Analyses of High-dimensional Stochastic Algorithms: Towards Adaptive Stepsizes

Andrew W. Mackenzie

Department of Mathematics and Statistics

McGill University, Montreal

December, 2024

A thesis submitted to McGill University in partial fulfillment of the

requirements of the degree of Master of Science

Table of Contents

Abstract

The aim of this work is to provide a precise characterization of how stochastic adaptive learning rate algorithms behave in high-dimensional settings, particularly focusing on their dynamics when both the number of parameters and samples are large. While adaptive learning rate strategies like AdaGrad and Adam have shown immense practical success in machine learning applications, our theoretical understanding of their behavior remains limited, especially regarding how they interact with the geometry of high-dimensional optimization problems.

We develop a framework for analyzing the exact dynamics of both the risk and learning rates for stochastic adaptive algorithms on a class of problems we call "high-dimensional linear composite functions." In this setting, we prove that as dimension grows, the training dynamics concentrate around deterministic curves described by a system of ordinary differential equations (ODEs). This concentration result enables us to derive precise predictions about how adaptive algorithms interact with problem geometry, leading to several insights: First, we show that exact line search (greedily minimizing the risk at each step) can perform arbitrarily worse than simpler strategies like Polyak stepsize when the data covariance exhibits strong anisotropy. Second, we prove that AdaGrad-Norm automatically discovers near-optimal stepsizes when initialized properly, though it pays a constant factor penalty related to the initial distance to optimality. Finally, we demonstrate how AdaGrad-Norm's behavior undergoes a phase transition on problems with power-law spectra - for easier problems it maintains a constant learning rate, while for harder problems it adopts a specific decay schedule that we characterize exactly.

Our framework provides one of the first theoretical approaches that can meaningfully distinguish between different adaptive algorithms that achieve minimax optimal rates. The resulting ODEs offer remarkably accurate predictions of algorithmic behavior even for medium-sized problems, as we verify through numerical experiments. This work takes important steps toward a more complete understanding of adaptive optimization in the high-dimensional regime that increasingly characterizes modern machine learning problems.

Along with our results, which may seem somewhat technical and specific, we provide a brief history of the various approaches taken to analyze SGD. This situates our work as part of an ongoing long-term project, one we hope will fully explain what is emerging as the most consequential algorithm of our time.

Abrégé

Ce travail vise à fournir une caractérisation précise du comportement des algorithmes stochastiques à taux d'apprentissage adaptatif dans des contextes de haute dimension, en se concentrant particulièrement sur leur dynamique lorsque le nombre de paramètres et d'échantillons est élevé. Bien que les stratégies de taux d'apprentissage adaptatif comme AdaGrad et Adam aient démontré un immense succès pratique dans les applications d'apprentissage automatique, notre compréhension théorique de leur comportement reste limitée, notamment concernant leur interaction avec la géométrie des problèmes d'optimisation en haute dimension.

Nous développons un cadre pour analyser les dynamiques exactes du risque et des taux d'apprentissage pour les algorithmes adaptatifs stochastiques sur une classe de problèmes que nous appelons "fonctions composites linéaires de haute dimension". Dans ce contexte, nous prouvons qu'à mesure que la dimension augmente, les dynamiques d'entraînement se concentrent autour de courbes déterministes décrites par un système d'équations différentielles ordinaires (EDO). Ce résultat de concentration nous permet de dériver des prédictions précises sur la façon dont les algorithmes adaptatifs interagissent avec la géométrie du problème, menant à plusieurs observations : Premièrement, nous montrons que la recherche linéaire exacte (minimisation gloutonne du risque à chaque étape) peut donner des résultats arbitrairement moins bons que des stratégies plus simples comme le pas de Polyak lorsque la covariance des données présente une forte anisotropie. Deuxièmement, nous prouvons qu'AdaGrad-Norm découvre automatiquement des pas quasi-optimaux lorsqu'il est correctement initialisé, bien qu'il paie une pénalité en facteur

constant liée à la distance initiale à l'optimalité. Enfin, nous démontrons comment le comportement d'AdaGrad-Norm subit une transition de phase sur des problèmes avec des spectres en loi de puissance - pour les problèmes plus simples, il maintient un taux d'apprentissage constant, tandis que pour les problèmes plus difficiles, il adopte un programme de décroissance spécifique que nous caractérisons exactement.

Notre cadre fournit l'une des premières approches théoriques permettant de distinguer significativement différents algorithmes adaptatifs qui atteignent des taux minimax optimaux. Les EDO résultantes offrent des prédictions remarquablement précises du comportement algorithmique même pour des problèmes de taille moyenne, comme nous le vérifions par des expériences numériques. Ce travail constitue des avancées importantes vers une compréhension plus complète de l'optimisation adaptative dans le régime de haute dimension qui caractérise de plus en plus les problèmes modernes d'apprentissage automatique.

Avec nos résultats, qui peuvent sembler quelque peu techniques et spécifiques, nous fournissons un bref historique des différentes approches utilisées pour analyser la SGD. Cela situe notre travail comme faisant partie d'un projet à long terme en cours, dont nous espérons qu'il expliquera pleinement ce qui s'impose comme l'algorithme le plus important de notre époque.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor, Courtney Paquette. Unfailingly encouraging, she would answer every last one of my questions, even when I stubbornly refused to believe her. Above and beyond technical math, she helped me with everything, both concrete and undefinable, needed for successful research, all while providing insight into the ins and outs of academia and industry. I'm not sure what magic she used, but there was never a day I would leave her office without being inspired anew.

Next, thank you to Elliot Paquette. His impromptu derivations have shown me what true mathematical mastery looks like, and his combination of reassurance, insight and sangfroid have given me yet another ideal to aspire to.

Thank you to my wonderful friend Tomer Moran. A year ahead of me in undergrad, he infected me with his love for real analysis; late-night presentations on PDEs in my basement turned into years of collaboration, in school and outside of it. Tomer has always had my back; be it in startups, New York City, or skiing trips, he has been a constant source of wisdom and energy.

Thank you to the best parents in the world. All else aside, not every high-schooler has the privilege of being taught linear algebra in a little home library - my dad is the one who, early on, sparked my love of real math. And my mom, who decided to homeschool us kids while working nights in emergency, was really the one who made all of this possible.

Thank you to my grandparents, who left their lives in Russia to provide their grandchildren with the sort of opportunities I have now. Thank you to my siblings, who've been

with me through thick and thin. And, finally, thank you to all the rest of my friends, who continue to make my life bright and unpredictable.

Contributions

This thesis is based largely on a recent paper, *The High Line: Exact Dynamics of Stochastic Adaptive Learning Algorithms* [17]. The main novelty in this paper is its general analysis of adaptive learning rates. Subject to some Lipschitz conditions, we provide an exact system of ODEs for SGD algorithms with stepsizes depending on the history of losses and gradient norms up to the current moment. Previous analogues required the stepsize to be determined in advance, limiting their applicability to the most efficient and widely-used variants of SGD.

Along with the theoretical foundations it introduces, this paper demonstrates the applicability of its methods by dissecting, among others, the AdaGrad-Norm algorithm. We categorize a variety of different possible convergence rates for AdaGrad-Norm, depending on the eigenvalues of the underlying data distribution, and identify a phase transition in the case where these eigenvalues follow a power law.

These results are part of a collaborative project; the main contributions from Andrew Mackenzie, the author of this thesis, are contained Section 5.3. He generalized the ODE in [15] from deterministic to adaptive stepsizes and provided the error bounds necessary to show concentration in this new setting. Andrew is also responsible for the numerical simulations involving AdaGrad-Norm (Figure 4.1) and the experiments on real data (Section 5.7.2).

Elizabeth Collins-Woodfin and Inbar Seroussi used the resulting ODE to analyze the behaviours of AdaGrad-Norm and Polyak stepsize, while Begoña García Malaxechebarría

investigated line search. Courtney and Elliot Paquette, the advisors of the author, provided intuition, high-level planning, and direction.

List of Figures

List of Tables

Chapter 1

Introduction

The current state of artificial intelligence (AI) is marked by a dynamic tension between empirical success and theoretical foundations, with practice often leading theory by several years, if not decades. While large-scale experiments drive many of the field's most significant discoveries, the underlying mathematical principles often remain unclear until years after their practical validation.

At the heart of modern AI lies a remarkably simple algorithm: stochastic gradient descent (SGD) [64]. Given a loss function $f(w)$ to minimize, SGD iteratively updates its estimate of the optimum via

$$w_{k+1} = w_k - \eta_k \widetilde{\nabla} f(w_k), \tag{1.1}$$

where η_k is a learning rate and $\widetilde{\nabla} f(w_k)$ is a stochastic gradient - a noisy but unbiased estimate of the true gradient $\nabla f(w_k)$. Despite its simplicity, this basic update rule, along with its adaptive variants, underlies virtually all of contemporary deep learning.

What puts this algorithm in such a unique position is the sheer scale at which it is currently run. One of the major discoveries of the past decade has been the fact that neural networks become predictably more capable as they grow larger. Consequently, modern architectures contain billions or even trillions of parameters, operating in spaces of correspondingly high dimension. SGD remains the only optimization algorithm capable of training these models effectively, but the computational resources required are immense,

to the point where companies are now negotiating nuclear power plant purchases for their datacenters.

At this scale, experimentation becomes prohibitively expensive, and the field's reliance on empirical heuristics becomes a liability. A particularly recent example of this theory-practice gap can be found in the evolution of our understanding of neural network scaling laws.

In 2022, researchers at DeepMind [34] demonstrated "optimal" compute trade-offs, or "scaling laws," for large language models, refining similar empirical work from OpenAI [37]. Very recently, theoretical work [60] derived these same scaling relationships from first principles. While the general power-law form remained constant, the new analysis included previously overlooked lower-order terms. These terms, difficult to detect experimentally, had potentially multi-million dollar implications for industrial AI training, throwing into stark relief the incompleteness of our understanding of SGD dynamics.

This thesis aims to contribute to closing this theory-practice gap by developing rigorous mathematical foundations for modern optimization techniques. Building on exact analyses of vanilla SGD and momentum methods, we extend these tools to adaptive learning rate algorithms such as AdaGrad-Norm [47]. This extension is particularly significant given that adaptive optimizers like AdamW [46] have become the de facto standard in contemporary deep learning practice.

To properly contextualize our contributions, we begin with a survey of both the mathematical and practical evolution of the field. We present several archetypal proofs from different eras, not only for their technical content but also to illustrate how theoretical focuses shifted as computational capabilities expanded. This dual perspective highlights how hardware constraints, practical needs, intuitive engineering and mathematical insights have shaped the development of optimization algorithms.

The technical core of our work draws heavily on random matrix theory and general high-dimensional probability. Given that these mathematical tools may be unfamiliar to some readers, we provide a self-contained introduction to the key concepts and techniques

needed to understand our main results. Finally, with this foundation established, we proceed to our analyses.

Chapter 2

History

Theoretical work on SGD has, unsurprisingly, evolved alongside its practical applications. The first major discussion of SGD [64], in 1951, made no reference to implementation details, or even the concept that the algorithm could be run on a computer. (Coincidentally, the first commercial computer, the UNIVAC I, was released that same year; it had a production volume of 46 units.) Due to the virtual absence of computational power, stochastic optimization was primarily of mathematical interest. Research at the time guaranteed convergence in the limit, and did not concern itself with rates.

Years passed; Moore's law advanced and research kept pace. In 1986, Geoffrey Hinton described how neural networks could be trained using backpropagation and gradient descent. By then, computers had moved from vacuum tubes to integrated circuits, and the leading chip, Intel's i386, sold for $299 and ran at around 4M instructions per second. While still largely academic, SGD was now regularly run in practice. Time complexity considerations were now inevitable.

In this context, the decade saw tight asymptotic bounds on the convergence rate of all versions of SGD, along with variants achieving this theoretical limit. In terms of implementation, a new momentum algorithm [67] achieved significant practical speedups. Most SGD runs, however, were one-offs, so analyses held the size of the network and dataset

constant when determining time complexity; only 30 years later would this assumption be revised.

A defining moment came in 2011, when a group, led by, once again, Geoffrey Hinton[1], demonstrated deep learning's practical viability. Training on a pair of NVIDIA GTX 580 GPUs, they dramatically improved the state-of-the-art on the ImageNet challenge with a convolutional neural network and momentum SGD[2]. With teraflops now at every researcher's disposal, algorithmic innovation surged: RMSProp [32], AdaGrad [23], and Adam [39] brought adaptive learning rates that transformed training dynamics. The importance of scale was dramatically validated by GPT-2 [63] and GPT-3 [12]'s capability jumps in 2020; optimization theory is still struggling to fully explain these dimension-dependent, average-case[3], nonconvex results.

As of late 2024, massive GPU clusters are being built: X.ai's Memphis datacenter, for example, houses 100K NVIDIA H100s, each offering a petaflop of float16 arithmetic. AI research shows signs of becoming an epoch-defining project and is increasingly facilitated by the very language models it produces. In a race between countries and labs, a significant proportion of discoveries are kept private, making it difficult to pinpoint the state of the art; the limits of scaling, in particular, remain a major open question.

2.1 Theoretical Beginnings

2.1.1 Robbins-Munro [1951]

The first mathematical foundations of stochastic gradient descent were laid by Herbert Robbins and Sutton Monro in their now-classic 1951 paper *A Stochastic Approximation Method* [64]. The paper mentions in passing that the idea for the algorithm presented was first suggested in Naval Ordnance Report No. 65, back in 1946. (The writer of said report just so happened to be J. W. Tukey. Some 20 years later, motivated by the need to

[1]Nobel Prize: well deserved.
[2]Unchanged since the 80s.
[3]As contrasted with worst-case; here the average is taken over the randomness in the gradient noise.

detect Soviet nuclear weapons tests from seismometer readings, he would rediscover and popularize the FFT algorithm.)

Robbins and Monro's key contribution was proving convergence for a general class of stochastic approximation methods. Consider finding a root $\theta^* \in \mathbb{R}$ of $M(\theta) = \alpha$, where $M(\theta) = \mathbb{E}[Y(\theta)]$ for some family of random variables $(Y(\theta))_{\theta \in \mathbb{R}}$. Rather than requiring direct access to $M(\theta)$, they proposed the iteration

$$\theta_{n+1} = \theta_n - a_n(Y(\theta_n) - \alpha), \tag{2.1}$$

where $\{a_n\}$ is a sequence of positive steps satisfying:

$$\sum a_n = \infty, \quad \sum a_n^2 < \infty. \tag{2.2}$$

Under these conditions on the step sizes $\{a_n\}$, assuming the random observations have bounded variance (i.e., $\sup_\theta \mathbb{E}[(Y(\theta) - M(\theta))^2] < \infty$), and some further assumptions on the behaviour of the function M (e.g., monotonicity and bounds on its growth), the paper proved, through fairly straightforward algebra, that θ_n would converge in L^2 to the root θ^*.

Phrased as it is, the algorithm above does not resemble standard SGD: there are no gradients mentioned; nothing is being minimized. Note, however, that minimizing a function $f(\theta)$ is, under certain conditions, equivalent to solving $f'(\theta) = 0$. Set $\alpha = 0$ and $M(\theta) = f'(\theta)$. Let $g(\theta)$ be a random variable such that $f'(\theta) = \mathbb{E}[g(\theta)]$. Then the Robbins-Monro update becomes

$$\theta_{n+1} = \theta_n - a_n g(\theta_n),$$

which is immediately recognizeable as single-variable SGD with stepsize schedule a_n.

Robbins and Monro did not, at this time, inquire further into the rates at which their algorithm would converge. Neither did they they search for the optimal values of a_n.

However, their analysis was groundbreaking in showing that noisy information was sufficient for exact optimization – a result that would kick off an entire new field.

2.1.2 Kiefer-Wolfowitz [1952]

While Robbins and Monro showed how to find roots given noisy function evaluations, Jack Kiefer and Jacob Wolfowitz [38] turned their attention to the problem of minimization, bringing the theory closer to what we now recognize as stochastic gradient descent. In cases where gradients were unavailable, they proposed an direct solution: approximate the gradient using finite differences. For a function $f(w)$, $w \in \mathbb{R}$, they suggested the iteration

$$w_{n+1} = w_n - a_n \frac{Y(w_n + c_n) - Y(w_n - c_n)}{2c_n}, \tag{2.3}$$

where $Y(w)$ is a noisy observation of $f(w)$, and $\{a_n\}, \{c_n\}$ are sequences of positive numbers converging to zero.

This scheme represents one of the earliest forms of what we would now recognize as SGD. The key difference from modern methods lies in the gradient computation: rather than using automatic differentiation (or "backpropagation"), Kiefer-Wolfowitz relies on finite differences, requiring two function evaluations per update. While theoretically sound, this approach becomes prohibitively expensive in high dimensions, requiring $2d$ evaluations for a d-dimensional problem.

2.1.3 Perceptron [1958] and ADALINE [1960]

The late 1950s saw the first marriage of stochastic optimization with neural computation, albeit in extremely simplified forms. Two key developments emerged almost simultaneously: Rosenblatt's Perceptron [66] and Widrow and Hoff's ADALINE (ADAptive LInear NEuron) [79]. Notably, both systems were implemented in custom-built hardware rather than general-purpose computers—a necessity in an era when digital computers were still rare and expensive.

The Mark I Perceptron, built at Cornell Aeronautical Laboratory, was a room-sized machine that used motor-driven potentiometers for weights and photocells for inputs. It implemented a simple binary threshold function[4],

$$\hat{y} = \text{sign}(w^T x) = \begin{cases} 1 & \text{if } w^T x > 0 \\ -1 & \text{otherwise} \end{cases} \tag{2.4}$$

Its training algorithm updated weights only on misclassified examples:

$$w_{t+1} = \begin{cases} w_t + \eta\, y_t x_t & \text{if } y_t(w_t^T x_t) \leq 0 \\ w_t & \text{otherwise} \end{cases} \tag{2.5}$$

where y_t is the true label. While theoretically guaranteed to converge for linearly separable data, the discontinuous threshold function made analysis difficult.

ADALINE, developed at Stanford and also implemented in specialized hardware, took a different approach. Instead of thresholding, it worked directly with the linear response $w^T x$, minimizing the squared error $(y - w^T x)^2$. This seemingly minor change had profound implications: the resulting system could be analyzed using classical optimization theory, and its stochastic gradient descent update

$$w_{t+1} = w_t + \eta(y_t - w_t^T x_t)x_t \tag{2.6}$$

became a template for future neural network training algorithms.[5]

These early systems, while primitive by modern standards, demonstrated two crucial points: first, that neural-inspired architectures could be trained using stochastic optimization, and second, that careful choice of objective function could make the difference between tractable and intractable analysis. Both insights would prove vital in the devel-

[4]Modern formulations have minor differences, but we stick with the original presentation here.
[5]Note that this is exactly linear regression.

opment of deep learning decades later, when the shift to general-purpose computers and eventually GPUs would enable far more complex architectures.

2.2 Digital Implementations

By the 1970s, vacuum tubes had given way to microprocessors, and Robbins-Monro's abstract algorithm could finally be tested at scale. A new generation of researchers, armed with DEC VAX minicomputers, gradually began exploring what worked in practice. The results were sometimes surprising: theoretical guarantees of asymptotic convergence offered little guidance on whether a network would train in hours or months, and seemingly minor implementation choices like momentum and learning rate schedules proved crucial. Thus a new era of theoretical work on SGD began.

2.2.1 Backpropagation [1986]

In 1986, Geoffrey Hinton, David Rumelhart, and Ronald Williams published their seminal work on backpropagation [67], which established practical methods for training neural networks using gradient descent. The paper introduced several key ideas that would become standard practice, including the momentum method for accelerating training.

The basic intuition behind momentum comes from physics: rather than having the gradient directly determine the update (as in standard SGD), momentum methods maintain a "velocity" term that is influenced by gradients. The resulting update equations are

$$v_{k+1} = \mu v_k + (1-\mu)\widetilde{\nabla} f(x_k) \tag{2.7}$$

$$x_{k+1} = x_k - \eta v_{k+1}, \tag{2.8}$$

where $\mu \in [0, 1)$ is the momentum coefficient. This modification has several advantages over vanilla SGD. First, it helps smooth out the high-frequency components of the gradient noise - if successive gradients point in different random directions, they tend to cancel out

in the velocity term. Second, and perhaps more importantly, momentum helps accelerate progress along low-curvature directions. In the presence of a long, narrow valley in the loss landscape, standard SGD must zigzag down the valley, while momentum allows the optimizer to build up speed in the correct direction.

Hinton and colleagues empirically demonstrated that momentum could dramatically speed up neural network training, particularly in the presence of "plateaus" - regions where the gradient is small but non-zero. While the theoretical understanding of momentum's benefits would only come later, this stood as one of the first attempts to accelerate SGD on a computer.

2.2.2 Nemirovski-Yudin [1983]

Even as faster SGD algorithms were discovered, proofs came out showing that stochastic gradient descent had its limits, no matter how sophisticated the algorithm used. In the late 1970s[6], Nemirovski and Yudin [52] introduced the random oracle model of SGD, which they used to prove upper bounds on the achievable convergence rate.

In this model, an optimization algorithm faces off against an adversarial oracle. The oracle is responsible for returning stochastic gradients while staying within certain constraints: the gradients it provides must be unbiased and have bounded variance. That is, if we are seeking to minimize an objective $f(x)$, the oracle's stochastic gradients $g(x)$ must satisfy

$$\mathbb{E}[g(x)] = \nabla f(x), \quad \mathbb{E}[\|g(x) - \nabla f(x)\|^2] \leq \sigma^2. \tag{2.9}$$

As an example of this style of proof, we show that, no matter the stepsize schedule, SGD can not converge at a rate faster than k^{-1}. To do this, we provide a "difficult" objective function with a corresponding random oracle.

Set the objective to optimize as the one-dimensional function $f(x) = \frac{1}{2}(x^2 + 1)$. For ξ chosen randomly from $\{-1, 1\}$, let the random oracle provide the stochastic gradient

[6]Originally published in Russian, their work was only translated in 1983.

$g(x) = x - \xi$. Note that both the bias and variance constraints are satisfied, as

$$\mathbb{E}[g(x)] = x = \nabla\left(\frac{1}{2}(x^2+1)\right) = \nabla f(x) \tag{2.10}$$

$$\mathbb{E}[\|g(x) - \nabla f(x)\|^2] = \mathbb{E}[g(x)^2] - \nabla f(x)^2 = (x^2+1) - x^2 = 1. \tag{2.11}$$

For a fixed stepsize schedule η_k, our SGD iterates are given by

$$x_{k+1} = x_k - \eta_k g(x_k) = x_k - \eta_k(x_k - \xi) = (1-\eta_k)x_k + \eta_k\xi \tag{2.12}$$

Taking squares and expectations:

$$\mathbb{E}[x_{k+1}^2] = (1-\eta_k)^2\,\mathbb{E}\left[x_k^2\right] + \eta_k^2\mathbb{E}\left[\xi_k^2\right] + 2\,(1-\eta_k)\,\eta_k\mathbb{E}\left[x_k\xi_k\right] = (1-\eta_k)^2\,\mathbb{E}\left[x_k^2\right] + \eta_k^2. \tag{2.13}$$

It is then straightforward to inductively show that, for some $c > 0$,

$$\mathbb{E}[x_k^2] \geq \frac{c}{k}. \tag{2.14}$$

To finish, note that

$$\mathbb{E}[f(x_k) - f(x^*)] = \mathbb{E}\left[\frac{1}{2}\left(x_k^2+1\right) - \frac{1}{2}\right] = \frac{1}{2}\,\mathbb{E}\left[x_k^2\right] = \frac{c}{2k}. \tag{2.15}$$

This means that to find a function value within ϵ of the optimal, we must have $\epsilon \geq \frac{c}{2k}$, or $k \geq \frac{c}{2\epsilon}$, giving us a clear limit of the speed at which this flavor of SGD can converge in the worst case.

This is, of course, a simplified proof. Firstly, we have assumed that the stepsizes η_k are deterministic; secondly, the bound we have obtained is not as tight as possible (though, in fact, this rate is optimal for strongly convex functions, of which f is a representative.) Using slightly more sophisticated choices of objective function and oracle, we can show that on convex functions, the worst case learning rate is on the order of $k^{-\frac{1}{2}}$ (a particularly

elegant proof can be found in Agarwal et. al [1]). The general approach, however, remains unchanged: we try to construct a maximally uninformative oracle while remaining within the given constraints.

2.2.3 Polyak-Juditsky [1992]

Along with their upper bounds on the speed of SGD, Nemirovski and Yudin [52] described an algorithm that would achieve the optimal $n^{\frac{1}{2}}$ rate, given only stochastic gradients. For pedagogical purposes, we describe a later, cleaner formulation of this optimal algorithm, due to Polyak and Juditsky [62].

As before, we wish to minimize a convex function f given stochastic gradients $g(x)$ satisfying the unbiasedness and bounded variance conditions:

$$\mathbb{E}[g(x)] = \nabla f(x), \quad \mathbb{E}[\|g(x) - \nabla f(x)\|^2] \leq \sigma^2. \tag{2.16}$$

For the sake of this proof, we assume we are working within a convex, closed, and bounded set $\mathcal{X}$ such that for any $x, x' \in \mathcal{X}$, $\|x - x'\| \leq D$. Let x^* be a minimizer of f in $\mathcal{X}$. We also assume the true gradients are bounded on $\mathcal{X}$, i.e., $\|\nabla f(x)\| \leq G$ for all $x \in \mathcal{X}$.

The algorithm we present achieves this minimization using iterate averaging. We run standard SGD iterates x_k defined as

$$x_{k+1} = \Pi_{\mathcal{X}}(x_k - \eta_k g(x_k)), \tag{2.17}$$

where $\Pi_{\mathcal{X}}$ is the projection onto the set $\mathcal{X}$. For simplicity in the analysis below, we often omit the projection, assuming the steps stay within $\mathcal{X}$ or noting that projection does not increase distance to points within the convex set $\mathcal{X}$ (specifically $\|\Pi_{\mathcal{X}}(y) - x^*\| \leq \|y - x^*\|$ for $x^* \in \mathcal{X}$). We define the averaged iterate as

$$\overline{x}_n = \frac{\sum_{k=1}^n \eta_k x_k}{\sum_{k=1}^n \eta_k}. \tag{2.18}$$

We will show that with an appropriate choice of stepsize η_k,

$$\mathbb{E}[f(\overline{x}_n) - f(x^*)] \leq O\left(\frac{1}{\sqrt{n}}\right). \tag{2.19}$$

The proof relies on analyzing the squared distance to the optimum and applying convexity properties.

First, let's bound the expected squared norm of the stochastic gradient conditioned on x_k:

$$\begin{aligned}
\mathbb{E}[\|g(x_k)\|^2 \mid x_k] &= \mathbb{E}[\|g(x_k) - \nabla f(x_k) + \nabla f(x_k)\|^2 \mid x_k] \\
&= \mathbb{E}[\|g(x_k) - \nabla f(x_k)\|^2 \mid x_k] + \|\nabla f(x_k)\|^2 \\
&\quad + 2\mathbb{E}[(g(x_k) - \nabla f(x_k))^\top \nabla f(x_k) \mid x_k] \\
&= \mathbb{E}[\|g(x_k) - \nabla f(x_k)\|^2 \mid x_k] + \|\nabla f(x_k)\|^2 && (2.20) \\
&\leq \sigma^2 + \|\nabla f(x_k)\|^2 && (2.21) \\
&\leq \sigma^2 + G^2, && (2.22)
\end{aligned}$$

where (2.20) follows because $\mathbb{E}[g(x_k) - \nabla f(x_k) \mid x_k] = \mathbb{E}[g(x_k) \mid x_k] - \nabla f(x_k) = \nabla f(x_k) - \nabla f(x_k) = 0$, making the cross-term zero. Step (2.21) uses the bounded variance assumption from (2.16), and (2.22) uses the assumed bound G on the true gradient norm.

Now consider the SGD update step (ignoring projection for simplicity here, as it only helps):

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - \eta_k g(x_k) - x^*\|^2 \\
&= \|x_k - x^*\|^2 - 2\eta_k g(x_k)^\top (x_k - x^*) + \eta_k^2 \|g(x_k)\|^2. && (2.23)
\end{aligned}$$

Taking the expectation conditioned on the history $\mathcal{F}_k$ (which determines x_k):

$$\begin{aligned}\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] &= \|x_k - x^*\|^2 - 2\eta_k \mathbb{E}[g(x_k) \mid \mathcal{F}_k]^\top (x_k - x^*) + \eta_k^2 \mathbb{E}[\|g(x_k)\|^2 \mid \mathcal{F}_k] \\ &= \|x_k - x^*\|^2 - 2\eta_k \nabla f(x_k)^\top (x_k - x^*) + \eta_k^2 \mathbb{E}[\|g(x_k)\|^2 \mid \mathcal{F}_k]. \end{aligned} \tag{2.24}$$

By convexity of f, we know $f(x^*) \geq f(x_k) + \nabla f(x_k)^\top (x^* - x_k)$, which implies $\nabla f(x_k)^\top (x_k - x^*) \geq f(x_k) - f(x^*)$. Using this and the bound (2.22) in (2.24):

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \leq \|x_k - x^*\|^2 - 2\eta_k (f(x_k) - f(x^*)) + \eta_k^2 (G^2 + \sigma^2). \tag{2.25}$$

Taking the total expectation (using the law of iterated expectations $\mathbb{E}[\cdot] = \mathbb{E}[\mathbb{E}[\cdot \mid \mathcal{F}_k]]$):

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq \mathbb{E}[\|x_k - x^*\|^2] - 2\eta_k \mathbb{E}[f(x_k) - f(x^*)] + \eta_k^2 (G^2 + \sigma^2). \tag{2.26}$$

Rearranging gives:

$$2\eta_k \mathbb{E}[f(x_k) - f(x^*)] \leq \mathbb{E}[\|x_k - x^*\|^2] - \mathbb{E}[\|x_{k+1} - x^*\|^2] + \eta_k^2 (G^2 + \sigma^2). \tag{2.27}$$

Summing this inequality (2.27) from $k = 1$ to n:

$$\begin{aligned}\sum_{k=1}^{n} 2\eta_k \mathbb{E}[f(x_k) - f(x^*)] &\leq \sum_{k=1}^{n} \left(\mathbb{E}[\|x_k - x^*\|^2] - \mathbb{E}[\|x_{k+1} - x^*\|^2]\right) + \sum_{k=1}^{n} \eta_k^2 (G^2 + \sigma^2) \\ &= \left(\mathbb{E}[\|x_1 - x^*\|^2] - \mathbb{E}[\|x_{n+1} - x^*\|^2]\right) + (G^2 + \sigma^2) \sum_{k=1}^{n} \eta_k^2 \\ &\leq \mathbb{E}[\|x_1 - x^*\|^2] + (G^2 + \sigma^2) \sum_{k=1}^{n} \eta_k^2, \end{aligned} \tag{2.28}$$

where the sum telescopes, and we used $\mathbb{E}[\|x_{n+1} - x^*\|^2] \geq 0$. Since $x_1, x^* \in \mathcal{X}$, we have $\|x_1 - x^*\|^2 \leq D^2$. Thus,

$$\sum_{k=1}^{n} 2\eta_k \mathbb{E}\left[f(x_k) - f(x^*)\right] \leq D^2 + (G^2 + \sigma^2) \sum_{k=1}^{n} \eta_k^2. \tag{2.29}$$

Now, consider the averaged iterate $\bar{x}_n$. By the convexity of f and Jensen's inequality for expectations:

$$\mathbb{E}[f(\bar{x}_n)] = \mathbb{E}\left[f\left(\frac{\sum_{k=1}^n \eta_k x_k}{\sum_{k=1}^n \eta_k}\right)\right] \leq \mathbb{E}\left[\frac{\sum_{k=1}^n \eta_k f(x_k)}{\sum_{k=1}^n \eta_k}\right] = \frac{\sum_{k=1}^n \eta_k \mathbb{E}[f(x_k)]}{\sum_{k=1}^n \eta_k}. \tag{2.30}$$

Subtracting $f(x^*)$ from both sides (note $f(x^*)$ is a constant):

$$\mathbb{E}[f(\bar{x}_n)] - f(x^*) \leq \frac{\sum_{k=1}^n \eta_k (\mathbb{E}[f(x_k)] - f(x^*))}{\sum_{k=1}^n \eta_k} = \frac{\sum_{k=1}^n \eta_k \mathbb{E}[f(x_k) - f(x^*)]}{\sum_{k=1}^n \eta_k}. \tag{2.31}$$

Combining (2.29) and (2.31):

$$\begin{aligned}\mathbb{E}[f(\bar{x}_n) - f(x^*)] &\leq \frac{1}{\sum_{k=1}^n \eta_k}\left(\frac{1}{2}\sum_{k=1}^n 2\eta_k \mathbb{E}\left[f(x_k) - f(x^*)\right]\right) \\ &\leq \frac{D^2 + (G^2+\sigma^2)\sum_{k=1}^n \eta_k^2}{2\sum_{k=1}^n \eta_k}.\end{aligned} \tag{2.32}$$

To minimize this bound, we choose a constant stepsize $\eta_k = \eta$ for all $k = 1, \ldots, n$. The bound (2.32) becomes:

$$\mathbb{E}[f(\bar{x}_n) - f(x^*)] \leq \frac{D^2 + (G^2+\sigma^2)n\eta^2}{2n\eta} = \frac{D^2}{2n\eta} + \frac{(G^2+\sigma^2)\eta}{2}. \tag{2.33}$$

This expression is minimized by balancing the two terms. The optimal constant step size for a fixed n is found by taking the derivative with respect to η and setting it to zero, or by setting the terms equal:

$$\frac{D^2}{2n\eta} = \frac{(G^2+\sigma^2)\eta}{2} \implies \eta^2 = \frac{D^2}{n(G^2+\sigma^2)} \implies \eta_{opt} = \frac{D}{\sqrt{n(G^2+\sigma^2)}}.$$

Setting $\eta_k = \eta_{opt}$ for all $k = 1, \ldots, n$, we substitute this back into (2.33):

$$\begin{aligned}
\mathbb{E}[f(\bar{x}_n) - f(x^*)] &\leq \frac{D^2}{2n} \frac{\sqrt{n(G^2+\sigma^2)}}{D} + \frac{(G^2+\sigma^2)}{2} \frac{D}{\sqrt{n(G^2+\sigma^2)}} \\
&= \frac{D\sqrt{G^2+\sigma^2}}{2\sqrt{n}} + \frac{D\sqrt{G^2+\sigma^2}}{2\sqrt{n}} \\
&= \frac{D\sqrt{G^2+\sigma^2}}{\sqrt{n}}. \qquad (2.34)
\end{aligned}$$

This demonstrates that Polyak-Juditsky averaging with an appropriately chosen constant stepsize achieves the optimal $O(1/\sqrt{n})$ convergence rate for convex stochastic optimization. (Note: A decreasing step size like $\eta_k \propto 1/\sqrt{k}$ can also achieve this rate asymptotically without needing to know n in advance).

Together, these results fully categorize the asymptotic learning rate for the minimax optimal stochastic gradient algorithm under convexity assumptions. The key words here, however, are “minimax" and “asymptotic"; for practical purposes, we will eventually wish to compare constants across algorithms, as well as finding average-case behaviour for a given distribution of target functions.

2.3 Deep Learning

In the early 2010s, deep learning underwent a dramatic transformation from academic curiosity to industrial workhorse. Three key developments enabled this shift. First, the introduction of ReLU activations [49] and careful initialization schemes largely solved the vanishing gradient problem that had previously made deep networks untrainable. Second, the emergence of CUDA [53] and general GPU computing provided the raw computational power needed to train large models. Most importantly, empirical results started showimg clear returns to scale - larger, deeper networks, trained on more data with more compute, reliably produced better results.

This new regime placed unprecedented demands on optimization algorithms. Networks grew from thousands to billions of parameters, training runs stretched from hours to months, and compute budgets ballooned into the millions of dollars. Small improvements in optimization efficiency could translate into massive cost savings. The field's focus shifted accordingly - while theoretical understanding remained important, the ability to reliably train ever-larger models became paramount.

This pressure led to the development of a new generation of optimizers, starting with AdaGrad [23] in 2011. These adaptive methods attempted to automatically tune learning rates across different layers and time scales, eliminating much of the manual scheduling that had previously been required. While their worst-case guarantees were often no better than vanilla SGD, their practical performance on the large, overparameterized models of modern deep learning proved consistently superior.

2.3.1 AdaGrad [2011]

The key insight behind AdaGrad, introduced by Duchi et al [23], was that each parameter in a neural network might require its own learning rate - parameters that receive larger or more noisy gradients should take smaller steps, while parameters that are updated rarely should take larger ones.

AdaGrad implements this by maintaining a running sum of squared gradients for each parameter:

$$g_{k,i} = \widetilde{\nabla_{x_i} f(x_k)} \tag{2.35}$$

$$v_{k,i} = v_{k-1,i} + g_{k,i}^2 \tag{2.36}$$

$$x_{k+1,i} = x_{k,i} - \frac{\eta}{\sqrt{v_{k,i} + b^2}} g_{k,i} \tag{2.37}$$

where i indexes the parameters, η is a global learning rate, and b^2 is a small constant added for numerical stability. This coordinate-wise scaling provides automatic regularization - frequently updated parameters get smaller effective learning rates, inversely proportional

to their standard deviations - and partially obviates the need for manual learning rate tuning.

A simpler variant, AdaGrad-Norm [47], uses a single adaptive scalar for the entire parameter vector:

$$v_{k+1} = v_k + \|\widetilde{\nabla} f(x_{k+1})\|^2 \tag{2.38}$$

$$x_{k+1} = x_k - \frac{\eta}{\sqrt{v_k + b^2}} \widetilde{\nabla} f(x_k). \tag{2.39}$$

While less flexible than the full version, AdaGrad-Norm retains many of the key theoretical properties while being significantly easier to analyze.

2.3.2 RMSProp [2012]

While AdaGrad's parameter-specific learning rates proved effective in many settings, its accumulation of squared gradients meant that learning rates would eventually become vanishingly small. Geoffrey Hinton proposed RMSProp [32] as a simple fix: replace the sum with an exponential moving average:

$$g_k = \widetilde{\nabla} f(x_k) \tag{2.40}$$

$$v_k = \beta v_{k-1} + (1 - \beta) g_k \odot g_k \tag{2.41}$$

$$x_{k+1} = x_k - \frac{\eta}{\sqrt{v_k + \epsilon}} \odot g_k \tag{2.42}$$

where β is typically set to 0.9, ϵ is a small constant for numerical stability, and $\odot$ denotes element-wise multiplication. This modification allows the algorithm to "forget" old gradients, maintaining responsiveness throughout training while still adapting to the local geometry of the loss surface.

The theoretical properties of RMSProp proved somewhat more difficult to analyze than those of AdaGrad. While empirical results showed clear benefits, especially in training deep neural networks, formal convergence guarantees remained elusive for several years

[20]. The key challenge lay in analyzing the interaction between the momentum-like exponential averaging and the adaptive learning rates.

2.3.3 Adam [2014]

Adam [40], introduced by Kingma and Ba in 2014, represents perhaps the most successful synthesis of ideas from momentum and adaptive methods. It maintains exponential moving averages of both the gradients (first moment) and their squares (second moment):

$$g_k = \widetilde{\nabla} f(x_k) \tag{2.43}$$

$$m_k = \beta_1 m_{k-1} + (1 - \beta_1) g_k \tag{2.44}$$

$$v_k = \beta_2 v_{k-1} + (1 - \beta_2) g_k \odot g_k \tag{2.45}$$

$$\hat{m}_k = \frac{m_k}{1 - \beta_1^k} \tag{2.46}$$

$$\hat{v}_k = \frac{v_k}{1 - \beta_2^k} \tag{2.47}$$

$$x_{k+1} = x_k - \eta \frac{\hat{m}_k}{\sqrt{\hat{v}_k} + \epsilon} \tag{2.48}$$

where β_1 and β_2 are decay rates (typically set to 0.9 and 0.999 respectively). The bias correction terms $(1 - \beta_i^k)$ ensure unbiased estimates of the moments early in training.

Adam's combination of momentum and adaptive learning rates proved particularly effective for training large neural networks. The momentum term helps navigate ravines in the loss landscape, while the adaptive rates allow different parameters to learn at different speeds. This robustness to hyperparameter choices made Adam especially valuable in practice - while it might not always achieve the best possible performance, it reliably achieves good performance across a wide range of architectures and problems.

Despite its practical success, Adam's theoretical properties remain somewhat mysterious. While convergence can be proven under certain conditions, examples exist where Adam fails to converge to the optimal solution. These counterexamples led to various modifications (AdaMax [39], AdamW [46], etc.) attempting to maintain Adam's practical

benefits while providing stronger theoretical guarantees. Nevertheless, vanilla Adam remains the default choice for many deep learning applications, particularly in natural language processing where models like GPT and BERT [22] have demonstrated its effectiveness at massive scale.

2.3.4 Cutting Through the Noise [2021-]

Even as hundreds of new techniques pushed the limits of deep learning, some fundamental questions about SGD remained mysterious. One among these was an exact description of average-case performance. Convergence rates were often known, but more comprehensive results remained out of reach.

In 2020, a group of researchers, led by Courtney and Elliot Paquette, set out to reduce the problem to its bare essentials [55]. They considered the simplest possible optimization problem – linear regression with random data – along with the simplest possible optimization algorithm – fixed-stepsize SGD. Drawing on a range of techniques from random matrix theory, high dimensional probability, complex analysis, and Itô calculus, they were able to fully predict the loss dynamics of SGD, without needing to know the trajectory of the individual parameters.

Behind this prediction were two key equations. The first, capturing the inherent randomness in finite-dimensional SGD, was a stochastic differential equation (SDE) driven by Brownian motion. The second, a deterministic Volterra equation, was perhaps more surprising.

Among the simplifying assumptions made in the paper, one had been conspicuously lacking: the dimensionality of the problem was allowed to grow arbitrarily large, rather than being fixed at some constant value. This, as it turned out, was a prescient choice. In the high dimensional limit, with appropriate scalings, SGD loss was proved to concentrate around a deterministic limit, described by the aforementioned by the Volterra equation. Now, characteristics of this Volterra equation could be directly translated into predictions

about SGD: in order to discover the optimal stepsize, one needed only determine how the stepsize affected the Volterra equation.

This approach proved powerful and easily extensible. Over the next several years, the same framework was adapted to problems with non-linear structure [15], minibatching [41], momentum methods [41], and more, discovering optimal hyperparameter choices, categorizing phase transitions, and ranking algorithms by average-case high-dimensional performance.

We now move into the 2010s by extending this framework to modern adaptive methods. We present a general framework for analyzing adaptive stepsizes, and categorize the behaviour of a selection of these algorithms, including AdaGrad-Norm.

Chapter 3

Concentration Inequality Background

A large part of what makes our analyses possible is the existence of relvant concentration inequalities. This chapter presents several fundamental technical results that facilitate our technical analysis by allowing us to conveniently bound random variables, be they matrix or scalar. More details can be found in [74].

Proposition 3.0.1 (Subgaussian properties)**.** *Let X be a random variable. Then the following properties are equivalent; the parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.*

(i) There exists $K_1 > 0$ such that

$$\mathbb{P}\{|X| \geq t\} \leq 2\exp(-t^2/K_1^2) \quad \textit{for all } t \geq 0.$$

(ii) There exists $K_2 > 0$ such that

$$\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p} \leq K_2\sqrt{p} \quad \textit{for all } p \geq 1.$$

(iii) There exists $K_3 > 0$ such that

$$\mathbb{E}\exp(X^2/K_3^2) \leq 2.$$

Definition 3.0.1 (Subgaussian random variables). *A random variable X that satisfies one of the equivalent properties i-iii in Proposition 3.0.1 is called a* subgaussian random variable. *The* subgaussian norm *of X, denoted $\|X\|_{\psi_2}$, is defined to be the smallest K_3 in property iii. In other words, we define*

$$\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \leq 2\}. \tag{2.13}$$

Throughout most of our later proofs, we will use the fact that a Gaussian random variable is, a fortiori, subgaussian; we do not deal with subgaussian random variables in their full generality. The subgaussian formulation, however, is useful in that it allows us to state several concentration results cleanly.

We now present two concentration inequalities, which allow us to bound deviations from our predicted deterministic limit for SGD loss.

Theorem 3.0.1 (Hanson-Wright Inequality). *Let $X = (X_1, \ldots, X_n)$ be a random vector with independent components X_i that satisfy $\mathbb{E}[X_i] = 0$ and $\|X_i\|_{\psi_2} \leq K$. Let A be an $n \times n$ matrix. Then for all $t > 0$:*

$$P(|X^T AX - \mathbb{E}[X^T AX]| > t) \leq 2\exp\left(-c\min\left\{\frac{t^2}{K^4\|A\|_F^2}, \frac{t}{K^2\|A\|_{op}}\right\}\right)$$

where $c > 0$ is an absolute constant, $\|A\|_F$ is the Frobenius norm, and $\|A\|_{op}$ is the operator norm.

The Azuma-Hoeffding inequality provides concentration bounds for martingales with bounded differences. This makes it especially valuable for analyzing iterative algorithms where each step depends on previous iterations.

Theorem 3.0.2 (Azuma-Hoeffding). *Let $\{X_k\}_{k=0}^n$ be a martingale sequence with $|X_k - X_{k-1}| \leq c_k$ almost surely. Then for all $t > 0$:*

$$P(|X_n - X_0| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sum_{k=1}^n c_k^2}\right).$$

Chapter 4

Dynamics of Stochastic Adaptive Learning Algorithms

4.1 Introduction

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

$$\min_{X\in\mathbb{R}^d} \Big\{ \mathcal{R}(X) \stackrel{\text{def}}{=} \mathbb{E}_{a,\epsilon}[f(\langle a, X\rangle, \langle a, X^\star\rangle, \epsilon)] \quad \text{for} \quad (a,\epsilon) \sim \mathcal{D}, X^\star \in \mathbb{R}^d \Big\}. \tag{4.1}$$

We suppose $a \sim \mathcal{N}(0, K)$ where $K \in \mathbb{R}^{d\times d}$ is a covariance matrix. We train (4.1) 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

[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 better-or-worse-performing.

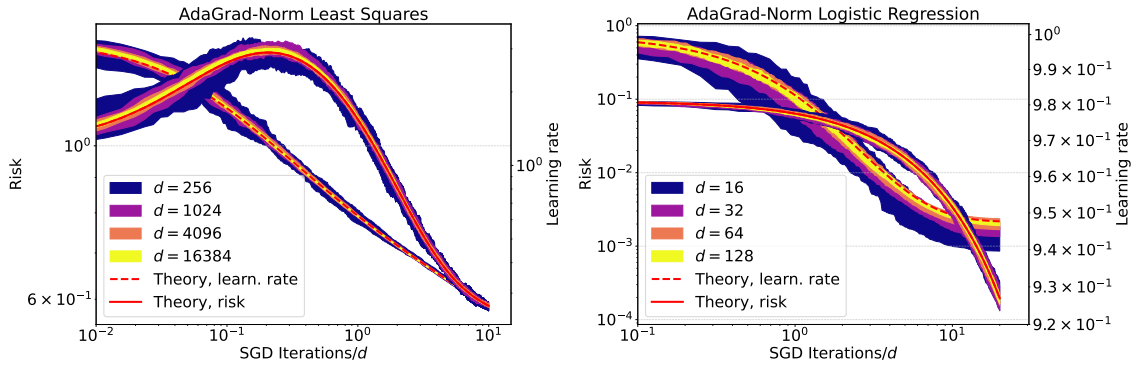

Figure 4.1: Concentration of learning rate and risk for AdaGrad-Norm on least squares with label noise $\omega = 1$ (left) and logistic regression with no noise (right). As dimension increases, both risk and learning rate concentrate around a deterministic limit (red) described by our ODE in Theorem 4.2.1. The initial risk increase (left) suggests the learning rate started too high, but AdaGrad-Norm adapts. Our ODEs predict this behavior. See Sec. 5.8 for simulation details.

algorithms on the least squares problem[2], the results of which appear in Table 4.1: exact line-search (idealistic) (Sec. 4.3) and AdaGrad-Norm (Sec. 4.4). We expect other losses and adaptive learning rates can be studied using this approach.

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 4.2), which we then analyze to show how the covariance spectrum influences the optimization. See Figure 4.1. 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 4.2.1).

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

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

problem) is $(\mathrm{Tr}(K)/d)^{-1}$ or the inverse of the average eigenvalue, as shown in the Polyak stepsize paper [61]. 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. 5.1).

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 [35] and D-Adaptation [21], 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 [78]. 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 4.1 and Sec. 4.4 for details.

Notation. Define $\mathbb{R}_+ = [0, \infty)$. We say a sequence $(E_d)_{d\geq 1}$ of events 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 $\mathbb{P}(E_d) \geq 1 - e^{-\omega(d)}$. We let $\mathbf{1}_A(x)$ be the indicator function of the set A: that is, 1 if $x \in A$ and 0 otherwise. For a matrix $A \in \mathbb{R}^{m\times d}$, we use $\|A\|_F$ 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,

Table 4.1: 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 4.4.4). For * (see Prop. 4.4.2), requires a good initialization on b, η.

Learning rate	**K assumption**	**Limiting γ_∞**	**Convergence rate**
AdaGrad-Norm(b,η) (see Sec. 4.4)	$\lambda_{\min} > C$	$\gamma_t \asymp \frac{\eta^2}{\frac{b}{\eta} + \frac{1}{4d}\mathrm{Tr}(K)\|X_0 - X^\star\|^2}$	$\log(\mathscr{R})^* \asymp -\lambda_{\min}\gamma_\infty t$
AdaGrad-Norm(b,η) Power law (see Sec. 4.4)	$\beta + \delta < 1$	$\gamma_t \asymp_{\delta,\beta} 1$	$\mathscr{R}(t) \asymp_{\delta,\beta} t^{\beta+\delta-2}$
	$\beta + \delta = 1$	$\gamma_t \asymp_{\delta,\beta} \frac{1}{\log(t+1)}$	$\mathscr{R}(t) \asymp_{\delta,\beta} \left(\frac{t}{\log(t+1)}\right)^{-1}$
	$1 < \beta + \delta < 2$	$\gamma_t \asymp_{\delta,\beta} t^{-1+\frac{1}{\beta+\delta}}$	$\mathscr{R}(t) \asymp_{\delta,\beta} t^{-\frac{2}{\beta+\delta}+1}$
Exact line search, idealized (see Sec. 4.3)	$\lambda_{\min} > C$	$\gamma_t \asymp \frac{\lambda_{\min}}{\mathrm{Tr}(K^2)/d}$	$\log(\mathscr{R}) \asymp -\lambda_{\min}\gamma_\infty t$
Polyak stepsize (see Sec. 4.3)	$\lambda_{\min} > C$	$\gamma_t = \frac{1}{\mathrm{Tr}(K)/d}$	$\log(\mathscr{R}) \asymp -\lambda_{\min}\gamma_\infty t$

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 $\asymp_\alpha$.

Related work. Some notable adaptive learning rates in the literature are AdaGrad-Norm [42,77,80], RMSprop [33], stochastic line search, stochastic Polyak stepsize [45], and

more recently DoG [35] and D-Adaptation [21]. In this work, we introduce a framework for analyzing these algorithms, and we strongly believe it can be used to analyze many more adaptive algorithms. We highlight below a nonexhaustive list of related work.

AdaGrad-Norm. AdaGrad, introduced by [23,47], updates the learning rate at each iteration using the stochastic gradient information. The single stepsize version [42,77,80], that depends on the norm of the gradient, (see Table 4.2 for the updates), has been shown to be robust to input parameters [44]. Several works have shown worst-case convergence guarantees [25,43,75,77]. 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 [81] (a similar idea appears in [80]), 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 [24, 59, 71] and stochastic Polyak stepsize (and their variants) [7,30,31,36,45,51,54,65]. Much of this research focuses on worst-case convergence guarantees for strongly convex and smooth functions (see e.g., [45]) and designing practical algorithms. In [72], 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 [9,10,68,69]. These results have been rigorously proven and extended to other models under the isotropic Gaussian assumption [2,3,6,18,19,27,76]. Extensions to multi-pass SGD with small mini-batches [57] as well as momentum [41] have also been studied. Other high-dimensional

limits leading to a different class of dynamics also exist [11, 13, 14, 26, 48]. Recently, significant contributions have been made in understanding the effects of a non-identity data covariance matrix on the training dynamics [5, 15, 16, 28, 29, 82]. The non-identity covariance modifies the optimization landscape and affects convergence properties, as discussed in [15]. This work extends the findings of [15] to stochastic adaptive algorithms, exploring the effect of non-identity covariance within these algorithms. Notably, Theorem 1.1 from [15] is restricted to deterministic learning rate schedules, limiting its applicability in many practical scenarios. In contrast, our Theorem 4.2.1 accommodates stochastic adaptive learning rates, aligning with widely used algorithms in practice.

4.1.1 Model Set-up

We suppose that a sequence of independent samples $\{(a_k, y_k)\}$ is provided, drawn from a data-generating distribution $\mathcal{D}$ over $\mathbb{R}^d \times \mathbb{R}$. Here, $a_k \in \mathbb{R}^d$ represents the input feature and $y_k \in \mathbb{R}$ is the corresponding target label. We assume a model where the target y_k is generated based on the feature a_k, a fixed (but unknown) ground truth signal $X^\star \in \mathbb{R}^d$, and some random label noise $\epsilon_k \in \mathbb{R}$. Specifically, the relationship is captured by the function Ψ in Equation (4.2), implying y_k is related to $\langle a_k, X^* \rangle$ and ϵ_k.

Since y_k is determined by a_k and ϵ_k (given the model structure and X^*), the fundamental random variables governing the distribution $\mathcal{D}$ are the feature a and the noise ϵ. Therefore, we place our distributional assumptions directly on these underlying variables:

Assumption 1 (Feature and Noise Distributions)**.** *The underlying random variables a (input feature) and ϵ (label noise) are assumed to be independent and normally distributed:*

- *The noise ϵ follows a normal distribution, $\epsilon \sim \mathcal{N}(0, \omega^2)$, where $\omega \in \mathbb{R}$.*
- *The feature a follows a multivariate normal distribution, $a \sim \mathcal{N}(0, K)$, where the covariance matrix $K \in \mathbb{R}^{d \times d}$ is positive semi-definite and bounded in operator norm independent of d, i.e., $\|K\|_{\text{op}} \leq C$ for some constant C.*

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

$$\mathcal{R}(X) \stackrel{\text{def}}{=} \mathbb{E}_{a,\epsilon}[\Psi(X; a, \epsilon)] \quad \text{for} \quad (a, \epsilon) \sim \mathcal{D}, \quad \text{and} \quad \Psi(X; a, \epsilon) = f(\langle a, X\rangle, \langle a, X^\star\rangle, \epsilon). \tag{4.2}$$

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 $B = B(W) = W^T K W$. Note that B is the covariance matrix of the Gaussian vector $(\langle a, X\rangle, \langle a, X^\star\rangle)$. Since $\mathcal{R}(X)$ and $I(B)$ (defined below) involve expectations over a, they depend on the distribution of this vector and thus are functions of B.

Assumption 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 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 (assuming sufficient regularity of f to exchange expectation and differentiation)*

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

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-behavedness 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'.

[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 4.2: 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. The deterministic equivalent γ_t utilizes quantities $\mathcal{B}(s)$ and $\mathscr{R}(s)$ derived from the ODE system described in Section 4.2.

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

Assumption 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* $I : \mathbb{R}^{2\times 2} \to \mathbb{R}$*. (Note that* $I(B)$ *is a scalar function related to the expected squared gradient magnitude, distinct from the* 2×2 *covariance matrix* B*.) We assume 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 [15, Appendix B]). We also note that Assumptions 3 and 4 are nearly satisfied for L-smooth objectives f (see Lemma 5.2.1), and a version of the main theorem holds under just this assumption (albeit with a weaker conclusion).

4.1.2 Algorithmic set-up

We apply *one-pass* or *streaming* SGD with an adaptive learning rate $\mathfrak{g}_k$ (SGD+AL) to minimize the risk $\mathcal{R}(X)$ defined in (4.2). Let $X_0 \in \mathbb{R}^d$ be an initial vector (random or non-random). Then SGD+AL iterates by selecting a *new*, independent 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

$$X_{k+1} = X_k - \frac{\mathfrak{g}_k}{d} \cdot \nabla_X \Psi(X_k; a_{k+1}, \epsilon_{k+1}) = X_k - \frac{\mathfrak{g}_k}{d} f'(\langle a_{k+1}, X_k \rangle; \langle a_{k+1}, X^\star \rangle, \epsilon_{k+1}) a_{k+1}, \quad (4.3)$$

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 the signal $X^\star$ (see Eq. (4.1)).

Assumption 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]

$$\begin{aligned} \mathfrak{g}_k &\stackrel{\text{def}}{=} \gamma(k, N_k(d \times \cdot), G_k(d \times \cdot), Q_k(d \times \cdot)), \text{ for } k \in \mathbb{N}, \text{ where for any } t \ge 0, \\ (N_k(t), G_k(t), Q_k(t)) &\stackrel{\text{def}}{=} \mathbf{1}_{\{t<k\}} \Big(W_{\lfloor t \rfloor}^T W_{\lfloor t \rfloor}, \tfrac{1}{d} \|\nabla_X \Psi(X_{\lfloor t \rfloor}; a_{\lfloor t \rfloor + 1}, \epsilon_{\lfloor t \rfloor + 1})\|^2, \mathcal{R}(X_{\lfloor t \rfloor}) \Big). \end{aligned} \quad (4.4)$$

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 $\mathcal{F}_k = \sigma(X^\star, X_0, \ldots, X_k)$:

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

With this, we impose the following learning rate condition.

[4] Note that cases where $\mathrm{Tr}(K^2)/d = o(d)$ can lead to dynamics that converge to full-batch gradient flow. While our theorem specifically addresses the scenario where the intrinsic dimension, $\mathrm{Dim}(K) \stackrel{\text{def}}{=} \mathrm{Tr}(K)\|K\|_{\mathrm{op}}/\|K\|_F^2$, 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 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 its last three arguments (the function spaces, equipped with the topology of uniform convergence on compact intervals). Moreover, for some constant* $C = C(\gamma) > 0$ *independent of* d *and* $\delta > 0$*,*

$$\mathbb{E}\left[|\gamma(k, f, G_k(d\times\cdot), q) - \gamma(k, f, \mathcal{G}_k(d\times\cdot), q)| \,|\, \mathcal{F}_k\right] \le Cd^{-\delta}(1 + \|f\|_\infty^\alpha + \|q\|_\infty^\alpha) \quad \textit{w.o.p.} \tag{4.5}$$

Finally, γ *satisfies the growth condition: there exists a constant* $\hat{C} = \hat{C}(\gamma) > 0$ *independent of* d *so that*

$$\gamma(k, f, g, q) \le \hat{C}(1 + \|f\|_\infty^\alpha + \|g\|_\infty^\alpha + \|q\|_\infty^\alpha). \tag{4.6}$$

The inequality (4.5) ensures that the learning rate concentrates around the mean behavior of the stochastic gradients. Many well-known adaptive stepsizes satisfy (4.4) and Assumption 6 including AdaGrad-Norm, DoG, D-Adaptation, and RMSProp (see Table 4.2, Sec. 5.1, and Sec. 5.3.3).

4.2 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.

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

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

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

$$\nabla h(\mathscr{B}(t)) \overset{\text{def}}{=} \begin{bmatrix} H_{1,t} & H_{2,t} \\ H_{2,t} & H_{3,t} \end{bmatrix}, \ \mathscr{N}(t) \overset{\text{def}}{=} \frac{1}{d} \sum_{i=1}^{d} \mathscr{V}_i(t), \ \mathscr{R}(t) \overset{\text{def}}{=} h(\mathscr{B}(t)), \quad \mathscr{I}(t) \overset{\text{def}}{=} I(\mathscr{B}(t)),$$

$$\text{and finally } \gamma_t \overset{\text{def}}{=} \gamma(t, \mathbf{1}_{\{\cdot \leq t\}} \mathscr{N}(\cdot), \tfrac{\text{Tr}(K)}{d} \mathbf{1}_{\{\cdot \leq t\}} \mathscr{I}(\cdot), \mathbf{1}_{\{\cdot \leq t\}} \mathscr{R}(\cdot)).$$

[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 [74].

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

$$\begin{aligned}
\mathrm{d}\mathscr{V}_{11,i}(t) &= -2\lambda_i\gamma_t\left(\mathscr{V}_{11,i}(t)H_{1,t} + H_{1,t}\mathscr{V}_{11,i}(t) + \mathscr{V}_{12,i}(t)H_{2,t} + H_{2,t}\mathscr{V}_{12,i}(t)\right) + \lambda_i\gamma_t^2\mathscr{I}(t), \\
\mathrm{d}\mathscr{V}_{12,i}(t) &= -2\lambda_i\gamma_t\left(H_{1,t}\mathscr{V}_{12,i}(t) + H_{2,t}\mathscr{V}_{22,i}(t)\right) \\
\mathrm{d}\mathscr{V}_{22,i}(t) &= 0 \quad (\text{since } X^\star \text{ is fixed})
\end{aligned} \tag{4.7}$$

with the initialization of $\mathscr{V}_i(0)$ given by $V_i(X_0) = d \cdot W_0^T\omega_i\omega_i^T W_0$. We finally state the deterministic dynamics for the risk and learning rate.

Theorem 4.2.1. *Under Assumptions 1, 2, 3, 4, 5, 6, then for any $\varepsilon \in (0, \frac{1}{2})$ and any $T > 0$*

$$\sup_{0\le t\le T}\left\|\begin{pmatrix}\mathcal{R}(X_{\lfloor td\rfloor})\\ \\ \mathfrak{g}_{\lfloor td\rfloor}\end{pmatrix} - \begin{pmatrix}\mathscr{R}(t)\\ \\ \gamma_t\end{pmatrix}\right\| < d^{-\varepsilon}, \quad \textit{w.o.p.} \tag{4.8}$$

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

In fact, we can derive deterministic dynamics for a large class of statistics which are linear combinations of $\mathscr{V}_i(t)$ and functions thereof (See Theorem 5.2.1, and Corollary 5.2.1).

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. 4.2.1. The equivalent deterministic dynamics are

$$\mathscr{D}^2(t) = \frac{1}{d}\sum_{i=1}^d \mathscr{D}_i^2(t) = \frac{1}{d}\sum_{i=1}^d (\mathscr{V}_{11,i}(t) - 2\mathscr{V}_{12,i}(t) + \mathscr{V}_{22,i}(t)), \tag{4.9}$$

where $\mathscr{D}_i^2(t)$ corresponds to $d \times (\langle X_{\lfloor td\rfloor} - X^\star, \omega_i\rangle)^2$.

Example: Least Squares. One canonical example of (4.2) 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

$$\min_{X\in\mathbb{R}^d}\left\{\mathcal{R}(X)=\tfrac{1}{2}\mathbb{E}_{a,\epsilon}[(\langle a,X-X^\star\rangle-\epsilon)^2]=\tfrac{1}{2}\omega^2+\tfrac{1}{2}(X-X^\star)^TK(X-X^\star)\right\}. \tag{4.10}$$

The pair of functions h (Assumption 3) and I (Assumption 4) can be evaluated simply:

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

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

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

where γ_s is the deterministic learning rate from the ODE system (4.7). This is a convolution Volterra equation with a convergence threshold of $\gamma_t<\frac{2d}{\mathrm{Tr}K}$ [16,56–58].

In the noiseless label case (i.e., $\omega=0$), 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 (4.7), we get the following deterministic equivalent ODE for the $\mathscr{D}_i^2$'s:

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}_i^2(t)=-2\gamma_t\lambda_i\mathscr{D}_i^2(t)+2\gamma_t^2\lambda_i\mathscr{R}(t). \tag{4.11}$$

We will perform a deep analysis of the dynamics of the learning rate on least squares (4.10), which will generalize to settings where the outer function f is strongly convex (see 5.4.1).

4.3 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

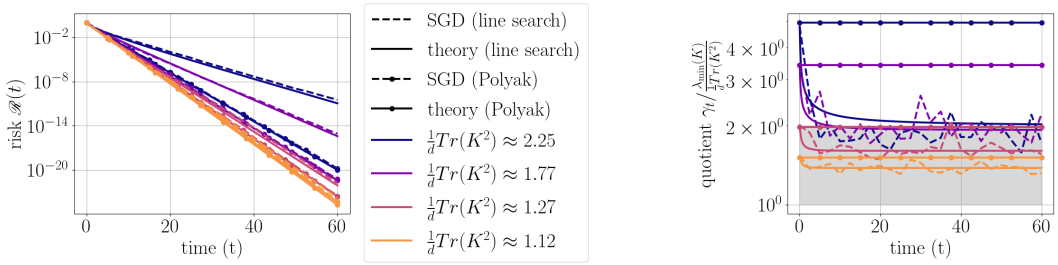

Figure 4.2: Comparison for Exact Line Search and Polyak Stepsize on a noiseless least squares problem. The left plot illustrates the convergence of the risk function, while the right plot depicts the convergence of the quotient $\gamma_t / \frac{\lambda_{\min}(K)}{\frac{1}{d}\operatorname{Tr}(K^2)}$ for Polyak stepsize and exact line search. Both plots highlight the implication of equation (4.13) 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 (4.13) is satisfied. See Appendix 5.8 for simulation details.

deterministic equivalents by Theorem 4.2.1. Thus, the question is: what choice of learning rate optimally 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 5.6.1 and 5.6.2 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 stability limit for the distance to optimality. We define $\bar{\gamma}_t^{\mathcal{P}}$ as the largest (continuous-time) learning rate such that the deterministic distance equivalent $\mathcal{D}(t)$ decreases, i.e., $\mathrm{d}\mathcal{D}(t) < 0$. Using the least squares ODE (4.11) (specifically,

summing it over i), this threshold is precisely

$$\bar{\gamma}_t^{\mathcal{D}} = \frac{(2\mathscr{R}(t) - \omega^2)}{\frac{\mathrm{Tr}(K)}{d}\mathscr{R}(t)} \quad \text{and} \quad \bar{\mathfrak{g}}_k^{\mathcal{D}} = \frac{(2\mathcal{R}(X_k) - \omega^2)}{\frac{\mathrm{Tr}(K)}{d}\mathcal{R}(X_k)}.^{7} \tag{4.12}$$

Without label noise ($\omega = 0$), equation (4.12) simplifies to $\bar{\gamma}_t^{\mathcal{D}} = \bar{\mathfrak{g}}_k^{\mathcal{D}} = \frac{2}{\mathrm{Tr}(K)/d}$, which corresponds to the exact stability threshold for convergence in the noiseless least squares case.

A greedy stepsize strategy aims to maximize the decrease in the distance to optimality at each iteration. This defines the *Polyak stepsize*, $\gamma_t^{\text{Polyak}} \in \arg\min_\gamma \mathrm{d}\mathcal{D}(t)$. For the specific case of least squares, this yields

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

The discrete-time version, $\mathfrak{g}_k^{\text{Polyak}}$, is known to yield the optimal fixed learning rate (up to absolute constant factors) for a noiseless target in a least squares problem [45, 61].[8]

Exact Line Search

In the context of risk, using (4.11) and noting that $\mathscr{R}(t) = \frac{1}{2d}\sum_{i=1}^d \lambda_i \mathscr{D}_i^2(t)$, we can find $\gamma_t^{\text{line}} \in \arg\min_\gamma \mathrm{d}\mathscr{R}(t)$; 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 4.2, (c.f. Appendix 5.6.1 for general losses). Because these come from ODEs, we can use ODE theory to give exact limiting values for the deterministic equivalent of $\mathfrak{g}_k^{\text{line}}$.

[7] Here, $\omega^2 = \mathbb{E}[\epsilon^2]$ is the variance of the label noise ϵ (see Assumption 1), and is distinct from the eigenvectors ω_i of the covariance matrix K mentioned elsewhere.

[8] The Polyak stepsize analyzed here, based on minimizing the ODE for $\mathcal{D}(t)$, coincides with the classic $\frac{\mathcal{R}(X_k)-\mathcal{R}(X^*)}{||\nabla\mathcal{R}(X_k)||^2}$ for least squares. We use the ODE-based definition derived from minimizing the distance decrease directly, which avoids an approximation step sometimes used in derivations [31].

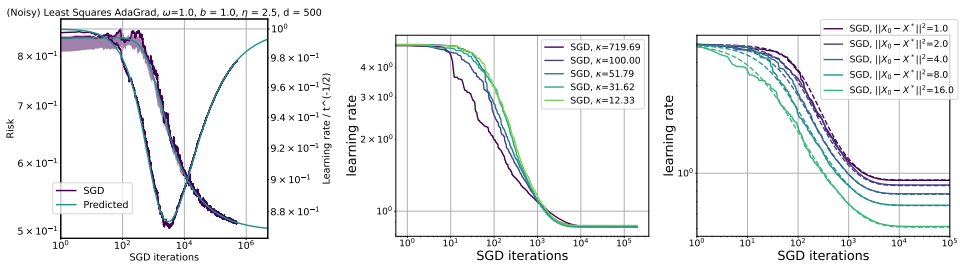

Figure 4.3: Quantities effecting AdaGrad-Norm learning rate. *(left):* Effect of noise ($\omega = 1.0$) on risk (left axis) and learning rate (right axis). Depicted is $\frac{\text{learning rate}}{\text{asymptotic l.r.}}$ so it approaches 1. *(Center, right)*: Noiseless least squares ($\omega = 0$). As predicted in Prop. 4.4.2, $\lim_{t\to\infty} \gamma_t$ 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 5.8 for simulation details.

Proposition 4.3.1. *[Limiting learning rate; line search on noiseless least squares] Consider the noiseless ($\omega = 0$) least squares problem* (4.10) *. Then the learning rate is always lower bounded by*

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

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. Then

$$\frac{\lambda_{\min}(K)}{\frac{1}{d}\mathrm{Tr}(K^2)} \leq \lim_{t\to\infty} \gamma_t^{\text{line}} \leq \frac{2\lambda_{\min}(K)}{\frac{1}{d}\mathrm{Tr}(K^2)}. \tag{4.13}$$

For a proof and explicit formula for $\lim_{t\to\infty} \gamma_t^{\text{line}}$, see Section 5.6.2. Hence, being greedy for the risk in a sufficiently anisotropic setting will badly underperform Polyak stepsize (see Fig. 4.2).

4.4 AdaGrad-Norm analysis

In this section, we analyze the behavior of AdaGrad-Norm learning rate in the least squares setting (see Sec. 5.4 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 4.3. 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 ($\omega^2 > 0$), 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 ($\omega = 0$) 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 4.4.1. *In the case of identity covariance ($K = I_d$) and no noise ($\omega = 0$), the risk solves the differential equation*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{R}(t) = \frac{\eta^2\mathscr{R}(t)}{b^2+2\int_0^t \mathscr{R}(s)\,\mathrm{d}s} - \frac{2\eta\mathscr{R}(t)}{\sqrt{b^2+2\int_0^t \mathscr{R}(s)\,\mathrm{d}s}}, \tag{4.14}$$

with $\mathscr{R}(0) = \frac{1}{2}\|X_0 - X^\star\|^2$.

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 4.4.2. *Consider the noiseless case ($\omega = 0$). Suppose $\frac{1}{d}\mathrm{Tr}(K) \le b/\eta$, and that $\int_0^\infty \mathscr{R}(s)\,\mathrm{d}s < \infty$. Let γ_s be the AdaGrad-Norm learning rate for least squares (Table 4.2). Then $\gamma_t \to \gamma_\infty > 0$ and $\gamma_\infty \asymp \frac{\eta^2}{\frac{b}{\eta}+\frac{\eta}{2d}\mathrm{Tr}(K)\mathscr{D}^2(0)}$.*

An analog of Proposition 4.4.2 for the strongly convex setting appears in Sec. 5.4 (see Prop. 5.4.1). We now consider two cases in which, as $d \to \infty$, there are eigenvalues of K arbitrarily close to 0.

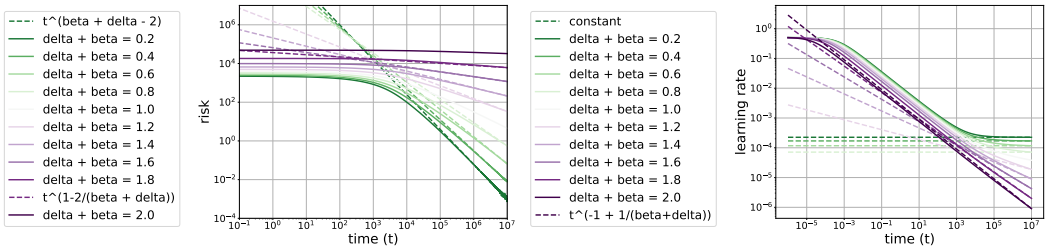

Figure 4.4: Power law covariance in AdaGrad Norm on a least squares problem. Ran exact predictions (ODE) for the risk and learning rate (solid lines). Dashed lines give the predictions from Prop. 4.4.4 which *match experimental results exactly*. **Phase transition as $\delta + \beta$ varies.** When $\delta + \beta < 1$ (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. The same phase transition occurs in the sublinear rate of the risk decay *(left)* (see Prop. 4.4.4).

Proposition 4.4.3. *Consider the noiseless case ($\omega = 0$). 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 4.4.4. *Consider the noiseless case ($\omega = 0$). 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[9] $\beta < 1$, and suppose that $\mathscr{D}_i^2(0) \sim \lambda_i^{-\delta}$ for $\delta \geq 0$. Then:*

- *For $1 > \beta + \delta$, there exists $\tilde{\gamma} > 0$ such that $\gamma_t \geq \tilde{\gamma}$ for all $t \geq 0$, and $\mathscr{R}(t) \asymp_{\delta,\beta} t^{\beta+\delta-2}$ for all $t \geq 1$.*
- *For $1 < \beta + \delta < 2$, $\gamma_t \asymp_{\delta,\beta} t^{-1+\frac{1}{\beta+\delta}}$, 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 $\delta = 0$

[9]Our result can be compared to existing findings for SGD under power-law distributions in [8,70,73]. 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 $\beta > 1$.

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. 4.4). 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. 5.4.

Chapter 5

Proofs, Examples, and Simulations

5.1 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

$$\mathfrak{g}_k = \frac{\eta}{\sqrt{b^2 + \frac{1}{d^2}\sum_{j=0}^{k} \|\nabla_X \Psi(X_j; a_{j+1}, \epsilon_{j+1})\|^2}}, \tag{5.1}$$

with the deterministic equivalent (see Section 4.2 and also 5.3.3) for this learning rate being

$$\gamma_t = \frac{\eta}{\sqrt{b^2 + \frac{\mathrm{Tr}(K)}{d}\int_0^t I(\mathscr{B}(s))\ \mathrm{d}s}}. \tag{5.2}$$

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

$$\gamma_t = \frac{\eta}{\sqrt{b^2 + \frac{2\mathrm{Tr}(K)}{d}\int_0^t \mathscr{R}(s)\ \mathrm{d}s}}. \tag{5.3}$$

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 4.2) is

$$\gamma_t = \frac{\eta}{\sqrt{b^2 e^{-\alpha t} + \frac{\mathrm{Tr}(K)}{d} \int_0^t e^{-\alpha(t-s)} I(\mathscr{B}(s))\, \mathrm{d}s}}. \tag{5.4}$$

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

$$\gamma_t = \frac{\eta}{\sqrt{b^2 e^{-\alpha t} + \frac{2\mathrm{Tr}(K)}{d} \int_0^t e^{-\alpha(t-s)} \mathscr{R}(s)\, \mathrm{d}s}}. \tag{5.5}$$

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

5.2 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 $\partial_x f$.

5.2.1 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:

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

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

Lemma 5.2.1. *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

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

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,

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

[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 $\epsilon \to 0$.

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

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

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 $M \mapsto \int_0^\infty e^{-tQ} M e^{-tQ} \, \mathrm{d}t$ 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

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

Hence

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

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

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

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

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

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 3 and 4 provided that $\|B^{-1}\|_{\text{op}}$ is bounded. Therefore, our concentration result Theorem 5.2.1 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 5.2.1, holds with t replaced with $t \wedge \hbar_M(B) \wedge \hbar_M(\mathscr{B})$. The corollaries of Theorem 5.2.1 also follow under this added stopping time.

In the next section, we prove this concentration theorem, Theorem 5.2.1.

5.2.2 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 3 and 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 $\mathscr{S}$:

$$\mathrm{d}\mathscr{S}(t,\cdot) = \mathcal{F}(z, \mathscr{S}(t,\cdot))\ \mathrm{d}t \tag{5.6}$$

$$\begin{aligned}\text{where } \mathcal{F}(z,\mathscr{S}(t,\cdot)) \stackrel{\text{def}}{=} -2\gamma_t\Bigg(&\left(\frac{-1}{2\pi i}\oint_\Omega \mathscr{S}(t,z)\ \mathrm{d}z\right)H(\mathscr{B}(t)) \\ &+ H^T(\mathscr{B}(t))\left(\frac{-1}{2\pi i}\oint_\Omega \mathscr{S}(t,z)\ \mathrm{d}z\right)\Bigg) \\ &+ \frac{\gamma_t^2}{d}\left[\begin{array}{c|c} \mathrm{Tr}(KR(z;K))I(\mathscr{B}(t)) & 0 \\ \hline 0 & 0 \end{array}\right] \\ &- \gamma_t(\mathscr{S}(t,z)(2zH(\mathscr{B}(t))) + (2zH^T(\mathscr{B}(t)))\mathscr{S}(t,z)).\end{aligned} \tag{5.7}$$

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

$$\gamma_t \text{ is defined in (4.7), and the initialization is } \quad \mathscr{S}(0,z) = W_0^\top R(z;K)W_0. \tag{5.8}$$

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 3 and Assumption 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 (4.7)) and proved in [15, Lemma 4.1].

Lemma 5.2.2 (Equivalence to coupled ODEs.)**.** *The unique solution of* (5.7) *with initial condition* (5.8) *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 (5.6) (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 $\Omega \stackrel{\text{def}}{=} \left\{z : |z| = \max\left\{1, 2\|K\|_{\text{op}}\right\}\right\}$. 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

$$\|A\|_\Omega = \max_{z \in \Omega} \|A(z)\|. \tag{5.9}$$

Definition 5.2.1 ((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 (5.6) 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 5.2.1 (Concentration of SGD and deterministic function $\mathscr{S}(t, z)$)**.** *Suppose the risk function $\mathcal{R}(X)$* (4.2) *satisfies Assumptions 2, 3, and 4. Suppose the learning rate satisfies Assumption 6, and the initialization X_0 and hidden parameters $X^\star$ satisfy Assumption 5. Moreover the data $a \sim \mathcal{N}(0, K)$ and label noise ϵ satisfy Assumption 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

$$\sup_{0\leq t\leq T\wedge\hat{\tau}_M(S(W,\cdot))\wedge\hat{\tau}_M(\mathscr{S})} \|S(W_{\lfloor td\rfloor},\cdot)-\mathscr{S}(t,\cdot)\|_\Omega \leq d^{-\varepsilon}, \tag{5.10}$$

where the deterministic function $\mathscr{S}(t,z)$ solves the integro-differential equation (5.6)*.*

Proof. By Proposition 5.3.1, 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, [15, Prop. 4.1], to conclude that there exists a $\varepsilon>0$ such that

$$\sup_{0\leq t\leq T\wedge\hat{\tau}_M} \|\mathscr{S}(t,z)-S(W_{td},z)\|_\Omega \leq d^{-\varepsilon}, \quad w.o.p, \tag{5.11}$$

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

Corollary 5.2.1. *Suppose the assumptions of Theorem 5.2.1 hold. Let f be an α-pseudo-Lipschitz function with $\alpha\leq 1$ and let q be a polynomial. Set*

$$\varphi(X)\overset{def}{=} f(W^Tq(K)W),\quad \phi(t)\overset{def}{=} f\left(\frac{-1}{2\pi i}\oint_\Omega q(z)\mathscr{S}(t,z)\,\mathrm{d}z\right),\quad \textit{where } \mathscr{S}(t,z) \textit{ solves } (5.6).$$

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

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

Proof. This is basically equivalent to [15, Corollary 4.2]. The only difference is that [15, Corollary 4.2] requires the boundedness of $\mathscr{N}$; however, since our function f is α-pseudo-Lipschitz with $\alpha\leq 1$, this boundedness follows from [15, Proposition 1.2], and the rest of the proof is identical to the one in [15]. □

Remark 5.2.1. *The learning rate $\mathfrak{g}_k$, technically, is not a function of $W^Tq(K)W$. However, Assumption 6 ensures that the learning rate concentrates around a function $W^Tq(K)W$. Therefore, Corollary 5.2.1 applies to the learning rate.*

5.3 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, $X_k \stackrel{\text{def}}{=} X_{\lfloor k \rfloor}$. Using the continuous parameter t, the iterates are related by $X_{td} = X_{\lfloor td \rfloor}$.

The paper [15] 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 5.3.1 (Net argument, [15], 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*

$$\sup_{0 \le t \le (\hat{\tau}_M \wedge T)} \|S(t, \cdot) - S(0, \cdot) - \int_0^t \mathcal{F}(\cdot, S(s, \cdot))\, \mathrm{d}s\|_{\Omega_\xi} \le \varepsilon \tag{5.12}$$

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 5.3.1 that $S(t, z)$ does indeed satisfy inequality (5.12).) We also cite the following lemma, which relates two stopping times used throughout this paper.

Lemma 5.3.2 (Stopping time, [15], 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 5.3.1. *Fix $M > 0$ and define the stopping time on $\|W_{td}\|$, $\vartheta = \vartheta_M$, by*

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

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 $c = C^{-1}$, $\hat{\tau}_M \leq \vartheta_{cM}$.

5.3.1 SGD-AL is an approximated solution

Proposition 5.3.1 (SGD-AL is an approximate solution)**.** *Fix a $T, M > 0$ and $0 < \varepsilon < \delta/8$, where δ is defined in Assumption 6. Then $S(W_{td}, z)$ is a $(d^{-\varepsilon}, M, T)$-approximate solution w.o.p., that is,*

$$\sup_{0\leq t\leq(T\wedge\tau_M)} \|S(W_{td}, z) - S(W_0, z) - \int_0^t \mathcal{F}(z, S(W_{sd}, z))\ \mathrm{d}s\|_\Omega \leq d^{-\varepsilon}. \tag{5.13}$$

Again, the proof is very similar to [15, 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 5.3.2, fills in the details of bounding these lower-order terms, so that the proof can proceed as in [15].

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

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

The normalization here (dividing by d) is chosen so that the I terms are all $O(1)$; this is formally shown in Lemma 5.3.5.

SGD-AL under the statistic

We follow the approach in [15, 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

$$\varphi(X_{k+1}) = \varphi(X_k) - \frac{\gamma(G_k)}{d}\nabla\varphi(X_k)^\top\Delta_k + \frac{\gamma(G_k)^2}{2d^2}\Delta_k^\top\nabla^2\varphi(X_k)\Delta_k. \tag{5.14}$$

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

$$\Delta\mathcal{M}_k^{\text{grad}}(\varphi) \stackrel{\text{def}}{=} \frac{1}{d}\left(-\gamma(G_k)I_3(k) + \mathbb{E}\left[\gamma(G_k)I_3(k)\,\middle|\,\mathcal{F}_k\right]\right), \tag{5.15}$$

$$\Delta\mathcal{M}_k^{\text{Hess}}(\varphi) \stackrel{\text{def}}{=} \frac{1}{2d}\left(\gamma(G_k)^2I_1(k) - \mathbb{E}\left[\gamma(G_k)^2I_1(k)\,\middle|\,\mathcal{F}_k\right]\right), \tag{5.16}$$

$$\mathbb{E}[\mathcal{E}_k^{\text{Hess}}(\varphi) \,|\, \mathcal{F}_k] \stackrel{\text{def}}{=} \frac{1}{2d}\left(\mathbb{E}\left[\gamma(G_k)^2I_1(k)\,\middle|\,\mathcal{F}_k\right] - \gamma(B_k)^2I_2(k)\right), \tag{5.17}$$

$$\mathbb{E}[\mathcal{E}_k^{\text{grad}}(\varphi) \,|\, \mathcal{F}_k] \stackrel{\text{def}}{=} \frac{1}{d}\left(-\mathbb{E}\left[\gamma(G_k)I_3(k)\,\middle|\,\mathcal{F}_k\right] + \gamma(B_k)\nabla\varphi(X_k)^\top\nabla\mathcal{R}(X_k)\right). \tag{5.18}$$

We can then write

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

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

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

giving us

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

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

$$\begin{aligned} \varphi(X_{td}) &= \varphi(X_0) - \int_0^t \gamma(B_{sd}) \nabla\varphi(X_{sd})^\top \nabla\mathcal{R}(X_{sd}) \, \mathrm{d}s \\ &\quad + \frac{1}{2d} \int_0^t \gamma(B_{sd})^2 \operatorname{Tr}(K \nabla^2 \varphi(X_{sd})) \mathbb{E}[f'(r_{sd})^2 \,|\, \mathcal{F}_{sd}] \, \mathrm{d}s \\ &\quad + \sum_{j=0}^{\lfloor td \rfloor - 1} \mathcal{E}_j^{\text{all}}(\varphi), \end{aligned} \tag{5.19}$$

$$\begin{aligned} \text{with} \quad \mathcal{E}_j^{\text{all}}(\varphi) &= \Delta\mathcal{M}_j^{\text{grad}}(\varphi) + \Delta\mathcal{M}_j^{\text{Hess}}(\varphi) \\ &\quad + \mathbb{E}[\mathcal{E}_j^{\text{Hess}}(\varphi) \,|\, \mathcal{F}_j] + \mathbb{E}[\mathcal{E}_j^{\text{grad}}(\varphi) \,|\, \mathcal{F}_j] \\ &\quad + \xi_{td}(\varphi). \end{aligned} \tag{5.20}$$

From here, we can proceed as in [15, 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 5.3.1. The appropriate stepsize, as a function of W_{td}, is

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

(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 [15, Proposition 5.2] that given the analogue of (5.19) for deterministic stepsize, $S(W_{td}, \cdot)$ satisfies

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

The only terms of (5.19) 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 (5.6) all we need is to bound the martingales and error terms contained in $\mathcal{E}_j^{\text{all}}$. Let $\Omega = \{z : |z| = \max\{1, 2\|K\|_{\text{op}}\}\}$, as previously. We thus have that for all $z \in \Omega$,

$$\sup_{0 \le t \le T \wedge \hat{\tau}_M} \left| S(W_{td}, z) - S(W_0, z) - \int_0^t \mathcal{F}(z, S(W_{sd}, z)) \, \mathrm{d}s \right| \le \sup_{0 \le t \le T \wedge \hat{\tau}_M} \|\mathcal{E}_{td}^{\text{all}}(S(\cdot, z))\|. \quad (5.21)$$

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 $\|K\|_{\text{op}}$. For all $z \in \Omega$, we note that $\hat{\tau}_M \le \vartheta_{cM}$ (see Lemma 5.3.2). Consequently, we evaluate the error with the stopped process $W_{td}^\vartheta \stackrel{\text{def}}{=} W_{d(t \wedge \vartheta)}$ instead of using $\hat{\tau}_M$. By Proposition 5.3.2, the proof of which we have deferred to Section 5.3.2, we have, for any $\hat{\delta} > 0$

$$\sup_{z \in \Omega_\xi} \sup_{0 \le t \le T \wedge \vartheta_{cM}} \|\mathcal{E}_{dt}^{\text{all}}(S(\cdot, z))\| \le d^{-\delta/4 + \hat{\delta}} \quad \text{w.o.p.} \quad (5.22)$$

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 5.3.1, finishes the proof after setting $\hat{\delta} = \delta/8$ and $\xi = \delta/8$. □

5.3.2 Error bounds

All the martingale and error terms (5.20) go to 0 as d grows. Formally,

Proposition 5.3.2. *Let the function f be defined as in Assumption 2. Let the statistic $S : [0, \infty) \times \mathbb{C} \to \mathbb{R}^{2\times 2}$ be defined as*

$$S(t, z) = W_{\lfloor td \rfloor}^{\top} R(z; K) W_{\lfloor td \rfloor}, \tag{5.23}$$

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 5.3.2.

Proof. This follows from combining Propositions 5.3.3, 5.3.4, 5.3.5, 5.3.6, and 5.3.7. □

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 5.3.3 (Hessian error term)**.** *Let f and S be defined as in Assumption 2 and* (5.23)*. 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 $k \leq (T \wedge \vartheta)d - 1$, set $\varphi(X) = S_{ij}(W, z)$ to be the ij-th entry of the matrix $S(W, z)$. Then

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

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 (4.6). 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 5.3.6). Since we know from Lemma 5.3.6 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 (4.5). 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, $\mathbb{E}\left[|I_1(k)|^2 \mid \mathcal{F}_k\right]^{\frac{1}{2}}$, is also bounded by a constant (see Lemma 5.3.5), and all together, we find that $|\mathcal{E}_1| \leq d^{\varepsilon-\delta/4}$ with overwhelming probability.

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

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

The first term is bounded by (4.6), since $\mathcal{G}_k$ and $\operatorname{Tr}(K)I(B_k)/d$ are bounded independent of d; the second term is bounded Cd^{-1} by Lemma 5.3.9, and the last term is bounded by a constant by Lemma 5.3.5.

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

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

By (4.6), 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 5.3.8 with $H = \nabla^2\varphi(X_k)$, and is thus bounded by Cd^{-1}. All together,

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

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

Proposition 5.3.4 (Gradient error term)**.** *Let f and S be defined as in Assumption 2 and* (5.23)*. Then, for any $z \in \Omega$, $\zeta > 0$ and $T > 0$, with overwhelming probability,*

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

Proof. We have

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

We then have

$$\begin{aligned} |\mathcal{E}_1| &\leq \mathbb{E}\left[|\gamma(G_k) - \gamma(\mathcal{G}_k)|^2 \,|\, \mathcal{F}_k\right]^{\frac{1}{2}} \cdot \mathbb{E}\left[|I_3(k)|^2 \,|\, \mathcal{F}_k\right]^{\frac{1}{2}} \\ &\leq \mathbb{E}\left[|\gamma(G_k) + \gamma(\mathcal{G}_k)|^3 \,|\, \mathcal{F}_k\right]^{\frac{1}{4}} \cdot \mathbb{E}\left[|\gamma(G_k) - \gamma(\mathcal{G}_k)| \,|\, \mathcal{F}_k\right]^{\frac{1}{4}} \cdot \mathbb{E}\left[|I_3(k)|^2 \,|\, \mathcal{F}_k\right]^{\frac{1}{2}}. \end{aligned}$$

Just as in the Hessian argument, (4.6) lets us bound $\mathbb{E}\left[|\gamma(G_k) + \gamma(\mathcal{G}_k)|^3 \,|\, \mathcal{F}_k\right]^{\frac{1}{4}}$ by d^{ε} w.o.p., (4.5) lets us bound $\mathbb{E}\left[|\gamma(G_k) - \gamma(\mathcal{G}_k)| \,|\, \mathcal{F}_k\right]^{\frac{1}{4}}$ by $d^{-\delta/4}$ w.o.p., and Lemma 5.3.5 lets us bound $\mathbb{E}\left[|I_3(k)|^2 \,|\, \mathcal{F}_k\right]^{\frac{1}{2}}$ by a constant, giving an overall bound of $|\mathcal{E}_1| \leq d^{-\delta/4+\varepsilon}$.

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

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

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

Proposition 5.3.5 (Gradient martingale)**.** *Let f and S be defined as in Assumption 2 and* (5.23)*. Then, for any $z \in \Omega$, $\zeta > 0$ and $T > 0$, with overwhelming probability,*

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

Proof. For notational convenience, set $\Delta\mathcal{M}_k = \Delta\mathcal{M}_{d(k/d \wedge \vartheta)}^{\text{grad}}$, 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 $F_k^\beta = \text{Proj}_\beta(F_k)$, that is, ensuring F_k stays in $[-\beta, \beta]$. Then $F_k^\beta - \mathbb{E}[F_k^\beta \mid \mathcal{F}_k]$ is in $[-2\beta, 2\beta]$, and so for the martingale $\mathcal{M}_k^\beta$ with increments $\Delta\mathcal{M}_k^\beta = F_k^\beta - \mathbb{E}[F_k^\beta \mid \mathcal{F}_k]$, Azuma's inequality tells us that

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

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

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

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

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

We bound this term in overwhelming probability. We have

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

The second term is superpolynomially small by Lemma 5.3.5; the first term is superpolynomially small by (4.6) and (5.3.6).

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

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 5.3.6 (Hessian martingale)**.** *Let f and S be defined as in Assumption 2 and* (5.23)*. Then, for any $z \in \Omega$, $\zeta > 0$ and $T > 0$, with overwhelming probability,*

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

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

$\mathcal{M}_k^\beta = F_k^\beta - \mathbb{E}[F_k^\beta \mid \mathcal{F}_k]$. As before, Azuma's inequality, with $\beta = d^{-1+\zeta/2}$, gives us

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

The error term is also quite similar:

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

We have

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

both of which are superpolynomially small by (4.6) and Lemma 5.3.5. For the expectation, we have

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

this product is superpolynomially small by (4.6), Lemma 5.3.6, and Lemma 5.3.5. Overall, we have, with overwhelming probability,

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

Taking the supremum, we obtain the desired result. □

Proposition 5.3.7 (Integral error term)**.** *Let f and S be defined as in Assumption 2 and* (5.23)*. Then, for $z \in \Omega$,*

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

Proof. We have, as above,

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

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 [74] for details.) When it exists, this norm is defined as

$$\|X\|_{\psi_2} \asymp \inf \left\{ V > 0 : \forall t > 0, \ \mathbb{P}(|X| > t) \leq 2e^{-t^2/V^2} \right\}. \tag{5.24}$$

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

Lemma 5.3.3 ([15], Lemma 5.3)**.** *There exist constants $c, C > 0$ such that*

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

Lemma 5.3.4 (Preliminary bounds)**.** *With f and Δ_k defined as above, for $\varepsilon > 0$ and $\lambda \geq 0$, we have*

$$f'(r_k) \leq d^\varepsilon \quad \textit{w.o.p. and} \quad \mathbb{E}[|f'(r_k)|^\lambda \,|\, \mathcal{F}_k] \leq C(\lambda), \tag{5.25}$$

$$\frac{\|\Delta_k\|^2}{d} \leq d^\varepsilon \quad \textit{w.o.p. and} \quad \mathbb{E}\left[\left(\frac{\|\Delta_k\|^2}{d} \right)^\lambda \,|\, \mathcal{F}_k \right] \leq C(\lambda). \tag{5.26}$$

Proof of (5.25) *in Lemma 5.3.4.* By [15, Lemma 3.4], if function f is α-pseudo-Lipschitz with Lipschitz constant $L(f)$ (as in (2)) and the noise ϵ is independent of a, then

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

Then

$$|f'(r_k)| \le C(\alpha)(L(f))(1 + |r_k| + |\epsilon|)^{\max\{1,\alpha\}}$$
$$\le C(\alpha)(L(f))(1 + |X_k^\top a_{k+1}| + |\epsilon|)^{\max\{1,\alpha\}}. \tag{5.27}$$

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 (5.27) is bounded w.o.p. by d^ε, and

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

for some constant $C(\lambda)$. □

Proof of (5.26) *in Lemma 5.3.4.* 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

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

Now, note that $\mathbb{E}[v_k^\top K v_k \,|\, \mathcal{F}_k] = \text{Tr}(K) \le d\|K\|_{\text{op}}$. 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:

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\|\Delta_k\|^2}{d}\right)^{\lambda} \mid \mathcal{F}_k\right] &\leq \mathbb{E}\left[\left(\frac{\|\sqrt{K}v_k\|^2}{d}\right)^{2\lambda} \mid \mathcal{F}_k\right]^{\frac{1}{2}} \cdot \mathbb{E}\left[f'(r_k)^{4\lambda} \mid \mathcal{F}_k\right]^{\frac{1}{2}} \\ &\leq \mathbb{E}\left[\left(\frac{\|K\|_{\mathrm{op}} \cdot \|v_k\|^2}{d}\right)^{2\lambda} \mid \mathcal{F}_k\right]^{\frac{1}{2}} \cdot \mathbb{E}\left[f'(r_k)^{4\lambda} \mid \mathcal{F}_k\right]^{\frac{1}{2}} \qquad (5.28)\end{aligned}$$

For the first term, we have

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

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 (5.28) is bounded by a constant by (5.25), as desired. □

Lemma 5.3.5 (Gradient and Hessian bounds)**.** *Setting*

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

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

$$|I_1(k)| \leq d^{\varepsilon} \quad \textit{w.o.p. and} \quad \mathbb{E}\left[|I_1(k)|^{\lambda} \mid \mathcal{F}_k\right] \leq C(\lambda), \qquad (5.29)$$

$$|I_2(k)| \leq C, \qquad (5.30)$$

$$|I_3(k)| \leq d^{\varepsilon} \quad \textit{w.o.p. and} \quad \mathbb{E}\left[|I_3(k)|^{\lambda} \mid \mathcal{F}_k\right] \leq C(\lambda). \qquad (5.31)$$

Proof of (5.29) *in Lemma 5.3.5.* Using the fact that $\|\nabla^2\varphi(X_k)\|_{\text{op}} \le \|S(W_k,\cdot)\|_\Omega$,

$$\begin{aligned}\frac{|\Delta_k^\top \nabla^2\varphi(X_k)\Delta_k|}{d} &\le \frac{\|S(W_k,\cdot)\|_\Omega\|\Delta_k\|^2}{d}\\ &\le \frac{C\|W_k\|^2\|\Delta_k\|^2}{d}. \qquad \text{(Lemma 5.3.3)}\end{aligned}$$

Now, $\|W_k\|$ is bounded by the stopping time. From Lemma 5.3.4, $\frac{\|\Delta_k\|^2}{d}$ is bounded by d^ε w.o.p., and every moment of this expression is bounded independent of d, as desired. □

Proof of (5.30) *in Lemma 5.3.5.* We have

$$\begin{aligned}\frac{|\operatorname{Tr}(\nabla^2\varphi(X_k)K)\mathbb{E}[f'(r_k)^2 \mid \mathcal{F}_k]|}{d} &\le \frac{d\|\nabla^2\varphi(X_k)K\|_{\text{op}}\cdot\mathbb{E}[f'(r_k)^2 \mid \mathcal{F}_k]}{d}\\ &\le \|\nabla^2\varphi(X_k)\|_{\text{op}}\cdot\|K\|_{\text{op}}\cdot\mathbb{E}[f'(r_k)^2 \mid \mathcal{F}_k]\\ &\le CM^2\,\mathbb{E}[f'(r_k)^2 \mid \mathcal{F}_k]. \qquad \text{(Lemma 5.3.3)}\end{aligned}$$

From Lemma 5.3.4, $\mathbb{E}[f'(r_k)^2 \mid \mathcal{F}_k]$ is bounded by a constant independent of d, as desired.

□

Proof of (5.31) *in Lemma 5.3.5.* We have

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

By Lemma 5.3.3, $\|\nabla\varphi(X_k)\| \le C\|W_k\| \le CM$ (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 (5.25), $f'(r_k)$ 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:

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

The first term is bounded by a constant independent of d, since subgaussian moments are bounded. The second term is bounded by Lemma 5.3.4, completing the proof. □

Lemma 5.3.6 (Infinity norm bounds)**.** *For G_k, N_k, Q_k as defined in 4.1.2, we have, for any $\varepsilon, \lambda > 0$, there exists $C > 0$ such that,*

$$\|G_k\|_\infty \leq d^\varepsilon \quad \textit{w.o.p. and} \quad \mathbb{E}[\|G_k\|_\infty^\lambda \mid \mathcal{F}_k] \leq d^\varepsilon \quad \textit{w.o.p.,} \tag{5.32}$$

$$\|N_k\|_\infty \leq C, \quad \|Q_k\|_\infty \leq C, \quad \|\mathcal{G}_k\|_\infty \leq C. \tag{5.33}$$

Proof. The first line, (5.32), follows from (5.26). For the first inequality, $\|G_k\|_\infty = \max_{0\leq j\leq k} \frac{\|\Delta_j\|^2}{d}$, 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,

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

as desired. The second line is more straightforward:

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

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 $\|Q_k\|_\infty$:

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

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 $\mathcal{G}_k$:

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

by Lemma 5.3.4. □

We now prove a concentration result that closely follows [15, Proposition 5.6].

Lemma 5.3.7 ([15], 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*

$$v \,|\, U^\top v \sim v - U(U^\top v) + UU^\top v, \tag{5.34}$$

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 5.3.8. *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*

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

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 [15], 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

$$\mathbb{E}[q(a_{k+1})f'(r_k)^2 \,|\, \hat{\mathcal{F}}_k] = \mathbb{E}[q(a_{k+1} - \mathbb{E}[a_{k+1} \,|\, \hat{\mathcal{F}}_k]) \,|\, \hat{\mathcal{F}}_k]\, \mathbb{E}_\epsilon[f'(r_k)^2] + q(\mathbb{E}[a_{k+1} \,|\, \hat{\mathcal{F}}_k])\, \mathbb{E}_\epsilon[f'(r_k)^2]. \tag{5.35}$$

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 5.3.7. By Assumption 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 $\sqrt{K} W_k \stackrel{\text{def}}{=} Q_k R_k$ 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 $\Pi_k \stackrel{\text{def}}{=} Q_k Q_k^T$. In distribution,

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

As R_k is invertible, by Lemma 5.3.7,

$$a_{k+1} \,|\, a_{k+1}^\top W_k \stackrel{\text{d}}{=} \sqrt{K} v_k \,|\, Q_k^T v_k \stackrel{\text{d}}{=} \sqrt{K}\big(v_k - \Pi_k v_k\big) + \sqrt{K} \Pi_k v_k. \tag{5.36}$$

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

$$\mathbb{E}[a_{k+1} \,|\, \hat{\mathcal{F}}_k] = \sqrt{K} \Pi_k v_k, \quad \text{where } v_k \sim N(0, I_d). \tag{5.37}$$

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

$$\begin{aligned} &(\mathbb{E}[(a_{k+1} - \mathbb{E}[a_{k+1} \,|\, \hat{\mathcal{F}}_k])(a_{k+1} - \mathbb{E}[a_{k+1} \,|\, \hat{\mathcal{F}}_k])^\top \,|\hat{\mathcal{F}}_k]) \\ &= \sqrt{K}(I_d - \Pi_k)\sqrt{K}, \quad \text{where } \Pi_k = Q_k Q_k^T. \end{aligned} \tag{5.38}$$

Next, using that $\mathbb{E}[H_k \mid \mathcal{F}_k] = H_k$, we expand (5.35) to get the leading order behavior

$$\begin{aligned}\mathbb{E}[q(a_{k+1})f'(r_k)^2 \mid \hat{\mathcal{F}}_k] &= \operatorname{Tr}(HK)\,\mathbb{E}_\epsilon[f'(r_k)^2] \\ &\quad - \operatorname{Tr}(H\sqrt{K}\Pi_k\sqrt{K})\,\mathbb{E}_\epsilon[f'(r_k)^2] \\ &\quad + q(\sqrt{K}\Pi_k v_k)\,\mathbb{E}_\epsilon[f'(r_k)^2].\end{aligned} \tag{5.39}$$

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

$$\mathbb{E}[q(a_{k+1})f'(r_k)^2 \mid \mathcal{F}_k] - \operatorname{Tr}(HK)\,\mathbb{E}[f'(r_k)^2 \mid \mathcal{F}_k] = \mathbb{E}[\mathcal{E}_k \mid \mathcal{F}_k], \tag{5.40}$$

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

$$\mathcal{E}_k = - \operatorname{Tr}(H\sqrt{K}\Pi_k\sqrt{K})\,\mathbb{E}_\epsilon[f'(r_k)^2] \tag{5.41}$$

$$+ q(\sqrt{K}\Pi_k v_k)\,\mathbb{E}_\epsilon[f'(r_k)^2]. \tag{5.42}$$

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

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

By (5.25), 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 (5.25) 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 5.3.9. *There is a constant C such that*

$$|\gamma(\mathcal{G}_k) - \gamma(B_k)| \le C d^{-1}.$$

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

$$\begin{aligned}
&|\gamma(\mathcal{G}_k) - \gamma(B_k)| \\
&\quad \le \|\mathcal{G}_k - \operatorname{Tr}(K) I(B_k)/d\|_\infty \times (1 + 2\|N_k\|_\infty^\alpha + \|\mathcal{G}_k\|_\infty^\alpha + \|\operatorname{Tr}(K) I(B_k)/d\|_\infty^\alpha + 2\|Q_k\|_\infty^\alpha) \\
&\quad \le C\|\mathcal{G}_k - \operatorname{Tr}(K) I(B_k)/d\|_\infty && \text{(Lemma 5.3.6)} \\
&\quad \le C d^{-1} \max_{0 \le j \le k} \left\| \mathbb{E}[a_{j+1}^\top a_{j+1} f'(r_j)^2 \,|\, \mathcal{F}_j] - \operatorname{Tr}(K)\, \mathbb{E}[f'(r_j)^2 \,|\, \mathcal{F}_j] \right\| \\
&\quad \le C d^{-1}, && \text{(Lemma 5.3.8)}
\end{aligned}$$

as desired. □

5.3.3 Specific learning rates

In this section, we confirm that AdaGrad-Norm satisfies Assumption 6. In the notation of Assumption 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 $g = G_k$:

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

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

Proposition 5.3.8 (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 5.3.2. *This is a stronger condition than the α-pseudo Lipschitz one in Assumption 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

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

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 5.3.9 (Boundedness)**.** *Suppose γ is AdaGrad-Norm. Then* (4.6)*, as part of Assumption 6, holds.*

Proof. This is immediate:

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

□

It remains to show that AdaGrad-Norm satisfies (4.5) in Assumption 6.

Proposition 5.3.10 (Concentration)**.** *Suppose γ is AdaGrad-Norm, with G_k and $\mathcal{G}_k$ being defined as before. Then Equation* (4.5)*, as part of Assumption 6, holds:*

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

Proof. Looking to remove the square roots, we have

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

For AdaGrad-Norm, we have

$$\begin{aligned}\left|\gamma(G_k)^2 - \gamma(\mathcal{G}_k)^2\right| &= \eta^2 \left| \frac{1}{b^2 + \frac{1}{d^2}\sum_{j=0}^{k} \|\Delta_j\|^2} - \frac{1}{b^2 + \frac{1}{d^2}\sum_{j=0}^{k} \mathbb{E}[\|\Delta_j\|^2 \,|\, \mathcal{F}_j]} \right| \\ &\leq \frac{\eta^2}{d^2 b^4} \cdot \left| \sum_{j=0}^{k} (\mathbb{E}[\|\Delta_j\|^2 \,|\, \mathcal{F}_j] - \|\Delta_j\|^2) \right|. \end{aligned} \tag{5.43}$$

We now bound the sum above. Set $F_i = \|\Delta_i\|^2/d$, $F_i^\beta = \text{Proj}_\beta(F_i)$, $\Delta\mathcal{M}_i = F_i - \mathbb{E}[F_i \,|\, \mathcal{F}_i]$, 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

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

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

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

We have

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

which superpolynomially small by (5.26). The expectation is similar:

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

The first expectation is bounded by a constant independent of d by (5.26), 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 (5.43), 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. □

5.4 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 4.4.

5.4.1 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 7 (Risk and loss minimizer)**.** *Suppose that*

$$X^\star \in \textit{arg min}_X \left\{ \mathcal{R}(X) = \mathbb{E}_{a,\epsilon}[f(\langle X, a\rangle, \langle X^\star, a\rangle), \epsilon] \right\}$$

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

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

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 5.4.1 ($\hat{L}$-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 $x, \hat{x}, y, z \in \mathbb{R}$*

$$f(\hat{x}, y, z) \leq f(x, y, z) + \langle f'(x, y, z), \hat{x} - x\rangle + \tfrac{\hat{L}(f)}{2}|\hat{x} - x|^2. \tag{5.44}$$

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 (5.44) holds with constant $\hat{L}$. Suppose $x^\star \in \arg\min_x\{f(x,y,z)\}$ exists. An immediate consequence of (5.44) is that

$$\frac{1}{2\hat{L}(f)}|f'(x,y,z)|^2 \leq f(x,y,z) - f(x^\star,y,z) \leq \frac{\hat{L}(f)}{2}|x-x^\star|^2. \tag{5.45}$$

Definition 5.4.2 (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 5.4.1. *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 7 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 4.2 (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. (5.73) for the distance to optimality, with $(x, x^\star) \sim \mathcal{N}(0, \mathscr{B})$,

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

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

$$\mathbb{E}_{a,\epsilon}\big[\langle x - x^\star, f'(x,x^\star)\rangle\big] \geq \hat{\mu}(f)\mathbb{E}_{a,\epsilon}[(x-x^\star)^2] = 2\hat{\mu}(f)\mathscr{R}(t), \tag{5.46}$$

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

$$\frac{1}{2(\hat{L}(f))^2}\mathbb{E}_{a,\epsilon}\big[(f'(x, x^\star)^2\big] \leq \frac{1}{2}\mathbb{E}_{a,\epsilon}\big[(x - x^\star)^2\big] = \mathscr{R}(t). \tag{5.47}$$

Thus by (5.46) and (5.47), 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 5.4.1, with $i(v) = I(\mathscr{B}(\Gamma^{-1}(v))) = \mathbb{E}_{a,\epsilon}\big[(f'(x, x^\star)^2\big]$ instead of the risk

$$\begin{aligned}\gamma_\infty &= \frac{\eta^2}{\frac{b}{\eta} + \frac{1}{2d}\operatorname{Tr}(K)\int_0^\infty i(v)\,\mathrm{d}v} \geq \frac{\eta^2}{\frac{b}{\eta} + \frac{1}{d}\operatorname{Tr}(K)(\hat{L}(f))^2\int_0^\infty r(v)\,\mathrm{d}v} \\ &\geq \frac{\eta^2}{\frac{b}{\eta} + \frac{1}{d}\operatorname{Tr}(K)\frac{(\hat{L}(f))^2\mathscr{D}^2(0)}{\left(2\hat{\mu}(f) - (\hat{L}(f))^2\frac{1}{d}\operatorname{Tr}(K)\gamma_0\right)}} = \frac{\eta^2}{\frac{b}{\eta} + \frac{\frac{1}{d}\operatorname{Tr}(K)(\hat{L}(f))^2}{2\hat{\mu}(f)(1-\zeta)}\mathscr{D}^2(0)}}.\end{aligned} \tag{5.48}$$

where the first inequality is by Eq. 5.47, 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 5.4.1. *Given γ_t as in Table 4.2 (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 4.2 (AdaGrad-Norm), changing variables to $u = \Gamma(t)$ and rearranging the terms:

$$b^2 + \tfrac{\operatorname{Tr}(K)}{d} \int_0^u \frac{i(v)}{g(v)} \, \mathrm{d}v = \frac{\eta^2}{g(u)^2}, \tag{5.49}$$

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. □

5.4.2 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,

$$\mathscr{R}(t) = F(\Gamma(t)) + \int_0^t \gamma_s^2 \mathcal{K}(\Gamma(t) - \Gamma(s)) \mathscr{R}(s) \, \mathrm{d}s \tag{5.50}$$

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

$$F(x) \stackrel{\text{def}}{=} \frac{1}{2d} \sum_{i=1}^{d} \lambda_i \mathscr{D}_i^2(0) e^{-2\lambda_i x}, \tag{5.51}$$

$$\mathcal{K}(x) \stackrel{\text{def}}{=} \frac{1}{d} \sum_{i=1}^{d} \lambda_i^2 e^{-2\lambda_i x}. \tag{5.52}$$

In the following we consider three cases, a strongly convex risk in which the spectrum of the eigenvalues is bounded from below (section 5.4.2). 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 5.4.2). Finally, power law spectrum supported on $[0, 1]$ with $d \to \infty$ (section 5.4.2).

Proofs for case of fixed d

Proof of Proposition 4.4.2. Define the composite functions $r(u) = \mathscr{R}(\Gamma^{-1}(u))$, and $g(u) = \gamma(\Gamma^{-1}(u))$. Integrating the formula for the risk:

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

Taking $t \to \infty$, we get

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

Using $\|\mathcal{K}\|_1 = \int_0^\infty \mathcal{K}(x)\,\mathrm{d}x < \gamma_0^{-1}$, and noting that by Eq. (5.52), and Eq. (5.51), we have that $\|F\|_1 = \frac{1}{4}\mathscr{D}^2(0)$, and $\|\mathcal{K}\|_1 = \frac{1}{2d}\operatorname{Tr}(K)$,

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

On the hand following Lemma 5.4.3, $\frac{1}{4}\mathscr{D}^2(0)(1+\frac{\gamma_0}{2d}\operatorname{Tr}(K)) \leq \|r\|_1$. Therefore, $\|r\|_1 \asymp \frac{1}{4}\mathscr{D}^2(0)$.

Next, rewriting the γ_t equation in Table 4.2 (AdaGrad-Norm for least squares) in terms of $g(u)$ (Lemma 5.4.1), we obtain

$$g(u) = \frac{\eta^2}{\frac{b}{\eta} + \frac{1}{d}\operatorname{Tr}(K)\int_0^u r(x)\,\mathrm{d}x} \tag{5.53}$$

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

$$\gamma_\infty = g(\infty) = \frac{\eta^2}{\frac{b}{\eta} + \frac{1}{d}\operatorname{Tr}(K)\|r\|_1} \asymp \frac{\eta^2}{\frac{b}{\eta} + \frac{1}{4d}\operatorname{Tr}(K)\mathscr{D}^2(0)}. \tag{5.54}$$

This then completes the proof. □

Remark 5.4.1. *We note that, on the Least square problem $\hat{L}(f) = \hat{\mu}(f) = 1$, therefore, the bound in Proposition 5.4.1 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 4.4.1. Using the equation for the distance to optimality (Eq. 4.11), we can derive an equation for the integral of the risk (with no target noise) which we denote by $g(t) = \int_0^t \mathscr{R}(s) \, \mathrm{d}s$:

$$g''(t) = -\gamma_t \sum_i \lambda_i^2 \mathscr{D}_i^2(t) + \gamma_t^2 \frac{\operatorname{Tr}(K^2)}{d} g'(t). \tag{5.55}$$

For $K = I_d$, this equation simplifies,

$$g''(t) = -2\gamma_t g'(t) + \gamma_t^2 \frac{\operatorname{Tr}(K^2)}{d} g'(t). \tag{5.56}$$

Plugging in the equation for the AdaGrad-Norm learning rate (Table 4.2) 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 4.4.2 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 4.4.3).

Proof of Proposition 4.4.3. 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 $C_1 > 0$,

$$\gamma_t = \frac{\eta}{\sqrt{b^2 + \frac{2}{d}\operatorname{Tr}(K)\int_0^t \mathscr{R}(s)ds}} \geq \frac{\eta}{\sqrt{b^2 + \frac{2}{d}\operatorname{Tr}(K)R_0 t}} \geq C_1 t^{-1/2}. \tag{5.57}$$

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. (4.11), and the definition of the risk $\mathscr{R}(t) = \frac{1}{2d}\sum_{i=1}^d \lambda_i \mathscr{D}_i^2(t)$,

$$\begin{aligned}\mathscr{R}(t) &= \mathscr{R}(0) - \frac{1}{d}\sum_{i=1}^d \lambda_i^2 \int_0^t \gamma_s \mathscr{D}_i(s)\mathrm{d}s + \frac{1}{d}\int_0^t \gamma_s^2 \operatorname{Tr}(K^2)\cdot \mathscr{R}(s)\mathrm{d}s \\ &\leq \mathscr{R}(0) - \int_0^t \gamma_s(2\alpha_s - \gamma_s \frac{1}{d}\operatorname{Tr}(K^2))\cdot \mathscr{R}(s)\mathrm{d}s + 2\int_0^t \gamma_s \mathscr{R}_2(s)\,\mathrm{d}s\end{aligned} \tag{5.58}$$

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 5.4.2 $\forall i$, $\mathscr{D}_i^2(t) \leq \max\left(\gamma_{t_1}\mathscr{R}(t_1), \mathscr{D}_i^2(0)\right) = c_0$ 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,

$$2\int_0^t \gamma_s \mathscr{R}_2(s)\,\mathrm{d}s \leq \frac{1}{d^2}\operatorname{Tr}(K^2)c_0 \int_0^t \gamma_s N_s\,\mathrm{d}s. \tag{5.59}$$

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,

$$\mathscr{R}(t) \leq \mathscr{R}(0) - \frac{1}{d}\operatorname{Tr}(K^2)\int_0^t \gamma_s^2 \cdot \mathscr{R}(s)\mathrm{d}s. \tag{5.60}$$

Using Gronwall's inequality,

$$\mathscr{R}(t) \leq \mathscr{R}(0)e^{-\frac{1}{d}\operatorname{Tr}(K^2)\int_0^t \gamma_s^2\,\mathrm{d}s} \leq \mathscr{R}(0)e^{-\frac{1}{d}\operatorname{Tr}(K^2)C_1^2 t} \tag{5.61}$$

where in the last transition we used the lower bound on the learning rate derived in Eq. (5.57). 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 $t > 0$,

$$\gamma_t \geq \frac{\eta}{\sqrt{b^2 + \frac{\frac{1}{d}\operatorname{Tr}(K)\mathscr{R}(0)}{\frac{1}{d}\operatorname{Tr}(K^2)C_1^2}}}. \tag{5.62}$$

□

Lemma 5.4.2. *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 (4.11), $\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}_i^2(0) \geq 0$. 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 (4.11), $\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}_i^2(0) < 0$. 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. (5.50)) to derive non-asymptotic bounds on the risk, which will be useful in Section 5.4.2 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 5.4.3. *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* $t > T$*,*

$$\int_0^t \mathcal{K}(s)\mathcal{K}(t-s) \, \mathrm{d}s \leq 2(1+\epsilon)\|\mathcal{K}\|_1 \mathcal{K}(t) \quad \textit{and} \quad 2\|\mathcal{K}\|_1(1+\epsilon)\gamma_0 < 1$$

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,

$$\begin{aligned}\mathcal{G}^2(f) = \mathcal{G}(\mathcal{G}(f))(t) &= \int_0^t \mathcal{K}(t-s)\mathcal{G}(f)(s)g(s)\,\mathrm{d}s \\ &= \int_0^t \mathcal{K}(t-s)\int_0^s \mathcal{K}(s-u)g(u)f(u)\,\mathrm{d}u g(s)\,\mathrm{d}s \\ &= \int_0^t \left(\int_u^t \mathcal{K}(t-s)\mathcal{K}(s-u)g(s)\,\mathrm{d}s\right) g(u)f(u)\,\mathrm{d}u \\ &\leq \int_0^t \mathcal{K}^{*2}(t-u)g(u)^2 f(u)\,\mathrm{d}u\end{aligned} \tag{5.63}$$

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 $q < 1$,

$$\begin{aligned}\mathscr{R}(t) &= F(t) + \sum_{j=1}^{\infty} \mathcal{G}^j(F)(t) \\ &\leq F(t) + \sum_{j=1}^{\infty}\int_0^t \mathcal{K}^{*(j)}(t-u)g(u)^j F(u)\,\mathrm{d}u \\ &\leq F(t) + \left(\sum_{j=0}^{\infty}(2\|\mathcal{K}\|_1\gamma_0(1+\varepsilon))^j - 1\right) C_1 \int_0^t \mathcal{K}(t-u)g(u)F(u)\,\mathrm{d}u \\ &\leq F(t) + \frac{q}{1-q}C_1(\mathcal{K} * (gF))(t)\end{aligned} \tag{5.64}$$

$$\tag{5.65}$$

where the third transition is by Lemma 5.4.4, with $C_1 = \frac{\mathcal{K}(0)}{\mathcal{K}(T)(2\epsilon+1)} + 1$, which then completes the proof. □

Lemma 5.4.4 (Lemma IV.4.7 in [4]). *Suppose $\mathcal{K}$ is monotonically decreasing, with $\|\mathcal{K}\|_1 < \infty$, and that there exists $T > 0$ such that $\forall t \geq T$, and $\epsilon \geq 0$,*

$$\int_0^t \mathcal{K}(s)\mathcal{K}(t-s)\,\mathrm{d}s \leq 2(1+\epsilon)\|\mathcal{K}\|_1\mathcal{K}(t). \tag{5.66}$$

Then,

$$\sup_{t\geq 0}\frac{\mathcal{K}^{*n}(t)}{\mathcal{K}(t)} \leq (2\|\mathcal{K}\|_1(1+\epsilon))^{n-1}\left(\frac{\mathcal{K}(0)}{\mathcal{K}(T)(2\epsilon+1)}+1\right) \tag{5.67}$$

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,

$$\frac{\mathcal{K}^{*(n+1)}(t)}{\mathcal{K}(t)(2\|\mathcal{K}\|_1)^n} = \frac{1}{\mathcal{K}(t)}\int_0^t \frac{\mathcal{K}(s)\mathcal{K}^{*n}(t-s)}{(2\|\mathcal{K}\|_1)^n}\,\mathrm{d}s \tag{5.68}$$

By the assumption of the Lemma, we know that there exists some $T > 0$ such that for $\forall t \geq T$

$$\int_0^t \frac{\mathcal{K}(s)\mathcal{K}(t-s)}{2\|\mathcal{K}\|_1}\,\mathrm{d}s \leq (1+\epsilon)\mathcal{K}(t). \tag{5.69}$$

Therefore, if $t \geq T$, we have

$$\begin{aligned}&\frac{1}{\mathcal{K}(t)}\int_0^t \frac{\mathcal{K}(s)\mathcal{K}^{*n}(t-s)}{(2\|\mathcal{K}\|_1)^n}\,\mathrm{d}s \\ &= \int_0^t \frac{\mathcal{K}(s)\mathcal{K}(t-s)}{2\|\mathcal{K}\|_1}\frac{\mathcal{K}^{*n}(t-s)}{\mathcal{K}(t-s)(2\|\mathcal{K}\|_1)^{n-1}}\,\mathrm{d}s \leq \alpha_n(1+\epsilon)\end{aligned} \tag{5.70}$$

On the other hand, if $t < T$,

$$\frac{1}{\mathcal{K}(t)}\int_0^t \frac{\mathcal{K}(s)\mathcal{K}^{*n}(t-s)}{(2\|\mathcal{K}\|_1)^n}\,\mathrm{d}s \leq \frac{\mathcal{K}(0)}{\mathcal{K}(T)}\frac{\|\mathcal{K}^{*n}(t)\|_1}{(2\|\mathcal{K}\|_1)^n} \leq \frac{\mathcal{K}(0)}{\mathcal{K}(T)2^n} \tag{5.71}$$

Taking supremum in Eq. (5.68), and combining the results of Eq. (5.71), and Eq. (5.70), 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 8. *$F(x) \asymp x^{-\kappa_1}$ and $\mathcal{K}(x) \asymp x^{-\kappa_2}$ for $x \geq 1$ with $\kappa_1 \geq 0$, $\kappa_2 > 1$*

Corollary 5.4.1 apply Lemma 5.4.3 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 5.4.2 shows the behavior of the learning rate in this setting. Finally, Lemma 5.4.5 shows that Assumption 8 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 5.4.1. *Suppose Assumption 8 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 5.4.3, we have

$$\begin{aligned} r(u) &\leq F(u) + C\int_0^u g(v)\mathcal{K}(u-v)F(v)\,\mathrm{d}v \\ &= F(u) + C\left(\int_0^{u/2} g(v)\mathcal{K}(u-v)F(v)\,\mathrm{d}v + \int_{u/2}^u g(v)\mathcal{K}(u-v)F(v)\,\mathrm{d}v\right) \\ &\leq F(u) + C_1 g(0)\left(\left(\frac{u}{2}\right)^{-\kappa_2}\int_0^{u/2} F(v)\,\mathrm{d}v + \left(\frac{u}{2}\right)^{-\kappa_1}\int_{u/2}^u \mathcal{K}(u-v)\,\mathrm{d}v\right) \\ &\leq F(u) + C_2(u^{-\kappa_2+1-\kappa_1} + u^{-\kappa_1}\|\mathcal{K}\|) \\ &= O(F(u)). \end{aligned} \tag{5.72}$$

Combining this upper bound with the lower bound from Lemma 5.4.3 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 5.4.2. *Suppose Assumption 8 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 5.4.1 and the formula for γ_t we get

$$g(u) \asymp \frac{\eta}{\sqrt{b^2 + \frac{2}{d}\mathrm{Tr}(K)\int_0^u \frac{(1+v)^{-\kappa_1}}{g(v)}\,\mathrm{d}v}}.$$

Let $I(u) = b^2 + \frac{2}{d}\mathrm{Tr}(K)\int_0^u \frac{(1+v)^{-\kappa_1}}{g(v)}\,\mathrm{d}v$ and observe that $g(u) \asymp \frac{1}{\sqrt{I(u)}}$ and $I'(u) = \frac{2}{d}\mathrm{Tr}(K)\frac{(1+u)^{-\kappa_1}}{g(u)}$. 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 5.4.1, we get

$$\gamma_t \asymp \frac{\eta}{\sqrt{b^2 + \frac{2}{d}\mathrm{Tr}(K)\int_0^t F(\Gamma(s))\,\mathrm{d}s}} \asymp (1+t)^{-(1-\kappa_1)/(2-\kappa_1)}.$$

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 5.4.1. □

Lemma 5.4.5. *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.* (5.66).

Proof. Following the definition in Eq. (5.52), and Eq. (5.51)

$$\begin{aligned} F(x) &= \frac{1-\beta}{2\lambda_{\max}^{1-\beta}}\int_0^{\lambda_{\max}} \lambda^{1-\beta-\delta}e^{-2\lambda x}\,\mathrm{d}\lambda \\ &= \frac{1-\beta}{2\lambda_{\max}^{1-\beta}(2x)^{2-\beta-\delta}}\int_0^{2\lambda_{\max}x} y^{1-\beta-\delta}e^{-y}\,\mathrm{d}y = \frac{1-\beta}{\lambda_{\max}^{1-\beta}2^{3-\beta-\delta}}\frac{\gamma(2-\beta-\delta, 2\lambda_{\max}x)}{x^{2-\beta-\delta}}. \end{aligned}$$

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. (5.66),

$$\begin{aligned}\int_0^t \mathcal{K}(s)\mathcal{K}(t-s) \, \mathrm{d}s &\leq \int_0^{t/2} \mathcal{K}(t)\mathcal{K}(t-s) \, \mathrm{d}s + \int_{t/2}^t \mathcal{K}(t)\mathcal{K}(t-s) \, \mathrm{d}s \\ &\leq \mathcal{K}(t/2) \left(\int_0^{t/2} \mathcal{K}(s) \, \mathrm{d}s + \int_{t/2}^t \mathcal{K}(t-s) \, \mathrm{d}s \right) \leq 2\mathcal{K}(t/2) \|\mathcal{K}\|_1\end{aligned}$$

by the power-law assumption for $t > T$, $\mathcal{K}(t/2) \asymp \mathcal{K}(t)$ which then complete the proof. □

Proof of Proposition 4.4.4. The proof is an immediate application of Corollary 5.4.2 with, $\kappa_1 = 2 - \beta - \delta$ as implied by Lemma 5.4.5. □

Remark 5.4.2. *This includes the case $\beta = 0$, which is the uniform measure on $[0, \lambda_{\max}]$.*

5.5 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 (4.9). Fixing $T > 0$ and any $\varepsilon \in (0, 1/2)$, we have by Theorem 4.2.1 (see also corollary 5.2.1 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 (4.7))

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}^2(t) = -2\gamma_t A(\mathscr{B}(t)) + \frac{\gamma_t^2}{d}\mathrm{Tr}(K) I(\mathscr{B}(t)), \quad \begin{cases} A(\mathscr{B}) = \mathbb{E}_{a,\epsilon}[\langle x - x^\star, f'(x \oplus x^\star)\rangle], \\ I(\mathscr{B}) = \mathbb{E}_{a,\epsilon}[f'(x \oplus x^\star)^2], \quad \text{where} \\ (x \oplus x^\star) \sim N(0, \mathscr{B}). \end{cases} \tag{5.73}$$

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

$$\bar{\mathfrak{g}}_k^{\mathscr{D}} \overset{\text{def}}{=} \frac{2\langle X_k - X^\star, \nabla\mathcal{R}(X_k)\rangle}{\frac{\mathrm{Tr}(K)}{d}\mathbb{E}_{a,\epsilon}[f'(\langle X_k, a\rangle; \langle X^\star, a\rangle, \epsilon)^2]} \text{ with the deterministic equiv. } \bar{\gamma}_t^{\mathscr{D}} = \frac{2A(\mathscr{B}(t))}{\frac{\mathrm{Tr}(K)}{d} I(\mathscr{B}(t))}. \tag{5.74}$$

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

$$\text{Polyak learning rate} \quad \mathfrak{g}_k^{\text{Polyak}} = \tfrac{1}{2}\bar{\mathfrak{g}}_k^{\mathscr{D}} \quad \text{and} \quad \text{deterministic equivalent} \quad \gamma_t^{\text{Polyak}} = \tfrac{1}{2}\bar{\gamma}_t^{\mathscr{D}}. \tag{5.75}$$

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 (5.74) and (5.75) 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 $\mathfrak{g}_k^{\text{Polyak}} \asymp \frac{1}{\mathrm{Tr}(K)/(d)} = \frac{1}{\text{avg. eig of } K}$.

[2]This is the idea of Polyak stepsize when the problem is deterministic.

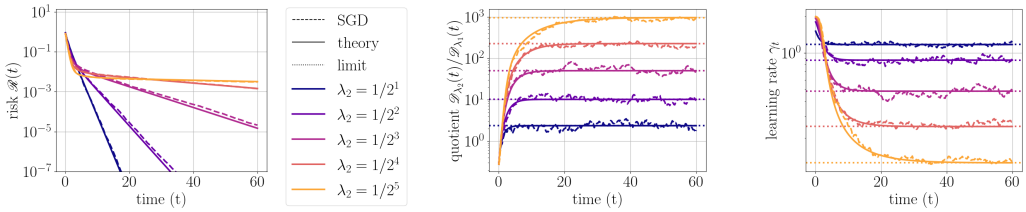

Figure 5.1: Convergence in Exact Line Search on a noiseless least squares problem. The plot on the left illustrates the convergence of the risk function, while the center and right plots depict the convergence of the quotient $\frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)}$ and the learning rate γ_t, respectively. Further details and formulas for the limiting behavior can be found in the Appendix 5.6.2. See Appendix 5.8 for simulation details.

In the case of least squares (see (4.10)), we get

$$\mathfrak{g}_k^{\text{Polyak}} = \frac{2\mathcal{R}(X_k) - \omega^2}{\frac{2\text{Tr}(K)}{d}\mathcal{R}(X_k)} \text{ and on a noiseless least squares, } \quad \mathfrak{g}_k^{\text{Polyak}} = \frac{1}{\frac{\text{Tr}(K)}{d}}.$$

The latter gives the best fixed learning rate for a noiseless target on a LS problem (as noted in [45, 55]).

5.6 Line Search

5.6.1 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 3 and 4),

$$\|\nabla\mathcal{R}(X)\|^2 = m(W^T K^2 W) \quad \text{and} \quad \text{Tr}(\nabla^2\mathcal{R}(X)K) = v(K). \tag{5.76}$$

Therefore using (4.7), 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

$$\mathrm{d}\mathscr{R} = -\gamma_t \mathscr{M}(t)\,\mathrm{d}t + \frac{\gamma_t^2}{d} v(K) I(\mathscr{B}(t)). \tag{5.77}$$

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

$$\text{risk threshold} \quad \bar{\mathfrak{g}}_k^{\mathscr{R}} = \frac{\|\nabla \mathcal{R}(X_k)\|^2}{\frac{\mathrm{Tr}(K\nabla^2 \mathcal{R}(X_k))}{d} I(W_k^T K W_k)} \text{ and deterministic equiv} \quad \bar{\gamma}_t^{\mathscr{R}} = \frac{\mathscr{M}(t)}{\frac{v(K)}{d} I(\mathscr{B}(t))}. \tag{5.78}$$

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

$$\mathfrak{g}_k^{\text{line}} = \tfrac{1}{2}\mathfrak{g}_k^{\mathscr{R}} \quad \text{and} \quad \text{deterministic equiv} \quad \gamma_t^{\text{line}} = \tfrac{1}{2}\gamma_t^{\mathscr{R}}.$$

5.6.2 Line Search on least squares

In this section, we provide a proof of Proposition 4.3.1, but, we show more than this including the exact limiting value for γ_t.

Proposition 5.6.1. *Consider the noiseless ($\omega = 0$) least squares problem* (4.10) *. Then the learning rate is always lower bounded by*

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

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

$$\lim_{k\to\infty} \gamma_t^{\text{line}} = \frac{2\left(\lambda_1^2 + \lambda_2^2 x\right)}{\left(\lambda_1 + \lambda_2 x\right)\left(\lambda_1^2 + \lambda_2^2\right)}, \tag{5.79}$$

where x is the positive real root of the second-degree polynomial

$$\mathcal{P}(x) = \lambda_1\lambda_2(x+1)(\lambda_2 x - \lambda_1) + (\lambda_2 - \lambda_1)^3 x. \tag{5.80}$$

This leads to

$$\frac{\lambda_{\min}(K)}{\frac{1}{d}\operatorname{Tr}(K^2)} \leq \lim_{t\to\infty} \gamma_t^{\text{line}} \leq \frac{2\lambda_{\min}(K)}{\frac{1}{d}\operatorname{Tr}(K^2)}. \tag{5.81}$$

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 (4.7), we derive

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}_\lambda(t) = -2\gamma_t\lambda\mathscr{D}_\lambda(t) + 2\gamma_t^2\lambda \times |\{\lambda = \lambda_i\}_{i=1}^{d}| \times \mathscr{R}(t)$$

for each distinct eigenvalue λ of K. Here $|\{\lambda = \lambda_i\}_{i=1}^{d}|$ is the number of eigenvalues of K that are equal to λ. It immediately follows by our construction of K that $|\{\lambda = \lambda_i\}_{i=1}^{d}| = \frac{d}{2}$. Thus, we establish the following system of ODEs

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}_{\lambda_1}(t) = -2\gamma_t\lambda_1\mathscr{D}_{\lambda_1}(t) + d\gamma_t^2\lambda_1\mathscr{R}(t) \\ \frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}_{\lambda_2}(t) = -2\gamma_t\lambda_2\mathscr{D}_{\lambda_2}(t) + d\gamma_t^2\lambda_2\mathscr{R}(t) \end{cases} \tag{5.82}$$

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

$$\mathcal{P}(x)=\lambda_1\lambda_2(x+1)(\lambda_2x-\lambda_1)+(\lambda_2-\lambda_1)^3x. \tag{5.83}$$

Solving for $x>0$, we derive the explicit formula

$$\begin{aligned}
&\lim_{t\to\infty}\frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)} \\
&=\frac{\lambda_1^3-2\lambda_1^2\lambda_2+2\lambda_1\lambda_2^2-\lambda_2^3+\sqrt{\lambda_1^6-4\lambda_1^5\lambda_2+8\lambda_1^4\lambda_2^2-6\lambda_1^3\lambda_2^3+8\lambda_1^2\lambda_2^4-4\lambda_1\lambda_2^5+\lambda_2^6}}{2\lambda_1\lambda_2^2}.
\end{aligned} \tag{5.84}$$

Given

$$\gamma_t^{\text{line}}=\frac{2\left(\lambda_1^2\mathscr{D}_{\lambda_1}(t)+\lambda_2^2\mathscr{D}_{\lambda_2}(t)\right)}{\left(\lambda_1\mathscr{D}_{\lambda_1}(t)+\lambda_2\mathscr{D}_{\lambda_2}(t)\right)\left(\lambda_1^2+\lambda_2^2\right)}=\frac{2\left(\lambda_1^2+\lambda_2^2\frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)}\right)}{\left(\lambda_1+\lambda_2\frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)}\right)\left(\lambda_1^2+\lambda_2^2\right)}, \tag{5.85}$$

we have

$$\lim_{t\to\infty}\gamma_t^{\text{line}}=\frac{2\left(\lambda_1^2+\lambda_2^2\lim_{t\to\infty}\frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)}\right)}{\left(\lambda_1+\lambda_2\lim_{t\to\infty}\frac{\mathscr{D}_{\lambda_2}(t)}{\mathscr{D}_{\lambda_1}(t)}\right)\left(\lambda_1^2+\lambda_2^2\right)}. \tag{5.86}$$

By substituting (5.84), we get

$$\begin{aligned}&\lim_{t\to\infty}\gamma_t^{\text{line}}\\&=\frac{\lambda_1^3+2\lambda_1^2\lambda_2+2\lambda_1\lambda_2^2+\lambda_2^3-\sqrt{\lambda_1^6-4\lambda_1^5\lambda_2+8\lambda_1^4\lambda_2^2-6\lambda_1^3\lambda_2^3+8\lambda_1^2\lambda_2^4-4\lambda_1\lambda_2^5+\lambda_2^6}}{\left(\lambda_1^2+\lambda_2^2\right)^2}.\end{aligned}\tag{5.87}$$

A direct calculation reveals that $\lambda_1>\lambda_2>0$ implies $\lim_{t\to\infty}\gamma_t^{\text{line}}\leq\frac{2\lambda_{\min}(K)}{\frac{1}{d}\text{Tr}(K^2)}$. □

Remark 5.6.1. *For the scenario where K has an arbitrary number n of distinct eigenvalues, equation* (4.13) *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*

$$\lim_{k\to\infty}\mathfrak{g}_k=\frac{n\left(\lambda_1^2+\lambda_2^2x_1+\cdots+\lambda_n^2x_{n-1}\right)}{\left(\lambda_1+\lambda_2x_1+\ldots\lambda_nx_{n-1}\right)\left(\lambda_1^2+\cdots+\lambda_n^2\right)},\tag{5.88}$$

where $x_1,\ldots,x_{n-1}>0$ satisfy a more intricate coupled system of $n-1$ equations.

5.7 Examples

Any single index model with α-pseudo Lipschitz ($\alpha\leq 1$) activation function is covered by our SGD+AL theory. In this section, we provide key learning problems within this family of models.

5.7.1 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 below formula,

$$\mathcal{R}(X)=\mathbb{E}_a\Big[-\langle X,a\rangle\cdot\frac{\exp(\langle X^\star,a\rangle)}{\exp(\langle X^\star,a\rangle)+1}+\log\left(\exp(\langle X,a\rangle)+1\right)\Big].\tag{5.89}$$

To study the ODE dynamics of SGD in Eq. (4.7) 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 [15] we obtain that

$$h(B) = -B_{21}\mathbb{E}_z\left[\frac{\exp(\sqrt{B_{22}}\cdot z)}{(1+\exp(\sqrt{B_{22}}\cdot z))^2}\right] + \mathbb{E}_w\big[\log(\exp(w\sqrt{B_{11}})+1)\big], \tag{5.90}$$

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 $g(x) \stackrel{\text{def}}{=} \frac{\exp(x)}{1+\exp(y)}$

$$I(B) = \frac{1}{2\pi\sqrt{\det(B)}}\int_{\mathbb{R}^2}(g(x)-g(y))^2\exp\left(-\frac{1}{2}\begin{pmatrix}x\\y\end{pmatrix}^T B^{-1}\begin{pmatrix}x\\y\end{pmatrix}\right)\mathrm{d}x\,\mathrm{d}y. \tag{5.91}$$

We note that the logistic regression is (μ,θ)–RSI (see 5.4.2) with $\mu = \frac{1}{\ell e^{\sqrt{4\theta}}}$ (see section 2.2 in [15]). Its Lipschitz constant is $\hat{L}(f) = 1$. Using Proposition 5.4.1 one can derive a lower bound on the limiting learning of AdaGrad Norm.

For more details and more examples, see [15].

5.7.2 CIFAR 5m

Finally, we include an example that uses real-world data, that is, the CIFAR 5m dataset [50]. 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

$$\arg\min_{X\in\mathbb{R}^d}\left\{\mathcal{R}(X):=\frac{1}{2n}\left\|FX-b\right\|^2=\frac{1}{2n}\sum_{i=1}^{n}\left(f_i\cdot X-b_i\right)^2\right\}, \tag{5.92}$$

where f_i is the ith row of F. The SGD we now consider is

$$X_{k+1}=X_k-\gamma_k\left(f_{i_{k+1}}\cdot X-b_{i_{k+1}}\right)f_{i_{k+1}},\quad \{i_k\}\ \text{iid}\ \text{Unif}(\{1,2,\cdots,n\}), \tag{5.93}$$

where γ_k is the usual AdaGrad-Norm stepsize, as in (5.1). Our empirical covariance matrix K (remembering that f_i is a row vector) is then

$$K=\mathbb{E}_{i\in[n],j\in[n]}\left[f_i^\top f_j\right]=\frac{1}{n}F^\top F. \tag{5.94}$$

We now use (5.50), 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.

5.8 Numerical simulation details

Here, for the sake of reproducibility, we provide more details for the figures that appear in the main paper.

Figure 4.1: 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 5.7. 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 4.2.1. The initial loss increase in the least squares problem suggesting

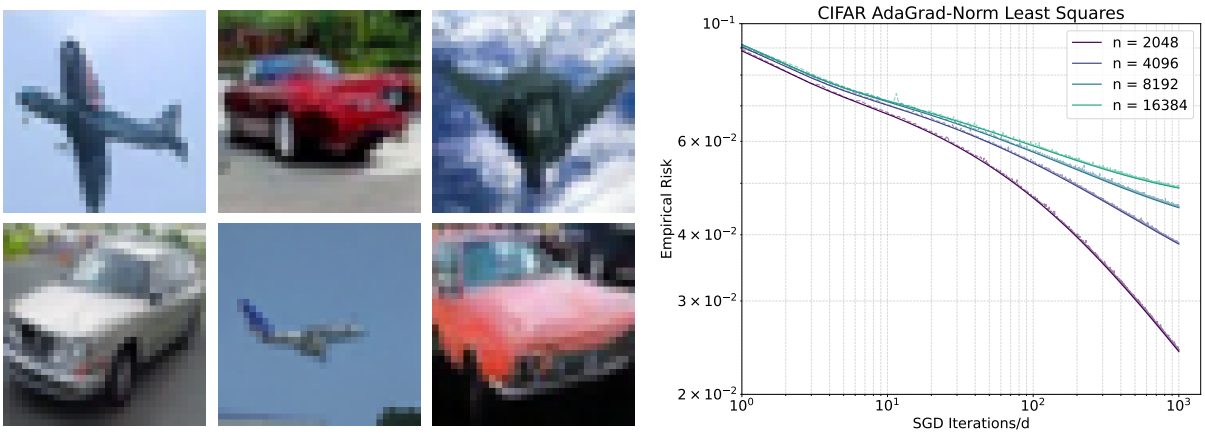

Figure 5.2: Predicting the training dynamics on a real dataset, CIFAR-5m [50], using multi-pass AdaGrad-Norm. This suggests the theory extends beyond Gaussian data and one-pass. Note that the curves look significantly different for different n; smaller values of n lead to an overparametrized problem, 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.

that the learning rate was initially too high, but AdaGrad-Norm naturally adapts and still the loss converges. Our ODEs predict this behavior.

Figure 4.2: 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 $\lambda_i(K) = \sqrt{\frac{d}{\sum_{i=1}^{d}\left(\frac{i}{d+1}\right)^{-2/s}}} \cdot \left(\frac{i}{d+1}\right)^{-1/s}, \quad i = 1, \ldots, d,$ 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 (4.13) 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 (4.13) is satisfied.

Figure 4.3: 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 $\frac{\eta}{\sqrt{b^2 + \frac{\mathrm{Tr}K\omega^2}{d}t}}$. Depicted is $\frac{\text{learning rate}}{\text{asymptotic l.r.}}$ so it approaches 1. *(center, right)*: Noiseless least squares setting $(\omega = 0)$. *(center)*: Prop. 4.4.2 predicts the avg. eig of K ($\mathrm{Tr}(K)/d$) as compared with $\lambda_{\max}$ affects the $\lim_{k\to\infty} \mathfrak{g}_k$. Indeed, this is true. We varied the $\kappa = \lambda_{\max}/\lambda_{\min}$ while keeping the $\mathrm{Tr}(K)/d$ and all other parameters fixed. All the learning rates behave identically verifying our theory about the effect of $\mathrm{Tr}(K)/d$ 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. 4.4.2. See Appendix 5.4 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 $\mathrm{Tr}(K)/d = 1$ 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 4.4: 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. 4.4.4.

Figure 5.1: 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 $K = \text{diag}(\lambda_i : i = 1, \ldots, 400)$. Further details and formulas for the limiting behavior can be found in the Appendix 5.6.2.

Figure 5.2 Convergence on CIFAR 5m [50]. 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 5.7.2.

Bibliography

[1] AGARWAL, A., BARTLETT, P. L., RAVIKUMAR, P., AND WAINWRIGHT, M. J. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization, 2011.

[2] ARNABOLDI, L., KRZAKALA, F., LOUREIRO, B., AND STEPHAN, L. Escaping mediocrity: how two-layer networks learn hard single-index models with SGD. *arXiv preprint arXiv:2305.18502* (2023).

[3] ARNABOLDI, L., STEPHAN, L., KRZAKALA, F., AND LOUREIRO, B. From high-dimensional and mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks. *arXiv preprint arXiv:2302.05882* (2023).

[4] ATHREYA, K. B., NEY, P. E., AND NEY, P. *Branching processes*. Courier Corporation, 2004.

[5] BALASUBRAMANIAN, K., GHOSAL, P., AND HE, Y. High-dimensional scaling limits and fluctuations of online least-squares SGD with smooth covariance. *arXiv preprint arXiv:2304.00707* (2023).

[6] BEN AROUS, G., GHEISSARI, R., AND JAGANNATH, A. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. In *Advances in Neural Information Processing Systems* (New York, 2022), vol. 35, Curran Associates, Inc., pp. 25349–25362.

[7] BERRADA, L., ZISSERMAN, A., AND KUMAR, M. P. Training neural networks for and by interpolation. In *International conference on machine learning* (2020), PMLR, pp. 799–809.

[8] BERTHIER, R., BACH, F., AND GAILLARD, P. Tight Nonparametric Convergence Rates for Stochastic Gradient Descent under the Noiseless Linear Model. In *Advances in Neural Information Processing Systems (NeurIPS)* (2020).

[9] BIEHL, M., AND RIEGLER, P. On-line learning with a perceptron. *Europhysics Letters 28*, 7 (1994), 525.

[10] BIEHL, M., AND SCHWARZE, H. Learning by on-line gradient descent. *Journal of Physics A: Mathematical and general 28*, 3 (1995), 643.

[11] BORDELON, B., AND PEHLEVAN, C. Learning Curves for SGD on Structured Features. In *International Conference on Learning Representations (ICLR)* (2022).

[12] BROWN, T., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J. D., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G., ASKELL, A., ET AL. Language models are few-shot learners. *Advances in Neural Information Processing Systems 33* (2020), 1877–1901.

[13] CELENTANO, M., CHENG, C., AND MONTANARI, A. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572* (2021).

[14] CHANDRASEKHER, K. A., PANANJADY, A., AND THRAMPOULIDIS, C. Sharp global convergence guarantees for iterative nonconvex optimization with random data. *Ann. Statist. 51*, 1 (2023), 179–210.

[15] COLLINS-WOODFIN, E., PAQUETTE, C., PAQUETTE, E., AND SEROUSSI, I. Hitting the high-dimensional notes: An ODE for SGD learning dynamics on GLMs and multi-index models. *arXiv preprint arXiv:2308.08977* (2023).

[16] COLLINS-WOODFIN, E., AND PAQUETTE, E. High-dimensional limit of one-pass SGD on least squares. *Electronic Communications in Probability 29* (2024), 1–15.

[17] COLLINS-WOODFIN, E., SEROUSSI, I., MALAXECHEBARRÍA, B. G., MACKENZIE, A. W., PAQUETTE, E., AND PAQUETTE, C. The high line: Exact risk and learning rate curves of stochastic adaptive learning rate algorithms, 2024.

[18] DAMIAN, A., NICHANI, E., GE, R., AND LEE, J. D. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. In *Advances in Neural Information Processing Systems* (2023), vol. 36, pp. 752–784.

[19] DANDI, Y., TROIANI, E., ARNABOLDI, L., PESCE, L., ZDEBOROVÁ, L., AND KRZAKALA, F. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. *arXiv preprint arXiv:2402.03220* (2024).

[20] DE, S., MUKHERJEE, A., AND ULLAH, E. Convergence guarantees for rmsprop and adam in non-convex optimization and an empirical comparison to nesterov acceleration, 2018.

[21] DEFAZIO, A., AND MISHCHENKO, K. Learning-rate-free learning by d-adaptation. In *International Conference on Machine Learning* (2023), PMLR, pp. 7449–7479.

[22] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[23] DUCHI, J., HAZAN, E., AND SINGER, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research 12*, 7 (2011).

[24] DVINSKIKH, D., OGALTSOV, A., GASNIKOV, A., DVURECHENSKY, P., TYURIN, A., AND SPOKOINY, V. Adaptive gradient descent for convex and non-convex stochastic optimization. *arXiv preprint arXiv:1911.08380* (2019).

[25] FAW, M., ROUT, L., CARAMANIS, C., AND SHAKKOTTAI, S. Beyond uniform smoothness: A stopped analysis of adaptive SGD. In *The Thirty Sixth Annual Conference on Learning Theory* (2023), PMLR, pp. 89–160.

[26] GERBELOT, C., TROIANI, E., MIGNACCO, F., KRZAKALA, F., AND ZDEBOROVÁ, L. Rigorous dynamical mean-field theory for stochastic gradient descent methods. *SIAM Journal on Mathematics of Data Science 6*, 2 (2024), 400–427.

[27] GOLDT, S., ADVANI, M., SAXE, A. M., KRZAKALA, F., AND ZDEBOROVÁ, L. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in neural information processing systems 32* (2019).

[28] GOLDT, S., LOUREIRO, B., REEVES, G., KRZAKALA, F., MÉZARD, M., AND ZDEBOROVÁ, L. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning* (New York, New York, USA, 2022), PMLR, pp. 426–471.

[29] GOLDT, S., MÉZARD, M., KRZAKALA, F., AND ZDEBOROVÁ, L. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X 10*, 4 (2020), 041044.

[30] GOWER, R. M., DEFAZIO, A., AND RABBAT, M. Stochastic polyak stepsize with a moving target. *arXiv preprint arXiv:2106.11851* (2021).

[31] HAZAN, E., AND KAKADE, S. Revisiting the polyak step size. *arXiv preprint arXiv:1905.00313* (2019).

[32] HINTON, G. Neural networks for machine learning - lecture 6a - overview of mini-batch gradient descent. Coursera Lecture Notes, 2012. Accessed: 2024.

[33] HINTON, G., SRIVASTAVA, N., AND SWERSKY, K. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on 14*, 8 (2012), 2.

[34] HOFFMANN, J., BORGEAUD, S., MENSCH, A., BUCHATSKAYA, E., CAI, T., RUTHERFORD, E., DE LAS CASAS, D., HENDRICKS, L. A., WELBL, J., CLARK, A., HENNIGAN, T., NOLAND, E., MILLICAN, K., VAN DEN DRIESSCHE, G., DAMOC, B., GUY, A., OSINDERO, S., SIMONYAN, K., ELSEN, E., RAE, J. W., VINYALS, O., AND SIFRE, L. Training compute-optimal large language models, 2022.

[35] IVGI, M., HINDER, O., AND CARMON, Y. DoG is SGD's best friend: A parameter-free dynamic step size schedule. *arXiv preprint arXiv:2302.12022* (2023).

[36] JIANG, X., AND STICH, S. U. Adaptive SGD with polyak stepsize and line-search: Robust convergence and variance reduction. *Advances in Neural Information Processing Systems 36* (2024).

[37] KAPLAN, J., MCCANDLISH, S., HENIGHAN, T., BROWN, T. B., CHESS, B., CHILD, R., GRAY, S., RADFORD, A., WU, J., AND AMODEI, D. Scaling laws for neural language models, 2020.

[38] KIEFER, J., AND WOLFOWITZ, J. Stochastic Estimation of the Maximum of a Regression Function. *Ann. Math. Statist. 23*, 3 (1952), 462–466.

[39] KINGMA, D. P., AND BA, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)* (2014).

[40] KINGMA, D. P., AND BA, J. Adam: A method for stochastic optimization, 2017.

[41] LEE, K., CHENG, A. N., PAQUETTE, C., AND PAQUETTE, E. Trajectory of Mini-Batch Momentum: Batch Size Saturation and Convergence in High Dimensions. *To Appear in NeurIPS 2022* (June 2022), 38pp.

[42] LEVY, K. Online to offline conversions, universality and adaptive minibatch sizes. *Advances in Neural Information Processing Systems 30* (2017).

[43] LEVY, K. Y., YURTSEVER, A., AND CEVHER, V. Online adaptive methods, universality and acceleration. *Advances in neural information processing systems 31* (2018).

[44] LI, X., AND ORABONA, F. On the convergence of stochastic gradient descent with adaptive stepsizes. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (2019), vol. 89 of *Proceedings of Machine Learning Research*, pp. 983–992.

[45] LOIZOU, N., VASWANI, S., LARADJI, I. H., AND LACOSTE-JULIEN, S. Stochastic polyak step-size for SGD: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics* (2021), PMLR, pp. 1306–1314.

[46] LOSHCHILOV, I., AND HUTTER, F. Decoupled weight decay regularization, 2019.

[47] MCMAHAN, H. B., AND STREETER, M. Adaptive bound optimization for online convex optimization. *arXiv preprint arXiv:1002.4908* (2010).

[48] MIGNACCO, F., KRZAKALA, F., URBANI, P., AND ZDEBOROVÁ, L. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. In *Advances in Neural Information Processing Systems* (2020), vol. 33, pp. 9540–9550.

[49] NAIR, V., AND HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning* (Madison, WI, USA, 2010), ICML'10, Omnipress, p. 807–814.

[50] NAKKIRAN, P., NEYSHABUR, B., AND SEDGHI, H. The Deep Bootstrap Framework: Good Online Learners are Good Offline Generalizers. In *International Conference on Learning Representations (ICLR)* (2021).

[51] NEDIĆ, A., AND BERTSEKAS, D. Convergence rate of incremental subgradient algorithms. *Stochastic optimization: algorithms and applications* (2001), 223–264.

[52] NEMIROVSKI, A., AND YUDIN, D. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, New York, 1983. Translated from Russian by E. R. Dawson.

[53] NICKOLLS, J., BUCK, I., GARLAND, M., AND SKADRON, K. Scalable parallel programming with cuda. In *ACM SIGGRAPH 2008 Classes* (New York, NY, USA, 2008), SIGGRAPH '08, Association for Computing Machinery.

[54] ORVIETO, A., LACOSTE-JULIEN, S., AND LOIZOU, N. Dynamics of SGD with stochastic polyak stepsizes: Truly adaptive variants and convergence to exact solution. *Advances in Neural Information Processing Systems 35* (2022), 26943–26954.

[55] PAQUETTE, C., LEE, K., PEDREGOSA, F., AND PAQUETTE, E. SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality. In *Proceedings of Thirty Fourth Conference on Learning Theory (COLT)* (2021), vol. 134, pp. 3548–3626.

[56] PAQUETTE, C., AND PAQUETTE, E. Dynamics of stochastic momentum methods on large-scale, quadratic models. In *Advances in Neural Information Processing Systems* (2021), vol. 34, pp. 9229–9240.

[57] PAQUETTE, C., PAQUETTE, E., ADLAM, B., AND PENNINGTON, J. Homogenization of SGD in high-dimensions: Exact dynamics and generalization properties. *arXiv e-prints* (May 2022), 64pp.

[58] PAQUETTE, C., PAQUETTE, E., ADLAM, B., AND PENNINGTON, J. Implicit regularization or implicit conditioning? exact risk trajectories of sgd in high dimensions. In *Advances in Neural Information Processing Systems* (New York, 2022), vol. 35, Curran Associates, Inc., pp. 35984–35999.

[59] PAQUETTE, C., AND SCHEINBERG, K. A stochastic line search method with expected complexity analysis. *SIAM J. Optim. 30*, 1 (2020), 349–376.

[60] PAQUETTE, E., PAQUETTE, C., XIAO, L., AND PENNINGTON, J. 4+3 phases of compute-optimal neural scaling laws, 2024.

[61] POLYAK, B. T. Introduction to optimization.

[62] POLYAK, B. T., AND JUDITSKY, A. B. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization 30*, 4 (1992), 838–855.

[63] RADFORD, A., WU, J., CHILD, R., LUAN, D., AMODEI, D., AND SUTSKEVER, I. Language models are unsupervised multitask learners. *OpenAI Blog 1*, 8 (2019), 9.

[64] ROBBINS, H., AND MONRO, S. A Stochastic Approximation Method. *Ann. Math. Statist.* (1951).

[65] ROLINEK, M., AND MARTIUS, G. L4: Practical loss-based stepsize adaptation for deep learning. *Advances in neural information processing systems 31* (2018).

[66] ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review 65*, 6 (nov 1958), 386–408.

[67] RUMELHART, D. E., HINTON, G. E., AND WILLIAMS, R. J. Learning representations by back-propagating errors. *Nature 323*, 6088 (1986), 533–536.

[68] SAAD, D., AND SOLLA, S. Dynamics of on-line gradient descent learning for multilayer neural networks. In *Advances in Neural Information Processing Systems* (1995), vol. 8, MIT Press.

[69] SAAD, D., AND SOLLA, S. A. Exact solution for on-line learning in multilayer neural networks. *Physical Review Letters 74*, 21 (1995), 4337.

[70] VARRE, A., PILLAUD-VIVIEN, L., AND FLAMMARION, N. Last iterate convergence of SGD for Least-Squares in the Interpolation regime. In *Advances in Neural Information Processing Systems (NeurIPS)* (2021).

[71] VASWANI, S., LARADJI, I., KUNSTNER, F., MENG, S. Y., SCHMIDT, M., AND LACOSTE-JULIEN, S. Adaptive gradient methods converge faster with over-parameterization (but you should do a line-search). *arXiv preprint arXiv:2006.06835* (2020).

[72] VASWANI, S., MISHKIN, A., LARADJI, I., SCHMIDT, M., GIDEL, G., AND LACOSTE-JULIEN, S. Painless Stochastic Gradient: Interpolation, Line-Search, and Convergence Rates. In *Advances in Neural Information Processing Systems (NeurIPS)* (2019), vol. 32, pp. 3732–3745.

[73] VELIKANOV, M., KUZNEDELEV, D., AND YAROTSKY, D. A view of mini-batch SGD via generating functions: conditions of convergence, phase transitions, benefit from negative momenta. In *International Conference on Learning Representations (ICLR)* (2023).

[74] VERSHYNIN, R. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, Cambridge, UK, 2018.

[75] WANG, B., ZHANG, H., MA, Z., AND CHEN, W. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual Conference on Learning Theory* (2023), PMLR, pp. 161–190.

[76] WANG, C., HU, H., AND LU, Y. A solvable high-dimensional model of GAN. In *Advances in Neural Information Processing Systems* (New York, 2019), vol. 32, Curran Associates, Inc.

[77] WARD, R., WU, X., AND BOTTOU, L. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *The Journal of Machine Learning Research 21*, 1 (2020), 9047–9076.

[78] WEI, A., HU, W., AND STEINHARDT, J. More than a toy: Random matrix models predict how real-world neural representations generalize. In *International Conference on Machine Learning* (2022), PMLR, pp. 23549–23588.

[79] WIDROW, B., AND HOFF, M. E. Adaptive switching circuits. *IRE WESCON Convention Record 4* (1960), 96–104.

[80] WU, X., WARD, R., AND BOTTOU, L. Wngrad: Learn the learning rate in gradient descent. *arXiv preprint arXiv:1803.02865* (2018).

[81] XIE, Y., WU, X., AND WARD, R. Linear convergence of adaptive stochastic gradient descent. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (2020), vol. 108 of *Proceedings of Machine Learning Research*, pp. 1475–1485.

[82] YOSHIDA, Y., AND OKADA, M. Data-dependence of plateau phenomenon in learning with neural network—statistical mechanical analysis. In *Advances in Neural Information Processing Systems* (New York, 2019), vol. 32, Curran Associates, Inc.