High-dimensional Limit of SGD for Adaptive Stepsize Algorithms and Diagonal Linear Networks

Begoña García Malaxechebarría

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Dmitriy Drusvyatskiy, Chair

Courtney Paquette

Maryam Fazel

Program Authorized to Offer Degree:
Mathematics

University of Washington

Abstract

High-dimensional Limit of SGD
for Adaptive Stepsize Algorithms
and Diagonal Linear Networks

Begoña García Malaxechebarría

Chair of the Supervisory Committee:
Professor and HDSI Faculty Fellow Dmitriy Drusvyatskiy
Halıcıoğlu Data Science Institute (HDSI)
University of California San Diego

Although stochastic gradient methods are widely used in practice, a complete theoretical understanding of their dynamics is still lacking. Classical analyses typically rely on asymptotic regimes with vanishing stepsizes, leading to continuous-time approximations such as deterministic gradient flow or small-noise diffusion models, which often fail to capture the rich behavior observed in practice. This thesis develops a unified framework for describing stochastic gradient descent (SGD) at finite stepsizes through high-dimensional limits, where algorithmic randomness gives rise to deterministic evolution laws for key quantities of interest.

We begin by studying adaptive stepsize methods in high-dimensional optimization problems, which we refer to as the *high line*. In this setting, we show that both the risk and the stepsize dynamics of one-pass SGD admit exact deterministic descriptions via a system of ordinary differential equations. This perspective enables a precise analysis of commonly used adaptive strategies. In particular, we demonstrate that idealized line search procedures can exhibit arbitrarily slow convergence compared to optimally tuned fixed stepsizes, even in simple least squares problems. We further characterize the long-time behavior of adaptive methods such as AdaGrad-Norm, showing that their stepsizes converge to explicit deterministic limits governed by the spectral properties of the data covariance, and uncover phase transitions under power-law eigenvalue distributions.

We then turn to diagonal linear networks as a canonical model for understanding neural optimization. In the high-dimensional setting, the trajectory of stochastic gradient descent admits a continuous-time stochastic representation, formalized through a stochastic differential equation, in which the deterministic and stochastic components of the dynamics are explicitly disentangled. This representation induces a closed deterministic evolution for a collection of low-dimensional summary statistics, including measures of risk and curvature. Leveraging this structure, we obtain a sharp description of the global behavior of the dynamics, establishing well-posedness and exponential convergence to zero risk with high probability.

Together, these results demonstrate that, in high-dimensional settings, stochastic gradient methods admit precise deterministic descriptions that capture both optimization performance and stepsize behavior. This perspective offers a new lens on algorithmic design and analysis, bridging stochastic optimization, high-dimensional probability, and dynamical systems.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

Table Number Page

ACKNOWLEDGMENTS

First, I would like to thank my advisor, Dmitriy Drusvyatskiy. When we met in the spring of 2023, I was deeply apathetic and determined to drop out the PhD program. After moving overseas and leaving my country in the fall of 2021, I had struggled for more than a year to adapt to my new life and find any motivation, something that eventually affected my self-esteem and took a toll on my mental health. Determined to master out, I decided to engage in applied courses that would best prepare me for industry. In one of these courses, Ranjay Krishna first exposed me to computer vision, and this experience awakened a deep-rooted passion for machine learning in me. I was very lucky that I decided to ask Rekha Thomas for her personal advice on my situation, and she suggested that I talk to Dima. That day, Dima asked me if I wanted to start working together, and I, completely overwhelmed by my circumstances, rejected his offer. But life sometimes has other plans for you, and it was not until I decided that I could allow myself to "fail" that things started to improve for me. When I began feeling better, I contacted Dima to ask whether he still wanted to read with me, and the rest is history: I have been working by his side since then, and he has helped turn me from a young student into the researcher I am today. I have learned so much technical knowledge from him, as well as academic best practices and how to read, write, and polish papers; but more importantly, he saved me in a moment when I was completely lost. He was able to see potential in me when I wanted to give up, and he had faith and trust in me when I did not even believe in myself. For that, I will be endlessly grateful.

I met my unofficial advisor, Courtney Paquette, during her visit to UW in late 2023. Since then, I have had no doubt that she was sent to me as a guardian angel. Her insatiable thirst for knowledge and unquestionable talent have helped me grow and inspired me over the years. She is the most successful human being I know and has been a role model for me

since the day we met. I am also beyond thankful for every single time she and Elliot have hosted me in Montréal, a place that will always be very special to me.

Next, I would like to thank my committee members, Maryam Fazel and Ranjay Krishna, for their continued support throughout this journey. It has been a pleasure to collaborate with Maryam and learn from her. My undergraduate advisor, Ángel Ferrández Izquierdo, is also a big reason why I could make it to the finish line. His great commitment to helping younger generations in the academic community is something that I deeply admire about him. He helped me succeed in my very first experience across the pond in San Diego and later encouraged me to pursue my PhD studies, always picking up my calls, listening to me with care, and grabbing a coffee with me during every single one of my visits home.

I would not be the person I am without the continued support and sacrifice of my family. Although family cannot be chosen, I was lucky enough to get the best one I could have hoped for. My grandparents have been an example of hard work and love that has shaped my personality, and my parents have raised me with the selfless values they established: putting family first, even before one's own health. I am also very grateful to them for giving me the best present I could ever have, my brother Iñaki, who has taught me one of the biggest lessons in life: unconditional love.

The family I chose has been extraordinary too. I like to think that I have become skilled in the art of surrounding myself with people of exceptional quality, and this skill has brought Ana, Inma, Hernán, Patri, Carmen, Jiayi, Apra, Diego, and Ángela, among others, into my life. There are also many people who, although no longer part of my everyday life, have stayed in my heart: Puchun, Reyes, Susana, Cristina, Jonathan, and Melissa. The Spanish community in Seattle has given me a sense of home, joy, and emotional stability, making my life here a little better every day. Grabbing a beer at a brewery with them has become my favorite Friday tradition; special thanks to Yara, Alfredo, Álvaro, Eva, Juanjo, Aimee, Cristina, Paco, Lucía, Rodrigo, and all the Alexes. Running has given me the discipline that the last few years of this adventure required, and the Brooks run club has kept me sane despite all the uphills. Special shoutout to Erin and AV, although I would run out of space if

I tried to mention all the big-hearted runners I have met there. I am also grateful for the math people this experience has given me, especially Juan, Yirong, and Sarafina. Finally, I will always carry my brief but unforgettable Montréal community in my heart: Vincent, Geneviève, Nina, Luna, Tess, Noah, Kelli, and the rest of the MRRC run club.

Finally, I want to thank my younger, more innocent self, who dared to embark on this challenging and wild life experience and was much braver than she could ever have imagined.

DEDICATION

to my abuelo, who taught me to love to learn;
and to my aitite, who taught me to learn to love

Chapter 1

INTRODUCTION

Modern machine learning operates in regimes where both the parameter dimension and the sample size are very large. In such settings, stochastic optimization algorithms often exhibit a remarkable form of concentration: despite their inherent randomness, their trajectories display stable and predictable behavior. In machine learning, this phenomenon appears in stochastic gradient descent (SGD), where training curves become stable, risk trajectories concentrate, and the dynamics are often reproducible across random data samples. These observations raise a fundamental question: why do high-dimensional stochastic optimization algorithms give rise to deterministic laws at the level of observable quantities?

This perspective is reminiscent of statistical physics, where microscopic randomness in large interacting systems gives rise to predictable macroscopic behavior through averaging effects. Mean-field equations, for example, describe the evolution of aggregate observables without tracking every random microscopic interaction. Despite their different motivations, both lines of work reveal the same underlying phenomenon: stochastic dynamics concentrate around deterministic limits in high dimensions. In this spirit, this thesis develops a mathematical framework for understanding stochastic optimization algorithms at scale.

Since the exact analysis of stochastic optimization algorithms is intractable, it is necessary to adopt reduced descriptions that capture their essential behavior. A classical approach is to approximate the discrete dynamics by continuous-time models. When the stepsize tends to zero, SGD is well approximated by gradient flow, a deterministic ordinary differential equation describing the evolution of the parameters. A refined description incorporates stochasticity through stochastic differential equations (SDEs), yielding stochastic gradient flow [75]. These diffusion approximations provide an explicit characterization of the noise in SGD and have been widely used in stochastic approximation theory. While powerful, they rely on vanishing stepsizes and are typically derived in fixed-dimensional settings. As a result, they do not

capture regimes with finite or large stepsizes, where qualitatively different phenomena emerge. In particular, they fail to describe structured noise effects such as transient growth in risk and rapid changes in curvature observed in modern practice.

In regimes with large stepsizes, SGD empirically exhibits stochastic effects not captured by classical theory, including transient growth in risk, rapid changes in curvature, and eventual stabilization. These are closely associated with phenomena such as edge-of-stability behavior, the catapult mechanism, and progressive sharpening. These observations highlight a gap in existing theory. Classical diffusion approximations are not valid beyond the small stepsize regime, and there is no clear small parameter governing the dynamics that would justify a perturbative analysis. This motivates the search for alternative asymptotic regimes.

In this work, we consider a complementary asymptotic regime in which the dimension tends to infinity while the stepsize remains fixed, a setting motivated by modern machine learning applications where both quantities can be large. In high dimensions, the dynamics of SGD exhibit concentration phenomena: key observable statistics evolve deterministically in the limit. This suggests the existence of a deterministic equivalent governing their evolution.

The behavior of SGD in this regime admits a multiscale description:

- **Microscopic level:** the discrete stochastic dynamics of SGD,
- **Mesoscopic level:** a homogenized stochastic differential equation describing the evolution of the parameters,
- **Macroscopic level:** a deterministic partial integro-differential equation governing the evolution of observable statistics.

These levels provide a unified framework linking stochastic optimization, diffusion approximations, and deterministic limits. The homogenized SDE provides a computationally efficient description of the stochastic dynamics of SGD, while the deterministic equivalent predicts the evolution of observable statistics without stochastic fluctuations. Together, these tools offer a reduced description of high-dimensional learning dynamics.

The main contributions of this dissertation are developed in Chapters 2 and 3, each corresponding to a research paper.

Chapter 2 studies the high-dimensional behavior of stochastic gradient descent for adaptive stepsize algorithms, with the goal of understanding how data-dependent stepsizes interact with stochastic dynamics in high dimensions. We show that similar high-dimensional limits can be derived in this setting, yielding effective descriptions of the dynamics and revealing new phenomena specific to adaptive methods.

Chapter 3 extends this framework to diagonal linear networks. In this setting, we show that the dynamics of SGD are accurately described by a homogenized stochastic differential equation, which captures the stochastic evolution of the parameters at finite stepsizes. Building on this representation, we derive a deterministic equivalent governing the evolution of observable statistics, providing a precise description of quantities such as the risk and curvature along the trajectory.

The organization of the dissertation is as follows. Chapter 1 provides an introduction and background. Chapter 2 focuses on adaptive stepsize algorithms and their high-dimensional behavior. Chapter 3 presents the high-dimensional limit of SGD for diagonal linear networks and develops the associated stochastic and deterministic descriptions.

The material in Chapters 2 and 3 is based on the following papers:

- Elizabeth Collins-Woodfin, Inbar Seroussi, Begoña García Malaxechebarría, Andrew W. Mackenzie, Elliot Paquette, and Courtney Paquette. The high line: Exact risk and stepsize curves of stochastic adaptive stepsize algorithms. *NeurIPS*, 2024

- Begoña García Malaxechebarría, Courtney Paquette, Maryam Fazel, and Dmitriy Drusvyatskiy. High-dimensional Limit of SGD for Diagonal Linear Networks. *Preprint arXiv:*, 2026

Together, these chapters develop a precise understanding of stochastic optimization dynamics in high-dimensional regimes at finite stepsizes.

1.1 High-dimensional Model Structure

We begin by describing the common high-dimensional framework underlying the results in Chapters 2 and 3. In both settings, we study SGD applied to optimization problems defined

through a population risk of the form

$$\mathcal{R}(x) = \mathbb{E}_{a,\varepsilon}\left[f\big(\langle\beta(x), a\rangle, \langle\beta^*, a\rangle, \varepsilon\big)\right], \tag{1.1}$$

where $a \in \mathbb{R}^d$ denotes high-dimensional Gaussian data with $a \sim \mathcal{N}(0, K)$, $\beta(x) \in \mathbb{R}^d$ is a parameter-dependent feature map, $\beta^* \in \mathbb{R}^d$ represents a target parameter, and f is an α-pseudo-Lipschitz loss (e.g., mean-squared error or logistic loss).

The covariance matrix $K \in \mathbb{R}^{d\times d}$ is assumed to have uniformly bounded operator norm, $\|K\|_{\mathrm{op}} \leqslant \bar{K}$ for some constant $\bar{K} > 0$ independent of d. In addition, we assume that the trace of K scales linearly with the dimension, $\mathrm{Tr}(K) = \Theta(d)$, reflecting that the total variance of the input distribution remains proportional to the ambient dimension. Our results are calibrated for this high-dimensional regime. When $\mathrm{Tr}(K)$ grows more slowly than d, different stepsize scalings may be required to obtain nontrivial limiting behavior.

In Chapter 2, the parameter map is linear, $\beta(x) = x \in \mathbb{R}^d$, and the data $(a, \varepsilon) \in \mathbb{R}^d \times \mathbb{R}$ include additive Gaussian noise with $\varepsilon \sim \mathcal{N}(0, \omega^2)$, while the covariance matrix K is general. This framework applies to several fundamental models in machine learning, including linear regression, simplified neural network training models, and multi-class logistic regression.

In Chapter 3, the parameter $\beta(x) \in \mathbb{R}^d$ is obtained through a component-wise quadratic feature map ψ, where $x \in \mathbb{R}^{2d}$. In this setting, the data are noiseless and consist only of $a \sim \mathcal{N}(0, K)$, with K assumed diagonal. Due to the high-dimensional scaling, the inner products $\langle\beta(x), a\rangle$ and $\langle\beta^*, a\rangle$ are normalized by $1/\sqrt{d}$ before being evaluated by f. A central example throughout the chapter is the two-layer diagonal linear network, where $x = (u, v)$ and $\beta(x) = u \odot v$. These models have been widely studied as a minimal nonconvex setting that captures phenomena such as implicit regularization and sparse recovery [38, 47, 75, 83].

1.2 Algorithm Formulation

To solve (1.1), we consider a *streaming* (or *one-pass*) stochastic gradient descent scheme. At each iteration, a fresh sample $a_{k+1} \sim \mathcal{N}(0, K)$ is drawn, and the iterates are updated according to the recurrence

$$x_{k+1} = x_k - \gamma_k \nabla_x \Psi(x_k; a_{k+1}),$$

where $\Psi(x, a)$ denotes the integrand in (1.1).

The choice of stepsize differs across the two settings. In Chapter 2, we consider adaptive stepsize schemes under mild conditions satisfied by methods such as AdaGrad-Norm, DoG, D-Adaptation, and RMSProp. In contrast, Chapter 3 focuses on bounded deterministic stepsize schedules, including fixed stepsizes and polynomial decay.

1.3 High-dimensional Diffusion Approximation for SGD

To analyze the high-dimensional behavior of SGD, we introduce a continuous-time stochastic process that captures its effective evolution. Namely, we define the *homogenized SGD*:

$$\mathrm{d}\mathfrak{X}_t = -\gamma(t)\,\nabla\mathcal{R}(\mathfrak{X}_t)\,\mathrm{d}t + \gamma(t)\sqrt{\mathbb{E}_a\left[\nabla f\big(\langle\beta(\mathfrak{X}_t), a\rangle\big)^2\right]}\,\nabla\beta(\mathfrak{X}_t)^\top\sqrt{K/d}\,\mathrm{d}\mathfrak{B}_t, \tag{1.2}$$

where $\mathfrak{X}_0 = x_0$ and $\mathfrak{B}_t$ is a standard Brownian motion in $\mathbb{R}^d$. To compare SGD with homogenized SGD, we embed the discrete iterates $\{x_k\}$ into continuous time by identifying the k-th iterate with time t through $k = \lfloor td \rfloor$. Under this scaling, one unit of time corresponds to d SGD iterations. We prove that, in a variety of settings, the SGD iterates track the homogenized SGD dynamics in an appropriate sense as d tends to infinity.

Homogenized SGD separates the deterministic drift, driven by the population risk $\mathcal{R}$, from a diffusion term that captures the effect of gradient noise. The diffusion structure reflects both the data distribution and the feature map β. This process also provides a convenient and effective description of SGD: it allows one to track the evolution of many quantities of interest through Itô calculus, often serves as a practical tool for numerical simulation via schemes such as Euler–Maruyama, and enables a natural route to identifying deterministic limits for observable statistics.

A key conceptual distinction between our framework and classical diffusion approximations lies in the regime under which the continuous-time description becomes accurate. Traditional diffusion approximations are derived at fixed dimension and rely on the small stepsize limit $\gamma \to 0$, where stochastic fluctuations vanish and the dynamics converge to a gradient flow or a perturbed-gradient-flow SDE.

In contrast, our approach operates in a high-dimensional regime where $d \to \infty$ while the stepsize remains fixed. In this setting, the effective dynamics are governed by high-dimensional

concentration rather than vanishing stepsizes. As a result, the diffusion coefficient is not calibrated through covariance matching, but instead arises from a high-dimensional limit theory. This leads to an SDE model that remains accurate at finite stepsizes and captures phenomena beyond the scope of classical diffusion approximations.

While the homogenized SGD provides a useful mesoscopic description of the dynamics, it is not always necessary for the analysis. In Chapter 2, the concentration of observable statistics can be established directly from the discrete-time dynamics, without passing through a continuous-time approximation. The SDE framework becomes particularly useful in Chapter 3, where it serves as an intermediate description linking the stochastic dynamics to a deterministic evolution of statistics and enables the derivation of precise stability and convergence guarantees.

In the next section, we rigorously show that the homogenized SGD provides an accurate description of the dynamics of SGD in high dimensions, even at finite stepsizes.

1.4 High-dimensional Concentration of SGD and its Diffusion Approximation

This section formalizes the main concentration phenomenon: in the high-dimensional regime, both streaming SGD and the homogenized SGD concentrate around the same deterministic dynamics. The key idea is that, under our structural assumptions, the learning dynamics can be described through a small collection of low-dimensional statistics.

1.4.1 Covariance Representation

To make this precise, we introduce the block matrix

$$W(x) := [\,\beta(x)\,|\,\beta^*\,|\,\mathbb{1}_d\,] \in \mathbb{R}^{d\times 3},$$

and define the associated covariance matrix of the random vector $W(x)^\top a \in \mathbb{R}^3$

$$B(x) := W(x)^\top K W(x) \in \mathbb{R}^{3\times 3},$$

(with an additional normalization by $1/d$ in Chapter 3).

Since the risk $\mathcal{R}(x)$ depends on x only through the Gaussian vector $W(x)^\top a$, the learning dynamics can be summarized by the matrix $B(x)$. In particular, there exists a function

$h : \mathbb{R}^{3\times 3} \to \mathbb{R}$ whose derivative ∇h is α-pseudo-Lipschitz and such that

$$\mathcal{R}(x) = h\big(B(x)\big).$$

Similarly, the second-moment quantity controlling the size of the stochastic gradient noise depends only on $B(x)$: there exists an α-pseudo-Lipschitz function $I : \mathbb{R}^{3\times 3} \to \mathbb{R}$ such that

$$\mathbb{E}_a\left[\nabla f\big(\langle \beta(x), a\rangle\big)^2\right] = I\big(B(x)\big).$$

This reduction shows that both the drift and diffusion of the learning dynamics are governed by a low-dimensional summary of the state. Next, we further encode this information through a resolvent-based matrix $S(x, z)$, which provides a convenient representation for tracking the evolution of these statistics.

1.4.2 Resolvent-based Representation

To obtain a tractable description of the learning dynamics, we further lift the low-dimensional statistics $B(x)$ into a resolvent-based representation.

In the setting of Chapter 2, for a complex parameter $z \in \mathbb{C}$, we define the matrix

$$\Omega(x, z) = R(z; K) := \big(z \cdot I_d - K\big)^{-1},$$

and introduce the associated 3×3 matrix

$$S(x, z) := W(x)^\top \Omega(x, z)\, W(x),$$

where $W(x)$ is the stacked matrix defined previously.

The key point is that the map $z \mapsto S(x, z)$ encodes all the information contained in $B(x)$. In particular, $B(x)$ can be recovered from $S(x, z)$ through a contour integral representation based on the Cauchy integral formula, and hence the evolution of $S(x, z)$ fully determines the learning curves and other observable statistics.

In the setting of Chapter 3, the nonlinear (quadratic) structure of the feature map ψ leads to a more intricate resolvent representation involving multiple spectral parameters. In this case, for a complex vector $z \in \mathbb{C}^4$, we define

$$\Omega(x, z) = R(z_1; \mathrm{diag}(u))R(z_2; \mathrm{diag}(v))R(z_3; \mathrm{diag}(\beta^*))R(z_4; K),$$

and the associated 3×3 matrix is defined by

$$S(x, z) = \frac{1}{d} W(x)^\top \Omega(x, z) W(x).$$

Thus, the representation is built from products of several resolvents, capturing higher-order interactions induced by the nonlinear parametrization. Despite this added complexity, the same principle applies: the resulting resolvent representation provides a compact description of the state of the system and governs the evolution of observable statistics.

In particular, we evaluate these quantities along both the discrete dynamics, writing $S(x_k, z)$ and $B(x_k)$, and the diffusion approximation, writing $S(\mathcal{X}_t, z)$ and $B(\mathcal{X}_t)$.

1.4.3 Concentration of the Dynamics

We now state the main concentration phenomenon at the level of the low-dimensional statistics. In the high-dimensional regime, the learning dynamics admit a deterministic limit characterized by the (local) solution of the partial integro-differential equation

$$\partial_t \mathscr{S}(t, \cdot) = \mathcal{F}\big(\cdot, \mathscr{S}(t, \cdot)\big), \qquad \mathscr{S}(0, z) = S(x_0, z), \tag{1.3}$$

where $\mathcal{F}$ is defined in (B.5) and involves contour integrals and derivatives in z.

To ensure a meaningful high-dimensional limit, we impose scaling assumptions on the initialization and target parameters. In Chapter 2, we assume that the Euclidean norms of x_0 and β^* are bounded independently of d. In Chapter 3, we instead assume that the entries of x_0 and β^* are uniformly bounded in d.

We also require that the stochastic dynamics remain stable over time. In Chapter 2, this is ensured by assuming $\alpha \leqslant 1$. In Chapter 3, stability can in some cases be verified directly from more primitive conditions. For instance, in certain regimes, sufficiently small fixed stepsizes guarantee that the risk converges exponentially fast with high probability.

Concentration of S (informal). Let $\mathscr{S}(t, z)$ denote the deterministic solution of (1.3). Then, for any fixed time horizon and uniformly over spectral parameters z ranging over contours that remain a fixed positive distance from the relevant spectra, both the discrete

SGD iterates and the homogenized SDE remain close to this deterministic trajectory:

$$S(x_{\lfloor td \rfloor}, z) \approx \mathscr{S}(t, z), \qquad S(\mathcal{X}_t, z) \approx \mathscr{S}(t, z),$$

uniformly over bounded time intervals with high probability.

This result shows that, despite the stochastic nature of the algorithm, the learning dynamics become effectively deterministic in high dimensions. Moreover, both SGD and its diffusion approximation follow the same limiting trajectory, providing a unified description at the level of observable statistics.

1.4.4 Other Statistics

The concentration of S transfers to any statistic that can be expressed through the same resolvent-based representation. Concretely, this includes quantities that can be written as contour integrals of functions of $S(x, z)$, or recovered from the low-dimensional statistics $B(x)$. The contour-integral representation is useful because uniform control of $S(x, z)$ along the contours implies control of the resulting statistic.

Concentration of φ (informal). Let $\varphi(x)$ be an admissible observable, meaning that it can be represented through the low-dimensional statistics $B(x)$ or through contour integrals involving $S(x, z)$. Then there exists a deterministic function $\phi(t)$ such that

$$\varphi(x_{\lfloor td \rfloor}) \approx \phi(t), \qquad \varphi(\mathcal{X}_t) \approx \phi(t),$$

uniformly over bounded time intervals with high probability.

Thus, the deterministic limit is not limited to the core statistics $B(x)$. It also governs many quantities of interest, such as risk, curvature, norms, and estimation error, whenever these quantities admit such a representation. In this sense, the resolvent dynamics provide a macroscopic description of the learning process.

1.5 Adaptive Stepsize Algorithms Analysis

We now apply the proposed framework to the analysis of adaptive stepsize algorithms. In particular, we consider two representative methods on the least squares problem: exact

line-search and AdaGrad-Norm. These examples illustrate how the framework can be used to analyze adaptive methods, and suggest that similar techniques extend to a broader class of losses and stepsize schemes.

1.5.1 Idealized Exact Line Search and Polyak Stepsize

We first illustrate the framework through two classical idealized stepsize strategies: exact line search and the Polyak stepsize. In deterministic optimization, these rules select the stepsize that maximally decreases either the objective function (exact line search) or a measure of distance to optimality (Polyak stepsize) at each iteration.

In the stochastic setting, we define analogous strategies by applying these principles to the deterministic limits of the learning dynamics. Since SGD concentrates around deterministic trajectories in high dimensions, we choose stepsizes that optimize the evolution of the corresponding deterministic quantities.

Let $\mathscr{D}(t)$ denote a deterministic measure of distance to optimality. A natural threshold to consider is the largest stepsize such that $\mathrm{d}\mathscr{D}(t) < 0$. The Polyak stepsize is then defined as the greedy stepsize strategy that maximizes the rate of decrease in the distance to optimality at each iteration:

$$\gamma_t^{\text{Polyak}} \in \arg\min_\gamma \mathrm{d}\mathscr{D}(t).$$

In the least squares setting, this yields a stepsize proportional to the maximal stable stepsize.

Polyak stepsize (informal). In the least squares setting, the Polyak stepsize is proportional to this stability threshold and, in particular, is given by a constant fraction of the maximal stepsize that ensures contraction. In the noiseless case, this threshold depends explicitly on the average eigenvalue of the covariance matrix, $\mathrm{Tr}(K)/d$, and recovers the optimal fixed stepsize up to constant factors.

Similarly, exact line search selects the learning rate that maximally decreases the deterministic risk $\mathscr{R}(t)$:

$$\gamma_t^{\text{line}} \in \arg\min_\gamma \mathrm{d}\mathscr{R}(t).$$

This corresponds to a greedy strategy that optimizes the instantaneous decrease of the risk.

Exact line search (informal). While exact line search optimizes the instantaneous decrease of the risk, its long-term behavior is sensitive to the spectral structure of the covariance matrix. In particular, in anisotropic settings, it can lead to suboptimal convergence rates compared to the Polyak stepsize, despite being greedy at each step.

These results highlight that optimal stepsize strategies in high dimensions are governed by global properties of the data distribution, rather than purely local optimization criteria. In particular, optimizing for risk and distance to optimality can lead to qualitatively different behaviors, especially when the covariance matrix exhibits strong spectral heterogeneity.

1.5.2 AdaGrad-Norm

We now analyze a practical adaptive stepsize scheme, AdaGrad-Norm, within the proposed framework. Unlike the idealized strategies considered previously, AdaGrad-Norm is implementable and widely used in stochastic optimization. Our analysis yields a precise characterization of its high-dimensional behavior, revealing how the stepsize dynamics depend on the interaction between noise and the spectral structure of the data.

Noisy regime (informal). In the presence of additive noise, the AdaGrad-Norm stepsize decays universally as

$$\gamma_t \asymp t^{-1/2},$$

independently of the covariance structure of the data. In particular, the asymptotic behavior is governed solely by the noise level and not by the spectrum of K.

Noiseless regime (informal). In the absence of noise, the behavior of the stepsize depends strongly on the spectral properties of the covariance matrix. When the spectrum of K is well-conditioned (e.g., $\lambda_{\min}(K)$ bounded away from zero), the stepsize remains bounded away from zero and converges to a positive constant determined by $\mathrm{Tr}(K)/d$ and the initialization.

Spectral phase transition (informal). When the covariance matrix exhibits a heavy-tailed spectrum, the stepsize can decay to zero at a rate determined by the interaction between the spectral distribution and the initialization. In particular, for power-law spectra, the

dynamics exhibit a phase transition: depending on the spectral exponent and initialization alignment, the stepsize either remains bounded or decays polynomially in time.

These results show that, unlike in the noisy regime, the behavior of adaptive stepsizes in high dimensions is governed by global spectral properties of the data. In particular, AdaGrad-Norm implicitly adapts to the effective dimensionality of the problem, leading to qualitatively different learning dynamics depending on the structure of the covariance matrix.

1.6 Diagonal Linear Networks Analysis

As a concrete application of our framework, we analyze diagonal linear networks under the parametrization $x = (u, v)$ and $\beta(x) = u^2 - v^2$. We show that, at sufficiently small but finite stepsizes, the stochastic dynamics satisfy strong stability properties.

1.6.1 High-probability Exponential Decay

When the stepsize is below a fixed threshold, the homogenized SDE remains well behaved and the risk decays exponentially fast.

High-probability exponential decay (informal). For sufficiently small learning rates, there exist constants $C, \mu > 0$ such that, with high probability,

$$\mathcal{R}(\mathcal{X}_t) \leq Ce^{-\mu t} \qquad \text{for all } t \geq 0.$$

In particular, the stochastic dynamics remain stable and non-explosive over time.

By the concentration results established earlier, the same exponential decay behavior holds for the discrete SGD iterates. Thus, in this regime, SGD converges to zero risk at an exponential rate with high probability.

This result shows that, despite the presence of stochastic noise, the learning dynamics are self-stabilizing: the same mechanism that drives the risk to zero also controls the fluctuations and confines the trajectory to a stable region of parameter space.

Concluding Remarks

Beyond convergence guarantees, the framework developed in this dissertation also enables the analysis of finer properties of the learning dynamics, including curvature-related quantities and transient phenomena such as progressive sharpening.

Overall, this dissertation develops a unified high-dimensional framework for stochastic optimization at finite stepsizes. By combining concentration methods with continuous-time approximations, the theory provides deterministic descriptions for SGD and adaptive methods beyond the classical small-stepsize regime, and characterizes how covariance structure, adaptive learning rates, and spectral properties shape the resulting dynamics.

Chapter 2

HIGH-DIMENSIONAL LIMIT OF SGD FOR ADAPTIVE STEPSIZE ALGORITHMS

Joint work with Elizabeth Collins-Woodfin, Inbar Seroussi, Andrew Mackenzie, Elliot Paquette, and Courtney Paquette [20]

Abstract. We develop a framework for analyzing the training and stepsize dynamics on a large class of high-dimensional optimization problems, which we call the high line, trained using one-pass stochastic gradient descent (SGD) with adaptive stepsizes. We give exact expressions for the risk and stepsize curves in terms of a deterministic solution to a system of ODEs. We then investigate in detail two adaptive stepsizes – an idealized exact line search and AdaGrad-Norm – on the least squares problem. When the data covariance matrix has strictly positive eigenvalues, this idealized exact line search strategy can exhibit arbitrarily slower convergence when compared to the optimal fixed stepsize with SGD. Moreover we exactly characterize the limiting stepsize (as time goes to infinity) for line search in the setting where the data covariance has only two distinct eigenvalues. For noiseless targets, we further demonstrate that the AdaGrad-Norm stepsize converges to a deterministic constant inversely proportional to the average eigenvalue of the data covariance matrix, and identify a phase transition when the covariance density of eigenvalues follows a power law distribution. We provide our code for evaluation at https://github.com/amackenzie1/highline2024.

2.1 Introduction

In deterministic optimization, adaptive stepsize strategies, such as line search (see [67], therein), AdaGrad-Norm [92], Polyak stepsize [76], and others were developed to provide stability and improve efficiency and adaptivity to unknown parameters. While the practical benefits for deterministic optimization problems are well-documented, much of our understanding of adaptive learning rate strategies for stochastic algorithms are still in their

infancy.

There are many adaptive learning rate strategies used in machine learning with many design goals. Some are known to adapt to stochastic gradient descent (SGD) gradient noise while others are robust to hyper-parameters (e.g., [8, 98]). Theoretical results for adaptive algorithms tend to focus on guaranteeing minimax-optimal rates, but this theory is not engineered to provide realistic performance comparisons; indeed many adaptive algorithms are minimax-optimal, and so more precise statements are needed to distinguish them. For instance, the exact learning rates (or rate schedules) to which these strategies converge are unknown, nor their dependence on the geometry of the problem. Moreover, we often do not know how these adaptive stepsizes compare with well-tuned constant or decaying fixed learning rate SGD, which can be viewed as a cost associated with selecting the adaptive strategy in comparison to tuning by hand.

In this work, we develop a framework for analyzing the exact dynamics of the risk and adaptive learning rate strategies for a wide class of optimization problems that we call *high-dimensional linear (high line) composite functions*. In this class, the objective function takes the form of an expected risk $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$ over high-dimensional data $(a, \epsilon) \sim \mathcal{D} \subset \mathbb{R}^d \times \mathbb{R}$ of a function $f : \mathbb{R}^3 \to \mathbb{R}$ composed with the linear functions $\langle X, a\rangle$, $\langle X^\star, a\rangle$. That is, we seek to solve

$$\min_{X\in\mathbb{R}^d} \left\{ \mathcal{R}(X) \stackrel{\text{def}}{=} \mathbb{E}_{a,\epsilon}[f(\langle a, X\rangle, \langle a, X^\star\rangle, \epsilon)] \quad \text{for} \quad (a,\epsilon) \sim \mathcal{D}, X^\star \in \mathbb{R}^d \right\}. \tag{2.1}$$

We suppose $a \sim \mathcal{N}(0, K)$ where $K \in \mathbb{R}^{d\times d}$ is the covariance matrix. We train (2.1) using (one-pass) stochastic gradient descent with adaptive learning rates, $\mathfrak{g}_k$ (SGD+AL). Our main goal is to give a framework for better[1] performance analysis of these adaptive methods. We then illustrate this framework by considering two adaptive learning rate algorithms on the least squares problem[2], the results of which appear in Table 2.1: exact line-search (idealistic) (Sec. 2.3) and AdaGrad-Norm (Sec. 2.4). We expect other losses and adaptive learning rates can be studied using this approach.

[1]More realistic, in that it deals with high-dimensional anisotropic loss geometries and more precise, in that it can distinguish minimax optimal algorithms as more-or-less performant.

[2]We extend some results to the general strongly convex setting.

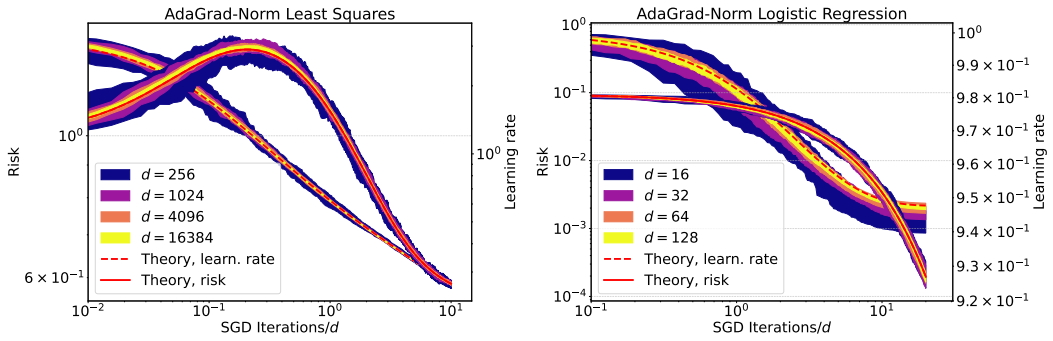

Figure 2.1: **Concentration of learning rate and risk for AdaGrad-Norm** on least squares with label noise $\omega = 1$ (left) and logistic regression with no noise (right). As dimension increases, both risk and learning rate concentrate around a deterministic limit (red) described by our ODE in Theorem 2.2.1. The initial risk increase (left) suggests the learning rate started too high, but AdaGrad-Norm adapts. Our ODEs predict this behavior. See Sec. A.8 for simulation details.

Main contributions. *Performance analysis framework.* We provide an equivalence of $\mathcal{R}(X_k)$ and learning rate $\mathfrak{g}_k$ under SGD+AL to deterministic functions $\mathscr{R}(t)$ and γ_t via solving a *deterministic* system of ODEs (see Section 2.2), which we then analyze to show how the covariance spectrum influences the optimization. See Figure 2.1. As the dimension d of the problem grows, the learning curves of $\mathcal{R}(X_k)$ become closer to $\mathscr{R}(t)$ and the curves concentrate around $\mathscr{R}(t)$ with probability better than any inverse power of d (See Theorem 2.2.1).

Greed can be arbitrarily bad in the presence of strong anisotropy (that is, $Tr(K)/d \ll Tr(K^2)/d$). Our analysis reveals that exact line search, which is to say optimally decreasing the risk at each step, can run arbitrarily slower than the best fixed learning rate for SGD on a least squares problem when $\lambda_{\min} \stackrel{\text{def}}{=} \lambda_{\min}(K) > C > 0$. The best fixed stepsize (least squares problem) is $(\mathrm{Tr}(K)/d)^{-1}$ or the inverse of the average eigenvalue, see Polyak stepsize [76]. Line search, on the other hand, converges to a fixed stepsize of order $\lambda_{\min}/(\mathrm{Tr}(K^2)/d)$. It can be that $\lambda_{\min}/(\mathrm{Tr}(K^2)/d) \ll (\mathrm{Tr}(K)/d)^{-1}$ making exact line search substantially underperform Polyak stepsize. We further explore this and, in the case where d-eigenvalues

of K take only two values $\lambda_1 > \lambda_2 > 0$, we give an exact expression as a function of λ_1 and λ_2 for the limiting behavior of γ_t as $t \to \infty$ (See Fig. A.1).

AdaGrad-Norm selects the optimal step-size, provided it has a warm start. In the absence of label noise and when the smallest eigenvalue of K satisfies $\lambda_{\min} > C > 0$, the learning rate converges to a deterministic constant that depends on the average condition number (like in Polyak) and scales inversely with $\frac{\mathrm{Tr}(K)}{d}\|X_0 - X^\star\|^2$. Therefore it attains automatically the optimal fixed stepsize in terms of the covariance *without* knowledge of $\mathrm{Tr}(K)$, but pays a penalty in the constant, namely $\|X_0 - X^\star\|^2$. If one knew $\|X_0 - X^\star\|^2$ then by tuning the parameters of AdaGrad-Norm one might achieve performance consistent with Polyak; this also motivates more sophisticated adaptive algorithms such as DoG [43] and D-Adaptation [26], which adaptively compensate and/or estimate $\|X_0 - X^\star\|^2$.

AdaGrad-Norm can use overly pessimistic decaying schedules on hard problems. Consider power law behavior for the spectrum of K and the signal $X^\star$. This is a natural setting as power law distributions have been observed in many datasets [93]. Here the learning rate and asymptotic convergence of K undergo a *phase transition*. For power laws corresponding to easier optimization problems, the learning rate goes to a constant and the risk decays at $t^{-\alpha_1}$. For harder problems, the learning rate decays like $t^{-\eta_1}$ and the risk decays at a different sublinear rate $t^{-\alpha_2}$. See Table 2.1 and Sec. 2.4 for details.

Notation. Define $\mathbb{R}_+ = [0, \infty)$. We say an event holds *with overwhelming probability, w.o.p.,* if there is a function $\omega : \mathbb{N} \to \mathbb{R}$ with $\omega(d)/\log d \to \infty$ so that the event holds with probability at least $1 - e^{-\omega(d)}$. We let $\mathbf{1}_A(x)$ be the indicator function of the set A where it is 1 if $x \in A$ and 0 otherwise. For a matrix $A \in \mathbb{R}^{m \times d}$, we use $\|A\|$ to denote the Frobenius norm and $\|A\|_{\mathrm{op}}$ to denote the operator-2 norm. If unspecified, we assume that the norm is the Frobenius norm. For normed vector spaces $\mathcal{A}$, $\mathcal{B}$ with norms $\|\cdot\|_{\mathcal{A}}$ and $\|\cdot\|_{\mathcal{B}}$, respectively, and for $\alpha \geq 0$, we say a function $F : \mathcal{A} \to \mathcal{B}$ is *α-pseudo-Lipschitz* with constant L if for any $A, \hat{A} \in \mathcal{A}$, we have

$$\|F(A) - F(\hat{A})\|_{\mathcal{B}} \leq L\|A - \hat{A}\|_{\mathcal{A}}(1 + \|A\|_{\mathcal{A}}^{\alpha} + \|\hat{A}\|_{\mathcal{A}}^{\alpha}).$$

We write $f(t) \asymp g(t)$ if there exist *absolute* constants $C, c > 0$ such that $c \cdot g(t) \leq f(t) \leq C \cdot g(t)$ for all t. If the constants depend on parameters, e.g., α, then we write $\asymp_\alpha$.

Table 2.1: **Summary of adaptive learning rates results on the least squares problem.** We summarize our results for line search and AdaGrad-Norm under various assumptions on the covariance matrix K. We denote $\lambda_{\min}$ the smallest non-zero eigenvalue of K and $\frac{\mathrm{Tr}(K)}{d}$ the average eigenvalue. Power law(δ, β) assumes the eigenvalues of K, $\{\lambda_i\}_{i=1}^d$, follow a power law distribution, that is, for $0 < \beta < 1$, $\lambda_i \sim (1-\beta)\lambda^{-\beta}\mathbf{1}_{(0,1)}$ for all $1 \leq i \leq d$ and $\langle X_0 - X^\star, \omega_i\rangle^2 \sim \lambda_i^{-\delta}$ where $\{\omega_i\}_{i=1}^d$ are eigenvectors of K (see Prop 2.4.4). For * (see Prop. 2.4.2), requires a good initialization on b, η.

Learning rate	**K assumption**	**Limiting γ_∞**	**Convergence rate**
AdaGrad-Norm(b,η) (see Sec. 2.4)	$\lambda_{\min} > C$	$\gamma_t \asymp \frac{\eta^2}{\frac{b}{\eta}+\frac{1}{4d}\mathrm{Tr}(K)\|X_0-X^\star\|^2}$	$\log(\mathscr{R})^* \asymp -\lambda_{\min}\gamma_\infty t$
AdaGrad-Norm(b,η) Power law (see Sec. 2.4)	$\beta+\delta<1$	$\gamma_t \asymp_{\delta,\beta} 1$	$\mathscr{R}(t) \asymp_{\delta,\beta} t^{\beta+\delta-2}$
	$\beta+\delta=1$	$\gamma_t \asymp_{\delta,\beta} \frac{1}{\log(t+1)}$	$\mathscr{R}(t) \asymp_{\delta,\beta} \left(\frac{t}{\log(t+1)}\right)^{-1}$
	$1<\beta+\delta<2$	$\gamma_t \asymp_{\delta,\beta} t^{-1+\frac{1}{\beta+\delta}}$	$\mathscr{R}(t) \asymp_{\delta,\beta} t^{-\frac{2}{\beta+\delta}+1}$
Exact line search, idealized (see Sec. 2.3)	$\lambda_{\min} > C$	$\gamma_t \asymp \frac{\lambda_{\min}}{\mathrm{Tr}(K^2)/d}$	$\log(\mathscr{R}) \asymp -\lambda_{\min}\gamma_\infty t$
Polyak stepsize (see Sec. 2.3)	$\lambda_{\min} > C$	$\gamma_t = \frac{1}{\mathrm{Tr}(K)/d}$	$\log(\mathscr{R}) \asymp -\lambda_{\min}\gamma_\infty t$

Related work. Some notable adaptive learning rates in the literature are AdaGrad-Norm [52, 92, 96], RMSprop [41], stochastic line search, stochastic Polyak stepsize [57], and more recently DoG [43] and D-Adaptation [26]. In this work, we introduce a framework for analyzing these algorithms, and we strongly believe it can be used to analyze many more of these adaptive algorithms. We highlight below a nonexhaustive list of related work.

AdaGrad-Norm. AdaGrad, introduced by [28, 61], updates the learning rate at each iteration using the stochastic gradient information. The single stepsize version [52, 92, 96], that depends on the norm of the gradient, (see Table 2.2 for the updates), has been shown to be robust to input parameters [55]. Several works have shown worst-case convergence

guarantees [32, 53, 89, 92]. A linear rate of $O(\exp(-\kappa T))$ is possible for μ-strongly convex, L-smooth functions (κ is the condition number μ/L). In [97] (similar idea in [96]), the authors show for strongly convex, smooth stochastic objectives (with additional assumptions) that the AdaGrad-Norm learning rate exhibits a two stage behavior – a burn in phase and then when it reaches the smoothness constant it self-stablizes.

Stochastic line search and Polyak stepsizes. Recently there has been renewed interest in studying stochastic line search [29, 70, 85] and stochastic Polyak stepsize (and their variants) [12, 37, 39, 45, 57, 66, 68, 77]. Much of this research focuses on worst-case convergence guarantees for strongly convex and smooth functions (see e.g., [57]) and designing practical algorithms. In [84], the authors provide a bound on the learning rate for Armijo line search in the finite sum setting with a rate of $L_{\max}/\text{avg. } \mu$ where avg. μ is the avg. strong convexity and $L_{\max}$ is the max. Lipschitz constant of the individual functions. In this work, we consider a slightly different problem. We work with the population loss and we note that the analogue to $L_{\max}$ for us would require that the samples a satisfy $\|aa^T\|_{\text{op}} \leq L_{\max}$ *for all* a; this fails to hold for $a \sim \mathcal{N}(0, K)$. Moreover, $L_{\max}$ could be much worse than $\mathbb{E}[\|aa^T\|_{\text{op}}]$.

Deterministic dynamics of stochastic algorithms in high-dimensions. The literature on deterministic dynamics for isotropic Gaussian data has a long history [14, 15, 78, 79]. These results have been rigorously proven and extended to other models under the isotropic Gaussian assumption [3, 4, 6, 24, 25, 34, 35, 91]. Extensions to multi-pass SGD with small mini-batches [73] as well as momentum [51] have also been studied. Other high-dimensional limits leading to a different class of dynamics also exist [16–18, 33, 62]. Recently, significant contributions have been made in understanding the effects of a non-identity data covariance matrix on the training dynamics [10, 20, 21, 35, 36, 100]. The non-identity covariance modifies the optimization landscape and affects convergence properties, as discussed in [21]. This work extends the findings of [21] to stochastic adaptive algorithms, exploring the effect of non-identity covariance within these algorithms. Notably, Theorem 1.1 from [21] is restricted to deterministic learning rate schedules, limiting its applicability in many practical scenarios. In contrast, our Theorem 2.2.1 accommodates stochastic adaptive learning rates, aligning with widely used algorithms in practice.

2.1.1 Model Set-up

We suppose that a sequence of independent samples $\{(a_k, y_k)\}$ drawn from a distribution $\mathcal{D} \subset \mathbb{R}^d \times \mathbb{R}$ is provided where y_k is the target. The target y_k is a function of some random label noise $\epsilon_k \in \mathbb{R}$ and the input feature a_k dotted with a ground truth signal $X^\star \in \mathbb{R}^d$, $\langle a_k, X^\star \rangle$. Therefore, the distribution of the data is only determined by the input feature and the noise, i.e., the pair (a, ϵ). In particular, we assume (a, ϵ) follows a distributional assumption.

Assumption 2.1.1 (Data and label noise)**.** *The samples $(a, \epsilon) \sim \mathcal{D}$ are normally distributed: $\epsilon \sim \mathcal{N}(0, \omega^2)$ where $\omega \in \mathbb{R}$, and $a \sim \mathcal{N}(0, K)$, with a covariance matrix $K \in \mathbb{R}^{d\times d}$ that is bounded in operator norm independent of d; i.e., $\|K\|_{op} \leq C$. Furthermore, a and ϵ are independent.*

For $a, X, X^\star \in \mathbb{R}^d$, $\epsilon \in \mathbb{R}$, and a function $f : \mathbb{R}^3 \to \mathbb{R}$, we seek to minimize an expected risk function $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$, which we refer to as the *high-dimensional linear composite*[3], of the form

$$\mathcal{R}(X) \stackrel{\text{def}}{=} \mathbb{E}_{a,\epsilon}[\Psi(X; a, \epsilon)] \quad \text{for} \quad (a, \epsilon) \sim \mathcal{D}, \quad \text{and} \quad \Psi(X; a, \epsilon) = f(\langle a, X\rangle, \langle a, X^\star\rangle, \epsilon). \tag{2.2}$$

In what follows, we use the matrix $W = [X|X^\star] \in \mathbb{R}^{d\times 2}$ that concatenates X and $X^\star$, and we shall let $B = B(W) = W^T K W$ be the covariance matrix of the Gaussian vector $(\langle a, X\rangle, \langle a, X^\star\rangle)$.

Assumption 2.1.2 (Pseudo-lipschitz f)**.** *The function $f : \mathbb{R}^3 \to \mathbb{R}$ is α-pseudo-Lipschitz with $\alpha \leq 1$.*

By assumption, $\mathcal{R}(X)$ involves an expectation over the correlated Gaussians $\langle a, X\rangle$ and $\langle a, X^\star\rangle$. We can express this as $\mathcal{R}(X) \stackrel{\text{def}}{=} h(B)$ for some well-behaved function $h : \mathbb{R}^{2\times 2} \to \mathbb{R}$.

Assumption 2.1.3 (Risk representation)**.** *There exists a function $h : \mathbb{R}^{2\times 2} \to \mathbb{R}$ such that $h(B) = \mathcal{R}(X)$ is differentiable and satisfies*

$$\nabla_X \mathcal{R}(X) = \mathbb{E}_{a,\epsilon} \nabla_X \Psi(X; a, \epsilon).$$

[3]Note that d need not be large to define this, but the structure allows us to consider d as a tunable parameter. Moreover, as we increase d, the analysis we do will be more meaningful.

Table 2.2: **Two adaptive learning rates considered in detail.** The stochastic adaptive learning rate, $\mathfrak{g}_k$, is the learning rate directly used in the update for SGD whereas the deterministic, γ_t, is the deterministic equivalent of $\mathfrak{g}_k$ after scaling.

Algorithm		**General update**	**Least squares**
AdaGrad-Norm(b, η) $b_0 = b \times d$	$\mathfrak{g}_k$	$b_k^2 = b_{k-1}^2 + \|\nabla \Psi(X_{k-1})\|^2$; $\mathfrak{g}_{k-1} = d \times \frac{\eta}{\lvert b_k \rvert}$	same
	γ_t	$\frac{\eta}{\sqrt{b^2 + \frac{\operatorname{Tr}(K)}{d} \int_0^t I(\mathscr{B}(s))\, \mathrm{d}s}}$	$\frac{\eta}{\sqrt{b^2 + \frac{2\operatorname{Tr}(K)}{d} \int_0^t \mathscr{R}(s)\, \mathrm{d}s}}$
Exact line search (idealized)	$\mathfrak{g}_k$	$\frac{\|\nabla \mathcal{R}(X_k)\|^2}{\frac{\operatorname{Tr}(\nabla^2 \mathcal{R}(X_k) K)}{d} \mathbb{E}_{a,\epsilon}[(f'(\langle a, X\rangle; \langle a, X^\star\rangle, \epsilon))^2]}$	$\frac{\|\nabla \mathcal{R}(X_k)\|^2}{\frac{2\operatorname{Tr}(K^2)}{d} \mathcal{R}(X_k)}$
	γ_t	$\arg\min_{\gamma} \mathrm{d}\mathscr{R}(t)$	$\frac{\sum_{i=1}^d \lambda_i^2 \mathscr{D}_i^2(t)}{2\operatorname{Tr}(K^2)\mathscr{R}(t)}$

Furthermore, h is continuously differentiable and its derivative ∇h is α-pseudo-Lipschitz for some $0 \leq \alpha \leq 1$, with constant $L(\nabla h)$.

The final assumption is the well-behavior of the Fisher information matrix of the gradients. The first coordinate of f is special, as the optimizer must be able to differentiate it. Thus, we treat $f(x, x^\star, \epsilon)$ as a function of a single variable with two parameters: $f(x, x^\star, \epsilon) = f(x; x^\star, \epsilon)$ and denote the (almost everywhere) derivative with respect to the first variable as f'.

Assumption 2.1.4 (Fisher matrix)**.** *Define $I(B) \stackrel{def}{=} \mathbb{E}_{a,\epsilon}[(f'(\langle a, X\rangle; \langle a, X^\star\rangle, \epsilon))^2]$ where the function $I : \mathbb{R}^{2\times 2} \to \mathbb{R}$. Furthermore, I is α-pseudo-Lipschitz with constant $L(I)$ for some $\alpha \leq 1$.*

A large class of natural regression problems fit within this framework, such as logistic regression and least squares (see [21, Appendix B]). We also note that Assumptions 2.1.3 and 2.1.4 are nearly satisfied for L-smooth objectives f (see Lemma A.2.1), and a version of the main theorem holds under just this assumption (albeit with a weaker conclusion).

2.1.2 Algorithmic set-up

We apply *one-pass* or *streaming* SGD with an adaptive learning rate $\mathfrak{g}_k$ (SGD+AL) to solve $\min_{X\in\mathbb{R}^d}\mathcal{R}(X)$, (2.2). Let $X_0 \in \mathbb{R}^d$ be an initial vector (random or non-random). Then SGD+AL iterates by selecting a *new* data point $(a_{k+1}, \epsilon_{k+1})$ such that $a_{k+1} \sim \mathcal{N}(0, K)$ and $\epsilon_{k+1} \sim \mathcal{N}(0, \omega^2)$ and makes the update

$$X_{k+1} = X_k - \frac{\mathfrak{g}_k}{d}\cdot\nabla_X \Psi(X_k; a_{k+1}, \epsilon_{k+1}) = X_k - \frac{\mathfrak{g}_k}{d} f'(\langle a_{k+1}, X_k\rangle; \langle a_{k+1}, X^\star\rangle, \epsilon_{k+1})a_{k+1}, \quad (2.3)$$

where $\mathfrak{g}_k > 0$ is a learning rate (see assumptions below)[4]. To perform our analysis, we place the following assumption on the initialization X_0 and signal $X^\star$.

Assumption 2.1.5 (Initialization and signal)**.** *The initialization point X_0 and the signal $X^\star$ are bounded independent of d, that is,* $\max\{\|X_0\|, \|X^\star\|\} \le C$ *for some C independent of d.*

Adaptive learning rate. Our analysis requires some mild assumptions on the learning rate. To this end, we define a learning rate function $\gamma : \mathbb{R}_+ \times D([0,\infty)) \times D([0,\infty)) \times D([0,\infty)) \to \mathbb{R}_+$ by[5]

$$\begin{aligned} \mathfrak{g}_k &\stackrel{\text{def}}{=} \gamma(k, N_k(d\times\cdot), G_k(d\times\cdot), Q_k(d\times\cdot)), \text{ for } k\in\mathbb{N}, \text{ where for any } t\ge 0, \\ (N_k(t), G_k(t), Q_k(t)) &\stackrel{\text{def}}{=} \mathbf{1}_{\{t<k\}}\Big((W_t)^T W_t, \tfrac{1}{d}\|\nabla_X\Psi(X_t; a_{t+1}, \epsilon_{t+1})\|^2, \mathcal{R}(X_t)\Big). \end{aligned} \quad (2.4)$$

In this definition, for functions taking integer arguments, we extend them to real-valued inputs by first taking the floor function of its argument. Note that the adaptive learning rates can depend on the whole history of stochastic iterates (N_k), gradients (G_k), and risk (Q_k) via this definition.

We also define a conditional expectation version of G_k where the filtration $\mathcal{F}_k = \sigma(X^\star, X_0, \ldots, X_k)$:

$$\mathcal{G}_k(t) \stackrel{\text{def}}{=} \mathbf{1}_{\{t<k\}}(\cdot)\frac{1}{d}\mathbb{E}[\|\nabla_X\Psi(X_t; a_{t+1}, \epsilon_{t+1})\|^2|\mathcal{F}_t] \quad \text{for } t\ge 0.$$

With this, we impose the following learning rate condition.

---

[4]Note that cases where $\frac{\mathrm{Tr}(K^2)}{d} = o(d)$ can lead to dynamics that converge to full-batch gradient flow. While our theorem specifically addresses the scenario where the intrinsic dimension, $\mathrm{Dim}(K) \stackrel{\text{def}}{=} \mathrm{Tr}(K)/\|K\|_{\mathrm{op}}$, satisfies $\mathrm{Dim}(K) = \Theta(d)$, other cases, such as $\mathrm{Dim}(K) = o(d)$, may require different learning rate scalings.

[5]$D([0,\infty))$ is the càdlàg function class on $[0,\infty)$.

Assumption 2.1.6 (Learning rate)**.** *The learning rate function* $\gamma : \mathbb{R}_+ \times D([0,\infty)) \times D([0,\infty)) \times D([0,\infty)) \to \mathbb{R}$ *is* α*-pseudo-Lipschitz with constant* $L(\gamma)$ *(independent of* d*) in* $D([0,\infty)) \times D([0,\infty)) \times D([0,\infty))$*. Moreover, for some constant* $C = C(\gamma) > 0$ *independent of* d *and* $\delta > 0$*,*

$$\mathbb{E}\left[|\gamma(k, f, G_k(d\,\times\cdot), q) - \gamma(k, f, \mathcal{G}_k(d\,\times\cdot), q)| \,|\, \mathcal{F}_k\right] \le Cd^{-\delta}(1 + \|f\|_\infty^\alpha + \|q\|_\infty^\alpha) \quad w.o.p. \tag{2.5}$$

Finally, γ *is bounded, i.e., there exists a constant* $\hat{C} = \hat{C}(\gamma) > 0$ *independent of* d *so that*

$$\gamma(k, f, g, q) \le \hat{C}(1 + \|f\|_\infty^\alpha + \|q\|_\infty^\alpha + \|g\|_\infty^\alpha). \tag{2.6}$$

The inequality (2.5) ensures that the learning rate concentrates around the mean behavior of the stochastic gradients. Many well-known adaptive stepsizes satisfy (2.4) and Assumption 2.1.6 including AdaGrad-Norm, DoG, D-Adaptation, and RMSProp (see Table 2.2, Sec. A.1, and Sec. A.3.3).

2.2 Deterministic dynamics for SGD with adaptive learning rates

Intuition for deriving dynamics: The risk $\mathcal{R}(X)$ and Fisher matrix can be evaluated solely in terms of the covariance matrix B. Thus, to know the evolution of the risk over time, it would suffice to know the evolution of B. Alas, except in the isotropic case where K is a multiple of the identity, the evolution of B is not autonomous (i.e., its time evolution depends on other unknown variables). However, if we let (λ_i, ω_i) be the eigenvalues and corresponding orthonormal eigenvectors of K, we can consider projections $V_i(X_k) = d \cdot W_k^T \omega_i \omega_i^T W_k$, and it turns out that these behave autonomously.

Example: Least Squares. One canonical example of (2.2) is least squares, where we aim to recover the target $X^\star$ given noisy observations $\langle a, X^\star\rangle + \epsilon$. In this case, the *least squares problem* is

$$\min_{X\in\mathbb{R}^d} \left\{\mathcal{R}(X) = \tfrac{1}{2}\,\mathbb{E}_{a,\epsilon}[(\langle a, X - X^\star\rangle - \epsilon)^2] = \tfrac{1}{2}\omega^2 + \tfrac{1}{2}(X - X^\star)^T K (X - X^\star)\right\}. \tag{2.7}$$

The pair of functions h (Assumption 2.1.3) and I (Assumption 2.1.4) can be evaluated simply:

$$h(B(W)) = \tfrac{1}{2} I(B(W)) = \tfrac{1}{2}(X - X^\star)^T K (X - X^\star) + \tfrac{1}{2}\omega^2.$$

The deterministic dynamics for the risk $\mathscr{R}(t)$ in this case can be simplified to:

$$\mathscr{R}(t) = \tfrac{1}{2}(X_0 - X^\star)^T K e^{-2K\int_0^t \gamma_s \,\mathrm{d}s}(X_0 - X^\star) + \tfrac{1}{2}\omega^2 + \tfrac{1}{d}\int_0^t \gamma_s^2 \operatorname{Tr}(K^2 e^{-2K\int_s^t \gamma_\tau \,\mathrm{d}\tau})\mathscr{R}(s)\,\mathrm{d}s.$$

This is a convolution Volterra equation with a convergence threshold of $\gamma_t < \frac{2d}{\operatorname{Tr} K}$ [20, 69, 72, 73].

In the noiseless label case (i.e., $\epsilon = 0$), the risk is given by $\mathscr{R}(t) = \frac{1}{2d}\sum_{i=1}^d \lambda_i \mathscr{D}_i^2(t)$. Using the ODEs in (2.9), we get the following deterministic equivalent ODE for the $\mathscr{D}_i^2$'s:

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}_i^2(t) = -2\gamma_t\lambda_i\mathscr{D}_i^2(t) + 2\gamma_t^2\lambda_i\mathscr{R}(t). \tag{2.8}$$

We will perform a deep analysis of the dynamics of the learning rate on least squares (2.7), which will generalize to settings where the outer function f is strongly convex (see A.4.1).

Deterministic dynamics. To derive deterministic dynamics, we make the following change to continuous time by setting

$$k \text{ iterations of SGD} = \lfloor td \rfloor, \quad \text{where } t \in \mathbb{R} \text{ is the continuous time parameter.}$$

This time change is necessary, as when we scale the size of the problem, more time is needed to solve the underlying problem. This scaling law scales SGD so all training dynamics live on the same space. One can solve a smaller d problem and scale it to recover the training dynamics of the larger problem.[6]

We now introduce a coupled system of differential equations, which will allow us to model the behaviour of our learning algorithms. For the ith (λ_i, ω_i)-eigenvalue/eigenvector of K, set

$$\mathscr{V}_i(t) \overset{\text{def}}{=} \begin{bmatrix} \mathscr{V}_{11,i}(t) & \mathscr{V}_{12,i}(t) \\ \mathscr{V}_{12,i}(t) & \mathscr{V}_{22,i}(t) \end{bmatrix} \text{ and averaging over } i,\ \mathscr{B}(t) \overset{\text{def}}{=} \frac{1}{d}\sum_{i=1}^d \lambda_i \mathscr{V}_i(t).$$

[6] Note that, holding time fixed, we perform $O(d)$ gradient updates for a problem of dimension d. For the problems considered here, this scaling leads to consistent dynamics, but there do exist related problems where a different scaling is more appropriate. For example, under random initialization, to capture the escape of phase retrieval from the high-dimensional saddle, $O(d \log d)$ iterations are needed; see for example [87].

The $\mathscr{V}_i(t)$ and $\mathscr{B}(t)$ are deterministic continuous analogues of $V_i(X_{td})$ and $B(X_{td})$ respectively. Define the following continuous analogues

$$\nabla h(\mathscr{B}(t)) \stackrel{\text{def}}{=} \begin{bmatrix} H_{1,t} & H_{2,t} \\ H_{2,t} & H_{3,t} \end{bmatrix}, \ \mathscr{N}(t) \stackrel{\text{def}}{=} \frac{1}{d}\sum_{i=1}^{d} \mathscr{V}_i(t), \ \mathscr{R}(t) \stackrel{\text{def}}{=} h(\mathscr{B}(t)), \quad \mathscr{I}(t) \stackrel{\text{def}}{=} I(\mathscr{B}(t)),$$
$$\text{and finally } \gamma_t \stackrel{\text{def}}{=} \gamma(t, 1_{\{\cdot \le t\}}\mathscr{N}(\cdot), \tfrac{\text{Tr}(K)}{d} 1_{\{\cdot \le t\}}\mathscr{I}(\cdot), 1_{\{\cdot \le t\}}\mathscr{R}(\cdot)).$$

We now introduce a system of coupled ODEs for each (λ_i, ω_i)-eigenvalue/eigenvector pair of K

$$\begin{aligned} \mathrm{d}\mathscr{V}_{11,i}(t) &= -2\lambda_i\gamma_t \left(\mathscr{V}_{11,i}(t)H_{1,t} + H_{1,t}\mathscr{V}_{11,i}(t) + \mathscr{V}_{12,i}(t)H_{2,t} + H_{2,t}\mathscr{V}_{12,i}(t)\right) + \lambda_i\gamma_t^2\mathscr{I}(t), \\ \mathrm{d}\mathscr{V}_{12,i}(t) &= -2\lambda_i\gamma_t \left(H_{1,t}\mathscr{V}_{12,i}(t) + H_{2,t}\mathscr{V}_{22,i}(t)\right) \end{aligned} \tag{2.9}$$

with the initialization of $\mathscr{V}_i(0)$ given by $V_i(X_0)$. We finally state the deterministic dynamics for the risk and learning rate.

Theorem 2.2.1. *Under Assumptions 2.1.1, 2.1.2, 2.1.3, 2.1.4, 2.1.5, 2.1.6, then for any $\varepsilon \in (0, \frac{1}{2})$ and any $T > 0$*

$$\sup_{0 \le t \le T} \left\| \begin{pmatrix} \mathcal{R}(X_{\lfloor td \rfloor}) \\ \mathfrak{g}_{\lfloor td \rfloor} \end{pmatrix} - \begin{pmatrix} \mathscr{R}(t) \\ \gamma_t \end{pmatrix} \right\| < d^{-\varepsilon}, \quad w.o.p. \tag{2.10}$$

The same statements hold comparing $W_{td}^T W_{td}$ to $\mathscr{N}(t)$ and $W_{td}^T K W_{td}$ to $\mathscr{B}(t)$.

In fact, we can derive deterministic dynamics for a large class of statistics which are linear combinations of $\mathscr{V}(t)$ and functions thereof (See Theorem A.2.4, and Corollary A.2.5).

One important corollary is a deterministic limit for the distance to optimality, $D^2(X_k) = \|X_k - X^\star\|^2$, which is a quadratic form of $W_k^T W_k$ and hence covered by Thm. 2.2.1. The equivalent deterministic dynamics are

$$\mathscr{D}^2(t) = \frac{1}{d}\sum_{i=1}^{d} \mathscr{D}_i^2(t) = \frac{1}{d}\sum_{i=1}^{d} (\mathscr{V}_{11,i}(t) - 2\mathscr{V}_{12,i}(t) + \mathscr{V}_{22,i}(t)), \tag{2.11}$$

where $\mathscr{D}_i^2(t)$ corresponds $D_i^2(X_k) \stackrel{\text{def}}{=} d \times (\langle X_k - X^\star, \omega_i \rangle)^2$.

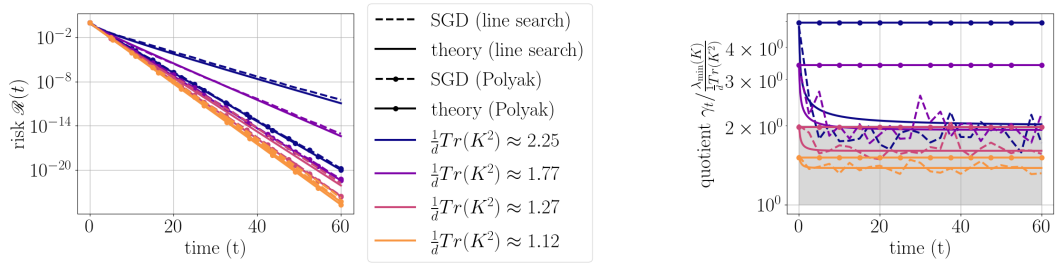

Figure 2.2: **Comparison for Exact Line Search and Polyak Stepsize** on a noiseless least squares problem. The left plot illustrates the convergence of the risk function, while the right plot depicts the convergence of the quotient $\gamma_t / \frac{\lambda_{\min}(K)}{\frac{1}{d}\operatorname{Tr}(K^2)}$ for Polyak stepsize and exact line search. Both plots highlight the implication of equation (2.13) in high-dimensional settings, where a broader spectrum of K results in $\frac{\lambda_{\min}(K)}{\frac{1}{d}\operatorname{Tr}(K^2)} \ll \frac{1}{\frac{1}{d}\operatorname{Tr}(K)}$, indicating slower risk convergence and poorer performance of exact line search (unmarked) as it deviates from the Polyak stepsize (circle markers) . The gray shaded region demonstrates that equation (2.13) is satisfied. See Appendix A.8 for simulation details.

2.3 Idealized Exact Line Search and Polyak Stepsize

In this section, we consider two classical idealized algorithms – *exact line search* and *Polyak stepsize*. In deterministic optimization, these learning rate strategies are chosen so that the function value (exact line search) or distance to optimality (Polyak) produces the largest decrease in function value (resp. distance to optimality) at the next iteration. For stochastic algorithms, we can ask this to hold for the deterministic equivalent to the risk $\mathscr{R}(t)$ (resp. distance to optimality, $\mathscr{D}(t)$) since we know that SGD is close to these deterministic equivalents. Thus, the question is: what choice of learning rate decreases the $\mathscr{R}(t)$ (*exact line search*) and/or $\mathscr{D}(t)$ (*Polyak stepsize*)? We will restrict to least squares in this section – see Appendix A.6.1 and A.6.2 for general functions as well as proofs for least squares. These are idealized algorithms because we can not implement them as they require distributional knowledge of a or $X^\star$. Despite this, they provide a basis for more practical algorithms.

Polyak Stepsize. A natural threshold to consider is the largest learning rate such that $\mathrm{d}\mathscr{D}(t) < 0$, which we denote by $\bar{\gamma}_t^{\mathscr{D}}$. Using the least squares ODE (2.8), this is precisely

$$\bar{\gamma}_t^{\mathcal{D}} = \frac{(2\mathscr{R}(t)-\omega^2)}{\frac{\mathrm{Tr}(K)}{d}\mathscr{R}(t)} \quad \text{and} \quad \bar{\mathfrak{g}}_k^{\mathscr{D}} = \frac{(2\mathcal{R}(X_k)-\omega^2)}{\frac{\mathrm{Tr}(K)}{d}\mathcal{R}(X_k)}. \tag{2.12}$$

Without label noise, (2.12) simplifies to $\bar{\gamma}_t^{\mathcal{D}} = \bar{\mathfrak{g}}_k^{\mathscr{D}} = \frac{2}{\mathrm{Tr}(K)/d}$, the exact threshold for convergence of least squares.

A greedy stepsize strategy would maximize the decrease in the distance to optimality at each iteration, denoted by us as *Polyak stepsize*, $\gamma_t^{\text{Polyak}} \in \arg\min_\gamma \mathrm{d}\mathscr{D}(t)$. In the case of least squares, this is

$$\gamma_t^{\text{Polyak}} = \tfrac{1}{2}\bar{\gamma}_t^{\mathcal{D}} \quad \text{and} \quad \mathfrak{g}_k^{\text{Polyak}} = \tfrac{1}{2}\bar{\mathfrak{g}}_k^{\mathcal{D}}.$$

The latter yields the optimal fixed learning rate (up to absolute constant factors) for a noiseless target on a least squares problem [57, 71]).[7]

Exact Line Search. In the context of risk, using (2.8) and noting that $\mathscr{R}(t) = \frac{1}{2d}\sum_{i=1}^d \lambda_i \mathscr{D}_i^2(t)$, we can find $\gamma_t^{\text{line}} \in \arg\min \mathrm{d}\mathscr{R}(t)$; i.e., the greedy learning rate that decreases the risk the most in the next iteration. We call this *exact line search*. Expressions for the learning rates are given in Table 2.2, (c.f. Appendix A.6.1 for general losses). Because these come from ODEs, we can use ODE theory to give exact limiting values for the deterministic equivalent of $\mathfrak{g}_k^{\text{line}}$.

Proposition 2.3.1. *[Limiting learning rate; line search on noiseless least squares] Consider the noiseless ($\omega = 0$) least squares problem (2.7) . Then the learning rate is always lower bounded by*

$$\frac{\lambda_{\min}(K)}{\frac{1}{d}\mathrm{Tr}(K^2)} \le \gamma_t^{\text{line}} \quad \textit{for all } t \ge 0.$$

Moreover, suppose K has only two distinct eigenvalues $\lambda_1 > \lambda_2 > 0$, i.e., K has $d/2$ eigenvalues equal to λ_1 eigenvalues and $d/2$ eigenvalues equal to λ_2. Then

$$\frac{\lambda_{\min}(K)}{\frac{1}{d}\mathrm{Tr}(K^2)} \le \lim_{t\to\infty} \gamma_t^{\text{line}} \le \frac{2\lambda_{\min}(K)}{\frac{1}{d}\mathrm{Tr}(K^2)}. \tag{2.13}$$

[7] The Polyak stepsize we analyze in this paper differs slightly from the "classic" stepsize in the literature, that is, $\frac{\mathscr{R}(X_k)-\mathscr{R}(X^*)}{||\nabla\mathscr{R}(X_k)||^2}$. Rather than using this form, we skip an approximation step in the derivation [39] and use the exactly optimal form. Both variations of the Polyak stepsize can be analyzed under our assumptions; the choice was admittedly somewhat arbitrary. (Note that in the case of least squares, the two stepsizes coincide.)

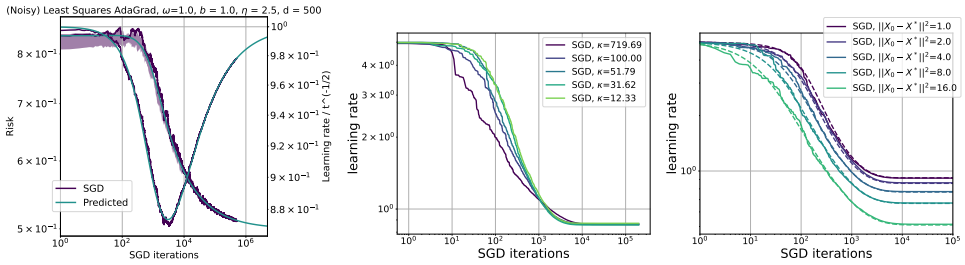

Figure 2.3: **Quantities effecting AdaGrad-Norm learning rate.** *(left):* Effect of noise ($\omega = 1.0$) on risk (left axis) and learning rate (right axis). Depicted is $\frac{\text{learning rate}}{\text{asymptotic}}$ so it approaches 1. *(Center, right)*: Noiseless least squares ($\omega = 0$). As predicted in Prop. 2.4.2, $\lim_{t\to\infty} \gamma_t$ depends on avg. eig. of K ($\mathrm{Tr}(K)/d$) and $\|X_0 - X^\star\|^2$ but not $\kappa = \lambda_{\max}/\lambda_{\min}$. See Appendix A.8 for simulation details.

For a proof and explicit formula for $\lim_{t\to\infty} \gamma_t^{\text{line}}$, see Section A.6.2. Hence, being greedy for the risk in a sufficiently anisotropic setting will badly underperform Polyak stepsize (see Fig. 2.2).

2.4 AdaGrad-Norm analysis

In this section, we analyze the behavior of AdaGrad-Norm learning rate in the least squares setting (see Sec. A.4 for general strongly convex functions). In the presence of additive noise, the AdaGrad-Norm learning rate decays like $t^{-1/2}$, regardless of the data covariance K. In contrast, the model with no noise exhibits a learning rate that depends on the spectrum of K, as illustrated in Figure 2.3. The learning rate is bounded below by a constant when $\lambda_{\min}(K) > 0$ is fixed as $d \to \infty$, and we quantify this lower bound. If the limiting spectral measure of K has unbounded density near 0 (e.g. power law spectrum), then the learning rate can approach zero and we quantify the rate of this convergence in the least squares setting as a function of spectral parameters.

For least squares with additive noise, the learning rate asymptotic $\gamma_t \asymp \eta/(b^2 + \frac{\omega^2}{d}\mathrm{Tr}(K)t)^{(1/2)}$ is the fastest decay that AdaGrad-Norm can exhibit. In contrast, the propositions below concern the noiseless case where, for various covariance

examples, the decay rate of γ_t changes. This is tightly connected to whether the risk is integrable or not. In the simple case of identity covariance, we obtain a closed formula for the trajectory of the integral of the risk and therefore also the learning rate.

Proposition 2.4.1. *In the case of identity covariance ($K = I_d$), the risk solves the differential equation*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{R}(t) = \frac{\eta^2 \mathscr{R}(t)}{b^2 + 2\int_0^t \mathscr{R}(s)\,\mathrm{d}s} - \frac{2\eta \mathscr{R}(t)}{\sqrt{b^2 + 2\int_0^t \mathscr{R}(s)\,\mathrm{d}s}}, \tag{2.14}$$

The solution $\int_0^t \mathscr{R}(s)\,\mathrm{d}s$ approaches (from below) a positive constant which yields a computable lower bound to which γ_t will converge. Generalizing this to a broader class of covariance matrices, we get the next proposition, which captures the dependence of γ_t on $\mathrm{Tr}(K)$.

Proposition 2.4.2. *Suppose $\frac{1}{d}\mathrm{Tr}(K) \leq b/\eta$, and that $\int_0^\infty \mathscr{R}(s)\gamma_s\,\mathrm{d}s < \infty$ with γ_s as in Table 2.2 (AdaGrad-Norm for least squares), then $\gamma_t \asymp \frac{1}{\frac{b}{\eta} + \frac{\eta^2}{4d}\mathrm{Tr}(K)\mathscr{D}^2(0)}$.*

An analog of Proposition 2.4.2 for the strongly convex setting appears in Sec. A.4 (see Prop. A.4.4). We now consider two cases in which, as $d \to \infty$, there are eigenvalues of K arbitrarily close to 0.

Proposition 2.4.3. *Assume that, for some $C > 0$, the number of eigenvalues of K below C is $o(d)$, and that $\langle X^\star, \omega_i \rangle = O(d^{-1/2})$ for all i, (i.e. $X^\star$ is not concentrated in any eigenvector direction). Then, with the initialization $X_0 = 0$, there exists some $\tilde{\gamma} > 0$ such that $\gamma_t > \tilde{\gamma}$ for all $t > 0$.*

Proposition 2.4.4. *Let K have a spectrum that converges as $d \to \infty$ to the power law measure $\rho(\lambda) = (1-\beta)\lambda^{-\beta}\mathbf{1}_{(0,1)}$, for some $\beta < 1$[8], and suppose that $\mathscr{D}_i^2(0) \sim \lambda_i^{-\delta}$ for $\delta \geq 0$. Then:*

- *For $1 > \beta + \delta$, there exists $\tilde{\gamma}$ such that $\gamma_t \geq \tilde{\gamma}$, and $\mathscr{R}(t) \asymp_{\delta,\beta} t^{\beta+\delta-2}$ for all $t \geq 1$.*

[8]Our result can be compared to existing findings for SGD under power-law distributions in [13, 82, 86]. While these works explore similar assumptions regarding the covariance matrix spectrum, they do not address the high-dimensional regime with diverging $\mathrm{Tr}(K)$, focusing primarily on $\beta > 1$.

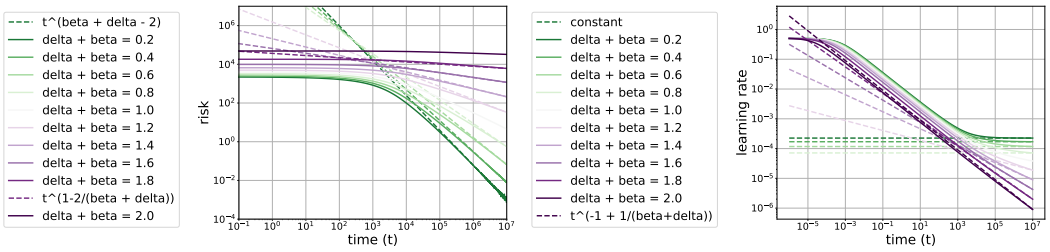

Figure 2.4: **Power law covariance in AdaGrad Norm** on a least squares problem. Ran exact predictions (ODE) for the risk and learning rate (solid lines). Dashed lines give the predictions from Prop. 2.4.4 which *match experimental results exactly.* **Phase transition as $\delta + \beta$ varies.** When $\delta + \beta < 1$ (green), the learning rate *(right)* is constant as $t \to \infty$. In contrast, when $2 > \delta + \beta > 1$ (purple), the learning rate decreases at a rate $t^{-1+1/(\beta+\delta)}$ with $\delta + \beta = 1$ (white) where the change occurs. Same phase transition occurs in the sublinear rate of the risk decay *(left)* (see Prop. 2.4.4).

- *For* $1 < \beta + \delta < 2$, $\gamma_t \asymp_{\delta,\beta} t_{\delta,\beta}^{-1+\frac{1}{\beta+\delta}}$, *and* $\mathscr{R}(t) \asymp_{\delta,\beta} t^{-\frac{2}{\beta+\delta}+1}$ *for all* $t \geq 1$.

- *For* $1 = \beta + \delta$, $\gamma_t \asymp_{\delta,\beta} \frac{1}{\log(t+1)}$, *and* $\mathscr{R}(t) \asymp_{\delta,\beta} (\frac{t}{\log(t+1)})^{-1}$ *for all,* $t \geq 1$.

This proposition shows non-trivial decay of the learning rate is dictated by the residuals (distance to optimality at initialization) and the spectrum of K. We note that $\delta = 0$ corresponds to uniform contribution of each mode (e.g. X_0 normally distributed). As the eigenmodes of the residuals become more localized, the decay of the learning rate is closer to the behaviour in the presence of additive noise. Furthermore, the scaling behaviour of the loss is affected by the structure of the AdaGrad-Norm algorithm (see Fig. 2.4). Lastly, constant stepsize SGD yields $\mathscr{R}(t) \asymp t^{\beta+\delta-2}$, with no transition occurring at $\beta + \delta = 1$.

Proofs of the above propositions, in a slightly more general setting, are deferred to Sec. A.4.

2.5 Conclusions and Limitations

This work studies stochastic adaptive optimization algorithms when data size and parameter size are large, allowing for nonconvex and nonlinear risk functions, as well as data with general covariance structure. The theory shows a concentration of the risk, the learning rate and other key functions to a deterministic limit, which is described by a set of ODEs. The theory is then used to derive the asymptotic behavior of the AdaGrad-Norm and idealized exact line search on strongly convex and least square problems, revealing the influence of the covariance matrix structure on the optimization. A potential extension of this work would be to study other adaptive algorithms such as D-adaptation, DOG, and RMSprop which are covered by the theory. Studying the asymptotic behavior of the risk and the learning rate may improve our understanding of the performance and scalability of these algorithms on more realistic data. Another important application of the theory would be to analyze the ODEs presented here on nonconvex problems.

The current form of the theory is limited to Gaussian data, though many parts of the proof can be extended easily beyond Gaussian data. The main ODE comparison theorem is also only tuned for analyzing problem setups where the trace of the covariance is on the order of the ambient dimension; when the trace of the covariance is much smaller than ambient dimension, other stepsize scalings of SGD are needed. In addition, the analysis is limited to the streaming stochastic adaptive methods. We conjecture that a similar deterministic equivalent holds also for multi-pass algorithms at least for convex problems. This has already been shown in the least square problem for SGD with a fixed deterministic learning rate [71, 74]. Lastly, numerical simulations on real datasets (e.g., CIFAR-5m) suggests that the predicted risk derived by our theory matches the empirical risk of multipass SGD beyond Gaussian data (see for example Figure A.2).

Acknowledgments and Disclosure of Funding

E. Collins-Woodfin was supported by Fonds de recherche du Québec – Nature et technologies (FRQNT) postdoctoral training scholarship and Centre de recherches mathématiques (CRM) Applied math postdoctoral fellowship. Research of B. García Malaxechebarría was in part funded by NSF DMS 2023166 (NSF TRIPODS II). Research by E. Paquette was supported by a Discovery Grant from the Natural Science and Engineering Council (NSERC). C. Paquette is a Canadian Institute for Advanced Research (CIFAR) AI chair, Quebec AI Institute (MILA) and a Sloan Research Fellow in Computer Science (2024). C. Paquette was supported by a Discovery Grant from the Natural Science and Engineering Research Council (NSERC) of Canada, NSERC CREATE grant Interdisciplinary Math and Artificial Intelligence Program (INTER-MATH-AI), Google research grant, and Fonds de recherche du Québec – Nature et technologies (FRQNT) New University Researcher's Start-Up Program. Additional revenues related to this work: C. Paquette has 20% part-time employment at Google DeepMind.

Chapter 3

HIGH-DIMENSIONAL LIMIT OF SGD FOR DIAGONAL LINEAR NETWORKS

Joint work with Courtney Paquette, Maryam Fazel, and Dmitriy Drusvyatskiy [59]

Abstract. Understanding the behavior of stochastic gradient methods is a central problem in modern machine learning. Recent work has highlighted diagonal linear networks as a simplified yet expressive setting for analyzing the optimization and generalization properties of neural models. In this work, we show that in the high-dimensional regime, stochastic gradient descent on diagonal linear networks is well-approximated by continuous dynamics governed by a stochastic differential equation (SDE), which explicitly decouples the drift from the gradient noise. We further derive a deterministic partial differential equation whose solution propagates the relevant state of the iterates and characterizes the time evolution of a broad class of observable statistics, including the risk, curvature, and other metrics for optimality. Finally, we show that, under a suitable parametrization, the stochastic dynamics are globally well posed and converge exponentially fast to zero risk with high probability, yielding a fully explicit non-asymptotic description of their long-time behavior. Numerical simulations corroborate our theoretical findings.

3.1 Introduction

Diagonal linear networks serve as an appealing and analytically tractable model for investigating various phenomena that are observed in deep learning. For example, diagonal linear networks display intriguing forms of implicit bias, whereby stochastic optimization algorithms tend to favor low-complexity solutions even in the absence of explicit regularization [2, 31, 75]. Nonetheless, analyzing SGD for diagonal linear networks is still challenging. In particular, classical analyses of SGD are frequently too coarse, relying on worst-case guarantees that hold only for very small stepsizes. In practice, however, the phenomena of interest emerge

only when using relatively large stepsizes. The central difficulty stems from the intrinsic randomness of SGD: the data used at each iteration are random, and the update rule itself is randomized through the sampling of indices that determine the stochastic gradient. This motivates the development of sharper analytical frameworks for SGD—ones that capture the evolution of key quantities such as risk and curvature, while remaining mathematically tractable.

One approach is to study gradient flow or diffusion-based approximations [75, 88]. Gradient flow emerges in the limit of a vanishing stepsize, while diffusion-based approximations incorporate stochasticity through an added noise term. Importantly, these methods can often only be rigorously justified as accurate approximations of SGD when the stepsize is extremely small ($\gamma \to 0$)—a regime that is largely unrealistic in practice.

Indeed, SGD for modern applications is typically run with relatively large stepsizes and in high-dimensional settings, where neither the stepsize nor the stochasticity can be treated as infinitesimal. The perspective adopted in this paper is to instead exploit high dimensionality—namely, the large number of parameters—as the primary simplifying mechanism. In the high-dimensional limit, concentration phenomena emerge, yielding exact, deterministic expressions for the risk (and other quantities of interest) at every iteration, even when the stepsize is large.

Specifically, in this work, we derive a continuous-time stochastic differential equation (SDE) that tracks discrete-time SGD in this high-dimensional, large stepsize regime for diagonal linear networks. Using this SDE as a foundation, we obtain a deterministic PDE that propagates the relevant state of the iterates and thereby predicts the time evolution of a broad class of key observable statistics. *To the best of our knowledge, our result provides the first deterministic description of the discrete-time SGD trajectory in high dimensions with fixed, non-vanishing stepsizes for diagonal linear networks, placing them among the first nonlinearly parametrized models admitting such a description beyond gradient-flow or stochastic-gradient-flow limits.*

This effectively allows for the analysis of SGD *without running it* (See Figure 3.1): solving the deterministic evolution provides learning curves without the need for expensive Monte Carlo simulations across mini-batch realizations. While the PDE offers theoretical insight,

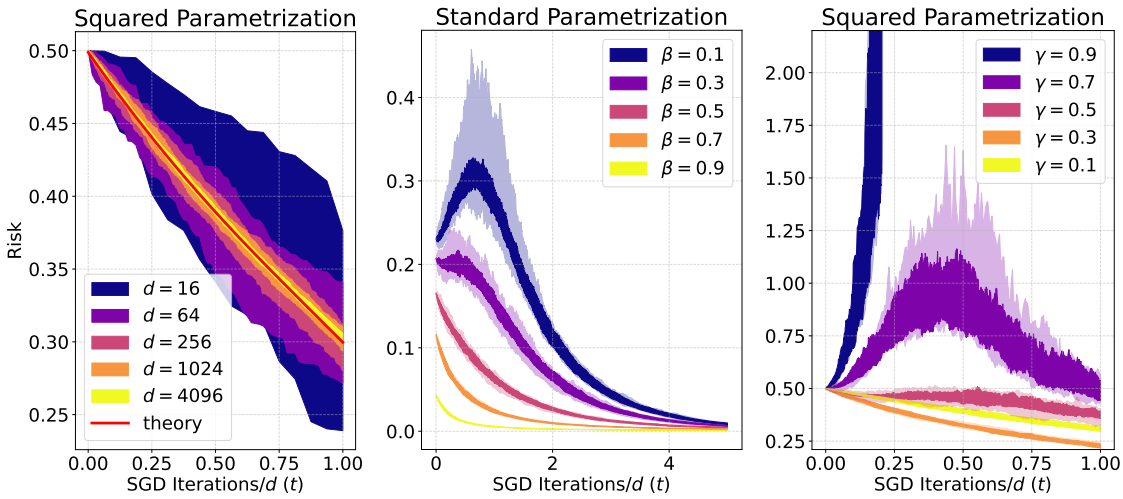

Figure 3.1: **Three views of empirical risk dynamics for SGD on a diagonal linear network.** *Left:* Covariance $K = I_d$. As d increases, the risk trajectory of SGD concentrates around a deterministic limit (red) described in Theorem 3.3.7. *Middle:* Power-law covariance spectrum. The homogenized SGD (transparent) from Theorem 3.3.7 closely tracks SGD (opaque) over a range of power-law exponents β in dimension $d = 10^3$. *Right:* Covariance $K = I_d$ in dimension $d = 10^3$. Varying the stepsize γ reveals distinct convergence/divergence regimes; the homogenized prediction remains accurate even for stepsizes above the convergence threshold. Left and right panels use the parametrization (B.5.4), whereas the middle one uses the parametrization (B.5.3). See Appendix B.6 for simulation details.

the SDE remains a powerful practical tool—it is significantly cheaper to simulate, and Itô calculus provides a direct route to closed evolution equations for many statistics of interest, which we validate numerically.

Setting the stage, our work targets a general class of problems where the risk exhibits the form:

$$\mathcal{R}(x) = \mathbb{E}_a \left[f\left(\frac{1}{\sqrt{d}} \begin{pmatrix} \psi(x) \\ \beta^* \end{pmatrix}^\top a \right) \right] \qquad \text{for } a \sim \mathcal{N}(0, K). \tag{3.1}$$

Here $K \in \mathbb{R}^{d\times d}$ is the covariance of the data, $\beta^* \in \mathbb{R}^d$ should be thought of as a ground truth vector, $\psi(x)$ maps the features $x \in \mathbb{R}^{2d}$ into $\mathbb{R}^d$, and f is a loss such as mean-square error

or logistic. The primary example of (3.1) is the mean-squared error of a two-layer *diagonal linear network*, given by

$$\min_{u,v\in\mathbb{R}^d} \frac{1}{2d}\,\mathbb{E}_a\langle u\odot v-\beta^*,a\rangle^2, \tag{3.2}$$

where we set $x=(u,v)$ with $u,v\in\mathbb{R}^d$ and $u\odot v$ is the Hadamard product. Thus, even though the network is linear in its prediction, the optimization variables enter the risk nonlinearly. This nonlinear parametrization is precisely what makes the high-dimensional closure problem substantially more delicate than in linear or generalized linear models. It is standard to check that (3.2) can be re-parametrized by a difference of squares [64, 75]:

$$\min_{u,v\in\mathbb{R}^d} \frac{1}{4d}\,\mathbb{E}_a\langle u^2-v^2-\beta^*,a\rangle^2. \tag{3.3}$$

Diagonal linear networks, in both formulations, have recently garnered attention due to their ability to recover sparse solutions [38, 47, 75, 83].

To solve (3.1), we run *streaming* stochastic gradient descent under a deterministic stepsize schedule γ_k, that is, at each iteration we generate a fresh sample of data points $a_{k+1}\sim\mathcal{N}(0,K)$ and update the iterates according to the rule

$$x_{k+1}=x_k-\gamma_k\,\nabla_x\Psi(x_k;a_{k+1}),$$

where $\Psi(x,a)$ is the integrand in equation (3.1). To analyze SGD, we introduce a stochastic differential equation, called the *homogenized SGD*:

$$\mathrm{d}\mathcal{X}_t=-\gamma(t)\,d\,\nabla\mathcal{R}(\mathcal{X}_t)\mathrm{d}t+\gamma(t)\sqrt{\mathbb{E}_a\left[\nabla f\left(\tfrac{1}{\sqrt{d}}\psi(\mathcal{X}_t)^\top a\right)^2\right]}\,\nabla\psi(\mathcal{X}_t)^\top\sqrt{K}\,\mathrm{d}\mathfrak{B}_t, \tag{3.4}$$

where the risk $\mathcal{R}$ and inner map ψ are defined in (3.1), the initial conditions are given by $\mathcal{X}_0=x_0$, and $\mathrm{d}\mathfrak{B}_t$ is the differential of a standard Brownian motion in $\mathbb{R}^d$. In this work, we rigorously show that the homogenized SGD behaves like SGD in high dimensions, even with stepsizes at or above the convergence threshold (see Fig. 3.1).

To enable a comparison between SGD and homogenized SGD, we extend the discrete-time iterate sequence $\{x_k\}$ to continuous time. The k-th iterate of SGD corresponds to the continuous time parameter t in homogenized SGD via $k=\lfloor td\rfloor$. Thus, when $t=1$, SGD has completed d updates.

Main Contributions. Our first main result shows that SGD and homogenized SGD are close in the sense that for a wide class of functions φ (statistics) the values $\varphi(x_{\lfloor td \rfloor})$ and $\varphi(\mathcal{X}_t)$ along the two trajectories are uniformly close. Two particularly informative examples are the risk $\mathcal{R}(x)$ and the scaled Hessian-trace statistic $\frac{1}{d}\operatorname{Tr}\big(\nabla^2\mathcal{R}(x)\big)$, a scalar measure of curvature (often interpreted as a notion of sharpness) along the optimization path. See Figures 3.1 and 3.2 for an illustration.

Theorem 3.1.1 (Informal)**.** *Under mild conditions, formalized in Theorem 3.3.7, for any statistic φ satisfying Assumption 3.3.6, any $\varepsilon \in (0, \frac{1}{2})$ and any $T > 0$, there exists a constant C (independent of d) such that with overwhelming probability[1] the estimate holds:*

$$\sup_{0 \le t \le T} \big|\varphi(x_{\lfloor td \rfloor}) - \varphi(\mathcal{X}_t)\big| \leqslant C d^{-\varepsilon}. \tag{3.5}$$

Thus, in the high-dimensional regime, SGD is accurately described by the homogenized SGD dynamics. In particular, the SDE approximation is both accurate and analytically tractable, providing a practical tool to study the behavior of SGD while allowing fixed (and potentially large) stepsizes. This contrasts with the classical small stepsize diffusion approximation, which instead analyzes the regime where the stepsize is infinitesimally small for any fixed dimension.

Our second main result goes a step further. We show that in high dimensions, the randomness becomes negligible at the level of φ. Specifically, the processes $\varphi(x_{\lfloor td \rfloor})$ and $\varphi(\mathcal{X}_t)$ concentrate uniformly over $t \in [0, T]$ around a *deterministic function* $\phi(t)$ (see Fig. 3.1). The function $\phi(t)$ is given explicitly in terms of the solution of a deterministic partial integro-differential equation (see (3.25) and (B.6) for details). In other words, $\phi(t)$ yields a deterministic prediction for φ for large d, allowing one to analyze these statistic curves without averaging over SGD noise.

Theorem 3.1.2 (Informal)**.** *Under mild conditions, formalized in Theorem 3.3.7, for any statistic φ satisfying Assumption 3.3.6, any $\varepsilon \in (0, \frac{1}{2})$ and any $T > 0$, there exists a constant*

[1] We say an event holds *with overwhelming probability (w.o.p.)* if there is a function $\omega\colon \mathbb{N} \to \mathbb{R}$ with $\omega(d)/\log d \to \infty$ so that the event holds with probability at least $1 - e^{-\omega(d)}$.

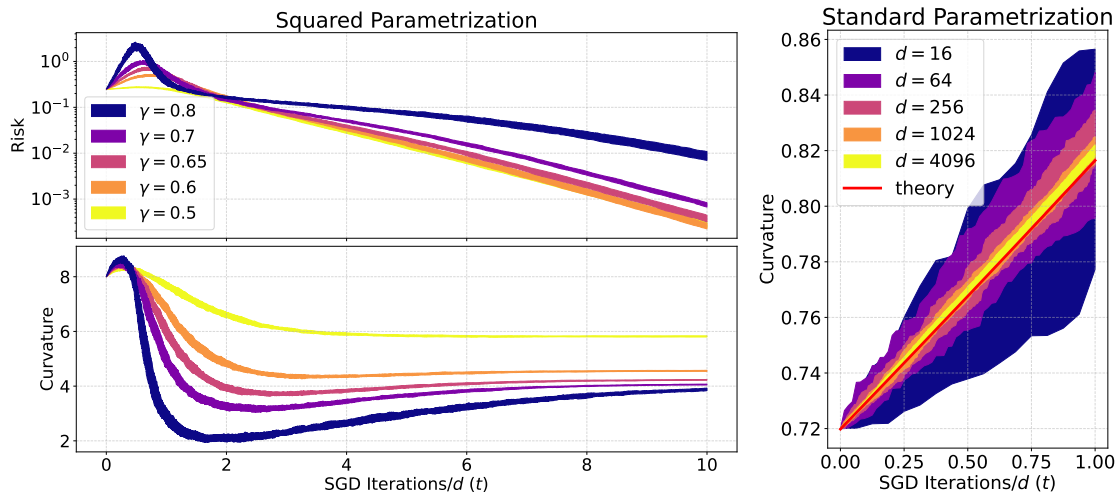

Figure 3.2: **Curvature dynamics for SGD on a diagonal linear network.** *Left:* The evolution of the curvature measured by the scaled trace of the Hessian $\frac{1}{d}\operatorname{Tr}(\nabla^2\mathcal{R})$ is shown alongside the empirical risk $\mathcal{R}$, illustrating "flat" progress in which the risk increases sharply accompanied by a marked drop in curvature as we vary the stepsize γ. *Right:* As the dimension d increases, the curvature dynamics of SGD concentrate around a deterministic limit (shown in red), as proven in Theorem 3.3.7. See Appendix B.6 for simulation details.

C (independent of d) such that with overwhelming probability the estimate holds:

$$\sup_{0\le t\le T}\left(\left|\varphi(x_{\lfloor td\rfloor}) - \phi(t)\right| + \left|\varphi(\mathcal{X}_t) - \phi(t)\right|\right) \leqslant Cd^{-\varepsilon} \tag{3.6}$$

where $\phi(t)$ is a deterministic function defined in (3.25) and (B.6).

Thus the function ϕ provides a tractable, deterministic description of the evolution of a broad class of statistics $\varphi(x)$ along the SGD trajectory in the high-dimensional regime, even for *large* stepsizes. We therefore expect this theorem to be useful for analyzing fine-grained properties of SGD such as instability, saddle point avoidance/escape, progressive sharpening, implicit bias, etc.

Lastly, our final result complements this picture by establishing global linear convergence of the stochastic dynamics themselves.

Theorem 3.1.3 (Informal)**.** *Consider homogenized SGD (3.4) for diagonal linear networks under the squared parametrization (3.3), initialized at $\mathcal{U}_{0,i} = \mathcal{V}_{0,i} = 1$ for $i = 1, \dots, d$. Then there exists a numerical constant $c > 0$ such that for any stepsize $\gamma \in (0, c)$, the risk $\mathcal{R}(\mathcal{X}_t)$ converges to zero exponentially fast. More precisely, for any $\delta \in (0, 1)$, there exist numerical constants $C, \mu > 0$ such that with probability at least $1 - \delta$ we have $\mathcal{R}(\mathcal{X}_t) \leqslant Ce^{-\mu t}$ for all $t \geqslant 0$.*

3.1.1 Literature Review

Continuous-time and stochastic-process approximations of SGD. A common approach to analyzing SGD is to replace the discrete-time iteration by a continuous-time process. Gradient flow arises in the infinitesimal-stepsize limit, but suppresses both finite-stepsize effects and stochasticity. Stochastic gradient flow and diffusion-type SDE approximations incorporate noise and can capture solution-selection phenomena not explained by gradient flow [75]. Related viewpoints include studying SGD augmented with label noise as a proxy for isolating stochastic effects during training [88]. These diffusion approximations are typically justified in small-stepsize limits, often at fixed dimension. In high-dimensional settings, stochastic-process descriptions can reveal additional regularization mechanisms, including biases induced by parameter-dependent noise [2] and repulsive interactions in eigenvalue dynamics that promote rank deficiency [81]. Beyond diffusion approximations, dynamical mean field theory (DMFT) provides another route to tracking high-dimensional training dynamics, including for SGD in Gaussian mixture classification and related batching regimes [17, 33, 62].

High-dimensional deterministic limits for SGD dynamics. Complementary to continuous-time approximations, a long line of work derives deterministic dynamical descriptions of learning in high-dimensional teacher–student and random-feature models. Early studies of multi-index and soft-committee models established ODE characterizations of the risk dynamics and showed how algorithmic choices, such as the stepsize, affect convergence and generalization [14, 15, 78, 79]. Building on this statistical-physics tradition, Goldt et al. [34] rigorously justified the resulting finite-dimensional ODE description for online SGD in

two-layer teacher–student networks with large input dimension and finitely many hidden units, including overparameterized students and the case where both layers are trained. Parallel mathematical work developed systematic scaling-limit techniques for high-dimensional online learning dynamics, often based on empirical-measure and martingale decompositions [90, 91]. More recent work emphasizes that generalization depends on the interaction between algorithm, architecture, and data distribution [34], and develops general techniques (e.g., martingale arguments) for proving high-dimensional limits [90, 91]. Comparisons across limiting regimes and refinements of these ODE descriptions are studied in [4]. Other variants show how changes in the geometry or constraints of the dynamics can lead to signal-locking and dimension-robust behavior [3, 6, 24]. Streaming and multi-pass SGD have also been analyzed through deterministic descriptions based on integral equations and homogenization-type methods [51, 63, 71, 73]. Recent high-dimensional analyses have also begun to address simplified attention/transformer-like architectures, including sequence single-index models and single-layer attention models trained by SGD or few-step gradient procedures [5, 11]; these works, however, address low-dimensional order-parameter dynamics, gradient-flow approximations, or few-step training rather than trajectory-level deterministic equivalents for discrete-time SGD in diagonal parametrizations.

Covariance structure and finite-stepsize effects. Many classical high-dimensional analyses assume isotropic data, but non-identity covariance can qualitatively change the training dynamics and is important when the population covariance is unknown. Prior work derives equations of motion under Gaussian-equivalence principles for multi-index and random-feature models [35], with extensions to deeper architectures [36]. Covariance structure has also been linked to plateau phenomena and slow phases in learning [100]. Recent work derives exact high-dimensional limits for linear regression and analyzes how non-identity covariance reshapes the optimization landscape and convergence behavior [20, 21]; related extensions to adaptive algorithms are studied in [10, 22]. These results are especially relevant at finite stepsizes, where stochasticity, data geometry, and algorithmic parameters can interact beyond the small-stepsize limits captured by diffusion approximations.

Implicit bias, flatness, and the role of stepsize. A related theme is the implicit bias of gradient-based optimization. In separable logistic regression, gradient descent converges in direction to the max-margin classifier [80], and in modern overparameterized settings SGD is often viewed as providing an implicit regularization mechanism that can improve generalization with little or no explicit regularization [101]. One influential hypothesis links this behavior to SGD's preference for flatter minima [40] and to empirical correlations between flatness and generalization [48]. Related evidence appears in structured recovery problems, where flat minima can coincide with exact recovery under suitable conditions [27]. Stepsize and batch size are key levers in this picture: larger stepsizes and smaller batch sizes have been argued to encourage exploration of flatter regions [44], and the batch-size-to-learning-rate ratio can affect generalization [40]. Initialization also plays an important role in determining the implicit bias: the initialization scale can control the transition between kernel/lazy and rich/active regimes [95], while the relative shape of the initialization can further affect the limiting solution selected by gradient methods [9]. Other proposed mechanisms include landscape-smoothing interpretations [49], random-walk models on random landscapes [42], and explanations based on batch variability and "catapult" effects [102]. At the same time, recent evidence cautions that sharpness and flatness can be tightly entangled with optimization hyperparameters such as the stepsize, and need not correlate monotonically with generalization across settings [1].

Diagonal linear neural networks. Diagonal linear networks provide a minimal yet expressive model for studying implicit bias, stochastic optimization, and solution selection in overparameterized settings. They allow precise comparisons between SGD and gradient descent or gradient flow, including regimes where SGD exhibits favorable implicit bias [75]. In sparse feature learning, empirical work suggests that large stepsizes alone may be insufficient for sparsification without stochasticity [2]. Complementary theory shows that, in sparse regression with diagonal linear networks, large stepsizes can systematically improve SGD while harming gradient descent [31].

3.1.2 High-dimensional Model Structure

We will consider objective functions $\mathcal{R}$ defined as the expectation of a composition of a simple outer function $f\colon \mathbb{R}^2 \to \mathbb{R}$ and an inner map $\psi\colon \mathbb{R}^{2d} \to \mathbb{R}^d$, given by

$$\mathcal{R}(x) := \mathbb{E}_a\, \Psi(x;a), \quad \text{for} \quad a \sim \mathcal{D} \subset \mathbb{R}^d, \quad \text{where} \quad \Psi(x;a) := f\left(\frac{1}{\sqrt{d}} \begin{pmatrix} \psi(x) \\ \beta^* \end{pmatrix}^\top a\right). \tag{3.7}$$

In the probabilistic analysis, it is important to treat f as a function of both components of its input $r = (r_1, r_2) \in \mathbb{R}^2$. However, in the optimization context, we often regard f as a function of a single real variable r_1, with the second component r_2 considered a fixed (but possibly random) parameter.

We impose the following mild Lipschitzness assumption on f. Throughout, the symbol $\|\cdot\|$ will denote the standard ℓ_2-norm on vectors and the Frobenius norm on matrices.

Assumption 3.1.4 (Pseudo-Lipschitz continuity of f)**.** The outer function $f\colon \mathbb{R}^2 \to \mathbb{R}$ is α-pseudo-Lipschitz with constant $L(f)$. That is, for all $r, \hat{r} \in \mathbb{R}^2$, it holds:

$$|f(r) - f(\hat{r})| \leqslant L(f)\, \|r - \hat{r}\|\, (1 + \|r\|^\alpha + \|\hat{r}\|^\alpha).$$

In words, pseudo-Lipschitzness stipulates polynomial growth of Lipschitz constants on balls with growing radii. Next we summarize our assumption on the data vector $a \sim \mathcal{D}$.

Assumption 3.1.5 (Data)**.** We consider data samples $a \sim \mathcal{D}$ drawn from a Gaussian distribution $\mathcal{N}(0, K)$, where the covariance $K \in \mathbb{R}^{d\times d}$ is diagonal. The entries of K are uniformly bounded, meaning its operator norm satisfies $\|K\|_{\mathrm{op}} \leqslant \bar{K}$, for some constant $\bar{K} > 0$ independent d. In addition, we assume that the trace of K scales linearly with the dimension d, that is, $\mathrm{Tr}(K) = \Theta(d)$.

Remark 3.1.6. The Gaussian assumption is used later both for concentration and for explicit conditional moment calculations. In particular, for a fixed parameter value x, the proof conditions the data vector a on its projection onto the low-dimensional subspace generated by $\psi(x)$ and β^*. Gaussian conditioning gives closed-form conditional means and covariances, which are used to identify the limiting drift and diffusion terms. The same assumption also

provides concentration bounds for the linear and quadratic forms appearing in the martingale estimates. Some of these concentration steps may extend to more general light-tailed data, but the explicit conditional moment calculations used to identify the limiting drift and diffusion explicitly rely on Gaussianity.

Remark 3.1.7. Observe that the trace condition $\mathrm{Tr}(K) = \Theta(d)$ reflects the assumption that the total variance of the input distribution remains proportional to the ambient dimension. Our main results are calibrated for this high-dimensional regime. In settings where the trace of K grows more slowly than d, alternative stepsize scalings for SGD may be required to obtain nontrivial limiting behavior.

Our main target application is that of a diagonal linear network, in either parametrization (3.2) and (3.3). In order to treat both parametrizations simultaneously, we assume that the inner function ψ has a special quadratic form.

Assumption 3.1.8 (Diagonal Linear Neural Networks)**.** The inner function $\psi\colon \mathbb{R}^{2d} \to \mathbb{R}^d$ is component-wise quadratic, meaning that each output coordinate for $i \in \{1, \dots, d\}$ is given by

$$\psi_i(u, v) = \begin{pmatrix} u_i & v_i \end{pmatrix} \mathcal{Q} \begin{pmatrix} u_i \\ v_i \end{pmatrix} + l^\top \begin{pmatrix} u_i \\ v_i \end{pmatrix} + c, \tag{3.8}$$

where $\mathcal{Q} = \begin{pmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{pmatrix} \in \mathbb{R}^{2\times 2}$ is symmetric, $l = \begin{pmatrix} l_1 \\ l_2 \end{pmatrix} \in \mathbb{R}^2$, and $c \in \mathbb{R}$.

Concretely, the formulations (3.2) and (3.3) correspond to setting $\psi(u, v) = u \odot v$ for $\mathcal{Q} = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}$, and $\psi(u, v) = u^2 - v^2$ for $\mathcal{Q} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, respectively; while setting l and c to be zero in both cases. Next, we assume that the ground truth vector β^* has bounded entries.

Assumption 3.1.9 (Parameter Scaling)**.** The entries of the signal $\beta^* \in \mathbb{R}^d$ are uniformly bounded, that is, $\|\beta^*\|_\infty \leqslant C$ for some $C > 0$ independent of d.

In what follows, for a parameter vector $x = (u, v) \in \mathbb{R}^{2d}$, we introduce the block matrix

$$W(x) := [\, \psi(x) \,|\, \beta^* \,|\, \mathbb{1}_d \,] \in \mathbb{R}^{d\times 3}, \tag{3.9}$$

and let $B(x)$ be the covariance matrix of the random vector $\frac{1}{\sqrt{d}}W(x)^\top a \in \mathbb{R}^3$, or more explicitly

$$B(x) \;:=\; \frac{1}{d}\, W(x)^\top K\, W(x) \in \mathbb{R}^{3\times 3}\,. \tag{3.10}$$

Since $\mathcal{R}(x)$ depends on x only through the Gaussian $\frac{1}{\sqrt{d}}W(x)^\top a$, the learning dynamics can be summarized by the matrix $B(x)$: there exists a function $h\colon\, \mathbb{R}^{3\times 3} \to \mathbb{R}$ such that the risk decomposes as $\mathcal{R}(x) = h\,(B(x))$. Explicitly, for both formulations (3.2) and (3.3), we obtain the same function

$$h\,(B) = \frac{1}{2}\,(B_{11} - B_{12} - B_{21} + B_{22})\,.$$

The next two assumptions simply require that (i) the risk is a smooth function of this summary (so we may differentiate through the expectation to obtain the gradient) and (ii) the corresponding second-moment quantity controlling the size of the stochastic gradient noise is equally well-behaved.

Assumption 3.1.10 (Risk Representation)**.** There is an open set $\mathcal{U} \subseteq \mathbb{R}^{3\times 3}$ such that $B(x_0) \in \mathcal{U}$, and provided that $B(x) \in \mathcal{U}$, the map $x \to \mathcal{R}(x) := h\,(B(x))$ is differentiable and satisfies

$$\nabla\mathcal{R}(x) = \mathbb{E}_a\, \nabla_x \Psi(x; a).$$

Furthermore, h is continuously differentiable on $\mathcal{U}$ and its derivative ∇h is α-pseudo-Lipschitz, that is, there is a constant $L(h) > 0$, so that for all $B, \hat{B} \in \mathcal{U}$, it holds:

$$\left\|\nabla h(B) - \nabla h(\hat{B})\right\| \leqslant L\left\|B - \hat{B}\right\|\left(1 + \|B\|^\alpha + \left\|\hat{B}\right\|^\alpha\right). \tag{3.11}$$

We note that the commutation of expectation and gradient holds trivially on $\mathcal{U} = \mathbb{R}^{3\times 3}$ once Ψ is continuously differentiable. Moreover, the choice of the Frobenius norm in (3.11) is essentially arbitrary, and it can be replaced by the operator norm due to equivalence of norms on $\mathbb{R}^{3\times 3}$.

Our arguments will rely on the evolution of the second moment of the gradient $\mathbb{E}_a\left[\nabla_{r_1} f\,(r)^2\right]$, where we set $r := \frac{1}{\sqrt{d}}\begin{pmatrix}\psi(x)\\ \beta^*\end{pmatrix}^\top a$. Since this quantity depends on x only through the Gaussian $\frac{1}{\sqrt{d}}W(x)^\top a$, analogous to the risk, there exists a function $I\colon\, \mathbb{R}^{3\times 3} \to \mathbb{R}$

satisfying $\mathbb{E}_a[\nabla_{r_1} f(r)^2] = I\left(B(x)\right)$. In particular, for the two formulations (3.2) and (3.3), the function I is simply

$$I(B) = 2h\left(B\right).$$

Assumption 3.1.11 (Pseudo-Lipschitzness of Square Gradients)**.** The function I is α-pseudo-Lipschitz with constant $L(I) > 0$, that is, for all $B, \hat{B} \in \mathcal{U}$,

$$\left|I(B) - I(\hat{B})\right| \leqslant L\left\|B - \hat{B}\right\| (1 + \|B\|^{\alpha} + \left\|\hat{B}\right\|^{\alpha}).$$

Remark 3.1.12. All the typical losses satisfy the requisite assumptions, including logistic regression and the mean square error (see Appendix B.5). We also note that Assumptions 3.1.10 and 3.1.11 are nearly satisfied for L-smooth objectives f, and a version of the main theorem holds under just this assumption (albeit with a weaker conclusion) (see [[21], Appendix B]).

3.1.3 Algorithm Formulation

We consider the widely used *streaming stochastic gradient descent* (SGD) algorithm. At each iteration k, the algorithm generates a fresh data point $a_{k+1} \sim \mathcal{N}(0, K)$ and updates the iterates $x_k \in \mathbb{R}^{2d}$ using a stepsize γ_k according to the recurrence

$$x_{k+1} = x_k - \gamma_k \nabla_x \Psi(x_k; a_{k+1}), \tag{3.12}$$

where ∇_x is the usual gradient operator with respect to the x variable. We impose the following assumption on the stepsize, which stipulates that γ_k is a bounded deterministic function of k/d.

Assumption 3.1.13 (Stepsize)**.** There exists a constant $\bar{\gamma} < \infty$ and a deterministic scalar function $\gamma\colon [0, \infty) \to [0, \infty)$ which is bounded by $\bar{\gamma}$ and satisfies the equation $\gamma_k = \gamma\left(\frac{k}{d}\right)$.

We always work in a formulation where the entries of the iterates $\{x_k\}$ remain bounded, independent of dimension. Within the class of high-dimensional representations, we note that the random variable $\frac{1}{\sqrt{d}}\left(\psi(x)\right)^{\top} a$ should not carry dimension dependence, as otherwise the outer function f (which can very well be non-linear) degenerates to its behavior at infinity.

The functions h and I will allow us to construct closed, deterministic dynamics that describe the high-dimensional limit of stochastic gradient descent. To condense the notation,

we use

$$r_k := \frac{1}{\sqrt{d}} \begin{pmatrix} \psi(x_k) \\ \beta^* \end{pmatrix}^\top a_{k+1} \in \mathbb{R}^2,$$

so that the SGD update (3.12) simplifies as follows:

$$x_{k+1} = x_k - \frac{\gamma_k}{\sqrt{d}} \nabla_{r_1} f(r_k) \left(\nabla \psi(x_k)\right)^\top a_{k+1}. \tag{3.13}$$

Note that the stepsize γ_k is scaled in a way that SGD will behave well across different dimensions; without the factor of $\sqrt{d}$ in (3.13), the algorithm would degenerate to pure noise or to gradient flow as dimension increases. For any fixed dimension, note however that the stepsize γ_k can be arbitrarily large (as long as it is bounded uniformly across dimensions).

3.2 High-dimensional Diffusion Approximation for SGD

We begin by introducing our stochastic differential equation (SDE), called *homogenized SGD*. Specifically, we approximate SGD by the diffusion $(\mathcal{X}_t)_{t \geqslant 0}$ solving the SDE:

$$\mathrm{d}\mathcal{X}_t = -\gamma(t) d \nabla \mathcal{R}(\mathcal{X}_t) \mathrm{d}t + \gamma(t) \sqrt{I\left(B(\mathcal{X}_t)\right)} \left(\nabla \psi(\mathcal{X}_t)\right)^\top \sqrt{K} \mathrm{d}\mathfrak{B}_t, \tag{3.14}$$

with initial condition $\mathcal{X}_0 = x_0$ and $\mathrm{d}\mathfrak{B}_t$ the differential of a standard Brownian motion in $\mathbb{R}^d$. In the diffusion coefficient in (3.14), the quantity $I(B(x))$ captures the effective scalar magnitude of the stochastic gradient noise, while $(\nabla \psi(x))^\top \sqrt{K}$ describes how this noise propagates through the nonlinear parametrization and the data covariance. This structure is not imposed by covariance matching; it arises from the high-dimensional concentration argument used to identify the limiting drift and diffusion.

Remark 3.2.1 (Non-diagonal covariance)**.** The homogenized SDE (3.14) admits a natural extension to general covariance matrices K in the mean-squared error setting introduced in Section B.5.1, where the diffusion is determined by the full covariance of the stochastic gradient. The key point is that in the diagonal K setting, the last two terms arising in the conditional covariance decomposition of the Hessian contribution are effectively negligible for the class of statistics considered in this work, whereas for non-diagonal covariance matrices these terms appear to remain non-negligible in the high-dimensional limit, contributing an

additional rank-one fluctuation term to the covariance structure of the noise. Indeed, an application of Isserlis/Wick formula yields

$$\mathbb{E}_a\Big[\langle\psi(\mathcal{X}_t)-\beta^\star,a\rangle^2 aa^\top\Big] = ((\psi(\mathcal{X}_t)-\beta^\star)^\top K(\psi(\mathcal{X}_t)-\beta^\star))K + (K(\psi(\mathcal{X}_t)-\beta^\star))(K(\psi(\mathcal{X}_t)-\beta^\star))^\top.$$

This leads formally to the SDE

$$\begin{aligned}\mathrm{d}\mathcal{X}_t =& -\gamma(t)\, d\,\nabla\mathcal{R}(\mathcal{X}_t)\,\mathrm{d}t \\ &+\gamma(t)\left(I\left(B(\mathcal{X}_t)\right)\left(\nabla\psi(\mathcal{X}_t)\right)^\top K\nabla\psi(\mathcal{X}_t) + d\,\nabla\mathcal{R}(\mathcal{X}_t)\left(\nabla\mathcal{R}(\mathcal{X}_t)\right)^\top\right)^{1/2}\mathrm{d}\mathfrak{B}_t,\end{aligned} \tag{3.15}$$

where in this case $\mathrm{d}\mathfrak{B}_t$ denotes the differential of a standard Brownian motion in $\mathbb{R}^{2d}$.

The main obstruction to extending our deterministic equivalent result to general covariance K is therefore not the SDE approximation itself, but rather the closure of the deterministic equations for the statistics. When K is non-diagonal, the covariance couples the coordinates through non-commuting resolvent terms, and the finite-dimensional PDE closure used in this work no longer applies. Nevertheless, our numerical experiments suggest that the same high-dimensional concentration phenomenon persists beyond the diagonal setting; see Figure 3.3.

The SDE (3.14) is useful because it describes the evolution of many quantities of interest by direct Itô calculus. Concretely, given any sufficiently regular statistic φ, Itô's formula gives a decomposition:

$$\mathrm{d}\varphi(\mathcal{X}_t) = \underbrace{\mathcal{L}\varphi(t,\mathcal{X}_t)}_{\text{drift}}\mathrm{d}t + \underbrace{\nabla\varphi(\mathcal{X}_t)^\top\sigma(t,\mathcal{X}_t)\,\mathrm{d}\mathfrak{B}_t}_{\text{Brownian / martingale term}}, \tag{3.16}$$

where $\mathcal{L}$ is the drift–diffusion operator (first-order drift contribution plus second-order Itô correction) and σ is the diffusion matrix coefficient of the SDE.

3.2.1 Relation to Prior SDE and Homogenization Approximations

A key conceptual distinction between our framework and classical diffusion (weak-approximation) approaches lies in the regime under which the continuous-time limit becomes accurate. Traditional diffusion approximations [23, 50, 54, 60] operate at fixed

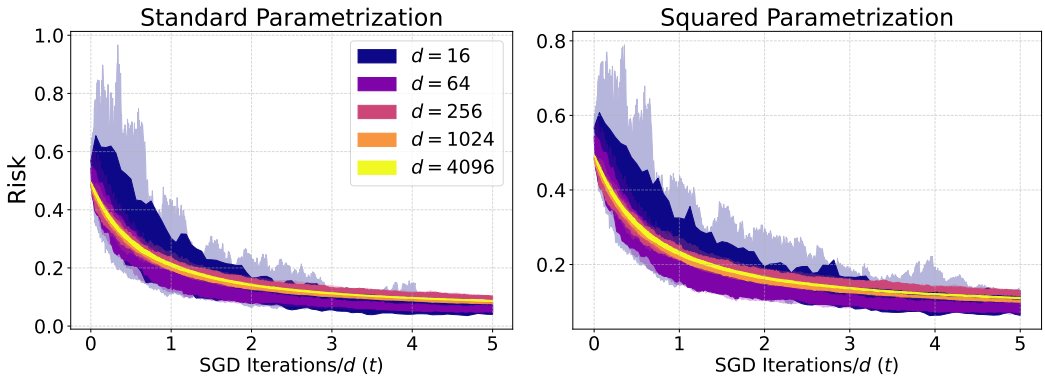

Figure 3.3: **Risk concentration of SGD and the homogenized SDE under non-diagonal covariance on a diagonal linear network.** As the dimension d increases, the risk trajectories of SGD (opaque) concentrate around the prediction of the non-diagonal homogenized SDE (3.15) (transparent), suggesting that the same high-dimensional concentration phenomenon persists beyond the diagonal covariance setting. The covariance matrix K is sampled from a Marchenko–Pastur ensemble. See Appendix B.6 for simulation details.

dimension with accuracy controlled by the small-stepsize limit $\gamma \to 0$ (after a time rescaling). In that setting, the drift dominates and the stochastic fluctuations vanish, yielding weak convergence of the discrete algorithm to its ODE/SDE formulation.

Our aim is not to provide a black-box homogenization theorem for arbitrary stochastic systems, but to obtain a closed finite-stepsize, high-dimensional description for a nonlinear overparameterized model. The diagonal linear network is a natural minimal target: it is simple enough to admit an exact resolvent closure, but already exhibits the interaction between nonconvex parametrization, stochastic gradients, and finite stepsize effects. Existing small-stepsize diffusion approximations and more general homogenization viewpoints do not directly yield a closed deterministic description for the nonlinear feature map $\psi(x)$ considered here; the closure instead requires tracking a multi-resolvent statistic and leads to the PDE system (B.6).

In the diagonal linear network setting defined in (3.3), [75] follows this diffusion viewpoint and studies SGD via stochastic gradient flow (SGF), a perturbed-gradient-flow SDE whose

diffusion term is calibrated so that an Euler discretization matches the covariance of the SGD noise:

$$\mathrm{d}\mathcal{X}_t = -d\nabla\mathcal{R}(\mathcal{X}_t)\mathrm{d}t + \sqrt{\gamma(t)I\left(B(\mathcal{X}_t)\right)}\,\left(\nabla\psi(\mathcal{X}_t)\right)^\top\sqrt{K}\,\mathrm{d}\mathfrak{B}_t. \tag{3.17}$$

The two SDEs are closely related in form, but they correspond to different limiting regimes. Homogenized SGD (3.14) has drift and noise both proportional to $\gamma(t)$, whereas the covariance-matched diffusion (3.17) has drift independent of $\gamma(t)$ and noise of order $\sqrt{\gamma(t)}$. Equivalently, a deterministic time change in (3.14) produces the algebraic scaling of (3.17). However, this does not make the two approximations equivalent for the discrete-time dynamics studied here: (3.17) is justified in the small-stepsize regime $\gamma \to 0$ at fixed dimension, while (3.14) is justified in the high-dimensional regime $d \to \infty$ at fixed stepsize. Thus the two SDEs describe different approximation limits of SGD. Our approach departs from covariance matching: our diffusion coefficient is derived from a high-dimensional limit theory [21], where approximation accuracy improves as $d \to \infty$ while the stepsize γ remains fixed. Here the effective dynamics of the algorithm are controlled by high-dimensional concentration rather than vanishing stepsizes. The resulting SDE model provides an increasingly precise description of the discrete-time risk curves as d grows. Empirically, we find that it tracks SGD more closely than the covariance-matched diffusion model at large stepsizes (see Figure 3.4), as our theory predicts. Shrinking γ drives the high-dimensional dynamics toward (3.17), whereas increasing d strengthens the approximation without changing the limiting dynamics.

3.3 High-dimensional Concentration of SGD and its Diffusion Approximation

This section formalizes the main concentration phenomenon: in the high-dimensional regime, both streaming SGD and the homogenized SGD introduced in Section 3.2 concentrate around the same deterministic dynamics. The key point is that, under our structural assumptions, the risk and the quantities that drive the update can be expressed through a small collection of empirical matrix statistics. We encode this information into a 3×3 resolvent-based matrix $S(x, z)$ (indexed by z on a fixed contour Γ). Knowing the map $z \mapsto S(x, z)$ determines $B(x)$ via a contour integral, and hence determines the learning curves and all other statistics of interest.

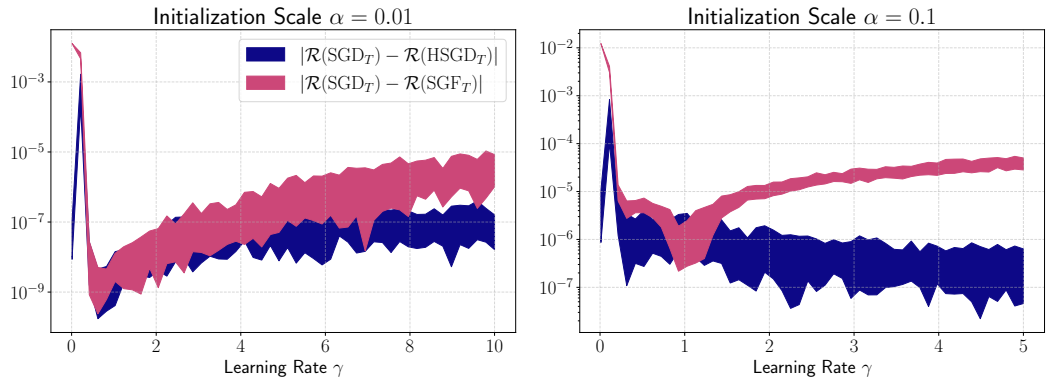

Figure 3.4: **Risk discrepancy between SGD and its continuous-time approximations on a diagonal linear network.** For each stepsize γ, we report the absolute difference between the empirical risk of SGD after $T \cdot d$ iterations (with $T = 20$) and two approximations: (i) homogenized SGD (HSGD) (3.14) (blue), and (ii) stochastic gradient flow (SGF) (3.17) (pink). As γ increases, HSGD is a more accurate approximation of SGD, whereas SGF degrades. Initialization scale α controls proximity to the saddle point $x = 0$: smaller α corresponds to a longer transient before learning accelerates. See Appendix B.6 for simulation details.

Let $z = (z_1, z_2, z_3, z_4) \in \mathbb{C}^4$ be such that

$$z_1 \notin \sigma(\mathrm{diag}(u)), \qquad z_2 \notin \sigma(\mathrm{diag}(v)), \qquad z_3 \notin \sigma(\mathrm{diag}(\beta^*)), \qquad z_4 \notin \sigma(K).$$

Equivalently, all resolvents appearing below are well-defined. In Remark 3.3.3, we specify a product domain on which this condition holds. Define the matrix

$$\Omega(x, z) := R\big(z_1; \mathrm{diag}(u)\big) R\big(z_2; \mathrm{diag}(v)\big) R\big(z_3; \mathrm{diag}(\beta^*)\big) R\big(z_4; K\big) \in \mathbb{C}^{d \times d}, \tag{3.18}$$

where $R(z; A) := (z I_d - A)^{-1}$ denotes the resolvent of A. We then set

$$S(x, z) := \frac{1}{d} W(x)^\top \Omega(x, z) W(x) \in \mathbb{C}^{3 \times 3}, \tag{3.19}$$

where $W(x)$ is the stacked matrix defined in (3.9). In particular, we evaluate these quantities either along SGD, writing $S(x_k, z)$ and $B(x_k)$, or along the diffusion, writing $S(\mathcal{X}_t, z)$ and $B(\mathcal{X}_t)$.

To control the resolvents uniformly over time (and to justify the fixed contour), we work on the high-probability event that the iterates remain uniformly bounded on the time interval of interest. This non-explosiveness condition is standard in high-dimensional/diffusion-limit arguments and is consistent with our numerical experiments, where we do not observe any such explosions. Moreover, we show in Theorem 3.3.9 that this holds automatically for all stepsizes γ below a small (explicit) numerical constant.

Assumption 3.3.1 (Non-explosiveness)**.** There exists $M > 0$ independent of d such that the initialization satisfies $\|x_0\|_\infty \leqslant M$, and such that, for every fixed $T > 0$, with overwhelming probability:

$$\sup_{0\leqslant k\leqslant \lfloor Td\rfloor} \|x_k\|_\infty \leqslant M \qquad \text{and} \qquad \sup_{0\leqslant t\leqslant T} \|\mathfrak{X}_t\|_\infty \leqslant M.$$

Remark 3.3.2. Assumption 3.3.1 is stated in a general form. In Appendix B.4, we verify that it holds in the isotropic squared-parametrization setting $\beta^* = \mathbb{1}_d$, $K = I_d$, and constant stepsize γ. More precisely, for sufficiently small γ, we establish a high-probability exponential decay of the risk. We then show that risk integrability yields uniform-in-time bounds on the coordinates of the solution and prevents them from approaching the origin. By the standard continuation criterion for SDEs, these bounds rule out finite-time explosion, thereby verifying Assumption 3.3.1 in this setting.

Remark 3.3.3. Under Assumption 3.3.1, the spectra of $\mathrm{diag}(u)$ and $\mathrm{diag}(v)$ remain inside the circle $C_{2M}(0) := \{z \in \mathbb{C} : \ |z| = 2M\}$ over the time intervals considered. Throughout the paper, we fix the product contour $\Gamma = \Gamma_1 \times \Gamma_2 \times \Gamma_3 \times \Gamma_4 \subset \mathbb{C}^4$ given by $\Gamma_1 = \Gamma_2 = C_{2M}(0)$ and

$$\Gamma_3 = \{z_3 : |z_3| = \max\{1, 2\|\beta^*\|_\infty\}\}, \qquad \Gamma_4 = \{z_4 : |z_4| = \max\{1, 2\|K\|_{\mathrm{op}}\}\}.$$

The contours are chosen in such a way that each Γ_i encloses the relevant spectrum with a fixed positive margin. Thus, by the Cauchy integral formula, B can be recovered from S via $B(x) = \frac{1}{(2\pi)^4} \oint_\Gamma z_4\, S(x, z)\, \mathrm{d}z$, and hence

$$B(x_k) = \frac{1}{(2\pi)^4} \oint_\Gamma z_4\, S(x_k, z)\, \mathrm{d}z, \quad \text{and} \quad B(\mathfrak{X}_t) = \frac{1}{(2\pi)^4} \oint_\Gamma z_4\, S(\mathfrak{X}_t, z)\, \mathrm{d}z. \tag{3.20}$$

In order to derive deterministic dynamics, we pass to a rescaled continuous-time parameter:

$$k \text{ iterations of SGD } = \lfloor td \rfloor, \quad \text{where } t \in \mathbb{R} \text{ is the continuous time parameter.}$$

This time change is natural, since when the size of the problem grows, more SGD iterations are needed to solve the underlying problem and make equivalent progress. This scaling law ensures all training dynamics live on the same space and are comparable across problem sizes.

Deterministic Limit Dynamics (PDE). Next, for each $z \in \Gamma \subset \mathbb{C}^4$, let $\mathscr{S}(t,z) \in \mathbb{C}^{3\times 3}$ denote the (local) solution of the partial integro-differential equation

$$\partial_t \mathscr{S}(t,\cdot) = \mathcal{F}\big(\cdot, \mathscr{S}(t,\cdot)\big), \qquad \mathscr{S}(0,z) = S(x_0,z), \tag{3.21}$$

where the function $\mathcal{F}$ is defined in (B.5) and involves contour integrals around Γ and derivatives in z of $\mathscr{S}$. We define the associated deterministic quantities

$$\mathscr{B}(t) := \frac{1}{(2\pi)^4} \oint_\Gamma z_4\, \mathscr{S}(t,z)\, \mathrm{d}z, \qquad \mathscr{R}(t) := h\left(\mathscr{B}(t)\right), \qquad \mathscr{I}(t) := I\left(\mathscr{B}(t)\right). \tag{3.22}$$

Since (3.21) is a fully nonlinear, nonlocal (integro-differential) parabolic PDE, we do not pursue global well-posedness. Throughout, $\mathscr{S}(t,z)$ denotes any solution defined up to $\mathscr{B}(t)$'s first exit time from the domain of validity of the coefficients $\mathcal{U}$ (or blow-up), whenever such a solution exists. In practice, we solve (3.21) numerically using standard discretization methods, and these computed solutions form the basis of the numerical simulations presented in this paper.

The next theorem states the concentration result: both sequences $B(x_{\lfloor td \rfloor})$ (SGD) and $B(\mathfrak{X}_t)$ (homogenized SGD) track the same deterministic curve $\mathscr{B}(t)$ while the dynamics remain in $\mathcal{U}$.

Theorem 3.3.4 (Learning Curves)**.** *Suppose that Assumptions 3.1.4, 3.1.5, 3.1.8, 3.1.9, 3.1.10, 3.1.11, 3.1.13, and 3.3.1 hold. Let $\mathscr{S}(t,z)$ solve (3.21) and define $\mathscr{B}(t)$ by (3.22).*

(i) Streaming SGD. *Let ϑ be the first time that either $\mathscr{B}(t)$ or $B(x_{\lfloor td \rfloor})$ exit $\mathcal{U}$. Then there exists an $\varepsilon > 0$ such that for any $T > 0$, with overwhelming probability, it holds:*

$$\sup_{0 \leqslant t \leqslant T \wedge \vartheta} \left\| \mathscr{B}(t) - B(x_{\lfloor td \rfloor}) \right\| \leqslant d^{-\varepsilon}. \tag{3.23}$$

(ii) Homogenized SDE. *For any $\eta > 0$, define the set $\mathcal{U}_\eta := \left\{ B \in \mathcal{U} : \inf_{\widehat{B} \notin \mathcal{U}} \|B - \widehat{B}\| \geqslant \eta \right\}$, and let ϑ be the first time that either $\mathscr{B}(t)$ or $B(\mathfrak{X}_t)$ exit $\mathcal{U}_\eta$. Then there exists $\varepsilon > 0$ such that for any $T > 0$, with overwhelming probability, it holds:*

$$\sup_{0 \leqslant t \leqslant T \wedge \vartheta} \|\mathscr{B}(t) - B(\mathfrak{X}_t)\| \leqslant d^{-\varepsilon}. \tag{3.24}$$

Remark 3.3.5. Theorem 3.3.4 extends the concentration framework in [21] to a larger class of risks, covering not only models such as multi-class logistic regression, but also mean-squared error over diagonal linear networks. The main additional difficulty is the presence of the inner nonlinear function ψ, through which the data enter as $\langle a, \psi(x) \rangle$ rather than as a linear inner product $\langle a, x \rangle$. This nonlinear parametrization leads, after applying Itô's formula, to products of multiple resolvents involving u, v, β^*, and K. This is precisely the step where the finite-dimensional ODE closure available in least-squares-type settings breaks down. As a result, the deterministic closure is no longer a finite-dimensional ODE as in [21], but instead takes the form of a partial integro-differential equation for the matrix-valued statistic $\mathscr{S}(t, z)$ with $z \in \mathbb{C}^4$. Thus the extension is not merely to more general test functions, but requires handling a different closure mechanism. The commutativity assumption is what allows these products to be reordered into a fixed multi-resolvent statistic; without it, the Itô expansion would generate noncommuting resolvent words that are not closed by the statistic $S(x, z)$. To retain tractability, we assume a commutativity structure, which in our setting leads to the restriction that K is diagonal rather than a general covariance matrix.

3.3.1 Other Statistics

Summarizing, we have shown that both $B(x_{\lfloor td \rfloor})$ and $B(\mathfrak{X}_t)$ concentrate around a deterministic limit. Now we pass to tracking various statistics $\varphi(x)$ along the SGD and homogenized SGD trajectories and showing that they are governed by the same deterministic dynamics in the large dimensional limit. The statistics that will satisfy this property are the ones satisfying the following assumption:

Assumption 3.3.6. The statistic $\varphi \colon \mathbb{R}^{2d} \to \mathbb{R}$ satisfies a composite structure,

$$\varphi(x) = g\left(\frac{1}{d} W^\top q_1(\mathrm{diag}(u)) q_2(\mathrm{diag}(v)) q_4(K) W \right),$$

where $g\colon \mathbb{R}^{3\times 3} \to \mathbb{R}$ is α-pseudo-Lipschitz on $\mathcal{U}$ and q_1, q_2, q_4 are polynomials.

This assumption covers all the typical statistics, including the risk, curvature, squared norms, and estimation errors that do not explicitly involve the covariance K (e.g., $\|\psi(x)\|_2^2$ and $\|\psi(x) - \beta^*\|_2^2$); ℓ^2–regularized objectives, and higher-order quantities built from derivatives of the risk. In order to describe the deterministic equivalent of φ we require its continuous-time analog:

$$\phi(t) := g\left(\frac{1}{(2\pi)^4}\oint_\Gamma q_1(z_1)q_2(z_2)q_4(z_4)\,\mathscr{S}(t,z)\,\mathrm{d}z\right). \tag{3.25}$$

The following theorem shows that $\phi(t)$–a deterministic function of a single variable–governs the evolution of φ both along the SGD and the homogenized SGD trajectories.

Theorem 3.3.7. *Suppose Assumptions 3.1.4, 3.1.5, 3.1.8, 3.1.9, 3.1.10, 3.1.11, 3.1.13, and 3.3.1 hold. Suppose further that $\mathcal{U} = \mathbb{R}^{3\times 3}$. For any function φ, which satisfies Assumption 3.3.6, and for any $T > 0$, and $\varepsilon \in (0, 1/2)$ there is a constant C (not depending on d) so that with overwhelming probability it holds:*

$$\sup_{0\leqslant t\leqslant T}\left(\left|\varphi(x_{\lfloor td\rfloor}) - \phi(t)\right| + \left|\varphi(\mathfrak{X}_t) - \phi(t)\right|\right) \leqslant Cd^{-\varepsilon}. \tag{3.26}$$

Intuition of the proof. We introduce the statistic $S(x,z)$ in (3.19) because the resolvents in $\Omega(x,z)$ allow us to represent the polynomial statistics from Assumption 3.3.6 by contour integrals, via Cauchy's integral formula. Thus, rather than tracking each statistic separately, it suffices to prove concentration for the single matrix-valued statistic S. The concentration is proved before taking contour integrals: for each time, the control holds simultaneously for all spectral parameters z on the fixed product contour from Remark 3.3.3. This uniform-in-z control is what allows the Cauchy integral representations to be applied afterward. The main step is to show that both $S(x_{\lfloor td\rfloor}, z)$ and $S(\mathfrak{X}_t, z)$ are approximate solutions of the same deterministic equation, whose solution is denoted by $\mathscr{S}(t,z)$. The commutativity assumptions ensure that the terms produced by the Doob/Itô expansions can be rewritten in terms of this same statistic S, yielding the closed PDE system (B.6). Stability of this PDE then implies that $S(x_{\lfloor td\rfloor}, z)$ and $S(\mathfrak{X}_t, z)$ both remain close to $\mathscr{S}(t,z)$. Finally, $B(x)$ follows from the contour representation in Remark 3.3.3, and any admissible statistic φ follows by applying the same contour-integral stability argument.

Remark 3.3.8. While the PDE (3.21) offers the most complete description of the high-dimensional dynamics, the SDE (3.14) is often the most effective tool in practice, both because it is cheaper to simulate numerically (e.g., via Euler–Maruyama time stepping) and because it lets us quickly identify the deterministic evolution of admissible statistics φ. Applying Itô's formula to $\varphi(\mathcal{X}_t)$ gives a drift term, including the second-order Itô correction, and a martingale term as in (3.16). The key point is that $\mathcal{L}\varphi(\mathcal{X}_t)$ can be rewritten in terms of the same class of admissible statistics (possibly after enlarging the class to achieve closure), producing a closed deterministic system. The Brownian term is not discarded at the level of the full trajectory $\mathcal{X}_t$; rather, our concentration estimates show that the resulting martingale term in the evolution of admissible statistics is negligible in the high-dimensional limit. Thus the deterministic equations for $\phi(t)$ are obtained by retaining the full drift of the Itô expansion, including the Itô correction, and controlling the martingale term at the level of statistics.

As a concrete application for our result, we show that when the stepsize is below a constant threshold, the SDE iterates remain bounded (hence Assumption 3.3.1 holds) and the risk decays exponentially fast to zero. Therefore using Theorem 3.3.7 we see that the same is true for SGD iterates themselves.

Theorem 3.3.9 (High-probability Exponential Decay)**.** *Consider homogenized SGD* (3.4) *for diagonal linear networks under the squared parametrization* (3.3)*, initialized at* $\mathcal{U}_{0,i} = \mathcal{V}_{0,i} = 1$ *for* $i = 1, \dots, d$*. Then for sufficiently small stepsize* $\gamma > 0$ *(below an explicit numerical constant) and for any* $\delta \in (0, 1)$*, there exist constants* $C, \mu > 0$ *(not dependent on* t *or* d*) such that with probability at least* $1 - \delta$*, it holds:*

$$\mathcal{R}(\mathcal{X}_t) \leqslant Ce^{-\mu t} \qquad \text{for all } t \geqslant 0. \tag{3.27}$$

In particular, Assumption 3.3.1 holds.

Our results also enable the analysis of finer statistics of the dynamics.

Potential further applications. Beyond convergence, our concentration framework can be used to analyze additional statistics of the dynamics. A promising interesting example

is $\varphi(x) = \frac{1}{d}\operatorname{Tr}(\nabla^2\mathcal{R}(x))$, which provides a tractable proxy for the *average local curvature* (and hence "sharpness") of the landscape at the current iterate. This quantity appears in [27], where flatness predicts good generalization, as well as in studies of the Sharpness-Aware Minimization (SAM) algorithm [94]. Tracking both $\mathcal{R}(x)$ and $\varphi(x)$ through their deterministic expressions gives a way to study *progressive sharpening*, a phenomenon that has recently garnered some attention [19, 44, 46, 56, 58, 99]. Empirically, one often observes transient phases where the risk temporarily increases before converging while the curvature decreases, indicating that the trajectory moves toward flatter regions of the landscape (cf. Figure 3.2). In this sense, $\frac{1}{d}\operatorname{Tr}(\nabla^2\mathcal{R})$ allows one to quantify how these phases depend on the learning rate. We see such behavior in Figure 3.2 and numerically can find the different ranges of stepsizes by using the deterministic equations.

Conclusion. We introduce a high-dimensional continuous-time description of discrete-time SGD for diagonal network models. Our framework develops an SDE, which we call homogenized SGD, that accurately tracks training dynamics at practical stepsizes. Our main technical contribution is a concentration argument that shows that both SGD and homogenized SGD converge in an appropriate sense to a deterministic PDE for a wide class of statistics, including risk and curvature. Unlike classical diffusion limits, which require vanishing stepsizes, our approximation becomes exact as the parameter dimension $d \to \infty$, even for large stepsizes.

An interesting next step, which we plan to pursue, is to leverage these deterministic dynamics to study key behaviors of diagonal linear networks, such as implicit bias and progressive sharpening. The framework we developed here provides a precise characterization of the risk and other statistics, enabling a detailed understanding of how problem-dependent quantities (e.g., average eigenvalue of the data covariance matrix) influence the effective stepsize and the convergence or divergence of SGD. In particular, using the deterministic equivalents developed here, we plan to investigate the progressive sharpening phenomenon, widely observed in practice in neural network training.

Acknowledgments and Disclosure of Funding

Research of B. García Malaxechebarría was in part funded by NSF DMS 2023166 (NSF TRIPODS II). C. Paquette was supported by a Discovery Grant from the Natural Science and Engineering Research Council (NSERC) of Canada, NSERC CREATE grant Interdisciplinary Math and Artificial Intelligence Program (INTER-MATH-AI), Google x Mila research grant, Fonds de recherche du Quebec Nature et technologies (FRQNT) NOVA Grant, and CIFAR AI Catalyst Grant. Additionally C. Paquette has a 20% part-time employment at Google DeepMind. The work of M. Fazel is supported in part by awards NSF TRIPODS II DMS 2023166, NSF CCF 2212261, NSF CCF 2312775, and the Moorthy Family Professorship at UW. She also works part-time at Amazon Inc. as an Amazon Scholar. Research of D. Drusvyatskiy was supported by NSF DMS-2306322, NSF DMS-2023166, and AFOSR FA9550-24-1-0092 awards.

Appendix A

APPENDIX FOR CHAPTER 2

Broader Impact Statement. The work presented in this paper is foundational research and it is not tied to any particular application. The set-up is on a simple well-studied high-dimensional linear composites (e.g., least squares, logistic regression, phase retrieval) with synthetic data and solved using known al gorithms, e.g., AdaGrad-Norm. We present deterministic dynamics for the training loss and adaptive stepsizes. The results are theoretical and we do not anticipate any direct ethical and societal issues. We believe the results will be used by machine learning practitioners and we encourage them to use it to build a more just, prosperous world.

A.1 SGD adaptive learning rate algorithms and stepsizes

In this section, we write down the explicit update rules for 2 different adaptive stochastic gradient descent algorithms.

Example: AdaGrad-Norm. We begin with AdaGrad-Norm (see Algorithm 1). Note by unraveling the recursion, we have that

$$\mathfrak{g}_k = \frac{\eta}{\sqrt{b^2 + \frac{1}{d^2}\sum_{j=0}^{k} \|\nabla_X \Psi(X_j; a_{j+1}, \epsilon_{j+1})\|^2}}, \tag{A.1}$$

with the deterministic equivalent (see Section 2.2 and also A.3.3) for this learning rate being

$$\gamma_t = \frac{\eta}{\sqrt{b^2 + \frac{\mathrm{Tr}(K)}{d}\int_0^t I(\mathscr{B}(s))\ \mathrm{d}s}}. \tag{A.2}$$

In the case of the least squares problem, the quantity $I(\mathscr{B}(t))$ is explicit and

$$\gamma_t = \frac{\eta}{\sqrt{b^2 + \frac{2\,\mathrm{Tr}(K)}{d}\int_0^t \mathscr{R}(s)\ \mathrm{d}s}}. \tag{A.3}$$

Algorithm 1 AdaGrad-Norm

Require: Initialize $\eta > 0$, $X_0 \in \mathbb{R}^d$, $b \in \mathbb{R}$ and set $b_0 = b \times d$

for $k = 1, 2, \ldots,$ **do**

Generate new sample $a_k \sim \mathcal{N}(0, K)$, $\epsilon_k \sim \mathcal{N}(0, \omega^2)$;

$b_k^2 \leftarrow b_{k-1}^2 + \|\nabla_X \Psi(X_{k-1}; a_k, \epsilon_k)\|^2$;

$\mathfrak{g}_{k-1} = d \times \frac{\eta}{|b_k|}$; ▷ updating learning rate

$X_k \leftarrow X_{k-1} - \frac{\mathfrak{g}_{k-1}}{d} \nabla_X \Psi(X_{k-1}; a_k, \epsilon_k)$; ▷ updating step with stochastic gradient

end for

Example: RMSprop-Norm We consider the "normed" version of RMSprop, that is, where there is only one learning rate parameter.

We consider Algorithm 2 where we put a factor of the learning into the exponential moving average for RMSprop. The deterministic equivalent for $\mathfrak{g}_k$ for Alg. 2 (see Section 2.2) is

$$\gamma_t = \frac{\eta}{\sqrt{b^2 e^{-\alpha t} + \frac{\mathrm{Tr}(K)}{d} \int_0^t e^{-\alpha(t-s)} I(\mathscr{B}(s))\, \mathrm{d}s}}. \tag{A.4}$$

In the case of the least squares problem, the quantity $I(\mathscr{B}(t))$ is explicit and

$$\gamma_t = \frac{\eta}{\sqrt{b^2 e^{-\alpha t} + \frac{2\,\mathrm{Tr}(K)}{d} \int_0^t e^{-\alpha(t-s)} \mathscr{R}(s)\, \mathrm{d}s}}. \tag{A.5}$$

A.2 The Dynamical nexus

In this section, we prove the main theorem on concentration of the risk curves and learning rates. We shall set some notation. In what follows, we again use $W = [X|X^\star] \in \mathbb{R}^{d\times 2}$. We also use $W^+ = [W|X_0] = [X|X^\star|X_0]$.

We shall also use the shorthand $r = \langle a, W\rangle$, and $x = \langle a, X\rangle$ so that $f(\langle a, X\rangle, \langle a, X^\star\rangle; \epsilon) = f(\langle a, W\rangle; \epsilon) = f(r; \epsilon)$.

We shall let $B = B(X) = W^T K W$ be the covariance matrix of the Gaussian vector r. We also write f' for the $\partial_x f$.

Algorithm 2 RMSprop-Norm, α Exponential Moving Average

Require: Initialize $\eta > 0$, $X_0 \in \mathbb{R}^d$, $b \in \mathbb{R}$ and set $b_0 = d \times b$, $\alpha > 0$ exponential moving avg.

$\mathfrak{g}_{-1} = d \times \frac{\eta}{b_0}$;

for $k = 1, 2, \ldots,$ **do**

Generate new sample $a_k \sim \mathcal{N}(0, K)$, $\epsilon_k \sim \mathcal{N}(0, \omega^2)$;

$b_k^2 \leftarrow \alpha \cdot b_{k-1}^2 + (1-\alpha)\|\nabla_X \Psi(X_{k-1}; a_k, \epsilon_k)\|^2$;

$\mathfrak{g}_{k-1} = d \times \frac{\eta}{|b_k|}$; ▷ updating learning rate

$X_k \leftarrow X_{k-1} - \frac{\mathfrak{g}_{k-1}}{d} \nabla_X \Psi(X_{k-1}; a_k, \epsilon_k)$; ▷ updating step with stochastic gradient

end for

A.2.1 Discussion of the assumptions on f

In this section we show how the assumptions we put on h and I are almost satisfied for L-smooth f. We say that f is L-smooth if:

$$\|\nabla f(r_1, \epsilon_1) - \nabla f(r_2, \epsilon_2)\| \leq L\sqrt{(\|r_1 - r_2\|^2 + \|\epsilon_1 - \epsilon_2\|^2)},$$

which we note implies f is α-pseudo Lipschitz with $\alpha = 1$.

Lemma A.2.1. *1. There exists a function $h : \mathbb{R}^{2\times 2} \to \mathbb{R}$ such that $h(B(X)) = \mathcal{R}(X)$ is differentiable and satisfies*

$$\nabla_X \mathcal{R}(X) = \mathbb{E}_{a,\epsilon} \nabla_X \Psi(X; a, \epsilon).$$

Furthermore, h is continuously differentiable on $\{B : \det B \neq 0\}$ and its derivative ∇h satisfies an estimate

$$\|\nabla h(B_1) - \nabla h(B_2)\| \leq (\sqrt{2}+1)L(f) \min\{\|B_1^{-1}\|_{op}, \|B_2^{-1}\|_{op}\}\|B_1 - B_2\|_F.$$

2. The function $I(B) = \mathbb{E}_{a,\epsilon}[(f'(\langle a, X\rangle; \langle a, X^\star\rangle, \epsilon))^2]$ satisfies an estimate

$$|I(B_1) - I(B_2)| \leq L(f)\sqrt{I(B_1) + I(B_2)} \min\{\|B_1^{-1}\|_{op}, \|B_2^{-1}\|_{op}\}\|B_1 - B_2\|_F.$$

Proof. To derive the existence of h, note that

$$\mathcal{R}(X) = \mathbb{E}(\mathbb{E}(f(\langle a, X\rangle, \langle a, X^\star\rangle, \epsilon)|\epsilon))$$

is an expectation of a Gaussian vector $r = (\langle a, X\rangle, \langle a, X^\star\rangle)$. This vector can be expressed as an image of an iid Gaussian vector z by representing $r = \sqrt{B}z$, and hence we have

$$h(B) \stackrel{\text{def}}{=} \mathbb{E}(\mathbb{E}(f(\sqrt{B}z, \epsilon)|\epsilon)).$$

As the function f is absolutely continuous with a Lipschitz gradient, we can differentiate under the integral sign and conclude

$$\nabla_X \mathcal{R}(X) = \nabla_X \mathbb{E}\, f(\langle a, X\rangle, \langle a, X^\star\rangle, \epsilon) = \mathbb{E}\, \nabla_X f(\langle a, X\rangle, \langle a, X^\star\rangle, \epsilon).$$

For the differentiability of h, suppose for the moment that f is C^2 with bounded second derivatives.[1] Setting $Q = \sqrt{B}$ the positive semi-definite square root of B, we have

$$\partial_{Q_{ij}} h(Q^2) = \mathbb{E}(\mathbb{E}(\partial_{Q_{ij}} f(Qz, \epsilon)|\epsilon)).$$

Then using the chain rule, and setting $\partial_i f$ to be the i-th partial derivative of f,

$$\partial_{Q_{ij}} h(Q^2) = \mathbb{E}(\mathbb{E}(z_j \partial_i f(Qz, \epsilon)|\epsilon)) = \mathbb{E}(\mathbb{E}([Q_{ij}\partial_i + Q_{jj}\partial_j]\partial_i f(Qz, \epsilon)|\epsilon)),$$

where we have applied Stein's Lemma. We conclude when $\det Q \neq 0$ by the implicit function theorem that h is differentiable and we have

$$\partial_{Q_{ij}} h(Q^2) = \sum \partial_{kl} h \partial_{Q_{ij}} (Q^2)_{kl} = \sum_l (\partial_{il} h) Q_{jl} + \sum_k (\partial_{kj} h) Q_{ik}.$$

As a matrix equation, this can be written as

$$(Dh)Q + Q(Dh) = JQ \quad \text{where} \quad J_{kl} = \mathbb{E}(\mathbb{E}((\partial_k \partial_l f)(Qz, \epsilon)|\epsilon)).$$

This is a linear equation in Dh. When $Q \succ 0$, we can define

$$A = \int_0^\infty e^{-tQ}(JQ)e^{-tQ}\, \mathrm{d}t,$$

[1] This condition can be removed in a standard way: one creates an f_ϵ which is an approximation to f formed by convolving with an isotropic Gaussian of variance ϵ. This is C^2 and has bounded second derivatives (as f was smooth). One then takes the limit as $\epsilon \to 0$.

and note

$$AQ + QA = -\int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t}\left(e^{-tQ}(JQ)e^{-tQ}\right)\mathrm{d}t = JQ.$$

Moreover, the mapping $M \mapsto \int_0^\infty e^{-tQ}Me^{-tQ}\,\mathrm{d}t$ defines a two-sided inverse for $M \mapsto MQ + QM$, and so $Dh = A$. Note that by symmetry of J, Q, and Dh

$$JQ = (Dh)Q + Q(Dh) = QJ,$$

and therefore

$$(Dh)Q + Q(Dh) = \frac{1}{2}(JQ + QJ),$$

and so taking inverses on both sides, $Dh = J$.

Undoing Stein's Lemma, we have $Q(Dh) = (Dh)Q = M$, where $M_{ij} = \mathbb{E}(\mathbb{E}(z_j\partial_i f(Qz,\epsilon)|\epsilon))$. From L-smoothness of f

$$\|M(Q_1) - M(Q_2)\| \le L\,\mathbb{E}(\|z\|\|Q_1z - Q_2z\|) \le \sqrt{2}L\|Q_1 - Q_2\|_F.$$

Hence

$$\begin{aligned}\|Dh(Q_1^2) - Dh(Q_2^2)\| &= \|Q_1^{-1}M(Q_1) - Q_2^{-1}M(Q_2)\| \\ &\le \|Q_1^{-1}\|_{op}\|M(Q_1) - Q_1Q_2^{-1}M(Q_2)\| \\ &\le \|Q_1^{-1}\|_{op}\left(\|M(Q_1) - M(Q_2)\| + \|(Q_2 - Q_1)Q_2^{-1}M(Q_2)\|\right).\end{aligned}$$

Note $Q_2^{-1}M(Q_2) = (Dh)(Q_2^2)$ is bounded by $L(f)$, and so we arrive at

$$\begin{aligned}\|Dh(Q_1^2) - Dh(Q_2^2)\| &\le (\sqrt{2}+1)L(f)\|Q_1^{-1}\|_{op}\|Q_1 - Q_2\|_F \\ &\le (\sqrt{2}+1)L(f)\|Q_1^{-2}\|_{op}\|Q_1^2 - Q_2^2\|_F.\end{aligned}$$

We note the bound is symmetric in Q_1 and Q_2, and by density of C^2 in space of $C^{1,lip}$, this holds for L-smooth f. This concludes the estimates for the derivative of h.

For the Fisher matrix, $I(B)$, from L-smoothness, we have again with $Q = \sqrt{B}$,

$$I(Q^2) = \mathbb{E}(\mathbb{E}((\partial_1 f(Qz,\epsilon))^2|\epsilon)).$$

Then

$$|I(Q_1^2) - I(Q_2^2)| \le \left|\mathbb{E}(\mathbb{E}((\partial_1 f(Q_1z,\epsilon))^2 - (\partial_1 f(Q_2z,\epsilon))^2|\epsilon))\right|.$$

Applying Cauchy-Schwarz and using the L-smoothness of f,

$$|I(Q_1^2) - I(Q_2^2)| \leq \sqrt{I(Q_1^2) + I(Q_2^2)} \times L(f)\|Q_1 - Q_2\|_F.$$

□

This lemma shows that an L-smooth function nearly satisfies Assumption 2.1.3 and 2.1.4 provided that $\|B^{-1}\|_{\text{op}}$ is bounded. Therefore, our concentration result Theorem A.2.4 and its Corollaries will hold provided we add a stopping time. Fix $M > 0$ and let

$$\hbar_M(B) \stackrel{\text{def}}{=} \inf\{t > 0 \,:\, \|B^{-1}\|_{\text{op}} > M\}.$$

Then the concentration of the risk under SGD to a deterministic function, Theorem A.2.4, holds with t replaced with $t \wedge \hbar_M(B) \wedge \hbar_M(\mathscr{B})$. The corollaries of Theorem A.2.4 also follow under this added stopping time.

In the next section, we prove this concentration theorem, Theorem A.2.4.

A.2.2 Integro-differential equation for $\mathscr{S}(t, z)$

A goal of this paper is to show that quadratic statistics $\varphi : \mathbb{R}^d \to \mathbb{R}$ applied to SGD converge to a deterministic function. This argument hinges on understanding the deterministic dynamics of one important statistic, defined as

$$S(W, z) = W^\top R(z; K) W,$$

applied to $W_{\lfloor td \rfloor}$ (SGD updates). Here $W = [X|X^\star]$ and $R(z; K) = (K - zI_d)^{-1}$ for $z \in \mathbb{C}$ is the resolvent of the matrix K. The statistic $S(W, z)$ is valuable because it encodes many other important quantities including $W^\top q(K) W$ for all polynomials q. We show that $S(W_{\lfloor td \rfloor}, z)$, is close to a deterministic function $(t, z) \mapsto \mathscr{S}(t, z)$ which satisfies an integro-differential equation.

To introduce the integro-differential equation, recall by Assumptions 2.1.3 and 2.1.4

$$\mathcal{R}(X) = h \circ B(W) \quad \text{and} \quad \mathbb{E}_{a,\epsilon}[f'(a^\top W)^2] = I \circ B(W) \quad \text{with} \quad B(W) = W^\top K W,$$

and α-pseudo-Lipschitz functions $h : \mathbb{R}^{2\times 2} \to \mathbb{R}$ differentiable and $I : \mathbb{R}^{2\times 2} \to \mathbb{R}$. It will be useful, throughout the remaining paper, to express ∇h explicitly as a 2×2 matrix, that is,

$$\nabla h \cong \left[\begin{array}{c|c} \nabla h_{11} & \nabla h_{12} \\ \hline \nabla h_{21} & \nabla h_{22} \end{array}\right].$$

With these recollections, the integro-differential equation is defined below.

Integro-Differential Equation for $\mathscr{S}(t,z)$. For any contour $\Omega \subset \mathbb{C}$ enclosing the eigenvalues of K, we have an expression for the derivative of $\mathscr{S}$:

$$\mathrm{d}\mathscr{S}(t,\cdot) = \mathcal{F}(z, \mathscr{S}(t,\cdot))\ \mathrm{d}t \tag{A.6}$$

$$\begin{aligned} \text{where } \mathcal{F}(z,\mathscr{S}(t,\cdot)) \stackrel{\text{def}}{=} & -2\gamma_t\left(\left(\frac{-1}{2\pi i}\oint_\Omega \mathscr{S}(t,z)\ \mathrm{d}z\right)H(\mathscr{B}(t)) \right. \\ & \left. + H^T(\mathscr{B}(t))\left(\frac{-1}{2\pi i}\oint_\Omega \mathscr{S}(t,z)\ \mathrm{d}z\right)\right) \\ & + \frac{\gamma_t^2}{d}\left[\begin{array}{c|c} \mathrm{Tr}(KR(z;K))I(\mathscr{B}(t)) & 0 \\ \hline 0 & 0 \end{array}\right] \\ & - \gamma_t(\mathscr{S}(t,z)(2zH(\mathscr{B}(t))) + (2zH^T(\mathscr{B}(t)))\mathscr{S}(t,z)). \end{aligned} \tag{A.7}$$

$$\text{Here } \mathscr{B}(t) = \frac{-1}{2\pi i}\oint_\Omega z\mathscr{S}(t,z)\ \mathrm{d}z, \quad H(\mathscr{B}) = \left[\begin{array}{c|c} \nabla h_{11}(\mathscr{B}) & 0 \\ \hline \nabla h_{21}(\mathscr{B}) & 0 \end{array}\right],$$

$$\gamma_t \text{ is defined in (2.9), and the initialization is } \quad \mathscr{S}(0,z) = W_0^\top R(z;K)W_0. \tag{A.8}$$

The functions $h : \mathbb{R}^{2\times 2} \to \mathbb{R}$ and $I : \mathbb{R}^{2\times 2} \to \mathbb{R}$ are defined in Assumption 2.1.3 and Assumption 2.1.4, respectively.

We first note that there is an actual solution to the integro-differential equation. This solution is the same as the ODEs defined in the introduction (see (2.9)) and proved in [21, Lemma 4.1].

Lemma A.2.2 (Equivalence to coupled ODEs.)**.** *The unique solution of* (A.7) *with initial*

condition (A.8) *is given by*

$$\mathscr{S}(t,z) = \frac{1}{d}\sum_{i=1}^{d} \frac{1}{\lambda_i - z}\mathscr{V}_i(t).$$

In this section, we will be working with approximate solutions to the integro-differential equation (A.6) (see below for specifics). For working with these solutions, we introduce some notation. We shall always work on a fixed contour Ω surrounding the spectrum of K, given by $\Omega \stackrel{\text{def}}{=} \{z : |z| = \max\{1, 2\|K\|_{\text{op}}\}\}$. We note that this contour is always distance at least $\frac{1}{2}$ from the spectrum of K. We define a norm, $\|\cdot\|_\Omega$, on a continuous function $A : \mathbb{C} \to \mathbb{R}$ as

$$\|A\|_\Omega = \max_{z\in\Omega} \|A(z)\|. \tag{A.9}$$

Definition A.2.3 ((ε, M, T)-approximate solution to the integro-differential equation)**.** For constants $M, T, \varepsilon > 0$, we call a continuous function $\mathcal{S} : [0,\infty) \times \mathbb{C} \to \mathbb{R}^{2\times 2}$ an (ε, M, T)-*approximate solution* of (A.6) if with

$$\hat{\tau}_M(\mathcal{S}) \stackrel{\text{def}}{=} \inf\left\{t \geq 0 \,:\, \|\mathcal{S}(t,\cdot)\|_\Omega > M\right\},$$

then

$$\sup_{0\leq t\leq(\hat{\tau}_M\wedge T)} \left\|\mathcal{S}(t,\cdot) - S(0,\cdot) - \int_0^t \mathcal{F}(\cdot, \mathcal{S}(s,\cdot))\, \mathrm{d}s\right\|_\Omega \leq \varepsilon$$

and $\mathcal{S}(0,\cdot) = W_0^\top R(\cdot, K)W_0$, where $W_0 = [X_0|X^\star]$ is the initialization of SGD.

We suppress the $\mathcal{S}$ in the notation for $\hat{\tau}_M$, that is, $\hat{\tau}_M = \hat{\tau}_M(\mathcal{S})$, when the function $\mathcal{S}$ is clear from the context.

We are now ready to state and prove one of our main results.

Theorem A.2.4 (Concentration of SGD and deterministic function $\mathscr{S}(t,z)$)**.** *Suppose the risk function* $\mathcal{R}(X)$ (2.2) *satisfies Assumptions 2.1.2, 2.1.3, and 2.1.4. Suppose the learning rate satisfies Assumption 2.1.6, and the initialization* X_0 *and hidden parameters* $X^\star$ *satisfy Assumption 2.1.5. Moreover the data* $a \sim \mathcal{N}(0, K)$ *and label noise* ϵ *satisfy Assumption 2.1.1. Let* $\{W_{\lfloor td\rfloor}\}$ *be generated from the iterates of SGD. Then there is an* $\varepsilon > 0$ *so that for any* $T, M > 0$ *and* d *sufficiently large, with overwhelming probability*

$$\sup_{0\leq t\leq T\wedge\hat{\tau}_M(S(W,\cdot))\wedge\hat{\tau}_M(\mathscr{S})} \|S(W_{\lfloor td\rfloor},\cdot) - \mathscr{S}(t,\cdot)\|_\Omega \leq d^{-\varepsilon}, \tag{A.10}$$

where the deterministic function $\mathscr{S}(t,z)$ *solves the integro-differential equation* (A.6)*.*

Proof. By Proposition A.3.4, for any M and T, we can find a $\tilde{\varepsilon} > 0$ such that the function $S(W_{td}, z)$ is an $(d^{-\tilde{\varepsilon}}, M, T)$-approximate solution. (For the deterministic function $\mathscr{S}$, it is an $(0, M, T)$-approximate solution by definition.) We now apply the stability result, [21, Prop. 4.1], to conclude that there exists a $\varepsilon > 0$ such that

$$\sup_{0 \le t \le T \wedge \hat{\tau}_M} \|\mathscr{S}(t, z) - S(W_{td}, z)\|_\Omega \le d^{-\varepsilon}, \quad w.o.p, \tag{A.11}$$

where $\hat{\tau}_M$ is shorthand for $\hat{\tau}_M(S(W, \cdot)) \wedge \hat{\tau}_M(\mathscr{S})$. The result immediately follows. □

Corollary A.2.5. *Suppose the assumptions of Theorem A.2.4 hold. Let f be an α-pseudo-Lipschitz function with $\alpha \le 1$ and let q be a polynomial. Set*

$$\varphi(X) \overset{def}{=} f(W^T q(K) W), \quad \phi(t) \overset{def}{=} f\left(\frac{-1}{2\pi i} \oint_\Omega q(z) \mathscr{S}(t, z)\, \mathrm{d}z\right), \quad \text{where } \mathscr{S}(t, z) \text{ solves (A.6).}$$

Then there is an $\varepsilon > 0$ such that for d sufficiently large, with overwhelming probability,

$$\sup_{0 \le t \le T} |\varphi(X_{td}) - \phi(t)| \le d^{-\varepsilon}.$$

Proof. This is basically equivalent to [21, Corollary 4.2]. The only difference is that [21, Corollary 4.2] requires the boundedness of $\mathscr{N}$; however, since our function f is α-pseudo-Lipschitz with $\alpha \le 1$, this boundedness follows from [21, Proposition 1.2], and the rest of the proof is identical to the one in [21]. □

Remark A.2.6. *The learning rate $\mathfrak{g}_k$, technically, is not a function of $W^T q(K) W$. However, Assumption 2.1.6 ensures that the learning rate concentrates around a function $W^T q(K) W$. Therefore, Corollary A.2.5 applies to the learning rate.*

A.3 SGD-AL is an approximate solution

We introduce a rescaling of time to relate the k-th iteration of SGD to the continuous time parameter t in the differential equation through the relationship $k = \lfloor td \rfloor$. Thus, when $t = 1$, SGD has done exactly d updates. Since the parameter t is continuous and the iteration counter k (integer) discrete, to simplify the discussion below, we *extend* k to continuous values through the floor operation, $X_k \overset{\text{def}}{=} X_{\lfloor k \rfloor}$. Using the continuous parameter t, the iterates are related by $X_{td} = X_{\lfloor td \rfloor}$.

The paper [21] provides a net argument showing that we do not need to work with every z on the contour Ω defining the integro-differential equation, but only polynomially many in d. Recall that $\Omega = \{z : |z| = \max\{2\|K\|_{\text{op}}, 1\}\}$. For a fixed $\xi > 0$, we say that Ω_ξ is a $d^{-\xi}$-*mesh of* Ω if $\Omega_\xi \subset \Omega$ and for every $z \in \Omega$ there exists a $\bar{z} \in \Omega_\xi$ such that $|z - \bar{z}| < d^{-\xi}$. We can achieve this with Ω_ξ having cardinality, $|\Omega_\xi| = C(|\Omega|)d^\xi$.

Lemma A.3.1 (Net argument, [21], Lemma 5.1)**.** *Fix $T, M > 0$ and let $\xi > 0$. Suppose Ω_ξ is a $d^{-\xi}$ mesh of Ω with $|\Omega_\xi| = C \cdot d^\xi$ and positive $C > 0$. Let the function $S(t, z) = S(W_{td}, z)$ satisfy*

$$\sup_{0 \le t \le (\hat{\tau}_M \wedge T)} \|S(t, \cdot) - S(0, \cdot) - \int_0^t \mathcal{F}(\cdot, S(s, \cdot))\, \mathrm{d}s\|_{\Omega_\xi} \le \varepsilon \tag{A.12}$$

with $\hat{\tau}_M = \inf\{t \ge 0 : \|S(t, \cdot)\|_\Omega > M\}$. Then S is a $(\varepsilon + C(M, T, \|K\|_{op})d^{-\xi}, M, T)$-approximate solution to the integro-differential equation, that is,

$$\sup_{0 \le t \le (\hat{\tau}_M \wedge T)} \|S(t, \cdot) - S(0, \cdot) - \int_0^t \mathcal{F}(\cdot, S(s, \cdot))\, \mathrm{d}s\|_{\Omega} \le \varepsilon + C \cdot d^{-\xi},$$

where $C = C(M, T, \|K\|_{op}, L(I), L(h))$ is a positive constant.

(We prove in Section A.3.1 that $S(t, z)$ does indeed satisfy inequality (A.12).) We also cite the following lemma, which relates two stopping times used throughout this paper.

Lemma A.3.2 (Stopping time, [21], Lemma 4.2)**.** *For a constant C depending on $\|K\|_{op}$, we have*

$$C \le \frac{\|S(W_{td}, \cdot)\|_\Omega}{\|W_{td}\|^2} \le 2.$$

Remark A.3.3. *Fix $M > 0$ and define the stopping time on $\|W_{td}\|$, $\vartheta = \vartheta_M$, by*

$$\vartheta_M(W_{td}) \stackrel{def}{=} \inf\left\{t \ge 0 : \|W_{td}\|^2 > M\right\}.$$

Due to the previous lemma, any stopping time $\hat{\tau}_M$ defined on $\|S(t, \cdot)\|_\Omega$ corresponds to a stopping time ϑ on $\|W_{td}\|$, that is, for $c = C^{-1}$, $\hat{\tau}_M \le \vartheta_{cM}$.

A.3.1 SGD-AL is an approximated solution

Proposition A.3.4 (SGD-AL is an approximate solution)**.** *Fix a $T, M > 0$ and $0 < \varepsilon < \delta/8$, where δ is defined in Assumption 2.1.6. Then $S(W_{td}, z)$ is a $(d^{-\varepsilon}, M, T)$-approximate solution*

w.o.p., that is,

$$\sup_{0\le t\le(T\wedge\tau_M)} \|S(W_{td},z) - S(W_0,z) - \int_0^t \mathcal{F}(z,S(W_{sd},z))\,\mathrm{d}s\|_\Omega \le d^{-\varepsilon}. \tag{A.13}$$

Again, the proof is very similar to [21, Prop. 5.2]. The one difference is that the martingales and error terms are slightly more involved, because of the non-deterministic stepsize we are using. The remainder of this section, along with section A.3.2, fills in the details of bounding these lower-order terms, so that the proof can proceed as in [21].

Shorthand notation

In the following sections, we will be using various versions of the stepsize γ. In order to simplify notation, we set

$$\begin{aligned}
\gamma(G_k) &= \gamma(k, N_k(d\times\cdot), G_k(d\times\cdot), Q_k(d\times\cdot)),\\
\gamma(\mathcal{G}_k) &= \gamma(k, N_k(d\times\cdot), \mathcal{G}_k(d\times\cdot), Q_k(d\times\cdot)),\\
\gamma(B_k) &= \gamma(k, N_k(d\times\cdot), \mathrm{Tr}(K)I(B_k(d\times\cdot))/d, Q_k(d\times\cdot)).
\end{aligned}$$

Further, setting $\Delta_k \overset{\text{def}}{=} f'(r_k)a_{k+1}$, define

$$I_1(k) \overset{\text{def}}{=} \Delta_k^\top \nabla^2\varphi(X_k)\Delta_k/d,\ I_2(k) \overset{\text{def}}{=} \mathrm{Tr}(\nabla^2\varphi(X_k)K)\,\mathbb{E}[f'(r_k)^2\,|\,\mathcal{F}_k]/d,\ I_3(k) \overset{\text{def}}{=} \nabla\varphi(X_k)^\top\Delta_k.$$

The normalization here (dividing by d) is chosen so that the I terms are all $O(1)$; this is formally shown in Lemma A.3.13.

SGD-AL under the statistic

We follow the approach in [21, Section 5.3] to rewrite the SGD adaptive learning rate update rule as an integral equation. Considering a quadratic function $\varphi:\mathbb{R}^d\to\mathbb{R}$ and performing Taylor expansion, we obtain

$$\varphi(X_{k+1}) = \varphi(X_k) - \frac{\gamma(G_k)}{d}\nabla\varphi(X_k)^\top\Delta_k + \frac{\gamma(G_k)^2}{2d^2}\Delta_k^\top\nabla^2\varphi(X_k)\Delta_k. \tag{A.14}$$

We will now relate this equation to its expectation by performing a Doob decomposition, involving the following martingale increments and error terms:

$$\Delta\mathcal{M}_k^{\mathrm{grad}}(\varphi) \stackrel{\mathrm{def}}{=} \frac{1}{d}\left(-\gamma(G_k)I_3(k) + \mathbb{E}\left[\gamma(G_k)I_3(k)\,\middle|\,\mathcal{F}_k\right]\right), \tag{A.15}$$

$$\Delta\mathcal{M}_k^{\mathrm{Hess}}(\varphi) \stackrel{\mathrm{def}}{=} \frac{1}{2d}\left(\gamma(G_k)^2 I_1(k) - \mathbb{E}\left[\gamma(G_k)^2 I_1(k)\,\middle|\,\mathcal{F}_k\right]\right), \tag{A.16}$$

$$\mathbb{E}[\mathcal{E}_k^{\mathrm{Hess}}(\varphi)\,|\,\mathcal{F}_k] \stackrel{\mathrm{def}}{=} \frac{1}{2d}\left(\mathbb{E}\left[\gamma(G_k)^2 I_1(k)\,\middle|\,\mathcal{F}_k\right] - \gamma(B_k)^2 I_2(k)\right), \tag{A.17}$$

$$\mathbb{E}[\mathcal{E}_k^{\mathrm{grad}}(\varphi)\,|\,\mathcal{F}_k] \stackrel{\mathrm{def}}{=} \frac{1}{d}\left(-\mathbb{E}\left[\gamma(G_k)I_3(k)\,\middle|\,\mathcal{F}_k\right] + \gamma(B_k)\nabla\varphi(X_k)^\top\nabla\mathcal{R}(X_k)\right). \tag{A.18}$$

We can then write

$$\begin{aligned}\varphi(X_{k+1}) = \varphi(X_k) - \frac{\gamma(B_k)}{d}\nabla\varphi(X_k)^\top\nabla\mathcal{R}(X_k) + \frac{\gamma(B_k)^2}{2d^2}\operatorname{Tr}(\nabla^2\varphi(X_k)K)\,\mathbb{E}[f'(r_k)^2\,|\,\mathcal{F}_k] \\ + \Delta\mathcal{M}_k^{\mathrm{grad}}(\varphi) + \Delta\mathcal{M}_k^{\mathrm{Hess}}(\varphi) + \mathbb{E}[\mathcal{E}_k^{\mathrm{Hess}}(\varphi)\,|\,\mathcal{F}_k] + \mathbb{E}[\mathcal{E}_k^{\mathrm{grad}}(\varphi)\,|\,\mathcal{F}_k].\end{aligned}$$

Extending X_k into continuous time by defining $X_t = X_{\lfloor t\rfloor}$, we sum up (integrate). For this, we introduce the forward difference

$$(\Delta\varphi)(X_j) \stackrel{\mathrm{def}}{=} \varphi(X_{j+1}) - \varphi(X_j),$$

giving us

$$\varphi(X_{td}) = \varphi(X_0) + \sum_{j=0}^{\lfloor td\rfloor-1}(\Delta\varphi)(X_j) \stackrel{\mathrm{def}}{=} \varphi(X_0) + \int_0^t d\cdot(\Delta\varphi)(X_{sd})\,\mathrm{d}s + \xi_{td},$$

where $|\xi_{td}| = \left|\int_{(\lfloor td\rfloor-1)/d}^t d\cdot\Delta\varphi(X_{sd})\,\mathrm{d}s\right| \le \max_{0\le j\le\lceil td\rceil}\{|\Delta\varphi(X_j)|\}$. With this, we obtain the Doob decomposition for SGD-AL:

$$\begin{aligned}\varphi(X_{td}) = \varphi(X_0) &- \int_0^t \gamma(B_{sd})\nabla\varphi(X_{sd})^\top\nabla\mathcal{R}(X_{sd})\,\mathrm{d}s \\ &+ \frac{1}{2d}\int_0^t \gamma(B_{sd})^2\operatorname{Tr}(K\nabla^2\varphi(X_{sd}))\,\mathbb{E}[f'(r_{sd})^2\,|\,\mathcal{F}_{sd}]\,\mathrm{d}s \\ &+ \sum_{j=0}^{\lfloor td\rfloor-1}\mathcal{E}_j^{\mathrm{all}}(\varphi),\end{aligned} \tag{A.19}$$

$$\begin{aligned}\text{with}\quad \mathcal{E}_j^{\mathrm{all}}(\varphi) = \Delta\mathcal{M}_j^{\mathrm{grad}}(\varphi) &+ \Delta\mathcal{M}_j^{\mathrm{Hess}}(\varphi) \\ &+ \mathbb{E}[\mathcal{E}_j^{\mathrm{Hess}}(\varphi)\,|\,\mathcal{F}_j] + \mathbb{E}[\mathcal{E}_j^{\mathrm{grad}}(\varphi)\,|\,\mathcal{F}_j] \\ &+ \xi_{td}(\varphi).\end{aligned} \tag{A.20}$$

From here, we can proceed as in [21, Section 5.3] to show that SGD-AL is an (ε, M, T)-approximated solution.

$S(W_{td}, z)$ is an approximate solution

Proof of Proposition A.3.4. The appropriate stepsize, as a function of W_{td}, is

$$\gamma_t = \gamma(td, N_{td}, \mathrm{Tr}(K)I(B_{td})/d, Q_{td}).$$

(Note that N, I and Q can all be found as functions of $S(W_{td}, \cdot)$ using contour integration.) It is shown in the proof of [21, Proposition 5.2] that given the analogue of (A.19) for deterministic stepsize, $S(W_{td}, \cdot)$ satisfies

$$S\left(W_{td}, z\right) = S\left(W_0, z\right) + \int_0^t \mathcal{F}\left(z, S\left(W_{sd}, z\right)\right) \mathrm{d}s + \sum_{i=0}^{\lfloor td \rfloor - 1} \mathcal{E}_j^{\mathrm{all}}(S).$$

The only terms of (A.19) that differ in our case are the martingale and error terms. Thus to show that $S(W_{td}, \cdot)$ is an approximate solution of the integro-differential equation (A.6) all we need is to bound the martingales and error terms contained in $\mathcal{E}_j^{\mathrm{all}}$. Let $\Omega = \{z : |z| = \max\{1, 2\|K\|_{\mathrm{op}}\}\}$, as previously. We thus have that for all $z \in \Omega$,

$$\sup_{0 \le t \le T \wedge \hat{\tau}_M} \left| S(W_{td}, z) - S(W_0, z) - \int_0^t \mathcal{F}(z, S(W_{sd}, z)) \, \mathrm{d}s \right| \le \sup_{0 \le t \le T \wedge \hat{\tau}_M} \|\mathcal{E}_{td}^{\mathrm{all}}(S(\cdot, z))\|. \tag{A.21}$$

Next, fix a constant $\xi > 0$. Let $\Omega_\xi \subset \Omega$ such that there exists a $\bar{z} \in \Omega_\xi$ such that $|z - \bar{z}| \le d^{-\xi}$ and the cardinality of Ω_ξ, $|\Omega_\xi| = Cd^\xi$ where $C > 0$ can depend on $\|K\|_{\mathrm{op}}$. For all $z \in \Omega$, we note that $\hat{\tau}_M \le \vartheta_{cM}$ (see Lemma A.3.2). Consequently, we evaluate the error with the stopped process $W_{td}^{\vartheta} \stackrel{\mathrm{def}}{=} W_{d(t \wedge \vartheta)}$ instead of using $\hat{\tau}_M$. By Proposition A.3.5, the proof of which we have deferred to Section A.3.2, we have, for any $\hat{\delta} > 0$

$$\sup_{z \in \Omega_\xi} \sup_{0 \le t \le T \wedge \vartheta_{cM}} \|\mathcal{E}_{dt}^{\mathrm{all}}(S(\cdot, z))\| \le d^{-\delta/4 + \hat{\delta}} \quad \text{w.o.p.} \tag{A.22}$$

We deduce that

$$\sup_{0 \le t \le T \wedge \hat{\tau}_M} \|S(W_{td}, z) - S(W_0, z) - \int_0^t \mathcal{F}(z, S(W_{sd}, z)) \, \mathrm{d}s\|_{\Omega_\xi} \le d^{\hat{\delta} - \delta/4} \quad \text{w.o.p.}$$

An application of the net argument, Lemma A.3.1, finishes the proof after setting $\hat{\delta} = \delta/8$ and $\xi = \delta/8$. □

A.3.2 Error bounds

All the martingale and error terms (A.20) go to 0 as d grows. Formally,

Proposition A.3.5. *Let the function f be defined as in Assumption 2.1.2. Let the statistic $S : [0, \infty) \times \mathbb{C} \to \mathbb{R}^{2\times 2}$ be defined as*

$$S(t, z) = W_{\lfloor td \rfloor}^{\top} R(z; K) W_{\lfloor td \rfloor}, \tag{A.23}$$

where $W = [X|X^\star]$. Then, for any $z \in \Omega$ and $T, M, \zeta > 0$, with overwhelming probability,

$$\sup_{0\le t\le T\wedge\vartheta} \left\|\mathcal{E}_{dt}^{all}(S(\cdot, z))\right\| \le d^{-\delta/4+\zeta},$$

where to suppress notation we use ϑ as shorthand for ϑ_{cM}, and c is the constant from Lemma A.3.2.

Proof. This follows from combining Propositions A.3.6, A.3.7, A.3.8, A.3.9, and A.3.10. □

The remainder of this subsection is devoted to proving these supporting propositions; throughout these proofs we will work with the stopping time ϑ as defined in the proposition above.

Bounds on the lower order terms in the gradient and hessian

Proposition A.3.6 (Hessian error term)**.** *Let f and S be defined as in Assumption 2.1.2 and (A.23). Then, for any $z \in \Omega$, $T > 0$ and $\zeta > 0$, with overwhelming probability,*

$$\sup_{0\le t\le T\wedge\vartheta} \sum_{k=0}^{\lfloor td \rfloor - 1} \left\|\mathbb{E}\left[\mathcal{E}_k^{Hess}(S(\cdot, z)) \mid \mathcal{F}_k\right]\right\| \le d^{-\delta/4+\zeta}.$$

Proof. For arbitrary $z \in \Omega$ and $k \le (T \wedge \vartheta)d - 1$, set $\varphi(X) = S_{ij}(W, z)$ to be the ij-th entry of the matrix $S(W, z)$. Then

$$\begin{aligned}
2d\,\mathbb{E}[\mathcal{E}_k^{\mathrm{Hess}}(\varphi) \,|\, \mathcal{F}_k] &= \mathbb{E}\left[\gamma(G_k)^2 I_1(k) \,|\, \mathcal{F}_k\right] - \gamma(B_k)^2 I_2(k) \\
&= \mathbb{E}[(\gamma(G_k)^2 - \gamma(\mathcal{G}_k)^2) I_1(k) \,|\, \mathcal{F}_k] \\
&\quad + (\gamma(\mathcal{G}_k)^2 - \gamma(B_k)^2)\,\mathbb{E}[I_1(k) \,|\, \mathcal{F}_k] + \gamma(B_k)^2\,\mathbb{E}[(I_1(k) - I_2(k) \,|\, \mathcal{F}_k] \\
&= \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3.
\end{aligned}$$

We look at $|\mathcal{E}_1|$ first.

$$\begin{aligned}
|\mathcal{E}_1| &= \left|\mathbb{E}\left[(\gamma(G_k)^2 - \gamma(\mathcal{G}_k)^2) I_1(k) \,|\, \mathcal{F}_k\right]\right| \\
&\leq \mathbb{E}\left[\left|(\gamma(G_k)^2 - \gamma(\mathcal{G}_k)^2)\right|^2 \,|\, \mathcal{F}_k\right]^{\frac{1}{2}} \cdot \mathbb{E}\left[\left|I_1(k)\right|^2 \mathcal{F}_k\right|\right]^{\frac{1}{2}} \\
&\leq \mathbb{E}\left[|\gamma(G_k) + \gamma(\mathcal{G}_k)|^{\frac{7}{2}} |\gamma(G_k) - \gamma(\mathcal{G}_k)|^{\frac{1}{2}} \,|\, \mathcal{F}_k\right]^{\frac{1}{2}} \cdot \mathbb{E}\left[\left|I_1(k)\right|^2 \mathcal{F}_k\right|\right]^{\frac{1}{2}} \\
&\leq \mathbb{E}\left[|\gamma(G_k) + \gamma(\mathcal{G}_k)|^7 \,|\, \mathcal{F}_k\right]^{\frac{1}{4}} \cdot \mathbb{E}\left[|\gamma(G_k) - \gamma(\mathcal{G}_k)| \,|\, \mathcal{F}_k\right]^{\frac{1}{4}} \cdot \mathbb{E}\left[|I_1(k)|^2 \,|\, \mathcal{F}_k\right]^{\frac{1}{2}}.
\end{aligned}$$

For the first term, we use (2.6). We have

$$\mathbb{E}\left[|\gamma(G_k) + \gamma(\mathcal{G}_k)|^7 \,|\, \mathcal{F}_k\right] \leq \hat{C}(\gamma) \cdot \mathbb{E}\left[|2 + 2\|N_k\|_\infty^\alpha + 2\|Q_k\|_\infty^\alpha + \|G_k\|_\infty^\alpha + \|\mathcal{G}_k\|_\infty^\alpha|^7 \,|\, \mathcal{F}_k\right].$$

All the terms inside the expectation, apart from $\|G_k\|_\infty^\alpha$, are deterministic with respect to $\mathcal{F}_k$ and bounded by a constant independent of d (see Lemma A.3.14). Since we know from Lemma A.3.14 that for any $\varepsilon > 0$, all moments of $\|G_k\|_\infty$ are bounded by d^ε w.o.p., we conclude

$$\mathbb{E}\left[|\gamma(G_k) + \gamma(\mathcal{G}_k)|^7 \,|\, \mathcal{F}_k\right] \leq d^\varepsilon \quad \text{w.o.p.}$$

For the second term, we use (2.5). Again, since $\|N_k\|_\infty$ and $\|Q_k\|_\infty$ are bounded due to our stopping time, we have

$$\mathbb{E}\left[|\gamma(G_k) - \gamma(\mathcal{G}_k)| \,|\, \mathcal{F}_k\right]^{\frac{1}{4}} \leq d^{-\delta/4}.$$

The last term, $\mathbb{E}\left[|I_1(k)|^2 \,|\, \mathcal{F}_k\right]^{\frac{1}{2}}$, is also bounded by a constant (see Lemma A.3.13), and all together, we find that $|\mathcal{E}_1| \leq d^{\varepsilon - \delta/4}$ with overwhelming probability.

Now let us consider $|\mathcal{E}_2|$:

$$|\mathcal{E}_2| = |(\gamma(\mathcal{G}_k)^2 - \gamma(B_k)^2)\,\mathbb{E}[I_1(k) \,|\, \mathcal{F}_k]| = |\gamma(\mathcal{G}_k) + \gamma(B_k)| \cdot |\gamma(\mathcal{G}_k) - \gamma(B_k)| \cdot |\,\mathbb{E}[I_1(k) \,|\, \mathcal{F}_k]|.$$

The first term is bounded by (2.6), since $\mathcal{G}_k$ and $\mathrm{Tr}(K)I(B_k)/d$ are bounded independent of d; the second term is bounded Cd^{-1} by Lemma A.3.17, and the last term is bounded by a constant by Lemma A.3.13.

Finally, consider $|\mathcal{E}_3|$:

$$|\mathcal{E}_3| = \gamma(B_k)^2 \cdot |\mathbb{E}[(I_1(k) - I_2(k) \,|\, \mathcal{F}_k]|\,.$$

By (2.6), the first term is bounded by $\hat{C}(\gamma)^2(1 + \|N_k\|_\infty^\alpha + \|Q_k\|_\infty^\alpha + \|\operatorname{Tr}(K)I(B_k)/d\|_\infty^\alpha)^2$. All of these terms are bounded by a constant independent of d (because of the stopping time.) The second term satisfies the assumptions of Lemma A.3.16 with $H = \nabla^2\varphi(X_k)$, and is thus bounded by Cd^{-1}. All together,

$$2d\,\mathbb{E}[\mathcal{E}_k^{\mathrm{Hess}}(\varphi)\,|\,\mathcal{F}_k] \le d^{-\delta/4+\varepsilon}.$$

Summing up to $k = Td$ and dividing through by $2d$, we obtain the desired bound. □

Proposition A.3.7 (Gradient error term)**.** *Let f and S be defined as in Assumption 2.1.2 and (A.23). Then, for any $z \in \Omega$, $\zeta > 0$ and $T > 0$, with overwhelming probability,*

$$\sup_{0\le t\le T\wedge\vartheta} \sum_{k=0}^{\lfloor td\rfloor-1} \left\|\mathbb{E}\left[\mathcal{E}_k^{grad}(S(\cdot,z))\mid\mathcal{F}_k\right]\right\| \le d^{-\delta/4+\zeta}.$$

Proof. We have

$$\begin{aligned} d\,\mathbb{E}[\mathcal{E}_k^{\mathrm{grad}}\,|\,\mathcal{F}_k] &= -\,\mathbb{E}[\gamma(G_k)\langle\nabla\varphi(X_k),\Delta_k\rangle\,|\,\mathcal{F}_k] + \gamma(B_k)\langle\nabla\varphi(X_k),\nabla R(X_k)\rangle \\ &= -\,\mathbb{E}[(\gamma(G_k)-\gamma(\mathcal{G}_k))I_3(k)\,|\,\mathcal{F}_k] - (\gamma(\mathcal{G}_k)-\gamma(B_k)\,\mathbb{E}[I_3(k)\,|\,\mathcal{F}_k] \\ &= \mathcal{E}_1 + \mathcal{E}_2. \end{aligned}$$

We then have

$$\begin{aligned} |\mathcal{E}_1| &\le \mathbb{E}\left[|\gamma(G_k)-\gamma(\mathcal{G}_k)|^2\mid\mathcal{F}_k\right]^{\frac12}\cdot\mathbb{E}\left[|I_3(k)|^2\mid\mathcal{F}_k\right]^{\frac12} \\ &\le \mathbb{E}\left[|\gamma(G_k)+\gamma(\mathcal{G}_k)|^3\mid\mathcal{F}_k\right]^{\frac14}\cdot\mathbb{E}\left[|\gamma(G_k)-\gamma(\mathcal{G}_k)|\mid\mathcal{F}_k\right]^{\frac14}\cdot\mathbb{E}\left[|I_3(k)|^2\mid\mathcal{F}_k\right]^{\frac12}. \end{aligned}$$

Just as in the Hessian argument, (2.6) lets us bound $\mathbb{E}\left[|\gamma(G_k)+\gamma(\mathcal{G}_k)|^3\mid\mathcal{F}_k\right]^{\frac14}$ by d^{ε} w.o.p., (2.5) lets us bound $\mathbb{E}\left[|\gamma(G_k)-\gamma(\mathcal{G}_k)|\mid\mathcal{F}_k\right]^{\frac14}$ by $d^{-\delta/4}$ w.o.p., and Lemma A.3.13 lets us bound $\mathbb{E}\left[|I_3(k)|^2\mid\mathcal{F}_k\right]^{\frac12}$ by a constant, giving an overall bound of $|\mathcal{E}_1| \le d^{-\delta/4+\varepsilon}$.

By the same argument as in the Hessian case, $|\mathcal{E}_2|$ is bounded by Cd^{-1}; in conclusion,

$$d\,\mathbb{E}[\mathcal{E}_k^{\mathrm{grad}}\,|\,\mathcal{F}_k] \le d^{\varepsilon-\delta/4}.$$

Summing and dividing through by d, we obtain the desired result with $\zeta = \varepsilon$. □

Proposition A.3.8 (Gradient martingale)**.** *Let f and S be defined as in Assumption 2.1.2 and (A.23). Then, for any $z \in \Omega$, $\zeta > 0$ and $T > 0$, with overwhelming probability,*

$$\sup_{0 \le t \le T \wedge \vartheta} \left\| \mathcal{M}^{\mathrm{grad}}_{\lfloor dt \rfloor}(S(\cdot, z)) \right\| \le d^{-1/2+\zeta}.$$

Proof. For notational convenience, set $\Delta\mathcal{M}_k = \Delta\mathcal{M}^{\mathrm{grad}}_{d(k/d \wedge \vartheta)}$, and $F_k = -\gamma(G_k) I_3(k)/d$, so that

$$\Delta\mathcal{M}_k = F_k - \mathbb{E}[F_k \,|\, \mathcal{F}_k].$$

Set $F_k^\beta = \mathrm{Proj}_\beta(F_k)$, that is, ensuring F_k stays in $[-\beta, \beta]$. Then $F_k^\beta - \mathbb{E}[F_k^\beta \,|\, \mathcal{F}_k]$ is in $[-2\beta, 2\beta]$, and so for the martingale $\mathcal{M}_k^\beta$ with increments $\Delta\mathcal{M}_k^\beta = F_k^\beta - \mathbb{E}[F_k^\beta \,|\, \mathcal{F}_k]$, Azuma's inequality tells us that

$$\mathbb{P}\left(|\mathcal{M}_k^\beta| \ge t \right) \le 2 \exp\left(\frac{-t^2}{2 \sum_{i=0}^k (2\beta)^2} \right) \le 2 \exp\left(\frac{-t^2}{2Td(2\beta)^2} \right).$$

Set $\beta = d^{-1+\zeta/2}$ and $t = d^{-1/2+\zeta}$; this becomes

$$\mathbb{P}\left(|\mathcal{M}_k^\beta| \ge d^{-1/2+\zeta} \right) \le 2 \exp\left(\frac{-d^\zeta}{8T} \right).$$

However, $\mathcal{M}_k^\beta$ is not quite the martingale we started with: there is still an error term,

$$\begin{aligned} |\mathcal{M}_k - \mathcal{M}_k^\beta| &= \left| \sum_{i=0}^k (F_k - \mathbb{E}[F_k \,|\, \mathcal{F}_k]) - (F_k^\beta - \mathbb{E}[F_k^\beta \,|\, \mathcal{F}_k]) \right| \\ &\le \sum_{i=0}^k \left| F_k - F_k^\beta \right| + \left| \mathbb{E}[F_k - F_k^\beta \,|\, \mathcal{F}_k] \right|. \end{aligned}$$

We bound this term in overwhelming probability. We have

$$\begin{aligned} \mathbb{P}\left(F_k - F_k^\beta \ne 0 \right) &= \mathbb{P}\left(|F_k| > \beta \right) \\ &= \mathbb{P}\left(|\gamma(G_k) I_3(k)/d| > d^{-1+\zeta/2} \right) \\ &\le \mathbb{P}\left(\gamma(G_k) \ge d^{\zeta/4} \right) + \mathbb{P}\left(|I_3(k)| \ge d^{\zeta/4} \right). \end{aligned}$$

The second term is superpolynomially small by Lemma A.3.13; the first term is superpolynomially small by (2.6) and (A.3.14).

$$\begin{aligned}\left|\mathbb{E}[F_k - F_k^\beta \,|\, \mathcal{F}_k]\right| &= \left|\mathbb{E}[(F_k - F_k^\beta)\mathbf{1}_{\{|F_k|>\beta\}} \,|\, \mathcal{F}_k]\right| \\ &\le \mathbb{E}[(F_k - F_k^\beta)^2 \,|\, \mathcal{F}_k]^{\frac{1}{2}} \cdot \mathbb{E}[\mathbf{1}^2_{\{|F_k|>\beta\}} \,|\, \mathcal{F}_k]^{\frac{1}{2}} \\ &\le 4\,\mathbb{E}[F_k^2 \,|\, \mathcal{F}_k]^{\frac{1}{2}} \cdot \mathbb{E}[\mathbf{1}_{\{|F_k|>\beta\}} \,|\, \mathcal{F}_k]^{\frac{1}{2}} \\ &\le 4d^{-1}\,\mathbb{E}[\gamma(G_k)^4 \,|\, \mathcal{F}_k]^{\frac{1}{4}} \cdot \mathbb{E}[I_3(k)^4 \,|\, \mathcal{F}_k]^{\frac{1}{4}} \cdot \mathbb{E}[\mathbf{1}_{\{|F_k|>\beta\}} \,|\, \mathcal{F}_k]^{\frac{1}{2}}.\end{aligned}$$

As before, the first and second expectations are bounded by constants, and the last expectation is just the probability that $|F_k| > \beta$, which we have already shown is superpolynomially small. So with overwhelming probability, we have

$$|\mathcal{M}_k - \mathcal{M}_k^\beta| = \left|\sum_{i=0}^{k}(F_k - \mathbb{E}[F_k \,|\, \mathcal{F}_k]) - (F_k^\beta - \mathbb{E}[F_k^\beta \,|\, \mathcal{F}_k])\right| \le d^{-1/2+\zeta}$$

(any power of d would have worked). Combining the error term and the projected martingale, we find that, with overwhelming probability,

$$|\mathcal{M}_k| \le d^{-1/2+\zeta}.$$

We can now take the maximum over k from 0 to Td using a union bound; this does not affect the overwhelming probability statement. □

Proposition A.3.9 (Hessian martingale)**.** *Let f and S be defined as in Assumption 2.1.2 and (A.23). Then, for any $z \in \Omega$, $\zeta > 0$ and $T > 0$, with overwhelming probability,*

$$\sup_{0\le t\le T\wedge\vartheta} \left\|\mathcal{M}^{\mathrm{Hess}}_{\lfloor td\rfloor}(S(\cdot, z))\right\| \le d^{-1/2+\zeta}.$$

Proof. The proof here is basically identical to the previous one. Again, set $F_k = \gamma(G_k)^2 I_1(k)/d$ and $F_k^\beta = \mathrm{Proj}_\beta(F_k)$, with their associated martingales being $\mathcal{M}_k = F_k - \mathbb{E}[F_k \,|\, \mathcal{F}_k]$ and $\mathcal{M}_k^\beta = F_k^\beta - \mathbb{E}[F_k^\beta \,|\, \mathcal{F}_k]$. As before, Azuma's inequality, with $\beta = d^{-1+\zeta/2}$, gives us

$$\mathbb{P}(\mathcal{M}_k^\beta \ge d^{-1/2+\zeta}) \le 2\exp\left(-\frac{d^\zeta}{8T}\right).$$

The error term is also quite similar:

$$|\mathcal{M}_k - \mathcal{M}_k^\beta| \le \sum_{i=0}^{k} |F_k - F_k^\beta| + |\,\mathbb{E}[F_k - F_k^\beta \,|\, \mathcal{F}_k]|.$$

We have

$$\mathbb{P}(F_k - F_k^\beta \neq 0) \le \mathbb{P}(\gamma(G_k)^2 \le d^{\zeta/4}) + \mathbb{P}(|I_2(k)| \le d^{\zeta/4}),$$

both of which are superpolynomially small by (2.6) and Lemma A.3.13. For the expectation, we have

$$|\,\mathbb{E}[F_k - F_k^\beta \,|\, \mathcal{F}_k]| \le 4d^{-1}\,\mathbb{E}[\gamma(G_k)^8 \,|\, \mathcal{F}_k]^{\frac{1}{4}} \cdot \mathbb{E}[I_1(k)^4 \,|\, \mathcal{F}_k]^{\frac{1}{4}} \cdot \mathbb{E}[\mathbf{1}_{\{|F_k|>\beta\}} \,|\, \mathcal{F}_k]^{\frac{1}{2}};$$

this product is superpolynomially small by (2.6), Lemma A.3.14, and Lemma A.3.13. Overall, we have, with overwhelming probability,

$$|\mathcal{M}_k| \le d^{-1/2+\zeta}.$$

Taking the supremum, we obtain the desired result. □

Proposition A.3.10 (Integral error term)**.** *Let f and S be defined as in Assumption 2.1.2 and (A.23). Then, for $z \in \Omega$,*

$$|\xi_{td}(S(\cdot, z))| \le d^{-1/2}.$$

Proof. We have, as above,

$$\begin{aligned} |\xi_{td}| &= \left| \int_{(\lfloor td \rfloor - 1)/d}^{t} d \cdot \Delta\varphi(X_{sd})\, \mathrm{d}s \right| \\ &\le \max_{0 \le j \le \lceil td \rceil} \{|\Delta\varphi(X_j)|\}, \end{aligned}$$

which is bounded by $d^{-1/2}$ w.o.p. by the boundedness of I_1, I_2, I_3, and $\gamma(B_k)$. □

General bounds

In this section, we make use of the subgaussian norm $\|\cdot\|_{\psi_2}$ of a random variable (see [87] for details.) When it exists, this norm is defined as

$$\|X\|_{\psi_2} \asymp \inf\left\{V > 0 : \forall t > 0,\ \mathbb{P}(|X| > t) \le 2e^{-t^2/V^2}\right\}. \tag{A.24}$$

In particular, Gaussian random variables have a well-defined subgaussian norm.

Lemma A.3.11 ([21], Lemma 5.3)**.** *There exist constants $c, C > 0$ such that*

$$c\|W\|^2 \le \|S(W,z)\|_\Omega \le C\|W\|^2, \quad \|\nabla_X S(W,z)\|_\Omega \le C\|W\|, \quad \text{and} \quad \|\nabla_X^2 S(W,z)\|_\Omega \le C.$$

Lemma A.3.12 (Preliminary bounds)**.** *With f and Δ_k defined as above, for $\varepsilon > 0$ and $\lambda \ge 0$, we have*

$$f'(r_k) \le d^\varepsilon \quad \textit{w.o.p. and} \quad \mathbb{E}[|f'(r_k)|^\lambda \,|\, \mathcal{F}_k] \le C(\lambda), \tag{A.25}$$

$$\frac{\|\Delta_k\|^2}{d} \le d^\varepsilon \quad \textit{w.o.p. and} \quad \mathbb{E}\left[\left(\frac{\|\Delta_k\|^2}{d}\right)^\lambda \,|\, \mathcal{F}_k\right] \le C(\lambda). \tag{A.26}$$

Proof of (A.25) *in Lemma A.3.12.* By [21, Lemma 3.4], if function f is α-pseudo-Lipschitz with Lipschitz constant $L(f)$ (as in (2.1.2)) and the noise ϵ is independent of a, then

$$\big|f'(r)\big| \le C(\alpha)(L(f))(1 + |r| + |\epsilon|)^{\max\{1,\alpha\}}.$$

Then

$$\begin{aligned}\big|f'(r_k)\big| &\le C(\alpha)(L(f))(1 + |r_k| + |\epsilon|)^{\max\{1,\alpha\}} \\ &\le C(\alpha)(L(f))(1 + |X_k^\top a_{k+1}| + |\epsilon|)^{\max\{1,\alpha\}}. \end{aligned} \tag{A.27}$$

Now, since a_{k+1} is Gaussian, we can write $a_{k+1} = \sqrt{K}v_k$, for a standard normal v_k. Then we see that $X_k^\top a_{k+1} = X_k^\top \sqrt{K} v_k$ is a single-variable Gaussian, with variance $|X_k^\top K X_k| \le \|X_k\|^2 \cdot \|K\|_{\text{op}}$ (bounded independently of d because of the stopping time on X_k). Similarly, ϵ is Gaussian and independent of a_{k+1}, so the expression (A.27) is bounded w.o.p. by d^ε, and

$$\mathbb{E}\left[\left(C(\alpha)(L(f))(1 + |X_k^\top a_{k+1}| + |\epsilon|)^{\max\{1,\alpha\}}\right)^\lambda \,\Big|\, \mathcal{F}_k\right] \le C(\lambda)$$

for some constant $C(\lambda)$. □

Proof of (A.26) *in Lemma A.3.12.* We can write $a_{k+1} = \sqrt{K}v_k$, where v_k is a standard d-dimensional normal vector. Then, by Hanson-Wright, we have

$$\begin{aligned}\mathbb{P}\left(\big|\|a_{k+1}\|^2 - \mathbb{E}[\|a_{k+1}\|^2 \,|\, \mathcal{F}_k]\big| \ge d\right) &= \mathbb{P}\left(\left|v_k^\top K v_k - \mathbb{E}[v_k^\top K v_k \,|\, \mathcal{F}_k]\right| \ge d\right) \\ &\le 2\exp\left(-\frac{cd^2}{\|K\|_F^2 + \|K\|_{\text{op}} d}\right) \\ &\le 2\exp\left(-\frac{cd^2}{d(\|K\|_{\text{op}} + \|K\|_{\text{op}}^2}\right) \\ &\le 2\exp\left(-Cd\right).\end{aligned}$$

Now, note that $\mathbb{E}[v_k^\top K v_k \,|\, \mathcal{F}_k] = \mathrm{Tr}(K) \le d\|K\|_{\mathrm{op}}$. Together, we get that $\|a_{k+1}\|^2 \le d^{1+\epsilon}$ with overwhelming probability. Then

$$\frac{\|\Delta_k\|^2}{d} = \frac{\|f'(r_k)a_{k+1}\|^2}{d} = \frac{\|a_{k+1}\|^2 f'(r_k)^2}{d},$$

which is bounded by $d^{2\varepsilon}$ w.o.p. Now for the expectation:

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\|\Delta_k\|^2}{d}\right)^\lambda \,|\, \mathcal{F}_k\right] &\le \mathbb{E}\left[\left(\frac{\|\sqrt{K}v_k\|^2}{d}\right)^{2\lambda} \,|\, \mathcal{F}_k\right]^{\frac{1}{2}} \cdot \mathbb{E}\left[f'(r_k)^{4\lambda} \,|\, \mathcal{F}_k\right]^{\frac{1}{2}} \\
&\le \mathbb{E}\left[\left(\frac{\|K\|_{\mathrm{op}} \cdot \|v_k\|^2}{d}\right)^{2\lambda} \,|\, \mathcal{F}_k\right]^{\frac{1}{2}} \cdot \mathbb{E}\left[f'(r_k)^{4\lambda} \,|\, \mathcal{F}_k\right]^{\frac{1}{2}} \qquad \text{(A.28)}
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\|K\|_{\mathrm{op}} \cdot \|v_k\|^2}{d}\right)^{2\lambda} \,|\, \mathcal{F}_k\right] &= \|K\|_{\mathrm{op}}^{2\lambda} \cdot \mathbb{E}\left[\left(\frac{\|v_k\|^2}{d}\right)^{2\lambda} \,|\, \mathcal{F}_k\right] \\
&\le \|K\|_{\mathrm{op}}^{2\lambda} \cdot \frac{1}{d}\sum_{i=0}^{d-1} \mathbb{E}\left[\left(\|v_k^i\|^2\right)^{2\lambda} \,|\, \mathcal{F}_k\right] \qquad \text{(Jensen's inequality)} \\
&= \|K\|_{\mathrm{op}}^{2\lambda} \cdot \mathbb{E}\left[\|v_k^0\|^{4\lambda} \,|\, \mathcal{F}_k\right], \qquad \text{(i.i.d. assumption)}
\end{aligned}$$

where we are using the notation v_k^i to refer to the ith component of the vector v_k. Now, since v_k^0 is just a standard Gaussian, all of its moments are bounded. The second term in (A.28) is bounded by a constant by (A.25), as desired. □

Lemma A.3.13 (Gradient and Hessian bounds)**.** *Setting*

$$I_1(k) \overset{def}{=} \Delta_k^\top \nabla^2\varphi(X_k)\Delta_k/d, \quad I_2(k) \overset{def}{=} \mathrm{Tr}(\nabla^2\varphi(X_k)K)\,\mathbb{E}[f'(r_k)^2 \,|\, \mathcal{F}_k]/d,$$
$$I_3(k) \overset{def}{=} \nabla\varphi(X_k)^\top \Delta_k,$$

for any $\varepsilon > 0$ and $\lambda \ge 0$, we have

$$|I_1(k)| \le d^\varepsilon \quad \textit{w.o.p. and} \quad \mathbb{E}\left[|I_1(k)|^\lambda \,|\, \mathcal{F}_k\right] \le C(\lambda), \qquad \text{(A.29)}$$

$$|I_2(k)| \le C, \qquad \text{(A.30)}$$

$$|I_3(k)| \le d^\varepsilon \quad \textit{w.o.p. and} \quad \mathbb{E}\left[|I_3(k)|^\lambda \,|\, \mathcal{F}_k\right] \le C(\lambda). \qquad \text{(A.31)}$$

Proof of (A.29) *in Lemma A.3.13.* Using the fact that $\|\nabla^2\varphi(X_k)\|_{\text{op}} \le \|S(W_k,\cdot)\|_\Omega$,

$$\begin{aligned}\frac{|\Delta_k^\top \nabla^2\varphi(X_k)\Delta_k|}{d} &\le \frac{\|S(W_k,\cdot)\|_\Omega \|\Delta_k\|^2}{d} \\ &\le \frac{C\|W_k\|^2\|\Delta_k\|^2}{d}. \qquad \text{(Lemma A.3.11)}\end{aligned}$$

Now, $\|W_k\|$ is bounded by the stopping time. From Lemma A.3.12, $\frac{\|\Delta_k\|^2}{d}$ is bounded by d^ε w.o.p., and every moment of this expression is bounded independent of d, as desired. □

Proof of (A.30) *in Lemma A.3.13.* We have

$$\begin{aligned}\frac{\left|\operatorname{Tr}(\nabla^2\varphi(X_k)K)\,\mathbb{E}[f'(r_k)^2 \,|\, \mathcal{F}_k]\right|}{d} &\le \frac{d\|\nabla^2\varphi(X_k)K\|_{\text{op}} \cdot \mathbb{E}[f'(r_k)^2 \,|\, \mathcal{F}_k]}{d} \\ &\le \|\nabla^2\varphi(X_k)\|_{\text{op}} \cdot \|K\|_{\text{op}} \cdot \mathbb{E}[f'(r_k)^2 \,|\, \mathcal{F}_k] \\ &\le CM^2\,\mathbb{E}[f'(r_k)^2 \,|\, \mathcal{F}_k]. \qquad \text{(Lemma A.3.11)}\end{aligned}$$

From Lemma A.3.12, $\mathbb{E}[f'(r_k)^2 \,|\, \mathcal{F}_k]$ is bounded by a constant independent of d, as desired. □

Proof of (A.31) *in Lemma A.3.13.* We have

$$|\nabla\varphi(X_k)^\top \Delta_k| \le |\nabla\varphi(X_k)^\top a_{k+1}| \cdot |f'(r_k)|.$$

By Lemma A.3.11, $\|\nabla\varphi(X_k)\| \le C\|W_k\| \le CM$ (since we are working under a stopping time), and so $\nabla\varphi(X_k)^\top a_{k+1}$ is subgaussian (and thus bounded by d^ε w.o.p.). By (A.25), $f'(r_k)$ is bounded by d^ε w.o.p., and so their product is bounded by $d^{2\varepsilon}$ w.o.p., as desired. Now for the expectation:

$$\begin{aligned}\mathbb{E}\left[|\nabla\varphi(X_k)^\top \Delta_k| \,|\, \mathcal{F}_k\right] &\le \mathbb{E}\left[|\nabla\varphi(X_k)^\top a_{k+1}| \cdot |f'(r_k)| \,|\, \mathcal{F}_k\right] \\ &\le \mathbb{E}\left[|\nabla\varphi(X_k)^\top a_{k+1}|^2 \,|\, \mathcal{F}_k\right]^{\frac{1}{2}} \cdot \mathbb{E}\left[f'(r_k)^2 \,|\, \mathcal{F}_k\right]^{\frac{1}{2}}\end{aligned}$$

The first term is bounded by a constant independent of d, since subgaussian moments are bounded. The second term is bounded by Lemma A.3.12, completing the proof. □

Lemma A.3.14 (Infinity norm bounds)**.** *For G_k, N_k, Q_k as defined in 2.1.2, we have, for any $\varepsilon, \lambda > 0$, there exists $C > 0$ such that,*

$$\|G_k\|_\infty \le d^\varepsilon \quad \textit{w.o.p. and} \quad \mathbb{E}[\|G_k\|_\infty^\lambda \,|\, \mathcal{F}_k] \le d^\varepsilon \quad \textit{w.o.p.,} \tag{A.32}$$

$$\|N_k\|_\infty \le C, \quad \|Q_k\|_\infty \le C, \quad \|\mathcal{G}_k\|_\infty \le C. \tag{A.33}$$

Proof. The first line, (A.32), follows from (A.26). For the first inequality, $\|G_k\|_\infty = \max_{0\le j\le k} \frac{\|\Delta_j\|^2}{d}$, which are all bounded by d^ε with overwhelming probability. A union bound tells us that the maximum is also bounded by d^ε w.o.p.. For the second inequality,

$$\begin{aligned}\mathbb{E}[|G_k|_\infty^\lambda \,|\, \mathcal{F}_k] &\le \mathbb{E}\left[\left(\frac{\|\Delta_k\|^2}{d}\right)^\lambda \,|\, \mathcal{F}_k\right] + \mathbb{E}\left[\max_{0\le j\le k-1}\left(\frac{\|\Delta_j\|^2}{d}\right)^\lambda \,|\, \mathcal{F}_k\right] \\ &\le \mathbb{E}\left[\left(\frac{\|\Delta_k\|^2}{d}\right)^\lambda \,|\, \mathcal{F}_k\right] + \max_{0\le j\le k-1}\left(\frac{\|\Delta_j\|^2}{d}\right)^\lambda \\ &\le d^\varepsilon, \qquad \text{(w.o.p.)}\end{aligned}$$

as desired. The second line is more straightforward:

$$\|N_k\|_\infty = \max_{0\le j\le k} \|(W_j^+)^\top W_j^+\|.$$

Now, $\|X^\star\|$ and $\|X_0\|$ are bounded independent of d, and $\|X_j\|$ is bounded by cM (because of the stopping time we are using.) Thus the maximum over j of their inner products are bounded by a constant. The same thing holds for $\|Q_k\|_\infty$:

$$\begin{aligned}\|Q_k\|_\infty &= \max_{0\le j\le k} \mathcal{R}(X_j) \\ &= \max_{0\le j\le k} h(W_j^\top K W_j).\end{aligned}$$

Since the derivative of h is pseudo-Lipschitz, h is continuous, and thus bounded for bounded arguments. And indeed, the argument to h is bounded:

$$\|W_j^\top K W_j\| \le \|W_j\|^2 \|K\|_{\mathrm{op}},$$

both of which are bounded independent of d. Finally, a similar argument applies to $\mathcal{G}_k$:

$$\|\mathcal{G}_k\|_\infty = \max_{0\le j\le k} \mathbb{E}\left[\frac{\|\Delta_j\|^2}{d} \,|\, \mathcal{F}_j\right] \le \max_{0\le j\le k} C = C$$

by Lemma A.3.12. □

We now prove a concentration result that closely follows [21, Proposition 5.6].

Lemma A.3.15 ([21], Lemma 5.2)**.** *Suppose $v \in \mathbb{R}^d$ is distributed $\mathcal{N}(0, I_d)$ and $U \in R^{d\times 2}$ has orthonormal columns. Then*

$$v \,|\, U^\top v \sim v - U(U^\top v) + UU^\top v, \tag{A.34}$$

where $v - U\left(U^T v\right) \sim N\left(0, I_d - UU^T\right)$ *and* $UU^T v \sim N\left(0, UU^T\right)$ *with* $v - U\left(U^T v\right)$ *independent of* $UU^T v$.

Lemma A.3.16. *For a matrix* $H = H_k$ *with bounded operator norm, or* $\|H\|_{op} < C$ *and* $\mathbb{E}[H_k \,|\, \mathcal{F}_k] = H_k$, *set* $q(a) = a^\top H a$. *Then*

$$\left|\mathbb{E}[q(a_{k+1}) f'(r_k)^2 \,|\, \mathcal{F}_k] - \operatorname{Tr}(KH)\, \mathbb{E}[f'(r_k)^2 \,|\, \mathcal{F}_k]\right| \leq C(H).$$

Note that the H *used here is not the same as the matrix used in the integro-differential equation.*

Proof. Many of the computations in this proof are taken directly from [21], but we repeat them here for completeness. We have $\mathcal{F}_k = \sigma(\{W_i\}_{i=0}^k)$; set $\hat{\mathcal{F}}_k = \sigma(\{W_i\}_{i=0}^k, \{r_i\}_{i=0}^k)$. A simple calculation shows that

$$\begin{aligned}\mathbb{E}[q(a_{k+1}) f'(r_k)^2 \,|\, \hat{\mathcal{F}}_k] = \mathbb{E}[q(a_{k+1} - \mathbb{E}[a_{k+1} \,|\, \hat{\mathcal{F}}_k]) \,|\, \hat{\mathcal{F}}_k]\, \mathbb{E}_\epsilon[f'(r_k)^2] \\ + q(\mathbb{E}[a_{k+1} \,|\, \hat{\mathcal{F}}_k])\, \mathbb{E}_\epsilon[f'(r_k)^2]. \end{aligned} \tag{A.35}$$

To compute the conditional mean $\mathbb{E}[a_{k+1} \,|\, \hat{\mathcal{F}}_k]$ and covariance $(a_{k+1} - \mathbb{E}[a_{k+1} \,|\, \hat{\mathcal{F}}_k])(a_{k+1} - \mathbb{E}[a_{k+1} \,|\, \hat{\mathcal{F}}_k])^\top$, we use Lemma A.3.15. By Assumption 2.1.1, we can write $a_{k+1} = \sqrt{K} v_k$, for $v_k \sim \mathcal{N}(0, I_d)$.

Now we perform a QR-decomposition on $\sqrt{K} W_k \overset{\text{def}}{=} Q_k R_k$ where $Q_k \in \mathbb{R}^{d \times 2}$ with orthonormal columns and $R_k \in \mathbb{R}^{2 \times 2}$ is upper triangular (and invertible). Set $\Pi_k \overset{\text{def}}{=} Q_k Q_k^T$. In distribution,

$$a_{k+1} \,|\, a_{k+1}^\top W_k \overset{\text{d}}{=} \sqrt{K} v_k \,|\, R_k^T Q_k^T v_k.$$

As R_k is invertible, by Lemma A.3.15,

$$a_{k+1} \,|\, a_{k+1}^\top W_k \overset{\text{d}}{=} \sqrt{K} v_k \,|\, Q_k^T v_k \overset{\text{d}}{=} \sqrt{K}\left(v_k - \Pi_k v_k\right) + \sqrt{K} \Pi_k v_k. \tag{A.36}$$

We note that $(I_d - \Pi_k) v_k \sim N(0, I_d - \Pi_k)$ and $\Pi_k v_k \sim N(0, \Pi_k)$ with $(I_d - \Pi_k) v_k$ independent of $\Pi_k v_k$. From this, we have that

$$\mathbb{E}[a_{k+1} \,|\, \hat{\mathcal{F}}_k] = \sqrt{K} \Pi_k v_k, \quad \text{where } v_k \sim N(0, I_d). \tag{A.37}$$

Moreover the conditional covariance of a_{k+1} is precisely

$$\begin{aligned}(\mathbb{E}[(a_{k+1} - \mathbb{E}[a_{k+1} \,|\, \hat{\mathcal{F}}_k])(a_{k+1} - \mathbb{E}[a_{k+1} \,|\, \hat{\mathcal{F}}_k])^\top \,|\hat{\mathcal{F}}_k]) \\ = \sqrt{K}(I_d - \Pi_k)\sqrt{K}, \quad \text{where } \Pi_k = Q_k Q_k^T.\end{aligned} \tag{A.38}$$

Next, using that $\mathbb{E}[H_k \,|\, \mathcal{F}_k] = H_k$, we expand (A.35) to get the leading order behavior

$$\begin{aligned}\mathbb{E}[q(a_{k+1}) f'(r_k)^2 \,|\, \hat{\mathcal{F}}_k] = \operatorname{Tr}(HK)\, \mathbb{E}_\epsilon[f'(r_k)^2] \\ - \operatorname{Tr}(H\sqrt{K}\Pi_k\sqrt{K})\, \mathbb{E}_\epsilon[f'(r_k)^2] \\ + q(\sqrt{K}\Pi_k v_k)\, \mathbb{E}_\epsilon[f'(r_k)^2].\end{aligned} \tag{A.39}$$

Taking the expectation with respect to $\mathcal{F}_k$, we obtain

$$\mathbb{E}[q(a_{k+1}) f'(r_k)^2 \,|\, \mathcal{F}_k] - \operatorname{Tr}(HK)\, \mathbb{E}[f'(r_k)^2 \,|\, \mathcal{F}_k] = \mathbb{E}[\mathcal{E}_k \,|\, \mathcal{F}_k], \tag{A.40}$$

where the error $\mathcal{E}_k$ is defined as

$$\mathcal{E}_k = - \operatorname{Tr}(H\sqrt{K}\Pi_k\sqrt{K})\, \mathbb{E}_\epsilon[f'(r_k)^2] \tag{A.41}$$

$$+ q(\sqrt{K}\Pi_k v_k)\, \mathbb{E}_\epsilon[f'(r_k)^2]. \tag{A.42}$$

The proof now turns to bounding the expectation of this error quantity.

$$\begin{aligned}|\operatorname{Tr}(H\sqrt{K}\Pi_k\sqrt{K})\, \mathbb{E}[f'(r_k)^2 \,|\, \mathcal{F}_k]| &= |\operatorname{Tr}(H\sqrt{K}\Pi_k\sqrt{K})| \cdot \mathbb{E}[f'(r_k)^2 \,|\, \mathcal{F}_k] \\ &\le \|H\|_{\mathrm{op}} \|K\|_{\mathrm{op}} |\operatorname{Tr}(\Pi_k)| \cdot \mathbb{E}[f'(r_k)^2 \,|\, \mathcal{F}_k] \\ &\le \|H\|_{\mathrm{op}} \|K\|_{\mathrm{op}} \cdot \operatorname{rank}(Q_k)\, \mathbb{E}[f'(r_k)^2 \,|\, \mathcal{F}_k] \\ &\le 2\|H\|_{\mathrm{op}} \|K\|_{\mathrm{op}}\, \mathbb{E}[f'(r_k)^2 \,|\, \mathcal{F}_k].\end{aligned}$$

By (A.25), the expectation is bounded by a constant, so this term is overall bounded by a constant. We move on to the next term in the error:

$$q(\sqrt{K}\Pi_k v_k) f'(r_k)^2 \le \|H\|_{\mathrm{op}} \|K\|_{\mathrm{op}} \|\Pi_k v_k\|^2 f'(r_k)^2.$$

Taking expectations and using Cauchy Schwarz, we obtain

$$\mathbb{E}[q(\sqrt{K}\Pi_k v_k) f'(r_k)^2 \,|\, \mathcal{F}_k] \le \|H\|_{\mathrm{op}} \|K\|_{\mathrm{op}} \cdot \sqrt{\mathbb{E}[\|\Pi_k v_k\|^4 \,|\, \mathcal{F}_k]} \cdot \sqrt{\mathbb{E}[f'(r_k)^4 \,|\, \mathcal{F}_k]}.$$

The first expectation is $\mathbb{E}[\|\Pi_k v_k\|^2 \,|\, \mathcal{F}_k] = \|\Pi_k\|_{\mathrm{F}}^4 = 8$, and the second is bounded by (A.25) as before. We thus conclude that $\mathbb{E}[\mathcal{E}_k \,|\, \mathcal{F}_k]$ is bounded by a constant depending on $\|H\|_{\mathrm{op}}$, completing the proof. □

Lemma A.3.17. *There is a constant C such that*

$$|\gamma(\mathcal{G}_k) - \gamma(B_k)| \leq Cd^{-1}.$$

Proof. Using the Lipschitz condition on the stepsize, we have

$$\begin{aligned}
&|\gamma(\mathcal{G}_k) - \gamma(B_k)| \\
&\quad \leq \|\mathcal{G}_k - \operatorname{Tr}(K)I(B_k)/d\|_\infty \times (1 + 2\|N_k\|_\infty^\alpha + \|\mathcal{G}_k\|_\infty^\alpha + \|\operatorname{Tr}(K)I(B_k)/d\|_\infty^\alpha + 2\|Q_k\|_\infty^\alpha) \\
&\quad \leq C\|\mathcal{G}_k - \operatorname{Tr}(K)I(B_k)/d\|_\infty && \text{(Lemma A.3.14)} \\
&\quad \leq Cd^{-1} \max_{0\leq j\leq k} \left\| \mathbb{E}[a_{j+1}^\top a_{j+1} f'(r_j)^2 \,|\, \mathcal{F}_j] - \operatorname{Tr}(K)\,\mathbb{E}[f'(r_j)^2 \,|\, \mathcal{F}_j] \right\| \\
&\quad \leq Cd^{-1}, && \text{(Lemma A.3.16)}
\end{aligned}$$

as desired. □

A.3.3 Specific learning rates

In this section, we confirm that AdaGrad-Norm satisfies Assumption 2.1.6. In the notation of Assumption 2.1.6, we have, for AdaGrad-Norm,

$$\gamma(td, f, g, q) = \frac{\eta}{\sqrt{b^2 + \int_0^\infty g(s)\ \mathrm{d}s}}.$$

Note that this reduces to the discrete stepsize if we plug in $g = G_k$:

$$\begin{aligned}
\gamma(td, f, G_k(d\times\cdot), q) &= \frac{\eta}{\sqrt{b^2 + \int_0^\infty G_k(ds)\ \mathrm{d}s}} \\
&= \frac{\eta}{\sqrt{b^2 + \int_0^\infty \left(1_{\{ds\leq k\}} \frac{1}{d} \sum_{i=0}^k \|\nabla_X \Psi(X_i; a_{i+1}, \epsilon_{i+1})\|^2 1_{[i,i+1)}(ds)\right) \mathrm{d}s}} \\
&= \frac{\eta}{\sqrt{b^2 + \int_0^\infty \left(1_{\{u\leq k\}} \frac{1}{d^2} \sum_{i=0}^k \|\nabla_X \Psi(X_i; a_{i+1}, \epsilon_{i+1})\|^2 1_{[i,i+1)}(u)\right) \mathrm{d}u}} \\
&= \frac{\eta}{\sqrt{b^2 + \frac{1}{d^2} \sum_{i=0}^k \|\nabla_X \Psi(X_i; a_{i+1}, \epsilon_{i+1})\|^2}},
\end{aligned}$$

which is exactly the discrete version of the AdaGrad-Norm stepsize.

Proposition A.3.18 (Lipschitz)**.** *For functions f, g, q such that $f(ds) = g(ds) = q(ds) = 0$ for $s > t$, the AdaGrad stepsize γ is Lipschitz. That is,*

$$|\gamma(td, f(d\times\cdot), g(d\times\cdot), q(d\times\cdot)) - \gamma(td, \hat{f}(d\times\cdot), \hat{g}(d\times\cdot), \hat{q}(d\times\cdot))| \le C(t,\gamma)(\|g-\hat{g}\|_\infty).$$

Remark A.3.19. *This is a stronger condition than the α-pseudo Lipschitz one in Assumption 2.1.6.*

Proof. To show this, we look at the derivative of the AdaGrad stepsize function. Setting $F(x) = \frac{\eta}{\sqrt{b^2+x}}$, we have

$$|F'(x)| = \frac{\eta}{2(b^2+x)^{3/2}} \le \frac{\eta}{2b^3}$$

for $x \in [0,\infty)$. We thus have

$$\begin{aligned}
&|\gamma(td, f(d\times\cdot), g(d\times\cdot), q(d\times\cdot)) - \gamma(td, \hat{f}(d\times\cdot), \hat{g}(d\times\cdot), \hat{q}(d\times\cdot))| \\
&= \left| \frac{\eta}{\sqrt{b^2 + \int_0^\infty g(ds)\ \mathrm{d}s}} - \frac{\eta}{\sqrt{b^2 + \int_0^\infty \hat{g}(ds)\ \mathrm{d}s}} \right| \\
&= \left| F\left(\int_0^\infty g(ds)\ \mathrm{d}s\right) - F\left(\int_0^\infty \hat{g}(ds)\ \mathrm{d}s\right) \right| \\
&\le \frac{\eta}{2b^3} \left| \int_0^\infty g(ds)\ \mathrm{d}s - \int_0^\infty \hat{g}(ds)\ \mathrm{d}s \right| \\
&\le \frac{\eta}{2b^3} \left| \int_0^t g(ds)\ \mathrm{d}s - \int_0^t \hat{g}(ds)\ \mathrm{d}s \right| \\
&\le \frac{\eta}{2b^3} (t \cdot \|g-\hat{g}\|_\infty) \\
&\le \frac{\eta t}{2b^3} \cdot \|g-\hat{g}\|_\infty,
\end{aligned}$$

where we were able to replace the ∞ with a t because $g(ds) = 0$ for $s > t$. We have thus obtained a Lipschitz constant $\frac{\eta t}{2b^3}$ depending only on t. □

Next we show that the AdaGrad-Norm is bounded.

Proposition A.3.20 (Boundedness)**.** *Suppose γ is AdaGrad-Norm. Then (2.6), as part of Assumption 2.1.6, holds.*

Proof. This is immediate:

$$\gamma(td, f, g, q) = \frac{\eta}{\sqrt{b^2 + \int_0^t g(s)\, ds}} \le \frac{\eta}{b}.$$

□

It remains to show that AdaGrad-Norm satisfies (2.5) in Assumption 2.1.6.

Proposition A.3.21 (Concentration)**.** *Suppose γ is AdaGrad-Norm, with G_k and $\mathcal{G}_k$ being defined as before. Then Equation (2.5), as part of Assumption 2.1.6, holds:*

$$\mathbb{E}[|\gamma(G_k) - \gamma(\mathcal{G}_k)| \,|\, \mathcal{F}_k] \le Cd^{-\delta}(1 + \|f\|_\infty^\alpha + \|q\|_\infty^\alpha).$$

Proof. Looking to remove the square roots, we have

$$|\gamma(G_k) - \gamma(\mathcal{G}_k)| \le |\gamma(G_k)^2 - \gamma(\mathcal{G}_k)^2|^{\frac{1}{2}}.$$

For AdaGrad-Norm, we have

$$\begin{aligned}\left|\gamma(G_k)^2 - \gamma(\mathcal{G}_k)^2\right| &= \eta^2 \left| \frac{1}{b^2 + \frac{1}{d^2}\sum_{j=0}^k \|\Delta_j\|^2} - \frac{1}{b^2 + \frac{1}{d^2}\sum_{j=0}^k \mathbb{E}[\|\Delta_j\|^2 \,|\, \mathcal{F}_j]} \right| \\ &\le \frac{\eta^2}{d^2 b^4} \cdot \left| \sum_{j=0}^k (\mathbb{E}[\|\Delta_j\|^2 \,|\, \mathcal{F}_j] - \|\Delta_j\|^2) \right|. \end{aligned} \tag{A.43}$$

We now bound the sum above. Set $F_i = \|\Delta_i\|^2/d$, $F_i^\beta = \mathrm{Proj}_\beta(F_i)$, $\Delta\mathcal{M}_i = F_i - \mathbb{E}[F_i \,|\, \mathcal{F}_i]$, and $\Delta\mathcal{M}_i^\beta = F_i^\beta - \mathbb{E}[F_i^\beta \,|\, \mathcal{F}_i]$. Then $|\Delta\mathcal{M}_i^\beta| \in [-2\beta, 2\beta]$, so Azuma's inequality gives us

$$\begin{aligned}\mathbb{P}\left(|\mathcal{M}_k^\beta| \ge t\right) &\le 2\exp\left(-\frac{-t^2}{2\sum_{i=0}^k (2\beta)^2}\right), \\ \mathbb{P}\left(|\mathcal{M}_k^\beta| \ge d^{1/2+\varepsilon}\right) &\le 2\exp\left(-\frac{-d^{1+2\varepsilon}}{2Td(2d^{\varepsilon/2})^2}\right) = \exp\left(-\frac{d^\varepsilon}{8T}\right).\end{aligned}$$

where we set $\beta = d^{\varepsilon/2}$. This is close to the bound we want: the error is

$$|\mathcal{M}_k - \mathcal{M}_k^\beta| \le \sum_{i=0}^k |F_i - F_i^\beta| + |\,\mathbb{E}[F_i - F_i^\beta \,|\, \mathcal{F}_i]|.$$

We have

$$\mathbb{P}(F_i - F_i^\beta \ne 0) = \mathbb{P}(|F_i| > \beta) = \mathbb{P}\left(\frac{\|\Delta_i\|^2}{d} > d^{\varepsilon/2}\right),$$

which superpolynomially small by (A.26). The expectation is similar:

$$\begin{aligned}|\,\mathbb{E}[F_i - F_i^\beta \,|\, \mathcal{F}_i]| &= |\,\mathbb{E}[(F_i - F_i^\beta)\mathbf{1}_{\{|F_i|>\beta\}} \,|\, \mathcal{F}_i]| \\ &\le \mathbb{E}[|F_i - F_i^\beta|^2 \,|\, \mathcal{F}_i]^{\frac{1}{2}} \cdot \mathbb{E}[\mathbf{1}_{\{|F_i|>\beta\}} \,|\, \mathcal{F}_i]^{\frac{1}{2}} \\ &\le 4\,\mathbb{E}[|F_i|^2 \,|\, \mathcal{F}_i]^{\frac{1}{2}} \cdot \mathbb{E}[\mathbf{1}_{\{|F_i|>\beta\}} \,|\, \mathcal{F}_i]^{\frac{1}{2}}.\end{aligned}$$

The first expectation is bounded by a constant independent of d by (A.26), and the second expectation is superpolynomially small by the same argument as above. We then have

$$|\mathcal{M}_k - \mathcal{M}_k^\beta| \leq d^{1/2+\varepsilon}$$

with overwhelming probability (note that this would be true for any power of d, by the definition of superpolynomially small.) We thus conclude that

$$|\mathcal{M}_k| \leq d^{1/2+\varepsilon}$$

with overwhelming probability. Multiplying by d, we find that

$$\left| \sum_{j=0}^{k} (\mathbb{E}[\|\Delta_j\|^2 \,|\, \mathcal{F}_j] - \|\Delta_j\|^2) \right| \leq d^{3/2+\varepsilon} \quad \text{w.o.p.}$$

Plugging this back into (A.43), we find that

$$\begin{aligned} \left|\gamma(G_k)^2 - \gamma(\mathcal{G}_k)^2\right| &\leq \frac{\eta^2}{d^2 b^4} d^{3/2+\varepsilon} \\ &\leq C d^{-1/2+\varepsilon} \end{aligned}$$

with overwhelming probability, and so, taking the square root,

$$|\gamma(G_k) - \gamma(\mathcal{G}_k)| \leq C d^{-1/4+\varepsilon/2} \quad \text{w.o.p,}$$

which is less than $d^{-1/4+\varepsilon}$ as d grows (we replaced the constant with an extra factor of $d^{\varepsilon/2}$.) Controlling the expectation via the boundedness of γ, we find that with $\delta = 1/8$,

$$\mathbb{E}[|\gamma(G_k) - \gamma(\mathcal{G}_k)| \,|\, \mathcal{F}_k] \leq d^{-\delta} \quad \text{w.o.p.,}$$

as desired. □

A.4 Proofs for AdaGrad-Norm analysis

In this section we provide proofs of the propositions related to AdaGrad-Norm in the least squares setting as well as the more general strongly convex setting. Statements of the propositions for least squares examples are found in Section 2.4.

A.4.1 Strongly convex setting

In order to derive the limiting learning rate in this case, we need the following assumption and some standard definitions of strong convexity.

Assumption A.4.1 (Risk and loss minimizer)**.** *Suppose that*

$$X^\star \in arg\,min_X\big\{\mathcal{R}(X) = \mathbb{E}_{a,\epsilon}[f(\langle X, a\rangle, \langle X^\star, a\rangle), \epsilon]\big\}$$

exists and has norm bounded independent of d. Then one has,

$$\langle X^\star, a\rangle \in arg\,min_x\{f(x, \langle X^\star, a\rangle, \epsilon)\}, \qquad \text{for almost surely } a \sim \mathcal{N}(0, K) \text{ and } \epsilon.$$

While at first, this assumption seems quite strong, in fact, in a typical student-teacher setup when label noise is 0 (i.e., $\epsilon = 0$), where the targets have the same model as the outputs, the assumption is satisfied. Our goal here is not to be exhaustive, but simply to illustrate that our framework admits a nontrivial and useful analysis and which gives nontrivial conclusions for the optimization theory of these problems.

Definition A.4.2 ($\hat{L}$-smoothness of outer function f)**.** *A function $f : \mathbb{R}^3 \to \mathbb{R}$ that is C^1-smooth (in the first variable) is called $\hat{L}(f)$-smooth if the following quadratic upper bound holds for any $x, \hat{x}, y, z \in \mathbb{R}$*

$$f(\hat{x}, y, z) \leq f(x, y, z) + \langle f'(x, y, z), \hat{x} - x\rangle + \tfrac{\hat{L}(f)}{2}|\hat{x} - x|^2. \tag{A.44}$$

Note that if $f' = \frac{\partial}{\partial x} f(x, y, z)$ is $\hat{L}(f)$-Lipschitz, i.e., $|f'(x, y, z) - f'(\hat{x}, y, z)| \leq \hat{L}(f)|x - \hat{x}|$, then the inequality (A.44) holds with constant $\hat{L}$. Suppose $x^\star \in \arg\min_x\{f(x, y, z)\}$ exists. An immediate consequence of (A.44) is that

$$\frac{1}{2\hat{L}(f)}|f'(x, y, z)|^2 \leq f(x, y, z) - f(x^\star, y, z) \leq \frac{\hat{L}(f)}{2}|x - x^\star|^2. \tag{A.45}$$

Definition A.4.3 (Restricted Secant Inequality)**.** *A function $f : \mathbb{R}^3 \to \mathbb{R}$ that is C^1-smooth (in the first variable) satisfies the (μ, θ)–restricted secant inequality (RSI) if, for any $x \in \mathbb{R}$ and $x^\star \in arg\,min_x\{f(x)\}$,*

$$\langle x - x^\star, f'(x)\rangle \geq \begin{cases} \mu|x - x^\star|^2, & \text{if } \max\{|x^\star|^2, |x - x^\star|^2\} \leq \theta, \\ 0, & \text{otherwise.} \end{cases}$$

If f satisfies the above for $\theta = \infty$, then we say f satisfies the μ–RSI.

Proposition A.4.4. *Let the outer function $f : \mathbb{R}^3 \to \mathbb{R}$ be a $\hat{L}(f)$-smooth function satisfying the RSI condition with $\hat{\mu}(f)$ with respect to $x \in \mathbb{R}$. Suppose $X^\star \in argmin_X\{\mathcal{R}(X)\}$ exists bounded, independent of d and Assumption A.4.1 holds and that $\gamma_0 = \frac{\eta}{b} = \frac{2\hat{\mu}(f)}{(\hat{L}(f))^2 \frac{1}{d}\operatorname{Tr}(K)}\zeta$, for some $\zeta \in (0,1)$, and that $\int_0^\infty \mathscr{R}(s)\gamma_s \,\mathrm{d}s < \infty$ with γ_s as in Table 2.2 (AdaGrad-Norm, general formula), then*

$$\gamma_\infty \geq \frac{\gamma_0 \eta^2}{1 + \frac{\zeta}{1-\zeta}\mathscr{D}^2(0)}.$$

Proof. Given the Eq. (A.73) for the distance to optimality, with $(x, x^\star) \sim \mathcal{N}(0, \mathscr{B})$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}^2(t) = -2\gamma_t \,\mathbb{E}_{a,\epsilon}[\langle x - x^\star, f'(x, x^\star)\rangle] + \frac{\gamma_t^2}{d}\operatorname{Tr}(K)\,\mathbb{E}_{a,\epsilon}[(f'(x, x^\star))^2]$$

By the RSI (with constant $\hat{\mu}(f)$) condition on f, we have that

$$\mathbb{E}_{a,\epsilon}\left[\langle x - x^\star, f'(x, x^\star)\rangle\right] \geq \hat{\mu}(f)\,\mathbb{E}_{a,\epsilon}[(x - x^\star)^2] = 2\hat{\mu}(f)\mathscr{R}(t), \tag{A.46}$$

where $x = \langle X, a\rangle$ and $x^\star = \langle X^\star, a\rangle$ and we note that x has t-dependence due to the t-dependence in $\mathscr{B}$. By $\hat{L}(f)$-smoothness,

$$\frac{1}{2\hat{L}(f)}(f'(x))^2 \leq \frac{\hat{L}(f)}{2}(x - x^\star)^2.$$

This implies that

$$\frac{1}{2(\hat{L}(f))^2}\,\mathbb{E}_{a,\epsilon}\left[(f'(x, x^\star)^2\right] \leq \frac{1}{2}\,\mathbb{E}_{a,\epsilon}\left[(x - x^\star)^2\right] = \mathscr{R}(t). \tag{A.47}$$

Thus by (A.46) and (A.47), we have that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}^2(t) \leq -\gamma_t\left(4\hat{\mu}(f) - 2(\hat{L}(f))^2\frac{1}{d}\operatorname{Tr}(K)\gamma_t\right)\mathscr{R}(t)$$

Which then yield:

$$\mathscr{D}^2(t) \leq \mathscr{D}^2(0) - 2\left(2\hat{\mu}(f) - (\hat{L}(f))^2\frac{1}{d}\operatorname{Tr}(K)\gamma_0\right)\int_0^t \mathscr{R}(s)\gamma_s \,\mathrm{d}s.$$

Changing variables $u = \Gamma(t) = \int_0^t \gamma_s \,\mathrm{d}s$, we have that $\int_0^\infty \mathscr{R}(t)\gamma_t \,\mathrm{d}t = \int_0^\infty r(u)\,\mathrm{d}u = \|r\|_1$. Rearranging the term in the above equation and taking $t \to \infty$. We obtain: $\|r\|_1 \leq \frac{\mathscr{D}^2(0)}{\left(2\hat{\mu}(f) - (\hat{L}(f))^2\frac{1}{d}\operatorname{Tr}(K)\gamma_0\right)}$, given that $\frac{2\hat{\mu}(f)}{(\hat{L}(f))^2\frac{1}{d}\operatorname{Tr}(K)} > \gamma_0$. Using Lemma A.4.5, with $i(v) =$

$I(\mathscr{B}(\Gamma^{-1}(v))) = \mathbb{E}_{a,\epsilon}\left[(f'(x,x^\star)^2\right]$ instead of the risk

$$\gamma_\infty = \frac{\eta^2}{\frac{b}{\eta} + \frac{1}{2d}\operatorname{Tr}(K)\int_0^\infty i(v)\,\mathrm{d}v} \geq \frac{\eta^2}{\frac{b}{\eta} + \frac{1}{d}\operatorname{Tr}(K)(\hat{L}(f))^2\int_0^\infty r(v)\,\mathrm{d}v} \tag{A.48}$$
$$\geq \frac{\eta^2}{\frac{b}{\eta} + \frac{1}{d}\operatorname{Tr}(K)\frac{(\hat{L}(f))^2\mathscr{D}^2(0)}{\left(2\hat{\mu}(f) - (\hat{L}(f))^2\frac{1}{d}\operatorname{Tr}(K)\gamma_0\right)}} = \frac{\eta^2}{\frac{b}{\eta} + \frac{\frac{1}{d}\operatorname{Tr}(K)(\hat{L}(f))^2}{2\hat{\mu}(f)(1-\zeta)}\mathscr{D}^2(0)}}.$$

where the first inequality is by Eq. A.47, and the last transition is by taking the initial learning rate to be $\gamma_0 = \frac{2\hat{\mu}(f)}{(\hat{L}(f))^2\frac{1}{d}\operatorname{Tr}(K)}\zeta$, for $\zeta \in (0,1)$. □

Lemma A.4.5. *Given γ_t as in Table 2.2 (AdaGrad-Norm), defining $g(u) = \gamma(\Gamma^{-1}(u))$, with $\Gamma(t) = \int_0^t \gamma_s\,\mathrm{d}s$, then $g(u) = \frac{\eta^2}{\frac{b}{\eta} + \frac{1}{2d}\operatorname{Tr}(K)\int_0^u i(v)\,\mathrm{d}v}$ with $i(v) = I(\mathscr{B}(\Gamma^{-1}(v)))$.*

Proof. Taking the square of both sides of the γ_t equation in Table 2.2 (AdaGrad-Norm), changing variables to $u = \Gamma(t)$ and rearranging the terms:

$$b^2 + \frac{\operatorname{Tr}(K)}{d}\int_0^u \frac{i(v)}{g(v)}\,\mathrm{d}v = \frac{\eta^2}{g(u)^2}, \tag{A.49}$$

such that $i(v) = I(\mathscr{B}(\Gamma^{-1}(v)))$. Taking derivative with respect to u, rearranging terms and integrating leads to the desired result. □

A.4.2 Least squares setting

To study the effect of the structured covariance matrix and cases in which the problem is not strongly convex, we will focus on the linear least square problem. In this setting, the continuum limit of the risk for the AdaGrad-Norm algorithm has the form of a convolutional integral Volterra equation,

$$\mathscr{R}(t) = F(\Gamma(t)) + \int_0^t \gamma_s^2 \mathcal{K}(\Gamma(t) - \Gamma(s))\mathscr{R}(s)\,\mathrm{d}s \tag{A.50}$$

where $\Gamma(t) := \int_0^t \gamma_s\,\mathrm{d}s$ with,

$$F(x) \overset{\text{def}}{=} \frac{1}{2d}\sum_{i=1}^d \lambda_i \mathscr{D}_i^2(0)e^{-2\lambda_i x}, \tag{A.51}$$

$$\mathcal{K}(x) \overset{\text{def}}{=} \frac{1}{d}\sum_{i=1}^d \lambda_i^2 e^{-2\lambda_i x}. \tag{A.52}$$

In the following we consider three cases, a strongly convex risk in which the spectrum of the eigenvalues is bounded from below (section A.4.2). A case in which the spectrum is not bounded from below as $d \to \infty$, but the number of eigenvalues below some fixed threshold is $o(d)$ (section A.4.2). Finally, power law spectrum supported on $[0, 1]$ with $d \to \infty$ (section A.4.2).

Proofs for case of fixed d

Proof of Proposition 2.4.2. Define the composite functions $r(u) = \mathscr{R}(\Gamma^{-1}(u))$, and $g(u) = \gamma(\Gamma^{-1}(u))$. Integrating the formula for the risk:

$$\begin{aligned}\int_0^t r(u)\,\mathrm{d}u &= \int_0^t F(u)\,\mathrm{d}u + \int_0^t \int_0^{\Gamma^{-1}(u)} \gamma_s^2 \mathcal{K}(u - \Gamma(s))\mathscr{R}(s)\,\mathrm{d}s\,\mathrm{d}u \\ &= \int_0^t F(u)\,\mathrm{d}u + \int_0^t \int_0^u \mathcal{K}(u-x)r(x)g(x)\,\mathrm{d}x\,\mathrm{d}u \\ &\le \int_0^t F(u)\,\mathrm{d}u + \gamma_0 \int_0^t r(x) \int_x^t \mathcal{K}(u-x)\,\mathrm{d}u\,\mathrm{d}x\end{aligned}$$

Taking $t \to \infty$, we get

$$\|r\|_1 \le \|F\|_1 + \gamma_0 \|\mathcal{K}\|_1 \|r\|_1.$$

Using $\|\mathcal{K}\|_1 = \int_0^\infty \mathcal{K}(x)\,\mathrm{d}x < \gamma_0^{-1}$, and noting that by Eq. (A.52), and Eq. (A.51), we have that $\|F\|_1 = \frac{1}{4}\mathscr{D}^2(0)$, and $\|\mathcal{K}\|_1 = \frac{1}{2d}\operatorname{Tr}(K)$,

$$\|r\|_1 \le \frac{\|F\|_1}{1 - \gamma_0\|\mathcal{K}\|_1} = \frac{\frac{1}{4}\mathscr{D}^2(0)}{1 - \frac{\gamma_0}{2d}\operatorname{Tr}(K)}.$$

On the hand following Lemma A.4.8, $\frac{1}{4}\mathscr{D}^2(0)(1 + \frac{\gamma_0}{2d}\operatorname{Tr}(K)) \le \|r\|_1$. Therefore, $\|r\|_1 \asymp \frac{1}{4}\mathscr{D}^2(0)$.

Next, rewriting the γ_t equation in Table 2.2 (AdaGrad-Norm for least squares) in terms of $g(u)$ (Lemma A.4.5), we obtain

$$g(u) = \frac{\eta^2}{\frac{b}{\eta} + \frac{1}{d}\operatorname{Tr}(K)\int_0^u r(x)\,\mathrm{d}x} \tag{A.53}$$

Taking $u \to \infty$, and using $\|r\|_1 \asymp \frac{1}{4}\mathscr{D}^2(0)$,

$$\gamma_\infty = g(\infty) = \frac{\eta^2}{\frac{b}{\eta} + \frac{1}{d}\operatorname{Tr}(K)\|r\|_1} \asymp \frac{\eta^2}{\frac{b}{\eta} + \frac{1}{4d}\operatorname{Tr}(K)\mathscr{D}^2(0)}. \tag{A.54}$$

This then completes the proof. □

Remark A.4.6. *We note that, on the Least square problem $\hat{L}(f) = \hat{\mu}(f) = 1$, therefore, the bound in Proposition A.4.4 yields* $\frac{\eta^2}{\frac{b}{\eta} + \frac{1}{2(1-\zeta)} \frac{1}{d} \operatorname{Tr}(K) \mathscr{D}^2(0)}$.

Proof of Proposition 2.4.1. Using the equation for the distance to optimality (Eq. 2.8), we can derive an equation for the integral of the risk (with no target noise) which we denote by $g(t) = \int_0^t \mathscr{R}(s)\, \mathrm{d}s$:

$$g''(t) = -\gamma_t \sum_i \lambda_i^2 \mathscr{D}_i^2(t) + \gamma_t^2 \frac{\operatorname{Tr}(K^2)}{d} g'(t). \tag{A.55}$$

For $K = I_d$, this equation simplifies,

$$g''(t) = -2\gamma_t g'(t) + \gamma_t^2 \frac{\operatorname{Tr}(K^2)}{d} g'(t). \tag{A.56}$$

Plugging in the equation for the AdaGrad-Norm learning rate (Table 2.2) leads to the desired result. We note that by using the equation for the learning rate, one can also derive a close equation for the learning rate itself. □

Vanishingly few eigenvalues near 0 as $d \to \infty$

We now consider the case where, as $d \to \infty$, there are eigenvalues of K arbitrarily close to 0. In Proposition 2.4.2 we saw a constant lower bound on γ_t when d is fixed (and thus there are finitely many eigenvalues within any fixed distance of 0). This can be extended to the case where we have some $C > 0$ such that the number of eigenvalues of K below C is $o(d)$ (see Proposition 2.4.3).

Proof of Proposition 2.4.3. Following the structure of the loss, after some time the risk starts to decrease, and therefore $\mathscr{R}(t) \leq R_0$ for and $t \geq 0$. Using these observations, we obtain a preliminary lower bound of $\gamma_t > C_1 t^{-1/2}$ (for $t > 0$), which enables us to deduce that $\mathscr{R}(t)$ is integrable and finally obtain a constant lower bound for γ_t. The details of this are below.

For $t \geq 0$ and some $C_1 > 0$,

$$\gamma_t = \frac{\eta}{\sqrt{b^2 + \frac{2}{d} \operatorname{Tr}(K) \int_0^t \mathscr{R}(s) ds}} \geq \frac{\eta}{\sqrt{b^2 + \frac{2}{d} \operatorname{Tr}(K) R_0 t}} \geq C_1 t^{-1/2}. \tag{A.57}$$

Next, to show that the risk is integrable, we divide the matrix K into two parts K_+, and K_-, such that the eigenvalues of K_+ are greater than some $\alpha_s > 0$ and the eigenvalues of

K_- are smaller than α_s where α_s is a decreasing function of s to be determined later. We then have that, following Eq. (2.8), and the definition of the risk $\mathscr{R}(t) = \frac{1}{2d}\sum_{i=1}^d \lambda_i \mathscr{D}_i^2(t)$,

$$\begin{aligned}\mathscr{R}(t) &= \mathscr{R}(0) - \frac{1}{d}\sum_{i=1}^d \lambda_i^2 \int_0^t \gamma_s \mathscr{D}_i(s)\mathrm{d}s + \frac{1}{d}\int_0^t \gamma_s^2 \operatorname{Tr}(K^2)\cdot \mathscr{R}(s)\mathrm{d}s \\ &\leq \mathscr{R}(0) - \int_0^t \gamma_s(2\alpha_s - \gamma_s \frac{1}{d}\operatorname{Tr}(K^2))\cdot \mathscr{R}(s)\mathrm{d}s + 2\int_0^t \gamma_s \mathscr{R}_2(s)\,\mathrm{d}s\end{aligned} \tag{A.58}$$

with $\mathscr{R}_2(s) = \frac{1}{2d}\sum_{i:\lambda_i\leq\alpha_s}\lambda_i\mathscr{D}_i^2(s)$. Next, choosing $\alpha_s = \gamma_s \frac{1}{d}\operatorname{Tr}(K^2)$, we show that the last term is of order $o_d(1)$. By Lemma A.4.7 $\forall i$, $\mathscr{D}_i^2(t) \leq \max\left(\gamma_{t_1}\mathscr{R}(t_1), \mathscr{D}_i^2(0)\right) = c_0$ where the bound c_0 comes from the assumption $\langle X^\star, \omega_i\rangle = O(d^{-1/2})$ and the initialization $X_0 = 0$. Therefore,

$$2\int_0^t \gamma_s \mathscr{R}_2(s)\,\mathrm{d}s \leq \frac{1}{d^2}\operatorname{Tr}(K^2)c_0\int_0^t \gamma_s N_s\,\mathrm{d}s. \tag{A.59}$$

where $N_s = \sum_{i=1}^d 1_{\lambda_i\leq\gamma_s\frac{1}{d}\operatorname{Tr}(K^2)}$. This implies that, if $\gamma_s N_s = o(d)$, then $2\int_0^t \gamma_s\mathscr{R}_2(s)\,\mathrm{d}s = o_d(1)$, provided that d is taken to be large before t.

We then have that up to $o_d(1)$ constant,

$$\mathscr{R}(t) \leq \mathscr{R}(0) - \frac{1}{d}\operatorname{Tr}(K^2)\int_0^t \gamma_s^2\cdot\mathscr{R}(s)\mathrm{d}s. \tag{A.60}$$

Using Gronwall's inequality,

$$\mathscr{R}(t) \leq \mathscr{R}(0)e^{-\frac{1}{d}\operatorname{Tr}(K^2)\int_0^t\gamma_s^2\,\mathrm{d}s} \leq \mathscr{R}(0)e^{-\frac{1}{d}\operatorname{Tr}(K^2)C_1^2 t} \tag{A.61}$$

where in the last transition we used the lower bound on the learning rate derived in Eq. (A.57). Thus, the risk is integrable, i.e. there is some C_3 such that

$$\int_0^t \mathscr{R}(s)\,\mathrm{d}s \leq \frac{\mathscr{R}(0)}{\frac{1}{d}\operatorname{Tr}(K^2)C_1^2}$$

for all $t > 0$. Finally, we plug this into the formula for γ_t and conclude that, for all $t > 0$,

$$\gamma_t \geq \frac{\eta}{\sqrt{b^2 + \frac{\frac{1}{d}\operatorname{Tr}(K)\mathscr{R}(0)}{\frac{1}{d}\operatorname{Tr}(K^2)C_1^2}}}. \tag{A.62}$$

□

Lemma A.4.7. *Assume that the risk is bounded and attains its maximum at time t_1. Then, for each i, we have $\mathscr{D}_i^2(t) \leq \max(\gamma_{t_1}\mathscr{R}(t_1), \mathscr{D}_i^2(0))$ for all $t \geq 0$.*

Proof. Case 1: Suppose that $\mathscr{D}_i^2(0) \leq \gamma_0 \mathscr{R}(0)$. Then, by equation (2.8), $\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}_i^2(0) \geq 0$. However, since $\mathscr{D}_i^2(t), \mathscr{R}(t)$ are continuous, this equation implies that $\mathscr{D}_i^2(t) \leq \gamma_t \mathscr{R}(t)$ for all t and thus $\mathscr{D}_i^2(t) \leq \gamma_{t_1} \mathscr{R}(t_1)$ for all t.

Case 2: Suppose that $\mathscr{D}_i^2(0) > \gamma_0 \mathscr{R}(0)$. Then, by equation (2.8), $\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}_i^2(0) < 0$. If $\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}_i^2(t) < 0$ for all t, then $\mathscr{D}_i^2(t) \leq \mathscr{D}_i^2(0)$ for all t. If at some point $\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}_i^2(t) > 0$, this implies $\mathscr{D}_i^2(t) \leq \gamma_t \mathscr{R}(t)$ and we are in Case 1. □

In the next section, we consider cases in which the risk is not integrable, an example of such case is when the spectrum of K is supported on the interval $[0, 1]$ or has power-law behavior near 0.

Power law behavior at $d \to \infty$

Non-asymptotic bound for the Convolutional Volterra In this section, we use the convolutional Volterra structure of the risk (Eq. (A.50)) to derive non-asymptotic bounds on the risk, which will be useful in Section A.4.2 to derive the asymptotic behavior of the risk and the learning rate under power law assumption on the spectrum of the covariance matrix and the discrepancy from the target at initialization.

Lemma A.4.8. *Let $\Gamma(t) := \int_0^t \gamma_s \, \mathrm{d}s$ and let*

$$\mathscr{R}(t) = F(\Gamma(t)) + \int_0^t \gamma_s^2 \mathcal{K}(\Gamma(t) - \Gamma(s)) \mathscr{R}(s) \, \mathrm{d}s$$

where $\gamma_t, \mathcal{K}$ are monotonically decreasing, with $\|\mathcal{K}\|_1 < \infty$. Then all t,

$$\mathscr{R}(t) \geq F(\Gamma(t)) + \int_0^t \gamma_s^2 \mathcal{K}(\Gamma(t) - \Gamma(s)) F(\Gamma(s)) \, \mathrm{d}s$$

If in addition, there exist $\epsilon > 0$ and $T > 0$ such that, for all $t > T$,

$$\int_0^t \mathcal{K}(s)\mathcal{K}(t-s) \, \mathrm{d}s \leq 2(1+\epsilon)\|\mathcal{K}\|_1 \mathcal{K}(t) \quad \text{and} \quad 2\|\mathcal{K}\|_1(1+\epsilon)\gamma_0 < 1$$

then for all t

$$\mathscr{R}(t) \leq F(\Gamma(t)) + C \int_0^t \gamma_s^2 \mathcal{K}(\Gamma(t) - \Gamma(s)) F(\Gamma(s)) \, \mathrm{d}s$$

for

$$C = \left(\frac{\mathcal{K}(0)}{\mathcal{K}(T)(2\epsilon + 1)} + 2 \right) \frac{1}{1 - 2\gamma(0)\|\mathcal{K}\|_1(1+\epsilon)}.$$

Proof. The lower bound holds trivially, using $\mathscr{R}(s) \geq F(\Gamma(s))$. For the upper bound, we start with the following change of variables:

$$\mathscr{R}(t) = F(\Gamma(t)) + \int_0^{\Gamma(t)} g(u)\mathcal{K}(\Gamma(t) - u))\mathscr{R}(u)\,\mathrm{d}u,$$

with $g(u) = \gamma_{\Gamma^{-1}(u)}$. Let us define the convolution map

$$\mathcal{G}(f)(\Gamma) = \mathcal{K} * (gf)(\Gamma) = \int_0^{\Gamma} \mathcal{K}(\Gamma - u)g(u)f(u)\,\mathrm{d}u.$$

Next we show that this map is contracting and in particular,

$$\begin{aligned}
\mathcal{G}^2(f) = \mathcal{G}(\mathcal{G}(f))(t) &= \int_0^t \mathcal{K}(t-s)\mathcal{G}(f)(s)g(s)\,\mathrm{d}s \qquad \text{(A.63)}\\
&= \int_0^t \mathcal{K}(t-s)\int_0^s \mathcal{K}(s-u)g(u)f(u)\,\mathrm{d}u g(s)\,\mathrm{d}s\\
&= \int_0^t \left(\int_u^t \mathcal{K}(t-s)\mathcal{K}(s-u)g(s)\,\mathrm{d}s\right) g(u)f(u)\,\mathrm{d}u\\
&\leq \int_0^t \mathcal{K}^{*2}(t-u)g(u)^2 f(u)\,\mathrm{d}u
\end{aligned}$$

where the third transition is since $u < s < t$. The last transition is by change of variables and the assumption that γ_t is a monotone decreasing function. Consecutive application of the convolution map will then yield by induction,

$$\mathcal{G}^j(f)(t) \leq \int_0^t \mathcal{K}^{*(j)}(t-u)g(u)^j f(u)\,\mathrm{d}u.$$

Therefore, expanding the loss and using the above upper bound, and denote by $q = 2(1+\varepsilon)\|\mathcal{K}\|_1\gamma_0$ such that $q < 1$,

$$\begin{aligned}
\mathscr{R}(t) &= F(t) + \sum_{j=1}^{\infty} \mathcal{G}^j(F)(t) \qquad \text{(A.64)}\\
&\leq F(t) + \sum_{j=1}^{\infty}\int_0^t \mathcal{K}^{*(j)}(t-u)g(u)^j F(u)\,\mathrm{d}u\\
&\leq F(t) + \left(\sum_{j=0}^{\infty}(2\|\mathcal{K}\|_1\gamma_0(1+\varepsilon))^j - 1\right) C_1 \int_0^t \mathcal{K}(t-u)g(u)F(u)\,\mathrm{d}u\\
&\leq F(t) + \frac{q}{1-q}C_1(\mathcal{K} * (gF))(t) \qquad \text{(A.65)}
\end{aligned}$$

where the third transition is by Lemma A.4.9, with $C_1 = \frac{\mathcal{K}(0)}{\mathcal{K}(T)(2\epsilon+1)} + 1$, which then completes the proof. □

Lemma A.4.9 (Lemma IV.4.7 in [7]). *Suppose $\mathcal{K}$ is monotonically decreasing, with $\|\mathcal{K}\|_1 < \infty$, and that there exists $T > 0$ such that $\forall t \geq T$, and $\epsilon \geq 0$,*

$$\int_0^t \mathcal{K}(s)\mathcal{K}(t-s)\,\mathrm{d}s \leq 2(1+\epsilon)\|\mathcal{K}\|_1\mathcal{K}(t). \tag{A.66}$$

Then,

$$\sup_{t\geq 0}\frac{\mathcal{K}^{*n}(t)}{\mathcal{K}(t)} \leq (2\|\mathcal{K}\|_1(1+\epsilon))^{n-1}\left(\frac{\mathcal{K}(0)}{\mathcal{K}(T)(2\epsilon+1)}+1\right) \tag{A.67}$$

Proof. Define $\alpha_n = \sup_{t\geq 0}\frac{\mathcal{K}^{*n}(t)}{\mathcal{K}(t)(2\|\mathcal{K}\|_1)^{n-1}}$, trivially $\alpha_1 = 1$. Consider the $n+1$ convolution,

$$\frac{\mathcal{K}^{*(n+1)}(t)}{\mathcal{K}(t)(2\|\mathcal{K}\|_1)^n} = \frac{1}{\mathcal{K}(t)}\int_0^t \frac{\mathcal{K}(s)\mathcal{K}^{*n}(t-s)}{(2\|\mathcal{K}\|_1)^n}\,\mathrm{d}s \tag{A.68}$$

By the assumption of the Lemma, we know that there exists some $T > 0$ such that for $\forall t \geq T$

$$\int_0^t \frac{\mathcal{K}(s)\mathcal{K}(t-s)}{2\|\mathcal{K}\|_1}\,\mathrm{d}s \leq (1+\epsilon)\mathcal{K}(t). \tag{A.69}$$

Therefore, if $t \geq T$, we have

$$\begin{aligned}&\frac{1}{\mathcal{K}(t)}\int_0^t \frac{\mathcal{K}(s)\mathcal{K}^{*n}(t-s)}{(2\|\mathcal{K}\|_1)^n}\,\mathrm{d}s \\ &= \int_0^t \frac{\mathcal{K}(s)\mathcal{K}(t-s)}{2\|\mathcal{K}\|_1}\frac{\mathcal{K}^{*n}(t-s)}{\mathcal{K}(t-s)(2\|\mathcal{K}\|_1)^{n-1}}\,\mathrm{d}s \leq \alpha_n(1+\epsilon)\end{aligned} \tag{A.70}$$

On the other hand, if $t < T$,

$$\frac{1}{\mathcal{K}(t)}\int_0^t \frac{\mathcal{K}(s)\mathcal{K}^{*n}(t-s)}{(2\|\mathcal{K}\|_1)^n}\,\mathrm{d}s \leq \frac{\mathcal{K}(0)}{\mathcal{K}(T)}\frac{\|\mathcal{K}^{*n}(t)\|_1}{(2\|\mathcal{K}\|_1)^n} \leq \frac{\mathcal{K}(0)}{\mathcal{K}(T)2^n} \tag{A.71}$$

Taking supremum in Eq. (A.68), and combining the results of Eq. (A.71), and Eq. (A.70), we obtain that,

$$\alpha_{n+1} \leq \frac{\mathcal{K}(0)}{\mathcal{K}(T)2^n} + \alpha_n(1+\epsilon)$$

Solving the above recursion equation,

$$\begin{aligned}\alpha_n &\leq \frac{\mathcal{K}(0)}{\mathcal{K}(T)}\sum_{k=0}^{n-2}\frac{1}{2^{n-k-1}}(1+\epsilon)^k + (1+\epsilon)^{n-1} = \frac{\mathcal{K}(0)}{\mathcal{K}(T)2^{n-1}}\frac{1-(2(1+\epsilon))^{n-1}}{1-2(1+\epsilon)} + (1+\epsilon)^{n-1} \\ &\leq (1+\epsilon)^{n-1}\left(\frac{\mathcal{K}(0)}{\mathcal{K}(T)(2\epsilon+1)}+1\right),\end{aligned}$$

rearranging the terms we arrived at the required result. □

Asymptotic analysis of the risk Here, we consider a family of models with $d \to \infty$, for which the following power law asymptotics assumption is satisfied:

Assumption A.4.10. *$F(x) \asymp x^{-\kappa_1}$ and $\mathcal{K}(x) \asymp x^{-\kappa_2}$ for $x \geq 1$ with $\kappa_1 \geq 0$, $\kappa_2 > 1$*

Corollary A.4.11 apply Lemma A.4.8 in the setting for which F, and $\mathcal{K}$ has a power law behavior asymptotically. It shows that the risk will then be dominated by F only. Corollary A.4.12 shows the behavior of the learning rate in this setting. Finally, Lemma A.4.13 shows that Assumption A.4.10 is a consequence of a power law spectrum near zero on the eigenvalues of the covariance matrix and a power law assumption on the projected discrepancy at initialization.

Corollary A.4.11. *Suppose Assumption A.4.10 is satisfied, then $\mathscr{R}(t) \asymp F(\Gamma(t))$.*

Proof. Define $g(u) = \gamma_{\Gamma^{-1}(u)}$ and $r(u) = \mathscr{R}(\Gamma^{-1}(u))$ and observe that $g(u)$ is a decreasing function. Then, from the upper bound in Lemma A.4.8, we have

$$\begin{aligned}
r(u) &\leq F(u) + C \int_0^u g(v)\mathcal{K}(u-v)F(v)\,\mathrm{d}v \\
&= F(u) + C\left(\int_0^{u/2} g(v)\mathcal{K}(u-v)F(v)\,\mathrm{d}v + \int_{u/2}^u g(v)\mathcal{K}(u-v)F(v)\,\mathrm{d}v\right) \\
&\leq F(u) + C_1 g(0)\left(\left(\frac{u}{2}\right)^{-\kappa_2}\int_0^{u/2} F(v)\,\mathrm{d}v + \left(\frac{u}{2}\right)^{-\kappa_1}\int_{u/2}^u \mathcal{K}(u-v)\,\mathrm{d}v\right) \\
&\leq F(u) + C_2(u^{-\kappa_2+1-\kappa_1} + u^{-\kappa_1}\|\mathcal{K}\|) \\
&= O(F(u)).
\end{aligned} \tag{A.72}$$

Combining this upper bound with the lower bound from Lemma A.4.8 and that $\kappa_2 > 1$, we conclude that $r(u) \asymp F(u)$ and $\mathscr{R}(t) \asymp F(\Gamma(t))$. □

Next, we derive the asymptotics of γ_t. There are three different cases, depending on whether the risk is integrable, which translates to a threshold with respect to the parameter κ_1.

Corollary A.4.12. *Suppose Assumption A.4.10 then the following asymptotics for the learning rate hold:*

- *For $\kappa_1 > 1$, there exists $\tilde{\gamma}$ such that $\gamma_t \geq \tilde{\gamma}$ and $\mathscr{R}(t) \asymp t^{-\kappa_1}$ for all $t \geq 0$.*
- *For $\kappa_1 < 1$, $\gamma_t \asymp t^{-(1-\kappa_1)/(2-\kappa_1)}$ and $\mathscr{R}(t) \asymp t^{-\frac{\kappa_1}{2-\kappa_1}}$ for all $t \geq 1$.*
- *For $\kappa_1 = 1$, $\gamma_t \asymp \frac{1}{\log(t+1)}$ and $\mathscr{R}(t) \asymp (\frac{t}{\log(t+1)})^{-\kappa_1}$ for all $t \geq 1$.*

Proof. Using the notations $g(u)$ and $r(u)$ defined above along with the change of variable $u = \Gamma(t)$, we get $\int_0^t \mathscr{R}(s)\,\mathrm{d}s = \int_0^u \frac{r(v)}{g(v)}\,\mathrm{d}v$. Combining this with Corollary A.4.11 and the formula for γ_t we get

$$g(u) \asymp \frac{\eta}{\sqrt{b^2 + \frac{2}{d}\operatorname{Tr}(K)\int_0^u \frac{(1+v)^{-\kappa_1}}{g(v)}\,\mathrm{d}v}}.$$

Let $I(u) = b^2 + \frac{2}{d}\operatorname{Tr}(K)\int_0^u \frac{(1+v)^{-\kappa_1}}{g(v)}\,\mathrm{d}v$ and observe that $g(u) \asymp \frac{1}{\sqrt{I(u)}}$ and $I'(u) = \frac{2}{d}\operatorname{Tr}(K)\frac{(1+u)^{-\kappa_1}}{g(u)}$. Thus, $I(u)$ satisfies $\frac{I'(u)}{\sqrt{I(u)}} \asymp (1+u)^{-\kappa_1}$ so we have

$$\sqrt{I(u)} - \sqrt{I(0)} \asymp \int_0^u (1+v)^{-\kappa_1}\,\mathrm{d}v.$$

In the case of $\kappa_1 > 1$, this implies $\sqrt{I(u)} \leq \sqrt{I(0)} + C\int(1+v)^{-\kappa_1}\,\mathrm{d}v$. This upper bound on $I(u)$ gives a corresponding lower bound on $g(u)$ and thus a lower bound on γ_t.

In the case of $\kappa_1 < 1$, we have $\sqrt{I(u)} - \sqrt{I(0)} \asymp (1+v)^{1-\kappa_1}$ so, for u sufficiently large, $g(u) \asymp (1+u)^{\kappa_1 - 1}$. To recover the asymptotic for γ_t, we observe that $\frac{\mathrm{d}}{\mathrm{d}u}\Gamma^{-1}(u) = \frac{1}{g(u)} \asymp (1+u)^{1-\kappa_1}$. Integrating both sides and changing back to t variables, we get $t \asymp (1+\Gamma(t))^{2-\kappa_1}$ (or equivalently $1 + \Gamma(t) \asymp t^{1/(2-\kappa_1)}$). Finally, plugging this into the formula for γ_t and applying Corollary A.4.11, we get

$$\gamma_t \asymp \frac{\eta}{\sqrt{b^2 + \frac{2}{d}\operatorname{Tr}(K)\int_0^t F(\Gamma(s))\,\mathrm{d}s}} \asymp (1+t)^{-(1-\kappa_1)/(2-\kappa_1)}.$$

In the case of $\kappa_1 = 1$, we follow a similar procedure as for $\kappa_1 < 1$ to show that $t \asymp \Gamma(t)\log(\Gamma(t))$ for sufficiently large t. This implies $\Gamma(t) \asymp t/\log(t)$ which gives the desired result after integration. The decay rate of the risk is then immediate using Corollary A.4.11. □

Lemma A.4.13. *Let K have a spectrum that converges as $d \to \infty$ to the power law measure $\rho(\lambda) = C\lambda^{-\beta}\mathbf{1}_{(0,\lambda_{\max})}$, with $C^{-1} = \frac{\lambda_{\max}^{1-\beta}}{1-\beta}$ for some $\beta < 1$, and $\lambda_{\max} > 0$, and suppose that*

$\mathscr{D}_i^2(0) \sim \lambda_i^{-\delta}$, then $F(t) \asymp t^{-\kappa_1}$, and $\mathcal{K}(t) \asymp t^{-\kappa_2}$, with $\kappa_1 = 2-\beta-\delta$, and $\kappa_2 = 3-\beta$. In addition, $\mathcal{K}(t) \asymp t^{-\kappa_2}$, satisfies Eq. (A.66).

Proof. Following the definition in Eq. (A.52), and Eq. (A.51)

$$\begin{aligned} F(x) &= \frac{1-\beta}{2\lambda_{\max}^{1-\beta}} \int_0^{\lambda_{\max}} \lambda^{1-\beta-\delta} e^{-2\lambda x} \, \mathrm{d}\lambda \\ &= \frac{1-\beta}{2\lambda_{\max}^{1-\beta}(2x)^{2-\beta-\delta}} \int_0^{2\lambda_{\max} x} y^{1-\beta-\delta} e^{-y} \, \mathrm{d}y = \frac{1-\beta}{\lambda_{\max}^{1-\beta} 2^{3-\beta-\delta}} \frac{\gamma(2-\beta-\delta, 2\lambda_{\max} x)}{x^{2-\beta-\delta}}. \end{aligned}$$

Similarly for $\mathcal{K}$,

$$\mathcal{K}(x) = \frac{1-\beta}{\lambda_{\max}^{1-\beta}} \int_0^{\lambda_{\max}} \lambda^{2-\beta} e^{-2\lambda x} \, \mathrm{d}\lambda = \frac{1-\beta}{\lambda_{\max}^{1-\beta} 2^{3-\beta}} \frac{\gamma(3-\beta, 2\lambda_{\max} x)}{x^{3-\beta}}.$$

with $\gamma(s,z) = \int_0^z x^{s-1} e^{-x} \, \mathrm{d}x$ is the incomplete gamma function. For large z, $\gamma(s,z) \asymp \Gamma(s)$, the complete gamma function. We therefore obtain $\kappa_1 = 2-\beta-\delta$, and $\kappa_2 = 3-\beta$. Next, we show that $\mathcal{K}(x) \asymp x^{-\kappa_2}$ satisfies Eq. (A.66),

$$\begin{aligned} \int_0^t \mathcal{K}(s)\mathcal{K}(t-s) \, \mathrm{d}s &\leq \int_0^{t/2} \mathcal{K}(t)\mathcal{K}(t-s) \, \mathrm{d}s + \int_{t/2}^t \mathcal{K}(t)\mathcal{K}(t-s) \, \mathrm{d}s \\ &\leq \mathcal{K}(t/2) \left(\int_0^{t/2} \mathcal{K}(s) \, \mathrm{d}s + \int_{t/2}^t \mathcal{K}(t-s) \, \mathrm{d}s \right) \leq 2\mathcal{K}(t/2) \|\mathcal{K}\|_1 \end{aligned}$$

by the power-law assumption for $t > T$, $\mathcal{K}(t/2) \asymp \mathcal{K}(t)$ which then complete the proof. □

Proof of Proposition 2.4.4. The proof is an immediate application of Corollary A.4.12 with, $\kappa_1 = 2-\beta-\delta$ as implied by Lemma A.4.13. □

Remark A.4.14. *This includes the case $\beta = 0$, which is the uniform measure on $[0, \lambda_{\max}]$.*

A.5 Polyak Stepsize

The distance to optimality of SGD is measured say by $D^2(X) = \|X - X^\star\|^2$. Let us consider the deterministic equivalent for the distance to optimality $\mathscr{D}^2(t)$ in (2.11). Fixing $T > 0$ and any $\varepsilon \in (0, 1/2)$, we have by Theorem 2.2.1 (see also corollary A.2.5 which show concentration for large class of statistics) that $\sup_{0 \leq t \leq T} |\|X_{\lfloor td \rfloor} - X^\star\|^2 - \mathscr{D}^2(t)| \leq d^{-\varepsilon}$, w.o.p. In this way, if we want to guarantee that the distance to optimality of SGD decreases, we need $\mathrm{d}\mathscr{D}^2(t) < 0$ with the maximum decrease being $\min_{\gamma_t} \mathrm{d}\mathscr{D}^2(t)$.

As it turns out, the evolution of $\mathscr{D}^2$ is particular simple, as it solves the differential equation (derived from the ODE in (2.9))

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}^2(t) = -2\gamma_t A(\mathscr{B}(t)) + \frac{\gamma_t^2}{d}\operatorname{Tr}(K) I(\mathscr{B}(t)), \quad \begin{cases} A(\mathscr{B}) = \mathbb{E}_{a,\epsilon}[\langle x - x^\star, f'(x \oplus x^\star)\rangle], \\ I(\mathscr{B}) = \mathbb{E}_{a,\epsilon}[f'(x \oplus x^\star)^2], \quad \text{where} \\ (x \oplus x^\star) \sim N(0, \mathscr{B}). \end{cases} \tag{A.73}$$

The distance to optimality threshold, $\bar{\gamma}_t^{\mathscr{D}}$, occurs precisely when $\mathrm{d}\mathscr{D}^2 < 0$. This choice of γ makes the ODE for the distance to optimality stable. By translating the relevant deterministic quantities in $\bar{\gamma}_t^{\mathscr{D}}$ back to SGD quantities, we get

$$\bar{\mathfrak{g}}_k^{\mathscr{D}} \overset{\text{def}}{=} \frac{2\langle X_k - X^\star, \nabla\mathcal{R}(X_k)\rangle}{\frac{\operatorname{Tr}(K)}{d}\mathbb{E}_{a,\epsilon}[f'(\langle X_k, a\rangle; \langle X^\star, a\rangle, \epsilon)^2]} \text{ with the deterministic equiv. } \bar{\gamma}_t^{\mathscr{D}} = \frac{2A(\mathscr{B}(t))}{\frac{\operatorname{Tr}(K)}{d} I(\mathscr{B}(t))}. \tag{A.74}$$

A greedy learning rate that maximizes the decrease at each iteration is simply given by $\mathfrak{g}_t^{\text{Polyak}} \in \arg\min \mathrm{d}\mathscr{D}^2(t)$. This has a closed form and we call this *Polyak stepsize*[2]. Again translating this back to SGD, we have

$$\text{Polyak learning rate} \quad \mathfrak{g}_k^{\text{Polyak}} = \tfrac{1}{2}\bar{\mathfrak{g}}_k^{\mathscr{D}} \quad \text{and} \quad \text{deterministic equivalent} \quad \gamma_t^{\text{Polyak}} = \tfrac{1}{2}\bar{\gamma}_t^{\mathscr{D}}. \tag{A.75}$$

In this context, the Polyak learning rate is impractical because we do not known $X^\star$. In spite of this, we can learn some things about this learning rate as it is the natural extension of Polyak learning rate to SGD.

The quantities $A(\mathscr{B})$ and $I(\mathscr{B})$ in (A.74) and (A.75) only depend on the low-dimensional function f and thus do not carry any covariance K or d dependence. Moreover, under additional assumptions on the function such as (strong) convexity, we can bound from below $A(\mathscr{B})/I(\mathscr{B})$. Thus, in terms covariance K and d, the Polyak stepsize $\mathfrak{g}_k^{\text{Polyak}} \asymp \frac{1}{\operatorname{Tr}(K)/(d)} = \frac{1}{\text{avg. eig of } K}$.

In the case of least squares (see (2.7)), we get

$$\mathfrak{g}_k^{\text{Polyak}} = \frac{2\mathcal{R}(X_k) - \omega^2}{\frac{2\operatorname{Tr}(K)}{d}\mathcal{R}(X_k)} \text{ and on a noiseless least squares, } \quad \mathfrak{g}_k^{\text{Polyak}} = \frac{1}{\frac{\operatorname{Tr}(K)}{d}}.$$

[2]This is the idea of Polyak stepsize when the problem is deterministic.

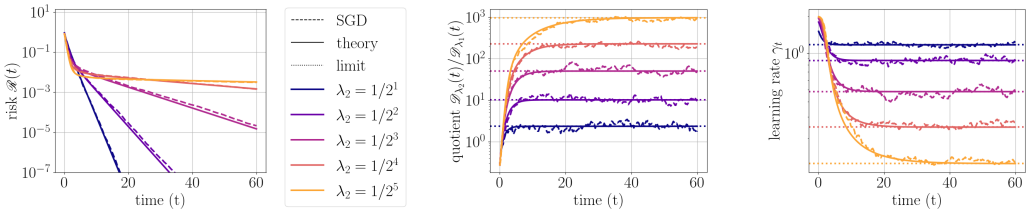

Figure A.1: **Convergence in Exact Line Search** on a noiseless least squares problem. The plot on the left illustrates the convergence of the risk function, while the center and right plots depict the convergence of the quotient $\frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)}$ and the learning rate γ_t, respectively. Further details and formulas for the limiting behavior can be found in the Appendix A.6.2. See Appendix A.8 for simulation details.

The latter gives the best fixed learning rate for a noiseless target on a LS problem (as noted in [57, 71]).

A.6 Line Search

A.6.1 General Line Search

Naturally, one can ask a similar question as in Polyak in the context of line search (i.e., decreasing risk at each iteration of SGD). First, by the structure of the risk (Assumption 2.1.3 and 2.1.4),

$$\|\nabla \mathcal{R}(X)\|^2 = m(W^T K^2 W) \quad \text{and} \quad \operatorname{Tr}(\nabla^2 \mathcal{R}(X) K) = v(K). \tag{A.76}$$

Therefore using (2.9), we have that the deterministic equivalent for $\|\nabla \mathcal{R}(X)\|^2$ is $\mathscr{M}(t) = \frac{1}{2}\sum_{i=1}^{d} m(\mathscr{V}_i(t)\lambda_i^2)$. In this case, the deterministic equivalent for the risk $\mathscr{R}$ satisfies the following ODE

$$\mathrm{d}\mathscr{R} = -\gamma_t \mathscr{M}(t)\,\mathrm{d}t + \frac{\gamma_t^2}{d} v(K) I(\mathscr{B}(t)). \tag{A.77}$$

From this, we get an immediate learning rate (stability) threshold for the risk, that is, $\bar{\mathfrak{g}}_k^{\mathscr{R}}$ is the largest learning rate for which SGD is guaranteed to decrease at each iteration, i.e., when the deterministic equivalent of $\mathscr{R}$ satisfies $\mathrm{d}\mathscr{R} < 0$ or equivalently after translating relevant

terms into SGD quantities

$$\text{risk threshold} \quad \bar{\mathfrak{g}}_k^{\mathscr{R}} = \frac{\|\nabla \mathcal{R}(X_k)\|^2}{\frac{\operatorname{Tr}(K\nabla^2\mathcal{R}(X_k))}{d} I(W_k^T K W_k)} \text{ and deterministic equiv} \quad \bar{\gamma}_t^{\mathscr{R}} = \frac{\mathscr{M}(t)}{\frac{v(K)}{d} I(\mathscr{B}(t))}. \tag{A.78}$$

The greediest approach, which we call *exact line search*, would choose the learning rate such that $\gamma_t^{\text{line}} \in \arg\min_\gamma \mathrm{d}\mathscr{R}$. In this case, we get

$$\mathfrak{g}_k^{\text{line}} = \tfrac{1}{2}\mathfrak{g}_k^{\mathscr{R}} \quad \text{and} \quad \text{deterministic equiv} \quad \gamma_t^{\text{line}} = \tfrac{1}{2}\gamma_t^{\mathscr{R}}.$$

A.6.2 Line Search on least squares

In this section, we provide a proof of Proposition 2.3.1, but, we show more than this including the exact limiting value for γ_t.

Proposition A.6.1. *Consider the noiseless ($\omega = 0$) least squares problem (2.7) . Then the learning rate is always lower bounded by*

$$\frac{\lambda_{\min}(K)}{\frac{1}{d}\operatorname{Tr}(K^2)} \leq \gamma_t^{\text{line}} \quad \textit{for all } t \geq 0.$$

Moreover, suppose K has only two distinct eigenvalues $\lambda_1 > \lambda_2 > 0$, i.e., K has $d/2$ eigenvalues equal to λ_1 eigenvalues and $d/2$ eigenvalues equal to λ_2. In this context, the exact limiting value of γ_t^{line} is given by

$$\lim_{k\to\infty} \gamma_t^{\text{line}} = \frac{2\left(\lambda_1^2 + \lambda_2^2 x\right)}{\left(\lambda_1 + \lambda_2 x\right)\left(\lambda_1^2 + \lambda_2^2\right)}, \tag{A.79}$$

where x is the positive real root of the second-degree polynomial

$$\mathcal{P}(x) = \lambda_1\lambda_2(x+1)(\lambda_2 x - \lambda_1) + (\lambda_2 - \lambda_1)^3 x. \tag{A.80}$$

This leads to

$$\frac{\lambda_{\min}(K)}{\frac{1}{d}\operatorname{Tr}(K^2)} \leq \lim_{t\to\infty} \gamma_t^{\text{line}} \leq \frac{2\lambda_{\min}(K)}{\frac{1}{d}\operatorname{Tr}(K^2)}. \tag{A.81}$$

Proof. We establish the inequality

$$\frac{\lambda_{\min}(K)}{\frac{1}{d}\operatorname{Tr}(K^2)} \leq \gamma_t^{\text{line}} \quad \text{for all } t \geq 0$$

by observing

$$\frac{1}{d}\sum_{i=1}^{d}\lambda_i^2\mathscr{D}_i^2(t) \geq 2\lambda_{\min}(K)\frac{1}{2d}\sum_{i=1}^{d}\lambda_i\mathscr{D}_i^2(t) = 2\lambda_{\min}(K)\mathscr{R}(t).$$

Now let us consider $K \sim \frac{1}{2}\lambda_1 + \frac{1}{2}\lambda_2$ for $\lambda_1 > \lambda_2 > 0$.

We define $\mathscr{D}_\lambda(t) \stackrel{\text{def}}{=} \sum_{\lambda_i=\lambda}^{d}\mathscr{D}_i^2(t)$. Utilizing the ODEs in (2.9), we derive

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}_\lambda(t) = -2\gamma_t\lambda\mathscr{D}_\lambda(t) + 2\gamma_t^2\lambda \times \left|\{\lambda = \lambda_i\}_{i=1}^{d}\right| \times \mathscr{R}(t)$$

for each distinct eigenvalue λ of K. Here $\left|\{\lambda = \lambda_i\}_{i=1}^{d}\right|$ is the number of eigenvalues of K that are equal to λ. It immediately follows by our construction of K that $\left|\{\lambda = \lambda_i\}_{i=1}^{d}\right| = \frac{d}{2}$. Thus, we establish the following system of ODEs

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}_{\lambda_1}(t) = -2\gamma_t\lambda_1\mathscr{D}_{\lambda_1}(t) + d\gamma_t^2\lambda_1\mathscr{R}(t) \\ \frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}_{\lambda_2}(t) = -2\gamma_t\lambda_2\mathscr{D}_{\lambda_2}(t) + d\gamma_t^2\lambda_2\mathscr{R}(t) \end{cases} \tag{A.82}$$

where $\mathscr{R}(t) = \frac{1}{2d}\left(\lambda_1\mathscr{D}_{\lambda_1}(t) + \lambda_2\mathscr{D}_{\lambda_2}(t)\right)$ and $\gamma_t^{\text{line}} = \frac{2\left(\lambda_1^2\mathscr{D}_{\lambda_1}(t)+\lambda_2^2\mathscr{D}_{\lambda_2}(t)\right)}{\left(\lambda_1\mathscr{D}_{\lambda_1}(t)+\lambda_2\mathscr{D}_{\lambda_2}(t)\right)\left(\lambda_1^2+\lambda_2^2\right)}$.

Since $\mathscr{D}_{\lambda_2}(t) \geq 0$ and $\lambda_1 > \lambda_2 > 0$, we infer that $\mathscr{R}(t) = \frac{1}{2d}\left(\lambda_1\mathscr{D}_{\lambda_1}(t) + \lambda_2\mathscr{D}_{\lambda_2}(t)\right) \geq \frac{1}{2d}\lambda_1\mathscr{D}_{\lambda_1}(t) \geq 0$. The structure of the exact line search algorithm ensures $\lim_{t\to\infty}\mathscr{R}(t) = 0$, hence $\lim_{t\to\infty}\mathscr{D}_{\lambda_1}(t) = 0$. Similarly, we deduce $\lim_{t\to\infty}\mathscr{D}_{\lambda_2}(t) = 0$.

By applying L'Hôpital's rule and substituting the expressions for γ_t^{line} and $\mathscr{R}(t)$ in terms of $\mathscr{D}_{\lambda_1}(t)$ and $\mathscr{D}_{\lambda_2}(t)$, we derive

$$\begin{aligned}
\lim_{t\to\infty}\frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)} &= \lim_{t\to\infty}\frac{\mathrm{d}\mathscr{D}_{\lambda_2}(t)}{\mathrm{d}\mathscr{D}_{\lambda_1}(t)} \\
&= \lim_{t\to\infty}\frac{-2\gamma_t\lambda_2\,\mathscr{D}_{\lambda_2}(t) + d\gamma_t^2\lambda_2\mathscr{R}(t)}{-2\gamma_t\lambda_1\,\mathscr{D}_{\lambda_1}(t) + d\gamma_t^2\lambda_1\mathscr{R}(t)} \\
&= \lim_{t\to\infty}\frac{-2\lambda_2\,\mathscr{D}_{\lambda_2}(t) + d\gamma_t\lambda_2\mathscr{R}(t)}{-2\lambda_1\,\mathscr{D}_{\lambda_1}(t) + d\gamma_t\lambda_1\mathscr{R}(t)} \\
&= \lim_{t\to\infty}\frac{\gamma_t\frac{\lambda_1\lambda_2}{2}\mathscr{D}_{\lambda_1}(t) + \lambda_2\mathscr{D}_{\lambda_2}(t)\left(\gamma_t\frac{\lambda_2}{2} - 2\right)}{\gamma_t\frac{\lambda_1\lambda_2}{2}\mathscr{D}_{\lambda_2}(t) + \lambda_1\mathscr{D}_{\lambda_1}(t)\left(\gamma_t\frac{\lambda_1}{2} - 2\right)} \\
&= \lim_{t\to\infty}\frac{\mathscr{D}_{\lambda_1}(t)^2\lambda_1^3\lambda_2 + \mathscr{D}_{\lambda_1}(t)\mathscr{D}_{\lambda_2}(t)(-\lambda_1\lambda_2^3 + \lambda_1^2\lambda_2^2 - 2\lambda_1^3\lambda_2) + \mathscr{D}_{\lambda_2}(t)^2(-\lambda_2^4 - 2\lambda_1^2\lambda_2^2)}{\mathscr{D}_{\lambda_1}(t)^2(-\lambda_1^4 - 2\lambda_1^2\lambda_2^2) + \mathscr{D}_{\lambda_1}(t)\mathscr{D}_{\lambda_2}(t)(-\lambda_1^3\lambda_2 + \lambda_1^2\lambda_2^2 - 2\lambda_1\lambda_2^3) + \mathscr{D}_{\lambda_2}(t)^2\lambda_1\lambda_2^3} \\
&= \frac{\lambda_1^3\lambda_2 + \lim_{t\to\infty}\frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)}(-\lambda_1\lambda_2^3 + \lambda_1^2\lambda_2^2 - 2\lambda_1^3\lambda_2) + \left(\lim_{t\to\infty}\frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)}\right)^2(-\lambda_2^4 - 2\lambda_1^2\lambda_2^2)}{(-\lambda_1^4 - 2\lambda_1^2\lambda_2^2) + \lim_{t\to\infty}\frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)}(-\lambda_1^3\lambda_2 + \lambda_1^2\lambda_2^2 - 2\lambda_1\lambda_2^3) + \left(\lim_{t\to\infty}\frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)}\right)^2\lambda_1\lambda_2^3}.
\end{aligned}$$

Therefore, $\lim_{t\to\infty} \frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)}$ is the positive real root of the second-degree polynomial

$$\mathcal{P}(x) = \lambda_1\lambda_2(x+1)(\lambda_2 x - \lambda_1) + (\lambda_2 - \lambda_1)^3 x. \tag{A.83}$$

Solving for $x > 0$, we derive the explicit formula

$$\lim_{t\to\infty} \frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)} = \frac{\lambda_1^3 - 2\lambda_1^2\lambda_2 + 2\lambda_1\lambda_2^2 - \lambda_2^3 + \sqrt{\lambda_1^6 - 4\lambda_1^5\lambda_2 + 8\lambda_1^4\lambda_2^2 - 6\lambda_1^3\lambda_2^3 + 8\lambda_1^2\lambda_2^4 - 4\lambda_1\lambda_2^5 + \lambda_2^6}}{2\lambda_1\lambda_2^2}. \tag{A.84}$$

Given

$$\gamma_t^{\text{line}} = \frac{2\left(\lambda_1^2\mathscr{D}_{\lambda_1}(t) + \lambda_2^2\mathscr{D}_{\lambda_2}(t)\right)}{\left(\lambda_1\mathscr{D}_{\lambda_1}(t) + \lambda_2\mathscr{D}_{\lambda_2}(t)\right)\left(\lambda_1^2+\lambda_2^2\right)} = \frac{2\left(\lambda_1^2 + \lambda_2^2\frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)}\right)}{\left(\lambda_1 + \lambda_2\frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)}\right)\left(\lambda_1^2+\lambda_2^2\right)}, \tag{A.85}$$

we have

$$\lim_{t\to\infty} \gamma_t^{\text{line}} = \frac{2\left(\lambda_1^2 + \lambda_2^2\lim_{t\to\infty}\frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)}\right)}{\left(\lambda_1 + \lambda_2\lim_{t\to\infty}\frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)}\right)\left(\lambda_1^2+\lambda_2^2\right)}. \tag{A.86}$$

By substituting (A.84), we get

$$\lim_{t\to\infty} \gamma_t^{\text{line}} = \frac{\lambda_1^3 + 2\lambda_1^2\lambda_2 + 2\lambda_1\lambda_2^2 + \lambda_2^3 - \sqrt{\lambda_1^6 - 4\lambda_1^5\lambda_2 + 8\lambda_1^4\lambda_2^2 - 6\lambda_1^3\lambda_2^3 + 8\lambda_1^2\lambda_2^4 - 4\lambda_1\lambda_2^5 + \lambda_2^6}}{\left(\lambda_1^2+\lambda_2^2\right)^2}. \tag{A.87}$$

A direct calculation reveals that $\lambda_1 > \lambda_2 > 0$ implies $\lim_{t\to\infty}\gamma_t^{\text{line}} \le \frac{2\lambda_{\min}(K)}{\frac{1}{d}\operatorname{Tr}(K^2)}$. □

Remark A.6.2. *For the scenario where K has an arbitrary number n of distinct eigenvalues, equation (2.13) remains valid. The proof parallels the one outlined above. However, in this case, the expression for $\lim_{k\to\infty}\mathfrak{g}_k$ is given by*

$$\lim_{k\to\infty}\mathfrak{g}_k = \frac{n\left(\lambda_1^2 + \lambda_2^2 x_1 + \cdots + \lambda_n^2 x_{n-1}\right)}{\left(\lambda_1 + \lambda_2 x_1 + \ldots \lambda_n x_{n-1}\right)\left(\lambda_1^2 + \cdots + \lambda_n^2\right)}, \tag{A.88}$$

where $x_1, \ldots, x_{n-1} > 0$ satisfy a more intricate coupled system of $n-1$ equations.

A.7 Examples

Any single index model with α-pseudo Lipschitz ($\alpha \le 1$) activation function is covered by our SGD+AL theory. In this section, we provide key learning problems within this family of models.

A.7.1 Binary logistic regression

We consider a binary logistic regression problem with $\epsilon = 0$ where we are trying to classify two classes. We will follow a Student-Teacher model, in which there exists a true vector $X^\star$ to be the true direction such that possible labels are, $y = \frac{\exp(\langle X^\star, a\rangle)}{\exp(\langle X^\star, a\rangle)+1}$. or $1-y$. In order to classify the data we minimize the KL-divergence between the label y and our estimate defined by the above formula,

$$\mathcal{R}(X) = \mathbb{E}_a\left[-\langle X, a\rangle \cdot \frac{\exp(\langle X^\star, a\rangle)}{\exp(\langle X^\star, a\rangle)+1} + \log\left(\exp(\langle X, a\rangle)+1\right)\right]. \tag{A.89}$$

To study the ODE dynamics of SGD in Eq. (2.9) one needs the deterministic risk $h(B)$, and $I(B) = \mathbb{E}_a[f'(\langle X, a\rangle, \langle X^\star, a\rangle)^2]$, with $B = W^T K W$. Following the computation in Appendix D example D.4 in [21] we obtain that

$$h(B) = -B_{21}\,\mathbb{E}_z\left[\frac{\exp(\sqrt{B_{22}}\cdot z)}{(1+\exp(\sqrt{B_{22}}\cdot z))^2}\right] + \mathbb{E}_w\left[\log(\exp(w\sqrt{B_{11}})+1)\right], \tag{A.90}$$

where $z, w \sim \mathcal{N}(0,1)$. The I function can also be computed explicitly by solving the following Gaussian integral, where we define $g(x) \overset{\text{def}}{=} \frac{\exp(x)}{1+\exp(y)}$

$$I(B) = \frac{1}{2\pi\sqrt{\det(B)}}\int_{\mathbb{R}^2}(g(x)-g(y))^2 \exp\left(-\frac{1}{2}\begin{pmatrix}x\\y\end{pmatrix}^T B^{-1}\begin{pmatrix}x\\y\end{pmatrix}\right)\mathrm{d}x\,\mathrm{d}y. \tag{A.91}$$

We note that, the logistic regression is (μ,θ)–RSI with $\mu = \frac{1}{\ell e^{\sqrt{4\theta}}}$ see section 2.2 in [21]. Its Lipschitz constant is $\hat{L}(f) = 1$. Using Proposition A.4.4 one can derive a lower bound on the limiting learning of AdaGrad Norm.

For more details and more examples, see [21].

A.7.2 CIFAR 5m

Finally, we include an example that uses real-world data, that is, the CIFAR 5m dataset [65]. Our theory does not explicitly deal with non-Gaussian distributions, but we find that the theoretical risk curves generalize cleanly to that case.

As we are now working with discrete data points rather than a distribution, the learning setup, while closely analogous to what was presented earlier, has some slight differences.

We start with a subset of the data consisting of n grayscale images, each of which is 32×32 pixels, that is, $A \in \mathbb{R}^{n \times 1024}$. We fill a vector $b \in \mathbb{R}^n$ with the corresponding labels (0 for an image of a plane, 1 for an image of a car.) We then randomly choose a matrix $W \in \mathbb{R}^{1024 \times d}$ with i.i.d. Gaussian entries to generate the features $F = \text{relu}(AW)$. We want to use least squares to predict the label from the features, i.e., find

$$\arg\min_{X \in \mathbb{R}^d} \left\{ \mathcal{R}(X) := \frac{1}{2n} \|FX - b\|^2 = \frac{1}{2n} \sum_{i=1}^{n} (f_i \cdot X - b_i)^2 \right\}, \tag{A.92}$$

where f_i is the ith row of F. The SGD we now consider is

$$X_{k+1} = X_k - \gamma_k \left(f_{i_{k+1}} \cdot X - b_{i_{k+1}} \right) f_{i_{k+1}}, \quad \{i_k\} \text{ iid Unif}(\{1, 2, \cdots, n\}), \tag{A.93}$$

where γ_k is the usual AdaGrad-Norm stepsize, as in (A.1). Our empirical covariance matrix K (remembering that f_i is a row vector) is then

$$K = \mathbb{E}_{i \in [n], j \in [n]} \left[f_i^\top f_j \right] = \frac{1}{n} F^\top F. \tag{A.94}$$

We now use (A.50), with the AdaGrad-Norm stepsize, to numerically simulate the SGD loss, which we then compare to the actual loss. Our theory matches empirical results very closely.

A.8 Numerical simulation details

Here we provide more details for the figures that appear in the main paper.

Figure 2.1: Concentration learning rate and risk for AdaGrad-Norm on a least squares problem with label noise $\omega = 1$ (left) and on a logistic regression problem with no label noise (right). For logistic, see Section A.7. 30 runs of AdaGrad-Norm with parameters $b = 1$ and $\eta = 1$ for each d; $X^\star \sim \mathcal{N}(0, I_d/d)$, $X_0 = 0$, and $K = I_d$. The shaded region represents a 90% confidence interval for the SGD runs. As the dimension increases, the risk and stepsize both concentrate around a deterministic limit (red). The deterministic limit is described by an ODE in Theorem 2.2.1. The initial loss increase in the least squares problem suggesting that the learning rate was initially too high, but AdaGrad-Norm naturally adapts and still the loss converges. Our ODEs predict this behavior.

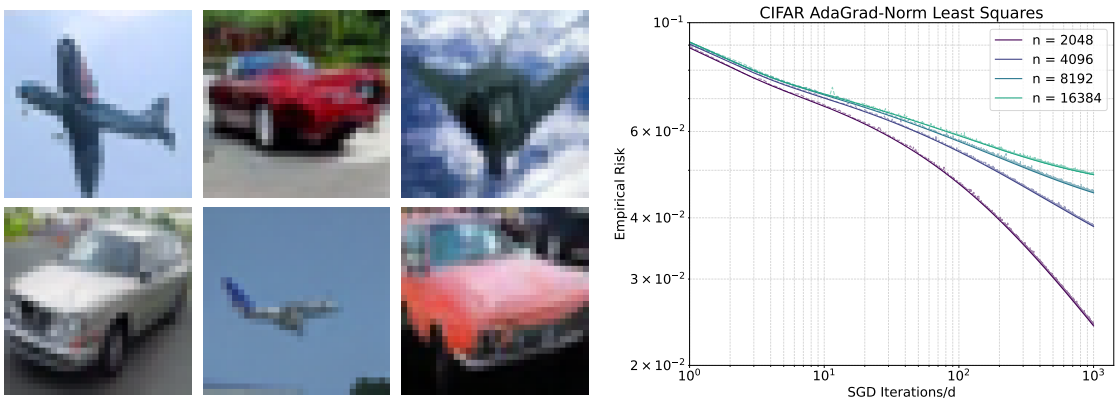

Figure A.2: **Predicting the training dynamics on a real dataset, CIFAR-5m [65], using multi-pass AdaGrad-Norm.** This suggests the theory extends beyond Gaussian data and one-pass. Note that the curves look significantly different for different n; smaller values of n lead to an overparametrized problem, allowing least squares to memorize datapoints, whereas for larger n, least squares must learn a general function mapping images of cars and airplanes to their respective labels.

Figure 2.2: Comparison for Exact Line Search and Polyak Stepsize on a noiseless least squares problem. The left plot illustrates the convergence of the risk function, while the right plot depicts the convergence of the quotient $\gamma_t / \frac{\lambda_{\min}(K)}{\frac{1}{d}\operatorname{Tr}(K^2)}$ for Polyak stepsize and exact line search. Both ODE theory and SGD results are presented, showing a close agreement between the two approaches. The covariance matrix K is generated such that the eigenvalues follow the expression $\lambda_i(K) = \sqrt{\frac{d}{\sum_{i=1}^{d}(\frac{i}{d+1})^{-2/s}}} \cdot \left(\frac{i}{d+1}\right)^{-1/s}, \quad i = 1, \ldots, d$, where $s > 2$ is a constant. As s approaches 2, the spectrum becomes more spread out, resulting in larger values of $\frac{1}{d}\operatorname{Tr}(K^2)$. Larger values of s correspond to smaller spreads in the spectrum. Additionally, $\operatorname{Tr}(K)/d = 1$ for all s. Both plots highlight the implication of equation (2.13) in high-dimensional settings, where a broader spectrum of K results in $\frac{\lambda_{\min}(K)}{\frac{1}{d}\operatorname{Tr}(K^2)} \ll \frac{1}{\frac{1}{d}\operatorname{Tr}(K)}$, indicating slower risk convergence and poorer performance of exact line search (unmarked) as it deviates from the Polyak stepsize (circle markers). The gray shaded region demonstrates that equation (2.13) is satisfied.

Figure 2.3: Quantities effecting AdaGrad-Norm learning rate. *(left):* The effect of adding noise to the targets ($\omega = 1.0$) to the risk (left axis) and learning rate (right axis). Ran AdaGrad-Norm($b = 1.0, \eta = 2.5$) on least squares problem with $d = 500$. X_0, $X^\star \sim \mathcal{N}(0, I_d/d)$. A single run of the SGD (solid line purple) matches exactly the prediction (ODE, teal). The shaded region represents 10 runs of SGD with 90% confidence interval. The learning rate decays at the exact predicted rate of $\frac{\eta}{\sqrt{b^2 + \frac{\operatorname{Tr} K\omega^2}{d} t}}$. Depicted is $\frac{\text{learning rate}}{\text{asymptotic}}$ so it approaches 1. *(center, right)*: Noiseless least squares setting ($\omega = 0$). *(center)*: Prop. 2.4.2 predicts the avg. eig of K ($\operatorname{Tr}(K)/d$) as compared with $\lambda_{\max}$ effects the $\lim_{k\to\infty} \mathfrak{g}_k$. Indeed, this is true. We varied the $\kappa = \lambda_{\max}/\lambda_{\min}$ while keeping the $\operatorname{Tr}(K)/d$ and all other parameters fixed. All the learning rates behave identically verifying our theory about the effect of $\operatorname{Tr}(K)/d$ on learning rates. *(right)*: Varying the learning rate of AdaGrad norm by $\|X_0 - X^\star\|^2$; our predictions (dashed) match and we see the inverse relationship predicted by Prop. 2.4.2. See Appendix A.4 for details. Additionally, we did the following.

- **Center plot**: AdaGrad with $b = 0.5$, $\eta = 2.5$ is run on the least squares problem with $d = 1000$ and $X_0, X^\star \sim \frac{1}{\sqrt{d}}\mathcal{N}(0, I)$. The covariance matrix K is generated so that the eigenvalues are

$$\lambda_i(K) = \sqrt{\frac{d}{\sum_{i=1}^{d} \left(\frac{i}{d+1}\right)^{-2/s}}} \cdot \left(\frac{i}{d+1}\right)^{-1/s}, \quad i = 1, \ldots, d.$$

  The constant $s > 2$. When s is near 2, the spectrum is more spread out, i.e., $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$ is large. Larger values of s mean smaller the spreads. Moreover $\operatorname{Tr}(K)/d = 1$ for all s. In the simulations, we used $s \in \{2.1, 3.0, 3.5, 4.0, 5.5\}$ and recorded the condition number κ.

- **Right plot**: Ran AdaGrad with $b = 0.5$, $\eta = 2.5$ on the least squares problem with $d = 1000$. $X^\star = 0$ and $X_0 \sim \sqrt{\frac{p}{d}}\mathcal{N}(0, I)$ where $p \in \{1, 2, 4, 8, 16\}$. In this way, $\|X_0 - X^\star\|^2 = p$.

Figure 2.4: Power law covariance in AdaGrad Norm on a least squares problem. Generated covariance K such that the density of eigenvalues are $(1 - \beta)\lambda^{-\beta}$ where $\beta = 0.2$

and set $X_0 = 0$. Choose $(X_i^\star)_{i=1}^d = (\lambda_i^{-\delta/2})_{i=1}^d$ where λ_i is the i-th eigenvalue of K and we vary $\delta \in (0, 1.8)$ so that $0 < \delta + \beta \leq 2$. Setting of Prop. 2.4.4.

Figure A.1: Convergence in Exact Line Search on a noiseless least squares problem. The plot on the left illustrates the convergence of the risk function, while the center and right plots depict the convergence of the quotient $\frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)}$ and the learning rate γ_t, respectively. Predictions from ODE theory are compared with results obtained from SGD, demonstrating close agreement between the two approaches. Initialization was performed randomly, with $X_0 \sim \mathcal{N}(0, I_d/d)$ and $X^\star \sim \frac{1}{\sqrt{d}}\mathbf{1}$, where $d = 400$. The covariance matrix K has two distinct eigenvalues $\lambda_1 = 1 > \lambda_2 > 0$, and was constructed by specifying the spectrum, with λ_i sampled from a discrete uniform distribution $\mathcal{U}\{1, \lambda_2\}$ for $i = 1, \ldots, d = 400$, and setting $K = \mathrm{diag}(\lambda_i : i = 1, \ldots, 400)$. Further details and formulas for the limiting behavior can be found in the Appendix A.6.2.

Figure A.2 Convergence on CIFAR 5m [65]. We train a classifier to distinguish between images of airplanes and cars. Fix $d = 2000$. Then for multiple values of n, we run AdaGrad-Norm with initialization $X_0 = 0$, $b = 0.1$ and $\eta = 5$, randomly sampling a datapoint from F at every step. Details of the setup can be found in Appendix A.7.2.

Appendix B

APPENDIX FOR CHAPTER 3

Outline of the paper. The remainder of the article is structured as follows:

1. Appendix B.1 collects notation and auxiliary tools used throughout the proofs. It fixes our conventions for complex-valued tensor products, coordinate contractions, and tensor norms; records derivative computations for the special functions ψ, Ψ, and S appearing in the proof of Theorem 3.3.7; and recalls the concentration and pseudo-Lipschitz estimates used in the probabilistic arguments.

2. Appendix B.2 develops the main dynamical argument. It introduces the partial integro-differential equation (B.6) and the notion of approximate solutions, proves a stability principle for these solutions, and applies it to the resolvent statistic S along SGD and homogenized SGD. This yields Theorem B.2.7; the result is then transferred to general statistics satisfying Assumption 3.3.6, yielding Theorem B.2.9 and its corollaries.

3. Appendix B.3 proves that the resolvent statistics $t \mapsto S(x_{\lfloor td \rfloor}, \cdot)$ and $t \mapsto S(\mathcal{X}_t, \cdot)$, associated respectively with SGD and homogenized SGD (3.14), are approximate solutions of the partial integro-differential equation (B.6). The proof uses Doob/Itô decompositions, a net argument over the fixed contour, and martingale and Taylor-error bounds.

4. Appendix B.4 studies the homogenized SDE in the isotropic squared-parameterization setting. It introduces an empirical entropy adapted to the coordinatewise dynamics, proves an exact entropy SDE and barrier estimates, and uses an exponential supermartingale argument to obtain high-probability global existence and exponential decay of the risk. The section also records consequences such as risk integrability and uniform separation from the saddle.

5. Appendix B.5 presents key examples illustrating our concentration risk framework.

6. Appendix B.6 provides additional details on the numerical simulations used to produce the figures in the main text.

B.1 Notation and Preliminaries

This appendix collects notation and auxiliary facts used throughout the proofs. We first fix basic notation and conventions for tensor products, coordinate contractions, and tensor norms for complex-valued tensors. We then record derivative identities for ψ, Ψ, and S, which are used in the derivation of the limiting dynamics. Finally, we recall the concentration and pseudo-Lipschitz estimates needed for the probabilistic bounds.

Basic notation. Throughout the paper, e_i denotes the i-th canonical coordinate vector, with its dimension inferred from context. Additionally, the matrix E_{ij} is defined as the outer product $e_i \cdot e_j^\top$.

For $x = (u, v)$, recall that each coordinate of the inner function ψ is given by

$$\psi_i(u, v) = \begin{pmatrix} u_i & v_i \end{pmatrix} \mathcal{Q} \begin{pmatrix} u_i \\ v_i \end{pmatrix} + l^\top \begin{pmatrix} u_i \\ v_i \end{pmatrix} + c,$$

where $\mathcal{Q} = \begin{pmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{pmatrix} \in \mathbb{R}^{2\times 2}$ is symmetric, $l = (l_1, l_2)^\top \in \mathbb{R}^2$, and $c \in \mathbb{R}$.

Equivalently, we have

$$\psi_i(u, v) = q_{11}u_i^2 + 2q_{12}u_iv_i + q_{22}v_i^2 + l_1u_i + l_2v_i + c.$$

B.1.1 Tensor Products and Contractions

We briefly fix the tensor-contraction conventions used throughout the proofs. The optimization variables are real, but several quantities, such as $S(x, z)$, are complex-valued because of the spectral parameter z. Consequently, we will use complex-valued tensors whose derivative directions are indexed by real coordinate spaces such as $\mathbb{R}^d$.

We order tensor indices so that derivative or ambient coordinates appear first, and observable coordinates appear last. For example, since $S(x,z) \in \mathbb{C}^{3\times 3}$, we identify

$$\nabla_u S(x,z) \in \mathbb{C}^{d\times 3\times 3} \cong \mathbb{C}^d \otimes \mathbb{C}^{3\times 3},$$

where the first index corresponds to the derivative with respect to u.

We use $\langle \cdot, \cdot \rangle_{\mathbb{R}^d}$ to denote contraction over the indicated d-dimensional coordinate. This contraction is bilinear and is defined by summing over the contracted coordinate, with no complex conjugation. Thus, if

$$A \in \mathbb{C}^{d\times m_1 \times \cdots \times m_k}, \qquad B \in \mathbb{C}^{d\times n_1 \times \cdots \times n_\ell},$$

then the expression

$$\langle A, B \rangle_{\mathbb{R}^d} \in \mathbb{C}^{m_1 \times \cdots \times m_k \times n_1 \times \cdots \times n_\ell}$$

is defined by

$$\big(\langle A, B \rangle_{\mathbb{R}^d}\big)_{\alpha_1,\ldots,\alpha_k,\beta_1,\ldots,\beta_\ell} := \sum_{i=1}^{d} A_{i,\alpha_1,\ldots,\alpha_k} B_{i,\beta_1,\ldots,\beta_\ell}.$$

In words, the contracted axes are summed over, and the uncontracted axes of the first tensor are listed before the uncontracted axes of the second tensor.

For instance, if

$$\nabla_u S(x,z) \in \mathbb{C}^{d\times 3\times 3}, \qquad \nabla_u \mathcal{R}(x) \in \mathbb{R}^d,$$

then the matrix

$$\langle \nabla_u S(x,z), \nabla_u \mathcal{R}(x) \rangle_{\mathbb{R}^d} \in \mathbb{C}^{3\times 3}$$

has entries

$$\big[\langle \nabla_u S(x,z), \nabla_u \mathcal{R}(x) \rangle_{\mathbb{R}^d}\big]_{ab} = \sum_{i=1}^{d} \partial_{u_i} S_{ab}(x,z)\, \partial_{u_i} \mathcal{R}(x).$$

Similarly, if

$$\nabla_u^2 S(x,z) \in \mathbb{C}^{d\times d\times 3\times 3}, \qquad M \in \mathbb{R}^{d\times d},$$

then contraction over the two derivative coordinates gives

$$\langle \nabla_u^2 S(x,z), M \rangle_{\mathbb{R}^{d\times d}} \in \mathbb{C}^{3\times 3},$$

with entries

$$\big[\langle \nabla_u^2 S(x,z), M\rangle_{\mathbb{R}^{d\times d}}\big]_{ab} = \sum_{i,j=1}^{d} \partial_{u_i}\partial_{u_j} S_{ab}(x,z)\, M_{ij}.$$

We reserve the unsubscripted notation $\langle\cdot,\cdot\rangle$ for the full Hilbert–Schmidt tensor inner product. Thus, for complex matrices or tensors of the same shape,

$$\langle A, B\rangle := \sum_{\alpha} \overline{A_\alpha} B_\alpha,$$

where α ranges over all tensor indices. In particular, for matrices,

$$\langle A, B\rangle = \mathrm{Tr}(A^* B),$$

where $A^* = \overline{A}^{\top}$ denotes the conjugate transpose. The corresponding unsubscripted norm is

$$\|A\| := \sqrt{\langle A, A\rangle}.$$

In contrast, a subscript on the bracket indicates a partial contraction over the specified coordinate space. For example, $\langle A, B\rangle_{\mathbb{R}^d}$ denotes contraction over the $\mathbb{R}^d$ coordinate. These partial contractions are bilinear coordinate contractions used in derivative and generator computations, and no complex conjugation is applied unless explicitly stated. Thus $\langle\cdot,\cdot\rangle_{\mathbb{R}^d}$ should be read as a contraction label, not as a Hermitian inner product on $\mathbb{C}^d$.

B.1.2 Norms on Tensors

We now define the tensor norms used throughout the paper. Unless otherwise specified, $\|\cdot\|$ denotes the Hilbert-space norm: the Euclidean norm for vectors and the Hilbert–Schmidt/Frobenius norm for matrices and tensors. As above, $\langle A, B\rangle$ denotes the full Hilbert–Schmidt inner product, and

$$\|A\| := \sqrt{\langle A, A\rangle}$$

denotes the associated Hilbert-space norm. Thus, for matrices, $\|A\|$ is the Frobenius norm unless another subscript, such as $\|\cdot\|_{\mathrm{op}}$, is specified. We use subscripts such as $\|\cdot\|_{\mathrm{op}}$, $\|\cdot\|_{\sigma}$, and $\|\cdot\|_{*}$ for the operator, injective, and nuclear norms, respectively.

For a matrix $A \in \mathbb{C}^{d\times d}$, the operator norm admits the variational representation

$$\|A\|_{\mathrm{op}} = \sup_{\substack{\|y\|_2=1\\ \|z\|_2=1}} \big|y^* A z\big|.$$

The injective tensor norm is the natural higher-order analogue of this formula. Let

$$A \in V_1 \otimes V_2 \otimes \cdots \otimes V_k.$$

We define

$$\|A\|_\sigma := \sup_{\substack{\|y_i\|_{V_i}=1 \\ i=1,\dots,k}} |\langle A,\, y_1 \otimes y_2 \otimes \cdots \otimes y_k \rangle| \,.$$

Equivalently, $\|A\|_\sigma$ is the largest correlation of A with a unit simple tensor. This norm is also known as the *injective tensor norm*. In the case $k = 2$, it reduces to the usual operator norm, up to the standard identification of matrices with order-two tensors.

By Cauchy–Schwarz, the Hilbert-space norm also has the variational representation

$$\|A\| = \sup_{\|B\| \leqslant 1} |\langle A, B \rangle| \,,$$

where the supremum is over tensors B of the same shape as A, and $\|B\| = \sqrt{\langle B, B \rangle}$ is the Hilbert–Schmidt norm.

Finally, we define the nuclear norm as the dual norm of the injective norm:

$$\|A\|_* := \sup_{\|B\|_\sigma \leqslant 1} |\langle A, B \rangle| \,.$$

For order-two tensors this agrees with the usual matrix nuclear norm. For higher-order tensors, this is the tensor nuclear norm, equivalently the projective tensor norm.

These norms satisfy the chain of inequalities

$$\|A\|_\sigma \leqslant \|A\| \leqslant \|A\|_*. \tag{B.1}$$

Indeed, the first inequality follows because unit simple tensors have Hilbert–Schmidt norm one. The second follows from the variational formula above and the fact that $\|B\|_\sigma \leq \|B\|$ for every tensor B.

B.1.3 Derivative Identities for Special Statistics

We record the derivative identities used in the derivation of the limiting dynamics. All tensor-valued derivatives are interpreted using the contraction conventions from the previous subsection.

Lemma B.1.1 (Jacobian of ψ for diagonal linear networks)**.** *Let ψ be as in Assumption 3.1.8. Then its Jacobian with respect to $x = (u, v)$ is*

$$\nabla\psi(u,v) = \begin{bmatrix} \nabla_u\psi(u,v) & \nabla_v\psi(u,v) \end{bmatrix} \in \mathbb{R}^{d\times 2d},$$

where

$$\nabla_u\psi(u,v) = 2q_{11}\,\mathrm{diag}(u) + 2q_{12}\,\mathrm{diag}(v) + l_1 I_d,$$

and

$$\nabla_v\psi(u,v) = 2q_{12}\,\mathrm{diag}(u) + 2q_{22}\,\mathrm{diag}(v) + l_2 I_d.$$

Lemma B.1.2 (Gradient of Ψ)**.** *Let*

$$r := \frac{1}{\sqrt{d}} \begin{pmatrix} \psi(x) \\ \beta^* \end{pmatrix}^\top a \in \mathbb{R}^2,$$

and define $\Psi(x; a) := f(r)$. Then we hav the expression

$$\nabla_x\Psi(x;a) = \frac{1}{\sqrt{d}}\,\nabla_{r_1} f(r)\,(\nabla\psi(x))^\top a \in \mathbb{R}^{2d},$$

where $\nabla_{r_1} f$ denotes the partial derivative of f with respect to its first coordinate.

Lemma B.1.3 (Derivative identities for S)**.** *Fix the product contour $\Gamma \subset \mathbb{C}^4$ from Remark 3.3.3, and let $z = (z_1, z_2, z_3, z_4) \in \Gamma$. We define*

$$\Omega(x,z) := R\big(z_1; \mathrm{diag}(u)\big) R\big(z_2; \mathrm{diag}(v)\big) R\big(z_3; \mathrm{diag}(\beta^*)\big) R\big(z_4; K\big) \in \mathbb{C}^{d\times d},$$

where

$$R(z; A) := (zI_d - A)^{-1}$$

denotes the resolvent of A. Since K is diagonal in our setting, all factors in $\Omega(x, z)$ are diagonal; in particular, $\Omega(x, z)^\top = \Omega(x, z)$.

Let us consider

$$W(x) := [\,\psi(x) \mid \beta^* \mid \mathbb{1}_d\,] \in \mathbb{R}^{d\times 3},$$

and set the matrix

$$S(x,z) := \frac{1}{d} W(x)^\top \Omega(x,z) W(x) \in \mathbb{C}^{3\times 3}.$$

Then

$$\nabla_u S(x,z), \nabla_v S(x,z) \in \mathbb{C}^{d\times 3\times 3},$$

and

$$\nabla_u^2 S(x,z), \nabla_{uv}^2 S(x,z), \nabla_v^2 S(x,z) \in \mathbb{C}^{d\times d\times 3\times 3}.$$

In the following displays, we write Ω *for* $\Omega(x,z)$*. Products involving block matrices are interpreted as tensor contractions over the d-dimensional coordinate: a* 3×3 *block matrix with vector-valued entries is identified with an element of* $\mathbb{C}^{d\times 3\times 3}$*, and a* 3×3 *block matrix with matrix-valued entries is identified with an element of* $\mathbb{C}^{d\times d\times 3\times 3}$*.*

The first derivatives of S *have the following form*

$$\begin{aligned}
\nabla_u S(x,z) = {}& \frac{1}{d}\nabla_u\psi(x)\Omega \begin{bmatrix} \psi(x) & \beta^* & \mathbb{1}_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{1}{d}\nabla_u\psi(x)\Omega \begin{bmatrix} \psi(x) & 0 & 0 \\ \beta^* & 0 & 0 \\ \mathbb{1}_d & 0 & 0 \end{bmatrix} \\
& + \frac{1}{d}\operatorname{diag}(\psi(x))R\big(z_1;\operatorname{diag}(u)\big)\Omega \begin{bmatrix} \psi(x) & \beta^* & \mathbb{1}_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
& + \frac{1}{d}\operatorname{diag}(\beta^*)R\big(z_1;\operatorname{diag}(u)\big)\Omega \begin{bmatrix} 0 & 0 & 0 \\ \psi(x) & \beta^* & \mathbb{1}_d \\ 0 & 0 & 0 \end{bmatrix} \\
& + \frac{1}{d}R\big(z_1;\operatorname{diag}(u)\big)\Omega \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \psi(x) & \beta^* & \mathbb{1}_d \end{bmatrix} \in \mathbb{C}^{d\times 3\times 3},
\end{aligned}$$

$$\nabla_v S(x,z) = \frac{1}{d}\nabla_v\psi(x)\Omega \begin{bmatrix} \psi(x) & \beta^* & \mathbb{1}_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{1}{d}\nabla_v\psi(x)\Omega \begin{bmatrix} \psi(x) & 0 & 0 \\ \beta^* & 0 & 0 \\ \mathbb{1}_d & 0 & 0 \end{bmatrix}$$

$$+\frac{1}{d}\operatorname{diag}(\psi(x))R\big(z_2;\operatorname{diag}(v)\big)\Omega\begin{bmatrix}\psi(x) & \beta^* & \mathbb{1}_d\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

$$+\frac{1}{d}\operatorname{diag}(\beta^*)R\big(z_2;\operatorname{diag}(v)\big)\Omega\begin{bmatrix}0 & 0 & 0\\ \psi(x) & \beta^* & \mathbb{1}_d\\ 0 & 0 & 0\end{bmatrix}$$

$$+\frac{1}{d}R\big(z_2;\operatorname{diag}(v)\big)\Omega\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ \psi(x) & \beta^* & \mathbb{1}_d\end{bmatrix}\in\mathbb{C}^{d\times 3\times 3}.$$

Moreover, the second derivatives of S are

$$\nabla_u^2 S(x,z)=\frac{2q_{11}}{d}\Omega\begin{bmatrix}\operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}+\frac{2q_{11}}{d}\Omega\begin{bmatrix}\operatorname{diag}(\psi(x)) & 0 & 0\\ \operatorname{diag}(\beta^*) & 0 & 0\\ I_d & 0 & 0\end{bmatrix}$$

$$+\frac{2}{d}\Omega\begin{bmatrix}(\nabla_u\psi(x))^2 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

$$+\frac{2}{d}\nabla_u\psi(x)R\big(z_1;\operatorname{diag}(u)\big)\Omega\begin{bmatrix}\operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

$$+\frac{2}{d}\nabla_u\psi(x)R\big(z_1;\operatorname{diag}(u)\big)\Omega\begin{bmatrix}\operatorname{diag}(\psi(x)) & 0 & 0\\ \operatorname{diag}(\beta^*) & 0 & 0\\ I_d & 0 & 0\end{bmatrix}$$

$$+\frac{2}{d}\operatorname{diag}(\psi(x))R\big(z_1;\operatorname{diag}(u)\big)^2\Omega\begin{bmatrix}\operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

$$+\frac{2}{d}\operatorname{diag}(\beta^*)R\big(z_1;\operatorname{diag}(u)\big)^2\Omega\begin{bmatrix}0 & 0 & 0\\ \operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d\\ 0 & 0 & 0\end{bmatrix}$$

$$+\frac{2}{d}R\big(z_1;\operatorname{diag}(u)\big)^2\Omega\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ \operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d\end{bmatrix}\in\mathbb{C}^{d\times d\times 3\times 3},$$

$$\nabla^2_{uv}S(x,z)=\frac{2q_{12}}{d}\Omega\begin{bmatrix}\operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}+\frac{2q_{12}}{d}\Omega\begin{bmatrix}\operatorname{diag}(\psi(x)) & 0 & 0\\ \operatorname{diag}(\beta^*) & 0 & 0\\ I_d & 0 & 0\end{bmatrix}$$

$$+\frac{2}{d}\Omega\begin{bmatrix}(\nabla_u\psi(x))(\nabla_v\psi(x)) & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

$$+\frac{1}{d}\nabla_v\psi(x)R\big(z_1;\operatorname{diag}(u)\big)\Omega\begin{bmatrix}\operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

$$+\frac{1}{d}\nabla_v\psi(x)R\big(z_1;\operatorname{diag}(u)\big)\Omega\begin{bmatrix}\operatorname{diag}(\psi(x)) & 0 & 0\\ \operatorname{diag}(\beta^*) & 0 & 0\\ I_d & 0 & 0\end{bmatrix}$$

$$+\frac{1}{d}\nabla_u\psi(x)R\big(z_2;\operatorname{diag}(v)\big)\Omega\begin{bmatrix}\operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

$$+\frac{1}{d}\nabla_u\psi(x)R\big(z_2;\operatorname{diag}(v)\big)\Omega\begin{bmatrix}\operatorname{diag}(\psi(x)) & 0 & 0\\ \operatorname{diag}(\beta^*) & 0 & 0\\ I_d & 0 & 0\end{bmatrix}$$

$$+\frac{1}{d}\operatorname{diag}(\psi(x))R\big(z_1;\operatorname{diag}(u)\big)R\big(z_2;\operatorname{diag}(v)\big)\Omega\begin{bmatrix}\operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

$$+\frac{1}{d}\operatorname{diag}(\beta^*)R\big(z_1;\operatorname{diag}(u)\big)R\big(z_2;\operatorname{diag}(v)\big)\Omega\begin{bmatrix}0 & 0 & 0\\ \operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d\\ 0 & 0 & 0\end{bmatrix}$$

$$+\frac{1}{d}R\big(z_1;\operatorname{diag}(u)\big)R\big(z_2;\operatorname{diag}(v)\big)\Omega\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ \operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d\end{bmatrix}\in\mathbb{C}^{d\times d\times 3\times 3},$$

$$\nabla_v^2 S(x,z)=\frac{2q_{22}}{d}\Omega\begin{bmatrix}\operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}+\frac{2q_{22}}{d}\Omega\begin{bmatrix}\operatorname{diag}(\psi(x)) & 0 & 0\\ \operatorname{diag}(\beta^*) & 0 & 0\\ I_d & 0 & 0\end{bmatrix}$$

$$+\frac{2}{d}\Omega\begin{bmatrix}(\nabla_v\psi(x))^2 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

$$+\frac{2}{d}\nabla_v\psi(x)R\big(z_2;\operatorname{diag}(v)\big)\Omega\begin{bmatrix}\operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

$$+\frac{2}{d}\nabla_v\psi(x)R\big(z_2;\operatorname{diag}(v)\big)\Omega\begin{bmatrix}\operatorname{diag}(\psi(x)) & 0 & 0\\ \operatorname{diag}(\beta^*) & 0 & 0\\ I_d & 0 & 0\end{bmatrix}$$

$$+\frac{2}{d}\operatorname{diag}(\psi(x))R\big(z_2;\operatorname{diag}(v)\big)^2\Omega\begin{bmatrix}\operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}$$

$$+\frac{2}{d}\operatorname{diag}(\beta^*)R\big(z_2;\operatorname{diag}(v)\big)^2\Omega\begin{bmatrix}0 & 0 & 0\\ \operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d\\ 0 & 0 & 0\end{bmatrix}$$

$$+\frac{2}{d}R\big(z_2;\operatorname{diag}(v)\big)^2\Omega\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ \operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d\end{bmatrix}\in\mathbb{C}^{d\times d\times 3\times 3}.$$

B.1.4 Concentration and Pseudo-Lipschitz Estimates

We use the subgaussian norm $\|\cdot\|_{\psi_2}$, which is equivalent up to universal constants to the optimal variance proxy in a Gaussian tail bound. Namely, for a real-valued random variable X,

$$\|X\|_{\psi_2}\asymp\inf\left\{V>0:\mathbb{P}\big(|X|>t\big)\leqslant 2e^{-t^2/V^2}\text{ for all }t>0\right\}. \tag{B.2}$$

Gaussian random variables are naturally subgaussian. Moreover, Gaussian measures satisfy the stronger property of dimension-free *Lipschitz concentration*, which gives cocnentration inequalities for nonlinear functions of Gaussian vectors. Specifically, let V_0 be a finite-dimensional Hilbert space, and let Z be a centered isotropic Gaussian vector in V_0. If

$g : V_0 \to \mathbb{R}$ is Lipschitz with constant $L(g)$, meaning that

$$|g(x) - g(y)| \leq L(g)\|x - y\|_{V_0} \qquad \text{for all } x, y \in V_0,$$

then we have

$$\|g(Z) - \mathbb{E}\, g(Z)\|_{\psi_2} \leq CL(f),$$

where $C > 0$ is an absolute universal constant. In particular, the bound does not depend on the dimension of V_0.

Pseudo-Lipschitz functions. In our setting, we will also work with functions which are not quite Lipschitz, in that they are locally Lipschitz (Lipschitz on compact sets) and moreover have polynomial growth of their Lipschitz on norm-balls.

Definition B.1.4 (Constant Pseudo-Lipschitz functions)**.** A function $f : V_0 \to V_1$ is called pseudo-Lipschitz of order α if there exists a constant $L = L(\alpha, f)$ such that, for all $x, y \in V_0$, we have

$$\|f(x) - f(y)\|_{V_1} \leq L\|x - y\|_{V_0} \left(1 + \|x\|_{V_0}^{\alpha} + \|y\|_{V_0}^{\alpha}\right).$$

We call L an α-pseudo-Lipschitz constant for f.

We will often works with outer functions and statistics whose gradients are α-pseudo-Lipschitz. To reduce pseudo-Lipschitz estimates to Lipschitz estimates on bounded sets, we will use projection onto norm balls. For $\beta > 0$, we define the *projection operator onto the ball of radius* β, denoted $\mathrm{Proj}_\beta : V_0 \to V_0$, by

$$\mathrm{Proj}_\beta(x) := \arg\min_{y \in \beta\mathbb{B}} \|x - y\|_{V_0}^2 = \begin{cases} x, & \text{if } \|x\|_{V_0} \leqslant \beta, \\ \beta\left(\frac{x}{\|x\|_{V_0}}\right), & \text{otherwise.} \end{cases},$$

where $\mathbb{B}$ denotes the unit ball in V_0.

Lemma B.1.5. *Suppose $f : V_0 \to V_1$ is α-pseudo-Lipschitz with constant L. Then the composition $f \circ \mathrm{Proj}_\beta$ is Lipschitz with constant $L(1 + 2\beta^\alpha)$.*

Proof. Projection onto a closed convex set is 1-Lipschitz. Therefore, for all $x, y \in V_0$, it holds

$$\begin{aligned}\left\|(f \circ \mathrm{Proj}_\beta)(x) - (f \circ \mathrm{Proj}_\beta)(y)\right\|_{V_1} &\le L \left\|\mathrm{Proj}_\beta(x) - \mathrm{Proj}_\beta(y)\right\|_{V_0} \\ &\quad \times \left(1 + \|\,\mathrm{Proj}_\beta(x)\|_{V_0}^{\alpha} + \|\,\mathrm{Proj}_\beta(y)\|_{V_0}^{\alpha}\right) \\ &\le L(1 + 2\beta^\alpha)\|x - y\|_{V_0},\end{aligned}$$

as we had to show. □

We will use the following growth estimate for moments of the first partial derivative of the α-pseudo-Lipschitz outer function f; see [21].

Lemma B.1.6 (Growth of $\nabla_{r_1} f$)**.** *Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is α-pseudo-Lipschitz with pseudo-Lipschitz constant $L(f)$, as in Assumption 3.1.4. Then, for any $p > 0$ and $r \in \mathbb{R}^2$, we have*

$$\left|\nabla_{r_1} f(r)\right|^p \le C(\alpha, p)\, L(f)^p \left(1 + \|r\|\right)^{\max\{1, \alpha p\}}, \tag{B.3}$$

where $\nabla_{r_1} f$ denotes the partial derivative of f with respect to its first coordinate.

Moreover, set $r = \frac{1}{\sqrt{d}} \begin{bmatrix} \psi(x) & \beta^* \end{bmatrix}^\top a \in \mathbb{R}^2$, and let

$$W(x) := [\,\psi(x) \mid \beta^* \mid \mathbb{1}_d\,] \in \mathbb{R}^{d \times 3}.$$

Then the following moment bound holds:

$$\begin{aligned}\mathbb{E}_a\left[\left|\nabla_{r_1} f(r)\right|^p\right] &\le C(\alpha, p)\, L(f)^p \left(1 + \frac{1}{\sqrt{d}} \|K\|_{\mathrm{op}}^{1/2} \|W(x)\|\right)^{\max\{1, \alpha p\}}, \\ \|1 + \|r\|\|_{\psi_2} &\le C\left(1 + \frac{1}{\sqrt{d}} \|K\|_{\mathrm{op}}^{1/2} \|W(x)\|\right).\end{aligned} \tag{B.4}$$

B.2 Dynamics of the Resolvent Statistic

The purpose of this section is to identify the statistic that mediates between the stochastic dynamics and their deterministic limit. Recall that our goal is to prove that, for every statistic $\varphi : \mathbb{R}^{2d} \to \mathbb{R}$ satisfying Assumption 3.3.6, the processes $\varphi(x_{\lfloor td \rfloor})$ and $\varphi(\mathcal{X}_t)$ are close in the sense that they converge to the same deterministic limit.

The central object in this comparison is the matrix-valued statistic

$$S(x, z) := \frac{1}{d} W(x)^\top \Omega(x, z) W(x) \in \mathbb{C}^{3 \times 3},$$

where

$$W(x) := [\,\psi(x) \mid \beta^* \mid \mathbb{1}_d\,] \in \mathbb{R}^{d\times 3},$$

and $\Omega(x,z)$ is the resolvent product defined in (3.18). We study this statistic along both trajectories, namely $S(\mathcal{X}_t, z)$ (homogenized SGD updates) and $S(x_{\lfloor td\rfloor}, z)$ (SGD updates), for z on the fixed product contour of Remark 3.3.3.

The argument has two main components. First, we show that the discrete and homogenized dynamics are close when tested against S, in the sense that $S(x_{\lfloor td\rfloor}, z)$ and $S(\mathcal{X}_t, z)$ remain close uniformly over the relevant time interval. Second, we show that $S(\mathcal{X}_t, z)$ is itself close to a deterministic limit $\mathscr{S}(t,z)$, where $\mathscr{S}$ solves the partial integro-differential equation (B.6). Combining these two steps yields a deterministic description of $S(x_{\lfloor td\rfloor}, z)$.

This statistic is powerful because it encodes enough spectral information to recover the deterministic limits of the broader class of statistics φ considered in Assumption 3.3.6. We make this reduction explicit in Section B.2.2. Thus, the dynamics of $S(x,z)$ serve as the dynamical nexus of the proof: once S is controlled, the limiting behavior of the other admissible statistics follows by the contour representations developed below. Beyond this, the dynamics of the mapping $S(x,z)$ itself often provide useful insights into analyzing the optimization trajectories of particular optimization problems. Indeed, properties of the solutions to which the algorithms converge can be derived by looking at the mapping $S(x,z)$.

B.2.1 Approximate Solutions and Stability

We begin by recalling the quantities entering the partial integro-differential equation. By Assumptions 3.1.10 and 3.1.11, we have

$$\mathcal{R}(x) := h\,(B(x)) \quad \text{and} \quad \mathbb{E}_a[\nabla_{r_1} f(r)^2] := I\,(B(x)) \quad \text{with} \quad B(x) = \frac{1}{d} W(x)^\top K W(x),$$

where $h, I : \mathbb{R}^{3\times 3} \to \mathbb{R}$ are differentiable and α-pseudo-Lipschitz. It will be useful to isolate the first-column components of the gradient of h. We therefore define

$$H\,(B(x)) = \left[\begin{array}{c|c|c} \nabla_{11} h\,(B(x)) & 0 & 0 \\ \hline \nabla_{21} h\,(B(x)) & 0 & 0 \\ \hline \nabla_{31} h\,(B(x)) & 0 & 0 \end{array}\right].$$

With this notation in place, we now introduce the partial integro-differential equation. In what follows, we will work with a function $\mathcal{F}\big(z, \mathcal{S}(t,\cdot)\big)$. Its explicit formula can be found in the displayed lines after Remark B.3.3, but will not be important for what follows; the only feature that matters is that it can be written as a sum of simple terms described below.

Partial Integro-differential Equation for $\mathscr{S}(t,z)$. Fix a product contour $\Gamma \subset \mathbb{C}^4$ as in Remark 3.3.3. For a multi-index $\zeta = (\zeta_1,\zeta_2,\zeta_3,\zeta_4)$ and a complex vector $z = (z_1,z_2,z_3,z_4) \in \Gamma$, write

$$z^\zeta := \prod_{i=1}^4 z_i^{\zeta_i}.$$

Let $\mathcal{F}\big(z, \mathcal{S}(t,\cdot)\big)$ be a finite sum of terms of the form

$$-2\gamma(t)\, H\big(\mathcal{B}(t)\big)^\top z^\zeta \left(\bigcirc_{i=1}^4 \mathcal{G}_i\right)\big(\mathcal{S}(t,\cdot)\big)$$

or (B.5)

$$C(Q)\,\gamma(t)^2\, I\big(\mathcal{B}(t)\big) z^\zeta \left(\bigcirc_{i=1}^4 \mathcal{G}_i\right)\big(\mathcal{S}(t,\cdot)\big),$$

where $\bigcirc_{i=1}^4 \mathcal{G}_i$ denotes successive composition,

$$\left(\bigcirc_{i=1}^4 \mathcal{G}_i\right)\big(\mathcal{S}(t,\cdot)\big) := (\mathcal{G}_1 \circ \mathcal{G}_2 \circ \mathcal{G}_3 \circ \mathcal{G}_4)\big(\mathcal{S}(t,\cdot)\big).$$

Here each $\mathcal{G}_i$ acts only on the z_i-coordinate and is chosen from the following three operators:

$$\mathcal{G}_i\big(\mathcal{S}(t,\cdot)\big) \in \left\{ \frac{1}{2\pi i}\oint_{\Gamma_i} q(w)\,\mathcal{S}(t,z)\,\mathrm{d}w,\ -\frac{\mathrm{d}}{\mathrm{d}z_i}\mathcal{S}(t,z),\ \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}z_i^2}\mathcal{S}(t,z) \right\}.$$

Finally,

$$\mathcal{B}(t) := \frac{1}{(2\pi)^4}\oint_{\Gamma} z_4\,\mathcal{S}(t,z)\,\mathrm{d}z.$$

We then define $\mathscr{S}$ as a solution to the evolution equation

$$\mathrm{d}\mathscr{S}(t,\cdot) = \mathcal{F}\big(z, \mathscr{S}(t,\cdot)\big)\,\mathrm{d}t, \tag{B.6}$$

with initial condition

$$\mathscr{S}(0,z) = S(x_0,z). \tag{B.7}$$

We next introduce a notion of approximate solution to (B.6). The point of this definition is that both $S(\mathcal{X}_t, z)$ and $S(x_{\lfloor td \rfloor}, z)$, which are functions of both homogenized SGD and SGD respectively, will be shown to satisfy the PDE up to a small error.

To measure errors uniformly on the fixed contour, we define a norm, $\|\cdot\|_\Gamma$, on a continuous function $S\colon \Gamma \subset \mathbb{C}^4 \to \mathbb{C}^{3\times 3}$ by

$$\|S\|_\Gamma = \max_{z\in\Gamma} \|S(z)\|.$$

The following lemma relates this contour norm to the parameter squared norm along both the homogenized and discrete trajectories.

Lemma B.2.1. *There exist constants $0 < c < C < \infty$, depending on $\|K\|_{\mathrm{op}}$, $\|\beta^*\|_\infty$, and Γ, such that for all $t \geqslant 0$,*

$$c \leqslant \frac{\|S(\mathcal{X}_t, \cdot)\|_\Gamma}{\frac{1}{d}\|\mathcal{W}_t\|^2}, \frac{\left\|S(x_{\lfloor td \rfloor}, \cdot)\right\|_\Gamma}{\frac{1}{d}\left\|W_{\lfloor td \rfloor}\right\|^2} \leqslant C.$$

Proof. For homogenized SGD, we have

$$\frac{1}{d}\|\mathcal{W}_t\|^2 = \frac{1}{(2\pi)^4}\oint_\Gamma \operatorname{Tr} S(\mathcal{X}_t, z)\,\mathrm{d}z \leqslant C\,\|S(\mathcal{X}_t,\cdot)\|_\Gamma.$$

On the other hand, by Neumann series and since $|z_i| > \|D_i\|_{\mathrm{op}}$ on each Γ_i, we know that

$$R(z_i; D_i) = (z_i \cdot I_d - D_i)^{-1} = \frac{1}{z_i}\left(I_d - \frac{1}{z_i}D_i\right)^{-1} = \frac{1}{z_i}\sum_{j=0}^{\infty}\left(\frac{1}{z_i}D_i\right)^j,$$

therefore

$$\|R(z_i; D_i)\|_{\mathrm{op}} \leqslant \frac{1}{|z_i|}\sum_{j=0}^{\infty}\left(\frac{1}{|z_i|}\|D_i\|_{\mathrm{op}}\right)^j = \frac{1}{|z_i|}\cdot\frac{1}{1-\frac{1}{|z_i|}\|D_i\|_{\mathrm{op}}} = \frac{1}{|z_i| - \|D_i\|_{\mathrm{op}}} \leqslant 2.$$

Thus we conclude

$$\|S(\mathcal{X}_t,\cdot)\|_\Gamma = \max_{z\in\Gamma}\left\|\frac{1}{d}\mathcal{W}_t^\top \Omega \mathcal{W}_t\right\| \leqslant \frac{1}{d}\|\mathcal{W}_t\|^2 \cdot \Pi_{i=1}^4 \max_{z_i\in\Gamma_i}\|R(z_i; D_i)\|_{\mathrm{op}} \leqslant \frac{2^4}{d}\|\mathcal{W}_t\|^2.$$

The same bounds hold for SGD with obvious changes. □

We will be working with *approximate solutions to the partial integro-differential equation* defined as:

Definition B.2.2 ((ε, M, T)-approximate solution to the partial integro-differential equation)**.** For constants $M, T, \varepsilon > 0$, we say that a continuous function

$$\mathcal{S}\colon \{t \geqslant 0\} \times \Gamma \subset \{t \geqslant 0\} \times \mathbb{C}^4 \to \mathbb{C}^{3\times 3}$$

is an (ε, M, T)-approximate solution to (B.6) if it satisfies the following properties.

Stopping time. Define the stopping time

$$\hat{\tau}_M(\mathcal{S}) := \inf\Big\{ t \geqslant 0 : \|\mathcal{S}(t,\cdot)\|_\Gamma > M \text{ or } \mathcal{B}(t) \notin \mathcal{U} \Big\}.$$

(i) **Lipschitz regularity of derivatives.** For any other (ε, M, T)–approximate solution $\tilde{\mathcal{S}}$ and for each coordinate z_i, the first and second partial derivatives $\frac{\mathrm{d}}{\mathrm{d}z_i}\mathcal{S}$ and $\frac{\mathrm{d}^2}{\mathrm{d}z_i^2}\mathcal{S}$ exist, and they depend Lipschitz–continuously on $\mathcal{S}$ in the sense that for any $s \geqslant 0$ and $z \in \Gamma \subset \mathbb{C}^4$, it holds:

$$\left\| \frac{\mathrm{d}}{\mathrm{d}z_i}\mathcal{S}(s \wedge \hat{\tau}_M^{\mathcal{S}}, z) - \frac{\mathrm{d}}{\mathrm{d}z_i}\tilde{\mathcal{S}}(s \wedge \hat{\tau}_M^{\tilde{\mathcal{S}}}, z) \right\|_\Gamma \leqslant C \left\| \mathcal{S}(s \wedge \hat{\tau}_M^{\mathcal{S}}, \cdot) - \tilde{\mathcal{S}}(s \wedge \hat{\tau}_M^{\tilde{\mathcal{S}}}, \cdot) \right\|_\Gamma,$$

$$\left\| \frac{\mathrm{d}^2}{\mathrm{d}z_i^2}\mathcal{S}(s \wedge \hat{\tau}_M^{\mathcal{S}}, z) - \frac{\mathrm{d}^2}{\mathrm{d}z_i^2}\tilde{\mathcal{S}}(s \wedge \hat{\tau}_M^{\tilde{\mathcal{S}}}, z) \right\|_\Gamma \leqslant C \left\| \mathcal{S}(s \wedge \hat{\tau}_M^{\mathcal{S}}, \cdot) - \tilde{\mathcal{S}}(s \wedge \hat{\tau}_M^{\tilde{\mathcal{S}}}, \cdot) \right\|_\Gamma.$$

(ii) **Approximate satisfaction of the PDE.** The following integral-form error bound holds:

$$\sup_{0 \leqslant t \leqslant (\hat{\tau}_M(\mathcal{S}) \wedge T)} \left\| \mathcal{S}(t,\cdot) - \mathcal{S}(0,\cdot) - \int_0^t \mathcal{F}\left(\cdot, \mathcal{S}(s,\cdot)\right) \mathrm{d}s \right\|_\Gamma \leqslant \varepsilon,$$

where the initial condition is $\mathcal{S}(0,\cdot) = S(x_0,\cdot)$, with x_0 the initialization of SGD.

We suppress the $\mathcal{S}$ in the notation for $\hat{\tau}_M$, that is $\hat{\tau}_M = \hat{\tau}_M(\mathcal{S})$, when the function $\mathcal{S}$ is clear from context.

Remark B.2.3. Consider

$$S(x,z) = \frac{1}{d} W(x)^\top \Omega(x,z) W(x) \in \mathbb{C}^{3\times 3}$$

for

$$\Omega = R\left(z_1; \operatorname{diag}(u)\right) \cdot R\left(z_2; \operatorname{diag}(v)\right) \cdot R\left(z_3; \operatorname{diag}(\beta^*)\right) \cdot R\left(z_4; K\right) \in \mathbb{C}^{d\times d}.$$

Differentiating the resolvent representation gives

$$\frac{\mathrm{d}^2}{\mathrm{d}z_i^2} S(x,z) = \frac{2}{d} W^\top R(z_i; D_i)^2 \Omega(x,z) W, \qquad \frac{\mathrm{d}^3}{\mathrm{d}z_i^3} S(x,z) = -\frac{6}{d} W^\top R(z_i; D_i)^3 \Omega(x,z) W,$$

so by Lemma B.2.1 and (B.40), we get

$$\left\| \frac{\mathrm{d}^2}{\mathrm{d}z_i^2} S(x,z) \right\|_\Gamma \leqslant C \left\| S(x,\cdot) \right\|_\Gamma \quad \text{and} \quad \left\| \frac{\mathrm{d}^3}{\mathrm{d}z_i^3} S(x,z) \right\|_\Gamma \leqslant C \left\| S(x,\cdot) \right\|_\Gamma .$$

Consequently, the processes $S(x_{\lfloor td \rfloor}, z)$ and $S(\mathcal{X}_t, z)$ satisfy condition (i), since $\|\mathcal{S}(s \wedge \hat{\tau}_M, \cdot)\|_\Gamma \leqslant M$ for any $s \geqslant 0$. Furthermore, in Section B.3, we prove that they also satisfy condition (ii) and hence that they are (ε, M, T)-approximate solutions. Note that we must extend the discrete time of SGD to a continuous time (see Section B.3 for details). Finally, it is clear by definition that the exact solution $\mathscr{S}$ of (B.6) is an $(0, M, T)$-approximate solution.

The first result is a *stability* statement: any two (ε, M, T)-approximate solutions remain uniformly close up to the stopping time.

Proposition B.2.4 (Stability)**.** *For all (ε, M, T)-approximate solutions $\mathcal{S}_1$ and $\mathcal{S}_2$, there exists a positive constant $C = C(L(h), L(I), \bar{\gamma}, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, M, \alpha, T)$ such that*

$$\sup_{0\leqslant t\leqslant T} \left\| \mathcal{S}_1(t \wedge \tau_M, \cdot) - \mathcal{S}_2(t \wedge \tau_M, \cdot) \right\|_\Gamma \leqslant C \cdot \varepsilon,$$

where $\tau_M = \min\{\hat{\tau}_M(\mathcal{S}_1), \hat{\tau}_M(\mathcal{S}_2)\}$.

Proof. First note that $\tau_M \leqslant \hat{\tau}_M(\mathcal{S}_1)$ and $\tau_M \leqslant \hat{\tau}_M(\mathcal{S}_2)$. Thus all estimates below are taken up to the common stopping time τ_M. Write $\mathcal{S}_1$ and $\mathcal{S}_2$ as

$$\begin{aligned} \mathcal{S}_1(t,\cdot) &= \mathcal{S}_1(0,\cdot) + \int_0^t \mathcal{F}\left(\cdot, \mathcal{S}_1(s,\cdot)\right) \mathrm{d}s + \varepsilon(\mathcal{S}_1) \\ \mathcal{S}_2(t,\cdot) &= \mathcal{S}_2(0,\cdot) + \int_0^t \mathcal{F}\left(\cdot, \mathcal{S}_2(s,\cdot)\right) \mathrm{d}s + \varepsilon(\mathcal{S}_2), \end{aligned} \tag{B.8}$$

where $\varepsilon(\mathcal{S}_1)$ and $\varepsilon(\mathcal{S}_2)$ are error terms from the (ε, M, T)-approximate solution inequality and we have for $j = 1, 2$ the estimate

$$\sup_{0 \leqslant t \leqslant (T \wedge \tau_M)} \|\varepsilon(\mathcal{S}_j)\|_\Gamma \leqslant \varepsilon.$$

We first prove the stability estimate under the following Lipschitz bound on $\mathcal{F}$. Suppose that there exists a constant $C = C(L(h), L(I), \bar{\gamma}, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, M, \alpha)$ such that, for all s,

$$\|\mathcal{F}(\cdot, \mathcal{S}_1(s \wedge \tau_M, \cdot)) - \mathcal{F}(\cdot, \mathcal{S}_2(s \wedge \tau_M, \cdot))\|_\Gamma \leqslant C \|\mathcal{S}_1(s \wedge \tau_M, \cdot) - \mathcal{S}_2(s \wedge \tau_M, \cdot)\|_\Gamma . \tag{B.9}$$

We verify (B.9) at the end of the proof. However, Equations (B.8) and (B.9) imply

$$\begin{aligned}
\sup_{0 \leqslant t \leqslant T \wedge \tau_M} \|\mathcal{S}_1(t, \cdot) - \mathcal{S}_2(t, \cdot)\|_\Gamma &\leqslant 2\varepsilon + \sup_{0 \leqslant t \leqslant T \wedge \tau_M} \int_0^t \|\mathcal{F}(\cdot, \mathcal{S}_1(s, \cdot)) - \mathcal{F}(\cdot, \mathcal{S}_2(s, \cdot))\|_\Gamma \, \mathrm{d}s \\
&= 2\varepsilon + \sup_{0 \leqslant t \leqslant T} \int_0^t \|\mathcal{F}(\cdot, \mathcal{S}_1(s \wedge \tau_M, \cdot)) - \mathcal{F}(\cdot, \mathcal{S}_2(s \wedge \tau_M, \cdot))\|_\Gamma \, \mathrm{d}s \\
&\leqslant 2\varepsilon + C \int_0^T \|\mathcal{S}_1(s \wedge \tau_M, \cdot) - \mathcal{S}_2(s \wedge \tau_M, \cdot)\|_\Gamma \, \mathrm{d}s,
\end{aligned}$$

where $C = C(L(h), L(I), \bar{\gamma}, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, M, \alpha)$ is a positive constant.

Define $Q_T := \sup_{0 \leqslant t \leqslant T} \|\mathcal{S}_1(t \wedge \tau_M, \cdot) - \mathcal{S}_2(t \wedge \tau_M, \cdot)\|_\Gamma$. Then one has that

$$Q_T = \sup_{0 \leqslant t \leqslant T \wedge \tau_M} \|\mathcal{S}_1(t, \cdot) - \mathcal{S}_2(t, \cdot)\|_\Gamma \leqslant 2\varepsilon + C \int_0^T Q_s \, \mathrm{d}s.$$

By an application of Gronwall's inequality,

$$\sup_{0 \leqslant t \leqslant T} \|\mathcal{S}_1(t \wedge \tau_M, \cdot) - \mathcal{S}_2(t \wedge \tau_M, \cdot)\|_\Gamma \leqslant 2\varepsilon e^{CT},$$

and the result is shown.

It remains to verify the Lipschitz estimate (B.9). We will do this in steps. First, define $\mathcal{B}_j(\cdot) = \frac{1}{(2\pi)^4} \oint_\Gamma z_4 \, \mathcal{S}_j(\cdot, z) \, \mathrm{d}z$ for $j = 1, 2$. We will use the shorthand $\mathcal{B}_j^{\tau_M}(s) = \mathcal{B}_j(s \wedge \tau_M)$ and $\mathcal{S}_j^{\tau_M}(s, \cdot) = \mathcal{S}_j(s \wedge \tau_M, \cdot)$. Now, by the α-pseudo-Lipschitzness of ∇h (Assumption 3.1.10),

we have

$$\begin{aligned}\left\|H(\mathcal{B}_1^{\tau_M}(s))^\top - H(\mathcal{B}_2^{\tau_M}(s))^\top\right\| &\leqslant L(h)\left(1+\|\mathcal{B}_1^{\tau_M}(s)\|^\alpha + \|\mathcal{B}_2^{\tau_M}(s)\|^\alpha\right)\|\mathcal{B}_1^{\tau_M}(s)-\mathcal{B}_2^{\tau_M}(s)\| \\ &\leqslant C(L(h),M,\alpha)\,\|\mathcal{B}_1^{\tau_M}(s)-\mathcal{B}_2^{\tau_M}(s)\| \\ \left\|H(\mathcal{B}_j^{\tau_M}(s))^\top\right\| &\leqslant L(h)\left(1+\left\|\mathcal{B}_j^{\tau_M}(s)\right\|^{\alpha+1}\right) \leqslant C(L(h),M,\alpha)\end{aligned} \tag{B.10}$$

since we have the expression

$$\left\|\mathcal{B}_j^{\tau_M}(s)\right\| = \left\|\frac{1}{(2\pi)^4}\oint_\Gamma z_4\,\mathcal{S}_j^{\tau_M}(s,z)\,\mathrm{d}z\right\| \leqslant C(|\Gamma|,\|K\|_{\mathrm{op}},\|\beta^*\|_\infty)\left\|\mathcal{S}_j^{\tau_M}(s,\cdot)\right\|_\Gamma \leqslant C\cdot M.$$

Here we used the stopping time τ_M explicitly. Similarly,

$$\begin{aligned}\|\mathcal{B}_1^{\tau_M}(s)-\mathcal{B}_2^{\tau_M}(s)\| &\leqslant C\oint_\Gamma |z_4|\,\|\mathcal{S}_1^{\tau_M}(s,\cdot)-\mathcal{S}_2^{\tau_M}(s,\cdot)\|_\Gamma\,\mathrm{d}\,|z| \\ &\leqslant C(|\Gamma|,\|K\|_{\mathrm{op}},\|\beta^*\|_\infty)\,\|\mathcal{S}_1^{\tau_M}(s,\cdot)-\mathcal{S}_2^{\tau_M}(s,\cdot)\|_\Gamma.\end{aligned}$$

Consequently, there exists a positive constant (independent of s) such that

$$\left\|H(\mathcal{B}_1^{\tau_M}(s))^\top - H(\mathcal{B}_2^{\tau_M}(s))^\top\right\| \leqslant C(L(h),M,\alpha,\|K\|_{\mathrm{op}},\|\beta^*\|_\infty)\,\|\mathcal{S}_1^{\tau_M}(s,\cdot)-\mathcal{S}_2^{\tau_M}(s,\cdot)\|_\Gamma. \tag{B.11}$$

Analogously,

$$\begin{aligned}\left\|\oint_{\Gamma_i} q(z_i)\mathcal{S}_1^{\tau_M}(s,z)\,\mathrm{d}z_i - \oint_{\Gamma_i} q(z_i)\mathcal{S}_2^{\tau_M}(s,z)\,\mathrm{d}z_i\right\| &\leqslant C(|\Gamma|,\|K\|_{\mathrm{op}},\|\beta^*\|_\infty)\cdot \\ &\quad \|\mathcal{S}_1^{\tau_M}(s,\cdot)-\mathcal{S}_2^{\tau_M}(s,\cdot)\|_\Gamma \\ \left\|\oint_{\Gamma_i} q(z_i)\mathcal{S}_j^{\tau_M}(s,z)\,\mathrm{d}z_i\right\| &\leqslant C(|\Gamma|,\|K\|_{\mathrm{op}},\|\beta^*\|_\infty)\left\|\mathcal{S}_j^{\tau_M}(s,\cdot)\right\|_\Gamma \\ &\leqslant C\cdot M.\end{aligned} \tag{B.12}$$

Lastly, by the definition of approximate solution (B.2.2), for every $1\leqslant i\leqslant 4$, we have

$$\begin{aligned}\left\|\frac{\mathrm{d}}{\mathrm{d}z_i}\mathcal{S}_1^{\tau_M}(s,z)-\frac{\mathrm{d}}{\mathrm{d}z_i}\mathcal{S}_2^{\tau_M}(s,z)\right\|_\Gamma &\leqslant C\,\|\mathcal{S}_1^{\tau_M}(s,\cdot)-\mathcal{S}_2^{\tau_M}(s,\cdot)\|_\Gamma \\ \left\|\frac{\mathrm{d}}{\mathrm{d}z_i}\mathcal{S}_j^{\tau_M}(s,z)\right\|_\Gamma &\leqslant C\left\|\mathcal{S}_j^{\tau_M}(s,\cdot)\right\|_\Gamma \leqslant C\cdot M,\end{aligned} \tag{B.13}$$

and

$$\begin{aligned}\left\|\frac{\mathrm{d}^2}{\mathrm{d}z_i^2}\mathcal{S}_1^{\tau_M}(s,z)-\frac{\mathrm{d}^2}{\mathrm{d}z_i^2}\mathcal{S}_2^{\tau_M}(s,z)\right\|_\Gamma &\leqslant C\left\|\mathcal{S}_1^{\tau_M}(s,\cdot)-\mathcal{S}_2^{\tau_M}(s,\cdot)\right\|_\Gamma \\ \left\|\frac{\mathrm{d}^2}{\mathrm{d}z_i^2}\mathcal{S}_j^{\tau_M}(s,z)\right\|_\Gamma &\leqslant C\left\|\mathcal{S}_j^{\tau_M}(s,\cdot)\right\|_\Gamma \leqslant C\cdot M.\end{aligned} \tag{B.14}$$

Therefore, since the composition of Lipschitz functions is Lipschitz, we have

$$\begin{aligned}\left\|\left(\bigcirc_{i=1}^4\mathcal{G}_i\right)\left(\mathcal{S}_1^{\tau_M}(s,\cdot)\right)-\left(\bigcirc_{i=1}^4\mathcal{G}_i\right)\left(\mathcal{S}_2^{\tau_M}(s,\cdot)\right)\right\|_\Gamma &\leqslant C\left\|\mathcal{S}_1^{\tau_M}(s,\cdot)-\mathcal{S}_2^{\tau_M}(s,\cdot)\right\|_\Gamma \\ \left\|\left(\bigcirc_{i=1}^4\mathcal{G}_i\right)\left(\mathcal{S}_j^{\tau_M}(s,\cdot)\right)\right\|_\Gamma &\leqslant C\left\|\mathcal{S}_j^{\tau_M}(s,\cdot)\right\|_\Gamma \leqslant C\cdot M.\end{aligned} \tag{B.15}$$

From equations (B.10), (B.11), and (B.15), it follows that there exists a positive constant $C = C(L(h), M,$
$\alpha, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, \bar{\gamma})$ such that

$$\begin{aligned}&\left\|2\gamma(s)H\left(\mathcal{B}_1^{\tau_M}(s)\right)^\top z^\zeta\left(\bigcirc_{i=1}^4\mathcal{G}_i\right)\left(\mathcal{S}_1^{\tau_M}(s,\cdot)\right)-2\gamma(s)H\left(\mathcal{B}_2^{\tau_M}(s)\right)^\top z^\zeta\left(\bigcirc_{i=1}^4\mathcal{G}_i\right)\left(\mathcal{S}_1^{\tau_M}(s,\cdot)\right)\right\|_\Gamma \\ &\qquad\leqslant C\cdot\left\|\mathcal{S}_1^{\tau_M}(s,\cdot)-\mathcal{S}_2^{\tau_M}(s,\cdot)\right\|_\Gamma.\end{aligned} \tag{B.16}$$

Next, by α-pseudo-Lipschitzness of the squared gradients (Assumption 3.1.11),

$$\begin{aligned}\left\|I(\mathcal{B}_1^{\tau_M}(s))-I(\mathcal{B}_2^{\tau_M}(s))\right\| &\leqslant C(L(I),M,\alpha)\left\|\mathcal{B}_1^{\tau_M}(s)-\mathcal{B}_2^{\tau_M}(s)\right\| \\ \left\|I(\mathcal{B}_j^{\tau_M}(s))\right\| &\leqslant L(I)\left(1+\left\|\mathcal{B}_j^{\tau_M}(s)\right\|^{\alpha+1}\right)\leqslant C(L(I),M,\alpha).\end{aligned} \tag{B.17}$$

Therefore, we deduce that

$$\begin{aligned}&\left\|C(Q)\gamma(s)^2I(\mathcal{B}_1^{\tau_M}(s))z^\zeta\left(\bigcirc_{i=1}^4\mathcal{G}_i\right)\left(\mathcal{S}_1^{\tau_M}(s,\cdot)\right)-C(Q)\gamma(s)^2I(\mathcal{B}_2^{\tau_M}(s))z^\zeta\left(\bigcirc_{i=1}^4\mathcal{G}_i\right)\left(\mathcal{S}_2^{\tau_M}(s,\cdot)\right)\right\|_\Gamma \\ &\qquad\leqslant C\cdot\left\|\mathcal{S}_1^{\tau_M}(s,\cdot)-\mathcal{S}_2^{\tau_M}(s,\cdot)\right\|_\Gamma,\end{aligned} \tag{B.18}$$

where $C = C(L(I), M, \alpha, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, \bar{\gamma})$ is a positive constant. The Lipschitz condition for $\mathcal{F}$ (B.9) holds after applying expressions (B.16) and (B.18). □

We now transfer the stability estimate from the resolvent statistic S to any statistic $\varphi = g \circ Q$ satisfying Assumption 3.3.6. Here we set

$$Q(x) := \frac{1}{d} W^\top q_1(\mathrm{diag}(u)) q_2(\mathrm{diag}(v)) q_4(K) W,$$

where q_1, q_2, q_4 are polynomials. For an approximate solution $\mathcal{S}_i$, define

$$\mathcal{Q}_i(t) = \frac{1}{(2\pi)^4} \oint_\Gamma q_1(z_1) q_2(z_2) q_4(z_4) \mathcal{S}_i(t, z) \, \mathrm{d}z.$$

The following proposition shows that given two approximate solutions, $\mathcal{S}_1$ and $\mathcal{S}_2$, the function $g \circ \mathcal{Q}_1(t)$ is close to $g \circ \mathcal{Q}_2(t)$. The pseudo-Lipschitzness of g, together with the boundedness imposed by the stopping time, allows us to control the difference between $g(\mathcal{Q}_1)$ and $g(\mathcal{Q}_2)$ by the contour distance between $\mathcal{S}_1$ and $\mathcal{S}_2$:

$$\sup_{0 \leqslant t \leqslant T} |(g \circ \mathcal{Q}_1)(t \wedge \tau_M) - (g \circ \mathcal{Q}_2)(t \wedge \tau_M)| \leqslant C \sup_{0 \leqslant t \leqslant T} \|\mathcal{S}_1(t \wedge \tau_M, \cdot) - \mathcal{S}_2(t \wedge \tau_M, \cdot)\|_\Gamma,$$

and then Proposition B.2.4 finishes the result.

Proposition B.2.5. *Suppose $\varphi \colon \mathbb{R}^{2d} \to \mathbb{R}$ is a statistic satisfying Assumption 3.3.6 such that $\varphi(x) = g \circ Q(x)$. Suppose $\mathcal{S}_1$ and $\mathcal{S}_2$ are (ε, M, T)-approximate solutions. Then there exists a positive constant $C = C(L(h), L(I), \bar{\gamma}, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, M, \alpha, T)$ such that*

$$\begin{aligned} \sup_{0 \leqslant t \leqslant T} \Bigg| & g\left(\frac{1}{(2\pi)^4} \oint_\Gamma q_1(z_1) q_2(z_2) q_4(z_4) \mathcal{S}_1^{\tau_M}(t, z) \, \mathrm{d}z \right) \\ & - g\left(\frac{1}{(2\pi)^4} \oint_\Gamma q_1(z_1) q_2(z_2) q_4(z_4) \mathcal{S}_2^{\tau_M}(t, z) \, \mathrm{d}z \right) \Bigg| \leqslant C \cdot \varepsilon \end{aligned}$$

where $\tau_M := \hat{\tau}_M(\mathcal{S}_1) \wedge \hat{\tau}_M(\mathcal{S}_2)$. Here $\mathcal{S}_i^{\tau_M}(t, \cdot) = \mathcal{S}_i(t \wedge \tau_M, \cdot)$.

Proof. Since $\tau_M \leqslant \hat{\tau}_M(\mathcal{S}_1)$ and $\tau_M \leqslant \hat{\tau}_M(\mathcal{S}_2)$, we can always work on the smaller time τ_M. We define $\mathcal{Q}_i(t) = \frac{1}{(2\pi)^4} \oint_\Gamma q_1(z_1) q_2(z_2) q_4(z_4) \mathcal{S}_i(t, z) \, \mathrm{d}z$ and the stopped process $\mathcal{Q}_i^{\tau_M}(t) = \mathcal{Q}_i(t \wedge \tau_M)$ for $i = 1, 2$. First, we observe

$$\begin{aligned} \|\mathcal{Q}_i^{\tau_M}(t)\| & \leqslant C \oint_\Gamma |q_1(z_1)| \, |q_2(z_2)| \, |q_4(z_4)| \, \|\mathcal{S}_i^{\tau_M}(t, z)\| \, \mathrm{d}z \\ & \leqslant C(\|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, \|q_1\|_{\Gamma_1}, \|q_2\|_{\Gamma_2}, \|q_4\|_{\Gamma_4}) \, \|\mathcal{S}_i^{\tau_M}(t, \cdot)\|_\Gamma \leqslant C \cdot M. \end{aligned} \tag{B.19}$$

Moreover, the map $\mathcal{S} \mapsto \mathcal{Q}$ is Lipschitz in the contour norm:

$$\begin{aligned}\|\mathcal{Q}_1^{\tau_M}(t) - \mathcal{Q}_2^{\tau_M}(t)\| &\leqslant C(\|q_1\|_{\Gamma_1}, \|q_2\|_{\Gamma_2}, \|q_4\|_{\Gamma_4}) \oint_\Gamma \|\mathcal{S}_1^{\tau_M}(t,z) - \mathcal{S}_2^{\tau_M}(t,z)\| \, \mathrm{d}\,|z| \\ &\leqslant C(\|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, \|q_1\|_{\Gamma_1}, \|q_2\|_{\Gamma_2}, \|q_4\|_{\Gamma_4}) \, \|\mathcal{S}_1^{\tau_M}(t,\cdot) - \mathcal{S}_2^{\tau_M}(t,\cdot)\|_\Gamma \,. \end{aligned} \tag{B.20}$$

Using the α-pseudo-Lipschitzness of g, together with the boundedness estimate (B.19) and the Lipschitz estimate (B.20), we obtain

$$\begin{aligned}|(g \circ \mathcal{Q}_1^{\tau_M})(t) - (g \circ \mathcal{Q}_2^{\tau_M})(t)| &\leqslant L(g) \, \|\mathcal{Q}_1^{\tau_M}(t) - \mathcal{Q}_2^{\tau_M}(t)\| \left(1 + \|\mathcal{Q}_1^{\tau_M}(t)\|^\alpha + \|\mathcal{Q}_2^{\tau_M}(t)\|^\alpha\right) \\ &\leqslant C \cdot \|\mathcal{S}_1^{\tau_M}(t,\cdot) - \mathcal{S}_2^{\tau_M}(t,\cdot)\|_\Gamma \,, \end{aligned}$$

where $C = C(\|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, M, \|q_1\|_{\Gamma_1}, \|q_2\|_{\Gamma_2}, \|q_4\|_{\Gamma_4}, L(g), \alpha)$ is a positive constant. Taking the supremum over all $0 \leqslant t \leqslant T$ and applying Proposition B.2.4 finishes the result. □

B.2.2 Main Argument of the Proof: Concentration of the Resolvent Statistic S

In this section, we prove concentration of both SGD and homogenized SGD under the resolvent statistic S around the deterministic solution $\mathscr{S}(t,z)$ of (B.6). We first prove a concentration result with a common stopping time, and then strengthen it to a one-sided stopping-time formulation in Theorem B.2.7. This resolvent-level result implies Theorem 3.3.4. The important statistic which will play a pivotal role is

$$S(x,z) = \frac{1}{d} W(x)^\top \Omega(x,z) W(x) \in \mathbb{C}^{3\times 3}, \tag{B.21}$$

as well as the function

$$B(x) = \frac{1}{d} W(x)^\top K W(x) \in \mathbb{R}^{3\times 3} \,. \tag{B.22}$$

Here

$$W(x) := [\, \psi(x) \mid \beta^* \mid \mathbb{1}_d \,] \in \mathbb{R}^{d\times 3}.$$

We will extend the iterates of SGD, $\{x_k\}$, defined on discrete time k to continuous time. This is so that we can compare SGD and homogenized SGD, $\{\mathfrak{X}_t\}$. We relate the k-th iterate of SGD to the continuous time parameter t in homogenized SGD through the relationship $k = \lfloor td \rfloor$. Thus, when $t = 1$, SGD has done exactly d updates.

We are now ready to state and prove one of our main results.

Theorem B.2.6 (Concentration under a common stopping time)**.** *Suppose the risk function* $\mathcal{R}(x)$ (3.7) *satisfies Assumptions 3.1.4, 3.1.10, and 3.1.11. Suppose the stepsize schedule satisfies 3.1.13, the iterates* x_k *and* $\mathfrak{X}_t$ *satisfy 3.3.1, and the hidden parameters* β^* *satisfy Assumption 3.1.9. Moreover the data* $a \sim \mathcal{N}(0, K)$ *satisfies Assumption 3.1.5. Write* $W_{\lfloor td \rfloor} := W(x_{\lfloor td \rfloor})$ *and* $\mathcal{W}_t := W(\mathfrak{X}_t)$ *initialized with* $\mathfrak{X}_0 = x_0$*. Then there is an* $\varepsilon > 0$ *so that for any* $T, M > 0$ *and* d *sufficiently large, with overwhelming probability,*

$$\begin{aligned} \sup_{0 \leqslant t \leqslant T \wedge \tau_M(S(x_{\lfloor td \rfloor}, \cdot), \mathscr{S}(t,\cdot))} \left\| S(x_{\lfloor td \rfloor}, \cdot) - \mathscr{S}(t, \cdot) \right\|_\Gamma &\leqslant d^{-\varepsilon}, \\ \sup_{0 \leqslant t \leqslant T \wedge \tau_M(S(\mathfrak{X}_t, \cdot), \mathscr{S}(t,\cdot))} \left\| S(\mathfrak{X}_t, \cdot) - \mathscr{S}(t, \cdot) \right\|_\Gamma &\leqslant d^{-\varepsilon}, \\ \sup_{0 \leqslant t \leqslant T \wedge \tau_M(S(x_{\lfloor td \rfloor}, \cdot), S(\mathfrak{X}_t, \cdot))} \left\| S(x_{\lfloor td \rfloor}, \cdot) - S(\mathfrak{X}_t, \cdot) \right\|_\Gamma &\leqslant d^{-\varepsilon}, \end{aligned} \tag{B.23}$$

where $\mathscr{S}$ *solves the partial integro-differential equation* (B.6)*, and*

$$\tau_M(\mathcal{S}_1, \mathcal{S}_2) = \min\{\hat{\tau}_M(\mathcal{S}_1), \hat{\tau}_M(\mathcal{S}_2)\}.$$

Proof. We will consider $\mathcal{S}_1(t,z) = S(x_{\lfloor td \rfloor}, z)$ and $\mathcal{S}_2(t,z) = S(\mathfrak{X}_t, z)$ and suppress the notation by setting $\tau_M(\mathcal{S}_1, \mathcal{S}_2) = \tau_M$. We also note that the cases when $\mathcal{S}_1(t,z) = S(x_{\lfloor td \rfloor}, z)$ and $\mathcal{S}_2(t,z) = \mathscr{S}(t,z)$ and $\mathcal{S}_1(t,z) = S(\mathfrak{X}_t, z)$ and $\mathcal{S}_2(t,z) = \mathscr{S}(t,z)$ follow an analogous proof, so for brevity, we do not present them.

By Proposition B.3.2, for some $\tilde{\varepsilon} > 0$, we have that $S(\mathfrak{X}_t, z)$ is an $(d^{-\tilde{\varepsilon}}, M, T)$-approximate solution with overwhelming probability. Moreover, by Proposition B.3.4, $S(x_{\lfloor td \rfloor}, z)$ is an $(d^{-\tilde{\varepsilon}}, M, T)$-approximate solution. (For the deterministic function $\mathscr{S}(t,z)$, it is an $(0, M+1, T)$-approximate solution by definition.) Applying the stability result, Proposition B.2.4, to the three pairs

$$(S(x_{\lfloor td \rfloor}, \cdot), S(\mathfrak{X}_t, \cdot)), \qquad (S(x_{\lfloor td \rfloor}, \cdot), \mathscr{S}(t,\cdot)), \qquad (S(\mathfrak{X}_t, \cdot), \mathscr{S}(t,\cdot))$$

gives the three stated bounds, after possibly decreasing ε. □

In the next theorem, we note that one can remove the condition that *both* processes must remain bounded in $\|\cdot\|_\Gamma$ and separated from $\mathcal{U}$, and reduce this to show that we only need *one* of the processes to satisfy these properties. In this way, we can show that SGD is well-behaved and then conclude that homogenized SGD must also be well-behaved.

For any (ε, M, T)-approximate solution $\mathcal{S}(t,\cdot)$, we define the stopping time

$$\hat{\tau}_{M,\eta}(\mathcal{S}) = \inf\{t \geqslant 0 : \|\mathcal{S}(t,\cdot)\|_\Gamma > M \text{ or } \sup_{\hat{\mathcal{B}} \notin \mathcal{U}} \left\| \mathcal{B}(t,\mathcal{S}) - \hat{\mathcal{B}} \right\| \leqslant \eta\}$$

where we set

$$\mathcal{B}(t,\mathcal{S}) = \frac{1}{(2\pi)^4} \oint_\Gamma z_4 \mathcal{S}(t,z)\,\mathrm{d}z.$$

Our main theorem requires that *only one* of the statistics stays bounded, and not, in particular, both. To define this, we introduce a stopping time

$$\Theta_{M,\eta}^{\mathcal{S}_1,\mathcal{S}_2} := \max_{i=1,2} \hat{\tau}_{M,\eta}(\mathcal{S}_i). \tag{B.24}$$

We note that $\hat{\tau}_{M,0} = \hat{\tau}_M$ with $\hat{\tau}_M$ defined in the (ε, M, T)-approximate solution definition.

Theorem B.2.7 (Concentration under a one-sided stopping time)**.** *Suppose the risk function* $\mathcal{R}(x)$ (3.7) *satisfies Assumptions 3.1.4, 3.1.10, and 3.1.11. Suppose the learning rate schedule satisfies 3.1.13, the iterates* x_k *and* $\mathfrak{X}_t$ *satisfy 3.3.1, and the hidden parameters* β^* *satisfy 3.1.9. Moreover the data* $a \sim \mathcal{N}(0,K)$ *satisfies Assumption 3.1.5. Let* $\Theta_{M,\eta}$ *be defined by* (B.24)*. Write* $W_{\lfloor td \rfloor} := W(x_{\lfloor td \rfloor})$ *and* $\mathcal{W}_t := W(\mathfrak{X}_t)$ *initialized with* $\mathfrak{X}_0 = x_0$*. Then there is an* $\varepsilon > 0$ *so that for any* $T, M, \eta > 0$ *and* d *sufficiently large, with overwhelming probability,*

$$\begin{aligned}
\sup_{0 \leqslant t \leqslant T \wedge \Theta_{M,\eta}^{S(x_{\lfloor td \rfloor},\cdot),\mathscr{S}(t,\cdot)}} \left\| S(x_{\lfloor td \rfloor},\cdot) - \mathscr{S}(t,\cdot) \right\|_\Gamma &\leqslant d^{-\varepsilon}, \\
\sup_{0 \leqslant t \leqslant T \wedge \Theta_{M,\eta}^{S(\mathfrak{X}_t,\cdot),\mathscr{S}(t,\cdot)}} \left\| S(\mathfrak{X}_t,\cdot) - \mathscr{S}(t,\cdot) \right\|_\Gamma &\leqslant d^{-\varepsilon}, \\
\sup_{0 \leqslant t \leqslant T \wedge \Theta_{M,\eta}^{S(x_{\lfloor td \rfloor},\cdot),S(\mathfrak{X}_t,\cdot)}} \left\| S(x_{\lfloor td \rfloor},\cdot) - S(\mathfrak{X}_t,\cdot) \right\|_\Gamma &\leqslant d^{-\varepsilon},
\end{aligned} \tag{B.25}$$

where $\mathscr{S}$ *solves the partial integro-differential equation* (B.6)*.*

Proof. Fix an $\eta > 0$. For two mappings $\mathcal{S}_1$ and $\mathcal{S}_2$, we define the stopping time

$$\tau_{M+1,0}^{\mathcal{S}_1,\mathcal{S}_2} = \min\{\hat{\tau}_{M+1,0}(\mathcal{S}_1), \hat{\tau}_{M+1,0}(\mathcal{S}_2)\}. \tag{B.26}$$

As in the previous theorem, we will consider $\mathcal{S}_1(t,z) = S(x_{\lfloor td \rfloor}, z)$ and $\mathcal{S}_2(t,z) = S(\mathfrak{X}_t, z)$ and suppress the notation by setting $\tau_{M,\eta}^{\mathcal{S}_1,\mathcal{S}_2} = \tau_{M,\eta}$. We also note that the cases when

$\mathcal{S}_1(t,z) = S(x_{\lfloor td \rfloor}, z)$ and $\mathcal{S}_2(t,z) = \mathscr{S}(t,z)$ and $\mathcal{S}_1(t,z) = S(\mathcal{X}_t, z)$ and $\mathcal{S}_2(t,z) = \mathscr{S}(t,z)$ follow an analogous proof so for brevity we do not present them.

By Theorem B.2.6, we have that

$$\sup_{0 \leqslant t \leqslant T \wedge \tau_{M+1,0}} \left\| S(x_{\lfloor td \rfloor}, \cdot) - S(\mathcal{X}_t, \cdot) \right\|_\Gamma \leqslant d^{-\varepsilon} \quad \text{w.o.p.} \tag{B.27}$$

The remaining component is to replace the stopping time $\tau_{M+1,0}$ which requires *both* statistics to have Γ-norm less than $M+1$ with $\Theta_{M,\eta}$ which only requires *one* of the statistics to remain in the good set. Denote the event that (B.27) occurs by A_ε and its complement by $A_\varepsilon^{\mathrm{C}}$. Then for sufficiently large d, we have

$$\mathbb{P}\left[\Theta_{M,\eta} > \tau_{M+1,0}\right] \leqslant \mathbb{P}\left[A_\varepsilon^{\mathrm{C}}\right]. \tag{B.28}$$

To see this, suppose $\Theta_{M,\eta} > \tau_{M+1,0}$. At time $t = \tau_{M+1,0}$, one of the following exits must occur: $\left\| S(x_{\lfloor td \rfloor}, \cdot) \right\|_\Gamma \geqslant M+1$ or $\sup_{\hat{B} \notin \mathcal{U}} \left\| B(x_{\lfloor td \rfloor}) - \hat{B} \right\| \leqslant 0$ or $\left\| S(\mathcal{X}_t, \cdot) \right\|_\Gamma \geqslant M+1$ or $\sup_{\hat{B} \notin \mathcal{U}} \left\| B(\mathcal{X}_t) - \hat{B} \right\| \leqslant 0$. On the other hand, since $\tau_{M+1,0} = t < \Theta_{M,\eta}$, then either $\left\| S(x_{\lfloor td \rfloor}, \cdot) \right\|_\Gamma \geqslant M$ or $\left\| S(\mathcal{X}_t, \cdot) \right\|_\Gamma \geqslant M$ and then $\sup_{\hat{B} \notin \mathcal{U}} \left\| B(x_{\lfloor td \rfloor}) - \hat{B} \right\| > \eta$ or $\sup_{\hat{B} \notin \mathcal{U}} \left\| B(\mathcal{X}_t) - \hat{B} \right\| > \eta$.

Now we consider cases. Suppose $\left\| S(\mathcal{X}_t, \cdot) \right\|_\Gamma \geqslant M+1$. Then $\left\| S(\mathcal{X}_t, \cdot) \right\|_\Gamma$ cannot be less than or equal to M so it must have been that $\left\| S(x_{\lfloor td \rfloor}, \cdot) \right\|_\Gamma \leqslant M$. Since $t = \tau_{M+1,0}$, working on the event that (B.27) occurs, we have that

$$\left\| S(\mathcal{X}_t, \cdot) \right\|_\Gamma \leqslant \left\| S(x_{\lfloor td \rfloor}, \cdot) - S(\mathcal{X}_t, \cdot) \right\|_\Gamma + \left\| S(x_{\lfloor td \rfloor}, \cdot) \right\|_\Gamma \leqslant d^{-\varepsilon} + M.$$

For sufficiently large d, then $\left\| S(\mathcal{X}_t, \cdot) \right\|_\Gamma < M+1$ which is a contradiction.

Suppose $\left\| S(x_{\lfloor td \rfloor}, \cdot) \right\|_\Gamma \geqslant M+1$. Then by reversing the roles of $x_{\lfloor td \rfloor}$ and $\mathcal{X}_t$ in the previous case, we see that this cannot occur.

Next suppose that $\sup_{\hat{B} \notin \mathcal{U}} \left\| B(\mathcal{X}_t) - \hat{B} \right\| \leqslant 0$. Then $\sup_{\hat{B} \notin \mathcal{U}} \left\| B(\mathcal{X}_t) - \hat{B} \right\|$ cannot be greater than η. Thus it had to be the case that $\sup_{\hat{B} \notin \mathcal{U}} \left\| B(x_{\lfloor td \rfloor}) - \hat{B} \right\| > \eta$. Now working on the event that (B.27) occurs, we have that

$$\left\| B(x_{\lfloor td \rfloor}) - \hat{B} \right\| \leqslant \left\| B(x_{\lfloor td \rfloor}) - B(\mathcal{X}_t) \right\| \leqslant C \cdot \sup_{z_4 \in \Gamma_4} |z_4| \cdot \left\| S(x_{\lfloor td \rfloor}, \cdot) - S(\mathcal{X}_t, \cdot) \right\|_\Gamma \leqslant \tilde{C} \cdot d^{-\varepsilon}.$$

where C and $\tilde{C}$ are positive constants. Hence for sufficiently large d, we have $\sup_{\hat{B} \notin \mathcal{U}} \left\| B(x_{\lfloor td \rfloor}) - \hat{B} \right\| < \eta$, a contradiction.

Lastly suppose $\sup_{\hat{B}\notin\mathcal{U}}\left\|B(x_{\lfloor td\rfloor})-\hat{B}\right\|\leqslant 0$. By reversing the roles of $x_{\lfloor td\rfloor}$ and $\mathcal{X}_t$, we reach the same conclusion as the previous case.

Hence the inequality (B.28) holds and thus, $\tau_{M+1,0}\geqslant\Theta_{M,\eta}$ with overwhelming probability. The result follows. □

We immediately get a corollary which shows that SGD and homogenized SGD concentrate around the deterministic function $\mathscr{S}(t,z)$ which is a solution to the partial integro-differential equation (B.6) provided Assumption 3.3.1 holds.

Corollary B.2.8 (Non-explosiveness and concentration)**.** *Suppose the assumptions of Theorem B.2.7 hold. Suppose, in addition, for a fixed $T>0$ and $\eta>0$ that*

$$\sup_{0\leqslant t\leqslant T}\sup_{\hat{B}\notin\mathcal{U}}\left\|B(x_{\lfloor td\rfloor})-\hat{B}\right\|>\eta\quad w.o.p. \tag{B.29}$$

Then there is an $\varepsilon>0$ so that for d sufficiently large, with overwhelming probability,

$$\sup_{0\leqslant t\leqslant T}\left\|S(x_{\lfloor td\rfloor},\cdot)-\mathscr{S}(t,\cdot)\right\|_{\Gamma}\leqslant d^{-\varepsilon}\quad\text{and}\quad\sup_{0\leqslant t\leqslant T}\left\|S(x_{\lfloor td\rfloor},\cdot)-S(\mathcal{X}_t,\cdot)\right\|_{\Gamma}\leqslant d^{-\varepsilon}, \tag{B.30}$$

and therefore

$$\sup_{0\leqslant t\leqslant T}\left\|\mathscr{S}(t,\cdot)-S(\mathcal{X}_t,\cdot)\right\|_{\Gamma}\leqslant 2d^{-\varepsilon}. \tag{B.31}$$

Proof. Define the following stopping time similar to $\Theta_{M,\eta}$ in (B.24) by

$$\tilde{\Theta}_{M,\eta}^{\mathcal{S}_1,\mathcal{S}_2}=\max_{i=1,2}\inf\{t\geqslant 0:\sup_{\hat{\mathcal{B}}\notin\mathcal{U}}\left\|\mathcal{B}(t,\mathcal{S}_i)-\hat{\mathcal{B}}\right\|\leqslant\eta\}. \tag{B.32}$$

Here we think of $\mathcal{S}_1$ as either homogenized SGD or $\mathscr{S}$ and $\mathcal{S}_2$ as SGD. The stopping time $\Theta_{M,\eta}^{\mathcal{S}_1,\mathcal{S}_2}$ from (B.24) is controlled by the non-explosiveness and separation assumptions on the SGD trajectory. By Lemma B.2.1 and Assumption 3.1.9, there exists some $C>0$ independent of d such that

$$\left\|S(x_{\lfloor td\rfloor},\cdot)\right\|_{\Gamma}\leqslant\frac{2^4}{d}\left\|W_{\lfloor td\rfloor}\right\|^2\leqslant 2^4\left(\left\|x_{\lfloor td\rfloor}\right\|_{\infty}^4+\left\|\beta^*\right\|_{\infty}^2+1\right)\leqslant C\left(\left\|x_{\lfloor td\rfloor}\right\|_{\infty}^4+1\right).$$

Consequently, this translates into

$$\{t\geqslant 0:\left\|S(x_{\lfloor td\rfloor},\cdot)\right\|_{\Gamma}>C\left(M^4+1\right)\}\subset\{t\geqslant 0:\left\|x_{\lfloor td\rfloor}\right\|_{\infty}>M\},$$

and so the infimum of the right-hand-side is smaller than the infimum of the left-hand-side. Moreover, we have by Assumption 3.3.1 that

$$T \leqslant \inf\{t \geqslant 0 : \left\|x_{\lfloor td \rfloor}\right\|_\infty > M\} \quad \text{w.o.p.}$$

Similarly we have that

$$T \leqslant \inf\{t \geqslant 0 : \sup_{\hat{B} \notin \mathcal{U}} \left\|B(x_{\lfloor td \rfloor}) - \hat{B}\right\| \leqslant \eta\} \quad \text{w.o.p.}$$

Thus, we have that

$$T \leqslant \tilde{\Theta}^{\mathcal{S}_1,\mathcal{S}_2}_{M,\eta} \leqslant \Theta^{\mathcal{S}_1,\mathcal{S}_2}_{C(M^4+1),\eta} \quad \text{w.o.p.},$$

where $\mathcal{S}_1$ is either $S(\mathcal{X}_t, \cdot)$ or $\mathscr{S}(t, \cdot)$. By Theorem B.2.7, we immediately get the result in (B.30). A simple triangle inequality gives the result in (B.31). □

Lastly, we make one final connection to Theorem 3.3.4, proving the result below.

Proof of Theorem 3.3.4. The result immediately follows from Theorem B.2.7 and Corollary B.2.8 after noting that

$$B(x) = \frac{1}{(2\pi)^4} \oint_\Gamma z_4 S(x, z)\, \mathrm{d}z \quad \text{and} \quad \mathscr{B}(t) = \frac{1}{(2\pi)^4} \oint_\Gamma z_4 \mathscr{S}(t, z)\, \mathrm{d}z$$

and Lipschitzness of the integral, that is,

$$\left\| \oint_\Gamma z_4 \mathcal{S}_1(t, \cdot)\, \mathrm{d}z - \oint_\Gamma z_4 \mathcal{S}_2(t, \cdot)\, \mathrm{d}z \right\| \leqslant C \cdot \|\mathcal{S}_1(t, \cdot) - \mathcal{S}_2(t, \cdot)\|_\Gamma \quad \text{for some positive} \quad C > 0.$$

□

B.2.3 Concentration of General Statistics

We now transfer the resolvent-level concentration result of Theorem B.2.7 to any statistic $\varphi : \mathbb{R}^{2d} \to \mathbb{R}$ satisfying Assumption 3.3.6. This gives Theorem B.2.9, which is a reformulation of Theorem 3.3.7. The result applies, in particular, to the risk and curvature curves, $\mathcal{R}(x)$ and $\mathrm{tr}(\nabla^2 \mathcal{R}(x))$, as well as to other generalization metrics covered by Assumption 3.3.6.

In this section, the statistics $\varphi\colon \mathbb{R}^{2d} \to \mathbb{R}$ of interest satisfy a composite structure

$$\varphi(x) = g\left(\frac{1}{d} W(x)^\top q_1(\operatorname{diag}(u)) q_2(\operatorname{diag}(v)) q_4(K) W(x)\right)$$

where $g\colon \mathbb{R}^{3\times 3} \to \mathbb{R}$ is α-pseudo-Lipschitz on $\mathcal{U}$ and q_1, q_2, q_4 are polynomials (see Assumption 3.3.6). Note that we assume that g is evaluated on the real-valued matrix recovered by the contour integral, which lies in $\mathcal{U}$ up to the stopping time.

Define

$$\mathscr{Q}(t) := \frac{1}{(2\pi)^4} \oint_\Gamma q_1(z_1) q_2(z_2) q_4(z_4) \mathscr{S}(t, z)\, \mathrm{d}z,$$

where $\mathscr{S}$ solves (B.6). The deterministic equivalent of $\varphi(x_{\lfloor td \rfloor})$ and $\varphi(\mathcal{X}_t)$ is then

$$\phi(t) := g(\mathscr{Q}(t)). \tag{B.33}$$

Thus we state our concentration theorem for $\varphi(x_{\lfloor td \rfloor})$ and $\varphi(\mathcal{X}_t)$.

Theorem B.2.9 (Concentration of general statistics)**.** *Suppose the Assumptions of Theorem B.2.7 hold. Suppose, in addition, the statistic satisfies a composite structure,*

$$\varphi(x) = g\left(\frac{1}{d} W(x)^\top q_1(\operatorname{diag}(u)) q_2(\operatorname{diag}(v)) q_4(K) W(x)\right)$$

where $g\colon \mathbb{R}^{3\times 3} \to \mathbb{R}$ is α-pseudo-Lipschitz on $\mathcal{U}$ and q_1, q_2, q_4 are polynomials (see Assumption 3.3.6). Then there is an $\varepsilon > 0$ so that for any $T, M, \eta > 0$ and d sufficiently large, with overwhelming probability,

$$\begin{aligned}
\sup_{0 \leqslant t \leqslant T \wedge \Theta_{M,\eta}^{S(x_{\lfloor td \rfloor}, \cdot), \mathscr{S}(t,\cdot)}} \left|\varphi(x_{\lfloor td \rfloor}) - \phi(t)\right| &\leqslant d^{-\varepsilon}, \\
\sup_{0 \leqslant t \leqslant T \wedge \Theta_{M,\eta}^{S(\mathcal{X}_t, \cdot), \mathscr{S}(t,\cdot)}} \left|\varphi(\mathcal{X}_t) - \phi(t)\right| &\leqslant d^{-\varepsilon}, \\
\sup_{0 \leqslant t \leqslant T \wedge \Theta_{M,\eta}^{S(x_{\lfloor td \rfloor}, \cdot), S(\mathcal{X}_t, \cdot)}} \left|\varphi(x_{\lfloor td \rfloor}) - \varphi(\mathcal{X}_t)\right| &\leqslant d^{-\varepsilon},
\end{aligned} \tag{B.34}$$

where ϕ is defined in (B.33) and the stopping time $\Theta_{M,\eta}^{\mathcal{S}_1, \mathcal{S}_2}$ is defined in (B.24).

Proof. As in the proof of Theorem B.2.7, we define the stopping time $\tau_{M+1,\eta}^{\mathcal{S}_1, \mathcal{S}_2}$ as in (B.26) and suppress the notation by setting $\tau_{M+1,\eta}^{\mathcal{S}_1, \mathcal{S}_2} = \tau_{M,\eta}$. We will consider the case when $\mathcal{S}_1(t, \cdot) = S(x_{\lfloor td \rfloor}, \cdot)$ and $\mathcal{S}_2(t, \cdot) = S(\mathcal{X}_t, \cdot)$. The other cases will follow by analogous proof.

By Proposition B.3.2, we have that $S(\mathcal{X}_t, z)$ is an $(d^{-\tilde{\varepsilon}}, M+1, T)$-approximate solution with overwhelming probability. Moreover, by Proposition B.3.4, the function $S(x_{\lfloor td \rfloor}, z)$ is an $(d^{-\tilde{\varepsilon}}, M+1, T)$-approximate solution. (For the deterministic function $\mathscr{S}$, it is a $(0, M+1, T)$-approximate solution by definition.) We observe that

$$\frac{1}{(2\pi)^4}\oint_\Gamma q_1(z_1)q_2(z_2)q_4(z_4)S(x_{\lfloor td \rfloor}, z)\,\mathrm{d}z = \frac{1}{d}W_{\lfloor td \rfloor}^\top q_1(\mathrm{diag}(u_{\lfloor td \rfloor}))q_2(\mathrm{diag}(v_{\lfloor td \rfloor}))q_4(K)W_{\lfloor td \rfloor},$$
$$\frac{1}{(2\pi)^4}\oint_\Gamma q_1(z_1)q_2(z_2)q_4(z_4)S(\mathcal{X}_t, z)\,\mathrm{d}z = \frac{1}{d}\mathcal{W}_t^\top q_1(\mathrm{diag}(_t))q_2(\mathrm{diag}(_t))q_4(K)\mathcal{W}_t.$$

We apply Proposition B.2.5 to conclude that there exists a $\varepsilon > 0$ such that

$$\sup_{0\leqslant t\leqslant T\wedge\tau_{M+1,0}} \left|\varphi(x_{\lfloor td \rfloor}) - \varphi(\mathcal{X}_t)\right| \leqslant d^{-\varepsilon} \quad \text{w.o.p.} \tag{B.35}$$

Using the same argument as in Theorem B.2.7, we can remove the stopping time $\tau_{M+1,0}$ and replace it with $\Theta_{M,0}$ for sufficiently large d. □

Lastly we formulate an immediate corollary which follows directly from the proofs of Theorem B.2.9 and Corollary B.2.8.

Corollary B.2.10. *Suppose the Assumptions of Theorem B.2.9 and Corollary B.2.8 hold. For any fixed $T > 0$, there exists $\varepsilon > 0$ such that, for d sufficiently large, with overwhelming probability,*

$$\sup_{0\leqslant t\leqslant T} \left|\varphi(x_{\lfloor td \rfloor}) - \phi(t)\right| \leqslant d^{-\varepsilon} \quad \textit{and} \quad \sup_{0\leqslant t\leqslant T} \left|\varphi(x_{\lfloor td \rfloor}) - \varphi(\mathcal{X}_t)\right| \leqslant d^{-\varepsilon}, \tag{B.36}$$

and therefore

$$\sup_{0\leqslant t\leqslant T} |\varphi(\mathcal{X}_t) - \phi(t)| \leqslant 2d^{-\varepsilon}. \tag{B.37}$$

Theorem 3.3.7 follows by applying Corollary B.2.10 to the statistic φ appearing in the theorem statement.

B.3 SGD and Homogenized SGD are Approximate Solutions

In this section, we show that the resolvent statistic $t \mapsto S(x_{\lfloor td \rfloor}, z)$ associated with SGD and the corresponding statistic $t \mapsto S(\mathcal{X}_t, z)$ associated with homogenized SGD satisfy the partial integro-differential equation (B.6) up to a small error. The argument uses a martingale

method in the spirit of diffusion approximation [30]: after applying a Doob decomposition to the statistic S, we show that the martingale terms are negligible and that the drift terms match the operator $\mathcal{F}$.

Recall the matrix-valued statistics

$$S(x,z) = \frac{1}{d} W(x)^\top \Omega(x,z) W(x) \in \mathbb{C}^{3\times 3}, \qquad B(x) = \frac{1}{d} W(x)^\top K W(x) \in \mathbb{R}^{3\times 3},$$

where $W(x) = [\,\psi(x) \mid \beta^* \mid \mathbb{1}_d\,]$ and $\Omega(x,z)$ is defined in (3.18).

We first show that both homogenized SGD and SGD on $S(\cdot, z)$ are (ε, M, T)-approximate solutions as defined in Definition B.2.2. Then by Proposition B.2.4, it is immediately implied that both homogenized SGD and SGD on $S(\cdot, z)$ are uniformly close. Finally, Proposition B.2.5 establishes that the same hold for any statistic $\varphi(x)$ satisfying Assumption 3.3.6. In order to show that both homogenized SGD and SGD on $S(\cdot, z)$ are (ε, M, T)-approximate solutions, we perform a Doob decomposition for both homogenized SGD and SGD and then show that both martingale terms are small.

To compare homogenized SGD with SGD, we rescale time by setting $k = \lfloor td \rfloor$. Thus, one unit of continuous time corresponds to d SGD updates, and we write

$$x_{td} = x_{\lfloor td \rfloor} \quad \text{(SGD)} \qquad \text{and} \qquad \mathcal{X}_t \quad \text{(HSGD)}.$$

Throughout this section, expressions such as $S(x_{\lfloor td \rfloor}, z)$ are understood under this identification, and scalar statistics are real-valued $\varphi : \mathbb{R}^{2d} \to \mathbb{R}$.

When applying the scalar Itô, Taylor, Doob decomposition, or martingale estimates to the complex-valued matrix statistic $S(\cdot, z) \in \mathbb{C}^{3\times 3}$, we do so coordinatewise. Namely, for $a, b \in \{1, 2, 3\}$ and fixed $z \in \Gamma$, define the real-valued functions

$$S_{ab}^{\mathrm{Re},z}(x) := \operatorname{Re} S_{ab}(x,z), \qquad S_{ab}^{\mathrm{Im},z}(x) := \operatorname{Im} S_{ab}(x,z).$$

All scalar identities and estimates below are applied to these real-valued functions. The corresponding complex matrix-valued identities are then obtained by recombining real and imaginary parts. For example,

$$\big(\mathcal{M}_t(S)(z)\big)_{ab} := \mathcal{M}_t\big(S_{ab}^{\mathrm{Re},z}\big) + i\,\mathcal{M}_t\big(S_{ab}^{\mathrm{Im},z}\big).$$

Since there are only finitely many entries, passing from scalar estimates for the real and imaginary parts to matrix estimates for S only changes constants.

Our first argument is a net argument showing that we do not need to work with every $z \in \Gamma \subset \mathbb{C}^4$, but only polynomially many in d. For this, recall Γ defined in Remark 3.3.3. For a fixed $\delta > 0$, we say that Γ_i^δ is a $d^{-\delta}$-mesh of Γ_i if $\Gamma_i^\delta \subset \Gamma_i$ and for every $z_i \in \Gamma_i$ there exists a $\bar{z}_i \in \Gamma_i^\delta$ such that $|z_i - \bar{z}_i| < d^{-\delta}$.

Lemma B.3.1 (Net argument)**.** *Fix $T, M > 0$ and let $\delta > 0$. For each $i = 1, \dots, 4$, let Γ_i^δ be a $d^{-\delta}$-mesh of Γ_i, chosen so that*

$$|\Gamma_i^\delta| \leqslant C_i d^\delta.$$

Set

$$\Gamma^\delta := \Gamma_1^\delta \times \Gamma_2^\delta \times \Gamma_3^\delta \times \Gamma_4^\delta.$$

Then

$$|\Gamma^\delta| = \prod_{i=1}^4 |\Gamma_i^\delta| \leqslant C_\Gamma d^{4\delta},$$

where $C_\Gamma := \prod_{i=1}^4 C_i$ depends only on the fixed contour Γ.

Let $S(t,z)$ denote either $S(x_{\lfloor td \rfloor}, z)$ or $S(\mathcal{X}_t, z)$, and suppose that

$$\sup_{0 \leqslant t \leqslant \hat{\tau}_M(S) \wedge T} \left\| S(t,\cdot) - S(0,\cdot) - \int_0^t \mathcal{F}\big(\cdot, S(s,\cdot)\big)\, \mathrm{d}s \right\|_{\Gamma^\delta} \leqslant \varepsilon, \tag{B.38}$$

where

$$\hat{\tau}_M(S) := \inf\left\{ t \geq 0 : \|S(t,\cdot)\|_\Gamma > M \text{ or } B(t,S) \notin \mathcal{U} \right\}, \qquad B(t,S) := \frac{1}{(2\pi)^4} \oint_\Gamma z_4 S(t,z)\, \mathrm{d}z.$$

Then the approximate PDE residual bound (i) in Definition B.2.2 holds on the full contour Γ with error $\varepsilon + Cd^{-\delta}$, that is,

$$\sup_{0 \leqslant t \leqslant \hat{\tau}_M(S) \wedge T} \left\| S(t,\cdot) - S(0,\cdot) - \int_0^t \mathcal{F}\big(\cdot, S(s,\cdot)\big)\, \mathrm{d}s \right\|_{\Gamma} \leq \varepsilon + Cd^{-\delta}.$$

Here $C = C(M, T, \Gamma, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, \bar{\gamma}, L(I), L(h))$ is positive and independent of d. Consequently, if S also satisfies the derivative regularity condition (i), then S is an $(\varepsilon + Cd^{-\delta}, M, T)$-approximate solution.

Proof. We consider only $S(t,z) = S(\mathcal{X}_t, z)$ as the same argument will also hold for SGD. We also will always work with the stopped process, that is, $S(t \wedge \hat{\tau}_M, z)$, where $\hat{\tau}_M = \inf\{t \geqslant 0 : \|S(t,\cdot)\|_\Gamma \geqslant M\}$. To simplify the notation, we suppress the $\hat{\tau}_M$ and use $S(t,z)$.

First, we state some resolvent identities. One such resolvent identity gives

$$\|R(z_i; D_i) - R(\bar{z}_i; D_i)\|_{\text{op}} \leqslant |z_i - \bar{z}_i| \, \|R(z_i; D_i) R(\bar{z}_i; D_i)\|_{\text{op}} \quad \text{for any } z_i, \bar{z}_i \in \Gamma_i. \tag{B.39}$$

Furthermore, by Neumann series and since $|z_i| > \|D_i\|_{\text{op}}$ on each Γ_i, we know that

$$R(z_i; D_i) = (z_i \cdot I_d - D_i)^{-1} = \frac{1}{z_i}\left(I_d - \frac{1}{z_i} D_i\right)^{-1} = \frac{1}{z_i} \sum_{j=0}^{\infty} \left(\frac{1}{z_i} D_i\right)^j,$$

so then

$$\|R(z_i; D_i)\|_{\text{op}} \leqslant \frac{1}{|z_i|} \sum_{j=0}^{\infty} \left(\frac{1}{|z_i|} \|D_i\|_{\text{op}}\right)^j = \frac{1}{|z_i|} \cdot \frac{1}{1 - \frac{1}{|z_i|}\|D_i\|_{\text{op}}} = \frac{1}{|z_i| - \|D_i\|_{\text{op}}} \leqslant C$$

and we immediately get

$$\sup_{z_i \in \Gamma_i} \|R(z_i; D_i)\|_{\text{op}} \leqslant C(M, \|K\|_{\text{op}}, \|\beta^*\|_\infty). \tag{B.40}$$

These bounds will be useful later in the proof.

Next, with these bounds, we can get estimates on quantities involving $S(t,\cdot)$ where t is fixed and z varies. Fix $z \in \Gamma$ and let $\bar{z} \in \Gamma_\delta$ be such that $|z_i - \bar{z}_i| < d^{-\delta}$ for each $i = 1, \cdots, 4$. Then, using Lemma B.2.1 (and the stopping time $\hat{\tau}_M$), we obtain

$$\begin{aligned} \|S(t,z) - S(t,\bar{z})\| &\leqslant \frac{C}{d} \|\mathcal{W}_t\|^2 \sum_{i=1}^{4} |z_i - \bar{z}_i| \, \|R(z_i; D_i)\|_{\text{op}} \, \|R(\bar{z}_i; D_i)\|_{\text{op}} \\ &\leqslant C \, \|S(\mathcal{X}_t, \cdot)\|_\Gamma \, d^{-\delta} \\ &\leqslant C \cdot M \cdot d^{-\delta}, \end{aligned} \tag{B.41}$$

where we used the boundedness of the product contour Γ in the last inequality.

Similarly, for any $z \in \Gamma$, we have

$$\|S(t,z)\| \leqslant \frac{1}{d} \|\mathcal{W}_t\|^2 \, \Pi_{i=1}^4 \, \|R(z_i; D_i)\|_{\text{op}} \leqslant C \, \|S(\mathcal{X}_t, \cdot)\|_\Gamma \leqslant C \cdot M.$$

Since $S(t,z)$ is a product of resolvents and $t \leqslant \hat{\tau}_M$, all finitely many z-derivatives appearing in the terms defining $\mathcal{F}$ are uniformly bounded on Γ. Hence each term

$$z^\zeta \left(\bigcirc_i \mathcal{G}_i\right) S(t,z)$$

is Lipschitz in z on Γ, with Lipschitz constant depending only on the stopped bounds and the fixed contour. Thus, since $z, \bar{z} \in \Gamma$ and the contour Γ is bounded,

$$\left\| z^{\zeta} \left(\bigcirc_{i=1}^{4} \mathcal{G}_i \right) (S(s,z)) - \bar{z}^{\zeta} \left(\bigcirc_{i=1}^{4} \mathcal{G}_i \right) (S(s,\bar{z})) \right\| \leqslant C \cdot M \cdot d^{-\delta}, \tag{B.42}$$

where we used equations (B.15) and (B.41).

Now we are ready to prove the main result of the proposition. For a fixed $t \leqslant \hat{\tau}_M$ and $z \in \Gamma$ with $\bar{z} \in \Gamma_\delta$ such that $|z_i - \bar{z}_i| \leqslant d^{-\delta}$ for each $i = 1, \cdots, 4$,

$$\begin{aligned}
&\left\| S(t,z) - S(0,z) - \int_0^t \mathcal{F}(z, S(s,\cdot))\, \mathrm{d}s \right\| \\
&\quad \leqslant \|S(t,z) - S(t,\bar{z})\| + \|S(0,z) - S(0,\bar{z})\| + \int_0^t \|\mathcal{F}(z, S(s,\cdot)) - \mathcal{F}(\bar{z}, S(s,\cdot))\|\, \mathrm{d}s \\
&\quad + \left\| S(t,\bar{z}) - S(0,\bar{z}) - \int_0^t \mathcal{F}(\bar{z}, S(s,\cdot))\, \mathrm{d}s \right\| \\
&\quad \leqslant CMd^{-\delta} \\
&\quad + 2\bar{\gamma} \int_0^t \|H(B(s))\| \cdot \left\| z^{\zeta} \left(\bigcirc_{i=1}^{4} \mathcal{G}_i \right) (S(s,z)) - \bar{z}^{\zeta} \left(\bigcirc_{i=1}^{4} \mathcal{G}_i \right) (S(s,\bar{z})) \right\| \mathrm{d}s \\
&\quad + C(Q)\bar{\gamma}^2 \int_0^t |I(B(s))| \cdot \left\| z^{\zeta} \left(\bigcirc_{i=1}^{4} \mathcal{G}_i \right) (S(s,z)) - \bar{z}^{\zeta} \left(\bigcirc_{i=1}^{4} \mathcal{G}_i \right) (S(s,\bar{z})) \right\| \mathrm{d}s \\
&\quad + \varepsilon,
\end{aligned} \tag{B.43}$$

where the omitted terms are bounded in the same way, since $\mathcal{F}$ is a finite sum of terms of the two forms in (B.5).

Here we used (B.41) to bound the first two terms in the first inequality and ε for the last term by the assumption (B.38) in the statement. In the difference $\mathcal{F}(z, S(s,\cdot)) - \mathcal{F}(\bar{z}, S(s,\cdot))$, the coefficients depending only on s and on $\mathcal{B}(s)$ are identical. Thus the difference is controlled by the variation in the monomial and coordinate-operator factor as z is replaced by $\bar{z}$.

As we have already shown that $z^{\zeta} \left(\bigcirc_{i=1}^{4} \mathcal{G}_i \right) (S(s,z))$ is Lipschitz in z, we only need to bound $|I(B(s))|$ and $\|H(B(s))\|$. We have already shown a uniform bound on $|I(B(s))|$ and $\|H(B(s))\|$ in the proof of Proposition B.2.4. Notably, we showed that for $s \leqslant \hat{\tau}_M$, we have

that $|I(B(s))| \leqslant C(L(I), M, \alpha)$ and $\|H(B(s))\| \leqslant C(L(h), M, \alpha)$.

Last, by taking the supremum over $z \in \Gamma$ and then the supremum over $0 \leqslant t \leqslant (\hat{\tau}_M \wedge T)$ on the left-hand-side of (B.43) and then using the bounds (B.41) and (B.42), yields the result. □

It remains to verify the approximate-solution property for the two processes

$$t \mapsto S(\mathfrak{X}_t, z) \qquad \text{and} \qquad t \mapsto S(x_{\lfloor td \rfloor}, z).$$

We do this by applying a Doob decomposition to each process and showing that the resulting martingale terms are negligible.

To do so, it will be convenient to work directly with the stopped process $x_{t \wedge \hat{\tau}_M}$ on the iterates. Since $\hat{\tau}_M$ is a time based on S-values, it is often difficult to apply to iterates of SGD and homogenized SGD, so we introduce equivalent stopping times

$$\begin{aligned} \vartheta_M &= \inf\{t \geqslant 0 : \frac{1}{d} \left\|W_{\lfloor td \rfloor}\right\|^2 > M \text{ or } B\left(x_{\lfloor td \rfloor}\right) \notin \mathcal{U}\} \quad \text{or} \\ \vartheta_M &= \inf\{t \geqslant 0 : \frac{1}{d} \left\|\mathcal{W}_t\right\|^2 > M \text{ or } B\left(\mathfrak{X}_t\right) \notin \mathcal{U}\}. \end{aligned} \tag{B.44}$$

We overload the notation ϑ_M to be either applied to SGD iterates, $x_{\lfloor td \rfloor}$ or homogenized SGD iterates, $\mathfrak{X}_t$, for which it will be made clear in the context which criterion is used. These stopping times are equivalent to $\hat{\tau}_M$ (see Lemma B.2.1). Moreover, we often drop the M so that $\vartheta = \vartheta_M$. It will be convenient to work with the stopped processes, $x^{\vartheta}_{\lfloor td \rfloor} = x_{\lfloor d(t \wedge \vartheta) \rfloor}$ and $\mathfrak{X}^{\vartheta}_t = \mathfrak{X}_{t \wedge \vartheta}$.

B.3.1 Homogenized SGD Under the Resolvent Statistic S

We first verify the approximate-solution property for the homogenized process. Specifically, we show that the resolvent statistic $t \mapsto S(\mathfrak{X}_t, z)$ satisfies the partial integro-differential equation (B.6) up to a martingale error, and that this martingale error is negligible uniformly on the fixed contour.

With this, we recall *homogenized SGD*

$$\mathrm{d}\mathfrak{X}_t = -\gamma(t) d \nabla \mathcal{R}(\mathfrak{X}_t) \mathrm{d}t + \gamma(t) \sqrt{I\left(B(\mathfrak{X}_t)\right)} \left(\nabla \psi(\mathfrak{X}_t)\right)^\top \sqrt{K}\, \mathrm{d}\mathfrak{B}_t,$$

where $\mathfrak{X}_t$ is a stochastic process taking values in $\mathbb{R}^{2d}$ with initial conditions $\mathfrak{X}_0 = x_0$, and $\mathrm{d}\mathfrak{B}_t$ is the differential of a standard Brownian motion in $\mathbb{R}^d$.

Along the homogenized trajectory, write

$$\mathcal{W}_t := [\,\psi\,(\mathfrak{X}_t)\mid \beta^* \mid \mathbb{1}_d\,] \in \mathbb{R}^{d\times 3}, \quad \text{and} \quad \rho_t := \frac{1}{\sqrt{d}} \begin{pmatrix} \psi(\mathfrak{X}_t) \\ \beta^* \end{pmatrix}^\top a \in \mathbb{R}^2\,.$$

We study the homogenized process through the resolvent statistic

$$x \in \mathbb{R}^{2d} \mapsto S(x,z) = \frac{1}{d} W(x)^\top \Omega(x,z) W(x) \in \mathbb{C}^{3\times 3} \quad \text{for} \quad z \in \Gamma \subset \mathbb{C}^4.$$

We will show that $S\,(\mathfrak{X}_t, z)$ is an approximate solution (B.2.2) to the partial integro-differential equation (B.6) which we state below.

Proposition B.3.2 (Homogenized SGD is an approximate solution)**.** *Fix $T, M > 0$ and $0 < \delta < 1/2$. Then $S\,(\mathfrak{X}_t, z)$ is a $(d^{-\delta}, M, T)$-approximate solution w.o.p., that is,*

$$\sup_{0\leqslant t\leqslant(\hat{\tau}_M\wedge T)} \left\| S(\mathfrak{X}_t,\cdot) - S(x_0,\cdot) - \int_0^t \mathcal{F}\,(\cdot, S(\mathfrak{X}_s,\cdot))\,\mathrm{d}s \right\|_\Gamma \leqslant d^{-\delta} \quad \textit{w.o.p.}$$

The proof of this Proposition is deferred to Section B.3.1.

Doob Decomposition for Homogenized SGD

We begin by applying Itô calculus to homogenized SGD under smooth statistics $\varphi : \mathbb{R}^{2d} \to \mathbb{R}$. Later, when applying the resulting identity to $S(\cdot, z)$, we apply it separately to $S_{ab}^{\mathrm{Re},z}$ and $S_{ab}^{\mathrm{Im},z}$ for each matrix entry.

Applying Itô's lemma, we deduce that

$$\begin{aligned}
\mathrm{d}\varphi(\mathfrak{X}_t) &= \langle \nabla\varphi(\mathfrak{X}_t), \mathrm{d}\mathfrak{X}_t\rangle + \frac{1}{2}\langle \nabla^2\varphi(\mathfrak{X}_t), (\mathrm{d}\mathfrak{X}_t)^{\otimes 2}\rangle \\
&= -\gamma(t) d \langle \nabla\varphi(\mathfrak{X}_t), \nabla\mathcal{R}(\mathfrak{X}_t)\rangle\,\mathrm{d}t + \gamma(t)\sqrt{I\,(B(\mathfrak{X}_t))}\langle \nabla\varphi(\mathfrak{X}_t), (\nabla\psi(\mathfrak{X}_t))^\top \sqrt{K}\,\mathrm{d}\mathfrak{B}_t\rangle \\
&+ \frac{\gamma(t)^2}{2} I\,(B(\mathfrak{X}_t))\,\langle \nabla^2\varphi(\mathfrak{X}_t), \left((\nabla\psi(\mathfrak{X}_t))^\top \sqrt{K}\,\mathrm{d}\mathfrak{B}_t\right)^{\otimes 2}\rangle.
\end{aligned} \tag{B.45}$$

We seek to simplify some of the terms in (B.45). For this, we flatten the second term in sum:

$$\langle \nabla\varphi(\mathfrak{X}_t), (\nabla\psi(\mathfrak{X}_t))^\top \sqrt{K}\,\mathrm{d}\mathfrak{B}_t\rangle = \langle (\nabla\psi(\mathfrak{X}_t))^\top \sqrt{K}, \nabla\varphi(\mathfrak{X}_t)\,(\mathrm{d}\mathfrak{B}_t)^\top\rangle_{\mathbb{R}^{2d\times d}}. \tag{B.46}$$

Next, we look at the second derivative term of φ,

$$\langle \nabla^2\varphi(\mathcal{X}_t), \left((\nabla\psi(\mathcal{X}_t))^\top \sqrt{K}\,\mathrm{d}\mathfrak{B}_t\right)^{\otimes 2}\rangle = \langle \nabla^2\varphi(\mathcal{X}_t), (\nabla\psi(\mathcal{X}_t))^\top K\nabla\psi(\mathcal{X}_t)\rangle\,\mathrm{d}t, \tag{B.47}$$

where we used the symmetry of $\sqrt{K}$. With this, we can now identify the martingale increment for homogenized SGD,

$$\begin{aligned}\mathrm{d}\varphi(\mathcal{X}_t) =& -\gamma(t)d\langle\nabla\varphi(\mathcal{X}_t), \nabla\mathcal{R}(\mathcal{X}_t)\rangle\,\mathrm{d}t \\ &+ \frac{\gamma(t)^2}{2} I\left(B(\mathcal{X}_t)\right)\langle\nabla^2\varphi(\mathcal{X}_t), (\nabla\psi(\mathcal{X}_t))^\top K\nabla\psi(\mathcal{X}_t)\rangle\,\mathrm{d}t + \mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}(\varphi),\end{aligned}$$

where $\mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}(\varphi) := \gamma(t)\sqrt{I\left(B(\mathcal{X}_t)\right)}\langle(\nabla\psi(\mathcal{X}_t))^\top\sqrt{K}, \nabla\varphi(\mathcal{X}_t)\left(\mathrm{d}\mathfrak{B}_t\right)^\top\rangle_{\mathbb{R}^{2d\times d}}$. By integrating, we derive the Doob decomposition for $\varphi(\mathcal{X}_t)$

$$\begin{aligned}\varphi(\mathcal{X}_t) =& \varphi(x_0) - \int_0^t \gamma(s)d\langle\nabla\varphi(\mathcal{X}_s), \nabla\mathcal{R}(\mathcal{X}_s)\rangle\,\mathrm{d}s \\ &+ \frac{1}{2}\int_0^t \gamma(s)^2 I\left(B(\mathcal{X}_s)\right)\langle\nabla^2\varphi(\mathcal{X}_s), (\nabla\psi(\mathcal{X}_s))^\top K\nabla\psi(\mathcal{X}_s)\rangle\,\mathrm{d}s + \int_0^t \mathrm{d}\mathcal{M}_s^{\mathrm{HSGD}}(\varphi).\end{aligned} \tag{B.48}$$

$S(\mathcal{X}_t, z)$ is an Approximate Solution, Proof of Proposition B.3.2

The goal in this section is to prove Proposition B.3.2, that is, show that

$$S(\mathcal{X}_t, z) = \frac{1}{d}\mathcal{W}_t^\top\Omega(x, z)\mathcal{W}_t \in \mathbb{C}^{3\times 3}$$

is an approximate solution (B.2.2) to the partial integro-differential equation in (B.6).

The first step is to derive a closed equation for $S(\mathcal{X}_t, z)$ using Itô calculus.

Itô calculus applied to $S(\mathcal{X}_t, z)$. Recall the expected risk $\mathcal{R}$ can be expressed as a composition, $\mathcal{R}(\mathcal{X}_t) = h\left(B(\mathcal{X}_t)\right)$, for some function $h\colon \mathbb{R}^{3\times 3} \to \mathbb{R}$ and

$$B(\mathcal{X}_t) = \frac{1}{d}\mathcal{W}_t^\top K\mathcal{W}_t \in \mathbb{R}^{3\times 3}.$$

A straightforward application of the chain rule shows that

$$\begin{aligned}\mathbb{R}^{2d} \ni \nabla\mathcal{R}(\mathcal{X}_t) &= \langle\nabla B(\mathcal{X}_t), \nabla h\left(B(\mathcal{X}_t)\right)\rangle_{\mathbb{R}^{3\times 3}} \\ &= \frac{2}{d}\left(\nabla h_{11}\cdot(\nabla\psi(\mathcal{X}_t))^\top K\psi(\mathcal{X}_t) + \nabla h_{21}\cdot(\nabla\psi(\mathcal{X}_t))^\top K\beta^* + \nabla h_{31}\cdot(\nabla\psi(\mathcal{X}_t))^\top K\mathbb{1}_d\right).\end{aligned}$$

Using the product rule for Itô derivatives, we obtain

$$\begin{aligned}
\mathrm{d}S =& -\gamma(t)d\langle\nabla S(\mathcal{X}_t,z),\nabla\mathcal{R}(\mathcal{X}_t)\rangle\,\mathrm{d}t+\mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}(S)\\
&+\frac{\gamma(t)^2}{2}I\left(B(\mathcal{X}_t)\right)\langle\nabla^2 S(\mathcal{X}_t,z),(\nabla\psi(\mathcal{X}_t))^\top K\nabla\psi(\mathcal{X}_t)\rangle\,\mathrm{d}t\\
=&-\gamma(t)d\left(\langle\nabla_u S(\mathcal{X}_t,z),\nabla_u\mathcal{R}(\mathcal{X}_t)\rangle+\langle\nabla_v S(\mathcal{X}_t,z),\nabla_v\mathcal{R}(\mathcal{X}_t)\rangle\right)\mathrm{d}t+\mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}(S)\\
&+\frac{\gamma(t)^2}{2}I\left(B(\mathcal{X}_t)\right)\Big[\langle\nabla_u^2 S(\mathcal{X}_t,z),K\left(\nabla_u\psi(\mathcal{X}_t)\right)^2\rangle\\
&+2\langle\nabla_{uv}^2 S(\mathcal{X}_t,z),\left(\nabla_u\psi(\mathcal{X}_t)\right)^\top K\nabla_v\psi(\mathcal{X}_t)\rangle+\langle\nabla_v^2 S(\mathcal{X}_t,z),K\left(\nabla_v\psi(\mathcal{X}_t)\right)^2\rangle\Big]\mathrm{d}t.
\end{aligned}\tag{B.49}$$

Remark B.3.3 (Applying scalar formulas to S)**.** The statistic $S(x,z)$ is complex-valued:

$$S(x,z)\in\mathbb{C}^{3\times 3}.$$

However, all scalar statistic formulas above were stated for real-valued functions $\varphi:\mathbb{R}^{2d}\to\mathbb{R}$. Therefore, whenever we apply these formulas to S, we apply them to the real-valued coordinate functions

$$S_{ab}^{\mathrm{Re},z}(x):=\operatorname{Re}S_{ab}(x,z),\qquad S_{ab}^{\mathrm{Im},z}(x):=\operatorname{Im}S_{ab}(x,z),$$

for each $a,b\in\{1,2,3\}$ and fixed $z\in\Gamma$.

For example, we define the complex-valued HSGD martingale associated with S entrywise by

$$\left(\mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}(S)(z)\right)_{ab}:=\mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}\left(S_{ab}^{\mathrm{Re},z}\right)+i\,\mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}\left(S_{ab}^{\mathrm{Im},z}\right)$$

for

$$\mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}(S_{ab}^{\mathrm{Re},z}):=\gamma(t)\sqrt{I\left(B(\mathcal{X}_t)\right)}\langle(\nabla\psi(\mathcal{X}_t))^\top\sqrt{K},\nabla S_{ab}^{\mathrm{Re},z}(\mathcal{X}_t,z)\left(\mathrm{d}\mathfrak{B}_t\right)^\top\rangle_{\mathbb{R}^{2d\times d}},$$

and analogously for $S_{ab}^{\mathrm{Im},z}$.

Thus matrix-valued stochastic identities involving S are shorthand for the collection of real-valued identities for the real and imaginary parts of its entries.

Similarly, we define

$$\mathcal{M}_t^{\mathrm{HSGD}}(S)=\int_0^t\mathrm{d}\mathcal{M}_s^{\mathrm{HSGD}}(S).$$

We consider the first term in the summation above, and plugging in $\nabla\mathcal{R}$ and using Lemma B.1.3, we have

$$\begin{aligned}
\mathbb{C}^{3\times 3} \ni \langle \nabla_u S(\mathcal{X}_t, z), \nabla_u \mathcal{R}(\mathcal{X}_t)\rangle_{\mathbb{R}^d} &= \frac{2}{d^2}(\nabla h_{11}\cdot\psi(\mathcal{X}_t) + \nabla h_{21}\cdot\beta^* + \nabla h_{31}\cdot \mathbb{1}_d)^\top\cdot \\
&\qquad (\nabla_u\psi(\mathcal{X}_t))^2 K\Omega \begin{bmatrix} \psi(\mathcal{X}_t) & \beta^* & \mathbb{1}_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
&\quad + \frac{2}{d^2}(\nabla h_{11}\cdot\psi(\mathcal{X}_t) + \nabla h_{21}\cdot\beta^* + \nabla h_{31}\cdot \mathbb{1}_d)^\top\cdot \\
&\qquad (\nabla_u\psi(\mathcal{X}_t))^2 K\Omega \begin{bmatrix} \psi(\mathcal{X}_t) & 0 & 0 \\ \beta^* & 0 & 0 \\ \mathbb{1}_d & 0 & 0 \end{bmatrix} \\
&\quad + \frac{2}{d^2}(\nabla h_{11}\cdot\psi(\mathcal{X}_t) + \nabla h_{21}\cdot\beta^* + \nabla h_{31}\cdot \mathbb{1}_d)^\top\cdot \\
&\qquad \nabla_u\psi(\mathcal{X}_t)\,\mathrm{diag}(\psi(\mathcal{X}_t))KR\,(z_1;\mathrm{diag}(\mathcal{U}_t))\,\Omega \begin{bmatrix} \psi(\mathcal{X}_t) & \beta^* & \mathbb{1}_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
&\quad + \frac{2}{d^2}(\nabla h_{11}\cdot\psi(\mathcal{X}_t) + \nabla h_{21}\cdot\beta^* + \nabla h_{31}\cdot \mathbb{1}_d)^\top\cdot \\
&\qquad \nabla_u\psi(\mathcal{X}_t)\,\mathrm{diag}(\beta^*)KR\,(z_1;\mathrm{diag}(\mathcal{U}_t))\,\Omega \begin{bmatrix} 0 & 0 & 0 \\ \psi(\mathcal{X}_t) & \beta^* & \mathbb{1}_d \\ 0 & 0 & 0 \end{bmatrix} \\
&\quad + \frac{2}{d^2}(\nabla h_{11}\cdot\psi(\mathcal{X}_t) + \nabla h_{21}\cdot\beta^* + \nabla h_{31}\cdot \mathbb{1}_d)^\top\cdot \\
&\qquad \nabla_u\psi(\mathcal{X}_t)KR\,(z_1;\mathrm{diag}(\mathcal{U}_t))\,\Omega \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \psi(\mathcal{X}_t) & \beta^* & \mathbb{1}_d \end{bmatrix} \\
&= \frac{2}{d}H^\top\cdot\frac{1}{d}\mathcal{W}_t^\top\,(\nabla_u\psi(\mathcal{X}_t))^2\,K\Omega\mathcal{W}_t \\
&\quad + \frac{2}{d^2}\mathcal{W}_t^\top\,(\nabla_u\psi(\mathcal{X}_t))^2\,K\Omega\mathcal{W}_t\cdot H \\
&\quad + \frac{2}{d}H^\top\cdot\frac{1}{d}\mathcal{W}_t^\top\nabla_u\psi(\mathcal{X}_t)\,\mathrm{diag}(\psi(\mathcal{X}_t))KR\,(z_1;\mathrm{diag}(\mathcal{U}_t))\,\Omega\mathcal{W}_t
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{2}{d} E_{21} \cdot H^\top \cdot \frac{1}{d} \mathcal{W}_t^\top \nabla_u \psi(\mathcal{X}_t) \operatorname{diag}(\beta^*) KR\left(z_1; \operatorname{diag}(\mathcal{U}_t)\right) \Omega \mathcal{W}_t \\
&+ \frac{2}{d} E_{31} \cdot H^\top \cdot \frac{1}{d} \mathcal{W}_t^\top \nabla_u \psi(\mathcal{X}_t) KR\left(z_1; \operatorname{diag}(\mathcal{U}_t)\right) \Omega \mathcal{W}_t,
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{C}^{3\times 3} \ni \left\langle \nabla_v S(\mathcal{X}_t, z), \nabla_v \mathcal{R}(\mathcal{X}_t) \right\rangle_{\mathbb{R}^d} &= \frac{2}{d} H^\top \cdot \frac{1}{d} \mathcal{W}_t^\top \left(\nabla_v \psi(\mathcal{X}_t)\right)^2 K \Omega \mathcal{W}_t \\
&\quad + \frac{2}{d^2} \mathcal{W}_t^\top \left(\nabla_v \psi(\mathcal{X}_t)\right)^2 K \Omega \mathcal{W}_t \cdot H \\
&+ \frac{2}{d} H^\top \cdot \frac{1}{d} \mathcal{W}_t^\top \nabla_v \psi(\mathcal{X}_t) \operatorname{diag}(\psi(\mathcal{X}_t)) KR\left(z_2; \operatorname{diag}(\mathcal{V}_t)\right) \Omega \mathcal{W}_t \\
&+ \frac{2}{d} E_{21} \cdot H^\top \cdot \frac{1}{d} \mathcal{W}_t^\top \nabla_v \psi(\mathcal{X}_t) \operatorname{diag}(\beta^*) KR\left(z_2; \operatorname{diag}(\mathcal{V}_t)\right) \Omega \mathcal{W}_t \\
&+ \frac{2}{d} E_{31} \cdot H^\top \cdot \frac{1}{d} \mathcal{W}_t^\top \nabla_v \psi(\mathcal{X}_t) KR\left(z_2; \operatorname{diag}(\mathcal{V}_t)\right) \Omega \mathcal{W}_t,
\end{aligned}
$$

for

$$
H\left(B(\mathcal{X}_t)\right) = \left[\begin{array}{c|c|c} \nabla_{11} h & 0 & 0 \\ \hline \nabla_{21} h & 0 & 0 \\ \hline \nabla_{31} h & 0 & 0 \end{array}\right].
$$

Similarly for the second term we have

$$
\begin{aligned}
\mathbb{C}^{3\times 3} \ni \left\langle \nabla_u^2 S(\mathcal{X}_t, z), K\left(\nabla_u \psi(x)\right)^2 \right\rangle_{\mathbb{R}^{d\times d}} &= \frac{2q_{11}}{d} \mathbb{1}_d^\top \cdot \left(\nabla_u \psi(x)\right)^2 K \Omega \begin{bmatrix} \psi(\mathcal{X}_t) & \beta^* & \mathbb{1}_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
&\quad + \frac{2q_{11}}{d} \mathbb{1}_d^\top \cdot \left(\nabla_u \psi(x)\right)^2 K \Omega \begin{bmatrix} \psi(\mathcal{X}_t) & 0 & 0 \\ \beta^* & 0 & 0 \\ \mathbb{1}_d & 0 & 0 \end{bmatrix} \\
&\quad + \frac{2}{d} \mathbb{1}_d^\top \cdot \left(\nabla_u \psi(x)\right)^3 KR\left(z_1; \operatorname{diag}(\mathcal{U}_t)\right) \Omega \begin{bmatrix} \psi(\mathcal{X}_t) & \beta^* & \mathbb{1}_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
&\quad + \frac{2}{d} \mathbb{1}_d^\top \cdot \left(\nabla_u \psi(x)\right)^3 KR\left(z_1; \operatorname{diag}(\mathcal{U}_t)\right) \Omega \begin{bmatrix} \psi(\mathcal{X}_t) & 0 & 0 \\ \beta^* & 0 & 0 \\ \mathbb{1}_d & 0 & 0 \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{2}{d}\begin{bmatrix}\operatorname{Tr}\left(\left(\nabla_u\psi(x)\right)^4 K\Omega\right) & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}\\
&+\frac{2}{d}(\psi(\mathcal{X}_t))^\top\cdot\left(\nabla_u\psi(x)\right)^2 KR\left(z_1;\operatorname{diag}(\mathcal{U}_t)\right)^2\Omega\begin{bmatrix}\psi(\mathcal{X}_t) & \beta^* & \mathbb{1}_d\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}\\
&+\frac{2}{d}(\beta^*)^\top\cdot\left(\nabla_u\psi(x)\right)^2 KR\left(z_1;\operatorname{diag}(\mathcal{U}_t)\right)^2\Omega\begin{bmatrix}0 & 0 & 0\\ \psi(\mathcal{X}_t) & \beta^* & \mathbb{1}_d\\ 0 & 0 & 0\end{bmatrix}\\
&+\frac{2}{d}\mathbb{1}_d^\top\cdot\left(\nabla_u\psi(x)\right)^2 KR\left(z_1;\operatorname{diag}(\mathcal{U}_t)\right)^2\Omega\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ \psi(\mathcal{X}_t) & \beta^* & \mathbb{1}_d\end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
&= 2q_{11}E_{13}\cdot\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)^2 K\Omega\mathcal{W}_t\\
&+2q_{11}\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)^2 K\Omega\mathcal{W}_t\cdot E_{31}\\
&+2E_{13}\cdot\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)^3 KR\left(z_1;\operatorname{diag}(\mathcal{U}_t)\right)\Omega\mathcal{W}_t\\
&+2\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)^3 KR\left(z_1;\operatorname{diag}(\mathcal{U}_t)\right)\Omega\mathcal{W}_t\cdot E_{31}\\
&+2E_{13}\cdot\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)^4 K\Omega\mathcal{W}_t\cdot E_{31}\\
&+2\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(x)\right)^2 KR\left(z_1;\operatorname{diag}(\mathcal{U}_t)\right)^2\Omega\mathcal{W}_t,
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{C}^{3\times3}\ni\left\langle\nabla^2_{uv}S(\mathcal{X}_t,z),\left(\nabla_u\psi(x)\right)^\top K\nabla_v\psi(x)\right\rangle_{\mathbb{R}^{d\times d}}=&\\
&2q_{12}E_{13}\cdot\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)\left(\nabla_v\psi(\mathcal{X}_t)\right)K\Omega\mathcal{W}_t\\
&+2q_{12}\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)\left(\nabla_v\psi(\mathcal{X}_t)\right)K\Omega\mathcal{W}_t\cdot E_{31}\\
&+E_{13}\cdot\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)\left(\nabla_v\psi(\mathcal{X}_t)\right)^2 KR\left(z_1;\operatorname{diag}(\mathcal{U}_t)\right)\Omega\mathcal{W}_t\\
&+\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)\left(\nabla_v\psi(\mathcal{X}_t)\right)^2 KR\left(z_1;\operatorname{diag}(\mathcal{U}_t)\right)\Omega\mathcal{W}_t\cdot E_{31}
\end{aligned}
$$

$$
\begin{aligned}
&+ E_{13} \cdot \frac{1}{d} \mathcal{W}_t^\top \left(\nabla_u \psi(\mathcal{X}_t)\right)^2 \left(\nabla_v \psi(\mathcal{X}_t)\right) KR\left(z_2; \operatorname{diag}(\mathcal{V}_t)\right) \Omega \mathcal{W}_t \\
&+ \frac{1}{d} \mathcal{W}_t^\top \left(\nabla_u \psi(\mathcal{X}_t)\right)^2 \left(\nabla_v \psi(\mathcal{X}_t)\right) KR\left(z_2; \operatorname{diag}(\mathcal{V}_t)\right) \Omega \mathcal{W}_t \cdot E_{31} \\
&+ 2E_{13} \cdot \frac{1}{d} \mathcal{W}_t^\top \left(\nabla_u \psi(\mathcal{X}_t)\right)^2 \left(\nabla_v \psi(\mathcal{X}_t)\right)^2 K \Omega \mathcal{W}_t \cdot E_{31} \\
&+ \frac{1}{d} \mathcal{W}_t^\top \left(\nabla_u \psi(\mathcal{X}_t)\right) \left(\nabla_v \psi(\mathcal{X}_t)\right) KR\left(z_1; \operatorname{diag}(\mathcal{U}_t)\right) R\left(z_2; \operatorname{diag}(\mathcal{V}_t)\right) \Omega \mathcal{W}_t,
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{C}^{3\times 3} \ni \left\langle \nabla_v^2 S(\mathcal{X}_t, z), K\left(\nabla_v \psi(x)\right)^2 \right\rangle_{\mathbb{R}^{d\times d}} &= 2q_{22} E_{13} \cdot \frac{1}{d} \mathcal{W}_t^\top \left(\nabla_v \psi(\mathcal{X}_t)\right)^2 K \Omega \mathcal{W}_t \\
&+ 2q_{22} \frac{1}{d} \mathcal{W}_t^\top \left(\nabla_v \psi(\mathcal{X}_t)\right)^2 K \Omega \mathcal{W}_t \cdot E_{31} \\
&+ 2E_{13} \cdot \frac{1}{d} \mathcal{W}_t^\top \left(\nabla_v \psi(\mathcal{X}_t)\right)^3 KR\left(z_2; \operatorname{diag}(\mathcal{V}_t)\right) \Omega \mathcal{W}_t \\
&+ 2 \frac{1}{d} \mathcal{W}_t^\top \left(\nabla_v \psi(\mathcal{X}_t)\right)^3 KR\left(z_2; \operatorname{diag}(\mathcal{V}_t)\right) \Omega \mathcal{W}_t \cdot E_{31} \\
&+ 2E_{13} \cdot \frac{1}{d} \mathcal{W}_t^\top \left(\nabla_v \psi(\mathcal{X}_t)\right)^4 K \Omega \mathcal{W}_t \cdot E_{31} \\
&+ 2 \frac{1}{d} \mathcal{W}_t^\top \left(\nabla_v \psi(x)\right)^2 KR\left(z_2; \operatorname{diag}(\mathcal{V}_t)\right)^2 \Omega \mathcal{W}_t.
\end{aligned}
$$

Now recall that

$$
\nabla \psi(\mathcal{X}_t) = \left[2q_{11} \operatorname{diag}(\mathcal{U}_t) + 2q_{12} \operatorname{diag}(\mathcal{V}_t) + l_1 I_d \; 2q_{12} \operatorname{diag}(\mathcal{U}_t) + 2q_{22} \operatorname{diag}(\mathcal{V}_t) + l_2 I_d\right] \in \mathbb{R}^{d\times 2d}.
$$

Accordingly, each term in (B.49) takes a form such as

$$
\begin{aligned}
&- 2\gamma(t) H^\top \cdot \frac{1}{d} \mathcal{W}_t^\top \operatorname{diag}(\mathcal{U}_t)^{m_1} R\big(z_1; \operatorname{diag}(\mathcal{U}_t)\big)^{\rho_1} \\
&\quad \cdot \operatorname{diag}(\mathcal{V}_t)^{m_2} R\big(z_2; \operatorname{diag}(\mathcal{V}_t)\big)^{\rho_2} \operatorname{diag}(\beta^*)^{m_3} R\big(z_3; \operatorname{diag}(\beta^*)\big) KR\big(z_4; K\big) \mathcal{W}_t, \\
\text{or} \quad &C(Q)\gamma(t)^2 I\big(B(\mathcal{X}_t)\big) \cdot \frac{1}{d} \mathcal{W}_t^\top \operatorname{diag}(\mathcal{U}_t)^{m_1} R\big(z_1; \operatorname{diag}(\mathcal{U}_t)\big)^{\rho_1} \\
&\quad \cdot \operatorname{diag}(\mathcal{V}_t)^{m_2} R\big(z_2; \operatorname{diag}(\mathcal{V}_t)\big)^{\rho_2} R\big(z_3; \operatorname{diag}(\beta^*)\big) KR\big(z_4; K\big) \mathcal{W}_t.
\end{aligned}
$$

Expanding these terms and applying the Cauchy integral formula along each contour Γ_i, we write

$$
q(D_i) = \frac{1}{2\pi i} \oint_{\Gamma_i} q(z_i) R\left(z_i; D_i\right) \, \mathrm{d}z_i,
$$

together with the standard resolvent identities,

$$
R\left(z_i; D_i\right)^2 = -\frac{\mathrm{d}}{\mathrm{d}z_i} R\left(z_i; D_i\right) \quad \text{and} \quad R\left(z_i; D_i\right)^3 = \frac{1}{2} \cdot \frac{\mathrm{d}^2}{\mathrm{d}z_i^2} R\left(z_i; D_i\right).
$$

It then follows that

$$\mathrm{d}S(\mathcal{X}_t, z) = \mathcal{F}\left(z, S(\mathcal{X}_t, z)\right)\mathrm{d}t + \mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}(S), \tag{B.50}$$

with $\mathcal{X}_0 = x_0$, where the summands of $\mathcal{F}\left(z, S(\mathcal{X}_t, z)\right)$ are of the form

$$\begin{gathered} -2\gamma(t)H^\top \cdot \frac{1}{d}\mathcal{W}_t^\top z^\zeta \left(\bigcirc_{i=1}^4 \mathcal{G}_i\right)\left(R\left(z_i; D_i\right)\right)\mathcal{W}_t \quad \text{or} \\ C(Q)\gamma(t)^2 I(B(\mathcal{X}_t)) \cdot \frac{1}{d}\mathcal{W}_t^\top z^\zeta \left(\bigcirc_{i=1}^4 \mathcal{G}_i\right)\left(R\left(z_i; D_i\right)\right)\mathcal{W}_t, \end{gathered} \tag{B.51}$$

for

$$\mathcal{G}_i(R\left(z_i; D_i\right)) = \frac{1}{2\pi i}\oint_{\Gamma_i} q(z_i)R\left(z_i; D_i\right)\,\mathrm{d}z_i, \quad -\frac{\mathrm{d}}{\mathrm{d}z_i}R\left(z_i; D_i\right), \quad \text{or} \quad \frac{1}{2}\cdot\frac{\mathrm{d}^2}{\mathrm{d}z_i^2}R\left(z_i; D_i\right). \tag{B.52}$$

We now prove Proposition B.3.2.

Proof of Proposition B.3.2. By Itô's Lemma, we have seen that

$$S(\mathcal{X}_t, \cdot) = S(x_0, \cdot) + \int_0^t \mathcal{F}\left(\cdot, S(\mathcal{X}_s, \cdot)\right)\mathrm{d}s + \int_0^t \mathrm{d}\mathcal{M}_s^{\mathrm{HSGD}}(S(\mathcal{X}_s, \cdot)).$$

Thus to show that $S(\mathcal{X}_t, \cdot)$ is an approximate solution of the partial integro-differential equation (B.6) it amounts to bounding the martingale term where C is a positive constant independent of d. For all $z \in \Gamma$, we note that for some constants $C, c > 0$ we have $\vartheta_{c\cdot M} \leqslant \hat{\tau}_M \leqslant \vartheta_{C\cdot M}$ (see Lemma B.2.1). Consequently, we can work with the stopped process $\mathcal{X}_t^\vartheta = \mathcal{X}_{t\wedge\vartheta}$ instead of using $\hat{\tau}_M$. We thus have that for all $z \in \Gamma$

$$\sup_{0\leqslant t\leqslant(\hat{\tau}_M\wedge T)}\left\|S(\mathcal{X}_t, z) - S(x_0, z) - \int_0^t \mathcal{F}\left(z, S(\mathcal{X}_s, z)\right)\,\mathrm{d}s\right\| \leqslant \sup_{0\leqslant t\leqslant(\vartheta_{C\cdot M}\wedge T)}\left\|\mathcal{M}_t^{\mathrm{HSGD}}(S(\cdot, z))\right\|.$$

Fix a constant $\delta > 0$. Let $\Gamma^\delta = \Gamma_1^\delta \times \Gamma_2^\delta \times \Gamma_3^\delta \times \Gamma_4^\delta$ where each Γ_i^δ is a $d^{-\delta}$-mesh of Γ_i with $|\Gamma^\delta| \leqslant C_\Gamma d^{4\delta}$ for positive $C_\Gamma > 0$ depending on $\|K\|_{\mathrm{op}}$ and $\|\beta^*\|_\infty$.

By the martingale error proposition, Proposition B.3.8, which we have deferred the proof to Section B.3.3, we have that for any $\hat{\delta} > 0$

$$\sup_{0\leqslant t\leqslant T}\left\|\mathcal{M}_{t\wedge\vartheta_{C\cdot M}}^{\mathrm{HSGD}}(S(\cdot, z))\right\| \leqslant CL(f)d^{\hat{\delta}/2-1/2} \quad \text{w.o.p.}$$

By a union bound over $z \in \Gamma^\delta$, using $|\Gamma^\delta| \leqslant C_\Gamma d^{4\delta}$, the martingale bound holds uniformly on Γ^δ with overwhelming probability:

$$\sup_{z\in\Gamma^\delta}\sup_{0\leqslant t\leqslant T}\left\|\mathcal{M}_{t\wedge\vartheta_{C\cdot M}}^{\mathrm{HSGD}}(S(\cdot, z))\right\| \leqslant CL(f)d^{\hat{\delta}/2-1/2} \quad \text{w.o.p.}$$

Consequently, we deduce that

$$\sup_{0\leqslant t\leqslant(\hat{\tau}_M\wedge T)}\left\|S(\mathcal{X}_t,z)-S(x_0,z)-\int_0^t\mathcal{F}\left(z,S(\mathcal{X}_s,z)\right)\,\mathrm{d}s\right\|_{\Gamma^\delta}$$

$$\leqslant\sup_{0\leqslant t\leqslant(\vartheta_{C\cdot M}\wedge T)}\left\|\mathcal{M}_t^{\mathrm{HSGD}}(S(\cdot,z))\right\|_{\Gamma^\delta}$$

$$\leqslant CL(f)d^{\hat{\delta}/2-1/2}\quad\text{w.o.p.}$$

An application of the net argument, Lemma B.3.1, finishes the proof after setting $\hat{\delta}=1-2\delta$.

The derivative regularity condition (i) is automatic for $t\mapsto S(\mathcal{X}_t,\cdot)$: for each fixed trajectory, the dependence on z is through a finite product of resolvents on the fixed contour Γ, and the resolvent derivative bounds from Remark B.2.3 give the required Lipschitz control up to $\hat{\tau}_M$. Thus the estimate above verifies condition (ii), and hence $S(\mathcal{X}_t,\cdot)$ is an $(d^{-\varepsilon},M,T)$-approximate solution. □

B.3.2 SGD Under the Resolvent Statistic S

In this section, we show that $S\left(x_{\lfloor td\rfloor},z\right)$ is an approximate solution (B.2.2) to the partial integro-differential equation (B.6) which we state below.

Proposition B.3.4 (SGD is an approximate solution)**.** *Fix a $T,M>0$ and $0<\delta<1/2$. Then $S\left(x_{\lfloor td\rfloor},z\right)$ is a $(d^{-\delta},M,T)$-approximate solution w.o.p., that is,*

$$\sup_{0\leqslant t\leqslant(\hat{\tau}_M\wedge T)}\left\|S(x_{\lfloor td\rfloor},z)-S(x_0,z)-\int_0^t\mathcal{F}\left(z,S(x_{\lfloor sd\rfloor},z)\right)\,\mathrm{d}s\right\|\leqslant d^{-\delta}\quad\textit{w.o.p.}$$

The proof of this Proposition is deferred to Section B.3.2.

Recall our iterates satisfy the recurrence

$$x_{k+1}=x_k-\gamma_k\,\nabla_x\Psi(x_k;a_{k+1})=x_k-\frac{\gamma_k}{\sqrt{d}}\,\nabla_{r_1}f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top\cdot a_{k+1}\tag{B.53}$$

with $r_k:=\frac{1}{\sqrt{d}}\begin{pmatrix}\psi(x_k)\\ \beta^*\end{pmatrix}^\top a_{k+1}$.

The derivation follows the martingale-decomposition strategy used in [21]. We first derive a one-step martingale decomposition for a smooth scalar statistic $\varphi:\mathbb{R}^{2d}\to\mathbb{R}$. We later

apply this decomposition to the real-valued coordinate functions $S_{ab}^{\mathrm{Re},z}$ and $S_{ab}^{\mathrm{Im},z}$ and then recombine the resulting identities to obtain the complex matrix-valued decomposition for $S(x_{\lfloor td \rfloor}, z)$.

Applying Taylor's expansion, we have

$$\begin{aligned}\varphi(x_{k+1}) &= \varphi(x_k) - \gamma_k \langle \nabla\varphi(x_k), \nabla_x \Psi(x_k; a_{k+1})\rangle + \frac{\gamma_k^2}{2}\langle \nabla^2\varphi(x_k), (\nabla_x \Psi(x_k; a_{k+1}))^{\otimes 2}\rangle + \dots \\ &= \varphi(x_k) - \frac{\gamma_k}{\sqrt{d}} \left\langle \nabla\varphi(x_k), \nabla_{r_1} f\left(r_k\right) \cdot \left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1}\right\rangle \\ &+ \frac{\gamma_k^2}{2d} \left\langle \nabla^2\varphi(x_k), \left(\nabla_{r_1} f\left(r_k\right) \cdot \left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1}\right)^{\otimes 2}\right\rangle + \mathcal{E}_k^{\mathrm{High}}(\varphi).\end{aligned} \tag{B.54}$$

We will show in Section B.3.3 that the third and higher-order terms, $\mathcal{E}_k^{\mathrm{High}}(\varphi)$, vanish when $d \to \infty$.

Doob Decomposition for SGD

Recall that, for a Hessian A and vector v,

$$\langle A, v^{\otimes 2}\rangle = v^\top A v.$$

To write the Doob decomposition, the idea is to condition on $r_k = \frac{1}{\sqrt{d}} \begin{pmatrix} \psi(x_k) \\ \beta^* \end{pmatrix}^\top a_{k+1}$ and $W_k = [\,\psi(x_k)\,|\,\beta^*\,|\,\mathbb{1}_d\,] \in \mathbb{R}^{d\times 3}$. For this, we will introduce some notation. Define the σ-algebras

$$\mathcal{F}_k := \sigma\left(\{W_i\}_{i=0}^k\right) \subset \mathcal{G}_k := \sigma\left(\{r_i\}_{i=0}^k, \{W_i\}_{i=0}^k\right).$$

Gradient term in Taylor expansion. First, consider the conditional expectation with respect to $\mathcal{F}_k$ of the gradient term in equation (B.54). Note we can safely assume the interchange of the gradient and the expectation in the risk formula $\mathcal{R}$ since a follows a Gaussian distribution, and generally, this assumption holds true. Therefore, we have

$$\frac{1}{\sqrt{d}}\, \mathbb{E}_{a_{k+1}}\left[\left\langle \nabla\varphi(x_k), \nabla_{r_1} f\left(r_k\right) \cdot \left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1}\right\rangle \middle| \mathcal{F}_k\right] = \langle \nabla\varphi(x_k), \nabla\mathcal{R}(x_k)\rangle.$$

By applying Doob's decomposition in this context, the gradient term takes the form

$$\frac{\gamma_k}{\sqrt{d}} \left\langle \nabla\varphi(x_k), \nabla_{r_1} f\left(r_k\right) \cdot \left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1}\right\rangle = \gamma_k \left\langle \nabla\varphi(x_k), \nabla\mathcal{R}(x_k)\right\rangle - \Delta\mathcal{M}_k^{\mathrm{Grad}}(\varphi) \tag{B.55}$$

where $\Delta\mathcal{M}_k^{\mathrm{Grad}}(\varphi)$ represents a martingale increment that tends to zero as d tends to infinity.

Hessian term in the Taylor expansion. Proceeding to the Hessian term in (B.54), we initiate the analysis by conditioning first on $\mathcal{G}_k$ and subsequently on $\mathcal{F}_k$, using the tower property. A simple computation leads to

$$\langle \nabla^2\varphi(x_k), \left(\nabla_{r_1} f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top\cdot a_{k+1}\right)^{\otimes 2}\rangle = \langle \mathcal{B}, a_{k+1}^{\otimes 2}\rangle,$$

where $\mathcal{B} := \nabla_{r_1} f\left(r_k\right)^2 \nabla\psi(x_k)\nabla^2\varphi(x_k)\left(\nabla\psi(x_k)\right)^\top$. To proceed, we must evaluate

$$\begin{aligned}\mathbb{E}_{a_{k+1}}\left[\langle \mathcal{B}, a_{k+1}^{\otimes 2}\rangle \middle| \mathcal{G}_k\right] &= \langle \mathcal{B}, \mathbb{E}_{a_{k+1}}\left[a_{k+1}^{\otimes 2}\middle|\mathcal{G}_k\right]\rangle \\ &= \langle \mathcal{B}, \mathbb{E}_{a_{k+1}}\left[\left(a_{k+1}-\mathbb{E}_{a_{k+1}}\left[a_{k+1}|\mathcal{G}_k\right]\right)^{\otimes 2}\middle|\mathcal{G}_k\right]\rangle \\ &\quad + \langle \mathcal{B}, \left(\mathbb{E}_{a_{k+1}}\left[a_{k+1}|\mathcal{G}_k\right]\right)^{\otimes 2}\rangle.\end{aligned} \tag{B.56}$$

To establish the conditional distribution of a_{k+1} given r_k, we invoke a standard conditioning lemma, see e.g. [21].

Lemma B.3.5 (Gaussian conditioning)**.** *Let $v\sim\mathcal{N}(0,I_d)$ and let $Q\in\mathbb{R}^{d\times 2}$ have orthonormal columns. Then, conditional on $Q^\top v$, we have*

$$v \overset{\mathcal{D}}{=} (I_d - QQ^\top)g + QQ^\top v$$

where $g\sim\mathcal{N}(0,I_d)$ is independent of $Q^\top v$. In particular,

$$\mathbb{E}[v \mid Q^\top v] = QQ^\top v, \qquad \mathrm{Cov}(v \mid Q^\top v) = I_d - QQ^\top.$$

By Assumption 3.1.5, we write $a_{k+1} = \sqrt{K}v_k$ with $v_k\sim\mathcal{N}(0,I_d)$, and express $\sqrt{K}\begin{pmatrix}\psi(x_k)\\ \beta^*\end{pmatrix}$ as Q_kR_k, where $Q_k\in\mathbb{R}^{d\times 2}$ is orthogonal and $R_k\in\mathbb{R}^{2\times 2}$ is upper triangular and invertible (obtained through QR decomposition, assuming $2\ll d$). Observe here $\Pi_k := Q_kQ_k^\top$ has rank 2.

Consequently, we can simplify $\begin{pmatrix}\psi(x_k)\\ \beta^*\end{pmatrix}^\top a_{k+1}$ as follows

$$\begin{pmatrix}\psi(x_k)\\ \beta^*\end{pmatrix}^\top a_{k+1} = \begin{pmatrix}\psi(x_k)\\ \beta^*\end{pmatrix}^\top \sqrt{K}v_k = R_k^\top Q_k^\top v_k.$$

Applying the conditioning lemma, we derive

$$
\begin{aligned}
a_{k+1}\Big|\begin{pmatrix}\psi(x_k)\\ \beta^*\end{pmatrix}^\top a_{k+1} &\overset{\mathcal{D}}{=} \sqrt{K}v_k|Q_k^\top v_k \\
&\overset{\mathcal{D}}{=} \sqrt{K}\left((I_d - Q_kQ_k^\top)g_k + Q_kQ_k^\top v_k\right) \\
&= \sqrt{K}\left(g_k - \Pi_k g_k\right) + \sqrt{K}\Pi_k v_k,
\end{aligned}
$$

where $g_k \sim \mathcal{N}(0, I_d)$ is independent of $Q_k^\top v_k$. From this, we have that

$$
\mathbb{E}_{a_{k+1}}\left[a_{k+1}|\mathcal{G}_k\right] = \sqrt{K}\Pi_k v_k \quad \text{where } v_k \sim \mathcal{N}(0, I_d). \tag{B.57}
$$

Moreover, the conditional covariance of a_{k+1} is precisely

$$
\mathbb{E}_{a_{k+1}}\left[\left(a_{k+1} - \mathbb{E}_{a_{k+1}}\left[a_{k+1}|\mathcal{G}_k\right]\right)^{\otimes 2}\Big|\mathcal{G}_k\right] = \sqrt{K}\left(I_d - \Pi_k\right)\sqrt{K}. \tag{B.58}
$$

Thus, from (B.56) we have

$$
\mathbb{E}_{a_{k+1}}\left[\langle\mathcal{B}, a_{k+1}^{\otimes 2}\rangle\big|\mathcal{G}_k\right] = \langle\mathcal{B}, K\rangle - \langle\mathcal{B}, \sqrt{K}\Pi_k\sqrt{K}\rangle + \langle\mathcal{B}, \left(\sqrt{K}\Pi_k v_k\right)^{\otimes 2}\rangle. \tag{B.59}
$$

We will later see, in Section B.3.3, that the term

$$
\mathcal{E}_k^{\text{Hess}}(\varphi) := -\frac{\gamma_k^2}{2d}\langle\mathcal{B}, \sqrt{K}\Pi_k\sqrt{K}\rangle + \frac{\gamma_k^2}{2d}\langle\mathcal{B}, \left(\sqrt{K}\Pi_k v_k\right)^{\otimes 2}\rangle
$$

is of lower order in expectation and will disappear as $d \to \infty$. So we may write

$$
\frac{\gamma_k^2}{2d}\mathbb{E}_{a_{k+1}}\left[\langle\mathcal{B}, a_{k+1}^{\otimes 2}\rangle\big|\mathcal{F}_k\right] = \frac{\gamma_k^2}{2d}\mathbb{E}_{a_{k+1}}\left[\langle\mathcal{B}, K\rangle|\mathcal{F}_k\right] + \mathbb{E}_{a_{k+1}}\left[\mathcal{E}_k^{\text{Hess}}(\varphi)\big|\mathcal{F}_k\right].
$$

Moreover, observe that

$$
\begin{aligned}
\frac{\gamma_k^2}{2d}\mathbb{E}_{a_{k+1}}\left[\langle\mathcal{B}, K\rangle|\mathcal{F}_k\right] &= \frac{\gamma_k^2}{2d}\mathbb{E}_{a_{k+1}}\left[\nabla_{r_1} f\left(r_k\right)^2\Big|\mathcal{F}_k\right]\langle\nabla\psi(x_k)\nabla^2\varphi(x_k)\left(\nabla\psi(x_k)\right)^\top, K\rangle \\
&= \frac{\gamma_k^2}{2d} I(B(x_k))\,\langle\nabla\psi(x_k)\nabla^2\varphi(x_k)\left(\nabla\psi(x_k)\right)^\top, K\rangle \\
&= \frac{\gamma_k^2}{2d} I(B(x_k))\,\langle\nabla^2\varphi(x_k), \langle K, \left(\nabla\psi(x_k)\right)^{\otimes 2}\rangle\rangle.
\end{aligned} \tag{B.60}
$$

By applying Doob's decomposition in this context, the Hessian term takes the form

$$
\begin{aligned}
&\frac{\gamma_k^2}{2d}\langle\nabla^2\varphi(x_k), \left(\nabla_{r_1} f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top\cdot a_{k+1}\right)^{\otimes 2}\rangle = \\
&\frac{\gamma_k^2}{2d} I(B(x_k))\,\langle\nabla^2\varphi(x_k), \langle K, \left(\nabla\psi(x_k)\right)^{\otimes 2}\rangle\rangle + \Delta\mathcal{M}_k^{\text{Hess}}(\varphi) + \mathbb{E}_{a_{k+1}}\left[\mathcal{E}_k^{\text{Hess}}(\varphi)\big|\mathcal{F}_k\right]
\end{aligned} \tag{B.61}
$$

where $\Delta\mathcal{M}_k^{\mathrm{Hess}}(\varphi)$ represents a martingale increment that tends to zero as d tends to infinity.

We have successfully identified the martingale increments of a *single* update of SGD, that is, by (B.55) and (B.61) in the Taylor expansion (B.54)

$$\begin{aligned}\varphi(x_{k+1}) &= \varphi(x_k) - \gamma_k \langle \nabla\varphi(x_k), \nabla\mathcal{R}(x_k)\rangle + \Delta\mathcal{M}_k^{\mathrm{Grad}}(\varphi) \\ &\quad + \frac{\gamma_k^2}{2d} I(B(x_k)) \langle \nabla^2\varphi(x_k), \langle K, (\nabla\psi(x_k))^{\otimes 2}\rangle\rangle + \Delta\mathcal{M}_k^{\mathrm{Hess}}(\varphi) \\ &\quad + \mathbb{E}_{a_{k+1}}\left[\mathcal{E}_k^{\mathrm{Hess}}(\varphi)\middle|\mathcal{F}_k\right] + \mathbb{E}_{a_{k+1}}\left[\mathcal{E}_k^{\mathrm{High}}(\varphi)\middle|\mathcal{F}_k\right],\end{aligned} \tag{B.62}$$

where the error terms look like

$$\begin{aligned}\Delta\mathcal{M}_k^{\mathrm{Grad}}(\varphi) &= \gamma_k \langle \nabla\varphi(x_k), \nabla\mathcal{R}(x_k)\rangle - \frac{\gamma_k}{\sqrt{d}} \langle \nabla\varphi(x_k), \nabla_{r_1} f(r_k) \cdot (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle \\ \Delta\mathcal{M}_k^{\mathrm{Hess}}(\varphi) &= \frac{\gamma_k^2}{2d}\langle \nabla^2\varphi(x_k), \left(\nabla_{r_1} f(r_k) \cdot (\nabla\psi(x_k))^\top \cdot a_{k+1}\right)^{\otimes 2}\rangle \\ &\quad - \frac{\gamma_k^2}{2d}\,\mathbb{E}_{a_{k+1}}\left[\langle \nabla^2\varphi(x_k), \left(\nabla_{r_1} f(r_k) \cdot (\nabla\psi(x_k))^\top \cdot a_{k+1}\right)^{\otimes 2}\rangle\middle|\mathcal{F}_k\right] \\ \mathcal{E}_k^{\mathrm{Hess}}(\varphi) &= -\frac{\gamma_k^2}{2d}\langle \nabla_{r_1} f(r_k)^2\, \nabla\psi(x_k)\nabla^2\varphi(x_k)\,(\nabla\psi(x_k))^\top, \sqrt{K}\Pi_k\sqrt{K}\rangle \\ &\quad + \frac{\gamma_k^2}{2d}\langle \nabla_{r_1} f(r_k)^2\, \nabla\psi(x_k)\nabla^2\varphi(x_k)\,(\nabla\psi(x_k))^\top, \left(\sqrt{K}\Pi_k v_k\right)^{\otimes 2}\rangle.\end{aligned}$$

We now pass from the one-step decomposition to the continuous-time embedding. For $t \geqslant 0$, write

$$n_t := \lfloor td \rfloor, \qquad t_d := \frac{n_t}{d}.$$

Thus $0 \leqslant t - t_d < 1/d$ and $x_{\lfloor td\rfloor} = x_{n_t}$.

Define the predictable drift integrand

$$A_j(\varphi) := -d\,\gamma_j \langle \nabla\varphi(x_j), \nabla\mathcal{R}(x_j)\rangle + \frac{\gamma_j^2}{2} I(B(x_j)) \left\langle \nabla^2\varphi(x_j), \left\langle K, (\nabla\psi(x_j))^{\otimes 2}\right\rangle\right\rangle.$$

Equivalently, the predictable part of one SGD step is $(1/d)A_j(\varphi)$. Let

$$A_\varphi^{(d)}(s) := A_{\lfloor sd\rfloor}(\varphi)$$

be its piecewise-constant interpolation. Then

$$\sum_{j=0}^{n_t-1} \frac{1}{d} A_j(\varphi) = \int_0^{t_d} A_\varphi^{(d)}(s)\,\mathrm{d}s.$$

Therefore,

$$\sum_{j=0}^{n_t-1} \frac{1}{d} A_j(\varphi) = \int_0^t A_\varphi^{(d)}(s)\,\mathrm{d}s + \xi_t^{\mathrm{mesh}}(\varphi),$$

where the mesh error is

$$\xi_t^{\mathrm{mesh}}(\varphi) := -\int_{t_d}^t A_\varphi^{(d)}(s)\,\mathrm{d}s.$$

Since $t - t_d \leqslant 1/d$, we have

$$\left|\xi_t^{\mathrm{mesh}}(\varphi)\right| \leqslant \frac{1}{d} \sup_{0\leqslant s\leqslant T} \left|A_\varphi^{(d)}(s)\right|.$$

Using the one-step decomposition above, we obtain

$$\begin{aligned}\varphi(x_{\lfloor td\rfloor}) = \varphi(x_0) + \int_0^t A_\varphi^{(d)}(s)\,\mathrm{d}s + \mathcal{M}_{\lfloor td\rfloor}^{\mathrm{Grad}}(\varphi) + \mathcal{M}_{\lfloor td\rfloor}^{\mathrm{Hess}}(\varphi) \\ + \sum_{j=0}^{\lfloor td\rfloor-1} \mathbb{E}_{a_{j+1}}\left[\mathcal{E}_j^{\mathrm{Hess}}(\varphi)\,\middle|\,\mathcal{F}_j\right] + \sum_{j=0}^{\lfloor td\rfloor-1} \mathbb{E}_{a_{j+1}}\left[\mathcal{E}_j^{\mathrm{High}}(\varphi)\,\middle|\,\mathcal{F}_j\right] + \xi_t^{\mathrm{mesh}}(\varphi). \quad \text{(B.63)}\end{aligned}$$

where

$$\mathcal{M}_{\lfloor td\rfloor}^{\mathrm{Grad}}(\varphi) := \sum_{j=0}^{\lfloor td\rfloor-1} \Delta\mathcal{M}_j^{\mathrm{Grad}}(\varphi), \qquad \mathcal{M}_{\lfloor td\rfloor}^{\mathrm{Hess}}(\varphi) := \sum_{j=0}^{\lfloor td\rfloor-1} \Delta\mathcal{M}_j^{\mathrm{Hess}}(\varphi).$$

In Section B.3.3, we prove that the error terms in (B.63) are negligible as $d \to \infty$. The other two terms in $\int_0^t A_\varphi^{(d)}(s)\,\mathrm{d}s$ survive the limit. Next, we show that SGD on S is an (ε, M, T)-approximate solution.

$S\left(x_{\lfloor td\rfloor}, z\right)$ is an Approximate Solution, Proof of Proposition B.3.4

The goal in this section is to prove Proposition B.3.4, that is, show that

$$S\left(x_{\lfloor td\rfloor}, z\right) = \frac{1}{d} W_{\lfloor td\rfloor}^\top \Omega\left(x_{\lfloor td\rfloor}, z\right) W_{\lfloor td\rfloor} \in \mathbb{C}^{3\times 3}$$

is an approximate solution to the partial integro-differential equation (B.6).

Proof of Proposition B.3.4. Applying the preceding real-valued identity to the real and imaginary parts of each matrix entry of $S(\cdot, z)$, and then recombining, yields

$$S\left(x_{\lfloor td\rfloor}, z\right) = S\left(x_0, z\right) + \int_0^t \mathcal{F}\left(z, S(x_{\lfloor sd\rfloor}, z)\right)\mathrm{d}s + \mathcal{M}_{\lfloor td\rfloor}^{\mathrm{Grad}}(S) + \mathcal{M}_{\lfloor td\rfloor}^{\mathrm{Hess}}(S)$$

$$+\sum_{j=0}^{\lfloor td\rfloor-1}\mathbb{E}_{a_{j+1}}\left[\mathcal{E}_j^{\mathrm{Hess}}(S)\,\middle|\,\mathcal{F}_j\right]+\sum_{j=0}^{\lfloor td\rfloor-1}\mathbb{E}_{a_{j+1}}\left[\mathcal{E}_j^{\mathrm{High}}(S)\,\middle|\,\mathcal{F}_j\right]+\xi_t^{\mathrm{mesh}}(S).$$

Thus to show that $S\left(x_{\lfloor td\rfloor},z\right)$ is an approximate solution of the partial integro-differential equation (B.6) it suffices to bound the martingales and error terms where C is a positive constant independent of d. We thus have for all $z\in\Gamma$, the estimate

$$\sup_{0\leqslant t\leqslant(\hat{\tau}_M\wedge T)}\left\|S\left(x_{\lfloor td\rfloor},z\right)-S\left(x_0,z\right)-\int_0^t\mathcal{F}\left(z,S(x_{\lfloor sd\rfloor},z)\right)\mathrm{d}s\right\|$$

$$\leqslant\sup_{0\leqslant t\leqslant(\hat{\tau}_M\wedge T)}\left\|\mathcal{M}_{\lfloor td\rfloor}^{\mathrm{Grad}}(S)\right\|+\sup_{0\leqslant t\leqslant(\hat{\tau}_M\wedge T)}\left\|\mathcal{M}_{\lfloor td\rfloor}^{\mathrm{Hess}}(S)\right\|$$

$$+\sup_{0\leqslant t\leqslant(\hat{\tau}_M\wedge T)}\left\|\sum_{j=0}^{\lfloor td\rfloor-1}\mathbb{E}_{a_{j+1}}\left[\mathcal{E}_j^{\mathrm{Hess}}(S)\middle|\mathcal{F}_j\right]\right\|+\sup_{0\leqslant t\leqslant(\hat{\tau}_M\wedge T)}\left\|\sum_{j=0}^{\lfloor td\rfloor-1}\mathbb{E}_{a_{j+1}}\left[\mathcal{E}_j^{\mathrm{High}}(S)\middle|\mathcal{F}_j\right]\right\|$$

$$+\sup_{0\leqslant t\leqslant(\hat{\tau}_M\wedge T)}\left\|\xi_t^{\mathrm{mesh}}(S)\right\|.$$

Next, fix a constant $\delta>0$. Let $\Gamma^\delta=\Gamma_1^\delta\times\Gamma_2^\delta\times\Gamma_3^\delta\times\Gamma_4^\delta$ where each Γ_i^δ is a $d^{-\delta}$-mesh of Γ_i with $|\Gamma^\delta|\leqslant C_\Gamma d^{4\delta}$ for positive $C_\Gamma>0$ depending on $\|K\|_{\mathrm{op}}$ and $\|\beta^*\|_\infty$. For all $z\in\Gamma$, we note that for some constants $C,c>0$ we have $\vartheta_{c\cdot M}\leqslant\hat{\tau}_M\leqslant\vartheta_{C\cdot M}$ (see Lemma B.2.1). Consequently, we evaluate the error with the stopped process $x_{td}^{\vartheta}=x_{td\wedge\vartheta}$ instead of using $\hat{\tau}_M$. By the martingale error propositions, Proposition B.3.9 and B.3.10, the proofs of which are deferred to Section B.3.3, we have for any $\hat{\delta}>0$ the estimates

$$\sup_{z\in\Gamma^\delta}\sup_{0\leqslant t\leqslant T}\left\|\mathcal{M}_{\lfloor(t\wedge\vartheta_{C\cdot M})d\rfloor}^{\mathrm{Grad}}(S(\cdot,z))\right\|<d^{-\frac{2+\alpha}{2}+\hat{\delta}}\quad\text{w.o.p.,}$$

$$\text{and}\quad\sup_{z\in\Gamma^\delta}\sup_{0\leqslant t\leqslant T}\left\|\mathcal{M}_{\lfloor(t\wedge\vartheta_{C\cdot M})d\rfloor}^{\mathrm{Hess}}(S(\cdot,z))\right\|<d^{-(2+\alpha)+\hat{\delta}}\quad\text{w.o.p.}$$

In addition, for the Hessian and higher order terms errors, by Propositions B.3.11 and B.3.12, the proofs of which are deferred to Sections B.3.3 and B.3.3, we have

$$\sup_{z\in\Gamma^\delta}\sup_{0\leqslant t\leqslant T}\sum_{j=0}^{\lfloor(t\wedge\vartheta_{C\cdot M})d\rfloor-1}\left\|\mathbb{E}_{a_{k+1}}\left[\mathcal{E}_j^{\mathrm{Hess}}(S)\middle|\mathcal{F}_j\right]\right\|\leqslant Cd^{-1}\quad\text{w.o.p.,}$$

$$\text{and}\quad\sup_{z\in\Gamma^\delta}\sup_{0\leqslant t\leqslant T}\sum_{j=0}^{\lfloor(t\wedge\vartheta_{C\cdot M})d\rfloor-1}\left\|\mathbb{E}_{a_{k+1}}\left[\mathcal{E}_j^{\mathrm{High}}(S)\middle|\mathcal{F}_j\right]\right\|\leqslant Cd^{-\frac{1}{2}}\quad\text{w.o.p.}$$

The mesh error is the contribution of the last fractional time interval:

$$\xi_t^{\text{mesh}}(S)(z) = -\int_{t_d}^{t} \mathcal{F}\big(z, S(x_{\lfloor sd \rfloor}, \cdot)\big)\,\mathrm{d}s, \qquad t_d = \frac{\lfloor td \rfloor}{d}.$$

Hence, since $0 \leqslant t - t_d \leqslant 1/d$,

$$\sup_{0\leqslant t\leqslant(\hat{\tau}_M \wedge T)} \left\|\xi_t^{\text{mesh}}(S)(\cdot)\right\|_{\Gamma} \leqslant \frac{1}{d} \sup_{0\leqslant s\leqslant(\hat{\tau}_M \wedge T)} \left\|\mathcal{F}\big(\cdot, S(x_{\lfloor sd \rfloor}, \cdot)\big)\right\|_{\Gamma}.$$

Last we show that in the stopped interval, the right-hand side is bounded by Cd^{-1}. For any $0 \leqslant s \leqslant (\hat{\tau}_M \wedge T)$, we have

$$\begin{aligned}\left\|\mathcal{F}(z, S(x_{\lfloor sd \rfloor}, z))\right\| \leqslant \bar{\gamma} C \left\|H\left(B(x_{\lfloor sd \rfloor})\right)\right\| \cdot \left\|z^{\zeta}\left(\textstyle\bigcirc_{i=1}^{4} \mathcal{G}_i\right)\left(S(x_{\lfloor sd \rfloor}, z)\right)\right\| \\ + \bar{\gamma}^2 C \left\|I\left(B(x_{\lfloor sd \rfloor})\right)\right\| \cdot \left\|z^{\zeta}\left(\textstyle\bigcirc_{i=1}^{4} \mathcal{G}_i\right)\left(S(x_{\lfloor sd \rfloor}, z)\right)\right\| + \cdots\end{aligned}$$

such that $B(x_{\lfloor sd \rfloor}) = \frac{1}{(2\pi)^4} \oint_{\Gamma} z_4\, S(x_{\lfloor sd \rfloor}, z)\,\mathrm{d}z$. Next, plugging equations (B.10), (B.15), (B.17), and

$$\left\|z^{\zeta}\left(\textstyle\bigcirc_{i=1}^{4} \mathcal{G}_i\right)\left(S(x_{\lfloor sd \rfloor}, z)\right)\right\| \leqslant C(|\Gamma|) \cdot \left\|\left(\textstyle\bigcirc_{i=1}^{4} \mathcal{G}_i\right)\left(S(x_{\lfloor sd \rfloor}, z)\right)\right\|,$$

we know there is a positive constant $C = C(L(h), L(I), \bar{\gamma}, \|K\|_{\text{op}}, \|\beta^*\|_{\infty}, M, \alpha)$, such that $\left\|\mathcal{F}\big(\cdot, S(x_{\lfloor sd \rfloor}, \cdot)\big)\right\|_{\Gamma} \leqslant C$. Therefore,

$$\sup_{0\leqslant t\leqslant(\hat{\tau}_M \wedge T)} \left\|\xi_t^{\text{mesh}}(S)(\cdot)\right\|_{\Gamma} \leqslant Cd^{-1}. \tag{B.64}$$

Consequently, combining all the errors, we deduce that for some $C > 0$, which does not depend on d,

$$\sup_{0\leqslant t\leqslant(\hat{\tau}_M \wedge T)} \left\|S\left(x_{\lfloor td \rfloor}, z\right) - S\left(x_0, z\right) - \int_0^t \mathcal{F}\left(z, S(x_{\lfloor sd \rfloor}, z)\right)\mathrm{d}s\right\|_{\Gamma^{\delta}} \leqslant Cd^{\hat{\delta}/2 - 1/2} \quad \text{w.o.p.}$$

An application of the net argument, Lemma B.3.1, finishes the proof after setting $\hat{\delta} = 1 - 2\delta$ for $\delta \in (0, 1/2)$.

The derivative regularity condition (i) is automatic for the continuous-time embedding $t \mapsto S(x_{\lfloor td \rfloor}, \cdot)$: for each fixed trajectory, the dependence on z is through a finite product of resolvents on the fixed contour Γ, and the resolvent derivative bounds from Remark B.2.3 give the required Lipschitz control up to $\hat{\tau}_M$. Thus the estimate above verifies condition (ii), and hence $S(x_{\lfloor td \rfloor}, \cdot)$ is an $(d^{-\varepsilon}, M, T)$-approximate solution. □

B.3.3 Error Bounds

Recall that we are interested in the statistic

$$S(x,z) := \frac{1}{d} W(x)^\top \Omega(x,z) W(x) \in \mathbb{C}^{3\times 3},$$

where

$$W(x) := [\,\psi(x) \mid \beta^* \mid \mathbb{1}_d\,] \in \mathbb{R}^{d\times 3}$$

and $\Omega(x,z)$ is defined in (3.18). Throughout this section, the contour Γ is as in Remark 3.3.3. We control the error terms arising in the comparison between SGD and homogenized SGD under S, with $\mathcal{F}$ given by the partial integro-differential equation (B.6). All estimates are taken on the stopped event where $\|x_{\lfloor td\rfloor}\|_\infty$ and $\|\mathfrak{X}_t\|_\infty$ remain uniformly bounded for $0 \leqslant t \leqslant T$ (see Assumption 3.3.1). Constants may depend on this bound, but not on d.

Before proceeding, we present some bounds on the derivatives of S.

Lemma B.3.6. *There exist constants $c, C > 0$ such that, for large enough d, the following hold:*

$$\frac{c}{d}\|W\|^2 \leqslant \|S(x,\cdot)\|_\Gamma \leqslant \frac{C}{d}\|W\|^2,$$

$$\|\nabla_x S(x,\cdot)\|_\Gamma \leqslant \frac{C(\|x\|_\infty)}{d}\|W\|,$$

$$\textit{and} \quad \left\|\nabla_x^2 S(x,\cdot)\right\|_\Gamma \leqslant \frac{C(\|x\|_\infty)}{d}.$$

Proof. The first result follows from the proof of Lemma B.2.1.

For the derivative, observe that

$$\|\mathrm{diag}(\psi(x))\|_{\mathrm{op}} = \|\psi(x)\|_\infty = \max_i |\psi_i(x)| \leqslant 2\|Q\|\,\|x\|_\infty^2 + \|l\|_1\,\|x\|_\infty + |c|\,. \tag{B.65}$$

Moreover, by Lemma B.1.1, we have

$$\begin{aligned}
\|\nabla_u \psi(x)\|_{\mathrm{op}} &= \|2q_{11}\,\mathrm{diag}(u) + 2q_{12}\,\mathrm{diag}(v) + l_1 I_d\|_{\mathrm{op}} \leqslant 2\sqrt{2}\,\|x\|_\infty + |l_1|\,,\\
\|\nabla_v \psi(x)\|_{\mathrm{op}} &= \|2q_{12}\,\mathrm{diag}(u) + 2q_{22}\,\mathrm{diag}(v) + l_2 I_d\|_{\mathrm{op}} \leqslant 2\sqrt{2}\,\|x\|_\infty + |l_2|\,.
\end{aligned} \tag{B.66}$$

Taking norms on Lemma B.1.3 and using that $\sup_{z_i \in \Gamma_i} \|R(z_i; D_i)\|_{\mathrm{op}} \leqslant 2$, the second result follows.

Analogously, for the Hessian, the bound immediately follows. □

Consequently, the same bounds hold for the real-valued coordinate functions $S_{ab}^{\mathrm{Re},z}$ and $S_{ab}^{\mathrm{Im},z}$, uniformly in $a, b \in \{1, 2, 3\}$ and $z \in \Gamma$.

To control the errors, we will need to make an *a priori* estimate that effectively shows that the iterates of homogenized SGD and SGD remain bounded. Thus, recall our definition, for fixed $M > 0$, of the stopping times

$$\begin{aligned} \vartheta_M &= \inf\{t \geqslant 0 : \frac{1}{d} \left\|W_{\lfloor td \rfloor}\right\|^2 > M \text{ or } B\left(x_{\lfloor td \rfloor}\right) \notin \mathcal{U}\} \quad \text{or} \\ \vartheta_M &= \inf\{t \geqslant 0 : \frac{1}{d} \|\mathcal{W}_t\|^2 > M \text{ or } B\left(\mathfrak{X}_t\right) \notin \mathcal{U}\}, \end{aligned} \tag{B.67}$$

depending on whether we are working with SGD iterates or homogenized SGD iterates. We often drop the M so that $\vartheta := \vartheta_M$. It will be convenient to work with the stopped processes, $W_{\lfloor td \rfloor}^{\vartheta} := W_{\lfloor (t \wedge \vartheta) d \rfloor}$ and $\mathcal{W}_t^{\vartheta} := \mathcal{W}_{t \wedge \vartheta}$.

Remark B.3.7. The stopping time $\hat{\tau}_M = \inf\{t \geqslant 0 : \left\|S(x_{\lfloor td \rfloor}, z)\right\|_{\Gamma} > M \text{ or } B(x_{\lfloor td \rfloor}) \notin \mathcal{U}\}$ and $\hat{\tau}_M = \inf\{t \geqslant 0 : \|S(\mathfrak{X}_t, z)\|_{\Gamma} > M \text{ or } B(\mathfrak{X}_t) \notin \mathcal{U}\}$ are related to ϑ_M by positive constants $c, C > 0$, $\vartheta_{c \cdot M} \leqslant \hat{\tau}_M \leqslant \vartheta_{C \cdot M}$ (see Lemma B.3.6).

In the remainder of this section, we establish a series of propositions that provide bounds on the martingale terms arising from both homogenized SGD and SGD. Throughout the proofs below, C denotes a positive constant independent of d. It may depend on $L(h)$, $L(I)$, $\bar{\gamma}$, T, $\|K\|_{\mathrm{op}}$, $\|\beta^*\|_{\infty}$, M, and α, and may change from line to line and not necessarily be the same as the constant C in Lemma B.3.6.

Homogenized SGD Martingale Error

In this section we control the martingale that arises in homogenized SGD, that is, for a test function $\varphi \colon \mathbb{R}^{2d} \to \mathbb{R}$, define

$$\mathcal{M}_t^{\mathrm{HSGD}}(\varphi) := \int_0^t \mathrm{d}\mathcal{M}_s^{\mathrm{HSGD}}(\varphi) = \int_0^t \gamma(s) \sqrt{I\left(B(\mathfrak{X}_s)\right)} \langle (\nabla \psi(\mathfrak{X}_s))^{\top} \sqrt{K}, \nabla \varphi(\mathfrak{X}_s) \left(\mathrm{d}\mathfrak{B}_s\right)^{\top} \rangle. \tag{B.68}$$

As introduced in Remark B.3.3, we are interested in controlling $\mathcal{M}_t^{\mathrm{HSGD}}(S(\cdot,z))$.

To control the fluctuations of this martingale, we need to control its quadratic variation, defined as follows. Consider a partition of time for $[0,t]$, that is, $0=t_0<t_1<\cdots<t_n=t$ such that the size of the partition $\Delta t=\max_i\{t_i-t_{i-1}\}\to 0$. We define for the continuous process Y,

$$[Y_t(n)]=\sum_{k=1}^{n}\left(Y_{t_k}-Y_{t_{k-1}}\right)^2.$$

If, for every partition of time $[0,t]$ such that $\Delta t\to 0$, the process $[Y_t(n)]$ converges in probability to a process $[Y_t]$ as $n\to\infty$, we call $[Y_t]$ the *quadratic variation* of Y. Using the quadratic variation of $\mathcal{M}_t^{\mathrm{HSGD}}$, we will show that the martingale arising from homogenized SGD is small.

Proposition B.3.8 (Homogenized SGD martingale small)**.** *Suppose $f\colon \mathbb{R}^2\to\mathbb{R}$ is α-pseudo-Lipschitz with constant $L(f)$ (see Assumption 3.1.4). Let the statistic $S\colon \mathbb{R}^{2d}\times\Gamma\subset\mathbb{R}^{2d}\times\mathbb{C}^4\to\mathbb{C}^{3\times 3}$ be defined as in (3.19). Then for each fixed $z\in\Gamma$ and each $T,\zeta>0$, there is some constant C such that, with overwhelming probability,*

$$\sup_{0\leqslant t\leqslant T}\left\|\mathcal{M}_{t\wedge\vartheta}^{\mathrm{HSGD}}(S(\cdot,z))\right\|\leqslant CL(f)d^{\zeta/2-1/2}. \tag{B.69}$$

Proof. Fix indices $i,j\in\{1,2,3\}$. Let φ denote either

$$\varphi(x)=S_{ij}^{\mathrm{Re},z}(x)=\operatorname{Re}S_{ij}(x,z)$$

or

$$\varphi(x)=S_{ij}^{\mathrm{Im},z}(x)=\operatorname{Im}S_{ij}(x,z).$$

We prove the scalar estimate for this real-valued φ. The corresponding matrix-valued estimate for $S(\cdot,z)$ follows by applying the same bound to all real and imaginary parts and taking a finite union bound over the 3×3 entries. First, we rewrite the martingale increment, $\mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}$, as

$$\mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}(\varphi)=\gamma(t)\sqrt{I\left(B(\mathfrak{X}_t)\right)}\langle\left(\nabla\psi(\mathfrak{X}_t)\right)^\top\sqrt{K},\nabla_x\varphi(\mathfrak{X}_t,z)\left(\mathrm{d}\mathfrak{B}_t\right)^\top\rangle_{\mathbb{R}^{2d\times d}}. \tag{B.70}$$

The quadratic variation of $\mathcal{M}_t^{\mathrm{HSGD}}$ is

$$\left[\mathcal{M}_t^{\mathrm{HSGD}}(\varphi)\right]=\int_0^t\gamma(s)^2\left|I\left(B(\mathfrak{X}_s)\right)\right|\left\|\langle\nabla_x\varphi(\mathfrak{X}_s,z),\nabla\psi(\mathfrak{X}_s)^\top\sqrt{K}\rangle_{\mathbb{R}^{2d}}\right\|^2\mathrm{d}s. \tag{B.71}$$

We need to compute $\sup_{0\leqslant t\leqslant T}\left[\mathcal{M}_t^{\mathrm{HSGD}}(\varphi)\right]$ and show that this quantity is small. In particular, we only need to show that the integrand is small. For this, we see that

$$\left\|\langle\nabla_x\varphi(\mathcal{X}_s,z),\nabla\psi(\mathcal{X}_s)^\top\sqrt{K}\rangle_{\mathbb{R}^{2d}}\right\|^2\leqslant\|K\|_{\mathrm{op}}\|\nabla\psi(\mathcal{X}_s)\|_{\mathrm{op}}^2\|\nabla_x\varphi(\mathcal{X}_s,z)\|^2.$$

Using (B.66), we have that $\|\nabla\psi(\mathcal{X}_s)\|_{\mathrm{op}}\leqslant C(\|\mathcal{X}_s\|_\infty)$. By Lemma B.3.6, we have a bound on $\|\nabla_x\varphi(\mathcal{X}_s,z)\|\leqslant\|\nabla_x S(\mathcal{X}_s,\cdot)\|_\Gamma\leqslant\frac{C(\|\mathcal{X}_s\|_\infty)}{d}\|\mathcal{W}_s\|$. From Lemma B.1.6, the growth condition on $|I(B(\mathcal{X}_s))|=\mathbb{E}_a[|\nabla_{r_1}f(\rho_s)|^2]$ yields

$$\begin{aligned}
&|I(B(\mathcal{X}_s))|\left\|\langle\nabla_x\varphi(\mathcal{X}_s,z),\nabla\psi(\mathcal{X}_s)^\top\sqrt{K}\rangle_{\mathbb{R}^{2d}}\right\|^2\\
&\quad\leqslant\frac{C(\|\mathcal{X}_s\|_\infty)}{d^2}(L(f))^2\|\mathcal{W}_s\|^2\left(1+\frac{1}{\sqrt{d}}\|K\|_{\mathrm{op}}^{1/2}\|\mathcal{W}_s\|\right)^{\max\{1,2\alpha\}}\\
&\quad\leqslant\frac{C(\|\mathcal{X}_s\|_\infty)}{d}(L(f))^2M\left(1+\sqrt{M}\right)^{\max\{1,2\alpha\}}.
\end{aligned}\tag{B.72}$$

Thus, (B.71) and (B.72), together

$$\sup_{0\leqslant t\leqslant T}\left[\mathcal{M}_{t\wedge\vartheta}^{\mathrm{HSGD}}(\varphi)\right]\leqslant C(\|\mathcal{X}_s\|_\infty)(L(f))^2\cdot\bar{\gamma}^2\cdot d^{-1}.\tag{B.73}$$

Using the fact that if $\sup_{0\leqslant t\leqslant T}\left[\mathcal{M}_{t\wedge\vartheta}^{\mathrm{HSGD}}(\varphi)\right]\leqslant b$ a.s. then $\mathbb{P}\left[\sup_{0\leqslant t\leqslant T}\left|\mathcal{M}_{t\wedge\vartheta}^{\mathrm{HSGD}}(\varphi)\right|>p\right]\leqslant\exp(-p^2/2b)$, for any $\zeta>0$ and $p=\sqrt{C}L(f)d^{\zeta/2-1/2}$,

$$\mathbb{P}\left[\sup_{0\leqslant t\leqslant T}\left|\mathcal{M}_{t\wedge\vartheta}^{\mathrm{HSGD}}(\varphi)\right|>p\right]\leqslant C\exp(-d^\zeta).$$

Applying the scalar bound to the real and imaginary parts of all nine entries of $S(\cdot,z)$ and taking a finite union bound gives the displayed matrix norm bound. Since the matrix dimension is fixed, this only changes the constant. □

Bounds on the Martingales $\mathcal{M}_k^{\mathrm{Grad}}$ and $\mathcal{M}_k^{\mathrm{Hess}}$

Now we move on to the martingale increments coming from SGD applied to test functions φ. Recall, the expressions for the martingale increments for any quartic statistics φ

$$\begin{aligned}
\Delta\mathcal{M}_k^{\mathrm{Grad}}(\varphi)&=\gamma_k\langle\nabla\varphi(x_k),\nabla\mathcal{R}(x_k)\rangle-\frac{\gamma_k}{\sqrt{d}}\langle\nabla\varphi(x_k),\nabla_{r_1}f(r_k)\cdot(\nabla\psi(x_k))^\top\cdot a_{k+1}\rangle\\
\Delta\mathcal{M}_k^{\mathrm{Hess}}(\varphi)&=\frac{\gamma_k^2}{2d}\langle\nabla^2\varphi(x_k),\left(\nabla_{r_1}f(r_k)\cdot(\nabla\psi(x_k))^\top\cdot a_{k+1}\right)^{\otimes2}\rangle
\end{aligned}$$

$$-\frac{\gamma_k^2}{2d}\,\mathbb{E}_{a_{k+1}}\left[\langle\nabla^2\varphi(x_k),\left(\nabla_{r_1}f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top\cdot a_{k+1}\right)^{\otimes 2}\rangle\middle|\mathcal{F}_k\right]$$

with

$$\mathcal{M}_k(\varphi)=\sum_{j=0}^{k-1}\Delta\mathcal{M}_j(\varphi).$$

Proposition B.3.9 (Gradient martingale)**.** *Suppose $f\colon\mathbb{R}^2\to\mathbb{R}$ is α-pseudo-Lipschitz with constant $L(f)$ (see Assumption 3.1.4). Let the statistic $S\colon\mathbb{R}^{2d}\times\Gamma\subset\mathbb{R}^{2d}\times\mathbb{C}^4\to\mathbb{C}^{3\times 3}$ be defined as in (3.19). Then for each fixed $z\in\Gamma$ and each $T,\zeta>0$, with overwhelming probability,*

$$\sup_{0\leqslant t\leqslant T}\left\|\mathcal{M}^{\mathrm{Grad}}_{\lfloor d(t\wedge\vartheta)\rfloor}(S(\cdot,z))\right\|\leqslant d^{-\frac{2+\alpha}{2}+\zeta}. \tag{B.74}$$

Proof. Fix indices $i,j\in\{1,2,3\}$. Let φ denote either

$$\varphi(x)=S_{ij}^{\mathrm{Re},z}(x)=\operatorname{Re}S_{ij}(x,z)$$

or

$$\varphi(x)=S_{ij}^{\mathrm{Im},z}(x)=\operatorname{Im}S_{ij}(x,z).$$

We prove the scalar estimate for this real-valued φ. The corresponding matrix-valued estimate for $S(\cdot,z)$ follows by applying the same bound to all real and imaginary parts and taking a finite union bound over the 3×3 entries. Throughout the proof of this proposition, we will be working on the stopped version of the martingale, $\mathcal{M}^{\mathrm{Grad}}_{\lfloor d(t\wedge\vartheta)\rfloor}$. However, to lighten the notation, we will suppress the ϑ dependence in the subscript as well as the φ and simply write $\mathcal{M}^{\mathrm{Grad}}_{\lfloor td\rfloor}:=\mathcal{M}^{\mathrm{Grad}}_{\lfloor d(t\wedge\vartheta)\rfloor}(\varphi)$. We have the martingale increments

$$\begin{aligned}
\Delta\mathcal{M}_k^{\mathrm{Grad}}&=\gamma_k\left\langle\nabla\varphi(x_k),\nabla\mathcal{R}(x_k)\right\rangle-\frac{\gamma_k}{\sqrt{d}}\left\langle\nabla\varphi(x_k),\nabla_{r_1}f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top\cdot a_{k+1}\right\rangle\\
&=\frac{\gamma_k}{\sqrt{d}}\,\mathbb{E}_{a_{k+1}}\left[\left\langle\nabla\varphi(x_k),\nabla_{r_1}f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top\cdot a_{k+1}\right\rangle\middle|\mathcal{F}_k\right]\\
&\quad-\frac{\gamma_k}{\sqrt{d}}\left\langle\nabla\varphi(x_k),\nabla_{r_1}f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top\cdot a_{k+1}\right\rangle\\
&=\frac{\gamma_k}{\sqrt{d}}\,\mathbb{E}_{a_{k+1}}\left[\left\langle\nabla\varphi(x_k),\left(\nabla\psi(x_k)\right)^\top\cdot a_{k+1}\right\rangle\nabla_{r_1}f\left(r_k\right)\middle|\mathcal{F}_k\right]\\
&\quad-\frac{\gamma_k}{\sqrt{d}}\left\langle\nabla\varphi(x_k),\left(\nabla\psi(x_k)\right)^\top\cdot a_{k+1}\right\rangle\nabla_{r_1}f\left(r_k\right).
\end{aligned}$$

We define $\mathcal{M}_k^{\mathrm{Grad},\beta}$ to be a new martingale with increments

$$\begin{aligned}\Delta\mathcal{M}_k^{\mathrm{Grad},\beta} &= \frac{\gamma_k}{\sqrt{d}}\,\mathbb{E}_{a_{k+1}}\left[\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\langle\nabla\varphi(x_k),(\nabla\psi(x_k))^\top\cdot a_{k+1}\rangle\nabla_{r_1}f\circ\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}(r_k)\middle|\mathcal{F}_k\right]\\ &\quad-\frac{\gamma_k}{\sqrt{d}}\,\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\langle\nabla\varphi(x_k),(\nabla\psi(x_k))^\top\cdot a_{k+1}\rangle\nabla_{r_1}f\circ\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}(r_k)\,,\end{aligned}$$

where we note that there are two projections and the projection of r_k is in both coordinates of $\mathbb{R}^2$, even though the gradient $\nabla_{r_1}f$ is only with respect to the first coordinate. We take $\beta = d^\zeta$ in the projection radius for some $\zeta > 0$ to be determined later. We will bound $\mathcal{M}_k^{\mathrm{Grad},\beta}$ first, and then bound the difference between $\mathcal{M}_k^{\mathrm{Grad}}$ and $\mathcal{M}_k^{\mathrm{Grad},\beta}$.

We begin by computing subgaussian bounds on the quantities that are going to be projected, namely r_k and $\langle\nabla\varphi(x_k),(\nabla\psi(x_k))^\top\cdot a_{k+1}\rangle$. For the purposes of this section, when we refer to a vector as *subgaussian*, we mean that its entries individually satisfy the stated subgaussian concentration bound. We can rewrite the quantities r_k and $\langle\nabla\varphi(x_k),(\nabla\psi(x_k))^\top\cdot a_{k+1}\rangle$ as

$$\begin{aligned}r_k = \frac{1}{\sqrt{d}}\begin{pmatrix}\psi(x_k)\\ \beta^*\end{pmatrix}^\top a_{k+1} = \frac{1}{\sqrt{d}}\begin{pmatrix}\psi(x_k)\\ \beta^*\end{pmatrix}^\top\sqrt{K}v_k \quad\text{and}\\ \langle\nabla\varphi(x_k),(\nabla\psi(x_k))^\top\cdot a_{k+1}\rangle = \langle\sqrt{K}\cdot\nabla\psi(x_k)\cdot\nabla\varphi(x_k),v_k\rangle,\end{aligned}\tag{B.75}$$

so r_k is $\frac{1}{\sqrt{d}}\left\|\sqrt{K}\right\|_{\mathrm{op}}\cdot\left\|\begin{pmatrix}\psi(x_k)\\ \beta^*\end{pmatrix}\right\|_{\mathrm{op}}$ - subgaussian and $\langle\nabla\varphi(x_k),(\nabla\psi(x_k))^\top\cdot a_{k+1}\rangle$ is $\left\|\sqrt{K}\right\|_{\mathrm{op}}\cdot\|\nabla\psi(x_k)\|_{\mathrm{op}}\cdot\|\nabla\varphi(x_k)\|_{\mathrm{op}}$ - subgaussian where

$$\|\nabla\varphi(x_k)\|_{\mathrm{op}} = \|\nabla_x S_{ij}(x_k,z)\|_{\mathrm{op}} \leqslant \|\nabla_x S(x_k,\cdot)\|_\Gamma \leqslant \frac{C(\|x_k\|_\infty)}{d}\,\|W_k\|\,.$$

by Lemma B.3.6, and $\left\|\begin{pmatrix}\psi(x_k)\\ \beta^*\end{pmatrix}\right\|_{\mathrm{op}}, \|\nabla\psi(x_k)\|_{\mathrm{op}} \leqslant C(\|x_k\|_\infty)$ by (B.65) and (B.66). Thus, since we are working on the stopped processes,

$$\|r_k\|_{\psi_2} = d^{-\frac{1}{2}}C(\|x_k\|_\infty) \quad\text{and}\quad \left\|\langle\nabla\varphi(x_k),(\nabla\psi(x_k))^\top\cdot a_{k+1}\rangle\right\|_{\psi_2} = d^{-\frac{1}{2}}C(\|x_k\|_\infty).\tag{B.76}$$

These subgaussian bounds will be used to bound the difference between $\mathcal{M}_k^{\mathrm{Grad}}$ and $\mathcal{M}_k^{\mathrm{Grad},\beta}$.

Furthermore, leveraging the projections and the fact that f is α-pseudo-Lipschitz, which implies the growth bound on $\nabla_{r_1} f(r)$ in Lemma B.1.6, we can derive the norm bounds

$$\left|\nabla_{r_1} f \circ \mathrm{Proj}_{d^{-\frac{1}{2}}\beta}(r_k)\right| \leqslant L(f)C(1+d^{-\frac{1}{2}}\beta)^{\max\{1,\alpha\}}, \tag{B.77}$$

$$\left|\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle\right| \leqslant d^{-\frac{1}{2}}\beta. \tag{B.78}$$

This gives us the bound

$$\left|\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle \nabla_{r_1} f \circ \mathrm{Proj}_{d^{-\frac{1}{2}}\beta}(r_k)\right| \leqslant L(f)Cd^{-\frac{2+\alpha}{2}}\beta^{2+\alpha} \tag{B.79}$$

and, since this is an almost sure bound, it holds for the expectation as well, and we get

$$\left|\Delta\mathcal{M}_k^{\mathrm{Grad},\beta}\right| \leqslant 2\gamma_k L(f)Cd^{-\frac{3+\alpha}{2}}\beta^{2+\alpha}.$$

Applying Azuma's inequality with the assumption $n = O(d)$, we obtain

$$\sup_{1\leqslant k\leqslant n} \mathbb{P}\left[\left|\mathcal{M}_k^{\mathrm{Grad},\beta}\right| > t\right] < 2\exp\left(\frac{-t^2}{2n\cdot(Cd^{-\frac{3+\alpha}{2}}\beta^{2+\alpha})^2}\right) \leqslant 2\exp\left(\frac{-t^2}{C'd^{-(2+\alpha)}\beta^{2(2+\alpha)}}\right).$$

Thus, with overwhelming probability,

$$\sup_{1\leqslant k\leqslant n}\left|\mathcal{M}_k^{\mathrm{Grad},\beta}\right| < d^{-\frac{2+\alpha}{2}}\beta^{3+\alpha}. \tag{B.80}$$

Finally, we bound the difference between $\{\mathcal{M}_k^{\mathrm{Grad}}\}_{k=1}^n$ and $\{\mathcal{M}_k^{\mathrm{Grad},\beta}\}_{k=1}^n$. For ease of notation, we write

$$\begin{aligned}
G_k &:= \frac{\gamma_k}{\sqrt{d}}\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle \nabla_{r_1} f(r_k)\\
G_{k,\beta} &:= \frac{\gamma_k}{\sqrt{d}}\,\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle \nabla_{r_1} f \circ \mathrm{Proj}_{d^{-\frac{1}{2}}\beta}(r_k).
\end{aligned}$$

The quantity we are trying to bound is

$$\begin{aligned}
\left|\Delta\mathcal{M}_k^{\mathrm{Grad}} - \Delta\mathcal{M}_k^{\mathrm{Grad},\beta}\right| &= \left|\left(G_k - \mathbb{E}_{a_{k+1}} G_k\right) - \left(G_{k,\beta} - \mathbb{E}_{a_{k+1}} G_{k,\beta}\right)\right|\\
&\leqslant |G_k - G_{k,\beta}| + \left|\mathbb{E}_{a_{k+1}}(G_k - G_{k,\beta})\right|.
\end{aligned}$$

First, we will show that $G_k - G_{k,\beta} = 0$ with overwhelming probability. Using the subgaussian bounds on r_k and $\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle$, we have

$$\mathbb{P}\left[G_k \neq G_{k,\beta}\right] \leqslant \mathbb{P}\left[\|r_k\| > d^{-\frac{1}{2}}\beta\right] + \mathbb{P}\left[\left|\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle\right| > d^{-\frac{1}{2}}\beta\right]$$

$$< 4\exp\left(-\frac{\beta^2}{2C}\right).$$

Since $\beta = d^{\zeta}$ for some $\zeta > 0$, the probability bounds above imply that $G_k - G_{k,\beta} = 0$ with overwhelming probability, and it remains to bound the difference in their expectations. For this, we have

$$\begin{aligned}\left|\mathbb{E}_{a_{k+1}}\left[G_k - G_{k,\beta}\right]\right| &= \left|\mathbb{E}_{a_{k+1}}\left[(G_k - G_{k,\beta})\cdot \mathbb{1}\{G_k \neq G_{k,\beta}\}\right]\right| \\ &\leqslant \left|\mathbb{E}_{a_{k+1}}\left[G_k \cdot \mathbb{1}\{G_k \neq G_{k,\beta}\}\right]\right| + \left|\mathbb{E}_{a_{k+1}}\left[G_{k,\beta}\cdot \mathbb{1}\{G_k \neq G_{k,\beta}\}\right]\right|.\end{aligned}$$

For $\left|\mathbb{E}_{a_{k+1}}\left[G_{k,\beta}\cdot \mathbb{1}\{G_k \neq G_{k,\beta}\}\right]\right|$, we have

$$\begin{aligned}\left|\mathbb{E}_{a_{k+1}}\left[G_{k,\beta}\cdot \mathbb{1}\{G_k \neq G_{k,\beta}\}\right]\right| &\leqslant \max|G_{k,\beta}|\cdot \mathbb{P}\left[G_k \neq G_{k,\beta}\right] \\ &\leqslant L(f)Cd^{-\frac{2+\alpha}{2}}\beta^{3+\alpha}\cdot 4\exp\left(-\frac{\beta^2}{2C}\right).\end{aligned}$$

For $\left|\mathbb{E}_{a_{k+1}}\left[G_k \cdot \mathbb{1}\{G_k \neq G_{k,\beta}\}\right]\right|$, we have

$$\begin{aligned}\left|\mathbb{E}_{a_{k+1}}\left[G_k \cdot \mathbb{1}\{G_k \neq G_{k,\beta}\}\right]\right| &\leqslant \mathbb{E}_{a_{k+1}}\left|G_k \cdot \mathbb{1}\{E_1\}\right| \\ &\quad + \mathbb{E}_{a_{k+1}}\left|G_k \cdot \mathbb{1}\{E_2\}\right| \\ &\quad + \mathbb{E}_{a_{k+1}}\left|G_k \cdot \mathbb{1}\{E_3\}\right|,\end{aligned}$$

$$\begin{aligned}\text{where } E_1 &:= \{\|r_k\| \leqslant d^{-\frac{1}{2}}\beta\} \cap \left\{\left|\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^{\top}\cdot a_{k+1}\rangle\right| > d^{-\frac{1}{2}}\beta\right\}, \\ E_2 &:= \{\|r_k\| > d^{-\frac{1}{2}}\beta\} \cap \left\{\left|\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^{\top}\cdot a_{k+1}\rangle\right| \leqslant d^{-\frac{1}{2}}\beta\right\}, \\ E_3 &:= \{\|r_k\| > d^{-\frac{1}{2}}\beta\} \cap \left\{\left|\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^{\top}\cdot a_{k+1}\rangle\right| > d^{-\frac{1}{2}}\beta\right\}.\end{aligned}$$

The term $\mathbb{E}_{a_{k+1}}\left|G_k \cdot \mathbb{1}\{E_1\}\right|$ can be bounded as

$$\mathbb{E}_{a_{k+1}}\left|G_k \cdot \mathbb{1}\{E_1\}\right| \leqslant L(f)C(d^{-\frac{1}{2}}\beta)^{\max\{1,\alpha\}}.$$

$$\mathbb{E}_{a_{k+1}}\left[\left|\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^{\top}\cdot a_{k+1}\rangle\right| \cdot \mathbb{1}\left\{\left|\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^{\top}\cdot a_{k+1}\rangle\right| > d^{-\frac{1}{2}}\beta\right\}\right],$$

where the expectation on the right-hand side is exponentially small due to being a tail of a sub-gaussian first moment (where β^2 is larger than the sub-gaussian variance and grows with

d). By similar reasoning, $\mathbb{E}_{a_{k+1}} |G_k \cdot \mathbb{1}\{E_2\}|$ is also exponentially small (using the growth bound on $\nabla_{r_1} f$). For $\mathbb{E}_{a_{k+1}} |G_k \cdot \mathbb{1}\{E_3\}|$, we have

$$\begin{aligned}
\mathbb{E}_{a_{k+1}} |G_k \cdot \mathbb{1}\{E_3\}| &\leqslant \frac{\gamma_k}{\sqrt{d}} \mathbb{E}_{a_{k+1}} \Big[\Big| \nabla_{r_1} f\left(r_k\right) \cdot \mathbb{1}\{\|r_k\| > d^{-\frac{1}{2}}\beta\} \Big| \cdot \\
&\qquad \Big| \langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1} \rangle \cdot \mathbb{1}\{ \Big| \langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1} \rangle \Big| > d^{-\frac{1}{2}}\beta \} \Big| \Big] \\
&\leqslant \frac{\gamma_k}{\sqrt{d}} \mathbb{E}_{a_{k+1}} \Big| \nabla_{r_1} f\left(r_k\right) \cdot \mathbb{1}\{\|r_k\| > d^{-\frac{1}{2}}\beta\} \Big|^2 \cdot \\
&\qquad \mathbb{E}_{a_{k+1}} \Big| \langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1} \rangle \cdot \mathbb{1}\{ \Big| \langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1} \rangle \Big| > d^{-\frac{1}{2}}\beta \} \Big|^2 .
\end{aligned}$$

This is a product of tails of Gaussian moments, which is again exponentially small. Thus, we conclude that, with overwhelming probability, $\sup_{1\leqslant k\leqslant n} \left| \Delta\mathcal{M}_k^{\mathrm{Grad}} - \Delta\mathcal{M}_k^{\mathrm{Grad},\beta} \right|$ is exponentially small and thus, taking $\beta = d^{\zeta}$, we conclude that, with overwhelming probability,

$$\sup_{1\leqslant k\leqslant n} \left| \mathcal{M}_k^{\mathrm{Grad}} \right| < d^{-\frac{2+\alpha}{2}+\zeta(3+\alpha)}.$$

Adjusting the value of ζ, recalling that all of this has been proved on the stopped process, and applying the scalar bound to the real and imaginary parts of all nine entries of $S(\cdot, z)$ and taking a finite union bound gives the displayed matrix norm bound. Since the matrix dimension is fixed, this only changes the constant. □

Proposition B.3.10 (Hessian martingale)**.** *Suppose $f\colon \mathbb{R}^2 \to \mathbb{R}$ is α-pseudo-Lipschitz with constant $L(f)$ (see Assumption 3.1.4). Let the statistic $S\colon \mathbb{R}^{2d} \times \Gamma \subset \mathbb{R}^{2d} \times \mathbb{C}^4 \to \mathbb{C}^{3\times 3}$ be defined as in (3.19). Then for each fixed $z \in \Gamma$ and each $T, \zeta > 0$, with overwhelming probability,*

$$\sup_{0\leqslant t\leqslant T} \left| \mathcal{M}^{\mathrm{Hess}}_{\lfloor d(t\wedge\vartheta)\rfloor}(S(\cdot, z)) \right| \leqslant d^{-(2+\alpha)+\zeta}. \tag{B.81}$$

Proof. As in the proof of the previous proposition, we will work on the stopped version of the martingale but will suppress the ϑ dependence in the subscript in order to lighten the notation. We also, as before, let φ denote either

$$\varphi(x) = S_{ij}^{\mathrm{Re},z}(x) = \operatorname{Re} S_{ij}(x, z)$$

or

$$\varphi(x) = S_{ij}^{\mathrm{Im},z}(x) = \operatorname{Im} S_{ij}(x, z).$$

We have the martingale increment

$$\begin{aligned}\Delta\mathcal{M}_k^{\mathrm{Hess}} &= \frac{\gamma_k^2}{2d}\langle\nabla^2\varphi(x_k), \left(\nabla_{r_1} f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top\cdot a_{k+1}\right)^{\otimes 2}\rangle \\ &\quad - \frac{\gamma_k^2}{2d}\,\mathbb{E}_{a_{k+1}}\left[\langle\nabla^2\varphi(x_k), \left(\nabla_{r_1} f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top\cdot a_{k+1}\right)^{\otimes 2}\rangle\middle|\mathcal{F}_k\right] \\ &= \frac{\gamma_k^2}{2d}\,\langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\right]^{\otimes 2}\rangle, a_{k+1}^{\otimes 2}\rangle\left(\nabla_{r_1} f\left(r_k\right)\right)^2 \\ &\quad - \frac{\gamma_k^2}{2d}\,\mathbb{E}_{a_{k+1}}\left[\langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\right]^{\otimes 2}\rangle, a_{k+1}^{\otimes 2}\rangle\left(\nabla_{r_1} f\left(r_k\right)\right)^2\middle|\mathcal{F}_k\right].\end{aligned}$$

We define $\mathcal{M}_k^{\mathrm{Hess},\beta}$ to be a new martingale with increments

$$\begin{aligned}\Delta\mathcal{M}_k^{\mathrm{Hess},\beta} &= \frac{\gamma_k^2}{2d}\,\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\right]^{\otimes 2}\rangle, a_{k+1}^{\otimes 2}\rangle\left(\nabla_{r_1} f\circ\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\left(r_k\right)\right)^2 \\ &\quad - \frac{\gamma_k^2}{2d}\,\mathbb{E}_{a_{k+1}}\left[\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\right]^{\otimes 2}\rangle, a_{k+1}^{\otimes 2}\rangle\left(\nabla_{r_1} f\circ\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\left(r_k\right)\right)^2\middle|\mathcal{F}_k\right].\end{aligned}$$

The approach here is similar to the procedure for bounding $\mathcal{M}_k^{\mathrm{Grad}}$. As we saw in the proof of the previous Proposition, r_k is $\frac{1}{\sqrt{d}}\left\|\sqrt{K}\right\|_{\mathrm{op}}\cdot\left\|\begin{pmatrix}\psi(x_k)\\ \beta^*\end{pmatrix}\right\|_{\mathrm{op}}$–subgaussian in each entry. To obtain a concentration bound for $\langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\right]^{\otimes 2}\rangle, a_{k+1}^{\otimes 2}\rangle$, we rewrite it as

$$\begin{aligned}\langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\right]^{\otimes 2}\rangle, a_{k+1}^{\otimes 2}\rangle &= \langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\right]^{\otimes 2}\rangle, \left(\sqrt{K}v_{k+1}\right)^{\otimes 2}\rangle \\ &= \langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\sqrt{K}\right]^{\otimes 2}\rangle, v_{k+1}^{\otimes 2}\rangle,\end{aligned}$$

where the vector v_k has independent standard gaussian entries. Since $\nabla^2\varphi(x_k)\in\mathbb{R}^{2d\times 2d}$ and $\left[\left(\nabla\psi(x_k)\right)^\top\sqrt{K}\right]^{\otimes 2}\in(\mathbb{R}^{2d\times d})^{\otimes 2}$, we get $\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\sqrt{K}\right]^{\otimes 2}\rangle\in\mathbb{R}^{d\times d}$.

Moreover, by Lemma B.3.6 and (B.66), we have

$$\left\|\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\sqrt{K}\right]^{\otimes 2}\rangle\right\|_{\mathrm{op}} \leqslant \|K\|_{\mathrm{op}}\cdot\|\nabla\psi(x_k)\|_{\mathrm{op}}^2\cdot\left\|\nabla^2\varphi(x_k)\right\|_{\mathrm{op}} \leqslant \frac{C(\|x_k\|_\infty)}{d},$$

so by Hanson-Wright inequality we obtain that, for large enough t,

$$\begin{aligned}&\mathbb{P}\left[\langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\sqrt{K}\right]^{\otimes 2}\rangle, v_{k+1}^{\otimes 2}\rangle > t\right] \\ &< 2\exp\left(-C\min\left\{\frac{t^2}{\left\|\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\sqrt{K}\right]^{\otimes 2}\rangle\right\|^2}, \frac{t}{\left\|\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\sqrt{K}\right]^{\otimes 2}\rangle\right\|_{\mathrm{op}}}\right\}\right).\end{aligned}$$

Using the fact that

$$\left\|\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\sqrt{K}\right]^{\otimes 2}\rangle\right\|^2 \leqslant d\left\|\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\sqrt{K}\right]^{\otimes 2}\rangle\right\|_{\mathrm{op}}^2 \leqslant \frac{C(\|x_k\|_\infty)}{d},$$

we conclude that

$$\mathbb{P}\left[\left|\langle\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\sqrt{K}\right]^{\otimes 2}\rangle, v_{k+1}^{\otimes 2}\rangle\right| > t\right] < 2\exp\left(-\frac{d\cdot\min\{t^2,t\}}{C}\right).$$

In particular, this tells us that, for any $\zeta > 0$,

$$\left|\langle\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\right]^{\otimes 2}\rangle, a_{k+1}^{\otimes 2}\rangle\right| \leqslant d^{-\frac{1}{2}+\zeta}$$

with overwhelming probability.

Having obtained concentration bounds for r_k and $\langle\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\right]^{\otimes 2}\rangle, a_{k+1}^{\otimes 2}\rangle$, we proceed to bound $\mathcal{M}_k^{\mathrm{Hess},\beta}$ and show that it is close to $\mathcal{M}_k^{\mathrm{Hess}}$. From the projections and the growth bound on $\nabla_{r_1} f$ in Lemma B.1.6, we can derive the norm bounds

$$\left|\nabla_{r_1} f\circ\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}(r_k)\right|^2 \leqslant \left(L(f)C(1+d^{-\frac{1}{2}}\beta)^{\max\{1,\alpha\}}\right)^2 \tag{B.82}$$

$$\left|\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\langle\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\right]^{\otimes 2}\rangle, a_{k+1}^{\otimes 2}\rangle\right| \leqslant d^{-\frac{1}{2}}\beta, \tag{B.83}$$

and thus

$$\left|\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\langle\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\right]^{\otimes 2}\rangle, a_{k+1}^{\otimes 2}\rangle\left(\nabla_{r_1} f\circ\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}(r_k)\right)^2\right| \leqslant (L(f)C)^2 d^{-\frac{3+2\alpha}{2}}\beta^{3+2\alpha}.$$

Since this is an almost sure bound, it holds for the expectation as well and we get

$$\left|\Delta\mathcal{M}_k^{\mathrm{Hess},\beta}\right| \leqslant \gamma_k^2 (L(f)C)^2 d^{-\frac{5+2\alpha}{2}}\beta^{3+2\alpha}.$$

Applying Azuma's inequality with $n = O(d)$, we obtain

$$\sup_{1\leqslant k\leqslant n}\mathbb{P}\left[\left|\mathcal{M}_k^{\mathrm{Hess},\beta}\right| > t\right] < 2\exp\left(\frac{-t^2}{2n\cdot\left(Cd^{-\frac{5+2\alpha}{2}}\beta^{3+2\alpha}\right)^2}\right) \leqslant 2\exp\left(\frac{-t^2}{C'd^{-2(2+\alpha)}\beta^{2(3+2\alpha)}}\right)$$

so, with overwhelming probability,

$$\sup_{1\leqslant k\leqslant n}\left|\mathcal{M}_k^{\mathrm{Hess},\beta}\right| < d^{-(2+\alpha)}\beta^{4+2\alpha}. \tag{B.84}$$

It remains only to bound the difference between $\{\mathcal{M}_k^{\text{Hess}}\}_{k=1}^n$ and $\{\mathcal{M}_k^{\text{Hess},\beta}\}_{k=1}^n$. This follows a very similar argument to what was in the proof of Proposition B.3.9, we write

$$G_k := \frac{\gamma_k^2}{2d}\left\langle\left\langle\nabla^2\varphi(x_k), \left[(\nabla\psi(x_k))^\top\right]^{\otimes 2}\right\rangle, a_{k+1}^{\otimes 2}\right\rangle \left(\nabla_{r_1} f\left(r_k\right)\right)^2$$
$$G_{k,\beta} := \frac{\gamma_k^2}{2d}\,\text{Proj}_{d^{-\frac{1}{2}}\beta}\left\langle\left\langle\nabla^2\varphi(x_k), \left[(\nabla\psi(x_k))^\top\right]^{\otimes 2}\right\rangle, a_{k+1}^{\otimes 2}\right\rangle \left(\nabla_{r_1} f \circ \text{Proj}_{d^{-\frac{1}{2}}\beta}\left(r_k\right)\right)^2.$$

The quantity we are trying to bound is

$$\begin{aligned}\left|\Delta\mathcal{M}_k^{\text{Hess}} - \Delta\mathcal{M}_k^{\text{Hess},\beta}\right| &= \left|\left(G_k - \mathbb{E}_{a_{k+1}} G_k\right) - \left(G_{k,\beta} - \mathbb{E}_{a_{k+1}} G_{k,\beta}\right)\right| \\ &\leqslant \left|G_k - G_{k,\beta}\right| + \left|\mathbb{E}_{a_{k+1}}\left(G_k - G_{k,\beta}\right)\right|.\end{aligned}$$

As in the proof of Proposition B.3.9, the first of the terms on the right-hand side is 0 with overwhelming probability, while the second is exponentially small. Computing the bound for $\left|\mathbb{E}_{a_{k+1}}\left(G_k - G_{k,\beta}\right)\right|$ is similar to what was done in the previous proof and is not repeated here. To see that $|G_k - G_{k,\beta}| = 0$ with overwhelming probability, we write

$$\begin{aligned}\mathbb{P}\left[G_k \neq G_{k,\beta}\right] &\leqslant \mathbb{P}[\|r_k\| > d^{-\frac{1}{2}}\beta] + \mathbb{P}[\left|\left\langle\left\langle\nabla^2\varphi(x_k), \left[(\nabla\psi(x_k))^\top\right]^{\otimes 2}\right\rangle, a_{k+1}^{\otimes 2}\right\rangle\right| > d^{-\frac{1}{2}}\beta] \\ &< 2\exp\left(-\frac{\beta^2}{2C}\right) + 2\exp\left(-\frac{\min\{\beta^2, d^{\frac{1}{2}}\beta\}}{2C}\right).\end{aligned}$$

Thus, $\left|\mathcal{M}_k^{\text{Hess}} - \mathcal{M}_k^{\text{Hess},\beta}\right|$ is exponentially small with overwhelming probability. Using (B.84) and setting β to be an arbitrarily small power of d, we obtain that, with overwhelming probability,

$$\sup_{1\leqslant k\leqslant n}\left|\mathcal{M}_k^{\text{Hess}}\right| < d^{-(2+\alpha)+\zeta(4+2\alpha)}.$$

Adjusting the value of ζ, recalling that all of this has been proved on the stopped process, and applying the scalar bound to the real and imaginary parts of all nine entries of $S(\cdot, z)$ and taking a finite union bound gives the displayed matrix norm bound. Since the matrix dimension is fixed, this only changes the constant. □

Bounds on the Higher Order Error Term in the Taylor Expansion, $\mathcal{E}_t^{\mathrm{High}}$

This section is devoted to showing that the high-order terms (third and higher-order terms derived in the Taylor expansion of the statistic φ) are negligible. Recall the third-order derivative in the Taylor expansion (B.54) is

$$-\frac{\gamma_k^3}{2}\int_0^1 (1-s)^2 \langle \nabla^3 \varphi(x_k - \gamma_k\, s\, \nabla_x \Psi(x_k; a_{k+1})), (\nabla_x \Psi(x_k; a_{k+1}))^{\otimes 3}\rangle \, \mathrm{d}s. \tag{B.85}$$

At first glance, these error terms look, in terms of d, large due to the tensor $(\nabla_x \Psi(x_k; a_{k+1}))^{\otimes 3}$, but the scaling will be sufficient to control this term.

Proposition B.3.11 (Higher-order error term)**.** *Suppose $f\colon \mathbb{R}^2 \to \mathbb{R}$ is α-pseudo-Lipschitz with constant $L(f)$ (see Assumption 3.1.4). Let the statistic $S\colon \mathbb{R}^{2d} \times \Gamma \subset \mathbb{R}^{2d} \times \mathbb{C}^4 \to \mathbb{C}^{3\times 3}$ be defined as in (3.19). Then for each fixed $z \in \Gamma$ and each $T > 0$, with overwhelming probability,*

$$\sup_{0\leqslant t\leqslant T} \sum_{k=0}^{\lfloor (t\wedge\vartheta)d\rfloor - 1} \left| \mathbb{E}_{a_{k+1}}\left[\mathcal{E}_k^{\mathrm{High}}(S(\cdot, z))\middle|\mathcal{F}_k\right]\right| \leqslant Cd^{-1/2}. \tag{B.86}$$

Proof. Fix indices $i, j \in \{1, 2, 3\}$. Let φ denote either

$$\varphi(x) = S_{ij}^{\mathrm{Re},z}(x) = \mathrm{Re}\, S_{ij}(x, z)$$

or

$$\varphi(x) = S_{ij}^{\mathrm{Im},z}(x) = \mathrm{Im}\, S_{ij}(x, z).$$

We prove the scalar estimate for this real-valued φ. The corresponding matrix-valued estimate for $S(\cdot, z)$ follows by applying the same bound to all real and imaginary parts and taking a finite union bound over the 3×3 entries. First, because ψ_i and the diagonal resolvents depend only on the coordinate pair (u_i, v_i), the third derivative of each entry of S is supported only on triples of derivative indices belonging to the same coordinate pair, so the inner product in (B.85) can be simplified as

$$\begin{aligned}
&\langle \nabla^3 \varphi(x_k - \gamma_k\, s\, \nabla_x \Psi(x_k; a_{k+1})), (\nabla_x \Psi(x_k; a_{k+1}))^{\otimes 3}\rangle \\
&\quad = \sum_{h,i,j=1}^{2d} \left(\nabla^3 \varphi(x_k - \gamma_k\, s\, \nabla_x \Psi(x_k; a_{k+1}))\right)_{hij} \left(\nabla_x \Psi(x_k)\right)_h \left(\nabla_x \Psi(x_k)\right)_i \left(\nabla_x \Psi(x_k)\right)_j
\end{aligned}$$

$$= \sum_{\substack{h=1 \\ i,j\in\{h,h+d\}}}^{d} \left(\nabla^3\varphi(x_k - \gamma_k\, s\, \nabla_x\Psi(x_k;a_{k+1}))\right)_{hij} \left(\nabla_x\Psi(x_k)\right)_h \left(\nabla_x\Psi(x_k)\right)_i \left(\nabla_x\Psi(x_k)\right)_j$$

$$+ \sum_{\substack{h=d+1 \\ i,j\in\{h,h-d\}}}^{2d} \left(\nabla^3\varphi(x_k - \gamma_k\, s\, \nabla_x\Psi(x_k;a_{k+1}))\right)_{hij} \left(\nabla_x\Psi(x_k)\right)_h \left(\nabla_x\Psi(x_k)\right)_i \left(\nabla_x\Psi(x_k)\right)_j .$$

Moreover, a simple computation shows that, for any $0 < s < 1$, we have

$$\begin{aligned}\left\|x_k - \gamma_k\, s\, \nabla_x\Psi(x_k;a_{k+1})\right\|_\infty &= \left\|x_k - s\frac{\gamma_k}{\sqrt{d}}\nabla_{r_1} f\left(r_k\right)\left(\nabla\psi(x_k)\right)^\top a_{k+1}\right\|_\infty \\ &\leqslant \left\|x_k\right\|_\infty + \frac{\bar\gamma}{\sqrt{d}}\left|\nabla_{r_1} f(r_k)\right| \left\|\left(\nabla\psi(x_k)\right)^\top a_{k+1}\right\|_\infty .\end{aligned}$$

If we write $a_{k+1} = \sqrt{K}v_k$ with $v_k \sim \mathcal{N}(0, I_d)$, then by (B.66) we have that, with overwhelming probability,

$$\left\|\left(\nabla\psi(x_k)\right)^\top a_{k+1}\right\|_\infty = \left\|\left(\nabla\psi(x_k)\right)^\top \sqrt{K}v_k\right\|_\infty \leqslant \left\|\nabla\psi(x_k)\right\|_{\mathrm{op}} \left\|\sqrt{K}\right\|_{\mathrm{op}} \|v_k\| \leqslant C(\|x_k\|_\infty)\sqrt{d}.$$

Additionally, by Lemma B.1.6,

$$\left|\nabla_{r_1} f(r_k)\right| \leqslant C(\alpha)L(f)\left(1 + \frac{1}{\sqrt{d}}\left\|K\right\|_{\mathrm{op}}^{1/2}\left\|W_k\right\|\right)^{\max\{1,\alpha\}} \qquad \text{w.o.p.}$$

Thus, for any $k \leqslant (t \wedge \vartheta)d$, we get

$$\left\|x_k - \gamma_k\, s\, \nabla_x\Psi(x_k;a_{k+1})\right\|_\infty \leqslant C(\|x_k\|_\infty),$$

where C is a constant depending on $\bar\gamma$, $\|K\|_{\mathrm{op}}$, $L(f)$, and α, but independent of d.

Next, we know

$$\left|\left(\nabla^3\varphi(x_k - \gamma_k\, s\, \nabla_x\Psi(x_k;a_{k+1}))\right)_{hij}\right| \leqslant \frac{1}{d}C(\|x_k - \gamma_k\, s\, \nabla_x\Psi(x_k;a_{k+1})\|_\infty) \leqslant \frac{C(\|x_k\|_\infty)}{d},$$

and so

$$\begin{aligned}&\left|\langle\nabla^3\varphi(x_k - \gamma_k\, s\, \nabla_x\Psi(x_k;a_{k+1})), \left(\nabla_x\Psi(x_k;a_{k+1})\right)^{\otimes 3}\rangle\right| \\ &\qquad \leqslant \frac{C(\|x_k\|_\infty)}{d} \sum_{\substack{h=1 \\ i,j\in\{h,h+d\}}}^{d} \left|\left(\nabla_x\Psi(x_k;a_{k+1})\right)_h \left(\nabla_x\Psi(x_k;a_{k+1})\right)_i \left(\nabla_x\Psi(x_k;a_{k+1})\right)_j\right| \\ &\qquad + \frac{C(\|x_k\|_\infty)}{d} \sum_{\substack{h=d+1 \\ i,j\in\{h,h-d\}}}^{2d} \left|\left(\nabla_x\Psi(x_k;a_{k+1})\right)_h \left(\nabla_x\Psi(x_k;a_{k+1})\right)_i \left(\nabla_x\Psi(x_k;a_{k+1})\right)_j\right|\end{aligned}$$

with overwhelming probability.

Furthermore, we can bound

$$\mathbb{E}_{a_{k+1}}\left[\left|\left(\nabla_x\Psi(x_k;a_{k+1})\right)_h\left(\nabla_x\Psi(x_k;a_{k+1})\right)_i\left(\nabla_x\Psi(x_k;a_{k+1})\right)_j\right|\middle|\mathcal{F}_k\right]$$
$$\leqslant \frac{1}{d\sqrt{d}}\left\|\nabla\psi(x_k)\right\|_{\mathrm{op}}^3\mathbb{E}_{a_{k+1}}\left[\left|\left(\nabla_{r_1}f\left(r_k\right)\right)^3\left(a_{k+1}\right)_h\left(a_{k+1}\right)_i\left(a_{k+1}\right)_j\right|\middle|\mathcal{F}_k\right],$$

where

$$\mathbb{E}_{a_{k+1}}\left[\left|\left(\nabla_{r_1}f\left(r_k\right)\right)^3\left(a_{k+1}\right)_h\left(a_{k+1}\right)_i\left(a_{k+1}\right)_j\right|\middle|\mathcal{F}_k\right]$$
$$\leqslant\sqrt{\mathbb{E}_{a_{k+1}}\left[\left(\nabla_{r_1}f\left(r_k\right)\right)^6\middle|\mathcal{F}_k\right]}\cdot\sqrt{\mathbb{E}_{a_{k+1}}\left[\left(a_{k+1}\right)_h^2\left(a_{k+1}\right)_i^2\left(a_{k+1}\right)_j^2\right]}$$
$$\leqslant\sqrt{C(\alpha)(L(f))^6\left(1+\frac{1}{\sqrt{d}}\left\|K\right\|_{\mathrm{op}}^{1/2}\left\|W_k\right\|\right)^{\max\{1,6\alpha\}}}\cdot\left\|K\right\|_{\mathrm{op}}^{3/2}.$$

Therefore, noting that $k\leqslant(t\wedge\vartheta)d$ and using (B.66), yields

$$\mathbb{E}_{a_{k+1}}\left[\left|\left(\nabla_x\Psi(x_k;a_{k+1})\right)_h\left(\nabla_x\Psi(x_k;a_{k+1})\right)_i\left(\nabla_x\Psi(x_k;a_{k+1})\right)_j\right|\middle|\mathcal{F}_k\right]\leqslant\frac{C(\left\|x_k\right\|_\infty)}{d\sqrt{d}}.$$

Consequently, for any $k\leqslant(t\wedge\vartheta)d$, the quantity

$$\mathbb{E}_{a_{k+1}}\left[\left|-\frac{\gamma_k^3}{2}\int_0^1(1-s)^2\langle\nabla^3\varphi(x_k-\gamma_k\,s\,\nabla_x\Psi(x_k;a_{k+1})),\left(\nabla_x\Psi(x_k;a_{k+1})\right)^{\otimes3}\rangle\,\mathrm{d}s\right|\middle|\mathcal{F}_k\right]$$

is bounded by $C(\left\|x_k\right\|_\infty)/(d\sqrt{d})$. Summing over at most Td indices gives

$$\sup_{0\leqslant t\leqslant T}\sum_{k=0}^{\lfloor(t\wedge\vartheta)d\rfloor-1}\left|\mathbb{E}_{a_{k+1}}\left[\mathcal{E}_k^{\mathrm{High}}(\varphi)\middle|\mathcal{F}_k\right]\right|\leqslant Cd^{-1/2}.$$

Applying the scalar bound to the real and imaginary parts of all nine entries of $S(\cdot,z)$ and taking a finite union bound gives the displayed matrix norm bound. Since the matrix dimension is fixed, this only changes the constant. Thus the third-order Taylor remainder in (B.54) vanish when d grows large after summing up over k. □

Bounds on the Lower Order Terms in the Hessian, $\mathcal{E}_t^{\mathrm{Hess}}$

We now bound the error term, $\sup_{0\leqslant t\leqslant T}\sum_{k=0}^{\lfloor(t\wedge\vartheta)d\rfloor-1}\left|\mathbb{E}_{a_{k+1}}\left[\mathcal{E}_k^{\mathrm{Hess}}\middle|\mathcal{F}_k\right]\right|$, in (B.63). For this, we utilize the operator norm and its dual norm, the nuclear norm.

Proposition B.3.12 (Hessian error term)**.** *Suppose $f\colon \mathbb{R}^2 \to \mathbb{R}$ is α-pseudo-Lipschitz with constant $L(f)$ (see Assumption 3.1.4). Let the statistic $S\colon \mathbb{R}^{2d} \times \Gamma \subset \mathbb{R}^{2d} \times \mathbb{C}^4 \to \mathbb{C}^{3\times 3}$ be defined as in (3.19). Then for each fixed $z \in \Gamma$ and each $T > 0$, with overwhelming probability,*

$$\sup_{0\leqslant t\leqslant T} \sum_{k=0}^{\lfloor (t\wedge\vartheta)d\rfloor - 1} \left|\mathbb{E}_{a_{k+1}}\left[\mathcal{E}_k^{\mathrm{Hess}}(S(\cdot,z))\middle|\mathcal{F}_k\right]\right| \leqslant C(L(f))^4 d^{-1}. \tag{B.87}$$

Proof. Fix indices $i, j \in \{1,2,3\}$. Let φ denote either

$$\varphi(x) = S_{ij}^{\mathrm{Re},z}(x) = \operatorname{Re} S_{ij}(x,z)$$

or

$$\varphi(x) = S_{ij}^{\mathrm{Im},z}(x) = \operatorname{Im} S_{ij}(x,z).$$

We prove the scalar estimate for this real-valued φ. The corresponding matrix-valued estimate for $S(\cdot,z)$ follows by applying the same bound to all real and imaginary parts and taking a finite union bound over the 3×3 entries. Define $\Pi_k := Q_k Q_k^\top$ and note that $\|\Pi_k\|^2 = \operatorname{rank}(\Pi_k) = 2$. Recall

$$\mathcal{E}_k^{\mathrm{Hess}}(\varphi) := -\frac{\gamma_k^2}{2d}\langle \mathcal{B}, \sqrt{K}\Pi_k\sqrt{K}\rangle + \frac{\gamma_k^2}{2d}\langle \mathcal{B}, \left(\sqrt{K}\Pi_k v_k\right)^{\otimes 2}\rangle$$

where $\mathcal{B} := \nabla_{r_1} f\,(r_k)^2\, \nabla\psi(x_k)\nabla^2\varphi(x_k)\,(\nabla\psi(x_k))^\top$. First, we consider the following term

$$\begin{aligned}
\left|\langle \mathcal{B}, \sqrt{K}\Pi_k\sqrt{K}\rangle\right| &= \left|\langle \nabla_{r_1} f\,(r_k)^2\, \nabla\psi(x_k)\nabla^2\varphi(x_k)\,(\nabla\psi(x_k))^\top, \sqrt{K}\Pi_k\sqrt{K}\rangle\right| \\
&\leqslant \left\|\sqrt{K}\Pi_k\sqrt{K}\right\|_* \left\|\nabla_{r_1} f\,(r_k)^2\, \nabla\psi(x_k)\nabla^2\varphi(x_k)\,(\nabla\psi(x_k))^\top\right\|_{\mathrm{op}} \\
&\leqslant \left\|\sqrt{K}\Pi_k\sqrt{K}\right\|_* \nabla_{r_1} f\,(r_k)^2\, \|\nabla\psi(x_k)\|_{\mathrm{op}}^2 \left\|\nabla^2\varphi(x_k)\right\|_{\mathrm{op}} \\
&\leqslant \|K\|_{\mathrm{op}} \|\Pi_k\|_* \nabla_{r_1} f\,(r_k)^2\, \|\nabla\psi(x_k)\|_{\mathrm{op}}^2 \left\|\nabla^2\varphi(x_k)\right\|_{\mathrm{op}}.
\end{aligned}$$

From Lemma B.1.6, we have

$$\mathbb{E}_{a_{k+1}}[|\nabla_{r_1} f(r_k)|^2\,|\mathcal{F}_k] \leqslant C(L(f))^2\left(1 + \frac{1}{\sqrt{d}}\|K\|_{\mathrm{op}}^{1/2}\|W_k\|\right)^{\max\{1,2\alpha\}}.$$

Using (B.66), we have that $\|\nabla\psi(x_k)\|_{\mathrm{op}} \leqslant C(\|x_k\|_\infty)$. Moreover, we also, by Lemma B.3.6, have $\left\|\nabla^2\varphi(x_k)\right\|_{\mathrm{op}} = \left\|\nabla_x^2 S_{ij}(x_k,z)\right\|_{\mathrm{op}} \leqslant \left\|\nabla_x^2 S(x_k,\cdot)\right\|_\Gamma \leqslant \frac{C(\|x_k\|_\infty)}{d}$. Since $k \leqslant (t\wedge\vartheta)d$, we

get

$$\mathbb{E}_{a_{k+1}}\left[\left|\langle \mathcal{B}, \sqrt{K}\Pi_k\sqrt{K}\rangle\right| |\mathcal{F}_k\right] \leqslant \frac{C(\|x_k\|_\infty)}{d}(L(f))^2\left(1+\frac{1}{\sqrt{d}}\|K\|_{\mathrm{op}}^{1/2}\|W_k\|\right)^{\max\{1,2\alpha\}}. \tag{B.88}$$

Similarly we get that

$$\begin{aligned}
\left|\langle \mathcal{B}, \left(\sqrt{K}\Pi_k v_k\right)^{\otimes 2}\rangle\right| &= \left|\langle \nabla_{r_1} f\left(r_k\right)^2 \nabla\psi(x_k)\nabla^2\varphi(x_k)\left(\nabla\psi(x_k)\right)^\top, \left(\sqrt{K}\Pi_k v_k\right)^{\otimes 2}\rangle\right| \\
&\leqslant \left\|\left(\sqrt{K}\Pi_k v_k\right)^{\otimes 2}\right\|_* \left\|\nabla_{r_1} f\left(r_k\right)^2 \nabla\psi(x_k)\nabla^2\varphi(x_k)\left(\nabla\psi(x_k)\right)^\top\right\|_{\mathrm{op}} \\
&\leqslant \left\|\sqrt{K}\Pi_k v_k\right\|^2 \nabla_{r_1} f\left(r_k\right)^2 \|\nabla\psi(x_k)\|_{\mathrm{op}}^2 \left\|\nabla^2\varphi(x_k)\right\|_{\mathrm{op}} \\
&\leqslant \|K\|_{\mathrm{op}} \|\Pi_k v_k\|^2 \nabla_{r_1} f\left(r_k\right)^2 \|\nabla\psi(x_k)\|_{\mathrm{op}}^2 \left\|\nabla^2\varphi(x_k)\right\|_{\mathrm{op}}.
\end{aligned}$$

Upon taking expectations, using Cauchy–Schwarz we have

$$\mathbb{E}_{a_{k+1}}[\|\Pi_k v_k\|^2 \nabla_{r_1} f(r_k)^2 \big| \mathcal{F}_k] \leqslant \left(\mathbb{E}[\|\Pi_k v_k\|^4 \big| \mathcal{F}_k]\right)^{1/2} \left(\mathbb{E}[|\nabla_{r_1} f(r_k)|^4 \big| \mathcal{F}_k]\right)^{1/2}.$$

Now by Lemma B.1.6 we know

$$\mathbb{E}_{a_{k+1}}[|\nabla_{r_1} f(r_k)|^4 \left|\mathcal{F}_k\right] \leqslant C(L(f))^4\left(1+\frac{1}{\sqrt{d}}\|K\|_{\mathrm{op}}^{1/2}\|W_k\|\right)^{\max\{1,4\alpha\}},$$

and $\mathbb{E}_{a_{k+1}}\left[\|\Pi_k v_k\|^4 \middle| \mathcal{F}_k\right] = 8$, as Π_k is a projection. Using Lemma B.3.6 on the growth of φ, yields

$$\mathbb{E}_{a_{k+1}}\left[\left|\langle \mathcal{B}, \left(\sqrt{K}\Pi_k v_k\right)^{\otimes 2}\rangle\right| \middle| \mathcal{F}_k\right] \leqslant \frac{C(\|x_k\|_\infty)}{d}(L(f))^4\left(1+\frac{1}{\sqrt{d}}\|K\|_{\mathrm{op}}^{1/2}\|W_k\|\right)^{\max\{1,4\alpha\}}. \tag{B.89}$$

As $k \leqslant (t \wedge \vartheta)d$, then $\frac{1}{\sqrt{d}}\|W_k\| \leqslant \sqrt{M}$. The result then immediately follows by combining (B.88) and (B.89) and summing up with the extra factor γ_k^2/d. Applying the scalar bound to the real and imaginary parts of all nine entries of $S(\cdot, z)$ and taking a finite union bound gives the displayed matrix norm bound. Since the matrix dimension is fixed, this only changes the constant. □

Note that since $|\Gamma^\delta| \leqslant C_\Gamma d^{4\delta}$, all of the same estimates hold uniformly over $z \in \Gamma^\delta$ by a union bound.

B.4 Entropy Barrier and High-probability Exponential Decay

This section analyzes the homogenized dynamics in the isotropic squared parameterization setting. We prove that, for sufficiently small constant stepsize, the risk decays exponentially with high probability and the dynamics exist globally. The argument is based on an entropy barrier tailored to the coordinatewise structure of the SDE. First, we derive an exact product identity for each coordinate pair $(\mathcal{U}_{t,i}, \mathcal{V}_{t,i})$ and introduce logarithmic coordinates in which the residual $\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1$ has an entropy representation. We then derive an exact SDE for the empirical entropy and prove deterministic barrier estimates comparing the entropy to both the coordinate size and the risk. These estimates yield an exponential supermartingale argument, giving explicit confidence bounds and a high-probability exponential decay theorem. Finally, we record dynamical consequences of this decay, including integrability of the risk and uniform separation from the saddle at the origin.

In this section, we specialize to the isotropic setting $\beta^* = \mathbb{1}_d$ and $a \sim \mathcal{N}(0, I_d)$, so that the risk in (3.3) becomes

$$\mathcal{R}(x) = \frac{1}{4d} \sum_{i=1}^{d} \left(u_i^2 - v_i^2 - 1\right)^2 .$$

We write $\mathcal{X}_t = (\mathcal{U}_t, \mathcal{V}_t)$ for the corresponding homogenized process. For constant stepsize $\gamma > 0$, the coordinatewise SDE in (3.4) takes the form

$$\mathrm{d}\mathcal{U}_{t,i} = -\gamma \mathcal{U}_{t,i} \left(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1\right) \mathrm{d}t + 2\gamma \sqrt{\mathcal{R}(\mathcal{X}_t)}\, \mathcal{U}_{t,i}\, \mathrm{d}\mathfrak{B}_{t,i}, \tag{B.90}$$

$$\mathrm{d}\mathcal{V}_{t,i} = \gamma \mathcal{V}_{t,i} \left(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1\right) \mathrm{d}t - 2\gamma \sqrt{\mathcal{R}(\mathcal{X}_t)}\, \mathcal{V}_{t,i}\, \mathrm{d}\mathfrak{B}_{t,i}, \tag{B.91}$$

for $i = 1, \ldots, d$, where $\mathfrak{B}_t$ is a standard Brownian motion in $\mathbb{R}^d$. We initialize the dynamics at

$$\mathcal{U}_{0,i} = \mathcal{V}_{0,i} = 1, \qquad i = 1, \ldots, d.$$

Since the coefficients are locally Lipschitz, the SDE admits a unique maximal local strong solution on a random interval $[0, \zeta)$, where $\zeta \in (0, \infty]$ denotes the explosion time.

B.4.1 Exact Product Identity and Logarithmic Coordinates

We first record an exact product identity for each coordinate pair.

Lemma B.4.1 (Exact product identity)**.** *For each $i = 1, \dots, d$ and all $0 \leqslant t < \zeta$, we have*

$$\mathcal{U}_{t,i}^2 \mathcal{V}_{t,i}^2 = \exp\Big(- 8\gamma^2 I_t \Big), \quad \textit{where} \quad I_t := \int_0^t \mathcal{R}(\mathcal{X}_s)\, \mathrm{d}s.$$

In particular, $\mathcal{U}_{t,i}^2 > 0$ and $\mathcal{V}_{t,i}^2 > 0$ for all $0 \leqslant t < \zeta$.

Proof. By Itô's formula,

$$\begin{aligned}
\mathrm{d}\mathcal{U}_{t,i}^2 &= \mathcal{U}_{t,i}^2 \Big(- 2\gamma(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1) + 4\gamma^2 \mathcal{R}(\mathcal{X}_t) \Big)\, \mathrm{d}t + 4\gamma \sqrt{\mathcal{R}(\mathcal{X}_t)}\, \mathcal{U}_{t,i}^2\, \mathrm{d}\mathfrak{B}_{t,i}, \\
\mathrm{d}\mathcal{V}_{t,i}^2 &= \mathcal{V}_{t,i}^2 \Big(+ 2\gamma(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1) + 4\gamma^2 \mathcal{R}(\mathcal{X}_t) \Big)\, \mathrm{d}t - 4\gamma \sqrt{\mathcal{R}(\mathcal{X}_t)}\, \mathcal{V}_{t,i}^2\, \mathrm{d}\mathfrak{B}_{t,i}.
\end{aligned}$$

Applying Itô's product rule to $\mathcal{U}_{t,i}^2 \mathcal{V}_{t,i}^2$, the martingale terms cancel and the quadratic covariation contributes

$$\mathrm{d}\langle \mathcal{U}_{\cdot,i}^2, \mathcal{V}_{\cdot,i}^2 \rangle_t = -16\gamma^2 \mathcal{R}(\mathcal{X}_t) \mathcal{U}_{t,i}^2 \mathcal{V}_{t,i}^2\, \mathrm{d}t,$$

so that

$$\mathrm{d}(\mathcal{U}_{t,i}^2 \mathcal{V}_{t,i}^2) = -8\gamma^2 \mathcal{R}(\mathcal{X}_t) \mathcal{U}_{t,i}^2 \mathcal{V}_{t,i}^2\, \mathrm{d}t.$$

Solving this scalar ODE and using $\mathcal{U}_{0,i}^2 \mathcal{V}_{0,i}^2 = 1$ gives

$$\mathcal{U}_{t,i}^2 \mathcal{V}_{t,i}^2 = \exp\Big(- 8\gamma^2 I_t \Big) > 0.$$

Because both factors are nonnegative, each must be strictly positive. □

Define

$$\rho(I_t) := \sqrt{1 + 4e^{-8\gamma^2 I_t}}.$$

Then

$$\frac{\rho(I_t) + 1}{2} > 0, \qquad \frac{\rho(I_t) - 1}{2} > 0, \qquad \frac{\rho(I_t) + 1}{2} \cdot \frac{\rho(I_t) - 1}{2} = e^{-8\gamma^2 I_t}.$$

By Lemma B.4.1, the product of the two normalizing factors matches $\mathcal{U}_{t,i}^2 \mathcal{V}_{t,i}^2$. Hence, for each i, we may define a unique real-valued process

$$\mathcal{Z}_{t,i} := \log\left(\frac{\mathcal{U}_{t,i}^2}{(\rho(I_t) + 1)/2} \right) = -\log\left(\frac{\mathcal{V}_{t,i}^2}{(\rho(I_t) - 1)/2} \right),$$

such that

$$\mathcal{U}_{t,i}^2 = \frac{\rho(I_t) + 1}{2} e^{\mathcal{Z}_{t,i}}, \qquad \mathcal{V}_{t,i}^2 = \frac{\rho(I_t) - 1}{2} e^{-\mathcal{Z}_{t,i}}.$$

For $c > 0$ and $x > 0$, define

$$f_c(x) := x - c - c\log(x/c),$$

and for $\rho \geqslant 1$ define

$$\Phi_\rho(z) := \rho(\cosh z - 1) + \sinh z - z.$$

Lemma B.4.2 (Logarithmic-coordinate identities)**.** *For every $0 \leqslant t < \zeta$ and every $i = 1, \dots, d$,*

$$\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1 = \rho(I_t)\sinh(\mathcal{Z}_{t,i}) + \cosh(\mathcal{Z}_{t,i}) - 1, \tag{B.92}$$

$$\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 = \rho(I_t)\cosh(\mathcal{Z}_{t,i}) + \sinh(\mathcal{Z}_{t,i}), \tag{B.93}$$

$$\Phi_{\rho(I_t)}(\mathcal{Z}_{t,i}) = f_{\frac{\rho(I_t)+1}{2}}(\mathcal{U}_{t,i}^2) + f_{\frac{\rho(I_t)-1}{2}}(\mathcal{V}_{t,i}^2). \tag{B.94}$$

Moreover, $\mathcal{U}_{t,i} > 0$ and $\mathcal{V}_{t,i} > 0$ for all $0 \leqslant t < \zeta$, and hence

$$\mathrm{d}\log \mathcal{U}_{t,i} = \Big(-\gamma(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1) - 2\gamma^2\mathcal{R}(\mathcal{X}_t)\Big)\,\mathrm{d}t + 2\gamma\sqrt{\mathcal{R}(\mathcal{X}_t)}\,\mathrm{d}\mathfrak{B}_{t,i}, \tag{B.95}$$

$$\mathrm{d}\log \mathcal{V}_{t,i} = \Big(+\gamma(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1) - 2\gamma^2\mathcal{R}(\mathcal{X}_t)\Big)\,\mathrm{d}t - 2\gamma\sqrt{\mathcal{R}(\mathcal{X}_t)}\,\mathrm{d}\mathfrak{B}_{t,i}. \tag{B.96}$$

Proof. The identities (B.92)–(B.94) follow directly from the parametrization

$$\mathcal{U}_{t,i}^2 = \frac{\rho(I_t)+1}{2}e^{\mathcal{Z}_{t,i}}, \qquad \mathcal{V}_{t,i}^2 = \frac{\rho(I_t)-1}{2}e^{-\mathcal{Z}_{t,i}}.$$

Since $\mathcal{U}_{0,i} = \mathcal{V}_{0,i} = 1$ and neither process can hit 0 before ζ by Lemma B.4.1, continuity gives strict positivity. The logarithmic SDEs then follow from Itô's formula applied to $\log \mathcal{U}_{t,i}$ and $\log \mathcal{V}_{t,i}$. □

For each coordinate, define the entropy density

$$h_{t,i} := \Phi_{\rho(I_t)}(\mathcal{Z}_{t,i}) = f_{\frac{\rho(I_t)+1}{2}}(\mathcal{U}_{t,i}^2) + f_{\frac{\rho(I_t)-1}{2}}(\mathcal{V}_{t,i}^2), \qquad 0 \leqslant t < \zeta. \tag{B.97}$$

The empirical entropy is then the average of these coordinate densities:

$$H_t := \frac{1}{d}\sum_{i=1}^{d} h_{t,i}. \tag{B.98}$$

Equivalently,

$$H_t = \frac{1}{d}\sum_{i=1}^{d} \Phi_{\rho(I_t)}(\mathcal{Z}_{t,i}) = \frac{1}{d}\sum_{i=1}^{d}\left[f_{\frac{\rho(I_t)+1}{2}}(\mathcal{U}_{t,i}^2) + f_{\frac{\rho(I_t)-1}{2}}(\mathcal{V}_{t,i}^2)\right].$$

At time $t = 0$,

$$\rho(I_0) = \sqrt{5}, \qquad \mathcal{Z}_{0,i} = -\log\Big(\tfrac{\sqrt{5}+1}{2}\Big),$$

and hence

$$h_{0,i} = 2 - \sqrt{5} + \log\Big(\tfrac{\sqrt{5}+1}{2}\Big) \qquad \text{for every } i = 1, \dots, d.$$

Therefore,

$$H_0 = 2 - \sqrt{5} + \log\Big(\tfrac{\sqrt{5}+1}{2}\Big),$$

which is independent of d.

B.4.2 Exact Dynamics of the Empirical Entropy

Recall that

$$H_t = \frac{1}{d}\sum_{i=1}^{d} h_{t,i}, \qquad h_{t,i} := \Phi_{\rho(I_t)}(\mathcal{Z}_{t,i}).$$

We next derive the SDE satisfied by the empirical entropy H_t.

Proposition B.4.3 (Empirical entropy SDE)**.** *For $0 \leqslant t < \zeta$,*

$$\mathrm{d}H_t = \left(-8\gamma\mathcal{R}(\mathcal{X}_t) + 4\gamma^2\mathcal{R}(\mathcal{X}_t)\left[\frac{1}{d}\sum_{i=1}^{d}(\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2) + \rho(I_t)\right]\right)\mathrm{d}t + \mathrm{d}M_t, \tag{B.99}$$

where

$$M_t = 4\gamma\frac{1}{d}\sum_{i=1}^{d}\int_0^t \mathbb{1}_{\{s<\zeta\}}\sqrt{\mathcal{R}(\mathcal{X}_s)}\,(\mathcal{U}_{s,i}^2 - \mathcal{V}_{s,i}^2 - 1)\,\mathrm{d}\mathfrak{B}_{s,i},$$

and

$$\mathrm{d}\langle M\rangle_t = \mathbb{1}_{\{t<\zeta\}}\frac{64\gamma^2}{d}\mathcal{R}(\mathcal{X}_t)^2\,\mathrm{d}t. \tag{B.100}$$

Proof. Since

$$\partial_x f_c(x) = 1 - \frac{c}{x}, \qquad \partial_{xx} f_c(x) = \frac{c}{x^2}, \qquad \partial_c f_c(x) = \log(c/x),$$

Itô's formula gives

$$\mathrm{d}f_{\frac{\rho(I_t)+1}{2}}(\mathcal{U}_{t,i}^2) = \left[\Big(\mathcal{U}_{t,i}^2 - \tfrac{\rho(I_t)+1}{2}\Big)\Big(-2\gamma(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1) + 4\gamma^2\mathcal{R}(\mathcal{X}_t)\Big) + 8\gamma^2\mathcal{R}(\mathcal{X}_t)\tfrac{\rho(I_t)+1}{2}\right]\mathrm{d}t$$

$$+ 4\gamma\sqrt{\mathcal{R}(\mathcal{X}_t)}\Big(\mathcal{U}_{t,i}^2 - \tfrac{\rho(I_t)+1}{2}\Big)\mathrm{d}\mathfrak{B}_{t,i} + \log\left(\frac{(\rho(I_t)+1)/2}{\mathcal{U}_{t,i}^2}\right)\mathrm{d}\Big(\tfrac{\rho(I_t)+1}{2}\Big),$$

and similarly

$$\mathrm{d}f_{\frac{\rho(I_t)-1}{2}}(\mathcal{V}_{t,i}^2) = \left[\Big(\mathcal{V}_{t,i}^2 - \tfrac{\rho(I_t)-1}{2}\Big)\Big(+ 2\gamma(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1) + 4\gamma^2\mathcal{R}(\mathcal{X}_t)\Big) + 8\gamma^2\mathcal{R}(\mathcal{X}_t)\tfrac{\rho(I_t)-1}{2}\right]\mathrm{d}t$$
$$- 4\gamma\sqrt{\mathcal{R}(\mathcal{X}_t)}\Big(\mathcal{V}_{t,i}^2 - \tfrac{\rho(I_t)-1}{2}\Big)\mathrm{d}\mathfrak{B}_{t,i} + \log\left(\frac{(\rho(I_t)-1)/2}{\mathcal{V}_{t,i}^2}\right)\mathrm{d}\Big(\tfrac{\rho(I_t)-1}{2}\Big).$$

Now we compute

$$\Big(\mathcal{U}_{t,i}^2 - \tfrac{\rho(I_t)+1}{2}\Big) - \Big(\mathcal{V}_{t,i}^2 - \tfrac{\rho(I_t)-1}{2}\Big) = \mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1,$$
$$\Big(\mathcal{U}_{t,i}^2 - \tfrac{\rho(I_t)+1}{2}\Big) + \Big(\mathcal{V}_{t,i}^2 - \tfrac{\rho(I_t)-1}{2}\Big) = \mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 - \rho(I_t).$$

Moreover, the two moving-target terms have the same finite-variation factor

$$\mathrm{d}\left(\frac{\rho(I_t)+1}{2}\right) = \mathrm{d}\left(\frac{\rho(I_t)-1}{2}\right) = \frac{1}{2}\,\mathrm{d}\rho(I_t),$$

and also we have

$$\mathcal{U}_{t,i}^2\mathcal{V}_{t,i}^2 = \frac{\rho(I_t)+1}{2}\cdot\frac{\rho(I_t)-1}{2}.$$

Therefore the two moving-target terms cancel:

$$\log\left(\frac{(\rho(I_t)+1)/2}{\mathcal{U}_{t,i}^2}\right) + \log\left(\frac{(\rho(I_t)-1)/2}{\mathcal{V}_{t,i}^2}\right) = 0.$$

Summing the two Itô identities therefore yields

$$\mathrm{d}\Phi_{\rho(I_t)}(\mathcal{Z}_{t,i}) = \Big(- 2\gamma(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1)^2 + 4\gamma^2\mathcal{R}(\mathcal{X}_t)(\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 + \rho(I_t))\Big)\mathrm{d}t$$
$$+ 4\gamma\sqrt{\mathcal{R}(\mathcal{X}_t)}(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1)\mathrm{d}\mathfrak{B}_{t,i}.$$

Averaging over i and using

$$\frac{1}{d}\sum_{i=1}^{d}(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1)^2 = 4\mathcal{R}(\mathcal{X}_t)$$

gives (B.99). The bracket identity (B.100) follows from independence of the Brownian motions:

$$\mathrm{d}\langle M\rangle_t = \frac{16\gamma^2}{d^2}\sum_{i=1}^{d}\mathcal{R}(\mathcal{X}_t)(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1)^2\,\mathrm{d}t = \frac{64\gamma^2}{d}\mathcal{R}(\mathcal{X}_t)^2\,\mathrm{d}t.$$

□

B.4.3 Barrier Estimates

We next derive deterministic barrier estimates. The first shows that the empirical entropy controls the average size of $\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2$. The second gives a dimension-free comparison between the entropy density $\Phi_{\rho(I_t)}(\mathcal{Z}_{t,i})$ and the squared residual,

$$(\rho(I_t)\sinh(\mathcal{Z}_{t,i}) + \cosh(\mathcal{Z}_{t,i}) - 1)^2\,,$$

provided we work below a coordinatewise entropy barrier.

Lemma B.4.4 (Pointwise control of $\mathcal{U}^2 + \mathcal{V}^2$ by the entropy)**.** *For every $\rho \geqslant 1$ and every $z \in \mathbb{R}$,*

$$\rho\cosh z + \sinh z \leqslant 2\Phi_\rho(z) + 2\rho\log 2. \tag{B.101}$$

Consequently, for all $0 \leqslant t < \zeta$,

$$\frac{1}{d}\sum_{i=1}^{d}(\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2) \leqslant 2H_t + 2\rho(I_t)\log 2 \leqslant 2H_t + 2\sqrt{5}\log 2. \tag{B.102}$$

Proof. It suffices to prove the claim for $\rho > 1$, since both sides of (B.101) are continuous in ρ and the case $\rho = 1$ follows by taking the limit $\rho \downarrow 1$.

Fix $c > 0$ and define

$$g_c(x) := f_c(x) - \frac{x}{2} = \frac{x}{2} - c - c\log(x/c), \qquad x > 0.$$

Then $g_c'(x) = \frac{1}{2} - \frac{c}{x}$, so g_c is minimized at $x = 2c$, where $g_c(2c) = -c\log 2$. Hence

$$f_c(x) \geqslant \frac{x}{2} - c\log 2 \qquad \text{for all } x > 0.$$

Applying this with

$$(c, x) = \left(\tfrac{\rho+1}{2},\, \tfrac{\rho+1}{2}e^{z}\right), \qquad (c, x) = \left(\tfrac{\rho-1}{2},\, \tfrac{\rho-1}{2}e^{-z}\right),$$

and using (B.94), we obtain

$$\begin{aligned}\Phi_\rho(z) &\geqslant \frac{1}{2}\left(\tfrac{\rho+1}{2}e^{z} + \tfrac{\rho-1}{2}e^{-z}\right) - \left(\tfrac{\rho+1}{2} + \tfrac{\rho-1}{2}\right)\log 2\\ &= \frac{1}{2}\big(\rho\cosh z + \sinh z\big) - \rho\log 2.\end{aligned}$$

Rearranging gives (B.101). Averaging with $(\rho, z) = (\rho(I_t), \mathcal{Z}_{t,i})$ and using $\rho(I_t) \leqslant \rho(0) = \sqrt{5}$ yields (B.102). □

Lemma B.4.5 (Coercivity under a coordinatewise entropy barrier)**.** *Fix $L_* > H_0$ and define*

$$K_{L_*} := \left\{(\rho, z) : 1 \leqslant \rho \leqslant \sqrt{5},\ \Phi_\rho(z) \leqslant L_*\right\}.$$

Then K_{L_*} is compact. Moreover, there exist constants

$$0 < m_{L_*} \leqslant M_{L_*} < \infty$$

depending only on L_* such that, for every $(\rho, z) \in K_{L_*}$,

$$m_{L_*}\Phi_\rho(z) \leqslant \big(\rho \sinh z + \cosh z - 1\big)^2 \leqslant M_{L_*}\Phi_\rho(z). \tag{B.103}$$

Proof. Since $\rho \in [1, \sqrt{5}]$ and $\Phi_\rho(z) \to \infty$ as $|z| \to \infty$ uniformly over that interval, K_{L_*} is compact by Heine–Borel. Consider

$$Q(\rho, z) := \frac{(\rho \sinh z + \cosh z - 1)^2}{\Phi_\rho(z)} \qquad (z \neq 0).$$

As $z \to 0$, we have

$$\rho \sinh z + \cosh z - 1 = \rho z + \frac{z^2}{2} + O(z^3),$$
$$\Phi_\rho(z) = \rho(\cosh z - 1) + \sinh z - z = \frac{\rho}{2}z^2 + O(z^3),$$

and therefore

$$Q(\rho, z) \to 2\rho \qquad \text{as } z \to 0.$$

Thus Q extends continuously across $\{z = 0\}$ by setting $Q(\rho, 0) := 2\rho$. Next,

$$\partial_z\big(\rho \sinh z + \cosh z - 1\big) = \rho \cosh z + \sinh z > 0 \qquad (\rho \geqslant 1),$$

so $\rho \sinh z + \cosh z - 1$ has the unique zero $z = 0$. Also,

$$\Phi'_\rho(z) = \rho \sinh z + \cosh z - 1, \qquad \Phi''_\rho(z) = \rho \cosh z + \sinh z,$$

with $\Phi_\rho(0) = \Phi'_\rho(0) = 0$ and $\Phi''_\rho(0) = \rho > 0$. Hence $\Phi_\rho(z) > 0$ for all $z \neq 0$. Therefore the continuous extension of Q is strictly positive on the compact set K_{L_*}, so it attains a positive minimum and finite maximum:

$$0 < m_{L_*} := \min_{K_{L_*}} Q \leqslant \max_{K_{L_*}} Q =: M_{L_*} < \infty.$$

This is exactly (B.103). □

Lemma B.4.6 (Risk–entropy comparison below the coordinate barrier)**.** *Fix $L_* > H_0$. On any time interval on which*

$$\max_{1\leqslant i\leqslant d} h_{t,i} < L_*,$$

we have

$$\frac{m_{L_*}}{4} H_t \leqslant \mathcal{R}(\mathcal{X}_t) \leqslant \frac{M_{L_*}}{4} H_t. \tag{B.104}$$

Proof. If $\max_i h_{t,i} < L_*$, then

$$(\rho(I_t), \mathcal{Z}_{t,i}) \in K_{L_*} \qquad \text{for every } i.$$

By Lemma B.4.5 and (B.92),

$$m_{L_*} h_{t,i} \leqslant (\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1)^2 \leqslant M_{L_*} h_{t,i}.$$

Averaging over i and using

$$\mathcal{R}(\mathcal{X}_t) = \frac{1}{4d} \sum_{i=1}^{d} (\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1)^2$$

gives (B.104). □

B.4.4 Exponential Decay with Explicit Confidence

We now prove exponential decay of the empirical entropy, and hence of the risk, using a coordinatewise entropy barrier.

Fix $H_* > H_0$ and $L_* > H_0$, and define the stopping time

$$\tau_{H,L} := \inf \left\{ t \in [0, \zeta) : H_t \geqslant H_* \text{ or } \max_{1\leqslant i\leqslant d} h_{t,i} \geqslant L_* \right\}, \tag{B.105}$$

with the convention $\inf \emptyset = \infty$.

Lemma B.4.7 (The entropy barrier is hit before explosion)**.** *Almost surely,*

$$\tau_{H,L} \leqslant \zeta.$$

Proof. Suppose, toward a contradiction, that $\zeta < \tau_{H,L}$ on some sample path. Then $H_t < H_*$ and $h_{t,i} < L_*$ for every i and all $t < \zeta$. Hence

$$(\rho(I_t), \mathcal{Z}_{t,i}) \in K_{L_*} \qquad \text{for all } i \text{ and all } t < \zeta.$$

Since K_{L_*} is compact, the logarithmic coordinates $\mathcal{Z}_{t,i}$ remain bounded on $[0,\zeta)$. By the parametrization

$$\mathcal{U}_{t,i}^2 = \frac{\rho(I_t)+1}{2}e^{\mathcal{Z}_{t,i}}, \qquad \mathcal{V}_{t,i}^2 = \frac{\rho(I_t)-1}{2}e^{-\mathcal{Z}_{t,i}},$$

the coordinates $\mathcal{U}_{t,i}$ and $\mathcal{V}_{t,i}$ remain bounded on every finite interval $[0,T] \subset [0,\zeta)$. Therefore the coefficients in (B.90)–(B.91) remain bounded and locally Lipschitz on a neighborhood of the attained path. The standard continuation criterion for SDEs then extends the solution beyond ζ, contradicting maximality. Thus $\tau_{H,L} \leqslant \zeta$ almost surely. □

On $[0,\tau_{H,L})$, Lemma B.4.6 gives

$$\frac{m_{L_*}}{4}H_t \leqslant \mathcal{R}(\mathcal{X}_t) \leqslant \frac{M_{L_*}}{4}H_t,$$

and Lemma B.4.4 gives

$$\frac{1}{d}\sum_{i=1}^{d}(\mathcal{U}_{t,i}^2+\mathcal{V}_{t,i}^2) \leqslant 2H_* + 2\sqrt{5}\log 2, \qquad t<\tau_{H,L}. \tag{B.106}$$

Define

$$\lambda_{H,L} := \gamma\Big(8 - 4\gamma(2H_* + 2\sqrt{5}\log 2 + \sqrt{5})\Big)\frac{m_{L_*}}{4}, \tag{B.107}$$

and

$$\sigma_{H,L}^2 := \frac{4\gamma^2 M_{L_*}^2}{d}. \tag{B.108}$$

Lemma B.4.8 (Drift and quadratic-variation bounds before the barrier)**.** *Assume*

$$0<\gamma<\frac{2}{2H_* + 2\sqrt{5}\log 2 + \sqrt{5}}, \tag{B.109}$$

so that $\lambda_{H,L}>0$*. Then, for* $t<\tau_{H,L}$*,*

$$\mathrm{d}H_t \leqslant -\lambda_{H,L}H_t\,\mathrm{d}t + \mathrm{d}M_t, \tag{B.110}$$

and

$$\mathrm{d}\langle M\rangle_t \leqslant \sigma_{H,L}^2 H_t^2\,\mathrm{d}t. \tag{B.111}$$

Proof. Starting from Proposition B.4.3 and using (B.106),

$$\mathrm{d}H_t \leqslant \Big(-8\gamma\mathcal{R}(\mathcal{X}_t) + 4\gamma^2\mathcal{R}(\mathcal{X}_t)(2H_* + 2\sqrt{5}\log 2 + \sqrt{5})\Big)\mathrm{d}t + \mathrm{d}M_t$$

$$= -\gamma\Big(8 - 4\gamma(2H_* + 2\sqrt{5}\log 2 + \sqrt{5})\Big)\mathcal{R}(\mathcal{X}_t)\,\mathrm{d}t + \mathrm{d}M_t.$$

Using the lower bound in (B.104) gives (B.110). Similarly, from (B.100) and the upper bound in (B.104),

$$\mathrm{d}\langle M\rangle_t = \frac{64\gamma^2}{d}\mathcal{R}(\mathcal{X}_t)^2\,\mathrm{d}t \leqslant \frac{64\gamma^2}{d}\left(\frac{M_{L_*}}{4}H_t\right)^2\mathrm{d}t = \frac{4\gamma^2 M_{L_*}^2}{d}H_t^2\,\mathrm{d}t.$$

This is (B.111). □

The next result gives exponential decay up to the coordinatewise barrier.

Theorem B.4.9 (Stopped exponential decay with explicit confidence)**.** *Let $H_*, L_* > H_0$, and define $\lambda_{H,L}$ and $\sigma^2_{H,L}$ by (B.107)–(B.108). Assume*

$$0 < \gamma < \frac{2}{2H_* + 2\sqrt{5}\log 2 + \sqrt{5}}.$$

Fix $\theta > 0$ such that

$$0 < \theta < \frac{2\lambda_{H,L}}{\sigma^2_{H,L}},$$

and define

$$\mu_{H,L}(\theta) := \lambda_{H,L} - \frac{\theta}{2}\sigma^2_{H,L} > 0.$$

Then

$$\mathbb{P}\left(\sup_{0\leqslant s\leqslant\tau_{H,L}} e^{\mu_{H,L}(\theta)s}H_s \geqslant H_*\right) \leqslant \left(\frac{H_0}{H_*}\right)^\theta. \tag{B.112}$$

Proof. Fix $0 < \varepsilon < H_0$ and define

$$\tau_\varepsilon := \inf\{t \in [0,\zeta) : H_t \leqslant \varepsilon\}, \qquad \tau := \tau_{H,L} \wedge \tau_\varepsilon.$$

On $[0,\tau]$, we have $H_t \in [\varepsilon, H_*]$, so Itô's formula gives

$$\log H_{t\wedge\tau} = \log H_0 + \int_0^{t\wedge\tau}\frac{1}{H_s}\,\mathrm{d}H_s - \frac{1}{2}\int_0^{t\wedge\tau}\frac{1}{H_s^2}\,\mathrm{d}\langle M\rangle_s.$$

By Lemma B.4.8,

$$\log H_{t\wedge\tau} \leqslant \log H_0 - \lambda_{H,L}(t\wedge\tau) - \frac{1}{2}\int_0^{t\wedge\tau}\frac{1}{H_s^2}\,\mathrm{d}\langle M\rangle_s + N_t,$$

where

$$N_t := \int_0^{t\wedge\tau}\frac{1}{H_s}\,\mathrm{d}M_s.$$

Moreover,

$$\langle N\rangle_t = \int_0^{t\wedge\tau} \frac{1}{H_s^2}\,\mathrm{d}\langle M\rangle_s \leqslant \sigma_{H,L}^2(t\wedge\tau).$$

Thus, for every $\theta > 0$,

$$\mathcal{E}_t := \exp\Big(\theta N_t - \frac{\theta^2}{2}\langle N\rangle_t\Big)$$

is a positive local martingale, hence a supermartingale. Since

$$\mu_{H,L}(\theta) = \lambda_{H,L} - \frac{\theta}{2}\sigma_{H,L}^2,$$

we obtain

$$\begin{aligned} e^{\theta\mu_{H,L}(\theta)(t\wedge\tau)} H_{t\wedge\tau}^\theta &\leqslant H_0^\theta \exp\Big(\theta N_t - \theta(\lambda_{H,L} - \mu_{H,L}(\theta))(t\wedge\tau)\Big) \\ &= H_0^\theta \exp\Big(\theta N_t - \frac{\theta^2}{2}\sigma_{H,L}^2(t\wedge\tau)\Big) \\ &\leqslant H_0^\theta \mathcal{E}_t. \end{aligned}$$

Ville's maximal inequality gives

$$\mathbb{P}\Big(\sup_{0\leqslant s\leqslant\tau} e^{\mu_{H,L}(\theta)s} H_s \geqslant H_*\Big) \leqslant \left(\frac{H_0}{H_*}\right)^\theta.$$

It remains to remove the auxiliary lower stopping time τ_ε. Define

$$\tau_0 := \inf\{t \in [0,\zeta) : H_t = 0\}, \qquad S_\varepsilon := \tau_{H,L} \wedge \tau_\varepsilon, \qquad S_0 := \tau_{H,L} \wedge \tau_0.$$

Since H_t is continuous and nonnegative, $S_\varepsilon \uparrow S_0$ pathwise as $\varepsilon \downarrow 0$. Therefore,

$$\left\{\sup_{0\leqslant s\leqslant S_\varepsilon} e^{\mu_{H,L}(\theta)s} H_s \geqslant H_*\right\} \uparrow \left\{\sup_{0\leqslant s\leqslant S_0} e^{\mu_{H,L}(\theta)s} H_s \geqslant H_*\right\}.$$

By continuity from below of probability, the estimate obtained above gives

$$\mathbb{P}\Big(\sup_{0\leqslant s\leqslant S_0} e^{\mu_{H,L}(\theta)s} H_s \geqslant H_*\Big) \leqslant \left(\frac{H_0}{H_*}\right)^\theta.$$

Finally, $H_t = 0$ is absorbing. Indeed, if $H_t = 0$, then every coordinate entropy density $h_{t,i}$ vanishes, hence $\mathcal{Z}_{t,i} = 0$ for every i. Therefore

$$\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1 = 0 \qquad \text{for all } i,$$

so $\mathcal{R}(\mathcal{X}_t) = 0$. The drift and diffusion coefficients in (B.90)–(B.91) then vanish, and by pathwise uniqueness the solution remains constant thereafter. Thus

$$\sup_{0 \leqslant s \leqslant S_0} e^{\mu_{H,L}(\theta)s} H_s = \sup_{0 \leqslant s \leqslant \tau_{H,L}} e^{\mu_{H,L}(\theta)s} H_s.$$

This proves (B.112). □

We now give a pathwise criterion for removing the coordinatewise barrier. Let

$$V_{H,L}(\theta) := \frac{M_{L_*}}{4} \frac{H_*}{\mu_{H,L}(\theta)}.$$

For $\eta \in (0,1)$, define

$$R_{H,L}(\theta,\eta) := \sqrt{2V_{H,L}(\theta) \log\left(\frac{4d}{\eta}\right)},$$

and

$$p_{H,L}(\theta) := \exp\left(-4\gamma^2 V_{H,L}(\theta)\right).$$

Set

$$A_{H,L}(\theta) := \operatorname{arcsinh}\left(\frac{1}{2p_{H,L}(\theta)}\right), \qquad B_{H,L}(\theta,\eta) := 2\cosh(A_{H,L}(\theta) + 8\gamma R_{H,L}(\theta,\eta)).$$

Finally, define

$$c_{H,L}(\theta,\eta) := \frac{p_{H,L}(\theta)^2}{B_{H,L}(\theta,\eta)}, \qquad \underline{c}_{H,L}(\theta) := \frac{\sqrt{1+4p_{H,L}(\theta)^2}-1}{2}, \qquad \bar{c} := \frac{\sqrt{5}+1}{2},$$

and

$$\mathfrak{L}_{H,L}(\theta,\eta) := 2 \sup_{\substack{c \in [\underline{c}_{H,L}(\theta), \bar{c}] \\ x \in [c_{H,L}(\theta,\eta), B_{H,L}(\theta,\eta)]}} f_c(x). \tag{B.113}$$

Lemma B.4.10 (Removal of the coordinatewise entropy barrier)**.** *Assume the hypotheses of Theorem B.4.9. Fix $\eta \in (0,1)$ and suppose that*

$$L_* > \mathfrak{L}_{H,L}(\theta,\eta). \tag{B.114}$$

Then

$$\mathbb{P}\Big(\zeta = \infty \text{ and } H_t \leqslant H_* e^{-\mu_{H,L}(\theta)t} \text{ for all } t \geqslant 0\Big) \geqslant 1 - \left(\frac{H_0}{H_*}\right)^{\theta} - \eta. \tag{B.115}$$

Moreover,

$$\mathbb{P}\left(\zeta = \infty \text{ and } \mathcal{R}(\mathcal{X}_t) \leqslant \frac{M_{L_*}}{4} H_* e^{-\mu_{H,L}(\theta)t} \text{ for all } t \geqslant 0\right) \geqslant 1 - \left(\frac{H_0}{H_*}\right)^{\theta} - \eta. \tag{B.116}$$

Proof. Let

$$E_H := \left\{ \sup_{0 \leqslant s \leqslant \tau_{H,L}} e^{\mu_{H,L}(\theta)s} H_s < H_* \right\}.$$

By Theorem B.4.9,

$$\mathbb{P}(E_H) \geqslant 1 - \left(\frac{H_0}{H_*}\right)^{\theta}.$$

On E_H, for all $t < \tau_{H,L}$,

$$H_t \leqslant H_* e^{-\mu_{H,L}(\theta)t}.$$

Using Lemma B.4.6, we get

$$\int_0^{t \wedge \tau_{H,L}} \mathcal{R}(\mathcal{X}_s)\, \mathrm{d}s \leqslant V_{H,L}(\theta).$$

For each coordinate define the martingale

$$N_{t,i} := \int_0^{t \wedge \tau_{H,L}} \sqrt{\mathcal{R}(\mathcal{X}_s)}\, \mathrm{d}\mathfrak{B}_{s,i}.$$

On E_H, its quadratic variation is bounded by $V_{H,L}(\theta)$. Therefore, by the reflection principle and a union bound over $i = 1, \ldots, d$,

$$\mathbb{P}\left(E_H \cap \left\{ \max_{1 \leqslant i \leqslant d} \sup_{t \geqslant 0} |N_{t,i}| > R_{H,L}(\theta, \eta) \right\} \right) \leqslant \eta.$$

Let E_N denote the complementary martingale event. On $E_H \cap E_N$, define

$$\widetilde{\mathcal{Z}}_{t,i} := \log\left(\frac{\mathcal{U}_{t,i}}{\mathcal{V}_{t,i}}\right).$$

Since

$$\mathcal{U}_{t,i}\mathcal{V}_{t,i} = \exp\left(-4\gamma^2 I_t\right),$$

we have, for $t < \tau_{H,L}$,

$$\mathcal{U}_{t,i}\mathcal{V}_{t,i} \geqslant p_{H,L}(\theta).$$

Moreover, Itô's formula gives

$$\mathrm{d}\widetilde{\mathcal{Z}}_{t,i} = \left(-4\gamma \mathcal{U}_{t,i}\mathcal{V}_{t,i} \sinh(\widetilde{\mathcal{Z}}_{t,i}) + 2\gamma\right) \mathrm{d}t + 4\gamma \sqrt{\mathcal{R}(\mathcal{X}_t)}\, \mathrm{d}\mathfrak{B}_{t,i}.$$

Thus the noise-compensated process

$$\mathcal{Y}_{t,i} := \widetilde{\mathcal{Z}}_{t,i} - 4\gamma N_{t,i}$$

satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{Y}_{t,i} = -4\gamma\mathcal{U}_{t,i}\mathcal{V}_{t,i}\sinh(\widetilde{\mathcal{Z}}_{t,i}) + 2\gamma.$$

A deterministic barrier argument gives, for every $t < \tau_{H,L}$,

$$|\widetilde{\mathcal{Z}}_{t,i}| \leqslant A_{H,L}(\theta) + 8\gamma R_{H,L}(\theta,\eta).$$

Consequently,

$$\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 = 2\mathcal{U}_{t,i}\mathcal{V}_{t,i}\cosh(\widetilde{\mathcal{Z}}_{t,i}) \leqslant B_{H,L}(\theta,\eta),$$

and, since

$$\mathcal{U}_{t,i}\mathcal{V}_{t,i} \geqslant p_{H,L}(\theta),$$

we also have

$$\mathcal{U}_{t,i}^2, \mathcal{V}_{t,i}^2 \geqslant c_{H,L}(\theta,\eta).$$

Moreover,

$$\frac{\rho(I_t)-1}{2} \geqslant \underline{c}_{H,L}(\theta), \qquad \frac{\rho(I_t)+1}{2} \leqslant \overline{c}.$$

Therefore, by the definition of $\mathfrak{L}_{H,L}(\theta,\eta)$, we have

$$h_{t,i} = f_{\frac{\rho(I_t)+1}{2}}(\mathcal{U}_{t,i}^2) + f_{\frac{\rho(I_t)-1}{2}}(\mathcal{V}_{t,i}^2) \leqslant \mathfrak{L}_{H,L}(\theta,\eta) < L_*.$$

Thus the coordinatewise entropy barrier cannot be hit on $E_H \cap E_N$. The entropy barrier $H_t = H_*$ cannot be hit either, because on E_H we have $H_t < H_* e^{-\mu_{H,L}(\theta)t} < H_*$ for all $t < \tau_{H,L}$. Hence $\tau_{H,L} = \infty$ on $E_H \cap E_N$. By Lemma B.4.7, this also implies $\zeta = \infty$.

The entropy decay estimate follows from the definition of E_H, and the risk decay follows from the upper bound in (B.104). The probability bound follows from

$$\mathbb{P}(E_H \cap E_N) \geqslant 1 - \left(\frac{H_0}{H_*}\right)^{\theta} - \eta.$$

□

B.4.5 High-probability Decay at Prescribed Confidence

The previous theorem gives an explicit confidence estimate. For fixed d, H_*, L_*, and γ, the admissible range of θ is finite. By choosing the stepsize sufficiently small, we can achieve a prescribed confidence level.

Corollary B.4.11 (High-probability decay for sufficiently small stepsize)**.** *Fix $H_*, L_* > H_0$, and $\delta \in (0,1)$. Define*

$$\theta_\delta := \frac{\log(2/\delta)}{\log(H_*/H_0)}$$

and

$$\bar{\gamma}_{H,L,\delta} := \frac{2}{2H_* + 2\sqrt{5}\log 2 + \sqrt{5} + \frac{2M_{L_*}^2}{d\, m_{L_*}}\theta_\delta}.$$

Assume

$$0 < \gamma < \bar{\gamma}_{H,L,\delta}.$$

Then

$$\mu_{H,L}(\delta) := m_{L_*}\gamma\Big(2 - \gamma(2H_* + 2\sqrt{5}\log 2 + \sqrt{5})\Big) - \frac{2\gamma^2 M_{L_*}^2}{d}\theta_\delta > 0.$$

Suppose further that the coordinatewise barrier is chosen so that

$$L_* > \mathfrak{L}_{H,L}\left(\theta_\delta, \frac{\delta}{2}\right), \tag{B.117}$$

where $\mathfrak{L}_{H,L}$ is defined in (B.113). Then

$$\mathbb{P}\Big(\zeta = \infty \text{ and } \mathcal{R}(\mathcal{X}_t) \leqslant \frac{M_{L_*}}{4} H_* e^{-\mu_{H,L}(\delta)t} \text{ for all } t \geqslant 0\Big) \geqslant 1 - \delta. \tag{B.118}$$

Proof. By construction,

$$\left(\frac{H_0}{H_*}\right)^{\theta_\delta} = \frac{\delta}{2}.$$

The condition $\gamma < \bar{\gamma}_{H,L,\delta}$ is exactly equivalent to

$$\theta_\delta < \frac{2\lambda_{H,L}}{\sigma_{H,L}^2},$$

where $\lambda_{H,L}$ and $\sigma_{H,L}^2$ are defined in (B.107)–(B.108). Moreover,

$$\mu_{H,L}(\delta) = \lambda_{H,L} - \frac{\theta_\delta}{2}\sigma_{H,L}^2.$$

Thus Lemma B.4.10 applies with $\theta = \theta_\delta$ and $\eta = \delta/2$. Substituting these values into (B.116) gives (B.118). □

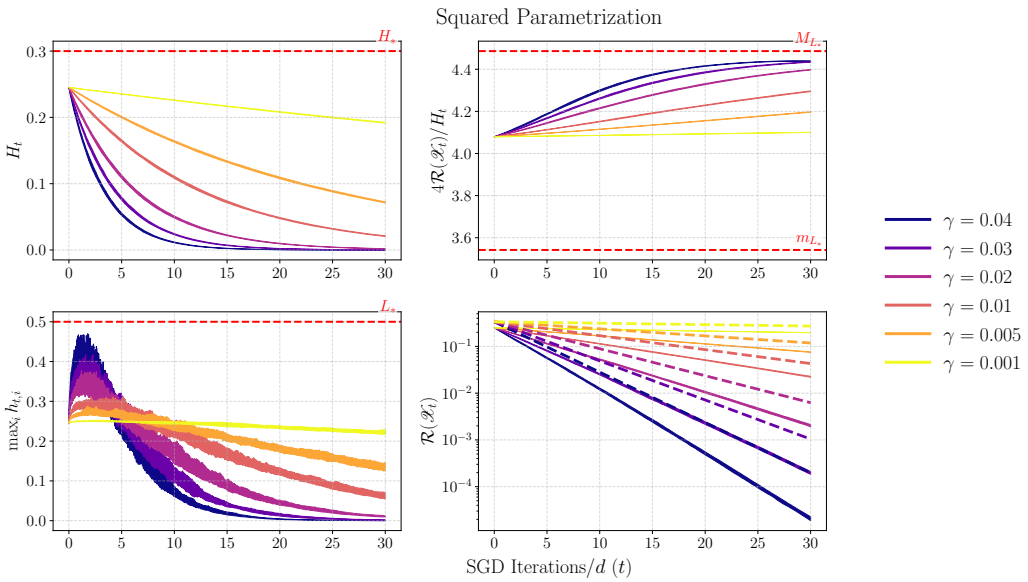

Figure B.1: **Coordinatewise entropy barriers and exponential risk decay on a diagonal linear network.** The figure illustrates the entropy-barrier mechanism from Appendix B.4. The top-left panel shows the empirical entropy H_t, while the bottom-left panel shows the largest coordinatewise entropy density $\max_i h_{t,i}$; the red dashed lines mark the barriers H_* and L_*. The top-right panel plots the risk–entropy ratio $4\mathcal{R}(\mathcal{X}_t)/H_t$, with red dashed lines marking empirical coordinatewise coercivity constants m_{L_*} and M_{L_*} estimated from the coordinate quotients $q_{t,i}$. The bottom-right panel shows the risk $\mathcal{R}(\mathcal{X}_t)$ on a logarithmic scale, together with exponential envelopes of the form $(M_{L_*}/4)H_* e^{-\mu t}$. We plot these envelopes for all tested stepsizes, including those beyond the theorem's certified small-stepsize regime, illustrating that the predicted exponential decay persists empirically beyond the sufficient condition.

B.4.6 Risk Integrability and Dynamical Consequences

We now record two consequences of the high-probability exponential decay: integrability of the risk and uniform separation from the saddle.

Lemma B.4.12 (Finite integral of the risk)**.** *On the event in Corollary B.4.11,*

$$\int_0^\infty \mathcal{R}(\mathcal{X}_s)\,\mathrm{d}s \leqslant \frac{M_{L_*}}{4}\frac{H_*}{\mu_{H,L}(\delta)} < \infty. \tag{B.119}$$

Proof. On the event in Corollary B.4.11,

$$\mathcal{R}(\mathcal{X}_t) \leqslant \frac{M_{L_*}}{4} H_* e^{-\mu_{H,L}(\delta)t} \qquad \text{for all } t \geqslant 0.$$

Therefore

$$\int_0^\infty \mathcal{R}(\mathcal{X}_s)\,\mathrm{d}s \leqslant \frac{M_{L_*}}{4} H_* \int_0^\infty e^{-\mu_{H,L}(\delta)s}\,\mathrm{d}s = \frac{M_{L_*}}{4}\frac{H_*}{\mu_{H,L}(\delta)}.$$

□

Non-explosion via Risk Integrability

Although Corollary B.4.11 already includes $\zeta = \infty$, the following pathwise argument explains why risk integrability prevents finite-time explosion.

For each coordinate define

$$\widetilde{\mathcal{Z}}_{t,i} := \log\left(\frac{\mathcal{U}_{t,i}}{\mathcal{V}_{t,i}}\right), \qquad 0 \leqslant t < \zeta.$$

By Lemma B.4.1,

$$\mathcal{U}_{t,i}\mathcal{V}_{t,i} = \exp\left(-4\gamma^2 I_t\right).$$

Applying Itô's formula to $\widetilde{\mathcal{Z}}_{t,i} = \log \mathcal{U}_{t,i} - \log \mathcal{V}_{t,i}$ gives

$$\mathrm{d}\widetilde{\mathcal{Z}}_{t,i} = \left(-4\gamma \mathcal{U}_{t,i}\mathcal{V}_{t,i}\sinh(\widetilde{\mathcal{Z}}_{t,i}) + 2\gamma\right)\mathrm{d}t + 4\gamma\sqrt{\mathcal{R}(\mathcal{X}_t)}\,\mathrm{d}\mathfrak{B}_{t,i}. \tag{B.120}$$

Lemma B.4.13 (Pathwise coordinate bound from risk integrability)**.** *Assume*

$$\int_0^\infty \mathcal{R}(\mathcal{X}_s)\,\mathrm{d}s < \infty.$$

Then, for every coordinate i,

$$\sup_{0\leqslant t<\zeta}\left(\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2\right) < \infty \qquad \textit{almost surely.}$$

Consequently, finite risk integral is incompatible with finite-time explosion.

Proof. Define the continuous martingale

$$N_{t,i} := \int_0^t \sqrt{\mathcal{R}(\mathcal{X}_s)}\, \mathrm{d}\mathfrak{B}_{s,i}.$$

Its quadratic variation satisfies

$$\langle N_i \rangle_\infty = \int_0^\infty \mathcal{R}(\mathcal{X}_s)\, \mathrm{d}s < \infty.$$

Hence $N_{t,i}$ converges almost surely as $t \to \infty$, and therefore

$$N_i^* := \sup_{t \geqslant 0} |N_{t,i}| < \infty \qquad \text{almost surely.}$$

Let

$$P_t := \mathcal{U}_{t,i}\mathcal{V}_{t,i} = \exp(-4\gamma^2 I_t).$$

Since $I_\infty < \infty$, there exists

$$p_\infty := \exp(-4\gamma^2 I_\infty) > 0$$

such that $P_t \geqslant p_\infty$ for all t. Define

$$\mathcal{Y}_{t,i} := \widetilde{\mathcal{Z}}_{t,i} - 4\gamma N_{t,i}.$$

By (B.120),

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{Y}_{t,i} = -4\gamma P_t \sinh(\widetilde{\mathcal{Z}}_{t,i}) + 2\gamma.$$

A deterministic barrier argument gives

$$|\widetilde{\mathcal{Z}}_{t,i}| \leqslant 8\gamma N_i^* + \operatorname{arcsinh}\left(\frac{1}{2p_\infty}\right) \qquad \text{for all } t.$$

Therefore,

$$\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 = 2P_t \cosh(\widetilde{\mathcal{Z}}_{t,i}) \leqslant 2\cosh\left(8\gamma N_i^* + \operatorname{arcsinh}\left(\frac{1}{2p_\infty}\right)\right) < \infty.$$

Thus all coordinates remain bounded on finite time intervals. Since the SDE coefficients are locally Lipschitz, the standard continuation criterion rules out finite-time explosion. □

Uniform Separation from the Saddle via Risk Integrability

We now show that risk integrability also prevents the dynamics from approaching the saddle at the origin.

Lemma B.4.14 (Uniform separation from the saddle)**.** *If*

$$\int_0^\infty \mathcal{R}(\mathcal{X}_s)\,\mathrm{d}s < \infty,$$

then, for every coordinate i and all $t \geqslant 0$,

$$\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 \geqslant m,$$

where

$$m := 2\exp\left(-4\gamma^2 \int_0^\infty \mathcal{R}(\mathcal{X}_s)\,\mathrm{d}s\right) > 0.$$

Proof. By the arithmetic–geometric mean inequality,

$$\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 \geqslant 2\mathcal{U}_{t,i}\mathcal{V}_{t,i}.$$

Using Lemma B.4.1,

$$\mathcal{U}_{t,i}\mathcal{V}_{t,i} = \exp\left(-4\gamma^2 \int_0^t \mathcal{R}(\mathcal{X}_s)\,\mathrm{d}s\right).$$

Therefore

$$\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 \geqslant 2\exp\left(-4\gamma^2 \int_0^t \mathcal{R}(\mathcal{X}_s)\,\mathrm{d}s\right) \geqslant 2\exp\left(-4\gamma^2 \int_0^\infty \mathcal{R}(\mathcal{X}_s)\,\mathrm{d}s\right) =: m.$$

Since the integral is finite, $m > 0$. □

Combining Corollary B.4.11 with Lemma B.4.14, we obtain that, with probability at least $1 - \delta$, the dynamics exist globally, the risk decays exponentially, and the coordinates remain uniformly separated from the saddle:

$$\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 \geqslant 2\exp\left(-\gamma^2 \frac{M_{L_*} H_*}{\mu_{H,L}(\delta)}\right) \qquad \text{for all } t \geqslant 0,\ i = 1, \ldots, d.$$

B.5 Examples

In this section, we provide further key learning problems within our family of models.

B.5.1 Mean-squared Error

One canonical example of (3.7) is the outer function $f\colon \mathbb{R}^2 \to \mathbb{R}$ given by $f(r) = \frac{1}{2}(r_1 - r_2)^2$.

In this case, the problem is

$$\min_{x\in\mathbb{R}^{2d}} \left\{ \mathcal{R}(x) = \frac{1}{2d}\,\mathbb{E}_a\langle \psi(x) - \beta^*, a\rangle^2 = \frac{1}{2d}(\psi(x) - \beta^*)^\top K(\psi(x) - \beta^*) \right\}, \tag{B.121}$$

and satisfies

$$I\left(B(x)\right) = \mathbb{E}_a[\nabla_{r_1} f(r)^2] = 2\mathcal{R}(x) = 2h\left(B(x)\right) = B(x)_{11} - B(x)_{12} - B(x)_{21} + B(x)_{22}.$$

Moreover,

$$\begin{aligned}\frac{1}{d}\operatorname{Tr}\left(\nabla^2\mathcal{R}(x)\right) = \frac{1}{d}\sum_{i=1}^{d} K_{ii}\left[\left(2q_{11}u_i + 2q_{12}v_i + l_1\right)^2 + \left(2q_{12}u_i + 2q_{22}v_i + l_2\right)^2\right] \\ + \frac{2}{d}(q_{11} + q_{22})\sum_{i=1}^{d} K_{ii}\left(\psi(x) - \beta^*\right)_i.\end{aligned}$$

Remark B.5.1. Note that both $\mathcal{R}(x)$ and $\frac{1}{d}\operatorname{Tr}\left(\nabla^2\mathcal{R}(x)\right)$ satisfy Assumption 3.3.6. Consequently, their SGD dynamics concentrate and can be analyzed via Theorem 3.3.7.

Example B.5.2. The simplest parametrization is to consider $\psi(x) = u$ which corresponds to the classical leas-squares framework,

$$\min_{u\in\mathbb{R}^d} \frac{1}{2d}\,\mathbb{E}_a\langle u - \beta^*, a\rangle^2. \tag{B.122}$$

Here, $\psi(u, v) = u$ for $\mathcal{Q} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, $l = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, and $c = 0$.

Example B.5.3 (Diagonal Linear Network, Standard Parametrization)**.** This formulation arises in the problem of minimizing the expected squared risk over a standard depth-two diagonal linear neural network, represented by

$$\min_{u,v\in\mathbb{R}^d} \frac{1}{2d}\,\mathbb{E}_a\langle u \odot v - \beta^*, a\rangle^2. \tag{B.123}$$

Here, $\psi(u, v) = u \odot v$ for $\mathcal{Q} = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}$, $l = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and $c = 0$.

Example B.5.4 (Diagonal Linear Network, Squared Parametrization)**.** The most relevant example is the squared parametrization of the linear regression model. The optimization problem associated with this parametrization is formulated as

$$\min_{u,v\in\mathbb{R}^d} \frac{1}{4d}\,\mathbb{E}_a\langle u^2 - v^2 - \beta^*, a\rangle^2. \tag{B.124}$$

Here, $\psi(u,v) = u^2 - v^2$ for $\mathcal{Q} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, $l = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and $c = 0$.

B.5.2 Binary Logistic Regression

We consider the setting of (3.7) with outer function $f\colon \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(r) = -r_1 \cdot \frac{\exp(r_2)}{\exp(r_2)+1} + \log(\exp(r_1)+1).$$

In this case, the problem takes the form

$$\min_{x\in\mathbb{R}^{2d}} \left\{ \mathcal{R}(x) = -\,\mathbb{E}_a\langle \psi(x), a\rangle/\sqrt{d} \cdot \frac{\exp(\langle\beta^*, a\rangle/\sqrt{d})}{\exp(\langle\beta^*, a\rangle/\sqrt{d})+1} + \log(\exp(\langle\psi(x), a\rangle/\sqrt{d})+1) \right\},$$

and, using a multivariate normal distribution as in [21], satisfies

$$\mathcal{R}(x) = h\left(B(x)\right) = \frac{1}{\pi}\int_{\mathbb{R}^2} \left(-\tilde{u} \cdot \frac{\exp(\tilde{v})}{\exp(\tilde{v})+1} + \log(\exp(\tilde{u})+1) \right) \exp\left(-\begin{pmatrix} u \\ v \end{pmatrix}^\top \begin{pmatrix} u \\ v \end{pmatrix} \right) \mathrm{d}u\mathrm{d}v$$

where $\begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} = \sqrt{2}L \begin{pmatrix} u \\ v \end{pmatrix}$ and $B(x) = LL^\top$.

Analogously, since

$$\nabla_{r_1} f(r) = -\frac{\exp(r_2)}{\exp(r_2)+1} + \frac{\exp(r_1)}{\exp(r_1)+1},$$

we obtain

$$I\left(B(x)\right) = \mathbb{E}_a[\nabla_{r_1} f(r)^2] = \frac{1}{\pi}\int_{\mathbb{R}^2} \left(-\frac{\exp(\tilde{v})}{\exp(\tilde{v})+1} + \frac{\exp(\tilde{u})}{\exp(\tilde{u})+1} \right)^2 \exp\left(-\begin{pmatrix} u \\ v \end{pmatrix}^\top \begin{pmatrix} u \\ v \end{pmatrix} \right) \mathrm{d}u\mathrm{d}v.$$

B.6 Numerical Simulation Details

Here we provide more details for the figures that appear in the main paper.

Figure 3.1: Three views of empirical risk dynamics for SGD on a diagonal linear network: concentration, spectral effects, and regime transitions. *Left:* Identity covariance $K = I_d$. As the dimension d increases, the risk trajectory of SGD concentrates around a deterministic limit (shown in red), as characterized in Theorem 3.3.7. *Middle:* Power-law covariance spectrum. Our homogenized SGD prediction (transparent) from Theorem 3.3.7 closely tracks SGD (opaque) over a range of power-law exponents β, at fixed dimension $d = 10^3$. *Right:* Identity covariance $K = I_d$ at fixed dimension $d = 10^3$. Varying the stepsize γ reveals distinct convergence/divergence regimes; the homogenized prediction remains accurate even for stepsizes above the convergence threshold.

The left and right panels use the squared parametrization (Example B.5.4), whereas the middle panel uses the standard parametrization (Example B.5.3). For the power-law spectrum, eigenvalues are generated via inverse-CDF sampling: draw $u_i \sim \mathcal{U}(0,1)$ and set $\lambda_i = u_i^{1/(1-\beta)}$, yielding i.i.d. $\lambda_i \in [0,1]$ with density $p(\lambda) = (1-\beta)\lambda^{-\beta}$. We set $\beta^* = \mathbb{1}_d$ and initialize SGD at $u_0 = v_0 = \alpha\,\mathbb{1}_d$ with $\alpha = 0.6$ (left and right), and at $u_0 = 2\,\mathbb{1}_d$, $v_0 = 0$ (middle). Curves are aggregated over 30 independent stochastic runs. The stepsize is fixed to $\gamma = 0.1$ (left) and $\gamma = 1.1$ (middle); in the right panel, γ is swept across the displayed range at fixed $d = 10^3$.

Numerical construction of the theoretical curve: The contour-integral evaluation used to recover the theory curve is very sensitive to the choice of contour radius. Under symmetric initialization ($u_0 = v_0$), a single contour $C_M(0)$ suffices; we empirically verify that $\|x_k\|_\infty < M$ along the trajectory and use $M = 1.3$ with $N = 100$ contour discretization points. If M is too small, the condition $\|x_k\|_\infty < M$ is violated, leading to incorrect contour integrals and a mismatch (notably at initialization) between SGD and the PDE prediction; if M is too large, the PDE solver may overflow. Accordingly, M must be calibrated for each initialization.

Figure 3.2: Curvature dynamics for SGD on a diagonal linear network. *Left:* The evolution of the curvature measured by the scaled trace of the Hessian $\frac{1}{d}\operatorname{Tr}(\nabla^2\mathcal{R})$ is shown alongside the empirical risk $\mathcal{R}$, illustrating "flat" progress in which the risk increases

sharply accompanied by a marked drop in curvature as we vary the stepsize γ. *Right:* As the dimension d increases, the curvature dynamics of SGD concentrate around a deterministic limit (shown in red), as proven in Theorem 3.3.7. The left panel corresponds to the squared parametrization (Example B.5.4), while the right one corresponds to the standard parametrization (Example B.5.3).

Note that in the standard parametrization, the scaled Hessian trace simplifies to $\mathrm{Tr}(\nabla^2\mathcal{R}(x)) = \frac{1}{d}\sum_{i=1}^{d} K_{ii}(u_i^2 + v_i^2)$. We set $\beta^* = \mathbb{1}_d$ and $K = I_d$, and initialize SGD at $u_0 = v_0 = \alpha\,\mathbb{1}_d$ with $\alpha = 1.0$ (left) and $\alpha = 0.6$ (right). Curves are aggregated over 30 independent stochastic runs. *Left:* fixed $d = 10^3$ with a sweep over γ. *Right:* fixed $\gamma = 0.1$ with varying d. We use $M = 1.0$ and $N = 300$ for the contour construction.

Figure 3.4: Risk discrepancy between SGD and its continuous-time approximations on a diagonal linear network. For each stepsize γ, we report the absolute difference between the empirical risk of SGD after $T{\cdot}d$ iterations (with $T = 20$) and two approximations: (i) homogenized SGD (HSGD) (3.14) (blue), and (ii) stochastic gradient flow (SGF) (3.17) (pink). As γ increases, the proposed HSGD—which captures large stepsize effects—provides a progressively more accurate approximation of SGD, whereas SGF is derived under a vanishing stepsize regime and thus degrades for larger γ. Initialization scale α controls proximity to the saddle point $x = 0$: smaller α corresponds to a longer transient before learning accelerates.

We consider the squared parametrization (Example B.5.4) with mean-squared error, in the sparse recovery setting of [75]. The ground-truth signal satisfies $\beta_i^* = 1$ for $1 \leqslant i \leqslant 5$ and $\beta_i^* = 0$ for $i > 5$, with covariance $K = I_d$ and dimension $d = 100$. SGD is initialized at $u_0 = v_0 = \alpha\,\mathbb{1}_d$, with $\alpha = 0.01$ (left) and $\alpha = 0.1$ (right), and results are aggregated over 30 independent stochastic runs. Both HSGD and SGF trajectories are simulated using the Euler–Maruyama method with time step $\mathrm{d}t = 2^{-8}$.

Figure B.1: Coordinatewise entropy barriers and exponential risk decay on a diagonal linear network. The figure illustrates the entropy-barrier mechanism developed in Appendix B.4. The left panels track the two quantities that define the controlled regime of the proof: the empirical entropy H_t and the largest coordinatewise entropy density $\max_i h_{t,i}$.

The red dashed lines mark the corresponding barriers H_* and L_*. Across the tested stepsizes, the trajectories remain below these barriers, indicating that the dynamics stay in the regime where the coordinatewise entropy-barrier argument applies.

The top-right panel displays the risk–entropy ratio $4\mathcal{R}(\mathcal{X}_t)/H_t$. The red dashed lines mark empirical estimates of the coordinatewise coercivity constants m_{L_*} and M_{L_*}. These constants are estimated from the coordinate quotients

$$q_{t,i} = \frac{(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1)^2}{h_{t,i}},$$

rather than directly from the averaged ratio $4\mathcal{R}(\mathcal{X}_t)/H_t$. Specifically, we pool the lower and upper coordinate-quotient summaries across all stepsizes and stochastic runs, restricted to times satisfying $H_{\min} \leqslant H_t \leqslant H_*$ and $\max_i h_{t,i} \leqslant L_*$, and then take robust quantiles to estimate m_{L_*} and M_{L_*}. The panel visualizes the theorem's risk–entropy comparison: while the coordinatewise barrier holds, the averaged ratio $4\mathcal{R}(\mathcal{X}_t)/H_t$ should remain controlled by the same coercivity constants.

The bottom-right panel shows the risk $\mathcal{R}(\mathcal{X}_t)$ on a logarithmic scale. The dashed curves are the exponential envelopes suggested by the theorem, of the form

$$\frac{M_{L_*}}{4} H_* e^{-\mu t},$$

using the empirical constants m_{L_*} and M_{L_*} from the coordinatewise quotient estimates. In the theorem, this bound is guaranteed under a sufficient small-stepsize condition ensuring a positive decay rate. In the experiment, we plot the same exponential envelope for all tested stepsizes, rather than suppressing curves outside the certified small-stepsize range. This highlights that the predicted exponential decay behavior appears to persist empirically even for stepsizes larger than those directly covered by the sufficient condition.

Shaded regions denote the 10th–90th percentile range over independent runs, and solid curves denote the median trajectory. We simulate the isotropic squared-parametrization setting with $\beta^* = \mathbb{1}_d$, covariance $K = I_d$, Gaussian covariates $a \sim \mathcal{N}(0, I_d)$, and initialization $u_0 = v_0 = \mathbb{1}_d$. The dimension is $d = 10^3$, the time horizon is $T = 30$, and results are aggregated over 30 independent stochastic runs for each constant stepsize $\gamma \in \{0.04, 0.03, 0.02, 0.01, 0.005, 0.001\}$. The horizontal axis is the rescaled time $t = k/d$, i.e., SGD iterations divided by the dimension.

BIBLIOGRAPHY

[1] Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A Modern Look at the Relationship between Sharpness and Generalization. In *Proceedings of the 40th International Conference on Machine Learning*, pages 840–902. PMLR, 2023.

[2] Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. SGD with Large Step Sizes Learns Sparse Features. In *Proceedings of the 40th International Conference on Machine Learning*, pages 903–925. PMLR, 2023.

[3] Luca Arnaboldi, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. Escaping mediocrity: How two-layer networks learn hard generalized linear models with SGD. In *OPT2023: 15th Annual Workshop on Optimization for Machine Learning*, 2023. arXiv preprint.

[4] Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks. In *Proceedings of Thirty Sixth Conference on Learning Theory*, pages 1199–1227. PMLR, 2023.

[5] Luca Arnaboldi, Bruno Loureiro, Ludovic Stephan, Florent Krzakala, and Lenka Zdeborova. Asymptotics of SGD in sequence-single index models and single-layer attention networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[6] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[7] Krishna B. Athreya and Peter Ney. *Branching Processes*. Dover Publications, Mineola, N.Y, 2004. ISBN 978-0-486-43474-2.

[8] Amit Attia and Tomer Koren. SGD with AdaGrad Stepsizes: Full Adaptivity with High Probability to Unknown Parameters, Unbounded Gradients and Affine Variance. In *Proceedings of the 40th International Conference on Machine Learning*, pages 1147–1171. PMLR, 2023.

[9] Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E. Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the Implicit Bias of Initialization Shape: Beyond Infinitesimal Mirror Descent. In *Proceedings of the 38th International Conference on Machine Learning*, pages 468–477. PMLR, 2021.

[10] Krishnakumar Balasubramanian, Promit Ghosal, and Ye He. High-dimensional scaling limits and fluctuations of online least-squares SGD with smooth covariance. *The Annals of Applied Probability*, 35(5), 2025. ISSN 1050-5164. doi: 10.1214/24-AAP2123.

[11] Nicholas Barnfield, Hugo Cui, and Yue M. Lu. High-dimensional analysis of single-layer attention for sparse-token classification. In *The Fourteenth International Conference on Learning Representations*, 2026.

[12] Leonard Berrada, Andrew Zisserman, and M. Pawan Kumar. Training Neural Networks for and by Interpolation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 799–809. PMLR, 2020.

[13] Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight Nonparametric Convergence Rates for Stochastic Gradient Descent under the Noiseless Linear Model. In *Advances in Neural Information Processing Systems*, volume 33, pages 2576–2586. Curran Associates, Inc., 2020.

[14] M Biehl and P Riegler. On-Line Learning with a Perceptron. *Europhysics Letters (EPL)*, 28(7):525–530, 1994. ISSN 0295-5075, 1286-4854. doi: 10.1209/0295-5075/28/7/012.

[15] M Biehl and H Schwarze. Learning by on-line gradient descent. *Journal of Physics*

A: Mathematical and General, 28(3):643–656, 1995. ISSN 0305-4470, 1361-6447. doi: 10.1088/0305-4470/28/3/018.

[16] Blake Bordelon and Cengiz Pehlevan. Learning Curves for SGD on Structured Features. In *International Conference on Learning Representations*, 2022.

[17] Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data, 2021.

[18] Kabir Aladin Chandrasekher, Ashwin Pananjady, and Christos Thrampoulidis. Sharp global convergence guarantees for iterative nonconvex optimization with random data. *The Annals of Statistics*, 51(1), 2023. ISSN 0090-5364. doi: 10.1214/22-AOS2246.

[19] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability. In *International Conference on Learning Representations*, 2021.

[20] Elizabeth Collins–Woodfin and Elliot Paquette. High-dimensional limit of one-pass SGD on least squares. *Electronic Communications in Probability*, 29(none), 2024. ISSN 1083-589X. doi: 10.1214/23-ECP571.

[21] Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the High-dimensional notes: An ODE for SGD learning dynamics on GLMs and multi-index models. *Information and Inference: A Journal of the IMA*, 13(4): iaae028, 2024. ISSN 2049-8772. doi: 10.1093/imaiai/iaae028.

[22] Elizabeth Collins-Woodfin, Inbar Seroussi, Begoña García Malaxechebarría, Andrew Mackenzie, Elliot Paquette, and Courtney Paquette. The High Line: Exact Risk and Learning Rate Curves of Stochastic Adaptive Learning Rate Algorithms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[23] Enea Monzio Compagnoni, Tianlin Liu, Rustem Islamov, Frank Norbert Proske, Antonio Orvieto, and Aurelien Lucchi. Adaptive Methods through the Lens of SDEs: Theoretical Insights on the Role of Noise. In *The Thirteenth International Conference on Learning Representations*, 2025.

[24] Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the Landscape Boosts the Signal for SGD: Optimal Sample Complexity for Learning Single Index Models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 752–784. Curran Associates, Inc., 2023.

[25] Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborova, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, Vienna, Austria, 2024. JMLR.org.

[26] Aaron Defazio and Konstantin Mishchenko. Learning-Rate-Free Learning by D-Adaptation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 7449–7479. PMLR, 2023.

[27] Lijun Ding, Dmitriy Drusvyatskiy, Maryam Fazel, and Zaid Harchaoui. Flat minima generalize for low-rank matrix recovery. *Information and Inference: A Journal of the IMA*, 13(2):iaae009, 2024. ISSN 2049-8772. doi: 10.1093/imaiai/iaae009.

[28] John Duchi, Elad Hazan, and Yoram Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12(61): 2121–2159, 2011.

[29] Darina Dvinskikh, Aleksandr Ogaltsov, Alexander Gasnikov, Pavel Dvurechensky, Alexander Tyurin, and Vladimir Spokoiny. Adaptive Gradient Descent for Convex and Non-Convex Stochastic Optimization, 2020.

[30] Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley Series in Probability and Mathematical Statistics. Wiley-

Interscience, Hoboken, NJ, 1986. ISBN 978-0-470-31665-8 978-0-470-31732-7. doi: 10.1002/9780470316658.

[31] Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (S)GD over Diagonal Linear Networks: Implicit bias, Large Stepsizes and Edge of Stability. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 29406–29448. Curran Associates, Inc., 2023.

[32] Matthew Faw, Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Beyond Uniform Smoothness: A Stopped Analysis of Adaptive SGD. In *Proceedings of Thirty Sixth Conference on Learning Theory*, pages 89–160. PMLR, 2023.

[33] Cédric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborová. Rigorous Dynamical Mean-Field Theory for Stochastic Gradient Descent Methods. *SIAM Journal on Mathematics of Data Science*, 6(2):400–427, 2024. doi: 10.1137/23M1594388.

[34] Sebastian Goldt, Madhu Advani, Andrew Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[35] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the Influence of Data Structure on Learning in Neural Networks: The Hidden Manifold Model. *Physical Review X*, 10(4):041044, 2020. ISSN 2160-3308. doi: 10.1103/PhysRevX.10.041044.

[36] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. The Gaussian equivalence of generative models for learning with shallow neural networks. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, pages 426–471. PMLR, 2022.

[37] Robert M. Gower, Aaron Defazio, and Michael Rabbat. Stochastic Polyak Stepsize with a Moving Target. In *OPT2021: 13th Annual Workshop on Optimization for Machine Learning*, 2021. arXiv preprint.

[38] Jeff Z. HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape Matters: Understanding the Implicit Bias of the Noise Covariance. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pages 2315–2357. PMLR, 2021.

[39] Elad Hazan and Sham Kakade. Revisiting the Polyak step size, 2019.

[40] Fengxiang He, Tongliang Liu, and Dacheng Tao. Control Batch Size and Learning Rate to Generalize Well: Theoretical and Empirical Evidence. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[41] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent, 2012.

[42] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: Closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[43] Maor Ivgi, Oliver Hinder, and Yair Carmon. DoG is SGD's Best Friend: A Parameter-Free Dynamic Step Size Schedule. In *Proceedings of the 40th International Conference on Machine Learning*, pages 14465–14499. PMLR, 2023.

[44] Stanisław Jastrzębski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amost Storkey. On the Relation Between the Sharpest Directions of DNN Loss and the SGD Step Length. In *International Conference on Learning Representations*, 2019.

[45] Xiaowen Jiang and Sebastian U Stich. Adaptive SGD with Polyak stepsize and Line-search: Robust Convergence and Variance Reduction. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 26396–26424. Curran Associates, Inc., 2023.

[46] Dayal Singh Kalra, Jean-Christophe Gagnon-Audet, Andrey Gromov, Ishita Mediratta, Kelvin Niu, Alexander H. Miller, and Michael Shvartsman. A Scalable Measure of Loss Landscape Curvature for Analyzing the Training Dynamics of LLMs, 2026.

[47] Varun Kanade, Patrick Rebeschini, and Tomas Vaškevičius. The statistical complexity of early-stopped mirror descent. *Information and Inference: A Journal of the IMA*, 12 (4):3010–3041, 2023. ISSN 2049-8772. doi: 10.1093/imaiai/iaad047.

[48] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *International Conference on Learning Representations*, 2017.

[49] Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An Alternative View: When Does SGD Escape Local Minima? In *Proceedings of the 35th International Conference on Machine Learning*, pages 2698–2707. PMLR, 2018.

[50] Harold J. Kushner and Dean S. Clark. *Weak Convergence for Unconstrained Systems*, volume 26, pages 106–157. Springer New York, New York, NY, 1978. ISBN 978-0-387-90341-5 978-1-4684-9352-8. doi: 10.1007/978-1-4684-9352-8_4.

[51] Kiwon Lee, Andrew Nicholas Cheng, Elliot Paquette, and Courtney Paquette. Trajectory of Mini-Batch Momentum: Batch Size Saturation and Convergence in High Dimensions. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[52] Kfir Levy. Online to Offline Conversions, Universality and Adaptive Minibatch Sizes. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[53] Kfir Y. Levy, Alp Yurtsever, and Volkan Cevher. Online Adaptive Methods, Universality and Acceleration. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[54] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic Modified Equations and Dynamics

of Stochastic Gradient Algorithms I: Mathematical Foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.

[55] Xiaoyu Li and Francesco Orabona. On the Convergence of Stochastic Gradient Descent with Adaptive Stepsizes. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 983–992. PMLR, 2019.

[56] Liming Liu, Zixuan Zhang, Simon Shaolei Du, and Tuo Zhao. A Minimalist Example of Edge-of-Stability and Progressive Sharpening. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[57] Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic Polyak Step-size for SGD: An Adaptive Learning Rate for Fast Convergence. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR, 2021.

[58] Lachlan Ewen MacDonald, Jack Valmadre, and Simon Lucey. On progressive sharpening, flat minima and generalisation, 2023.

[59] Begoña García Malaxechebarría, Courtney Paquette, Maryam Fazel, and Dmitriy Drusvyatskiy. High-dimensional limit of sgd for diagonal linear networks, 2026.

[60] Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and Scaling Rules for Adaptive Gradient Algorithms. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[61] H. Brendan McMahan and Matthew Streeter. Adaptive Bound Optimization for Online Convex Optimization, 2010.

[62] Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification. In *Advances in Neural Information Processing Systems*, volume 33, pages 9540–9550. Curran Associates, Inc., 2020.

[63] Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A. Erdogdu. Neural Networks Efficiently Learn Low-Dimensional Representations with SGD. In *The Eleventh International Conference on Learning Representations*, 2023.

[64] Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, and Daniel Soudry. Implicit Bias of the Step Size in Linear Diagonal Neural Networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16270–16295. PMLR, 2022.

[65] Preetum Nakkiran, Behnam Neyshabur, and Hanie Sedghi. The Deep Bootstrap Framework: Good Online Learners are Good Offline Generalizers, 2020.

[66] Angelia Nedić and Dimitri Bertsekas. Convergence Rate of Incremental Subgradient Algorithms. In Panos M. Pardalos, Donald Hearn, Stanislav Uryasev, and Panos M. Pardalos, editors, *Stochastic Optimization: Algorithms and Applications*, volume 54, pages 223–264. Springer US, Boston, MA, 2001. ISBN 978-1-4419-4855-7 978-1-4757-6594-6. doi: 10.1007/978-1-4757-6594-6_11.

[67] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2006. ISBN 978-0-387-30303-1. doi: 10.1007/978-0-387-40065-5.

[68] Antonio Orvieto, Simon Lacoste-Julien, and Nicolas Loizou. Dynamics of SGD with Stochastic Polyak Stepsizes: Truly Adaptive Variants and Convergence to Exact Solution. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[69] Courtney Paquette and Elliot Paquette. Dynamics of Stochastic Momentum Methods on Large-scale, Quadratic Models. In *Advances in Neural Information Processing Systems*, volume 34, pages 9229–9240. Curran Associates, Inc., 2021.

[70] Courtney Paquette and Katya Scheinberg. A Stochastic Line Search Method with Expected Complexity Analysis. *SIAM Journal on Optimization*, 30(1):349–376, 2020. ISSN 1052-6234, 1095-7189. doi: 10.1137/18M1216250.

[71] Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3548–3626. PMLR, 2021.

[72] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Implicit regularization or implicit conditioning? exact risk trajectories of SGD in high dimensions. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 978-1-7138-7108-8.

[73] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of SGD in high-dimensions: Exact dynamics and generalization properties. *Mathematical Programming*, 214(1-2):1–90, 2025. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-024-02171-3.

[74] Elliot Paquette and Courtney Paquette. High-dimensional Optimization, 2022.

[75] Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit Bias of SGD for Diagonal Linear Networks: A Provable Benefit of Stochasticity. In *Advances in Neural Information Processing Systems*, volume 34, pages 29218–29230. Curran Associates, Inc., 2021.

[76] B. T. Polyak. *Introduction to optimization*. Translations series in mathematics and engineering. Optimization Software, Publications Division, New York, 1987. ISBN 978-0-911575-14-9.

[77] Michal Rolinek and Georg Martius. L4: Practical loss-based stepsize adaptation for

deep learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[78] David Saad and Sara Solla. Dynamics of On-Line Gradient Descent Learning for Multilayer Neural Networks. In *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995.

[79] David Saad and Sara A. Solla. Exact Solution for On-Line Learning in Multilayer Neural Networks. *Physical Review Letters*, 74(21):4337–4340, 1995. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.74.4337.

[80] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

[81] Aditya Varre, Margarita Sagitova, and Nicolas Flammarion. SGD vs GD: Rank Deficiency in Linear Networks. In *High-Dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024.

[82] Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of SGD for Least-Squares in the Interpolation regime. In *Advances in Neural Information Processing Systems*, volume 34, pages 21581–21591. Curran Associates, Inc., 2021.

[83] Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit Regularization for Optimal Sparse Recovery. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[84] Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless Stochastic Gradient: Interpolation, Line-Search, and Convergence Rates. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[85] Sharan Vaswani, Issam Laradji, Frederik Kunstner, Si Yi Meng, Mark Schmidt, and Simon Lacoste-Julien. Adaptive gradient methods converge faster with over-

parameterization (but you should do a line-search). In *OPT2020: 12th Annual Workshop on Optimization for Machine Learning*, 2020. arXiv preprint.

[86] Maksim Velikanov, Denis Kuznedelev, and Dmitry Yarotsky. A view of mini-batch SGD via generating functions: Conditions of convergence, phase transitions, benefit from negative momenta. In *The Eleventh International Conference on Learning Representations*, 2023.

[87] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[88] Loucas Pillaud Vivien, Julien Reygner, and Nicolas Flammarion. Label noise (stochastic) gradient descent implicitly solves the Lasso for quadratic parametrisation. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 2127–2159. PMLR, 2022.

[89] Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of AdaGrad for Non-convex Objectives: Simple Proofs and Relaxed Assumptions. In *Proceedings of Thirty Sixth Conference on Learning Theory*, pages 161–190. PMLR, 2023.

[90] Chuang Wang, Jonathan Mattingly, and Yue M. Lu. Scaling Limit: Exact and Tractable Analysis of Online Learning Algorithms with Applications to Regularized Regression and PCA, 2017.

[91] Chuang Wang, Hong Hu, and Yue Lu. A Solvable High-Dimensional Model of GAN. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[92] Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad Stepsizes: Sharp Convergence Over Nonconvex Landscapes. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6677–6686. PMLR, 2019.

[93] Alexander Wei, Wei Hu, and Jacob Steinhardt. More Than a Toy: Random Matrix Models Predict How Real-World Neural Representations Generalize. In *Proceedings of*

the 39th International Conference on Machine Learning, pages 23549–23588. PMLR, 2022.

[94] Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How Sharpness-Aware Minimization Minimizes Sharpness? In *The Eleventh International Conference on Learning Representations*, 2023.

[95] Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and Rich Regimes in Overparametrized Models. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

[96] Xiaoxia Wu, Rachel Ward, and Léon Bottou. WNGrad: Learn the Learning Rate in Gradient Descent, 2020.

[97] Yuege Xie, Xiaoxia Wu, and Rachel Ward. Linear Convergence of Adaptive Stochastic Gradient Descent. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 1475–1485. PMLR, 2020.

[98] Junchi YANG, Xiang Li, Ilyas Fatkhullin, and Niao He. Two Sides of One Coin: The Limits of Untuned SGD and the Power of Adaptive Methods. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 74257–74288. Curran Associates, Inc., 2023.

[99] Geonhui Yoo, Minhak Song, and Chulhee Yun. Understanding Sharpness Dynamics in NN Training with a Minimalist Example: The Effects of Dataset Difficulty, Depth, Stochasticity, and More. In *Forty-Second International Conference on Machine Learning*, 2025.

[100] Yuki Yoshida and Masato Okada. Data-Dependence of Plateau Phenomenon in Learning with Neural Network — Statistical Mechanical Analysis. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[101] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[102] Libin Zhu, Chaoyue Liu, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Catapults in SGD: Spikes in the training loss and their impact on generalization through feature learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, Vienna, Austria, 2024. JMLR.org.