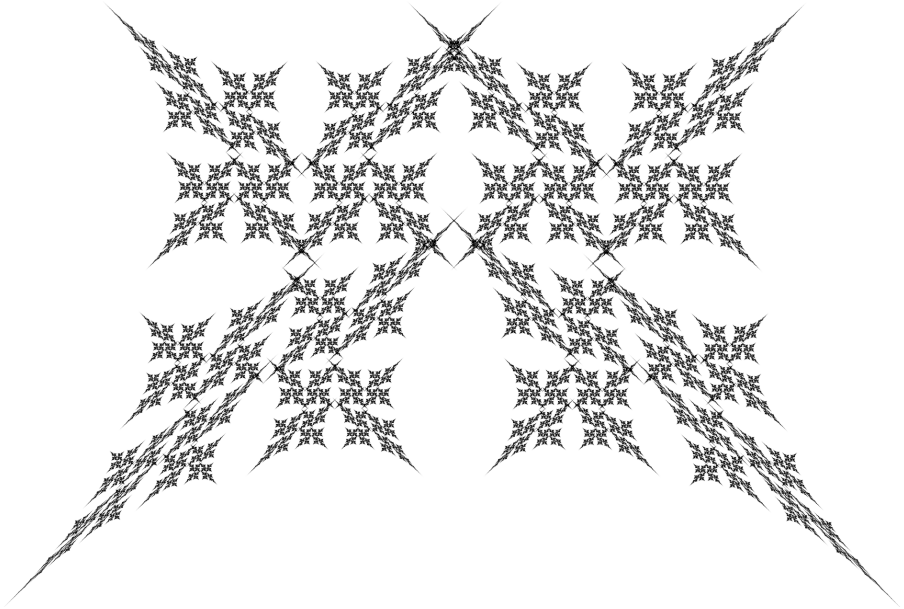
UNIVERSITÄT BAYREUTH

Adaptive Step Sizes for Stochastic Gradient Descent



Von der Universität Bayreuth
zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr.rer.nat)
genehmigte Abhandlung

von

Frederik Köhne
aus Homburg

1. Gutachter: Prof. Dr. Anton Schiela
2. Gutachter: Prof. Dr. Péter Koltai

Tag der Einreichung: 11.02.2026
Tag des Kolloquiums: 24.04.2026

The figure on the cover page shows the *attractor* of SGD with constant step size $\alpha = 1$ applied to the stochastic optimization problem:

$$\min_{x \in \mathbb{R}^2} F(x) = \frac{1}{2} \left(f_1(x) + f_2(x) \right)$$

where

$$f_i(x) = \frac{1}{2}(x - y_i)^T A_i (x - y_i)$$

with

$$A_i = \begin{pmatrix} 0.8892 & (-1)^i \cdot 0.6742 \\ (-1)^i \cdot 0.6742 & 0.8892 \end{pmatrix} \qquad \text{and} \qquad y_i = \begin{pmatrix} (-1)^i \cdot 1.3262 \\ -0.2473 \end{pmatrix}.$$

This attractor can also be viewed as the support of the invariant measure of SGD, as discussed in Section 3.4 and Chapter 7. The image was generated using an online tool for fractal rendering[1] applied to the corresponding iterated functions system given by the transition functions

$$\varphi_i(x) = x - \nabla f_i(x) = (I - A_i)x + A_i y_i.$$

[1] Accessed on October 7, 2025, under `https://sirxemic.github.io/ifs-animator/`.

Danksagung

An dieser Stelle möchte ich einigen Personen meinen besonderen Dank aussprechen, da ohne sie die Anfertigung dieser Arbeit nicht möglich gewesen wäre. Sie alle haben über vier Jahre entscheidend dazu beigetragen, dass diese Arbeit entstehen konnte oder waren auf andere Art und Weise daran beteiligt, dass ich meine Faszination für die Mathematik entwickeln und erleben durfte.

Zunächst gilt mein Dank natürlich meinem Doktorvater Prof. Dr. Anton Schiela, für die Ideengebung und Betreuung meiner Promotion. Anton war für Probleme aller Art immer direkt ansprechbar und hat stets mit Leidenschaft nach Lösungsansätzen gesucht, ohne aber dabei den Weg für meine eigenen Ideen und Ansätze einzuschränken. Die fachlichen Diskussionen an der Tafel oder über mit Stift und Papier, aber auch die persönlichen Gespräche an der Kaffeemaschine des Lehrstuhls werden mir immer in positiver Erinnerung bleiben. Ebenso die zahlreichen gemeinsamen Dienstreisen, auf denen neben den Tagungen immer auch etwas Zeit für sonstigen Austausch blieb.

Prof. Dr. Péter Koltai danke ich für die aufmerksame Lektüre meiner Dissertation und seine konstruktiven Verbesserungsvorschläge als Zweitgutachter. Die Diskussionen, auch mit Dr. Julia Slipantschuk, über Mittelungsprozesse, invariante Maße und iterierte Funktionensysteme haben mein Interesse für das Thema geweckt und dazu beigetragen, dass ich entsprechende Resultate in der Dissertation verwenden konnte.

Weiter gilt ein besonderer Dank Prof. Dr. Roland Herzog für die angenehme Zusammenarbeit im Projekt und den mehreren dabei entstandenen Papern. Roland stand für fachliche, aber auch taktische und technische Fragen zur Verfügung und hat mir einiges seiner technischen Expertise weitergeben können. Zusammen mit Anton hat Roland mit dem gemeinsamen Projektantrag erst dafür gesorgt, dass meine Promotionsstelle finanziert werden konnte.

Ebenfalls gilt ein besonderer Dank Leonie Kreis, die bei Roland im gleichen Projekt wie ich promovierte. Bei zahlreichen gegenseitigen Besuchen in Heidelberg und Bayreuth war es neben den fachlichen Diskussionen schön, einen Gesprächspartner in einer vergleichbaren Situation zu den persönlichen Fragen der wissenschaftlichen Welt gefunden zu haben.

Mein Dank gilt dem gesamten Team des Lehrstuhls für Angewandte Mathematik an der Universität Bayreuth. Prof. Dr. Lars Grüne war nicht nur mein Zweitbetreuer, sondern hat mit seiner ruhigen und professionellen Art stets für eine ausgesprochen angenehme Arbeitsatmosphäre am Lehrstuhl gesorgt. Meine geschätzten Kollegen Laura Weigl, Lisa Krügel, Jonas Schießl und Mario Sperl haben für eine unvergessliche Zeit gesorgt, nicht nur im Büro. Besonderer Dank gilt an dieser Stelle Dr. Robert Baier, für seinen unermüdlichen Einsatz für den Lehrstuhl, dafür dass er bei allen Hilferufen stets zur Stelle war und Problemen mit einer Genauigkeit auf den Grund geht, die ihresgleichen sucht. Und schließlich danke ich Sigrid Kinder und Andrea Groll, die für allerlei spontane Anfragen stets erreichbar waren und mich präzise durch die Stromschnellen der Universitätsverwaltung zu führen wussten.

An dieser Stelle möchte ich auch Markus Schröder, meinem Lehrer im Mathematik Leistungskurs am Gymnasium der Stadt Warstein danken, dafür, mir die Begeisterung für die Schönheit der Mathematik erfolgreich vermittelt zu haben.

Robin danke ich für die gemeinsame Studienzeit in Bayreuth. Sowohl die mathematischen Diskussionen als auch die Momente abseits der Mathematik haben meine Studienzeit nachhaltig geprägt.

Tobias hat als mein Mitbewohner und guter Freund die letzten Jahre meiner Zeit in Bayreuth, und schließlich das Kolloquium zu meiner Doktorarbeit, entscheidend bereichert.

Besonderer Dank gilt meiner Familie, insbesondere meinen Eltern Sigrun und Georg. Nicht nur habt ihr mir die Freiheit gegebenen, mich über Jahre mit der Mathematik beschäftigen zu dürfen, ihr habt auch dabei immer an mich geglaubt, ohne genau nachzuvollziehen, womit ich mich genau beschäftige. Ihr habt mir das Gefühl gegeben, stolz auf mich zu sein, und mich darin bestärkt, diesen Weg zu gehen. Ohne diese Unterstützung wäre diese Doktorarbeit sicher nicht entstanden. Dank gilt auch meinem Großvater Norbert, der mich früh mit seinen persönlichen Erzählungen für die Wissenschaft sensibilisiert hat und sich stets für die Inhalte meiner Forschung interessiert zeigte.

Schließlich möchte ich meiner Freundin Sonja ganz besonders danken. Für die gesamte gemeinsame Zeit während des Studiums in Bayreuth. Dafür, dass Du Dich im Studium mit mir über meinen Spaß an der Mathematik freuen konntest, auch wenn Du ihn nicht immer teilen konntest. Dafür, dass Du gerade in den schweren Phasen der Promotion an mich geglaubt hast und mich stets motiviert hast weiter zu machen. Während des Finalisierens der Dissertation und der Vorbereitung des Kolloquiums in Hamburg hast Du viel auf Dich genommen, um mir die Zeit und den Raum zu verschaffen, den ich brauchte. Für all das bin ich Dir unendlich dankbar.

Zusammenfassung

Stochastische Optimierungsprobleme treten in verschiedenen Anwendungsfällen auf. In den letzten Jahren hat die rasante Entwicklung von Techniken des maschinellen Lernens, und damit verbunden das Training von Künstlicher Intelligenz, das Interesse an solchen Problemen noch einmal verstärkt. Ein weit verbreiteter und konzeptionell sehr einfacher Algorithmus zum Lösen solcher stochastischen Optimierungsprobleme ist das stochastische Gradientenverfahren (Stochastic Gradient Descent (SGD)). Dieses Verfahren kann als eine Adaption des klassischen (deterministischen) Gradientenverfahren verstanden werden, bei dem der Gradient mit einer stochastischen Approximation an den Gradienten ersetzt wird. Trotz seiner Ähnlichkeit zum deterministischen Gegenpart, verhält sich das stochastische Gradientenverfahren teils deutlich anders. Dies gilt insbesondere bei der Wahl der Schrittweiten. Während beim deterministischen Verfahren konstante Schrittweiten für Konvergenz in der Regel ausreichend sind, ist dies beim stochastischen Gradientenverfahren nicht der Fall. Hier muss, je nach Fall, die Schrittweitensteuerung die Unsicherheit in den Suchrichtungen berücksichtigen und Schrittweiten gegebenenfalls anpassen.

Um diesem Problem zu begegnen, befasst sich diese Arbeit mit der Konstruktion und Analyse einer adaptiven Schrittweitensteuerung für das stochastische Gradientenverfahren, welche insbesondere auf die Unsicherheit in der Suchrichtung versucht einzugehen. Diese adaptive Schrittweitensteuerung beruht auf einer Schrittweitenwahl, für die sich zwar optimale Konvergenzraten zeigen lassen, die allerdings in der Praxis nicht berechenbar ist. Als Lösung dieses Problems werden bestimmte Größen identifiziert, die während des Durchlaufs des Algorithmus beobachtet und mittels eines geeigneten Mittelungsprozess gemittelt werden, um verlässliche Schätzer für die zu verwendenden Schrittweiten zu erhalten. Neben der Analyse von SGD unter Verwendung der theoretischen, nicht berechenbaren Schrittweiten, befasst sich diese Arbeit insbesondere mit der Analyse der Schätzungsprozesse und dem Verhalten der aus den geschätzten Größen ermittelten Schrittweiten. Diese Analyse beruht zum einen auf einer detaillierten Konvergenztheorie des Mittelungsprozesses, zum anderen auf dem Langzeitverhalten von SGD mit konstanten Schrittweiten, welches mit Hilfe so-genannter invarianter Maße beschrieben werden kann. Neben der theoretischen Analyse zeigen wir in numerischen Experimenten den adaptiven Charakter und das Konvergenzverhalten von SGD unter Verwendung der geschätzten Schrittweiten.

Abstract

Stochastic optimization problems arise in various applications. In recent years, the rapid development of machine learning techniques, and the associated training of artificial intelligence, has further increased interest in such problems. A widely used and conceptually very simple algorithm for solving such stochastic optimization problems is the Stochastic Gradient Descent (SGD) method. This method can be understood as an adaptation of the classical (deterministic) gradient method in which the gradient is replaced by a stochastic approximation of the gradient. Despite its similarity to its deterministic counterpart, the stochastic gradient method behaves in part quite differently. This is particularly true for the choice of step sizes. While constant step sizes are usually sufficient for convergence in the deterministic method, this is not the case for the stochastic gradient method. Here, depending on the situation, the step size control must take the uncertainty in the search directions into account and adjust the step sizes if necessary.

To address this problem, this work deals with the construction and analysis of an adaptive step size control for the stochastic gradient method, which in particular aims to account for the uncertainty in the search direction. This adaptive step size control is based on a step size rule for which optimal convergence rates can be shown, but which is not computable in practice. As a solution to this problem, certain quantities are identified that can be observed during the execution of the algorithm and averaged by means of a suitable averaging process in order to provide reliable estimators for the step sizes to be used. In addition to the analysis of SGD using the theoretical, non-computable step sizes, this work focuses in particular on the analysis of the estimation processes and the behavior of the step sizes derived from the estimated quantities. This analysis is based, on the one hand, on a detailed convergence theory of the averaging process, and, on the other hand, on the long-term behavior of SGD with constant step sizes, which can be described using so-called invariant measures. In addition to the theoretical analysis, we demonstrate the adaptive character and the convergence behavior of SGD using the estimated step sizes in numerical experiments.

Contents

1 Introduction

Stochastic optimization problems arise in many applications in which the objective function is affected by noise. Such noise may originate from inherent randomness in the modeled phenomenon or from limitations in computational resources. In the latter case, a computationally simpler stochastic approximation of a complex objective function is often used, for example by subsampling a large data set, as is standard in machine learning and training of artificial intelligence. Formally, these problems are described by a target function F that is the *mean* of a *parameterized distribution* of functions:

$$F(x) = \int_\Omega f_\omega(x) \, \mathrm{d}\mathbb{P}(\omega).$$

Here, Ω denotes a suitable measurable space equipped with a probability measure $\mathbb{P}$.

Stochastic Gradient Descent (SGD) is a simple yet powerful first-order method widely used to solve such problems. Assuming access to a noisy but unbiased estimator of the gradient ∇F of the target function F, SGD mimics the classical gradient descent method by replacing the true gradient with a stochastic sample. Formally, one assumes that

$$\nabla F(x) = \int_\Omega \nabla f_\omega(x) \, \mathrm{d}\mathbb{P}(\omega),$$

and that $\nabla f_\omega(x)$ can be sampled with $\omega \sim \mathbb{P}$. In each iteration n of SGD, a new sample ω_n is drawn and the update

$$x_{n+1} = x_n - \alpha_n \nabla f_{\omega_n}(x_n)$$

is performed. The functions f_ω are referred to as the *sampled functions*. This algorithm dates back until the 1950s, where it was studied in [60] as a stochastic approximation algorithm. Its conceptual simplicity and scalability have led to numerous variants and to its central role in modern machine learning.

In case of the deterministic, or classical, gradient descent method, the choice of the step size α_n is relatively straightforward. For L-smooth and strongly convex functions, simple constant step sizes $\alpha_n = \alpha$ guarantee linear convergence of the algorithm, if α is chosen sufficiently small. Under mere convexity, one still obtains convergence rates

of order $O(\frac{1}{n})$. More generally, in the deterministic setting, the step sizes compensate for the *nonlinearity* of F, quantifying how well a linear model locally approximates the function.

In contrast to the simplicity of the deterministic case, the selection of step sizes is significantly more subtle in stochastic optimization. Besides accounting for nonlinearity, step sizes must additionally accommodate noise in the search direction. Depending on the behavior of the noise near the solution, reduction of step sizes, or even convergence to zero, may be required to ensure convergence. In other scenarios decaying step sizes unnecessarily slow down the speed of convergence. Consequently, the choice of step sizes heavily influences the speed of convergence of the resulting algorithm. Optimal step sizes depend on the problem itself, as in the deterministic case, but also on local problem characteristics, motivating the use of adaptive step sizes. A core difficulty arises because all practically accessible information is noisy. Not only are stochastic gradients noisy, but most computable indicators of convergence, such as function values or gradient norms, are not available exactly in practice. Thus, adaptive step size schemes must be robust to noise while relying only on limited, indirect observations of the algorithm's progress. For example, in the deterministic setting, the computationally available value of $\|\nabla F(x_n)\|$ is a good surrogate for the progress of the algorithm. Such a convergence monitor is, in general, not available in the stochastic setting.

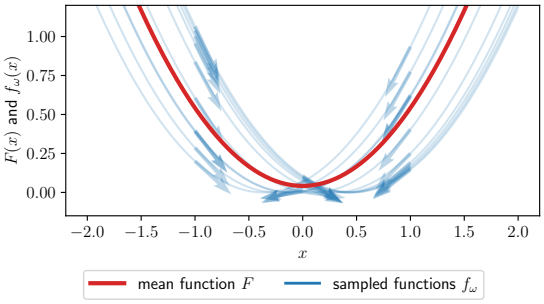


Figure 1.1: Visualization of a simple stochastic optimization problem, in which SGD is not stationary at the solution, demonstrating the need for decaying step sizes.

To illustrate the difficulties in step size selection and the dependency on local characteristic of the problem, we consider a one–dimensional stochastic optimization problem in which each sampled function depends on a single random parameter. For each realization of the parameter ω, drawn independently from a uniform distribution on the interval $[-\frac{1}{2}, \frac{1}{2}]$, we define the sampled function $f_\omega(x) = \frac{1}{2}(x-\omega)^2$, so that $F(x) = \frac{1}{2}\left(x^2 + \frac{1}{12}\right)$, minimized in $x^\star = 0$. This construction yields a family of convex quadratic functions whose minimizers coincide with the corresponding sampled parameters.

To visualize the variability induced by the randomness in ω, we generate multiple samples and plot the corresponding functions f_ω as well as their mean $F(x)$, which is our objective, in Figure 1.1. Additionally, we evaluate the gradient $\nabla_x f_\omega(x) = 2(x-\omega)$ at

the query points $x = -1$, $x = 0$, and $x = 1$, see Figure 1.1 At each query point and for each sampled function f_ω, we plot a normalized arrow indicating the negative gradient direction and the descent on the sampled function, expected from taking this gradient step. Of particular interest are the evaluations at $x = x^\star = 0$. It becomes evident that SGD is not stationary at the optimizer of the expected objective, but instead produces a nonzero update in random directions due to the inherent sampling noise.

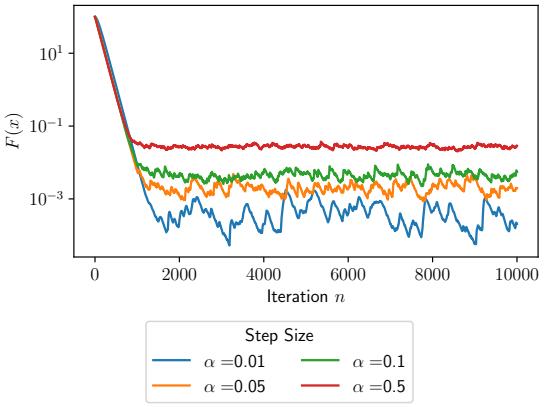


Figure 1.2: Performance of SGD on the example problem using different step sizes.

These random updates in the vicinity of the minimizer are scaled by the step size α, and their magnitude therefore increases proportionally with α. As a consequence, larger step sizes yield less accurate approximations of the optimal solution, whereas smaller step sizes reduce the influence of these random fluctuations and thereby lead to a more accurate approximation of the minimizer. To illustrate this effect, we have applied SGD to this simple test problem, with different constant step sizes. The results of this experiment are depicted in Figure 1.2, where the performance of SGD, measured in the value of the objective $F(x_n)$ is plotted for different step sizes used. The individual plots are smoothed for enhanced readability. Evidently, all step sizes lead to stagnating behavior of the algorithm, with larger step sizes leading to a larger stagnation level. This effect will be discussed in more detail in Chapter 3.

To address these challenges in step size selection, this work focusses on the development and analysis of an adaptive step size scheme for SGD, by identifying a theoretical (in general non-computable) step size rule. This step sizes rule depends on the current iterate x_n, or, more general, on the variable x, and thus provides a map $\alpha^\star : x \mapsto \alpha^\star(x)$. We have illustrated this local dependency, as well as the performance of SGD using this step size rule in Figure 1.3, applied to the example problem discussed above. In this scenario it can be shown that

$$\alpha^\star(x) = 1 - \frac{1}{12\,x^2 + 1}.$$

In addition to the ideal step sizes, we propose an estimation technique that aims at approximating the step sizes scheme. This estimation technique is based upon three quantities that are observed during the execution of the algorithm. These observations are smoothed in a suitable way, using a technique we refer to as p-EMA, which is

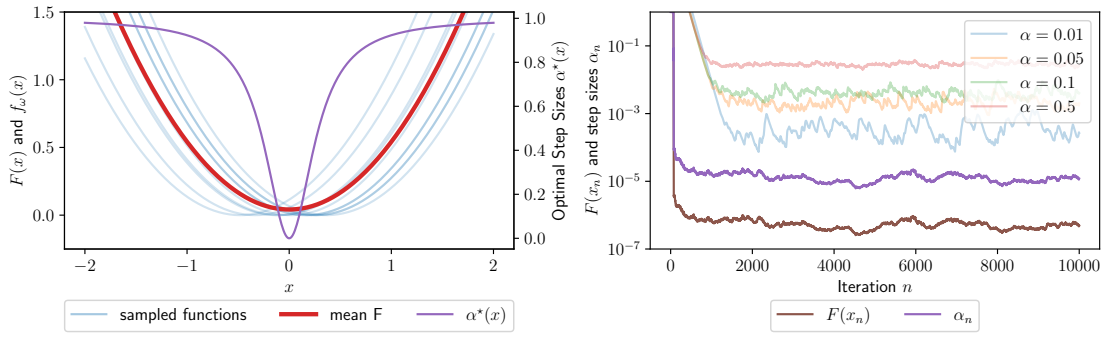


Figure 1.3: Illustration of the ideal step sizes developed and approximated in this work. The figure on the left shows the mean function and some sampled functions as in Figure 1.1. Additionally, the ideal step sizes $\alpha^\star$ in dependence of x is plotted. The figure on the right shows the performance of SGD using the step sizes $\alpha_n = \alpha^\star(x_n)$, displaying both, $F(x_n)$ and α_n. For reference, the performance of the constant step sizes from Figure 1.2 is shown as well.

particularly suited for de-noising noisy observations made along trajectories of evolving systems.

Before summarizing related work, we highlight the main contributions of this dissertation:

- identification of an ideal (non-computable) step size rule for SGD based on local problem characteristics,
- development of estimators for this ideal rule using observable noisy quantities,
- introduction and analysis of the p-EMA smoothing scheme for robust estimation,
- convergence results for SGD with ideal step sizes and convergence results for the estimated step sizes.

The remainder of this introduction reviews existing adaptive strategies. We then provide a reading guide to help navigate the structure of the dissertation.

Known Adaptive Step Size Strategies

It is well known that the performance of SGD, as well as convergence guarantees, crucially depend on the step sizes employed. Therefore, different approaches for making the step sizes of SGD adaptive have been developed. We briefly summarize them in what follows.

This following discussion of the literature is partly taken from our article [41].

Polyak-Type Strategies

Polyak-type strategies aim to adapt the well known Polyak step sizes for classical gradient schemes, first presented in [57], to the stochastic setting. A common assumption is that the minimum, or at least a lower bound to the minimum of the sampled function, is known. In [47], the authors derive convergence properties of SGD with Polyak-type step sizes for the interpolating setting (no noise at the minimizer) and convergence to a neighborhood of the minimizer for the non-interpolating case (noise present at the minimizer); see Definition 3.5 for a definition of the two settings. [33] extended the work of [47] to obtain convergence, even in the non-interpolating setting. For strongly convex target functions, both works obtain linear convergence in the interpolating setting. The latter work also shows sublinear convergence of order $O(\frac{1}{\sqrt{k}})$ in the non-interpolating setting.

Line Search Strategies

Line search strategies aim to apply the concept of line search from classical optimization to stochastic optimization. These strategies typically involve the repeated evaluation of the sampled function at various candidates for the next iterate until a desired decrease is observed. A direct adaptation of the well-established Armijo line search is documented in [67]. Convergence theory for line search methods must often consider the noise introduced by the sampled function. A theory that controls this noise can be found in [55]. Both works achieve linear convergence in the strongly convex, interpolating regime. A significant limitation of line search methods is the repeated evaluation of the sampled function at each iteration, which can become computationally expensive.

(Diagonal) Scaling Methods

Another class of commonly used adaptive methods can be classified as diagonal scaling methods, which gather information from past iterations to develop a step size strategy where each dimension of the input space has its unique step size. It is also possible to interpret these as methods that employ a *diagonal* preconditioning matrix to the derivative, in order to obtain the search direction, where the preconditioning matrix needs not to be constant over time. Prominent examples of these methods include RMSProp[1], Adagrad ([19]), Adadelta ([70]), and Adam ([36]), as well as its numerous variants. For these classes of algorithms it remains, however, unclear how the choices

[1]Proposed in unpublished work [29] by Geoffrey Hinton et al.; see also [61].

of scalings are related the convergence of the algorithms. In [66] the authors propose to use line search methods to set up the step size for Adagrad. All these methods have in common that they scale the gradient in every dimension of the state space with the inverse of the sum of the squared gradient values in the respective directions seen so far. A similar scaling, although not component-wise but globally over all dimensions, has recently been introduced in different works [17, 31, 49, 50]. The key idea behind such techniques is to distinguish between interpolating and non-interpolating regimes, see Definition 3.5, as in the non-interpolating case the cumulative gradient norms of noisy gradients tends to infinity, and consequently the corresponding step sizes tend to zero. While this is theoretically valid, the scaling factor, which is a sum over past observations, tends to be large close to the minimizer. This is due to large observations at the beginning of the algorithm and the usually large number of iterations. However, close to the minimizer the increments to this sum are small. As a result, the divergence of the sum is slowed down, and subsequently the adjustment of the step sizes. To overcome this limitation, we introduce *local estimators* that allow for more responsive step size adaptation.

Trust Region Methods

Another line of research focuses on trust region methods. Here, adaptivity stems from selecting the trust region radius based on previous iterations. Examples of such work can be found in [9] and [15].

Factorization of Step Sizes

A core property of the step sizes proposed in Chapter 4 is their factorization into two components: one that is agnostic to the nonlinearity of the problem, and another that accounts for the stochasticity of the problem. A related factorization—albeit one that depends on *a priori* chosen constants—was also considered in [65]. In contrast, our method will depend on local properties of the noise.

Other Adaptations of SGD

The aforementioned extensions, as well as the method we propose in Chapter 4, aim at reducing the impact of the uncertainty in the search direction by adapting the step size accordingly. Another approach is to reduce the noise in the search direction itself. Such approaches usually aggregate information on the gradient over several iterations, or require the occasional evaluation of the true gradient ∇F. Such extensions include variance reduction techniques such as SVRG in [35], SAG in [62] or SAGA in [16], which aim at reducing the noise in the search direction. Momentum schemes, based

on the heavy ball method [56] are a popular choice and can be considered a variance reduction scheme as well. However, their theoretical advantage is not as clear as in the deterministic case ([63, 46]).

How to Read This Work

This dissertation is largely self-contained. As a consequence, it includes several preparatory sections that introduce concepts which are needed only occasionally or serve mainly as background. Moreover, the convergence theory of the proposed estimators requires substantial technical development that is not necessary for readers primarily interested in the motivation and ideas behind the adaptive step sizes. The purpose of this guide is therefore to provide an overview of the contents of each chapter and to indicate which parts may be safely skipped on a first reading.

Chapter 2 reviews standard material. It may be used as a reference if certain notions are unfamiliar, but can be omitted on a first pass.

Chapter 3 recalls the stochastic gradient method. It further provides insights into the performance of different step size strategies and motivates the introduction of adaptive step sizes. The final Section 3.4 describes the long-term behavior of SGD with constant step sizes. Some of the basic insights on SGD are used occasionally throughout this work. The results on the long-term behavior are particularly relevant for the convergence theory of the estimators in Chapter 7. Readers already familiar with SGD may safely skip most parts of this chapter. Section 3.4 is not required for understanding the construction of the adaptive step sizes, but it provides essential technical foundations for the convergence theory of the estimators and is therefore recommended for readers who wish to fully understand the developments in Chapter 7.

Chapter 4 presents the *ideal* step sizes and our approach to estimating them during the execution of SGD. This chapter is central to the dissertation and focuses on the conceptual ideas rather than detailed theoretical results.

Chapter 5 contains the convergence theory for the ideal step sizes and thereby motivates the incorporation of estimators to approximate these step sizes. The technical details of the proofs are not required for the remainder of the work.

Chapter 6 develops the theory of the p-EMA averaging process, an adaption of the classic exponential moving average (EMA), used in the construction of the adaptive step sizes. Together with the invariant measures discussed in Section 3.4, this theory underpins the convergence analysis for the estimators in Chapter 7. The key contributions of this chapter are Theorem 6.9 and Corollary 6.13. The technical details, in particular those in the proof of Theorem 6.9, are not necessary for understanding the implications of the results in Chapter 7.

Finally, Chapter 7 presents the convergence results for the estimators, whose central result is stated in Theorem 7.8. To fully appreciate these results, it is helpful to recall the invariant measure from Section 3.4. The discussions in Sections 7.2.2 and 7.2.3 are central for understanding the justification of the estimated step sizes, while the technical material in Section 7.1 may be omitted on a first reading.

Finally, in Chapter 8 we apply the developed algorithm to test problems and discuss numerical results.

2 Preliminaries

2.1 General Notation

In this work, $\mathbb{N}$ denotes the positve natural numbers

$$\mathbb{N} = \{1, 2, \dots\},$$

and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. For $n, m \in \mathbb{N}_0 \cup \{\infty\}$ we define

$$[n : m] := \{k \in \mathbb{N}_0 \mid n \leq k \leq m\}.$$

By $\mathbb{R}$ we denote the set of real numbers.

2.2 Convex and Smooth Functions

During this work we will consider special classes of functions. In this section, we will give the relevant definitions and some elementary properties. All results are well known, we repeat them here for completeness. We consider a real, nontrivial Hilbert space $\mathcal{H} \neq \{0\}$ with dual $\mathcal{H}^*$ and norm $\|\cdot\|_{\mathcal{H}}$. The dual norm will be denoted with $\|\cdot\|_{\mathcal{H}^*}$, the dual paring of $\phi \in \mathcal{H}^*$ and $x \in \mathcal{H}$ by (ϕ, x), and the inner product of $x, y \in \mathcal{H}$ by $\langle x, y\rangle$.

Some of the following, well known, results are taken from the reference [52], where $\mathcal{H} = \mathbb{R}^n$ is considered. If we refer to the respective result in the reference, the proof is simple and can directly be adjusted to our setting, where we consider a real, possibly infinite dimensional Hilbert space $\mathcal{H}$.

We recall the notion of (Fréchet-) Differentiability in our setting (see [5, Definition 2.56]):

Definition 2.1. *A map $f : \mathcal{H} \to \mathbb{R}$ is called differentiable, if, for any $x \in \mathcal{H}$, there is $f'(x) \in \mathcal{H}^*$, such that*

$$\lim_{y \to x} \frac{f(y) - f(x) - (f'(x), y - x)}{\|x - y\|_{\mathcal{H}}} = 0.$$

Definition 2.2. *Consider $\mu, L \geq 0$. A map $f : \mathcal{H} \to \mathbb{R}$ is called*

1. *convex, if*
$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$
for all $x, y \in \mathcal{H}$ *and* $\lambda \in [0,1]$.

2. μ*-strongly convex, if* f *is differentiable with derivative* $f : \mathcal{H} \to \mathcal{H}^*$ *such that*
$$f(y) \geq f(x) + \big(f'(x), y - x\big) + \frac{\mu}{2} \|x - y\|_{\mathcal{H}}^2$$
for all $x, y \in \mathcal{H}$.

3. L*-smooth, if* f *is differentiable with derivative* $f' : \mathcal{H} \to \mathcal{H}^*$ *such that*
$$\big\|f'(x) - f'(y)\big\|_{\mathcal{H}^*} \leq L \|x - y\|_{\mathcal{H}}$$
for all $x, y \in \mathcal{H}$.

4. (μ, L)*-feasible, if* f *is* μ*-strongly convex and* L*-smooth.*

Remark 2.3. *It is not necessary for a function to be differentiable to define the notion of strong convexity. The characterization*
$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) - \lambda(1-\lambda)\frac{\mu}{2} \|x - y\|_{\mathcal{H}}^2 \tag{2.1}$$
for all $x, y \in \mathcal{H}$ *and* $\lambda \in [0,1]$ *is equivalent to the definition in Definition 2.2 and does not require differentiability. For the finite dimensional case this is proven in [52, Theorem 2.1.9], and the proof can be directly adjusted to our setting. In the infinite-dimensional setting the result also follows from [5, Proposition 17.7] together with [5, Proposition 10.8].*

Remark 2.4. *We explicitly allow for* $\mu = 0$ *in the definition of strong convexity. A function is* 0*-stronlgy convex, if and only if it is convex.*

The following result is fundamental to the theory on gradient based optimization algorithms, and can be interpreted as a quantification of the first-order approximation error. It is commonly used to show descent inequalities for gradient based methods.

Lemma 2.5. *Let* $f : \mathcal{H} \to \mathbb{R}$ *be differentiable. Then,* f *is* L*-smooth, if and only if*
$$f(y) \leq f(x) + \big(f'(x), y - x\big) + \frac{L}{2} \|x - y\|_{\mathcal{H}}^2 \qquad \text{for all} \qquad x, y \in \mathcal{H}. \tag{2.2}$$

Proof. See [52, Theorem 2.1.5] for the finite dimensional case and [5, Theorem 18.15] for the infinite dimensional case. □

The following result summarizes some basic properties of convex and L-smooth functions.

Proposition 2.6. *1. Let f be differentiable and μ-strongly convex, then:*

$$\big(f'(x) - f'(y), x - y\big) \geq \mu \|x - y\|_{\mathcal{H}}^2$$

for all $x, y \in \mathcal{H}$.

2. If $f : \mathcal{H} \to \mathbb{R}$ is L-smooth, then:

$$\big(f'(x) - f'(y), x - y\big) \leq L \|x - y\|_{\mathcal{H}}^2$$

for all $x, y \in \mathcal{H}$.

3. If f is μ-strongly convex and L-smooth, then $\mu \leq L$.

Proof. 1. Consider arbitrary $x, y \in \mathcal{H}$. Then, by the definition of strong convexity we have:

$$f(y) \geq f(x) + \big(f'(x), y - x\big) + \frac{\mu}{2} \|x - y\|_{\mathcal{H}}^2$$

and

$$f(x) \geq f(y) + \big(f'(y), x - y\big) + \frac{\mu}{2} \|x - y\|_{\mathcal{H}}^2 .$$

Adding the two inequalities thus gives:

$$0 \geq \big(f'(x) - f'(y), y - x\big) + \mu \|x - y\|_{\mathcal{H}}^2 ,$$

and therefore the result.

2. This is proven completely analogous to the first statement, using the bounds:

$$f(y) \leq f(x) + \big(f'(x), y - x\big) + \frac{L}{2} \|x - y\|_{\mathcal{H}}^2$$

and

$$f(x) \leq f(y) + \big(f'(y), x - y\big) + \frac{L}{2} \|x - y\|_{\mathcal{H}}^2$$

which follow from Lemma 2.5.

3. Follows directly from the other two claims.

□

The following property, sometimes referred to as *co-coercivity*, provides a similar bound as the first item in Proposition 2.6, but in dependence of L, and does not require strong convexity.

Proposition 2.7. *Suppose that f is L-smooth and convex (not necessarily strongly convex). Then, for all $x, y \in \mathcal{H}$ we have*

$$\big(f'(x) - f'(y), x - y\big) \geq \frac{1}{L} \big\|f'(x) - f'(y)\big\|_{\mathcal{H}^*}^2 \tag{2.3}$$

Proof. See [24, Lemma 2.29] for the finite dimensional case of [5, Theorem 18.15] for the infinite-dimensional case. □

Strongly convex functions (with $\mu > 0$) exhibit a unique minimizer, not only in the finite-dimensional setting:

Lemma 2.8. *Suppose that $f : \mathcal{H} \to \mathbb{R}$ is differentiable and μ-strongly convex for some $\mu > 0$. Then f has a unique minimizer $x^\star$.*

Proof. See [5, Corollary 11.17]. □

Remark 2.9. *Lemma 2.8 can be found, e.g. in [52, Theorem 2.2.6] as well. However, there the technique of the proof is confined to finite dimensions.*

An important consequence of strong convexity is the so-called Polyak-Lojasiewicz inequality: It provides a lower bound on the growth of the derivative in terms of the functional value and is commonly used to obtain linear convergence rates for gradient based optimization methods.

Lemma 2.10. *Suppose that $f : \mathcal{H} \to \mathbb{R}$ is differentiable and μ-strongly convex. Then:*

$$\left\| f'(x) \right\|_{\mathcal{H}^*}^2 \geq 2\mu \left(f(x) - f(x^\star) \right),$$

where $x^\star$ denotes the unique minimizer of f.

Proof. See, e.g., [24, Lemma 2.18] for a proof in the finite dimensional setting which directly can be applied to our setting. □

2.3 Gradients

Each Hilbert space with given inner product induces an isometric isomorphism $R : \mathcal{H} \to \mathcal{H}^*$, also known as the Riesz-Isomorphism. It is uniquely determined by the condition:

$$\phi(x) = (\phi, x) = \left\langle R^{-1}\phi, x \right\rangle \qquad \text{for all } \phi \in \mathcal{H}^*, x \in \mathcal{H}.$$

Iterative first order methods, like the (stochastic) gradient method require a search direction based on the first derivative of the target function F at x. This search direction is formally defined as the solution δ_x to the steepest descent problem

$$\min_{\delta_x \in \mathcal{H}} \left(F'(x), \delta_x \right) + \frac{1}{2} \left\| \delta_x \right\|_{\mathcal{H}}^2,$$

which is easily verified to be $\delta_x = -R^{-1}F'(x)$. The gradient is the direction of steepest ascent, so one has:

$$\nabla F(x) = R^{-1}F'(x) \in \mathcal{H}.$$

Thus, the gradient is the Riesz-Representer of the derivative. The following identities hold true:

$$\left\|F'(x)\right\|_{\mathcal{H}^*}^2 = \big(F'(x), \nabla F(x)\big) = \|\nabla F(x)\|_{\mathcal{H}}^2 .$$

2.4 Some Measure Theoretic Results

The probabilistic nature of stochastic gradient algorithms necessitates an explicit treatment of probability. This section introduces basic definitions and collects the probabilistic results used throughout this work. Unless stated otherwise, the Hilbert space $\mathcal{H}$ will be equipped with the Borel σ-Algebra $\mathcal{B}(\mathcal{H})$ generated by the collection of open sets in $\mathcal{H}$.

2.4.1 Products of Probability Spaces

Throughout this work we will consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, i.e. a set Ω, equipped with a σ-algebra $\mathcal{A}$ and a probability measure $\mathbb{P} : \mathcal{A} \to [0, 1]$. To model a sequence of events, one considers infinite products of probability spaces. We denote by

$$\Omega^{\mathbb{N}} := \{(\omega_n)_{n\in\mathbb{N}} \mid \omega_n \in \Omega \text{ for all } n \in \mathbb{N}\}$$

the set of all sequences of elements from Ω. $\Omega^{\mathbb{N}}$ can also be seen as the infinite Cartesian product

$$\Omega^{\mathbb{N}} = \Omega \times \Omega \times \cdots$$

By $\mathcal{A}^{\mathbb{N}}$ we denote the smallest σ-Algebra on $\Omega^{\mathbb{N}}$, such that all coordinate maps

$$p_i : \Omega^{\mathbb{N}} \to \Omega, \qquad (\omega_n)_{n\in\mathbb{N}} \mapsto \omega_i$$

are measurable. $\mathcal{A}^{\mathbb{N}}$ is generated by the system:

$$\{A_1 \times A_2 \times \cdots \mid A_n \in \mathcal{A} \text{ for all } n \in \mathbb{N}\},$$

see [38, Theorem 14.12 (i)], and is therefore also the smallest σ-Algebra on $\Omega^{\mathbb{N}}$, which contains all Cartesian products of measurable sets. By the well known Ionescu-Tulcea Theorem (see e.g. [38, Theorem 14.35]) there exists a unique probability measure $\mathbb{P}^{\mathbb{N}}$ on the measurable space $(\Omega^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}})$ that satisfies

$$\mathbb{P}^{\mathbb{N}}\left(A_1 \times \cdots \times A_n \times \Omega^{\mathbb{N}}\right) = \prod_{k=1}^{n} \mathbb{P}(A_k)$$

for all n and $A_1, \ldots, A_n \in \mathcal{A}$.

2.4.2 Robbins-Siegmund Lemma

An important tool to establish almost sure convergence results for stochastic optimization algorithms is the *Robbins-Siegmnund Lemma*, which was proven in [59]. Here, we will first state the result in its general form, and subsequently present how it is nowadays applied to the analysis of stochastic optimization algorithms. To state the result in its general form, we will first recall some basic properties from probability theory.

Conditional Expectations

We briefly recall the definition and important properties of the *conditional expectation*. For more details, we refer, e.g., to the textbook [38, Chapter 8]. Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, a σ-algebra $\mathcal{F} \subset \mathcal{A}$ and a random variable X on Ω.

Definition 2.11 ([38, Definition 8.11])**.** *A random variable Y is called* conditional expectation *of X given $\mathcal{F}$, if*

1. *Y is $\mathcal{F}$-measurable*
2. *For any set $A \in \mathcal{F}$, we have*

$$\int_A X \, \mathrm{d}\mathbb{P} = \int_A Y \, \mathrm{d}\mathbb{P}.$$

In this case we write $Y = \mathbb{E}[X \mid \mathcal{F}]$.

The following theorem states that there always exists a unique conditional expectation.

Theorem 2.12 ([38, Theorem 8.12])**.** *In the setting described above, $Y = \mathbb{E}[X \mid \mathcal{F}]$ exists and is unique up to equality almost surely.*

The following properties of the conditional expectation are of interest for us:

1. If X is $\mathcal{F}$-measurable, then $\mathbb{E}[X \mid \mathcal{F}] = X$.
2. The conditional expectation is linear in X
3. The conditional expectation is monotone: If $X_1 \leq X_2$ almost surely, then also $\mathbb{E}[X_1 \mid \mathcal{F}] \leq \mathbb{E}[X_2 \mid \mathcal{F}]$ almost surely.

The first of these claims is trivial, the other two can be found, e.g. in [38, Theorem 8.14].

Filtrations

Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and a sequence of σ-algebras $(\mathcal{F}_n)_{n\in\mathbb{N}}$ over Ω, such that

1. $\mathcal{F}_k \subset \mathcal{A}$ for all $k \in \mathbb{N}$ and
2. $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all $k \in \mathbb{N}$.

Such a sequence of σ-algebras is called a *filtration* (see, e.g., [38, Definition 9.9]).[1]

The result of Robbins and Siegmund [59] now reads:

Lemma 2.13. *Consider a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ and sequences of non-negative, real-valued random variables $(z_n)_{n\in\mathbb{N}}, (\beta_n)_{n\in\mathbb{N}}, (v_n)_{n\in\mathbb{N}}, (w_n)_{n\in\mathbb{N}}$, such that for each n, z_n, β_n, v_n and w_n are $\mathcal{F}_n$-measurable. Suppose that*

$$\mathbb{E}\left[z_{n+1} \mid \mathcal{F}_n\right] \leq z_n(1+\beta_n) + v_n - w_n,$$

and further that $\sum_{n=1}^{\infty} \beta_n < \infty$ and $\sum_{n=1}^{\infty} v_n < \infty$. Then, z_n converges almost surely to a finite limit and

$$\sum_{n=1}^{\infty} w_n < \infty \qquad \textit{almost surely.}$$

Application to Iterative Stochastic Algorithms

Lemma 2.13 allows us, as we will see later in Chapter 5, to prove almost sure convergence results for stochastic optimization algorithms. Iterative stochastic optimization algorithms generally are of the form

$$x_{n+1} = \varphi_{\omega_n}(x_n).$$

Thus, the next iterate x_{n+1} depends on the initial iterate x_0 and $\omega_0, \ldots, \omega_n \in \Omega$. If we fix x_0 and consider the sequence $\boldsymbol{\omega} = (\omega_n)_{n\in\mathbb{N}_0}$ as an element of the measurable space $\left(\Omega^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}}\right)$ and write

$$x_n = x_n(\boldsymbol{\omega}),$$

we can consider x_n as a mapping $\Omega^{\mathbb{N}} \to \mathcal{H}$. Now consider the σ-algebras $\mathcal{F}_n = \sigma(x_1, \ldots, x_n)$ over $\Omega^{\mathbb{N}}$ generated by the first n of these mappings. This is obviously a filtration in the measure space $\left(\Omega^{\mathbb{N}}, \bigotimes_{n\in\mathbb{N}} \mathcal{A}\right)$ and often referred to as the *filtration generated* by $(x_n)_{n\in\mathbb{N}}$ ([38, Chapter 17.1]) or the *natural filtration* ([39, Chapter 4.5]).

[1] The property $\mathcal{F}_k \subset \mathcal{A}$ is not necessary to define the notion of a filtration. If, as in this work, we seek to discuss probabilities of events from the σ-algebras $\mathcal{F}_k$, it is sensible to consider a common super-σ-algebra, where a suitable probability measure is defined.

Given a *quantity of interest*, i.e. a map $Q : \mathcal{H} \to \mathbb{R}$, we can consider $Q(x_{n+1})$ and compare it to $Q(x_n)$ in order to get an idea of the quality of the step from x_n to x_{n+1}. Often, something is known about the expectation of $Q(x_{n+1})$ with respect to only the last sample ω_n. If one considers $\omega_0, \dots, \omega_{n-1}$ fixed, this is just a scalar. However, if one considers this as a function of $\omega_0, \dots, \omega_{n-1}$, this is precisely the conditional expectation

$$\mathbb{E}\left[Q(x_{n+1}) \mid \mathcal{F}_n\right],$$

see Lemma 2.15 below. In our applications, we will have a bound of the form

$$Q(x_{n+1}) \leq Q(x_n) + \widetilde{Q}(\omega_n, x_n).$$

Using monotonicity and linearity of the conditional expectation, and the fact that the mapping $\boldsymbol{\omega} \mapsto Q(x_n)$ is $\mathcal{F}_n$-measurable, we see that:

$$\mathbb{E}\left[Q(x_{n+1}) \mid \mathcal{F}_n\right] \leq Q(x_n) + \mathbb{E}\left[\widetilde{Q}(\omega_n, x_n) \mid \mathcal{F}_n\right].$$

A bound of this form is useful, if further bounds or explicit expressions for the second addend on the right-hand side $\mathbb{E}\left[\widetilde{Q}(\omega_n, x_n) \mid \mathcal{F}_n\right]$ are available as it will be the case in our analysis. We conclude this small detour to probability theory with a result characterizing the conditional expectation in this particular setting.

Lemma 2.14. *Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and the corresponding product space $(\Omega^{\mathbb{N}}, \bigotimes_{n\in\mathbb{N}} \mathcal{A}, \mathbb{P}^{\mathbb{N}})$. Consider a Banach space $\mathcal{H}$ and a measurable mapping $f : \Omega^{\mathbb{N}} \to \mathcal{H}$, such that for some $n \in \mathbb{N}$, we have*

$$f(\boldsymbol{\omega}) = f_n(\omega_1, \dots, \omega_n)$$

for some $f_n : \Omega^n \to \mathcal{H}$. Then, if $A \in \sigma(f)$, it has the form:

$$A = A_n \times \Omega^{\mathbb{N}},$$

where $A_n \in \sigma(f_n)$.

Proof. It holds σ-algebra $\sigma(f) = \left\{f^{-1}(B) \mid B \in \mathcal{B}(\mathcal{H})\right\}$. Let $A \in \sigma(f)$. There exists a set $B \in \mathcal{B}(\mathcal{H})$ such that $A = f^{-1}(B)$. Since $f(\boldsymbol{\omega}) = f_n(\omega_1, \dots, \omega_n)$, we can write:

$$f^{-1}(B) = \{\boldsymbol{\omega} \in \Omega^{\mathbb{N}} : f_n(\omega_1, \dots, \omega_n) \in B\}.$$

Consequently, $\boldsymbol{\omega} \in A$ if and only if $(\omega_1, \dots, \omega_n) \in f_n^{-1}(B) =: A_n$ and therefore $A = A_n \times \Omega^{\mathbb{N}}$. □

Lemma 2.15. *Consider a measurable mapping $Q : \Omega \times \mathcal{H} \to \mathbb{R}$ and denote as above $\mathcal{F}_n = \sigma(x_1, \dots, x_n)$ as σ-algebra over $\Omega^{\mathbb{N}}$. Then:*

$$\mathbb{E}\left[\boldsymbol{\omega} \mapsto Q(\omega_n, x_n) \mid \mathcal{F}_n\right] = \boldsymbol{\omega} \mapsto \int_\Omega Q(\omega, x_n)\, \mathrm{d}\mathbb{P}(\omega) \tag{2.4}$$

Proof. Denote the random variable on the right-hand side with Y. Then, Y is obviously $\mathcal{F}_n$ measurable, as it depends on $\boldsymbol{\omega}$ only through x_n. Consider any $A \in \mathcal{F}_n$. Then, by Lemma 2.14, we have that:

$$\begin{aligned}\int_A Q(\omega_n, x_n)\,\mathrm{d}\mathbb{P}^{\mathbb{N}}(\boldsymbol{\omega}) &= \int_{A_n}\int_{\Omega^{\mathbb{N}}} Q(\omega_n, x_n)\,\mathrm{d}\mathbb{P}^{\mathbb{N}}(\omega_n, \omega_{n+1}, \dots)\,\mathrm{d}\mathbb{P}^k(\omega_0, \dots, \omega_{n-1}) \\ &= \int_{A_n}\int_{\Omega} Q(\omega, x_n)\,\mathrm{d}\mathbb{P}(\omega)\,\mathrm{d}\mathbb{P}(\omega_0, \dots, \omega_{n-1}) \\ &= \int_A Y\,\mathrm{d}\mathbb{P}^{\mathbb{N}}\end{aligned}$$

□

Lemma 2.15 shows that conditional expectations w.r.t. the so-called *natural filtration* $\mathcal{F}_n$ can be computed by just integrating over the single additional source of randomness ω_n. We could interpret the right-hand side of (2.4) in two ways: On the one hand, this is a mapping which depends on the first n entries of $\boldsymbol{\omega}$, namely $\omega_0, \dots, \omega_{n-1}$, through x_n. On the other hand, one could see this as scalar value, if one considers x_n to be a fixed quantity. In the convergence analysis of SGD, in particular in Chapters 4 and 5, we will use the second point of view for deriving certain bounds - but keep in mind that this corresponds to the conditional expectation w.r.t. to the natural filtration, and therefore results like the Robbins-Siegmund-Lemma Lemma 2.13 can be applied.

2.4.3 Transfer of Convergence Rates

Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a sequence of non-negative, real valued random variables $(X_n)_{n\in\mathbb{N}}$. In many cases we will derive convergence rates of the form

$$\mathbb{E}\left[X_n\right] \leq b_n, \tag{2.5}$$

where, for each n, b_n is a suitable, deterministic bound satisfying $b_n \to 0$, for $n \to \infty$. Clearly, this bound gives information about the mean $\mathbb{E}\left[X_n\right]$. However, it also provides information about the *trajectories* $X_n(\omega)$ for almost every $\omega \in \Omega$, as the following simple consequence of the Borel-Cantelli lemma (see [38, Theorem 2.7]) shows:

Theorem 2.16. *Suppose* (2.5) *holds for a sequence of non-negative random variables* $(X_n)_{n\in\mathbb{N}}$. *Suppose further, that* $(s_n)_{n\in\mathbb{N}}$ *is a sequence of positive numbers, such that*

$$\sum_{n=1}^{\infty} \frac{1}{s_n} < \infty$$

Then, for almost every $\omega \in \Omega$ *there is* $N = N(\omega)$, *such that*

$$X_n(\omega) \leq s_n b_n \qquad \text{for all } n \geq N.$$

Proof. For $n \in \mathbb{N}$ consider the sets

$$A_n := \{\omega \in \Omega \mid X_n > s_n b_n\}$$

These sets can be written as:

$$A_n = X_n^{-1}((s_n b_n, \infty))$$

and are thus measurable, i.e. $A_n \in \mathcal{A}$. We have

$$\mathbb{P}(A_n) = \int_{A_n} 1 \, \mathrm{d}\mathbb{P} \leq \frac{1}{s_n b_n} \int_{A_n} X_n \, \mathrm{d}\mathbb{P} \leq \frac{1}{s_n b_n} \int_{\Omega} X_n \, \mathrm{d}\mathbb{P} \leq \frac{1}{s_n}.$$

Thus:

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$$

By the Borel-Cantelli lemma (see [38, Theorem 2.7]) we thus get:

$$\mathbb{P}(A) = 0, \qquad \text{where } A = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n.$$

Consequently, $\mathbb{P}(A^c) = 1$. We have:

$$\omega \in A \iff \forall\, m \in \mathbb{N}\, \exists\, n \geq m : X_n(\omega) > s_n b_n$$

And thus:

$$\omega \in A^c \iff \exists\, m \in \mathbb{N}\, \forall\, n \geq m : X_n(\omega) \leq s_n b_n,$$

which concludes the proof. □

2.4.4 Tightness

The concept of tightness and the properties presented here will be used to show the existence of invariant measures in Section 3.4.

Definition 2.17. *Consider a set of probability measures $\mathcal{M}$ on some metric space $\mathcal{H}$. $\mathcal{M}$ is called* tight*, if, for any $\varepsilon > 0$, there is a compact set $K_\varepsilon \in \mathcal{H}$, such that for any $\mu \in \mathcal{M}$ it holds*

$$\mu(K_\varepsilon) \geq 1 - \varepsilon.$$

Remark 2.18. *The concept of tightness as defined in Definition 2.17 is referred to as* uniform tightness *in the reference [10, Defintion 8.6.1].*

Tight sets of measures are particular useful, as they are precompact in the topology of weak convergence, as Prokhorov's theorem shows:

Theorem 2.19. *Consider a complete, separable metric space $\mathcal{H}$ and let $\mathcal{M}$ be a tight set of probability measures on $\mathcal{H}$. Then, for any sequence $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{M}$, there is a subsequence n_k and a probability measure μ, such that*

$$\int_{\mathcal{H}} f \,\mathrm{d}\mu_{n_k} \to \int_{\mathcal{H}} f \,\mathrm{d}\mu$$

for any bounded and continuous function $f : \mathcal{H} \to \mathbb{R}$.

Proof. This is a simplified version of the more general result [10, Theorem 8.6.2]. See also the original work in [58]. □

Every singleton, and consequently every finite set of probability measures, is tight:

Lemma 2.20. *Consider a complete, separable metric space $\mathcal{H}$ and a probability measure μ on $\mathcal{H}$. Then, for every $\varepsilon > 0$ there is a compact set $K_\varepsilon \subset \mathcal{H}$, such that*

$$\mu(K_\varepsilon) \geq 1 - \varepsilon.$$

Proof. Again, this is a consequence of [10, Theorem 8.6.2], see also [27, Lemma 4.14]. □

3 Stochastic Gradients

The stochastic gradient descent algorithm is a natural and popular choice for the numerical solution of stochastic optimization problems. Although more than 70 years have passed since it was introduced in the 1950s in [60], it is nowadays the foundation of the driving optimization algorithms behind the success of modern machine learning applications. This is mainly due to its simplicity, which makes it well-suited to such problems and allows for many modification and extensions, which have been proposed in the last decades. In this chapter, we will define the class of stochastic optimization problems (SOPs) in Definition 3.1 and relevant subclasses in Definition 3.2. Based on the problem definition, we will present the stochastic gradient descent algorithm (SGD) in Section 3.2. We will discuss commonly used bounds on the variance in the search direction and their asymptotic behavior for diminishing strong convexity. Leveraging these bounds, in Section 3.3 we present some well known convergence properties of the stochastic gradient algorithm applied to such SOPs, when either constant step sizes, or step sizes satisfying a certain decay condition are employed. Finally, in Section 3.4 we focus on the long-term behavior of the iterates of SGD with constant step sizes. The long-term behavior can be described by a probability measure on the space of iterates $\mathcal{H}$, which will play a crucial role in the convergence analysis in Chapter 7.

3.1 Stochastic Optimization Problems

In this section, we will formally define stochastic optimization problems and illustrate their occurrence in machine learning applications.

Definition 3.1. *Suppose that* $(\Omega, \mathcal{A}, \mathbb{P})$ *is a probability space. Consider a mapping*

$$f : \mathcal{H} \times \Omega \to \mathbb{R},$$

such that

1. *The mapping* $x \mapsto f_\omega(x) := f(x, \omega)$ *is continuously differentiable, i.e. there is a continuous mapping*

$$f'_\omega : \mathcal{H} \to \mathcal{H}^*,$$

such that

$$\lim_{n\to\infty} \frac{f_\omega(x+h_n) - f_\omega(x) - (f'_\omega(x), h_n)}{\|h_n\|_{\mathcal{H}}} = 0$$

for any $(h_n)_{n\in\mathbb{N}} \subset \mathcal{H}$ *with* $h_n \to 0$.

2. *The mapping* $\omega \mapsto f(x,\omega)$ *is* $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$ *measurable and integrable for any* $x \in \mathcal{H}$.

Define the mean function F *by:*

$$F(x) := \mathbb{E}_\omega\left[f_\omega(x)\right] = \int_\Omega f(x,\omega)\,\mathrm{d}\mathbb{P}(w).$$

Suppose that F *is continuously differentiable and for any* $\delta_x \in \mathcal{H}$ *we have*

$$\left(F'(x), \delta_x\right) = \int_\Omega \left(f'_\omega(x), \delta_x\right)\,\mathrm{d}\mathbb{P}(\omega).$$

Then, the problem of finding $x^\star \in \arg\min_{x\in\mathcal{H}} F(x)$ *is called* ***stochastic optimization problem*** *(SOP) induced by* f. *We will refer to this problem as* $\mathrm{SOP}(f)$.

Obviously, the problem in this form is not always well posed, as we might have $\arg\min_{x\in\mathcal{H}} F(x) = \emptyset$. To exclude such cases, and to even obtain a uniquely solvable problem, we impose the following set of assumptions.

Definition 3.2. *Consider a stochastic optimization problem* $\mathrm{SOP}(f)$ *induced by* $f : \mathcal{H}\times\Omega\to\mathbb{R}$. *Consider* $0 \le \mu \le L < \infty$.

1. $\mathrm{SOP}(f)$ *is said to be* (μ, L)*-feasible, if its mean function* F *is* (μ, L)*-feasible (see Definition 2.2).*

2. $\mathrm{SOP}(f)$ *is said to be strongly* (μ, L)*-feasible, if* f_ω *is* (μ, L)*-feasible for* $\mathbb{P}$*-almost every* $\omega \in \Omega$.

3. *Suppose there are measurable functions* $\mu : \Omega \to \mathbb{R}_{\ge 0}, \omega \mapsto \mu_\omega$ *and* $L : \Omega \to \mathbb{R}_{\ge 0}, \omega \mapsto L_\omega$, *such that*

   - f_ω *is* (μ_ω, L_ω)*-feasible for* $\mathbb{P}$*-almost every* $\omega \in \Omega$ *and*
   - $L_{\max} := \operatorname{ess\,sup}_{\omega\in\Omega} L_\omega < \infty$.

   Then $\mathrm{SOP}(f)$ *is said to be pointwise* (μ_ω, L_ω)*-feasible.*

The following result summarizes immediate consequences of the definition.

Lemma 3.3. *Consider an SOP induced by* $f : \mathcal{H}\times\Omega\to\mathbb{R}$. *Then:*

1. *If* $\mathrm{SOP}(f)$ *is strongly* (μ, L)*-feasible, then* $\mathrm{SOP}(f)$ *is also pointwise* (μ_ω, L_ω)*-feasible with* $\mu_\omega = \mu$ *and* $L_\omega = L$.

2. *If* $\mathrm{SOP}(f)$ *is pointwise* (μ_ω, L_ω)*-feasible, then* $\mathrm{SOP}(f)$ *is strongly* (μ, L)*-feasible with* $\mu = \operatorname{ess\,inf}_{\omega\in\Omega} \mu_\omega$ *and* $L = L_{\max}$.

3. *If* $\mathrm{SOP}(f)$ *is pointwise* (μ_ω, L_ω)*-feasible, then* $\mathrm{SOP}(f)$ *is* (μ, L)*-feasible with*

$$\mu = \int_\Omega \mu_\omega \,\mathrm{d}\mathbb{P}(\omega) \geq 0 \qquad \text{and} \qquad L = \int_\Omega L_\omega \,\mathrm{d}\mathbb{P}(\omega) \leq L_{\max}.$$

4. *If* $\mathrm{SOP}(f)$ *is strongly* (μ, L)*-feasible, then* $\mathrm{SOP}(f)$ *is* (μ, L)*-feasible.*

Proof. Items 1 and 2 are clear from definition. To show Item 3 we note that

$$\begin{aligned} F(y) &= \int_\Omega f_\omega(y) \,\mathrm{d}\mathbb{P}(\omega) \\ &\geq \int_\Omega f_\omega(x) + \big(f'_\omega(x), y - x\big) + \frac{\mu_\omega}{2} \|x - y\|^2_{\mathcal{H}} \,\mathrm{d}\mathbb{P}(\omega) \\ &= F(x) + \big(F'(x), y - x\big) + \frac{\mu}{2} \|x - y\|^2_{\mathcal{H}}, \end{aligned}$$

showing μ-strong convexity of F. A similar computation using the characterization in Lemma 2.5 shows that F satisfies (2.2) with $L = \mathbb{E}_\omega[L_\omega]$ and is thus L-smooth. Item 4 in turn follows from Items 1 and 3. □

SOPs in Machine Learning

Stochastic Optimization Problems play a crucial role in modern machine learning. A classical task is to fit a parameter-dependent model to a distribution of observations. Consider some space of parameters Θ, a space of inputs X and a space of outputs Y. For every $\theta \in \Theta$, h_θ is a mapping $h_\theta : X \to Y$. Suppose that inputs and outputs follow a joined probability distribution $\mathbb{P}$ on $X \times Y$. Then, the task is to find a solution to

$$\min_{\theta\in\Theta} R(\theta) = \int_{X\times Y} \ell(h_\theta(x), y) \,\mathrm{d}\mathbb{P}(x, y). \tag{3.1}$$

Here, $\ell : Y \times Y \to \mathbb{R}$ is the so called *loss*-function, describing how close the *prediction* $h_\theta(x)$ is to the desired output y. A common choice for regression problems is $\ell(y_1, y_2) = \|y_1 - y_2\|^2_Y$, if Y is a normed space. For classification tasks, other loss functions are used, for example the so-called cross-entropy loss function, see e.g. [8, Section 4.3.2]. See also [21] for a recent overview of commonly used loss functions. This problem fits in our framework of stochastic optimization problems by selecting $\Omega = X \times Y$ with probability measure $\mathbb{P}$ and $f_\omega(\theta) = \ell(h_\theta(x), y)$ for $\omega = (x, y)$. However, the resulting functions are usually non-convex and, depending on the precise architecture of the model h_θ, might even be non-differentiable.

The function R in (3.1) is often referred to as the *expected risk* ([11, Section 2.3], [26, Chapter 8]). Usually, the distribution $\mathbb{P}$ is unknown, and only a finite number of samples (x_i, y_i), $i \in [1 : N]$ are drawn from the distribution $\mathbb{P}$. In this case, instead of solving (3.1), one solves the *finite sum* optimization problem

$$\min_{\theta \in \Theta} \widehat{R}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(h_\theta(x_i), y_i). \tag{3.2}$$

Equation (3.2) can be seen as a Monte Carlo approximation to (3.1). The target function $\widehat{R}$ is also referred to as the *empirical risk* ([11, Section 2.3], [26, Chapter 8]). This problem formulation also fits in the framework of stochastic optimization as presented above: One chooses $\Omega = [1 : N]$ with the uniform distribution as probability measure and $f_\omega(\theta) = \ell(h_\theta(x_i), y_i)$ for $\omega = i$.

Mini-batching

The iterative schemes to solve SOPs we will discuss below will require to sample one $\omega \sim \mathbb{P}$ in each iteration. For the sake of computational efficiency and stability, in practice often a certain number M (the *batch size*) of samples is drawn and the mean of these M samples is used as the sampled function. This approach can also be regarded as a classical SOP in the context we have presented: One chooses $\widehat{\Omega} = \Omega^M$ with $\widehat{\mathbb{P}} = \mathbb{P}^M$ and $f_{\widehat{\omega}}^{(M)} = \frac{1}{M} \sum_{j=1}^{M} f_{\omega_j}$. Consequently, we will not consider mini-batching explicitly in this work. Rather, it is contained in our results as a special case, when the distribution is chosen correctly. Note that mini-batching leaves the target function F unchanged.

3.2 Stochastic Gradient Descent

Let us now consider a simple, iterative numerical scheme, tailored to solve stochastic optimization problems. This scheme, usually referred to as the Stochastic Gradient method or Stochastic Gradient Descent (SGD), is a direct adaptation of the classical Gradient Descent method to the stochastic setting. Classical gradient methods to solve the optimization problem

$$\min_{x \in \mathcal{H}} \; F(x)$$

for smooth F use an update of the form

$$x^+ = x - \alpha \nabla F(x),$$

where $\alpha > 0$ is some positive step size. In the case of a stochastic optimization problem, where the target function F is of the form

$$F(x) = \int_\Omega f_\omega(x) \, \mathrm{d}\mathbb{P}(\omega),$$

the gradient $\nabla F(x)$ is usually not computable or too expensive to evaluate. Instead, it is assumed that *sampling* from the distribution $\mathbb{P}$ is possible. The stochastic gradient descent algorithms leverages this by sampling $\omega \sim \mathbb{P}$ in each iteration and updating

$$x^+ = x - \alpha \nabla f_\omega(x). \tag{3.3}$$

Consequently, x^+ is a random variable, depending on the previous iterate x, the step size α and ω. In Algorithm 1 the plain SGD algorithm is given in pseudocode.

Algorithm 1: Stochastic Gradient Descent (SGD)

Input: SOP induced by $f : \mathcal{H} \times \Omega \to \mathbb{R}$. Initial iterate x_0. Sequence of step sizes $(\alpha_n)_{n\in\mathbb{N}}$

for $n \geq 0$ **do**

  Sample $\omega_n \sim \mathbb{P}$

  $x_{n+1} \leftarrow x_n - \alpha_n \nabla f_{\omega_n}(x_n)$

end

3.2.1 Uncertainty in Search Directions

The main difference to the classical gradient descent algorithm is the uncertainty in the search directions introduced by the sampling in the stochastic gradient method. It is usually quantified by

$$\mathbb{V}_\omega \left[f'_\omega(x) \right] := \mathbb{E}_\omega \left[\left\| f'_\omega(x) - F'(x) \right\|^2_{\mathcal{H}^*} \right] = \mathbb{E}_\omega \left[\left\| \nabla f_\omega(x) - \nabla F(x) \right\|^2_{\mathcal{H}} \right].$$

Although denoted (and also referred to in this work) as the variance, this quantity is, strictly speaking, the trace of the corresponding covariance. An important property is

$$\mathbb{V}_\omega \left[f'_\omega(x) \right] = \mathbb{E}_\omega \left[\left\| f'_\omega(x) \right\|^2_{\mathcal{H}^*} \right] - \left\| F'(x) \right\|^2_{\mathcal{H}^*} = \mathbb{E}_\omega \left[\left\| \nabla f_\omega(x) \right\|^2_{\mathcal{H}} \right] - \left\| \nabla F(x) \right\|^2_{\mathcal{H}}. \tag{3.4}$$

The variance is a function of the current state x. It is a deterministic quantity, if x is deterministic. The variance, and especially its behavior at the potential limit of stochastic gradient descent (SGD), plays a central role in step-size selection.

Definition 3.4. *Consider an SOP with a unique minimizer $x^\star$. By V_0, we denote the variance at the minimizer, i.e.*

$$V_0 := \mathbb{V}_\omega \left[f'_\omega(x^\star) \right].$$

Consider a SOP with unique minimizer $x^\star$. If $V_0 > 0$, we have

$$\mathbb{E}_\omega \left[\left\| f'_\omega(x^\star) \right\|^2 \right] > 0,$$

implying that $f'_\omega(x^\star) \neq 0$ occurs with positive probability. Thus, even if the SGD algorithm would exactly reach the minimizer $x^\star$, almost surely, in some later iteration (assuming SGD is run for infinitely many iterations) a non-zero search direction would be chosen, and the algorithm would divert from the minimizer. This highlights two key challenges in controlling the step size in SGD. First one has to deal with the fact that stationary points (of the true target function) are not stationary points of the algorithm as well. Second, in general it is hard to detect that the algorithm has reached a stationary point or is close to one. A good step size strategy for SGD has to deal with these uncertainties, and will still yield convergence results.

The following definition formalizes the setting described above.

Definition 3.5. *Consider an SOP with a unique minimizer $x^\star$. We will refer to the problem as belonging to the*

- interpolating setting, *if $V_0 = 0$.*
- non-interpolating setting, *if $V_0 > 0$.*

Thus, the discussion above deals with the non-interpolating setting. Problems in this setting are usually much harder to solve and require a careful selection of step sizes due to the difficulties mentioned.

It is worth noting that a global minimizer $x^\star$ is not necessarily also a minimizer of the variance, i.e. there might be $\bar{x} \in \mathcal{H}$ such that $\mathbb{V}_\omega\left[f'_\omega(\bar{x})\right] < V_0$. Clearly, this can only be the case in the non-interpolating setting. To illustrate this, we consider a simple example:

Example 3.6. *Let $f_\omega(x) := \omega x^2 + x$, where ω is uniformly distributed on $\Omega = [1, 2]$ and $x \in \mathcal{H} = \mathbb{R}$. The mean function F is therefore given by $F(x) = \frac{3}{2}x^2 + x$. Obviously, this SOP is strongly $(2, 4)$-feasible and exhibits a unique minimizer $x^\star = -\frac{1}{3}$. We obtain for the variance at the minimizer:*

$$V_0 = \int_1^2 \left(-\frac{2}{3}\omega + 1\right)^2 \mathrm{d}\omega = \frac{1}{27}.$$

However, for $\bar{x} = 0$, we have $f'_\omega(\bar{x}) = 1$ for every $\omega \in [1, 2]$, therefore $\mathbb{V}_\omega\left[f'_\omega(\bar{x})\right] = 0$.

Mini-batching (see the short discussion at the end of Section 3.1) is often employed to reduce the variance of the problem. While it can reduce the variance, the interpolating property is invariant under the incorporation of mini-batching:

Proposition 3.7. *Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, a normed space $\mathcal{H}$ and an SOP induced by $f : \mathcal{H} \times \Omega \to \mathbb{R}$ with unique minimizer $x^\star$. Further, consider $M \in \mathbb{N}$ and the*

SOP induced by

$$f^{(M)} : \mathcal{H} \times \Omega^M \to \mathbb{R}$$
$$(x, \omega_1, \ldots, \omega_M) \mapsto \frac{1}{M} \sum_{i=1}^{M} f_{\omega_i}(x).$$

Then, $\mathrm{SOP}(f^{(M)})$ *is interpolating, if and only if* $\mathrm{SOP}(f)$ *is interpolating.*

Proof. $x^\star$ is the unique solution of $\mathrm{SOP}(f^{(M)})$ as well. Therefore, the well known property of that averaging scales the variance, in our context

$$\mathbb{V}_{\omega_1,\ldots,\omega_M} \left[{f^{(M)}_{\omega_1,\ldots,\omega_M}}'(x^\star) \right] = \frac{1}{M} \mathbb{V}_\omega \left[f'_\omega(x^\star) \right],$$

implies the result. □

3.2.2 Variance Bounds

It will become evident later, that the variance is a crucial ingredient to step size control in SGD. Therefore, several approaches exist in the literature to bound the variance a-priori. One common approach is found in [11] and bounds the variance in terms of a constant $\widetilde{V}_0$ and allows for growth proportional to $\|F'(x)\|^2_{\mathcal{H}^*}$:

$$\mathbb{V}_\omega \left[f'_\omega(x) \right] \leq \widetilde{V}_0 + V_1 \left\| F'(x) \right\|^2_{\mathcal{H}^*}. \tag{3.5}$$

If there is a unique global minimizer $x^\star$, then $F'(x^\star) = 0$, and therefore $\widetilde{V}_0$ describes the noise at the minimizer, similar to our constant V_0 from Definition 3.4. However, $\widetilde{V}_0$ might differ from V_0. Another approach replaces $\|F'(x)\|^2_{\mathcal{H}^*}$ in (3.5) by $\Delta_x := F(x) - F(x^\star)$ and $\mathbb{V}_\omega \left[f'_\omega(x) \right]$ by $\mathbb{E}_\omega \left[\|f'_\omega(x)\|^2_{\mathcal{H}^*} \right] = \mathbb{V}_\omega \left[f'_\omega(x) \right] + \|F'(x)\|^2_{\mathcal{H}^*}$:

$$\mathbb{E}_\omega \left[\left\| f'_\omega(x) \right\|^2_{\mathcal{H}^*} \right] \leq \widetilde{V}_0 + V_1 \Delta_x. \tag{3.6}$$

If F is convex and differentiable, (3.6) implies

$$\mathbb{E}_\omega \left[\left\| f'_\omega(x) \right\|^2_{\mathcal{H}^*} \right] \leq \widetilde{V}_0 + V_1 \left(F'(x), x - x^\star \right) \tag{3.7}$$

for any stationary point $x^\star$ of F. In many cases the variance bounds (3.5) and (3.6) follow from convexity and smoothness assumptions, as Propositions 3.8 and 3.9 below show.

Proposition 3.8 (Conditions for (3.5))**.**

a) Consider a (μ, L)-feasible SOP ($\mu > 0$) such that f_ω is L_ω-smooth for some square-integrable function $\omega \mapsto L_\omega$. Then the variance assumption (3.5) holds with

$$\widetilde{V}_0 = 2\,\mathbb{E}_\omega\left[\left\|f'_\omega(x^\star)\right\|_{\mathcal{H}^*}^2\right] \quad \textit{and} \quad V_1 = 2\,\frac{\mathbb{E}_\omega\left[L_\omega^2\right]}{\mu^2} - 1.$$

b) Consider a pointwise (μ_ω, L_ω)-feasible SOP such that F is μ-strongly convex for some $\mu > 0$. Then the variance assumption (3.5) holds with

$$\widetilde{V}_0 = 2\,\mathbb{E}_\omega\left[\left\|\nabla f_\omega(x^\star)\right\|_{\mathcal{H}}^2\right] \quad \textit{and} \quad V_1 = 2\,\frac{L_{\max}}{\mu} - 1.$$

Proof. a) First, we compute:

$$\begin{aligned}
\left\|f'_\omega(x) - F'(x)\right\|_{\mathcal{H}^*}^2 &= \left\|f'_\omega(x) - f'_\omega(x^\star)\right\|_{\mathcal{H}^*}^2 \\
&\quad + 2\left\langle f'_\omega(x) - f'_\omega(x^\star), f'_\omega(x^\star) - F'(x^\star)\right\rangle_{\mathcal{H}^*} \\
&\quad + 2\left\langle f'_\omega(x) - f'_\omega(x^\star), F'(x^\star) - F'(x)\right\rangle_{\mathcal{H}^*} + \left\|f'_\omega(x^\star) - F'(x^\star)\right\|_{\mathcal{H}^*}^2 \\
&\quad + 2\left\langle f'_\omega(x^\star) - F'(x^\star), F'(x^\star) - F'(x)\right\rangle_{\mathcal{H}^*} + \left\|F'(x^\star) - F'(x)\right\|_{\mathcal{H}^*}^2 \\
&\leq 2\left\|f'_\omega(x) - f'_\omega(x^\star)\right\|_{\mathcal{H}^*}^2 + 2\left\|f'_\omega(x^\star)\right\|_{\mathcal{H}^*}^2 + \left\|F'(x)\right\|_{\mathcal{H}^*}^2 \\
&\quad - 2\left\langle f'_\omega(x) - f'_\omega(x^\star), F'(x)\right\rangle_{\mathcal{H}^*} - 2\left\langle f'_\omega(x^\star), F'(x)\right\rangle_{\mathcal{H}^*}.
\end{aligned}$$

Thus, taking the expectation yields:

$$\begin{aligned}
\mathbb{V}_\omega\left[f'_\omega(x)\right] &\leq 2\,\mathbb{E}_\omega\left[L_\omega^2\right]\|x - x^\star\|_X^2 + 2\,\mathbb{E}_\omega\left[\left\|f'_\omega(x^\star)\right\|_{\mathcal{H}^*}^2\right] - \left\|F'(x)\right\|_{\mathcal{H}^*}^2 \\
&\leq \left(2\,\frac{\mathbb{E}_\omega\left[L_\omega^2\right]}{\mu^2} - 1\right)\left\|F'(x)\right\|_{\mathcal{H}^*}^2 + 2\,\mathbb{E}_\omega\left[\left\|f'_\omega(x^\star)\right\|_{\mathcal{H}^*}^2\right].
\end{aligned}$$

b) As in the proof of Item a), we get

$$\mathbb{V}_\omega\left[f_\omega(x)\right] \leq 2\,\mathbb{E}_\omega\left[\left\|f'_\omega(x) - f'_\omega(x^\star)\right\|_{\mathcal{H}^*}^2\right] + 2\,\mathbb{E}_\omega\left[\left\|f'_\omega(x^\star)\right\|_{\mathcal{H}^*}^2\right] - \left\|F'(x)\right\|_{\mathcal{H}^*}^2.$$

As $L_\omega \leq L_{\max}$ holds for almost every ω, we can use the following bound, which can be found in [35, Equation (8)][1]:

$$\mathbb{E}_\omega\left[\left\|f'_\omega(x) - f'_\omega(x^\star)\right\|_X^2\right] \leq 2\,L_{\max}\,(F(x) - F(x^\star)).$$

Strong convexity yields $2\,\mu\,(F(x) - F(x^\star)) \leq \|F'(x)\|_{\mathcal{H}^*}^2$ (see Lemma 2.10), and we thus obtain:

$$\mathbb{E}_\omega\left[\left\|f'_\omega(x) - f'_\omega(x^\star)\right\|_X^2\right] \leq \frac{L_{\max}}{\mu}\left\|F'(x)\right\|_{\mathcal{H}^*}^2.$$

[1] The authors in [35] establish the bound in the case $\mathcal{H} = \mathbb{R}^d$. Their proof can directly be carried over to our setting.

We conclude for the variance:

$$\mathbb{V}_\omega \left[f'_\omega(x) \right] \leq 2 \left(\frac{L_{\max}}{\mu} - 1 \right) \left\| F'(x) \right\|_{\mathcal{H}^*}^2 + 2\, \mathbb{E}_\omega \left[\left\| f'_\omega(x^\star) \right\|_{\mathcal{H}^*}^2 \right]$$

□

Proposition 3.9 (Conditions for (3.6))**.** *Consider a pointwise (μ_ω, L_ω)-feasible SOP and a global minimizer $x^\star$ of F. For $x \in \mathcal{H}$ it holds:*

$$\begin{aligned} \mathbb{E}_\omega \left[\left\| f'_\omega(x) \right\|_{\mathcal{H}^*}^2 \right] &\leq 4\, L_{\max}\, \Delta_x + 2\, V_0 \\ &\leq 4\, L_{\max}\, \left(F'(x), x - x^\star \right) + 2\, V_0 \end{aligned}$$

and thus

$$\begin{aligned} \mathbb{V}_\omega \left[f'_\omega(x) \right] &\leq 4\, L_{\max}\, \Delta_x + 2\, V_0 - \left\| F'(x) \right\|_{\mathcal{H}^*}^2 \\ &\leq 4\, L_{\max}\, \left(F'(x), x - x^\star \right) + 2\, V_0 - \left\| F'(x) \right\|_{\mathcal{H}^*}^2 . \end{aligned}$$

Proof. The bound

$$\mathbb{E}_\omega \left[\left\| f'_\omega(x) \right\|_{\mathcal{H}^*}^2 \right] \leq 4\, L_{\max}\, \Delta_w + 2\, V_0$$

can be found in [24, Lemma 4.20], see also [35, Equation 8]. In these references, the authors consider the finite sum setting and the case $X = \mathbb{R}^d$, equipped with the standard Euclidean inner product. Since their arguments apply directly to our setting, we omit the detailed proof. Due to convexity we have $\Delta_w \leq (F'(x), w - w^\star)$, which implies the second bound. The bound on the variance follows from $\mathbb{V}_\omega \left[f'_\omega(x) \right] = \mathbb{E}_\omega \left[\| f'_\omega(x) \|_{\mathcal{H}^*}^2 \right] - \| F'(x) \|_{\mathcal{H}^*}^2$. □

Asymptotic Behavior of the Variance Bounds

The following discussion is mostly taken from our paper [41].

Propositions 3.8 and 3.9 show that variance bounds (3.5) and (3.6) can be deduced from smoothness and convexity assumptions. The constants introduced by Proposition 3.9 do not depend on the strong convexity constant μ, and it is easy to observe that they can't be significantly improved: Simple examples show that for L-smooth F, asymptotically at infinity, a factor of L is necessary between $\| F'(x) \|^2$ and $(F'(x), x - x^\star)$. On the other hand, the constants introduced by Proposition 3.8 contain the factor $\frac{1}{\mu}$. In light of the convergence results in Section 3.3.1, in particular Proposition 3.14, this would require step sizes to depend on the strong-convexity constant μ. This seems unnecessary and,

as a numerical example will show in Example 3.17, leads to drastically reduced speed of convergence, i.e. to too conservative step sizes. However, the bound itself can't be improved, the factor $\frac{1}{\mu}$ is necessarily introduced by the term in which we measure the growth of the variance, i.e. $\|F'(x)\|_{\mathcal{H}^*}^2$ in (3.5), as the following results show. In both (Propositions 3.11 and 3.12), we consider the smallest possible constant V_1, for which the bound (3.5) is satisfied for a given $\widetilde{V}_0$.

Definition 3.10. *Given a stochastic optimization problem* SOP(f) *following Definition 3.1 and* $\widetilde{V}_0 \geq 0$*, let*

$$V_1(\widetilde{V}_0) := \sup \left\{ \frac{\mathbb{V}_\omega \left[\nabla f_\omega(x)\right] - V_0}{\|\nabla F(x)\|_X^2} \,\middle|\, \mathbb{V}_\omega \left[\nabla f_\omega(x)\right] > \widetilde{V}_0,\ x \neq x^\star \right\}$$

denote the smallest possible constant V_1 *such that the variance assumption* (3.5) *is met.*

Proposition 3.11. *Suppose that* $\mathcal{P}(\mu, L)$ *is the set of all* (μ, L)*-feasible stochastic optimization problems* SOP(f)*. Then for any* $\mu \in (0, 1)$ *we have*

$$\sup_{\mathrm{SOP}(f) \in \mathcal{P}(\mu,1)} \inf_{\widetilde{V}_0 \in \mathbb{R}} V_1(\widetilde{V}_0) = \infty.$$

Proof. For $\gamma > 0$ and $\beta > 2$, let $\mathrm{Par}(\gamma, \beta)$ be the Pareto distribution with parameters γ and β. When $\omega \sim \mathrm{Par}(\gamma, \beta)$, then we have $\mathbb{E}_\omega [\omega] = \gamma \frac{\beta}{\beta-1}$ and $\mathbb{V}_\omega [\omega] = \gamma^2 \frac{\beta}{(\beta-2)(\beta-1)^2}$; see, e. g., [44, Chapter 23]. Thus with the choice $A_\omega := \begin{pmatrix} \omega & 0 \\ 0 & 1 \end{pmatrix}$ and $f_\omega(x) := \frac{1}{2} x^T A_\omega x$, f_ω is μ_ω-strongly convex and L_ω-smooth with $\mu_\omega = \min(\omega, 1) \geq \min(\omega, \mu)$, and $L_\omega = \max(\omega, 1)$. By definition, we have

$$F(x) = \frac{1}{2} x^T \begin{pmatrix} \gamma \frac{\beta}{\beta-1} & 0 \\ 0 & 1 \end{pmatrix} x.$$

Consequently, F is $\gamma \frac{\beta}{\beta-1}$-strongly convex. For $\gamma = \mu$ and $\beta = 2 + \varepsilon$ with arbitrary but sufficiently small ε we get that F is μ-strongly convex, and thus the corresponding stochastic optimization problem is $(\mu, 1)$-feasible.

Further, choosing $x = \begin{pmatrix} s \\ 0 \end{pmatrix}$ with some scaling parameter s, we observe

1. $\mathbb{V}_\omega [f'_\omega(x)] = s^2 \mu^2 \frac{\beta}{(\beta-2)(\beta-1)^2} = s^2 \mu^2 \frac{2+\varepsilon}{\varepsilon(1+\varepsilon)}$.
2. $\|F'(x)\|_X^2 = s^2 \mu^2 \frac{\beta^2}{(\beta-1)^2} = s^2 \mu^2 \left(\frac{2+\varepsilon}{1+\varepsilon}\right)^2$.

Thus, selecting $s = 2\,\frac{\varepsilon(1+\varepsilon)}{\mu\widetilde{V}_0(2+\varepsilon)}$, we obtain w with $\mathbb{V}_\omega\left[F'_\omega(x)\right] > \widetilde{V}_0$ and

$$\frac{\mathbb{V}_\omega\left[f'_\omega(x)\right] - \widetilde{V}_0}{\|F'(x)\|_X^2} \geq \frac{1}{4\,\varepsilon}.$$

Since ε was arbitrary, this proves the result. □

Thus, the constant V_1 in bounds of the type of (3.5) can become arbitrarily large for certain distributions. In the proof of Proposition 3.11, we used a heavy-tailed distribution to let $\mathbb{V}_\omega\left[L_\omega\right]$ grow arbitrarily, which leads to the variance of the gradient growing arbitrarily, while $F'(x)$ remains bounded.

Such behavior can not occur if we consider strongly (μ, L)-feasible problems. This can be seen easily by observing that $\mathbb{V}_\omega\left[L_\omega\right] = \mathbb{E}_\omega\left[\left|L_\omega - \mathbb{E}_\omega\left[L_\omega\right]\right|^2\right] < L^2$, since $\mathbb{E}_\omega\left[L_\omega\right] \in [\mu, L]$ and therefore $|L - \mathbb{E}_\omega\left[L_\omega\right]| \leq L$. Recalling Proposition 3.8, we see that for strongly (μ, L)-feasible problems, we obtain the stronger bound $V_1 \leq 2\,\frac{L}{\mu} - 1$. However, this still becomes arbitrarily large with $\mu \to 0$. This also is not a flaw in the results, but rather a necessary consequence, as the following result demonstrates.

Proposition 3.12. *Suppose that $\mathcal{P}_\star(\mu, L)$ is the set of all strongly (μ, L)-feasible stochastic optimization problems $\mathrm{SOP}(f)$. Then for any $\mu \leq \frac{1}{2}$ we have*

$$\sup_{\mathrm{SOP}(f)\in\mathcal{P}_\star(\mu,1)} \inf_{\widetilde{V}_0\in\mathbb{R}} V_1(\widetilde{V}_0) \geq \frac{1}{64\mu}.$$

In particular, the bound in Proposition 3.8, Item b) is asymptotically sharp.

Proof. This result is proved by a family of stochastic optimization problems that are strongly $(\mu, 1)$-feasible and satisfy

$$\inf_{\widetilde{V}_0\in\mathbb{R}} V_1(\widetilde{V}_0) \geq \frac{(1-\mu)^3}{2\,\mu\,(2-\mu)^2},$$

which implies the result for any $\mu \leq \frac{1}{2}$.

For $\mu \in (0, \frac{1}{2}]$ and $\alpha \in (0, 2\pi)$ let

$$A_1 := \begin{pmatrix} \mu\cos^2(\alpha) + \sin^2(\alpha) & \frac{(1-\mu)\sin(2\alpha)}{2} \\ \frac{(1-\mu)\sin(2\alpha)}{2} & \mu\sin^2(\alpha) + \cos^2(\alpha) \end{pmatrix}$$

and

$$A_2 := \begin{pmatrix} \mu\cos^2(\alpha) + \sin^2(\alpha) & -\frac{(1-\mu)\sin(2\alpha)}{2} \\ -\frac{(1-\mu)\sin(2\alpha)}{2} & \mu\sin^2(\alpha) + \cos^2(\alpha) \end{pmatrix}.$$

Then,

$$A = \frac{1}{2}(A_1 + A_2) = \begin{pmatrix} \mu\cos^2(\alpha) + \sin^2(\alpha) & 0 \\ 0 & \mu\sin^2(\alpha) + \cos^2(\alpha) \end{pmatrix}.$$

As is easily checked, A_1 and A_2 have the eigenvalues μ and 1, and A has the eigenvalues $\mu\cos^2(\alpha) + \sin^2(\alpha)$ and $\mu\sin^2(\alpha) + \cos^2(\alpha)$.

For $i = 1, 2$, let $f_i(x) := \frac{1}{2}x^T A_1 x$ and $F(x) := \frac{1}{2}(f_1(x) + f_2(x))$. The corresponding SOP (with $\Omega = \{1, 2\}$ and $\mathbb{P}$ being the uniform distribution on Ω) is strongly (μ, L)-feasible. Trivially, $\nabla f_i(x) = A_i x$ and $\nabla F(x) = Ax$. When fixing $\alpha = \arcsin(\sqrt{\mu})$ and choosing $x = s\begin{pmatrix} \frac{1}{\mu} \\ 0 \end{pmatrix}$ for a scaling parameter $s > 0$, we observe

- $\mathbb{V}_\omega\left[f'_\omega(x)\right] = \frac{s^2(1-\mu)^3}{\mu}$ and
- $\|F'(x)\|_X^2 = s^2(\mu - 2)^2$.

Thus, choosing $s := \sqrt{2}\sqrt{\frac{\widetilde{V}_0\mu}{(1-\mu)^3}}$ provides us with a vector w with $\mathbb{V}_\omega\left[f'_\omega(x)\right] > \widetilde{V}_0$. Therefore,

$$V_1(\widetilde{V}_0) \geq \frac{\mathbb{V}_\omega\left[f'_\omega(x)\right] - \widetilde{V}_0}{\|F'(x)\|_X^2} = \frac{(1-\mu)^3}{2\,\mu\,(2-\mu)^2}.$$

□

Remark 3.13. *Proposition 3.12 describes the behavior of the constant V_1 in (3.5) for small μ. Thus, the assumption $\mu \leq \frac{1}{2}$ is not a major restriction. Reviewing the proof, we see that we could allow for arbitrary L (instead of $L = 1$) in Proposition 3.12 with the restriction $\mu \leq \frac{L}{2}$.*

3.3 Step Sizes for SGD

The performance of SGD is heavily influenced by the step sizes α_n which are used, and the selection of good step size schemes is much more involved compared to the deterministic setting. In this section, we will elaborate the behavior of SGD with different approaches to select the step sizes. We will present convergence results using the variance bounds discussed in Section 3.2.2 for different, well known step size schemes. These approaches to step sizes selection discussed here are well known in the literature. We collect them here to give a comprehensive overview, and also to highlight the challenges in the development of adaptive step sizes.

3.3.1 Constant Step Sizes

Let us begin by considering the most simple case: SGD with constant step sizes. Here, we use $\alpha_n = \alpha$ for some fixed $\alpha > 0$ and all $n \in \mathbb{N}_0$.

Consider a (μ, L)-feasible SOP, a current iterate $x_n \in \mathcal{H}$ and a step size α. Then, for the next iterate x_{n+1}, given by (3.3), we have due to L-smoothness of F (see Lemma 2.5):

$$F(x_{n+1}) \leq F(x_n) - \alpha F'(x_n) \nabla f_{\omega_n}(x_n) + \frac{\alpha^2 L}{2} \left\| \nabla f_{\omega_n}(x_n) \right\|^2 .$$

Again, x_{n+1} is a random variable depending on ω_n, and consequently $F(x_{n+1})$ is a random variable as well. Taking the expectation with respect to ω_n, and assuming x_n and α to be given and fixed, we get:

$$\begin{aligned} \mathbb{E}_{\omega_n} \left[F(x_{n+1}) \right] &\leq F(x_n) - \alpha \left\| F'(x_n) \right\|_{\mathcal{H}^*}^2 + \frac{\alpha^2 L}{2} \mathbb{E}_{\omega_n} \left[\left\| f'_{\omega_n}(x_n) \right\|_{\mathcal{H}^*}^2 \right] && (3.8) \\ &= F(x_n) + \alpha \left(\frac{\alpha L}{2} - 1 \right) \left\| F'(x_n) \right\|_{\mathcal{H}^*}^2 + \frac{\alpha^2 L}{2} \mathbb{V}_{\omega_n} \left[f'_{\omega_n}(x_n) \right] . && (3.9) \end{aligned}$$

In order to obtain descent from this bound, one clearly needs $\frac{\alpha L}{2} - 1 < 0$, or equivalently $\alpha < \frac{2}{L}$. Then, descent in functional value is expected, as long as

$$\left(1 - \frac{\alpha L}{2} \right) \left\| F'(x_n) \right\|_{\mathcal{H}^*}^2 > \frac{\alpha L}{2} \mathbb{V}_{\omega_n} \left[f'_{\omega_n}(x_n) \right] .$$

For fixed α and close to the minimizer, this can only hold if, with $F'(x_n) \to 0$, we also have $\mathbb{V}_{\omega_n} \left[f'_{\omega_n}(x_n) \right] \to 0$, i.e. in the interpolating setting.

The following results are well known in the literature and characterize the convergence behavior of SGD with constant step sizes under the different variance bounds discussed in Section 3.2.2. The first result explores the (stronger) assumption (3.5) and describes the behavior for convex and non-convex problems. The second result uses (3.6) and shows convergence in the convex case.

Proposition 3.14 (Convergence with the variance bound (3.5))**.** *Consider an SOP such that* (3.5) *is satisfied.*

a) Suppose that the SOP is (μ, L)-feasible for some $\mu > 0$. Then, for any constant, positive step size α satisfying

$$\alpha \leq \frac{1}{L(1 + V_1)},$$

we have for the iterates (x_n) generated by SGD:

$$\mathbb{E}_{\omega_0, \ldots, \omega_{n-1}} \left[F(x_n) - F(x^\star) \right] \leq (1 - \mu\alpha)^n \left(F(x_0) - F(x^\star) - \frac{\alpha L \widetilde{V}_0}{2\mu} \right) + \frac{\alpha L \widetilde{V}_0}{2\mu}. \tag{3.10}$$

b) Suppose that the mean F is L-smooth and bounded from below by $F_{\inf} > -\infty$. Then, for any constant, positive step size α satisfying

$$\alpha \leq \frac{1}{L(1+V_1)},$$

we have for the iterates (x_n) generated by SGD, for any $N \in \mathbb{N}$:

$$\mathbb{E}_{\omega_0,\dots,\omega_{N-1}}\left[\sum_{n=1}^{N} \left\|F'(x_n)\right\|_{\mathcal{H}^*}^2\right] \leq N\alpha L\widetilde{V}_0 + \frac{2(F(x_0)-F_{\inf})}{\alpha}.$$

Proof. See [11, Theorems 4.6 and 4.8]. □

Proposition 3.15 (Convergence with the variance bound (3.6))**.** *Consider an SOP such that* (3.6) *is satisfied and such that F is μ-strongly convex for some $\mu > 0$. Then for any constant step size α satisfying*

$$\alpha \leq \frac{1}{V_1},$$

we have for the iterates (x_n) generated by SGD:

$$\mathbb{E}_{\omega_0,\dots,\omega_{n-1}}\left[\|x_n - x^\star\|_{\mathcal{H}}^2\right] \leq (1-\mu\alpha)^n\left(\|x_0 - x^\star\|_{\mathcal{H}}^2 - \frac{\alpha\widetilde{V}_0}{\mu}\right) + \frac{\alpha\widetilde{V}_0}{\mu}. \tag{3.11}$$

Proof. We have

$$\begin{aligned}\mathbb{E}_{\omega_{n-1}}\left[\|x_n - x^\star\|_{\mathcal{H}}\right] = \|x_{n-1} - x^\star\|_{\mathcal{H}}^2 \\ - 2\alpha F'(x_{n-1})(x_{n-1}-x^\star) + \alpha^2\mathbb{E}_{\omega_{n-1}}\left[\left\|f'_{\omega_{n-1}}(x_{n-1})\right\|_{\mathcal{H}^*}^2\right].\end{aligned} \tag{3.12}$$

By (3.6) and convexity we have:

$$\begin{aligned}\mathbb{E}_{\omega_{n-1}}\left[\left\|f'_{\omega_{n-1}}(x_{n-1})\right\|_{\mathcal{H}^*}^2\right] &\leq \widetilde{V}_0 + V_1(F(x_{n-1}) - F(x^\star)) \\ &\leq \widetilde{V}_0 + V_1\left(F'(x_{n-1}), x_{n-1}-x^\star\right).\end{aligned}$$

Thus, by inserting into (3.12) we get:

$$\mathbb{E}_{\omega_{n-1}}\left[\|x_n - x^\star\|_{\mathcal{H}}^2\right] \leq \|x_{n-1}-x^\star\|_{\mathcal{H}}^2 + \left(F'(x_{n-1}), x_{n-1}-x^\star\right)\left(\alpha^2 V_1 - 2\alpha\right) + \alpha^2\widetilde{V}_0. \tag{3.13}$$

For $\alpha \leq \frac{1}{V_1}$, we have $\left(\alpha^2 V_1 - 2\alpha\right) \leq -\alpha$. Further, due to strong convexity we have $(F'(x_{n-1}), x_{n-1}-x^\star) \geq \mu\|x_{n-1}-x^\star\|_{\mathcal{H}}^2$ (see Proposition 2.6). Therefore:

$$\mathbb{E}_{\omega_{n-1}}\left[\|x_n - x^\star\|_{\mathcal{H}}^2\right] \leq (1-\alpha\mu)\|x_{n-1}-x^\star\|_{\mathcal{H}}^2 + \alpha^2\widetilde{V}_0. \tag{3.14}$$

Subtracting $\frac{\alpha\widetilde{V}_0}{\mu}$ from both sides gives

$$\mathbb{E}_{\omega_{n-1}}\left[\|x_n - x^\star\|_{\mathcal{H}}^2 - \frac{\alpha\widetilde{V}_0}{\mu}\right] \leq (1-\alpha\mu)\left(\|x_{n-1} - x^\star\|_{\mathcal{H}}^2 - \frac{\alpha\widetilde{V}_0}{\mu}\right). \tag{3.15}$$

Iterating this inequality while taking expectation over $\omega_0, \dots \omega_{n-1}$ now gives the result. □

In the non-interpolating case the constant $\widetilde{V}_0$ in a variance bound of the form (3.5) or (3.6) has to be positive. Therefore, for $n \to \infty$, the right-hand sides of the bounds (3.10) and (3.11) do not converge to zero, but tend towards a positive threshold proportional to α and $\widetilde{V}_0$ at a linear rate of $(1-\mu\alpha)$. Consequently, in the strongly convex case, both, Propositions 3.14 and 3.15, show convergence up to a stagnation level. With a larger step size, the initial convergence rate is faster, as $(1-\mu\alpha)$ is smaller, but the threshold reached by SGD with this given step size is also larger. SGD with smaller step sizes reaches a smaller stagnation level, but at a slower convergence rate. This convergence behavior is typical for stochastic optimization in the non-interpolating setting and can easily be observed in a simple numerical experiment:

Example 3.16. *We consider SGD applied to a simple SOP: Fix $N = 100$ and consider symmetric, positive definite $A_k \in \mathbb{R}^{d\times d}$ and $b_k \in \mathbb{R}^d$, where $d = 50$, set $f_k(x) = \frac{1}{2}x^T A_k x^t + b_k^t x$ and $F(x) = \frac{1}{N}\sum_{k=1}^N f_k(x)$. The eigenvalues of each A_k are selected to be equally spaced in the interval $[0.05, 1]$, rending the SOP strongly $[0.05, 1]$-feasible. Then, minimizing F can be seen as a SOP in the sense of Definition 3.1. When applying SGD to this SOP with different (sufficiently small) constant step sizes, we would expect, by Propositions 3.14 and 3.15, that smaller step sizes lead to slower initial convergence, but will also lead to a smaller remaining suboptimally gap (see* (3.10) *and* (3.11)*). This phenomenon can be observed in this numerical experiment: In Figure 3.1 the performance (in terms of the functional value $F(x_n) - F(x^\star)$) is displayed for SGD with different constant step sizes applied to the same problem. The curves are smoothed for better visualization. The same initialization has been used in the different runs of the algorithm. It is evident that, as the step size α decreases, the level of suboptimality achieved also declines. In addition, the speed of convergence to this level of suboptimality is reduced for smaller values of α, which is in line with the theoretical expectations.*

Example 3.17 (Dependence of constant step sizes on the convexity parameter μ)**.** *Proposition 3.14 gives a bound on the step size α, which guarantees convergence to a stagnation level as discussed in Example 3.16. Additionally, Proposition 3.8 shows that the variance bound* (3.5)*, which is a condition for Proposition 3.14 to hold, holds with $V_1 \sim \frac{1}{\mu}$ in the case of Proposition 3.8 b), or even with $V_1 \sim \frac{1}{\mu^2}$ in the case of Proposition 3.8 a). The bound on the step size present in Proposition 3.14 is proportional to $\frac{1}{V_1}$,*

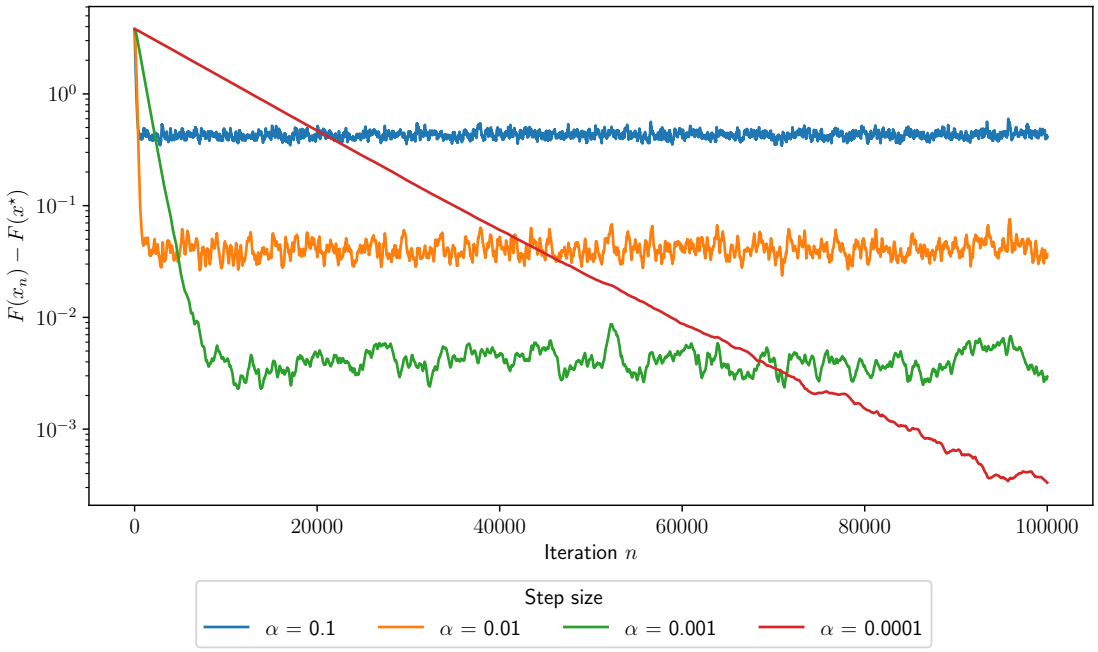


Figure 3.1: Performance of different constant step sizes. See Example 3.16 for further explanation.

implying that the step size α needs to decrease, when the convexity constant decreases. A simple experiment reveals that this behavior of α is too conservative in practice. We have performed the experiment described in Example 3.16, and selected the smallest eigenvalue of all A_k to be μ and the largest to be $L = 1$. By doing so, we obtain a strongly (μ, L)-feasible SOP. The results of our experiment are depicted in Figure 3.2. The figure shows a comparison of different step sizes, in dependency of the convexity parameter μ for the SOP in the proof of Proposition 3.12. SGD's relative progress is plotted, with higher values indicating better performance. According to Proposition 3.14, a step size of $\frac{1}{LV_1}$ should be employed. As shown in the proof of Proposition 3.12, V_1 grows at a rate of $\frac{1}{\mu}$ in this example. Therefore, keeping $L = 1$ fixed would result in a step size proportional to μ. However, this approach appears to be too conservative, if a fixed number of iterations is performed. Indeed, Proposition 3.15 together with Proposition 3.9 shows that the step size does not need to depend on the convexity parameter μ. In the experiment shown in Figure 3.2, the smaller step sizes resulting from the possible dependency of α on μ, result in slower initial convergence. Of course, smaller step sizes will reach a lower stagnation level eventually, (see the discussion in Example 3.16), but the time until it is reached grows with smaller step sizes. The observation that the curve for the step sizes $\alpha = 0.1 \cdot \mu$ and $\alpha = 0.1 \cdot \mu^2$ is higher than the curve for step sizes $\alpha = 0.1$ can be explained by the smaller stagnation level that these step sizes reach. In these cases, SGD is already

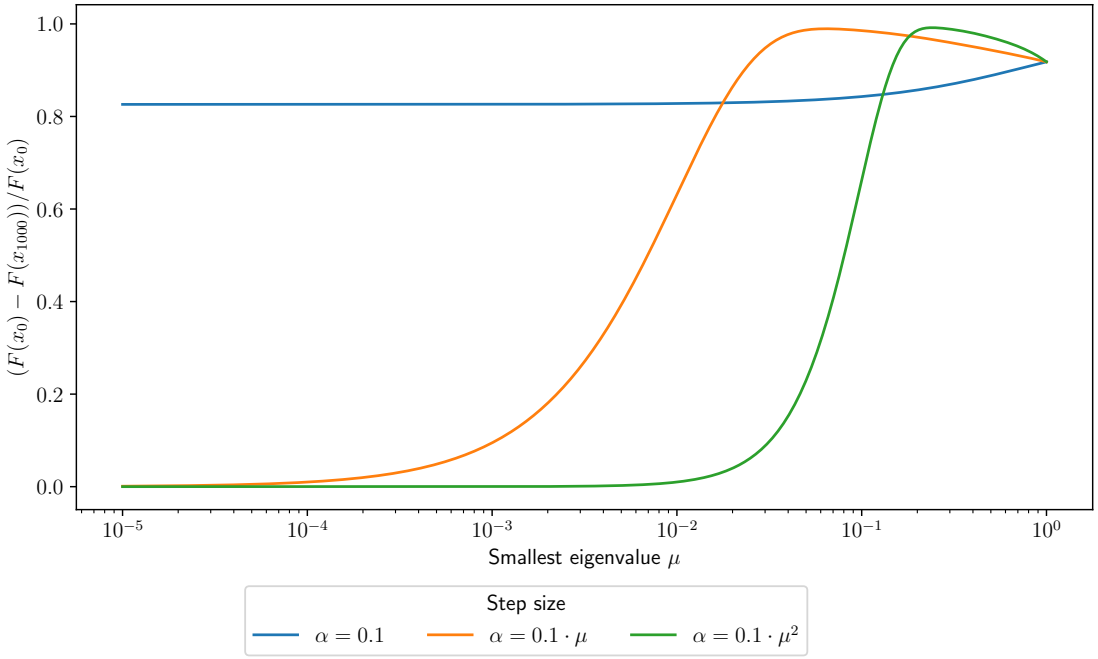


Figure 3.2: A step size $\sim \mu$ is too conservative. For more details see the discussion in Example 3.17.

stagnating with all step sizes, and the smaller step sizes $\alpha = 0.1 \cdot \mu$ *and* $\alpha = 0.1 \cdot \mu^2$ *have a smaller stagnation level than the step sizes* $\alpha = 0.1$*. For smaller values of* μ*, and consequently smaller values of the step sizes dependent on* μ*, the stagnation level is not yet reached due to the reduced speed of convergence with the smaller step sizes.*

3.3.2 Robbins Monroe Step Sizes

The results of the previous section indicate that, to obtain convergence for SGD, the step sizes need to decrease to zero at an appropriate rate. This rate of descent of the step sizes must not be too fast: If we have

$$\sum_{n=0}^{\infty} \alpha_n < \infty,$$

the algorithm is restricted to a bounded neighborhood of the initialization, and consequently global convergence cannot be expected. On the other hand, if

$$\sum_{n=0}^{\infty} \alpha_n^2 = \infty,$$

one often runs into statistical problems, as the noise to which is the algorithm is prone to during its complete (infinitely long) iteration might be unbounded. Consequently, a common class of a-priori step size strategies, which avoids these problems and goes back to [60] is obtained by assuming

$$\sum_{n=0}^{\infty} \alpha_n = \infty \qquad \text{and} \qquad \sum_{n=0}^{\infty} \alpha_n^2 < \infty. \tag{3.16}$$

Step sizes satisfying this condition are often referred to as *Robbins-Monroe* step sizes. For example, step sizes of the form

$$\alpha_n = \alpha \frac{1}{n^p}$$

for $p \in (\frac{1}{2}, 1]$ are Robbins-Monroe step sizes.

For further analysis of this kind of step sizes, the following result is useful:

Lemma 3.18. *Consider a sequence $\alpha_n \in (0, 1)$, such that*

$$\sum_{n=0}^{\infty} \alpha_n = \infty.$$

Then, for any $K \in \mathbb{N}_0$ we have:

$$\prod_{k=K}^{n} (1 - \alpha_k) \to 0, \qquad n \to \infty$$

Proof. First note that $\sum_{k=K}^{n} \alpha_k \to \infty$ for $n \to \infty$. We will use the well known inequality $\log(1 + x) \leq x$ for all $x > -1$. We compute

$$\begin{aligned} 0 \leq \prod_{k=K}^{n} (1 - \alpha_k) &= \exp\left(\log\left(\prod_{k=K}^{n} (1 - \alpha_k)\right)\right) \\ &= \exp\left(\sum_{k=K}^{n} \log(1 - \alpha_k)\right) \\ &\leq \exp\left(-\sum_{k=K}^{n} \alpha_k\right). \end{aligned}$$

As the sum in the argument diverges to ∞, this expression vanishes with $n \to \infty$. □

If the step sizes satisfy(3.16), the following can be obtained:

Proposition 3.19 (Convergence with the variance bound (3.5))**.** *Consider an SOP, such that* (3.5) *is satisfied and F is L-smooth. Suppose that the step sizes α_n of SGD satisfy* (3.16) *and $\alpha_n \leq \frac{1}{L(1+V_1)}$ for every $n \in \mathbb{N}_0$. Then:*

a) ***(Convergence almost surely)*** *If $F_{\inf} = \inf_{x\in\mathcal{H}} F(x) > -\infty$, then almost surely:*

$$\sum_{n=0}^{\infty} \alpha_n \left\|F'(x_n)\right\|_{\mathcal{H}^*}^2 < \infty.$$

If F is additionally μ-strongly convex with global minimizer $x^\star$, then:

$$\sum_{n=0}^{\infty} \alpha_n \left\|x_n - x^\star\right\|_{\mathcal{H}}^2 < \infty \qquad \text{and} \qquad x_n \to x^\star \ \text{ almost surely.}$$

b) ***(Convergence in expectation)*** *If the problem is (μ, L)-feasible, then:*

$$\mathbb{E}_{\omega_0,\dots,\omega_{n-1}} \left[F(x_n) - F(x^\star)\right] \to 0, \qquad n \to \infty.$$

Proof.

a) From (3.8) and (3.5) we infer:

$$\mathbb{E}_{\omega_n} \left[F(x_{n+1}) - F_\star\right] \leq F(x_n) - F_\star - \frac{\alpha_n}{2} \left\|F'(x_n)\right\|_{\mathcal{H}^*}^2 + \frac{\alpha_n^2 L}{2} \widetilde{V}_0.$$

Thus, applying Lemma 2.13, we conclude that $F(x_n) - F_\star$ convergences to a finite limit almost surely, and that

$$\sum_{n=0}^{\infty} \alpha_n \left\|F'(x_n)\right\|_{\mathcal{H}^*}^2 < \infty$$

almost surely. From L-smoothness and μ-strongly convex we infer

$$\sum_{n=0}^{\infty} \alpha_n \left\|x_n - x^\star\right\|_{\mathcal{H}}^2 < \infty \qquad \text{and} \qquad \sum_{n=0}^{\infty} \alpha_n \left(F(x_n) - F_\star\right) < \infty$$

almost surely. As $(\alpha_n)_{n\in\mathbb{N}}$ is *not* summable, and $F(x_n) - F_\star$ converges almost surely, the latter implies that $F(x_n) - F_\star \to 0$ almost surely. By strong convexity of F this also implies $\|x_n - x^\star\|_{\mathcal{H}^*}^2 \to 0$.

b) Denote $d_n = \mathbb{E}_{\omega_0,\dots,\omega_{n-1}} \left[F(x_n) - F(x^\star)\right]$. From (3.8) and strong convexity we infer:

$$d_{n+1} \leq (1 - \mu\alpha_n) d_n + \frac{\alpha_n^2 L}{2} \widetilde{V}_0.$$

Iterating this bound gives:

$$d_{n+1} \leq \left(\prod_{k=0}^{n} (1 - \mu\alpha_k)\right) d_0 + \frac{L\widetilde{V}_0}{2} \sum_{k=0}^{n} \left(\prod_{s=k+1}^{n} (1 - \mu\alpha_s)\right) \alpha_k^2. \tag{3.17}$$

The first addend goes to zero by Lemma 3.18. To show that the second vanishes as well, consider any $\varepsilon > 0$. As we have that $\sum_{n=0}^{\infty} \alpha_n^2 < \infty$, there is $N_1 \in \mathbb{N}$, such that

$$\sum_{n=N_1}^{\infty} \alpha_n^2 \leq \frac{\varepsilon}{2}$$

Denote $\alpha_{\max} = \sup_{n\in\mathbb{N}} \alpha_n < \infty$. Then, by Lemma 3.18, there is $N_2 \geq N_1$, such that

$$\prod_{s=N_1}^{N_2} (1 - \mu\alpha_s) \leq \frac{\varepsilon}{2\alpha_{\max}^2 N_1}$$

Then, for $n \geq N_2$ we have:

$$\begin{aligned}
\sum_{k=0}^{n} \left(\prod_{s=k+1}^{s} (1 - \mu\alpha_s) \right) \alpha_k^2 &\leq \sum_{k=0}^{N_1-1} \left(\prod_{s=k+1}^{n} (1 - \mu\alpha_s) \right) \alpha_k^2 + \frac{\varepsilon}{2} \\
&\leq N_1 \alpha_{\max}^2 \prod_{s=N_1}^{N_2} (1 - \mu\alpha_s) + \frac{\varepsilon}{2} \\
&\leq \varepsilon.
\end{aligned}$$

This shows that the right-hand side of (3.17) converges to zero for $n \to \infty$, and thus concludes the proof.

□

Proposition 3.20 (Convergence with the variance bound (3.6))**.** *Consider an SOP such that* (3.6) *is satisfied and F is convex. Suppose that the sequence of step sizes satisfies* (3.16) *and that $\alpha_n \leq \frac{1}{V_1}$ for all $n \in \mathbb{N}_0$.*

a) ***Convergence almost surely*** *For any stationary point $x^\star$ of F we have almost surely:*

$$\sum_{n=0}^{\infty} \alpha_n F'(x_n)(x_n - x^\star) < \infty.$$

If F is additionally μ-strongly convex for $\mu > 0$, then

$$\sum_{n=0}^{\infty} \alpha_n \left\| x_n - x^\star \right\|_{\mathcal{H}}^2 < \infty \qquad \text{and} \qquad x_n \to x^\star \ \text{ almost surely.}$$

b) ***Convergence in expectation*** *If F is additionally μ-strongly convex for $\mu > 0$, then*

$$\mathbb{E}_{\omega_0,\ldots,\omega_{n-1}} \left[\left\| x_n - x^\star \right\|_{\mathcal{H}}^2 \right] \to 0, \qquad n \to \infty.$$

Proof. a) Consider an arbitrary stationary point $x^\star$. We have (see also (3.13) in the proof of Proposition 3.15)

$$\mathbb{E}_{\omega_n}\left[\|x_{n+1} - x^\star\|_{\mathcal{H}}^2\right] \leq \|x_n - x^\star\|_{\mathcal{H}} \alpha_n + (\alpha_n V_1 - 2)\left(F'(x_n), x_n - x^\star\right) + \alpha_n^2 \widetilde{V}_0.$$

We proceed analogous to the proof of Proposition 3.19 and apply Lemma 2.13 and obtain convergence almost surely of $\|x_n - x^\star\|$ to a finite limit, and that

$$\sum_{n=0}^{\infty} \alpha_n F'(x_n)(x_n - x^\star) < \infty \qquad \text{almost surely.}$$

If F is additionally μ-strongly convex this implies

$$\sum_{n=0}^{\infty} \alpha_n \|x_n - x^\star\|_{\mathcal{H}}^2 < \infty$$

almost surely. As $_{n\in\mathbb{N}}$ is *not* summable and $\|x_n - x^\star\|_{\mathcal{H}}^2$ converges almost surely, this implies that $\|x_n - x^\star\|_{\mathcal{H}}^2 \to 0$ almost surely.

b) Denote $d_n = \mathbb{E}_{\omega_0,\ldots,\omega_{n-1}}\left[\|x_n - x^\star\|_{\mathcal{H}}^2\right]$. Then we have (see (3.15) in the proof of Proposition 3.15):

$$d_{n+1} \leq (1 - \mu\alpha_n) d_n + \alpha_n^2 \widetilde{V}_0.$$

Thus, the remaining proof can be performed completely analogous to the proof of Proposition 3.19 b).

□

The following table gives an overview over the different scenarios that we have covered.

	Step Sizes	
Variance Bound	**Constant**	**Robbins-Monro** (3.16)
(3.5)	Proposition 3.14	Proposition 3.19
(3.6)	Proposition 3.15	Proposition 3.20

Table 3.1: Overview of the convergence results discussed in this chapter

3.4 Behavior of SGD with Constant Step Sizes

The previous sections have demonstrated two important insights:

1. SGD with diminishing step sizes enjoys convergence guarantees under mild assumptions on the decay of the step sizes.

2. SGD with constant step sizes does (in general) not converge to the minimizer in the non-interpolating setting.

In this section, the second case will be investigated more thoroughly. Until now, we have only shown that the iterates converge to *some neighborhood* of the minimizer, specified by the convergence results in Propositions 3.14 and 3.15. While these results provide a bound on the size of this neighborhood (in the mean), they do not specify this neighborhood any further. However, numerical results indicate that, after a certain number of iterations, the initialization seems to be forgotten by the algorithm, and the iterates seem to follow a certain pattern. Let us consider a motivating example of a simple SOP with two functions:

$$\Omega = \{1, 2\} \qquad \text{with} \qquad \mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \frac{1}{2},$$

and

$$f_i(x) = \frac{1}{2}(x - y_i)^T A_i (x - y_i), \qquad i = 1, 2,\ x \in \mathbb{R}^2$$

where

$$A_i = \begin{pmatrix} 3 & (-1)^i \cdot 2 \\ (-1)^i \cdot 2 & 3 \end{pmatrix} \qquad \text{and} \qquad y_i = \begin{pmatrix} (-1)^i \\ 0 \end{pmatrix}. \tag{3.18}$$

We have applied SGD to this problem for 1×10^5 iterations with different constant step sizes and focus on the iterates of the algorithm. In Figure 3.3 we display the results of this experiment with four different step sizes. Each of the subplots shows the contour lines of the two functions f_1 and f_2, and the iterates of SGD with the specified step size. The black cross indicates the position of the minimizer $x^\star$ of F. The iterates are displayed as dots with low opacity, such that areas where many iterates accumulate are visible by stronger coloring. The same initialization $x_0 = (0, -1.5)$ has been used in all cases. It was observed that the choice of initialization has no implications on the result of this experiment, and the same pattern forms independently of the initialization. A possible explanation for this independence and the patterns seen in Figure 3.3 is that, over numerous iterations, the iterates of SGD follow a certain distribution. This distribution does depend on the step sizes, but not on the initialization.

It turns out that, under certain assumptions, the *long-term* distribution of the iterates of SGD can indeed be described by a measure $\nu^\star$, depending on the step sizes α. While this observation is of interest just by itself, we will investigate this long-term behavior for a certain reason: The adaptive step sizes, developed below in Chapter 4, will be based on *estimators*, computed from observations made along the trajectory of SGD. In order to develop an understanding for the resulting algorithm, we will analyze the behavior of these estimators along trajectories of constant step size SGD. We will see,

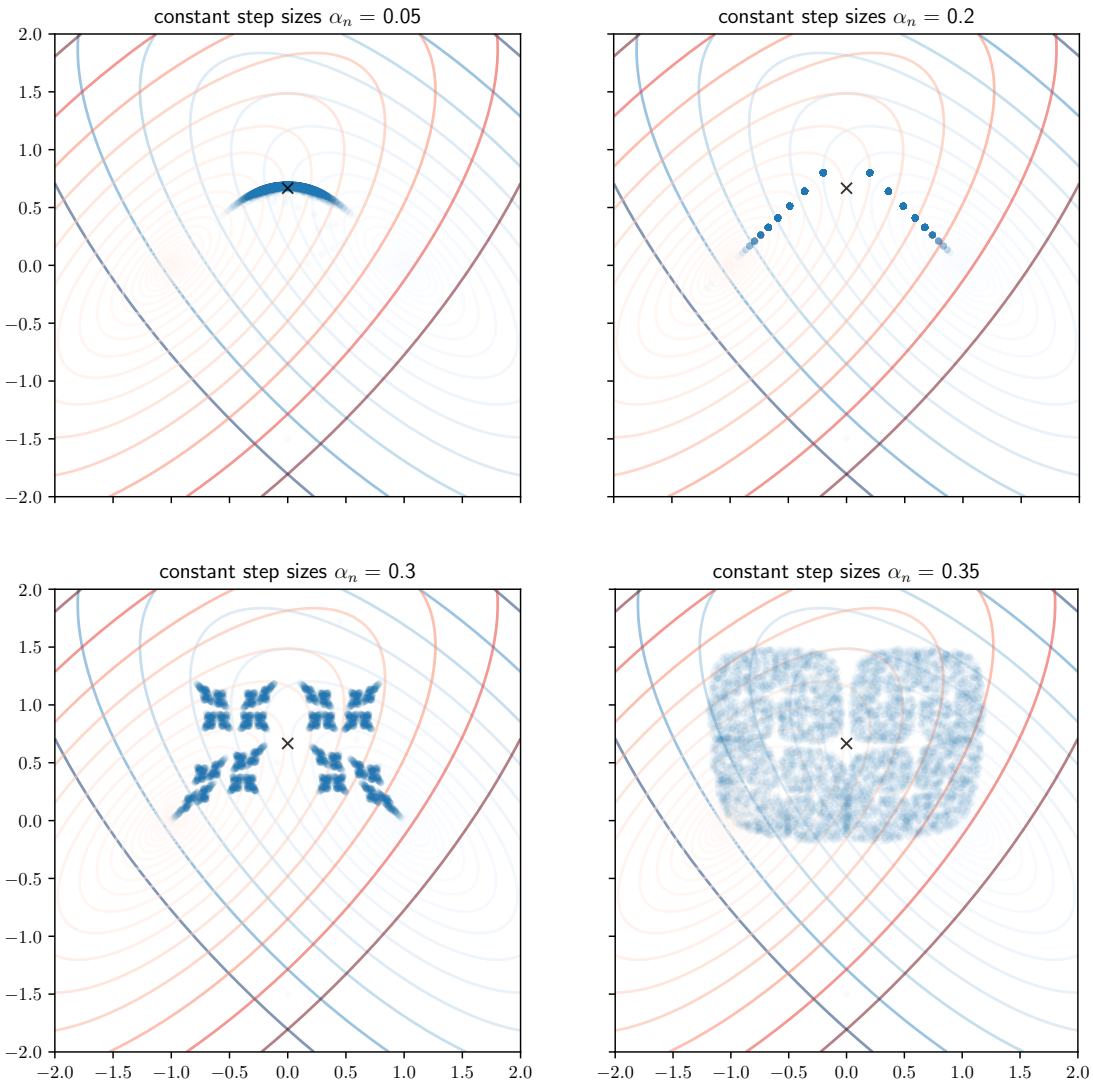


Figure 3.3: Iterates of SGD with different constant step sizes.

that the long-term behavior of the iterates and consequently also of the estimators can be described by means of this measure.

The existence of such a measure, which is invariant under the dynamics of SGD, has been studied in the literature in the context of Markov Chains ([1, 7, 27]) and, more recently, also in the context of SGD, which can be considered as a Markov Chain ([2, 13, 14, 18, 48, 64]). As there is no reference known to the author, which establishes existence and uniqueness of such a measure in our setting, we will derive these results in what follows for the readers' convenience.

Consider some fixed step size $\alpha > 0$, assumed to be sufficiently small. Then, the dynamics of SGD can be described using the transition function

$$\varphi(x, \omega) = x - \alpha \nabla f_\omega(x),$$

such that for the SGD iteration it holds:

$$x_{n+1} = \varphi(x_n, \omega_n).$$

This transition function, together with the probability measure $\mathbb{P}$ on Ω induces the Markov kernel

$$\begin{aligned} p : \mathcal{H} \times \mathcal{B}(\mathcal{H}) &\to \mathbb{R} \\ (x, A) &\mapsto \mathbb{P}\left(\{\omega \in \Omega \mid \varphi(x, \omega) \in A\}\right). \end{aligned}$$

Intuitively, $p(x, A)$ is the probability of SGD going from x to a point in A in one iteration. By construction, $p(x, \cdot\,)$ is a probability measure on $\mathcal{H}$ for every $x \in \mathcal{H}$, and for every $A \in \mathcal{B}(\mathcal{H})$, $p(\,\cdot\,, A)$ is a measurable function. We have the identity

$$p(x, A) = \mathbb{P}\left(\varphi(x, \cdot\,)^{-1}(A)\right),$$

showing that $p(x, \cdot)$ can be viewed as the push-forward measure of $\mathbb{P}$ under $\varphi(x, \cdot\,)$.

The Markov kernel p induces the Markov operator $\mathcal{P}$ defined by:

$$\mathcal{P}\nu = \int_{\mathcal{H}} p(x, \cdot\,)\, \mathrm{d}\nu(x),$$

acting on the signed Borel measure ν on $\mathcal{H}$. $\mathcal{P}$ maps probability measures to probability measures. For a probability measure ν_0 on $\mathcal{H}$, describing the distribution of an initial iterate x_0 of SGD, the probability measure $\nu_1 := \mathcal{P}\nu_0$ describes the distribution of the next iterate x_1 of SGD.

The stationary, or invariant, distribution $\nu^\star$ of SGD with step sizes α is a probability measure, which is invariant under the dynamics of SGD, formally:

$$\mathcal{P}\nu^\star = \nu^\star. \tag{3.19}$$

We highlight that, in general, all functions and operators discussed so far, namely the transition function φ, the Markov kernel p and the Markov operator $\mathcal{P}$ depend on the step size α. Consequently, also the invariant measure of SGD depends on the step size in general. For ease of notation we simply write $\varphi, p, \mathcal{P}$, and $\nu^\star$, and have in mind that they depend on α.

In what follows, we will show that there is a unique probability measure $\nu^\star$ satisfying (3.19). For the remainder of this chapter we will make the following assumptions:

Assumption A1. *The state space $\mathcal{H}$ is finite-dimensional.*

Assumption A2. *The variance bound* (3.5) *holds.*

Assumption A3. *The* $\mathrm{SOP}(f)$ *is (μ, L)-feasible for some $0 < \mu \leq L < \infty$ and pointwise (μ_ω, L_ω)-feasible (in the sense of Definition 3.2).*

Assumption A4. *The step size is sufficiently small, more precisely we assume*

$$\alpha < \min\left(\frac{2}{L(1+V_1)}, \frac{1}{L_{\max}}\right),$$

where V_1 is the constant from (3.5) *and $L_{\max}$ is defined in Definition 3.2.*

Existence

Existence of invariant measures of Markov Operators is well studied in the literature. An often very handy first reference are the lecture notes [27]. Here, we will use the classic result, that Markov operators which have the so-called *Feller* property and exhibit a suitable Lyapunov function have at least one invariant measure. This result can be found also in the aforementioned lecture notes in [27, Theorem 4.21], however, here we will use a similar result stated in Corollary 4.23 of the text book [7].

To show the Feller property in the sense of [7, Chapter 1.2], we need to show that for any bounded and continuous function $g : \mathcal{H} \to \mathbb{R}$ the function defined by

$$h : x \mapsto \int_{\mathcal{H}} g(y)\, \mathrm{d}p(x, \cdot)(y)$$

is also bounded and continuous. The boundedness of h is a trivial consequence of the boundedness of g and the property $p(x, \mathcal{H}) = 1$. To see that h is also continuous, consider some $x \in \mathcal{H}$ a sequence $(x_n)_{n\in\mathbb{N}} \subset \mathcal{H}$ with $x_n \to x$. We note that, by construction of $p(x, \cdot)$ we have the representation:

$$h(x_n) = \int_\Omega g(x_n - \alpha \nabla f_\omega(x_n))\, \mathrm{d}\mathbb{P}(\omega). \tag{3.20}$$

By continuity of g and ∇f_ω we have for almost every $\omega \in \Omega$ that

$$g(x_n - \alpha \nabla f_\omega(x_n)) \to g(x - \alpha \nabla f_\omega(x)).$$

Further, the sequence $g(x_n - \alpha \nabla f_\omega(x_n))$ is uniformly bounded in ω, as g is bounded. Thus, by the dominated convergence theorem, one gets:

$$h(x_n) \to \int_\omega g(x - \alpha \nabla f_\omega(x)) \, \mathrm{d}\mathbb{P}(\omega),$$

which equals $h(x)$ with the same argument which was used to justify the representation in (3.20). Consequently, the Markov operator $\mathcal{P}$ satisfies the Feller property in the sense of [7]. To employ Corollary 4.23 from [7], it thus remains to show that there is a proper (in the sense of [7, Chapter 4.1.1]) function $V : \mathcal{H} \to \mathbb{R}$, some $0 \leq \rho < 1$ and $\kappa \in \mathbb{R}$, such that:

$$\int_{\mathcal{H}} V(y) \, \mathrm{d}p(x, \cdot)(y) \leq \rho V(x) + \kappa$$

for all $x \in \mathcal{H}$. Such a function is referred to as a *Lyapunov* function for the Markov operator $\mathcal{P}$. We define $V(x) := F(x) = \int_\Omega f_\omega(x) \, \mathrm{d}\mathbb{P}(\omega)$, then due to L-smoothness (see Lemma 2.5):

$$\begin{aligned}
\int_{\mathcal{H}} V(y) \, \mathrm{d}p(x, \cdot)(y) &= \int_\Omega F(x - \alpha \nabla f_\omega(x)) \, \mathrm{d}\mathbb{P}(\omega) \\
&\leq \int_\Omega F(x) - \alpha \left(F'(x), \nabla f_\omega(x) \right) + \frac{\alpha^2 L}{2} \left\| f'_\omega(x) \right\|_{\mathcal{H}^*}^2 \, \mathrm{d}\mathbb{P}(\omega) \\
&= F(x) + \left(\frac{\alpha^2 L}{2} - \alpha \right) \left\| F'(x) \right\|_{\mathcal{H}^*}^2 + \frac{\alpha^2 L}{2} \mathbb{V}_\omega \left[\nabla f_\omega(x) \right]
\end{aligned}$$

Using Assumption A2, we have $\mathbb{V}_\omega \left[\nabla f_\omega(x) \right] \leq \widetilde{V}_0 + V_1 \left\| F'(x) \right\|_{\mathcal{H}^*}^2$ for some constants $\widetilde{V}_0$ and V_1. Consequently, we get, using Assumption A4 and $\left\| F'(x) \right\|_{\mathcal{H}^*}^2 \geq 2\mu \left(F(x) - F(x^\star) \right)$ (see Lemma 2.10):

$$\begin{aligned}
F(x) + \left(\frac{\alpha^2 L}{2} - \alpha \right) \left\| F'(x) \right\|_{\mathcal{H}^*}^2 + \frac{\alpha^2 L}{2} \mathbb{V}_\omega \left[\nabla f_\omega(x) \right] & \\
&\leq F(x) + \left(\alpha^2 \frac{L(1 + V_1)}{2} - \alpha \right) \left\| F'(x) \right\|_{\mathcal{H}^*}^2 + \frac{\alpha^2 L}{2} \widetilde{V}_0 \\
&\leq \rho F(x) + 2\mu \left(\alpha - \alpha^2 \frac{L(1 + V_1)}{2} \right) F(x^\star) + \frac{\alpha^2 L}{2} \widetilde{V}_0.
\end{aligned}$$

Here, we have defined $\rho = 1 - 2\mu \left(\alpha - \alpha^2 \frac{L(1+V_1)}{2} \right) < 1$. We also have $\rho \geq 0$, as the following computation verifies.

$$\rho = 1 - 2\mu\alpha + \alpha^2 \mu L (1 + V_1) \geq 1 - 2\mu\alpha + \alpha^2 \mu^2 = (1 - \mu\alpha)^2 \geq 0. \tag{3.21}$$

This shows that indeed $V = F$ is a suitable Lyapunov function. It is proper, as $\mathcal{H}$ is finite dimensional by Assumption A1 and F is strongly convex.

Remark 3.21. *One could replace Assumption A2 with:*

Assumption A5. *The bound* (3.6) *holds.*

In this case, the step size bound $\alpha < 4\frac{\mu}{L}$ implies that F is a Lyapunov function. Similarly, one could also use $V(x) = \|x - x^\star\|_{\mathcal{H}}^2$ as a Lyapunov function with an appropriate step size bound.

In any case, we have justified the use of [7, Corollary 4.23] and thus get:

Corollary 3.22. *Assume Assumption A1 and (Assumption A2 or Assumption A5). If the step size α is sufficiently small, there exists a probability measure $\nu^\star$ on $\mathcal{H}$, which is invariant under $\mathcal{P}$:*

$$\mathcal{P}\nu^\star = \nu^\star.$$

Uniqueness

In this section we show that strong convexity of the target function F implies uniqueness of the invariant measure. More precisely, this is achieved by considering a certain contracting property (contracting on average) of the SGD dynamics, which is implied by strong convexity of F. We follow the line of the proof of [27, Theorem 4.25] with minor adjustments to our setting.

For any $x, y \in \mathcal{H}$ and $\omega \in \Omega$ we have:

$$\begin{aligned}
&\|x - \alpha\nabla f_\omega(x) - (y - \alpha\nabla f_\omega(y))\|_{\mathcal{H}}^2 \\
&\quad = \|x - y\|_{\mathcal{H}}^2 - 2\alpha\left(f'_\omega(x) - f'_\omega(y), x - y\right) + \alpha^2 \left\|f'_\omega(x) - f'_\omega(y)\right\|_{\mathcal{H}^*}^2 \\
&\quad \leq \|x - y\|_{\mathcal{H}}^2 + (\alpha^2 L_{\max} - 2\alpha)\left(f'_\omega(x) - f'_\omega(y), x - y\right) \\
&\quad \leq \|x - y\|_{\mathcal{H}}^2 - \alpha\left(f'_\omega(x) - f'_\omega(y), x - y\right)
\end{aligned}$$

Here, we used Proposition 2.7 and $\alpha \leq \frac{1}{L_{\max}}$. Consequently, we get:

$$\begin{aligned}
&\int_\Omega \|x - \alpha\nabla f_\omega(x) - (y - \alpha\nabla f_\omega(y))\|_{\mathcal{H}}^2 \, \mathrm{d}\mathbb{P}(\omega) \\
&\qquad\qquad \leq \|x - y\|_{\mathcal{H}}^2 - \alpha\left(F'(x) - F'(y), x - y\right) \\
&\qquad\qquad \leq (1 - \mu\alpha)\|x - y\|_{\mathcal{H}}^2 = \rho\|x - y\|_{\mathcal{H}}^2,
\end{aligned} \tag{3.22}$$

denoting $\rho = (1 - \mu\alpha) \in [0, 1)$.

Now consider two invariant probability measures ν_1 and ν_2 on $\mathcal{H}$. Consider two independent random variables x_0 and y_0, distributed according to ν_1 and ν_2, respectively. Define the projection maps $G_i : \mathcal{H} \times \mathcal{H} \to \mathcal{H}$ by $G_1(x, y) = x$ and $G_2(x, y) = y$. Finally,

denote by π_n the joint distribution of (x_n, y_n), where x_n and y_n are obtained by SGD with initialization x_0 and y_0. Then, by invariance of ν_i, we have

$$\pi_n \circ G_i^{-1} = \nu_i, \qquad i = 1, 2,\ n \geq 0.$$

We wish to show that the sequence π_n is tight. By Lemma 2.20, ν_1 and ν_2 are tight, such that there are compact sets $K_1, K_2 \subset \mathcal{H}$, such that $\nu_i(K_i) \geq 1 - \varepsilon$, or, equivalently, $\nu_i(\mathcal{H} \setminus K_i) \leq \varepsilon$. We have:

$$\pi_n(K_1 \times K_2) = 1 - \pi_n((\mathcal{H} \times \mathcal{H}) \setminus (K_1 \times K_2)). \tag{3.23}$$

Now using

$$(\mathcal{H} \times \mathcal{H}) \setminus (K_1 \times K_2) \subset \big((\mathcal{H} \setminus K_1) \times \mathcal{H}\big) \cup \big(\mathcal{H} \times (\mathcal{H} \setminus K_2)\big),$$

we get from (3.23):

$$\begin{aligned}\pi_n(K_1 \times K_2) &\geq 1 - \pi_n\big((\mathcal{H} \setminus K_1) \times \mathcal{H}\big) - \pi_n\big(\mathcal{H} \times (\mathcal{H} \setminus K_2)\big) \\ &= 1 - \nu_1(\mathcal{H} \setminus K_1) - \nu_2(\mathcal{H} \setminus K_2) = 1 - 2\varepsilon.\end{aligned}$$

This shows that π_n is tight. Therefore, by Theorem 2.19, there is a subsequence n_k and a probability measure π on $\mathcal{H} \times \mathcal{H}$, such that

$$\int_{\mathcal{H}\times\mathcal{H}} g \,\mathrm{d}\pi_{n_k} \to \int_{\mathcal{H}\times\mathcal{H}} g \,\mathrm{d}\pi, \qquad k \to \infty \tag{3.24}$$

for any bounded and continuous function $g : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$.

Define the function $h(x, y) = \min(1, \|x - y\|_{\mathcal{H}}^2)$. We then have

$$\begin{aligned}\mathbb{E}_{\omega_{n-1}}\left[h(x_n, y_n)\right] &= \int_\Omega \min(1, \|x_n - \alpha\nabla f_\omega(x_n) - (y_n - \alpha\nabla f_\omega(y_n))\|_{\mathcal{H}}^2) \,\mathrm{d}\mathbb{P}(\omega) \\ &\leq \min\left(1, \int_\Omega \|x_n - \alpha\nabla f_\omega(x_n) - (y_n - \alpha\nabla f_\omega(y_n))\|_{\mathcal{H}}^2 \,\mathrm{d}\mathbb{P}\right) \\ &\leq \min\left(1, \rho\,\|x_{n-1} - y_{n-1}\|_{\mathcal{H}}^2\right),\end{aligned}$$

where we have used (3.22) in the last inequality. By iterating we deduce that

$$\mathbb{E}_{0:n-1}\left[h(x_n, y_n)\right] \leq \min(1, \rho^n\,\|x_0 - y_0\|_{\mathcal{H}}^2). \tag{3.25}$$

Also note that

$$\int_{\mathcal{H}\times\mathcal{H}} \mathbb{E}_{0:n-1}\left[h(x_n, y_n)\right] \,\mathrm{d}\pi_0 = \int_{\mathcal{H}\times\mathcal{H}} h \,\mathrm{d}\pi_n.$$

The map h is bounded and continuous. Using (3.24), we get:

$$\int_{\mathcal{H}\times\mathcal{H}} h \,\mathrm{d}\pi = \lim_{k\to\infty} \int_{\mathcal{H}\times\mathcal{H}} h \,\mathrm{d}\pi_{n_k} \leq \lim_{k\to\infty} \int_{\mathcal{H}\times\mathcal{H}} \min(1, \rho^{n_k}\,\|x - y\|_{\mathcal{H}}) \,\mathrm{d}\pi_0(x, y) = 0 \tag{3.26}$$

by the dominated convergence theorem. This implies

$$\int_{\mathcal{H}\times\mathcal{H}} h \,\mathrm{d}\pi = 0, \tag{3.27}$$

as $h \geq 0$. Denote by $D = \{(x, x) \mid x \in \mathcal{H}\}$ the *diagonal* in $\mathcal{H} \times \mathcal{H}$. Further, for $\varepsilon > 0$, by $D_\varepsilon = \{(x, y) \in \mathcal{H} \times \mathcal{H} \mid \|x - y\|_{\mathcal{H}}^2 \leq \varepsilon\}$ the tube with radius $\sqrt{\varepsilon}$ around the diagonal. We then have for every $0 < \varepsilon < 1$:

$$\begin{aligned}
\pi(D_\varepsilon) &= 1 - \int_{\|x-y\|_{\mathcal{H}}^2 > \varepsilon} \mathrm{d}\pi(x, y) \\
&= 1 - \frac{1}{\varepsilon} \int_{\|x-y\|_{\mathcal{H}}^2 > \varepsilon} \varepsilon \,\mathrm{d}\pi(x, y) \\
&\geq 1 - \frac{1}{\varepsilon} \int_{\|x-y\|_{\mathcal{H}}^2 > \varepsilon} h(x, y) \,\mathrm{d}\pi(x, y) \\
&\geq 1 - \frac{1}{\varepsilon} \int_{\mathcal{H}\times\mathcal{H}} h(x, y) \,\mathrm{d}\pi(x, y) = 1
\end{aligned}$$

In the first inequality, we have used that $\varepsilon < 1$ and $\varepsilon < \|x - y\|_{\mathcal{H}}^2$ on $\{\|x - y\|_{\mathcal{H}}^2 > \varepsilon\}$, implying $\varepsilon \leq h(x, y)$. In the last equality we have used (3.27). We have $D = \bigcap_{n=1}^{\infty} D_{\frac{1}{n}}$, therefore $\pi(D) = 1$. Finally, we can conclude for any $A \in \mathcal{B}(\mathcal{H})$:

$$\nu_1(A) = \pi(A \times \mathcal{H}) = \pi((A \times \mathcal{H}) \cap D) = \pi(A \times A) = \nu_2(A),$$

showing uniqueness of the invariant measure. The last equality can be justified by repeating the previous steps with interchanged roles of the first and second component. We have thus just shown:

Theorem 3.23. *Assume Assumption A3 and $\alpha \leq \frac{2}{L_{\max}}$. Then, there is at most one probability measure $\nu^\star$, such that*

$$\mathcal{P}\nu^\star = \nu^\star.$$

Remark 3.24. *The contraction property in Equation* (3.25) *allows us to deduce that two different trajectories of SGD will come arbitrary close to each other. This phenomenon, also known as* synchronization*, will play a role in the convergence theory of the estimators in Chapter 7.*

Finally, we get:

Corollary 3.25. *Assume Assumption A1, Assumption A3 and (Assumption A2 or Assumption A5). Then, for sufficiently small step size $\alpha > 0$, there is a unique probability measure $\nu^\star$, such that*

$$\mathcal{P}\nu^\star = \nu^\star.$$

Practical access to the invariant measure $\nu^\star$ is possible via so-called *observables*. For any $\nu^\star$-integrable function $h : \mathcal{H} \to \mathbb{R}$ results like the Birkhoff ergodic theorem (see, e.g.[28, Corollary 2.5.2]) ensure $\nu^\star$-alomst-sure convergence of the arithmetic mean of observations $h(y_n)$ made along the trajectory of a Markov Chain (y_n) to the mean of h:

$$\frac{1}{n}\sum_{k=1}^{n} h(y_k) \to \int_{\mathcal{H}} h(y)\,\mathrm{d}\nu^\star(y).$$

We will later use a similar result on the averaging process we use in the estimation of the adaptive step sizes and obtain convergence of certain averaged quantities to a mean with respect to the invariant measure $\nu^\star$.

4 Adaptive Step Sizes for SGD

The convergence results presented in Chapter 3 ensure convergence of SGD towards a stagnation level, if a sufficiently small constant step size is employed. Intuitively, taking Figure 3.1 into account, a good step size strategy has to decrease the step sizes to zero gradually at an appropriate rate to ensure convergence, at least in the non-interpolating setting. This intuition is further supported by the fact the Robbins-Monroe step sizes (Section 3.3.2) which decrease to zero, lead to convergence to the true minimizer (Propositions 3.19 and 3.20). From a mathematical perspective, the usage of Robbins-Monroe step sizes can be unsatisfactory: On the one hand, the guaranteed converges rates are in general not optimal. On the other hand, the step sizes are not problem-agnostic. Consequently, they also decrease to zero, if it is not necessary, which slows down the convergence in the interpolating setting. In this chapter, we devise a method to determine step sizes that do not only depend on the nonlinearity of the problem ($L_{\max}$) (as the constant step sizes in Section 3.3.1 do), but also take into account the local noise. As a consequence, the step sizes will decrease to zero adaptively in the non-interpolating setting, and will stay bounded from below by a positive constant in the interpolating setting. The suggested step sizes are proven to lead to convergence of order at least $O(1/n)$ in the non-interpolating setting and to linear convergence in the interpolating setting. Therefore, the method matches the best known convergence rates in the respective settings as they can be found, e.g. in [11, Theorems 4.6 and 4.7]. The adaptive step size rule derived below in Section 4.1 is, in the first place, theoretical. It involves quantities, that are in general not available, for example the smoothness constant L. To obtain a practical usable step size rule, we develop estimators for the quantities involved in Section 4.2. We will also present convergence results: For SGD with the *ideal* step sizes from Section 4.1 in Chapter 5, and for the estimators and the step sizes build upon these estimators in Chapter 7.

4.1 A noise adaptive step size rule

In this section, we present a noise adaptive step size rule, to which we will refer in the remainder of this work as the *ideal* step sizes. The idea behind these step sizes is simple: We use the bound for L-smooth functions from Lemma 2.5 (applied to F) and consider

its expectation with respect to the selection of the current sampled function f_{ω_n}. We then minimize this expression for the step sizes α_n.

Consider SGD applied to a (μ, L)-feasible SOP. From $x_{n+1} = x_n - \alpha_n \nabla f_\omega(x_n)$ we infer

$$\begin{aligned}\mathbb{E}_{\omega_n}\left[F(x_{n+1})\right] &\leq \mathbb{E}_{\omega_n}\left[F(x_n) - \alpha_n \left(F'(x_n), f'_{\omega_n}(x_n)\right) + \frac{L\,\alpha_n^2}{2} \left\|f'_{\omega_n}(x_n)\right\|_{\mathcal{H}^*}^2\right] \\ &= F(x_n) - \alpha_n \left\|F'(x_n)\right\|_{\mathcal{H}^*}^2 + \frac{L\,\alpha_n^2}{2}\,\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|_{\mathcal{H}^*}^2\right]. \end{aligned} \tag{4.1}$$

A straightforward idea is to minimize the right-hand side with respect to the step size α. This yields a step size, which maximizes the *expected* descent in the current iteration:

$$\alpha_n = \frac{\|F'(x_n)\|_{\mathcal{H}^*}^2}{L\,\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|_{\mathcal{H}^*}^2\right]} = \frac{\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|_{\mathcal{H}^*}^2\right] - \mathbb{V}_{\omega_n}\left[f'_{\omega_n}(x_n)\right]}{L\,\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|_{\mathcal{H}^*}^2\right]}. \tag{4.2}$$

Rewriting the step sizes in (4.2) further yields the following equivalent expression.

$$\alpha_n = \frac{1}{L}\left(1 - \frac{\mathbb{V}_{\omega_n}\left[f'_{\omega_n}(x_n)\right]}{\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|_{\mathcal{H}^*}^2\right]}\right).$$

This expression highlights that two factors determine good step sizes. On the one hand, this is the nonlinearity of the problem, described by L. On the other hand, we have the factor

$$\left(1 - \frac{\mathbb{V}_{\omega_n}\left[f'_{\omega_n}(x_n)\right]}{\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|_{\mathcal{H}^*}^2\right]}\right) \in [0, 1],$$

describing the local stochasticity of the problem. While the former is a well known factor in step sizes for deterministic gradient methods, see, for example [6, Chapter 4], the latter is intrinsic to the stochastic optimization setting. Recall that $\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|_{\mathcal{H}^*}^2\right] = \|F'(x_n)\|_{\mathcal{H}^*}^2 + \mathbb{V}_{\omega_n}\left[f'_{\omega_n}(x_n)\right]$. Thus, if the variance is relatively small compared to $\|F'(x_n)\|_{\mathcal{H}^*}^2$, we have step sizes close to $\frac{1}{L}$, as we would expect in the deterministic setting, and the factor equals 1 if the variance vanishes. Conversely, if the variance becomes relatively large compared to $\|F'(x_n)\|_{\mathcal{H}^*}^2$, as it would be the case close to a minimizer in the non-interpolating setting, the step sizes also become small. Consequently, the step sizes in (4.2) are not necessarily bounded away from zero and will become arbitrary small (i.e. tend to zero) on non-interpolating problems. This behavior of step sizes is not needed in comparable deterministic (or interpolating) settings, where it would slow down convergence, but is crucial in the non-interpolating setting in stochastic optimization to ensure convergence.

In contrast to the global Lipschitz constant L, the quantities

$$\mathbb{V}_{\omega_n}\left[f'_{\omega_n}(x_n)\right] \qquad \text{and} \qquad \mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|^2_{\mathcal{H}^*}\right]$$

depend on the current iterate x_n and are thus of a local nature. As we whish to model the noise in the current search direction (which is a local quantity), and in particular the behavior of the noise as we approach the minimizer $x^\star$, this is to be expected. Note that the quantities $\mathbb{V}_{\omega_n}\left[f'_{\omega_n}(x_n)\right]$ and $\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|^2_{\mathcal{H}^*}\right]$ do *not* depend on the current sample ω_n.

4.2 Estimation techniques

The practical use of the step sizes in (4.2) is limited by the lack of knowledge of the involved quantities. As a remedy, in this section we propose estimation techniques designed to yield local estimators for the smoothness constant L, the local variance in the search direction $\mathbb{V}_{\omega_n}\left[f'_{\omega_n}(x_n)\right]$ and the second moment of the noisy search direction $\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|^2_{\mathcal{H}^*}\right]$, which can subsequently be used to estimate the step sizes. These techniques are build upon an averaging technique we refer to as p-EMA, which is introduced in the next paragraph. The estimation techniques we present are heuristically motivated and, as a consequence of their construction, only provide biased estimators for the quantities they aim to estimate. Numerically, we will see that the step sizes derived from these techniques provide the expected convergence behavior, namely the convergence rates of the ideal step sizes (4.2), derived below in Theorem 5.1. The estimators we develop will also be analyzed theoretically in Chapter 7. There, we will present a theory which describes the convergence behavior of the estimators and the resulting estimated step sizes, if the estimation process is applied to SGD with constant step sizes. The important insights from the latter provide a better theoretical understanding of why these estimated step sizes are working in our experiments. However, they do not provide a rigorous proof of any convergence rates for SGD with the proposed step sizes.

4.2.1 Definition of p-EMA

The three estimators for the three quantities described above will all follow the same idea: We identify computable quantities (observables) and employ an averaging process, which takes noisy evaluations of the observables and yields a suitable de-noised estimate for a desired quantity of interest. More formally, we will consider the following recursive scheme: Suppose that a current estimate $\widehat{\tau}_n$ is given, and we make the observation $\widetilde{\tau}_n$ in the n-th Iteration of SGD. Then, we update the estimate as follows:

$$\widehat{\tau}_{n+1} = \gamma_n \widehat{\tau}_n + (1-\gamma_n)\widetilde{\tau}_n, \tag{4.3}$$

where we use

$$\gamma_n = 1 - \frac{1}{(n+1)^p}$$

for some $p \in (\frac{1}{2}, 1)$. We refer to this method of averaging the observations $\widetilde{\tau}_n$ as p-EMA. It is easily verified that $\widehat{\tau}_n$ is a convex combination of the initialization $\widehat{\tau}_0$ and all observables $\widetilde{\tau}_k$ for $0 \leq k \leq n-1$, where the weights on observables with larger k are larger. A more sophisticated discussion of the rationals behind this choice, as well as a stochastic convergence theory for this process is presented in Chapter 6.

4.2.2 Nonlinearity Estimation

We start by considering an estimator for the Lipschitz constant L of the gradient of F. To this end note that (the smallest possible) L is the smallest constant L satisfying

$$F(x+\delta_x) \leq F(x) + \big(F'(x), \delta_x\big) + \frac{L}{2}\|\delta_x\|_{\mathcal{H}}^2 \qquad \text{for all } x, \delta_x \in \mathcal{H}.$$

Rearranging this term gives

$$L \geq 2\frac{F(x+\delta_x) - F(x) - (F'(x), \delta_x)}{\|\delta_x\|_{\mathcal{H}}^2}. \tag{4.4}$$

For a general SOP, the term on the right-hand side is still not computable for any given δ_x, as $F(x)$, $F(x+\delta_x)$ and $(F'(x), \delta_x)$ are not computable. However, in the n-th iteration of SGD, the quantities $f_{\omega_n}(x_n)$ and $\delta_x = -\alpha_n \nabla f_{\omega_n}(x_n)$ are available. Suppose that additional evaluation of $f_{\omega_n}(x_{n+1}) = f_{\omega_n}(x_n + \delta_x)$ is possible. Noting that $\big(f'_{\omega_n}(x), \delta_x\big) = -\alpha \big\|f'_{\omega_n}(x)\big\|_{\mathcal{H}^*}^2$, a lower bound on the Lipschitz constant L_{ω_n} of f_{ω_n} is given by

$$\widetilde{L}_n = 2\frac{f_{\omega_n}(x_{n+1}) - f_{\omega_n}(x_n) + \alpha_n \big\|f'_{\omega_n}(x_n)\big\|_{\mathcal{H}^*}^2}{\alpha_n^2 \big\|f'_{\omega_n}(x_n)\big\|_{\mathcal{H}^*}^2}. \tag{4.5}$$

The values $\widetilde{L}_n$ can be evaluated during the execution of SGD with the computational overhead of evaluating $f_{\omega_n}(x_{n+1})$ in each iteration. We then employ p-EMA to average these observations and to obtain an estimate $\widehat{L}_n$ for the (local) Lipschitz constant of the gradient.

Remark 4.1. *This approach to estimate the Lipschitz constant L of the gradient is by far not new. The idea to use normalized deviations from a quadratic model as estimates for the remainder of the Taylor expansion can be found for example in [68, Equations 11 and 12], where a similar estimate as in* (4.5) *is used to estimate the Lipschitz constant of the second derivative. In [51, Section 6.1.2] a normalized deviation from a linear model, as in our case, is used as an estimator for the Lipschitz constant of the derivative in the context of step size section. In the context of first order methods for machine learning it can also be found in [23, Appendix A.2].*

As it is evident form (4.4), $\widetilde{L}_n$ from (4.5) is underestimating the true Lipschitz constant L_{ω_n} of f_{ω_n}. From this perspective, it might seem appealing to use the maximum of all observations $\widetilde{L}_n$ made so far as an estimator for L. In practice however, we have observed, that this seems overly cautions and leads to unnecessarily small step sizes. We therefore did not consider this possibility for estimating L in the present work, although it might be a valid alternative, especially if a more conservative step size strategy is desired.

The additional evaluation of $f_{\omega_n}(x_{n+1})$ is the main factor determining the computational cost introduced by this estimator. This quantity would usually not be evaluated in the execution of SGD. In machine learning applications, where gradients are computed via backpropagation, the computational cost of a standard iteration typically consists of approximately one-third for the forward pass and about two-thirds for the backward pass [3, 32]. Consequently, one additional forward pass per iteration (to evaluate $f_{\omega_n}(x_{n+1})$) increases the computational costs of the algorithm by roughly 33%. The cost of evaluating $\left\|f'_{\omega_n}(x_n)\right\|^2_{\mathcal{H}^*} = \left(f'_{\omega_n}(x_n), \nabla f_{\omega_n}(x_n)\right)$ is negligible in practical scenarios, where $f'_{\omega_n}(x_n)$ and $\nabla f_{\omega_n}(x_n)$ are computed anyway so the evaluation of $\left\|f'_{\omega_n}(x_n)\right\|^2_{\mathcal{H}^*}$ comes at the cost of one Euclidean inner product.

The values $\widetilde{L}_n$ are scalars, so that employing p-EMA to these values also comes with negligible additional cost.

Initialization

p-EMA requires setting an initialization. In p-EMA we have $\widehat{\tau}_1 = \widetilde{\tau}_0$, hence the initialization is determined by the first observation $\widetilde{\tau}_0$. For the case of the Lipschitz constant L, there are two options that can be considered:

1. If an initial step size α_0 is given, then one could use $\widehat{L}_1 = \widehat{L}_0 = \widetilde{L}_0$, where $\widetilde{L}_0$ is obtained as in (4.5).

2. If no initial step size is given, we can't directly evaluate (4.5). In Section 4.3 below we discuss the possibility to use a line search method, to find an initial step size. If such a line search method is employed and yields an initial step size α_0, then, in the light of the results on constant step sizes in Section 3.3.1, it is sensible to use $\widehat{L}_1 = \widehat{L}_0 = \frac{1}{\alpha_0}$ as an initialization.

In our numerical experiments in Chapter 8 we have used the second approach.

4.2.3 Estimation of $\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|^2_{\mathcal{H}^*}\right]$

The idea behind the estimator for the second moment of the noisy search direction is quite straightforward. We wish to estimate an expected value, which can be achieved by averaging observations drawn from the corresponding distribution. In this case, this would correspond to drawing several samples $f_{\bar{\omega}_1}, \ldots, f_{\bar{\omega}_K}$ for some $K \in \mathbb{N}$ and using an average like

$$\frac{1}{K}\sum_{k=1}^{K}\left\|f'_{\bar{\omega}_k}(x_n)\right\|^2_{\mathcal{H}^*}$$

as an approximation for $\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|^2_{\mathcal{H}^*}\right]$. This however comes with large additional computational costs and increases the per-iteration cost by a factor of K, as K additional forward- and backward passes are necessary. As a remedy, we might assume that the distribution of the two random variables (in ω)

$$\left\|f'_{\omega}(x_{n-s})\right\|^2_{\mathcal{H}^*} \qquad \text{and} \qquad \left\|f'_{\omega}(x_n)\right\|^2_{\mathcal{H}^*}$$

does not differ much, if $s \in \mathbb{N}$ is small, as in this case, the difference $x_{n-s} - x_n$ is small and f'_ω is (Lipschitz) continuous. Consequently, we might also use recent observations

$$\left\|f'_{\omega_{n-s}}(x_{n-s})\right\|^2_{\mathcal{H}^*}$$

in the averaging approach. This is again achieved by employing p-EMA to the observations

$$\widetilde{g}_n = \left\|f'_{\omega}(x_n)\right\|^2_{\mathcal{H}^*},$$

to obtain an approximation $\widehat{g}_n$ to $\mathbb{E}_\omega\left[\|f'_\omega(x_n)\|^2_{\mathcal{H}^*}\right]$. It is worth noticing, that in this way, older observations (larger s) are assigned a smaller weight compared to younger observations (smaller s). See the discussion in Chapter 6 for more details.

The computational cost introduced by this estimator is negligible, as $\left\|f'_{\omega_n}(x_n)\right\|^2_{\mathcal{H}^*}$ is computed already by the estimation process for L and can be reused. Application of p-EMA to the scalar values $\widetilde{g}_n$ also comes with negligible computational cost.

4.2.4 Variance Estimation

The remaining quantity in (4.2) we need to estimate is the variance $\mathbb{V}_\omega\left[f'_\omega(x_n)\right]$ at the current iterate x_n. By definition, we have

$$\mathbb{V}_\omega\left[\nabla f'_\omega(x_n)\right] = \mathbb{E}_\omega\left[\left\|f'_\omega(x_n) - F'(x_n)\right\|^2_{\mathcal{H}^*}\right].$$

Thus, using similar ideas as in Section 4.2.3, one would be tempted to apply p-EMA to the observations

$$\left\|f'_{\omega_n}(x_n) - F'(x_n)\right\|^2_{\mathcal{H}^*}.$$

This however would require the knowledge of $F'(x_n)$, which can't be assumed in the stochastic optimization framework. Instead, we propose the following approach to estimate the local variance. Consider one SGD step performed with step size α_n. Note that when α_n is sufficiently small, we have

$$\mathbb{E}_{\omega_n}\left[f_{\omega_n}(x_{n+1})\right] < \mathbb{E}_{\omega_{n+1}}\left[f_{\omega_{n+1}}(x_{n+1})\right] = F(x_{n+1}),$$

so the term on the left provides a biased estimate of the true functional value at the iterate x_{n+1}. This is because the search direction $\delta_x = x_{n+1} - x_n$ is selected to minimize f_{ω_n}, not F. By comparing the unbiased estimator $f_{\omega_{n+1}}(x_{n+1})$ to the biased estimator $f_{\omega_n}(x_{n+1})$, we can determine a notion of the local variance.

In order to quantify the above heuristic, recall that for sufficiently smooth functions we have

$$f(x+\delta_x) = f(x) + \left(f'(x), \delta_x\right) + O(\|\delta_x\|_{\mathcal{H}}^2).$$

Applying this expansion to $f_{\omega_n}(x_{n+1})$ and $f_{\omega_{n+1}}(x_{n+1})$, we obtain

$$f_{\omega_n}(x_{n+1}) = f_{\omega_n}(x_n) - \alpha_n \left\|f'_{\omega_n}(x_n)\right\|_{\mathcal{H}^*}^2 + O(\alpha_n^2)$$

and

$$f_{\omega_{n+1}}(x_{n+1}) = f_{\omega_{n+1}}(x_n) - \alpha_n \left(f'_{\omega_{n+1}}(x_n), \nabla f_{\omega_n}(x_n)\right) + O(\alpha_n^2).$$

We thus obtain

$$\begin{aligned} &f_{\omega_{n+1}}(x_{n+1}) - f_{\omega_n}(x_{n+1}) \\ &= f_{\omega_{n+1}}(x_n) - f_{\omega_n}(x_n) + \alpha_n \left\|f'_{\omega_n}(x_n)\right\|_{\mathcal{H}^*}^2 - \alpha_n \left(f'_{\omega_{n+1}}(x_n), \nabla f_{\omega_n}(x_n)\right) + O(\alpha_n^2) \end{aligned} \tag{4.6}$$

for the difference. We have $\mathbb{E}_{\omega_{n+1}}\left[f_{\omega_{n+1}}(x_n)\right] = \mathbb{E}_{\omega_n}\left[f_{\omega_n}(x_n)\right] = F(x_n)$. Since ω_n and ω_{n+1} are independent, we obtain, taking the expectation w.r.t. both, ω_n and ω_{n+1}:

$$\mathbb{E}_{\omega_n,\omega_{n+1}}\left[\left(f'_{\omega_{n+1}}(x_n), \nabla f_{\omega_n}(x_n)\right)\right] = \left\|F'(x_n)\right\|_{\mathcal{H}^*}^2.$$

We thus get, taking the expectation of (4.6):

$$\begin{aligned} \mathbb{E}_{\omega_n,\omega_{n+1}}\left[f_{\omega_{n+1}}(x_{n+1}) - f_{\omega_n}(x_{n+1})\right] &= \alpha_n \mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|_{\mathcal{H}^*}^2\right] \\ &\quad - \alpha_n \left\|F'(x_n)\right\|_{\mathcal{H}^*}^2 + O(\alpha_n^2) \\ &= \alpha_n \mathbb{V}_{\omega_n}\left[f'_{\omega_n}(x_n)\right] + O(\alpha_n^2). \end{aligned}$$

Neglecting the second order term, we obtain a way to estimate the variance

$$\mathbb{E}_{\omega_n,\omega_{n+1}}\left[f_{\omega_{n+1}}(x_{n+1}) - f_{\omega_n}(x_{n+1})\right] \approx \alpha_n \mathbb{V}_{\omega_n}\left[f'_{\omega_n}(x_n)\right].$$

Motivated by this, in the $(n+1)$-th iteration, after evaluating $f_{\omega_{n+1}}(x_{n+1})$, we evaluate

$$\widetilde{\sigma}_n^2 = \frac{f_{\omega_{n+1}}(x_{n+1}) - f_{\omega_n}(x_{n+1})}{\alpha_n} \tag{4.7}$$

and then apply p-EMA to obtain an estimation for the local variance. As in the case of the estimator for the second moment of the noisy search direction, this estimator incurs virtually no additional computational cost, as it only requires evaluating the finite difference term

$$\frac{f_{\omega_{n+1}}(x_{n+1}) - f_{\omega_n}(x_{n+1})}{\alpha_n}$$

and updating p-EMA. This arises because $f_{\omega_n}(x_{n+1})$ is already computed as part of the smoothness constant estimation (see Section 4.2.2), and $f_{\omega_{n+1}}(x_{n+1})$ will be computed in iteration $n+1$ regardless.

4.2.5 Alternative estimation of the variance

One drawback of the usage of (4.7) as observation for the variance estimate is the fact that this term might evaluate to values outside the interval $\left[0, \left\|f'_{\omega_n}(x_n)\right\|^2_{\mathcal{H}^*}\right]$, and consequently the resulting variance estimate might be negative or exceed the estimated value $\widehat{g}_n$ for $\mathbb{E}_\omega\left[\|f'_\omega(x_n)\|^2_{\mathcal{H}^*}\right]$. Such an estimate leads to the undesirable behavior of step sizes exceeding $\frac{1}{L}$ in the former case and negative step sizes in the latter. A simple remedy is to use $f_{\omega_n}(x_n)$ instead of $f_{\omega_{n+1}}(x_{n+1})$ in (4.7) and thus use

$$\widetilde{\sigma}_n^2 = \frac{f_{\omega_n}(x_n) - f_{\omega_n}(x_{n+1})}{\alpha_n} \tag{4.8}$$

as observations for the variance estimation process. While this selection is not supported by the motivation in Section 4.2.4, it offers similar convergence properties. We will discuss this in more detail in Section 7.2.1 and present numerical results in Section 7.2.1 and Chapter 8.

The benefit from using the observations from (4.8) is that they satisfy

$$\widetilde{\sigma}_n^2 \in \left[0, \left\|f'_{\omega_n}(x_n)\right\|^2_{\mathcal{H}^*}\right]$$

and therefore the resulting variance estimate satisfies

$$\widehat{\sigma}_n^2 \in [0, \widehat{g}_n].$$

Consequently, the estimated step sizes are from the interval $\left[0, \frac{1}{\widehat{L}_n}\right]$, which is desirable. Our numerical experiments in Chapter 8 on stochastic quadratic problems indicate that indeed both approaches perform comparably well, while the alternative selection of variance observations (4.8) offers more stable estimates for the variance. However, we also

observe that the step sizes emerging from the alternative variance estimation become small at significantly faster rate, and thus, in some cases, reduce the speed of convergence. We refer to the experiments presented in Section 8.4.

4.3 Practical Use of the Estimators

So far, we have developed estimators for the three key quantities in L, $\mathbb{V}_\omega\left[f'_\omega(x_n)\right]$, and $\mathbb{E}_\omega\left[\|f'_\omega(x_n)\|^2_{\mathcal{H}^*}\right]$, that are heuristically motivated. These estimators provide us with estimations $\widehat{L}_n$, $\widehat{g}_n$ and $\widehat{\sigma}^2_n$ for the three quantities at each iteration. Thus, recalling (4.2), one would use the step sizes

$$\widehat{\alpha}_n = \frac{1}{\widehat{L}_n}\left(1 - \frac{\widehat{\sigma}^2_n}{\widehat{g}_n}\right). \tag{4.9}$$

However, if done directly, this could cause instabilities of the algorithm. This is due to the fact that the estimators might be unreliable and provide false estimation, in particular at early stages of the run of the algorithm. To prevent undesirable behavior, we might impose the following *safeguards*, which aim to prevent the algorithm from using either too large or negative step sizes. These safeguards are, as the estimators, heuristically motivated. We shall see in the numerical experiments in Chapter 8 that the algorithm which uses our estimators and these safeguards is able to achieve the expected rate of convergence, that is guaranteed by Theorem 5.1 for the optimal step sizes (4.2), that the estimators aim to approximate.

Start of the Algorithm

When starting the algorithm, we are only given an initialization x_0 and no further information on the problem at hand. Thus, our best guess for a step size for the first iteration can be obtained by performing a simple line search on the first sampled function f_{ω_0}. The resulting step size from this line search can be utilized in two ways: First, obviously, it can be used as a good initial step size for the first iteration. Second, the inverse of this step also serves as a good initialization for the estimation process for the smoothness constant L. Initializations for the remaining two estimation processes, namely for $\mathbb{E}_\omega\left[\|f'_\omega(x_n)\|^2_{\mathcal{H}^*}\right]$ and $\mathbb{V}_\omega\left[f'_\omega(x_n)\right]$, can be motivated in the following ways: In the case of $\mathbb{E}_\omega\left[\|f'_\omega(x_n)\|^2_{\mathcal{H}^*}\right]$ we simply use $\widehat{g}_1 = \widehat{g}_0 = \left\|f'_{\omega_0}(x_0)\right\|^2_{\mathcal{H}^*}$, as this represents our best knowledge of this quantity at this point. In case of the variance $\mathbb{V}_\omega\left[f'_\omega(x_n)\right]$ we have no initial knowledge at this point. Recalling that the variance estimate is required for convergence to optimality, and not for initial descent, it seems appropriate to use $\widehat{\sigma}^2_1 = \widehat{\sigma}^2_0 = 0$ as an initialization. As another, more conservative, approach one could also

use the first variance estimate as initialization. When the classical observations (4.7) are employed, the problem of evaluating $f_{\omega_1}(x_1)$ arises. For the purpose of (stochastically) analyzing the algorithm, it is desirable, that the current step size α_n does not depend on the current sample ω_n. Consequently, the variance estimate can only be used from the third step ($n = 2$) onwards. When employing the alternative variance estimators discussed in Section 4.2.5 one does not encounter these problems and might use the variance estimate from the second iteration onwards.

Unreliable and Unstable Estimates

Another problem that arises, particularly at the start of the algorithm, are unreliable and unstable estimates.

The former is caused by the limited number of information gathered so far, which is, in principle, unavoidable. However, we might discuss possible ways to algorithmically handle this lack of information. One possibility is to enforce a descent on every sample. To check the iterates for descent on the current sample comes with no additional cost, as we evaluate $f_{\omega_n}(x_{n+1})$ for the estimation of the smoothness constant L and the variance. If no descent on the sample is observed, this indicates that at least one of the following has occurred:

1. The current step size is too large, leading to unstable iteration and possibly even to exploding gradients and/or iterates.
2. The current sample is an outlier, in the sense that it would require a smaller step size, while the current step size remains a good step size for the majority of possible sampled functions at the given iterate.

To deal with both possibilities at once, we propose to

1. Reject the current step, i.e. set $x_{n+1} = x_n$ and
2. perform a line search on the next sample $f_{\omega_{n+1}}$.

The use of the sample $f_{\omega_{n+1}}$ instead of f_{ω_n} for the line search provides (with high probability) a suitable step size in the former case, while preventing the step size from being unnecessarily reduced in the second case. As an additional safeguard, a maximum number of line search iterations on a single sample could be imposed, before continuing the line search on another sample. This further reduces the *risk* of unnecessarily small step sizes.

The latter, unstable estimates, are typically induced by the averaging process p-EMA itself, whose design intentionally places greater emphasis on recent observations to react faster to new observations during the early stages of the algorithm. As becomes apparent

from the weights assigned to young observations in p-EMA, see Figures 6.1 and 6.2, and from the discussion in the introduction of Chapter 6, this behavior is intrinsic to the method. One possible remedy is to alter the definition of p-EMA (4.3) in the following way:

$$\widehat{\tau}_{n+1} = \gamma_n \widehat{\tau}_n + (1 - \gamma_n)\widetilde{\tau}_n \tag{4.10}$$

with

$$\gamma_n = \max\left(\gamma_{\min}, 1 - \frac{1}{(n+1)^p}\right),$$

with some $\gamma_{\min} \in (0, 1)$, which should be selected close to 1. In this way the averaging procedure is damped in its sensitivity at early iterations, while the idea of assigning younger observations larger weights is preserved. Additionally, this *altered* version of p-EMA still enjoys the same convergence guarantees as p-EMA, as we will have $\gamma_n = 1 - \frac{1}{(1+n)^p}$ after a finite number of iterations.

The following ideas might also be incorporated:

Initialization Phase

To counteract unreliable estimators at the beginning of the algorithm, one could set a number n_{init} of iterations, and do not update the step size using the estimators until this number of iterations is reached. This allows for some amount of information to be gathered by the estimators, to provide a more reliable, and also more stable estimate for the step size. Taking into account the convergence results for the estimators under constant step sizes in Chapter 7, this can yield a more suitable estimate for the step size.

Smoothing of Step Sizes

One may also introduce an additional smoothing procedure applied to the estimated step sizes. Such a procedure should not interfere with the convergence of the estimators, and hence with the convergence of the estimated step sizes, but can moderate overly rapid adjustments in their magnitude. For instance, one could take a moving average of the last n_{avg} proposed step sizes, or employ an exponential moving average scheme; see Chapter 6 for a detailed discussion of these approaches. In principle, one could even apply the p-EMA in this context, although doing so would diminish the algorithm's responsiveness, particularly in its later stages.

It should be emphasized that the preceding discussion is purely heuristic and that the underlying ideas have not yet been subjected to rigorous analysis. In our numerical experiments, however, the proposed techniques at least mitigate the influence of unreliable and unstable estimates.

5 Convergence Analysis: Ideal Step Sizes

In this chapter we derive convergence results for SGD with the ideal step sizes introduced in Chapter 4. We show that the step sizes defined in (4.2) yield optimal convergence rates in both the interpolating and the non-interpolating regime. Although these step sizes are not accessible in practical settings, the estimators developed in Section 4.2 allow us to approximate them. Convergence guarantees for the ideal step sizes therefore motivate the use of their approximations in practice. We establish convergence results for the estimated step sizes in Chapter 7.

The following two theorems present the main convergence results of this chapter. We provide results on convergence in expectation in Theorem 5.1 and almost sure convergence in Theorem 5.2. Each theorem covers the non-interpolating case in a) and the interpolating case in b). The proofs rely on several auxiliary lemmas that are stated at the end of this chapter.

Theorem 5.1 (Convergence in Expectation)**.** *Consider a (μ, L)-feasible SOP with $\mu > 0$. Suppose that the iterates $(x_n)_{n\in\mathbb{N}}$ are generated by SGD with step sizes as specified in (4.2). Denote $V_0 = \mathbb{E}_\omega \left[\|f'_\omega(x^\star)\|^2_{\mathcal{H}^*}\right]$ and $D_n = F(x_n) - F(x^\star)$. Then, it holds:*

a) In case $V_0 > 0$, suppose that for all $\omega \in \Omega$, f_ω is L_ω-smooth for some $(\omega \mapsto L_\omega) \in L_2(\Omega)$. Then

$$C := \sup_{n\in\mathbb{N}} \mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|^2_{\mathcal{H}^*}\right] < \infty$$

almost surely, and there exists $n_0 \in \mathbb{N}$ such that

$$\mathbb{E}_{0:n-1}\left[D_n\right] \leq \frac{L}{\mu^2}\frac{C}{n - n_0}$$

for all $n > n_0$.

b) In case $V_0 = 0$, and if the SOP is pointwise (μ_ω, L_ω)-feasible, we have

$$\mathbb{E}_{0:n-1}\left[D_n\right] \leq \theta^n D_0$$

with $\theta = 1 - \frac{\mu^2}{2\,L\,L_{\max}}$.

Theorem 5.2 (Almost Sure Convergence)**.** *Consider a (μ, L)-feasible SOP with $\mu > 0$. Suppose that the iterates $(x_n)_{n\in\mathbb{N}}$ are generated by SGD with step sizes as specified in (4.2).*

a) Suppose that f_ω is L_ω smooth for some mapping $(\omega \mapsto L_\omega) \in L_2(\Omega)$. Then it holds almost surely:

$$\sum_{n=1}^{\infty} \left\|F'(x_n)\right\|_{\mathcal{H}^*}^4 \;<\; \infty.$$

In particular, we have $x_n \to x^\star$ and $\sup_{n\in\mathbb{N}} \mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|_{\mathcal{H}^*}^2\right] < \infty$ almost surely.

b) If the problem is interpolating and pointwise (μ_ω, L_ω)-feasible, then for any sequence of positive numbers (s_n) with $\sum_{n=1}^{\infty} \frac{1}{s_n} < \infty$, and almost every trajectory of SGD, there exists $N \in \mathbb{N}$, possibly depending on the trajectory of SGD, such that

$$D_n \leq s_n \theta^n D_0, \qquad \text{for all } n \geq N.$$

Here $\theta < 1$ is defined as in Theorem 5.1 b).

In order to prove these theorems, we first state several auxiliary lemmas.

Lemma 5.3. *Consider a (μ, L)-feasible SOP with $\mu > 0$. Consider the step from x_n to x_{n+1} using the step size α_n as specified in (4.2). Denote $D_n = F(x_n) - F(x^\star)$. If $x_n \neq x^\star$, then the following chain of inequalities holds:*

$$\begin{aligned}
\mathbb{E}_{\omega_n}\left[D_{n+1}\right] &\leq D_n - \frac{\alpha_n}{2}\left\|F'(x_n)\right\|_{\mathcal{H}^*}^2 && (5.1)\\
&\leq (1 - \mu\,\alpha_n)\, D_n && (5.2)\\
&\leq D_n \exp\left(-\mu\,\alpha_n\right)\\
&\leq D_n \exp\left(-2\,\frac{\mu^2}{L}\,\frac{D_n}{\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|_{\mathcal{H}^*}^2\right]}\right).
\end{aligned}$$

Proof. The first inequality follows by inserting the step sizes (4.2) into the bound (4.1). The second inequality follows from the strong convexity of F, which implies

$$\left\|F'(x_n)\right\|_{\mathcal{H}^*} \geq 2\mu(F(x_n) - F(x^\star)),$$

as shown in Lemma 2.10. The third inequality follows from the property $1 - x \leq \exp(-x)$ for all $x \in \mathbb{R}$. The last inequality again uses strong convexity of F and the definition of α_n in (4.2). □

The next two results provide a rate of convergence (to zero) for sequences $(d_n)_{n\in\mathbb{N}}$ of positive numbers satisfying a nonlinear recursive bound of the type

$$d_{n+1} \leq d_n \exp\left(-c_n d_n\right)$$

where the sequence $(c_n)_{n\in\mathbb{N}}$ is bounded from below by a positive constant $c > 0$. The first result (Lemma 5.4) is a technical preparation, the second (Lemma 5.5) states the result.

Lemma 5.4. *For every $n \in \mathbb{N}$, the inequality*

$$\frac{1}{n}\exp\left(-\frac{2\log(2)}{n+1}\right) \leq \frac{1}{n+1}$$

holds. Here, log *denotes the natural logarithm.*

Proof. The statement is equivalent to:

$$2\log(2) \geq (n+1)\log\left(1+\frac{1}{n}\right).$$

Consider the function $f(x) = (x+1)\log\left(1+\frac{1}{x}\right)$ for $x > 0$. Its derivative satisfies

$$f'(x) = \log\left(1+\frac{1}{x}\right) - \frac{1}{x} \leq 0.$$

Thus, for $n \in \mathbb{N}$, we have $2\log(2) = f(1) \geq f(n) = (n+1)\log\left(1+\frac{1}{n}\right)$, and the claim follows. □

Lemma 5.5. *Let $(d_n)_{n\in\mathbb{N}}$ and $(c_n)_{n\in\mathbb{N}}$ be sequences such that $d_n > 0$ and $c_n \geq c > 0$ for all n. Suppose that for all $n \in \mathbb{N}$, we have*

$$d_{n+1} \leq d_n \exp(-c_n\, d_n).$$

Then there exists $n_0 \in \mathbb{N}$ such that for all $n > n_0$,

$$d_n \leq \frac{2\log(2)}{c\,(n-n_0)}.$$

Proof. We provide a proof by induction. Suppose that $d_n \leq \frac{2\log(2)}{c\,(n-n_0)}$ for some $n > n_0$ for some $n_0 \in \mathbb{N}$. If already $d_n \leq \frac{2\log(2)}{c\,(n+1-n_0)}$, then the claim for d_{n+1} follows immediately. Otherwise, applying Lemma 5.4 yields

$$d_{n+1} \leq \frac{2\log(2)}{c\,(n-n_0)}\exp\left(-\frac{2\log(2)}{n+1-n_0}\right) \leq \frac{2\log(2)}{c\,(n+1-n_0)}.$$

It remains to show that there exists an initial $n_0 \in \mathbb{N}$ such that $d_{n_0+1} \leq \frac{2\log(2)}{c}$ holds. Suppose, to the contrary, that $d_n > \frac{2\log(2)}{c}$ holds for all $n \in \mathbb{N}$. Then

$$d_{n+1} \leq d_n \exp(-2\log(2)) = \frac{d_n}{4}.$$

Iterating yields $d_n \leq \frac{d_0}{4^n}$. Consequently, for $n \geq \frac{1}{\log(4)}\log\left(\frac{c\,d_0}{2\log(2)}\right)$ it holds that $d_n \leq \frac{2\log(2)}{c}$, showing that an appropriate n_0 exists. □

The following simple result deals with a special form of converging series and will be used in the proof of Theorem 5.2.

Lemma 5.6. *Let $c_0, c_1 > 0$ and consider a sequence of non-negative numbers $(a_n)_{n\in\mathbb{N}}$, such that*

$$\sum_{n=1}^{\infty} \frac{a_n^2}{c_0 + c_1 a_n} < \infty.$$

Then

$$\sum_{n=1}^{\infty} a_n^2 < \infty.$$

Proof. The given conditions imply $a_n \to 0$. Hence, there is $n_0 \in \mathbb{N}$, such that $a_n \leq 1$ for $n \geq n_0$. For such n we have

$$\frac{a_n^2}{c_0 + c_1 a_n} \geq \frac{a_n^2}{c_0 + c_1}$$

and the claim follows. □

We are now in the position to prove Theorems 5.1 and 5.2. First, we will present a proof for Theorem 5.2 a), as this result will then be used in the proof in Theorem 5.1. Finally, we present the proof of Theorem 5.2 b), as it is implied by Theorem 5.1.

Proof of Theorem 5.2 a). From (5.1) and the definition of α_n in (4.2) we obtain:

$$\mathbb{E}_{\omega_n}\left[D_{n+1}\right] \leq D_n - \frac{\|F'(x_n)\|_{\mathcal{H}^*}^4}{2L\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|_{\mathcal{H}^*}^2\right]}$$

Consider the two sequences of random variables $(z_n)_{n\in\mathbb{N}}$ and $(w_n)_{n\in\mathbb{N}}$ defined as

$$z_n = D_n \qquad \text{and} \qquad w_n = \frac{\|F'(x_n)\|_{\mathcal{H}^*}^4}{2L\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|_{\mathcal{H}^*}^2\right]}.$$

Then, z_n and w_n are $\mathcal{F}_n$-measurable, where $(\mathcal{F}_n)_{n\in\mathbb{N}}$ is the natural filtration of SGD discussed in Section 2.4.2. Hence we can apply the Robbins-Siegmund Lemma Lemma 2.13) (with $\beta_n = v_n = 0$ and obtain almost surely:

$$\sum_{n=1}^{\infty} \frac{\|F'(x_n)\|_{\mathcal{H}^*}^4}{2L\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|_{\mathcal{H}^*}^2\right]} < \infty \tag{5.3}$$

Using Proposition 3.8 a), we see that there is a constant $V_1 < \infty$, such that (3.5) holds, i.e. we have:

$$\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|_{\mathcal{H}^*}^2\right] \leq 2V_0 + (V_1 - 1)\left\|F'(x_n)\right\|_{\mathcal{H}^*}^2, \tag{5.4}$$

where $V_0 = \mathbb{E}_\omega \left[\|f'_\omega(x^\star)\|^2_{\mathcal{H}^*} \right]$. In particular, we get from (5.3) and (5.4):

$$\sum_{n=1}^{\infty} \frac{\|F'(x_n)\|^4_{\mathcal{H}^*}}{V_1 \|F'(x_n)\|^2_{\mathcal{H}^*} + V_0} < \infty$$

almost surely. By Lemma 5.6 we conclude:

$$\sum_{n=1}^{\infty} \left\|F'(x_n)\right\|^4_{\mathcal{H}^*} < \infty.$$

almost surely, which implies $\|F'(x_n)\|_{\mathcal{H}^*} \to 0$ and thus $x_n \to x^\star$. □

Proof of Theorem 5.1. a) By Theorem 5.2 a) we have

$$C := \sup_{n\in\mathbb{N}} \mathbb{E}_{\omega_n} \left[\left\|f'_{\omega_n}(x_n)\right\|^2_{\mathcal{H}^*} \right] < \infty \qquad \text{almost surely}$$

From (5.2) in Lemma 5.3 we obtain:

$$\mathbb{E}_{\omega_n} \left[D_{n+1}\right] \leq (1 - \mu\alpha_n) D_n. \tag{5.5}$$

The step sizes α_n from (4.2) satisfy almost surely:

$$\alpha_n = \frac{\|F'(x_n)\|^2_{\mathcal{H}^*}}{L\, \mathbb{E}_\omega \left[\|f'_\omega(x_n)\|^2_{\mathcal{H}^*} \right]} \geq \frac{\|F'(x_n)\|^2_{\mathcal{H}^*}}{C\, L} \geq \frac{2\,\mu\, D_n}{C\, L}$$

Thus, by (5.5)

$$\mathbb{E}_{0:n} \left[D_{n+1}\right] \leq \mathbb{E}_{0:n-1} \left[D_n \left(1 - \frac{2\,\mu^2}{C\, L} D_n \right) \right]$$

The map $x \mapsto x(1-cx)$ is concave for any non-negative c. Using Jensen's inequality (cf. [20, Theorem 1.5.1]), we thus obtain:

$$\begin{aligned} \mathbb{E}_{0:n-1} \left[D_n \left(1 - \frac{2\,\mu^2}{CL}\, D_n \right) \right] &\leq \mathbb{E}_{0:n-1} \left[D_n\right] \left(1 - \frac{2\,\mu^2}{CL} \mathbb{E}_{0:n-1} \left[D_n\right] \right) \\ &\leq \mathbb{E}_{0:n-1} \left[D_n\right] \exp\left(-\frac{2\,\mu^2}{CL} \mathbb{E}_{0:n-1} \left[D_n\right] \right). \end{aligned}$$

Thus, denoting $d_n := \mathbb{E}_{0:n-1} \left[D_n\right]$, we have:

$$d_{n+1} \leq d_n \exp\left(-\frac{2\,\mu^2}{CL}\, d_n \right)$$

and the results follows from Lemma 5.5.

b) By Proposition 3.9, we have

$$\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(x_n)\right\|^2_{\mathcal{H}^*}\right] \leq 4\,L_{\max}\,D_n + 2\,V_0.$$

Using $V_0 = 0$ and $\|F'(x_n)\|^2 \geq 2\,\mu D_n$ (see Lemma 2.10) we obtain:

$$\alpha_n = \frac{\|F'(w_n)\|^2_{\mathcal{H}^*}}{L\,\mathbb{E}_{\omega_n}\left[\left\|f'_{\omega_n}(w_n)\right\|^2_{\mathcal{H}^*}\right]} \geq \frac{2\,\mu D_n}{4\,L\,L_{\max}\,D_n} \geq \frac{\mu}{2\,L\,L_{\max}}.$$

Denoting $d_n = \mathbb{E}_{0:n-1}\left[D_n\right]$, (5.2) yields:

$$d_{n+1} \leq \mathbb{E}_{0:n-1}\left[(1-\mu\,\alpha_n)\,D_n\right] \leq \left(1 - \frac{\mu^2}{2\,L\,L_{\max}}\right) d_n$$

and thus the claimed linear convergence.

□

Proof of Theorem 5.2b). The result follows from Item b) of Theorem 5.1 combined with Theorem 2.16. □

6 Smoothing Techniques and Convergence of p-EMA

The estimators we have considered in Section 4.2 are based on noisy observations. To obtain meaningful quantities from these noisy observations, we need to smooth them in a suitable way. The observations in Section 4.2 are made along the trajectory of the stochastic gradient algorithm, and it is reasonable to assume that older observations are less informative about a local quantity than more recent ones. The smoothing (or averaging) process should hence take into account that the amount of information about a current state decays with the age of the observations. Therefore, in a weighted average over all observations, older observations should generally be assigned less weight. However, more recent observations are also prone to noise. To achieve strong convergence results for the averaged quantities, it is therefore necessary that the weights assigned to all observations (and in particular to the youngest) vanish over time.

Suppose we have noisy observations $(\widetilde{\tau}_k)_{k\in\mathbb{N}_0}$ of an unknown target $\tau(k)$, which may vary with k, and that at time n we have access to the first n observations $\widetilde{\tau}_0, \dots, \widetilde{\tau}_{n-1}$. We aim at computing a de-noised estimate $\widehat{\tau}_n$, depending only on these first n observations. For smoothing, the classical arithmetic mean or an exponential smoothing process is often employed. In the former case, the estimate $\widehat{\tau}_n^{\text{class}}$ is given by

$$\widehat{\tau}_n^{\text{class}} = \frac{1}{n}\sum_{k=0}^{n-1} \widetilde{\tau}_k, \tag{6.1}$$

which assigns every observation $\widetilde{\tau}_k$ the same weight $\frac{1}{n}$. This weight decreases over time, but is constant in k for any fixed n. Stochastic properties of the arithmetic mean have been studied extensively over many decades. If the observations $\widetilde{\tau}_k$ are independent, identically distributed (iid) random variables with finite variance, the classical law of large numbers applies (see, e.g., [10, Theorem 10.10.22]) and yields almost sure convergence of $\widehat{\tau}_n^{\text{class}}$ to the mean $\mathbb{E}[\widetilde{\tau}_1]$. More generally, results such as the Birkhoff ergodic theorem (see, e.g., [43, Theorem 2.1.3] or [28, Corollary 2.5.2]) allow for strong convergence results, if $\widetilde{\tau}_k$ are observations made along the trajectory of an ergodic dynamical system.

Assigning each observation the same weight $\frac{1}{n}$ is reasonable if all observations contain the same amount of information about the target we whish to estimate. If, however, the target or the mean of the observations $\widetilde{\tau}_k$ changes over time and only becomes stationary eventually, then more recent observations generally contain more information about the target than older ones. It thus seems reasonable to assign larger weights to more recent observations in an averaging process and consider *weighted* averages instead. A frequently employed approach is exponential moving averaging (EMA), used in the signal processing literature since the 1950s ([12, 30, 69]). Here, an initial guess $\widehat{\tau}_0^{\text{EMA}}$ and a parameter $\gamma \in (0,1)$ are fixed, and the current estimate is updated according to:

$$\widehat{\tau}_{n+1}^{\text{EMA}} = \gamma \widehat{\tau}_n^{\text{EMA}} + (1-\gamma)\widetilde{\tau}_n \tag{6.2}$$

Clearly, the weight of $\widetilde{\tau}_k$ in $\widehat{\tau}_n^{\text{EMA}}$ decays exponentially with $n-k$ and is proportional to γ^{n-k}. By adjusting γ, one can control the sensitivity of the estimate in (6.2) to changes in the distribution of the observations. Larger values of γ place less weight on the current observation $\widetilde{\tau}_n$, and more weight on the previous estimate $\widehat{\tau}_n^{\text{EMA}}$, which yields a more stable estimate, which in turn is less responsive to changes in the distribution of the observations. Smaller values of γ lead to faster adaptation of the estimate $\widehat{\tau}_n^{\text{EMA}}$ to such changes, but comes at the cost of increased noise in the estimate.

In any case, the weight assigned to the last observation $\widetilde{\tau}_n$ in the estimate $\widehat{\tau}_n^{\text{EMA}}$ is $1-\gamma$ and thus constant in n. Consequently, the noise in the observations is only scaled down by this factor and is therefore still present in the de-noised estimate $\widehat{\tau}_n^{\text{EMA}}$, preventing strong convergence guarantees for the estimate. To obtain strong convergence guarantees, it is therefore essential that the weights (in particular on the youngest observations) vanish over time, as is the case for the classical arithmetic mean (6.1). An appealing way to combine the virtues of both the arithmetic mean (6.1) and EMA (6.2) is to consider an update of the form (6.2), but with time dependent parameters $\gamma = \gamma_n$. To get vanishing weights on the youngest, most recent observation $\widetilde{\tau}_n$ in the estimate $\widehat{\tau}_{n+1}^{\text{EMA}}$, we need $\gamma_n \to 1$. We might choose

$$\gamma_n = 1 - \frac{1}{(n+1)^p}, \qquad n \geq 0,$$

for some $p \in (\frac{1}{2}, 1)$. We will refer to this averaging process as p-EMA. Formally, with a chosen initial guess $\widehat{\tau}_0^{p\text{-EMA}}$, we perform the updates:

$$\begin{aligned} \widehat{\tau}_{n+1}^{p\text{-EMA}} &= \gamma_n \widehat{\tau}_n^{p\text{-EMA}} + (1-\gamma_n)\widetilde{\tau}_n \\ &= \left(1 - \frac{1}{(n+1)^p}\right) \widehat{\tau}_n^{p\text{-EMA}} + \frac{1}{(n+1)^p}\widetilde{\tau}_n. \end{aligned} \tag{6.3}$$

With p-EMA we achieve the aforementioned goals:

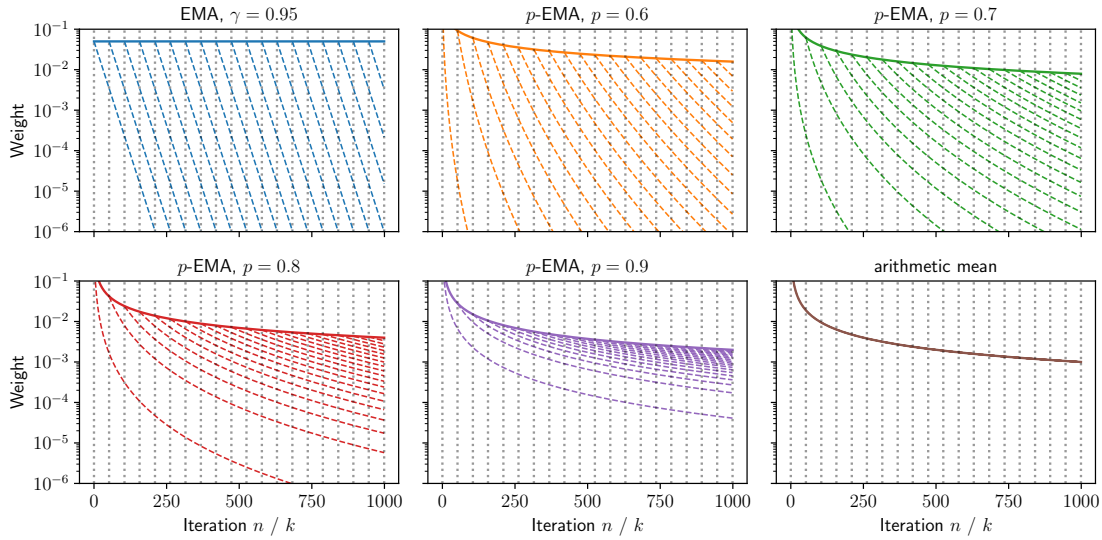


Figure 6.1: Comparison of the weights in the different averaging procedures. Weight on the youngest observation $\widetilde{\tau}_n$ in $\widehat{\tau}_{n+1}$ (solid) and weight of $\widetilde{\tau}_k$ in later averages $\widehat{\tau}_n$ (dashed).

1. The weight on the youngest observation $\widetilde{\tau}_{n+1}$ in (6.3) vanishes as $n \to \infty$, enabling the noise in $\widehat{\tau}_n^{p\text{-EMA}}$ to vanish as well.
2. For fixed n, the weight on $\widetilde{\tau}_k$ in $\widehat{\tau}_n^{p\text{-EMA}}$ is monotonically increasing in $k \leq n$, thereby assigning larger weights to more recent observations compared to older ones.

Note that the two conditions above do not contradict each other. In the first condition, n varies, whereas in the second condition n is fixed and k varies.

In each of the averaging techniques discussed so far (arithmetic mean, EMA, and p-EMA), the estimate $\widehat{\tau}_{n+1}$ is a *convex combination* of the observations $\widehat{\tau}_0, \widetilde{\tau}_1, \ldots, \widetilde{\tau}_n$. However, the techniques differ in the distribution of weights, which is visualized in Figures 6.1 and 6.2. In Figure 6.1 each subplot shows the development of the weight assigned to the youngest, most recent observation $\widetilde{\tau}_n$ in $\widehat{\tau}_{n+1}$ (solid line) in the corresponding averaging technique. Additionally, for selected values of k, each dashed line indicates the weight on $\widetilde{\tau}_k$ in $\widehat{\tau}_{n+1}$ where k is the index, where the dashed line emerges from the solid line. The selected values of k are indicated by dotted vertical lines. There are no dashed lines visible in the case of the arithmetic mean, since all observations are assigned the same weight and the dashed lines thus coincide with the solid curve. The solid curve is constant in the case of EMA, as the youngest observation always has the constant weight $1-\gamma$. Figure 6.2 shows the behavior of the weights from the perspective

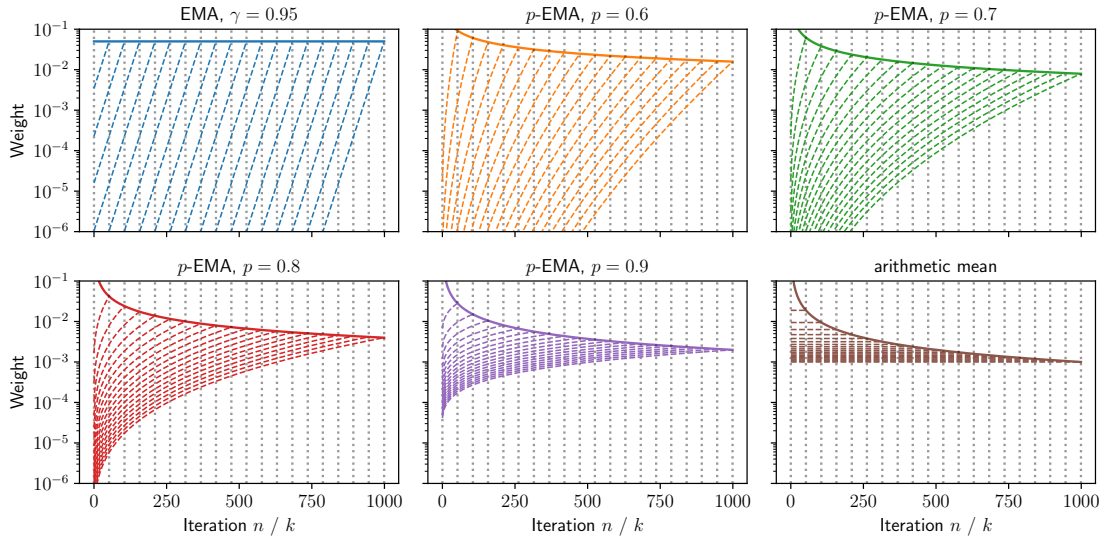


Figure 6.2: Comparison of the weights in the different averaging procedures. Weight on the youngest observation $\widetilde{\tau}_n$ in $\widetilde{\tau}_{n+1}$ (solid) and weights on previous observation $\widetilde{\tau}_k$ in given $\widehat{\tau}_n$ (dashed).

of the current estimate $\widehat{\tau}_{n+1}$. The solid line again shows the weight assigned to $\widetilde{\tau}_n$ in $\widehat{\tau}_{n+1}$. For selected values of n, the dashed lines show the weight assigned to the observation $\widetilde{\tau}_k$ in the weighted average $\widehat{\tau}_{n+1}$. The selected values of n are indicated by dotted vertical lines. In the case of p-EMA for $p \in (\frac{1}{2}, 1)$, the weight assigned to $\widetilde{\tau}_k$ in $\widehat{\tau}_{n+1}$ is strictly increasing in $k \leq n$, and it is constant in k in the case of the arithmetic mean. Qualitatively, we see that p-EMA yields an averaging technique that lies *between* EMA and the arithmetic mean.

Note that, due to the definition of p-EMA in (6.3), the initialization $\widehat{\tau}_0^{p\text{-EMA}}$ has weight $\gamma_0 = 0$ in $\widehat{\tau}_1^{p\text{-EMA}}$. Consequently, the initialization has no impact on subsequent estimates. By considering the alternative $\gamma_n = 1 - \frac{1}{(n+2)^p}$, this can easily be avoided. We consider the definition as above, as it simplifies computations, however the convergence results derived below in this chapter also apply to the alternative $\gamma_n = 1 - \frac{1}{(n+2)^p}$.

In the remainder of this chapter we present a rigorous stochastic convergence analysis for p-EMA in a general setting. This general framework will be specialized to our setting in Chapter 7, where we establish convergence for the estimators introduced in Chapter 4. We begin with the convergence analysis in Section 6.1 and then, in Section 6.2, address the necessity of the restriction $p \in (\frac{1}{2}, 1)$ that appears in the theoretical results. This chapter is based on the author's recent preprint [40].

6.1 Convergence of p-EMA

In this section we provide the convergence analysis for p-EMA in a general form. We consider a probability space $(\Gamma, \mathcal{G}, \pi)$ consisting of a set Γ, a σ-algebra $\mathcal{G}$ and a probability measure $\pi : \mathcal{G} \to [0, 1]$. First, we introduce the notion of an *averaging scheme* and show convergence results for this abstract class of weights. Later, we show that p-EMA induces an averaging scheme in this sense.

6.1.1 Averaging Schemes

Definition 6.1. *By Ψ_c we denote the set of all monotonically increasing functions $\psi : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$, such that*

$$\sum_{n=1}^{\infty} \frac{1}{n\psi(n)} < \infty.$$

For example, for $\varepsilon > 0$ we have $(x \mapsto x^{\varepsilon}) \in \Psi_c$ and $(x \mapsto \log^{1+\varepsilon}(x)) \in \Psi_c$. In the following, for a sequence $(b_n)_{n\in\mathbb{N}}$, we denote $A_n = \sum_{k=1}^{n} b_n$.

Definition 6.2. *A non-decreasing sequence $(b_n)_{n\in\mathbb{N}}$ of positive numbers is called an* averaging scheme*, if there exists $\psi \in \Psi_c$ such that for n sufficiently large*

$$b_n \leq \frac{A_n}{\psi(A_n)}.$$

Intuitively this definition ensures that in the weighted average $\frac{1}{A_n}\sum_{k=1}^{n} b_k\widetilde{\tau}_k$ the weight on the most recent observation is not too large compared to the cumulative previous weights. For example, the arithmetic mean induces an averaging scheme with $b_n = 1$, $A_n = n$, and $\psi(n) = n$.

Before we establish convergence results along trajectories under suitable dynamics, we focus on the more general case of dependent random variables. Our result on averaging schemes, stated in Theorem 6.4 below, is a consequence of generalized law of large numbers found in [42, Theorem 1], which is stated here without proof. For a sequence $(b_n)_{n\in\mathbb{N}} \subset \mathbb{R}_{>0}$ and a sequence of random variables $(X_n)_{n\in\mathbb{N}}$, we denote

$$S_n = \sum_{k=1}^{n} b_k X_k \qquad \text{and} \qquad A_n = \sum_{k=1}^{n} b_k.$$

Theorem 6.3. *Consider a sequence of non-negative random variables X_n. Let $(b_n)_{n\in\mathbb{N}}$ be a sequence of positive numbers. Suppose that the following conditions hold: $A_n \to \infty$ as $n \to \infty$,*

$$\sum_{k=m}^{n} b_k \mathbb{E}\left[X_k\right] \leq C \sum_{k=m}^{n} b_k \tag{6.4}$$

for all sufficiently large $n - m$, where C is a constant, and

$$\mathbb{E}\left[|S_n - \mathbb{E}S_n|^2\right] = O\left(\frac{A_n^2}{\psi(A_n)}\right) \tag{6.5}$$

for some function $\psi \in \Psi_c$.

Then

$$\frac{S_n - \mathbb{E}S_n}{A_n} \to 0 \quad \textit{almost surely.}$$

Proof. See [42, Theorem 1]. □

We now state a convergence result for averaging schemes.

Theorem 6.4. *Consider a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ on $(\Gamma, \mathcal{G}, \pi)$ such that:*

1. $\mathbb{E}[X_n] = \mathbb{E}[X_1] =: \eta$ *for all* $n \in \mathbb{N}$.
2. $\left|\mathbb{E}[X_n X_m] - \eta^2\right| = \rho(|n-m|)$ *for some function* $\rho : \mathbb{N}_0 \to \mathbb{R}_{\geq 0}$.
3. $\sum_{m=0}^{\infty} \rho(m) < \infty$.
4. $X_n \geq c$ *almost surely for some* $c \in \mathbb{R}$ *and all* $n \in \mathbb{N}$.

Further, suppose $(b_n)_{n\in\mathbb{N}}$ is an averaging scheme. Then:

$$\frac{1}{A_n}\sum_{k=1}^{n} b_k X_k \to \eta \qquad \pi - \textit{almost surely.}$$

Proof. By considering $X_n - c$ instead of X, we can assume that $X_n \geq 0$. We seek to apply Theorem 6.3. For this, we have to verify Equations (6.4) and (6.5). In our setting $\mathbb{E}[X_n] = \mathbb{E}[X_1]$ implies (6.4) with equality ($C = 1$). Denote $\eta = \mathbb{E}[X_1]$. To verify (6.5)

we compute:

$$\begin{aligned}
\mathbb{E}\left[|S_n - \mathbb{E}[S_n]|^2\right] &= \mathbb{E}\left[|S_n - A_n\eta|^2\right] \\
&= \sum_{k,\ell=1}^{n} b_k b_\ell \left(\int X_k X_\ell d\mu - \eta^2\right) \\
&\leq 2\sum_{m=0}^{n-1} \rho(m) \sum_{i=1}^{m} b_i b_{m-i} \\
&\leq 2\sum_{m=0}^{\infty} \rho(m) b_n A_n \\
&= 2\sum_{m=0}^{\infty} \rho(m) \frac{b_n}{A_n} A_n^2 \\
&\leq 2\sum_{m=0}^{\infty} \rho(m) \frac{A_n^2}{\psi(A_n)},
\end{aligned}$$

which verifies (6.5), as the sum is finite by assumption. Thus, by Theorem 6.3:

$$\frac{S_n - \mathbb{E}[S_n]}{A_n} \to 0 \qquad \text{almost surely.}$$

We have $\mathbb{E}[S_n] = A_n \mathbb{E}[X_1]$ and therefore $\frac{S_n - \mathbb{E}[S_n]}{A_n} = \frac{S_n}{A_n} - \mathbb{E}[X_1]$. Thus, $\frac{S_n}{A_n} \to \mathbb{E}[X_1]$ almost surely. □

The conditions 2. and 3. in Theorem 6.4 require that the correlations of the random variables decay in a summable way and are central to the proof of the result. Intuitively, these two conditions quantify how fast the random variables X_n and X_m *become* independent as $|n - m|$ grows. If the random variables $(X_n)_{n\in\mathbb{N}}$ are pairwise independent, the conditions are satisfied with $\rho(m) = 0$ for $n \geq 1$. In Chapter 7 we will verify that the observations for the estimators in Section 4.2.4 satisfy these conditions under certain assumptions with $\rho(m) = C\theta^m$ for some constant C and $\theta < 1$.

6.1.2 Properties of Averaging Schemes

The following properties of averaging schemes are largely elementary. We will need them in the analysis of the estimators in Chapter 7.

Lemma 6.5. *Let $(b_n)_{n\in\mathbb{N}}$ be an averaging scheme. For a sequence $(\tau_n)_{n\in\mathbb{N}}$ of observations, denote by $(b(\tau)_n)_{n\in\mathbb{N}}$ the sequence defined by:*

$$b(\tau)_n = \frac{1}{A_n} \sum_{k=1}^{n} b_k \tau_k,$$

i.e. the weighted average obtained by the averaging scheme applied to the first n observations. Then the following statements hold:

1. $A_n = \sum_{k=1}^{n} b_k \to \infty, n \to \infty$.
2. *The map* $(\tau_n)_{n\in\mathbb{N}} \mapsto (b(\tau)_n)_{n\in\mathbb{N}}$ *is linear.*
3. *If* $\tau_n \to 0$, *then* $b(\tau)_n \to 0$.
4. *If* (u_n) *is a second sequence of observations, then*

$$|u_n - \tau_n| \to 0 \qquad \textit{implies} \qquad |b(u)_n - b(\tau)_n| \to 0.$$

Proof. 1. By the definition of an averaging scheme the sequence (b_n) is a sequence of non-decreasing positive numbers.

2. This is an immediate consequence of the definition.

3. Let $\varepsilon > 0$. Then, there is N_0, such that $|\tau_n| < \varepsilon$ for all $n \geq n_0$. We obtain for $n \geq n_0$:

$$|b(\tau)_n| \leq \frac{1}{A_n} \left| \sum_{k=1}^{n_0} b_k \tau_k \right| + \varepsilon \frac{1}{A_n} \sum_{k=n_0+1}^{n} b_k$$

The first term on the right-hand side is a constant times $\frac{1}{A_n}$. As $A_n \to \infty$, this term vanishes with $n \to \infty$. The second term is bounded by ε.

4. This follows from 2. and 3.

□

6.1.3 p-EMA Induces an Averaging Scheme

In this subsection, we show that Theorem 6.4 can be applied to p-EMA. Recall the definition of p-EMA given in (6.3).

$$\widehat{\tau}_{n+1} = \gamma_n \widehat{\tau}_n + (1 - \gamma_n)\widetilde{\tau}_n \tag{6.6}$$

with $\gamma_n = 1 - \frac{1}{(n+1)^p}$. For any initialization $\widehat{\tau}_0$ we get, by explicitly unwrapping the recursion (6.6):

$$\begin{aligned}
\widehat{\tau}_{n+1} &= \left(1 - \frac{1}{(n+1)^p}\right)\widehat{\tau}_n + \frac{1}{(n+1)^p}\widetilde{\tau}_n \\
&= \left(1 - \frac{1}{(n+1)^p}\right)\left(1 - \frac{1}{n^p}\right)\widehat{\tau}_{n-1} \\
&\quad + \left(1 - \frac{1}{(n+1)^p}\right)\frac{1}{n^p}\widetilde{\tau}_{n-1} + \frac{1}{(n+1)^p}\widetilde{\tau}_n \\
&\vdots \\
&= \widehat{\tau}_0 \prod_{k=1}^{n+1}\left(1 - \frac{1}{k^p}\right) + \sum_{k=0}^{n}\widetilde{\tau}_k \frac{1}{(k+1)^p}\prod_{s=k+2}^{n+1}\left(1 - \frac{1}{s^p}\right) \qquad (6.7) \\
&= \sum_{k=0}^{n}\beta_{k+1}^{(n+1)}\widetilde{\tau}_k
\end{aligned}$$

with

$$\beta_k^{(n+1)} = k^{-p}\prod_{s=k+1}^{n+1}\left(1 - \frac{1}{s^p}\right).$$

Note that $\prod_{k=1}^{n+1}\left(1 - \frac{1}{k^p}\right) = 0$ and therefore the first addend in (6.7) vanishes. A crucial step is to factor $\beta_k^{(n+1)}$ into a part depending only on $n+1$, and a part only depending on k. By expanding the product we obtain:

$$\beta_k^{(n+1)} = \left[\prod_{s=2}^{n+1}\left(1 - \frac{1}{s^p}\right)\right]\left[k^{-p}\prod_{s=2}^{k}\left(1 - \frac{1}{s^p}\right)^{-1}\right] \qquad (6.8)$$

We define β_k to be the right factor in (6.8):

$$\beta_k := k^{-p}\prod_{s=2}^{k}\left(1 - \frac{1}{s^p}\right)^{-1} \qquad (6.9)$$

The estimate is $\widehat{\tau}_n$ is a convex combination of the observations $\widetilde{\tau}_0, \ldots, \widetilde{\tau}_{n-1}$. Therefore, we have $\sum_{k=1}^{n}\beta_k^{(n)} = 1$, and thus

$$\Lambda_n := \sum_{k=1}^{n}\beta_k = \left[\prod_{s=2}^{n+1}\left(1 - \frac{1}{s^p}\right)\right]^{-1}.$$

Thus, a candidate for an averaging scheme is $(\beta_n)_{n\in\mathbb{N}}$, and we can write the p-EMA average $\widehat{\tau}_{n+1}^{p\text{-EMA}}$ in the form

$$\widehat{\tau}_n = \frac{1}{\Lambda_n}\sum_{k=1}^{n}\beta_k\widetilde{\tau}_{k-1}.$$

We note that we have $\frac{\beta_n}{\Lambda_n} = n^{-p}$, as this is the weight of $\widetilde{\tau}_{n-1}$ in $\widehat{\tau}_n$ in p-EMA. In this weighted average, there is, as desired, more weight on younger observations:

Lemma 6.6. *For $p < 1$, the sequence $(\beta_n)_{n\in\mathbb{N}}$ is monotonically increasing.*

Proof. Since $\beta_k \neq 0$ for all k, it suffices to show that $\frac{\beta_k}{\beta_{k+1}} < 1$ for all k. By (6.9) we have

$$\frac{\beta_k}{\beta_{k+1}} = \frac{k^{-p}}{(k+1)^{-p}}\left(1 - \frac{1}{(k+1)^p}\right) = \frac{(k+1)^p - 1}{k^p}.$$

The expression above is < 1 if and only if

$$(k+1)^p < k^p + 1,$$

which holds for all $p < 1$. □

We will use the following notation.

Definition 6.7. *Let $\mathcal{S}$ be some set and $f, g : \mathcal{S} \to \mathbb{R}$ be two functions. Then, we write*

$$f(s) \lesssim g(s),$$

if there is a uniform constant $c \in \mathbb{R}_{\geq 0}$, independent of $s \in \mathcal{S}$, such that $f(s) \leq cg(s)$ for all $s \in \mathcal{S}$.

The following lemma prepares parts of the proof the main result of this section.

Lemma 6.8. *Consider a differentiable function $h : \mathbb{R} \to \mathbb{R}$ such that $h' \geq 0$. Then, for any $n, m \in \mathbb{N}$ we have*

$$\sum_{k=n}^{m} h(k) \leq \int_n^{m+1} h(x)\,\mathrm{d}x.$$

Proof. This follows immediately from a rectangular approximation of the integral. We compute

$$\begin{aligned}\int_n^{m+1} h(x)\,\mathrm{d}x &= \sum_{k=n}^{m} \int_k^{k+1} h(x)\,\mathrm{d}x \\ &\geq \sum_{k=n}^{m} \int_k^{k+1} h(k)\,\mathrm{d}x \\ &= \sum_{k=n}^{m} h(k).\end{aligned}$$

□

The following result, whose proof is surprisingly involved, shows that for appropriate values of p the sequence $(\beta_n)_{n\in\mathbb{N}}$ defined in (6.9) is indeed an averaging scheme.

Theorem 6.9. *For $p \in (\frac{1}{2}, 1]$, there exists $\varepsilon > 0$ such that, for n sufficiently large,*

$$\beta_n \leq \frac{\Lambda_n}{\log^{1+\varepsilon}(\Lambda_n)}.$$

In particular, $(\beta_n)_{n\in N}$ is an averaging scheme in the sense of Definition 6.2 with $\psi(x) = \log^{1+\varepsilon}(x)$.

We comment on the necessity of the condition $p \in (\frac{1}{2}, 1]$ in Section 6.2.

Proof. Note that $\frac{\beta_n}{\Lambda_n} = n^{-p}$, since $\frac{\beta_n}{\Lambda_n}$ is the weight of $\widetilde{\tau}_n$ in $\widehat{\tau}_n$ obtained by p-EMA. We will show that

$$\lim_{n\to\infty} \frac{\log(\Lambda_n)}{n^{\frac{p}{1+\varepsilon}}} = 0, \tag{6.10}$$

for any $\varepsilon \in \left(0, \frac{2p-1}{1-p}\right)$ if $p < 1$, and for any $\varepsilon > 0$ when $p = 1$. This yields the claim, as $\Lambda_n \to \infty, n \to \infty$. The proof of (6.10) will be given in multiple lemmas and is structured as follows.

1. We derive a differentiable function $\widetilde{\Lambda} : \mathbb{R}_{>0} \to \mathbb{R}$, such that $\Lambda_n \lesssim c_a + \widetilde{\Lambda}(n)$ for some constant c_a (Lemma 6.10).

2. We show that the limit in (6.10) agrees with the limit

$$\lim_{y\to\infty} \frac{1+\varepsilon}{p} \frac{(y+1)^{-p}}{y^{\frac{p}{1+\varepsilon}-1}} \frac{g(y)}{c_a + \widetilde{\Lambda}(y)},$$

   where $g : \mathbb{R}_{>0} \to \mathbb{R}$ is a suitable function. We show that the second factor converges to zero (Lemma 6.11), while the third factor converges to one (Lemma 6.12).

Lemma 6.10. *Define the mapping*

$$\begin{aligned} \widetilde{\Lambda} : \mathbb{R}_{>0} &\to \mathbb{R} \\ y &\mapsto \int_2^{y+1} s^{-p} \exp\left(\int_2^{s+1} \log\left(\frac{\tau^p}{\tau^p - 1}\right) \mathrm{d}\tau\right) \mathrm{d}s. \end{aligned}$$

Then $\widetilde{\Lambda}$ is monotonically increasing, and there exists an additive constant c_a such that

$$\Lambda_n \lesssim c_a + \widetilde{\Lambda}(n).$$

Proof of Lemma 6.10. Observe that the mapping $y \mapsto \frac{y^p}{y^p-1}$ is monotonically decreasing in $y > 1$. The same therefore holds for $\log\left(\frac{y^p}{y^p-1}\right)$. In particular, we have

$$\sum_{j=2}^{k} \log\left(\frac{j^p}{j^p-1}\right) \leq \log\left(\frac{2^p}{2^p-1}\right) + \int_2^k \log\left(\frac{\tau^p}{\tau^p-1}\right) \mathrm{d}\tau.$$

Thus,

$$\prod_{j=2}^{k} \frac{j^p}{j^p-1} = \exp\left(\sum_{j=2}^{k} \log\left(\frac{j^p}{j^p-1}\right)\right) \lesssim \exp\left(\int_2^k \log\left(\frac{\tau^p}{\tau^p-1}\right) \mathrm{d}\tau\right).$$

Consider the function

$$h(y) := y^{-p} \exp\left(\int_2^y \log\left(\frac{\tau^p}{\tau^p-1}\right) \mathrm{d}\tau\right).$$

We compute

$$h'(y) = y^{-p}\left(-py^{-1} + \log\left(\frac{y^p}{y^p-1}\right)\right) \exp\left(\int_2^y \log\left(\frac{\tau^p}{\tau^p-1}\right) \mathrm{d}\tau\right).$$

Using $\log(1+x) \geq \frac{x}{1+x}$ and $\frac{y^p}{y^p-1} = 1 + \frac{1}{y^p-1}$, one obtains

$$-py^{-1} + \log\left(1 + \frac{1}{y^p-1}\right) \geq -py^{-1} + \frac{1}{y^p-1}\frac{1}{1+\frac{1}{y^p-1}} = -py^{-1} + y^{-p}$$

and thus $h'(y) \geq 0$ for all sufficiently large y. Furthermore,

$$\beta_k = k^{-p} \prod_{s=2}^{k} \left(1 - \frac{1}{s^p}\right)^{-1} = k^{-p} \prod_{s=2}^{k} \frac{s^p}{s^p-1} \lesssim h(k).$$

Therefore, by Lemma 6.8, there exists a constant c_a such that

$$\Lambda_n \lesssim \sum_{k=1}^{n} h(k) \leq c_a + \int_2^{n+1} h(y)\,\mathrm{d}y = c_a + \widetilde{\Lambda}(n).$$

□

Next, we state an additional technical lemma.

Lemma 6.11. *It holds:*

$$\lim_{y\to\infty} \frac{(y+1)^{-p}}{y^{\frac{p}{1+\varepsilon}-1}} = 0$$

Proof of Lemma 6.11. The claim is immediate in the case $p = 1$. Otherwise, the claimed convergence is equivalent to the convergence of

$$\frac{y^{-p}}{y^{\frac{p}{1+\varepsilon}-1}} = y^{-p(1+\frac{1}{1+\varepsilon})+1}$$

to zero, and consequently equivalent to $-p(1+\frac{1}{1+\varepsilon})+1 < 0$. We compute:

$$\begin{aligned} & -p\left(1+\tfrac{1}{1+\varepsilon}\right)+1 && < 0 \\ \iff \quad & p(2+\varepsilon) && > 1+\varepsilon \\ \iff \quad & \varepsilon(1-p) && < 2p-1 \end{aligned}$$

By assumption, we have $0 < \varepsilon < \frac{2p-1}{1-p}$, and therefore the last inequality holds, which completes the proof. □

Lemma 6.12. *Define*

$$g(y) = \exp\left(\int_2^{y+1} \log\left(\frac{\tau^p}{\tau^p - 1}\right) \mathrm{d}\tau\right),$$

so that $\widetilde{\Lambda}(y) = \int_2^{y+1} s^{-p} g(s)\,\mathrm{d}s$. *Then*

$$\lim_{y\to\infty} \frac{g(y)}{\widetilde{\Lambda}(y)} = 1. \tag{6.11}$$

Proof of Lemma 6.12. We note that

$$\log\left(\frac{y^p}{y^p-1}\right) = \log\left(1+\frac{1}{y^p-1}\right).$$

Therefore, using $\frac{1}{1+x} \le \log(1+x)$,

$$\frac{1}{y^p} = \frac{1}{y^p-1}\frac{1}{1+\frac{1}{y^p-1}} \le \log\left(\frac{y^p}{y^p-1}\right) \le \frac{1}{y^p-1}$$

Using this, we see that:

$$g(y) \gtrsim \exp\left(\int_2^{y+1} \tau^{-p}\,\mathrm{d}\tau\right) \to \infty, \qquad y \to \infty.$$

We have $\lim_{y\to\infty} \widetilde{\Lambda}(y) = \infty$ as well, and therefore examine the limit

$$\lim_{y\to\infty} \frac{g'(y)}{\widetilde{\Lambda}'(y)} \tag{6.12}$$

which agrees with the limit in (6.11) by L'Hôpital's rule. We compute

$$g'(y) = \log\left(\frac{(y+1)^p}{(y+1)^p - 1}\right) g(y)$$

and

$$\widetilde{\Lambda}'(y) = (y+1)^{-p} g(y+1).$$

Since $g(y+1) - g(y) \to 0$ as $y \to \infty$, and $g(y) \to \infty$, we obtain

$$\frac{g(y)}{g(y+1)} \to 1, \qquad y \to \infty. \tag{6.13}$$

It is well known that

$$\lim_{y\to\infty} y \log\left(1 + \frac{1}{y}\right) = 1,$$

therefore:

$$\frac{\log\left(\frac{(y+1)^p}{(y+1)^p - 1}\right)}{(y+1)^{-p}} \to 1, \qquad y \to \infty.$$

Combining the above limits yields

$$\frac{g'(y)}{\widetilde{\Lambda}'(y)} = \frac{\log\left(\frac{(y+1)^p}{(y+1)^p - 1}\right)}{(y+1)^{-p}} \frac{g(y)}{g(y+1)} \to 1, \qquad y \to \infty,$$

This concludes the proof of Lemma 6.12. □

Define $\widehat{\Lambda}(y) = c_a + \widetilde{\Lambda}(y)$. Then, by Lemma 6.10 there exists a constant c_m (due to $\lesssim$ in the results above) such that

$$0 \le \frac{\log(\Lambda_n)}{n^{\frac{p}{1+\varepsilon}}} \le \frac{\log(c_m) + \log(\widehat{\Lambda}(n))}{n^{\frac{p}{1+\varepsilon}}}. \tag{6.14}$$

We have $\log(\widehat{\Lambda}(y)) \to \infty$ and $y^{\frac{p}{1+\varepsilon}} \to \infty$ as $y \to \infty$. Thus, the limit on the right-hand side of (6.14) exists if the limit

$$\lim_{y\to\infty} \frac{\widehat{\Lambda}'(y)}{\widehat{\Lambda}(y)\frac{p}{1+\varepsilon} y^{\frac{p}{1+\varepsilon}-1}}$$

exists, and in this case the limits agree by L'Hôpital's rule. We note that

$$\widehat{\Lambda}'(y) = \widetilde{\Lambda}'(y) = (y+1)^{-p} g(y+1).$$

Therefore,

$$\frac{\widehat{\Lambda}'(y)}{\widehat{\Lambda}(y)\frac{p}{1+\varepsilon} y^{\frac{p}{1+\varepsilon}-1}} = \frac{1+\varepsilon}{p} \frac{(y+1)^{-p}}{y^{\frac{p}{1+\varepsilon}-1}} \frac{g(y+1)}{c_a + \widetilde{\Lambda}(y)}.$$

The first factor is constant. The second factor tends to zero by Lemma 6.11. The third factor tends to one by Lemma 6.12 and (6.13), since $g(y)$ and $\widetilde{\Lambda}(y)$ both diverge to infinity and c_a is a constant. Hence, by (6.14), we conclude that

$$\lim_{n\to\infty} \frac{\log(\Lambda_n)}{n^{\frac{p}{1+\varepsilon}}} = 0,$$

which completes the proof of Theorem 6.9. □

We finally obtain the following consequence.

Corollary 6.13. *Consider a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ on $(\Gamma, \mathcal{G}, \pi)$ such that:*

1. $\mathbb{E}[X_n] = \mathbb{E}[X_1] =: \eta$ *for all* $n \in \mathbb{N}$.
2. $\mathbb{E}[X_n X_m] - \eta^2 = \rho(|n-m|)$ *for some function* $\rho : \mathbb{N}_0 \to \mathbb{R}$.
3. $\sum_{m=0}^{\infty} |\rho(m)| < \infty$.
4. $X_n \geq c$ *almost surely for some* $c \in \mathbb{R}$ *and all* $n \in \mathbb{N}$.

Let $(\widehat{X}_n)_{n\in\mathbb{N}}$ be the sequence obtained by applying p-EMA with $p \in (\frac{1}{2}, 1]$ to the sequence $(X_n)_{n\in\mathbb{N}}$. Then

$$\widehat{X}_n \to \eta \qquad \textit{almost surely.}$$

Proof. Apply Theorem 6.4 together with Theorem 6.9. □

6.2 On the Condition $p \in (\frac{1}{2}, 1]$

We have imposed the condition $p \in (\frac{1}{2}, 1]$, in order to show that p-EMA induces an averaging scheme in the sense of Definition 6.2, and thus obtain convergence from Theorem 6.4. In fact, in the case $p = 1$ we have $\beta_n \equiv 1$ and $\Lambda_n = n$. Thus,

$$\tau_n = \frac{1}{n} \sum_{k=1}^{n} X_k.$$

In this case, almost sure convergence is already known from classical results such as the strong law of large numbers or Birkhoff's ergodic theorem.

In this section, we discuss the necessity of the restriction $p \in (\frac{1}{2}, 1]$. We provide formal counterexamples to Corollary 6.13 for independent and identically distributed (iid) random variables for $p > 1$ and $p < \frac{1}{2}$.

The behavior of the weights of p-EMA with p outside the interval $(\frac{1}{2}, 1]$ is depicted in Figures 6.3 and 6.4 (see also Figures 6.1 and 6.2 and the description of the figures in the introduction to this chapter). In the case $p > 1$, we observe in Figure 6.3 that older observations are assigned a *larger* weight compared to younger observations, which is counterintuitive. This corresponds to monotonically decreasing dashed lines in Figure 6.4 for the cases $p > 1$. In the case $p > 1$ we also note that, at any iteration, the sum of weights assigned to *all* subsequent observations remains uniformly bounded. We will use this fact in Section 6.2.1 to show that we no longer have almost sure convergence, even for iid observations.

In the case $p < \frac{1}{2}$, the distribution of the weights shown in Figures 6.3 and 6.4 qualitatively not differs significantly from the case $p \in (\frac{1}{2}, 1]$ in Figures 6.1 and 6.2. We will see in Section 6.2.2, that the weights assigned to younger observations do not decay fast enough, while the weights on older observations decay too quickly to allow for almost sure convergence. This corresponds to dashed lines that are too *steep* in Figures 6.3 and 6.4.

The case $p = \frac{1}{2}$ remains unclear. In this case, the weights assigned to the most recent observations are not square summable, a property that is evident for $p > \frac{1}{2}$. We believe that this property is crucial for almost sure convergence of the averaging technique. However, the counterexample for the case $p < \frac{1}{2}$ provided in Section 6.2.2 does not apply to this boundary case.

6.2.1 The case $p > 1$

If $p > 1$, almost sure convergence can no longer be expected, even if all observations are iid. To see this, first observe that for any bounded sequence of observations $\widetilde{\tau}_n$, the sequence of differences $|\tau_{n+1} - \tau_n|$ is summable:

$$\begin{aligned} |\tau_{n+1} - \tau_n| &= \frac{|\Lambda_n S_{n+1} - \Lambda_{n+1} S_n|}{\Lambda_n \Lambda_{n+1}} \\ &= \frac{|\Lambda_n (S_n + \beta_{n+1}\widetilde{\tau}_n) - (\Lambda_n + \beta_{n+1}) S_n|}{\Lambda_n \Lambda_{n+1}} \\ &\leq \frac{\beta_n \, |\widetilde{\tau}_n|}{\Lambda_n} + \frac{\beta_{n+1}}{\Lambda_{n+1}} \frac{|S_n|}{\Lambda_n} \end{aligned}$$

If $(\widetilde{\tau}_n)$ is bounded, then so is $\frac{S_n}{\Lambda_n}$. The identity $\frac{\beta_n}{\Lambda_n} = n^{-p}$ implies that $|\tau_{n+1} - \tau_n|$ is summable if $p > 1$. A concrete counterexample where almost sure convergence fails can be constructed as follows. Choose N_0, such that

$$\sum_{n=N_0+1}^{\infty} n^{-p} < \frac{1}{4}.$$

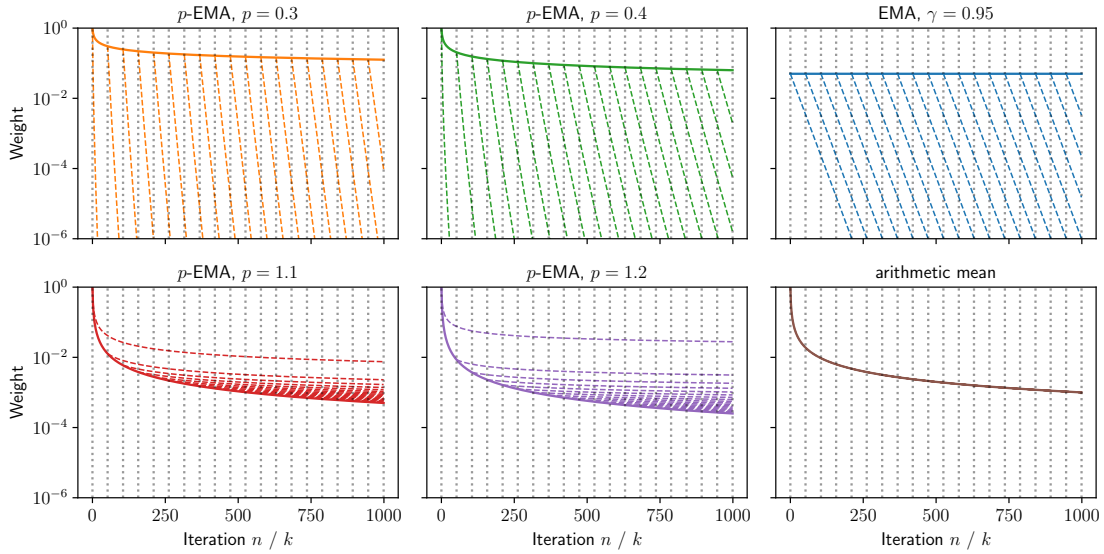


Figure 6.3: Comparison of weights for p-EMA with p outside the admissible interval $(\frac{1}{2}, 1]$. Weight on the youngest observation $\widetilde{\tau}_n$ in $\widetilde{\tau}_{n+1}$ (solid) and weight of $\widetilde{\tau}_k$ in later averages $\widehat{\tau}_n$ (dashed).

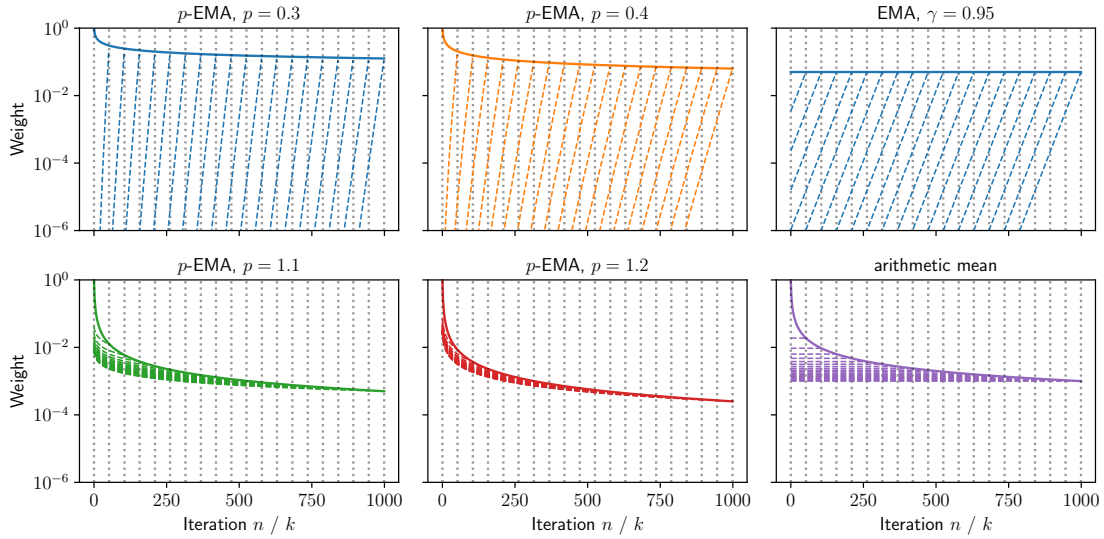


Figure 6.4: Comparison of weights for p-EMA with p outside the admissible interval $(\frac{1}{2}, 1]$. Weight on the youngest observation $\widetilde{\tau}_n$ in $\widetilde{\tau}_{n+1}$ (solid) and weights on previous observation $\widetilde{\tau}_k$ in given $\widehat{\tau}_n$ (dashed).

Consider a sequence of iid random variables X_n, such that

$$P(X_n = 1) = P(X_n = -1) = \frac{1}{2}.$$

Then the event

$$A = \{X_1 = \cdots = X_{N_0} = 1\}$$

has probability $2^{-N_0} > 0$. However, on A we do not have convergence of $\frac{S_n}{\Lambda_n}$ to $\mathbb{E}[X_1] = 0$. Indeed,

$$\frac{S_n}{\Lambda_n} = \tau_{N_0} + \tau_n - \tau_{N_0} \geq 1 - 2\sum_{n=N_0+1}^{\infty} n^{-p} > \frac{1}{2} \qquad \forall n > N_0.$$

6.2.2 The case $p < \frac{1}{2}$

For $p < \frac{1}{2}$, there exists $s > 3$, such that $p(1-s) > -1$. Consider a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ taking values in $[1, \infty)$, independently and identically distributed according to the density function

$$f(x) = \frac{1}{I_s} x^{-s},$$

where $I_s = \int_1^\infty x^{-s}\,\mathrm{d}x$. As $s > 3$, these random variables have finite first and second moment. In particular, they satisfy the assumptions of Theorem 6.4 and Corollary 6.13. Then we have

$$P(X_{n-1} \geq 2n^p) = \frac{1}{I_s}\int_{2n^p}^\infty x^{-s}\,\mathrm{d}x = \frac{1}{I_s}\frac{2^{p(1-s)}}{s-1} n^{p(1-s)}.$$

From $p(1-s) > -1$ we conclude that

$$\sum_{n=1}^{\infty} P(X_{n-1} \geq 2n^p) = \frac{2^{p(1-s)}}{I_s}\frac{1}{s-1}\sum_{n=1}^{\infty} n^{p(1-s)} = \infty.$$

All events $A_n := \{X_{n-1} \geq 2n^p\}$ are independent. Thus, by the second Borel-Cantelli lemma (see [38, Theorem 2.7]), infinitely many of them occur with probability one. We further have

$$\eta = \mathbb{E}[X_{n-1}] = \mathbb{E}[X_1] = \frac{1}{I_s}\int_1^\infty x^{1-s}\,\mathrm{d}x = \frac{s-1}{s-2} = 1 + \frac{1}{s-2} < 2.$$

However, on A_n we have, for the estimate $\widehat{\tau}_n$ obtained by p-EMA with observations X_n,

$$\widehat{\tau}_n = \gamma_n\widehat{\tau}_{n-1} + (1-\gamma_n)X_{n-1} \geq 1 - \frac{1}{n^p} + \frac{1}{n^p}X_{n-1} \geq 1 - \frac{1}{n^p} + \frac{1}{n^p}2n^p = 3 - \frac{1}{n^p}.$$

Here we used that $\widehat{\tau}_{n-1} \geq 1$, since all observations satisfy $X_k \geq 1$ (see Lemma 6.5), and that $X_{n-1} \geq 2n^p$ on A_n. Thus, $\widehat{\tau}_n$ will leave any sufficiently small ε-ball around $\eta < 2$ with probability one infinitely often, which contradicts almost sure convergence.

7 Convergence Analysis: Estimated Step Sizes

Besides the convergence result for the theoretical algorithm using the ideal step sizes discussed in Chapter 5, convergence properties of the *practical* algorithm using our estimators are of considerable interest. In this chapter, we present a convergence theory for the estimators in a specific scenario, which substantiates our motivation for employing these estimators. In this theory, the invariant measure $\nu^\star$ defined in Section 3.4 plays a central role, together with the convergence guarantees for the smoothing technique p-EMA described in Chapter 6 that we use to obtain the estimators. The interpretation of the results is that the adaptive step sizes indeed recognize the non-interpolating setting and eventually reduce the step sizes, while keeping them bounded away from zero in the interpolating setting.

We begin with some technical preparations in the next section, which establishes the concept of synchronizing trajectories of SGD. As a consequence, we will be able to justify the summable decay of correlation assumed in Theorem 6.4 and Corollary 6.13 and obtain convergence of the estimated quantities. Furthermore, we will characterize the resulting limits of $\widehat{g}_n$ and $\widehat{\sigma}_n^2$ defined in Section 4.2 and the resulting estimated step size $\widehat{\alpha}_n$.

7.1 Technical Preparations

Synchronization of Trajectories

Consider two initial points x_0 and $\widetilde{x}_0$ and the corresponding trajectories given by the recursions

$$x_{n+1} = x_n - \alpha \nabla f_{\omega_n}(x_n) \qquad \text{and} \qquad \widetilde{x}_{n+1} = \widetilde{x}_n - \alpha \nabla f_{\omega_n}(\widetilde{x}_n),$$

respectively. Here, in each iteration, $\omega_n \sim \mathbb{P}$ is sampled, and *both* updates are performed with *the same* sampled function. We refer to such a pair as a *pair of simultaneous trajectories*. If we treat x_0 and $\widetilde{x}_0$ as fixed, we may view x_n and $\widetilde{x}_n$ as functions of $\omega_0, \dots, \omega_{n-1}$. The following result shows that the expected distance between the trajectories decays at a linear rate.

Theorem 7.1. *Consider a pointwise (μ_ω, L_ω)-feasible SOP, and suppose that F is μ-strongly convex for some $\mu > 0$. Then, for any $\alpha \leq \frac{1}{L_{\max}}$, we have*

$$\mathbb{E}_{0:n-1}\left[\|x_n - \widetilde{x}_n\|_{\mathcal{H}}^2\right] \leq (1-\mu\alpha)^n \|x_0 - \widetilde{x}_0\|_{\mathcal{H}}^2 .$$

Proof. Using Equation (3.22) we obtain

$$\begin{aligned}
&\mathbb{E}_{0:n-1}\left[\|x_n - \widetilde{x}_n\|_{\mathcal{H}}^2\right] \\
&\quad \leq \mathbb{E}_{0:n-2}\left[\|x_{n-1} - \widetilde{x}_{n-1}\|_{\mathcal{H}}^2 - \alpha\left(F'(x_{n-1}) - F'(\widetilde{x}_{n-1}), x_{n-1} - \widetilde{x}_{n-1}\right)\right] \\
&\quad \leq (1-\mu\alpha)\mathbb{E}_{0:n-2}\left[\|x_{n-1} - \widetilde{x}_{n-1}\|_{\mathcal{H}}^2\right].
\end{aligned}$$

Iterating this bound proves the claim. □

Theorem 7.1 provides a bound in expectation. For individual pairs of simultaneous trajectories, we obtain the following almost sure statement.

Corollary 7.2. *For almost every pair of simultaneous trajectories of SGD, and for every sequence of positive numbers $(s_n)_{n\in\mathbb{N}}$ such that $\sum_{n=1}^{\infty} \frac{1}{s_n} < \infty$, there exists $N \in \mathbb{N}$ such that*

$$\|x_n - \widetilde{x}_n\|_{\mathcal{H}}^2 \leq s_n(1-\mu\alpha)^n \|x_0 - \widetilde{x}_0\|_{\mathcal{H}}^2 \qquad \text{for all } n \geq N.$$

Proof. Apply Theorem 2.16 to the convergence established in Theorem 7.1. □

Such behavior is known as *synchronization* or *synchronization by noise* in the context of random dynamical systems, and has been studied, for example, in [22, 53, 54]. The two results above demonstrate that each pair of simultaneous trajectories converges toward one another at a linear rate. This is particularly relevant, as it quantifies the speed of *mixing* induced by the dynamics of SGD. Informally, the results show that the initialization is forgotten exponentially fast, and that the iterates are largely determined by the sampled functions f_{ω_n} used during the iteration of SGD. This perspective will be refined in Theorem 7.3 below.

Notation

We now introduce notation used throughout the remainder of the section. We denote

$$\overline{\Omega} = \mathcal{H} \times \Omega^{\mathbb{N}}.$$

Each element $z = (x_0, \omega_0, \omega_1, \dots) \in \overline{\Omega}$ can be interpreted as one trajectory of SGD via the transition map

$$\begin{aligned} \theta : \overline{\Omega} &\to \overline{\Omega} \\ z &\mapsto \theta z = (x - \alpha \nabla f_{\omega_0}(x), \omega_1, \omega_2, \dots). \end{aligned} \tag{7.1}$$

Recall that, under suitable assumptions (see Corollary 3.22), there exists a unique probability measure $\nu^\star$ on $\mathcal{H}$ that is invariant under the dynamics of SGD. Equipping $\overline{\Omega}$ with the probability measure $\pi = \nu^\star \times \mathbb{P}^{\mathbb{N}}$, we have that π is invariant under θ ([1, Theorem 2.1.7]).

Given $z = (x, \omega_0, \dots) \in \mathcal{H} \times \Omega^{\mathbb{N}}$, we write z_1 for the component $x \in \mathcal{H}$ and z_2 for the component $\boldsymbol{\omega} = (\omega_0, \omega_1, \dots) \in \Omega^{\mathbb{N}}$. For $l, k \in \mathbb{N} \cup \{\infty\}$, we denote

$$z_{2,[l:k]} = (\omega_i)_{l \le i \le k-1}.$$

Finally, we denote by

$$\begin{aligned} x_n : \mathcal{H} \times \Omega^n &\to \mathcal{H} \\ (x_0, \omega_0, \dots, \omega_{n-1}) &\mapsto x_n(x_0, \omega_0, \dots, \omega_{n-1}) \end{aligned}$$

the mapping which maps the initialization x_0 and the sequence $(\omega_i)_{i \in \mathbb{N}}$ to the nth iterate of SGD. We have

$$x_n(x_0, \omega_0, \dots, \omega_{n-1}) = \left(\theta^n(x_0, \omega_0, \dots)\right)_1 .$$

Decay of Correlations

The next result forms the foundation for the convergence of the estimators by identifying a class of functions exhibiting a summable (linear) decay of correlations under SGD. Later, we will use this result to show that the observables we have used for the estimators possess summable auto-correlations.

Theorem 7.3. *Let $\nu^\star$ denote the invariant measure of SGD with constant step size α. Consider $g, h \in L_2(\overline{\Omega})$, where $\overline{\Omega}$ is equipped with the probability measure $\pi := \nu^\star \times \mathbb{P}^{\mathbb{N}}$. Define the* coefficient of correlation *as*

$$\mathrm{Cor}_n(g, h) = \int_{\overline{\Omega}} g \cdot (h \circ \theta^n) \, \mathrm{d}\pi - \int_{\overline{\Omega}} g \, \mathrm{d}\pi \int_{\overline{\Omega}} h \, \mathrm{d}\pi.$$

Suppose additionally that $h(z)$ is η-Hölder continuous in its first argument for some $\eta \in (0, 1]$, i.e., for $z = (x, \omega_0, \omega_1, \dots)$ and $\widetilde{z} = (\widetilde{x}, \omega_0, \omega_1, \dots)$ we have

$$|h(z) - h(\widetilde{z})| \le C_h \, \|x - \widetilde{x}\|_{\mathcal{H}}^{\eta}$$

for almost every x and $\widetilde{x}$.

Further, suppose that for some $k \in \mathbb{N}$, the function g depends only on the first k entries of $(\omega_0, \omega_1, \dots)$, i.e. there exists a function $\widetilde{g} : \mathcal{H} \times \Omega^k \to \mathbb{R}$ such that

$$g(z) = \widetilde{g}(x, \omega_0, \dots, \omega_{k-1}), \qquad z = (x, \omega_0, \dots)$$

for every $z \in \overline{\Omega}$. Suppose also that $\widetilde{g}$ is essentially bounded, meaning that there exists $\|\widetilde{g}\|_{L_\infty}$ such that

$$|\widetilde{g}| \leq \|\widetilde{g}\|_{L_\infty} \qquad (\nu^\star \times \mathbb{P}^n) \text{ - almost everywhere.}$$

Then for all $n \geq k$, we have

$$|\mathrm{Cor}_n(g, h)| \leq (1 - \mu\alpha)^{\frac{(n-k)\eta}{2}} \|\widetilde{g}\|_{L_\infty} C_h C_{k,\eta},$$

where

$$C_{k,\eta} = \int_{\mathcal{H}\times\Omega^k} \int_{\mathcal{H}\times\Omega^k} \left\| x_k(x, \boldsymbol{\omega}_{[0:k]}) - x_k(\bar{x}, \overline{\boldsymbol{\omega}}_{[0:k]}) \right\|_{\mathcal{H}}^{\eta} \\ \mathrm{d}(\nu^\star \times \mathbb{P}^k)(\bar{x}, \overline{\boldsymbol{\omega}}_{[0:k]})\, \mathrm{d}(\nu^\star \times \mathbb{P}^k)(x, \boldsymbol{\omega}_{[0:k]}).$$

Proof. We first compute

$$\begin{aligned} \int_{\overline{\Omega}} g \cdot (h \circ \theta^n)\, \mathrm{d}\pi &= \int_{\mathcal{H}\times\Omega^{\mathbb{N}}} \widetilde{g}(z_1, z_{2,[0:k]}) h((\theta^n z)_1, z_{2,[n:\infty]})\, \mathrm{d}\pi(z) \\ &= \int_{\mathcal{H}\times\Omega^k} \widetilde{g}(x, \boldsymbol{\omega}_{[0:k]}) \int_{\Omega^{n-k}} \int_{\Omega^{\mathbb{N}}} h(x_n(x, (\boldsymbol{\omega}_{[0:k]}, \boldsymbol{\omega}_{[k:n]})), \boldsymbol{\omega}_{[n:\infty]}) \\ &\qquad \mathrm{d}\mathbb{P}^{\mathbb{N}}(\boldsymbol{\omega}_{[n:\infty]})\, \mathrm{d}\mathbb{P}^{n-k}(\boldsymbol{\omega}_{[k:n]})\, \mathrm{d}(\nu^\star \times \mathbb{P}^k)(x, \boldsymbol{\omega}_{[0:k]}). \end{aligned}$$

By invariance of π, we obtain

$$\begin{aligned} \int_{\overline{\Omega}} h\, \mathrm{d}\pi &= \int_{\overline{\Omega}} h \circ \theta^n\, \mathrm{d}\pi \\ &= \int_{\mathcal{H}\times\Omega^k} \int_{\Omega^{n-k}} \int_{\Omega^{\mathbb{N}}} h(x_n(\bar{x}, (\overline{\boldsymbol{\omega}}_{[0:k]}, \boldsymbol{\omega}_{[k:n]})), \boldsymbol{\omega}_{[n:\infty]}) \\ &\qquad \mathrm{d}\mathbb{P}^{\mathbb{N}}(\boldsymbol{\omega}_{[n:\infty]})\, \mathrm{d}\mathbb{P}^{n-k}(\boldsymbol{\omega}_{[k:n]})\, \mathrm{d}(\nu^\star \times \mathbb{P}^k)(\bar{x}, \overline{\boldsymbol{\omega}}_{[0:k]}). \end{aligned}$$

Where $\bar{x}$ and $\overline{\boldsymbol{\omega}}_{[0:k]}$ are notations for variables in $\mathcal{H}$ and Ω^k, respectively. Since $g(z) = \widetilde{g}(x, \boldsymbol{\omega}_{[0:k]})$, we also have

$$\int_{\overline{\Omega}} g\, \mathrm{d}\pi = \int_{\mathcal{H}\times\Omega^k} \widetilde{g}(x, \boldsymbol{\omega}_{[0:k]})\, \mathrm{d}(\nu^\star \times \mathbb{P}^k)(x, \boldsymbol{\omega}_{[0:k]}).$$

Combining the integrals where possible, we obtain for the difference

$$\begin{aligned}
&\int_{\overline{\Omega}} g \cdot h \circ \theta^n \,\mathrm{d}\pi - \int_{\overline{\Omega}} g \,\mathrm{d}\pi \int_{\overline{\Omega}} h \,\mathrm{d}\pi \\
&\quad = \int_{\mathcal{H}\times\Omega^k} \widetilde{g}(x, \boldsymbol{\omega}_{[0:k]}) \int_{\mathcal{H}\times\Omega^k} \int_{\Omega^{n-k}} \int_{\Omega^{\mathbb{N}}} \qquad (7.2)\\
&\qquad\qquad h(x_n(x, (\boldsymbol{\omega}_{[0:k]}, \boldsymbol{\omega}_{[k:n]})), \boldsymbol{\omega}_{[n:\infty]}) - h(x_n(\bar{x}, (\overline{\boldsymbol{\omega}}_{[0:k]}, \boldsymbol{\omega}_{[k:n]})), \boldsymbol{\omega}_{[n:\infty]}) \\
&\qquad\qquad \mathrm{d}\mathbb{P}^{\mathbb{N}}(\boldsymbol{\omega}_{[n:\infty]}) \,\mathrm{d}\mathbb{P}^{n-k}(\boldsymbol{\omega}_{[k:n]}) \,\mathrm{d}(\nu^\star \times \mathbb{P}^k)(\bar{x}, \overline{\boldsymbol{\omega}}_{[0:k]}) \,\mathrm{d}(\nu^\star \times \mathbb{P}^k)(x, \boldsymbol{\omega}_{[0:k]}).
\end{aligned}$$

Since h is η-Hölder continuous in the first argument, we have

$$|h(x, \boldsymbol{\omega}) - h(\bar{x}, \boldsymbol{\omega})| \leq C_h \left\|x - \bar{x}\right\|_{\mathcal{H}}^{\eta}.$$

Applying this pointwise and using boundedness of g, we obtain

$$\begin{aligned}
|\mathrm{Cor}_n(g, h)| \leq C_h \left\|\widetilde{g}\right\|_{L_\infty(\overline{\Omega})} &\int_{\mathcal{H}\times\Omega^k} \int_{\mathcal{H}\times\Omega^k} \int_{\Omega^{n-k}} \\
&\left\|x_n(x, (\boldsymbol{\omega}_{[0:k]}, \boldsymbol{\omega}_{[k:n]})) - x_n(\bar{x}, (\overline{\boldsymbol{\omega}}_{[0:k]}, \boldsymbol{\omega}_{[k:n]}))\right\|_{\mathcal{H}}^{\eta} \\
&\mathrm{d}\mathbb{P}^{n-k}(\boldsymbol{\omega}_{[k:n]}) \,\mathrm{d}(\nu^\star \times \mathbb{P}^k)(\bar{x}, \overline{\boldsymbol{\omega}}_{[0:k]}) \,\mathrm{d}(\nu^\star \times \mathbb{P}^k)(x, \boldsymbol{\omega}_{[0:k]}).
\end{aligned}$$

For the inner integral we obtain, using Hölder's inequality (see, e.g., [10, Theorem 2.11.1]) and Theorem 7.1:

$$\begin{aligned}
&\int_{\Omega^{n-k}} \left\|x_n(x, (\boldsymbol{\omega}_{[0:k]}, \boldsymbol{\omega}_{[k:n]})) - x_n(\bar{x}, (\overline{\boldsymbol{\omega}}_{[0:k]}, \boldsymbol{\omega}_{[k:n]}))\right\|_{\mathcal{H}}^{\eta} \,\mathrm{d}\mathbb{P}^{n-k}(\boldsymbol{\omega}_{[k:n]}) \\
&\quad \leq \left(\int_{\Omega^{n-k}} \left\|x_n(x, (\boldsymbol{\omega}_{[0:k]}, \boldsymbol{\omega}_{[k:n]})) - x_n(\bar{x}, (\overline{\boldsymbol{\omega}}_{[0:k]}, \boldsymbol{\omega}_{[k:n]}))\right\|_{\mathcal{H}}^{2} \,\mathrm{d}\mathbb{P}^{n-k}(\boldsymbol{\omega}_{[k:n]})\right)^{\frac{\eta}{2}} \\
&\quad \leq (1-\mu\alpha)^{\frac{(n-k)\eta}{2}} \left\|x_k(x, \boldsymbol{\omega}_{[0:k]}) - x_k(\bar{x}, \overline{\boldsymbol{\omega}}_{[0:k]})\right\|_{\mathcal{H}}^{\eta}.
\end{aligned}$$

This concludes the proof. □

Remark 7.4. *The key idea in the proof of Theorem 7.3 is that g only depends on the first k entries of the $\boldsymbol{\omega}$ variable. This ensures that the values $\boldsymbol{\omega}_{[k:n]}$ appearing in the difference* (7.2) *are* identical *for both terms of h and allows us to apply the contraction property from Theorem 7.1 to the iterations from k to n. We will see that the observables for the estimators from Section 4.2 only depend on the first one (in case of the estimator of the second moment of the noisy gradient) or two (in case of the estimator for the variance) entries of the $\boldsymbol{\omega}$ variable.*

Observables from Section 4.2

We now turn to the estimators for the variance $\widehat{\sigma}_n^2$ and for the second moment of the noisy gradients $\widehat{g}_n$. To mitigate noise in the estimators, we introduced the p-EMA averaging

scheme in Section 4.2.1 and Chapter 6. We have already established a convergence theory showing that p-EMA averages converge under suitable conditions; see Theorems 6.4 and 6.9. A central assumption in Theorem 6.4 concerns the rate of mixing of the underlying stochastic process, quantified through the decay of correlations of the observables. We intend to use the results from Theorem 7.1 to show that SGD indeed induces the required decay of correlations for the observables used in the estimation process.

We consider the case where SGD is run with a constant step size α and the estimators are evaluated as described in Chapter 4 (but not used for step size selection). Of particular interested in this chapter are the convergence properties of the estimators for the variance and the second moment of the noisy gradients.

Again, we consider the probability space $\mathcal{H} \times \Omega^{\mathbb{N}}$, equipped with the probability measure $\nu^\star \times \mathbb{P}^{\mathbb{N}}$. On this probability space we can model the estimators for the variance and the second moment of the noisy gradient as follows. For $z = (x_0, \omega_0, \omega_1, \dots)$ define

$$v(z) = \frac{1}{\alpha} \left(f_{\omega_1}(x_0 - \alpha \nabla f_{\omega_0}(x_0)) - f_{\omega_0}(x_0 - \alpha \nabla f_{\omega_0}(x_0)) \right).$$

Then the observation $\widetilde{\sigma}_k^2 = \frac{f_{\omega_{k+1}}(x_{k+1}) - f_{\omega_k}(x_{k+1})}{\alpha}$ can be expressed as:

$$\widetilde{\sigma}_k^2 = v(\theta^{k+1} z), \tag{7.3}$$

where θ is defined in (7.1).

Similarly, define

$$g(z) = \left\| f'_{\omega_0}(x_0) \right\|_{\mathcal{H}^*}^2,$$

so that the observations $\widetilde{g}_k = \left\| f'_{\omega_k}(x_k) \right\|_{\mathcal{H}^*}^2$ satisfy

$$\widetilde{g}_k = g(\theta^k z). \tag{7.4}$$

For the remainder of this section we impose the following assumptions.

Assumption A6. *The SOP is (μ_ω, L_ω)-feasible and F is μ-strongly convex. The step size $\alpha > 0$ is sufficiently small.*

Assumption A6 enables us to apply Theorem 7.1. The next assumption requires the existence of unique invariant measure. We discussed sufficient conditions for existence and uniqueness of $\nu^\star$ in Section 3.4.

Assumption A7. *There exists a unique probability measure $\nu^\star$ on $\mathcal{H}$ that is invariant under SGD.*

Assumption A8. *Let $\nu^\star$ denote the invariant probability measure of SGD. We assume that there exists a compact set $K \subset \mathcal{H}$ such that $\nu^\star(K) = 1$.*

Remark 7.5. *Assumption A8 is not restrictive for pointwise (μ_j, L_j)-feasible finite-sum problems*

$$F(x) = \frac{1}{J}\sum_{j=1}^{J} f_j(x)$$

on $\mathcal{H} = \mathbb{R}^d$ with step size $\alpha \leq \frac{1}{L_{\max}}$. In this case, SGD can be interpreted as an iterated function system (IFS), i.e. a finite set of contractions

$$\varphi_j(x) = x - \alpha \nabla f_j(x), \qquad 1 \leq j \leq J.$$

Using Proposition 2.7 in the first, $\alpha \leq \frac{1}{L_{\max}}$ in the second and Proposition 2.6 in the last inequality we have for ally $x, y \in \mathcal{H}$ and $j \in [1:J]$

$$\begin{aligned}
\|\varphi_j(x) - \varphi_j(y)\|_{\mathcal{H}}^2 &= \|x-y\|_{\mathcal{H}}^2 - 2\alpha\left(f_j'(x) - f_j'(y), x-y\right) + \alpha^2 \left\|f_j'(x) - f_j'(y)\right\|_{\mathcal{H}^*}^2 \\
&\leq \|x-y\|_{\mathcal{H}}^2 + (\alpha^2 L_{\max} - 2\alpha)\left(f_j'(x) - f_j'(y), x-y\right) \\
&\leq \|x-y\|_{\mathcal{H}}^2 - \alpha\left(f_j'(x) - f_j'(y), x-y\right) \\
&\leq (1-\mu\alpha)\|x-y\|_{\mathcal{H}}^2 .
\end{aligned}$$

Hence, each φ_j is a contraction.[1] *By [45, Theorem 2.63], the IFS admits a unique, nonempty compact* attractor *$K \subset \mathcal{H}$ satisfying*

$$K = \bigcup_{j=1}^{J} \varphi_j(K). \tag{7.5}$$

and $\nu^\star(K) = 1$.

Assumption A9. *We assume that*

$$C_\infty := \max\left(\operatorname*{ess\,sup}_{\substack{x\sim\nu^\star \\ \omega\sim\mathbb{P}}} \left\|f_\omega'(x)\right\|_{\mathcal{H}}, \operatorname*{ess\,sup}_{\substack{x\sim\nu^\star \\ \omega\sim\mathbb{P}}} |f_\omega(x)|\right) < \infty.$$

Remark 7.6. *For finite-sum problems, Assumption A9 follows directly from Assumption A8.*

With these assumptions in place, we now prove the following.

Lemma 7.7. *There exist constants C_v and C_g such that:*

1. $|\mathrm{Cor}_n(v,v)| \leq C_v(1-\mu\alpha)^n$

[1] A similar computation appears in the proof of Theorem 7.1.

2. $|\mathrm{Cor}_n(g,g)| \leq C_g(1-\mu\alpha)^n$

Proof. 1. The function $v(z)$ depends only on the first two entries of the ω-part of z. Moreover, we have

$$\begin{aligned}
|v(z)| &= \frac{1}{\alpha}\left|f_{\omega_1}(x-\alpha\nabla f_{\omega_0}(x)) - f_{\omega_0}(x-\alpha\nabla f_{\omega_0}(x))\right| \\
&\leq \frac{1}{\alpha}\left(|f_{\omega_1}(x)| + \alpha\left\|f'_{\omega_0}(x)\right\|_{\mathcal{H}^*}\left\|f'_{\omega_1}(x)\right\|_{\mathcal{H}^*} + \frac{\alpha^2 L_{\max}}{2}\left\|f'_{\omega_1}(x)\right\|_{\mathcal{H}^*}^2\right) \\
&\quad + \frac{1}{\alpha}\left(|f_{\omega_0}(x)| + \alpha\left\|f'_{\omega_0}(x)\right\|_{\mathcal{H}^*}^2 + \frac{\alpha^2 L_{\max}}{2}\left\|f'_{\omega_0}(x)\right\|_{\mathcal{H}^*}^2\right) \\
&\leq \frac{2}{\alpha}\left(C_\infty + \alpha C_\infty^2 + \frac{\alpha^2 L_{\max}}{2}C_\infty^2\right),
\end{aligned}$$

where the last inequality holds almost everywhere. Hence, v is essentially bounded.

Furthermore, for $z=(x,\omega_0,\omega_1,\dots)$ and $\bar{z}=(\bar{x},\omega_0,\omega_1,\dots)$, convexity of f_ω gives

$$\begin{aligned}
\alpha\,|v(z)-v(\bar{z})| &\leq \left|f_{\omega_1}(x-\alpha\nabla f_{\omega_0}(x)) - f_{\omega_1}(\bar{x}-\alpha\nabla f_{\omega_0}(\bar{x}))\right| \\
&\quad + \left|f_{\omega_0}(x-\alpha\nabla f_{\omega_0}(x)) - f_{\omega_0}(\bar{x}-\alpha\nabla f_{\omega_0}(\bar{x}))\right| \\
&\leq \left\|f'_{\omega_1}(x-\alpha\nabla f_{\omega_0}(x))\right\|_{\mathcal{H}^*}\left(\|x-\bar{x}\|_{\mathcal{H}} + \alpha\left\|f'_{\omega_0}(x)-f'_{\omega_0}(\bar{x})\right\|_{\mathcal{H}}\right) \\
&\quad + \left\|f'_{\omega_0}(x-\alpha\nabla f_{\omega_0}(x))\right\|_{\mathcal{H}^*}\left(\|x-\bar{x}\|_{\mathcal{H}} + \alpha\left\|f'_{\omega_0}(x)-f'_{\omega_0}(\bar{x})\right\|_{\mathcal{H}}\right) \\
&\leq 2C_\infty(1+\alpha L_{\max})^2\|x-\bar{x}\|_{\mathcal{H}}\,.
\end{aligned}$$

The claim then follows from Theorem 7.3.

2. Again, we verify the assumptions of Theorem 7.3. We have almost surely

$$|g(z)| = \left\|f'_{\omega_0}(x)\right\|_{\mathcal{H}^*}^2 \leq C_\infty^2,$$

so g is essentially bounded. Moreover, for $z=(x,\omega_0,\omega_1,\dots)$ and $\bar{z}=(\bar{x},\omega_0,\omega_1,\dots)$, we get

$$\begin{aligned}
|g(z)-g(\bar{z})| &= \left|\left\|f'_{\omega_0}(x)\right\|_{\mathcal{H}^*}^2 - \left\|f'_{\omega_0}(\bar{x})\right\|_{\mathcal{H}^*}^2\right| \\
&\leq \left\|f'_{\omega_0}(x)+f'_{\omega_0}(\bar{x})\right\|_{\mathcal{H}^*}\left\|f'_{\omega_0}(x)-f'_{\omega_0}(\bar{x})\right\|_{\mathcal{H}^*} \\
&\leq 2C_\infty L_{\max}\|x-\bar{x}\|_{\mathcal{H}}\,.
\end{aligned}$$

Thus, the result follows again from Theorem 7.3.

□

7.2 Convergence of the Estimators

With the results of the previous section at hand, we are now able to present a convergence theory for the estimated step sizes. It is important to note that the following discussion does not provide a convergence theory for the algorithm with the adaptive step sizes. Instead, we present asymptotic convergence results for the estimators, assuming they are computed from observations made along the trajectory of SGD with constant step sizes. Despite this limitation, the results offer meaningful insight into the long-term behavior of the algorithm. In particular, they clarify how the estimated step sizes behave if the adaptive step sizes were to stagnate at some fixed level, in which case the algorithm effectively behaves like SGD with constant step sizes.

Recall that for non-interpolating problems it is necessary to reduce the step sizes to zero as iterations progress in order to obtain convergence. In this section, we demonstrate, both numerically and theoretically, that for non-interpolating problems, if SGD is run with a constant step size α, and the adaptive step sizes are evaluated as described in Chapter 4 (but not used), then the estimated step sizes eventually converge to a value less than or equal to $\frac{\alpha}{2}$. This observation suggests that on non-interpolating problems the adaptive step sizes cannot remain constant indefinitely and must eventually decrease.

The following result characterizes the limits of the estimators. Recall that we work under the assumptions Assumptions A6 to A9.

Theorem 7.8. *Consider SGD with constant step size α applied to some initialization $x_0 \in \mathcal{H}$.*

1. *Denote by $\widehat{\sigma}_n^2$ the sequence obtained by p-EMA applied to the observations*

$$\widetilde{\sigma}_n^2 = \frac{f_{\omega_{n+1}}(x_{n+1}) - f_{\omega_n}(x_{n+1})}{\alpha}.$$

   Then

$$\widehat{\sigma}_n^2 \to \widehat{\sigma}_\infty^2 := \frac{1}{\alpha} \int_{\mathcal{H}} \left(F(x) - \int_\Omega f_\omega(x - \alpha \nabla f_\omega(x)) \, \mathrm{d}\mathbb{P}(\omega) \right) \mathrm{d}\nu^\star(x)$$

   for $\nu^\star$-almost every x_0 and $\mathbb{P}^{\mathbb{N}}$-almost every realization of SGD.

2. *Denote by $\widehat{g}_n$ the sequence obtained by p-EMA applied to the observations*

$$\widetilde{g}_n = \left\| f'_{\omega_n}(x_n) \right\|_{\mathcal{H}^*}^2 .$$

   Then:

$$\widehat{g}_n \to \widehat{g}_\infty := \int_{\mathcal{H}} \int_\Omega \left\| f'_\omega(x) \right\|_{\mathcal{H}^*}^2 \mathrm{d}\mathbb{P}(\omega) \, \mathrm{d}\nu^\star(x)$$

   for $\nu^\star$-almost every x_0 and $\mathbb{P}^{\mathbb{N}}$-almost every realization of SGD.

Proof. Using the representations in (7.3) and (7.4), we obtain

$$\mathbb{E}_{\nu^\star \times \mathbb{P}^{\mathbb{N}}}\left[\widetilde{\sigma}_n^2\right] = \widehat{\sigma}_\infty^2 \qquad \text{and} \qquad \mathbb{E}_{\nu^\star \times \mathbb{P}^{\mathbb{N}}}\left[\widetilde{g}_n\right] = \widehat{g}_\infty.$$

Further, Lemma 7.7 shows summable decay of correlations of the quantities of interest. We have $\widehat{g}_n \geq 0$, and $\widehat{\sigma}_n^2$ is bounded almost surely by Assumption A9. By Theorem 6.9, p-EMA is an averaging scheme. Therefore, we can apply Theorem 6.4, which yields the result. □

Remark 7.9. *Theorem 7.8 establishes convergence of the estimators for almost every trajectory whose initialization lies in the support of the invariant measure. Intuitively, this means that the stated convergence has to be expected when SGD is already in its stationary regime at the beginning. This is undesirable from a practical standpoint, since in applications we do not start within the support of the invariant measure but instead at some arbitrary initialization that may be far from it. The iterates subsequently converge towards the attractor, and observations made along this transient phase must be taken into account. Fortunately, the convergence result can be extended to arbitrary initializations.*
Consider an arbitrary initialization x_0 and an initialization $\bar{x}_0 \in \mathcal{H}$ for which the convergence statements in Theorem 7.8 hold true. By Corollary 7.2 we have

$$|x_n - \bar{x}_n| \to 0$$

almost surely, where $\bar{x}_n$ denotes the iterates of SGD initialized at $\bar{x}_0$. Both observables (v and g) are continuous in x, and therefore it follows from Lemma 6.5 and the fact that p-EMA induces an averaging scheme (Theorem 6.9) that the statement of Theorem 7.8 holds for arbitrary initialization x_0.

Remark 7.10. *We emphasize that the limits of $\widehat{\sigma}_n^2$ and $\widehat{g}_n$ are simply the expectations of their defining observables. In the case of $\widetilde{\sigma}_n^2$, the observable depends on the state x_n, the* current *ω_n and the* next *ω_{n+1}. The integral*

$$\int_\Omega f_{\omega_{n+1}}(x_{n+1}) \, \mathrm{d}\mathbb{P}(\omega_{n+1})$$

equals $F(x_{n+1})$ by definition of F as the expectation of f_ω, and by the independence of ω_{n+1} from x_{n+1}.

We obtain the following characterization of interpolating SOPs:

Corollary 7.11. *Denote by $x^\star \in \mathcal{H}$ the unique solution to the SOP and by $\delta_{x^\star}$ the Dirac-measure on $\mathcal{H}$ centered at $x^\star$. Denote by $\widehat{\sigma}_\infty^2$ and $\widehat{g}_\infty$ the limits from Theorem 7.8. Assume that $\alpha \leq \frac{1}{4L_{\max}}$. Then the following assertions are equivalent.*

1. *The SOP is interpolating, that is,*

$$V_0 = \int_\Omega \left\| f'_\omega(x^\star) \right\|^2_{\mathcal{H}^*} \, \mathrm{d}\mathbb{P}(\omega) = 0.$$

2. $\nu^\star = \delta_{x^\star}$.

3. $\widehat{\sigma}^2_\infty = 0$.

4. $\widehat{g}_\infty = 0$.

Proof. We show:

$$1. \Leftrightarrow 2. \qquad \text{and} \qquad 2. \Rightarrow 3. \Rightarrow 4. \Rightarrow 2.$$

$\boxed{1. \Rightarrow 2.}$ By the interpolation property, (3.14) in the proof of Proposition 3.15 and the invariance property we have for arbitrary $x \in \mathcal{H}$.

$$\begin{aligned} 0 \leq \int_{\mathcal{H}} \|x - x^\star\|^2_{\mathcal{H}} \, \mathrm{d}\nu^\star(x) &= \int_{\mathcal{H}} \int_\Omega \|x - \alpha \nabla f_\omega(x) - x^\star\|^2_{\mathcal{H}} \, \mathrm{d}\mathbb{P}(\Omega) \, \mathrm{d}\nu^\star(x) \\ &\leq (1 - \mu\alpha) \int_{\mathcal{H}} \|x - x^\star\|^2_{\mathcal{H}} \, \mathrm{d}\nu^\star(x). \end{aligned}$$

Thus

$$\int_{\mathcal{H}} \|x - x^\star\|^2_{\mathcal{H}} \, \mathrm{d}\nu^\star(x) = 0,$$

which implies $\nu^\star = \delta_{x^\star}$.

$\boxed{2. \Rightarrow 1.}$ Using the invariance property and $\nu^\star = \delta_{x^\star}$, we compute

$$\begin{aligned} 0 &= \int_{\mathcal{H}} \|x - x^\star\|^2_{\mathcal{H}} \, \mathrm{d}\nu^\star(x) \\ &= \int_{\mathcal{H}} \int_\Omega \|x - \alpha \nabla f_\omega(x) - x^\star\|^2_{\mathcal{H}} \, \mathrm{d}\mathbb{P}(\omega) \, \mathrm{d}\nu^\star(x) \\ &= \int_{\mathcal{H}} \left(\|x - x^\star\|^2_{\mathcal{H}} + \int_\omega -2\alpha \left(f'(x), x - x^\star \right) + \alpha^2 \left\| f'_\omega(x) \right\|^2_{\mathcal{H}^*} \, \mathrm{d}\mathbb{P}(\omega) \right) \mathrm{d}\nu^\star(x) \\ &= \int_{\mathcal{H}} \|x - x^\star\|^2_{\mathcal{H}} \, \mathrm{d}\nu^\star(x) + \alpha^2 \int_\Omega \left\| f'_\omega(x^\star) \right\|^2_{\mathcal{H}^*} \\ &= \alpha^2 \int_\Omega \left\| f'_\omega(x^\star) \right\|^2_{\mathcal{H}^*} \end{aligned}$$

Since $\alpha > 0$, this implies $\int_\Omega \|f'_\omega(x^\star)\|^2_{\mathcal{H}^*} = 0$, which is the interpolating property.

$\boxed{3. \Rightarrow 4.}$ For almost every ω, f_ω is $L_{\max}$-smooth, hence

$$f_\omega(x - \alpha \nabla f_\omega(x)) \leq f_\omega(x) + \left(\frac{\alpha^2 L_{\max}}{2} - \alpha \right) \left\| f'_\omega(x) \right\|^2_{\mathcal{H}^*}.$$

Therefore,

$$\begin{aligned}\widehat{\sigma}_\infty^2 &= \frac{1}{\alpha}\int_{\mathcal{H}}\left(F(x) - \int_\Omega f_\omega(x-\alpha\nabla f_\omega(x))\,\mathrm{d}\mathbb{P}(\omega)\right)\mathrm{d}\nu^\star(x)\\ &\geq \frac{1}{\alpha}\int_{\mathcal{H}} F(x) - F(x) + \alpha\left(1-\frac{\alpha L_{\max}}{2}\right)\left\|f'_\omega(x)\right\|_{\mathcal{H}^*}^2\,\mathrm{d}\nu^\star(x) \qquad (7.6)\\ &= \left(1-\frac{\alpha L_{\max}}{2}\right)\widehat{g}_\infty.\end{aligned}$$

Since $\widehat{g}_\infty \geq 0$ and $\alpha \leq \frac{1}{L_{\max}}$, the implication $\widehat{\sigma}_\infty^2 = 0 \Rightarrow \widehat{g}_\infty = 0$ follows.

4. $\Rightarrow$ 2. Using the invariance, we compute

$$\begin{aligned}\int_{\mathcal{H}} F(x)\,\mathrm{d}\nu^\star(x) &= \int_{\mathcal{H}}\int_\Omega F(x-\alpha\nabla f'_\omega(x))\mathbb{P}(\omega)\,\mathrm{d}\nu^\star(x)\\ &\leq \int_{\mathcal{H}} F(x) - \int_\Omega \alpha\left(F'(x),\nabla f_\omega(x)\right) + \frac{\alpha^2 L}{2}\left\|f'_\omega(x)\right\|_{\mathcal{H}^*}^2\\ &\qquad\qquad \mathrm{d}\mathbb{P}(\omega)\,\mathrm{d}\nu^\star(x)\\ &= \int_{\mathcal{H}} F(x)\,\mathrm{d}\nu^\star(x) - \alpha\int_{\mathcal{H}}\left\|F'(x)\right\|_{\mathcal{H}^*}^2\,\mathrm{d}\nu^\star(x) + \frac{\alpha^2 L}{2}\widehat{g}_\infty.\end{aligned}$$

Hence,

$$\widehat{g}_\infty \geq \frac{2}{\alpha L}\int_{\mathcal{H}}\left\|F'(x)\right\|_{\mathcal{H}^*}^2\,\mathrm{d}\nu^\star(x).$$

Therefore, strong convexity of F yields

$$\widehat{g}_\infty = 0 \qquad \Rightarrow \qquad \int_{\mathcal{H}}\|x-x^\star\|_{\mathcal{H}}^2\,\mathrm{d}\nu^\star(x) = 0 \qquad \Rightarrow \qquad \nu^\star = \delta_{x^\star}.$$

2. $\Rightarrow$ 3. From 2. $\Rightarrow$ 1. we already know that

$$\int_\Omega \left\|f'_\omega(x^\star)\right\|_{\mathcal{H}^*}^2\,\mathrm{d}\mathbb{P}(\omega) = 0.$$

f_ω is convex for almost every ω. Hence,

$$f_\omega(x-\alpha\nabla f_\omega(x)) \geq f_\omega(x) - \alpha\left\|f'_\omega(x)\right\|_{\mathcal{H}^*}^2.$$

Thus

$$\begin{aligned}\widehat{\sigma}_\infty^2 &= \frac{1}{\alpha}\int_{\mathcal{H}}\left(F(x) - \int_\Omega f_\omega(x-\alpha\nabla f_\omega(x))\,\mathrm{d}\mathbb{P}(\omega)\right)\mathrm{d}\nu^\star(x)\\ &\leq \int_{\mathcal{H}}\int_\Omega \left\|f'_\omega(x)\right\|_{\mathcal{H}^*}^2\,\mathrm{d}\mathbb{P}(\omega)\,\mathrm{d}\nu^\star(x)\\ &= \int_\Omega \left\|f'_\omega(x^\star)\right\|_{\mathcal{H}^*}^2\,\mathrm{d}\mathbb{P}(\omega)\\ &= 0,\end{aligned}$$

where we used $\nu^\star = \delta_{x^\star}$ in the second to last step. Finally, (7.6) implies $\widehat{\sigma}_\infty^2 \geq 0$, so that we get $\widehat{\sigma}_\infty^2 = 0$.

□

7.2.1 Alternative Variance Estimation

In Equation (4.8) in Section 4.2.5 we discussed the possibility of using

$$\widetilde{\sigma}_k^2 = \frac{f_{\omega_k}(x_k) - f_{\omega_k}(x_{k+1})}{\alpha_k} \tag{7.7}$$

as observations for the variance estimation instead of (4.7). As noted there, this alternative formulation has the appealing property that $\widehat{\sigma}_n \in [0, \widehat{g}_n]$, which in turn yields step sizes satisfying

$$\widehat{\alpha}_n \in \left[0, \frac{1}{\widehat{L}_n}\right].$$

In fact, the motivation originally provided for our approach to estimate the variance (Section 4.2.4) does not apply to the observations in Equation (7.7). However, with the convergence theory developed above, we can still justify their use. Indeed, the almost sure limit of the $\widehat{\alpha}_n$ with the alternative variance estimation agrees with the classical variance estimation obtained from the observations (4.7).

Corollary 7.12. *Consider the setting of Theorem 7.8. Denote by $\widehat{\sigma}_n^2$ the sequence obtained by p-EMA applied to the observations*

$$\widetilde{\sigma}_n^2 = \frac{f_{\omega_n}(x_n) - f_{\omega_{n+1}}(x_{n+1})}{\alpha}$$

Then

$$\widehat{\sigma}_n^2 \to \widehat{\sigma}_\infty^2 = \frac{1}{\alpha} \int_{\mathcal{H}} \left(F(x) - \int_\Omega f_\omega(x - \alpha \nabla f_\omega(x)) \, \mathrm{d}\mathbb{P}(\omega) \right) \mathrm{d}\nu^\star(x),$$

for $\nu^\star$-almost every x_0 and $\mathbb{P}^{\mathbb{N}}$-almost every realization of SGD.

Proof. The proof proceeds exactly as in Theorem 7.8, using a suitably modified version of Lemma 7.7, which can be established analogously for the alternative observables from (7.7). □

Remark 7.13. *Remark 7.9 applies to Corollary 7.12 as well.*

7.2.2 Consequences for the Step Sizes – The Non-Interpolating Case

The convergence of the estimators under constant step sizes allows us to determine the behavior of the corresponding estimated step sizes. We emphasize that in this analysis, SGD is run with *constant* step sizes, and the estimated step sizes are evaluated, but not used. Let us assume that we have a sufficiently accurate estimate $\widehat{L}$ of the Lipschitz constant L and focus on the estimators for the $\widetilde{\sigma}_n^2$ and for the second moment of the noisy gradients $\widetilde{g}_n$. Recalling the representation of the estimated step sizes

$$\widehat{\alpha}_n = \frac{1}{\widehat{L}_n}\left(1 - \frac{\widehat{\sigma}_n^2}{\widehat{g}_n}\right) \tag{7.8}$$

from (4.9), we may, assuming access to $\widehat{L}$, consider instead the more informed

$$\widehat{\alpha}_n^{\widehat{L}} := \frac{1}{\widehat{L}}\left(1 - \frac{\widehat{\sigma}_n^2}{\widehat{g}_n}\right).$$

We therefore focus on the expression inside the parentheses, which is intended to account for the variance in the noisy search directions. This term is particularly relevant in the non-interpolating case, where its purpose is to drive the step sizes downward as the iterates approach the minimizer $x^\star$.

By Theorem 7.8 and Remark 7.9, the estimators $\widehat{\sigma}_n^2$ and $\widehat{g}_n$ converge almost surely to limits $\widehat{\sigma}_\infty^2$ and $\widehat{g}_\infty$, respectively. By Corollary 7.11, in the non-interpolating case these limits satisfy

$$\widehat{\sigma}_\infty^2 > 0 \qquad \text{and} \qquad \widehat{g}_\infty > 0.$$

Consequently, the estimated step sizes $\widehat{\alpha}_n^{\widehat{L}}$ converge almost surely towards the limit

$$\widehat{\alpha}_\infty^{\widehat{L}} = \frac{1}{\widehat{L}}\left(1 - \frac{\widehat{\sigma}_\infty^2}{\widehat{g}_\infty}\right).$$

In (7.6) we established the inequality

$$\widehat{\sigma}_\infty^2 \geq \left(1 - \frac{\alpha L_{\max}}{2}\right)\widehat{g}_\infty,$$

which yields the bound

$$\widehat{\alpha}_\infty^{\widehat{L}} = \frac{1}{\widehat{L}}\left(1 - \frac{\widehat{\sigma}_\infty^2}{\widehat{g}_\infty}\right) \leq \frac{1}{\widehat{L}}\left(1 - \left(1 - \frac{\alpha L_{\max}}{2}\right)\right) = \frac{\alpha}{2}\frac{L_{\max}}{\widehat{L}}. \tag{7.9}$$

One interpretation of this result is that, provided $\widehat{L}$ is a sufficiently accurate estimate of $L_{\max}$, the estimated step sizes eventually become significantly smaller than the current constant step size used in the run of SGD. This occurs when the iterates are distributed

according to the invariant measure $\nu^\star$, which itself depends on α. In such a regime, the algorithm is stagnating, indicating the need for a reduction in step size.

It is worth noting that the bound in (7.6) remains valid, if $L_{\max}$ is interpreted *locally* with respect to $\nu^\star$ as

$$L_{\max}^{\mathrm{loc}} := \operatorname*{ess\,sup}_{\substack{x\sim\nu^\star \\ \omega\sim\mathbb{P}}} 2\frac{f_\omega(x-\alpha f_\omega(x)) - f_\omega(x) + \alpha \left\|f'_\omega(x)\right\|_{\mathcal{H}^*}^2}{\alpha^2 \left\|f'_\omega(x)\right\|_{\mathcal{H}^*}^2}.$$

Indeed, when using the estimator of L proposed in Section 4.2.2, and taking the maximum of its observations, the resulting estimate converges almost surely to $L_{\max}^{\mathrm{loc}}$, provided that the run of SGD is started within the support of the invariant measure. However, as already discussed in Section 4.2.2, this approach leads to overly cautious step sizes in practice.

Numerically, we illustrate the convergence shown in Theorem 7.8 using the example introduced earlier on page 40, see also Figure 3.3. Recall that in that example we applied SGD to a simple SOP consisting of quadratic functions on $\mathbb{R}^2$, each drawn with equal probability. During the run of SGD, we evaluated the estimators using p-EMA

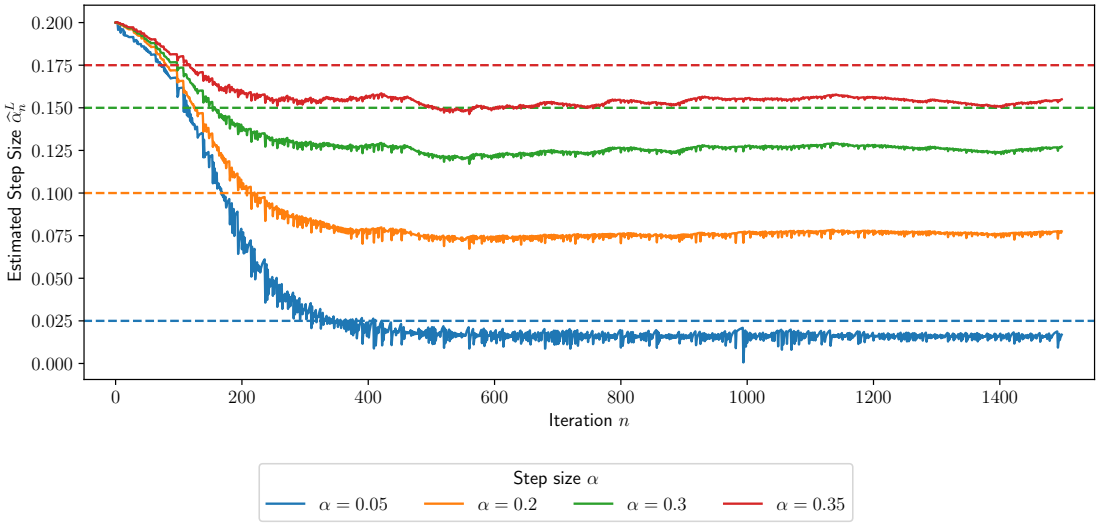


Figure 7.1: Convergence of the estimated step sizes during the run of SGD with constant step size α, variance estimation due to (4.7).

with $p = 0.75$, and initializations $\widehat{\sigma}_0^2 = 0$ and $\widehat{g}_0 = \left\|f'_{\omega_0}(x_0)\right\|_{\mathcal{H}^*}^2$. We plot the current estimates for $\widehat{\alpha}_n^L$ as solid lines over the iteration n with colors indicating the used step sizes. The plots in Figure 7.1 corresponds to the *classical* variance estimation (4.7), while the plots in Figure 7.2 depict the behavior under the alternative approach (4.8) discussed in Sections 4.2.5 and 7.2.1 For reference, each figure includes dashed horizontal

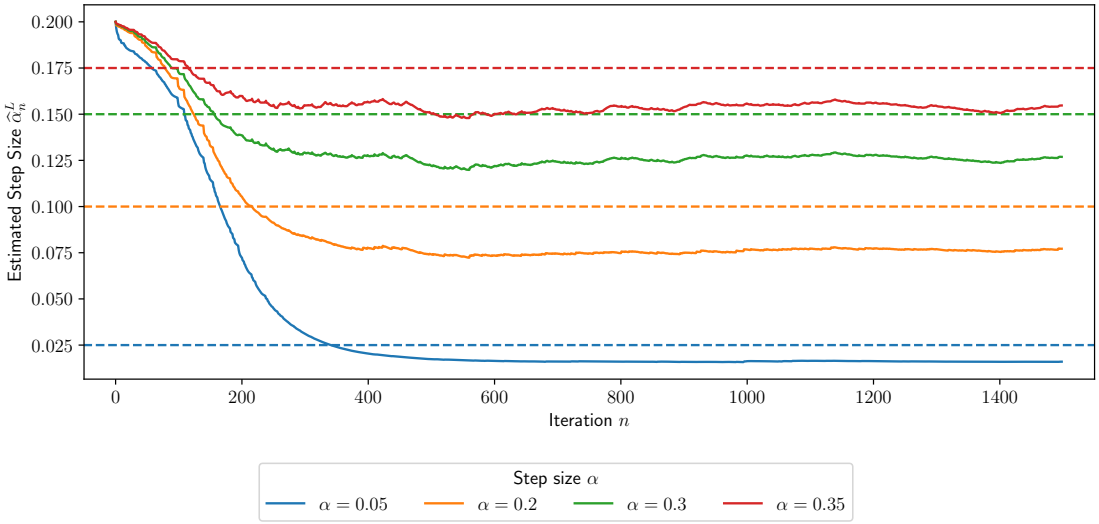


Figure 7.2: Convergence of the estimated step sizes during the run of SGD with constant step size α, variance estimation due to (4.8).

lines representing the values $\frac{\alpha}{2}$ corresponding to the constant step sizes employed. For all four step sizes tested in this experiment, we observe that the estimated step sizes converge to limits that lie clearly below the threshold $\frac{\alpha}{2}$, as described by (7.9).

7.2.3 Consequences for the Step Sizes – The Interpolating Case

In contrast to the non-interpolating setting, it is not necessary to decrease the step sizes to zero in the interpolating case. Indeed, Propositions 3.14 and 3.15 with $V_0 = 0$ show that sufficiently small constant step sizes already ensure convergence. In fact, unnecessarily reducing the step size slows down the rate of convergence. As we have shown in Corollary 7.11, both estimators, $\widehat{\sigma}_n^2$ for the variance and $\widehat{g}_n$ for the second moment of the noisy gradient, converge to zero almost surely. Consequently, the limit of the estimated step sizes in (7.8), assuming such a limit exists, depends critically on the respective rates at which $\widehat{\sigma}_n^2$ and $\widehat{g}_n$ decay.

To illustrate that the step sizes stay bounded away from zero in this case, we consider the alternative variance estimation due to (4.8). Assuming both averaging processes, for $\widehat{\sigma}_n^2$ and $\widehat{g}_n$, use p-EMA with the same parameter p, the following worst-case discussion shows that the estimated step sizes stay bounded away from zero, provided that the variance estimator is initialized by setting $\widetilde{\sigma}_0^2 = 0$, as discussed in Section 4.3. We may assume that the initial observation, and thus the initialization, for the estimator for the second moment of the noisy gradient $\widetilde{g}_0$ is positive, as otherwise we would have

started the algorithm at the true minimizer $x^\star$. Further, as discussed in Section 4.2.5, the observations for the alternative variance estimation satisfy

$$\widetilde{\sigma}_n^2 \leq \widetilde{g}_n, \tag{7.10}$$

hence we have for the estimates obtained from p-EMA:

$$\widehat{\sigma}_n^2 = \frac{1}{\Lambda_n}\sum_{k=2}^{n}\beta_k\widetilde{\sigma}_{k-1}^2 \leq \frac{1}{\Lambda_n}\sum_{k=2}^{n}\beta_k\widetilde{g}_{k-1} = \widehat{g}_n - \frac{1}{\Lambda_n}\beta_1\widetilde{g}_0,$$

and therefore

$$\frac{\widehat{\sigma}_n^2}{\widehat{g}_n} \leq 1 - \frac{\beta_1\widetilde{g}_0}{\Lambda_n\widehat{g}_n}.$$

Next, we show that $\Lambda_n\widehat{g}_n \leq K$ for some constant K independent of n, assuming that $\widetilde{g}_n \leq C_0\rho^n$ for some $\rho \in (0,1)$. The latter assumption is justified, as we expect linear convergence for sufficiently small step sizes in the interpolating setting. First, recall from (6.9) that

$$\beta_k = k^{-p}\prod_{s=2}^{k}\frac{s^p}{s^p-1}.$$

Consider $\varepsilon \in (0, 1-\rho)$. Then, there exists $k_0 \in \mathbb{N}$ such that

$$\frac{s^p}{s^p-1} \leq \frac{1}{\rho+\varepsilon} \qquad \text{for all } s > k_0.$$

Consequently, for $k > k_0$,

$$\beta_k \leq (\rho+\varepsilon)^{-k}k^{-p}(\rho+\varepsilon)^{k_0}\prod_{s=2}^{k_0}\frac{s^p}{s^p-1}.$$

Denoting $C = (\rho+\varepsilon)^{k_0}\prod_{s=2}^{k_0}\frac{s^p}{s^p-1}$, we obtain for $n > k_0$:

$$\Lambda_n\widehat{g}_n = \sum_{k=1}^{n}\beta_k\widetilde{g}_{k-1} \leq \sum_{k=1}^{k_0}\beta_k\widetilde{g}_{k-1} + CC_0\sum_{k=k_0+1}^{n}k^{-p}\left(\frac{\rho}{\rho+\varepsilon}\right)^k.$$

The first term on the right-hand side is a constant, and the second can be bounded independently of n, showing that indeed $\Lambda_n\widehat{g}_n \leq K$ for some constant K independent of n. This implies that

$$\frac{\widehat{\sigma}_n^2}{\widehat{g}_n} \leq 1 - \frac{\beta_1\widetilde{g}_0}{\Lambda_n\widehat{g}_n} \leq 1 - \frac{\beta_1\widetilde{g}_0}{K}$$

and thus

$$1 - \frac{\widehat{\sigma}_n^2}{\widehat{g}_n} \geq \frac{\beta_1\widetilde{g}_0}{K} > 0.$$

We emphasize that the above discussion considers a worst-case scenario for the decay of the step sizes and bounds them away from zero by leveraging the initialization of the variance estimator. Typically, there is a gap between $\widetilde{\sigma}_n^2$ and $\widetilde{g}_n$, and (7.10) is not close to equality. In such situations, the lower bound on the step sizes can be significantly larger than in the worst-case scenario above, where equality in (7.10) is effectively assumed.

The above discussion applies only to the alternative variance estimation, as it relies on the bound (7.10). In our numerical experiments, however, we did not observe significant differences between the two variance estimation approaches in the interpolating setting.

8 Numerical Results

In this chapter we present additional numerical results that complement the examples discussed earlier in Chapters 3 and 5 to 7. In those chapters we primarily considered a very simple class of SOPs described in the motivation of Section 3.4. Here, we focus on a more complex class of SOPs introduced in the following section. These SOPs are defined in much higher dimension (here: 50) and they exhibit substantially stronger anisotropy than the previously considered examples. Additionally, these problems are *not* finite sum problems, and the sampled functions f_ω are not necessarily convex. The same class of SOPs was also investigated in [41]. After describing the class of problems, we evaluate the performance of the different step size strategies discussed in this work on this class of problems. In particular, we consider constant step sizes (discussed in Section 3.3.1), Robbins–Monroe step sizes (discussed in Section 3.3.2), and finally the adaptive step sizes introduced in Chapter 4.

8.1 A Test Case: Quadratic SOPs

Given an orthogonal matrix $S \in \mathbb{R}^{d\times d}$ and a diagonal matrix $D = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$, we construct an SOP as follows. We set the mean Hessian to $A := S^\mathsf{T} D S$ and choose a noise level $\sigma_A > 0$. In every iteration of SGD we sample a random matrix $\boldsymbol{\omega} \in \mathbb{R}^{d\times d}$ whose entries ω_{ij} are drawn independently from the uniform distribution on $[-\sigma_A, \sigma_A]$. We then set $W_\omega := \boldsymbol{\omega}^\mathsf{T}\boldsymbol{\omega} - \frac{2}{3}\sigma_A^3 \mathsf{Id}$. This choice ensures that $\mathbb{E}_{\boldsymbol{\omega}}[W_\omega] = 0$. We set $A_{\boldsymbol{\omega}} = A + W_\omega$.

For $b \in \mathbb{R}^d$, we choose a noise level $\sigma_b \geq 0$ and sample each entry of b_ω from the uniform distribution on $[-\sigma_b, \sigma_b]$.

We then consider the problem of minimizing the expected value of

$$f_\omega(x) = \frac{1}{2} x^\mathsf{T} A_\omega x + b_\omega^\mathsf{T} x + c, \tag{8.1}$$

where the constant c is selected such that $F(x) = \frac{1}{2}x^\mathsf{T}Ax + b^\mathsf{T}x + c$ satisfies $F(x^\star) = 0$ at the minimizer $x^\star = -A^{-1}b$.

The eigenvalues $\lambda_1, \dots, \lambda_d$ of A allow us to control the strong convexity constant μ of F, given by the smallest eigenvalue, and the Lipschitz constant of ∇F, given by the

largest eigenvalue. We explore two different ways of varying the condition number of the problem.

1. We fix $\mu = 1$ and consider different values of $L > 1$.
2. We fix $L = 1$ and consider different values of $\mu < 1$.

In both cases we choose the eigenvalues according to

$$\lambda_i = \left(\sqrt{\mu} + \frac{(i-1)(\sqrt{L} - \sqrt{\mu})}{d-1} \right)^2, \qquad 1 \leq i \leq d,$$

so that $\lambda_1 = \mu$, $\lambda_d = L$ and $\mu < \lambda_i < L$ for $1 \leq i \leq d$. This construction yields problems of condition number $\kappa = L/\mu \geq 1$. We refrain from using linearly spaced eigenvalues, because the nonlinear spacing produces more anisotropic and thus more challenging test problems.

The non-interpolating case corresponds to the case $\mathbb{E}_\omega \left[\|\nabla f_\omega(x^\star)\|_2^2 \right] > 0$ which occurs precisely when $\sigma_b > 0$. In contrast, choosing $\sigma_b = 0$ yields interpolating problems. This setup allows us to generate four classes of test cases.

- We choose either $\sigma_b > 0$ for a non-interpolating problem or $\sigma_b = 0$ for an interpolating proplem.
- We either fix L or μ, and vary the remaining extreme eigenvalue.

Strictly speaking, this produces more than four individual test problems, since for each class we consider several values of the free extreme eigenvalue. We group the experiments according to which parameter is fixed, resulting in four groups of test scenarios.

We apply each step size strategy to every test case five times with different random initializations and random seeds for sampling. In each figure we display the mean of the corresponding quantity across these five runs.

8.2 Performance of Constant Step Sizes

In this section we apply SGD with constant step sizes to the four groups of test problems described in Section 8.1. We present the results for the interpolating cases in Figures 8.2 and 8.4, and for the non-interpolating cases in Figures 8.1 and 8.3. For each experiment we plot the suboptimality $F(x_n) - F(x^\star)$ on the left, the squared distance to the minimizer $\|x_n - x^\star\|^2$ in the center, and the *suggested* step size $\widehat{\alpha}_n$ on the right. We emphasize that the suggested step size is not used by the algorithm in these experiments. Instead, we run SGD with the fixed step size $\alpha = \frac{1}{L}$ and merely evaluate the estimated

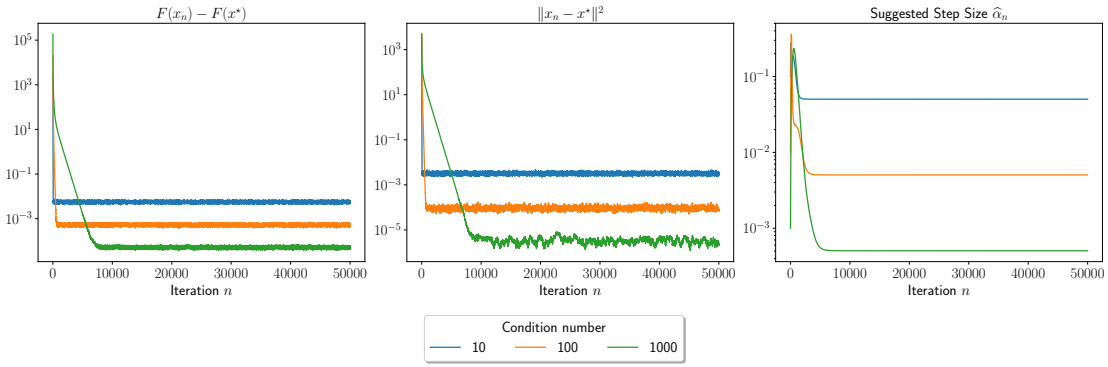


Figure 8.1: SGD with constant step sizes $\alpha = \frac{1}{L}$. Non-interpolating problem, $\mu = 1$ fixed, L varying.

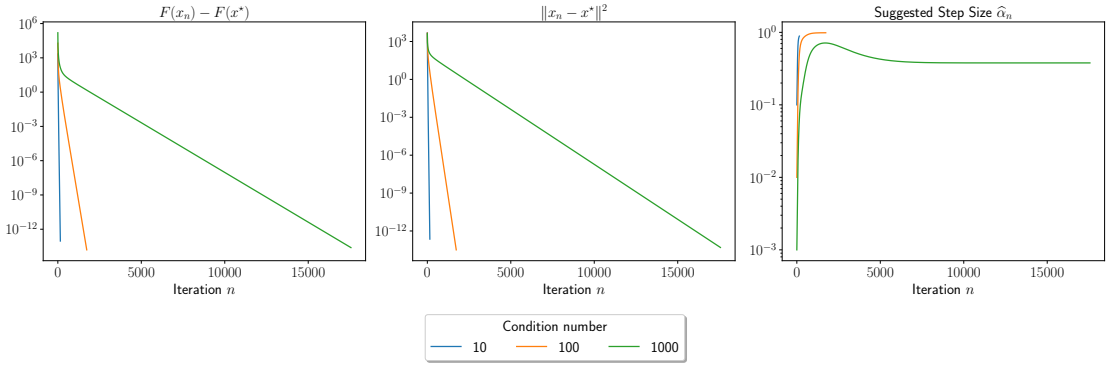


Figure 8.2: SGD with constant step sizes $\alpha = \frac{1}{L}$. Interpolating problem, $\mu = 1$ fixed, L varying.

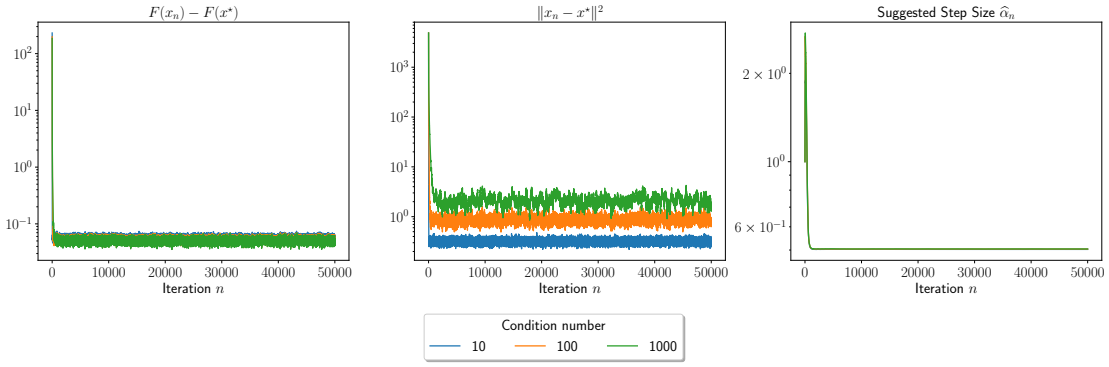


Figure 8.3: SGD with constant step sizes $\alpha = \frac{1}{L}$. Non-interpolating problem, $L = 1$ fixed, μ varying.

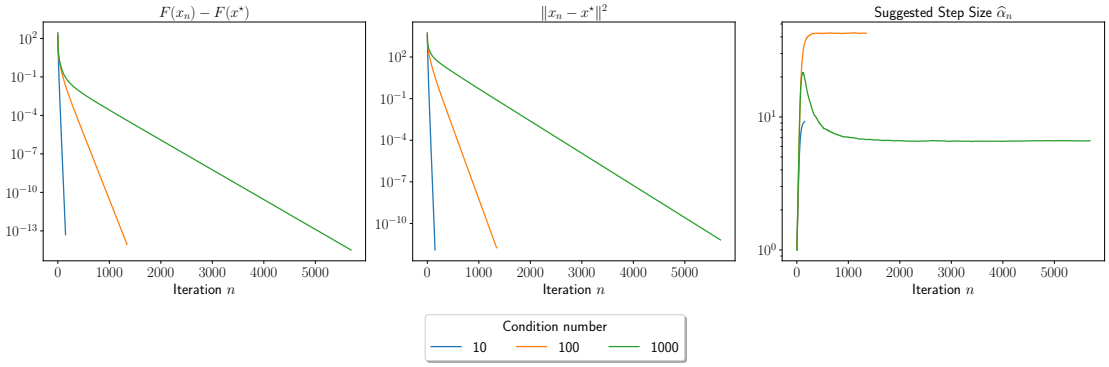


Figure 8.4: SGD with constant step sizes $\alpha = \frac{1}{L}$. Interpolating problem, $L = 1$ fixed, μ varying.

step sizes to demonstrate the convergence behavior described in Chapter 7. For these experiments we employ the alternative variance estimator introduced in Section 4.2.5.

The results agree with the theoretical predictions in Section 3.3.1 as well as with the convergence analysis of the estimators presented in Chapter 7. In particular, in the interpolating case we observe convergence of the iterates to the solution, while in the non-interpolating case a positve suboptimality gap remains. In the non-interpolating regime, we further observe that the suggested step sizes converge as described in Chapter 7.

Contrary to the theory and experiments in Chapter 7, we did not use the *true* value of L as the estimate, but used the estimator from Section 4.2.2. We also observe the characteristic convergence rates: In the interpolating case we obtain linear convergence of both, the functional values and the iterates, in the non-interpolating case we observe linear convergence up to the remaining suboptimality gap. Note that the horizontal axis is substantially shorter in the interpolating case: We have terminated the algorithm, once we have reached $F(x_n) - F(x^\star) < 1 \times 10^{-15}$. The plots do not appear to reach this threshold, because the displayed values represent the mean across five independent runs. The number of iterations required to reach 10^{-15} differs across runs. We therefore only plot the averages up to the minimum number of iterations required by any run to reach the threshold. Since other runs have not yet reached the threshold at that iteration, their larger residuals dominate the mean, causing the curve to remain above 10^{-15}.

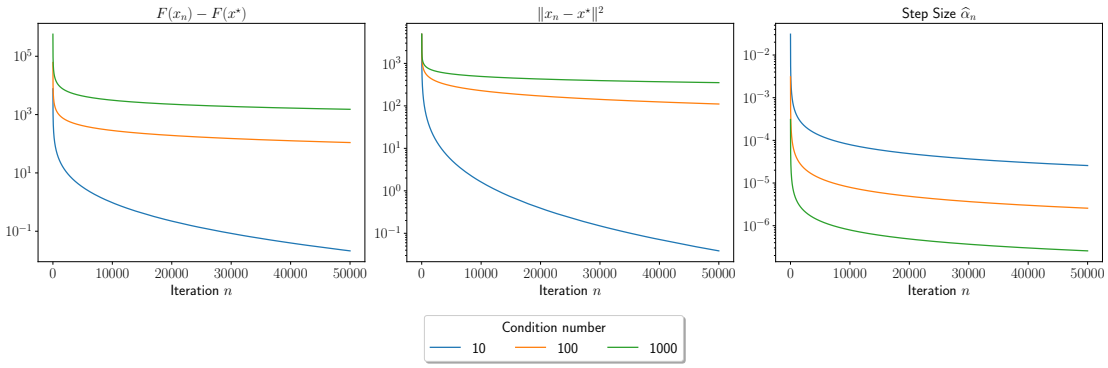


Figure 8.5: SGD with Robbins–Monroe Step Sizes (see Section 3.3.2) $\alpha = \frac{1}{2 \cdot L \cdot n^{0.7}}$. Non-interpolating problem, $\mu = 1$ fixed, L varying.

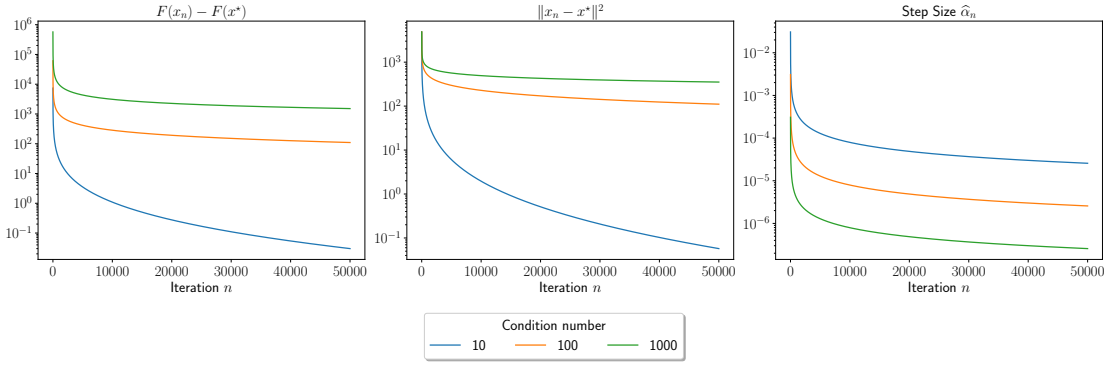


Figure 8.6: SGD with Robbins–Monroe Step Sizes (see Section 3.3.2) $\alpha = \frac{1}{2 \cdot L \cdot n^{0.7}}$. Interpolating problem, $\mu = 1$ fixed, L varying.

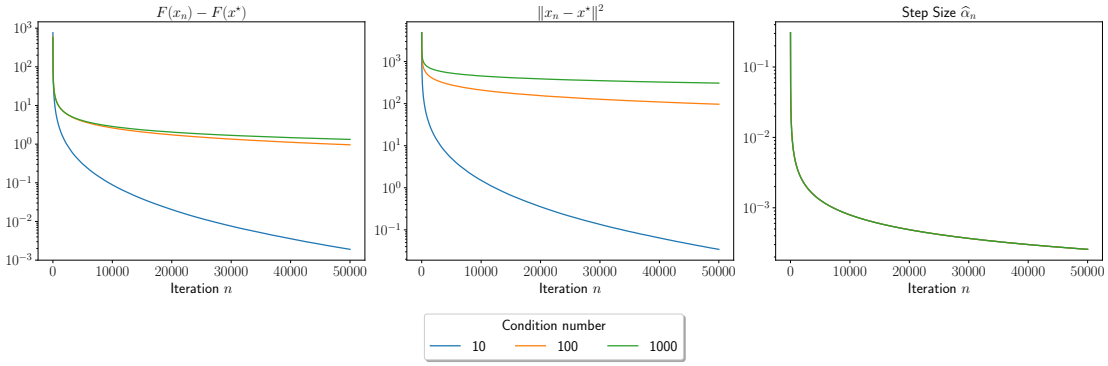


Figure 8.7: SGD with Robbins–Monroe Step Sizes (see Section 3.3.2) $\alpha = \frac{1}{2 \cdot L \cdot n^{0.7}}$. Non-interpolating problem, $L = 1$ fixed, μ varying.

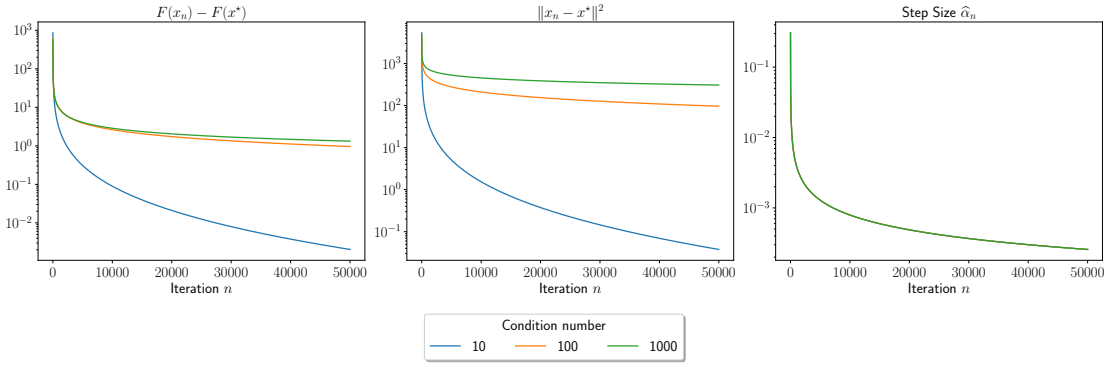


Figure 8.8: SGD with Robbins–Monroe Step Sizes (see Section 3.3.2) $\alpha = \frac{1}{2 \cdot L \cdot n^{0.7}}$. Interpolating problem, $L = 1$ fixed, μ varying.

8.3 Performance of Robbins–Monroe Step Sizes

In this section we present the results obtained using a particular instance of Robbins–Monroe step sizes. For the experiments we choose

$$\alpha_n = \frac{1}{2 \cdot L \cdot n^{0.7}}.$$

Clearly, these step sizes satisfy the condition (3.16). We present the results for the interpolating problems in Figures 8.6 and 8.8, and for the non-interpolating problems in Figures 8.5 and 8.7. In each experiment we plot the suboptimally $F(x_n) - F(x^\star)$ on the left, the squared distance to the minimizer $\|x_n - x^\star\|^2$ in the center and the used Robbins–Monroe step size α_n on the right. In all experiments we observe convergence of both the iterates and the function values, as predicted by the theory. However, in contrast to the experiments with constant step sizes, we do not observe linear convergence for the non-interpolating problems. This is a consequence of the decay of the step sizes, which slows convergence and is unnecessary in the interpolating regime. These observations support the discussion in Section 7.2.2, which highlights that forcing step sizes to converge to zero may substantially reduce the convergence speed for interpolating problems. Conversely, in non-interpolating settings, maintaining step sizes bounded away from zero prevents the iterates from converging to the optimum.

8.4 Performance of Adaptive Step Sizes

We now discuss the performance of the adaptive step size strategies developed in Chapter 4. We consider both the classical variance estimator derived in (4.7) and the alternative estimator introduced in (4.8). For the numerical experiments in this section we

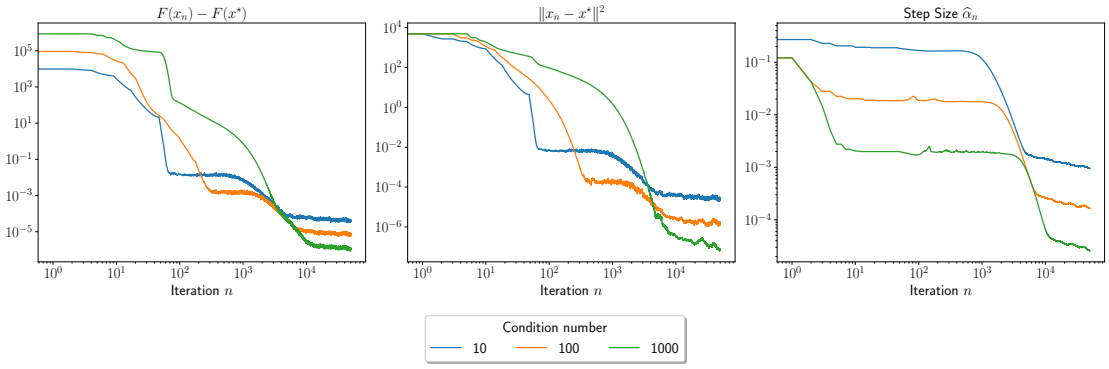


Figure 8.9: SGD with adaptive Step Sizes. Non-interpolating problem, $\mu = 1$ fixed, L varying. Classical variance estimation due to Section 4.2.5.

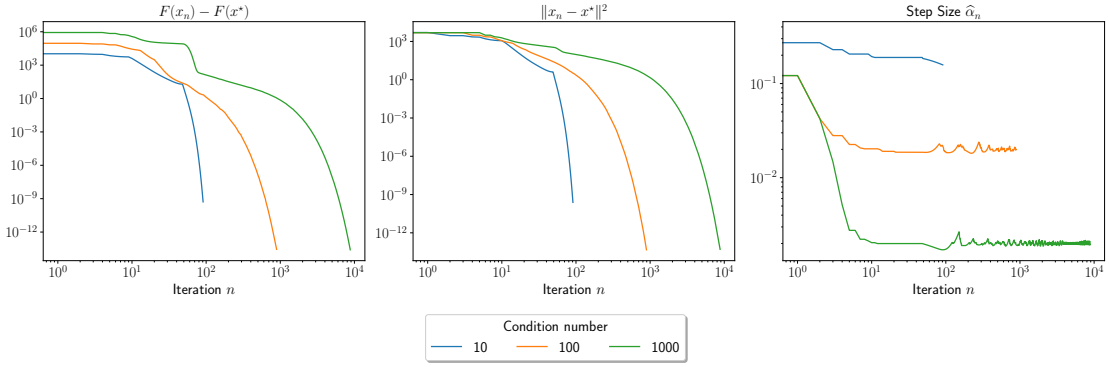


Figure 8.10: SGD with adaptive Step Sizes. Interpolating problem, $\mu = 1$ fixed, L varying. Classical variance estimation due to Section 4.2.5.

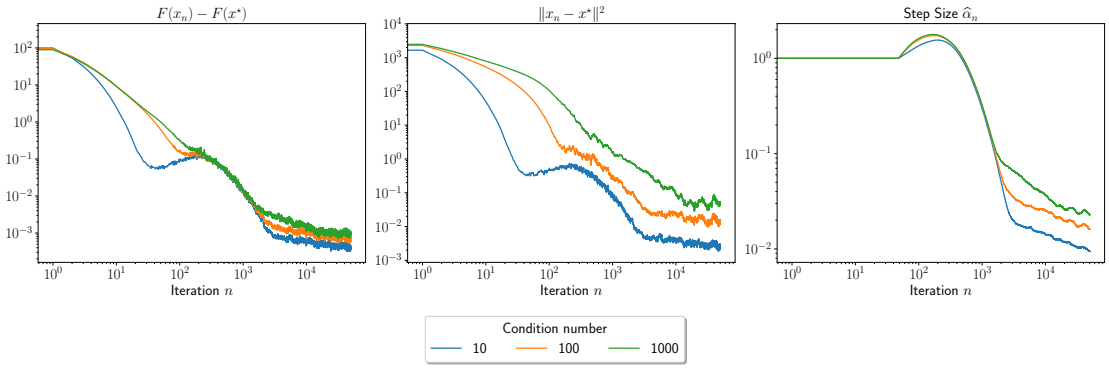


Figure 8.11: SGD with adaptive Step Sizes. Non-interpolating problem, $L = 1$ fixed, μ varying. Classical variance estimation due to Section 4.2.5.

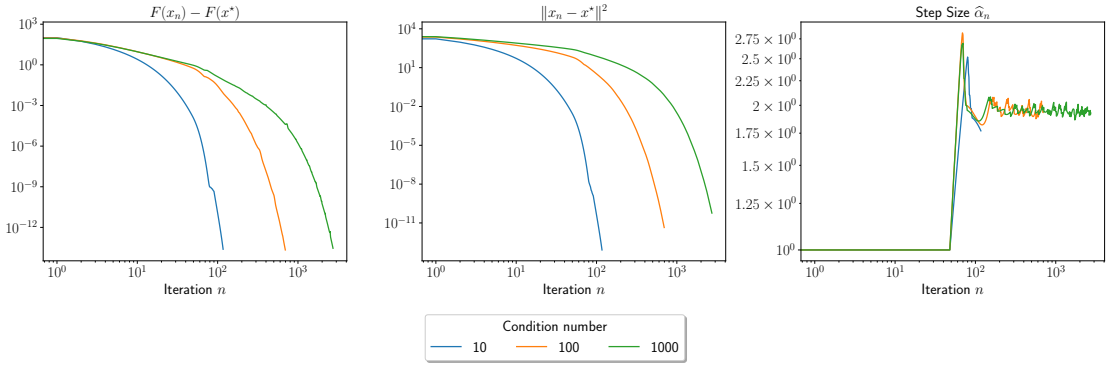


Figure 8.12: SGD with adaptive Step Sizes. Interpolating problem, $L = 1$ fixed, μ varying. Classical variance estimation due to Section 4.2.5.

incorporate several practical enhancements to improve the robustness and effectiveness of the algorithm, as discussed in Section 4.3.

In particular, during the execution of the algorithm we apply the following modifications.

- If a step results in an increase of the sampled functional value, we reject the step and perform a line search on the next sample.
- We use p-EMA with a lower bound on the γ-value as described in (4.10). We choose $p = 0.75$ and $\gamma_{\min} = 0.95$.
- We additionally smooth the estimated step sizes obtained from the estimators using classical EMA (see Equation (6.2)) with parameter $\gamma = 0.95$, i.e. we set

$$\widehat{\alpha}_{n+1} = \gamma \widehat{\alpha}_n + (1 - \gamma)\widetilde{\alpha}_n,$$

  where $(\widehat{\alpha}_n)$ denote the step sizes used, and $(\widetilde{\alpha}_n)$ the step sizes obtained from the estimators.
- We employ an initialization phase of $n_{\text{init}} = 1000$ iterations during which the algorithm does not yet use the step sizes suggested by the estimators. During this phase step sizes are adapted only via sample-wise line searches on samples, which are performed at the start of the algorithm and in the case of ascent on samples.

The results for the classical variance estimation due to Equation (4.7) are depicted in Figures 8.9 to 8.12. The results for the alternative variances estimation as discussed in Section 4.2.5 are depicted in Figures 8.13 to 8.16. For better visualization of the asymptotic effects of the step size selection, we display the horizontal axis on a logarithmic scale.

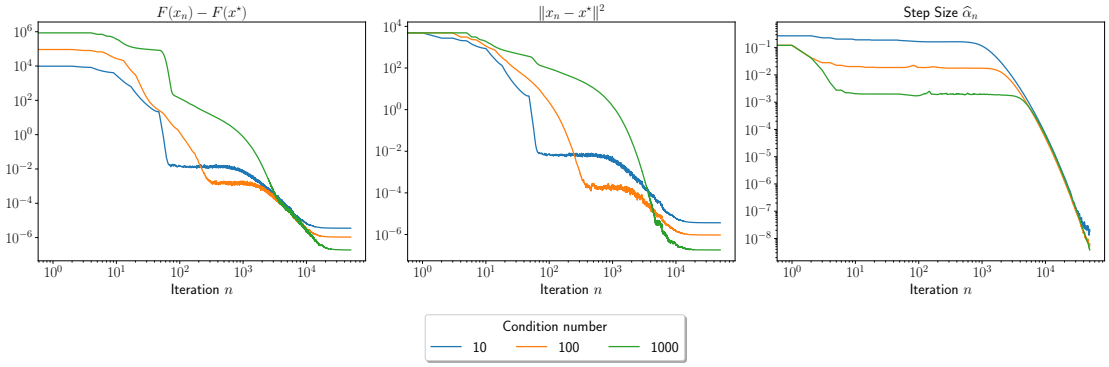


Figure 8.13: SGD with adaptive Step Sizes. Non-interpolating problem, $\mu = 1$ fixed, L variable. Alternative variance estimation due to Section 4.2.5.

Let us first focus on the results obtained using the classical variance estimation scheme. In the interpolating setting, shown in Figures 8.10 and 8.12, we observe that the adaptive step sizes behave approximately like $\frac{1}{L}$. This indicates that the estimator for the Lipschitz constant of the gradient yields a suitable estimate, and the variance correction factor

$$1 - \frac{\widehat{\sigma}_n^2}{\widehat{g}_n}$$

does not converge to zero. As a consequence, we observe linear convergence of both the function values and the iterates. In the non-interpolating regime, shown in Figures 8.9 and 8.11 we first observe that the step sizes indeed are reduced gradually as the algorithm progresses. Initially, after approximately ten iterations, the step sizes behave like $\frac{1}{L}$. This behavior is apparent both in the experiments with variable L (Figure 8.9), where different condition numbers produce visibly different initial step sizes, and in the experiments with fixed L (Figure 8.11), where all initial step sizes cluster near $\frac{1}{L} = 1$. Once the algorithm begins to stagnate, the step sizes decrease, which in turn enables further reductions in both the functional value and the distance to the minimizer.

We now turn to the results obtained with the alternative variance estimator introduced in Section 4.2.5. Although we have shown theoretically and demonstrated numerically in Chapter 7 that this estimator yields the same asymptotic limit for the estimated step sizes when SGD is run with constant step sizes, the behaviour changes markedly when the estimator is used inside the adaptive algorithm.

In the non-interpolating regime, illustrated in Figures 8.13 and 8.15, the alternative variance estimator leads to noticeably different behaviour than the classical one. In these experiments the step sizes are again reduced to zero, which is a desired qualitative feature. However, the reduction occurs significantly faster than in the classical variance

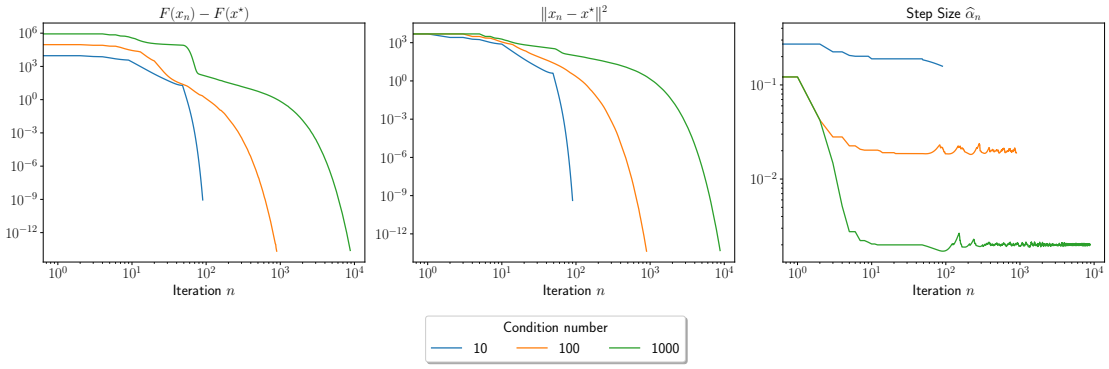


Figure 8.14: SGD with adaptive Step Sizes. Interpolating problem, $\mu = 1$ fixed, L variable. Alternative variance estimation due to Section 4.2.5.

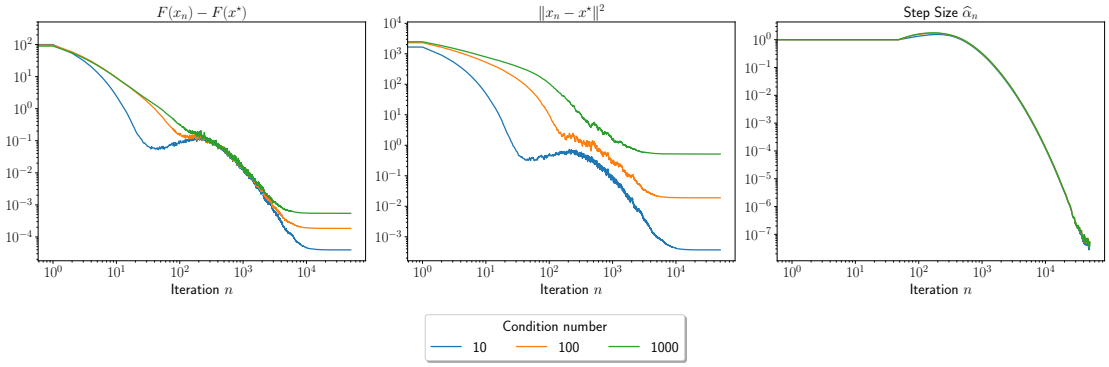


Figure 8.15: SGD with adaptive Step Sizes. Non-interpolating problem, $L = 1$ fixed, μ variable. Alternative variance estimation due to Section 4.2.5.

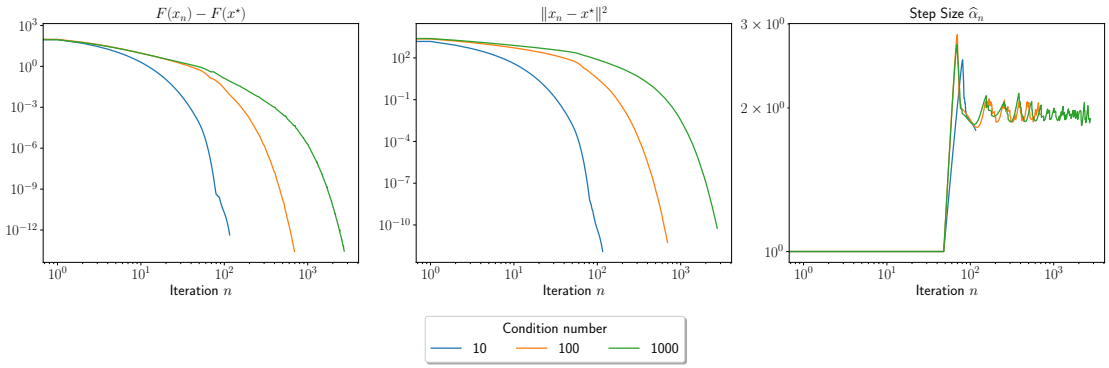


Figure 8.16: SGD with adaptive Step Sizes. Interpolating problem, $L = 1$ fixed, μ variable. Alternative variance estimation due to Section 4.2.5.

estimation approach. This rapid decay is evident from a direct comparison of the step size curves in Figures 8.13 and 8.15 with those in Figures 8.9 and 8.11. The resulting small step sizes subsequently lead to a seemingly stagnating algorithm in Figures 8.13 and 8.15, as very little progress can be made with step sizes of order 1×10^{-5} or smaller. In contrast, in the interpolating regime the alternative variance estimation behaves similar to the classical variance estimation: We observe linear convergence of the iterates and the functional values, and step sizes of approximately $\frac{1}{L}$.

Further Numerical Experiments

We have also tested the adaptive step size algorithm on standard image classification tasks. In these settings the resulting SOP is typically highly non-convex and, due to the activation function used in modern neural network architectures, is usually non-differentiable. Despite this lack of differentiability, SGD is routinely applied in practice. For the use of our adaptive step sizes, the non-convexity of the mean objective F poses substantial difficulties. On non-convex problems we have observed that SGD using the ideal step size defined in Equation (4.2) tends to drive the iterates toward a local minimum usually located near the initialization. Consequently, the method loses the desirable ability of SGD to escape *bad* local minima and saddle points. This phenomenon is well documented in the literature, see for example [25, 34, 37].

To mitigate this behavior, additional globalization strategies must be incorporated into the algorithm. We refer to [41] for a description of such heuristic modifications and for corresponding numerical experiments.

8.5 Complete Algorithm in Pseudocode

For reference, we provide below the algorithm used in the numerical experiments in pseudocode.

For readability, we have divided the algorithm into several parts. Algorithm 3 describes the initialization performed at the beginning. This initialization phase is based on the simple line search presented in Algorithm 2. Algorithm 4 presents one iteration of our method, while Algorithm 5 summarizes the overall procedure, including the handling of ascent on a single sample. We have selected the following parameters and strategies for our implementation. For more detailed explanations, we refer to Section 4.3.

- We set the parameter p of the p-EMA to $p = 0.75$.

- We set the minimum value $\gamma_{\min}$ for the initially damped p-EMA (see Equation (4.10)) to $\gamma_{\min} = 0.95$.
- We set the number of iterations performed before using the estimators for step-size selection to $n_{\text{init}} = 50$.
- We set the initial step size to $\alpha_0 = 1$.
- We employed EMA to further smooth the estimated step sizes, using the parameter $\gamma = 0.99$.

Algorithm 2: Sample-wise Line Search (`SimpleLineSearch`)

Input:

- Sample-wise objective f_ω
- Current iterate x
- Initial step size α
- Step reduction factor $\eta \in (0, 1)$
- Max. number of inner iterations $n_{\max}$

$n \leftarrow 0$

$\mathsf{success} \leftarrow$ false

while $f_\omega(x - \alpha \nabla f_\omega(x)) > f_\omega(x)$ ***and*** $n < n_{\max}$ **do**

  $\alpha \leftarrow \eta\alpha$

  $n \leftarrow n + 1$

end

if $f_\omega(x - \alpha \nabla f_\omega(x)) \leq f_\omega(x)$ **then**

  $\mathsf{success} \leftarrow$ true

end

return $(\alpha, \mathsf{success})$

Algorithm 3: Stochastic Initialization using Line Search (`Initialize`)

Input:

- Initial iterate x_0
- Initial step size estimate α
- Probability measure $\mathbb{P}$
- Maximum line search iterations $n_{\max}$

Sample $\omega \sim \mathbb{P}$

$(\alpha, \mathsf{success}) \leftarrow$ `SimpleLineSearch`(f_ω, x_0, α)

while $\neg\mathsf{success}$ **do**

  Sample $\omega \sim \mathbb{P}$

  $(\alpha, \mathsf{success}) \leftarrow$ `SimpleLineSearch`(f_ω, x_0, α)

end

$\omega_0 \leftarrow \omega$

$\alpha_0 \leftarrow \alpha$

$x_1 \leftarrow x_0 - \alpha_0 \nabla f_{\omega_0}(x_0)$

$\widehat{L} \leftarrow 1/\alpha_0$

$\widehat{\sigma}^2 \leftarrow 0$

$\widehat{g} \leftarrow \left\| f'_{\omega_0}(x_0) \right\|^2_{\mathcal{H}^*}$

return $(x_1, \alpha_0, \omega_0, \widehat{L}, \widehat{\sigma}^2, \widehat{g})$

Algorithm 4: One Iteration of Adaptive SGD (`PerformIteration`)

Input:

- Current state $(x_n, \alpha_n, \omega_{n-1}, \widehat{L}, \widehat{\sigma}^2, \widehat{g})$
- Probability measure $\mathbb{P}$
- Maximum line search iterations $n_{\max}$
- Parameter p for p-EMA

Sample $\omega_n \sim \mathbb{P}$

$x_{n+1} \leftarrow x_n - \alpha_n \nabla f_{\omega_n}(x_n)$

if $f_{\omega_n}(x_{n+1}) > f_{\omega_n}(x_n)$ **then**

  $x_{n+1} \leftarrow x_n$

  Sample $\omega \sim \mathbb{P}$

  $(\alpha_n, \mathsf{success}) \leftarrow \texttt{SimpleLineSearch}(f_\omega, x_n, \alpha_n)$

  while $\neg\mathsf{success}$ **do**

    Sample $\omega \sim \mathbb{P}$

    $(\alpha_n, \mathsf{success}) \leftarrow \texttt{SimpleLineSearch}(f_\omega, x_n, \alpha_n)$

  end

  return $(x_{n+1}, \alpha_n, \omega_n, \widehat{L}, \widehat{\sigma}^2, \widehat{g})$

end

Update $\widehat{\sigma}^2$ using p-EMA with observation

$$\frac{f_{\omega_n}(x_n) - f_{\omega_{n-1}}(x_n)}{\alpha_n}$$

Update $\widehat{L}$ using p-EMA with observation

$$2\frac{f_{\omega_n}(x_{n+1}) - f_{\omega_n}(x_n) + \alpha_n \left\| f'_{\omega_n}(x_n) \right\|^2_{\mathcal{H}^*}}{\alpha_n^2 \left\| f'_{\omega_n}(x_n) \right\|^2_{\mathcal{H}^*}}$$

Update $\widehat{g}$ using p-EMA with observation

$$\left\| f'_{\omega_n}(x_n) \right\|^2_{\mathcal{H}^*}$$

if $n > n_{\text{init}}$ **then**

  Compute the step size α_{n+1} using EMA with α_n and observation

$$\left\| f'_{\omega_n}(x_n) \right\|^2_{\mathcal{H}^*}$$

end

return $(x_{n+1}, \alpha_{n+1}, \omega_n, \widehat{L}, \widehat{\sigma}^2, \widehat{g})$

Algorithm 5: Adaptive SGD (Full Procedure)

Input:

- Initial iterate x_0
- Initial step size estimate α
- Probability measure $\mathbb{P}$

$(x_1, \alpha_1, \widehat{L}, \widehat{\sigma}^2, \widehat{g}) \leftarrow \texttt{Initialize}(x_0, \alpha, \dots)$

for $n \geq 1$ **do**

  $(x_{n+1}, \alpha_{n+1}, \omega_n, \widehat{L}, \widehat{\sigma}^2, \widehat{g}) \leftarrow \texttt{PerformIteration}(x_n,\ \alpha_n,\ \dots)$

end

Conclusion and Outlook

This work has addressed the issue of selecting step sizes for Stochastic Gradient Descent (SGD). We have identified a step size rule (*ideal step sizes*) leading to optimal convergence rates of SGD on convex problems, in both, the interpolating and non-interpolating setting. We have proposed a computable version of this step size control that adapts the step sizes online, during the run of the algorithm, based on information gathered during the pervious iterations. While the convergence theory for the ideal step sizes yields optimal convergence rates, we were not able to show convergence rates for SGD with the adaptive step sizes. Instead, we have considered the behavior of the estimated step sizes, if the SGD is run with a constant step size, and the adaptive step sizes are evaluated, but not used. We have demonstrated, both theoretically and numerically, that the adaptive step size scheme we propose is able to effectively distinguish between the interpolating and non-interpolating setting. In the former, the suggested step sizes remain bounded away from zero, while in the latter they are reduced to zero, which crucial in this case to ensure convergence.

A more refined algorithm and analysis might be able to leverage these results to obtain theoretical convergence guarantees with computable adaptive step sizes. One possibility might be to run SGD with a constant step size, until stagnation of the algorithm is observed, and then consider the estimated step size. In the non-interpolating setting, the estimated step size will be smaller than the current step size eventually, and using this smaller step size for the next iterations will enable the algorithm to progress further, until stagnation is observed again. Clearly this raises several additional questions: How to detect stagnation? How long to wait, if stagnation is detected, until the estimated step size is considered reliable? Should there be a minimum number of iterations between two reductions in the step size? Further research picking up our algorithmic ideas might deal with such concerns.

The restriction to convex problems is another limitation of the present work. In many applications, the target function F is non-convex. While it is possible to show that, using the ideal step sizes, SGD yields iterates x_n such that the gradients of $\nabla F(x_n)$ converge to zero, the limit point of x_n (if existent) might be a local minimizer with undesirable properties. Additionally, in the non-convex setting the theory on the convergence of the estimators does not hold as presented it in this work. For example, the invariant

measure might not be unique, the attractor might be consists of several components, located around several local minimizers, and transition between the components might or might not be possible.[1] We have mentioned some of the possible approaches to deal with this non-convexity in the discussion at the end of Chapter 8 and in our Paper [41], however, an adaptation to non-convex problems could clearly benefit from more sophisticated ideas.

Despite these concerns, the concepts in the present work seem to perform well on numerical test cases, even beyond theoretical boundaries. Additionally, the presented estimation techniques are merely one possibility to obtain an approximation for the ideal step sizes. If other estimates for each of the three key quantities are available, they can simply be used instead of our estimators. Such an approach is documented in the recent Preprint [4], where a different estimate for the variance is available. Further, the general theory on the smoothing technique p-EMA, and the resulting insights on the behavior of the estimators under SGD with constant step size could be utilized to develop and analyze different approaches to evaluate the progress and behavior of SGD during its execution, not only in the case of step size selection.

[1] Such a scenario is considered in [64].

Bibliography

[1] Ludwig Arnold. *Random Dynamical Systems*. Springer Monographs in Mathematics. Springer Berlin Heidelberg, 1998. ISBN: 9783662128787. DOI: 10.1007/978-3-662-12878-7.

[2] Waïss Azizian et al. *What is the long-run distribution of stochastic gradient descent? a large deviations analysis*. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria. (2024).

[3] Lukas Balles, Cedric Archambeau, and Giovanni Zappella. *A Negative Result on Gradient Matching for Selective Backprop*. 2023. arXiv: 2312.05021 [cs.LG]. URL: https://arxiv.org/abs/2312.05021.

[4] Niklas Baumgarten and David Schneiderhan. *Multilevel Stochastic Gradient Descent for Optimal Control Under Uncertainty*. 2025. arXiv: 2506.02647 [math.OC].

[5] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer International Publishing, 2017. ISBN: 9783319483115. DOI: 10.1007/978-3-319-48311-5.

[6] Amir Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. Society for Industrial and Applied Mathematics, Oct. 2014. ISBN: 9781611973655. DOI: 10.1137/1.9781611973655.

[7] Michel Benaïm and Tobias Hurth. *Markov Chains on Metric Spaces: A Short Course*. Universitext. Springer International Publishing, 2022. ISBN: 9783031118227. DOI: 10.1007/978-3-031-11822-7.

[8] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2008, p. 738. ISBN: 9780387310732.

[9] Jose Blanchet et al. *Convergence Rate Analysis of a Stochastic Trust-Region Method via Supermartingales*. In: INFORMS Journal on Optimization 1(2) (Apr. 2019). Pp. 92–119. DOI: 10.1287/ijoo.2019.0016.

[10] Vladimir I. Bogachev. *Measure Theory*. Springer Berlin Heidelberg, 2007. ISBN: 9783540345145. DOI: 10.1007/978-3-540-34514-5.

[11] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. *Optimization Methods for Large-Scale Machine Learning*. In: SIAM Review 60(2) (2018). Pp. 223–311. DOI: 10.1137/16M1080173.

[12] Robert Goodell Brown. *Exponential Smoothing for Predicting Demand*. 1956. URL: https://www.industrydocuments.ucsf.edu/docs/jzlc0130/.

[13] Alexander Camuto et al. *Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms*. June 2021. arXiv: 2106.04881 [stat.ML].

[14] Zaiwei Chen, Shancong Mou, and Siva Theja Maguluri. *Stationary Behavior of Constant Stepsize SGD Type Algorithms: An Asymptotic Characterization*. In: Proceedings of the ACM on Measurement and Analysis of Computing Systems 6(1) (Feb. 2022). Pp. 1–24. DOI: 10.1145/3508039.

[15] Frank E. Curtis and Rui Shi. *A fully stochastic second-order trust region method*. In: Optimization Methods and Software 37(3) (Nov. 2020). Pp. 844–877. DOI: 10.1080/10556788.2020.1852403.

[16] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. *SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives*. In: Advances in Neural Information Processing Systems. (2014). URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/ede7e2b6d13a41ddf9f4bdef84fdc737-Paper.pdf.

[17] Aaron Defazio and Konstantin Mishchenko. *Learning-rate-free learning by D-Adaptation*. In: Proceedings of the 40th International Conference on Machine Learning. Pp. 7449–7479. (2023). URL: https://proceedings.mlr.press/v202/defazio23a.html.

[18] Aymeric Dieuleveut, Alain Durmus, and Francis Bach. *Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains*. In: The Annals of Statistics 48(3) (2020). pp. 1348–1382. URL: https://www.jstor.org/stable/26931514.

[19] John Duchi, Elad Hazan, and Yoram Singer. *Adaptive subgradient methods for online learning and stochastic optimization*. In: Journal of Machine Learning Research 12(61) (2011). Pp. 2121–2159. URL: http://jmlr.org/papers/v12/duchi11a.html.

[20] Rick Durrett. *Probability. Theory and Examples*. 5th ed. Cambridge University Press, 2019. DOI: 10.1017/9781108591034.

[21] Omar Elharrouss et al. *Task-based Loss Functions in Computer Vision: A Comprehensive Review*. 2025. arXiv: 2504.04242 [cs.LG].

[22] Franco Flandoli, Benjamin Gess, and Michael Scheutzow. *Synchronization by noise*. In: Probability Theory and Related Fields 168(3–4) (May 2016). Pp. 511–556. DOI: 10.1007/s00440-016-0716-2.

[23] Curtis Fox et al. *Nonmonotone line searches operate at the edge of stability*. In: OPT 2024: Optimization for Machine Learning. (2024). URL: https://openreview.net/forum?id=8XD0PMTNZm.

[24] Guillaume Garrigos and Robert M. Gower. *Handbook of Convergence Theorems for (Stochastic) Gradient Methods*. Jan. 2023. arXiv: 2301.11235 [math.OC].

[25] Rong Ge et al. *Escaping From Saddle Points - Online Stochastic Gradient for Tensor Decomposition*. In: Annual Conference Computational Learning Theory. (2015). URL: https://api.semanticscholar.org/CorpusID:11513606.

[26] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[27] Martin Hairer. *Ergodic properties of Markov processes*. Lecture notes. 2006. URL: https://hairer.org/notes/Markov.pdf.

[28] O. Hernández-Lerma. *Markov Chains and Invariant Probabilities*. Birkhäuser Basel, 2003, p. 206. ISBN: 9783034894081.

[29] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. *Lecture Notes: Neural Networks for Machine Learning. Lecture 6a: Overview of mini-batch gradient descent*. 2012. URL: https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

[30] Charles C. Holt. *Forecasting seasonals and trends by exponentially weighted moving averages*. In: International Journal of Forecasting 20(1) (2004). Pp. 5–10. DOI: 10.1016/j.ijforecast.2003.09.015.

[31] Maor Ivgi, Oliver Hinder, and Yair Carmon. *DoG is SGD's best friend: a parameter-free dynamic step size schedule*. In: Proceedings of the 40th International Conference on Machine Learning. Pp. 14465–14499. (2023). URL: https://proceedings.mlr.press/v202/ivgi23a.html.

[32] Angela H. Jiang et al. *Accelerating Deep Learning by Focusing on the Biggest Losers*. 2019. arXiv: 1910.00762 [cs.LG]. URL: https://arxiv.org/abs/1910.00762.

[33] Xiaowen Jiang and Sebastian U. Stich. *Adaptive SGD with Polyak stepsize and line-search: robust convergence and variance reduction*. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. (2023). URL: https://papers.nips.cc/paper_files/paper/2023/file/540eb9e0ee35d525231c3fd22d1dcbf2-Paper-Conference.pdf.

[34] Chi Jin et al. *How to escape saddle points efficiently*. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. Pp. 1724–1732. Sydney, NSW, Australia. (2017).

[35] Rie Johnson and Tong Zhang. *Accelerating stochastic gradient descent using predictive variance reduction*. In: Advances in Neural Information Processing Systems. (2013). URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/ac1dd209cbcc5e5d1c6e28598e8cbbe8-Paper.pdf.

[36] Diederik P. Kingma and Jimmy Ba. *Adam: a method for stochastic optimization*. In: 3rd International Conference on Learning Representations, ICLR 2015. San Diego. (2015).

[37] Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. *An Alternative View: When Does SGD Escape Local Minima?* In: Proceedings of the 35th International Conference on Machine Learning. Pp. 2698–2707. (July 2018). URL: https://proceedings.mlr.press/v80/kleinberg18a.html.

[38] Achim Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer International Publishing, 2020. ISBN: 9783030564025. DOI: 10.1007/978-3-030-56402-5.

[39] Oliver Knill. *Probability theory and stochastic processes with applications*. Overseas Press India Private Limitied, 2009, p. 373. ISBN: 9788189938406.

[40] Frederik Köhne and Anton Schiela. *An Exponential Averaging Process with Strong Convergence Properties*. May 2025. arXiv: 2505.10605 [stat.ML].

[41] Frederik Köhne et al. *Adaptive Step Sizes for Preconditioned Stochastic Gradient Descent*. Nov. 2023. arXiv: 2311.16956 [math.OC].

[42] V. M. Korchevsky and V. V. Petrov. *On the strong law of large numbers for sequences of dependent random variables*. In: Vestnik St. Petersburg University: Mathematics 43(3) (Sept. 2010). Pp. 143–147. DOI: 10.3103/s1063454110030040.

[43] Ulrich Krengel. *Ergodic Theorems*. Berlin, New York: De Gruyter, 1985. ISBN: 9783110844641. DOI: 10.1515/9783110844641.

[44] Kalimuthu Krishnamoorthy. *Handbook of Statistical Distributions with Applications*. Statistics: Textbooks and Monographs. Chapman & Hall/CRC, Boca Raton, FL, 2006. DOI: 10.1201/9781420011371.

[45] Herb Kunze et al. *Fractal-Based Methods in Analysis*. Springer US, 2012. ISBN: 9781461418917. DOI: 10.1007/978-1-4614-1891-7.

[46] Yanli Liu, Yuan Gao, and Wotao Yin. *An improved analysis of stochastic gradient descent with momentum*. In: Advances in Neural Information Processing Systems. Red Hook, NY, USA. (2020).

[47] Nicolas Loizou et al. *Stochastic Polyak step-size for SGD: an adaptive learning rate for fast convergence*. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics. Pp. 1306–1314. (2021). URL: https://proceedings.mlr.press/v130/loizou21a.html.

[48] Ibrahim Merad and Stéphane Gaïffas. *Convergence and concentration properties of constant step-size SGD through Markov chains*. In: Electronic Journal of Statistics 19(2) (2025). Pp. 5843–5894. DOI: 10.1214/25-EJS2471.

[49] Konstantin Mishchenko and Aaron Defazio. *Prodigy: An Expeditiously Adaptive Parameter-Free Learner*. In: Proceedings of the 41st International Conference on Machine Learning. Pp. 35779–35804. (July 2024). URL: https://proceedings.mlr.press/v235/mishchenko24a.html.

[50] Konstantin Mishchenko and Aaron Defazio. *Prodigy: an expeditiously adaptive parameter-free learner*. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria. (2024).

[51] Georg Müller. "Optimal Control of Time-Discretized Contact Problems". PhD thesis. University of Bayreuth, 2019. urn: urn:nbn:de:bvb:703-epub-4379-0.

[52] Yurri Nesterov. *Introductory lectures on convex optimization. a basic course*. Kluwer Academic Publishers, 2004, p. 236. ISBN: 1402075537.

[53] Julian Newman. *Necessary and sufficient conditions for stable synchronization in random dynamical systems*. In: Ergodic Theory and Dynamical Systems 38 (2014). Pp. 1857–1875. URL: https://api.semanticscholar.org/CorpusID:119326733.

[54] Julian Newman. *Synchronisation of almost all trajectories of a random dynamical system*. In: Discrete & Continuous Dynamical Systems - A 40(7) (2020). Pp. 4163–4177. DOI: 10.3934/dcds.2020176.

[55] Courtney Paquette and Katya Scheinberg. *A stochastic line search method with expected complexity analysis*. In: SIAM Journal on Optimization 30(1) (Jan. 2020). Pp. 349–376. DOI: 10.1137/18m1216250.

[56] B. T. Poljak. *Some methods of speeding up the convergence of iterative methods*. In: Žurnal Vyčislitel' noĭ Matematiki i Matematičeskoĭ Fiziki 4 (1964). Pp. 791–803.

[57] Boris T. Polyak. *Introduction to Optimization*. New York: Optimization Software, Inc., 1987.

[58] Yu. V. Prokhorov. *Convergence of Random Processes and Limit Theorems in Probability Theory*. In: Theory of Probability & Its Applications 1(2) (Jan. 1956). Pp. 157–214. DOI: 10.1137/1101016.

[59] H. Robbins and D. Siegmund. "A Convergence Theorem for Non Negative Almost Supermartingales and Some Applications". In: *Herbert Robbins Selected Papers*. Springer New York, 1985, pp. 111–135. ISBN: 9781461251101. DOI: 10.1007/978-1-4612-5110-1_10.

[60] Herbert Robbins and Sutton Monro. *A Stochastic Approximation Method*. In: The Annals of Mathematical Statistics 22(3) (Sept. 1951). Pp. 400–407. DOI: 10.1214/aoms/1177729586.

[61] Sebastian Ruder. *An overview of gradient descent optimization algorithms*. Sept. 2016. arXiv: 1609.04747.

[62] Mark Schmidt, Nicolas Le Roux, and Francis Bach. *Minimizing finite sums with the stochastic average gradient*. In: Mathematical Programming 162(1–2) (June 2016). Pp. 83–112. DOI: 10.1007/s10107-016-1030-6.

[63] Othmane Sebbouh, Robert M Gower, and Aaron Defazio. *Almost sure convergence rates for Stochastic Gradient Descent and Stochastic Heavy Ball*. In: Proceedings of Thirty Fourth Conference on Learning Theory. Pp. 3935–3971. (Aug. 2021). URL: https://proceedings.mlr.press/v134/sebbouh21a.html.

[64] David Shirokoff and Philip Zaleski. *Convergence of Markov Chains for Constant Step-size Stochastic Gradient Descent with Separable Functions*. Sept. 2024. arXiv: 2409.12243 [math.OC].

[65] Sharan Vaswani, Benjamin Dubois-Taine, and Reza Babanezhad. *Towards Noise-adaptive, Problem-adaptive (accelerated) Stochastic Gradient Descent*. In: Proceedings of the 39th International Conference on Machine Learning. Pp. 22015–22059. (July 2022). URL: https://proceedings.mlr.press/v162/vaswani22a.html.

[66] Sharan Vaswani et al. *Adaptive gradient methods converge faster with over-parameterization (but you should do a line-search)*. June 2020. arXiv: 2006.06835.

[67] Sharan Vaswani et al. *Painless stochastic gradient: interpolation, line-search, and convergence rates*. In: Advances in Neural Information Processing Systems. (May 2019). URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/2557911c1bf75c2b643afb4ecbfc8ec2-Paper.pdf.

[68] Martin Weiser, Peter Deuflhard, and Bodo Erdmann. *Affine conjugate adaptive Newton methods for nonlinear elastomechanics*. In: Optimization Methods & Software 22(3) (2007). Pp. 413–431. DOI: 10.1080/10556780600605129.

[69] Peter R. Winters. *Forecasting Sales by Exponentially Weighted Moving Averages*. In: Management Science 6(3) (1960). Pp. 324–342. URL: http://www.jstor.org/stable/2627346.

[70] Matthew D. Zeiler. *ADADELTA: an adaptive learning rate method*. Dec. 2012. arXiv: 1212.5701.

Own Publications

- Frederik Köhne et al. *Adaptive Step Sizes for Preconditioned Stochastic Gradient Descent*. Nov. 2023. arXiv: 2311.16956 [math.OC].
- Frederik Köhne et al. *L^∞-error Bounds for Approximations of the Koopman Operator by Kernel Extended Dynamic Mode Decomposition*. In: SIAM Journal on Applied Dynamical Systems 24(1) (2025). Pp. 501–529. DOI: 10.1137/24M1650120.
- Roland Herzog et al. *Metric Frobenius norms and inner products of matrices and linear maps*. In: Linear Algebra and its Applications 727 (2025). Pp. 112–128. DOI: 10.1016/j.laa.2025.08.005.
- Frederik Köhne and Anton Schiela. *An Exponential Averaging Process with Strong Convergence Properties*. May 2025. arXiv: 2505.10605 [stat.ML].
- Leonie Kreis et al. *SensLI: Sensitivity-Based Layer Insertion for Neural Networks*. Nov. 2025. arXiv: 2311.15995 [cs.LG].
- Evelyn Herberg, Roland Herzog, and Frederik Köhne. *Time regularization in optimal time variable learning*. In: PAMM 24(1) (2024). DOI: 10.1002/pamm.202300299.

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet habe.

Weiterhin erkläre ich, dass ich die Hilfe von gewerblichen Promotionsberatern bzw. -vermittlern oder ähnlichen Dienstleistern weder bisher in Anspruch genommen habe, noch künftig in Anspruch nehmen werde.

Zusätzlich erkläre ich hiermit, dass ich keinerlei frühere Promotionsversuche unternommen habe.

Hamburg, den 28. Mai 2026 Frederik Köhne