}_t$ can be captured by another deterministic quantity $\Omega(t)$ that depends on Ψ. Then we provide motivating examples that satisfy our assumptions - random features, training loss, and generalization errors. One caveat: Our analysis assumed $\delta = 0$ in 3.1, however, our analyis easily generalizes to the $\delta > 0$ case.

In the rest of the chapter, we demonstrate using concentration of measure results, that the martingale error terms $\mathcal{E}$ are indeed small. □

Chapter 4

Exact dynamics: Polyak heavy-ball

This chapter shows Theorem 3 applied to Polyak heavy-ball. In particular, we address the open problem of quantifying convergence rates of Polyak heavy-ball in terms of batch-size as mentioned in Subsection 2.3.1. Also, we will explore the algorithmic intuition behind the solution of the Volterra equation ψ (see (2.16) that captures the dynamics of Polyak heavy-ball. Moreover, we show how to select optimal hyperparameters momentum and stepsize. Lastly, we provide numerical experiments to illustrate these pheneomena. We will present the results and defer their proofs to the last section of this chapter.

4.1 Setting and assumptions

We consider the *unregularized* least squares problem when the number of samples (n) and features (d) are large:

$$\arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ f(\boldsymbol{x}) = \sum_{i=1}^{n} f_i(x) \stackrel{\text{def}}{=} \frac{1}{2}\sum_{i=1}^{n} (\boldsymbol{a}_i\boldsymbol{x} - b_i)^2 \right\}, \quad \text{with } \boldsymbol{b} \stackrel{\text{def}}{=} \boldsymbol{A}\widetilde{\boldsymbol{x}} + \boldsymbol{\eta}, \tag{4.1}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ is a data matrix whose i-th row is denoted by $\boldsymbol{a}_i \in \mathbb{R}^d$, $\widetilde{\boldsymbol{x}} \in \mathbb{R}^d$ is the signal vector, and $\boldsymbol{\eta} \in \mathbb{R}^n$ is a source of noise. The target $\boldsymbol{b} = \boldsymbol{A}\widetilde{\boldsymbol{x}} + \boldsymbol{\eta}$ comes from a generative model corrupted by noise. We let $\sigma_1^2 \geq \cdots \geq \sigma_n^2 \geq 0$ be the eigenvalues of the matrix $\boldsymbol{A}\boldsymbol{A}^T$

with $\sigma^2_{\max}$ and $\sigma^2_{\min}$ the largest and smallest (nonzero) eigenvalues, and $\boldsymbol{\eta} \in \mathbb{R}^n$ is a source of noise. The target $\boldsymbol{b} = \boldsymbol{A}\widetilde{\boldsymbol{x}} + \boldsymbol{\eta}$ comes from a generative model corrupted by noise. We let $\sigma_1^2 \geq \cdots \geq \sigma_n^2 \geq 0$ be the eigenvalues of the matrix $\boldsymbol{A}\boldsymbol{A}^T$ with $\sigma^2_{\max}$ and $\sigma^2_{\min}$ the largest and smallest (nonzero) eigenvalues.

$$\arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \Big\{ f(\boldsymbol{x}) = \sum_{i=1}^{n} f_i(x) \stackrel{\text{def}}{=} \frac{1}{2}\sum_{i=1}^{n} (\boldsymbol{a}_i\boldsymbol{x} - b_i)^2 \Big\}, \quad \text{with } \boldsymbol{b} \stackrel{\text{def}}{=} \boldsymbol{A}\widetilde{\boldsymbol{x}} + \boldsymbol{\eta}, \tag{4.2}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ is a data matrix whose i-th row is denoted by $\boldsymbol{a}_i \in \mathbb{R}^d$, $\widetilde{\boldsymbol{x}} \in \mathbb{R}^d$ is the signal vector, and $\boldsymbol{\eta} \in \mathbb{R}^n$ is a source of noise. The target $\boldsymbol{b} = \boldsymbol{A}\widetilde{\boldsymbol{x}} + \boldsymbol{\eta}$ comes from a generative model corrupted by noise. We let $\sigma_1^2 \geq \cdots \geq \sigma_n^2 \geq 0$ be the eigenvalues of the matrix $\boldsymbol{A}\boldsymbol{A}^T$ with $\sigma^2_{\max}$ and $\sigma^2_{\min}$ the largest and smallest (nonzero) eigenvalues.

We apply SGD with momentum (SGD+M) with mini-batches to the finite sum, quadratic problem (4.2). Let $\boldsymbol{x}_0 \in \mathbb{R}^d$ be randomly selected following Assumption 10 and $\boldsymbol{x}_1$ be generated from SGD without momentum, i.e., $\boldsymbol{x}_1 = \boldsymbol{x}_0 - \gamma \sum_{i\in B_0} \nabla f_i(\boldsymbol{x}_0)$. SGD+M iterates by selecting uniformly at random a subset $B_k \subseteq \{1, 2, \cdots, n\}$ of cardinality β and makes the update

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \gamma \sum_{i\in B_k} \nabla f_i(\boldsymbol{x}_k) + \Delta(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}) \\ &= \boldsymbol{x}_k - \gamma \boldsymbol{A}^T \boldsymbol{P}_k(\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}) + \Delta(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}), \quad \text{where} \quad \boldsymbol{P}_k \stackrel{\text{def}}{=} \sum_{i\in B_k} \mathbb{e}_i \mathbb{e}_i^T, \end{aligned} \tag{4.3}$$

with $\boldsymbol{P}_k$ a random orthogonal projection matrix and $\mathbb{e}_i$ the i-th standard basis vector. Here $\gamma > 0$ is the learning rate parameter, Δ is the momentum parameter, and the function f_i is the i-th element of the sum in (4.2).

When the stochastic gradient in (4.3) is replaced with the full-gradient $\nabla f(\boldsymbol{x})$, the resulting algorithm with learning rate and momentum optimally chosen yields the celebrated algorithm, heavy-ball momentum (a.k.a. Polyak momentum) [49]. The optimal

learning rate and momentum parameters are explicitly given by

$$\gamma = \frac{4}{(\sqrt{\sigma_{\max}^2} + \sqrt{\sigma_{\min}^2})^2} \quad \text{and} \quad \Delta = \left(\frac{\sqrt{\sigma_{\max}^2} - \sqrt{\sigma_{\min}^2}}{\sqrt{\sigma_{\max}^2} + \sqrt{\sigma_{\min}^2}} \right)^2. \tag{4.4}$$

It is well-known that heavy-ball is an optimal algorithm on the least squares problem in that it converges linearly at a rate of $\mathcal{O}(1/\sqrt{\kappa})$ (see [11,47]).

To perform our analysis we make the following explicit assumptions on the signal $\widetilde{\boldsymbol{x}}$, the noise $\boldsymbol{\eta}$, and the data matrix $\boldsymbol{A}$.

Assumption 10 (Initialization, signal, and noise)**.** *The initial vector $\boldsymbol{x}_0 \in \mathbb{R}^d$ is chosen so that $\boldsymbol{x}_0 - \widetilde{\boldsymbol{x}}$ is independent of the matrix $\boldsymbol{A}$. The noise vector $\boldsymbol{\eta} \in \mathbb{R}^n$ is centered and has i.i.d. entries, independent of $\boldsymbol{A}$. The signal and noise are normalized so that*

$$\mathbb{E}\|\boldsymbol{x}_0 - \widetilde{\boldsymbol{x}}\|_2^2 = R\frac{d}{n}, \quad \textit{and} \quad \mathbb{E}[\|\boldsymbol{\eta}\|_2^2] = \widetilde{R}.$$

Each row $\boldsymbol{a}_i \in \mathbb{R}^{d\times 1}$ is centered and is normalized so that $\mathbb{E}\|\boldsymbol{a}_i\|_2^2 = 1$ for all i. We suggest as a central example the *Gaussian random least squares* setup where each entry of $\boldsymbol{A}$ is sampled independently from a standard normal distribution with variance $\frac{1}{d}$.

4.2 Deterministic Equivalent of Polyak heavy-ball

With these assumptions, we can give an explicit representation of the loss values on a least squares problem at the iterates generated by SGD+M algorithm. We show in this section (see Theorem 7): for any $T > 0$,

$$\sup_{0\le t\le T} |f(\boldsymbol{x}_t) - \psi(t)| \to 0 \quad \text{in probability,}$$

where ψ solves (**??**). We begin by discussing the forcing and the noise terms of $\psi(t)$ and their relationship to SGD+M.

Forcing term: problem instance information. The forcing term represents the mean (with respect to expectation over the mini-batches) behavior of SGD+M. In fact, the forcing term is the loss f under full-batch gradient descent with momentum Δ but with learning rate $\gamma\zeta$. For a *small* learning rate γ, the forcing term $F(t)$ in (2.16) governs the dynamics of $\psi(t)$.

Let $\boldsymbol{w}_t \stackrel{\text{def}}{=} \boldsymbol{A}\boldsymbol{x}_t - \boldsymbol{b}$ and observe that $1/2\|\boldsymbol{w}_t\|_2^2$ is the loss $f(\boldsymbol{x}_t)$. One way to get the dynamics of f is by deriving a recurrence for $w_{t,j}^2$ (for each $j \in [n]$) which we get from the updates of SGD+M. The recurrence is as followed: define $\tilde{\mathcal{X}}_{t,j} \stackrel{\text{def}}{=} (w_{t,j}^2\ w_{t-1,j}^2\ w_{t,j}w_{t-1,j})^T$ and there exists a matrix $\boldsymbol{M}_j$ (see Subsection 4.5.3) so that

$$\tilde{\mathcal{X}}_{t+1,j} = \boldsymbol{M}_j\tilde{\mathcal{X}}_{t,j} + (\text{Error}). \tag{4.5}$$

The forcing term at iteration t is given by applying a linear recurrence $\boldsymbol{M}_j$ operator $t-1$ times on a vector containing initialization information at each $j \in [n]$ and then summed up for the first coordinate.

It is clear that the *maximum* of the eigenvalues of the operator $\boldsymbol{M}_j$ is essential to analyze the convergence behavior of $F(t)$. We denote $\lambda_{1,j} = \Delta$ and $\lambda_{i,j}$, $i = 2, 3$, the eigenvalues of $\boldsymbol{M}_j$ and note that $\Delta \leq \max_{i=2,3}|\lambda_{i,j}|$ (see (4.38) for an explicit formula of $\lambda_{i,j}$ that depends only on γ, Δ, and eigenvalues of $\boldsymbol{A}\boldsymbol{A}^T$). Let $\lambda_{2,j}$ be the eigenvalue of $\boldsymbol{M}_j$ with the biggest modulus and let

$$\lambda_{2,\max} \stackrel{\text{def}}{=} \max_j |\lambda_{2,j}|. \tag{4.6}$$

From this, an equation for the forcing term $F(t)$ can be made explicit in terms of $\lambda_{i,j}$, see Theorem 7 below and Subsection 4.5.1. We can conclude that $F(t) = \mathcal{O}(\lambda_{2,\max}^t)$.

Kernel term: noise from the algorithm. The convolution term in (**??**) is due to the inherent stochasticity of SGD+M. More specifically, it is given by

$$\begin{aligned}\gamma^2\zeta(1-\zeta)\sum_{k=0}^{t} H_2(t-k)\psi(k), \quad \text{where} \quad \Omega_j \stackrel{\text{def}}{=} 1-\gamma\zeta\sigma_j^2+\Delta,\\ \text{and} \quad H_2(t) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^{n}\frac{2\sigma_j^4}{\Omega_j^2-4\Delta}\Big(-\Delta^{t+1}+\frac{1}{2}\lambda_{2,j}^{t+1}+\frac{1}{2}\lambda_{3,j}^{t+1}\Big).\end{aligned} \tag{4.7}$$

The presence of ψ (training loss) is due to the fact that the noise generated by the k-th stochastic gradient is proportional to $\psi(k)$ (training loss), and the function $H_2(t-k)$ represents the progress of the algorithm in sending this extra noise to 0. Observe (4.7) scales quadratically in the learning rate γ. Hence for *large* learning rates, (4.7) dominates the decay behaviour of ψ. Further details discussed in Section 4.3.1.

We now state the main result:

Theorem 7 (Concentration of SGD+M)**.** *Suppose Assumptions 10 and hold with the learning rate $\gamma < \frac{1+\Delta}{\zeta\sigma_{max}^2}$ and the batch size satisfies $\beta/n = \zeta$ for some $\zeta > 0$. Let the constant $T \in \mathbb{N}$. Then there exists $C > 0$ such that for any $c > 0$, there exists $D > 0$ satisfying*

$$\Pr\left[\sup_{0\le t\le T, t\in\mathbb{N}} |f(\boldsymbol{x}_t)-\psi(t)| > n^{-C}\right] \le Dn^{-c}, \tag{4.8}$$

for sufficiently large $n \in \mathbb{N}$. The function ψ is the solution to the Volterra equation

$$\psi(t+1) = \underbrace{\frac{R}{2}h_1(t+1)+\frac{\widetilde{R}}{2}h_0(t+1)}_{\textit{forcing}} + \underbrace{\sum_{k=0}^{t}\gamma^2\zeta(1-\zeta)H_2(t-k)\psi(k)}_{\textit{noise}}, \quad \psi(0)=f(\boldsymbol{x}_0), \tag{4.9}$$

where for $k = 0, 1$, $i = 2, 3$, and $j \in [n]$, $\kappa_{i,j} \stackrel{\textit{def}}{=} \lambda_{i,j}\Omega_j/(\lambda_{i,j}+\Delta)$, and

$$h_k(t) = \frac{1}{n}\sum_{j=1}^{n}\frac{2(\sigma_j^2)^k}{\Omega_j^2-4\Delta}\left(-\Delta\gamma\zeta(\sigma_j^2)\cdot\Delta^t+\frac{1}{2}(\kappa_{2,j}-\Delta)^2\cdot(\lambda_{2,j})^t+\frac{1}{2}(\kappa_{3,j}-\Delta)^2\cdot(\lambda_{3,j})^t\right).$$

For a more detailed description on h_0, h_1, H_2 as well as the proof of Theorem 7 and Corollary 4, see Subsection 4.5.1. The expression of ψ highlights how the algorithm, learning rate, batch size, momentum, and noise levels interact with each other to produce different dynamics. Note that the learning rate assumption will be necessary for the solution to the Volterra equation to be convergent, see Proposition 12. When $\Delta \to 0$, we obtain the Volterra equation for SGD with mini-batching.

Corollary 4 (Concentration of SGD, no momentum)**.** *Under the same setting as Theorem 7 and when $\Delta = 0$, the function values $f(\boldsymbol{x}_t)$ converge to $\psi(t)$ as in* (4.8) *where now the limit ψ is a solution to the Volterra equation*

$$\psi(t+1) = \frac{R}{2}h_1(t+1) + \frac{\widetilde{R}}{2}h_0(t+1) + \sum_{k=0}^{t} \gamma^2\zeta(1-\zeta)h_2(t-k)\psi(k). \tag{4.10}$$

where for $k = 0, 1, 2$,

$$h_k(t) = \frac{1}{n}\sum_{j=1}^{n} \sigma_j^{2k}(1-\gamma\zeta\sigma_j^2)^{2t}.$$

Remark. Note that $H_2(t)$ reduces to $h_2(t)$ in $\Delta = 0$ case. Also when the limit $\zeta \to 0$ and when we scale time by t/ζ, we have that $(1-\gamma\zeta\sigma_j^2)^{2t/\zeta} \to e^{-2\gamma\sigma_j^2 t}$. This coincides with the result from [42, Theorem 1]. Indeed, this shows not only how our dynamics of SGD+M includes the no momentum case (i.e. SGD), but also how the dynamics of SGD+M differ from SGD.

4.3 Convolution Volterra analysis

In this section, we outline how to utilize the Volterra equation (4.9) to produce a complexity analysis of SGD+M. For additional details and proofs in this section, see Subsection 4.7.

We begin by establishing sufficient conditions for the convergence of the solution to the Volterra equation (4.9). Our Volterra equation can be seen as the *renewal equation* ([8]).

Let us translate (4.9) into the form of the renewal equation as follows:

$$\psi(t+1) = F(t+1) + (\tilde{K} * \psi)(t), \tag{4.11}$$

where $(f * g)(t) = \sum_{k=0}^{\infty} f(t-k)g(k)$. Let the *kernel norm* be $\|\tilde{K}\| = \sum_{t=0}^{\infty} \tilde{K}(t)$. By [8, Proposition 7.4], we see that $\|\tilde{K}\| < 1$ is necessary for our solution to the Volterra equation to be convergent. Indeed, we have the following result.

Proposition 11. *If the norm $\|\tilde{K}\| < 1$, the algorithm is convergent in that*

$$\psi(\infty) \overset{def}{=} \lim_{t\to\infty} \psi(t) = \frac{\frac{\widetilde{R}}{2}(\max\{1 - \frac{d}{n}, 0\})}{1 - \|\tilde{K}\|}. \tag{4.12}$$

Note that the noise factor $\widetilde{R}$ and the matrix dimension ratio d/n appear in the limit. Proposition 11 formulates the limit behaviour of the objective function in both the over-determined and the under-determined case of least squares. When under-determined, the ratio $d/n \geq 1$ and the limiting $\psi(\infty)$ is 0; otherwise the limit loss value is strictly positive. The result (4.12) only makes sense when the noise term K satisfies $||K|| < 1$; the next proposition illustrates the conditions on the learning rate and the trace of the eigenvalues of $\boldsymbol{A}\boldsymbol{A}^T$ such that the kernel norm is less than 1.

Proposition 12 (Convergence threshold)**.** *Under the learning rate condition $\gamma < \frac{1+\Delta}{\zeta\sigma_{max}^2}$ and trace condition $\frac{(1-\zeta)\gamma}{1-\Delta} \cdot \frac{1}{n} tr(\boldsymbol{A}\boldsymbol{A}^T) < 1$, the kernel norm $\|\tilde{K}\| < 1$, i.e., $\sum_{t=0}^{\infty} \tilde{K}(t) < 1$.*

The *learning rate condition* quantifies an upper bound of good learning rates by the largest eigenvalue of the covariance matrix $\sigma_{\max}^2$, batch size ζ, and the momentum parameter Δ. The *trace condition* illustrates a constraint on the growth of $\sigma_{\max}^2$. Moreover, for a full batch gradient descent model ($\zeta = 1$), the trace condition can be dropped and we get the classical learning rate condition for gradient descent.

4.3.1 The Malthusian exponent and complexity

The rate of convergence of ψ is essentially the worse of two terms – the forcing term $F(t)$ and a discrete time convolution $\sum_{k=0}^{t} \psi(k)K(t-k)$ which depends on the kernel K. Intuitively, the forcing term captures the behavior of the expected value of SGD+M and the discrete time convolution captures the slowdown in training due to noise created by the algorithm. Note that $F(t)$ is always a lower bound for $\psi(t)$, but it can be that $\psi(t)$ is exponentially (in t) larger than $F(t)$ owing to the convolution term. This occurs when something called the *Malthusian exponent*, denoted Ξ, of the convolution Volterra equation exists. The Malthusian exponent Ξ is given as the unique solution to

$$\gamma^2\zeta(1-\zeta)\sum_{t=0}^{\infty}\Xi^t H_2(t) = 1, \qquad \text{if the solution exists.} \tag{4.13}$$

The Malthusian exponent enters into the complexity analysis in the following way:

Theorem 8 (Asymptotic rates)**.** *The inverse of the Malthusian exponent always satisfies* $\Xi^{-1} > \lambda_{2,\max}$ *for finite* n*. Moreover, for some* $C > 0$*, the convergence rate for SGD+M is*

$$\psi(t) - \psi(\infty) \le C\max\{\lambda_{2,\max}, \Xi^{-1}\}^t \quad \textit{and} \lim_{t\to\infty}(\psi(t)-\psi(\infty))^{1/t} = \max\{\lambda_{2,\max}, \Xi^{-1}\}. \tag{4.14}$$

Thus to understand the rates of convergence, it is necessary to understand the Malthusian exponent as a function of γ and Δ.

4.3.2 Two regimes for the Malthusian exponent

On the one hand, the Malthusian exponent Ξ comes from the stochasticity of the algorithm itself. On the other hand, $\lambda_{2,\max}(\gamma, \Delta, \zeta)$ is determined completely by the problem instance information — the eigenspectrum of $\boldsymbol{A}\boldsymbol{A}^T$. (Note we want to emphasize the dependence of $\lambda_{2,\max}$ on learning rate, momentum, and batch fraction.) Let $\sigma^2_{\max}$ and $\sigma^2_{\min}$ denote the maximum and minimum *nonzero* eigenvalues of $\boldsymbol{A}\boldsymbol{A}^T$, respectively. For a fixed

batch size, the optimal parameters $(\gamma_\lambda, \Delta_\lambda)$ of $\lambda_{2,\max}$ are

$$\gamma_\lambda = \frac{1}{\zeta}\left(\frac{2}{\sqrt{\sigma^2_{\max}} + \sqrt{\sigma^2_{\min}}}\right)^2 \quad \text{and} \quad \Delta_\lambda = \left(\frac{\sqrt{\sigma^2_{\max}} - \sqrt{\sigma^2_{\min}}}{\sqrt{\sigma^2_{\max}} + \sqrt{\sigma^2_{\min}}}\right)^2. \tag{4.15}$$

In the full batch setting, i.e. $\zeta = 1$, these optimal parameters γ_λ and Δ_λ for $\lambda_{2,\max}$ are exactly the Polyak momentum parameters (4.4). Moreover, in this setting, there is no stochasticity so the Malthusian exponent disappears and the convergence rate (4.14) is $\lambda_{2,\max}$. We observe from (4.15) that for all fixed batch sizes, the optimal momentum parameter, Δ_λ, is independent of batch size. The only dependence on batch size appears in the learning rate. At first it appears that for small batch fractions, one can take large learning rates, but in that case, the inverse of the Malthusian exponent Ξ^{-1} dominates the convergence rate of SGD+M (4.14) and you cannot take γ and Δ to be as in (4.15) (See Figure 4.1).

We will define two subsets of parameter space, the *problem constrained regime* and the *algorithmically constrained regime* (or stochastically constrained regime). The problem constrained regime is for some tolerance $\varepsilon > 0$

$$\text{problem constrained regime} \quad \stackrel{\text{def}}{=} \{(\gamma, \Delta) \, : 1 - \sqrt{\Xi} < (1 - \sqrt{\lambda^{-1}_{2,\max}})(1 - \varepsilon)\}. \tag{4.16}$$

The remainder we call the *algorithmically constrained* regime. To explain the tolerance: for finite n, it transpires that we always have $\Xi^{-1} > \lambda_{2,\max}$, but it could be vanishingly close to $\lambda_{2,\max}$ as a function of n. Hence we introduce the tolerance to give the correct qualitative behavior in finite n.

Proposition 13. *If the learning rate $\gamma \leq \min(\frac{1+\Delta}{\zeta\sigma^2_{max}}, \frac{(1-\sqrt{\Delta})^2}{\zeta\sigma^2_{\min}})$, with the trace condition $\frac{8(1-\zeta)\gamma}{1-\Delta} \cdot \frac{1}{n} tr(\boldsymbol{A}^T\boldsymbol{A}) < 1$, then (γ, Δ) is in the problem constrained regime with $\varepsilon = 1/2$.*

Therefore by (4.14), we have that

$$\psi(t) - \psi(\infty) \leq D\left(\frac{4\lambda_{2,\max}}{(1 + \sqrt{\lambda_{2,\max}})^2}\right)^t \quad \text{for some } D > 0; \tag{4.17}$$

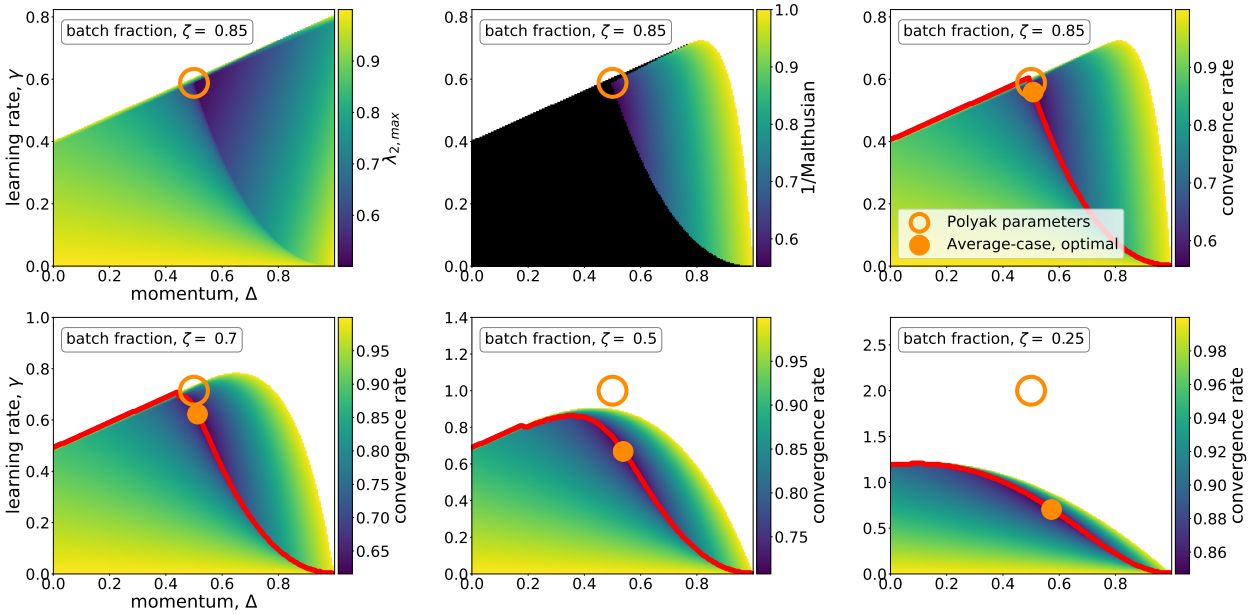

Figure 4.1: Different convergence rate regions: problem constrained regime versus algorithmically constrained regime for Gaussian random least squares problem with $(n = 2000 \times d = 1000)$. Plots are functions of momentum (x-axis) and learning rate (y-axis). Analytic expression for $\lambda_{2,\max}$ (see (4.6), (4.38)) – convergence rate of forcing term $F(t)$ – given in (top row, column 1) represents the problem constrained region. (top row, column 2) plots 1/(Malthusian exponent) ((4.13), for details see Subsection 4.8); black region is where the Malthusian exponent Ξ does not exist. This represents the algorithmically constrained region. Finally, (top row, column 3 and bottom row) plots convergence rate of SGD+M $= \max\{\lambda_{2,\max}, \Xi^{-1}\}$, (see (4.14)), for various batch fractions. When the Malthusian exponent does not exist (black), $\lambda_{2,\max}$ takes over the convergence rate of SGD+M; otherwise the noise in the algorithm (i.e. Malthusian exponent Ξ) dominates. Optimal parameters that maximize $\lambda_{2,\max}$ denoted by Polyak parameters (orange circle, (4.15)) and the optimal parameters for SGD+M (orange dot); below red line is the problem constrained region; otherwise the algorithmic constrained region. When batch fractions $\zeta = 0.85$ and $\zeta = 0.7$ (top row and bottom row, column 1) (i.e., large batch), the SGD+M convergence rate is the deterministic momentum rate of $1/\sqrt{\kappa}$. As the batch fraction decreases ($\zeta = 0.25$), the convergence rate becomes that of SGD and the optimal parameters of SGD+M and Polyak parameters are quite far from each other. The Malthusian exponent (algorithmically constrained region) starts to control the SGD+M rate as batch fraction $\to 0$.

we note that the expression in the parenthesis is $1 - \frac{1}{2}(1 - \lambda_{2,\max}) + \mathcal{O}((1 - \lambda_{2,\max})^2)$.

In the problem constrained regime, it is worthwhile to note that the overall convergence rate is the same as full batch momentum with adjusted learning rate, i.e., the batch size does not play an important role as long as we are in the problem constrained regime:

Proposition 14 (Concentration of SGD + M, full batch). *Suppose $\zeta = 1$ and Assumptions 10 hold with the learning rate $\gamma < \frac{1+\Delta}{\sigma_{max}^2}$. If we let $\boldsymbol{x}_t^{full}$ denote the iterates of full-batch gradient descent with momentum (GD+M), then*

$$\sup_{0 \le t \le T} \left| f\left(\boldsymbol{x}_t^{full}\right) - \psi_{full}(t) \right| \xrightarrow[n \to \infty]{\Pr} 0, \quad \textit{where} \quad \psi_{full}(t+1) = \frac{R}{2} h_1(t+1) + \frac{\tilde{R}}{2} h_0(t+1). \tag{4.18}$$

The functions h_1 and h_0 are defined in Theorem 7 with $\zeta = 1$.

In particular, let γ_{full} denote the learning rate for full batch GD+M, and $\gamma, \zeta < 1$ for the learning rate and batch fraction in SGD+M with corresponding ψ in Theorem 7. Then when $\gamma_{full} = \gamma\zeta$ is satisfied, ψ and ψ_{full} share the same convergence rate in the problem constrained regime.

4.4 Performance of SGD+M: implicit conditioning ratio (ICR)

We define the average condition number, condition number, and the implicit conditioning ratio to be

$$\bar{\kappa} \overset{\text{def}}{=} \frac{\frac{1}{n}\sum_{j \in [n]} \sigma_j^2}{\sigma_{\min}^2} < \frac{\sigma_{\max}^2}{\sigma_{\min}^2} \overset{\text{def}}{=} \kappa \qquad \text{and} \qquad \text{ICR} \overset{\text{def}}{=} \frac{\bar{\kappa}}{\sqrt{\kappa}}, \tag{4.19}$$

respectively.

Large vs. small batch regime. We refer to the *large batch* regime where $\zeta \ge$ ICR and the *small batch* regime where $\zeta \le$ ICR.

We begin by giving a rate guarantee that holds in the problem constrained regime, for a specific choice of γ and Δ.

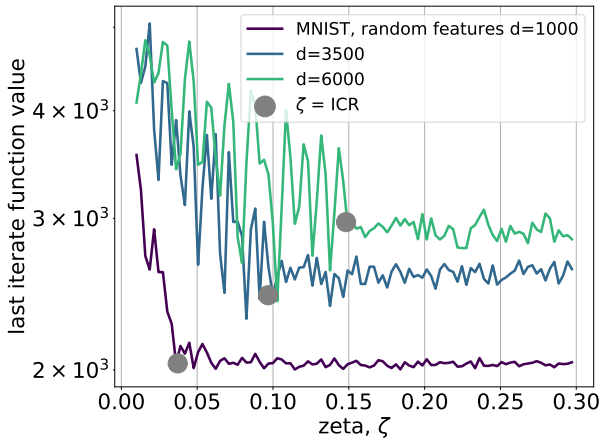

Figure 4.2: For each value of the batch fraction, ζ, we run SGD+M for 50 iterations on (normalized) MNIST data set using a random features set-up with Gaussian weight matrix $\boldsymbol{W} \in \mathbb{R}^{784\times d}$ (see App. 4.8 for details) and targets odd/even. We record the function value of the last iterate. The momentum and learning rate parameters are set to be near-optimal (4.20). Gray dot is the computed ICR value. At the predicted $\zeta = \text{ICR}$ (gray dot), there is a change in the behavior of the last iterate. For $\zeta \leq \text{ICR}$, the value of the last iterate monotonically decreases until it hits the ICR. For $\zeta \geq \text{ICR}$, we see no improvement in the value of the last iterate. This agrees with the theory that the convergence rate does not change.

Proposition 15 (Good momentum parameters)**.** *Suppose the learning rate and momentum satisfy*

$$\gamma = \frac{(1-\sqrt{\Delta})^2}{\zeta\sigma_{\min}^2} \text{ and } \Delta = \max\left\{\left(\frac{1-\frac{\mathcal{C}}{\bar{\kappa}}}{1+\frac{\mathcal{C}}{\bar{\kappa}}}\right), \left(\frac{1-\frac{1}{\sqrt{2\kappa}}}{1+\frac{1}{\sqrt{2\kappa}}}\right)\right\}^2, \text{ where } \mathcal{C} \stackrel{def}{=} \zeta/(8(1-\zeta)). \quad (4.20)$$

Then $\lambda_{2,\max} = \Delta$ and for some $C > 0$, the convergence rate for SGD+M is

$$\psi(t) - \psi(\infty) \leq C \cdot \Delta^t = C \cdot \max\left\{\left(\frac{1-\frac{\mathcal{C}}{\bar{\kappa}}}{1+\frac{\mathcal{C}}{\bar{\kappa}}}\right), \left(\frac{1-\frac{1}{\sqrt{2\kappa}}}{1+\frac{1}{\sqrt{2\kappa}}}\right)\right\}^{2t}. \quad (4.21)$$

Remark 6. *We note that for all Δ satisfying $\frac{(1-\sqrt{\Delta})^2}{\zeta\sigma_{\min}^2} \leq \frac{(1+\sqrt{\Delta})^2}{2\zeta\sigma_{\max}^2}$ with the learning rate γ as in* (4.20)*, we have that $\lambda_{2,\max} = \Delta$. By minimizing the Δ (i.e., by finding the fastest convergence rate), we get the formula for the momentum parameter in* (4.20)*.*

The exact tradeoff in convergence rates (4.21) occurs when

$$\frac{\mathcal{C}}{\bar{\kappa}} = \frac{1}{\sqrt{2\kappa}}, \quad \text{or} \quad \zeta = \frac{\frac{8}{\sqrt{2}}\text{ICR}}{1 + \frac{8}{\sqrt{2}}\text{ICR}}. \tag{4.22}$$

As $\zeta \leq 1,$ this condition is only nontrivial when ICR $\ll 1$, in which case $\zeta = \frac{8}{\sqrt{2}}$ICR, up to vanishing errors. This is illustrated on the MNIST data set in Figure 4.2.

Large batch ($\zeta \geq$ ICR). In this regime SGD+M's performance matches the performance of the heavy-ball algorithm with the Polyak momentum parameters (up to absolute constants). More specifically with the choices of γ and Δ in Proposition 15, the linear rate of convergence of SGD+M is $1 - \frac{c}{\sqrt{\kappa}}$ for an absolute c. Note that ζ does not appear in the rate, and in particular there is no gain in convergence rate by increasing the batch fraction.

Small batch ($\zeta \leq$ ICR). In the small batch regime, the value of $\mathcal{C}$ is relatively small and the first term is dominant in (4.21), and so the linear rate of convergence of SGD+M is $1 - \frac{c\zeta}{\bar{\kappa}}$ for some absolute constant $c > 0$. In this regime, there is a benefit in increasing the batch fraction, and the rate increases linearly with the fraction. We note that on expanding the choice of constants in small ζ the choices made in Proposition 15 are

$$\Delta \approx 1 - \frac{\zeta}{8\bar{\kappa}} \quad \text{and} \quad \gamma \approx \frac{\zeta}{256\bar{\kappa}^2\sigma_{\min}^2}.$$

This rate can also be achieved by taking $\Delta = 0$, i.e. mini-batch SGD with no momentum. Moreover, it is not possible to beat this by using momentum; we show the following lower bound:

Proposition 16. *If $\zeta \leq \min\{\frac{1}{2}, ICR\}$ then there is an absolute constant $C > 0$ so that for convergent (γ, Δ) (those satisfying Proposition 12), $\sqrt{\lambda_{2,\max}} \geq 1 - \frac{C\zeta}{\bar{\kappa}}$.*

This is a lower bound on the rate of convergence by Theorem 8.

A parallel argument of Proposition 15 holds for SGD without momentum.

Proposition 17 (Good learning rate for SGD without momentum)**.** *Suppose the learning rate γ satisfies*

$$\gamma = \min\left\{\frac{1}{\zeta\sigma_{\max}^2}, \frac{1}{8(1-\zeta)\cdot\frac{1}{n}tr(\boldsymbol{A}^T\boldsymbol{A})}\right\}. \tag{4.23}$$

Then $\lambda_{2,\max} = (1-\gamma\zeta\sigma_{\min})^2$ and for some $C > 0$, the convergence rate for SGD without momentum is

$$\psi(t) - \psi(\infty) \leq C\cdot\lambda_{2,\max}^t = C\cdot\max\left\{1-\frac{\mathcal{C}}{\bar{\kappa}}, 1-\frac{1}{\kappa}\right\}^{2t}, \tag{4.24}$$

where $\mathcal{C} \overset{def}{=} \zeta/(8(1-\zeta))$.

For details on the proof of Proposition 17, see Subsection 4.7. The exact tradeoff in convergence rates occurs when

$$\frac{1}{\zeta\sigma_{\max}^2} = \frac{1}{8(1-\zeta)\cdot\frac{1}{n}\text{tr}(\boldsymbol{A}^T\boldsymbol{A})}, \quad \text{or} \quad \zeta = \frac{8\times\widehat{\text{ICR}}}{1+8\times\widehat{\text{ICR}}}, \quad \text{where} \quad \widehat{\text{ICR}} \overset{\text{def}}{=} \frac{\bar{\kappa}}{\kappa}. \tag{4.25}$$

As $\zeta \leq 1$, this condition is only nontrivial when $\widehat{\text{ICR}} \ll 1$, in which case $\zeta = 8\times\widehat{\text{ICR}}$, up to vanishing errors. When we are in the large batch setting ($\zeta \gtrsim \widehat{\text{ICR}}$), the linear rate of convergence of SGD is $1-\frac{c\zeta}{\bar{\kappa}}$ for some absolute constant $c > 0$.

On the other hand, when in the small batch regime, i.e., $\zeta \lesssim \widehat{\text{ICR}}$, the convergence rate is fixed by $1-\frac{1}{\kappa}$. Note that ζ does not appear in the convergence rate, so there is no loss in convergencec rate by decreasing the batch fraction.

As a result, SGD converges more slowly in the large batch setting than in the small batch setting (see [27]).

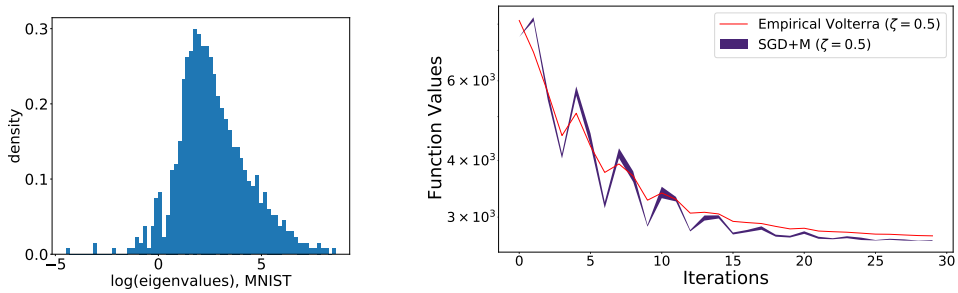

Figure 4.3: SGD+M vs. Theory on even/odd MNIST. MNIST ($60,000 \times 28 \times 28$ images) [31] is reshaped into a single matrix of dimension $60,000 \times 784$ (preconditioned to have centered rows of norm-1), representing 60,000 samples of 10 digits. The target $\boldsymbol{b}$ satisfies $b_i = 0.5$ if the i^{th} sample is an odd digit and $b_i = -0.5$ otherwise. SGD+M was run 10 times with $(\Delta = 0.8, \gamma = 0.001, \zeta = 0.5)$ and the empirical Volterra was run once with $(R = 11,000, \tilde{R} = 5300)$. The 10^{th} to 90^{th} percentile interval is displayed for the loss values of 10 runs of SGD+M. While MNIST data set does not satisfy our eigenvalue assumption on the data matrix, the solution to the Volterra equation on MNIST data set captures the dynamics of SGD+M. See Subsection 4.8 for more details.

4.5 Proofs of results

This section is organized into 4 subsections as follows:

1. Subsection 4.5.1 derives the Volterra equation and proves the main concentration for the dynamics of SGD+M (Theorem 7).

2. We show in Subsection 4.6.1 that the error terms associated with concentration of measure on the high-dimensional orthogonal group disappear in the large-n limit.

3. Subsection 4.7 derives main results including Proposition 13 and speed up of convergence rate of SGD+M (Proposition 15) in the large batch regime, as well as the lower bound convergence rate in the small batch regime (Proposition 16). We also provide a proof of Proposition 17 in this section.

4. Subsection 4.8 contains details on the numerical simulations.

Remark 7. *The analysis is this section assumes left orthogonal invariance on the data. However, this assumption is unnecessary and can be circumvented by proving the results in the manner similar to the proof of Theorem 3. In fact, numerical simulations support the hypothesis that* (4.26) *can be weakened and that the theory herein can be applied to other ensembles without this orthogonal invariance property. See Figure 4.3.*

Assumption 11 (Orthogonal invariance)**.** *Let $\boldsymbol{A}$ be a random $n \times d$ matrix. Suppose these random matrices satisfy a left orthogonal invariance condition: Let $\boldsymbol{O} \in \mathbb{R}^{n\times n}$ be an orthogonal matrix. Then the matrix $\boldsymbol{A}$ is orthogonally left invariant in the sense that*

$$\boldsymbol{O}\boldsymbol{A} \overset{law}{=} \boldsymbol{A}. \tag{4.26}$$

This assumption implies that the left singular vectors of A are uniformly distributed on the sphere which is the strongest form of eigenvector delocalization; many distributions of random matrices including some sparse ones (such as random regular graph adjacency matrices) are known to have some form of eigenvector delocalization. The classic example of a random matrix which has left orthogonal invariance is the sample covariance matrix, $\boldsymbol{Z}\sqrt{\boldsymbol{\Sigma}}$, for an i.i.d. Gaussian matrix $\boldsymbol{Z}$ and any covariance matrix $\boldsymbol{\Sigma}$.

4.5.1 Derivation of the dynamics of Polyak momentum (SGD+M)

In this section, we establish the fundamental of the proof of Theorem 7. Let us state the theorem in full detail first.

Theorem 9 (Theorem 7, detailed version)**.** *Suppose Assumptions 10 and 11 hold with the learning rate $\gamma < \frac{1+\Delta}{\zeta\sigma_{max}^2}$ and the batch size satisfies $\beta/n = \zeta$ for some $\zeta > 0$. Let the constant $T \in \mathbb{N}$. Then there exists $C > 0$ such that for any $c > 0$, there exists $D > 0$ satisfying*

$$\Pr\left[\sup_{0\le t\le T, t\in\mathbb{N}} |f(\boldsymbol{x}_t) - \psi(t)| > n^{-C}\right] \le Dn^{-c}, \tag{4.27}$$

for sufficiently large $n \in \mathbb{N}$. $\boldsymbol{x}_t$ are the iterates of SGD+M and the function ψ is the solution to the Volterra equation

$$\psi(t+1) = \underbrace{\frac{R}{2}h_1(t+1) + \frac{\tilde{R}}{2}h_0(t+1)}_{\textit{forcing}} + \underbrace{\sum_{k=0}^{t} \gamma^2\zeta(1-\zeta)H_2(t-k)\psi(k)}_{\textit{noise}}, \textit{ and } \psi(0) = f(\boldsymbol{x}_0), \tag{4.28}$$

where for $k = 0, 1$,

$$h_k(t) = \frac{1}{n}\sum_{j=1}^{n} \frac{2(\sigma_j^2)^k}{\Omega_j^2 - 4\Delta}\left(-\Delta\gamma\zeta(\sigma_j^2)\cdot\Delta^t + \frac{1}{2}(\kappa_{2,j}-\Delta)^2\cdot(\lambda_{2,j})^t + \frac{1}{2}(\kappa_{3,j}-\Delta)^2\cdot(\lambda_{3,j})^t\right),$$

and

$$H_2(t) = \frac{1}{n}\sum_{j=1}^{n} \frac{2\sigma_j^4}{\Omega_j^2 - 4\Delta}\Big(-\Delta^{t+1} + \frac{1}{2}\lambda_{2,j}^{t+1} + \frac{1}{2}\lambda_{3,j}^{t+1}\Big).$$

Here $\Omega_j, \lambda_{2,j}, \lambda_{3,j}, \kappa_{2,j}, \kappa_{3,j}, j \in [n]$ are defined as

$$\Omega_j = 1 - \gamma\zeta\sigma_j^2 + \Delta,\ \kappa_{2,j} = \frac{\lambda_{2,j}\Omega_j}{\lambda_{2,j}+\Delta},\ \kappa_{3,j} = \frac{\lambda_{3,j}\Omega_j}{\lambda_{3,j}+\Delta},\ \textit{and}$$
$$\lambda_{2,j} = \frac{-2\Delta + \Omega_j^2 + \sqrt{\Omega_j^2(\Omega_j^2-4\Delta)}}{2},\ \lambda_{3,j} = \frac{-2\Delta + \Omega_j^2 - \sqrt{\Omega_j^2(\Omega_j^2-4\Delta)}}{2}.$$

4.5.2 Change of basis

Consider the singular value decomposition of $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices, i.e. $\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{V}^T\boldsymbol{V} = \mathbf{I}$ and $\boldsymbol{\Sigma}$ is the $n \times d$ singular value matrix with diagonal entries $\mathbf{diag}(\sigma_j), j = 1, \ldots, n$ (in the case $n > d$, we extend the set of singular values so that $\sigma_{d+1} = \cdots = \sigma_n = 0$). We define the spectral weight vector $\boldsymbol{\nu}_k \overset{\text{def}}{=} \boldsymbol{V}^T(\boldsymbol{x}_k - \widetilde{\boldsymbol{x}})$, which therefore evolves like

$$\boldsymbol{\nu}_{k+1} = \boldsymbol{\nu}_k - \gamma\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{P}_k(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_k - \boldsymbol{\eta}) + \Delta(\boldsymbol{\nu}_k - \boldsymbol{\nu}_{k-1}). \tag{4.29}$$

Moreover, we can define

$$\boldsymbol{w}_k := \boldsymbol{\Sigma}\boldsymbol{\nu}_k - \boldsymbol{U}^T\boldsymbol{\eta}, \tag{4.30}$$

so that

$$f(\boldsymbol{x}_t) = \frac{1}{2}\|\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{U}^T\boldsymbol{\eta}\|_2^2 = \frac{1}{2}\sum_{j=1}^{n} \boldsymbol{w}_{t,j}^2. \tag{4.31}$$

Then (4.29) can be translated as

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \gamma\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{P}_k\boldsymbol{U}\boldsymbol{w}_k + \Delta(\boldsymbol{w}_k - \boldsymbol{w}_{k-1}). \tag{4.32}$$

From this point, we focus on the evolution of $\boldsymbol{w}$ rather than the iterates $\boldsymbol{x}$.

4.5.3 Evolution of f

Now we would like to demonstrate the recurrence relation of $\boldsymbol{w}_k$ and eventually that of $f(t)$, which will lead to a Volterra equation and error terms in a large scale. First, for $j \in [n]$ and $t \in \mathbb{N}$, (4.32) implies that

$$w_{t+1,j} = w_{t,j} - \gamma\sigma_j^2 \sum_l w_{t,l}(\sum_{i\in B_t} U_{ij}U_{il}) + \Delta(w_{t,j} - w_{t-1,j}), \tag{4.33}$$

where $B_t = B$ denotes a randomly chosen mini-batch at the t-th iteration, whose size is given by $\beta \leq n$. We interchangeably use the notation of B_t and B, because it is independently chosen at each iteration. By taking squares on both sides, we have

$$\begin{aligned}
w_{t+1,j}^2 &= \left(w_{t,j} - \gamma\sigma_j^2 \sum_{l\in[n]} w_{t,l}(\sum_{i\in B_t} U_{ij}U_{il}) + \Delta(w_{t,j} - w_{t-1,j}) \right)^2 \\
&= w_{t,j}^2 + \gamma^2\sigma_j^4\big(\sum_{l\in[n]} w_{t,l}(\sum_{i\in B_t} U_{ij}U_{il})\big)^2 - 2\gamma\sigma_j^2 w_{t,j}\sum_{l\in[n]} w_{t,l}(\sum_{i\in B_t} U_{ij}U_{il}) \\
&\quad + \Delta^2(w_{t,j} - w_{t-1,j})^2 + 2\Delta w_{t,j}(w_{t,j} - w_{t-1,j}) \\
&\quad - 2\gamma\sigma_j^2\Delta\sum_{l\in[n]} w_{t,l}(\sum_{i\in B_t} U_{ij}U_{il})(w_{t,j} - w_{t-1,j}).
\end{aligned}$$

Now let us denote the following error caused by mini-batching, i.e.,

$$\mathbb{E}_B^{(l,j)} \stackrel{\text{def}}{=} \sum_{i\in B} U_{il}U_{ij} - \frac{\beta}{n}\delta_{l,j}. \tag{4.34}$$

where $\delta_{l,j}$ is the Kronecker-delta symbol, meaning

$$\text{For } l,j \in [n],\ \delta_{l,j} = 1 \quad \text{if } l = j, \text{ and } 0 \text{ otherwise.}$$

Then the iteration on w_{t+1}^2 reduces to

$$\begin{aligned}
w_{t+1,j}^2 &= w_{t,j}^2(1+\Delta^2+2\Delta) + w_{t-1,j}^2\Delta^2 + w_{t,j}w_{t-1,j}(-2\Delta^2-2\Delta) \\
&\quad - 2\gamma\sigma_j^2 w_{t,j}\sum_{l\in[n]} w_{t,l}(\mathbb{E}_B^{(l,j)} + \frac{\beta}{n}\delta_{l,j}) \\
&\quad - 2\gamma\sigma_j^2\Delta\sum_{l\in[n]} w_{t,l}(w_{t,j}-w_{t-1,j})(\mathbb{E}_B^{(l,j)} + \frac{\beta}{n}\delta_{l,j}) + \gamma^2\sigma_j^4\big(\sum_{l\in[n]} w_{t,l}(\sum_{i\in B} U_{ij}U_{il})\big)^2 \\
&= w_{t,j}^2(1+\Delta^2+2\Delta-2\gamma\sigma_j^2\frac{\beta}{n}-2\Delta\gamma\sigma_j^2\frac{\beta}{n}) + w_{t-1,j}^2\Delta^2 \\
&\quad + w_{t,j}w_{t-1,j}(-2\Delta^2-2\Delta+2\Delta\gamma\sigma_j^2\frac{\beta}{n}) \\
&\quad + \underbrace{\gamma^2\sigma_j^4\big(\sum_{l\in[n]} w_{t,l}(\sum_{i\in B} U_{ij}U_{il})\big)^2}_{\stackrel{\text{def}}{=}①} + \underbrace{\left(-2\gamma\sigma_j^2 w_{t,j}\sum_{l\in[n]}\mathbb{E}_B^{(l,j)} w_{t,l}\right)}_{\stackrel{\text{def}}{=}\mathbb{E}_{B,1}^{(j)}(t)} \\
&\quad + \underbrace{\left(-2\gamma\sigma_j^2\Delta\sum_{l\in[n]}\mathbb{E}_B^{(l,j)} w_{t,l}(w_{t,j}-w_{t-1,j})\right)}_{\stackrel{\text{def}}{=}\mathbb{E}_{B,2}^{(j)}(t)}.
\end{aligned}$$

When it comes to ①, we can decompose it into its expectation over the mini-batch B and the error generated by it. By applying the technique from [43, Lemma 8], we have

$$\begin{aligned}\mathbb{E}[①|\mathcal{F}_t] &= \gamma^2\sigma_j^4\left[\frac{\beta(\beta-1)}{n(n-1)}w_{t,j}^2+\left(\frac{\beta}{n}-\frac{\beta(\beta-1)}{n(n-1)}\right)\sum_{i\in[n]}U_{ij}^2\left(\sum_{l\in[n]}U_{il}w_{t,l}\right)^2\right]\\ &= \Gamma_j^2 w_{t,j}^2+\frac{(1-\zeta)\gamma\sigma_j^2\Gamma_j}{n}\sum_{l\in[n]}w_{t,l}^2+\mathbb{E}_{KL}^{(j)}(t)+\mathbb{E}_{beta}^{(j)}(t),\end{aligned}$$

where

$$\Gamma_j \stackrel{\text{def}}{=} \gamma\zeta\sigma_j^2,$$

$$\mathbb{E}_{beta}^{(j)}(t) \stackrel{\text{def}}{=} \gamma^2\sigma_j^4\left[\left(\frac{\beta(\beta-1)}{n(n-1)}-\zeta^2\right)w_{t,j}^2+\left(-\frac{\beta(\beta-1)}{n(n-1)}+\zeta^2\right)\sum_{i\in[n]}U_{ij}^2\left(\sum_{l\in[n]}U_{il}w_{t,l}\right)^2\right],$$

$$\text{and}\quad \mathbb{E}_{KL}^{(j)}(t) \stackrel{\text{def}}{=} \gamma^2\sigma_j^4(\zeta-\zeta^2)\sum_{i\in[n]}(U_{ij}^2-\frac{1}{n})\left(\sum_{l}U_{il}w_{t,l}\right)^2.$$

Note that $\mathbb{E}_{beta}^{(j)}(t)$ is generated by the error between $\beta(\beta-1)/(n(n-1))$ and $\zeta^2=\beta^2/n^2$, whereas $\mathbb{E}_{KL}^{(j)}(t)$ is generated by the replacement of U_{ij}^2 by $1/n$; In Subsection 4.6, we establish that this error can be bounded by the *Key Lemma* (this is where the acronym "KL" comes from). Let $\mathbb{E}_{B^2}^{(j)}(t) \stackrel{\text{def}}{=} ① - \mathbb{E}[①|\mathcal{F}_t]$.

Then observe

$$① = \Gamma_j^2 w_{t,j}^2+\frac{(1-\zeta)\gamma\sigma_j^2\Gamma_j}{n}\sum_{l\in[n]}w_{t,l}^2+\mathbb{E}_{B^2}^{(j)}(t)+\mathbb{E}_{beta}^{(j)}(t)+\mathbb{E}_{KL}^{(j)}(t).$$

Therefore, we obtain

$$\begin{aligned}w_{t+1,j}^2 = \Omega_j^2 w_{t,j}^2+\Delta^2 w_{t-1,j}^2-2\Delta\Omega_j w_{t,j}w_{t-1,j}+\frac{(1-\zeta)\gamma\sigma_j^2\Gamma_j}{n}\sum_{l\in[n]}w_{t,l}^2\\ +\mathbb{E}_{beta}^{(j)}(t)+\mathbb{E}_{KL}^{(j)}(t)+\mathbb{E}_{B}^{(j)}(t),\end{aligned} \tag{4.35}$$

where

$$\mathbb{E}_B^{(j)}(t) \stackrel{\text{def}}{=} \mathbb{E}_{B^2}^{(j)}(t) + \mathbb{E}_{B,1}^{(j)}(t) + \mathbb{E}_{B,2}^{(j)}(t).$$

Similarly, we have

$$\begin{aligned} w_{t+1,j}w_{t,j} &= w_{t,j}(w_{t,j} - \gamma\sigma_j^2 \sum_{l\in[n]} w_{t,l}(\sum_{i\in B_t} U_{ij}U_{il}) + \Delta(w_{t,j} - w_{t-1,j})) \\ &= w_{t,j}^2 - \gamma\sigma_j^2 w_{t,j} \sum_{l\in[n]} w_{t,l}(\mathbb{E}_B^{(l,j)} + \frac{\beta}{n}\delta_{l,j}) + \Delta w_{t,j}(w_{t,j} - w_{t-1,j}) \\ &= \Omega_j w_{t,j}^2 - \Delta w_{t,j}w_{t-1,j} - \underbrace{\gamma\sigma_j^2 w_{t,j} \sum_l \mathbb{E}_B^{(l,j)} w_{t,l}}_{=\frac{1}{2}\mathbb{E}_{B,1}^{(j)}(t)}, \end{aligned} \tag{4.36}$$

where $\Omega_j \stackrel{\text{def}}{=} 1 - \Gamma_j + \Delta$.

Therefore, (4.35) and (4.36) imply

$$\begin{pmatrix} w_{t+1,j}^2 \\ w_{t,j}^2 \\ w_{t+1,j}w_{t,j} \end{pmatrix} = \underbrace{\begin{pmatrix} \Omega_j^2 & \Delta^2 & -2\Delta\Omega_j \\ 1 & 0 & 0 \\ \Omega_j & 0 & -\Delta \end{pmatrix}}_{\stackrel{\text{def}}{=}\boldsymbol{M}_j} \underbrace{\begin{pmatrix} w_{t,j}^2 \\ w_{t-1,j}^2 \\ w_{t,j}w_{t-1,j} \end{pmatrix}}_{\stackrel{\text{def}}{=}\tilde{\mathcal{X}}_{t,j}} + \underbrace{\begin{pmatrix} \tilde{N}_{t,j} + \mathbb{E}_1^{(j)}(t) \\ 0 \\ \mathbb{E}_2^{(j)}(t) \end{pmatrix}}_{\stackrel{\text{def}}{=}\tilde{\mathcal{Y}}_{t,j}}, \tag{4.37}$$

where

$$\tilde{N}_{t,j} \stackrel{\text{def}}{=} \frac{(1-\zeta)\gamma\sigma_j^2\Gamma_j}{n} \sum_l w_{t,l}^2 = \varphi_j^{(n)} \sum_l w_{t,l}^2, \text{ with } \varphi_j^{(n)} \stackrel{\text{def}}{=} \frac{(1-\zeta)\gamma\sigma_j^2\Gamma_j}{n},$$

$$\mathbb{E}_1^{(j)}(t) \stackrel{\text{def}}{=} \mathbb{E}_{beta}^{(j)}(t) + \mathbb{E}_{KL}^{(j)}(t) + \mathbb{E}_B^{(j)}(t), \text{ and}$$

$$\mathbb{E}_2^{(j)}(t) \stackrel{\text{def}}{=} -\frac{1}{2}\mathbb{E}_{B,1}^{(j)}(t).$$

Let us rewrite (4.37) as

$$\begin{aligned}\tilde{\mathcal{X}}_{t+1,j} &= \boldsymbol{M}_j \tilde{\mathcal{X}}_{t,j} + \tilde{\mathcal{Y}}_{t,j} \\ &= \boldsymbol{M}_j^2 \tilde{\mathcal{X}}_{t-1,j} + \boldsymbol{M}_j \tilde{\mathcal{Y}}_{t-1,j} + \tilde{\mathcal{Y}}_{t,j} \\ &= \boldsymbol{M}_j^t \tilde{\mathcal{X}}_{1,j} + \sum_{k=1}^{t} \boldsymbol{M}_j^{t-k} \tilde{\mathcal{Y}}_{k,j}.\end{aligned}$$

The eigendecomposition of $\boldsymbol{M}_j$ is given by $\boldsymbol{M}_j = \boldsymbol{X}_j \Lambda_j \boldsymbol{X}_j^{-1}$, $\Lambda_j = \mathbf{diag}(\lambda_{1,j}, \lambda_{2,j}, \lambda_{3,j})$, $\boldsymbol{X}_j = (x_{1,j}, x_{2,j}, x_{3,j})^T$ where

$$\lambda_{1,j} = \Delta, \lambda_{2,j} = \frac{-2\Delta + \Omega_j^2 + \sqrt{(\Omega_j^2)(\Omega_j^2 - 4\Delta)}}{2}, \lambda_{3,j} = \frac{-2\Delta + \Omega_j^2 - \sqrt{(\Omega_j^2)(\Omega_j^2 - 4\Delta)}}{2},$$
$$\text{and } \boldsymbol{X}_j = \begin{pmatrix} \Delta & \lambda_{2,j} & \lambda_{3,j} \\ 1 & 1 & 1 \\ \frac{\Omega_j}{2} & \kappa_{2,j} & \kappa_{3,j} \end{pmatrix} \text{ with } \kappa_{i,j} = \frac{\lambda_i \Omega_j}{\lambda_i + \Delta}. \tag{4.38}$$

Also, its inverse, assuming $\det \boldsymbol{X}_j \neq 0$, is given by

$$\boldsymbol{X}_j^{-1} = \frac{2}{\Omega_j^2 - 4\Delta} \begin{pmatrix} -1 & -\Delta & \Omega_j \\ \frac{1}{2} & \frac{\lambda_{3,j}}{2} & -\kappa_{3,j} \\ \frac{1}{2} & \frac{\lambda_{2,j}}{2} & -\kappa_{2,j} \end{pmatrix}.$$

Note that for each $j \in [n]$ and $i \in \{2, 3\}$, $\kappa_{i,j}$ satisfies

- $\kappa_{i,j}^2 = \lambda_{i,j}$ and $\kappa_{i,j} = \sqrt{\lambda_{i,j}}$ when $\Omega_j \geq 0$,
- $\kappa_{2,j} + \kappa_{3,j} = \Omega_j$, and
- $\kappa_{2,j}\kappa_{3,j} = \Delta$.

This implies that

$$\begin{aligned}\boldsymbol{M}_j^{t-k}\tilde{\mathcal{Y}}_{k,j} &= \boldsymbol{X}_j\Lambda_j^{t-k}\boldsymbol{X}_j^{-1}\begin{pmatrix}\tilde{N}_{k,j}+\mathbb{E}_1^{(j)}(k)\\0\\\mathbb{E}_2^{(j)}(k)\end{pmatrix}\\ &= \boldsymbol{X}_j\cdot\frac{2}{\Omega_j^2-4\Delta}\begin{pmatrix}-\lambda_{1,j}^{t-k}\left(\tilde{\mathcal{N}}_{k,j}+\mathbb{E}_1^{(j)}(k)\right)+\Omega_j\lambda_{1,j}^{t-k}\mathbb{E}_2^{(j)}(k)\\ \frac{1}{2}\lambda_{2,j}^{t-k}\left(\tilde{\mathcal{N}}_{k,j}+\mathbb{E}_1^{(j)}(k)\right)-\kappa_{3,j}\lambda_{2,j}^{t-k}\mathbb{E}_2^{(j)}(k)\\ \frac{1}{2}\lambda_{3,j}^{t-k}\left(\tilde{\mathcal{N}}_{k,j}+\mathbb{E}_1^{(j)}(k)\right)-\kappa_{2,j}\lambda_{3,j}^{t-k}\mathbb{E}_2^{(j)}(k)\end{pmatrix}.\end{aligned}$$

In particular, if we just focus on the (first coordinate of $\tilde{\mathcal{X}}_{t+1,j}) = w_{t+1,j}^2$, we have

$$\begin{aligned}w_{t+1,j}^2 &= (\boldsymbol{M}_j^t\tilde{\mathcal{X}}_{1,j})_1+\frac{2}{\Omega_j^2-4\Delta}\sum_{k=1}^{t}(-\lambda_{1,j}\cdot\lambda_{1,j}^{t-k}+\frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k}+\frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})\varphi_j^{(n)}\sum_{l\in[n]}w_{k,l}^2\\ &+\frac{2}{\Omega_j^2-4\Delta}\sum_{k=1}^{t}(-\lambda_{1,j}\cdot\lambda_{1,j}^{t-k}+\frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k}+\frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})\mathbb{E}_1^{(j)}(k)\\ &+\frac{2}{\Omega_j^2-4\Delta}\sum_{k=1}^{t}(\Omega_j\lambda_{1,j}\cdot\lambda_{1,j}^{t-k}-\kappa_{3,j}\lambda_{2,j}\cdot\lambda_{2,j}^{t-k}-\kappa_{2,j}\lambda_{3,j}\cdot\lambda_{3,j}^{t-k})\mathbb{E}_2^{(j)}(k)\end{aligned}$$

(Here $(\cdot)_1$ denotes the first coordinate of a vector). Summing over $j\in[n]$ and dividing both sides by 2 gives

$$\begin{aligned}\frac{1}{2}\sum_{j=1}^{n}w_{t+1,j}^2 &= \frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{M}_j^t\tilde{\mathcal{X}}_{1,j})_1\\ &+\sum_{k=1}^{t}\left(\sum_{j=1}^{n}\frac{2\varphi_j^{(n)}}{\Omega_j^2-4\Delta}(-\Delta\cdot\lambda_{1,j}^{t-k}+\frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k}+\frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})\right)f(k)\\ &+\sum_{k=1}^{t}\left(\sum_{j=1}^{n}\frac{1}{\Omega_j^2-4\Delta}(-\Delta\cdot\lambda_{1,j}^{t-k}+\frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k}+\frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})\mathbb{E}_1^{(j)}(k)\right)\\ &+\sum_{k=1}^{t}\left(\sum_{j=1}^{n}\frac{1}{\Omega_j^2-4\Delta}(\Omega_j\lambda_{1,j}\cdot\lambda_{1,j}^{t-k}-\kappa_{3,j}\lambda_{2,j}\cdot\lambda_{2,j}^{t-k}-\kappa_{2,j}\lambda_{3,j}\cdot\lambda_{3,j}^{t-k})\mathbb{E}_2^{(j)}(k)\right).\end{aligned}\tag{4.39}$$

Note that $\sum_{j=1}^{n}(\boldsymbol{M}_j^t\tilde{\mathcal{X}}_{1,j})_1$ describes the *forcing* term (see Section **??**). In order to analyze this term, observe

$$
\begin{aligned}
\tilde{\mathcal{X}}_{1,j} = \begin{pmatrix} w_{1,j}^2 \\ w_{0,j}^2 \\ w_{1,j}w_{0,j} \end{pmatrix} &= \begin{pmatrix} (1-\Gamma_j)^2 w_{0,j}^2 + \varphi_j^{(n)} \sum_l w_{0,l}^2 + \mathbb{E}_{beta}^{(j)}(0) + \mathbb{E}_{KL}^{(j)}(0) + \mathbb{E}_{B}^{(j)}(0) \\ w_{0,j}^2 \\ (1-\Gamma_j) w_{0,j}^2 - \frac{1}{2}\mathbb{E}_{B,1}^{(j)}(0) \end{pmatrix} \\
&= \begin{pmatrix} (1-\Gamma_j)^2 \left(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n}\right) + 2\varphi_j^{(n)} f(0) + \mathbb{E}_{beta}^{(j)}(0) + \mathbb{E}_{KL}^{(j)}(0) + \mathbb{E}_{B}^{(j)}(0) + (1-\Gamma_j)^2 \mathbb{E}_{w_0}^{(j)} \\ \sigma_j^2 \frac{R}{n} + \frac{\tilde{R}}{n} + \mathbb{E}_{w_0}^{(j)} \\ (1-\Gamma_j)\left(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n}\right) + (1-\Gamma_j)\mathbb{E}_{w_0}^{(j)} - \frac{1}{2}\mathbb{E}_{B,1}^{(j)}(0) \end{pmatrix} \\
&= \left(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n}\right) \begin{pmatrix} (1-\Gamma_j)^2 \\ 1 \\ (1-\Gamma_j) \end{pmatrix} + \begin{pmatrix} 2\varphi_j^{(n)} f(0) + \mathbb{E}_{1}^{(j)}(0) \\ 0 \\ 0 \end{pmatrix} + \mathbb{E}_{w_0}^{(j)} \begin{pmatrix} (1-\Gamma_j)^2 \\ 1 \\ (1-\Gamma_j) \end{pmatrix} + \mathbb{E}_{2}^{(j)}(0) \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbb{E}_{w_0}^{(j)} &\stackrel{\text{def}}{=} w_{0,j}^2 - \mathbb{E}[w_{0,j}^2] = w_{0,j}^2 - \left(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n}\right), \\
\mathbb{E}_{1}^{(j)}(0) &= \mathbb{E}_{beta}^{(j)}(0) + \mathbb{E}_{KL}^{(j)}(0) + \mathbb{E}_{B}^{(j)}(0), \quad \text{and } \mathbb{E}_{2}^{(j)}(0) = -\frac{1}{2}\mathbb{E}_{B,1}^{(j)}(0).
\end{aligned} \tag{4.40}
$$

Therefore, by using the eigendecomposition of $\boldsymbol{M}_j$ again, the first coordinate of $\boldsymbol{M}_j^t \tilde{\mathcal{X}}_{1,j}$ is given by

$$
\begin{aligned}
(\boldsymbol{M}_j^t \tilde{\mathcal{X}}_{1,j})_1 = \Bigg[& \boldsymbol{X}_j \Lambda_j^t \begin{pmatrix} -(1-\Gamma_j)^2 - \Delta + \Omega_j(1-\Gamma_j) \\ \frac{1}{2}(1-\Gamma_j)^2 + \frac{\lambda_{3,j}}{2} - \kappa_{3,j}(1-\Gamma_j) \\ \frac{1}{2}(1-\Gamma_j)^2 + \frac{\lambda_{2,j}}{2} - \kappa_{2,j}(1-\Gamma_j) \end{pmatrix} \cdot \frac{2}{\Omega_j^2 - 4\Delta} \Big(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n} + \mathbb{E}_{w_0}^{(j)} \Big) \\
& + \boldsymbol{X}_j \Lambda_j^t \begin{pmatrix} -1 \\ 1/2 \\ 1/2 \end{pmatrix} \cdot \frac{2}{\Omega_j^2 - 4\Delta} \Big(2\varphi_j^{(n)} f(0) + \mathbb{E}_1^{(j)}(0) \Big) + \boldsymbol{X}_j \Lambda_j^t \begin{pmatrix} \Omega_j \\ -\kappa_{3,j} \\ -\kappa_{2,j} \end{pmatrix} \cdot \frac{2}{\Omega_j^2 - 4\Delta} \mathbb{E}_2^{(j)}(0) \Bigg]_1 \\
= & \frac{2(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n})}{\Omega_j^2 - 4\Delta} \left(-\Delta\Gamma_j \cdot \lambda_{1,j}^{t+1} + \frac{1}{2}(1-\Gamma_j-\kappa_{3,j})^2 \cdot \lambda_{2,j}^{t+1} + \frac{1}{2}(1-\Gamma_j-\kappa_{2,j})^2 \cdot \lambda_{3,j}^{t+1} \right) \\
& + \frac{2\mathbb{E}_{w_0}^{(j)}}{\Omega_j^2 - 4\Delta} \left(-\Delta\Gamma_j \cdot \lambda_{1,j}^{t+1} + \frac{1}{2}(1-\Gamma_j-\kappa_{3,j})^2 \cdot \lambda_{2,j}^{t+1} + \frac{1}{2}(1-\Gamma_j-\kappa_{2,j})^2 \cdot \lambda_{3,j}^{t+1} \right) \\
& + \left(\frac{2\varphi_j^{(n)}}{\Omega_j^2 - 4\Delta} (-\lambda_{1,j}^{t+1} + \frac{1}{2} \cdot \lambda_{2,j}^{t+1} + \frac{1}{2} \cdot \lambda_{3,j}^{t+1}) \right) 2f(0) \\
& + \left(\frac{2}{\Omega_j^2 - 4\Delta} (-\lambda_{1,j}^{t+1} + \frac{1}{2} \cdot \lambda_{2,j}^{t+1} + \frac{1}{2} \cdot \lambda_{3,j}^{t+1}) \right) \mathbb{E}_1^{(j)}(0) \\
& + \left(\frac{2}{\Omega_j^2 - 4\Delta} (\Omega_j \Delta \cdot \Delta^t - \kappa_{3,j}\lambda_{2,j} \cdot \lambda_{2,j}^t - \kappa_{2,j}\lambda_{3,j} \cdot \lambda_{3,j}^t) \right) \mathbb{E}_2^{(j)}(0).
\end{aligned}
$$

Simple algebra shows $1 - \Gamma_j - \kappa_{3,j} = (\Omega_j - \Delta) - (\Omega_j - \kappa_{2,j}) = \Delta - \kappa_{2,j}$, and similarly, $1 - \Gamma_j - \kappa_{2,j} = \Delta - \kappa_{3,j}$. Hence, we conclude that

$$
f(t+1) = \frac{R}{2} h_1(t+1) + \frac{\tilde{R}}{2} h_0(t+1) + \sum_{k=0}^{t} \gamma^2 \zeta(1-\zeta) H_2(t-k) f(k) + \mathbb{E}(t).
$$

Here for $k = 0, 1$,

$$
h_k(t) = \frac{1}{n} \sum_{j=1}^{n} \frac{2(\sigma_j^2)^k}{\Omega_j^2 - 4\Delta} \left(-\Delta\Gamma_j \cdot \Delta^t + \frac{1}{2}(\Delta - \kappa_{2,j})^2 \cdot \lambda_{2,j}^t + \frac{1}{2}(\Delta - \kappa_{3,j})^2 \cdot \lambda_{3,j}^t \right),
$$

and

$$
H_2(t) = \frac{1}{n} \sum_{j=1}^{n} \frac{2\sigma_j^4}{\Omega_j^2 - 4\Delta} \Big(-\lambda_{1,j}^{t+1} + \frac{1}{2}\lambda_{2,j}^{t+1} + \frac{1}{2}\lambda_{3,j}^{t+1} \Big).
$$

Also, the error term $\mathbb{E}(t)$ is defined as

$$\mathbb{E}(t) \stackrel{\text{def}}{=} \mathbb{E}_{IC}(t) + \mathbb{E}_{beta}(t) + \mathbb{E}_{KL}(t) + \mathbb{E}_{M}(t), \tag{4.41}$$

where

$$\begin{aligned}
\mathbb{E}_{IC}(t) &\stackrel{\text{def}}{=} \sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta} \left(-\Delta\Gamma_j \cdot \Delta^{t+1} + \frac{1}{2}(\Delta - \kappa_{2,j})^2 \cdot \lambda_{2,j}^{t+1} + \frac{1}{2}(\Delta - \kappa_{3,j})^2 \cdot \lambda_{3,j}^{t+1} \right) \mathbb{E}_{w_0}^{(j)}, \\
\mathbb{E}_{beta}(t) &\stackrel{\text{def}}{=} \sum_{k=0}^{t} \left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta} (-\Delta \cdot \Delta^{t-k} + \frac{\lambda_{2,j}}{2} \cdot \lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2} \cdot \lambda_{3,j}^{t-k}) \mathbb{E}_{beta}^{(j)}(k) \right), \\
\mathbb{E}_{KL}(t) &\stackrel{\text{def}}{=} \sum_{k=0}^{t} \left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta} (-\Delta \cdot \Delta^{t-k} + \frac{\lambda_{2,j}}{2} \cdot \lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2} \cdot \lambda_{3,j}^{t-k}) \mathbb{E}_{KL}^{(j)}(k) \right), \text{ and} \\
\mathbb{E}_{M}(t) &\stackrel{\text{def}}{=} \sum_{k=0}^{t} \left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta} (-\Delta \cdot \Delta^{t-k} + \frac{\lambda_{2,j}}{2} \cdot \lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2} \cdot \lambda_{3,j}^{t-k}) \mathbb{E}_{B}^{(j)}(k) \right) \\
&\quad + \sum_{k=0}^{t} \left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta} (\Omega_j \Delta \cdot \Delta^{t-k} - \kappa_{3,j}\lambda_{2,j} \cdot \lambda_{2,j}^{t-k} - \kappa_{2,j}\lambda_{3,j} \cdot \lambda_{3,j}^{t-k}) \mathbb{E}_{2}^{(j)}(k) \right).
\end{aligned}$$

A few comments on the naming of errors: IC in $\mathbb{E}_{IC}(t)$ stands for *initial condition*. This error is generated from the initial bias on $w_{0,j}^2$. On the other hand, M in $\mathbb{E}_{M}(t)$ stands for *Martingale*; the error is an accumulation of martingales over each time iteration. We deal with these errors in detail in following sections. And note that Theorem 9 can be proved once we control the error $\mathbb{E}(t)$ *with overwhelming probability*.

4.6 Estimates based on concentration of measure on the high–dimensional orthogonal group

In this section, we give a high-level overview of the errors and how to bound them with overwhelming probability. Recall that we have the following error pieces:

$$\mathbb{E}(t) \stackrel{\text{def}}{=} \mathbb{E}_{IC}(t) + \mathbb{E}_{beta}(t) + \mathbb{E}_{KL}(t) + \mathbb{E}_{M}(t). \tag{4.42}$$

In order to bound the errors, we follow the methods that are used in [43]: we would like to make an *a priori* estimate that shows the function values remain bounded. Thus, we define the stopping time, for any fixed $\theta > 0$ and *large enough* $n \in \mathbb{N}$, by

$$\vartheta \stackrel{\text{def}}{=} \inf\left\{t \geq 0 : \|\boldsymbol{w}_t\| (= \|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta}\|) > n^{\theta}\right\}.$$

We then need to show:

Lemma 15. *For any $\theta > 0$, and for any $T > 0$, $\vartheta > T$ with overwhelming probability.*

Proof. From (4.32), we have

$$\boldsymbol{w}_{k+1} = \left((1+\Delta)\mathbf{I}_n - \gamma\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{P}_k\boldsymbol{U}\right)\boldsymbol{w}_k - \Delta\boldsymbol{w}_{k-1},$$

where $\mathbf{I}_n$ denotes an identity matrix of dimension $n \times n$. Therefore, by taking norm on both sides and applying triangle inequality, we have

$$\|\boldsymbol{w}_{k+1}\| \leq \left(1 + \Delta + \gamma\|\boldsymbol{\Sigma}\|_2^2\right)\|\boldsymbol{w}_k\| + \Delta\|\boldsymbol{w}_{k-1}\|.$$

Let $C := 1 + 2\Delta + \gamma\|\boldsymbol{\Sigma}\|_2^2$ and $\epsilon > 0$ is small enough so that $C^T \cdot n^{\epsilon} \leq n^{\theta}$. By induction hypothesis, if we are given $\|\boldsymbol{w}_l\| \leq C^l n^{\epsilon}$ for $l = 0, \cdots, k < T$, we have

$$\|\boldsymbol{w}_{k+1}\| \leq (1 + 2\Delta + \gamma\sigma_{\max}^2)C^k n^{\epsilon} \leq C^{k+1}n^{\epsilon},$$

and this finishes the proof once we check the initial conditions, i.e., $\|\boldsymbol{w}_0\|, \|\boldsymbol{w}_1\|$ are small enough with overwhelming probability. Observe, for any $\epsilon > 0$ and sufficiently large n,

$$\|\boldsymbol{w}_0\|^2 = \sum_{j\in[n]} \Big(\sigma_j \nu_{0,j} - (\boldsymbol{U}^T\boldsymbol{\eta})_j\Big)^2 \leq 2(\sigma_{\max}^2\|\boldsymbol{\nu}_0\|_2^2 + \|\boldsymbol{\eta}\|_2^2) = \mathcal{O}(1) \leq n^{\epsilon},$$

w.o.p. by assumption 10. Similarly, $\boldsymbol{w}_1$ is generated by the following formula

$$\boldsymbol{w}_1 = \left(\mathbf{I}_n - \gamma\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{P}_k\boldsymbol{U}\right)\boldsymbol{w}_0,$$

and applying norm on both sides gives

$$\|\boldsymbol{w}_1\| \leq (1+\gamma\sigma_{\max}^2)\|\boldsymbol{w}_0\| \leq (1+\gamma\sigma_{\max}^2)n^{\epsilon} \leq Cn^{\epsilon}.$$

□

We will need the result in what follows. Also, as an input, we work with the stopped process defined for any $t \geq 0$ by $\boldsymbol{w}_t^{\vartheta} \stackrel{\text{def}}{=} \boldsymbol{w}_{t\wedge\vartheta}$. Moreover, we condition on $\boldsymbol{\Sigma}$ going forward.

4.6.1 Control of the errors from the Initial Conditions

In this section, we focus on controlling the errors generated by the initial conditions:

$$\mathbb{E}_{IC}(t) = \sum_{j=1}^{n} \frac{1}{\Omega_j^2 - 4\Delta}\left(-\Delta\Gamma_j \cdot \lambda_{1,j}^{t+1} + \frac{1}{2}(\Delta - \kappa_{2,j})^2 \cdot \lambda_{2,j}^{t+1} + \frac{1}{2}(\Delta - \kappa_{3,j})^2 \cdot \lambda_{3,j}^{t+1}\right)\mathbb{E}_{w_0}^{(j)},$$

where

$$\mathbb{E}_{w_0}^{(j)} = w_{0,j}^2 - \mathbb{E}[w_{0,j}^2] = w_{0,j}^2 - \Big(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n}\Big).$$

The next Proposition shows that the error $\mathbb{E}_{IC}(t)$ can be bounded w.o.p.

Proposition 18. *For any $T > 0$ and for any $\epsilon > 0$, with overwhelming probability,*

$$\max_{0\leq t\leq T} |\mathbb{E}_{IC}(t)| \leq n^{\epsilon - 1/2}.$$

Proof. The proof is similar to that of [43, Lemma 10]. We rely on Chebyshev's inequality and the law of total probability to control the error. Fix $t \in [T]$ and let

$$C^{(j)}(t) \stackrel{\text{def}}{=} \frac{1}{\Omega_j^2 - 4\Delta}\left(-\Delta\Gamma_j \cdot \lambda_{1,j}^{t+1} + \frac{1}{2}(\Delta - \kappa_{2,j})^2 \cdot \lambda_{2,j}^{t+1} + \frac{1}{2}(\Delta - \kappa_{3,j})^2 \cdot \lambda_{3,j}^{t+1}\right),$$

and

$$W(t) \stackrel{\text{def}}{=} \sum_{j=1}^{n} C^{(j)}(t) w_{0,j}^2,$$

so that $\mathbb{E}_{IC}(t) = W(t) - \mathbb{E}[W(t)]$. From [43, Lemma 10], we know that the vector $\boldsymbol{\nu}_0^2$ follows the Dirichlet distribution (recall $\boldsymbol{\nu}_k = \boldsymbol{V}^T(\boldsymbol{x}_k - \widetilde{\boldsymbol{x}})$), and in particular, $\mathbb{E}(\nu_{0,j}^4) \leq \mathcal{O}(n^{-2})$ leads to $\mathbb{E}(w_{0,j}^4) \leq \mathcal{O}(n^{-2})$ (also recall $\boldsymbol{w}_k = \boldsymbol{\Sigma}\boldsymbol{\nu}_k - \boldsymbol{U}^T\boldsymbol{\eta}$, (4.30)). Therefore, the (conditional) variance of $W(t)$ is bounded by

$$\begin{aligned}
\text{Var}\Big[W(t)\Big] &= \mathbb{E}\left[\left(\sum_{j=1}^{n} C^{(j)}(t) w_{0,j}^2 - \sum_{j=1}^{n} C^{(j)}(t)\Big(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n}\Big)\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n} C^{(j)}(t)\Big(n w_{0,j}^2 - \big(R\sigma_j^2 + \tilde{R}\big)\Big)\right)^2\right] \\
&\leq \frac{1}{n^2}\mathbb{E}\left[\sum_{j=1}^{n}\big(C^{(j)}(t)\big)^2\Big(n w_{0,j}^2 - \big(R\sigma_j^2 + \tilde{R}\big)\Big)^2\right] \\
&= \frac{1}{n}\left[\frac{1}{n}\sum_{j=1}^{n}\big(C^{(j)}(t)\big)^2\Big(n^2\mathbb{E}[w_{0,j}^4] - \big(R\sigma_j^2 + \tilde{R}\big)^2\Big)\right] = \mathcal{O}\Big(\frac{1}{n}\Big),
\end{aligned}$$

where the Cauchy-Schwarz inequality was used in the second last line. Therefore, for $\epsilon > 0$, Chebyshev inequality gives

$$\Pr\left[\left|\sum_{j=1}^{n} C^{(j)}(t) w_{0,j}^2 - \sum_{j=1}^{n} C^{(j)}(t)\Big(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n}\Big)\right| \geq n^{\epsilon - 1/2}\right] \leq \frac{1}{n^{2\epsilon-1}}\text{Var}\Big[W(t)\Big] \xrightarrow[n\to\infty]{} 0.$$

Now applying the law of total probability (over $t = 1, \cdots, T$) to this gives the claim. □

4.6.2 Control of the beta errors

In this section, we control the errors generated by the difference of $\frac{\beta(\beta-1)}{n(n-1)}$ and $\zeta^2 = (\frac{\beta}{n})^2$. For $t \in [T \wedge \vartheta]$, recall

$$\mathbb{E}_{beta}(t) = \sum_{k=0}^{t} \left(\sum_{j=1}^{n} \frac{1}{\Omega_j^2 - 4\Delta} (-\Delta \cdot \lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2} \cdot \lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2} \cdot \lambda_{3,j}^{t-k}) \mathbb{E}_{beta}^{(j)}(k) \right),$$

with

$$\mathbb{E}_{beta}^{(j)}(t) = \gamma^2 \sigma_j^4 \left[\left(\frac{\beta(\beta-1)}{n(n-1)} - \zeta^2 \right) w_{t,j}^2 + \left(-\frac{\beta(\beta-1)}{n(n-1)} + \zeta^2 \right) \sum_i U_{ij}^2 \left(\sum_l U_{il} w_{t,l} \right)^2 \right].$$

First of all, note that

$$\delta \overset{\text{def}}{=} \frac{\beta(\beta-1)}{n(n-1)} - \zeta^2 = \frac{\beta}{n} \cdot \frac{(\beta-1)n - \beta(n-1)}{n(n-1)} = \frac{\zeta(\zeta-1)}{n-1} = \mathcal{O}(n^{-1}).$$

Then we can show the following:

Proposition 19. *For any $T > 0$ and for any $\epsilon > 0$, with overwhelming probability,*

$$\max_{0 \le t \le T \wedge \vartheta} |\mathbb{E}_{beta}(t)| \le n^{\alpha - 1/2},$$

for some $1/4 > \alpha > \epsilon$.

Proof. Let

$$C^{(j)}(t,k) \overset{\text{def}}{=} \frac{\gamma^2 \sigma_j^4}{\Omega_j^2 - 4\Delta} (-\Delta \cdot \lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2} \cdot \lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2} \cdot \lambda_{3,j}^{t-k}).$$

Then $C^{(j)}(t,k), j \in [n]$ are uniformly bounded by our assumptions, and we have

$$\mathbb{E}_{beta}(t) = \sum_{k=0}^{t} \sum_{j=1}^{n} C^{(j)}(t,k) \left[\delta w_{t,j}^2 - \delta \sum_i U_{ij}^2 \left(\sum_l U_{il} w_{t,l} \right)^2 \right].$$

Now Lemma 15 (boundedness on the norm of $\boldsymbol{w}_t$) and Lemma 17 (uniform boundedness on the coordinates of $\boldsymbol{U}\boldsymbol{w}_t$) gives

$$\mathbb{E}_{beta}(t \wedge \vartheta) \leq C\delta\big(\|\boldsymbol{w}_t^{\vartheta}\|^2 + n \cdot \max_i (\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_i^2\big) = \mathcal{O}(n^{2\alpha-1}),$$

for some $C > 0$, which shows our claim. □

4.6.3 Control of the Key lemma errors

In this section, we show that $\mathbb{E}_{KL}(t)$ can be bounded with overwhelming probability. The following Key Lemma from [43, Lemma 14] will be useful in the following:

Lemma 16 (Key Lemma)**.** *For any $T > 0$ and for any $\epsilon > 0$, for some $\{C^{(j)}(t)\}, j \in [n], 0 \leq t \leq T$ that are uniformly bounded, with overwhelming probability*

$$\max_{1\leq i\leq n} \max_{0\leq t\leq T} \left| \sum_{j=1}^{n} C^{(j)}(t) \left(\left(\mathbb{e}_j^T \boldsymbol{U}^T \mathbb{e}_i \right)^2 - \frac{1}{n} \right) \right| \leq n^{\epsilon - 1/2}.$$

Given this lemma, combined with the Key Lemma, we can bound the error $\mathbb{E}_{KL}(t)$ with overwhelming probability.

Proposition 20. *For any $T > 0$ and for any $\epsilon > 0$, with overwhelming probability,*

$$\max_{0\leq t\leq T\wedge\vartheta} |\mathbb{E}_{KL}(t)| \leq n^{\epsilon - 1/2}.$$

Proof. By definition, we have

$$\mathbb{E}_{KL}(t) = \sum_{k=0}^{t} \left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta} (-\lambda_{1,j} \cdot \lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2} \cdot \lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2} \cdot \lambda_{3,j}^{t-k}) \mathbb{E}_{KL}^{(j)}(k) \right),$$

with

$$\mathbb{E}_{KL}^{(j)}(t) = \gamma^2 \sigma_j^4 (\zeta - \zeta^2) \sum_{i\in[n]} (U_{ij}^2 - \frac{1}{n}) \left(\sum_{l\in[n]} U_{il} w_{t,l} \right)^2.$$

Thus for a sufficiently small $\tilde{\epsilon} > 0$ and some $C > 0$, and by applying Lemma 16 and Lemma 15,

$$|\mathbb{E}_{KL}^{(n)}(t \wedge \vartheta)| \leq C \sum_{k=0}^{t} \sum_{i=1}^{n} \left(e_i^T \boldsymbol{U} \boldsymbol{w}_t^{\vartheta}\right)^2 \cdot n^{\tilde{\epsilon}-1/2} \leq CTn^{\tilde{\epsilon}-1/2} \cdot \|\boldsymbol{w}_t^{\vartheta}\|^2 \leq n^{\epsilon-1/2}.$$

□

4.6.4 Control of the Martingale error

In this section, we bound the error caused by Martingale terms. Recall that

$$\begin{aligned}\mathbb{E}_M(t) = &\sum_{k=0}^{t} \left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta} (-\Delta \cdot \Delta^{t-k} + \frac{\lambda_{2,j}}{2} \cdot \lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2} \cdot \lambda_{3,j}^{t-k}) \mathbb{E}_B^{(j)}(k) \right) \\ &+ \sum_{k=0}^{t} \left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta} (\Omega_j \Delta \cdot \Delta^{t-k} - \kappa_{3,j}\lambda_{2,j} \cdot \lambda_{2,j}^{t-k} - \kappa_{2,j}\lambda_{3,j} \cdot \lambda_{3,j}^{t-k}) \mathbb{E}_2^{(j)}(k) \right),\end{aligned}$$

where

$$\mathbb{E}_B^{(j)}(t) = \mathbb{E}_{B^2}^{(j)}(t) + \mathbb{E}_{B,1}^{(j)}(t) + \mathbb{E}_{B,2}^{(j)}(t),$$

with

$$\begin{aligned}&\mathbb{E}_B^{(j)}(t) = \mathbb{E}_{B^2}^{j}(t) + \mathbb{E}_{B,1}^{(j)}(t) + \mathbb{E}_{B,2}^{(j)}(t), \text{ and } \mathbb{E}_2^{(j)}(t) = -\frac{1}{2}\mathbb{E}_{B,1}^{(j)}(t), \text{ with} \\ &\mathbb{E}_{B,1}^{(j)}(t) = -2\gamma\sigma_j^2 w_{t,j} \sum_{l\in[n]} \mathbb{E}_B^{(l,j)} w_{t,l}, \ \mathbb{E}_B^{(l,j)} = \sum_{i\in B} U_{il}U_{ij} - \zeta\delta_{l,j}, \\ &\mathbb{E}_{B,2}^{(j)}(t) = -2\gamma\sigma_j^2\Delta \sum_{l\in[n]} \mathbb{E}_B^{(l,j)} w_{t,l}(w_{t,j} - w_{t-1,j}), \text{ and} \\ &\mathbb{E}_{B^2}^{(j)}(t) = \underbrace{\gamma^2\sigma_j^4 \big(\sum_{l\in[n]} w_{t,l} (\sum_{i\in B} U_{ij}U_{il}) \big)^2}_{\stackrel{\text{def}}{=} ①} - \mathbb{E}[①].\end{aligned}$$

In view of the expression of $\mathbb{E}_M(t)$, we define

$$\mathbb{E}_{B,1}(t) \stackrel{\text{def}}{=} \sum_{k=0}^{t} \left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta} (-\Delta \cdot \lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2} \cdot \lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2} \cdot \lambda_{3,j}^{t-k}) \mathbb{E}_{B,1}^{(j)}(k) \right),$$

and

$$\mathbb{E}_{B^2}(t) \stackrel{\text{def}}{=} \sum_{k=0}^{t} \left(\sum_{j\in[n]} \frac{1}{\Omega_j^2 - 4\Delta} (-\Delta \cdot \lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2} \cdot \lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2} \cdot \lambda_{3,j}^{t-k}) \mathbb{E}_{B^2}^{(j)}(k) \right).$$

Then it is easy to see that controlling these two terms will lead to the control of the entire Martingale error. Control of $\mathbb{E}_{B,2}(t)$, which can be defined similarly to $\mathbb{E}_{B,1}(t)$, can be done with exactly the same as that of $\mathbb{E}_{B,1}(t)$. As for the second term of $\mathbb{E}_{M}(t)$ which includes $\mathbb{E}_{2}^{(j)}(t)$, our analysis will show that the coefficients won't play an important rule in the control of the error; so that term can be controlled for the same reason as $\mathbb{E}_{B,1}(t)$.

We organize the proof as follows. First, we introduce a proposition from [9] that gives an overwhelming probability concentration for sampling with replacement. Also, we claim that $\{\boldsymbol{U}\boldsymbol{w}_t\}, t \in [T \wedge \vartheta]$ is uniformly distributed with overwhelming probability over different coordinates. This lemma will lead to bounding the "first-order" error $\mathbb{E}_{B,1}(t)$ (similarly for $\mathbb{E}_{B,2}(t)$). As for bounding the "second-order" error $\mathbb{E}_{B^2}(t)$, we will use the Hanson-Wright inequality for sampling without replacement [2].

Control of $\mathbb{E}_{B,1}(t)$

The Martinagle error originates from randomly sampling a mini-batch at every iteration. We begin by presenting the following Bernstein-type concentration result for sampling without replacement so that we see that randomness does not deviate too much from the "expectation".

Proposition 21 (Proposition 1.4, [9]). *Let $\mathcal{X} = (x_1, \cdots, x_n)$ be a finite population of n points and $X_1, \cdots, X_\beta$ be a random sample drawn without replacement from $\mathcal{X}$. Let*

$$a = \min_{1\le i\le n} x_i \text{ and } b = \max_{1\le i\le n} x_i.$$

Also let

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

be the mean and variance of $\mathcal{X}$, respectively. Then for all $\epsilon > 0$,

$$\mathbb{P}\left(\frac{1}{\beta}\sum_{i=1}^{\beta} X_i - \mu \geq \epsilon\right) \leq \exp\left(-\frac{\beta\epsilon^2}{2\sigma^2 + (2/3)(b-a)\epsilon}\right).$$

Now we can show that $\boldsymbol{U}\boldsymbol{w}_t$ is more or less uniformly distributed over coordinates.

Lemma 17. $\max_k |(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_k| = \mathcal{O}(n^{\alpha-1/2})$ *with overwhelming probability for some* $1/4 > \alpha > \epsilon$.

Proof. We show a more general result, which is

$$\begin{aligned} MB^{(t)} &\overset{\text{def}}{=} \max_{1\leq k\leq n}\max_{1\leq m\leq n} |B_{k,m}^{(t)}| = \mathcal{O}(n^{\alpha(t)-1/2}) \text{ w.o.p.,} \\ \text{where } B_{k,m}^{(t)} &\overset{\text{def}}{=} \sum_{j=1}^{m} U_{kj} w_{t,j}^{\vartheta} \text{ and } 1/4 > \alpha(T\wedge\vartheta) > \alpha((T\wedge\vartheta)-1) > \cdots > \alpha(0) > \epsilon. \end{aligned} \tag{4.43}$$

Note that $B_{k,n}^{(t)} = (\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_k$, so $\max_{1\leq k\leq n} |(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_k| \leq MB^{(t)}$. One approach is to apply the Proposition 21 and the induction hypothesis. Note that the initial condition for the induction hypothesis will be treated later. From (4.33), we have

$$w_{t+1,j}^{\vartheta} = w_{t,j}^{\vartheta} - \gamma\sigma_j^2 \sum_{l\in[n]} w_{t,l}^{\vartheta}\Big(\sum_{i\in B_{t+1}} U_{ij}U_{il}\Big) + \Delta(w_{t,j}^{\vartheta} - w_{t-1,j}^{\vartheta}).$$

By multiplying U_{kj} and summing over $j = 1, \cdots, m$, on both sides, we have

$$B_{k,m}^{(t+1)} = B_{k,m}^{(t)} - \underbrace{\gamma\sum_{j=1}^{m}\sigma_j^2 U_{kj}\sum_{l\in[n]} w_{t,l}^{\vartheta}\Big(\sum_{i\in B_{t+1}} U_{ij}U_{il}\Big)}_{\overset{\text{def}}{=}①} + \Delta(B_{k,m}^{(t)} - B_{k,m}^{(t-1)}). \tag{4.44}$$

Let

$$X_{i,k,m}^{(t)} \overset{\text{def}}{=} \sum_{j=1}^{m}\sigma_j^2 U_{kj}U_{ij}\sum_{l\in[n]} U_{il}w_{t,l}^{\vartheta} = (\boldsymbol{U}\boldsymbol{\Sigma}_m^2\boldsymbol{U}^T)_{ik}(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_i,$$

where $\boldsymbol{\Sigma}_m = diag(\sigma_1^2, \cdots, \sigma_m^2, 0, \cdots, 0)$, so that ① $= \sum_{i\in B_{t+1}} X_{i,k,m}^{(t)}$. Note that we can assume that $k \notin B_{t+1}$ so that $k \neq i$, because we can deal with the term

$X_{k,k,m}^{(t)} = (\boldsymbol{U}\boldsymbol{\Sigma}_m^2\boldsymbol{U}^T)_{kk}(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_k$ separately. In order to use Proposition 21, we evaluate

$$\mu = \frac{1}{n}\sum_{i\in[n]} X_{i,k,m}^{(t)} = \frac{1}{n}\sum_{j=1}^{m}\sigma_j^2 U_{kj}\sum_{l\in[n]} w_{t,l}^{\vartheta}\delta_{j,l} = \frac{1}{n}\sum_{j=1}^{m}\sigma_j^2 U_{kj} w_{t,j}^{\vartheta},$$

and

$$\sigma^2 = \frac{1}{n}\sum_i (X_{i,k,m}^{(t)})^2 - \mu^2 = \frac{1}{n}\sum_i (\boldsymbol{U}\boldsymbol{\Sigma}_m^2\boldsymbol{U}^T)_{ik}^2(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_i^2 - \mu^2.$$

Now observe,

1. As for μ, by applying Abel's inequality,

$$|\mu| \leq \frac{1}{n}\sigma_{\max}^2 \max_m |B_{k,m}^{(t)}| \leq \frac{1}{n}\sigma_{\max}^2 M B^{(t)}.$$

2. When it comes to controlling σ^2, by using Lemma 15,

$$\sigma^2 \leq \frac{1}{n}\left(\max_{i\neq k} |(\boldsymbol{U}\boldsymbol{\Sigma}_m^2\boldsymbol{U}^T)_{ik}|^2\right)\|\boldsymbol{U}\boldsymbol{w}_t^{\vartheta}\|_2^2 + \mu^2 \leq n^{-1+2\theta}\left(\max_{i\neq k} |(\boldsymbol{U}\boldsymbol{\Sigma}_m^2\boldsymbol{U}^T)_{ik}|^2\right).$$

   When $i \neq k$, by referring to [43, Lemma 25],

$$|(\boldsymbol{U}\boldsymbol{\Sigma}_m^2\boldsymbol{U}^T)_{ik}| = |\sum_{j=1}^{m}\sigma_j^2 U_{ij}U_{kj}| = \mathcal{O}(n^{-1/2+\epsilon}) \text{ w.o.p.}$$

   Therefore, we have, with overwhelming probability,

$$\sigma^2 = \mathcal{O}(n^{-2+2\theta+2\epsilon}).$$

3. Observe,

$$b = \max_i X_{i,k,m}^{(t)} = \max_i (\boldsymbol{U}\boldsymbol{\Sigma}_m^2\boldsymbol{U}^T)_{ik}(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_i \leq \mathcal{O}(n^{\epsilon})\max_i |(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_i| \leq \mathcal{O}(n^{\epsilon})\cdot M B^{(t)}$$
$$= \mathcal{O}(n^{\epsilon+\alpha(t)-1/2}) \text{ w.o.p., and similar for } a = \min_i |X_{i,k,m}^{(t)}|.$$

Now applying Proposition 21 gives

$$\mathbb{P}\left(\frac{1}{\beta}\sum_{i=1}^{\beta} X_{i,k,m} - \mu \geq t\right) \leq \exp\left(-\frac{\beta t^2}{2\sigma^2 + (2/3)(b-a)t}\right),$$

where the concentration with overwhelming probability is attained for $t = n^{-3/2+\alpha'(t)}$, $\alpha'(t) > \alpha(t) > \theta + \epsilon$, and therefore

$$\mathbb{P}\left(\sum_{i=1}^{\beta} X_{i,k,m}^{(t)} - \beta\mu \geq \tilde{\epsilon}\right) \to 0 \text{ as } n \to \infty \text{ when } \tilde{\epsilon} = n^{-1/2+\alpha'(t)}.$$

So applying this to (4.44) gives

$$B_{k,m}^{(t+1)} = B_{k,m}^{(t)} - \mathbb{1}_{i=k} \cdot X_{k,k,m}^{(t)} - \left(\beta\mu + \mathcal{O}(n^{-1/2+\alpha'(t)})\right) + \Delta(B_{k,m}^{(t)} - B_{k,m}^{(t-1)}),$$

or

$$\begin{aligned} |B_{k,m}^{(t+1)}| &\leq MB^{(t)} + \sigma_{\max}^2 MB^{(t)} + \left(\frac{\beta}{n}\sigma_{\max}^2 MB^{(t)} + \mathcal{O}(n^{-1/2+\alpha'(t)})\right) \\ &\quad + \Delta(MB^{(t)} + MB^{(t-1)}) \\ &\leq C^{(k,m)}\mathcal{O}(n^{-1/2+\alpha'(t)}), \end{aligned}$$

for some $C^{(k,m)} > 0$. Now taking maximum on k and m gives

$$MB^{(t+1)} \leq \left(\max_{k,m} C^{(k,m)}\right)\mathcal{O}(n^{-1/2+\alpha'(t)}) = \mathcal{O}(n^{-1/2+\alpha(t+1)}) \text{ w.o.p.},$$

for some $\alpha(t+1) > \alpha'(t)$. Now once we show that the initial value $MB^{(0)}$ is small enough, by the induction hypothesis, we prove the theorem. Note that as $n \to \infty$, we can always make the increment $\alpha(t+1) - \alpha(t), t \in [T \wedge \vartheta - 1]$ small enough so that $\alpha(T \wedge \vartheta) < 1/4$.

Now it suffices to check the initial condition, i.e., $MB^{(0)}$ is small enough:

Claim. $MB^{(0)} = \max_k \max_m |B_{k,m}^{(0)}| = \mathcal{O}(n^{\alpha(0)-1/2})$ w.o.p., $\quad \alpha(0) > \theta + \epsilon$.

First note that $w_{0,j} = \sigma_j \nu_{0,j} - (\boldsymbol{U}^T \boldsymbol{\eta})_j, \boldsymbol{\nu}_t = \boldsymbol{V}^T(\boldsymbol{x}_t - \tilde{\boldsymbol{x}})$. Therefore

$$B^{(0)}_{k,m} = \sum_{j=1}^{m} U_{kj}(\sigma_j \nu_{0,j} - \sum_{l\in[n]} U_{lj}\eta_l) = \underbrace{\sum_{j=1}^{m} \sigma_j U_{kj}\nu_{0,j}}_{\stackrel{\text{def}}{=}①} - \underbrace{\sum_{j=1}^{m} U_{kj}(\sum_{l\in[n]} U_{lj}\eta_l)}_{\stackrel{\text{def}}{=}②}.$$

We first show that $B_{k,m} = B^{(0)}_{k,m}$ for a fixed k and m attains the desired error order. As for ①, we show that $f_m(\boldsymbol{U}_k) \stackrel{\text{def}}{=} ①$ is a Lipschitz function on S^{n-1}: observe, for $\boldsymbol{U}_k, \boldsymbol{U}'_k \in S^{n-1}$,

$$\begin{aligned} f_m(\boldsymbol{U}_k) - f_m(\boldsymbol{U}'_k) &= \sum_{j=1}^{m} \sigma_j (U_{kj} - U'_{kj})\nu_{0,j} \\ &\le \sqrt{\sum_{j=1}^{m} \sigma_j^2 \nu_{0,j}^2} \sqrt{\sum_{j=1}^{m} (U_{kj} - U'_{kj})^2} \le C\|\boldsymbol{U}_k - \boldsymbol{U}'_k\|_2, \end{aligned}$$

for some $C > 0$. Therefore, the concentration result for Lipschitz function ([**?**, Ex 5.1.12]) gives

$$\Pr\{|f_m(\boldsymbol{U}_k) - \mathbb{E} f_m(\boldsymbol{U}_k)| \ge t\} \le 2\exp(-cnt^2),$$

and the overwhelming probability concentration is attained for $t = n^{-1/2+\epsilon}$, $\epsilon > 0$.

As for ②, observe that

$$② = \sum_{j=1}^{n} g_j(t)(\boldsymbol{a}^T \boldsymbol{U})_j (\boldsymbol{b}^T \boldsymbol{U})_j,$$

where $g_j(t) = 1$ for $1 \le j \le m$ and 0 otherwise, $\boldsymbol{a} = e_k$, and $\boldsymbol{b} = \boldsymbol{\eta}$. Given $\boldsymbol{\eta}$ fixed, we have $\mathbb{E}_\eta[②|\boldsymbol{\eta}] = \frac{m}{n}\eta_k$. Therefore, by [43, Lemma 25], $② = \frac{m}{n}\eta_k + \mathcal{O}(n^{\epsilon-1/2})$ w.o.p. As $\max_k |\eta_k| \le n^{\epsilon-1/2}$ w.o.p. ($f(x) = \max_i |x_i|,\ x \in S^{n-1}$ is a Lipschitz function on S^{n-1} with Lipschitz constant 1), we conclude that $② = \mathcal{O}(n^{\epsilon-1/2})$ w.o.p. Therefore $B^{(0)}_{k,m} = \mathcal{O}(n^{\alpha(0)-1/2})$ w.o.p. for arbitrarily small enough $\epsilon + \theta < \alpha(0) < 1/4$ and taking maximum over k and m shows our claim. □

Above lemma leads to the control of $\mathbb{E}_{B,1}(t)$. Note that control of $\mathbb{E}_{B,2}(t)$ can be done very similarly to $\mathbb{E}_{B,1}(t)$.

Proposition 22 (Error bound for $\mathbb{E}_{B,1}(t)$)**.**

$$\max_{0\le t\le T\wedge\vartheta} |\mathbb{E}_{B,1}(t)| = \mathcal{O}(n^{\alpha'-1/2}) \textit{ w.o.p.},$$

where $1/2 > \alpha' > \alpha$*, with* α *from Lemma 17.*

Proof. Our strategy is to apply Proposition 21 as well as Lemma 17. Recall that

$$\begin{aligned}\mathbb{E}_{B,1}(t) &= \sum_{k=0}^{t}\left(\sum_{j\in[n]}\frac{1}{\Omega_j^2-4\Delta}(-\Delta\cdot\lambda_{1,j}^{t-k}+\frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k}+\frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})\mathbb{E}_{B,1}^{(j)}(k)\right),\\ &= \sum_{k=0}^{t}\left(\sum_{j\in[n]}C^{(j)}(t,k)w_{t,j}\sum_{l\in[n]}\mathbb{E}_{B}^{(l,j)}w_{t,l}\right),\end{aligned}$$

where $C^{(j)}(t,k)\overset{\text{def}}{=}\frac{1}{\Omega_j^2-4\Delta}(-\Delta\cdot\lambda_{1,j}^{t-k}+\frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k}+\frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})\cdot(-2\gamma\sigma_j^2)$. Let us define

$$X_i^{(t,k)}\overset{\text{def}}{=}\sum_{j\in[n]}C^{(j)}(t,k)U_{ij}w_{t,j}^{\vartheta}\sum_{l\in[n]}U_{il}w_{t,l}^{\vartheta},\text{ and }\mu_{(t,k)}=\frac{1}{n}\sum_{i\in[n]}X_i^{(t,k)}=\frac{1}{n}\sum_{j\in[n]}C^{(j)}(t,k)(w_{t,j}^{\vartheta})^2,$$

so that $\mathbb{E}_{B,1}(t\wedge\vartheta)=\sum_{k=0}^{t}\left(\sum_{i\in B}X_i^{(t,k)}-\beta\mu_{(t,k)}\right)$. Let $\sigma_{(t,k)}^2$ be the variance of $X_i^{(t,k)}$:

$$\sigma_{(t,k)}^2\overset{\text{def}}{=}\frac{1}{n}\sum_{i\in[n]}\left(\sum_{j\in[n]}C^{(j)}(t,k)U_{ij}w_{t,j}^{\vartheta}\sum_{l}U_{il}w_{t,l}^{\vartheta}\right)^2-\left(\frac{1}{n}\sum_{j\in[n]}C^{(j)}(t,k)(w_{t,j}^{\vartheta})^2\right)^2.$$

In order to determine its order, note that

$$\frac{1}{n}\sum_{i\in[n]}\left(\sum_{j\in[n]}C^{(j)}(t,k)U_{ij}w_{t,j}^{\vartheta}\sum_{l}U_{il}w_{t,l}^{\vartheta}\right)^2=\frac{1}{n}\sum_{i\in[n]}(\boldsymbol{U}\boldsymbol{\Sigma}_C\boldsymbol{w}_t^{\vartheta})_i^2(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_i^2,$$

where $\boldsymbol{\Sigma}_C\overset{\text{def}}{=}\mathbf{diag}\{C^{(j)}(t,k)\}_{j\in[n]}$. By applying Lemma 17, we have with overwhelming probability

$$\sigma_{(t,k)}^2\le\max_i\|(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_i\|^2\frac{1}{n}\|\boldsymbol{U}\boldsymbol{\Sigma}_C\boldsymbol{w}_t^{\vartheta}\|_2^2\le\max_i\|(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_i\|^2\frac{1}{n}\|\boldsymbol{\Sigma}_C\|_2^2\|\boldsymbol{w}_t^{\vartheta}\|_2^2\le\mathcal{O}(n^{2\alpha(t)+2\theta-1}).$$

Now Proposition 21 gives

$$\Pr\left(\frac{1}{\beta}\sum_{i=1}^{\beta} X_i^{(t,k)} - \mu_{(t,k)} \geq \tilde{\epsilon}\right) \leq \exp\left(-\frac{\beta\tilde{\epsilon}^2}{2\sigma_{(t,k)}^2 + (2/3)(b-a)\tilde{\epsilon}}\right),$$

where, by using Cauchy-Schwarz's inequality and applying Lemma 17 again,

$$\begin{aligned} b &= \max_{1\leq i\leq n}\left(\sum_{j\in[n]} C^{(j)}(t,k)U_{ij}w_{t,j}^{\vartheta}\sum_{l} U_{il}w_{t,l}^{\vartheta}\right) \\ &\leq \max_i |(\boldsymbol{U}\boldsymbol{w}_t^{\vartheta})_i| \sqrt{\sum_{j\in[n]}(C^{(j)}(t,k))^2(w_{t,j}^{\vartheta})^2}\sqrt{\sum_{j\in[n]} U_{ij}^2} = \mathcal{O}(n^{\alpha(t)+\theta-1/2}) \text{ w.o.p.} \end{aligned}$$

So by applying the same argument used in Proposition 21, and applying the union bound, we have

$$\mathbb{P}\left(|\mathbb{E}_{B,1}(t)| \geq \tilde{\epsilon}\right) \leq T\cdot\mathbb{P}\left(\left|\sum_{i=1}^{\beta} X_i^{(t,k)} - \beta\mu_{(t,k)}\right| \geq c\tilde{\epsilon}\right) \searrow 0 \text{ as } n\to\infty \text{ when } \tilde{\epsilon} = n^{-1/2+\alpha'(t)},$$

for $c = 1/t$ and any $1/2 > \alpha'(t) > \alpha(t) + \theta$. Note that θ can be taken as small as possible. Now taking maximum over $t, 0 \leq t \leq T\wedge\vartheta, t\in\mathbb{N}$, gives the claim, with $\alpha' \stackrel{\text{def}}{=} \alpha'(T\wedge\vartheta)$. □

Control of $\mathbb{E}_{B^2}^{(j)}(t)$

This section deals with controlling the error $\mathbb{E}_{B^2}^{(j)}(t)$. Recall that

$$\begin{aligned} \mathbb{E}_{B^2}(t) &= \sum_{k=0}^{t}\left(\sum_{j\in[n]}\frac{1}{\Omega_j^2 - 4\Delta}(-\lambda_{1,j}\cdot\lambda_{1,j}^{t-k} + \frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k} + \frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})\mathbb{E}_{B^2}^{(j)}(k)\right) \\ &= \sum_{k=0}^{t}\left(\sum_{j\in[n]} C^{(j)}(t,k)\left(\big(\sum_{l\in[n]} w_{t,l}(\sum_{i\in B_k} U_{ij}U_{il})\big)^2 - \mathbb{E}\Big[\big(\sum_{l\in[n]} w_{t,l}(\sum_{i\in B_k} U_{ij}U_{il})\big)^2\Big|\mathcal{F}_k\Big]\right)\right), \end{aligned}$$

where $C^{(j)}(t,k) \stackrel{\text{def}}{=} \frac{\gamma^2\sigma_j^4}{\Omega_j^2-4\Delta}(-\lambda_{1,j}\cdot\lambda_{1,j}^{t-k}+\frac{\lambda_{2,j}}{2}\cdot\lambda_{2,j}^{t-k}+\frac{\lambda_{3,j}}{2}\cdot\lambda_{3,j}^{t-k})$. Observe that the expression in the summand of k can be translated as a *quadratic form*:

$$\sum_{j\in[n]} C^{(j)}(t,k)\big(\sum_{l\in[n]} w_{k,l}(\sum_{i\in B_k} U_{ij}U_{il})\big)^2 = \sum_{j\in[n]} C^{(j)}(t,k)\left(e_j^T\boldsymbol{U}^T\boldsymbol{P}_k\boldsymbol{U}\boldsymbol{w}_k\right)^2$$
$$= (\boldsymbol{U}\boldsymbol{w}_k)^T\boldsymbol{P}_k\boldsymbol{U}\boldsymbol{\Sigma}_C\boldsymbol{U}^T\boldsymbol{P}_k(\boldsymbol{U}\boldsymbol{w}_k),$$

where $\boldsymbol{\Sigma}_C \stackrel{\text{def}}{=} \mathbf{diag}\{C^{(j)}(t,k)\}_{j\in[n]}$. Let $\boldsymbol{X}_k \stackrel{\text{def}}{=} \boldsymbol{P}_k(\boldsymbol{U}\boldsymbol{w}_k)$ and $\boldsymbol{D} \stackrel{\text{def}}{=} \boldsymbol{U}\boldsymbol{\Sigma}_C\boldsymbol{U}^T$. Note that, for a fixed time t and k, and conditioned on $\boldsymbol{U}$, $\boldsymbol{D}$ is a fixed symmetric matrix and $\boldsymbol{X}_k$ has a randomness only depending on $\boldsymbol{P}_k$. Therefore, our error $\mathbb{E}_{B^2}(t)$ can be expressed as

$$\mathbb{E}_{B^2}(t) = \sum_{k=0}^{t}\left(\boldsymbol{X}_k^T\boldsymbol{D}\boldsymbol{X}_k - \mathbb{E}[\boldsymbol{X}_k^T\boldsymbol{D}\boldsymbol{X}_k|\mathcal{F}_k]\right). \tag{4.45}$$

As we did in the previous section, in view of union bounds, it suffices to impose bounds on each summand of (4.45) at $k = 0,\cdots,t$. In order to have the *Hanson-Wright* type concentration for our expression, we introduce the concept of *Convex concentration property*.

Definition 18 (Convex concentration property, [2]). *Let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^n$. We will say that $\boldsymbol{X}$ has the convex concentration property with constant K if for every $1-$Lipschitz convex function $\varphi:\mathbb{R}^n\to\mathbb{R}$. we have $\mathbb{E}[\varphi(\boldsymbol{X})]<\infty$ and for every $t>0$,*

$$Pr(|\varphi(\boldsymbol{X})-\mathbb{E}\varphi(\boldsymbol{X})|\geq t)\leq 2\exp(-t^2/K^2).$$

Remark 8. *By a simple scaling, the previous remark can extend to $x_1,\cdots,x_n\in[a,b]$, in which case K in the definition above will be replaced by $K(b-a)$.*

What is interesting for us is that vectors obtained via sampling without replacement follow the convex concentration property ([2, Remark 2.3]). More precisely, if $x_1,\cdots,x_n\in[0,1]$ and for $m\leq n$ the random vector $\boldsymbol{X}=(X_1,\cdots,X_m)$ is obtained by sampling without replacement m numbers from the set $\{x_1,\cdots,x_n\}$, then $\boldsymbol{X}$ satisfies the convex

concentration property with an absolute constant K. In this sense, the following lemma ([2, Theorem 2.5]) will be useful to us.

Lemma 18 (Hanson-Wright concentration for sampling without replacement)**.** *Let $\boldsymbol{X}$ be a mean zero random vector in $\mathbb{R}^n$. If $\boldsymbol{X}$ has the convex concentration property with constant K, then for any $n \times n$ matrix $\boldsymbol{A}$ and every $t > 0$,*

$$\mathbb{P}(|\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} - \mathbb{E}\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}| \geq t) \leq 2\exp\left(-\frac{1}{C}\min\left(\frac{t^2}{2K^4\|\boldsymbol{A}\|_{HS}^2}, \frac{t}{K^2\|\boldsymbol{A}\|}\right)\right),$$

for some universal constant C.

Remark 9. *The assumption that $\boldsymbol{X}$ is centered is introduced just to simplify the statement of the theorem. Note that if $\boldsymbol{X}$ has the convex concentration property with constant K, then so does $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \mathbb{E}\boldsymbol{X}$. Moreover, observe,*

$$\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} = (\tilde{\boldsymbol{X}} + \mathbb{E}\boldsymbol{X})^T\boldsymbol{A}(\tilde{\boldsymbol{X}} + \mathbb{E}\boldsymbol{X}) = \tilde{\boldsymbol{X}}^T\boldsymbol{A}\tilde{\boldsymbol{X}} + \tilde{\boldsymbol{X}}^T\boldsymbol{A}(\mathbb{E}\boldsymbol{X}) + (\mathbb{E}\boldsymbol{X})^T\boldsymbol{A}\tilde{\boldsymbol{X}} + (\mathbb{E}\boldsymbol{X})^T\boldsymbol{A}(\mathbb{E}\boldsymbol{X}),$$

and $\mathbb{E}[\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}] = \mathbb{E}[\tilde{\boldsymbol{X}}^T\boldsymbol{A}\tilde{\boldsymbol{X}}] + \mathbb{E}[\boldsymbol{X}]^T\boldsymbol{A}\mathbb{E}[\boldsymbol{X}]$. *This implies*

$$\begin{aligned}
&\boldsymbol{P}(|\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}]| \geq t) \\
&\quad \leq \boldsymbol{P}(|\tilde{\boldsymbol{X}}^T\boldsymbol{A}\tilde{\boldsymbol{X}} - \mathbb{E}[\tilde{\boldsymbol{X}}^T\boldsymbol{A}\tilde{\boldsymbol{X}}]| \geq t/3) + 2\boldsymbol{P}(|\tilde{\boldsymbol{X}}^T\boldsymbol{A}(\mathbb{E}\boldsymbol{X}) - \mathbb{E}\big[\tilde{\boldsymbol{X}}^T\boldsymbol{A}(\mathbb{E}\boldsymbol{X})\big]| \geq t/3) \\
&\quad \leq 2\exp\left(-\frac{1}{C}\min\left(\frac{t^2}{2\cdot 9K^4\|\boldsymbol{A}\|_{HS}^2}, \frac{t}{3K^2\|\boldsymbol{A}\|}\right)\right) + 2\cdot 2\exp\left(-\frac{t^2}{9K^2\|\boldsymbol{A}(\mathbb{E}\boldsymbol{X})\|_2^2}\right).
\end{aligned}$$

Finally, we can bound the error $\mathbb{E}_{B^2}(t)$ using Lemma 18.

Proposition 23. *For any $\epsilon > 0$, we have*

$$\max_{0\leq t\leq T\wedge\vartheta} |\mathbb{E}_{B^2}(t)| = \mathcal{O}(n^{-1/2+2\tilde{\alpha}}) \text{ w.o.p.},$$

where $1/4 > \tilde{\alpha} > \alpha$, with α from Lemma 17.

Proof. Recall that

$$\mathbb{E}_{B^2}(t) = \sum_{k=0}^{t} \left(\boldsymbol{X}_k^T \boldsymbol{D} \boldsymbol{X}_k - \mathbb{E}[\boldsymbol{X}_k^T \boldsymbol{D} \boldsymbol{X}_k | \mathcal{F}_k] \right),$$

and we apply Lemma 18 to each summand of $\mathbb{E}_{B^2}(t \wedge \vartheta)$. More precisely,

- K is replaced by $K \cdot M_k$, where $M_k \stackrel{\text{def}}{=} \max_{l \in [n]} |(\boldsymbol{U}\boldsymbol{w}_k^{\vartheta})_l| = \mathcal{O}(n^{\alpha(k)-1/2})$, by Lemma 17.

- Observe that

$$\|\boldsymbol{D}\|_{HS}^2 \leq \|\boldsymbol{\Sigma}_C\|_{HS}^2 = \mathcal{O}(n),$$

  and

$$\|\boldsymbol{D}\| = \|\boldsymbol{\Sigma}_C\| = \mathcal{O}(1).$$

- $\mathbb{E}\boldsymbol{X}_k = (\mu_1, \cdots, \mu_n)$ where $\mu_l = \frac{\beta}{n}(\boldsymbol{U}\boldsymbol{w}_k^{\vartheta})_l, l \in [n]$, so that $\|\boldsymbol{D}\mathbb{E}\boldsymbol{X}\|_2 \leq \|\boldsymbol{D}\|_2 \|\mathbb{E}\boldsymbol{X}\|_2 \leq \mathcal{O}(n^{\theta})$.

Therefore, by using Lemma 18, we have

$$\begin{aligned}
&\boldsymbol{P}(|\boldsymbol{X}_k^T \boldsymbol{D} \boldsymbol{X}_k - \mathbb{E}\boldsymbol{X}_k^T \boldsymbol{D} \boldsymbol{X}_k| \geq \tilde{\epsilon} | \mathcal{F}_k) \\
&\quad \leq 2 \exp\left(-\frac{1}{C} \min\left(\frac{\tilde{\epsilon}^2}{2 \cdot 9 M_k^4 K^4 \|\boldsymbol{D}\|_{HS}^2}, \frac{\tilde{\epsilon}}{3 M_k^2 K^2 \|\boldsymbol{D}\|} \right) \right) \\
&\qquad + 2 \cdot 2 \exp\left(-\frac{\tilde{\epsilon}^2}{M_k^2 K^2 \|\boldsymbol{D}(\mathbb{E}\boldsymbol{X}_k)\|_2^2} \right),
\end{aligned}$$

and for $\tilde{\epsilon} = n^{2\tilde{\alpha}(k)-1/2}, 1/4 > \tilde{\alpha}(k) > \alpha(k)$, we obtain the desired concentration result. Now taking union bound over $k = 0, \cdots, T \wedge \vartheta$ gives the desired result, with $\tilde{\alpha} \stackrel{\text{def}}{=} \tilde{\alpha}(T \wedge \vartheta)$. □

4.6.5 Proof of Theorem 9

Proof of Theorem 9. We have observed that Proposition 18, Proposition 19, Proposition 20, Proposition 22 and Proposition 23 imply that there exists $C > 0$ such that for any $c > 0$,

there exists $D > 0$ such that

$$\Pr\left[\sup_{0\leq t\leq T\wedge\vartheta, t\in\mathbb{N}} |\mathbb{E}(t)| > n^{-C}\right] < Dn^{-c}.$$

Now combining this result with Lemma 15 proves the Theorem. □

4.7 Proof of Main Results

In this section, we prove various statements from Section 4.3. First, we analyze assumptions on the learning rate γ so that the kernel K is convergent (Proposition 12). Second, we define the Malthusian exponent and show under which conditions the convergence rate of our algorithm is determined by $\lambda_{2,\max}$ (Proposition 13). Third, We find an optimal set of learning rate and momentum parameter so that the SGD+M outperforms SGD in the large batch regime (Proposition 15). Lastly, we show the lower bound of the convergence rate of SGD+M in the small batch regime (Proposition 16).

4.7.1 Learning rate assumption and kernel bound

First, we show that the kernel $\tilde{K}$ is always a nonnegative function, regardless of whether the eigenvalues $\{\lambda_{2,j}, \lambda_{3,j}\}, j \in [n]$ are real or complex values.

Lemma 19 (Positivity of the kernel)**.** *The kernel function satisfies $\tilde{K}(t) \geq 0$ for any $t \geq 0$.*

Proof. Fix $j \in [n]$ and let

$$H_{2,j}(t) \stackrel{\text{def}}{=} \frac{2\sigma_j^4}{\Omega_j^2 - 4\Delta}\Big(-\Delta\cdot\Delta^t + \frac{1}{2}\lambda_{2,j}\cdot\lambda_{2,j}^t + \frac{1}{2}\lambda_{3,j}\cdot\lambda_{3,j}^t\Big)$$

be the j-th summand of $H_2(t)$. We address two cases. In the first case, assume $\Omega_j^2 - 4\Delta \geq 0$. Then $\lambda_{2,j}$ and $\lambda_{3,j}$ are positive real numbers and one can easily verify that $\lambda_{2,j} \geq \Delta \geq \lambda_{3,j}$

and $\lambda_{2,j}\lambda_{3,j} = \Delta^2$. By the arithmetic-geometric inequality, we have

$$H_{2,j}(t) \geq \frac{2\sigma_j^4}{\Omega_j^2 - 4\Delta}\Big(-\Delta^{t+1} + \sqrt{\lambda_{2,j}^{t+1}\lambda_{3,j}^{t+1}}\Big) = \frac{2\sigma_j^4}{\Omega_j^2 - 4\Delta}\Big(-\Delta^{t+1} + \Delta^{t+1}\Big) = 0.$$

In the second case, we assume $\Omega_j^2 - 4\Delta < 0$. In this case, $\lambda_{2,j}$ and $\lambda_{3,j}$ are complex conjugates with magnitude Δ, and therefore we have the relation

$$\lambda_{2,j}^t = \Delta^t e^{i\theta_j t}, \quad \text{and } \lambda_{3,j}^t = \Delta^t e^{-i\theta_j t},$$

for some $\theta_j \in \mathbb{R}$. By Euler's formula, we obtain

$$-\Delta^{t+1} + \frac{1}{2}\Big(\lambda_{2,j}^{t+1} + \lambda_{3,j}^{t+1}\Big) = -\Delta^{t+1} + \Delta^{t+1}\cos(\theta_j t) \leq 0.$$

and combined with the condition $\Omega_j^2 - 4\Delta < 0$ gives $H_{2,j}(t) \geq 0$. Hence these two cases give the claim. □

The next proposition establishes that, under an upper bound on the learning rate, the maximum of the eigenvalues $\{\lambda_{2,j}\}$ for $j \in [n]$ has its magnitude less than one. Let $\lambda_{2,\max} \stackrel{\text{def}}{=} \max_j |\lambda_{2,j}|$. A simple computation shows that when $\lambda_{2,j}$ is complex then $|\lambda_{2,j}| = \Delta$. In particular, when all the eigenvalues $\lambda_{2,j}$ are complex numbers, $\lambda_{2,\max} = \Delta$. Otherwise, $\lambda_{2,\max} > \Delta$. Recall again that $\sigma_{\max}^2$ and $\sigma_{\min}^2$ be the largest and smallest (nonzero) eigenvalue of $\boldsymbol{A}\boldsymbol{A}^T$, respectively.

Proposition 24. *If* $\gamma < \frac{2(1+\Delta)}{\zeta\sigma_{max}^2}$ *and* $0 \leq \Delta < 1$, *then* $\lambda_{2,\max} < 1$.

Proof. First observe that

$$\gamma < \frac{2(1+\Delta)}{\zeta\sigma_{\max}^2} \iff \Omega_{\min} \stackrel{\text{def}}{=} 1 - \gamma\zeta\sigma_{\max}^2 + \Delta > -1 - \Delta,$$

so we conclude $\Omega_j > -1 - \Delta$ for all $j \in [n]$. Note that Ω_j increases as σ_j decreases. Fix $j \in [n]$. First, when Ω_j is non-positive, i.e.

$$0 \geq \Omega_j > -1 - \Delta,$$

this implies $0 \leq \Omega_j < (1+\Delta)^2$. Second, let $\Omega_j \geq 0$. Then by the definition of $\Omega_j = 1 - \gamma\zeta\sigma_j^2 + \Delta$, and as $\sigma_j^2 > 0$, we have $\Omega_j \leq 1 + \Delta$, or $\Omega_j^2 < (1+\Delta)^2$. So in both cases, we have

$$\Omega_j^2 < (1+\Delta)^2. \tag{4.46}$$

Then plugging in (4.46) into the expression of $\lambda_{2,j}$ gives

$$\begin{aligned}
|\lambda_{2,j}| = \left|\frac{-2\Delta + \Omega_j^2 + \sqrt{\Omega_j^2(\Omega_j^2 - 4\Delta)}}{2}\right| &< \left|\frac{\Delta^2 + 1\sqrt{(1+\Delta)^2(\Delta^2 - 2\Delta + 1)}}{2}\right| \\
&= \frac{\Delta^2 + 1\sqrt{(1+\Delta)^2(\Delta-1)^2}}{2} \\
&= \frac{\Delta^2 + 1 + (1+\Delta)(1-\Delta)}{2} \\
&= 1,
\end{aligned}$$

where the second last inequality comes from the constraint $0 \leq \Delta < 1$. □

Now we are ready to prove Proposition 12.

Proof of Proposition 12

Proof. Note that $\gamma < \frac{1+\Delta}{\zeta\sigma_{\max}^2}$ implies not only $\lambda_{2,\max} < 1$ from Proposition 24, but also $\Omega_j > 0$ for all $j \in [n]$. Let $\tilde{C}_j \stackrel{\text{def}}{=} \gamma^2\zeta(1-\zeta)\sigma_j^4/(\Omega_j^2 - 4\Delta)$ for the following. Using the the fact that

$\lambda_{2,j}\lambda_{3,j} = \Delta^2$ and $\lambda_{2,j} + \lambda_{3,j} = -2\Delta + \Omega_j^2$, we have

$$\begin{aligned}
\sum_{t=0}^{\infty} \tilde{K}(t) &= \sum_{t=0}^{\infty} \frac{1}{n}\bigg(\sum_{j=1}^{n} \tilde{C}_j(-2\Delta \cdot \Delta^t + \lambda_{2,j} \cdot \lambda_{2,j}^t + \lambda_{3,j} \cdot \lambda_{3,j}^t)\bigg) \\
&= \frac{1}{n}\sum_{j=1}^{n} \tilde{C}_j\bigg(-2\frac{\Delta}{1-\Delta} + \frac{\lambda_{2,j}}{1-\lambda_{2,j}} + \frac{\lambda_{3,j}}{1-\lambda_{3,j}}\bigg) \\
&= \frac{1}{n}\sum_{j=1}^{n} \tilde{C}_j\bigg(\frac{-2\Delta}{1-\Delta} + \frac{-2\Delta + \Omega_j^2 - 2\Delta^2}{1 + 2\Delta - \Omega_j^2 + \Delta^2}\bigg) \\
&= \frac{1}{n}\sum_{j=1}^{n} \frac{(1-\zeta)\zeta\gamma^2\sigma_j^4}{\Omega_j^2 - 4\Delta} \cdot \frac{(1+\Delta)(\Omega_j^2 - 4\Delta)}{(1-\Delta)(1+\Delta+\Omega_j)(1+\Delta-\Omega_j)} \\
&= \frac{1}{n}\sum_{j=1}^{n} \frac{(1-\zeta)\gamma\sigma_j^2}{\Omega_j^2 - 4\Delta} \cdot \frac{(1+\Delta)(\Omega_j^2 - 4\Delta)}{(1-\Delta)(1+\Delta+\Omega_j)} \\
&= \frac{1}{n}\sum_{j=1}^{n} \frac{(1-\zeta)\gamma\sigma_j^2(1+\Delta)}{(1-\Delta)(1+\Delta+\Omega_j)} \\
&\leq \frac{1}{n}\sum_{j=1}^{n} \frac{(1-\zeta)\gamma\sigma_j^2}{1-\Delta} = \frac{(1-\zeta)\gamma}{1-\Delta} \cdot \frac{1}{n}\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{A}) < 1,
\end{aligned}$$

where $\Omega_j > 0$ was used in the last inequality. □

When the norm of the kernel is less than 1, we can specify the limit of the solution $\psi(t)$ to the Volterra equation when $t \to \infty$, as Proposition 11 states.

Proof of Proposition 11

Proof. This is immediate from [8, Proposition 7.4]. In particular, from our expression of the renewal equation (4.11), we have

$$\psi(t) \to \frac{F(\infty)}{1 - \|\tilde{K}\|} \quad \text{as} \quad t \to \infty.$$

Now the proof is done once we evaluate the limit of $F(t) = \frac{R}{2}h_1(t) + \frac{\tilde{R}}{2}h_0(t)$. Note that $\lim_{t\to\infty} h_1(t) = 0$. On the other hand, as for $h_0(t)$, if $n > d$, $\sigma_j = 0$ for $j = d+1, \cdots, n$. And for such j's satisfying $\sigma_j = 0$, we can easily verify that $\lambda_{2,j} = 1, \lambda_{3,j} = \Delta^2, \Omega_j =$

$1+\Delta, \kappa_{2,j}=1, \kappa_{3,j}=\Delta$. Therefore,

$$\lim_{t\to\infty} h_0(t) = \lim_{t\to\infty}\left\{\frac{1}{n}\sum_{j=d+1}^{n}\frac{2}{\Omega_j^2-4\Delta}\big(0+\frac{1}{2}(1-\Delta)^2\cdot 1+0\big)\right\} = \frac{n-d}{n} = 1-r,$$

and this proves the claim. □

4.7.2 Malthusian exponent and convergence rate

In this section, we show that the Malthusian exponent Ξ is always smaller than $\lambda_{2,\max}^{-1}$ for a finite dimension n. Also, in the problem constrained regime we show that SGD+M shares the same convergence rate with full batch gradient descent with momentum with adjusted learning rate.

Proposition 25. *The Malthusian exponent defined in* (4.13) *satisfies*

$$\Xi < (\lambda_{2,\max})^{-1}$$

when the dimension n is finite.

Proof. It suffices to observe that the convergence rate of $H_2(t)$ is determined by $\lambda_{2,\max}$; if all $\lambda_{2,j}, j\in[n]$, are real numbers, then we can easily show that $\lambda_{2,j}>\Delta>\lambda_{3,j}$. Therefore $\lambda_{2,\max}$ takes over the convergence rate of $H_2(t)$. If, for some $j\in[n]$, $\lambda_{2,j}$ and $\lambda_{3,j}$ are both complex numbers, observe that $|\lambda_{2,j}|=|\lambda_{3,j}|=\Delta$. In that case, if we let $\lambda_{2,j}=\Delta\exp(i\theta_j)$ for some $\theta_j\in\mathbb{R}$, $\lambda_{3,j}=\Delta\exp(-i\theta_j)$ then

$$-\Delta^{t+1}+\frac{1}{2}\lambda_{2,j}^{t+1}+\frac{1}{2}\lambda_{3,j}^{t+1} = -\Delta^{t+1}+\frac{1}{2}\Delta^{t+1}\cdot 2\cos(i(t+1)\theta_j) = \Delta^{t+1}(-1+\cos(i(t+1)\theta_j)).$$

Therefore, Δ is the governing convergence rate of such j-th summand of $H_2(t)$ and the overall convergence rate of $H_2(t)$ is still determined by $\lambda_{2,\max}$. If all $\lambda_{2,j}, j\in[n]$, are complex numbers then the observation above shows that the governing convergence rate of $H_2(t)$ should be $\Delta=\lambda_{2,\max}$ and this proves our claim. □

When $\lambda_{2,\max}$ takes over the convergence behavior of SGD+M, we can easily see that its convergence dynamics is nothing but its analogue with full batch size but with an adjusted learning rate. This can be easily obtained by $\zeta = 1$ in Theorem 7, but we provide a statement for full batch SGD+M and its proof for completeness.

Proof of Proposition 14

Proof. Basically, we follow the same arguments introduced in 4.5.3, but with $\zeta = 1$; so we would not have any errors generated by selecting mini-batches. In other words, $\mathbb{E}_B^{(l,j)} = 0$. This implies the following, which is an analogue of (4.37),

$$\begin{pmatrix} w_{t+1,j}^2 \\ w_{t,j}^2 \\ w_{t+1,j}w_{t,j} \end{pmatrix} = \underbrace{\begin{pmatrix} \Omega_j^2 & \Delta^2 & -2\Delta\Omega_j \\ 1 & 0 & 0 \\ \Omega_j & 0 & -\Delta \end{pmatrix}}_{=\boldsymbol{M}_j} \begin{pmatrix} w_{t,j}^2 \\ w_{t-1,j}^2 \\ w_{t,j}w_{t-1,j} \end{pmatrix}. \tag{4.47}$$

This implies $w_{t+1,j}^2 = (\boldsymbol{M}_j^t \tilde{\mathcal{X}}_{1,j})_1$ and following the same arguments in 4.5.3 gives

$$\begin{aligned}(\boldsymbol{M}_j^t \tilde{\mathcal{X}}_{1,j})_1 = &\frac{2(\frac{R}{n}\sigma_j^2 + \frac{\tilde{R}}{n})}{\Omega_j^2 - 4\Delta}\left(-\Delta\Gamma_j \cdot \lambda_{1,j}^{t+1} + \frac{1}{2}(1-\Gamma_j-\kappa_{3,j})^2 \cdot \lambda_{2,j}^{t+1} + \frac{1}{2}(1-\Gamma_j-\kappa_{2,j})^2 \cdot \lambda_{3,j}^{t+1}\right) \\ &+ \frac{2\mathbb{E}_{w_0}^{(j)}}{\Omega_j^2 - 4\Delta}\left(-\Delta\Gamma_j \cdot \lambda_{1,j}^{t+1} + \frac{1}{2}(1-\Gamma_j-\kappa_{3,j})^2 \cdot \lambda_{2,j}^{t+1} + \frac{1}{2}(1-\Gamma_j-\kappa_{2,j})^2 \cdot \lambda_{3,j}^{t+1}\right).\end{aligned}$$

Therefore, this leads to

$$f(t+1) = \frac{R}{2}h_1(t+1) + \frac{\tilde{R}}{2}h_0(t+1) + \mathbb{E}(t),$$

with the error term $\mathbb{E}(t) = \mathbb{E}_{IC}(t)$. Now taking $n \to \infty$ combined with Proposition 18 gives (4.18). Note that the convergence rate of $\psi_{\text{full}}(t)$ is determined by $\lambda_{2,\max}^{(\text{full})} :=$

$\max_j |\lambda_{2,j}^{(\text{full})}|$, where

$$\lambda_{2,j}^{(\text{full})} = \frac{-2\Delta + (\Omega_j^{(\text{full})})^2 + \sqrt{(\Omega_j^{(\text{full})})^2((\Omega_j^{(\text{full})})^2 - 4\Delta)}}{2},\ \Omega_j^{(\text{full})} \stackrel{\text{def}}{=} 1 - \gamma_{\text{full}}\sigma_j^2 + \Delta.$$

And observing that $\lambda_{2,j}^{(\text{full})} = \lambda_{2,j}$ if $\gamma_{\text{full}} = \gamma\zeta$ gives our conclusion. □

4.7.3 Choice of optimal learning rate and momentum

In this section, we prove Proposition 13 which states a sufficient condition for a set of learning rate and momentum parameters to be in the problem-constrained regime. We also offer the proof of Proposition 15, which gives an optimal learning rate and momentum so that SGD+M outperforms SGD in terms of convergence rate. Finally, the proof of Proposition 16 will be given as well.

Proof of Proposition 13

Remark on the assumption. The first assumption on the learning rate, i.e., $\gamma \leq \frac{1+\Delta}{\zeta\sigma_{\max}^2}$ implies that $\Omega_j \geq 0$ for all $j \in [n]$. On the other hand, the second condition, i.e., $\gamma \leq \frac{(1-\sqrt{\Delta})^2}{\zeta\sigma_{\min}^2}$, implies that $\Omega_{\max} \geq 2\sqrt{\Delta}$. Note that when $\Omega_{\max} = 2\sqrt{\Delta}$, $\lambda_{2,\max} = \frac{1}{2}\big(-2\Delta + \Omega_{\max}^2 + \sqrt{\Omega_{\max}^2(\Omega_{\max}^2 - 4\Delta)}\big) = \Delta$.

Proof. First recall that $\varphi_j^{(n)} = \frac{(1-\zeta)\gamma\sigma_j^2\Gamma_j}{n}$ and observe that, for $1 < \Upsilon < \lambda_{2,\max}^{-1}$,

$$\begin{aligned}
\tilde{K}(\Upsilon) \stackrel{\text{def}}{=} \sum_{t=0}^{\infty} \Upsilon^t K(t) &= \sum_{t=0}^{\infty} \left(\sum_{j=1}^{n} \frac{\varphi_j^{(n)}}{\Omega_j^2 - 4\Delta} (-2\Delta \cdot (\Upsilon\lambda_{1,j})^t + \lambda_{2,j} \cdot (\Upsilon\lambda_{2,j})^t + \lambda_{3,j} \cdot (\Upsilon\lambda_{3,j})^t) \right) \\
&= \sum_{j=1}^{n} \frac{\varphi_j^{(n)}}{\Omega_j^2 - 4\Delta} \left(\frac{-2\Delta}{1-\Upsilon\Delta} + \frac{\lambda_{2,j}}{1-\Upsilon\lambda_{2,j}} + \frac{\lambda_{3,j}}{1-\Upsilon\lambda_{3,j}} \right) \\
&= \sum_{j=1}^{n} \frac{\varphi_j^{(n)}}{\Omega_j^2 - 4\Delta} \left(\frac{-2\Delta}{1-\Upsilon\Delta} + \frac{-2\Delta + \Omega_j^2 - 2\Upsilon\Delta^2}{1 + \Upsilon(2\Delta - \Omega_j^2) + \Upsilon^2\Delta^2} \right) \\
&= \sum_{j=1}^{n} \frac{(1-\zeta)\zeta\gamma^2\sigma_j^4}{n} \left(\frac{(1+\Upsilon\Delta)}{(1-\Upsilon\Delta)(1-\Upsilon(-2\Delta+\Omega_j^2) + \Upsilon^2\Delta^2)} \right) \\
&= \sum_{j=1}^{n} \frac{C\zeta\gamma\sigma_j^4}{n} \left(\frac{(1-\Delta)(1+\Upsilon\Delta)}{(1-\Upsilon\Delta)(1+\Upsilon\Delta+\sqrt{\Upsilon}\Omega_j)(1+\Upsilon\Delta-\sqrt{\Upsilon}\Omega_j)} \right),
\end{aligned}$$

where $C = (1-\zeta)\gamma/(1-\Delta)$. Observe, as $\Omega_j \geq 0$,

$$\tilde{K}(\Upsilon) \leq \frac{C\zeta\gamma}{n} \cdot \frac{(1-\Delta)}{(1-\Upsilon\Delta)} \sum_{j=1}^{n} \frac{\sigma_j^4}{1+\Upsilon\Delta-\sqrt{\Upsilon}\Omega_j}. \tag{4.48}$$

Let us analyze the denominator of the summand first. Let $f_j(x) := 1 + x^2\Delta - x\Omega_j, 1 < x < \sqrt{\Delta^{-1}}$. Then the denominator in the summand is $f_j(\sqrt{\Upsilon})$. Especially, $f_{\min}(x) := \min_j f_j(x) = 1 + x^2\Delta - x\Omega_{\max}, \Omega_{\max} = 1 - \gamma\zeta{\sigma_{\min}}^2 + \Delta$. Note that $f_{\min}(x)$ is a quadratic function of x and the solution to $f_{\min}(x) = 0$ is $x = \sqrt{\lambda_{2,\max}^{-1}}$ (the other root $\sqrt{\lambda_{3,\max}^{-1}}$ exceeds the valid domain of x). Also, observe that this is where the assumption $\Omega_{\max} \geq 2\sqrt{\Delta}$ is used.

Note that $f_j(1) = \gamma\zeta\sigma_j^2$. Simple algebra shows that for $1 < x < \alpha < \beta$, $c_1(x-\alpha)^2 \leq c_2(x-\alpha)(x-\beta)$ where $c_1, c_2 > 0$ satisfies $c_1(1-\alpha)^2 = c_2(1-\alpha)(1-\beta)$, i.e. two functions coincide at $x = 1$ and $x = \alpha$. If $\lambda_{2,j} \geq 0$, or $\Omega_j^2 - 4\Delta \geq 0$, then the argument above gives

$$f_{\min}\Big(\frac{1+\sqrt{\lambda_{2,\max}^{-1}}}{2}\Big) \geq \frac{\gamma\zeta\sigma_{\min}^2}{4}.$$

Now for any $j \in [n]$, note that $f_j(x) - f_{\min}(x) = x\gamma\zeta(\sigma_j^2 - \sigma_{\min}^2)$ is an increasing function of $x \in \mathbb{R}$. So observe,

$$\begin{aligned} f_j\Big(\frac{1+\sqrt{\lambda_{2,\max}^{-1}}}{2}\Big) &\geq f_{\min}\Big(\frac{1+\sqrt{\lambda_{2,\max}^{-1}}}{2}\Big) + \gamma\zeta(\sigma_j^2 - \sigma_{\min}^2) \\ &\geq \frac{\gamma\zeta\sigma_{\min}^2}{4} + \frac{1}{4}\gamma\zeta(\sigma_j^2 - \sigma_{\min}^2) = \frac{1}{4}\gamma\zeta\sigma_j^2. \end{aligned}$$

Therefore, when $\sqrt{\Upsilon} = \frac{1+\sqrt{\lambda_{2,\max}^{-1}}}{2}$, (4.48) gives

$$\tilde{K}(\Upsilon) \leq \frac{4C}{n} \cdot \frac{(1-\Delta)}{(1-\Upsilon\Delta)} \sum_{j=1}^{n} \sigma_j^2.$$

Moreover, in order to bound the denominator $(1 - \Upsilon\Delta)$ on the right-hand side, if we define $g(x) \stackrel{\text{def}}{=} 1 - \Delta x^2$, g is a decreasing function on $[1, \sqrt{\Delta^{-1}}]$ and

$$g(\frac{1+\sqrt{\lambda_{2,\max}^{-1}}}{2}) \geq g(\frac{1+\sqrt{\Delta^{-1}}}{2}) \geq \frac{1-\Delta}{2},$$

by considering a linear line passing through $(1, 1-\Delta)$ and $(\sqrt{\Delta^{-1}}, 0)$ that lies below g. Therefore,

$$\tilde{K}(\Upsilon) \leq \frac{4C}{n} \cdot \frac{(1-\Delta)}{(1-\Upsilon\Delta)} \cdot \sum_{j=1}^{n} \sigma_j^2 \leq \frac{8(1-\zeta)\gamma}{(1-\Delta)} \frac{1}{n} \mathsf{tr}(\boldsymbol{A}^T\boldsymbol{A}).$$

□

Proof of Proposition 15

Proof. First, when the assumption $\frac{(1-\sqrt{\Delta})^2}{\zeta\sigma_{\min}^2} \leq \frac{(1+\sqrt{\Delta})^2}{2\zeta\sigma_{\max}^2}$ is met, we have

$$\frac{1-\sqrt{\Delta}}{1+\sqrt{\Delta}} \leq \frac{1}{\sqrt{2\kappa}}.$$

Solving this inequality with respect to Δ gives

$$\Delta \geq \left(\frac{1 - \frac{1}{\sqrt{2\kappa}}}{1 + \frac{1}{\sqrt{2\kappa}}} \right)^2.$$

Furthermore, from Proposition 13, when $\gamma = \frac{(1-\sqrt{\Delta})^2}{\zeta \sigma_{\min}^2}$, observe that $\lambda_{2,\max} = \Delta$ and

$$\frac{8(1-\zeta)\gamma}{(1-\Delta)} \frac{1}{n} \text{tr}(\boldsymbol{A}^T \boldsymbol{A}) = \frac{8(1-\zeta)(1-\sqrt{\Delta})^2}{\zeta \sigma_{\min}^2 (1-\Delta)} \frac{1}{n} \text{tr}(\boldsymbol{A}^T \boldsymbol{A}) = \frac{8(1-\zeta)}{\zeta} \cdot \frac{1-\sqrt{\Delta}}{1+\sqrt{\Delta}} \bar{\kappa} < 1.$$

Therefore, this condition implies

$$\frac{1-\sqrt{\Delta}}{1+\sqrt{\Delta}} < \frac{\mathcal{C}}{\bar{\kappa}},$$

where $\mathcal{C} = \mathcal{C}(\zeta) \stackrel{\text{def}}{=} \zeta/(8(1-\zeta))$ and solving this inequality gives

$$\sqrt{\Delta} > \frac{1 - \frac{\mathcal{C}}{\bar{\kappa}}}{1 + \frac{\mathcal{C}}{\bar{\kappa}}}.$$

□

Next, we present the proof of Proposition 16.

Proof of Proposition 16

Proof. For brevity and clarity, we define the following quantities:

$$\gamma_1 \stackrel{\text{def}}{=} \frac{1+\Delta}{\zeta \sigma_{\max}^2}, \quad \gamma_2 \stackrel{\text{def}}{=} \frac{(1-\sqrt{\Delta})^2}{\zeta \sigma_{\min}^2}, \text{ and } \gamma_3 \stackrel{\text{def}}{=} \frac{1}{\bar{\kappa} \sigma_{\min}^2} \cdot \frac{1-\Delta}{1-\zeta}.$$

Note that the assumptions on the learning rate γ in Proposition 12 imply that $\gamma \leq \min(\gamma_1, \gamma_3)$.

First, let us assume that $\gamma \geq \gamma_2$. Recall that this condition implies that $\Omega_{\max}^2 - 4\Delta \leq 0$ and therefore $\lambda_{2,\max} = \Delta$. In this case, $\gamma_2 \leq \gamma \leq \gamma_3$ implies that

$$\frac{(1-\sqrt{\Delta})^2}{\zeta\sigma_{\min}^2} \leq \frac{1}{\bar{\kappa}\sigma_{\min}^2} \cdot \frac{1-\Delta}{1-\zeta} \Rightarrow \frac{1-\sqrt{\Delta}}{1+\sqrt{\Delta}} \leq \frac{\zeta}{(1-\zeta)\bar{\kappa}}, \text{ or}$$
$$\sqrt{\Delta} \geq \frac{1-\frac{\zeta}{(1-\zeta)\bar{\kappa}}}{1+\frac{\zeta}{(1-\zeta)\bar{\kappa}}}.$$

So, combining the condition $\zeta \leq 1/2$ with the above inequality gives the claim. Therefore, for the following arguments, we assume that $\gamma \leq \gamma_2$. It is worthwhile to note that by the definition of $\lambda_{2,\max}$ and $\Omega_{\max} = 1 - \gamma\zeta\sigma_{\min}^2 + \Delta$, we know that $\lambda_{2,\max}$ is an increasing function of $\Omega_{\max}$ when $\Omega_{\max}^2 - 4\Delta \geq 0$ and $\Omega_{\max} \geq 0$ and $\Omega_{\max}$ is a decreasing function of γ. Therefore, $\lambda_{2,\max}$ attains its minimum at the maximum feasible learning rate γ.

First, let as assume that $\gamma \leq \gamma_3 \leq \gamma_1$. Then $\lambda_{2,\max}$ attains its minimum at $\gamma = \gamma_3$ and

$$\Omega_{\max} \geq 1 + \Delta - \zeta\sigma_{\min}^2 \cdot \frac{1}{\bar{\kappa}\sigma_{\min}^2} \cdot \frac{1-\Delta}{1-\zeta} = 1 + \Delta - \frac{\zeta}{(1-\zeta)\bar{\kappa}}(1-\Delta).$$

By observing that

$$\sqrt{\lambda_{2,\max}} = \frac{\Omega_{\max} + \sqrt{\Omega_{\max}^2 - 4\Delta}}{2},$$

we have

$$\sqrt{\lambda_{2,\max}} \geq \frac{1 + \Delta - c_1(1-\Delta) + \sqrt{1 + \Delta - c_1(1-\Delta)^2 - 4\Delta}}{2} =: f_1(\Delta),$$

where $c_1 \overset{\text{def}}{=} \frac{\zeta}{(1-\zeta)\bar{\kappa}} < 1$. One can easily verify that f_1 is an increasing function of $\Delta, 0 \leq \Delta < 1$, so we conclude that

$$\sqrt{\lambda_{2,\max}} \geq \sqrt{\lambda_{2,\max}}|_{\Delta=0} = 1 - c_1,$$

and we obtain the claim with the condition $\zeta \leq 1/2$.

Second, now we assume that $\gamma \le \gamma_1 \le \gamma_3$. Then $\lambda_{2,\max}$ attains its minimum at $\gamma = \gamma_1$ and

$$\Omega_{\max} \ge 1 + \Delta - \frac{1+\Delta}{\sigma_{\max}^2} \cdot \sigma_{\min}^2 = (1+\Delta)(1 - \frac{1}{\kappa}).$$

Therefore, for the same argument as above, we have

$$\sqrt{\lambda_{2,\max}} \ge \frac{(1+\Delta)(1-c_2) + \sqrt{(1+\Delta)^2(1-c_2)^2 - 4\Delta}}{2} =: f_2(\Delta),$$

where $c_2 \stackrel{\text{def}}{=} 1/\kappa$. On the other hand, the condition $\gamma_1 \le \gamma_3$ gives

$$\begin{aligned} \frac{1+\Delta}{\zeta\sigma_{\max}^2} \le \frac{1}{\bar{\kappa}\sigma_{\min}^2} \cdot \frac{1-\Delta}{1-\zeta} \Rightarrow \frac{1-\Delta}{1+\Delta} \ge \frac{\bar{\kappa}}{\kappa} \cdot \frac{1-\zeta}{\zeta}, \text{ or} \\ \Delta \le \frac{1 - \frac{\bar{\kappa}}{\kappa} \cdot \frac{1-\zeta}{\zeta}}{1 + \frac{\bar{\kappa}}{\kappa} \cdot \frac{1-\zeta}{\zeta}} =: \Delta_*. \end{aligned} \tag{4.49}$$

Let us define $c_3 \stackrel{\text{def}}{=} \frac{\bar{\kappa}}{\kappa} \cdot \frac{1-\zeta}{\zeta} < 1$. Then it suffices to show that $\sqrt{\lambda_{2,\max}} \ge 1 - D\frac{c_2}{c_3}$ for some $D > 0$.

Simple algebra shows that f_2 is a concave function on $[0, \Delta_u]$ where $\Delta_u \stackrel{\text{def}}{=} \frac{1-\sqrt{2c_2 - c_2^2}}{1-c_2}$ makes the radical in the numerator of f_2 vanish. Also, one can verify that $f_2(0) = 1 - c_2 \ge \frac{1-c_2+\sqrt{-2c_2+c_2^2+c_3^2}}{1+c_3} = f_2(\Delta_*)$ and $\Delta_* \le \Delta_u$, so that $f_2(\Delta) \ge f_2(\Delta_*)$ on $[0, \Delta_*]$. Hence, it suffices to show that $f_2(\Delta_*) \ge 1 - D\frac{c_2}{c_3}$ for some $D > 0$. Observe,

$$\begin{aligned} f_2(\Delta_*) &= \frac{1 - c_2 + \sqrt{-2c_2 + c_2^2 + c_3^2}}{1 + c_3} \\ &= \frac{1 - c_2 + c_3\sqrt{1 - \frac{2c_2 - c_2^2}{c_3^2}}}{1 + c_3} \\ &\ge \frac{1 - c_2 + c_3(1 - \frac{2c_2 - c_2^2}{c_3^2})}{1 + c_3} \\ &= 1 - \frac{\frac{c_2}{c_3}(2 + c_3 - c_2)}{1 + c_3} \\ &\ge 1 - 3\frac{c_2}{c_3}, \end{aligned}$$

and we finish the proof. □

Proof of Proposition 17

Proof. First, you could easily verify the following (by just setting $\Delta = 0$ in the proof of the proposition), which is an analogue of Proposition 13 for SGD without momentum.

Corollary 5. *If the learning rate $\gamma \leq \frac{1}{\zeta\sigma_{max}^2}$, with the trace condition $8(1-\zeta)\gamma \cdot \frac{1}{n}tr(\boldsymbol{A}^T\boldsymbol{A}) < 1$, then γ is in the problem constrained regime with $\varepsilon = 1/2$.*

Note that when $\Delta = 0$, $\lambda_{1,j} = 0, \lambda_{2,j} = \Omega_j^2 = (1-\gamma\zeta\sigma_j)^2, \lambda_{3,j} = 0$ so that $\lambda_{2,\max} = \Omega_{\max}^2 = (1-\gamma\zeta\sigma_{\min})^2$. Besides, the assumption on γ makes the learning rate reside in the problem constrained region by Corollary 5. Observe, when $\gamma = 1/(\zeta\sigma_{\max}^2)$,

$$\lambda_{2,\max} = (1-\gamma\zeta\sigma_{\min}^2)^2 = \Big(1-\frac{1}{\kappa}\Big)^2.$$

On the other hand, when $\gamma = \big(8(1-\zeta)\cdot\frac{1}{n}\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{A})\big)^{-1}$, we have

$$\lambda_{2,\max} = (1-\gamma\zeta\sigma_{\min}^2)^2 = \Big(1-\frac{\zeta}{8(1-\zeta)\bar{\kappa}}\Big)^2 = \Big(1-\frac{\mathcal{C}}{\bar{\kappa}}\Big)^2.$$

□

4.8 Numerical Simulations

To illustrate our theoretical results, we compare SGD+M's dynamics to (4.28) on moderately sized problems ($n \approx 1000$) under the setting of section 4.1. Moreover, the dynamics were also compared using the MNIST data set. Finally, heat maps were displayed to illustrate the interplay between the algorithmic and problem constraints. For all MNIST experiments the hyperparameters R and $\tilde{R}$ were found by running a grid-search. For simulated data experiments, we fixed R and $\tilde{R}$ and generated the data according to assumption 1.1.

Random least squares. In all simulations of the Gaussian random least squares problem, the initial weight vector $\boldsymbol{x}_0$ is set to zero and the signal and noise vectors $\tilde{\boldsymbol{x}}$ and $\boldsymbol{\eta}$ are set to $N(0, \frac{R}{n}\mathbf{I})$ and $N(0, \frac{\tilde{R}}{n}\mathbf{I})$ respectively with $\tilde{R} = R = 1$. Moreover, $\boldsymbol{A}$ is constructed by independently sampling its entries $A_{ij} \sim N(0, 1)$ then row-normalized. Similarly, $\boldsymbol{b}$ is first sampled $\boldsymbol{b} \sim N(0, \frac{\tilde{R}d}{n}\mathbf{I})$ then the i-th entry of $\boldsymbol{b}$ is divided by the norm of the i-th row of $\boldsymbol{A}$. The objective function in which we run SGD+M in all cases is the least squares objective function $f(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||^2$.

Random features (RF). In Figure 4.2, we generate the data matrix $\boldsymbol{A}$ using a random features set-up. The model was introduced by [6] as a randomized approach for scaling kernel methods to large data sets, and has seen a surge of interest in recent years as a way to study the generalization properties of neural networks [4, 25, 33, 37]. RF is a way to increase the number of parameters without changing the data set for a least-squares problem.

In this model, the entries of $\boldsymbol{A}$ are the result of a matrix multiplication composed with a (potentially non-linear) activation function $\sigma : \mathbb{R} \to \mathbb{R}$:

$$A_{ij} \stackrel{\text{def}}{=} \sigma\left(\frac{[\boldsymbol{X}\boldsymbol{W}]_{ij}}{\sqrt{n_0}}\right), \quad \text{where } \boldsymbol{X} \in \mathbb{R}^{n \times n_0} \text{ and } \boldsymbol{W} \in \mathbb{R}^{n_0 \times d}. \tag{4.50}$$

The entries of $\boldsymbol{W}$ (in Fig 4.2) are i.i.d. with zero mean and variance 1. The data matrix $\boldsymbol{X}$ is the MNIST data set where each row of $\boldsymbol{X}$ is an image (i.e., $n_0 = 784$). In these experiments, the activation function σ is the normalized ReLU function $\sigma(\cdot) = (\max\{0, \cdot\} - 1/\sqrt{2\pi})/\sqrt{0.5 - 1/(2\pi)})$; it is normalized so that σ applied to a standard Gaussian outputs a mean 0 and variance 1 random variable (not necessarily Gaussian).

Empirical Volterra equation. We assume that we have access to the eigenvalues of the matrix $\boldsymbol{A}\boldsymbol{A}^T$. The empirical Volterra equation (4.28) were computed using a dynamic programming approach by using as inputs the eigenvalues of $\boldsymbol{A}\boldsymbol{A}^T$. First, the values of $h_0(t), h_1(t), H_2(t)$ were computed and stored for values of $t \in [T]$. Then a dynamic

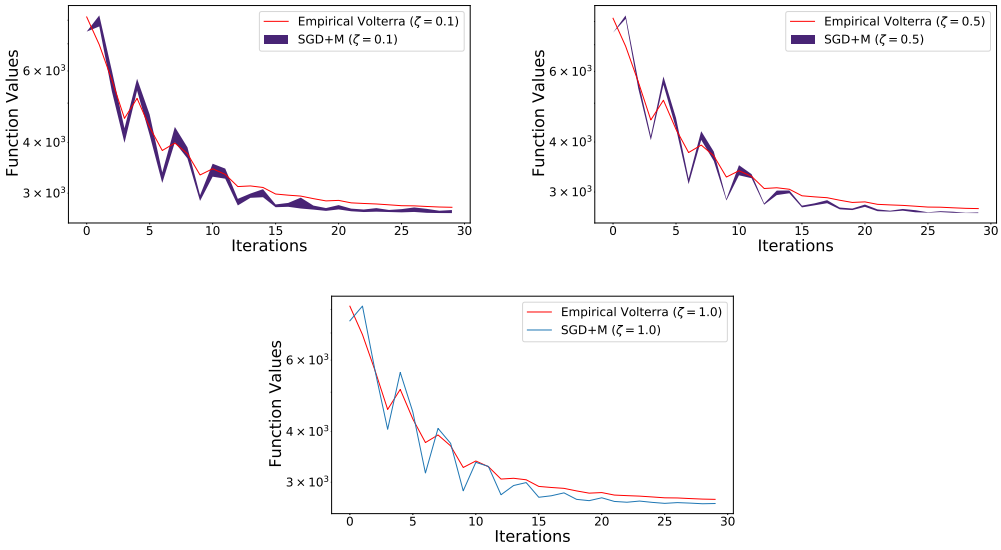

Figure 4.4: SGD+M vs. Theory on even/odd MNIST. MNIST ($60,000 \times 28 \times 28$ images) [31] is reshaped into a single matrix of dimension $60,000 \times 784$ (preconditioned to have centered rows of norm-1), representing 60,000 samples of 10 digits. The target $\boldsymbol{b}$ satisfies $\boldsymbol{b}_i = 0.5$ if the i^{th} sample is an odd digit and $\boldsymbol{b}_i = -0.5$ otherwise. SGD+M was run 10 times with $\Delta = 0.8$, various values of ζ, and learning rates $\gamma = 0.005, 0.001, 0.0005$ (left to right, top to bottom) and empirical Volterra was run once with $(R = 11,000, \tilde{R} = 5300)$. The R and $\tilde{R}$ values were found by running a grid-search. The 10^{th} to 90^{th} percentile interval is displayed for the loss values of 10 runs of SGD+M. Volterra predicts the convergent behavior of SGD+M in this setting.

programming approach is used to compute $\psi(t)$ for values of $t \in [T]$. The discrete convolution operation in (4.28) is computed by an array reversal and Numpy dot product.

Volterra equation with Marchenko-Pastur distribution. In this setting, we use the theoretical limiting distribution for a large class of random matrices. In a celebrated work by [36], when the entries of $(n \times d)$ matrix $\boldsymbol{A}$ are drawn from a common, mean 0, variance $1/d$ distribution with fourth moment $\mathcal{O}(d^{-2})$ (e.g., Gaussian $N(0, \frac{1}{d})$), it is known that the distribution of eigenvalues of $\boldsymbol{A}\boldsymbol{A}^T$ converges to the Marchenko-Pastur law

$$d\mu_{MP}(\lambda) \overset{\text{def}}{=} \delta_0(\lambda)\max\{1-r,0\} + \frac{r\sqrt{(\lambda-\lambda^-)(\lambda^+-\lambda)}}{2\pi\lambda}1_{[\lambda^-,\lambda^+]},$$
$$\text{where} \quad \lambda^- \overset{\text{def}}{=} (1-\sqrt{\frac{1}{r}})^2 \quad \text{and} \quad \lambda^+ \overset{\text{def}}{=} (1+\sqrt{\frac{1}{r}})^2. \tag{4.51}$$

For these experiments, we generated the data matrix $\boldsymbol{A}$ with entries $N(0, 1/d)$. Instead of using the eigenvalues of $\boldsymbol{A}\boldsymbol{A}^T$ in the Volterra equation (4.28), we used the Marchenko-Pastur distribution directly. We used a Chebyshev quadrature rule to approximate the integrals with respect to the Marchenko-Pastur distribution that arise in (4.28). Similar to the finite case, the integrand is computed using dynamic-programming. However, the implementation of the quadrature rule ignores the point mass at 0 so we manually add this at the end.

Volterra simulations remarks. Despite the numerical approximations to the integral, the resulting solution to the Volterra equation ψ models the true behavior of SGD+M remarkably well. Notably, the fit of the Volterra equation to SGD+M is extremely accurate across various learning rates, batch sizes, and momentum parameters as long as the learning rate condition is satisfied. In Figure 1, the red line corresponds to the Volterra equation with Marchenko-Pastur distribution with values $R = \tilde{R} = 1$. Also, we opted to shade the 10^{th} to 90^{th} percentile instead of an α-confidence interval for an easier read. One can observe the exact same dynamics in either case.

Heat maps. The heat maps (Figures 4.1, 4.5, and 4.7) illustrate when the convergence rate is dictated by the problem, ($\lambda_{2,\max} \geq \Xi^{-1}$) or by the algorithm ($\lambda_{2,\max} < \Xi^{-1}$). The white regions of the heat maps represent divergent behaviour ($\lambda_{2,\max} > 1$). The threshold, denoted by the red line, describes the boundary for two different regimes. Any non-white point above or to the right of the threshold lies in the algorithmic constraint setting. Conversely, all non-white points lying below or to the left of the threshold lies in the problem constraint setting.

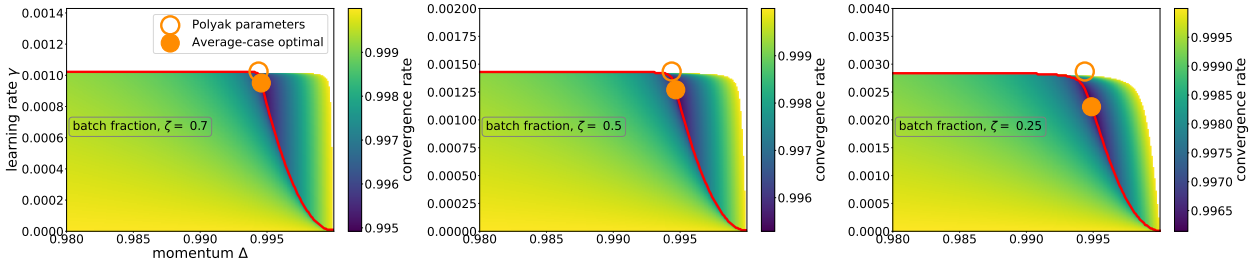

Figure 4.5: Different convergence rate regions for MNIST dataset. Plots are functions of momentum (x-axis) and learning rate (y-axis). Optimal parameters that maximize $\lambda_{2,\max}$ denoted by Polyak parameters (orange circle, (4.15)) and the optimal parameters for SGD+M (orange dot); below red line is the problem constrained region; otherwise the algorithmic constrained region. The MNIST data set is standardized. As the batch fraction decreases (left $\zeta = 0.7$ to right $\zeta = 0.25$), the optimal parameters of SGD+M and Polyak parameters are quite far from each other. The Malthusian exponent (algorithmically constrained region) starts to control the SGD+M rate as batch fraction $\to 0$.

The heat maps are generated by computing $\lambda_{2,\max}$ and Ξ (when it exists) across values of (Δ, γ). Here $\lambda_{2,\max}$ is obtained by calculating

$$\lambda_{2,\max} = \frac{-2\Delta + \Omega_{\max}^2 + \sqrt{\Omega_{\max}^2(\Omega_{\max}^2 - 4\Delta)}}{2}, \; \Omega_{\max} = 1 - \gamma\zeta\sigma_{\min}^2 + \Delta, \text{ and } \sigma_{\min}^2 = \left(1 - \sqrt{\frac{1}{r}}\right)^2.$$

In order to compute Ξ, recall that Ξ is the solution of

$$\tilde{K}(\Xi) \stackrel{\text{def}}{=} \sum_{t=0}^{\infty} \Xi^t K(t) = 1, \tag{4.52}$$

when it exists. One can show (4.52) is equal to (see Appendix 4.7.3 for detail)

$$\sum_{j=1}^{n} \frac{\zeta(1-\zeta)\gamma^2\sigma_j^4}{n} \left(\frac{(1+\Xi\Delta)}{(1-\Xi\Delta)(1+\Xi\Delta+\sqrt{\Xi}\Omega_j)(1+\Xi\Delta-\sqrt{\Xi}\Omega_j)} \right) = 1, \tag{4.53}$$

which is computed using the Chebyshev quadrature rule.

For a given (Δ, γ), we are interested in the algorithmic case ($1 \leq \Xi \leq \lambda_{2,\max}^{-1}$) so if $\lambda_{2,\max}^{-1} < 1$ we assign a Nan value to Ξ. Otherwise, because of monotonicity of $\tilde{K}$ in (4.52), we perform a binary search starting with initial endpoints 1 and $\lambda_{2,\max}^{-1}$ to find the solution

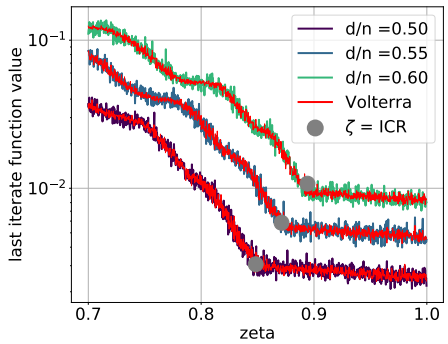

Figure 4.6: Convergence behavior near the ICR. For each value of batch fraction ζ, run SGD+M for 20 (left) and 50 (right) iterations (colored lines – blue, green, and purple) and record the function value of the last iterate. The momentum and learning rate parameters are set to be near-optimal (see (4.20)). Gray dot is the computed ICR (4.22), ζ value. Data matrix $\boldsymbol{A} \in \mathbb{R}^{d,n}$ Gaussian entries, $\tilde{\boldsymbol{x}} \sim N(\mathbf{0}, 1/n\mathbf{I}_d)$, $\boldsymbol{x}_0 = \mathbf{0}$ (R = 1.0), and $\boldsymbol{\eta} \sim N(\mathbf{0}, 0.0001/n\mathbf{I}_n)$ ($\tilde{R} = 0.0001$). Different colored lines (blue, green, purple) correspond to running SGD+M with different values of the ratio d/n. At the predicted $\zeta =$ ICR (gray dot), there is a noticeable change in the behavior of the last iterate. For ζ values less than the ICR, the value of the last iterate gets smaller as ζ increases. Then the batch fraction ζ hits the ICR and we see little to no improvement in the value of the last iterate. This agrees *exactly* with our theory for batch fraction saturation (Prop 15 and Prop. 16). For $\zeta \geq$ ICR, the convergence rate does not change; thus the values of the last iterates are approximately all equal in this regime. For $\zeta <$ ICR, our theory predicts the convergence rate improves as $\zeta \to$ ICR, $\mathcal{O}(\zeta/\bar{\kappa})$. Hence the value of the last iterate decreases here. Moreover (left), SGD+M dynamics match the predicted last value given by the Volterra equation (red) (see Thm 7).

Ξ satisfying (4.52). Finally, with Ξ^{-1} and $\lambda_{2,\max}$ computed for a given (Δ, γ), we plot the maximum of the two.

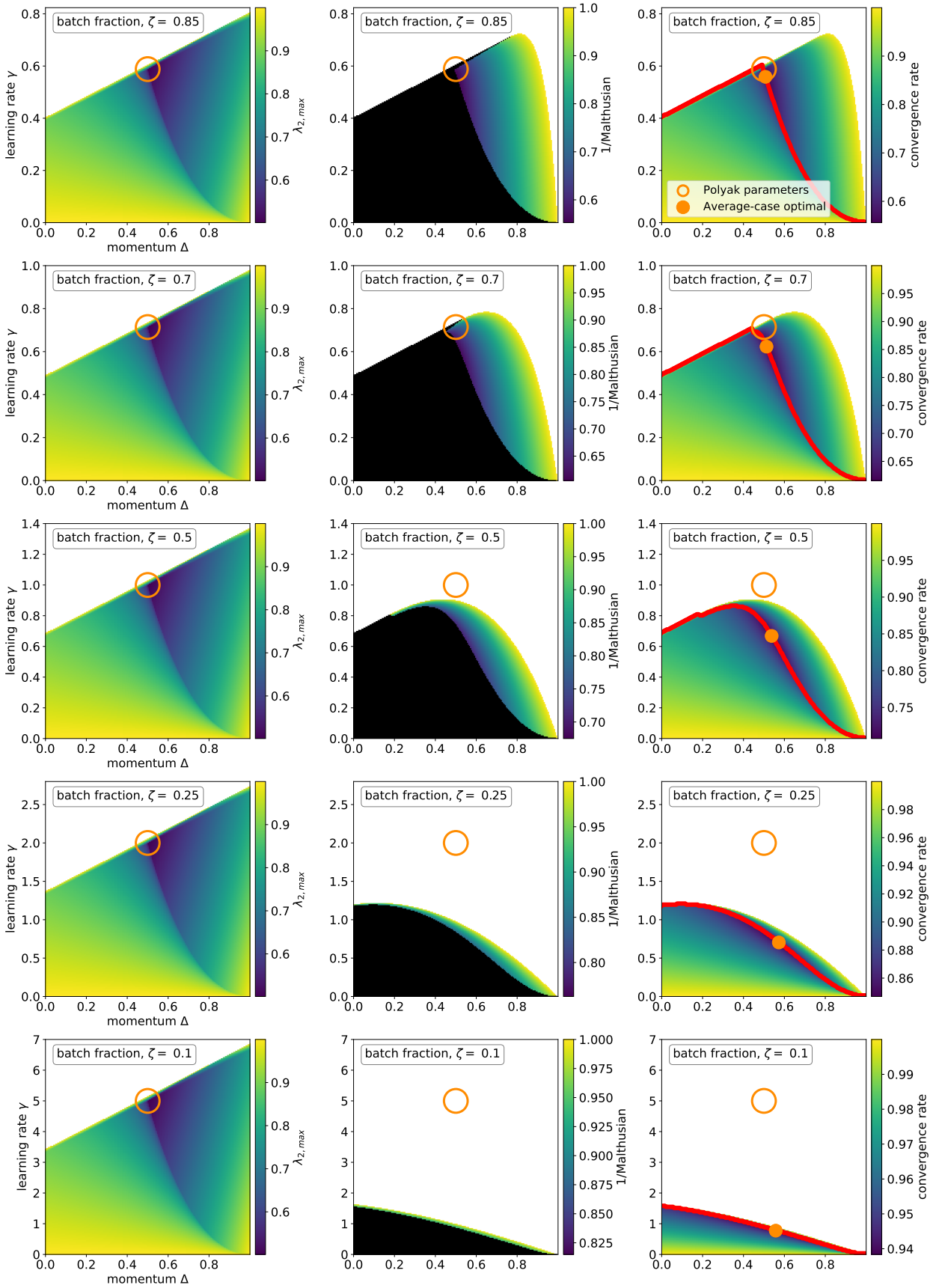

Figure 4.7: Convergence rate regions for Gaussian random least squares. Same set-up as in Figure 4.1 but for a wider range of batch fractions.

Bibliography

[1] Chatgpt description. `https://openai.com/blog/chatgpt/`. Accessed: 2023-01-20.

[2] ADAMCZAK, R. A note on the Hanson-Wright inequality for random vectors with dependencies. *Electronic Communications in Probability 20*, none (2015), 1 – 13.

[3] ADLAM, B., AND PENNINGTON, J. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *Proceedings of the 37th International Conference on Machine Learning* (13–18 Jul 2020), H. D. III and A. Singh, Eds., vol. 119 of *Proceedings of Machine Learning Research*, PMLR, pp. 74–84.

[4] ADLAM, B., AND PENNINGTON, J. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning* (2020).

[5] AGRESTI, A. Foundations of linear and generalized linear models.

[6] ANFINSEN, C. Principles that govern the folding of protein chains. *Science (New York, N.Y.) 181*, 4096 (July 1973), 223–230.

[7] ANFINSEN, C. Principles that govern the folding of protein chains. *Science (New York, N.Y.) 181*, 4096 (July 1973), 223–230.

[8] ASMUSSEN, S. *Applied probability and queues*, second ed., vol. 51 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 2003. Stochastic Modelling and Applied Probability.

[9] BARDENET, R., AND M., O.-A. Concentration inequalities for sampling without replacement. *Bernoulli 21*, 3 (2015), 1361 – 1385.

[10] BILLINGSLEY, P. *Probability and Measure*, second ed. John Wiley and Sons, 1986.

[11] BOLLAPRAGADA, R., CHEN, T., AND WARD, R. On the fast convergence of minibatch heavy ball momentum, 2022.

[12] BOTTOU, L., CURTIS, F. E., AND NOCEDAL, J. Optimization methods for large-scale machine learning, 2016.

[13] BOYD, S., AND VANDENBERGHE, L. *Convex optimization*. Cambridge university press, 2004.

[14] BRUNNER, H. *Volterra Integral Equations: An Introduction to Theory and Applications*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2017.

[15] BUBECK, S. Convex optimization: Algorithms and complexity, 2015.

[16] CHANDRASEKHER, K. A., PANANJADY, A., AND THRAMPOULIDIS, C. Sharp global convergence guarantees for iterative nonconvex optimization: A gaussian process perspective, 2021.

[17] COUILLET, R., AND LIAO, Z. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022.

[18] DEIFT, P., AND TROGDON, T. Universality in numerical computation with random data: Case Studies, Analytical Results, and Some Speculations. *Abel Symposia 13*, 3 (2018), 221–231.

[19] DUPUY, C., ARAVA, R., GUPTA, R., AND RUMSHISKY, A. An efficient dp-sgd mechanism for large scale nlp models, 2022.

[20] DURRETT, R. *Probability: Theory and Examples*, 5 ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[21] ENGELI, M., GINSBURG, T., RUTISHAUSER, H., AND STIEFEL, E. Refined iterative methods for computation of the solution and the eigenvalues of self-adjoint boundary value problems. *Mitt. Inst. Angew. Math. Zürich 8* (1959), 107.

[22] GOLUB, G., AND VARGA, R. Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods. *Numerische Mathematik* (1961).

[23] GOLUB, G. H., AND VAN LOAN, C. F. *Matrix Computations*, third ed. The Johns Hopkins University Press, 1996.

[24] GRONWALL, T. H. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Ann. of Math. (2) 20*, 4 (1919), 292–296.

[25] HASTIE, T., MONTANARI, A., ROSSET, S., AND TIBSHIRANI, R. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560* (2019).

[26] INABA, H. *The Stable Population Model*. Springer Singapore, Singapore, 2017, pp. 1–74.

[27] JAIN, P., KAKADE, S., KIDAMBI, R., NETRAPALLI, P., AND SIDFORD, A. Accelerating Stochastic Gradient Descent for Least Squares Regression. In *Proceedings of the 31st Conference On Learning Theory (COLT)* (2018), vol. 75 of *Proceedings of Machine Learning Research*, PMLR, pp. 545–604.

[28] JASTRZEBSKI, S., KENTON, Z., ARPIT, D., BALLAS, N., FISCHER, A., BENGIO, Y., AND STORKEY, A. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623* (2017).

[29] JUMPER, J. M., EVANS, R., PRITZEL, A., GREEN, T., FIGURNOV, M., RONNEBERGER, O., TUNYASUVUNAKOOL, K., BATES, R., ZÍDEK, A., POTAPENKO, A., BRIDGLAND, A., MEYER, C., KOHL, S. A. A., BALLARD, A., COWIE, A., ROMERA-PAREDES, B., NIKOLOV, S., JAIN, R., ADLER, J., BACK, T., PETERSEN, S., REIMAN, D. A., CLANCY, E., ZIELINSKI, M., STEINEGGER, M., PACHOLSKA, M., BERGHAMMER, T., BODENSTEIN, S., SILVER, D., VINYALS, O., SENIOR, A. W., KAVUKCUOGLU, K., KOHLI, P., AND HASSABIS, D. Highly accurate protein structure prediction with alphafold. *Nature 596* (2021), 583 – 589.

[30] KIDAMBI, R., NETRAPALLI, P., JAIN, P., AND KAKADE, S. M. On the insufficiency of existing momentum schemes for stochastic optimization, 2018.

[31] LECUN, Y., CORTES, C., AND BURGES, C. "MNIST" handwritten digit database, 2010.

[32] LEE, K., CHENG, A., PAQUETTE, E., AND PAQUETTE, C. Trajectory of mini-batch momentum: Batch size saturation and convergence in high dimensions. In *Advances in Neural Information Processing Systems* (2022), S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, Curran Associates, Inc., pp. 36944–36957.

[33] LIAO, Z., AND COUILLET, R. The Dynamics of Learning: A Random Matrix Approach. *Proceedings of the 35th International Conference on Machine Learning (ICML)* (2018).

[34] LJUNG, L. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control 22*, 4 (1977), 551–575.

[35] MA, S., BASSILY, R., AND BELKIN, M. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning, 2018.

[36] MARÄENKO, V. A., AND PASTUR, L. A. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik 1*, 4 (apr 1967), 457.

[37] MEI, S., AND MONTANARI, A. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics 75* (06 2021).

[38] MONTANARI, A., AND SAEED, B. N. Universality of empirical risk minimization. In *Proceedings of Thirty Fifth Conference on Learning Theory* (02–05 Jul 2022), P.-L. Loh and M. Raginsky, Eds., vol. 178 of *Proceedings of Machine Learning Research*, PMLR, pp. 4310–4312.

[39] NAKKIRAN, P., KAPLUN, G., BANSAL, Y., YANG, T., BARAK, B., AND SUTSKEVER, I. Deep double descent: where bigger models and more data hurt*. *Journal of Statistical Mechanics: Theory and Experiment 2021*, 12 (dec 2021), 124003.

[40] NESTEROV, Y. On an approach to the construction of optimal methods of minimization of smooth convex functions. *Ekonom. i. Mat. Metody 24* (1988), 509–517.

[41] NESTEROV, Y. *Introductory lectures on convex optimization*. Springer, 2004.

[42] PAQUETTE, C., LEE, K., PEDREGOSA, F., AND PAQUETTE, E. SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality. In *Proceedings of Thirty Fourth Conference on Learning Theory (COLT)* (2021), vol. 134 of *Proceedings of Machine Learning Research*, pp. 3548–3626.

[43] PAQUETTE, C., LEE, K., PEDREGOSA, F., AND PAQUETTE, E. Sgd in the large: Average-case analysis, asymptotics, and stepsize criticality. In *Proceedings of Thirty Fourth Conference on Learning Theory* (15–19 Aug 2021), M. Belkin and S. Kpotufe, Eds., vol. 134 of *Proceedings of Machine Learning Research*, PMLR, pp. 3548–3626.

[44] PAQUETTE, C., AND PAQUETTE, E. Dynamics of stochastic momentum methods on large-scale, quadratic models. In *Advances in Neural Information Processing Systems* (2021), M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., pp. 9229–9240.

[45] PAQUETTE, C., PAQUETTE, E., ADLAM, B., AND PENNINGTON, J. Homogenization of sgd in high-dimensions: Exact dynamics and generalization properties, 2022.

[46] PAQUETTE, C., VAN MERRIËNBOER, B., AND PEDREGOSA, F. Halting Time is Predictable for Large Models: A Universality Property and Average-case Analysis. *Foundations of Computational Mathematics* (2022).

[47] PEDREGOSA, F. A hitchhiker's guide to momentum, 2021.

[48] PEDREGOSA, F., AND SCIEUR, D. Acceleration through spectral density estimation. In *Proceedings of the 37th International Conference on Machine Learning* (13–18 Jul 2020), H. D. III and A. Singh, Eds., vol. 119 of *Proceedings of Machine Learning Research*, PMLR, pp. 7553–7562.

[49] POLYAK, B. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics 04* (1964).

[50] ROBBINS, H., AND MONRO, S. A Stochastic Approximation Method. *Ann. Math. Statist.* (1951).

[51] SILVER, D., HUANG, A., MADDISON, C. J., GUEZ, A., SIFRE, L., VAN DEN DRIESSCHE, G., SCHRITTWIESER, J., ANTONOGLOU, I., PANNEERSHELVAM, V., LANCTOT, M., DIELEMAN, S., GREWE, D., NHAM, J., KALCHBRENNER, N., SUTSKEVER, I., LILLICRAP, T., LEACH, M., KAVUKCUOGLU, K., GRAEPEL, T., AND HASSABIS, D. Mastering the game of Go with deep neural networks and tree search. *Nature 529*, 7587 (jan 2016), 484–489.

[52] SILVER, D., SCHRITTWIESER, J., SIMONYAN, K., ANTONOGLOU, I., HUANG, A., GUEZ, A., HUBERT, T., BAKER, L., LAI, M., BOLTON, A., CHEN, Y., LILLICRAP, T., HUI, F., SIFRE, L., DRIESSCHE, G., GRAEPEL, T., AND HASSABIS, D. Mastering the game of go without human knowledge. *Nature 550* (10 2017), 354–359.

[53] SONG, S., CHAUDHURI, K., AND SARWATE, A. D. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing* (2013), pp. 245–248.

[54] VERSHYNIN, R. High-dimensional probability.

[55] WISHART, J. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika 20A*, 1/2 (1928), 32–52.

[56] YU, D., KAMATH, G., KULKARNI, J., LIU, T.-Y., YIN, J., AND ZHANG, H. Individual privacy accounting for differentially private stochastic gradient descent, 2023.

[57] ZHANG, G., LI, L., NADO, Z., MARTENS, J., SACHDEVA, S., DAHL, G. E., SHALLUE, C. J., AND GROSSE, R. Which algorithmic choices matter at which batch sizes? insights from a noisy quadratic model, 2019.