Interpolation, Growth Conditions, and Stochastic Gradient Descent

by

Aaron Mishkin

B.Sc., University of British Columbia, 2018

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Computer Science)

The University of British Columbia
(Vancouver)

September 2020

© Aaron Mishkin, 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Interpolation, Growth Conditions, and Stochastic Gradient Descent

submitted by **Aaron Mishkin** in partial fulfillment of the requirements for the degree of **Master of Science** in **Computer Science**.

Examining Committee:

Mark Schmidt, Computer Science
Supervisor

Nicholas J. A. Harvey, Computer Science
Supervisory Committee Member

Abstract

Current machine learning practice requires solving huge-scale empirical risk minimization problems quickly and robustly. These problems are often highly under-determined and admit multiple solutions which exactly fit, or *interpolate*, the training data. In such cases, stochastic gradient descent has been shown to converge without decreasing step-sizes or averaging, and can achieve the fast convergence of deterministic gradient methods. Recent work has further shown that stochastic acceleration and line-search methods for step-size selection are possible in this restricted setting. Although pioneering, existing analyses for first-order methods under interpolation have two major weaknesses: they are restricted to the finite-sum setting, and, secondly, they are not tight with the best deterministic rates. To address these issues, we extend the notion of interpolation to stochastic optimization problems with general, first-order oracles. We use the proposed framework to analyze stochastic gradient descent with a fixed step-size and with an Armijo-type line-search, as well as Nesterov's accelerated gradient method with stochastic gradients. In nearly all settings, we obtain faster convergence with a wider range of parameters. The improvement for stochastic Nesterov acceleration is comparable to dividing by the square-root of the condition number and addresses criticism that existing rates were not truly "accelerated". In the case of convex functions, our convergence rates for stochastic gradient descent — both with and without the stochastic Armijo line-search — recover the best-known rates in the deterministic setting. We also provide a simple extension to ℓ_2-regularized minimization, which opens the path to proximal-gradient methods and non-smooth optimization under interpolation.

Lay Summary

A major trend in machine learning is the use of flexible models which can exactly fit large quantities of data. For example, deep learning approaches can "memorize" datasets, meaning they achieve nearly perfect predictions on the samples used to fit the model. In this case, we say that the model *interpolates* the dataset. Interpolating models are particularly interesting from an optimization perspective because they can be fit very quickly using stochastic gradient methods. This contrasts with general models, where stochastic gradient methods are notoriously slow. In this thesis, we develop a rigorous definition of interpolation and study the speed of stochastic gradient methods for interpolating models. Our approach is more general than existing analyses and covers standard model-fitting using a dataset as a special case. For many model classes, we show stochastic gradient methods permit a wider range of parameters and are faster than previously known when interpolation is satisfied.

Preface

This thesis was conceived and written solely by the author, Aaron Mishkin. The basis for the theoretical work was developed in collaboration with Sharan Vaswani and Frederik Kunstner over the last two years. The work here is original and unpublished with the exception of the stochastic Armijo line-search, which was published in Vaswani et al. (2019b). A. Mishkin is a coauthor of this paper.

The breakdown of contributions for results included from Vaswani et al. (2019b) is as follows: the idea to investigate stochastic line-search techniques was proposed by S. Vaswani. S. Vaswani also conceived of the stochastic Armijo line-search and proved the original form of Lemma 8. A. Mishkin and S. Vaswani later revised the lemma statement to the version stated in Chapter 4. The proof technique for Theorem 11 (Chapter 4) was developed and suggested by S. Vaswani, while the specific result was proved by A. Mishkin.

Many of the unpublished results in this thesis have also benefited from collaboration. The extension of interpolation to general stochastic oracles in Chapter 2 was conducted by the author alone, while the connection between interpolation and weak/strong growth was developed in collaboration with S. Vaswani, and F. Kunstner. The improved convergence theorems in Chapters 3 and 5 build on previous work by Vaswani et al. (2019a) and were proved solely by the author.

Theorem 9 in Chapter 4, which improves the dependency on the strong-convexity parameter from $\bar{\mu}$ to μ, was suggested by S. Vaswani and F. Kunstner with reference to a result for the stochastic Polyak step-size by Loizou et al. (2020). A. Mishkin proved the theorem alone. The analysis of stochastic gradient descent for ℓ_2-regularized functions satisfying interpolation in Chapter 6 was inspired by a conversation with Eduard Gorbunov.

All figures are the original product of the author, A. Mishkin.

Table of Contents

List of Tables

List of Figures

Glossary

AGD accelerated gradient descent

PL Polyak-Łojasiewicz

R-SAGD reformulated stochastic accelerated gradient descent

SAGD stochastic accelerated gradient descent

SFO stochastic first-order oracle

SGD stochastic gradient descent

Acknowledgments

I am extremely grateful to my supervisor, Mark Schmidt, for fostering an intellectually curious and collaborative environment. His group at the University of British Columbia was an ϵ-optimal setting for this research. Sharan Vaswani convinced me to join the world of stochastic optimization in 2018 and none of the work herein would have come to be without his support. I am also greatly indebted to my friend and colleague Frederik Kunstner, with whom I have shared innumerable conversations on interpolation and stochastic gradient descent. Finally, Si Yi (Cathy) Meng, Betty Shea, and Jonathan Lavington provided many excellent comments and suggestions throughout the writing of this thesis.

Chapter 1

Introduction

Stochastic first-order methods are the most popular optimization algorithms in modern machine learning. In particular, stochastic gradient descent (SGD) (Robbins and Monro, 1951) and its adaptive variants (Duchi et al., 2011; Kingma and Ba, 2015; Tieleman and Hinton, 2012; Zeiler, 2012) are widely used in large-scale supervised learning, where they are frequently referred to as fundamental "workhorse" algorithms (Assran et al., 2019; Grosse and Salakhutdinov, 2015; Qian et al., 2019). The main advantage of these methods for machine learning is that they only use the gradient of a single or small sub-sample of training examples to update the model parameters at each iteration. The computational cost of SGD (and variants) is thus independent of the training set size and scales to very large datasets and models. This property is also why stochastic first-order methods are the dominant approach to training highly expressive models, such as deep neural networks (Bengio, 2012; Zhang et al., 2017) and non-parametric kernels (Belkin et al., 2019b; Liang and Rakhlin, 2018).

Despite their successes, stochastic first-order methods suffer from two well known problems. The step-size, momentum term, and other algorithmic hyper-parameters must be carefully tuned to obtain good performance (Bengio, 2012; Choi et al., 2019; Li and Orabona, 2019; Schaul et al., 2013); and they converge slowly compared to deterministic (Nesterov, 2004) or variance-reduced algorithms (Defazio et al., 2014; Johnson and Zhang, 2013; Le Roux et al., 2012) even when well-tuned. Hyper-parameter tuning for SGD is the focus of intense research, with approaches ranging from adaptive step-sizes inspired by online learning (Li and Orabona, 2019; Luo et al., 2019; Orabona and Tommasi, 2017) to

meta-learning (Almeida et al., 1998; Baydin et al., 2018; Plagianakos et al., 2001; Schraudolph, 1999; Shao and Yip, 2000; Sutton, 1992) and heuristics for online estimation (Rolinek and Martius, 2018; Schaul et al., 2013; Tan et al., 2016). In contrast, the slow convergence of first-order methods in the general stochastic setting cannot be improved, with tight lower-bounds showing $\Theta(\epsilon^{-4})$ iteration complexity for convergence to an ϵ-approximate stationary point (Arjevani et al., 2019; Drori and Shamir, 2019). This compares poorly to deterministic methods, which are $\Theta(\epsilon^{-2})$ for the same problem class (Carmon et al., 2019).

Increasing experimental and theoretical evidence shows that the slow optimization speed of stochastic first-order methods is mitigated when training over-parameterized models (Arora et al., 2018; Ma et al., 2018; Zou and Gu, 2019). For example, variance-reduced algorithms typically underperform SGD in this setting despite using increased memory or computation (Defazio and Bottou, 2019; Ma et al., 2018). A key property of over-parameterized problems is that they satisfy *interpolation*, meaning the model can exactly fit all of the available training data (Belkin et al., 2019b). While this may seem strong, interpolation has been observed in practice for popular methods such as boosting (Schapire et al., 1997), kernel learning (Belkin et al., 2019b), and over-parameterized neural networks (Belkin et al., 2019a; Zhang and Yin, 2013). Under additional assumptions, interpolation is a sufficient condition for the strong (Schmidt and Le Roux, 2013; Solodov, 1998; Tseng, 1998) or weak (Bassily et al., 2018; Vaswani et al., 2019a) growth conditions, which imply the second-moment of the stochastic gradients is bounded by a linear function of either the gradient-norm, or the optimality gap, respectively. A form of automatic reduction in gradient noise occurs (Liu and Belkin, 2020) when strong or weak growth is satisfied, explaining why variance reduction may be unnecessary.

The connection between over-parameterization and fast stochastic optimization has led to a wave of interest in analyzing first-order methods under interpolation and weak/strong growth. A number of works have shown that SGD obtains the fast convergence rate of deterministic gradient methods for interpolating models (Bassily et al., 2018; Cevher and Vu, 2019; Jain et al., 2018; Schmidt and Le Roux, 2013; Vaswani et al., 2019a), while related research has established accelerated rates under an additional requirement for convexity (Jain et al., 2018; Liu and Belkin, 2020; Vaswani et al., 2019a). Sub-sampled second-order methods have also been explored and proved to have local quadratic convergence for specific function classes (Meng et al., 2020). These positive results show the interpolation setting is restrictive enough to break the $\Omega(\epsilon^{-4})$ barrier for stochastic first-order methods and attain the optimal $\Theta(\epsilon^{-1})$ complexity (up to problem-specific constants) for finding stationary

points of smooth, convex functions (Arjevani and Shamir, 2016; Nemirovsky and Nesterov, 1985).

Despite these rapid advances, theoretical rates and practical performance for SGD under interpolation still rely on carefully selected hyper-parameters. A number of approaches from the general stochastic setting are rapidly being adopted to tackle this problem. For instance, several works have considered a stochastic version of the Polyak step-size, which does not require knowledge of the smoothness or convexity parameters (Berrada et al., 2019; Loizou et al., 2020). Stochastic line-searches using the classic Armijo condition (Armijo, 1966) have also been proposed and shown to obtain fast convergence under interpolation (Vaswani et al., 2019b). Very recently, adaptive variants of SGD have been analyzed both with and without the Armijo line-search (Vaswani et al., 2020).

This thesis analyzes a core group of first-order methods in stochastic optimization under interpolation. We consider constant step-size SGD, SGD with a stochastic Armijo line-search, and a version of Nesterov's accelerated gradient descent with stochastic gradients (Nesterov, 2004). In nearly all cases, we show that existing analyses can be tightened to yield faster convergence rates with a larger range of step-sizes. In the case of acceleration, the improvement is comparable to dividing by the square-root of the condition number and addresses criticisms that previous analyses yield inferior rates to those of SGD in some circumstances (Liu and Belkin, 2020).

The thesis is organized as follows: In Chapter 2, we formalize interpolation and the strong/weak growth conditions in the context of general stochastic oracles. Connections between interpolation, smoothness properties of the stochastic oracle, and growth of the stochastic gradients are derived. Chapter 3 analyzes the complexity of SGD with a fixed step-size, drawing comparisons with previous work as well as convergence rates in the deterministic setting. Asymptotic convergence of SGD with a constant step-size to (i) stationary points of general non-convex functions, and (ii) minimizers of convex functions is shown under the strong and weak growth conditions. Chapters 4 and 5 then address the convergence of SGD with a stochastic Armijo line-search and stochastic accelerated gradient descent, respectively. Finally, Chapter 6 leaves the basic interpolation setting behind and considers structural minimization with interpolating functions as components. Convergence to an explicit neighborhood is derived for L_2-regularized problems as a special case.

1.1 Preliminaries and Assumptions

This work considers the problem of minimizing a continuous function $f : \mathbb{R}^d \to \mathbb{R}$ under interpolation conditions. We assume that f is bounded below with at least one finite minimizer. That is, there exists at least one $w^* \in \mathbb{R}^d$ such that

$$f(w) \geq f(w^*) \qquad \forall w \in \mathbb{R}^d.$$

For functions with multiple minimizers, we write $\mathcal{X}^* = \arg\min_{w \in \mathbb{R}^d} f(w)$ and denote the projection of a point onto the optimal set as $\Pi_{\mathcal{X}^*}(w)$. Additionally, f is required to be differentiable and L-smooth, meaning the mapping $w \mapsto \nabla f(w)$ is an L-Lipschitz function,

$$\|\nabla f(w) - \nabla f(v)\| \leq L\|w - v\| \qquad \forall\, w, v \in \mathbb{R}^d.$$

This is equivalent to the existence of the following quadratic bound on f:

$$\frac{L}{2}\|v - w\|^2 \geq |f(v) - f(w) - \langle \nabla f(w), v - w \rangle| \qquad \forall\, w, v \in \mathbb{R}^d. \qquad (L\text{-Smoothness})$$

Often, only the upper-bound given by L-smoothness is used,

$$f(v) \leq f(w) + \langle \nabla f(w), v - w \rangle + \frac{L}{2}\|v - w\|^2 \quad \forall\, w, v \in \mathbb{R}^d,$$

which is sometimes called one-sided Lipschitz-smoothness. At times, we will further assume that f is convex or μ-strongly-convex,

$$f(v) \geq f(w) + \langle \nabla f(w), v - w \rangle \qquad \forall\, w, v \in \mathbb{R}^d, \qquad (\text{Convexity})$$
$$f(v) \geq f(w) + \langle \nabla f(w), v - w \rangle + \frac{\mu}{2}\|v - w\|^2 \quad \forall\, w, v \in \mathbb{R}^d. \quad (\mu\text{-Strong-Convexity})$$

Convexity holds for many problems in machine learning, including linear and logistic regression.

Convexity can be relaxed to a more limited property called invexity. Formally, we say a differentiable function f is invex if all stationary points are also global minimizers of f (Ben-Israel and Mond, 1986), meaning

$$\nabla f(w) = 0 \implies f(w) \leq f(v) \qquad \forall\, v \in \mathbb{R}^d.$$

Invexity is formally weaker than convexity and follows immediately from the first-order conditions for convexity given above. The analogue of strong-convexity for invex functions is the Polyak-Łojasiewicz (PL) condition (Karimi et al., 2016). A function is said to be μ-PL if

$$\frac{1}{2\mu}\|\nabla f(w)\|^2 \geq f(w) - f(w^*),$$

holds for all $w \in \mathbb{R}^d$. Again, the PL condition is weaker than strong convexity; a μ-strongly-convex function is μ-PL, but the converse does not hold.

In several cases, it will be useful to interpret our results in the context of finite-sum functions. The function f is said to be a finite-sum if it can be written as

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w),$$

where the individual (or sub-) functions $f_i : \mathbb{R}^d \to \mathbb{R}$ may be strongly-convex, convex, or non-convex depending on the problem. Such objective functions arise naturally in empirical risk minimization, where f_i typically corresponds to a single training example (x_i, y_i) in a training set. For example, the classic least-squares regression objective can be written as

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} \left(\langle w, x_i\rangle - y_i\right)^2,$$

which is finite-sum with sub-functions $f_i(w) = (\langle w, x_i\rangle - y_i)^2$.

1.2 Related Work

Interpolation: Existing work focuses on interpolation in the context of finite-sum objectives. In this setting, Bassily et al. (2018) define interpolation in terms of sequences converging to global minima of f. They say interpolation holds if, for every sequence (w_k) satisfying $\lim_{k\to\infty} f(w_k) = f(w^*)$, we also have

$$\forall i \in [n], \quad \lim_{k\to\infty} f_i(w_k) = f(w^*).$$

Berrada et al. (2019) propose ϵ-interpolation, which requires the sub-functions to be lower-

bounded and ϵ sub-optimal for all $w^* \in \arg\min f$:

$$\forall i \in [n], \quad \inf_w f_i(w) \geq B \quad \text{and} \quad f_i(w^*) - B \leq \epsilon.$$

A larger body of work considers interpolation to hold when the optimal points w^* are also stationary points or global minimizers of each sub-function f_i (Loizou et al., 2020; Meng et al., 2020; Vaswani et al., 2019a,b, 2020),

$$w^* \in \arg\min_w f(w) \implies w^* \in \arg\min_w f_i(w) \quad \forall i \in [n],$$

or

$$w^* \in \arg\min_w f(w) \implies \nabla f_i(w^*) = 0 \quad \forall i \in [n].$$

We will focus on interpolation conditions of this form, which we extend to general stochastic optimization problems.

Growth Conditions: The strong growth condition was first proposed in the context of incremental gradient methods by Solodov (1998) and Tseng (1998) as a bound on the squared-norm of the stochastic gradients. Suppose f is finite-sum and P_i is an arbitrary distribution used to sub-sample the finite sum. Then strong growth with parameter $\rho > 0$ may be written as

$$\text{Maximal Strong Growth:} \quad \|\nabla f_i(w)\|^2 \leq \rho \|\nabla f(w)\|^2,$$

where the inequality holds almost-surely with respect to P_i. This definition was later used by Schmidt and Le Roux (2013) to derive linear convergence rates for SGD with a constant step-size. Vaswani et al. (2019a) propose a modified version of strong growth which holds in expectation,

$$\text{Strong Growth:} \quad \mathbb{E}_{P_i}\left[\|\nabla f_i(w)\|^2\right] \leq \rho \|\nabla f(w)\|^2.$$

We call this latter condition strong growth and refer to the earlier definition as maximal strong growth. Vaswani et al. (2019a) also propose the following weak growth condition as

a natural relaxation of strong growth:

$$\text{Weak Growth}: \quad \mathbb{E}_{P_i}\left[\|\nabla f_i(w)\|^2\right] \leq 2\rho L\left(f(w) - f(w^*)\right).$$

Cevher and Vu (2019) refer to strong growth simply as "the growth condition" and suggest strong growth with an additive noise parameter as an alternative relaxation,

$$\text{Strong Growth + Noise}: \quad \mathbb{E}_{P_i}\left[\|\nabla f_i(w)\|^2\right] \leq \rho\|\nabla f(w)\|^2 + \sigma^2,$$

which they also call weak growth. For clarity, we refer to this assumption as strong growth with additive noise. For unbiased P_i, strong growth with noise is slightly weaker than assuming a bound on the variance of the stochastic gradients (Ghadimi and Lan, 2012; Khaled and Richtárik, 2020). Cevher and Vu (2019) study the convergence of proximal-gradient methods under strong growth with additive noise and also prove that strong growth is both *sufficient and necessary* for SGD to converge linearly.

Stochastic Acceleration: Many authors have considered accelerating stochastic gradient methods. Schmidt et al. (2011) provide sufficient conditions on gradient noise for acceleration of proximal-gradient methods. D'Aspremont (2008) and Devolder et al. (2014) investigate accelerated gradient methods under the assumption of approximate oracles with deterministic, bounded errors and derive rates for convergence and error accumulation in this setting. Honorio (2012) consider accelerated proximal-gradient methods under biased stochastic oracles and show that accumulated errors prevent a high-probability convergence guarantee. In contrast, Cohen et al. (2018) assume access to an unbiased oracle with additive gradient noise and propose a noise-resistant acceleration scheme. Very recently, Chen and Kolar (2020) analyze accelerated methods for strongly-convex functions under strong growth with additive noise.

An alternative approach to stochastic acceleration splits the convergence rate into stochastic and deterministic parts. Using this framework, multiple authors have shown that acceleration schemes speed-up convergence for the deterministic component and achieve nearly optimal rates in the stochastic approximation setting (Ghadimi and Lan, 2012, 2013; Hu et al., 2009). For finite-sum functions, variance reduction techniques can be combined with acceleration to improve convergence on the stochastic component (Allen-Zhu, 2017, 2018; Defazio, 2016; Kovalev et al., 2020; Shang et al., 2018; Tang et al., 2018). Such

methods have iteration complexity of $O((n + \sqrt{n\kappa})\log(1/\epsilon))$, where n is the number of sub-functions and κ is the condition number. This is as fast as deterministic acceleration up to the additional factor of $\sqrt{n}$. Accelerated primal-dual methods have been proposed in the same setting under an additional requirement for each f_i to be Lipschitz-smooth (Lan and Zhou, 2018; Zhang and Lin, 2015). Alternative approaches also leveraging finite-sum structure are based on the proximal-point algorithm and include accelerated SDCA (Shalev-Shwartz and Zhang, 2014), Catalyst (Lin et al., 2015), and accelerated APPA (Frostig et al., 2015).

Several works have recently considered acceleration under interpolation. The most similar to the investigation here is that of Vaswani et al. (2019a), who analyze a slightly altered version of Nesterov's accelerated gradient descent under the strong growth condition. Liu and Belkin (2020) propose a different modification, called MaSS, and analyze its convergence for convex quadratics as well as strongly-convex functions with additional assumptions. These assumptions imply strong growth, but yield hard-to-compare rates. Jain et al. (2018) prove accelerated rates for squared-losses under interpolation using a tail-averaging scheme. Finally, Assran and Rabbat (2020) study the stochastic approximation setting and prove that accelerated gradient descent may fail to accelerate even when interpolation is satisfied.

Line Search: Line-search is a classic technique to set the step-size in deterministic optimization (see Nocedal and Wright (1999)), but extensions to stochastic optimization are non-obvious. Mahsereci and Hennig (2017) use a Gaussian process model to define probabilistic Wolfe conditions (Wolfe, 1969, 1971); however, the convergence of SGD with this procedure is not known. Fridovich-Keil and Recht (2019) propose line-search conditions based on golden-section search (Avriel and Wilde, 1968), but again only provide empirical evidence for SGD. Paquette and Scheinberg (2020), Krejic and Krklec (2013), and Ogaltsov et al. (2019) prove convergence of SGD with the Armijo condition in the general stochastic setting with several caveats. Paquette and Scheinberg (2020) require adaptive batch-sizes, while Krejic and Krklec (2013) assume the stochastic gradients are bounded and Ogaltsov et al. (2019) use explicit knowledge of an upper-bound on the variance. Other authors consider similar approaches based on multiple function and/or gradient evaluations at each iteration (Byrd et al., 2012; De et al., 2016; Friedlander and Schmidt, 2012). Such approaches also use growing batch-sizes to ensure convergence. The work in the thesis follows Vaswani et al. (2019b), who investigated stochastic versions of the Armijo line-search under

interpolation; we provide tighter analyses in a more general setting.

Asymptotic Convergence: The original paper by Robbins and Monro (1951) establishes asymptotic, almost-sure convergence for SGD to the zero of a convex function if the stochastic gradients are bounded and a decreasing step-size is used. Highly general analyses have since shown almost-sure convergence for non-convex functions when strong growth with additive noise is satisfied (Bertsekas and Tsitsiklis, 2000; Bottou, 1991). Alternative work directly derives these conditions from properties of strongly-convex or non-convex functions, also for the purpose of proving almost-sure convergence (Lei et al., 2019; Nguyen et al., 2018). Recently, asymptotic convergence was shown with Adagrad-style step-sizes instead of a fixed, decreasing schedule (Li and Orabona, 2019).

Chapter 2

Interpolation and Growth Conditions

Interpolation is a technique from numerical analysis where a function is approximated by exactly fitting a sample of its evaluations using a tractable class of approximators. This class is often piece-wise linear, piece-wise cubic, or polynomial, but can also be non-parametric. Even over-parameterized neural networks can be used, assuming it is possible to solve the resulting non-convex optimization problem. This is because the defining aspect of interpolation methods is not the choice of approximating function, but that the approximation attains the *true* function value at all points in the sample.

This basic notion of interpolation can be directly carried over to supervised machine learning. The regression case is nearly identical — intuitively, a model interpolates a set of training examples if it predicts exactly the true target for each example. The analogy to numerical analysis is complete if we assume that the data are generated by an underlying deterministic mapping. For classification problems, interpolation can be defined in two obvious ways. We might say a model satisfies interpolation if it predicts the correct class for each training example, or alternatively, if it attains zero loss for all training examples. The latter interpretation is dependent on the choice of objective function and starts to hint at the role of interpolation in optimization. Note that in the regression example, the two notions of interpolation coincide for the natural choice of squared loss on all examples.

These intuitive definitions for interpolation mean approximately the same thing: "to fit the data exactly". While this agrees with the spirit of the methodology and broader

scientific usage of the term, a rigorous definition is necessary to analyze the complexity of first-order optimizers under interpolation. This chapter formalizes interpolation in the context of stochastic optimization. In particular, we do the following:

1. Define the concept of a stochastic first-order oracle (SFO) $\mathcal{O}$ and propose a new Lipschitz-smoothness property for SFOs, which we call individually-smoothness.

2. Give three specific definitions of interpolation in the context of a stochastic first-order oracle and objective function pair $(f, \mathcal{O})$. These are the first definitions applicable to the general stochastic setting.

3. Characterize the relationships between the three models of interpolation as well as connections to strong and weak growth. In particular, we show that interpolation implies weak growth if $\mathcal{O}$ is individually-smooth; this relaxes existing sufficient conditions, which further assume f is convex and finite-sum (Vaswani et al., 2019a).

We begin with the discussion of stochastic oracles.

2.1 Stochastic Oracles

The defining feature of stochastic optimization algorithms is that they cannot directly access the value or gradient of the objective function f. Instead, they obtain noisy function and gradient evaluations through a SFO $\mathcal{O}$. At each iteration k, $\mathcal{O}$ outputs a stochastic function $f(w, z_k)$ and gradient $\nabla f(w, z_k)$ at query point w, where z_k is a random variable with distribution μ_k supported on $\mathcal{Z}_k \subseteq \mathbb{R}^m$, $m > 0$.[1] We assume that $f(w, \cdot)$ is a deterministic Borel function, meaning the stochasticity in $f(w, z_k)$ stems only from the random variable z_k. Similarly, $\nabla f(w, \cdot)$ is taken to be a deterministic Borel function related to $f(w, \cdot)$ through the standard differentiation operator,

$$\nabla f(w, \cdot) = \left[\frac{\partial}{\partial_{w_1}} f(w, \cdot), \ldots, \frac{\partial}{\partial_{w_d}} f(w, \cdot) \right]^\top.$$

Unless stated otherwise, the oracle outputs are always taken to be unbiased, implying that

$$\mathbb{E}_{z_k}\left[f(w, z_k) \right] = f(w) \quad \text{and} \quad \mathbb{E}_{z_k}\left[\nabla f(w, z_k) \right] = \nabla f(w),$$

[1] It will be sufficient to treat the measure μ_k as a prob. density function p_k for nearly all of our purposes.

for all k.

The definition of $\mathcal{O}$ is non-stationary in the sense that it allows the distributions μ_k and their support $\mathcal{Z}_k$ to change across iterations. As such, it will be necessary to refer to the support of the entire stochastic process (z_k),

$$\mathcal{Z} = \bigcup_{k \in \mathbb{N}} \mathcal{Z}_k.$$

For simplicity, $\mathcal{Z}$ is called the support of $\mathcal{O}$. Note that a statement which holds point-wise over $\mathcal{Z}$ holds almost-surely for all z_k. These two requirements are equivalent, since (by definition) every $z \in \mathcal{Z}$ is in the support of some μ_k. We will make use of this less cumbersome notation.

Some results will further require that the stochastic function $f(\cdot, z_k)$ is Lipschitz-smooth for all outcomes in $\mathcal{Z}$. In this case, we say $\mathcal{O}$ satisfies *individual smoothness* with parameter $L_{\max}$.

Definition 1 (Individual Smoothness)**.** *An SFO $\mathcal{O}$ is called L_{max} individually-smooth if the stochastic gradient mapping $w \mapsto \nabla f(w, z)$ is L_z-Lipschitz with $L_z \leq L_{max}$ for all $z \in \mathcal{Z}$.*

Individual smoothness implies the quadratic upper bound

$$f(v, z) \leq f(w, z) + \langle \nabla f(w, z), v - w \rangle + \frac{L_{\max}}{2} \|v - w\|^2 \quad \forall v, w \in \mathbb{R}^d,$$

holds point-wise over $\mathcal{Z}$ and thus almost-surely for each z_k. Note that the existence of an $L_{\max}$ individually-smooth, unbiased SFO for f implies that f is L-smooth with $L \leq L_{\max}$ as an immediately corollary.

Corollary 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex, differentiable function. If there exists an unbiased, L_{max} individually-smooth SFO $\mathcal{O}$ for f, then f is L-smooth with $L \leq L_{max}$. Alternatively, if $\mathcal{O}$ is individually-convex and biased with finite support $\mathcal{Z}$ and finite partial derivatives $\frac{\partial}{\partial w_j} f(w, z)$ for each $z \in \mathcal{Z}$, then $\mathbb{E}_{z_k}[f(\cdot, z_k)]$ is L_{max}-smooth for all k.*

See Section A.1 for proof. In several cases, we will also make use of individual strong-convexity.

Definition 2 (Individual Strong-Convexity)**.** *A SFO $\mathcal{O}$ is called μ_{max} individually-strongly-convex if the function $f(\cdot, z)$ is μ_z-strongly-convex with $\mu_z \leq \mu_{max}$ for all $z \in \mathcal{Z}$. If $\mu_{max} = 0$, then we say $\mathcal{O}$ is individually-convex.*

Individually-smooth oracles are wide-spread throughout machine learning in the form of finite-sum functions. Recall that finite-sum functions are a particular class of structured optimization problem where the objective f is the sum of n sub-functions. The following example demonstrates individual smoothness in the context of least-squares regression.

Example 1 (Least Squares Regression: Individual Smoothness)**.** *The classic least-squares regression problem*

$$w^* = \arg\min \frac{1}{n} \sum_{i=1}^{n} \left(\langle w, x_i \rangle - y_i \right)^2,$$

has a finite-sum structure with $f_i(w) = \left(\langle w, x_i \rangle - y_i \right)^2$*. The mini-batch SFO uniformly sub-samples* $b \leq n$ *examples to obtain the following function and gradient estimators:*

$$f(w, z) = \frac{1}{b} \sum_{i \in z} \left(\langle w, x_i \rangle - y_i \right)^2 \quad \textit{and} \quad \nabla f(w, z) = \frac{1}{b} \sum_{i \in z} 2 \left(\langle w, x_i \rangle - y_i \right) x_i,$$

where $z \in \mathcal{Z} = \{A \subseteq \{1, \ldots, n\} : |A| = b\}$ *is the set of sampled indices. Notice that* $f(\cdot, z)$ *is convex and* L_z*-smooth with* $L_z = \frac{2}{b} \| \sum_{i \in z} x_z x_z^\top \|_{op}^2$*, where* $\| \cdot \|_{op}$ *is the operator norm. This implies* $\mathcal{O}$ *is individually-convex and* L_{max} *individually-smooth with* $L_{max} \leq 2 \max_{i \in [n]} \|x_i\|^2$*.*

Example 1 illustrates the *mini-batch* oracle commonly used in machine learning. The mini-batch oracle is unbiased, simple to compute, and by far the most common SFO in machine learning. If the sub-functions f_i are each individually L_i-smooth, then the mini-batch SFO satisfies individual-smoothness with $L_{\max} = \max_{i \in [n]} L_i$. We call general SFOs individually-smooth specifically to invite comparison with this natural property of the mini-batch oracle.

2.2 Interpolation

Now we formalize interpolation in the context of stochastic optimization problems. Unlike previous work, we consider general SFOs and do not require f to be finite-sum or $\mathcal{O}$ to be the mini-batch oracle (cf. Vaswani et al. (2019a) or Bassily et al. (2018)). Instead, three different notions of interpolation are developed as a joint property of the objective and oracle. As with individual smoothness, interpolation is shown to have a simple realization for finite-sum functions that satisfies the intuition previously developed for machine learning problems.

The basic requirement of interpolation is that the oracle outputs $f(\cdot, z)$ and $\nabla f(\cdot, z)$ resemble the true function at key target points. The axes of variation in the definitions to follow are (a) the specific target points and (b) the type of "matching" required. Unlike classic interpolation, matching here refers to mutual optimality or mutual stationarity, rather than a requirement for equal function values; enforcing $f(w^*) = f(w^*, z)$ may not be useful from an optimization perspective, since $\nabla f(w^*, z)$ could be non-zero. The three definitions are as follows:

Definition 3 (Interpolation: Minimizers)**.** *A function-oracle pair $(f, \mathcal{O})$ satisfies minimizer interpolation if for all $z \in \mathcal{Z}$,*

$$f(w') \leq f(w)\ \forall w \in \mathbb{R}^d \implies f(w', z) \leq f(w, z)\ \forall w \in \mathbb{R}^d.$$

Definition 4 (Interpolation: Stationary Points)**.** *A function-oracle pair $(f, \mathcal{O})$ satisfies stationary-point interpolation if for all $z \in \mathcal{Z}$,*

$$\nabla f(w') = 0 \implies \nabla f(w', z) = 0.$$

Definition 5 (Interpolation: Mixed)**.** *A function-oracle pair $(f, \mathcal{O})$ satisfies mixed interpolation if for all $z \in \mathcal{Z}$,*

$$f(w') \leq f(w)\ \forall w \in \mathbb{R}^d \implies \nabla f(w', z) = 0.$$

In words, Definition 3 states that global minimizers of the objective f must be global minimizers of the stochastic functions given by the SFO at every iteration k. In contrast, Definition 4 puts the same requirement on stationary points of f, while Definition 5 merely demands that minimizers of f are stationary points of $f(\cdot, z)$ for all outcomes z.

For general functions and SFOs, the relationship between these models of interpolation is limited to the following: minimizer interpolation and stationary-point interpolation are stronger than mixed interpolation. However, all three definitions are equivalent when f and $f(\cdot, z)$ are invex for all $z \in \mathcal{Z}$. This is stated in the following lemma.

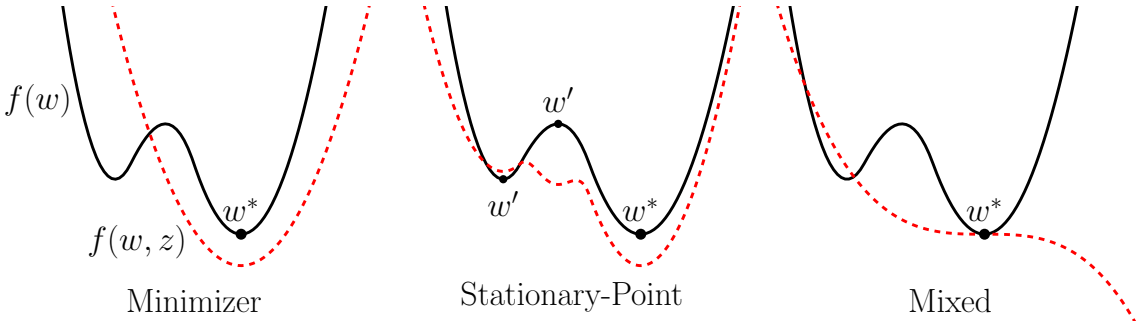

Figure 2.1: Oracle samples satisfying the three different models of interpolation for the same non-convex function. The dashed red-line is the graph of $f(w,z)$ for a single, fixed $z \in \mathcal{Z}$. Stationary-points are denoted as w', while the global minimizer is marked as w^*. Note that $f(w,z)$ is permitted to have many "extra" stationary points when $\mathcal{O}$ satisfies stationary-point interpolation. Similarly, mixed interpolation does not prevent $f(w,z)$ from being unbounded below.

Lemma 1. *Let $(f,\mathcal{O})$ be an arbitrary function-SFO pair. Then only the following relationships hold between models of interpolation:*

$$\textit{Minimizer Interpolation } (\textit{Definition } 3) \implies \textit{Mixed Interpolation } (\textit{Definition } 5)$$
$$\textit{and}$$
$$\textit{Stationary-Point Interpolation } (\textit{Definition } 4) \implies \textit{Mixed Interpolation } (\textit{Definition } 5).$$

However, if f is invex and $\mathcal{O}$ is such that $f(\cdot,z)$ is invex for all $z \in \mathcal{Z}$, then minimizer, stationary-point, and mixed interpolation are equivalent.

This result follows immediately from first-order optimality conditions and the equivalence of stationary points and global minimizers for invex functions. For completeness, a short proof with counter-examples for the implications which do not hold is given in Appendix A.

Let us return to the finite-sum setting and mini-batch oracle to better understand the three definitions of interpolation. In particular, let f and $\mathcal{O}$ be such that

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w), \quad f(w,z_k) = f_{z_k}(w), \quad \nabla f(w,z_k) = \nabla f_{z_k}(w),$$

where μ_k is a uniform distribution over the support set $\mathcal{Z} = \{1,\dots,n\}$. The function-oracle pair $(f,\mathcal{O})$ satisfy minimizer interpolation if the individual functions f_i are globally min-

imized at every global minimum of f. The differences between minimizer and stationary-point interpolation are readily apparent for finite-sums of non-convex functions, as shown in Figure 2.1. The following example further narrows this to least-squares regression, where interpolation implies the entire training set can be fit exactly.

Example 2 (Least Squares Regression: Interpolation)**.** *Consider the least-squares problem*

$$w^* = \arg\min \frac{1}{n} \sum_{i=1}^{n} \left(\langle w, x_i \rangle - y_i \right)^2,$$

with individually-smooth and convex mini-batch oracle

$$f(w, z) = \frac{1}{|z|} \sum_{i \in z} \left(\langle w, x_i \rangle - y_i \right)^2 \quad \textit{and} \quad \nabla f(w, z) = \frac{1}{|z|} \sum_{i \in z} 2 \left(\langle w, x_i \rangle - y_i \right) x_i.$$

The stochastic functions $f(\cdot, z)$ are invex, as is f, meaning minimizer, stationary-point, and mixed interpolation are equivalent. If $(f, \mathcal{O})$ satisfies minimizer interpolation, then first-order optimality implies

$$2 \left(\langle w^*, x_i \rangle - y_i \right) x_i = 0 \qquad \forall i \in [n].$$

Further requiring $x_i \neq 0$ for all $i \in [n]$ guarantees $\langle w^*, x_i \rangle = y_i$ and our abstract definitions of interpolation recover the original meaning from numerical analysis.

There are several natural ways to establish interpolation. Least-squares regression satisfies interpolation when $y \in (\text{span}\left(\{x_i\}_{i=1}^n\right))$. This occurs, for example, when the data matrix is full-rank and $n \leq d$ (Hastie et al., 2009). Similarly, interpolation is satisfied for linear classifiers on separable datasets using the squared-hinge loss (Vaswani et al., 2019a). In the general setting of SFOs, the following lemma establishes simple conditions for $(f, \mathcal{O})$ to satisfy minimizer interpolation.

Lemma 2. *Let $(f, \mathcal{O})$ be a function-oracle pair. If $\mathcal{O}$ is unbiased and*

$$f(w, z) \geq f(w^*) \quad \forall w \in \mathbb{R}^d, \forall z \in \mathcal{Z},$$

holds, then $(f, \mathcal{O})$ satisfies minimizer interpolation.

See Section A.2 for proof.

Lemma 2 gives a convenient mechanism for checking if interpolation holds. It is most useful for finite-sum functions, where the sufficient conditions enforce $f_i(w^*) = f(w^*)$ for each i. It is also highly related to the ϵ-interpolation concept proposed by Berrada et al. (2019), which requires $|f(w^*, z) - f(w^*)| \leq \epsilon$ for all $z \in \mathcal{Z}$, where $\epsilon > 0$. This alternative model of interpolation is not sufficient for exact convergence (Berrada et al., 2019) and we do not explore ϵ-interpolation any further in this work.

2.3 Growth conditions

Minimizer, stationary-point, and mixed interpolation all constrain the stochastic oracle to resemble the true objective at a set of target points. Now we show that for individually-smooth oracles, interpolation further implies that the stochastic gradients must be well-behaved globally. This is made concrete by the notion of *growth conditions*, which constrain the stochasticity of $\nabla f(w, z_k)$ in terms of $\|\nabla f(w)\|$. In particular, we prove that the weak (Vaswani et al., 2019a) and strong (Schmidt et al., 2011) growth conditions hold with specific and simple constants when $\mathcal{O}$ is individually-smooth and satisfies minimizer interpolation. But, first we give some brief background on regularity conditions for first-order methods.

Regularity conditions on the stochastic gradients have a long history in stochastic optimization. The first analysis of SGD by Robbins and Monro (1951) required a uniform bound on the norm of the stochastic gradients

$$\|\nabla f(w, z_k)\| \leq C,$$

in order to prove convergence. This "bounded gradients" assumption is rarely satisfied for objective functions in machine learning. For example, taking $\|w\| \to \infty$ in Example 1 shows it does not hold even for simple least-squares problems. A more realistic alternative is bounded variance, which requires

$$\mathbb{E}_{z_k}\left[\|\nabla f(w, z_k) - \nabla f(w)\|^2\right] \leq \sigma^2,$$

at each iteration k. Bounded variance, unbiasedness of $\mathcal{O}$, and independence of the z_k variables (ie. $z_k \perp\!\!\!\perp z_j$ for $k \neq j$) collectively define the stochastic approximation setting, which has been widely studied; see Kushner and Yin (1997) for more details. Yet, it is also

simple to show that bounded variance fails for least-squares with a mini-batch oracle.[2]

A far more realistic model is obtained by relaxing bounded variance to

$$\mathbb{E}_{z_k}\left[\|\nabla f(w, z_k)\|^2\right] \leq \rho\|\nabla f(w)\|^2 + \sigma^2,$$

where $\rho, \sigma \geq 0$. We call this condition "strong growth with additive noise" for reasons which will soon be clear. Note that it is strictly weaker than bounded variance, which is recovered when $\rho = 1$. Strong growth with additive noise was used as early as the classical analysis of stochastic optimization algorithms by Polyak and Tsypkin (1973), and continues to appear in current work (Bertsekas and Tsitsiklis, 2000; Khaled and Richtárik, 2020; Nguyen et al., 2018).

The first growth condition we discuss was proposed by Tseng (1998) and Solodov (1998), who used a version of strong growth with noise where $\sigma^2 = 0$ is fixed. In comparison, their condition is also strengthened to require the norm of the stochastic gradient to be almost surely bounded by that of the true gradient. Following Khaled and Richtárik (2020), we call this condition maximal strong growth.

Definition 6 (Maximal Strong Growth)**.** *A function-oracle pair $(f, \mathcal{O})$ satisfies maximal strong growth with parameter ρ if*

$$\|\nabla f(w, z)\|^2 \leq \rho\|\nabla f(x)\|^2,$$

holds for all $w \in \mathbb{R}^d$ and $z \in \mathcal{Z}$.

The "maximal" moniker is justified by the fact that Tseng and Solodov's condition is equivalent to

$$\max_E \int_E \|\nabla f(w, z)\|^2 d\mu_k(z) \leq \rho\|\nabla f(w)\|,$$

where E is any event with non-zero probability under the measure μ_k. It is important to note that maximal strong growth is much more restrictive than strong growth with additive noise. The former condition immediately implies $\mathcal{O}$ satisfies stationary-point interpolation, while $\sigma^2 \gg 0$ in the latter allows interpolation to be violated to an arbitrary degree.

Maximal strong growth was originally suggested in the context of incremental gradient methods and is impractical in the general setting due to the almost-sure requirement. The

[2]Take $\|w\| \to \infty$ in a problem instance where $\|x_i\| \neq \|x_j\|$ for at least two examples x_i, x_j to see that there can be no bound on the variance.

following, far more practical variation, was proposed by Vaswani et al. (2019a) and is simply called strong growth.

Definition 7 (Strong Growth)**.** *A function-oracle pair $(f, \mathcal{O})$ satisfies strong growth with parameter ρ if*

$$\mathbb{E}_{z_k}\left[\|\nabla f(w, z_k)\|^2\right] \leq \rho \|\nabla f(w)\|^2,$$

holds for all $k \geq 0$ and $w \in \mathbb{R}^d$.

It is straightforward to show that strong growth is indeed a weaker condition than maximal strong growth, which we do in the following lemma.

Lemma 3 (Formulations of Strong Growth)**.** *Let $(f, \mathcal{O})$ be a function-oracle pair. If $(f, \mathcal{O})$ satisfies maximal strong growth, then it also satisfies strong growth. However, strong growth does not imply maximal strong growth for general $\mathcal{O}$.*

The proof of lemma is given in Section A.3 and illustrates the impracticality of maximal strong growth, which is not even satisfied for multiplicative Gaussian noise. However, a sufficient condition for the equivalence of maximal strong and strong growth is finite $\mathcal{Z}$.

Lemma 4. *Let $(f, \mathcal{O})$ be a function-SFO pair satisfying the strong growth condition with constant ρ. Moreover, assume that the support $\mathcal{Z}$ of $\mathcal{O}$ is finite and each z_k admits probability mass function p_k. Then $(f, \mathcal{O})$ also satisfies maximal strong growth.*

See Section A.3 for proof.

An important consequence of Lemma 4 is the equivalence of the maximal strong growth and strong growth conditions for finite-sum optimization. This nicely reflects the original use of maximal strong growth for incremental gradient methods. Finally, we complete our statement of growth conditions with weak growth, which was recently proposed by Vaswani et al. (2019a) as a further relaxation of strong growth.

Definition 8 (Weak Growth Condition)**.** *Let f be an L-smooth function an $\mathcal{O}$ an SFO. The pair $(f, \mathcal{O})$ satisfies weak growth with parameter α if*

$$\mathbb{E}_{z_k}\left[\|\nabla f(w, z_k)\|^2\right] \leq \alpha L(f(w) - f(w^*)),$$

holds for all $k \geq 0$ and $w \in \mathbb{R}^d$.

If f is μ-PL, then the weak growth condition implies strong growth holds with parameter $\rho \leq \frac{\alpha L}{\mu}$ (Vaswani et al., 2019a, Proposition 1). In fact, the converse relation holds if f is Lipschitz-smooth, as we now show.

Lemma 5. *Let $(f, \mathcal{O})$ be a function-oracle pair satisfying the strong growth condition with parameter ρ. If f is L-smooth, then $(f, \mathcal{O})$ satisfies weak growth with parameter $\alpha \leq 2\rho L$.*

See Section A.3 for proof.

We now give the exact relationships between the strong/weak growth conditions and interpolation. As mentioned above, maximal strong growth immediately implies stationary-point interpolation. Strong growth with $w = w^*$ gives

$$\mathbb{E}_k \left[\|\nabla f(w^*, z_k)\|^2 \right] = 0,$$

for all k, after which non-negativity of the Euclidean norm guarantees $\nabla f(w^*, z_k) = 0$ almost-surely and stationary-point interpolation holds. Replicating this last argument under weak growth shows that the weak growth condition implies mixed interpolation. The reverse implications are more involved; we establish them formally in the following lemmas.

Lemma 6 (Interpolation and Weak Growth)**.** *Let f be an L-smooth function and $\mathcal{O}$ an L_{max} individually-smooth SFO. If $(f, \mathcal{O})$ satisfies minimizer interpolation, then the pair also satisfies the weak growth condition with parameter $\alpha \leq \frac{L_{max}}{L}$.*

See Section A.3 for proof.

Lemma 6 is similar to the sufficient conditions for weak growth derived by Vaswani et al. (2019a), who additionally require f to be convex and finite-sum. Thus, our result applies to a much larger class of functions and SFOs than previous work. We can derive similarly relaxed sufficient conditions for strong growth using the relationship between the strong and weak growth parameters. The following lemma does exactly this.

Lemma 7 (Interpolation and Strong Growth)**.** *Let f be a L-smooth μ-PL function and $\mathcal{O}$ an L_{max} individually-smooth SFO. If $(f, \mathcal{O})$ satisfies minimizer interpolation, then the pair also satisfies the strong growth condition with parameter $\rho \leq \frac{L_{max}}{\mu}$.*

Assumptions	Weak Growth	Strong Growth
Ind. Smooth	$\alpha \leq \frac{L_{\max}}{L}$	—
μ-PL + Ind. Smooth	$\alpha \leq \frac{L_{\max}}{L}$	$\rho \leq \frac{L_{\max}}{\mu}$
Strong Growth	$\alpha \leq 2\rho L$	$\rho = \rho$

Table 2.1: Relationship between minimizer interpolation and parameters of the weak and strong growth conditions. Weak growth is guaranteed to hold when $(f, \mathcal{O})$ satisfies minimizer interpolation and $\mathcal{O}$ is $L_{\max}$ individually-smooth. Strong growth holds if f additionally satisfies the PL condition, which is implied by strong-convexity.

See Section A.3 for proof.

Lemmas 6 and 7 show that minimizer interpolation implies global regularity of the stochastic gradients when $\mathcal{O}$ is individually-smooth. As a by-product, we obtain worst-case bounds on the strong and weak growth parameters and demonstrate the "automatic variance reduction" property (Liu and Belkin, 2020), which guarantees $\mathbb{E}_k\|\nabla f(w_k, z_k)\|^2 \to 0$ if $w_k \to w^*$ in the interpolation setting. Table 2.1 summarizes the relationships between minimizer interpolation and the weak/strong growth conditions.

2.4 Conclusions

We have now completed our goal of formalizing interpolation for general stochastic optimization problems. To do this, we leveraged the concept of an SFO, which provides a mechanism for optimization algorithms to query (unbiased) stochastic function and gradient values at each iteration. Moreover, we also showed that a class of SFOs satisfying a natural Lipschitz-smoothness property, called individual smoothness, satisfy the weak/strong growth conditions when minimizer interpolation holds (Lemmas 6 and 7). These oracles are well-behaved globally despite the inherently local nature of minimizer interpolation. As we shall see in the coming chapters, such global regularity is sufficient for the convergence of constant step-size SGD, as well as effective line-search techniques and acceleration.

Chapter 3

Stochastic Gradient Descent

The discussion in the previous chapter formalized interpolation for general stochastic optimization problems and derived connections between interpolation and the weak/strong growth conditions. Now, we turn to the main interest of this work: the complexity of iterative algorithms for $(f, \mathcal{O})$ when interpolation is satisfied. This chapter analyzes the convergence of SGD for strongly-convex and convex functions, while the next two chapters tackle SGD with the Armijo line-search (Chapter 4) and stochastic acceleration (Chapter 5). In particular, this chapter establishes the following non-asymptotic results for SGD with a fixed step-size:

1. Linear convergence *in-expectation* for strongly-convex f when $(f, \mathcal{O})$ satisfies strong growth; this rate is tight with the best-known deterministic rates when $\rho = 1$.

2. *Almost sure* linear convergence for strongly-convex f and individually-smooth and strongly-convex SFO $\mathcal{O}$.

3. Sub-linear convergence for convex f when $(f, \mathcal{O})$ satisfies weak growth; our proof is simpler than existing analyses and permits a larger step-size.

4. Faster sub-linear convergence for convex f when $(f, \mathcal{O})$ satisfies weak growth *and* $\mathcal{O}$ is individually-smooth; this rate is tight with the deterministic case when $\alpha = 1$.

Section 3.3 at the end of this chapter leaves the finite-time regime and considers asymptotic, almost-sure convergence of SGD with a fixed step-size under strong and weak growth, respectively. The following is proved:

Stochastic Gradient Descent

0. Choose an initial point $w_0 \in \mathbb{R}^d$.

1. For each iteration $k \geq 0$:

(a) Query $\mathcal{O}$ for $\nabla f(w_k, z_k)$.

(b) Update input as

$$w_{k+1} = w_k - \eta \nabla f(w_k, z_k).$$

Figure 3.1: Stochastic gradient descent with a fixed step-size η. Note that only one query to the stochastic oracle is needed per-iteration.

1. Almost-sure convergence to a stationary point when f is a general non-convex function and $(f, \mathcal{O})$ satisfies strong growth.

2. Almost-sure convergence to a global minimum when f is convex and $(f, \mathcal{O})$ satisfies weak growth.

This last result is particularly interesting because it concerns convergence of the last input generated by SGD; we shall see that such results are not straightforward in the non-asymptotic regime, where convergence is instead shown for an averaged input.

Now, let us briefly introduce SGD with a fixed step-size before diving into the analysis. The basic procedure is given in Figure 3.1; the key components of the algorithm are (i) the use of a stochastic gradient $\nabla f(w_k, z_k)$ queried from the oracle at every iteration, (ii) the fixed step-size $\eta > 0$, and (iii) the sequence of inputs (w_k) generated by the algorithm, which are called the *iterates*. The non-asymptotic rates in this chapter will be derived by analyzing the sequence of distances to a minimizer, $\left(\|w_k - w^*\|^2\right)$. The minimizer w^* is unique for strongly-convex functions and we show $w_k \to w^*$ in the iteration limit. In contrast, we only establish $f(\bar{w}_K) \to f(w^*)$ for convex functions, where $\bar{w}_K = \frac{1}{K}\sum_{k=0}^{K-1} w_k$. Chapter 4 discusses these proof techniques in greater detail and with specific reference to the case where η is itself a random variable that depends on $\nabla f(w_k, z_k)$. We will see that this introduces significant challenges compared to SGD with a fixed and deterministic

step-size.

3.1 Convergence for Strongly-Convex Functions

The analysis of fixed step-size SGD for strongly-convex functions is divided into two sub-sections based on the properties of $\mathcal{O}$. First, we consider general SFOs satisfying strong growth. Then, the setting is restricted to individually-smooth and convex SFOs, where we show that existing convergence results can be slightly improved. Finally, we analyze the case when $\mathcal{O}$ is both individually-smooth and individually-strongly-convex. This setting degenerates to a simple deterministic optimization problem if $\nabla f(\cdot, z)$ is directly accessible to the optimizer, which is true for mini-batch oracles.

3.1.1 General Oracles

We first establish the convergence rate of SGD for strongly-convex f when $(f, \mathcal{O})$ satisfies the strong growth condition. Recall from Lemma 7 that this is more general than assuming $\mathcal{O}$ is $L_{\max}$ individually-smooth and $(f, \mathcal{O})$ satisfies minimizer interpolation. Furthermore, in the case that individual smoothness and minimizer interpolation do hold, we are guaranteed $\rho \leq \frac{L_{\max}}{\mu}$. This value should be kept in mind, as it informs the worst-case rate that can be obtained in the following theorem.

Theorem 1. *Let f be a μ-strongly-convex, L-smooth function and $\mathcal{O}$ a SFO such that $(f, \mathcal{O})$ satisfies the strong growth condition with parameter ρ. Then stochastic gradient descent with fixed step-size $\eta \leq \frac{2}{\rho(\mu+L)}$ converges as*

$$\mathbb{E}\left[\|w_K - w^*\|^2\right] \leq \left(1 - \frac{2\eta\mu L}{\mu + L}\right)^K \|w_0 - w^*\|^2.$$

See Section B.1 for proof.

We compare this convergence rate to the original result given by Schmidt and Le Roux (2013, Section 6) in Table 3.1. Our analysis allows for a larger step-size and establishes asymptotically faster convergence. The improvement is most significant for ill-conditioned

problems, where $\mu \ll L$ implies $\frac{2L}{\mu+L} \approx 2$. Finally, this result is tight in the sense that when $\rho = 1$, which holds in the deterministic setting, it recovers the best known convergence rate for gradient descent on strongly-convex functions (Bubeck, 2015, Theorem 3.12).

When $(f, \mathcal{O})$ satisfies minimizer interpolation and $\mathcal{O}$ is individually-smooth, the complexity given by Theorem 1 can be worse than that achieved under the weak-growth condition. To see this, consider when the worst-case values $\rho = \frac{L_{\max}}{\mu}$ and $\alpha = \frac{L_{\max}}{L}$ are attained in the interpolation setting. If $\eta = \frac{2}{\rho(\mu+L)} = \frac{2\mu}{L_{\max}(\mu+L)}$, then Theorem 1 guarantees

$$\mathbb{E}\left[\|w_K - w^*\|^2\right] \leq \left(1 - \frac{4\mu^2 L}{L_{\max}\left(\mu + L\right)^2}\right)^K \|w_0 - w^*\|^2.$$

In contrast, Vaswani et al. (2019a, Theorem 5) show

$$\mathbb{E}\left[\|w_K - w^*\|^2\right] \leq \left(1 - \frac{\mu}{L_{\max}}\right)^K \|w_0 - w^*\|^2,$$

with a bigger step-size of $\eta = \frac{1}{L_{\max}}$. Noting $\frac{4\mu L}{(\mu+L)^2} \leq 1$ — with equality when $\mu = L$ — establishes that the rate given in this work is slower.

The discrepancy in convergence rates for smooth, interpolating oracles emerges from the use of smoothness and strong-convexity in the proof of Theorem 1. The worst-case value for ρ is proved using $L_{\max}$-smoothness of $f(\cdot, z_k)$ and μ-strong-convexity of f (see Lemma 7). Thus, bounding $\mathbb{E}_{z_k}\left[\|\nabla f(w_k, z_k)\|^2\right]$ using strong growth and then following the typical, deterministic proof strategy equates to using smoothness and strong-convexity *twice* and leads to an unnecessarily loose bound.

The above conclusion only holds when $(f, \mathcal{O})$ satisfies interpolation and $\mathcal{O}$ is individually smooth. The two convergence speeds cannot be directly compared when $\mathcal{O}$ is more general because of the cyclic relationship between the strong/weak growth parameters. Recall from Lemma 5 that strong growth implies weak growth with constant $\alpha \leq 2\rho L$, while weak growth implies strong growth with constant $\rho \leq \alpha \frac{L}{\mu}$. As a result, no order can be established on α and ρ and it is not clear which rate is superior.

3.1.2 Individually Smooth and Convex Oracles

It is possible to improve upon the result from Vaswani et al. (2019a) when $\mathcal{O}$ further satisfies individual convexity. In particular, we are able to obtain the same convergence

Assumptions	Max. Step-Size		Best Convergence Rate	
	Ours	SLR	Ours	SLR
—	$\eta \leq \frac{2}{\rho(\mu+L)}$	$\eta \leq \frac{2}{\rho L}$	$O\left(\frac{\rho(\mu+L)^2}{4\mu L}\log\left(\frac{1}{\epsilon}\right)\right)$	$O\left(\frac{\rho L}{\mu}\log\left(\frac{1}{\epsilon}\right)\right)$
Ind. Smooth & Convex	$\eta < \frac{2}{L_{\max}}$	N/A	$O\left(\frac{L_{\max}}{\mu}\log\left(\frac{1}{\epsilon}\right)\right)$	N/A
Ind. Smooth & μ_z-SC	$\eta \leq \frac{2}{\mu_{\max}+L_{\max}}$	N/A	$O\left(\frac{\mu_{\max}+L_{\max}}{4\delta_{\min}}\log\left(\frac{1}{\epsilon}\right)\right)$	N/A

Table 3.1: Comparison of iteration complexities of fixed step-size SGD for μ-SC f under the strong growth condition. For each result, we report the complexity obtained with the optimal step-size according to that analysis. Recall from Theorem 3 that $\delta_{\min} = \min_{z\in\mathcal{Z}} \frac{\mu_z L_z}{\mu_z + L_z}$. The first row shows our results for general $\mathcal{O}$ are tighter than SLR (Schmidt and Le Roux, 2013) because they allow for larger step-sizes.

speed with a looser bound on the step-size. A case analysis in Section B.4 shows that the step-size permitted by Vaswani et al. (2019a, Theorem 5) is $\eta \leq \frac{\mu+L}{\alpha L^2}$, which is $\frac{\mu+L}{LL_{\max}}$ in the worst-case. If $\mu < L$, then $\frac{\mu+L}{LL_{\max}} < \frac{2}{L_{\max}}$ and the following theorem improves upon the original result.

Theorem 2. *Let f be a μ-strongly-convex, L-smooth function and $\mathcal{O}$ an L_{max} individually-smooth and convex SFO such that $(f, \mathcal{O})$ satisfies minimizer interpolation. Then stochastic gradient descent with fixed step-size $\eta < \frac{2}{L_{max}}$ converges as*

$$\mathbb{E}\left[\|w_K - w^*\|^2\right] \leq \left(1 - \mu\,\eta\,(2 - \eta L_{max})\right)^K \|w_0 - w^*\|^2.$$

See Section B.1 for proof.

A key feature of Theorem 2 is that it requires only the full function f to be strongly-convex; the stochastic functions $f(\cdot, z)$ may be merely convex, as is typically the case in the finite-sum setting. To illustrate this, consider the linear regression problem

$$\min_{w\in R^d} \sum_{i=0}^{n} \left(\langle w, x_i\rangle - y_i\right)^2,$$

which satisfies interpolation if $y \in \text{span}(\{x_i\}_{i=0}^n)$. The objective f is μ-strongly-convex if

$n \geq d$ and the data matrix is full-rank, while the individual functions $f_i(w) = (\langle w, x_i \rangle - y_i)^2$ are *not* strongly-convex unless $d = 1$.[3] Learning problems of this form occur in non-parametric kernel regression, such as when using radial basis functions with a small bandwidth (Hastie et al., 2009). However, in the unlikely case that $\mathcal{O}$ is also individually-strongly-convex, we can show that SGD will converge almost surely, as established in the following theorem.

Theorem 3. *Let f be a μ-strongly-convex, L-smooth function and $\mathcal{O}$ an L_{max} individually-smooth and μ_{max}-strongly-convex SFO such that $(f, \mathcal{O})$ satisfies minimizer interpolation. Then stochastic gradient descent with fixed step-size $\eta \leq \frac{2}{\mu_{max}+L_{max}}$ converges almost surely at the rate*

$$\|w_K - w^*\|^2 \leq (1 - 2\eta\,\delta_{min})^K \|w_0 - w^*\|^2,$$

where $\delta_{min} = \min_{z \in \mathcal{Z}} \frac{\mu_z L_z}{\mu_z + L_z}$.

See Section B.1 for proof.

Theorem 3 is not surprising; strongly-convex functions have only one minimizer, meaning that gradient-descent on a single stochastic function $f(\cdot, z_k)$ is sufficient to recover the global solution to the optimization problem. Choosing the best-conditioned sub-function function yields convergence to w^* as $O\left(\exp\left\{-2\eta\,\delta_{\max} K\right\}\right)$, where $\delta_{\max} = \max_{z \in \mathcal{Z}} \frac{\mu_z L_z}{\mu_z + L_z}$ (Bubeck, 2015). Notice that we have exchanged worst-case performance for best-case ($\delta_{\min}$ vs $\delta_{\max}$) by optimizing $f(\cdot, z)$ directly. SGD is clearly sub-optimal when $\mathcal{O}$ is individually-strongly-convex and the optimization procedure has direct access to $f(\cdot, z)$ for each $z \in \mathcal{Z}$, such as in the finite-sum setting. This illustrates the dangers of assuming individual strong-convexity and interpolation hold simultaneously.

3.2 Convergence for Convex Functions

Convex functions are significantly more interesting than strongly-convex functions when interpolation is satisfied. Minimizer interpolation for convex functions implies $\mathcal{X}^* \subseteq \mathcal{X}_z^*$ — the optimal set for f is a subset of the optimal set for $f(\cdot, z)$. This condition intuitively feels milder than the requirement for a single shared optimal point w^*. Moreover, assuming

[3] Consider w and $w' = w + v$, where v is orthogonal to x_i, to see the f_i are not strongly-convex when $d > 1$.

individual convexity does not lead to degenerate optimization problems such as those seen in the previous section. Convex functions also require a major shift in analysis; now we shall show concentration of the optimality gap $f(w)-f(w^*)$, rather than shrinking distance to a specific minimizer, $\|w-w^*\|^2$.

We first establish the convergence rate of SGD for convex functions under the weak growth condition. As before, it is strictly more general to analyze the complexity of SGD under weak growth than in the setting where $\mathcal{O}$ is individually-smooth and minimizer interpolation holds. Remember that Lemma 6 guarantees $\alpha \leq \frac{L_{\max}}{L}$. The following result improves on that given by Vaswani et al. (2019a) by constant factors and allows for a larger step-size (see Table 3.2). Moreover, the proof in Section B.2 is simpler and far shorter.

Theorem 4. *Let f be a convex, L-smooth function and $\mathcal{O}$ a SFO such that $(f, \mathcal{O})$ satisfies the weak growth condition with parameter α. Then stochastic gradient descent with fixed step-size $\eta < \frac{1}{\alpha L}$ converges as*

$$\mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{1}{2\eta\left(1-\eta\alpha L\right)K}\|w_0 - w^*\|^2,$$

where $\bar{w}_K = \frac{1}{K}\sum_{k=0}^{K-1} w_k$ and $w^* = \Pi_{\mathcal{X}^*}(w_0)$.

In fact, an even larger step-size and faster convergence rate can be obtained when $\mathcal{O}$ is individually-smooth. We show this now.

Theorem 5. *Let f be a convex, L-smooth function and $\mathcal{O}$ a SFO such that $(f, \mathcal{O})$ satisfies the weak growth condition with parameter α. Moreover, suppose $\mathcal{O}$ is L_{max} individually-smooth. Then stochastic gradient descent with fixed step-size $\eta < \frac{1}{\alpha L} + \frac{1}{L_{max}}$ converges as*

$$\mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{1}{2\eta\,\delta\,K}\|w_0 - w^*\|^2,$$

where $\bar{w}_K = \frac{1}{K}\sum_{k=0}^{K-1} w_k$, $w^* = \Pi_{\mathcal{X}^*}(w_0)$, and $\delta = \min\left\{1, 1+\alpha L\left(\frac{1}{L_{max}} - \eta\right)\right\}$.

See Section B.2 for proof.

Theorem 5 is tight with the deterministic case in the following sense: if $f(\cdot, z) = f$ for each $z \in \mathcal{Z}$, then $\alpha = 1$, $L_{\max} = L$, and the rate given is comparable to the best known results in the deterministic setting (cf. Bubeck (2015, Theorem 3.3)). This result also further illustrates the benefits of directly assuming the weak growth condition, since the maximum

Assumptions	Max. Step-Size		Best Convergence Rate	
	Ours	VBS	Ours	VBS
—	$\eta < \frac{1}{\alpha L}$	$\eta < \frac{1}{2\alpha L}$	$O\left(\frac{\alpha L}{\epsilon}\right)$	$O\left(\frac{4(1+\alpha)L}{\epsilon}\right)$
Convex + Ind. Smooth	$\eta \leq \frac{1}{\alpha L} + \frac{1}{L_{\max}}$	N/A	$O\left(\frac{L_{\max}}{2\epsilon}\right)$	N/A

Table 3.2: Comparison of iteration complexities of fixed step-size SGD for convex f under the weak growth condition. For each result, we report the complexity obtained with the optimal step-size according to that analysis. Our results are tighter than VBS (Vaswani et al., 2019a) and allow for larger step-sizes.

step-size satisfies $\frac{1}{\alpha L} + \frac{1}{L_{\max}} \geq \frac{2}{L_{\max}}$, where $\frac{2}{L_{\max}}$ is the condition obtained when deriving weak growth directly from individual smoothness and minimizer interpolation.

3.3 Almost Sure Convergence

Now we briefly change paradigms and consider the asymptotic behavior of SGD with a fixed step-size. Our goal in this section is to show the almost-sure (with probability 1) convergence of SGD to a minimizer or stationary point of f when $(f, \mathcal{O})$ satisfy weak or strong growth, respectively. In the latter scenario, this means we want to show the random variable $\lim_{k\to\infty} \|\nabla f(w_k)\|^2$ exists and is almost-surely zero. We will need some tools from measure-theoretic probability to accomplish this.

To start, observe that each iterate w_k can be written as a deterministic Borel function of the stochastic gradients $\{\nabla f(w_k, z_k)\}_{k=0}^K$ by unrolling the SGD update. Formally, we also assume a probability space $(\Omega, \mathcal{F}, P)$ in the background and say the sequence (w_k) is *adapted* to the filtration generated by the stochastic gradients,

$$\mathcal{F}_K = \sigma\left(\bigcup_{k=0}^{K-1} \sigma\nabla f(w_k, z_k)\right).$$

The sequences of function and gradient values are functions of (w_k) and so are also adapted to $(\mathcal{F}_k)$. See Çınlar (2011) for additional details on filtrations.

Our main tool to show convergence of sequences of random variables will be super-

martingale theory (Çınlar, 2011). Supermartingales are one of two classic tools used to analyze the convergence of SGD, the other being Lyapunov functions (Bertsekas and Tsitsiklis, 2000). In particular, recent authors have made use of convergence of discrete-time, positive supermartingales (Bertsekas, 2011; Nguyen et al., 2018). This theorem is due to Neveu (1975) and will be the cornerstone of our analyses; we reproduce it here for convenience.

Theorem 6 (Convergence of Positive Supermartingales)**.** *Let (Y_k), (X_k), and (A_k) be discrete, non-negative random processes indexed by $k \in \mathbb{N}$ and adapted to the filtration $(\mathcal{F}_k)$. Suppose that*

$$\forall k \in \mathbb{N},\ \mathbb{E}\left[Y_{k+1} \mid \mathcal{F}_k\right] \leq Y_k - X_k + A_k, \qquad \text{and} \qquad \sum_{k=0}^{\infty} A_k < \infty,$$

almost surely. Then the sequence (Y_k) converges almost surely to a non-negative random variable Y_∞ and $\sum_{k=0}^{\infty} X_k < \infty$ almost surely.

With the necessary measure-theoretic background complete, we are now ready to study the almost-sure convergence of stochastic gradient descent.

Theorem 7. *Let f be a convex, L-smooth function with at least one finite minimizer and $\mathcal{O}$ an L_{max} individually-smooth SFO such that $(f, \mathcal{O})$ satisfies the weak growth condition with parameter α. Then the sequence $(f(w_k))$ generated by stochastic gradient descent with fixed step-size $\eta < \frac{1}{\alpha L} + \frac{1}{L_{max}}$ converges to the optimal function value $f(w^*)$ almost surely.*

The proof is given in Section B.3 and follows a straightforward argument. First, we establish that the sequence of distances to an arbitrary finite minimizer $\left(\|w_k - w^*\|^2\right)$ satisfies the conditions of Theorem 6. As a by-product, we establish that $\sum_{k=0}^{\infty} \left(f(w_k) - f(w^*)\right)$ is convergent (although $w_k \to w^*$ does not necessarily hold almost surely) and then deduce

$$\lim_{k \to \infty} f(w_k) \overset{\text{a.s.}}{=} f(w^*).$$

It is straightforward to prove an alternative version of Theorem 7 which does not require individual smoothness. In this case, convergence is established for $\eta < \frac{1}{\alpha L}$ using the same progress condition as in Theorem 4, namely Equation B.1.

Theorem 7 holds for the *final* iterate generated by the SGD procedure. This should be contrasted with Theorems 4 and 5, which apply only to the averaged iterate $\bar{w}_k$. While the existence of an asymptotic result suggests that non-asymptotic, final-iterate convergence

for constant step-size SGD under the weak growth condition is possible, we do not establish such a result in this work. Convergence (or non-convergence) of the final iterate remains an interesting and surprisingly simple open problem in optimization under interpolation.

The final result of this chapter shows almost-sure convergence to a stationary point for general non-convex functions f such that $(f, \mathcal{O})$ satisfy the strong growth condition. The proof is presented in Section B.3 and follows a similar structure to the proof of Theorem 7.

Theorem 8. *Let f be an L-smooth function with at least one finite minimizer w^* and $\mathcal{O}$ a SFO such that $(f, \mathcal{O})$ satisfies the strong growth condition with parameter ρ. Then the sequence of gradient-norms $\left(\|\nabla f(w_k)\|^2\right)$ generated by stochastic gradient descent with fixed step-size $\eta < \frac{2}{\rho L}$ converges to 0 almost surely.*

3.4 Conclusions

Theorem 8 ends our study of fixed step-size SGD for strongly-convex and convex minimization under the strong/weak growth conditions. Building on work by Schmidt and Le Roux (2013), we established a faster linear convergence rate for μ-strongly-convex functions when using a larger step-size that requires knowledge of μ; this result attains the deterministic rate when $\rho = 1$ (Bubeck, 2015). Unfortunately, the subsequent discussion showed that Theorem 1 can still be slower then the corresponding result of Vaswani et al. (2019a) when minimizer interpolation holds and the worst-case values of α and ρ are attained.

Inspired by this discrepancy, we then leveraged additional properties of $\mathcal{O}$, such as individual smoothness and convexity, to derive convergence rates that match those of Vaswani et al. (2019a), but permit a larger range of step-sizes (Theorem 2). We also proved that SGD converges almost-surely if $\mathcal{O}$ is individually-strongly-convex (Theorem 3). In this case, the optimization problem is degenerate, meaning it can be solved by directly minimizing any $f(\cdot, z)$. Moreover, the conditioning of the optimization problem can be improved from worst-case to best-case if we have direct access to $f(\cdot, z)$ for all z.

The remainder of the chapter established improved convergence rates for convex functions (Theorems 4 and 5). These theorems improve over existing rates by constant factors. Individual smoothness of $\mathcal{O}$ again allowed us to show SGD converges with a slightly wider range of larger step-sizes. Finally, we concluded with asymptotic convergence results for SGD applied to convex and non-convex functions (Theorems 7 and 8).

Chapter 4

Line Search

A major weakness of constant step-size SGD is that the problem constants must be known a priori in order to select a step-size for which convergence is guaranteed. Consider Theorem 1 as an example. A fast linear rate of convergence is guaranteed only when $\eta \leq \frac{2}{\rho(\mu+L)}$, but optimization speed is sub-optimal for $\eta \ll \frac{2}{\rho(\mu+L)}$. Meeting this requirement demands knowledge of the strong-convexity and smoothness constants, μ and L, as well as the strong growth parameter ρ. In practice, we are unlikely to have knowledge of these coefficients or to possess straightforward means for estimating them. This leaves a dilemma: over-estimate the optimal step-size and risk divergence, or use a small step-size that yields slow convergence.

This chapter tackles the problem of step-size selection by augmenting SGD with a stochastic Armijo line-search algorithm that *automatically* finds a suitable step-size at each iteration. Unlike meta-optimization schemes such as grid search, random search, or Bayesian optimization, line-search techniques do not require multiple, costly executions of the optimization procedure (Snoek et al., 2012). Accordingly, practical line-search implementations are typically very fast in comparison to meta-optimization approaches (Nocedal and Wright, 1999). Moreover, as we will show in the following sections, the convergence speed of SGD with a stochastic line-search is tight with that of SGD with the best fixed step-size when minimizer interpolation is satisfied.

The work in this chapter was originally conducted by the author as part of Vaswani et al. (2019b). The proposal to augment SGD with the stochastic Armijo line-search, the algorithmic procedure in Figure 4.2, Lemma 8, and Theorem 11 are reproduced from this

work. In contrast, Theorems 9 and 10 are new, unpublished improvements over the originally published theorems.

4.1 Background

Line-search algorithms are a classic solution to the problem of step-size selection in the deterministic setting. The basic notion is to search along the gradient direction (or, more generally, a descent direction) for a step-size that satisfies a desirable property. A common rule is the Armijo condition,

$$f(w_k + \eta_k v_k) \leq f(w_k) + c \cdot \eta_k \langle \nabla f(w_k), v_k \rangle , \tag{4.1}$$

which demands sufficient decrease in the function value when taking a step along search direction v_k. The vector v_k is required to satisfy $\langle \nabla f(w_k), v_k \rangle < 0$, in which case it is called a *descent direction*. The (locally) steepest descent direction is the negative gradient $v_k = -\nabla f(w_k)$ (Nesterov, 2004). In comparison, the negative stochastic gradient $-\nabla f(w_k, z_k)$ direction used by SGD is only guaranteed to be a descent direction in expectation.

A simple mechanism for enforcing the Armijo condition is iteratively decreasing the step-size, or *backtracking*, from a maximal step-size $\eta_{\max}$ until the condition is satisfied. Lipschitz-smoothness of f guarantees that for all w there exists sufficiently small η_k such that Equation 4.1 holds (Nocedal and Wright, 1999); we will establish a specific case of this for SGD in the stochastic setting (Lemma 8). Practical implementations of backtracking reduce the step-size as $\eta_k \leftarrow \beta \cdot \eta_k$ when the Armijo condition does not hold, where $\beta \in (0, 1)$ is a tunable parameter. Alternatively, cubic or other interpolation schemes can be used to model and approximately minimize the one-dimensional function $g(\eta_k) = f(w_k - \eta_k v_k)$; see Nocedal and Wright (1999) for more details. In this work we assume that backtracking can be accomplished exactly, meaning that η_k is the largest step-size less than $\eta_{\max}$ for which Equation 4.1 is satisfied.

The Armijo line-search as given above is not suitable for stochastic optimization problems. Remember that the optimizer does not have access to exact function and gradient evaluations — it can only access noisy estimates $f(w, z_k)$ and $\nabla f(w, z_k)$. Even when exact evaluations are available, as in the finite-sum setting, the cost of computing function and gradient values is much larger than the cost of querying $\mathcal{O}$. An alternative to the full

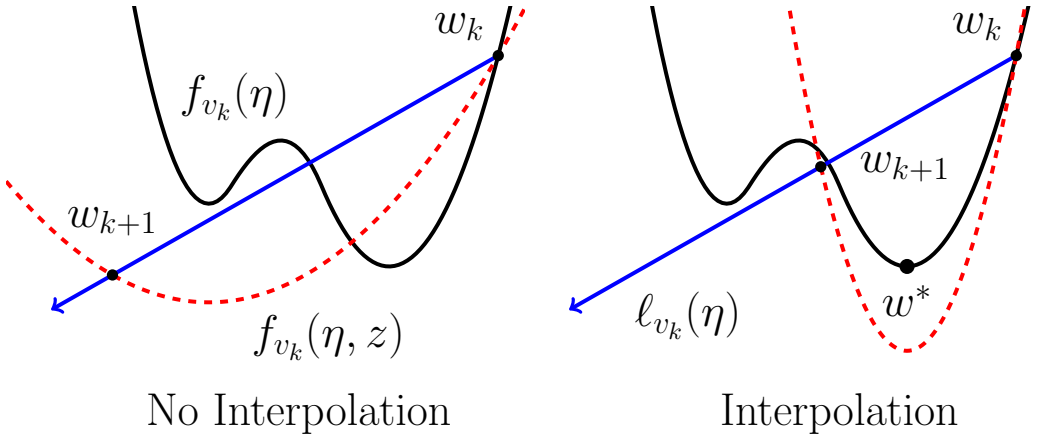

Figure 4.1: Progress when using the stochastic Armijo line-search with and without minimizer interpolation. The function f_{v_k} is the restriction of f to the descent direction $v_k = -\nabla f(w_k)$, while ℓ_{v_k} is the Armijo condition as a function of the step-size, η. The dashed red line shows the restriction of the stochastic function $f(w, z)$ to v_k. The next iterate w_{k+1} is obtained by choosing the largest step-size satisfying the line-search. When interpolation is not satisfied, we "overshoot" significantly and $f(w_{k+1}) > f(w_k)$ as a result. In contrast, minimizer interpolation ensures that the iteration makes progress towards w^*. Note that $f(w_k, z) = f(w_k)$ for convenience only; this need not hold in general.

Armijo line-search is the following stochastic version of the Armijo condition:

$$f(w_k + \eta_k v_k, z_k) \leq f(w_k, z_k) + c \cdot \eta_k \langle \nabla f(w_k, z_k), v_k \rangle . \tag{4.2}$$

This condition is identical to Equation 4.1, but uses stochastic function and gradient evaluations queried from the SFO. Intuitively, the stochastic Armijo condition requires sufficient decrease on $f(\cdot, z_k)$, rather than f. Note that practical backtracking on Equation 4.2 will require $f(\cdot, z_k)$ to be evaluated multiple times in the same iteration. This is straightforward in the finite-sum setting, but may be impossible when z_k reflects an inherently noisy measurement process.

The main issue with stochastic line-search conditions is that the oracle queries may not be representative of the true function. That is, progress as measured by $f(\cdot, z_k)$ may be uninformative or even lead to ascent in f. Figure 4.1 illustrates such a situation, as well as why minimizer interpolation may preclude this problem. Intuitively, progress on the stochastic functions $f(\cdot, z_k)$ should be sufficient for progress on f when $\mathcal{X}^* \subseteq \bigcap_{z \in \mathcal{Z}} \mathcal{X}_z^*$ —

SGD with Armijo Line-Search

0. Choose an initial point $w_0 \in \mathbb{R}^d$.

1. For each iteration $k \geq 0$:

(a) Query $\mathcal{O}$ for $f(w_k, z_k)$, $\nabla f(w_k, z_k)$.

(b) Set $\eta_k = \eta_{\max}$ and backtrack until

$$f(w_k - \eta_k \nabla f(w_k, z_k), z_k) \leq f(w_k, z_k) - c \cdot \eta_k \|\nabla f(w_k, z_k)\|^2.$$

(c) Update input as

$$w_{k+1} = w_k - \eta_k \nabla f(w_k, z_k).$$

Figure 4.2: Stochastic gradient descent with the stochastic Armijo line-search, as proposed by Vaswani et al. (2019b). We assume that the backtracking procedure can be evaluated exactly, meaning the maximal step-size η_k satisfying the Armijo is selected. Note that $\mathcal{O}$ must be queried for additional function values $f(w_{k+1}, z_k)$ during backtracking.

i.e. when $(f, \mathcal{O})$ satisfies minimizer interpolation The remainder of this chapter formalizes this intuition and develops non-asymptotic convergence rates for SGD with the stochastic Armijo line-search. In particular, we establish the following results under minimizer interpolation:

1. Linear convergence for strongly-convex f and $L_{\max}$ individually-smooth and convex $\mathcal{O}$; moreover, this rate is tight with fixed step-size SGD.

2. Sub-linear convergence for convex f and $L_{\max}$ individually-smooth and convex $\mathcal{O}$; again, this rate is tight with the fixed step-size case.

3. Sub-linear convergence to a stationary point for non-convex f when $(f, \mathcal{O})$ satisfies strong growth. Note that this result is reproduced from Vaswani et al. (2019b).

Now, let us introduce a formal definition of SGD with the Armijo line-search before proceeding with our analysis. Figure 4.2 presents the basic procedure of the algorithm.

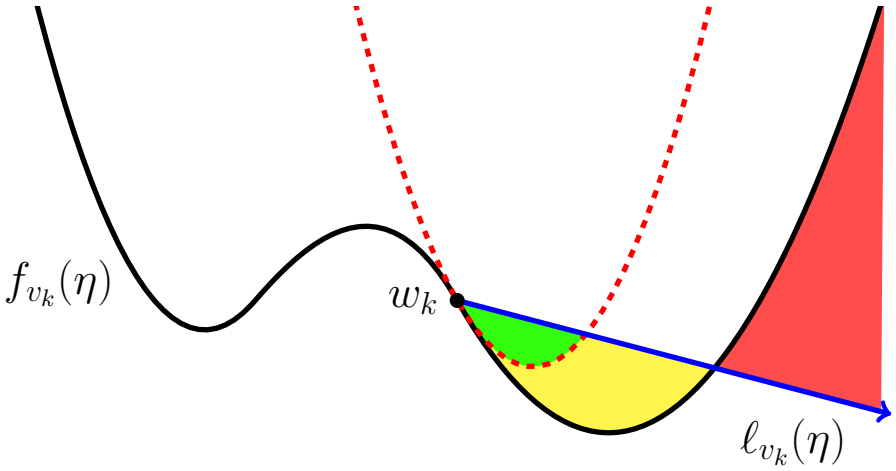

Figure 4.3: Graphical depiction of Lemma 8. The function f_{v_k} is the restriction of f to the descent direction $v_k = -\nabla f(w_k)$, while ℓ_{v_k} is the Armijo condition as a function of the step-size, η. The dashed red line shows the upper-bound on f_{v_k} due to Lipschitz-smoothness. We can see that a large range of step-sizes make sufficient progress (yellow region) and will be accepted by the Armijo condition; step-sizes for which the graph of ℓ_{v_k} lies below f_{v_k} (red region) do not make enough progress and will be rejected. Smoothness guarantees that ℓ_{v_k} is above f_{v_k} for a non-empty interval of step-sizes (green region), which will always be accepted.

There are several key differences from fixed step-size SGD, which are noted as follows: (i) the step-size η_k is defined per-iteration and initialized at $\eta_{\max}$, (ii) η_k is then chosen by exact backtracking on Equation 4.2 evaluated at $v_k = \nabla f(w_k, z_k)$, and (iii) the backtracking procedure implicitly requires additional queries to $\mathcal{O}$ for $f(w_k - \eta_k \nabla f(w_k, z_k), z_k)$. While Figure 4.2 uses an idealized backtracking procedure, in practice the hyper-parameter β may need to be tuned to avoid excessive oracle queries.

4.2 Convergence for Strongly-Convex Functions

The first step of our analysis is to control the step-size η_k using smoothness and the line-search condition. Lower-bounding the step-size relies on smoothness of $f(\cdot, z_k)$, which is why the results in this chapter require $\mathcal{O}$ to be individually-smooth. This fact also motivates our transition from assuming the strong/weak growth conditions to directly reasoning with

individually-smooth oracles where $(f, \mathcal{O})$ satisfies minimizer interpolation. The following lemma is reproduced from Vaswani et al. (2019b); the proof in Appendix C is modified from the original for greater clarity.

Lemma 8. *Let f be an L-smooth function and $\mathcal{O}$ an L_{max} individually-smooth SFO such that $(f, \mathcal{O})$ satisfies minimizer interpolation. Then the maximum possible step-size returned by the stochastic Armijo line-search constrained to lie in the $(0, \eta_{max}]$ range satisfies the following inequalities:*

$$\min\left\{\frac{2(1-c)}{L_{max}}, \eta_{max}\right\} \leq \eta_k \leq \frac{f(w_k, z_k) - f(w^*, z_k)}{c\|\nabla f(w_k, z_k)\|^2}.$$

Lemma 8 is the main tool needed to analyze SGD with the Armijo line-search in the convex setting. Specifically, it allows us to establish the following progress condition, which will be at the core of our proofs for convex and strongly convex functions.

Lemma 9. *Let f be a convex, L-smooth function and $\mathcal{O}$ an L_{max} individually-smooth and convex SFO such that $(f, \mathcal{O})$ satisfy minimizer interpolation. Then stochastic gradient descent using the Armijo line-search with $c \geq \frac{1}{2}$ satisfies the following inequality:*

$$f(w_k) - f(w^*) \leq \frac{1}{2}\max\left\{\frac{L_{max}}{2(1-c)}, \frac{1}{\eta_{max}}\right\}\left(\|w_k - w^*\|^2 - \mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right]\right),$$

where $w^* \in \mathcal{X}^*$ is arbitrary.

See Appendix C for proof.

Lemma 9 is the ideal inequality for (a) applying μ-strong-convexity to obtain a recursion for $\|w_k - w^*\|^2$ and a linear convergence rate, or (b) summing over iterations to obtain sub-linear convergence of the average iterate $\bar{w}_K$. Indeed, these are exactly the strategies used to prove Theorems 9 and 10. We start with the strongly-convex case, with for the following theorem proof given in Section C.1.

Theorem 9. *Let f be a μ-strongly-convex, L-smooth function and $\mathcal{O}$ an L_{max} individually-smooth and convex SFO such that $(f, \mathcal{O})$ satisfies minimizer interpolation. Then stochastic gradient descent using the Armijo line-search with $c \geq \frac{1}{2}$ converges as*

$$\mathbb{E}\left[\|w_K - w^*\|^2\right] \leq \left(1 - \mu\min\left\{\frac{2(1-c)}{L_{max}}, \eta_{max}\right\}\right)^K \|w_0 - w^*\|^2.$$

Setting $c = \frac{1}{2}$ and $\eta_{\max} = \infty$ in Theorem 9 gives

$$\mathbb{E}\left[\|w_K - w^*\|^2\right] \leq \left(1 - \frac{\mu}{L_{\max}}\right)^K \|w_0 - w^*\|^2,$$

which is exactly the rate obtained for fixed step-size SGD in the same setting (cf. Theorem 2). The convergence speed is also the same as the worst-case under the weak growth condition when interpolation and individual smoothness hold (Vaswani et al., 2019a, Theorem 5). While this latter result does not use individual convexity, we (intuitively) need this additional assumption when using a line-search in order for the step-size to yield reliable progress towards the global minimizer. Indeed, general non-convex $f(\cdot, z_k)$ can satisfy minimizer interpolation and still have stochastic gradients $\nabla f(\cdot, z_k)$ which point away from the minimizer w^*. This problem is presented formally in Section 4.4.1, which discusses challenges in establishing convergence for non-convex functions.

The $c \geq \frac{1}{2}$ requirement in Theorem 9 is notable because it is the opposite of the constraint on c required for Newton or quasi-Newton methods (cf. Nocedal and Wright (1999, Theorem 3.6)). The key issue here is that the Armijo line-search with $c > \frac{1}{2}$ can exclude unit step-lengths (particularly for quadratic functions), which must be accepted in some neighborhood of w^* for Newton and quasi-Newton methods to obtain local superlinear convergence. Unit step-sizes are not required for SGD, where the stochastic Armijo line-search is mainly used to obtain sufficient progress at each iteration. Moreover, if $\mathcal{O}$ is individually-smooth and $c \in (0, 1)$, then Lemma 8 guarantees SGD will satisfy Equation 4.2 for some $\eta_k > 0$. As such, larger c values — which demand greater decrease at each iteration — arise naturally in our analysis.

Table 4.1 contrasts Theorem 9 with the other known result for SGD with Armijo line-search on strongly-convex functions (Vaswani et al., 2019b, Theorem 1). In particular, note that the convergence rate presented here depends on the strong-convexity constant of the overall function, μ, instead of the expected constant $\bar{\mu} = \mathbb{E}_{z_k}[\mu_{z_k}]$. This is a nice theoretical improvement for the following reason: our analysis permits all stochastic functions to be convex only, while that of Vaswani et al. (2019b) requires $f(\cdot, z)$ to be strongly-convex for at least one $z \in \mathcal{Z}$. Such a condition is unlikely to hold in practice and can lead to degenerate problems (see the discussion in Section 3.1.2). Theorem 9 is motivated by the work of Loizou et al. (2020), who previously obtained a similar improvement in the context of the stochastic Polyak step-size.

4.3 Convergence for Convex Functions

Now we derive a sub-linear convergence rate for SGD with the stochastic Armijo line-search when f is convex, $\mathcal{O}$ is individually-smooth, and minimizer interpolation is satisfied. Convergence is established for the averaged iterate $\bar{w}_K$ similarly to the case of constant step-size SGD; non-asymptotic, final-iterate rates are also an open problem when η_k is chosen by line-search. The proof of the following theorem follows almost immediately from Lemma 9 as briefly described in the previous section; it can be found in Section C.2.

Theorem 10. *Let f be a convex, L-smooth function and $\mathcal{O}$ an L_{max} individually-smooth and convex SFO such that $(f, \mathcal{O})$ satisfies minimizer interpolation. Then stochastic gradient descent using the Armijo line-search with $c \geq \frac{1}{2}$ converges as*

$$\mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{1}{2\,K} \max\left\{\frac{L_{max}}{2(1-c)}, \frac{1}{\eta_{max}}\right\} \|w_0 - w^*\|^2,$$

where $\bar{w}_K = \sum_{k=0}^{K-1} w_k$ and $w^* = \Pi_{\mathcal{X}^*}(w_0)$.

This result is tight in the following sense: when $c = \frac{1}{2}$ and $\eta_{\max} = \infty$, Theorem 10 yields

$$\mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{L_{\max}}{2\,K} \|w_0 - w^*\|^2.$$

This rate is identical to that for constant step-size SGD with $\eta = \frac{1}{L_{\max}}$ if we assume the worst-case value for the weak-growth parameter α in Theorem 4. If $f(\cdot, z) = f$ for each $z \in \mathcal{Z}$, then $L_{\max} = L$ and this rate is comparable to the best known convergence results for gradient descent on convex functions (Bubeck, 2015).

Table 4.1 compares Theorem 10 with the original result from Vaswani et al. (2019b, Theorem 2). The sub-linear rate established here is faster by constant factors and has a simpler proof. Furthermore, our result holds for all $c \geq \frac{1}{2}$, while $c > \frac{1}{2}$ must hold strictly for their theorem. The unnecessary strictness of the original result prevents the natural case where $c = \frac{1}{2}$, $v_k = \nabla f(w_k, z_k)$ and Equation 4.2 becomes equivalent to the quadratic upper-bound implied by Lipschitz-smoothness.

Assumptions	Minimum c		Convergence Rate	
	Ours	VML+	Ours	VML+
μ-SC	$c \geq \frac{1}{2}$	$c \geq \frac{1}{2}$	$O\left(\frac{L_{\max}}{\mu} \log\left(\frac{1}{\epsilon}\right)\right)$	$O\left(\frac{L_{\max}}{\bar{\mu}} \log\left(\frac{1}{\epsilon}\right)\right)$
Convex	$c \geq \frac{1}{2}$	$c > \frac{1}{2}$	$O\left(\frac{L_{\max}}{2\,\epsilon}\right)$	$O\left(\frac{3L_{\max}}{\epsilon}\right)$

Table 4.1: Comparison of iteration complexities for SGD with the stochastic Armijo line-search. We omit results for non-convex functions, which are identical between the two works. Results are shown for $\eta_{\max} = \infty$ and $c = \frac{1}{2}$, excepting the convex case of VML+ (Vaswani et al., 2019b) where $c = \frac{2}{3}$ is used as suggested by the authors. The constant $\bar{\mu} = \min_k \{\mathbb{E}_{z_k}[\mu_{z_k}]\}$ requires that at least one stochastic function in the support of z_k is strongly convex for each iteration k; in contrast, our strongly-convex proof relies only on μ, the parameter of the true function.

4.4 Convergence for Non-Convex Functions

The complexity of SGD with the stochastic Armijo line-search for general non-convex functions is particularly challenging to determine. As discussed below, proofs which analyze the sequence of distances to a minimizer $\left(\|w_k - w^*\|^2\right)$ fail without convexity. This includes the technique used to prove Theorems 9 and 10. Instead, convergence to a stationary point is typically established through the quadratic majorant provided by L-smoothness of f. The following theorem shows that such convergence does indeed hold, but mandates a severe upper-bound on the line-search. A major weakness of our result is that the setting of $\eta_{\max}$ requires explicit knowledge of ρ and $L_{\max}$.

Theorem 11. *Let f be an L-smooth function and $\mathcal{O}$ an L_{max} individually-smooth SFO such that $(f, \mathcal{O})$ satisfies the strong growth condition with parameter ρ. Then stochastic gradient descent using the Armijo line-search with $c > 1 - \frac{L_{max}}{\rho L}$ and $\eta_{max} < \frac{2}{\rho L}$ converges as*

$$\min_{k \in [K]} \|\nabla f(w_k)\|^2 \leq \frac{1}{\delta\, K} \left(f(w_0) - f(w^*)\right),$$

where $\delta = \left(\eta_{max} + \frac{2(1-c)}{L_{max}}\right) - \rho\left(\eta_{max} - \frac{2(1-c)}{L_{max}} + L\eta_{max}^2\right)$.

See Section C.3 for proof.

The step-size constraint $\eta_k \leq \eta_{\max} < \frac{2}{\rho L}$ forces the Armijo line-search to behave like constant step-size SGD. Indeed, Lemma 14 shows that $\eta < \frac{2}{\rho L}$ is exactly the condition required for constant step-size SGD to make guaranteed progress for an L-smooth function such that $(f, \mathcal{O})$ satisfies strong growth. It is not clear if the upper-bound on $\eta_{\max}$ is a fundamental property of SGD for non-convex functions, or an artifact of the proof for Theorem 11. In either case, it is worthwhile to see how the requirement on $\eta_{\max}$ emerges.

4.4.1 Challenges in the Analysis

A main object in our proofs so far has been the inner-product

$$\eta_k \langle \nabla f(w_k, z_k), w_k - w^* \rangle.$$

When η_k is independent of z_k, linearity of expectation gives the following:

$$\mathbb{E}_{z_k} \left[\eta_k \langle \nabla f(w_k, z_k), w_k - w^* \rangle \right] = \eta_k \langle \nabla f(w_k), w_k - w^* \rangle .$$

It is now straightforward to use convexity of f to control this term, which is how the proofs for Chapter 3 proceed. However, the stochastic Armijo line-search yields step-sizes which are corollated with $\nabla f(w_k, z_k)$ and thus are not independent of z_k. A straighforward solution is to use individual convexity and minimizer interpolation to obtain

$$\eta_k \langle \nabla f(w_k, z_k), w_k - w^* \rangle \geq \eta_k \left(f(w_k, z_k) - f(w^*, z_k) \right) \geq 0.$$

Lemma 8 then provides the necessary tool to lower-bound η_k, which is how the proofs for this chapter proceed. The major challenge of non-convex functions is now apparent: the inner-product is not guaranteed to be non-negative[4] and the proof cannot proceed by bounding η_k. Intuitively, disentangling η_k and $\nabla f(w_k, z_k)$ becomes the key to establishing convergence in the non-convex setting.

The successful convergence proof for non-convex functions starts from L-smoothness of f, rather than "going through iterates" as do the proofs for Theorems 9 and 10. Again, a correlated inner-product $\eta_k \langle \nabla f(w_k), \nabla f(w_k, z_k) \rangle$ is encountered; the main innovation is

[4]Non-negativity of $\langle \nabla f(w), w - w^* \rangle$ is sometimes called monotonicity of the gradient and is equivalent to convexity if it holds for all w (Bubeck, 2015).

to expand this as

$$
\begin{aligned}
-2\eta_k \left\langle \nabla f(w_k), \nabla f(w_k, z_k) \right\rangle &= \eta_k \left(\|\nabla f(w_k, z_k) - \nabla f(w_k)\|^2 - \|\nabla f(w_k)\|^2 - \|\nabla f(w_k, z_k)\|^2 \right) \\
&\leq \eta_{\max} \|\nabla f(w_k, z_k) - \nabla f(w_k)\|^2 - \eta_{\min} \big(\|\nabla f(w_k)\|^2 \\
&\qquad + \|\nabla f(w_k, z_k)\|^2 \big),
\end{aligned}
$$

where the inequality stems from separately bounding η_k on each positive and negative terms. This bound is worst-case, but separates the step-size and stochastic gradient, allowing the proof to proceed. As a side-effect, we require a tight upper-bound on the maximum step-size: $\eta_{\max} < \frac{2}{\rho L}$.

4.5 Conclusions

The main focus of this chapter was augmenting SGD with an automatic mechanism for selecting η, the step-size parameter. In particular, we sought to develop a tuning-free approach that obtains comparable convergence speed to SGD with a fixed step-size (as established in Chapter 3) *without* restarts or a meta-optimization procedure. This was achieved by combining SGD with the stochastic Armijo line-search, as proposed by Vaswani et al. (2019b). We then developed new convergence results for strongly-convex (Theorem 9) and convex (Theorem 10) functions that improve over those given by Vaswani et al. (2019b) by constant factors. These rates show that SGD with the stochastic Armijo line-search obtains a converge rate for (strongly-) convex functions that is tight with fixed step-size SGD despite requiring no knowledge of the problem constants. Finally, we considered general non-convex functions and showed SGD with the stochastic Armijo line-search converges to a stationary point if an aggressive upper-bound on the step-size is enforced (Theorem 11).

Chapter 5

Acceleration

The focus in Chapters 3 and 4 was on relatively simple first-order methods: SGD with a fixed step-size and with a stochastic line-search, respectively. Now, we move on to Nesterov's famous accelerated gradient descent (AGD) algorithm (Nesterov, 2004), which is guaranteed faster convergence in the deterministic setting. First, a small degree of background for accelerated methods is developed, with a particular focus on why stochastic acceleration is an interesting question for oracles satisfying interpolation. Then, the following convergence results are established for stochastic accelerated gradient descent (SAGD):

1. $O\left(\sqrt{\frac{\rho L}{\mu}}\log\left(\frac{1}{\epsilon}\right)\right)$ iteration complexity for strongly convex f when $(f, \mathcal{O})$ satisfies strong growth; this rate is tight with the deterministic analysis when $\rho = 1$.
2. $O\left(\sqrt{\frac{\rho L}{\epsilon}}\right)$ iteration complexity for convex f when $(f, \mathcal{O})$ satisfies strong growth; again, this rate is tight with AGD when $\rho = 1$.

These convergence rates are tighter than existing work by a factor of $\sqrt{\rho}$ (Vaswani et al., 2019a, Theorems 1-2), which we show is comparable to the speed-up obtained by AGD over gradient descent in the deterministic setting. The chapter ends with a brief discussion of acceleration under the weak growth condition, which is an open question.

5.1 Background

Recall that Theorems 1 and 4 establish fast linear and sub-linear convergence rates for SGD that are comparable with analyses for deterministic problems (e.g. gradient descent). Unfortunately, the classic work by Nemirovsky and Yudin (1983) shows that the convergence of gradient descent is not tight with lower-bounds for convex, Lipschitz-smooth functions with deterministic oracles. Subsequent developments by Nesterov (Nemirovsky and Nesterov, 1985; Nesterov, 1983, 1988) defined a series of first-order methods (terminating with AGD) achieving the optimal $O(\frac{1}{K^2})$ convergence in the setting analyzed by Nemirovsky and Yudin. Algorithms with this rate are called *accelerated* and subsequent research has generated a large number of accelerated algorithms for deterministic problems. That acceleration literature firmly places gradient descent as a sub-optimal algorithm and also casts the analyses in Chapter 3 in a new light. In particular, the obvious parallels between SGD under interpolation and deterministic gradient descent hint that SAGD may be a faster algorithm than SGD in restricted settings.

In fact, restricting the SFO is necessary for any hope of proving an accelerated rate for SAGD. Black-box accelerated methods which do not leverage structural information about f rapidly accumulate errors when used with general stochastic oracles $\mathcal{O}$ (Devolder et al., 2014; Schmidt et al., 2011). Such error accumulation prevents acceleration, as shown by tight $\Omega(\epsilon^{-2})$ and $\Omega(\epsilon^{-1})$ lower bounds for the iteration complexity of convex and strongly-convex minimization with general SFOs, respectively (Agarwal et al., 2012; Nemirovsky and Yudin, 1983; Raginsky and Rakhlin, 2011). The key question is whether or not the interpolation setting is restrictive enough to exclude these lower bounds and permit acceleration; the goal of this chapter is to extend the estimating sequences framework (Nesterov, 2004) to show that SAGD does in fact achieve an accelerated convergence rate when the strong growth condition holds.

We now introduce AGD in greater detail before discussing its analysis using estimating sequences. Figure 5.1 presents the classic procedural definition for AGD with stochastic gradients; the major differences to SGD are (i) AGD introduces a secondary sequence of points (y_k) that are calculated by extrapolating from the primary sequence (w_k), (ii) the primary sequence is updated using stochastic gradient steps from the extrapolation points y_k, rather than the proceeding iterate w_k, and (iii) the extrapolation step-size is computed from a quadratic equation that depends on μ and L, where $\mu = 0$ for convex functions.

Stochastic AGD (SAGD)

0. Choose an initial point $w_0 = y_0 \in \mathbb{R}^d$ and $\alpha_0 \in (0, 1)$.

1. For each iteration $k \geq 0$:

(a) Query $\mathcal{O}$ for $\nabla f(w_k, z_k)$.

(b) Update input as

$$w_{k+1} = w_k - \eta \nabla f(y_k, z_k).$$

(c) Compute $\alpha_{k+1}^2 = (1 - \alpha_k)\alpha_k^2 + \frac{\mu}{L}\alpha_{k+1}$.

(d) Set $\beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}}$ and extrapolate as

$$y_{k+1} = w_{k+1} + \beta_k(w_{k+1} - w_k).$$

Figure 5.1: Classical form of AGD with stochastic gradients. This algorithm is equivalent to that proposed and analyzed by Vaswani et al. (2019a).

The procedure for selecting the β_k step-sizes is particularly unintuitive and has lead some authors to describe the fast convergence of AGD as an "algebraic trick" (Allen Zhu and Orecchia, 2017). Partly for this reason, we shall re-express AGD as alternating steps of gradient descent on f and a new function $\phi : \mathbb{R}^d \to \mathbb{R}$.

5.2 Estimating Sequences

Let (w_k) be the sequence of iterates produced by SGD. Lemma 14 shows that the following descent condition holds if f is L-smooth and $(f, \mathcal{O})$ satisfies strong growth:

$$\mathbb{E}_{z_k}\left[f(w_{k+1})\right] \leq f(w_k) - \eta\left(1 - \frac{\rho L \eta}{2}\right)\|\nabla f(w_k)\|^2.$$

Choosing $\eta = \frac{1}{\rho L}$ minimizes the right-hand-side, in which case SGD can be viewed as iterative minimization of the quadratic upper-bound on f given by smoothness and strong

growth. This procedure decreases f at each iteration by assuming *worst-case* curvature and noise at every iterate w_k. Intuitively, SGD is sub-optimal exactly because it employs global worst-case bounds regardless of past knowledge — e.g. the $(f(w_k, z_k))$ and $(\nabla f(w_k, z_k))$ sequences.

We can instead consider an algorithm which builds local approximations to f based on all past information accumulated by the procedure. Sufficiently accurate approximations can replace or augment the worst-case curvature assumption and allow for faster convergence. This intuition is formalized by the notion of estimating sequences (Nesterov, 2004).

Definition 9 (Estimating Sequences)**.** *The two sequences $(\lambda_k)_{k=0}^{\infty}$ and $(\phi_k)_{k=0}^{\infty}$ are called estimating sequences for $f : \mathbb{R}^d \to \mathbb{R}$ if (i) $\lambda_k \geq 0$ for all $k \in \mathbb{N}$; (ii) $\lim_{k\to\infty} \lambda_k = 0$; and (iii) for all $k \in \mathbb{N}$ and $w \in \mathbb{R}^d$, the functions $\phi_k : \mathbb{R}^d \to \mathbb{R}$ satisfy*

$$\phi_k(w) \leq (1 - \lambda_k) f(w) + \lambda_k \phi_0(w).$$

If ϕ_0 is chosen so that $\phi_0(w) \approx f(w)$ in a neighborhood of w_0, then $(\phi_k)_{k=0}^{\infty}$ matches our intuition of a sequence of improving, local approximations to f. This is particularly true since the conditions on λ_k guarantee $\lim_{k\to\infty} \phi_k \leq f$ point-wise — i.e. ϕ eventually becomes a *global* lower bound on f. Finally, and most importantly, if $f(w_k) \leq \inf_x \phi_k(x)$ for all k, then we obtain

$$\begin{aligned} f(w_k) \leq \phi_k(w^*) &\leq (1 - \lambda_k) f(w^*) + \phi_0(w^*) \\ \implies f(w_k) - f(w^*) &\leq \lambda_k \left(\phi_0(w^*) - f(w^*)\right), \end{aligned}$$

and the convergence rate of λ_k controls convergence of w_k to w^*. Establishing this last property will be the core of our analysis. Let us now fix a choice of estimating sequences following Nesterov (2004, Lemma 2.2.2).

Lemma 10. *Suppose f is a μ-strongly-convex function (with $\mu = 0$ in the convex case). Let $(\alpha_k)_{k=0}^{\infty}$ be a sequence of real numbers such that $\alpha_k \in (0,1)$ and $\sum_{i=1}^{\infty} \alpha_k = \infty$. Let $(y_k)_{k=0}^{\infty}$ be an arbitrary of sequence of points in $\mathbb{R}^n$. Then the following pair of sequences are estimating sequences for f:*

$$\begin{aligned}\lambda_{k+1} &= (1-\alpha_k)\lambda_k \\ \phi_{k+1}(w) &= (1-\alpha_k)\phi_k(w) + \alpha_k\left(f(y_k) + \langle \nabla f(y_k), w - y_k\rangle + \frac{\mu}{2}\|w-y_k\|^2\right),\end{aligned}$$

where $\lambda_0 = 1$ and $\phi_0(x) = \phi_0^* + \frac{\gamma_0}{2}\|x - v_0\|^2$ with arbitrary $\gamma_0 \geq 0$, $v_0 \in \mathbb{R}^n$ and $\phi_0^* \in \mathbb{R}$.

Lemma 10 provides a recipe for generating local approximations of f by updating ϕ_k with a global minorant of f "centered" at y_k. This minorant is quadratic when f is strongly-convex and linear when f is convex (since $\mu = 0$). An important property of the update is that it preserves the canonical form

$$\phi_k(w) = \phi_k^* + \frac{\gamma_k}{2}\|w - v_k\|^2, \tag{5.1}$$

where $\gamma_{k+1} = (1-\alpha_k)\,\gamma_k + \alpha_k\mu$, $v_k = \frac{1}{\gamma_{k+1}}\left((1-\alpha_k)v_k + \alpha_k\mu\, y_k - \alpha_k\nabla f(y_k)\right)$, and

$$\phi_{k+1}^* = (1-\alpha_k)\phi_k^* + \alpha_k\Bigg(f(y_k) - \frac{\alpha_k}{2\gamma_{k+1}}\|\nabla f(y_k)\|^2 + \frac{(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\big(\frac{\mu}{2}\|y_k - v_k\|^2 + \langle \nabla f(y_k), v_k - y_k\rangle\big)\Bigg).$$

See Nesterov (2004, Lemma 2.2.3) for proof. While cumbersome, the canonical form for ϕ_k allows AGD to be reformulated as alternating descent steps on ϕ_k and f. We do this now.

Figure 5.2 re-expresses SAGD in the notation of estimating sequences. We call the resulting algorithm R-SAGD: reformulated stochastic accelerated gradient descent. Note that the order of the (y_k) and (w_k) sequences has been reversed for convenience; this does not affect the algorithm. Formulating SAGD as an estimating sequence procedure is particularly nice for two reasons: (1) the extrapolation step

$$y_{k+1} = w_{k+1} + \beta_k(w_{k+1} - w_k),$$

Reformulated SAGD (R-SAGD)

0. Choose an initial point $w_0 = y_0 \in \mathbb{R}^d$ and $\gamma_0 \geq 0$.

1. For each iteration $k \geq 0$:

(a) Compute $\alpha_k^2 = \frac{\gamma_{k+1}}{\rho L} = \frac{(1-\alpha_k)\gamma_k + \alpha_k \mu}{\rho L}$.

(b) Extrapolate as

$$y_k = w_k - \frac{\alpha_k}{\gamma_k + \alpha_k \mu} \nabla \phi_k(w_k).$$

(c) Query $\mathcal{O}$ for $\nabla f(y_k, z_k)$.

(d) Update input as

$$w_{k+1} = w_k - \frac{1}{\rho L} \nabla f(y_k, z_k).$$

Figure 5.2: Reformulation of AGD as alternating steps of gradient descent on ϕ and SGD on f.

takes on an intuitive interpretation as gradient descent on ϕ_{k+1}, and (2) if

$$\mathbb{E}[\inf \phi_k(x)] \geq \mathbb{E}[f(w_k)],$$

holds for all k, then convergence of SAGD is determined entirely by convergence of (λ_k). We will show this second property shortly, but first we formally state the equivalence of R-SAGD and SAGD, with proof given in Section D.1.

Lemma 11. *Let (λ_k) and (ϕ_k) be as defined in Lemma 10. Then the R-SAGD and SAGD algorithms are equivalent.*

Now we establish the main result of this section; convergence for convex and μ-strongly-convex functions will follow almost immediately in Sections 5.3 and 5.4. The proof relies on the expected decrease condition given by Lemma 14, which holds for L-smooth f when $(f, \mathcal{O})$ satisfies strong growth. Intuitively, this condition shows that SGD obtains similar per-

iteration improvement (in expectation) to deterministic gradient descent if strong growth holds. Finally, note that the expectations in the lemma below are taken with respect to the entire sequence of random variables $(z_t)_{t=0}^k$.

Lemma 12. *Let f be a μ-strongly-convex, L-smooth function (with $\mu = 0$ in the convex case) and $\mathcal{O}$ a SFO such that $(f, \mathcal{O})$ satisfies strong growth with parameter ρ. If $\phi_0^* = f(w_0)$ and $\gamma_0 \geq 0$ is independent of the random process (z_k), then for all $k \in \mathcal{N}$ R-SAGD satisfies*

$$\mathbb{E}[\inf_w \phi_k(w)] = \mathbb{E}\left[\phi_k^*\right] \geq \mathbb{E}[f(w_k)],$$

See Section D.1 for proof.

5.3 Convergence for Strongly-Convex Functions

It is now straightforward to prove R-SAGD converges at an accelerated rate. Earlier, we showed that Lemma 12 and the fact that (λ_k) and (ϕ_k) are estimating sequences for f together imply

$$f(w_k) - f(w^*) \leq \lambda_k \left(\phi_0(w^*) - f(w^*)\right) = \lambda_k \left(f(w_0) - f(w^*) + \frac{\gamma_0}{2}\|w_0 - w^*\|^2\right).$$

This is *almost* an explicit convergence rate. The last step of the analysis is to select an appropriate value for γ_0 and analyze convergence of the λ_k sequence, which is done in the following theorem.

Theorem 12. *Let f be a μ-strongly-convex, L-smooth function with $\mu > 0$ and $\mathcal{O}$ a SFO such that $(f, \mathcal{O})$ satisfies strong growth with parameter ρ. Moreover, choose $\gamma_0 = \mu$, $v_0 = w_0$, and $\phi_0^* = f(w_0)$. Then R-SAGD converges as*

$$\mathbb{E}\left[f(w_K)\right] - f(w^*) \leq \left(1 - \sqrt{\frac{\mu}{\rho L}}\right)^K \left(f(w_0) - f(w^*) + \frac{\mu}{2}\|w^* - w_0\|^2\right).$$

See Section D.2 for proof.

Accelerated convergence of SAGD is an immediate corollary of the equivalence of R-SAGD and SAGD (Lemma 11 and Theorem 12). The only catch here is that α_0 must be selected to correspond to $\gamma_0 = \mu$. It is easy to see that choosing $\alpha_0 = \sqrt{\frac{\mu}{\rho L}}$ for SAGD is identical to $\gamma_0 = \mu$ in R-SAGD since $\alpha_0^2 = \left((1 - \alpha_0)\gamma_0 + \alpha_0\mu\right)/(\rho L)$.

Corollary 2. *Let f be a μ-strongly-convex, L-smooth function with $\mu > 0$ and $\mathcal{O}$ a SFO such that $(f, \mathcal{O})$ satisfies strong growth with parameter ρ. Moreover, choose $\alpha_0 = \sqrt{\frac{\mu}{\rho L}}$. Then SAGD converges as*

$$\mathbb{E}\left[f(w_K)\right] - f(w^*) \leq \left(1 - \sqrt{\frac{\mu}{\rho L}}\right)^K \left(f(w_0) - f(w^*) + \frac{\mu}{2}\|w^* - w_0\|^2\right).$$

Corollary 2 looks like a potential positive answer to the problem of whether or not acceleration is possible in the interpolation setting. SAGD's $O\left(\exp\left\{-\sqrt{\frac{\mu}{\rho L}}\,K\right\}\right)$ convergence certainly improves on that of constant step-size SGD. Moreover, the form of improvement — taking the square-root of the condition number — is identical to that obtained by AGD over deterministic gradient descent (Nesterov, 2004). In this sense, SAGD is a true accelerated algorithm.

However, SAGD is, in general, not optimal for L-smooth, strongly-convex functions. For example, if $\mathcal{O}$ is $L_{\max}$ individually-smooth, then $\rho \leq \frac{L_{\max}}{\mu}$ and the worst-case complexity is $O\left(\exp\left\{-\sqrt{\mu^2/L_{\max}L}\,K\right\}\right)$. This is roughly equivalent to gradient descent, implying SAGD is not accelerated. Overall, we conclude SAGD is an accelerated method only relative to the complexity of other stochastic algorithms under interpolation.

The analysis of SAGD for strongly-convex functions presented here is strictly tighter than prior work by Vaswani et al. (2019a, Theorem 1), who also studied convergence under strong growth. Table 5.1 compares the two complexities; the major difference is that ρ is squared in the rate obtain by Vaswani et al. Recalling $\rho = {}^{L_{\max}}/_{\mu}$ in the worst-case implies that our result is faster by a multiple of the condition number for individually-smooth $\mathcal{O}$ satisfying minimizer interpolation, which is comparable to the improvement obtained by AGD over deterministic gradient descent. This tighter rate addresses the criticism of Liu and Belkin (2020), who showed that the convergence speed shown by Vaswani et al. could be slower than SGD in some circumstances.

5.4 Convergence for Convex Functions

Our final result for SAGD addresses convergence for convex functions under the strong growth condition. The following theorem shows that R-SAGD obtains an accelerated $O\left(\frac{\rho L}{K^2}\right)$ complexity. Invoking the correspondence between R-SAGD and SAGD leads an identical result for the latter algorithm as a corollary. This improves on prior analyses by a factor of $\sqrt{\rho}$ (Vaswani et al., 2019a, Theorem 2). See Table 5.1 for additional details.

Assumptions	Convergence Rate	
	Ours	VBS
μ-SC	$O\left(\sqrt{\frac{\rho L}{\mu}}\log\left(\frac{1}{\epsilon}\right)\right)$	$O\left(\sqrt{\frac{\rho^2 L}{\mu}}\log\left(\frac{1}{\epsilon}\right)\right)$
Convex	$O\left(\sqrt{\frac{\rho L}{\epsilon}}\right)$	$O\left(\sqrt{\frac{\rho^2 L}{\epsilon}}\right)$

Table 5.1: Comparison of iteration complexities for stochastic acceleration schemes under the strong growth condition. For each result, we report the complexity obtained with the optimal step-size and momentum parameter according to the analysis. Our results are tighter than VBS (Vaswani et al., 2019a) by a factor of $\sqrt{\rho}$. Liu and Belkin (2020) also consider stochastic acceleration under interpolation, but make substantially different assumptions about f and thus obtain difficult-to-compare rates. Accordingly, we omit their algorithm, MaSS, from this table.

Theorem 13. *Let f be a convex, L-smooth function and $\mathcal{O}$ a SFO such that $(f, \mathcal{O})$ satisfies strong growth with parameter ρ. Moreover, choose $\gamma_0 = 2\rho L$, $v_0 = w_0$, and $\phi_0^* = f(w_0)$. Then R-SAGD converges as*

$$\mathbb{E}\left[f(w_K)\right] - f(w^*) \leq \frac{2}{(K+1)^2}\left(f(w_0) - f(w^*) + \rho L\|w_0 - w^*\|^2\right).$$

The proof is given in Section D.3. Again, we have the following corollary from the equivalence of R-SAGD and SAGD. Note that the choice of $\alpha_0 = \sqrt{2} - 1$ for SAGD is identical to the choice of $\gamma_0 = 2\rho L$ in R-SAGD, since $\alpha_0^2 = (1 - \alpha_0)\gamma_0/(\rho L)$.

Corollary 3. *Let f be a convex, L-smooth function and $\mathcal{O}$ a SFO such that $(f, \mathcal{O})$ satisfies strong growth with parameter ρ. Moreover, choose $\alpha_0 = \sqrt{2} - 1$. Then R-SAGD converges as*

$$\mathbb{E}\left[f(w_K)\right] - f(w^*) \leq \frac{2}{(K+1)^2}\left(f(w_0) - f(w^*) + \rho L\|w_0 - w^*\|^2\right).$$

5.5 Acceleration under Weak Growth

A disappointing aspect of Theorem 13 is that it uses the strong growth condition, rather than weak growth. Recall that sufficient conditions for strong growth are minimizer

interpolation, individual-smoothness, and the PL condition (Lemma 7). Thus, requiring strong growth in this case has the *flavour* of requiring f to be strongly-convex, but only leads to sub-linear convergence. This is quite disappointing, but it is not obvious that the assumption can be relaxed.

The natural condition for convex functions used throughout this work is the weak growth condition. The main issue using weak growth in Theorem 13 is that we lose access to the progress condition

$$\mathbb{E}_{z_k}\left[f(w_{k+1})\right] \leq f(w_k) - \eta_k\left(1 - \frac{\rho L \eta}{2}\right)\|\nabla f(w_k)\|^2,$$

from Lemma 14, which was a critical piece of our analysis for SAGD. The direct proof using estimating sequences collapses without obvious recourse.

5.6 Conclusions

This chapter has tackled the challenging question of stochastic Nesterov acceleration under interpolation-type conditions. Chapters 3 and 4 showed that SGD converges nearly as quickly as deterministic gradient descent when minimizer interpolation is satisfied. These positive results suggested that stochastic acceleration might also be possible in a similarly restricted setting. Following Vaswani et al. (2019a), we investigated a straightforward modification of AGD to use stochastic gradients. However, unlike Vaswani et al. (2019a), we develop an argument based on estimating sequences (Nesterov, 2004). This allowed us to cast SAGD as an alternating minimization procedure on f and a sequence of curvature estimates, ϕ_k. Careful analysis of this procedure showed that SAGD obtains accelerated rates of convergence in both the strongly-convex and convex settings (Theorems 12 and 13) when strong growth holds. Our theorems improve over Vaswani et al. (2019a) by a factor of $\sqrt{\rho}$, which is equal to $\sqrt{\frac{L_{\max}}{\mu}}$ in the worst case.

Chapter 6

Beyond Interpolation

So far we have investigated the performance of stochastic gradient methods for unconstrained minimization of an L-smooth function f given access to a SFO $\mathcal{O}$ such that $(f, \mathcal{O})$ satisfies interpolation or a growth condition. This chapter extends these analyses to L_2-regularized minimization of f,

$$\min_{w \in \mathbb{R}^d} F(w) := f(w) + \frac{\lambda}{2}\|w\|^2, \tag{6.1}$$

where $\lambda > 0$ is a tuning parameter which controls the degree of regularization. To maintain consistency with common notation, w^* will refer to the global minimizer of this regularized problem, while minimizers of f will now be denoted as w^+.

The canonical SFO for L_2-regularized minimization evaluates the regularization term exactly, giving the following stochastic function and gradient evaluations:

$$F(w, z_k) = f(w, z_k) + \frac{\lambda}{2}\|w\|^2 \qquad \nabla F(w, z_k) = \nabla f(w, z_k) + \lambda w.$$

We call this the L_2-regularization of $\mathcal{O}$ and denote it as $\mathcal{O}_2$. Except in degenerate circumstances,[5] the pair $(F, \mathcal{O}_2)$ will *not* satisfy interpolation or weak/strong growth even when $(f, \mathcal{O})$ does. This means the analyses from previous chapters cannot be applied directly to regularized problems.

There is still significant reason to believe the L_2-regularization of f can be minimized

[5] For example, when f is minimized at $w^* = 0$.

efficiently when $(f, \mathcal{O})$ satisfy the weak growth condition. For example, $f(w^*) - f(w^+)$ will be small when $\lambda \ll \mu$ and so we might expect $(F, \mathcal{O}_2)$ to approximately satisfy weak growth. On the other hand, $\mathcal{O}_2$ will become nearly deterministic as $\lambda \to \infty$ and $\frac{\lambda}{2}\|w\|^2$ dominates f. Intuition indicates that it is only when λ is moderate that minimizing F will be challenging. The next lemma formalizes these observations into a version of weak growth which holds for $(F, \mathcal{O}_2)$.

Lemma 13. *Let f be an L-smooth function with at least one finite minimizer w^+ and $\mathcal{O}$ a SFO such that $(f, \mathcal{O})$ satisfies the weak growth condition with parameter ρ. Then the L_2-regularized problem $(F, \mathcal{O}_2)$ satisfies the following inequality for all $w \in \mathbb{R}^d$ and $k \geq 0$:*

$$\mathbb{E}_{z_k}\left[\|\nabla F(w, z_k)\|^2\right] \leq 4\max\{\rho L, \lambda\}\left(F(w) - F(w^*) - \frac{\lambda - L}{2}\|w^* - w^+\|^2 + \frac{\lambda}{2}\|w^+\|^2\right),$$

where w^* minimizes the regularized function F. Moreover, this can be improved to

$$\mathbb{E}_{z_k}\left[\|\nabla F(w, z_k)\|^2\right] \leq 4\max\{\rho L, \lambda\}\left(F(w) - F(w^*) - \frac{\lambda + \mu}{2}\|w^* - w^+\|^2 + \frac{\lambda}{2}\|w^+\|^2\right),$$

if f is μ-strongly-convex (with $\mu = 0$ giving the convex case).

See Appendix E for proof.

6.1 Convergence for L_2-Regularized Convex Functions

Let us use Lemma 13 to show that SGD with a constant step-size converges linearly to a *neighborhood* of w^* when f is convex or strongly-convex. Approximate convergence to a region around w^* is the best that can be obtained without an averaging scheme or decreasing step-size due to the irreducible $\frac{\lambda}{2}\|w^+\|^2 + \frac{\lambda+\mu}{2}\|w^* - w^+\|^2$ term in the noise-bound. The following analysis assumes μ-strong-convexity, but permits $\mu = 0$ in order to capture convex functions. In the case that $\mu = 0$, we are still able to obtain linear convergence because the regularized function F is $\mu + \lambda$-strongly-convex and $L + \lambda$-smooth. Proof of the following theorem is given in Section E.1.

Theorem 14. *Let f be a μ-strongly-convex (with $\mu = 0$ in the convex case), L-smooth function with at least one finite minimizer w^+ and $\mathcal{O}$ a SFO such that the pair $(f, \mathcal{O})$ satisfies the weak growth condition with constant ρ. Then stochastic gradient descent with constant step-size $\eta \leq \frac{\mu+\lambda}{\max\{\rho L,\lambda\}((\mu+\lambda)+(L+\lambda))}$ obtains the following convergence rate for the L_2-regularized problem $(F, \mathcal{O}_2)$:*

$$\mathbb{E}\left[\|w_K - w^*\|^2\right] \leq \left(1 - \frac{2\eta(\mu+\lambda)(L+\lambda)}{(\mu+\lambda)+(L+\lambda)}\right)^K \|w_0 - w^*\|^2 + \gamma\lambda\|w^+\|^2 - \gamma\left(\lambda+\mu\right)\|w^* - w^+\|^2,$$

where $\gamma = \frac{\eta \max\{\rho L,\lambda\}[(\mu+\lambda)+(L+\lambda)]}{(\mu+\lambda)(L+\lambda)}$.

Theorem 14 shows that the size of the neighborhood to which SGD converges is determined by the degree of regularization, which fits with the intuition from the previous section. We make the following observations about the effect of λ on the region of convergence: (i) the minimizer w^* depends on λ, with $w^* = w^+$ when $\lambda = 0$ — it is easy to see that the neighborhood terms collapse to 0 and convergence mirrors Theorem 1 when $\lambda = 0$ holds; (ii) as $\lambda \to \infty$, $w^* \to 0$ and the neighborhood again collapses to 0; and (iii) when f is convex and η attains is maximal value, the convergence rate of SGD has the special case

$$\mathbb{E}\left[\|w_K - w^*\|^2\right] \leq \left(1 - \frac{2\lambda^2(L+\lambda)}{\max\left\{\rho L,\lambda\right\}\left(L+2\lambda\right)^2}\right)^K \|w_0 - w^*\|^2 + \frac{\lambda}{(L+\lambda)}\|w^+\|^2 - \frac{\lambda}{(L+\lambda)}\|w^* - w^+\|^2.$$

Crudely assuming $\frac{\lambda}{\lambda+L} \approx 1$ shows that the volume of the neighborhood largely depends on the convergence of w^* to 0 as λ grows.

6.2 Conclusions

This chapter characterized the complexity of SGD for L_2-regularized problems under interpolation. Theorem 14 is thus a small but promising step towards the analysis of general structural minimization problems. For example, consider generalizing Equation 6.1 to

$$F(w) = f(w) + g(w),$$

where f is an L-smooth function with SFO $\mathcal{O}$ such that $(f, \mathcal{O})$ satisfies interpolation and g is a potentially non-smooth regularizer or indicator function. This problem class includes L_1-regularized minimization, as well as minimization with (convex) constraints. An interesting avenue of research is extending the analysis above to this setting through the proximal-gradient method. Although Cevher and Vu (2019) considered such problems, they considered convergence under strong growth plus noise only and did not derive specific expressions for the neighbourhood size.

Chapter 7

Conclusion

This thesis develops the convergence theory of stochastic first-order methods under interpolation conditions. Unlike existing work which is confined to the finite-sum setting, we propose a highly general model for interpolation that applies to black-box objective functions and stochastic first-order oracles (SFOs). The subsequent discussion relates interpolation to the strong and weak growth conditions. Specifically, we provide upper-bounds on the weak and strong growth parameters for interpolating oracles that satisfy an intuitive Lipschitz-smoothness property. Informally, this means that *global* regularity of the stochastic gradients is guaranteed from a *local* correspondence between the oracle and objective as long as the stochastic gradient mapping is smooth in the model parameters. Such a property is satisfied for smooth supervised learning problems.

The theoretical analysis of first-order methods focuses on stochastic gradient descent (SGD) and a stochastic version of Nesterov's accelerated gradient descent. For SGD, we analyze the iteration complexity with a fixed step-size as well as with a stochastic Armijo line-search. In both cases, our results are tighter than existing convergence rates and allow for a wider range of parameters. For convex functions, we recover the best-known convergence rates in the deterministic setting both with and without the Armijo line-search. In all cases, specific care is taken to understand the impact of additional assumptions, such as smoothness or convexity of the oracle.

Our study of stochastic accelerated gradient descent (SAGD) uses the estimating sequences framework developed by Nesterov (2004). Estimating sequences allow us to cast SAGD as alternating descent procedure on the objective and a sequence of improving lo-

cal approximations, avoiding the "analytical tricks" used in many accelerated convergence proofs. Moreover, we are able to derive tighter convergence rates than existing work using this framework. In the case of smooth, interpolating oracles, the improvement in iteration complexity is comparable to dividing by the square-root of the condition number (Vaswani et al., 2019a).

Although this thesis attempts to be a comprehensive analysis of stochastic gradient methods under interpolation, the scope is often too narrow and several important methods are not studied. We briefly discuss the role of growth conditions in structural optimization in Chapter 6. The emphasis falls on L_2-regularized problems where the regularized function satisfies weak growth. In this limited setting, we derive linear convergence to a neighbourhood of the optimal solution and precisely characterize the volume in terms of the regularization parameter. The far more general and interesting problem of composite smooth/non-smooth optimization where the smooth component satisfies interpolation remains completely unaddressed. Deriving exact convergence neighborhoods for proximal-gradient methods in this setting is an interesting open problem.

Open problems in optimization under interpolation are not limited to the largely unexplored area of structural optimization. Fundamental questions *not* answered in this thesis or in the literature include the following:

1. Is there a finite-time convergence rate for the final iterate generated by constant step-size SGD when the objective is convex and weak growth is satisfied?

2. If so, can this rate be extended to SGD with the stochastic Armijo line-search?

3. Does SAGD obtain an accelerated convergence rate for convex functions under the weak growth condition?

4. Can analyses for the stochastic Armijo line-searches be extended to SAGD?

5. Can tighter convergence rates be established for SAGD when additional assumptions are made on the oracle (e.g. smoothness or convexity) as they can for SGD?

Many of these questions build on the problems tackled in this thesis, but require new tools or insights to answer. The analyses contributed here are merely a first step towards a larger understanding of optimization under interpolation.

Bibliography

A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Trans. Inf. Theory*, 58(5):3235–3249, 2012.

Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *J. Mach. Learn. Res.*, 18:221:1–221:51, 2017. URL http://jmlr.org/papers/v18/16-410.html.

Z. Allen-Zhu. Katyusha X: Practical momentum method for stochastic sum-of-nonconvex optimization. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 179–185. PMLR, 2018.

Z. Allen Zhu and L. Orecchia. Linear Coupling: An ultimate unification of gradient and mirror descent. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017*, volume 67 of *LIPIcs*, pages 3:1–3:22. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

L. B. Almeida, T. Langlois, J. D. Amaral, and A. Plakhov. Parameter adaptation in stochastic optimization. *On-Line Learning in Neural Networks, Publications of the Newton Institute*, pages 111–134, 1998.

Y. Arjevani and O. Shamir. On the iteration complexity of oblivious first-order optimization algorithms. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 908–916, 2016.

Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

L. Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966.

S. Arora, N. Cohen, and E. Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of the 35th International Conference on*

Machine Learning, ICML 2018, volume 80 of *Proceedings of Machine Learning Research*, pages 244–253. PMLR, 2018.

M. Assran and M. Rabbat. On the convergence of Nesterov's accelerated gradient method in stochastic settings. *arXiv preprint arXiv:2002.12414*, 2020.

M. Assran, N. Loizou, N. Ballas, and M. Rabbat. Stochastic gradient push for distributed deep learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 344–353. PMLR, 2019.

M. Avriel and D. J. Wilde. Golden block search for the maximum of unimodal functions. *Management Science*, 14(5):307–319, 1968.

R. Bassily, M. Belkin, and S. Ma. On exponential convergence of SGD in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

A. G. Baydin, R. Cornish, D. Martínez-Rubio, M. Schmidt, and F. Wood. Online learning rate adaptation with hypergradient descent. In *6th International Conference on Learning Representations, ICLR 2018*. OpenReview.net, 2018.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019a.

M. Belkin, A. Rakhlin, and A. B. Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, volume 89 of *Proceedings of Machine Learning Research*, pages 1611–1619. PMLR, 2019b.

A. Ben-Israel and B. Mond. What is invexity? *The ANZIAM Journal*, 28(1):1–9, 1986.

Y. Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade - Second Edition*, volume 7700 of *Lecture Notes in Computer Science*, pages 437–478. Springer, 2012.

L. Berrada, A. Zisserman, and M. P. Kumar. Training neural networks for and by interpolation. *arXiv preprint arXiv:1906.05661*, 2019.

D. P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, 2010(1-38):3, 2011.

D. P. Bertsekas and J. N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

L. Bottou. *Une approche théorique de l'apprentissage connexionniste et applications à la reconnaissance de la parole*. PhD thesis, 1991.

S. Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3-4):231–357, 2015.

R. H. Byrd, G. M. Chin, J. Nocedal, and Y. Wu. Sample size selection in optimization methods for machine learning. *Math. Program.*, 134(1):127–155, 2012.

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, pages 1–50, 2019.

V. Cevher and B. C. Vu. On the linear convergence of the stochastic gradient method with constant step-size. *Optim. Lett.*, 13(5):1177–1187, 2019.

Y.-L. Chen and M. Kolar. Understanding accelerated stochastic gradient descent via the growth condition. *arXiv preprint arXiv:2006.06782*, 2020.

D. Choi, C. J. Shallue, Z. Nado, J. Lee, C. J. Maddison, and G. E. Dahl. On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv:1910.05446*, 2019.

E. Çınlar. *Probability and stochastics*, volume 261. Springer Science & Business Media, 2011.

M. Cohen, J. Diakonikolas, and L. Orecchia. On acceleration with noise-corrupted gradients. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1018–1027. PMLR, 2018.

A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM J. Optim.*, 19 (3):1171–1183, 2008.

S. De, A. Yadav, D. Jacobs, and T. Goldstein. Big batch SGD: Automated inference using adaptive batch sizes. *arXiv preprint arXiv:1610.05792*, 2016.

A. Defazio. A simple practical accelerated method for finite sums. In *Advances in Neural Information Processing Systems 29: NeurIPS 2016*, pages 676–684, 2016.

A. Defazio and L. Bottou. On the ineffectiveness of variance reduced optimization for deep learning. In *Advances in Neural Information Processing Systems 32: NeurIPS 2019*, pages 1753–1763, 2019.

A. Defazio, F. R. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems 27: NeurIPS 2014*, pages 1646–1654, 2014.

O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75, 2014.

Y. Drori and O. Shamir. The complexity of finding stationary points with stochastic gradient descent. *arXiv preprint arXiv:1910.01845*, 2019.

J. C. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, 2011.

S. Fridovich-Keil and B. Recht. Choosing the step size: Intuitive line search algorithms with efficient convergence. In *The 11th Workshop on Optimization for Machine Learning (OPT 2019)*, 2019.

M. P. Friedlander and M. Schmidt. Hybrid deterministic-stochastic methods for data fitting. *SIAM J. Scientific Computing*, 34(3), 2012.

R. Frostig, R. Ge, S. M. Kakade, and A. Sidford. Un-regularizing: Approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2540–2548, 2015.

S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM J. Optim.*, 22(4):1469–1492, 2012.

S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: Shrinking procedures and optimal algorithms. *SIAM J. Optim.*, 23(4):2061–2089, 2013.

R. B. Grosse and R. Salakhutdinov. Scaling up natural gradient by sparsely factorizing the inverse Fisher matrix. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2304–2313, 2015.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition.* Springer Series in Statistics. Springer, 2009.

J. Honorio. Convergence rates of biased stochastic optimization for learning sparse Ising models. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*. icml.cc / Omnipress, 2012.

C. Hu, J. T. Kwok, and W. Pan. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems 22: NeurIPS 2009*, pages 781–789. Curran Associates, Inc., 2009.

P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference On Learning Theory, COLT 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 545–604. PMLR, 2018.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26: NeurIPS 2013*, pages 315–323, 2013.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016*, volume 9851 of *Lecture Notes in Computer Science*, pages 795–811. Springer, 2016.

A. Khaled and P. Richtárik. Better theory for SGD in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

D. Kovalev, S. Horváth, and P. Richtárik. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Algorithmic Learning Theory, ALT 2020*, volume 117 of *Proceedings of Machine Learning Research*, pages 451–467. PMLR, 2020.

N. Krejic and N. Krklec. Line search methods with variable sample size for unconstrained optimization. *J. Comput. Appl. Math.*, 245:213–231, 2013.

H. J. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*, volume 35 of *Applications of Mathematics*. Springer, 1997.

G. Lan and Y. Zhou. An optimal randomized incremental gradient method. *Math. Program.*, 171(1-2):167–215, 2018.

N. Le Roux, M. Schmidt, and F. R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems 25: NeurIPS 2012*, pages 2672–2680, 2012.

Y. Lei, T. Hu, G. Li, and K. Tang. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

X. Li and F. Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, volume 89 of *Proceedings of Machine Learning Research*, pages 983–992. PMLR, 2019.

T. Liang and A. Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387*, 2018.

H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems 28: NeurIPS 2015*, pages 3384–3392, 2015.

C. Liu and M. Belkin. Accelerating SGD with momentum for over-parameterized learning. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020.

N. Loizou, S. Vaswani, I. Laradji, and S. Lacoste-Julien. Stochastic Polyak step-size for SGD: An adaptive learning rate for fast convergence. *arXiv preprint arXiv:2002.10542*, 2020.

L. Luo, Y. Xiong, Y. Liu, and X. Sun. Adaptive gradient methods with dynamic bound of learning rate. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019.

S. Ma, R. Bassily, and M. Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3331–3340. PMLR, 2018.

M. Mahsereci and P. Hennig. Probabilistic line searches for stochastic optimization. *J. Mach. Learn. Res.*, 18:119:1–119:59, 2017.

S. Y. Meng, S. Vaswani, I. H. Laradji, M. Schmidt, and S. Lacoste-Julien. Fast and furious convergence: Stochastic second order methods under interpolation. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020*, volume 108 of *Proceedings of Machine Learning Research*, pages 1375–1386. PMLR, 2020.

A. S. Nemirovsky and Y. E. Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(2):21–30, 1985.

A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics, 1983.

Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. In *Doklady an USSR*, volume 269, pages 543–547, 1983.

Y. Nesterov. On an approach to the construction of optimal methods of minimization of smooth convex functions. *Ekonomika i Mateaticheskie Metody*, 24(3):509–517, 1988.

Y. E. Nesterov. *Introductory Lectures on Convex Optimization - A Basic Course*, volume 87 of *Applied Optimization*. Springer, 2004.

J. Neveu. *Discrete-parameter martingales*, volume 10. Elsevier, 1975.

L. M. Nguyen, P. H. Nguyen, M. van Dijk, P. Richtárik, K. Scheinberg, and M. Takác. SGD and Hogwild! convergence without the bounded gradients assumption. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3747–3755. PMLR, 2018.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 1999.

A. Ogaltsov, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and V. Spokoiny. Adaptive gradient descent for convex and non-convex stochastic optimization. *arXiv preprint arXiv:1911.08380*, 2019.

F. Orabona and T. Tommasi. Training deep networks without learning rates through coin betting. In *Advances in Neural Information Processing Systems 30: NeurIPS 2017*, pages 2160–2170, 2017.

C. Paquette and K. Scheinberg. A stochastic line search method with expected complexity analysis. *SIAM J. Optim.*, 30(1):349–376, 2020.

V. Plagianakos, G. Magoulas, and M. Vrahatis. Learning rate adaptation in stochastic gradient descent. In *Advances in convex analysis and global optimization*, pages 433–444. Springer, 2001.

B. Polyak and Y. Z. Tsypkin. Pseudogradient adaptation and training algorithms. *Automation and Remote Control*, 34:45–67, 1973.

X. Qian, P. Richtárik, R. M. Gower, A. Sailanbayev, N. Loizou, and E. Shulgin. SGD with arbitrary sampling: General analysis and improved rates. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209. PMLR, 2019.

M. Raginsky and A. Rakhlin. Information-based complexity, feedback and dynamics in convex programming. *IEEE Trans. Inf. Theory*, 57(10):7036–7056, 2011.

H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22 (3):400–407, 09 1951.

M. Rolinek and G. Martius. L4: practical loss-based step-size adaptation for deep learning. In *Advances in Neural Information Processing Systems 31: NeurIPS 2018*, pages 6434–6444, 2018.

R. E. Schapire, Y. Freund, P. Barlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997)*, pages 322–330. Morgan Kaufmann, 1997.

T. Schaul, S. Zhang, and Y. LeCun. No more pesky learning rates. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 343–351, 2013.

M. Schmidt and N. Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

M. Schmidt, N. Le Roux, and F. R. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems 24: NeurIPS 2011*, pages 1458–1466, 2011.

N. N. Schraudolph. Local gain adaptation in stochastic gradient descent. 1999.

S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 64–72, 2014.

F. Shang, L. Jiao, K. Zhou, J. Cheng, Y. Ren, and Y. Jin. ASVRG: Accelerated proximal SVRG. In *Proceedings of The 10th Asian Conference on Machine Learning, ACML 2018*, volume 95 of *Proceedings of Machine Learning Research*, pages 815–830. PMLR, 2018.

S. Shao and P. P. Yip. Rates of convergence of adaptive step-size of stochastic approximation algorithms. *Journal of mathematical analysis and applications*, 244(2):333–347, 2000.

J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 2960–2968, 2012.

M. V. Solodov. Incremental gradient algorithms with stepsizes bounded away from zero. *Comp. Opt. and Appl.*, 11(1):23–35, 1998.

R. S. Sutton. Gain adaptation beats least squares. In *Proceedings of the 7th Yale workshop on adaptive and learning systems*, volume 161168, 1992.

C. Tan, S. Ma, Y. Dai, and Y. Qian. Barzilai-Borwein step size for stochastic gradient descent. In *Advances in Neural Information Processing Systems 29: NeurIPS 2016*, pages 685–693, 2016.

J. Tang, M. Golbabaee, F. Bach, and M. E. Davies. Rest-Katyusha: Exploiting the solution's structure via scheduled restart schemes. In *Advances in Neural Information Processing Systems 31: NeurIPS 2018*, pages 427–438, 2018.

T. Tieleman and G. Hinton. Lecture 6.5-RMSProp: Divide the gradient by a running average of its recent magnitude. *Coursera: Neural networks for machine learning*, 2012.

P. Tseng. An incremental gradient(-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8(2):506–531, 1998.

S. Vaswani, F. Bach, and M. W. Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019*, volume 89 of *Proceedings of Machine Learning Research*, pages 1195–1204. PMLR, 2019a.

S. Vaswani, A. Mishkin, I. H. Laradji, M. Schmidt, G. Gidel, and S. Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *Advances in Neural Information Processing Systems 32: NeurIPS 2019*, pages 3727–3740, 2019b.

S. Vaswani, F. Kunstner, I. Laradji, S. Y. Meng, M. Schmidt, and S. Lacoste-Julien. Adaptive gradient methods converge faster with over-parameterization (and you can do a line-search). *arXiv preprint arXiv:2006.06835*, 2020.

P. Wolfe. Convergence conditions for ascent methods. *SIAM review*, 11(2):226–235, 1969.

P. Wolfe. Convergence conditions for ascent methods. II: Some corrections. *SIAM review*, 13(2):185–188, 1971.

M. D. Zeiler. AdaDelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net, 2017.

H. Zhang and W. Yin. Gradient methods for convex minimization: Better rates under weaker conditions. *arXiv preprint arXiv:1303.4645*, 2013.

Y. Zhang and X. Lin. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 353–361, 2015.

D. Zou and Q. Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems 32: NeurIPS 2019*, pages 2053–2062, 2019.

Appendix A

Interpolation and Growth Conditions: Proofs

A.1 Stochastic Oracles

Corollary 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex, differentiable function. If there exists an unbiased, L_{max} individually-smooth SFO $\mathcal{O}$ for f, then f is L-smooth with $L \leq L_{max}$. Alternatively, if $\mathcal{O}$ is individually-convex and biased with finite support $\mathcal{Z}$ and finite partial derivatives $\frac{\partial}{\partial w_j} f(w, z)$ for each $z \in \mathcal{Z}$, then $\mathbb{E}_{z_k}[f(\cdot, z_k)]$ is L_{max}-smooth for all k.*

Proof. We first consider when $\mathcal{O}$ is unbiased. In this case, $L_{\max}$ individual smoothness of $\mathcal{O}$ implies

$$\begin{aligned} & f(v, z_k) \leq f(w, z_k) + \langle \nabla f(w, z_k), v - w \rangle + \frac{L_{\max}}{2}\|v - w\|^2 \\ \implies & \mathbb{E}_{z_k}[f(v, z_k)] \leq \mathbb{E}_{z_k}[f(w, z_k)] + \langle \mathbb{E}_{z_k}[\nabla f(w, z_k)], v - w \rangle + \frac{L_{\max}}{2}\|v - w\|^2 \\ \implies & f(v) \leq f(w) + \langle \nabla f(w), v - w \rangle + \frac{L_{\max}}{2}\|v - w\|^2, \end{aligned}$$

Smoothness also gives the lower-bound,

$$\begin{aligned} & f(v, z_k) \geq f(w, z_k) + \langle \nabla f(w, z_k), v - w \rangle - \frac{L_{\max}}{2}\|v - w\|^2 \\ \implies & \mathbb{E}_{z_k}[f(v, z_k)] \geq \mathbb{E}_{z_k}[f(w, z_k)] + \langle \mathbb{E}_{z_k}[\nabla f(w, z_k)], v - w \rangle + \frac{L_{\max}}{2}\|v - w\|^2 \\ \implies & f(v) \geq f(w) + \langle \nabla f(w), v - w \rangle - \frac{L_{\max}}{2}\|v - w\|^2. \end{aligned}$$

We conclude that

$$\frac{L_{\max}}{2}\|v - w\|^2 \geq |f(v) - f(w) - \langle \nabla f(w), v - w \rangle|,$$

which, along with convexity, is a necessary and sufficient condition for f to be $L_{\max}$-smooth (Nesterov, 2004, Theorem 2.1.5).

Now, assume $\mathcal{O}$ is biased, meaning $\mathbb{E}_{z_k}\left[\nabla f(w_k, z_k)\right] \neq \nabla f(w_k)$, but $\mathcal{Z}$ is finite. In this case, $\frac{\partial}{\partial w_j} f(w, z_k) \leq \max_{z \in \mathcal{Z}} |\frac{\partial}{\partial w_j}(w, z)|$ almost surely for each dimension $j \in d$. The maximum of each partial derivative over $\mathcal{Z}$ is finite by assumption and thus $\nabla f(w, z_k)$ is dominated by a constant vector. Starting again from individual smoothness,

$$\begin{aligned} f(v, z_k) &\leq f(w, z_k) + \langle \nabla f(w, z_k), v - w \rangle + \frac{L_{\max}}{2}\|v - w\|^2 \\ \implies \mathbb{E}_{z_k}\left[f(v, z_k)\right] &\leq \mathbb{E}_{z_k}\left[f(w, z_k)\right] + \langle \mathbb{E}_{z_k}\left[\nabla f(w, z_k)\right], v - w \rangle + \frac{L_{\max}}{2}\|v - w\|^2 \\ &\leq \mathbb{E}_{z_k}\left[f(w, z_k)\right] + \langle \nabla \mathbb{E}_{z_k}\left[f(w, z_k)\right], v - w \rangle + \frac{L_{\max}}{2}\|v - w\|^2, \end{aligned}$$

where the order of the gradient and expectation operators was exchanged by appealing to the dominated convergence theorem (Çınlar, 2011, Theorem 4.16). An identical derivation using the lower-bound from smoothness yields

$$\mathbb{E}_{z_k}\left[f(v, z_k)\right] \geq \mathbb{E}_{z_k}\left[f(w, z_k)\right] + \langle \nabla \mathbb{E}_{z_k}\left[f(w, z_k)\right], v - w \rangle - \frac{L_{\max}}{2}\|v - w\|^2,$$

which implies

$$\frac{L_{\max}}{2}\|v - w\|^2 \geq |\mathbb{E}_{z_k}\left[f(v, z_k)\right] - \mathbb{E}_{z_k}\left[f(w, z_k)\right] - \langle \nabla \mathbb{E}_{z_k}\left[f(w, z_k)\right], v - w \rangle|,$$

Convexity of $f(\cdot, z_k)$ point-wise over $\mathcal{Z}$ implies that $\mathbb{E}_{z_k}\left[f(\cdot, z_k)\right]$ is also convex. Thus, $\mathbb{E}_{z_k}\left[f(w, z_k)\right]$ is $L_{\max}$-smooth in w by Lemma 2.1.5 of Nesterov (2004). □

A.2 Interpolation

Definition 3 (Interpolation: Minimizers)**.** *A function-oracle pair $(f, \mathcal{O})$ satisfies minimizer interpolation if for all $z \in \mathcal{Z}$,*

$$f(w') \leq f(w)\ \forall w \in \mathbb{R}^d \implies f(w', z) \leq f(w, z)\ \forall w \in \mathbb{R}^d.$$

Definition 4 (Interpolation: Stationary Points)**.** *A function-oracle pair $(f, \mathcal{O})$ satisfies stationary-point interpolation if for all $z \in \mathcal{Z}$,*

$$\nabla f(w') = 0 \implies \nabla f(w', z) = 0.$$

Definition 5 (Interpolation: Mixed)**.** *A function-oracle pair $(f, \mathcal{O})$ satisfies mixed interpolation if for all $z \in \mathcal{Z}$,*

$$f(w') \leq f(w)\ \forall w \in \mathbb{R}^d \implies \nabla f(w', z) = 0.$$

Lemma 1. *Let $(f, \mathcal{O})$ be an arbitrary function-SFO pair. Then only the following relationships hold between models of interpolation:*

$$\textit{Minimizer Interpolation}\ (Definition\ 3) \implies \textit{Mixed Interpolation}\ (Definition\ 5)$$
$$\textit{and}$$
$$\textit{Stationary-Point Interpolation}\ (Definition\ 4) \implies \textit{Mixed Interpolation}\ (Definition\ 5).$$

However, if f is invex and $\mathcal{O}$ is such that $f(\cdot, z)$ is invex for all $z \in \mathcal{Z}$, then minimizer, stationary-point, and mixed interpolation are equivalent.

Proof.
(**i**) Minimizer Interpolation $\implies$ Mixed Interpolation:
Let $w' \in \arg\min_w f(w)$. The stochastic functions $f(w, z)$ are minimized at w' for all $z \in \mathcal{Z}$ by minimizer interpolation. First-order optimality conditions thus imply $\nabla f(w', z) = 0$, which implies mixed interpolation holds.

(**ii**) Stationary-Point Interpolation $\implies$ Mixed Interpolation:
Let $w' \in \arg\min_w f(w)$. First-order optimality conditions ensure w' is a stationary point f and by stationary-point interpolation, w' is also a stationary point of $f(\cdot, z)$ for all $z \in \mathcal{Z}$. Once again, mixed interpolation holds.

(**iii**) Mixed Interpolation, Stationary-Point Interpolation $\not\implies$ Minimizer Interpolation:
We construct a simple counter example. Define the finite-sum function

$$f(w) = \frac{1}{2}\left(f_1(w) + f_2(w)\right) = \frac{3}{2}w^2 - \frac{1}{2}w^2,$$

and consider the oracle $\mathcal{O}$ which returns

$$f(w, z_k) = f_{z_k}(w) \qquad \nabla f(w_k, z_k) = \nabla f_{z_k}(w),$$

where z_k is supported on $\{1, 2\}$. The stochastic functions f_1 and f_2 are only stationary at the global minimum $f(0) = 0$, but f_2 is maximized at this point. So, both stationary-point and mixed interpolation hold, but minimizer interpolation does not.

(**iv**) Minimizer Interpolation $\not\implies$ Stationary-Point Interpolation:
This also follows from a simple finite-sum function. Consider

$$f(w) = \frac{1}{2}\left(f_1(w) + f_2(w)\right) = -\frac{1}{2}\cos(w) - \frac{1}{2}\cos(\frac{w}{2}),$$

and consider the sub-sampling oracle $\mathcal{O}$ as above. The global minimizers of f are

$$\mathcal{X}_0 = \left\{(-1)^t\,(4\pi)t : t \in \{0, 1, \dots\}\right\}.$$

The stochastic functions $f_1(w)$ and $f_2(w)$ are also globally minimized at every point in $\mathcal{X}_0$, so minimizer interpolation holds. However, stationary-point interpolation does not hold as f has infinitely many stationary points which are not stationary points of f_1 or f_2.

Finally, suppose that f is invex and $\mathcal{O}$ is such that $f(\cdot, z)$ is invex for $z \in \mathcal{Z}$. The equivalence of all three definitions follows immediately, since all stationary points of invex functions are global minima. □

Lemma 2. *Let $(f, \mathcal{O})$ be a function-oracle pair. If $\mathcal{O}$ is unbiased and*

$$f(w, z) \geq f(w^*) \quad \forall w \in \mathbb{R}^d, \forall z \in \mathcal{Z},$$

holds, then $(f, \mathcal{O})$ satisfies minimizer interpolation.

Proof. Since $f(w, z) \geq f(w^*)$, the optimality gap $f(w, z) - f(w^*)$ must be non-negative for all w and z. Using the fact that $\mathcal{O}$ is unbiased,

$$\mathbb{E}_{z_k}\left[f(w^*, z_k) - f(w^*)\right] = f(w^*) - f(w^*) = 0,$$

and thus $f(w^*, z_k) = f(w^*)$ almost surely for all k. This is equivalent to $f(w^*, \cdot) = f(w^*)$ point-wise over $\mathcal{Z}$. So, $f(w, z) \geq f(w^*) = f(w^*, z)$ for all $w \in \mathbb{R}^d$ and w^* is a global minimizer of $f(\cdot, z)$. We conclude $(f, \mathcal{O})$ satisfies minimizer interpolation. □

A.3 Growth Conditions

Lemma 3 (Formulations of Strong Growth)**.** *Let $(f, \mathcal{O})$ be a function-oracle pair. If $(f, \mathcal{O})$ satisfies maximal strong growth, then it also satisfies strong growth. However, strong growth does not imply maximal strong growth for general $\mathcal{O}$.*

Proof. Suppose that $(f, \mathcal{O})$ satisfy maximal strong growth with constant $\rho_{\max}$. Then,

$$\begin{aligned} \|\nabla f(w, z_k)\|^2 &\leq \rho_{\max}\|\nabla f(w)\|^2 \\ \implies \mathbb{E}_{z_k}\left[\|\nabla f(w, z_k)\|^2\right] &\leq \mathbb{E}_{z_k}\left[\rho_{\max}\|\nabla f(w)\|^2\right] \\ &= \rho_{\max}\|\nabla f(w)\|^2, \end{aligned}$$

which completes the forward direction.

Now, let us show there are function-oracle pairs which satisfy strong growth, but not maximal strong growth. Let $f(w) = \frac{1}{2}w^2$ and consider $\mathcal{O}$ such that $z_k \sim \mathcal{N}(1, 1)$, and

$$f(w, z_k) = \frac{z_k}{2}w^2, \quad \nabla f(w, z_k) = z_k \cdot w,$$

for all k. A simple calculation shows that this oracle is unbiased,

$$\begin{aligned}\mathbb{E}_{z_k}\left[f(w,z_k)\right] &= \mathbb{E}_{z_k}\left[\frac{z_k}{2}w^2\right] = \frac{1}{2}w^2 = f(w),\\ \mathbb{E}_{z_k}\left[\nabla f(w,z_k)\right] &= \mathbb{E}_{z_k}\left[z_k \cdot w\right] = w = \nabla f(w).\end{aligned}$$

It is also trivial to verify that strong growth holds with $\rho = 2$,

$$\mathbb{E}_{z_k}\left[\|\nabla f(w,z_k)\|^2\right] = \mathbb{E}_{z_k}\left[z_k^2 w^2\right] = 2w^2.$$

For maximal strong growth to hold, we require $c \in \mathbb{R}$ such that

$$z_k^2 \cdot w^2 \leq c \cdot w^2 \implies z_k^2 \leq c,$$

almost surely. But, $z_k^2 > c$ with non-zero probability for any $c \in \mathbb{R}$, so maximal strong growth does not hold for $(f, \mathcal{O})$. □

Lemma 4. *Let $(f, \mathcal{O})$ be a function-SFO pair satisfying the strong growth condition with constant ρ. Moreover, assume that the support $\mathcal{Z}$ of $\mathcal{O}$ is finite and each z_k admits probability mass function p_k. Then $(f, \mathcal{O})$ also satisfies maximal strong growth.*

Proof. We have

$$\begin{aligned}\rho\|\nabla f(w)\|^2 &\geq \mathbb{E}_{z_k}\left[\|\nabla f(w,z_k)\|^2\right]\\ &= \sum_{z\in\mathcal{Z}} \|\nabla f(w,z)\|^2 p_k(z)\\ &\geq \|\nabla f(w,z)\|^2 p_k(z),\end{aligned}$$

for all $z \in \mathcal{Z}$. For $\tilde{z} \in \mathcal{Z}$ such that $p_k(\tilde{z}) > 0$, we have

$$\begin{aligned}&\frac{\rho}{p_k(\tilde{z})}\|\nabla f(w)\|^2 \geq \|\nabla f(w,\tilde{z})\|^2\\ \implies &\frac{\rho}{p_k^*}\|\nabla f(w)\|^2 \geq \max\left\{\|\nabla f(w,\tilde{z})\|^2 : p(\tilde{z}) > 0\right\},\end{aligned}$$

where $p_k^* = \min\{p_k(\tilde{z}) : p_k(z) > 0\}$. We conclude that maximal strong growth holds with $\rho' = \max_k \frac{\rho}{p_k^*}$. □

Lemma 5. *Let $(f, \mathcal{O})$ be a function-oracle pair satisfying the strong growth condition with parameter ρ. If f is L-smooth, then $(f, \mathcal{O})$ satisfies weak growth with parameter $\alpha \leq 2\rho L$.*

Proof. Lemma 18 implies

$$\|\nabla f(w)\|^2 \leq 2L\left(f(w) - f(w^*)\right).$$

Comining this with the definition of strong growth,

$$\begin{aligned}\mathbb{E}_{z_k}\left[\|\nabla f(w, z_k)\|^2\right] &\leq \rho\|\nabla f(w)\|^2 \\ &\leq 2L\rho\left(f(w) - f(w^*)\right),\end{aligned}$$

which shows that the weak growth condition holds $\alpha = 2L\rho$. □

Lemma 6 (Interpolation and Weak Growth)**.** *Let f be an L-smooth function and $\mathcal{O}$ an L_{max} individually-smooth SFO. If $(f, \mathcal{O})$ satisfies minimizer interpolation, then the pair also satisfies the weak growth condition with parameter $\alpha \leq \frac{L_{max}}{L}$.*

Proof. Starting from $L_{\max}$ individual-smoothness of $\mathcal{O}$,

$$f(u, z_k) \leq f(w, z_k) + \langle \nabla f(w, z_k), u - w\rangle + \frac{L_{\max}}{2}\|u - w\|^2$$

Choosing $u = w - \frac{1}{L_{\max}}\nabla f(w, z_k)$,

$$\begin{aligned}f(u, z_k) &\leq f(w, z_k) - \frac{1}{L_{\max}}\langle \nabla f(w, z_k), \nabla f(w, z_k)\rangle + \frac{L_{\max}}{2L_{\max}^2}\|\nabla f(w, z_k)\|^2 \\ &= f(w, z_k) - \frac{1}{2L_{\max}}\|\nabla f(w, z_k)\|^2.\end{aligned}$$

Noting that $f(u, z_k) \geq f(w^*, z_k)$ by minimizer interpolation and taking expectations with respect to z_k gives the following:

$$\begin{aligned}f(w^*, z_k) &\leq f(w, z_k) - \frac{1}{2L_{\max}}\|\nabla f(w, z_k)\|^2 \\ \implies \mathbb{E}_{z_k}\left[f(w^*, z_k)\right] &\leq f(w) - \frac{1}{2L_{\max}}\mathbb{E}_{z_k}\left[\|\nabla f(w, z_k)\|^2\right] \\ \implies f(w^*) &\leq f(w) - \frac{1}{2L_{\max}}\mathbb{E}_{z_k}\left[\|\nabla f(w, z_k)\|^2\right] \\ \implies \mathbb{E}_{z_k}\left[\|\nabla f(w, z_k)\|^2\right] &\leq 2L_{\max}\left(f(w^*) - f(w)\right) \\ &= 2\left(\frac{L_{\max}}{L}\right)L\left(f(w^*) - f(w)\right).\end{aligned}$$

We conclude that weak growth holds with $\alpha \leq \frac{L_{\max}}{L}$. □

Lemma 7 (Interpolation and Strong Growth)**.** *Let f be a L-smooth μ-Polyak-Łojasiewicz (PL) function and $\mathcal{O}$ an L_{max} individually-smooth SFO. If $(f, \mathcal{O})$ satisfies minimizer interpolation, then the pair also satisfies the strong growth condition with parameter $\rho \leq \frac{L_{max}}{\mu}$.*

Proof. Lemma 6 implies that f satisfies the weak growth condition with parameter

$$\alpha \leq \frac{L_{\max}}{L}.$$

Vaswani et al. (2019a, Proposition 1) now implies that f satisfies strong growth with parameter

$$\rho \leq \frac{\alpha L}{\mu} \leq \frac{L_{\max}}{\mu}.$$

This concludes the proof. □

Appendix B

Stochastic Gradient Descent: Proofs

B.1 Convergence for Strongly Convex Functions

Lemma 14. *Let f be an L-smooth function satisfying the strong growth condition with parameter ρ. Then stochastic gradient descent with fixed step-size η satisfies the following expected decrease condition:*

$$\mathbb{E}_{z_k}\left[f(w_{k+1})\right] \leq f(w_k) - \eta\left(1 - \frac{\rho L \eta}{2}\right)\|\nabla f(w_k)\|^2.$$

Proof. Starting from L-smoothness of f,

$$\begin{aligned} f(w_{k+1}) &\leq f(w_k) + \langle \nabla f(w_k), w_{k+1} - w_k\rangle + \frac{L}{2}\|w_{k+1} - w_k\|^2 \\ &= f(w_k) - \eta\left\langle \nabla f(w_k), \nabla f(w_k, z_k)\right\rangle + \frac{L\eta^2}{2}\|\nabla f(w_k, z_k)\|^2 \end{aligned}$$

Taking expectations with respect to z_k:

$$\begin{aligned} \implies \mathbb{E}_{z_k}[f(w_{k+1})] &\leq f(w_k) - \eta\left\langle \mathbb{E}_{z_k}\left[\nabla f(w_k, z_k)\right], \nabla f(w_k)\right\rangle + \frac{L\eta^2}{2}\mathbb{E}_{z_k}\left[\|\nabla f(w_k, z_k)\|^2\right] \\ &= f(w_k) - \eta\|\nabla f(w_k)\|^2 + \frac{L\eta^2}{2}\mathbb{E}_{z_k}\left[\|\nabla f(w_k, z_k)\|^2\right]. \end{aligned}$$

Using the strong growth condition,

$$\begin{aligned} \implies \mathbb{E}_{z_k}\left[f(w_{k+1})\right] &\leq f(w_k) - \eta\|\nabla f(w_k)\|^2 + \frac{\rho L\eta^2}{2}\|\nabla f(w_k)\|^2 \\ &\leq f(w_k) - \eta\left(1 - \frac{\rho L\eta}{2}\right)\|\nabla f(w_k)\|^2. \end{aligned}$$

□

Theorem 1. *Let f be a μ-strongly-convex, L-smooth function and $\mathcal{O}$ a SFO such that $(f, \mathcal{O})$ satisfies the strong growth condition with parameter ρ. Then stochastic gradient descent with fixed step-size $\eta \leq \frac{2}{\rho(\mu+L)}$ converges as*

$$\mathbb{E}\left[\|w_K - w^*\|^2\right] \leq \left(1 - \frac{2\eta\mu L}{\mu + L}\right)^K \|w_0 - w^*\|^2.$$

Proof.

$$\begin{aligned}\|w_{k+1} - w^*\|^2 &= \|w_k - \eta\nabla f(w_k, z_k) - w^*\|^2 \\ &= \eta^2\|\nabla f(w_k, z_k)\|^2 - 2\eta\left\langle\nabla f(w_k, z_k), w_k - w^*\right\rangle + \|w_k - w^*\|^2.\end{aligned}$$

Taking expectations with respect to z_k,

$$\begin{aligned}\implies \mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] &= \eta^2\mathbb{E}_{z_k}\left[\|\nabla f(w_k, z_k)\|^2\right] - 2\eta\mathbb{E}_{z_k}\left[\langle\nabla f(w_k, z_k), w_k - w^*\rangle\right] + \|w_k - w^*\|^2 \\ &= \eta^2\mathbb{E}_{z_k}\left[\|\nabla f(w_k, z_k)\|^2\right] - 2\eta\left\langle\nabla f(w_k), w_k - w^*\right\rangle + \|w_k - w^*\|^2.\end{aligned}$$

Now we use the strong growth condition to control $\mathbb{E}_{z_k}\left[\|\nabla f(w_k, z_k)\|^2\right]$, which yields

$$\mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] \leq \eta^2\rho\|\nabla f(w_k)\|^2 - 2\eta\left\langle\nabla f(w_k), w_k - w^*\right\rangle + \|w_k - w^*\|^2.$$

Coercivity of the gradient (Lemma 21) implies

$$\begin{aligned}\mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] &\leq \eta^2\rho\|\nabla f(w_k)\|^2 - 2\eta\left(\frac{\mu L}{\mu + L}\|w_k - w^*\|^2 + \frac{1}{\mu + L}\|\nabla f(w_k)\|^2\right) \\ &\qquad + \|w_k - w^*\|^2 \\ &= \eta\left(\eta\rho - \frac{2}{\mu + L}\right)\|\nabla f(w_k)\|^2 + \left(1 - \frac{2\eta\mu L}{\mu + L}\right)\|w_k - w^*\|^2.\end{aligned}$$

If $\eta \leq \frac{2}{\rho(\mu+L)}$ then $\eta\rho - \frac{2}{\mu+L} \leq 0$ and we obtain

$$\mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] \leq \left(1 - \frac{2\eta\mu L}{\mu + L}\right)\|w_k - w^*\|^2.$$

Taking expectations and recursing on this inequality,

$$\implies \mathbb{E}\left[\|w_{k+1} - w^*\|^2\right] \leq \left(1 - \frac{2\eta\mu L}{\mu + L}\right)^{k+1}\|w_0 - w^*\|^2.$$

□

Theorem 2. *Let f be a μ-strongly-convex, L-smooth function and $\mathcal{O}$ an L_{max} individually-smooth and convex SFO such that $(f, \mathcal{O})$ satisfies minimizer interpolation. Then stochastic gradient descent with fixed step-size $\eta < \frac{2}{L_{max}}$ converges as*

$$\mathbb{E}\left[\|w_K - w^*\|^2\right] \leq (1 - \mu\,\eta\,(2 - \eta L_{max}))^K \|w_0 - w^*\|^2.$$

Proof.

$$\begin{aligned}\|w_{k+1} - w^*\|^2 &= \|w_k - \eta \nabla f(w_k, z_k) - w^*\|^2 \\ &= \eta^2 \|\nabla f(w_k, z_k)\|^2 - 2\eta \left\langle \nabla f(w_k, z_k), w_k - w^* \right\rangle + \|w_k - w^*\|^2.\end{aligned}$$

Recall $f(\cdot, z_k)$ is convex, $L_{\max}$-smooth, and $\nabla f(w^*, z_k) = 0$ by interpolation. By Lemma 20, we have

$$\|\nabla f(w_k, z_k)\|^2 \leq L_{\max} \left\langle \nabla f(w_k, z_k), w_k - w^* \right\rangle.$$

Applying this to the above yields

$$\begin{aligned}\|w_{k+1} - w^*\|^2 &\leq \eta^2 L_{\max} \left\langle \nabla f(w_k, z_k), w_k - w^* \right\rangle - 2\eta \left\langle \nabla f(w_k, z_k), w_k - w^* \right\rangle + \|w_k - w^*\|^2 \\ \|w_{k+1} - w^*\|^2 &= -\eta\,(2 - \eta L_{\max}) \left\langle \nabla f(w_k, z_k), w_k - w^* \right\rangle + \|w_k - w^*\|^2.\end{aligned}$$

Taking expectations with respect to z_k:

$$\begin{aligned}\implies \mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] &\leq -\eta\,(2 - \eta L_{\max})\,\mathbb{E}_{z_k}\left[\langle \nabla f(w_k, zk), w_k - w^* \rangle\right] + \|w_k - w^*\|^2 \\ &= -\eta\,(2 - \eta L_{\max}) \left\langle \nabla f(w_k), w_k - w^* \right\rangle + \|w_k - w^*\|^2.\end{aligned}$$

If $\eta \leq \frac{2}{L_{\max}}$, then $2 - \eta L_{\max} \geq 0$. Combing this with $\langle \nabla f(w_k), w_k - w^* \rangle \geq \mu \|w_k - w^*\|^2$ by Lemma 22 allows us to obtain the following:

$$\begin{aligned}\mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] &\leq -\mu\,\eta\,(2 - \eta L_{\max}) \|w_k - w^*\|^2 + \|w_k - w^*\|^2 \\ &= (1 - \mu\,\eta\,(2 - \eta L_{\max})) \|w_k - w^*\|^2.\end{aligned}$$

Taking expectations and recursing on this inequality yields the final result,

$$\implies \mathbb{E}\left[\|w_{k+1} - w^*\|\right]^2 \leq (1 - \mu\,\eta\,(2 - \eta L_{\max}))^{k+1} \|w_0 - w^*\|^2.$$

□

Theorem 3. *Let f be a μ-strongly-convex, L-smooth function and $\mathcal{O}$ an L_{max} individually-smooth and μ_{max}-strongly-convex SFO such that $(f, \mathcal{O})$ satisfies minimizer interpolation. Then stochastic gradient descent with fixed step-size $\eta \leq \frac{2}{\mu_{max}+L_{max}}$ converges almost surely at the rate*

$$\|w_K - w^*\|^2 \leq (1 - 2\eta\,\delta_{min})^K \|w_0 - w^*\|^2,$$

where $\delta_{min} = \min_{z\in\mathcal{Z}} \frac{\mu_z L_z}{\mu_z + L_z}$.

Proof.

$$\begin{aligned}\|w_{k+1} - w^*\|^2 &= \|w_k - \eta\nabla f(w_k, z_k) - w^*\|^2 \\ &= \eta^2\|\nabla f(w_k, z_k)\|^2 - 2\eta\,\langle \nabla f(w_k, z_k), w_k - w^*\rangle + \|w_k - w^*\|^2.\end{aligned}$$

Coercivity of the stochastic gradient (Lemma 21) implies

$$\begin{aligned}\|w_{k+1} - w^*\|^2 \leq \eta^2\|\nabla f(w_k, z_k)\|^2 - 2\eta\left(\frac{\mu_z L_z}{\mu_z + L_z}\|w_k - w^*\|^2 + \frac{1}{\mu_z + L_z}\|\nabla f(w_k, z_k)\|^2\right) \\ + \|w_k - w^*\|^2\end{aligned}$$

Let $\delta_{\min} = \min_{z\in\mathcal{Z}} \frac{\mu_z L_z}{\mu_z + L_z}$. Then we have

$$\begin{aligned}\|w_{k+1} - w^*\|^2 &\leq \eta^2\|\nabla f(w_k, z_k)\|^2 - 2\eta\left(\delta_{\min}\|w_k - w^*\|^2 + \frac{1}{\mu_{\max} + L_{\max}}\|\nabla f(w_k, z_k)\|^2\right) \\ &\quad + \|w_k - w^*\|^2 \\ &= \eta\left(\eta - \frac{2}{\mu_{\max} + L_{\max}}\right)\|\nabla f(w_k, z_k)\|^2 + (1 - 2\eta\,\delta_{\min})\|w_k - w^*\|^2.\end{aligned}$$

If $\eta \leq \frac{2}{(\mu_{\max}+L_{\max})}$ then $\eta - \frac{2}{\mu_{\max}+L_{\max}} \leq 0$ and we obtain

$$\|w_{k+1} - w^*\|^2 \leq (1 - 2\eta\,\delta_{\min})\|w_k - w^*\|^2.$$

Recursing on this inequality,

$$\implies \|w_{k+1} - w^*\|^2 \leq (1 - 2\eta\,\delta_{\min})^{k+1}\|w_0 - w^*\|^2.$$

□

B.2 Convergence for Convex Functions

Lemma 15. *Let f be a convex, L-smooth function and $\mathcal{O}$ an L_{max} individually-smooth SFO such that $(f, \mathcal{O})$ satisfies minimizer interpolation. Then stochastic gradient descent with step-size η satisfies the following inequality:*

$$f(w_k) - f(w^*) \leq \frac{1}{2\eta\,\delta}\left(\|w_k - w^*\|^2 - \mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right]\right),$$

where $\delta = \min\left\{1, 1 + \alpha L\left(\frac{1}{L_{max}} - \eta\right)\right\}$ and $w^* = \Pi_{\mathcal{X}^*}(w_0)$. Furthermore, if $\eta \leq \frac{1}{L_{max}}$, then

$$f(w_k) - f(w^*) \leq \frac{1}{2\eta}\left(\|w_k - w^*\|^2 - \mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right]\right).$$

Proof. Let $w^* = \Pi_{\mathcal{X}^*}(w_0)$. We have

$$\begin{aligned}\|w_{k+1} - w^*\|^2 &= \|w_k - \eta_k \nabla f(w_k, z_k) + -w^*\|^2 \\ &= \eta^2\|\nabla f(w_k, z_k)\|^2 - 2\eta\,\langle \nabla f(w_k, z_k), w_k - w^*\rangle + \|w_k - w^*\|^2.\end{aligned}$$

The weak growth condition implies $\nabla f(w^*, z) = 0$ for all $z \in \mathcal{Z}$. Lemma 19 at w_k and w^* states

$$f(w_k, z_k) - f(w^*, z_k) \leq \langle \nabla f(w_k, z_k), w_k - w^*\rangle - \frac{1}{2L_{\max}}\|\nabla f(w_k, z_k)\|^2.$$

Substituting this into the above gives

$$\begin{aligned}\|w_{k+1} - w^*\|^2 &\leq \eta^2\|\nabla f(w_k, z_k)\|^2 - 2\eta\left(f(w_k, z_k) - f(w^*, z_k) + \frac{1}{2L_{\max}}\|\nabla f(w_k, z_k)\|^2\right) \\ &\qquad + \|w_k - w^*\|^2 \\ &\leq \left(\eta^2 - \frac{\eta}{L_{\max}}\right)\|\nabla f(w_k, z_k)\|^2 - 2\eta\left(f(w_k, z_k) - f(w^*, z_k)\right) + \|w_k - w^*\|^2.\end{aligned}$$

Taking expectations with respect to z_k:

$$\begin{aligned}\mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] &\leq \left(\eta^2 - \frac{\eta}{L_{\max}}\right)\mathbb{E}_{z_k}\left[\|\nabla f(w_k, z_k)\|^2\right] - 2\eta\mathbb{E}_{z_k}\left[f(w_k, z_k) - f(w^*, z_k)\right] \\ &\qquad + \|w_k - w^*\|^2, \\ &\leq \left(\eta^2 - \frac{\eta}{L_{\max}}\right)\mathbb{E}_{z_k}\left[\|\nabla f(w_k, z_k)\|^2\right] - 2\eta\left(f(w_k) - f(w^*)\right) + \|w_k - w^*\|^2.\end{aligned}$$

Case 1: If $\eta \leq \frac{1}{L_{\max}}$ then $\left(\eta^2 - \frac{\eta}{L_{\max}}\right)\left(f(w_k, z_k) - f(w^*, z_k)\right) \leq 0$ by interpolation and we obtain

$$\begin{aligned}&\mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] \leq -2\eta\left(f(w_k) - f(w^*)\right) + \|w_k - w^*\|^2 \\ &\implies f(w_k) - f(w^*) \leq \frac{1}{2\eta}\left[\|w_k - w^*\|^2 - \mathbb{E}_{z_k}\|w_{k+1} - w^*\|^2\right].\end{aligned}$$

Case 2: If $\eta > \frac{1}{L_{\max}}$ then $\eta^2 - \frac{\eta}{L_{\max}} > 0$ and the weak growth condition implies

$$\begin{aligned}\mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] &\leq 2\eta\alpha L\left(\eta - \frac{1}{L_{\max}}\right)(f(w_k) - f(w^*)) - 2\eta\left(f(w_k) - f(w^*)\right) + \|w_k - w^*\|^2 \\ &= -2\eta\left(1 + \alpha L\left(\frac{1}{L_{\max}} - \eta\right)\right)(f(w_k) - f(w^*)) + \|w_k - w^*\|^2.\end{aligned}$$

If $\eta < \frac{1}{\alpha L} + \frac{1}{L_{\max}}$, then $1 + \alpha L\left(\frac{1}{L_{\max}} - \eta\right) > 0$,

$$\implies f(w_k) - f(w^*) \leq \frac{1}{2\eta\left(1 + \alpha L\left(\frac{1}{L_{\max}} - \eta\right)\right)}\left(\|w_k - w^*\|^2 - \mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right]\right).$$

Let us combine the cases by taking the worst-case bound. Let $\delta = \min\left\{1, 1 + \alpha L\left(\frac{1}{L_{\max}} - \eta\right)\right\}$ to obtain:

$$f(w_k) - f(w^*) \leq \frac{1}{2\eta\,\delta}\left(\|w_k - w^*\|^2 - \mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right]\right).$$

Note that this bound is tight with Case 1 since $1 + \alpha L\left(\frac{1}{L_{\max}} - \eta\right) \geq 1$ when $\eta \leq \frac{1}{L_{\max}}$. □

Theorem 4. *Let f be a convex, L-smooth function and $\mathcal{O}$ a SFO such that $(f, \mathcal{O})$ satisfies the weak growth condition with parameter α. Then stochastic gradient descent with fixed step-size $\eta < \frac{1}{\alpha L}$ converges as*

$$\mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{1}{2\eta\left(1 - \eta\alpha L\right)K}\|w_0 - w^*\|^2,$$

where $\bar{w}_K = \frac{1}{K}\sum_{k=0}^{K-1} w_k$ and $w^* = \Pi_{\mathcal{X}^*}(w_0)$.

Proof.

$$\begin{aligned}\|w_{k+1} - w^*\|^2 &= \|w_k - \eta\nabla f(w_k, z_k) - w^*\|^2 \\ &= \eta^2\|\nabla f(w_k, z_k)\|^2 - 2\eta\left\langle\nabla f(w_k, z_k), w_k - w^*\right\rangle + \|w_k - w^*\|^2.\end{aligned}$$

Taking expectations with respect to z_k,

$$\begin{aligned}\mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] &= \eta^2\mathbb{E}_{z_k}\left[\|\nabla f(w_k, z_k)\|^2\right] - 2\eta\mathbb{E}_{z_k}\left[\left\langle\nabla f(w_k, z_k), w_k - w^*\right\rangle\right] + \|w_k - w^*\|^2 \\ &= \eta^2\mathbb{E}_{z_k}\left[\|\nabla f(w_k, z_k)\|^2\right] - 2\eta\left\langle\nabla f(w_k), w_k - w^*\right\rangle + \|w_k - w^*\|^2.\end{aligned}$$

By convexity of f and the weak growth condition,

$$\begin{aligned}\mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] &\leq \eta^2\mathbb{E}_{z_k}\left[\|\nabla f(w_k, z_k)\|^2\right] - 2\eta\left(f(w_k) - f(w^*)\right) + \|w_k - w^*\|^2 \\ &\leq 2\eta^2\alpha L\left(f(w_k) - f(w^*)\right) - 2\eta\left(f(w_k) - f(w^*)\right) + \|w_k - w^*\|^2 \\ &= -2\eta\left(1 - \eta\alpha L\right)\left(f(w_k) - f(w^*)\right) + \|w_k - w^*\|^2. \qquad \text{(B.1)}\end{aligned}$$

Rearranging the expression to put the optimality gap on the left-hand side,

$$2\eta\left(1-\eta\alpha L\right)\left(f(w_k)-f(w^*)\right) \leq \|w_k - w^*\|^2 - \mathbb{E}_{z_k}\left[\|w_{k+1}-w^*\|^2\right].$$

If $\eta < \frac{1}{\alpha L}$ then $1-\eta\alpha L > 0$,

$$\implies f(w_k) - f(w^*) \leq \frac{1}{2\eta\left(1-\eta\alpha L\right)}\left(\|w_k - w^*\|^2 - \mathbb{E}_{z_k}\left[\|w_{k+1}-w^*\|^2\right]\right).$$

Taking expectations and summing from $k=0$ to $K-1$ now gives

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[f(w_k)\right] - f(w^*) &\leq \frac{1}{2\eta\left(1-\eta\alpha L\right)\,K}\sum_{k=0}^{K-1}\left(\mathbb{E}\left[\|w_k-w^*\|^2\right] - \mathbb{E}\left[\|w_{k+1}-w^*\|^2\right]\right) \\
&= \frac{1}{2\eta\left(1-\eta\alpha L\right)\,K}\left(\|w_0-w^*\|^2 - \mathbb{E}\left[\|w_K-w^*\|^2\right]\right) \\
&\leq \frac{1}{2\eta\left(1-\eta\alpha L\right)\,K}\|w_0-w^*\|^2.
\end{aligned}$$

Noting $\frac{1}{K}\sum_{k=0}^{K-1} f(w_k) \geq f(\bar{w}_K)$ by convexity leads to the final result,

$$\mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{1}{2\eta\left(1-\eta\alpha L\right)\,K}\|w_0-w^*\|^2.$$

□

Theorem 5. *Let f be a convex, L-smooth function and $\mathcal{O}$ a SFO such that $(f,\mathcal{O})$ satisfies the weak growth condition with parameter α. Moreover, suppose $\mathcal{O}$ is L_{max} individually-smooth. Then stochastic gradient descent with fixed step-size $\eta < \frac{1}{\alpha L} + \frac{1}{L_{max}}$ converges as*

$$\mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{1}{2\eta\,\delta\,K}\|w_0-w^*\|^2,$$

where $\bar{w}_K = \frac{1}{K}\sum_{k=0}^{K-1} w_k$, $w^* = \Pi_{\mathcal{X}^*}(w_0)$, and $\delta = \min\left\{1, 1+\alpha L\left(\frac{1}{L_{max}}-\eta\right)\right\}$.

Proof. Starting from Lemma 15,

$$f(w_k) - f(w^*) \leq \frac{1}{2\eta\,C}\left(\|w_k-w^*\|^2 - \mathbb{E}_{z_k}\left[\|w_{k+1}-w^*\|^2\right]\right)$$

Taking expectations and summing from $k=0$ to $K-1$ now gives

$$\begin{aligned}
\implies \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[f(w_k)-f(w^*)\right] &\leq \frac{1}{2\eta\,\delta\,K}\sum_{k=0}^{K-1}\left(\mathbb{E}\left[\|w_k-w^*\|^2\right] - \mathbb{E}\left[\|w_{k+1}-w^*\|^2\right]\right) \\
&= \frac{1}{2\eta\,\delta\,K}\left(\|w_0-w^*\|^2 - \mathbb{E}\left[\|w_K-w^*\|^2\right]\right) \\
&\leq \frac{1}{2\eta\,\delta\,K}\|w_0-w^*\|^2
\end{aligned}$$

Noting $\frac{1}{K}\sum_{k=0}^{K-1} f(w_k) \geq f(\bar{w}_K)$ by convexity leads to the final result,

$$\implies \mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{1}{2\eta\,\delta\,K}\|w_0 - w^*\|^2.$$

□

B.3 Almost Sure Convergence

Theorem 7. *Let f be a convex, L-smooth function with at least one finite minimizer and $\mathcal{O}$ an L_{max} individually-smooth SFO such that $(f, \mathcal{O})$ satisfies the weak growth condition with parameter α. Then the sequence $(f(w_k))$ generated by stochastic gradient descent with fixed step-size $\eta < \frac{1}{\alpha L} + \frac{1}{L_{max}}$ converges to the optimal function value $f(w^*)$ almost surely.*

Proof. Lemma 15 gives the decrease condition

$$\mathbb{E}\left[\|w_{k+1} - w^*\|^2 \mid \mathcal{F}_k\right] \leq \|w_k - w^*\|^2 - 2\eta\delta\left(f(w_k) - f(w^*)\right),$$

where $\delta \geq 1$ since $\eta < \frac{1}{\alpha L} + \frac{1}{L_{\max}}$ The conditions of Theorem 6 are satisfied with $A_k = 0$ for all k and the sequence $\left(\|w_k - w^*\|^2\right)$ converges to a non-negative random variable $\lim_{k\to\infty}\|w_k - w^*\|^2$ almost surely. The theorem also guarantees

$$\begin{aligned} &\sum_{k=0}^{\infty} 2\eta\delta\left(f(w_k) - f(w^*)\right) < \infty \\ \implies &\sum_{k=0}^{\infty}\left(f(w_k) - f(w^*)\right) < \infty \\ \implies &\lim_{k\to\infty} f(w_k) \overset{\text{a.s.}}{=} f(w^*). \end{aligned} \tag{B.2}$$

That is, stochastic gradient descent converges to the optimal function value. □

Theorem 8. *Let f be an L-smooth function with at least one finite minimizer w^* and $\mathcal{O}$ a SFO such that $(f, \mathcal{O})$ satisfies the strong growth condition with parameter ρ. Then the sequence of gradient-norms $\left(\|\nabla f(w_k)\|^2\right)$ generated by stochastic gradient descent with fixed step-size $\eta < \frac{2}{\rho L}$ converges to 0 almost surely.*

Proof. Lemma 14 gives the decrease condition

$$\mathbb{E}\left[f(w_{k+1}) - f(w^*) \mid \mathcal{F}_k\right] \leq \left(f(w_k) - f(w^*)\right) - \eta\left(1 - \frac{\rho L\eta}{2}\right)\|\nabla f(w_k)\|^2.$$

Since $\eta < \frac{2}{\rho L}$,

$$\eta\left(1 - \frac{\eta\rho L}{2}\right)\|\nabla f(w_k)\|^2 > 0,$$

and the conditions of Theorem 6 are satisfied with $A_k = 0$ for all k. The sequence $(f(w_k) - f(w^*))$ converges to a non-negative random variable. Of more interest is that

$$\sum_{k=0}^{\infty} \eta \left(1 - \frac{\eta\rho L}{2}\right) \|\nabla f(w_k)\|^2 < \infty$$
$$\implies \sum_{k=0}^{\infty} \|\nabla f(w_k)\|^2 < \infty, \tag{B.3}$$

almost surely. Accordingly, the sequence of gradient norms satisfies

$$\lim_{k\to\infty} \|\nabla f(w_k)\|^2 \overset{\text{a.s.}}{=} 0,$$

and we conclude that the sequence of gradients converges to a stationary point.

□

B.4 Additional Lemmas

Lemma 16. *Consider the setting of Theorem 5 from Vaswani et al. (2019a): f is μ-strongly-convex, L-smooth, and $(f, \mathcal{O})$ satisfies weak growth with parameter α. Then Theorem 5 can be modified to allow for a maximum step-size of $\eta \leq \frac{\mu+L}{\alpha L^2}$ and the following convergence rate:*

$$\mathbb{E}\left[\|w_{K+1} - w^*\|^2\right] \leq \left(1 - \mu\,\eta + \max\left\{\eta\mu\left(\eta\alpha L - 1\right), \eta L\left(\eta\alpha L - 1\right)\right\}\right)^K \|w_0 - w^*\|^2.$$

Moreover, this is tight with the original result when $\eta = \frac{1}{\alpha L}$.

Proof. Starting from the seventh line of the proof of Theorem 5 from Vaswani et al. (2019a),

$$\mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] \leq 2\eta\left(\eta\alpha L - 1\right)\left(f(w_k) - f(w^*)\right) + \left(1 - \mu\,\eta\right)\|w_k - w^*\|^2.$$

Case 1: $\eta \leq \frac{1}{\alpha L}$. Then $2\eta\left(\eta\alpha L - 1\right)\left(f(w_k) - f(w^*)\right) \leq 0$ and Lemma 18 implies

$$\mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] \leq \eta\mu\left(\eta\alpha L - 1\right)\|w_k - w^*\|^2 + \left(1 - \mu\,\eta\right)\|w_k - w^*\|^2.$$

Case 2: $\frac{\mu+L}{\alpha L^2} > \eta > \frac{1}{\alpha L}$. Then the other direction of Lemma 18 implies

$$\mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] \leq \eta L\left(\eta\alpha L - 1\right)\|w_k - w^*\|^2 + \left(1 - \mu\,\eta\right)\|w_k - w^*\|^2.$$

Note that the upper bound on η is required to make progress in this case. We combine cases by taking the worst-case bound as follows:

$$\begin{aligned}\implies \mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] &\leq \left(1 - \mu\,\eta\right)\|w_k - w^*\|^2 + \max\left\{\eta\mu\left(\eta\alpha L - 1\right), \eta L\left(\eta\alpha L - 1\right)\right\}\|w_k - w^*\|^2 \\ &= \left(1 - \mu\,\eta + \max\left\{\eta\mu\left(\eta\alpha L - 1\right), \eta L\left(\eta\alpha L - 1\right)\right\}\right)\|w_k - w^*\|^2,\end{aligned}$$

Taking expectations and recursing on the inequality gives the final results,

$$\implies \mathbb{E}\left[\|w_{k+1} - w^*\|^2\right] \leq \left(1 - \mu\,\eta + \max\left\{\eta\mu\left(\eta\alpha L - 1\right), \eta L\left(\eta\alpha L - 1\right)\right\}\right)^{k+1} \|w_0 - w^*\|^2.$$

Tightness with the original result from Vaswani et al. (2019a) is immediate when $\eta = \frac{1}{\alpha L}$ in the last expression above. □

Appendix C

Line Search: Proofs

Lemma 8. *Let f be an L-smooth function and $\mathcal{O}$ an L_{max} individually-smooth SFO such that $(f, \mathcal{O})$ satisfies minimizer interpolation. Then the maximum possible step-size returned by the stochastic Armijo line-search constrained to lie in the $(0, \eta_{max}]$ range satisfies the following inequalities:*

$$\min\left\{\frac{2(1-c)}{L_{max}}, \eta_{max}\right\} \leq \eta_k \leq \frac{f(w_k, z_k) - f(w^*, z_k)}{c\|\nabla f(w_k, z_k)\|^2}.$$

Proof. Let $\tilde{\eta}_k$ be the step-size returned by the exact Armijo line-search. The back-tracking constraint $\eta_k \in (0, \eta_{\max}]$ is equivalent to choosing $\eta_k = \min\{\tilde{\eta}_k, \eta_{\max}\}$. Starting from individual smoothness of $\mathcal{O}$,

$$\begin{aligned} f(w_{k+1}, z_k) &\leq f(w_k, z_k) - \langle \nabla f(w_k, z_k), w_{k+1} - w_k \rangle + \frac{L_{\max}}{2}\|w_{k+1} - w_k\|^2 \\ &= f(w_k, z_k) - \tilde{\eta}_k\|\nabla f(w_k, z_k)\|^2 + \frac{\tilde{\eta}_k^2 L_{\max}}{2}\|\nabla f(w_k, z_k)\|^2 \\ &= f(w_k, z_k) - \tilde{\eta}_k\left(1 - \frac{L_{\max}\tilde{\eta}_k}{2}\right)\|\nabla f(w_k, z_k)\|^2. \end{aligned}$$

This implies that the stochastic Armijo condition,

$$f(w_{k+1}, z_k) \leq f(w_k, z_k) - c \cdot \tilde{\eta}_k\|\nabla f(w_k, z_k)\|^2,$$

is guaranteed to hold whenever

$$\begin{aligned} c\,\tilde{\eta}_k &\leq \tilde{\eta}_k\left(1 - \frac{L_{\max}\tilde{\eta}_k}{2}\right) \\ \implies \tilde{\eta}_k &\leq \frac{2(1-c)}{L_{\max}}. \end{aligned}$$

Recalling $\tilde{\eta}_k$ is the maximal step-size satisfying the Armijo condition leads to the lower-bound $\eta_k \geq \min\left\{\frac{2(1-c)}{L_{\max}}, \eta_{\max}\right\}$.

Let us now upper-bound η_k. The argument above established that the Armijo condition is satisfied

for any step-size $\eta \leq \frac{2(1-c)}{L_{\max}}$. Thus, it must be satisfied for $\eta_k = \min\{\tilde{\eta}_k, \eta_{\max}\}$ and we obtain

$$\begin{aligned} f(w_{k+1}, z_k) &\leq f(w_k, z_k) - c \cdot \eta_k \|\nabla f(w_k, z_k)\|^2 \\ \implies \eta_k &\leq \frac{f(w_k, z_k) - f(w_{k+1}, z_k)}{c\|\nabla f(w_k, z_k)\|^2}. \end{aligned}$$

Minimizer interpolation implies $f(w_{k+1}, z_k) \geq f(w^*, z_k)$ and thus

$$\eta_k \leq \frac{f(w_k, z_k) - f(w^*, z_k)}{c\|\nabla f(w_k, z_k)\|^2}.$$

This completes the proof. □

Lemma 9. *Let f be a convex, L-smooth function and $\mathcal{O}$ an L_{max} individually-smooth and convex SFO such that $(f, \mathcal{O})$ satisfy minimizer interpolation. Then stochastic gradient descent using the Armijo line-search with $c \geq \frac{1}{2}$ satisfies the following inequality:*

$$f(w_k) - f(w^*) \leq \frac{1}{2} \max\left\{\frac{L_{max}}{2(1-c)}, \frac{1}{\eta_{max}}\right\} \left(\|w_k - w^*\|^2 - \mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right]\right),$$

where $w^* \in \mathcal{X}^*$ is arbitrary.

Proof. Let $w^* \in \mathcal{X}^*$ be arbitrary. Then,

$$\begin{aligned} \|w_{k+1} - w^*\|^2 &= \|w_k - \eta_k \nabla f(w_k, z_k) - w^*\|^2 \\ &= \eta_k^2 \|\nabla f(w_k, z_k)\|^2 - 2\eta_k \langle \nabla f(w_k, z_k), w_k - w^* \rangle + \|w_k - w^*\|^2. \end{aligned}$$

Minimizer interpolation implies $\nabla f(w^*, z) = 0$ for all $z \in \mathcal{Z}$. We may thus use Lemma 19 at w_k and w^* to obtain

$$\begin{aligned} \|w_{k+1} - w^*\|^2 &\leq \eta_k^2 \|\nabla f(w_k, z_k)\|^2 - 2\eta_k \left(f(w_k, z_k) - f(w^*, z_k) + \frac{1}{2L_{\max}} \|\nabla f(w_k, z_k)\|^2\right) \\ &\qquad + \|w_k - w^*\|^2 \\ &= \eta_k \left(\eta_k - \frac{1}{L_{\max}}\right) \|\nabla f(w_k, z_k)\|^2 - 2\eta_k \left(f(w_k, z_k) - f(w^*, z_k)\right) \\ &\qquad + \|w_k - w^*\|^2. \end{aligned}$$

Our analysis now proceeds in cases on η_k.

Case 1: $\eta_k \leq \frac{1}{L_{\max}}$. Then $\eta_k \left(\eta_k - \frac{1}{L_{\max}}\right) \|\nabla f(w_k, z_k)\|^2 \leq 0$,

$$\begin{aligned} \implies \|w_{k+1} - w_k\|^2 &\leq -2\eta_k \left(f(w_k, z_k) - f(w^*, z_k)\right) + \|w_k - w^*\|^2 \\ &\leq -2 \min\left\{\frac{2(1-c)}{L_{\max}}, \eta_{\max}\right\} \left(f(w_k) - f(w^*)\right) + \|w_k - w^*\|^2. \quad \text{(Lemma 8)} \end{aligned}$$

Case 2: $\eta_k > \frac{1}{L_{\max}}$. Then $\eta_k \left(\eta_k - \frac{1}{L_{\max}}\right) \|\nabla f(w_k, z_k)\|^2 \geq 0$ and we may apply Lemma 8 to obtain

$$\begin{aligned}\|w_{k+1} - w^*\|^2 &\leq \left(\frac{\eta_k}{c} - \frac{1}{cL_{\max}}\right)(f(w_k, z_k) - f(w^*, z_k)) - 2\eta_k (f(w_k, z_k) - f(w^*, z_k)) \\ &\quad + \|w_k - w^*\|^2 \\ &= \eta_k \left(\frac{1}{c} - 2\right)(f(w_k, z_k) - f(w^*, z_k)) - \frac{1}{cL_{\max}}(f(w_k, z_k) - f(w^*, z_k)) \\ &\quad + \|w_k - w^*\|^2.\end{aligned}$$

If $c \geq \frac{1}{2}$ then $\eta_k(\frac{1}{c} - 2)(f(w_k, z_k) - f(w^*, z_k)) \leq 0$ and

$$\|w_{k+1} - w^*\|^2 \leq -\frac{1}{cL_{\max}}(f(w_k, z_k) - f(w^*, z_k)) + \|w_k - w^*\|^2.$$

We may combine cases by taking the loosest bound as follows:

$$\implies \|w_{k+1} - w^*\|^2 \leq -\min\left\{\frac{4(1-c)}{L_{\max}}, 2\eta_{\max}, \frac{1}{cL_{\max}}\right\}(f(w_k, z_k) - f(w^*, z_k)) + \|w_k - w^*\|^2.$$

Noting $\frac{1}{cL_{\max}} \geq \frac{4(1-c)}{L_{\max}}$ for $c \in [0.5, 1]$ (with equality holding when $c = \frac{1}{2}$) implies

$$\|w_{k+1} - w^*\|^2 \leq -2\min\left\{\frac{2(1-c)}{L_{\max}}, \eta_{\max}\right\}(f(w_k, z_k) - f(w^*, z_k)) + \|w_k - w^*\|^2.$$

Taking expectations with respect to z_k:

$$\begin{aligned}\implies \mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] &\leq -2\min\left\{\frac{2(1-c)}{L_{\max}}, \eta_{\max}\right\}\mathbb{E}_{z_k}\left[f(w_k, z_k) - f(w^*, z_k)\right] + \|w_k - w^*\|^2 \\ &= -2\min\left\{\frac{2(1-c)}{L_{\max}}, \eta_{\max}\right\}(f(w_k) - f(w^*)) + \|w_k - w^*\|^2.\end{aligned}$$

Rearranging the expression so the optimality gap is on the left-hand side completes the proof,

$$\implies f(w_k) - f(w^*) \leq \frac{1}{2}\max\left\{\frac{L_{\max}}{2(1-c)}, \frac{1}{\eta_{\max}}\right\}\left(\|w_k - w^*\|^2 - \mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right]\right).$$

□

C.1 Convergence for Strongly-Convex Functions

Theorem 9. *Let f be a μ-strongly-convex, L-smooth function and $\mathcal{O}$ an L_{max} individually-smooth and convex SFO such that $(f, \mathcal{O})$ satisfies minimizer interpolation. Then stochastic gradient descent using the Armijo line-search with $c \geq \frac{1}{2}$ converges as*

$$\mathbb{E}\left[\|w_K - w^*\|^2\right] \leq \left(1 - \mu \min\left\{\frac{2(1-c)}{L_{max}}, \eta_{max}\right\}\right)^K \|w_0 - w^*\|^2.$$

Proof. Starting from Lemma 9,

$$\begin{aligned} f(w_k) - f(w^*) &\leq \frac{1}{2}\max\left\{\frac{L_{\max}}{2(1-c)}, \frac{1}{\eta_{\max}}\right\}\left(\|w_k - w^*\|^2 - \mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right]\right) \\ \implies \mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] &\leq \|w_k - w^*\|^2 - 2\min\left\{\frac{2(1-c)}{L_{\max}}, \eta_{\max}\right\}\left(f(w_k) - f(w^*)\right) \end{aligned}$$

Strong-convexity implies $f(w_k) - f(w^*) \geq \frac{\mu}{2}\|w_k - w^*\|^2$ (Lemma 17). Using this inequality yields

$$\begin{aligned} \mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] &\leq \|w_k - w^*\|^2 - \mu\min\left\{\frac{2(1-c)}{L_{\max}}, \eta_{\max}\right\}\|w_k - w^*\|^2 \\ &= \left(1 - \mu\min\left\{\frac{2(1-c)}{L_{\max}}, \eta_{\max}\right\}\right)\|w_k - w^*\|^2. \end{aligned}$$

Taking expectations and recursing on the inequality,

$$\implies \mathbb{E}\left[\|w_{k+1} - w^*\|^2\right] \leq \left(1 - \mu\min\left\{\frac{2(1-c)}{L_{\max}}, \eta_{\max}\right\}\right)^{k+1}\|w_0 - w^*\|^2.$$

□

C.2 Convergence for Convex Functions

Theorem 10. *Let f be a convex, L-smooth function and $\mathcal{O}$ an L_{max} individually-smooth and convex SFO such that $(f, \mathcal{O})$ satisfies minimizer interpolation. Then stochastic gradient descent using the Armijo line-search with $c \geq \frac{1}{2}$ converges as*

$$\mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{1}{2\,K}\max\left\{\frac{L_{max}}{2(1-c)}, \frac{1}{\eta_{max}}\right\}\|w_0 - w^*\|^2,$$

where $\bar{w}_K = \sum_{k=0}^{K-1} w_k$ and $w^* = \Pi_{\mathcal{X}^*}(w_0)$.

Proof. Starting from Lemma 9 with $w^* = \Pi_{\mathcal{X}^*}(w_0)$,

$$f(w_k) - f(w^*) \leq \frac{1}{2}\max\left\{\frac{L_{\max}}{2(1-c)}, \frac{1}{\eta_{\max}}\right\}\left(\|w_k - w^*\|^2 - \mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right]\right).$$

Taking expectations and summing from $k = 0$ to $K - 1$,

$$\begin{aligned}
\implies \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[f(w_k)\right] - f(w^*) &\leq \frac{1}{2\,K}\max\left\{\frac{L_{\max}}{2(1-c)}, \frac{1}{\eta_{\max}}\right\}\sum_{k=0}^{K-1}\mathbb{E}\left[\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2\right] \\
&= \frac{1}{2\,K}\max\left\{\frac{L_{\max}}{2(1-c)}, \frac{1}{\eta_{\max}}\right\}\left(\|w_0 - w^*\|^2 - \mathbb{E}\left[\|w_K - w^*\|^2\right]\right) \\
&\leq \frac{1}{2\,K}\max\left\{\frac{L_{\max}}{2(1-c)}, \frac{1}{\eta_{\max}}\right\}\|w_0 - w^*\|^2.
\end{aligned}$$

Noting $\frac{1}{K}\sum_{k=0}^{K-1} f(w_k) \geq f(\bar{w}_K)$ by convexity leads to the final result,

$$\implies \mathbb{E}\left[f(\bar{w}_K)\right] - f(w^*) \leq \frac{1}{2\,K}\max\left\{\frac{L_{\max}}{2(1-c)}, \frac{1}{\eta_{\max}}\right\}\|w_0 - w^*\|^2.$$

□

C.3 Convergence for Non-Convex Functions

Theorem 11. *Let f be an L-smooth function and $\mathcal{O}$ an L_{max} individually-smooth SFO such that $(f, \mathcal{O})$ satisfies the strong growth condition with parameter ρ. Then stochastic gradient descent using the Armijo line-search with $c > 1 - \frac{L_{max}}{\rho L}$ and $\eta_{max} < \frac{2}{\rho L}$ converges as*

$$\min_{k\in[K]}\|\nabla f(w_k)\|^2 \leq \frac{1}{\delta\,K}\left(f(w_0) - f(w^*)\right),$$

where $\delta = \left(\eta_{max} + \frac{2(1-c)}{L_{max}}\right) - \rho\left(\eta_{max} - \frac{2(1-c)}{L_{max}} + L\eta_{max}^2\right)$.

Proof. Firstly, note that for any vectors $a, b \in \mathbb{R}^d$,

$$\begin{aligned}
\|a - b\|^2 &= \|a\|^2 + \|b\|^2 - 2\left\langle a, b\right\rangle \\
\implies -\left\langle a, b\right\rangle &= \frac{1}{2}\left(\|a - b\|^2 - \|a\|^2 - \|b\|^2\right).
\end{aligned} \tag{C.1}$$

Let $\Delta_k = f(w_{k+1}) - f(w_k)$. Starting from L-smoothness of f:

$$\begin{aligned}
\Delta_k &\leq \left\langle\nabla f(w_k), w_{k+1} - w_k\right\rangle + \frac{L}{2}\|w_{k+1} - w_k\|^2 \\
&= -\eta_k\left\langle\nabla f(w_k), \nabla f(w_k, z_k)\right\rangle + \frac{L\eta_k^2}{2}\|\nabla f(w_k, z_k)\|^2.
\end{aligned}$$

Using Equation C.1 on $-\langle \nabla f(w_k), \nabla f(w_k, z_k) \rangle$,

$$\begin{aligned}
\implies \Delta_k &\leq \frac{\eta_k}{2} \left(\|\nabla f(w_k) - \nabla f(w_k, z_k)\|^2 - \|\nabla f(w_k)\|^2 - \|\nabla f(w_k, z_k)\|^2 \right) \\
&\quad + \frac{L\eta_k^2}{2} \|\nabla f(w_k, z_k)\|^2 \\
\implies 2\Delta_k &\leq \eta_k \|\nabla f(w_k) - \nabla f(w_k, z_k)\|^2 - \eta_k \left(\|\nabla f(w_k)\|^2 + \|\nabla f(w_k, z_k)\|^2 \right) \\
&\quad + L\eta_k^2 \|\nabla f(w_k, z_k)\|^2.
\end{aligned}$$

Using $\min\left\{ \frac{2(1-c)}{L_{\max}}, \eta_{\max} \right\} = \eta_{\min} \leq \eta_k \leq \eta_{\max}$ from Lemma 8 and taking expectations with respect to z_k:

$$\begin{aligned}
2\Delta_k &\leq \eta_{\max} \|\nabla f(w_k) - \nabla f(w_k, z_k)\|^2 - \eta_{\min} \left(\|\nabla f(w_k)\|^2 + \|\nabla f(w_k, z_k)\|^2 \right) \\
&\quad + L\eta_{\max}^2 \|\nabla f(w_k, z_k)\|^2 \\
\implies 2\mathbb{E}_{z_k}\left[\Delta_k\right] &\leq \eta_{\max} \mathbb{E}_{z_k}\left[\|\nabla f(w_k) - \nabla f(w_k, z_k)\|^2 \right] - \eta_{\min} \mathbb{E}_{z_k}\left[\|\nabla f(w_k)\|^2 + \|\nabla f(w_k, z_k)\|^2 \right] \\
&\quad + L\eta_{\max}^2 \mathbb{E}_{z_k}\left[\|\nabla f(w_k, z_k)\|^2 \right] \\
&= \eta_{\max} \mathbb{E}\left[\|\nabla f(w_k, z_k)\|^2 \right] - \eta_{\max} \|\nabla f(w_k)\|^2 - \eta_{\min} \mathbb{E}_{z_k}\left[\|\nabla f(w_k)\|^2 + \|\nabla f(w_k, z_k)\|^2 \right] \\
&\quad + L\eta_{\max}^2 \mathbb{E}_{z_k}\left[\|\nabla f(w_k, z_k)\|^2 \right]
\end{aligned}$$

Collecting terms and applying the strong growth condition,

$$\begin{aligned}
2\mathbb{E}_{z_k}\left[\Delta_k\right] &\leq \left(\eta_{\max} - \eta_{\min} + L\eta_{\max}^2 \right) \mathbb{E}_{z_k}\left[\|\nabla f(w_k, z_k)\|^2 \right] - \left(\eta_{\max} + \eta_{\min} \right) \|\nabla f(w_k)\|^2 \\
&\leq \rho \left(\eta_{\max} - \eta_{\min} + L\eta_{\max}^2 \right) \|\nabla f(w_k)\|^2 - \left(\eta_{\max} + \eta_{\min} \right) \|\nabla f(w_k)\|^2 \\
&= - \left(\left(\eta_{\max} + \eta_{\min} \right) - \rho \left(\eta_{\max} - \eta_{\min} + L\eta_{\max}^2 \right) \right) \|\nabla f(w_k)\|^2.
\end{aligned}$$

Assuming $\delta = \left(\eta_{\max} + \eta_{\min} \right) - \rho \left(\eta_{\max} - \eta_{\min} + L\eta_{\max}^2 \right) > 0$,

$$\implies \|\nabla f(w_k)\|^2 \leq \frac{2}{\delta} \mathbb{E}_{z_k}\left[-\Delta_k \right].$$

Taking expectations and summing from $k = 0$ to $K - 1$,

$$\begin{aligned}
\implies \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[\|\nabla f(w_k)\|^2 \right] &\leq \frac{2}{\delta K} \sum_{k=0}^{K-1} \mathbb{E}\left[-\Delta_k \right] \\
&= \frac{2}{\delta K} \sum_{k=0}^{K-1} \mathbb{E}\left[f(w_k) - f(w_{k+1}) \right] \\
&= \frac{2}{\delta K} \mathbb{E}\left[f(w_0) - f(w_{k+1}) \right] \\
&\leq \frac{2}{\delta K} \left(f(w_0) - f(w^*) \right) \\
\implies \min_{k \in [K-1]} \mathbb{E}\left[\|\nabla f(w_k)\|^2 \right] &\leq \frac{2}{\delta K} \left(f(w_0) - f(w^*) \right).
\end{aligned}$$

It remains to show that $\delta > 0$ holds. Our analysis proceeds in cases.

Case 1: $\eta_{\max} \leq \frac{2(1-c)}{L_{\max}}$. Then $\eta_{\min} = \eta_{\max}$ and

$$\begin{aligned}\delta &= (\eta_{\max} + \eta_{\max}) - \rho\left(\eta_{\max} - \eta_{\max} + L\eta_{\max}^2\right) \\ &= 2\eta_{\max} - \rho L \eta_{\max}^2 > 0 \\ \implies \eta_{\max} &< \frac{2}{\rho L}.\end{aligned}$$

Case 2: $\eta_{\max} > \frac{2(1-c)}{L_{\max}}$. Then $\eta_{\min} = \frac{2(1-c)}{L_{\max}}$ and

$$\delta = \left(\eta_{\max} + \frac{2(1-c)}{L_{\max}}\right) - \rho\left(\eta_{\max} - \frac{2(1-c)}{L_{\max}} + L\eta_{\max}^2\right).$$

This is a concave quadratic in $\eta_{\max}$ and is strictly positive when

$$\eta_{\max} \in \left(0, \frac{(1-\rho) + \sqrt{(\rho-1)^2 + {}^{[8\rho(1+\rho)L(1-c)]}/_{L_{\max}}}}{2L\rho}\right).$$

To avoid contradiction with the case assumption $\frac{2(1-c)}{L_{\max}} < \eta_{\max}$, we require

$$\begin{aligned}\frac{(1-\rho) + \sqrt{(\rho-1)^2 + {}^{[8\rho(1+\rho)L(1-c)]}/_{L_{\max}}}}{2L\rho} &> \frac{2(1-c)}{L_{\max}} \\ \implies \frac{8\rho(1+\rho)L(1-c)}{L_{\max}} &> \left(\frac{4L\rho}{L_{\max}} + (\rho-1)\right)^2 - (\rho-1)^2 \\ &= \frac{16L^2\rho^2(1-c)^2}{L_{\max}^2} + \frac{8L\rho(\rho-1)(1-c)}{L_{\max}} \\ \implies \frac{L_{\max}}{\rho L} &> (1-c) \\ \implies c &> 1 - \frac{L_{\max}}{\rho L}.\end{aligned}$$

The line-search requires $c \in (0,1)$. Noting that $\rho \geq 1$ by definition, we have $\frac{L_{\max}}{\rho L} > 0$ as long as $L, L_{\max} > 0$. The Lipschitz constants are strictly positive when f is bounded-below and non-zero. We obtain the non-empty constraint set

$$c \in \left(1 - \frac{L_{\max}}{\rho L}, 1\right).$$

Substituting the maximum value for c into the upper-bound on $\eta_{\max}$ yields a similar requirement,

$$\eta_{\max} \in \left(0, \frac{2}{\rho L}\right).$$

This completes the second case.

Putting the two cases together gives the final constraints on c and $\eta_{\max}$ as

$$c \geq 1 - \frac{L_{\max}}{\rho L} \qquad \eta_{\max} < \frac{2}{\rho L}.$$

We note that the upper and lower bounds on η_k are consistent since

$$\eta_{\min} = \min\left\{\frac{2(1-c)}{L_{\max}}, \eta_{\max}\right\} < \min\left\{\frac{2L_{\max}}{\rho L L_{\max}}, \eta_{\max}\right\} = \max\left\{\frac{2}{\rho L}, \eta_{\max}\right\} \leq \frac{2}{\rho L},$$

where the last inequality follows from the bound on $\eta_{\max}$. In particular, taking $c \to 1$ and $\eta_{\max} \to \frac{2}{\rho L}$ yields an adaptive step-size $\eta_k \in (0, \frac{2}{\rho L})$. □

Appendix D

Acceleration: Proofs

D.1 Estimating Sequences

Lemma 11. *Let (λ_k) and (ϕ_k) be as defined in Lemma 10. Then the reformulated stochastic accelerated gradient descent (R-SAGD) and SAGD algorithms are equivalent.*

Proof. We expand the updated for y_k using the canonical form of ϕ_k as follows:

$$\begin{aligned} y_k &= w_k - \frac{\alpha_k}{\gamma_k + \alpha_k \mu} \nabla \phi_k(w_k) \\ &= w_k - \frac{\alpha_k}{\gamma_k + \alpha_k \mu} \nabla \left(\phi_k^* + \frac{\gamma_k}{2} \|w_k - v_k\|^2 \right) \\ &= w_k - \frac{\alpha_k \gamma_k}{\gamma_k + \alpha_k \mu} (w_k - v_k) \\ &= \left(\frac{(1-\alpha_k)\gamma_k + \alpha_k \mu}{\gamma_k + \alpha_k \mu} \right) w_k + \frac{\alpha_k \gamma_k}{\gamma_k + \alpha_k \mu} v_k. \end{aligned}$$

Recalling $\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k \mu$ gives

$$\begin{aligned} y_k &= \frac{\gamma_{k+1}}{\gamma_k + \alpha_k \mu} w_k + \frac{\alpha_k \gamma_k}{\gamma_k + \alpha_k \mu} v_k \\ &= \frac{1}{\gamma_k + \alpha_k \mu} \left(\gamma_{k+1} w_k + \alpha_k \gamma_k v_k \right). \end{aligned}$$

This is identical to update Step (b) of **Constant Step-Size Scheme I** given by Nesterov (2004, Eq. 2.2.19). The equivalence of **Constant Step-Size Scheme I** and accelerated gradient descent (AGD) (called **Constant Step-Size Scheme II**) is established by Nesterov (2004, Page 92) — albeit with a different step-size for the gradient step — which completes the proof. □

Lemma 12. *Let f be a μ-strongly-convex, L-smooth function (with $\mu = 0$ in the convex case) and $\mathcal{O}$ a SFO such that $(f, \mathcal{O})$ satisfies strong growth with parameter ρ. If $\phi_0^* = f(w_0)$ and $\gamma_0 \geq 0$ is independent of the random process (z_k), then for all $k \in \mathcal{N}$ R-SAGD satisfies*

$$\mathbb{E}[\inf_w \phi_k(w)] = \mathbb{E}\left[\phi_k^*\right] \geq \mathbb{E}[f(w_k)],$$

Proof. First we establish a few preliminaries. The condition $\gamma_0 \perp\!\!\!\perp (z_k)$ implies the $(\gamma_k)_{k=0}^{\infty}$, $(\alpha_k)_{k=0}^{\infty}$, and $(\lambda_k)_{k=0}^{\infty}$ sequences evolve *independently* of the stochastic processes $(y_k)_{k=0}^{\infty}$, $(w_k)_{k=0}^{\infty}$, and $(v_k)_{k=0}^{\infty}$. This will be necessary later in the proof. Next, invoking Lemma 14 with $\eta = \frac{1}{\rho L}$ yields

$$\mathbb{E}_{z_k}\left[f(w_{k+1})\right] \leq f(w_k) - \frac{1}{2\rho L}\|\nabla f(w_k)\|^2.$$

Taking expectations with respect to $z_0, \ldots, z_{k-1}$,

$$\implies \mathbb{E}\left[f(w_{k+1})\right] \leq \mathbb{E}\left[f(w_k)\right] - \frac{1}{2\rho L}\mathbb{E}\left[\|\nabla f(w_k)\|^2\right]. \tag{D.1}$$

Now we move on to the main argument, which proceeds by induction.

The choice of $\phi_0^* = f(x_0)$ ensures $\inf \phi_0(w) = f(w_0)$ deterministically, which is the base case for induction. The inductive assumption is $\mathbb{E}[\inf \phi_k(w)] \geq \mathbb{E}[f(w_k)]$; let us use this to show

$$\mathbb{E}\left[\inf_w \phi_{k+1}(w)\right] \geq \mathbb{E}\left[f(w_{k+1})\right].$$

Recall the explicit form of the minimizer $\inf \phi_{k+1}(w) = \phi_{k+1}^*$ is

$$\begin{aligned}\phi_{k+1}^* = (1-\alpha_k)\phi_k^* + \alpha_k f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(y_k)\|^2 \\ + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y_k - v_k\|^2 + \langle \nabla f(y_k), v_k - y_k\rangle\right)\end{aligned}$$

Taking expectations with respect to $z_0, \ldots, z_k$ and using $\gamma_0 \perp\!\!\!\perp (z_k)$ plus linearity of expectation:

$$\begin{aligned}\mathbb{E}[\phi_{k+1}^*] &= (1-\alpha_k)\mathbb{E}[\phi_k^*] + \mathbb{E}\left[\alpha_k f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(y_k)\|^2\right] \\ &\quad + \mathbb{E}\left[\frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y_k - v_k\|^2 + \langle \nabla f(y_k), v_k - y_k\rangle\right)\right] \\ &\geq (1-\alpha_k)\mathbb{E}\left[f(w_k)\right] + \mathbb{E}\left[\alpha_k f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(y_k)\|^2\right] \\ &\quad + \mathbb{E}\left[\frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y_k - v_k\|^2 + \langle \nabla f(y_k), v_k - y_k\rangle\right)\right],\end{aligned}$$

where the inequality follows from the inductive assumption. Convexity of f implies $f(w_k) \geq f(y_k) + \langle \nabla f(y_k), w_k - y_k \rangle$. Recalling $\frac{\alpha_k^2}{\gamma_{k+1}} = \frac{1}{\rho L}$ (see Figure 5.2) allows us to obtain

$$\begin{aligned}
\mathbb{E}[\phi_{k+1}^*] &\geq (1-\alpha_k)\mathbb{E}[(f(y_k) + \langle \nabla f(y_k), w_k - y_k \rangle)] + \mathbb{E}\left[\alpha_k f(y_k) - \frac{1}{2\rho L}\|\nabla f(y_k)\|^2\right] \\
&\qquad + \mathbb{E}\left[\frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y_k - v_k\|^2 + \langle \nabla f(y_k), v_k - y_k \rangle\right)\right] \\
&= \mathbb{E}\left[(1-\alpha_k)f(y_k) + \alpha_k f(y_k)\right] + (1-\alpha_k)\mathbb{E}[\langle \nabla f(y_k), w_k - y_k \rangle] - \mathbb{E}\left[\frac{1}{2\rho L}\|\nabla f(y_k)\|^2\right] \\
&\qquad + \mathbb{E}\left[\frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y_k - v_k\|^2 + \langle \nabla f(y_k), v_k - y_k \rangle\right)\right] \\
&= \mathbb{E}\left[f(y_k) - \frac{1}{2\rho L}\|\nabla f(y_k)\|^2\right] + (1-\alpha_k)\mathbb{E}\Bigg[\langle \nabla f(y_k), w_k - y_k \rangle \\
&\qquad + \frac{\alpha_k\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y_k - v_k\|^2 + \langle \nabla f(y_k), v_k - y_k \rangle\right)\Bigg]
\end{aligned}$$

Equation D.1 now implies

$$\mathbb{E}[\phi_{k+1}^*] \geq \mathbb{E}\left[f(w_{k+1})\right] + (1-\alpha_k)\mathbb{E}\left[\langle \nabla f(y_k), w_k - y_k \rangle + \frac{\alpha_k\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y_k - v_k\|^2 + \langle \nabla f(y_k), v_k - y_k \rangle\right)\right].$$

The remainder of the proof is largely unchanged from the deterministic case. The definition of R-SAGD gives $w_k - y_k = \frac{\alpha_k}{\gamma_k + \alpha_k\mu}\nabla\phi(w_k)$, which we use to obtain

$$\begin{aligned}
\mathbb{E}\left[\phi_{k+1}^*\right] &\geq \mathbb{E}[f(w_{k+1})] + (1-\alpha_k)\mathbb{E}\Bigg[\frac{\alpha_k}{\gamma_k + \alpha_k\mu}\langle \nabla f(y_k), \nabla\phi_k(w_k) \rangle + \frac{\alpha_k\gamma_k}{\gamma_{k+1}}\Big(\frac{\mu}{2}\|y_k - v_k\|^2 \\
&\qquad + \langle \nabla f(y_k), v_k - y_k \rangle\Big)\Bigg]
\end{aligned}$$

Noting that $\nabla\phi_k(w_k) = \gamma_k(w_k - v_k)$ by Equation 5.1 and $v_k - y_k = \frac{\gamma_{k+1}}{\gamma_k + \alpha_k\mu}(v_k - w_k)$ gives

$$\begin{aligned}
\mathbb{E}\left[\phi_{k+1}^*\right] &\geq \mathbb{E}[f(w_{k+1})] + (1-\alpha_k)\mathbb{E}\Bigg[\frac{\alpha_k\gamma_k}{\gamma_k + \alpha_k\mu}\langle \nabla f(y_k), w_k - v_k \rangle + \frac{\alpha_k\gamma_k}{\gamma_{k+1}}\Big(\frac{\mu}{2}\|y_k - v_k\|^2 \\
&\qquad + \langle \nabla f(y_k), v_k - y_k \rangle\Big)\Bigg] \\
&= \mathbb{E}[f(w_{k+1})] + (1-\alpha_k)\mathbb{E}\Bigg[\frac{\alpha_k\gamma_k}{\gamma_k + \alpha_k\mu}\langle \nabla f(y_k), w_k - v_k \rangle + \frac{\alpha_k\gamma_k}{\gamma_{k+1}}\Big(\frac{\mu}{2}\|y_k - v_k\|^2 \\
&\qquad + \frac{\gamma_{k+1}}{\gamma_k + \alpha_k\mu}\langle \nabla f(y_k), v_k - w_k \rangle\Big)\Bigg] \\
&= \mathbb{E}[f(w_{k+1})] + \frac{\mu\alpha_k(1-\alpha_k)\gamma_k}{2\gamma_{k+1}}\mathbb{E}\left[\|y_k - v_k\|^2\right] \\
&\geq \mathbb{E}[f(w_{k+1})].
\end{aligned}$$

since $\frac{\mu\alpha_k(1-\alpha_k)\gamma_k}{2\gamma_{k+1}} \geq 0$. We conclude that $\mathbb{E}[\inf \phi_k(w)] \geq \mathbb{E}[f(w_k)]$ holds for all $k \in \mathbb{N}$ by induction. □

D.2 Convergence for Strongly-Convex Functions

Theorem 12. *Let f be a μ-strongly-convex, L-smooth function with $\mu > 0$ and $\mathcal{O}$ a SFO such that $(f, \mathcal{O})$ satisfies strong growth with parameter ρ. Moreover, choose $\gamma_0 = \mu$, $v_0 = w_0$, and $\phi_0^* = f(w_0)$. Then R-SAGD converges as*

$$\mathbb{E}\left[f(w_K)\right] - f(w^*) \leq \left(1 - \sqrt{\frac{\mu}{\rho L}}\right)^K \left(f(w_0) - f(w^*) + \frac{\mu}{2}\|w^* - w_0\|^2\right).$$

Proof. We obtain the following inequality immediately:

$$\begin{aligned}
\mathbb{E}\left[f(w_k)\right] &\leq \mathbb{E}[\inf_x \phi_k(x)] && \text{(Lemma 12)} \\
&= \mathbb{E}[\phi_k(w^*)] \\
&\leq (1 - \lambda_k)\, f(w^*) + \lambda_k \phi_0(w^*) && \text{(Lemma 10)} \\
\implies \mathbb{E}\left[f(w_k)\right] - f(w^*) &\leq \lambda_k(\phi_0(w^*) - f(w^*)) \\
&= \lambda_k \left(f(w_0) - f(w^*) + \frac{\mu}{2}\|w_0 - w^*\|^2\right). && \text{(D.2)}
\end{aligned}$$

The convergence rate of R-SAGD is determined by the speed at which λ_k converges to 0, which we shall now derive. The canonical form of ϕ_k (Equation 5.1) gives the following expression for γ_k:

$$\gamma_k = (1 - \alpha_{k-1})\gamma_{k-1} + \alpha_{k-1}\mu.$$

The choice $\gamma_0 = \mu$ now yields

$$\gamma_k = (1 - \alpha_{k-1})\mu + \alpha_{k-1}\mu = \mu,$$

by induction on k. The definition of R-SAGD (Figure 5.2) gives

$$\begin{aligned}
\alpha_k^2 &= \frac{\gamma_{k+1}}{\rho L} = \frac{\mu}{\rho L}, \\
\implies \alpha_k &= \sqrt{\frac{\mu}{\rho L}}.
\end{aligned}$$

Finally, the choice of estimating sequences in Lemma 10 gives $\lambda_0 = 1$ and

$$\begin{aligned}
\lambda_k &= (1 - \alpha_k)\, \lambda_{k-1} \\
&= \left(1 - \sqrt{\frac{\mu}{\rho L}}\right) \lambda_{k-1}.
\end{aligned}$$

for every $k > 0$. Recursing on this inequality gives

$$\begin{aligned}
\lambda_k &= \left(1 - \sqrt{\frac{\mu}{\rho L}}\right)^k \lambda_0 \\
&= \left(1 - \sqrt{\frac{\mu}{\rho L}}\right)^k.
\end{aligned}$$

Substituting this expression into Equation D.2 yields the final result:

$$\mathbb{E}\left[f(w_k)\right] - f(w^*) \leq \left(1 - \sqrt{\frac{\mu}{\rho L}}\right)^k \left(f(w_0) - f(w^*) + \frac{\mu}{2}\|w^* - w_0\|^2\right).$$

□

D.3 Convergence for Convex Functions

Theorem 13. *Let f be a convex, L-smooth function and $\mathcal{O}$ a SFO such that $(f, \mathcal{O})$ satisfies strong growth with parameter ρ. Moreover, choose $\gamma_0 = 2\rho L$, $v_0 = w_0$, and $\phi_0^* = f(w_0)$. Then R-SAGD converges as*

$$\mathbb{E}\left[f(w_K)\right] - f(w^*) \leq \frac{2}{(K+1)^2}\left(f(w_0) - f(w^*) + \rho L\|w_0 - w^*\|^2\right).$$

Proof. We obtain the following inequality immediately:

$$\begin{aligned}
\mathbb{E}\left[f(w_k)\right] &\leq \mathbb{E}[\inf_x \phi_k(x)] && \text{(by Lemma 12)} \\
&\leq \mathbb{E}[\phi_k(w^*)] \\
&\leq (1-\lambda_k)\, f(w^*) + \lambda_k \phi_0(w^*) && \text{(by Definition 9)} \\
\implies \mathbb{E}\left[f(w_k)\right] - f(w^*) &\leq \lambda_k(\phi_0(w^*) - f(w^*)) \\
&= \lambda_k\left(f(w_0) - f(w^*) + \rho L\|w_0 - w^*\|^2\right). && \text{(D.3)}
\end{aligned}$$

We can see that the convergence rate of R-SAGD is controlled by the λ_k sequence. Let us analyze its rate of convergence to 0. The canonical form of ϕ_k (Equation 5.1) gives the following expression for γ_k:

$$\begin{aligned}
\gamma_k &= (1 - \alpha_{k-1})\gamma_{k-1} + \alpha_{k-1}\mu \\
\gamma_k &= (1 - \alpha_{k-1})\gamma_{k-1},
\end{aligned}$$

since $\mu = 0$. The definition of R-SAGD (Figure 5.2) and our choice of estimating sequences (Lemma 10 gives

$$\begin{aligned}
\alpha_k^2 &= \frac{\gamma_{k+1}}{\rho L}, \\
\lambda_k &= (1 - \alpha_k)\lambda_{k-1}.
\end{aligned}$$

Now we must upper-bound λ_k, given that $\lambda_0 = 2\tilde{L}$. Let $\tilde{L} = \rho L$. Invoking Lemma 2.2.4 of Nesterov

(2004) with $\tilde{L}$ implies

$$\begin{aligned}\lambda_k &\leq \frac{4\tilde{L}}{\gamma_0 (k+1)^2} \\ &= \frac{2}{(k+1)^2}.\end{aligned}$$

Substituting this into Equation D.3 yields the final result:

$$\mathbb{E}\left[f(w_k)\right] - f(w^*) \leq \frac{2}{(k+1)^2}\left(f(w_0) - f(w^*) + \rho L \|w_0 - w^*\|^2\right).$$

□

Appendix E

Beyond Interpolation: Proofs

Lemma 13. *Let f be an L-smooth function with at least one finite minimizer w^+ and $\mathcal{O}$ a SFO such that $(f, \mathcal{O})$ satisfies the weak growth condition with parameter ρ. Then the L_2-regularized problem $(F, \mathcal{O}_2)$ satisfies the following inequality for all $w \in \mathbb{R}^d$ and $k \geq 0$:*

$$\mathbb{E}_{z_k}\left[\|\nabla F(w, z_k)\|^2\right] \leq 4\max\{\rho L, \lambda\}\left(F(w) - F(w^*) - \frac{\lambda - L}{2}\|w^* - w^+\|^2 + \frac{\lambda}{2}\|w^+\|^2\right),$$

where w^* minimizes the regularized function F. Moreover, this can be improved to

$$\mathbb{E}_{z_k}\left[\|\nabla F(w, z_k)\|^2\right] \leq 4\max\{\rho L, \lambda\}\left(F(w) - F(w^*) - \frac{\lambda + \mu}{2}\|w^* - w^+\|^2 + \frac{\lambda}{2}\|w^+\|^2\right),$$

if f is μ-strongly-convex (with $\mu = 0$ giving the convex case).

Proof. By definition of the L_2-regularized oracle $\mathcal{O}_2$,

$$\mathbb{E}_{z_k}\left[\|\nabla F(w, z_k)\|^2\right] = \mathbb{E}_{z_k}\left[\|\nabla f(w, z_k) + \lambda w\|^2\right].$$

Young's inequality for products and the weak growth condition imply

$$\begin{aligned}\mathbb{E}_{z_k}\left[\|\nabla F(w, z_k)\|^2\right] &\leq \mathbb{E}_{z_k}\left[2\|\nabla f(w, z_k)\|^2 + 2\lambda^2\|w\|^2\right] \\ &\leq 4\rho L\left(f(w) - f(w^+)\right) + 2\lambda^2\|w\|^2.\end{aligned}$$

We now proceed in cases depending on the degree of regularization. If $\lambda \leq \rho L$,

$$\begin{aligned}\implies \mathbb{E}_{z_k}\left[\|\nabla F(w, z_k)\|^2\right] &\leq 4\rho L\left(f(w) - f(w^+) + \frac{\lambda^2}{2\rho L}\|w\|^2\right) \\ &\leq 4\rho L\left(f(w) - f(w^+) + \frac{\lambda}{2}\|w\|^2\right) \\ &= 4\rho L\left(F(w) - f(w^+)\right).\end{aligned}$$

On the other hand, if $\lambda > \rho L$,

$$\begin{aligned}\implies \mathbb{E}_{z_k}\left[\|\nabla F(w, z_k)\|^2\right] &\leq 4\lambda\left(f(w) - f(w^+) + \frac{\lambda}{2}\|w\|^2\right) \\ &= 4\lambda\left(F(w) - f(w^+)\right).\end{aligned}$$

Combining the two cases yields

$$\mathbb{E}_{z_k}\left[\|\nabla F(w, z_k)\|^2\right] \leq 4\max\left\{\rho L, \lambda\right\}\left(F(w) - f(w^+)\right).$$

Let $\delta = \mu$ if f is μ-strongly-convex and set $\delta = -L$ otherwise. Strong-convexity (or smoothness if $\delta = -L$) implies $f(w^+) \geq f(w^*) + \langle \nabla f(w^*), w^+ - w^* \rangle + \frac{\delta}{2}\|w^+ - w^*\|^2$. Subsituting this into the bound gives

$$\mathbb{E}_{z_k}\left[\|\nabla F(w, z_k)\|^2\right] \leq 4\max\left\{\rho L, \lambda\right\}\left(F(w) - f(w^*) - \left\langle \nabla f(w^*), w^+ - w^* \right\rangle - \frac{\delta}{2}\|w^+ - w^*\|^2\right)$$

Recall that w^* minimizes the regularized function F. First-order optimality conditions imply $\nabla f(w^*) = -\lambda w^*$ and we obtain

$$\begin{aligned}\mathbb{E}_{z_k}\left[\|\nabla F(w, z_k)\|^2\right] &\leq 4\max\left\{\rho L, \lambda\right\}\left(F(w) - f(w^*) + \left\langle \lambda w^*, w^+ - w^* \right\rangle - \frac{\delta}{2}\|w^+ - w^*\|^2\right) \\ &= 4\max\left\{\rho L, \lambda\right\}\left(F(w) - f(w^*) - \lambda\|w^*\|^2 + \lambda\left\langle w^*, w^+ \right\rangle - \frac{\delta}{2}\|w^+ - w^*\|^2\right) \\ &= 4\max\left\{\rho L, \lambda\right\}\left(F(w) - F(w^*) - \frac{\lambda}{2}\|w^*\|^2 + \lambda\left\langle w^*, w^+ \right\rangle - \frac{\delta}{2}\|w^+ - w^*\|^2\right) \\ &= 4\max\left\{\rho L, \lambda\right\}\left(F(w) - F(w^*) - \frac{\lambda + \delta}{2}\|w^* - w^+\|^2 + \frac{\lambda}{2}\|w^+\|^2\right).\end{aligned}$$

Substituting in the appropriate value for δ completes the proof. □

E.1 Convergence for L_2-Regularized Convex Functions

Theorem 14. *Let f be a μ-strongly-convex (with $\mu = 0$ in the convex case), L-smooth function with at least one finite minimizer w^+ and $\mathcal{O}$ a SFO such that the pair $(f, \mathcal{O})$ satisfies the weak growth condition with constant ρ. Then stochastic gradient descent with constant step-size $\eta \leq \frac{\mu+\lambda}{\max\{\rho L,\lambda\}((\mu+\lambda)+(L+\lambda))}$ obtains the following convergence rate for the L_2-regularized problem $(F, \mathcal{O}_2)$:*

$$\mathbb{E}\left[\|w_K - w^*\|^2\right] \leq \left(1 - \frac{2\eta(\mu+\lambda)(L+\lambda)}{(\mu+\lambda)+(L+\lambda)}\right)^K \|w_0 - w^*\|^2 + \gamma\lambda\|w^+\|^2 - \gamma\left(\lambda+\mu\right)\|w^* - w^+\|^2,$$

where $\gamma = \frac{\eta \max\{\rho L,\lambda\}[(\mu+\lambda)+(L+\lambda)]}{(\mu+\lambda)(L+\lambda)}$.

Proof.

$$\begin{aligned}\|w_{k+1} - w^*\|^2 &= \|w_k - \eta\nabla F(w_k, z_k) - w^*\|^2 \\ &= \eta^2\|\nabla F(w_k, z_k)\|^2 - 2\eta\left\langle\nabla F(w_k, z_k), w_k - w^*\right\rangle + \|w_k - w^*\|^2.\end{aligned}$$

Taking expectations with respect to $\mathbb{E}_{z_k}$,

$$\begin{aligned}\implies \mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] &= \eta^2\mathbb{E}_{z_k}\left[\|\nabla F(w_k, z_k)\|^2\right] - 2\eta\mathbb{E}_{z_k}\left[\langle\nabla F(w_k, z_k), w_k - w^*\rangle\right] + \|w_k - w^*\|^2 \\ &= \eta^2\mathbb{E}_{z_k}\left[\|\nabla F(w_k, z_k)\|^2\right] - 2\eta\left\langle\nabla F(w_k), w_k - w^*\right\rangle + \|w_k - w^*\|^2.\end{aligned}$$

Let $C = 2\eta^2\max\left\{\rho L, \lambda\right\}\left(\lambda\|w^+\|^2 - \left(\lambda+\mu\right)\|w^* - w^+\|^2\right)$. Then Lemma 13 implies

$$\begin{aligned}\mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] \leq 4\eta^2\max\left\{\rho L, \lambda\right\}\left(F(w_k) - F(w^*)\right) - 2\eta\left\langle\nabla F(w_k), w_k - w^*\right\rangle \\ + \|w_k - w^*\|^2 + C\end{aligned}$$

Using $(\mu+\lambda)$-strong-convexity of F,

$$\implies \mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] \leq \frac{2\eta^2\max\left\{\rho L, \lambda\right\}}{\mu+\lambda}\|\nabla F(w_k)\|^2 - 2\eta\left\langle\nabla F(w_k), w_k - w^*\right\rangle + \|w_k - w^*\|^2 + C.$$

Coercivity of the gradient (Lemma 21) now implies

$$\begin{aligned}\mathbb{E}_{z_k}\left[\|w_{k+1} - w^*\|^2\right] &\leq \frac{2\eta^2\max\left\{\rho L, \lambda\right\}}{\mu+\lambda}\|\nabla F(w_k)\|^2 - 2\eta\left(\frac{(\mu+\lambda)(L+\lambda)}{(\mu+\lambda)+(L+\lambda)}\|w_k - w^*\|^2\right. \\ &\qquad \left. + \frac{1}{(\mu+\lambda)+(L+\lambda)}\|\nabla F(w_k)\|^2\right) + \|w_k - w^*\|^2 + C \\ &= 2\eta\left(\frac{\eta\max\left\{\rho L, \lambda\right\}}{\mu+\lambda} - \frac{1}{(\mu+\lambda)+(L+\lambda)}\right)\|\nabla F(w_k)\|^2 \\ &\qquad + \left(1 - \frac{2\eta(\mu+\lambda)(L+\lambda)}{(\mu+\lambda)+(L+\lambda)}\right)\|w_k - w^*\|^2 + C.\end{aligned}$$

If $\eta \leq \frac{\mu+\lambda}{\max\{\rho L,\lambda\}((\mu+\lambda)+(L+\lambda))}$, then we obtain

$$\mathbb{E}_{z_k}\left[\|w_{k+1}-w^*\|^2\right] \leq \left(1-\frac{2\eta(\mu+\lambda)(L+\lambda)}{(\mu+\lambda)+(L+\lambda)}\right)\|w_k-w^*\|^2+C.$$

Let $\delta = \frac{2\eta(\mu+\lambda)(L+\lambda)}{(\mu+\lambda)+(L+\lambda)}$. Taking expectations and recursing on this inequality,

$$\begin{aligned}
\implies \mathbb{E}\left[\|w_{k+1}-w^*\|^2\right] &\leq (1-\delta)^{k+1}\|w_0-w^*\|^2 + C\sum_{l=0}^{k}(1-\delta)^l \\
&= (1-\delta)^{k+1}\|w_0-w^*\|^2 + C\left(\frac{1-(1-\delta)^{k+1}}{\delta}\right) \\
&\leq (1-\delta)^{k+1}\|w_0-w^*\|^2 + \frac{C}{\delta}.
\end{aligned}$$

Letting $\gamma = \frac{\eta\max\{\rho L,\lambda\}[(\mu+\lambda)+(L+\lambda)]}{(\mu+\lambda)(L+\lambda)}$ and subsituting in the value of δ gives the final result:

$$\begin{aligned}
\implies \mathbb{E}\left[\|w_{k+1}-w^*\|^2\right] \leq \left(1-\frac{2\eta(\mu+\lambda)(L+\lambda)}{(\mu+\lambda)+(L+\lambda)}\right)^{k+1} &\|w_0-w^*\|^2 + \gamma\lambda\|w^+\|^2 \\
&- \gamma\left(\lambda+\mu\right)\|w^*-w^+\|^2.
\end{aligned}$$

□

Appendix F

Useful Lemmas

Lemma 17. *Let f be a μ-strongly-convex, L-smooth function. Then, for all $w \in \mathbb{R}^d$, the optimality gap and distance to the minimizer can be related as follows:*

$$\frac{\mu}{2}\|w - w^*\|^2 \leq f(w) - f(w^*) \leq \frac{L}{2}\|w - w^*\|^2.$$

The proof of Lemma 17 follows immediately from the definitions of Lipschitz-smoothness and strong convexity evaluated at w and w^*.

Lemma 18. *Let f be an L-smooth function. Then f satisfies the following inequality for all $w \in \mathbb{R}^d$ as follows:*

$$\frac{1}{2L}\|\nabla f(w)\|^2 \leq f(w) - f(w^*).$$

Similarly, if f is μ-strongly-convex, then f satisfies the Polyak-Łojasiewicz condition,

$$f(w) - f(w^*) \leq \frac{1}{2\mu}\|\nabla f(w)\|^2 \qquad \forall\, w \in \mathbb{R}^d.$$

See Karimi et al. (2016) for proof.

Lemma 19. *Let f be a convex, L-smooth function. Then f satisfies the following inequality for all $w, v \in \mathbb{R}^d$:*

$$f(w) - f(v) \leq \langle \nabla f(w), w - v\rangle - \frac{1}{2L}\|\nabla f(w) - \nabla f(v)\|^2.$$

See Bubeck (2015, Lemma 3.5) for proof.

Lemma 20. *Let f be a convex, L-smooth function. Then f satisfies the following inequality:*

$$\langle \nabla f(w) - \nabla f(v), w - v\rangle \geq \frac{1}{L_{max}}\|\nabla f(w) - \nabla f(v)\|^2.$$

See Bubeck (2015, Eq. 3.6) for proof.

Lemma 21 (Coercivity of the Gradient)**.** *Let f be a μ-strongly-convex, L-smooth function. Then f satisfies the following inequality:*

$$\langle \nabla f(w) - \nabla f(v), w - v \rangle \geq \frac{\mu L}{\mu + L} \|w - v\|^2 + \frac{1}{\mu + L} \|\nabla f(w) - \nabla f(v)\|^2.$$

See Bubeck (2015, Lemma 3.11) for proof.

Lemma 22. *Let f be a μ-strongly-convex, L-smooth function. Then f satisfies the following inequality:*

$$\langle \nabla f(w) - \nabla f(v), w - v \rangle \geq \mu \|w - v\|^2.$$

See Nesterov (2004, Theorem 2.1.9) for proof.