



universität  
wien

# MAGISTERARBEIT / MASTER'S THESIS

Titel der Magisterarbeit / Title of the Master's Thesis

**“A non-asymptotic analysis of averaged stochastic gradient descent for least squares regression”**

verfasst von / submitted by

**Manuel Müller, BSc**

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of  
**Magister der Sozial- und Wirtschaftswissenschaften (Mag. rer. soc. oec.)**

Wien, 2021 / Vienna, 2021

Studienkennzahl lt. Studienblatt /  
degree programme code as it appears on  
the student record sheet:

A 066 951

Studienrichtung lt. Studienblatt /  
degree programme as it appears on  
the student record sheet:

Magisterstudium Statistik

Betreut von / Supervisor:

Univ.-Prof. Mag. Dr. Hannes Leeb



# Abstract

*Stochastic gradient descent* (SGD) is a very popular tool in optimization and machine learning due to its versatility and computational efficiency. While asymptotic properties of SGD were soon proven after its emergence in 1951 (see Robbins and Monro [1951]), recent results have characterized the behavior of SGD in non-asymptotic settings. The results by Bach and Moulines [2013] belong to this line of work and provide an optimal convergence rate of order  $O(1/n)$  of SGD for the problem of least squares regression, when combining constant step sizes with *Polyak–Ruppert averaging* [Ruppert, 1988, Polyak and Juditsky, 1992].

Over the course of this thesis, the relevant definitions, assumptions and algorithms are presented and discussed in extensive detail. The result by Bach and Moulines [2013] is analyzed and an overview over the related literature is given, in order to provide insight into the status quo at the time of the publication of the paper by Bach and Moulines [2013] as well as into subsequent analyses and improvements. In particular, this will shed light on the fact that, while the usage of SGD with constant step sizes and Polyak–Ruppert averaging shows desirable behavior for the problem of least squares regression, it will not produce equally appealing results for other problems. This shows that problem-specific analyses of SGD (and its variants) are necessary in order to make full use of the possible performance of SGD.

# Zusammenfassung

*Stochastic Gradient Descent* (SGD, dt. stochastisches Gradientenverfahren) ist wegen seiner breiten Anwendbarkeit sowie rechnerischen Effizienz eine sehr beliebte Technik im Feld der Optimierung und des Maschinellen Lernens. Das asymptotische Verhalten von SGD wurde recht bald nach seiner Entstehung 1951 (siehe Robbins and Monro [1951]) ausführlich erforscht. Jüngere Literatur befasste sich nun mit nicht-asymptotischen Resultaten. Dazu gehört insbesondere das Werk von Bach and Moulines [2013], in dem die Autoren eine optimale Konvergenzrate von  $O(1/n)$  für den Anwendungsfall der Kleinst-Quadrate-Regression durch die Verwendung einer konstanten Schrittweite in Kombination mit Mittelung der erzeugten Punkte (s.g. *Polyak–Ruppert averaging*, siehe Ruppert [1988], Polyak and Juditsky [1992]) erzielen.

Im Rahmen dieser Masterarbeit werden zuerst die notwendigen Definitionen, Annahmen und Algorithmen eingeführt und diskutiert. In weiterer Folge wird das konkrete Resultat von Bach and Moulines [2013] analysiert. Abschließend ist ein Überblick über die Literatur inkludiert, der den Status Quo zum Zeitpunkt der Arbeit von Bach and Moulines [2013], sowie die weiterfolgenden Analysen und Folgeresultate erläutert. Insbesondere wird durch diesen Überblick klar, dass SGD mit konstanter Schrittweite und Mittelung der erzeugten Punkte zwar erstrebenswerte Resultate für das Problem der Kleinst-Quadrate-Regression erzeugt, für andere Probleme jedoch weniger geeignet ist. Das zeigt insbesondere, dass es notwendig ist, SGD (und seine Varianten) im Kontext der jeweiligen Problemstellung zu analysieren, um optimale Resultate zu erhalten.

# Danksagung

Der erste und größte Dank gebührt zweifellos meinen Eltern Manfred und Ilse, die mir durch ihre Unterstützung nicht nur das Studium ermöglicht haben, sondern auch dafür gesorgt haben, dass meine Studienzeit unbeschwert durch existenzielle Sorgen war. Natürlich umfasst ihr Einfluss aber deutlich mehr. So habe ich von ihnen zum Beispiel meine Begeisterung für Knobelien und das Lösen von Problemen mit auf den Weg bekommen. Genau wie meine Eltern waren auch meine Schwestern Miriam und Melina immer für mich da, wenn ich eine Auszeit vom Studium gebraucht habe und haben mich so unwissentlich sehr unterstützt.

Ein anderer wichtiger Ausgleich zum Studieren war für mich meine Tätigkeit als Tutor bzw. Studienassistent für Prof. Hudec, Prof. Schachinger und Prof. Leeb. Allen dreien bin ich dafür dankbar, dass sie mir die Möglichkeit gegeben haben, meine Faszination für Statistik durch die Tutorien mit anderen Studierenden zu teilen und meine Freude am Erklären auszuleben. Prof. Leeb bin ich aber auch in vielen anderen Belangen dankbar. Nicht nur für die ausdauernde Betreuung dieser Arbeit, sondern auch für die tatkräftige Unterstützung bei meiner Suche nach einer Stelle für mein Doktoratsstudium und die wertvollen Ratschläge für meine nächsten (oder vielmehr ersten) Schritte im Forschungsbetrieb. Ebenso bin ich Ivana Milovic und Prof. Pflug, sowie insbesondere Prof. Pötscher und Prof. Bomze für die aktive Unterstützung bei meinen Bewerbungen auf Doktoratsstellen dankbar.

Da diese Masterarbeit auch den Abschluss meines Studiums an der Universität Wien bedeutet, möchte ich mich bei den vielen StudienkollegInnen bedanken, die meine Studienzeit ausgezeichnet haben. Allen voran zählen dazu Christian, dessen Wissensdrang und Diskursfreudigkeit ohnegleichen sind, Xaver, der stets für gute Laune, Interesse und Faszination sorgt, und natürlich Bella, die jede Gruppe durch ihre Hilfsbereitschaft und ihr Engagement prägt.

Ein ganz besonderer Platz in meiner Danksagung gehört natürlich zu guter Letzt meiner Freundin Consti. Auch wenn ich ihretwegen immer aufpassen muss, nicht über das Wort „Konstante“ zu stolpern, haben ihre selbstverständliche Unterstützung und die vielen schönen Momenten in den letzten Jahren dazu beigetragen, dass ich die Studienzeit in so positiver Erinnerung behalten werde.



# Contents

|                                                                                 |            |
|---------------------------------------------------------------------------------|------------|
| <b>Abstract</b>                                                                 | <b>i</b>   |
| <b>Zusammenfassung</b>                                                          | <b>ii</b>  |
| <b>Danksagung</b>                                                               | <b>iii</b> |
| <b>1. Introduction</b>                                                          | <b>1</b>   |
| <b>2. Background</b>                                                            | <b>5</b>   |
| 2.1. Notation . . . . .                                                         | 5          |
| 2.2. Basic examples of applications . . . . .                                   | 6          |
| 2.3. Basic definitions . . . . .                                                | 6          |
| 2.4. Common assumptions . . . . .                                               | 8          |
| 2.4.1. Properties of the domain $\Theta$ . . . . .                              | 8          |
| 2.4.2. Smoothness . . . . .                                                     | 8          |
| 2.4.3. Convexity . . . . .                                                      | 9          |
| 2.4.4. Strong convexity . . . . .                                               | 9          |
| 2.4.5. The condition number for smooth, strongly convex functions . . . . .     | 11         |
| 2.4.6. Performance measures for smooth, strongly convex functions . . . . .     | 12         |
| 2.5. A characterization of large-scale learning . . . . .                       | 12         |
| 2.6. Gradient Descent (GD) . . . . .                                            | 14         |
| 2.6.1. Basic performance results of GD . . . . .                                | 15         |
| 2.7. Stochastic Gradient Descent (SGD) . . . . .                                | 16         |
| 2.7.1. SGD for empirical risk minimization . . . . .                            | 16         |
| 2.7.2. SGD for risk minimization . . . . .                                      | 17         |
| 2.7.3. Variants of SGD . . . . .                                                | 18         |
| 2.8. Comparison of GD and SGD . . . . .                                         | 22         |
| 2.9. The concept of optimality . . . . .                                        | 25         |
| 2.9.1. Optimality and problem complexity in oracle-based optimization . . . . . | 27         |
| 2.9.2. Optimality in empirical risk minimization and M-estimation . . . . .     | 30         |
| 2.9.3. The ERM for least squares regression . . . . .                           | 31         |
| 2.10. Concluding remark . . . . .                                               | 32         |

|                                                                            |           |
|----------------------------------------------------------------------------|-----------|
| <b>3. Non-asymptotic analysis of least squares regression</b>              | <b>35</b> |
| 3.1. Revisiting the example of least squares regression . . . . .          | 35        |
| 3.2. The non-asymptotic bound by Bach and Moulines [2013] . . . . .        | 37        |
| 3.3. Discussion . . . . .                                                  | 39        |
| 3.3.1. Discussion of the assumptions . . . . .                             | 39        |
| 3.3.2. Discussion of the bound . . . . .                                   | 43        |
| 3.4. Proof outline . . . . .                                               | 45        |
| <b>4. Related results and literature review</b>                            | <b>49</b> |
| 4.1. Introduction and common goal . . . . .                                | 49        |
| 4.2. Preceding literature . . . . .                                        | 50        |
| 4.2.1. Origins . . . . .                                                   | 50        |
| 4.2.2. Subsequent asymptotic analysis . . . . .                            | 52        |
| 4.2.3. Iterate averaging in stochastic approximation . . . . .             | 52        |
| 4.2.4. Constant step sizes for least squares regression . . . . .          | 53        |
| 4.2.5. Non-asymptotic viewpoint . . . . .                                  | 54        |
| 4.3. The result by Bach and Moulines [2013] . . . . .                      | 57        |
| 4.4. Subsequent literature . . . . .                                       | 57        |
| 4.4.1. Improving the sublinear convergence of the bias . . . . .           | 58        |
| 4.4.2. Linear convergence of the bias . . . . .                            | 60        |
| 4.4.3. Geometrically decaying averaging weights . . . . .                  | 62        |
| 4.4.4. Characterization through Markov chains . . . . .                    | 63        |
| 4.5. Further related results . . . . .                                     | 64        |
| 4.5.1. Choice of the step size sequence . . . . .                          | 65        |
| 4.5.2. Detection of the stationary oscillation phase in practice . . . . . | 66        |
| 4.5.3. Finite horizon . . . . .                                            | 67        |
| <b>5. Discussion</b>                                                       | <b>69</b> |
| <b>References</b>                                                          | <b>71</b> |
| <b>List of Algorithms</b>                                                  | <b>75</b> |
| <b>List of Figures</b>                                                     | <b>75</b> |

|                                                                                                  |           |
|--------------------------------------------------------------------------------------------------|-----------|
| <b>Appendices</b>                                                                                | <b>77</b> |
| <b>A. An additional result</b>                                                                   | <b>79</b> |
| A.1. Asymptotic equalities in Bottou [2010]                                                      | 79        |
| <b>B. The proof of Theorem 1 by Bach and Moulines [2013]</b>                                     | <b>81</b> |
| B.1. Lemmata                                                                                     | 81        |
| B.2. Assumptions                                                                                 | 90        |
| B.3. The theorem of interest                                                                     | 92        |
| B.4. Properties of $\eta_n = \theta_n - \theta_*$ and $\bar{\eta}_n = \bar{\theta}_n - \theta_*$ | 92        |
| B.5. Construction of the upper bound on the expected excess risk                                 | 96        |
| B.5.1. Construction of the bias term                                                             | 97        |
| B.5.2. Construction of the variance term                                                         | 100       |
| B.5.3. Completion of the upper bound on the expected excess risk                                 | 111       |
| B.6. Improvement of the bound                                                                    | 111       |



# 1. Introduction

Optimization problems naturally occur in machine learning and statistics. The most common occurrence lies in the minimization of either a loss or the expectation of a loss. Specific examples for this are the maximum likelihood estimation approach commonly employed in statistics or the minimization of the training and/or test error in the supervised learning framework. The typical setting in supervised learning assumes data in the form of i.i.d. observations  $(y_i, x_i) \in Y \times X$ ,  $i \geq 1$ , with  $Y \subseteq \mathbb{R}$ ,  $X \subseteq \mathbb{R}^{d_X}$  and  $d_X \in \mathbb{N}$ . The former element of each observation (i.e.  $y_i$ ) is the response variable, which we would like to predict by the latter element (i.e.  $x_i$ ), which is the explanatory variable. The suitability of the prediction is quantified through a loss function, which is often defined per observation, i.e.  $l : Y \times Y \rightarrow \mathbb{R}$ . The goal is to minimize the *risk*  $\mathbb{E} [l(y_1, g(x_1))]$  of a function  $g$  over a suitable class of candidate predictive functions  $\mathcal{G}$  mapping from  $X$  to  $Y$ .

Throughout this thesis we will deal with a parameterized class of functions, meaning that  $\mathcal{G}$  is of the form  $\mathcal{G} = \mathcal{G}_\Theta = \{g_\theta \mid g_\theta : X \rightarrow Y, \theta \in \Theta\}$  for  $\Theta \subseteq \mathbb{R}^d$ ,  $d \in \mathbb{N}$ . This is opposed to the case of infinitely dimensional domains, which appear in nonparametric statistical methods (see for example Dieuleveut and Bach [2016]). As a result, in this thesis the task simplifies to finding the *risk minimizer*  $\theta_* \in \Theta$ , given as a solution of

$$\underset{\theta \in \Theta}{\text{minimize}} \mathbb{E} [l(y_1, g_\theta(x_1))], \quad (1.1)$$

if one should exist. For simplicity, we will assume integrability of  $l(g_\theta(x_1), y_1)$  for all  $\theta \in \Theta$  for now. Note that the measure with respect to which the expected value is calculated does not depend on  $\theta$ .

In practice, the distribution of the data will typically be unknown, meaning that (1.1) cannot be solved directly, as the expectation cannot be calculated. The value of the risk therefore has to be estimated, which is commonly done via the *empirical risk* defined through the corresponding sample average. This gives rise to the *empirical risk minimizer*  $\hat{\theta}_n = \hat{\theta}_n((y_1, x_1), \dots, (y_n, x_n)) \in \Theta$ , which is a random variable solving

$$\underset{\theta \in \Theta}{\text{minimize}} \frac{1}{n} \sum_{i=1}^n l(y_i, g_\theta(x_i)). \quad (1.2)$$

## 1. Introduction

Again, we will assume for now that such a quantity  $\hat{\theta}_n$  exists. Sufficient conditions will be introduced at a later time.

In statistics, using  $\hat{\theta}_n$  as an estimator for  $\theta_*$  is studied in the context of *M-estimators* (see for example van der Vaart [2000]). For many common problems, such estimators and their properties are well known. While they are often times known to be efficient with respect to the amount of available data  $n$  (e.g. maximum likelihood estimation), solving (1.2) can indeed be cumbersome. This holds especially true in the context of large-scale learning, where  $n$  is large. In such applications, it might be preferable (and in a sense even more efficient) to only approximately solve (1.2) even if exact minimization is possible, as will be discussed in detail in the following chapter.

One very basic algorithm that can achieve approximate minimization is *stochastic gradient descent* (*SGD*) which dates back to Robbins and Monro [1951] and Kiefer and Wolfowitz [1952]. It can be used to solve both (1.1) and (1.2) through iterative updates of an estimator sequence based on estimates of the (sub-)gradient of the target function. SGD and its variants enjoy high popularity in the machine learning community due to their versatile applicability and good empirical performance. In particular, in large-scale learning settings, SGD is competitive with more complicated algorithms and sometimes even the empirical risk minimizer itself (e.g. in the case of least squares regression), when the computational cost is taken into account.

Due to the the need for efficient algorithms in modern applications with high-dimensional search spaces and an abundance of data, a substantial amount of research has been undertaken in order to characterize the behavior of the resulting estimator sequences from a theoretical point of view. Of particular interest are, of course, performance guarantees.

This thesis will shed some light on these performance guarantees with a special focus on non-asymptotic guarantees, which have been of particular interest in recent years.

Since there is a wide range of machine learning methods that are trained through the optimization problems (1.1) and (1.2), performance guarantees for SGD need to hold under different assumptions regarding the loss function  $l$  and the risk  $f(\theta) := \mathbb{E} [l(y_1, g_\theta(x_1))]$ . Such assumptions are typically concerned with the smoothness of  $f(\theta)$  (e.g. the existence of Lipschitz-continuous gradients on  $\Theta$ ) and the type of convexity of  $f(\theta)$  (i.e. non-convexity, general convexity, strict convexity or strong convexity on  $\Theta$ ).

Making such assumptions often yields two issues. Firstly, of course, they restrict the applicability of the result. For example, the risk of logistic regression is not strongly convex and the one for neural networks is in general not convex at all. Secondly, algorithms like SGD typically include hyperparameters, for which the optimal choice may depend on unknown problem specific properties relating to the assumptions. For example, it might not suffice to know that the gradients of  $f$  are Lipschitz-continuous, as one also needs to know for which constant  $L$  they are Lipschitz-continuous. More details on this as well as the necessary definitions are given over the

course of this thesis.

Indeed, misconceptions of such quantities can drastically worsen the performance of an algorithm and can render it basically unusable. As a result, there is significant interest in relaxing some of the popular assumptions and making the choice of the hyperparameters more robust with respect to the knowledge of problem specific properties.

One particularly exciting result in that regard is presented by Bach and Moulines [2013] who achieve an optimal convergence rate of  $O(1/n)$  of the expected risk induced by a special variant of SGD for the case of least squares regression, which also holds in a non-asymptotic setting as they prove an upper bound of the type  $C/n$  for a constant  $C > 0$ . In particular, they managed to make SGD competitive with the ordinary least squares (OLS) estimator, irrespective of possible high correlation between explanatory variables, which made previously known bounds quickly uninformative. As a result, their work makes a highly important contribution to the question of how one can make the most of SGD when applied to the problem of least squares regression.

The structure of the thesis will be as follows. In Chapter 2, basic definitions as well as important considerations regarding large-scale learning are given. In Chapter 3 the mentioned seminal result by Bach and Moulines [2013] is presented and discussed in more detail. Lastly, in Chapter 4 an overview over the existing literature is provided.

For the sake of better readability, some technical details can be found in the Appendix.



## 2. Background

### 2.1. Notation

In this thesis, all vectors are column vectors. Let  $d, d_X \in \mathbb{N}$  and  $\Theta \subseteq \mathbb{R}^d$ . Let  $x$  be a random vector with values in  $\mathbb{R}^{d_X}$  and  $y$  a real random variable. Let  $(y_i, x_i^T)^T$ ,  $i \geq 1$ , be a sequence of i.i.d. random vectors such that  $(y_1, x_1^T)^T \sim (y, x^T)^T$ . Define  $l : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  as the known loss function. For parameterized functions  $g_\theta$  in a class  $\mathcal{G} = \mathcal{G}_\Theta = \{g_\theta \mid g_\theta : \mathbb{R}^{d_X} \rightarrow \mathbb{R}, \theta \in \Theta\}$  let  $f(\theta) := \mathbb{E}[l(y, g_\theta(x))]$  denote the risk for all  $\theta \in \Theta$  and  $\hat{f}_n(\theta) := n^{-1} \sum_{i=1}^n l(y_i, g_\theta(x_i))$  denote the empirical risk for all  $\theta \in \Theta$  and  $n \in \mathbb{N}$ . Here, it is implicitly assumed that the loss function  $l$ , the function class  $\mathcal{G}_\Theta$  and the distribution of  $(y, x^T)^T$  is such that  $f(\theta)$  is well defined for all  $\theta \in \Theta$ , i.e.  $l(y, g_\theta(x))$  is integrable. Further, if it uniquely exists, the minimizer of  $f$  will be denoted by  $\theta_*$  and the one of  $\hat{f}_n$  by  $\hat{\theta}_n$ . The function  $F : \Theta \rightarrow \mathbb{R}$  is used to present results that hold true in general and particularly for any  $F \in \{f, \hat{f}_n\}$ . If it uniquely exists, the minimizer of  $F$  over  $\Theta$  will be denoted by  $t_*$ . Naturally, often we will have  $t_* = \theta_*$  or  $t_* = \hat{\theta}_n$ .

In the following chapters, sequences  $(t_k)_{k \geq 0}$ ,  $(\theta_k)_{k \geq 0}$  and  $(\tilde{\theta}_k)_{k \geq 0}$  will appear among others. The first represents a sequence achieved through gradient descent, the second a sequence resulting from stochastic gradient descent and the last is a generic sequence that can represent either.

Furthermore, let  $\|\cdot\|_2$  denote either the Euclidean norm or the corresponding induced matrix norm. For a quadratic, positive semi-definite, real matrix  $A$  and a real random vector  $Z$  of suitable dimension and  $\mathbb{E}[\|Z\|_2^2] < \infty$ , let  $\|Z\|_A = \sqrt{\mathbb{E}[Z^T A Z]}$ . For any matrix  $A \in \mathbb{R}^{d \times d}$  with real eigenvalues, let  $\lambda_i = \lambda_i(A)$ ,  $i \in \{1, \dots, d\}$ , denote its eigenvalues.  $\lambda_{\max}(A)$  is the largest and  $\lambda_{\min}(A)$  the smallest eigenvalue. Furthermore,  $I_d$  denotes the  $d$ -dimensional identity matrix, for which the subscript will mostly be dropped, as the dimension is typically clear from the context. For deterministic sequences  $(a_n)_{n \geq 1}$  and  $(b_n)_{n \geq 1}$ , the symbol  $\sim$  is used to denote asymptotic equality up to constants, i.e.  $a_n \sim b_n$  is equivalent to  $\lim_{n \rightarrow \infty} a_n/b_n = r$  for some constant  $r \in (0, \infty)$ . Furthermore,  $a_n = O(b_n)$  is used to denote  $\limsup_{n \rightarrow \infty} |a_n/b_n| < \infty$  and  $a_n = \Omega(b_n)$  to denote  $\liminf_{n \rightarrow \infty} |a_n/b_n| > 0$ . If and only if both  $a_n = O(b_n)$  and  $a_n = \Omega(b_n)$  hold,  $a_n = \Theta(b_n)$  is used as a shorthand notation.

In order to avoid convoluted notation, the symbols  $c$ ,  $C$  and  $\rho$  are used to denote constants on multiple occasions with different meanings. Similarly  $\gamma_k$  can denote different step size sequences. As a result, the values of the given quantities may differ between sections and they should be understood in the context in which they are presented.

## 2. Background

### 2.2. Basic examples of applications

**Example 1** (Least Squares Regression). Let  $d \geq 1$  be an arbitrary integer and  $d_X = d$ . Let furthermore  $\Theta = \mathbb{R}^d$ . We consider the class of linear model functions, i.e. the class of functions given through  $g_\theta(x) = \theta^T x$ . The quadratic loss is given through  $l(y, y') = \frac{1}{2}(y - y')^2$ . As a result, the risk is  $f(\theta) = \mathbb{E} \left[ \frac{1}{2}(y - x^T \theta)^2 \right]$  and the empirical risk for a dataset of size  $n$  is  $\hat{f}_n(\theta) = \frac{1}{2n} \sum_{i=1}^n (y_i - x_i^T \theta)^2$ . If the empirical risk minimizer (ERM) is unique, it is equal to the ordinary least squares (OLS) estimator.

Typically, some regularity conditions are imposed on the distribution of  $(y, x)$ . For example, for the risk to be well defined for all  $\theta \in \mathbb{R}^d$ , we need finite second moments of  $y$  and  $x$ . We can further assume without loss of generality that the second moment matrix  $H = \mathbb{E} [xx^T]$  is not only positive semi-definite, but indeed positive definite. The risk function of least squares regression allows for the simplification to the following quadratic function

$$f(\theta) = \frac{1}{2} \theta^T H \theta - \theta^T \mathbb{E} [yx] + \frac{1}{2} \mathbb{E} [y^2]$$

and will be of much interest later.

**Example 2** (Logistic Regression). Let again  $d \in \mathbb{N}$ ,  $d = d_X$  and  $\Theta = \mathbb{R}^d$ . Again, the class of model functions are those that can be written as  $g_\theta(x) = \theta^T x$ . Other than in the previous example, let here  $y \in \{-1, 1\}$  and define the loss function through  $l(y, y') = \log(1 + \exp(-yy'))$ . The loss incurred through an observation  $(y_i, x_i)$  and a choice of parameter  $\theta \in \mathbb{R}^d$  is then given by  $\log(1 + \exp(-y_i \cdot x_i^T \theta))$ .

### 2.3. Basic definitions

While the notion of *convergence rate* is intuitive, a rigorous definition is needed to understand the optimization literature. An introduction to the most important concepts is, for example, given in Nocedal and Wright [2006, Appendix A.2].

**Definition.** Let  $p \in \mathbb{N}$  and let  $(a_k)_{k \geq 1}$  be a sequence in  $\mathbb{R}^p$  converging to  $a \in \mathbb{R}^p$ . Let  $\|\cdot\|$  denote the norm of interest, e.g. the Euclidean norm. The convergence is ...

(a) ... Q-linear if there is some  $r \in (0, 1)$  such that for all sufficiently large  $k$ :

$$\frac{\|a_{k+1} - a\|}{\|a_k - a\|} \leq r.$$

(b) ... Q-superlinear if

$$\lim_{k \rightarrow \infty} \frac{\|a_{k+1} - a\|}{\|a_k - a\|} = 0.$$

(c) ... Q-sublinear if

$$\lim_{k \rightarrow \infty} \frac{\|a_{k+1} - a\|}{\|a_k - a\|} = 1.$$

While the prefix "Q-" for "quotient" is sometimes used to distinguish between this definition and the definition of "root convergence" marked by the prefix "R-", it is typically omitted in the literature. For that reason, it will be omitted in this thesis as well (e.g. "linear convergence" refers to "Q-linear convergence"). For details, see Nocedal and Wright [2006, Appendix A.2].

The naming of the different types of convergence is not immediately obvious. However, consider the case of  $p = 1$ ,  $a = 0$  and  $a_k > 0$  for all  $k \geq 1$ , as this will be the situation of interest in this thesis. Now, define  $b_k := \log(a_k)$  for all  $k \geq 1$  and note that this sequence  $(b_k)_{k \geq 1}$  diverges to  $-\infty$ . Note the following property of the quotient of interest in the definition above:

$$\frac{\|a_{k+1} - a\|}{\|a_k - a\|} = \frac{a_{k+1}}{a_k} = \frac{e^{b_{k+1}}}{e^{b_k}} = e^{b_{k+1} - b_k}.$$

Reformulating the definition above for the logarithm  $b_k$  of the sequence  $a_k$  gives the following.

- (a) The series  $a_k$  converges Q-linearly if and only if there is some  $s < 0$ , such that for all sufficiently large  $k \geq 1$ :

$$b_{k+1} - b_k \leq s,$$

i.e.  $b_k$  diverges to  $-\infty$  at least as quickly as a linear function with slope  $s$ .

- (b) The series  $a_k$  converges Q-superlinearly if and only if

$$b_{k+1} - b_k \rightarrow -\infty$$

for  $k \rightarrow \infty$ , i.e.  $b_k$  diverges increasingly fast towards  $-\infty$ .

- (c) The series  $a_k$  converges Q-sublinearly if and only if

$$b_{k+1} - b_k \rightarrow 0$$

for  $k \rightarrow \infty$ , i.e.  $b_k$  diverges increasingly slowly towards  $-\infty$ .

As a result, the naming convention for the convergence rates becomes intuitive, when relating it to the behavior of the logarithm of the sequence.

Consider the following important example.

**Example 3.** Sequences of the form  $a_k = 1/k^c$  for some  $c > 0$  converge sublinearly to 0 for  $k \rightarrow \infty$ , while sequences of the form  $a'_k = \rho^k$  for some  $\rho \in (0, 1)$  converge linearly to 0 for  $k \rightarrow \infty$ . These specific examples will reappear several times throughout this thesis.

## 2. Background

Note that in the literature the term "rate of convergence" can have different meanings. For example, consider a frequently observed case in statistics, where we have a deterministic sequence  $(a_n)_{n \geq 1}$  and a sequence of random vectors  $(Z_n)_{n \geq 1}$  such that for some constant vector  $c$ , it holds that  $a_n(Z_n - c)$  converges in distribution to a normal distribution  $\mathcal{N}(0, V)$  for  $n \rightarrow \infty$  and some matrix  $V \succ 0$ . As noted in Kushner and Yin [2003, Chapter 11], the sequence  $a_n$ , taken together with the matrix  $V$ , is also often considered to give the "rate of convergence".

## 2.4. Common assumptions

In the following subsections, some concepts typically appearing in assumptions in the optimization literature are defined and basic consequences are discussed. These results are well-known in the literature.

### 2.4.1. Properties of the domain $\Theta$

A common assumption on  $\Theta$  is that it is a compact, convex, non-empty subset of  $\mathbb{R}^d$ . In this case, SGD is implemented with additional projection steps, which will be introduced in Subsection 2.7.3.

An alternative assumption on  $\Theta$ , which is prevalent when dealing with cases like least squares regression, is that  $\Theta$  spans the entire real space, i.e.  $\Theta = \mathbb{R}^d$ .

### 2.4.2. Smoothness

An important property of the target function  $F$  is its smoothness.

**Definition.** We define a function  $F : \Theta \rightarrow \mathbb{R}$  to be  *$L$ -smooth (on  $\Theta$ )* if there is an open set  $\Theta_+ \supseteq \Theta$ , such that the gradients of  $F$  exist in every  $\theta \in \Theta_+$  and are Lipschitz-continuous with constant  $L > 0$ . Formally, we call  $F$   *$L$ -smooth* if there is an  $L > 0$ , such that  $\forall \theta, \theta' \in \Theta_+$

$$\|\nabla F(\theta) - \nabla F(\theta')\|_2 \leq L \|\theta - \theta'\|_2.$$

Note that if  $\Theta$  is open (e.g. if  $\Theta = \mathbb{R}^d$ ), we can simply set  $\Theta_+ = \Theta$ .

A particularly illustrative consequence of smoothness is that for all  $\theta, \theta' \in \Theta$

$$F(\theta) \leq F(\theta') + (\nabla F(\theta'))^T (\theta - \theta') + \frac{L}{2} \|\theta - \theta'\|_2^2 \quad (2.1)$$

holds. For a proof, see for example Bottou et al. [2018]. One can show that for twice continuously differentiable functions, the function is  $L$ -smooth if the moduli of the eigenvalues of the Hessian matrix are upper bounded by  $L$  on  $\Theta$ .

The prefix " $L$ -" will sometimes be omitted in the following.

*Remark.* It should be pointed out that the definition of smoothness typically provided in the literature is slightly simpler. Specifically, the set  $\Theta_+$  was introduced here in order to avoid

having to define the gradient of  $F$  on the border of  $\Theta$  if it is compact. This is typically not commented on in the literature (see for example Nemirovski et al. [2009]). If  $\Theta$  is open, we simply have  $\Theta_+ = \Theta$  and the definition provided here coincides with the one that is commonly used in the literature.

### 2.4.3. Convexity

**Definition.** A function  $F : \Theta \rightarrow \mathbb{R}$  is said to be convex (on  $\Theta$ ) if for all  $\theta, \theta' \in \Theta$  and for all  $\lambda \in [0, 1]$  it holds that

$$F(\lambda\theta + (1 - \lambda)\theta') \leq \lambda F(\theta) + (1 - \lambda)F(\theta').$$

If  $\nabla F(\theta)$  exists for all  $\theta \in \Theta$  (as is implied by smoothness of  $F$ , for example), the definition is equivalent to requiring that for all  $\theta, \theta' \in \Theta$

$$F(\theta) \geq F(\theta') + (\nabla F(\theta'))^T(\theta - \theta'),$$

i.e. the function is on  $\Theta$  lower bounded by its tangent plane in any point in  $\Theta$ .

Strict convexity simply requires the two inequalities to be strict for  $\theta \neq \theta'$ .

### 2.4.4. Strong convexity

Strong convexity is a natural extension of convexity, in which the lower bound is not a tangent plane, but rather a tangent paraboloid. It is a condition even more restrictive than strict convexity. At several points in this thesis, it is necessary to put emphasis on the fact that a function is convex but might not be strictly convex or strongly convex. In such cases, the function will be described to satisfy "general convexity".

**Definition.** A function  $F : \Theta \rightarrow \mathbb{R}$  is said to be  $\mu$ -strongly convex (on  $\Theta$ ) for a constant  $\mu > 0$  if the function

$$\theta \mapsto F(\theta) - \frac{\mu}{2}\|\theta\|_2^2$$

is convex [Bubeck, 2015].

Again, the prefix " $\mu$ -" will sometimes be omitted in the following.

If  $\nabla F(\theta)$  exists for every  $\theta \in \Theta$  (as is implied by smoothness of  $F$ ), it is easy to see that the definition is equivalent to  $F$  satisfying for all  $\theta, \theta' \in \Theta$

$$F(\theta) \geq F(\theta') + (\nabla F(\theta'))^T(\theta - \theta') + \frac{\mu}{2}\|\theta - \theta'\|_2^2, \quad (2.2)$$

which means that the function can be lower bounded by a paraboloid of minimal curvature  $\mu$  at every point in  $\Theta$ . As pointed out by Bubeck [2015], differentiability is not strictly necessary, as one can replace the gradient by any subgradient. Note further that for  $\Theta = \mathbb{R}^d$ , strong convexity implies the existence of a unique minimizer  $\theta_*$ .

## 2. Background

Inequality (2.2) yields that for all  $\theta, \theta' \in \Theta$

$$(\nabla F(\theta) - \nabla F(\theta'))^T(\theta - \theta') \geq \mu \|\theta - \theta'\|_2,$$

which follows from using (2.2) twice; once as is and once with switched  $\theta$  and  $\theta'$  and summing up each side of the two resulting inequalities separately. Using the Cauchy–Schwarz inequality, the weaker version

$$\|\nabla F(\theta) - \nabla F(\theta')\|_2 \geq \mu \|\theta - \theta'\|_2$$

follows, which in particular implies that if  $\nabla F(\theta_*) = 0$  (e.g.  $F$  is differentiable and the optimum  $\theta_*$  is an interior point of  $\Theta$ ), we have

$$\|\nabla F(\theta)\|_2 \geq \mu \|\theta - \theta_*\|_2. \quad (2.3)$$

This last inequality tells us that, for strongly convex functions, the norm of the gradient increases the farther we are away from the optimum. This property is of high importance for gradient based optimization methods and an important difference to regular convexity or even strict convexity. Some illustrative examples are provided below.

Lastly, another equivalent definition of  $\mu$ -strong convexity for twice continuously differentiable functions is that the eigenvalues of the Hessian matrix are uniformly lower bounded by  $\mu > 0$  on  $\Theta$ .

In the following, some short illustrative examples that consider the univariate case  $\Theta = \mathbb{R}$  are given, unless otherwise noted. Here, the gradient of  $F$  is simply the first derivative  $F'$ .

**Example 4.** *The function  $F : \theta \mapsto |\theta|$  is convex, but not strictly convex (let alone strongly convex). In particular, for  $\theta \neq 0$ , the norm of the gradient is equal to 1, no matter the distance to  $\theta_* = 0$ .*

**Example 5.** *The function  $F : \theta \mapsto \theta^2$  is  $\mu$ -strongly convex with  $\mu = 2$  by definition, as  $F(\theta) - \frac{2}{2}\theta^2 = 0$  is convex. Similar to a constant function being trivially convex, a quadratic function is the most simple strongly convex function. Note that  $F'(\theta) = 2\theta$ , which means that the norm of the gradient increases the further one moves away from the optimum  $\theta_* = 0$ .*

**Example 6.** *The difference between strong convexity and strict convexity is not as obvious. One simple example is the function  $F : (0, \infty) \rightarrow (0, \infty)$  defined through  $F : \theta \mapsto \theta^{-1}$ . Since the function is twice continuously differentiable, we see by  $F''(\theta) = 2\theta^{-3} > 0$  that the function is strictly convex. However, for  $\theta$  increasing towards infinity, the second derivative decreases towards 0, meaning that no positive constant  $\mu$  is a lower bound on the second derivative. As a result,  $F$  is not strongly convex. Even simpler, one could argue that  $F$  is not strongly convex due to not having a minimizer.*

**Example 7.** Another, more subtle example to illustrate the difference between strict convexity and strong convexity is the hyperbola defined through  $F : \theta \mapsto \sqrt{1 + \theta^2}$ , which has a unique minimizer  $\theta_* = 0$ . Due to the function being twice continuously differentiable, strict convexity follows again from the fact that  $F''(\theta) > 0$  for all  $\theta \in \mathbb{R}$ , but again the second derivative converges towards 0 for  $\theta \rightarrow \pm\infty$ . Since  $F$  has a unique minimizer  $\theta_*$ , we can also employ the necessary condition (2.3), to show that  $F$  is not strongly convex. Since  $F'(\theta) = \theta/\sqrt{1 + \theta^2}$ , we would require  $\mu > 0$  to satisfy  $|\theta|/\sqrt{1 + \theta^2} \geq \mu|\theta|$  for all  $\theta \in \mathbb{R}$ , which is impossible, as  $1/\sqrt{1 + \theta^2} \rightarrow 0$  for  $\theta \rightarrow \pm\infty$ . Geometrically, the problem lies in the fact that if one moves away from the optimum  $\theta_*$ , the norm of the gradient is converging towards 1, instead of steadily increasing.

**Example 8.** An important example is that of an  $l_2$ -penalized risk function with penalization parameter  $c > 0$ . Given a convex risk function  $f$ , it follows immediately from the definition of strong convexity that the  $l_2$ -penalized risk  $\theta \mapsto f(\theta) + \frac{c}{2}\|\theta\|_2^2$  is  $c$ -strongly convex. As we will see in several situations later, strong convexity can lead to much faster convergence rates than convexity alone, meaning that  $l_2$ -regularization can help to speed up the estimation process.

### 2.4.5. The condition number for smooth, strongly convex functions

**Definition.** Following Golub and Van Loan [2013, Section 2.6], let the (2-norm) condition number  $\kappa_2(A)$  of a square matrix  $A \in \mathbb{R}^{d \times d}$  be defined by

$$\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2$$

and set equal to  $\infty$  if  $A$  is not invertible.

If  $A$  is real, symmetric and positive definite, this corresponds to the ratio of the largest to the smallest eigenvalue. In other words, for such an  $A$  we have

$$\kappa_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

In the optimization literature, the condition number of the problem is associated with the condition number of the Hessian  $H(\theta)$  of  $F$  at an arbitrary point  $\theta \in \Theta$ , if it exists. Specifically, for twice differentiable,  $L$ -smooth,  $\mu$ -strongly convex  $F$ , we have noted that  $L \geq \lambda_{\max}(H(\theta))$  and  $\lambda_{\min}(H(\theta)) \geq \mu$  for all  $\theta \in \Theta$ . The ratio  $\kappa := L/\mu$  is referred to as the *condition number of  $F$*  in this context [Nesterov, 2004]. Note that we have for all  $\theta \in \Theta$  that

$$\kappa = \frac{L}{\mu} \geq \frac{\lambda_{\max}(H(\theta))}{\lambda_{\min}(H(\theta))} = \kappa_2(H(\theta)).$$

If  $H(\theta) = H$  is constant on  $\Theta$  (i.e. if  $F$  is quadratic or affine), we have  $L = \lambda_{\max}(H)$  and  $\mu = \lambda_{\min}(H)$  thus providing  $\kappa = \kappa_2(H)$ .

## 2. Background

### 2.4.6. Performance measures for smooth, strongly convex functions

Functions that are strongly convex and smooth satisfy the well-known and particularly useful set of inequalities

$$\forall \theta \in \Theta : \quad \frac{\mu}{2} \|\theta - \theta_*\|_2^2 \leq F(\theta) - F(\theta_*) \leq \frac{L}{2} \|\theta - \theta_*\|_2^2, \quad (2.4)$$

where  $\theta_* := \operatorname{argmin}_{\theta \in \Theta} F(\theta)$  is assumed to be an interior point of  $\Theta$ . The proof follows immediately from the inequalities presented in (2.1) and (2.2), since  $\nabla F(\theta_*) = 0$ .

This result is particularly useful as it gives a connection between the distance from the optimizer  $\theta_*$  in the domain and the suboptimality of the function value compared to the optimum value  $F(\theta_*)$ .

## 2.5. A characterization of large-scale learning

The issues summarized in this section are presented in a slightly more general setting in the insightful papers Bottou and Bousquet [2008] and Bottou [2010].

In order to illustrate the problems arising in large-scale learning settings, consider an arbitrary algorithm used to learn  $\theta_*$  which minimizes  $f(\theta) = \mathbb{E}[l(y, g_\theta(x))]$  for  $g_\theta \in \mathcal{G}_\Theta$ . Let  $f_0$  be the minimal achievable risk (e.g. for the square loss, the risk resulting from using the model function  $\tilde{g}(x) = \mathbb{E}[y|x]$ ). Typically, we will not achieve this benchmark and  $\inf_{\theta \in \Theta} f(\theta) > f_0$ , as the optimal model function will often not lie in the considered set  $\mathcal{G}_\Theta$ . No matter the learning algorithm and no matter the sample size or computation time, we will get the following deterministic *approximation error*:

$$\varepsilon_{\text{app}} = f(\theta_*) - f_0,$$

which only depends on the choice of  $\mathcal{G}_\Theta$  [Bottou and Bousquet, 2008]. An increase in  $\varepsilon_{\text{app}}$  follows from defining a more restricted class  $\mathcal{G}_\Theta$  or through decreasing the size of  $\Theta$ . For example, for the class of linear models, where functions in  $\mathcal{G}_\Theta$  are of the form  $g_\theta(x) = x^T \theta$  (see examples 1 and 2) restricting some coordinates of  $\theta$  to 0 and consequently omitting elements of the vector  $x$ , can lead to an increase in the approximation error.

Note that as explained in the introduction (see Chapter 1), the minimization of  $f(\theta)$  over  $\Theta$  will typically not be feasible, as  $f$  is unknown. Instead, one needs to estimate the solution  $\theta_*$ , for example by estimation of the function  $f(\theta)$  through  $\hat{f}_n(\theta) = n^{-1} \sum_{i=1}^n l(y_i, g_\theta(x_i))$ . Let  $\hat{\theta}_n$  denote a minimizer of  $\hat{f}_n$  over  $\Theta$ , i.e. an empirical risk minimizer (ERM), which will be assumed to exist almost surely. Note that  $\hat{\theta}_n$  depends on the data and is thus random. Using  $\hat{\theta}_n$  as an

estimator for  $\theta_*$  incurs the so-called *estimation error*, which is defined by

$$\varepsilon_{\text{est}} = \mathbb{E} \left[ f(\hat{\theta}_n) - f(\theta_*) \right]$$

and depends on the sample size  $n$  and the class  $\mathcal{G}_\Theta$  [Bottou and Bousquet, 2008].

Using  $\hat{\theta}_n$  as an estimator for  $\theta_*$  is now a feasible approach with a performance that will depend on the choice of  $\mathcal{G}_\Theta$  and the sample size  $n$ . The total error incurred by the ERM-approach is given by [Bottou and Bousquet, 2008]

$$\mathbb{E} \left[ f(\hat{\theta}_n) - f_0 \right] = \varepsilon_{\text{app}}(\mathcal{G}_\Theta) + \varepsilon_{\text{est}}(\mathcal{G}_\Theta, n)$$

and as the authors outline, minimizing this quantity is the goal in small-scale settings and leads to a trade-off in the choice of  $\mathcal{G}_\Theta$ , since a more restrictive choice will increase the first error but decrease the second.

The authors argue that in practice, the calculation of  $\hat{\theta}_n$  takes time, which has to be taken into account. Furthermore, exact calculation of  $\hat{\theta}_n$  might not be necessary, since  $\varepsilon_{\text{app}}$  and  $\varepsilon_{\text{est}}$  will persist anyway. This leads to the idea of only approximately calculating  $\hat{\theta}_n$  by defining some tolerance  $\delta > 0$  and using an estimator  $\tilde{\theta}_n(\delta)$  which satisfies  $\hat{f}_n(\tilde{\theta}_n(\delta)) \leq \hat{f}_n(\hat{\theta}_n) + \delta$  almost surely. This approximate optimization approach incurs the additional *optimization error*, defined by Bottou and Bousquet [2008] through

$$\varepsilon_{\text{opt}} = \mathbb{E} \left[ f(\tilde{\theta}_n(\delta)) - f(\hat{\theta}_n) \right].$$

The total error incurred by using an approximation of the ERM is then given by

$$\varepsilon = \mathbb{E} \left[ f(\tilde{\theta}_n(\delta)) - f_0 \right] = \varepsilon_{\text{app}}(\mathcal{G}_\Theta) + \varepsilon_{\text{est}}(\mathcal{G}_\Theta, n) + \varepsilon_{\text{opt}}(\mathcal{G}_\Theta, n, \delta).$$

Bottou and Bousquet [2008] argue further that, typically, using more data points  $n$  and decreasing the optimization tolerance  $\delta$ , will lead to smaller values of  $\varepsilon_{\text{est}}(\mathcal{G}_\Theta, n) + \varepsilon_{\text{opt}}(\mathcal{G}_\Theta, n, \delta)$  and are thus wanted. However, increases in  $n$  and decreases in  $\delta$  typically lead to more time  $T$  being needed to calculate  $\tilde{\theta}_n(\delta)$ . As a result, a trade-off between accuracy  $\delta$  in the optimization and usage of more data points  $n$  exists if the time  $T$  is limited. In addition, the trade-off in the approximation and the estimation error (resulting from the choice of  $\mathcal{G}_\Theta$ ) is still present.

More formally, Bottou and Bousquet [2008] and Bottou [2010] show that the trade-offs can be viewed as the problem of minimizing the given error subject to limitations on the used portion of the available data (i.e.  $n \leq n_{\text{max}}$ ) and the computation time (i.e.  $T \leq T_{\text{max}}$ ). They define *small-scale learning problems* as those, in which the bound on the sample size is active at the solution, while *large-scale learning problems* are defined as those, in which the bound on the computation time is active. For modern applications with an abundance of data and large function

## 2. Background

classes  $\mathcal{G}_\Theta$  (e.g. through large values of the dimension  $d$  of  $\Theta$ ), we find ourselves in the large-scale learning setting. As a consequence, algorithms that provide good approximate minimizers  $\tilde{\theta}_n(\delta)$  in a timely manner are highly important.

Deterministic gradient descent (GD) and stochastic gradient descent (SGD) are such algorithms and are very popular. They will be introduced in the following sections. After that, in Section 2.8, one method by Bottou [2010] to compare different procedures with varying computational cost is given, which uses asymptotic approximations of the time needed (in terms of the number of computations), in order to reach a certain optimization error  $\varepsilon_{\text{opt}}$  or respectively total error  $\varepsilon_{\text{app}} + \varepsilon_{\text{est}} + \varepsilon_{\text{opt}}$ .

Furthermore, there has been a lot of work to provide upper bounds on the sum  $\varepsilon_{\text{est}} + \varepsilon_{\text{opt}} = \mathbb{E} \left[ f(\tilde{\theta}_n(\delta)) - f(\theta_*) \right]$ , which is known as the *expected excess risk*, given a function class  $\mathcal{G}_\Theta$ . Comparing and analyzing such bounds for variants of SGD with the same runtime complexity will be the main interest of the subsequent chapters.

## 2.6. Gradient Descent (GD)

The method of (*deterministic*) *gradient descent*, *full gradient descent* or sometimes *batch gradient descent* (GD) originating from Cauchy [1847] is a versatile tool for finding the minimum of a function (should it exist), given information about the (sub-)gradient of the function. As opposed to the stochastic gradient descent method introduced in the next section, exact knowledge of the (sub-)gradient is required, i.e. no noise may be present.

The procedure is quite simple. Assume we have a function  $F$ , which we wish to minimize over  $\Theta$ , such as in the problems given in the introduction. Under the assumption that we can query  $\nabla F(t)$  (or alternatively a subgradient) at every point  $t \in \Theta$ , gradient descent is given through the iteration

$$t_{k+1} = t_k - \gamma_{k+1} \nabla F(t_k) \quad \text{for } k \in \{0, 1, 2, \dots\}$$

for some starting point  $t_0 \in \Theta$  and a *step size sequence* (or *learning rate sequence*)  $(\gamma_k)_{k \geq 1}$  with  $\gamma_k > 0$ . In certain situations, additional modifications can be necessary in order to ensure that  $t_k \in \Theta$  for all  $k$ .

The output of the algorithm is a deterministic sequence  $(t_k)_{k \geq 0}$  that is typically truncated and the last iterate is taken as the approximate minimizer of  $F$ .

While we will in practice not be able to use GD for the test error (i.e. the risk,  $F = f$ ) due to the lack of access to its gradient, one can use GD to approximately minimize the training error (i.e. the empirical risk,  $F = \hat{f}_n$ ). In the later case,  $(t_k)_{k \geq 0}$  is no longer a deterministic sequence, due to  $\hat{f}_n$  being stochastic. However, no new randomness is introduced through the algorithm. Lastly, note that GD fits in the framework of the preceding section by choosing  $\tilde{\theta}_n(\delta) = t_k$  for a suitable  $k = k(\delta)$ .

### 2.6.1. Basic performance results of GD

In Bubeck [2015] the author provides some well-known results on the convergence behavior of GD under different assumptions on  $F$  with minimizer  $t_*$ . The results can theoretically be applied to the risk  $F = f$ ,  $t_* = \theta_*$  or the empirical risk  $F = \hat{f}_n$ ,  $t_* = \hat{\theta}_n$ , even though the former will typically not be feasible. Let the learning rate sequence  $(\gamma_k)_{k \geq 1}$  be constant in the following, i.e.  $\gamma_k = \gamma$  for all  $k \geq 1$ .

Firstly, for an  $L$ -smooth, convex (but not necessarily strongly convex) function  $F$  with minimum  $t_*$ , GD satisfies the following inequality for  $\gamma = 1/L$  and all  $k \in \mathbb{N}$  [Bubeck, 2015]:

$$F(t_k) - F(t_*) \leq \frac{2L\|t_0 - t_*\|_2^2}{k}.$$

If we now assume that  $F$  is  $\mu$ -strongly convex and  $L$ -smooth, we get the following performance bound for  $\gamma = 1/L$  [Bubeck, 2015]:

$$\|t_k - t_*\|_2^2 \leq \left(1 - \frac{1}{\kappa}\right)^k \|t_0 - t_*\|_2^2 \leq e^{-k/\kappa} \|t_0 - t_*\|_2^2, \quad (2.5)$$

where  $\kappa = L/\mu$  is the condition number of the problem.

We see that this gives a linear rate of convergence for the strongly-convex case. Bubeck [2015] explains this speedup by inequality (2.3), which implies that the norm of the gradient is bounded below by the difference to the optimum. As a result, if the iterate  $t_k$  is far away from the optimum  $t_*$ , the gradient  $\nabla F(t_k)$  has to be sufficiently large (in terms of its norm), leading to a large step when producing  $t_{k+1}$ . However, the drawback of the result is the dependence on  $\kappa$ . Especially if the strong convexity is very weak, i.e.  $\mu$  is close to 0, arbitrarily large values of  $\kappa$  can occur. Here, considering the first inequality in (2.5),  $\kappa$  modulates the strength of the given contraction. Since  $\kappa \in [1, \infty)$ , the factor satisfies  $(1 - 1/\kappa) \in [0, 1)$ . Note that the larger  $\kappa$  is, the closer the factor will be to 1 and thus the smaller will be the decrease introduced by each additional step. For comparison with the result in the non-strongly convex case, note that we can use the second inequality in (2.4), in order to translate the result to a bound on the excess risk:

$$F(t_k) - F(t_*) \leq \frac{L}{2} e^{-k/\kappa} \|t_0 - t_*\|_2^2.$$

Lastly, Bubeck [2015] points out that with a slightly more elaborate step size, one can reduce the role of  $\kappa$  by a constant factor 1/4. Specifically, for  $\gamma = 2/(\mu + L)$ , they show that

$$F(t_k) - F(t_*) \leq \frac{L}{2} \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \|t_0 - t_*\|_2^2 \leq \frac{L}{2} e^{-4k/(\kappa+1)} \|t_0 - t_*\|_2^2.$$

Note that interestingly, it is not recommendable to increase  $\gamma$  to  $2/L$ , as will be illustrated for the case of a quadratic target function in Subsection 3.3.1.

## 2. Background

It should be noted that for smooth, strongly convex functions, a linear convergence rate is the best we can hope for. In the case of smooth, convex functions, gradient descent can generally not be guaranteed to perform better than  $O(1/k)$ , while other methods can achieve  $O(1/k^2)$  [Nemirovski, 1994]. This will be discussed shortly in Subsection 2.9.1.

### 2.7. Stochastic Gradient Descent (SGD)

Deterministic gradient descent is often not appropriate. One reason for this can be that  $\nabla F(\theta)$  cannot be queried. For example, for the risk minimization problem (1.1),  $\nabla f(\theta)$  is simply unknown. In addition, even if  $\nabla F(\theta)$  can be calculated at any  $\theta \in \Theta$ , as for example in the empirical risk minimization problem (1.2), it might be costly and we would like to replace  $\nabla F(\theta)$  by a quantity that is similar, yet easier to compute. In both cases,  $\nabla F(\theta)$  is frequently replaced by a noisy estimator. In general, such an estimator can differ between iterations. Given such estimators  $\phi_k(\theta)$ ,  $k \geq 1$ , which are typically assumed to have mean  $\nabla F(\theta)$  for all  $\theta \in \Theta$ , *stochastic gradient descent* (SGD) is defined through

$$\theta_k = \theta_{k-1} - \gamma_k \phi_k(\theta_{k-1}) \quad (2.6)$$

for all  $k \geq 1$ , some starting point  $\theta_0 \in \Theta$  and a step size sequence  $(\gamma_k)_{k \geq 1}$  with  $\gamma_k > 0$ . For the application to (empirical) risk minimization considered here, we often have a function  $\psi$  such that  $\phi_k(\theta) = \psi(\theta, y_{i_k}, x_{i_k})$ . This means that differences in the functions  $(\phi_k)_{k \geq 1}$  stem from the fact that different sample points  $(y_{i_k}, x_{i_k})$  may be used to calculate it at different iterations.

In this section, we will look at SGD for empirical risk minimization and risk minimization. Afterwards, basic asymptotic convergence results of SGD will be presented and compared to performing GD on the empirical risk minimization problem. Lastly, some basic adaptations of SGD are presented.

#### 2.7.1. SGD for empirical risk minimization

For the empirical risk minimization problem of (1.2), we could simply use GD. To that end, we would use the gradients of  $\hat{f}_n$ , i.e.

$$\nabla \hat{f}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta} l(y_i, g_{\theta}(x_i)).$$

The problem with that approach is that we would have to evaluate  $n$  gradients of length  $d$  at each step, which can be computationally cumbersome. If the gradients  $\nabla_{\theta} l(y_i, g_{\theta}(x_i))$  are similar (or even identical) for all  $i$ , then evaluating all of them and calculating the averaged gradient serves little purpose. This is explained in Bottou et al. [2018], where the authors propose the

following thought experiment. Imagine you have a sample  $S = (x_1, \dots, x_n)$  of size  $n$  and another sample  $S_{10}$  which consists of the same elements as  $S$ , but each element occurs ten times, i.e.  $S_{10} = (x_1, x_1, \dots, x_1, x_2, x_2, \dots, x_2, \dots, x_n, x_n, \dots, x_n)$ . The length of  $S_{10}$  is therefore  $10n$ . As a result, the averaged gradient will be the same, no matter whether we have  $S$  or  $S_{10}$ , but for the latter case we require ten times the number of gradient evaluations needed for the former case. This leads to the idea of stochastic gradient descent, where we only evaluate one gradient in each iteration. If the index  $i$  of the respective gradient is chosen uniformly at random in each iteration, the applications to  $S$  and  $S_{10}$  are indistinguishable in the sense that the stochastic processes  $(\theta_k)_{k \in \mathbb{N}}$  resulting from the iterations have the same distribution. In both cases we notably need only one gradient evaluation per iteration, meaning that the cost of each iteration does not depend on the sample size  $n$ .

Formalizing this idea in the framework of (2.6), we can choose  $i_k$  uniformly at random from  $\{1, \dots, n\}$  at each iteration  $k$  and use  $\psi(\theta, y, x) = \nabla_{\theta} l(y, g_{\theta}(x))$ , such that we have

$$\phi_k(\theta) = \nabla_{\theta} l(y_{i_k}, g_{\theta}(x_{i_k})).$$

Sources of randomness are the data points and the randomly selected indices. Conditionally on the data points,  $\phi_k(\theta)$  is obviously unbiased for  $\nabla \hat{f}_n(\theta)$  for all  $\theta \in \Theta$ , since  $i_k$  is chosen uniformly at random from  $\{1, \dots, n\}$ .

In a non-truncated algorithmic form, we can describe SGD for the empirical risk minimization problem as follows [Schmidt et al., 2017].

---

**Algorithm 1:** Stochastic gradient descent for empirical risk minimization

---

```

1 Choose  $\theta_0 \in \Theta$  and  $(\gamma_k)_{k \geq 1}$ 
2 for  $k \in \{1, 2, \dots\}$  do
3   | Choose  $i_k \sim \mathcal{U}(\{1, \dots, n\})$  independently from previous iterations
4   | Set  $\theta_k = \theta_{k-1} - \gamma_k \nabla_{\theta} l(y_{i_k}, g_{\theta}(x_{i_k}))|_{\theta=\theta_{k-1}}$ 
5 end
```

---

Note that if  $\Theta \neq \mathbb{R}^d$ , adaptations might be necessary in order to ensure that  $\theta_k \in \Theta$  for all  $k$ . Details are given in Subsection 2.7.3.

To summarize, in the case of empirical risk minimization SGD can be used to decrease the computational burden of each step at the cost of introducing additional randomness.

## 2.7.2. SGD for risk minimization

In risk minimization, calculation of  $\nabla f(\theta)$  will typically not merely be cumbersome, but will simply be impossible, as information on the distribution of the data points would be required.

## 2. Background

We therefore need to estimate  $\nabla f(\theta) = \nabla_{\theta} \mathbb{E} [l(y, g_{\theta}(x))]$  at each step. Under mild regularity conditions on the loss, one can show that  $\nabla_{\theta} \mathbb{E} [l(y, g_{\theta}(x))] = \mathbb{E} [\nabla_{\theta} l(y, g_{\theta}(x))]$  for all  $\theta \in \Theta$ , making  $\phi_k(\theta) = \nabla_{\theta} l(y_k, g_{\theta}(x_k))$  appealing unbiased estimators for  $\nabla f(\theta)$ .

If  $f(\theta)$  should be minimized based on the data, it is clear that any finite sample size  $n$  will in general be insufficient. Instead of dealing with this issue by using empirical risk minimization, one can use the approach of online learning if there is an abundance of data, as is the case in the large-scale setting described earlier. Algorithm 1 then naturally extends to the following non-truncated algorithm [Bottou et al., 2018].

---

**Algorithm 2:** Stochastic gradient descent for online learning

---

```

1 Choose  $\theta_0 \in \Theta$  and  $(\gamma_k)_{k \geq 1}$ 
2 for  $k \in \{1, 2, \dots\}$  do
3   | Generate a data point  $(y_k, x_k)$  independently from the previous iterations
4   | Set  $\theta_k = \theta_{k-1} - \gamma_k \nabla_{\theta} l(y_k, g_{\theta}(x_k))|_{\theta=\theta_{k-1}}$ 
5 end
```

---

Again, adaptations might be necessary to make sure that  $\theta_k \in \Theta$  for all  $k$  if  $\Theta \neq \mathbb{R}^d$ .

Here, instead of sampling from a finite data set, we generate a new data point at each iteration and use it only once. As a result, we only perform a single pass through the data, while in the previous algorithm several passes through the data were allowed. Algorithm 2 therefore needs to be stopped, when no new data points can be generated, i.e. all available data has been used. However, by the definition of large-scale learning, the algorithm will typically be stopped before that due to time constraints.

Lastly, note that it is in line with the notation used so far to denote  $\theta_k$  by  $\theta_n$  for online learning SGD, since we use  $k$  data points in order to calculate  $\theta_k$ .

In order to get a first impression of the performance of SGD, note that with a properly tuned step size sequence, SGD can decrease the expected excess risk at a rate of  $O(1/\sqrt{n})$  if the target function is convex, but not strongly convex, and a rate of  $O(1/n)$  if strong convexity is assumed. Interestingly, for SGD smoothness does not affect the rate, while it does for GD [Bach and Moulines, 2013].

### 2.7.3. Variants of SGD

There are several ways in which SGD can be adapted through algorithmic primitives. Some of them will be of special interest in the following chapters and will therefore be introduced here. It should be noted that the first two adaptations might be required due to the structure of the

problem, while the subsequent adaptations can be used voluntarily in order to achieve a certain behavior of the estimator sequence  $(\theta_k)_{k \geq 1}$ , such as better convergence rates or higher robustness to the randomness in the data.

**Projection steps.** If  $\Theta \neq \mathbb{R}^d$ , it might be necessary to ensure that the iterates  $\theta_k$ ,  $k \geq 1$ , of SGD lie within  $\Theta$ , as expressions like  $\phi_{k+1}(\theta_k)$  might otherwise not be defined and convexity or smoothness assumptions might not hold outside of  $\Theta$ . Typically, if  $\Theta \neq \mathbb{R}^d$ , it is assumed that  $\Theta \subseteq \mathbb{R}^d$  is non-empty, convex and compact. In the context of SGD,  $\theta_k \in \Theta$  can be guaranteed by replacing the update step in (2.6) with

$$\theta_k = \pi_{\Theta}(\theta_{k-1} - \gamma_k \phi_k(\theta_{k-1})),$$

where  $\pi_{\Theta} : \mathbb{R}^d \rightarrow \Theta$  given by  $\pi_{\Theta}(\theta) := \operatorname{argmin}_{\theta' \in \Theta} \|\theta - \theta'\|_2$  denotes the projection onto  $\Theta$  with respect to the Euclidean norm (see for example Nemirovski et al. [2009]). Since  $\Theta$  is non-empty, closed and convex, the expression is well defined.

**Subgradient method.** If the target function  $F$  is not differentiable on  $\Theta$ , one needs to adapt the stochastic gradient descent method to the *stochastic subgradient descent* method. The only change is that the random direction used in the update step can formally no longer be required to be an unbiased estimator of the gradient of  $F$  at the given point. Instead, the expected value of the random direction has to be a subgradient of  $F$ . For an example see Nemirovski et al. [2009], where the authors start without any smoothness assumptions. This adaptation of SGD is a straight-forward generalization, as the set of subgradients only contains one vector, namely the gradient, if  $F$  is differentiable.

**Mini-batching.** For the application to the empirical risk minimization problem (1.2), GD and SGD are both applicable. While GD uses a costly, but exact term for the direction of the steepest descent of the type  $n^{-1} \sum_{i=1}^n \nabla_{\theta} l(y_i, g_{\theta}(x_i))$ , SGD uses cheaper but inexact estimators  $\nabla_{\theta} l(y_{i_k}, g_{\theta}(x_{i_k}))$  with  $i_k$  being chosen at random from  $\{1, \dots, n\}$  (see Algorithm 1). A compromise is given by using a mini-batch approach, which naturally extends to the case of online SGD (see Algorithm 2).

For this approach, a mini-batch size  $b \in \{1, \dots, n\}$  is defined, which determines the number of samples to be used for each step. In the finite sample setting, this leads to drawing a subset  $J_k = \{i_k^{(1)}, \dots, i_k^{(b)}\}$  of size  $b$  without replacement from the indices at iteration  $k$ , instead of just one  $i_k$ . The update step for Algorithm 1 is then given by [Bottou et al., 2018]

$$\theta_k = \theta_{k-1} - \gamma_k \frac{1}{b} \sum_{i \in J_k} \nabla_{\theta} l(y_i, g_{\theta_{k-1}}(x_i)).$$

Note that GD is recovered for  $b = n$  if the indices in  $J_k$  are sampled from  $\{1, \dots, n\}$  without replacement.

## 2. Background

For the case of online learning, mini-batching amounts to generating  $b$  observations

$$S_k = \left\{ \left( y_k^{(1)}, x_k^{(1)} \right), \dots, \left( y_k^{(b)}, x_k^{(b)} \right) \right\}$$

at iteration  $k$ , each of which is an i.i.d. copy of  $(y, x)$ . The following update step in Algorithm 2 is then used [Jain et al., 2018b]:

$$\theta_k = \theta_{k-1} - \gamma_k \frac{1}{b} \sum_{(y,x) \in S_k} \nabla_{\theta} l(y, g_{\theta_{k-1}}(x)).$$

Again, SGD is recovered for  $b = 1$ .

In the spirit of keeping the runtime complexity unaffected, it should be noted that each iteration of mini-batching SGD uses  $b$  data points and thus has a computational cost of  $O(b)$ . As a result, if we want to evaluate the benefit of using mini-batching with respect to the estimator achieved through  $k$  iterations of regular SGD, it is only fair to do so with respect to the  $\lfloor k/b \rfloor$  iteration through mini-batching SGD, as this iterate is the last that does not exceed the computational cost or the number of samples used by regular SGD after  $k$  steps. Such an approach is used for example in Jain et al. [2018b].

Some performance results for SGD with mini-batching are provided in Chapter 4.

**(Tail-)Averaging.** One important adaption of SGD algorithms is the idea of iterate averaging (see for example Polyak and Juditsky [1992]). Formally, this means that if  $(\theta_k)_{k \geq 0}$  is a sequence of estimators achieved through SGD after starting at  $\theta_0$ , the averaged iterate is defined through

$$\bar{\theta}_k := \frac{1}{k+1} \sum_{i=0}^k \theta_i.$$

The procedure of creating the sequence of estimators  $(\bar{\theta}_k)_{k \geq 0}$  is called *averaged SGD* and will be of specific interest in this thesis.

One frequently employed generalization of averaged SGD is *tail-averaged SGD* (see for example Nemirovski et al. [2009]). The idea is simple; instead of averaging all previous iterates of SGD, only the last few are averaged. More specifically, tail-averaged SGD is given through

$$\bar{\theta}_j^k := \frac{1}{k-j+1} \sum_{i=j}^k \theta_i$$

for all  $k \geq j \geq 0$ . Note that averaged SGD is recovered for  $j = 0$ . Note further that  $j$  does not have to be a fixed constant, but is allowed to depend on  $k$ . For example,  $j = \lfloor k/2 \rfloor$  gives the tail-averaged SGD estimator that uses only the last half of the iterations of SGD.

Lastly, one further generalization is *weighted tail-averaged SGD* (see, again, for example Nemirovski et al. [2009]). For some weights  $\nu_i \geq 0$ ,  $i \in \{j, j+1, \dots, k\}$ , with  $\sum_{i=j}^k \nu_i = 1$ , weighted

tail-averaged SGD is given through

$$\bar{\theta}_j^k := \sum_{i=j}^k \nu_i \theta_i$$

where  $k \geq j \geq 0$  and the values of  $\nu_i$  might, of course, depend on  $j$  and  $k$ .

Note that averaging is carried out in parallel, meaning that in the  $k$ -th step, the gradient estimator  $\phi_k$  is evaluated at  $\theta_{k-1}$  and not at  $\bar{\theta}_{k-1}$  (or  $\bar{\theta}_j^{k-1}$ ). There are other approaches to averaging in SGD, such as Bather’s algorithm (see Schwabe and Walk [1996]), where the average over all previous iterates is taken as the point for the next iteration step. This different version of averaging will not be examined further in this thesis.

Specific benefits of (tail-)averaging will be discussed in Chapters 3 and 4. To give some basic intuition, using the averaged sequence  $\bar{\theta}_k$  (or  $\bar{\theta}_j^k$ ) instead of  $\theta_k$  can help to reduce the variability in the sequence. A different justification of using tail-averaging is given in Nemirovski [1994], where the author argues that other than in deterministic optimization, we do not know which of the iterates in the trajectory  $(\theta_0, \theta_1, \dots, \theta_k)$  actually results in the lowest value of the target function. However, the expectation that later iterates tend to be better motivates a moving window tail-averaging approach of taking the average over the second half of the iterates.

**Further variants.** There is a long list of further variants one could include here. Examples include implicit SGD [Toulis and Airoidi, 2017], stochastic variance reduced gradient descent (SVRG) [Frostig et al., 2015] or Bather’s algorithm mentioned in the previous paragraph [Schwabe and Walk, 1996]. A further popular extension of SGD is the use of momentum terms, where the update step from  $\theta_{k-1}$  to  $\theta_k$  is extended by an additive momentum term  $\beta_k(\theta_{k-1} - \theta_{k-2})$ , for a sequence  $\beta_k > 0$ ,  $k \geq 1$  (see for example Roux et al. [2012] or Dieuleveut et al. [2017]). As pointed out by Roux et al. [2012], another variant is to decrease the step size if the inner product of subsequent gradient estimates is negative, which is a modification known as *acceleration*.<sup>1</sup> Some explanation of why this is sensible is given in Chapter 4.

A particularly interesting algorithm for solving the empirical risk minimization problem (1.2) was proposed by Roux et al. [2012] (see also Schmidt et al. [2017]), where the authors extend Algorithm 1 by storing and averaging gradient estimates in a manner different to mini-batching. For the training error, i.e. the suboptimality in the empirical risk minimization problem, the authors recover the linear convergence rate known from deterministic GD without losing the favorable computational properties of SGD. As the authors point out, the convergence does of course not translate to the testing error, i.e. the risk minimization problem, which is why this result will be of no further interest for this thesis.

---

<sup>1</sup>Roux et al. [2012] note that the term *acceleration* has a different meaning in the context of deterministic gradient descent.

## 2.8. Comparison of GD and SGD

For a smooth optimization problem with properly tuned step sizes, GD will achieve faster convergence than SGD in terms of the numbers of steps. This poses the question, why one would ever use SGD if  $\nabla F(\theta)$  can be queried at every  $\theta \in \Theta$ . Indeed, even in those cases, SGD can be preferable over GD, due to the lower complexity.

This section continues the ideas of Section 2.5. There, it was discussed that in order to find the minimizer  $\theta_*$  of the risk  $f$  over  $\Theta$ , one might approximately minimize the empirical risk  $\hat{f}_n$  through the usage of an iterative algorithm, such as GD or SGD. GD will converge faster to the empirical risk minimizer  $\hat{\theta}_n$ , namely linearly under strong convexity assumptions, whereas SGD can only hope for a sublinear convergence of rate  $O(1/k)$  in this setting. However, GD is computationally much more cumbersome than SGD and  $\hat{f}_n$  and  $\hat{\theta}_n$  are only estimators for  $f$  and  $\theta_*$  anyway. Therefore, using a lot of computational resources to get a very precise estimator for  $\hat{\theta}_n$  might be in vain. This idea will be presented in a little more rigor in this section, in order to argue why it is sensible to consider the properties of SGD instead of focusing only on GD.

The guiding results presented in this section hold under certain regularizing assumptions, namely linear parameterization of the functions  $g_\theta \in \mathcal{G}_\Theta$ , boundedness of  $\Theta$ , almost sure boundedness of the inputs  $(y_i, x_i)$ ,  $i \geq 1$ , and Lipschitz-continuity of  $l(y, \cdot)$  for all  $y$  [Bottou and Bousquet, 2008]. These assumptions are admittedly quite restrictive and are imposed by Bottou and Bousquet [2008] to keep their paper short.

Consider situations in which GD can minimize the empirical risk at a linear convergence rate of  $O(\rho^k)$  for some  $\rho \in [0, 1)$  and finite-sample SGD (i.e. Algorithm 1) at a sublinear rate of  $O(1/k)$  (e.g. situations with a smooth and strongly convex target function).

As argued in Bottou [2010] and Bottou et al. [2018], it follows that in order to reach some small predefined tolerance  $\delta$  (in expectation with respect to the selected indices  $i_k$ ,  $k \geq 1$ ), GD needs proportional to  $\log(1/\delta)$  iterations, while SGD needs proportional to  $1/\delta$  iterations. The fact that GD needs less iterations than SGD is not surprising, given that each iteration information from all data points  $n$  is used in GD, while SGD only uses information from one data point. In return, the number of calculations needed for GD to achieve tolerance  $\delta$  is proportional to  $n \log(1/\delta)$ , while this quantity is still only proportional to  $1/\delta$  for SGD [Bottou et al., 2018]. As a result, for sufficiently large values of  $n$ , GD will take more computation time to achieve tolerance  $\delta$  than SGD, making SGD more appealing for large  $n$  and fixed  $\delta$ . Conversely, if small values of  $\delta$  are the main concern and  $n$  is moderate, GD is more suitable than SGD.

This shows that already for the task of finding the empirical risk minimizer, SGD can be preferable. The statement becomes much stronger for the task of risk minimization.

In order to make conclusions for the risk minimization problem given in (1.1), uniform laws of large numbers for linearly parametrized functions can be employed, so that the effects of  $\delta$ -

optimal minimization of the empirical risk function  $\hat{f}_n$  can be carried over to the risk function  $f$ . Bottou and Bousquet [2008] provide one such law which is given through

$$\mathbb{E} \left[ \sup_{\theta \in \Theta} |f(\theta) - \hat{f}_n(\theta)| \right] \leq c \sqrt{\frac{d}{n} \log \left( \frac{n}{d} \right)},$$

where  $c > 0$  is a constant and  $n > d$  is implicitly assumed, or simply

$$\mathbb{E} \left[ \sup_{\theta \in \Theta} |f(\theta) - \hat{f}_n(\theta)| \right] = O \left( \sqrt{\frac{\log(n)}{n}} \right).$$

Bottou et al. [2018] show that this can be used to achieve an upper bound for  $\varepsilon_{\text{est}} + \varepsilon_{\text{opt}}$  by using the following decomposition into differences to the empirical risk:

$$\begin{aligned} \varepsilon_{\text{est}} + \varepsilon_{\text{opt}} &= \mathbb{E} \left[ f(\hat{\theta}_n) - f(\theta_*) \right] + \mathbb{E} \left[ f(\tilde{\theta}_n(\delta)) - f(\hat{\theta}_n) \right] \\ &= \mathbb{E} \left[ f(\tilde{\theta}_n(\delta)) - f(\theta_*) \right] \\ &= \mathbb{E} \left[ f(\tilde{\theta}_n(\delta)) - \hat{f}_n(\tilde{\theta}_n(\delta)) \right] + \mathbb{E} \left[ \hat{f}_n(\tilde{\theta}_n(\delta)) - \hat{f}_n(\hat{\theta}_n) \right] \\ &\quad + \mathbb{E} \left[ \hat{f}_n(\hat{\theta}_n) - \hat{f}_n(\theta_*) \right] + \mathbb{E} \left[ \hat{f}_n(\theta_*) - f(\theta_*) \right]. \end{aligned}$$

For the first and the fourth expected value, the arguments within the functions are identical, making the application of the given uniform law of large numbers possible. The difference in the second expectation is almost surely smaller or equal to  $\delta$  by construction of the approximate ERM  $\tilde{\theta}_n(\delta)$ . Lastly, the difference in the third expectation is almost surely negative, since the ERM  $\hat{\theta}_n$  minimizes  $\hat{f}_n(\theta)$ . This gives the result:

$$\varepsilon_{\text{est}} + \varepsilon_{\text{opt}} = \delta + O \left( \sqrt{\frac{\log(n)}{n}} \right).$$

Bottou [2010] states that this bound can typically be improved to the following asymptotic equality for large  $n$ :

$$\varepsilon_{\text{est}} + \varepsilon_{\text{opt}} \sim \delta + \left[ \frac{\log(n)}{n} \right]^\alpha$$

for some  $\alpha \in [1/2, 1]$  depending on the assumptions. Here, the symbol  $\sim$  denotes asymptotic equality up to constants, as introduced in Section 2.1. For example, Bottou et al. [2018] note that for strongly convex loss functions or when the distribution of the data satisfies certain assumptions, the statement  $\varepsilon_{\text{est}} + \varepsilon_{\text{opt}} = \delta + O(\log(n)/n)$  holds true (i.e.  $\alpha = 1$ ).

Bottou [2010] argues further that since the convergence speed of a sum is determined by its slowest term, each component ( $\varepsilon_{\text{app}}$ ,  $\varepsilon_{\text{est}}$  and  $\varepsilon_{\text{opt}}$ ) of the error  $\varepsilon = \mathbb{E} \left[ f(\tilde{\theta}_n(\delta)) - f(\theta_*) \right]$  should decrease at the same rate of  $[\log(n)/n]^\alpha$ . This implies that  $\delta$  should be chosen in dependence of  $n$  (i.e.  $\delta = \delta_n$ ) and decreased at this rate. Analogously,  $\mathcal{G}_\Theta$  should be chosen (in dependence of

## 2. Background

$n$ ) such that  $\varepsilon_{\text{opt}}$  can decrease at this rate. If the trade-off discussed in Section 2.5 is solved with respect to this rule, the asymptotic (in  $n \rightarrow \infty$ ) behavior of  $\varepsilon = \varepsilon_n$  becomes

$$\varepsilon \sim \left[ \frac{\log(n)}{n} \right]^\alpha,$$

as noted by Bottou [2010]. To summarize:

$$\varepsilon \sim \left[ \frac{\log(n)}{n} \right]^\alpha \sim \delta.$$

Finally, the authors reach the following conclusions with respect to the cost of reaching a predefined error  $\varepsilon$  through GD and through SGD.

First, consider GD. Recall that GD needs proportional to  $n \log(1/\delta)$  computations to reach tolerance  $\delta$ . Reformulating this result on the computational cost in terms of  $\varepsilon$  by using the above asymptotic equalities, leads to a computational cost proportional to

$$\frac{1}{\alpha \varepsilon^{1/\alpha}} \log^2 \left( \frac{1}{\varepsilon} \right)$$

of reaching an error of  $\varepsilon$  in the risk problem when using GD [Bottou, 2010]. The proof for this last result is given in Appendix A.1.

Lastly, consider the same reformulation in terms of  $\varepsilon$  for SGD. There, the cost to reach tolerance  $\delta$  was given by  $1/\delta$ , which was notably independent of  $n$  (for fixed  $\delta$ ). Using the asymptotic inequality  $\varepsilon \sim \delta$  shows that, asymptotically, the computational cost of reaching an error of  $\varepsilon$  in the problem (1.1) when using SGD is proportional to  $1/\varepsilon$  [Bottou, 2010].

Comparing the asymptotic (in  $n$ ) cost  $1/(\alpha \varepsilon^{1/\alpha}) \log^2(1/\varepsilon)$  for GD and  $1/\varepsilon$  for SGD illustrates that for large  $n$  (i.e. in the large-scale learning setting) GD is computationally more cumbersome for all  $\alpha \in [1/2, 1]$  than SGD.

To summarize the results so far, in Section 2.5, it was argued why approximate calculation of the ERM (i.e. with tolerance  $\delta > 0$ ) can be preferable over exact optimization, depending on the algorithm. This section now shows that for problems with sufficiently large  $n$ , under optimal tuning of  $\delta = \delta_n$  and  $\mathcal{G}_\Theta$ , SGD is preferable to GD with respect to the risk minimization problem and possibly even the empirical risk minimization problem. However, these results give by no means a complete picture of the problems involved.

Lastly, some caveats with respect to the results presented here need to be mentioned:

- The appropriate choice of the step size needed to ensure the results from this section often times depends critically on unknown quantities.
- The presented quantities proportional to the asymptotic cost omit constant factors. Most notably, the condition number  $\kappa$  does not appear. Bottou and Bousquet [2008] give some

insight here and show that (under adequate tuning of the step sizes), the cost for GD includes a factor  $d^2\kappa$  and the one for SGD a factor  $d\kappa^2$ . This can make GD interesting again for high-dimensional settings and can be problematic for both (but especially SGD) if the condition number  $\kappa$  is large.

- The results presented in this section give the asymptotic behavior, but do not yield much information on the behavior of the algorithms for finite  $n$ .

These issues and their importance in the context of least squares regression will reappear in Chapters 3 and 4.

## 2.9. The concept of optimality

There has been a lot of work with one common goal: finding an algorithm that behaves “optimally” while still achieving the computational cost per iteration of SGD. This has led to a vast amount of literature that deals with adaptations of SGD like the ones presented in Subsection 2.7.3, limitations on the permitted learning rate  $(\gamma_k)_{k \geq 1}$  and restrictions on  $F$  and  $\Theta$ . The question is, how the (sub-)optimality of different algorithms can be quantified if their complexity does not need to be accounted for due to being equal or simply of no interest.

In order to analyze the behavior of an algorithm outputting approximate minimizers  $(\tilde{\theta}_k)_{k \geq 0}$  on its search for a minimizer  $t_*$  (or a set of minimizers  $T_* = \{t \in \Theta : F(t) = \inf_{\theta \in \Theta} F(\theta)\}$ ), one first needs a measure to quantify how good one specific point  $\tilde{\theta}_k$  is. In the literature, different performance measures exist.

In this thesis, we mostly deal with the excess risk of an estimator  $\tilde{\theta}_k \in \Theta$

$$f(\tilde{\theta}_k) - f(\theta_*)$$

and in particular its expectation, i.e. the quantity

$$\mathbb{E} \left[ f(\tilde{\theta}_k) - f(\theta_*) \right],$$

where  $f : \Theta \rightarrow \mathbb{R}$  is the risk function and  $\theta_*$  minimizes  $f$  over  $\Theta$ .

For a general optimization problem  $F$ , this quantity, i.e.

$$\mathbb{E} \left[ F(\tilde{\theta}_k) - F(t_*) \right],$$

is known as the *accuracy* in the optimization literature. The expectation vanishes of course if  $\tilde{\theta}_k$  is deterministic (e.g. in GD). Here,  $t_*$  minimizes  $F$  over  $\Theta$ . One significant advantage of this measure is that it allows for multiple optima and we can even define the accuracy if we only have a reference value  $F_*$  which is not realized at any point in  $\Theta$  (e.g. as could happen when defining  $F_* = \inf_{t \in \Theta} F(t)$  for a non-compact  $\Theta \subseteq \mathbb{R}^d$  or a discontinuous  $F$ ). As discussed in Nemirovsky

## 2. Background

and Yudin [1983], normalization might be necessary for comparability, as otherwise rescaling of the target function  $F$  could be used to improve the accuracy.

A different measure that is of particular interest in the context of risk minimization (i.e.  $F = f$ ,  $t_* = \theta_*$ ) is the mean squared error

$$\mathbb{E} \left[ \left\| \tilde{\theta}_k - \theta_* \right\|_2^2 \right],$$

where we do in fact need exactly one minimizer  $\theta_* \in \Theta$  of  $f$  for this expression to be meaningful. As discussed in Section 2.4 through the inequalities in (2.4), the mean squared error and the expected excess risk only differ in constants depending on  $\mu$  and  $L$  if the target function is  $L$ -smooth and  $\mu$ -strongly convex.

Lastly, some are not satisfied with results in expectation and require instead concentration bounds. For example for the case of least squares regression, in Bach and Moulines [2013], the authors not only provide a non-asymptotic upper bound on the expected excess risk, but also give upper bounds in high probability on the excess risk.

We will from now on mainly focus on the accuracy (i.e. the expected excess risk). Note that this quantity (as well as the other ones just discussed) is a measure of suitability of an algorithm for one specific target function, given a specific input (e.g. observed data points). Typically, however, we are confronted with a class of target functions and different types of input. We would then like the algorithm to perform well on the entire class, in a sense which will be specified in the next paragraph.

First, however, note that this goal is by no means too ambitious. Take the simple case of least squares regression presented in Example 1. The target function here is  $f(\theta) = \frac{1}{2} \theta^T \mathbb{E} \left[ x_1 x_1^T \right] \theta - \theta^T \mathbb{E} [x_1 y_1] + \frac{1}{2} \mathbb{E} [y_1^2]$  and the source of information are i.i.d. observations  $(x_i, y_i)$ ,  $i \geq 1$ . The specific target function  $f$  and the available information now depend on the joint distribution of  $x_i$  and  $y_i$ , which we do not know and do not wish to restrict too much. As a result, we would like our algorithm to perform well for all instances of the least squares regression problems resulting from different joint distributions.

As Jain et al. [2018a] put it, there are at least two approaches to the notion of performing well on the entire class of problems in this context. One can view these two approaches as the different points of view by statistics and optimization, at the intersection of which the problems at hand lie.

One group aims at finding an algorithm that produces a series of estimators which recover (at least asymptotically) the properties of the known statistically optimal estimator with respect to the chosen measure of optimality (e.g. accuracy or expected excess risk) for every single instance in the considered class of problems.

The other group aims at finding an algorithm that performs well for the worst problem in the considered class. This is formalized through the notion of *minimax-optimality*, i.e. the maximum

accuracy over the class of problems should be as low as possible. For several classes of problems, the limits of what is possible are given by Nemirovsky and Yudin [1983], who define notions of problem complexity and whose work is still of high relevance and interest today.

In the following subsections, these two notions of optimality are described further and selected benchmarks from the literature are given. Lastly, it should be noted that numerical optimization algorithms, such as variants of (stochastic) gradient descent, typically require the user to set some hyperparameters, such as the starting point or the step size, which can be difficult. Therefore, another goal is to have the algorithm depend on these hyperparameters in a benign way.

### 2.9.1. Optimality and problem complexity in oracle-based optimization

The field of problem complexity in optimization was pioneered by Nemirovsky and Yudin [1983], where the authors create the necessary framework to evaluate the laboriousness of function classes, based on iterative algorithms. Parts of their work are summarized in the lecture notes by Nemirovski [1994]. In short, Nemirovsky and Yudin [1983, Chapter 1] formalize how for a given accuracy  $\nu > 0$ , a given problem class and a given oracle, one can determine the minimum number of oracle calls required in order to be able to guarantee  $\nu$ -optimality. They achieve this by defining a complexity measure  $N(\nu)$  on a problem class, which denotes the minimum number of oracle calls for which an iterative method exists that almost surely needs no longer to reach the accuracy  $\nu$  over all problems in the considered class. Nemirovsky and Yudin [1983] further construct lower and upper bounds of  $N(\nu)$  for a variety of cases. To that end, they define the term *method* rather broadly, as any iterative search rule that has access to all previously acquired information to decide on the next step. This acquired information includes the previous iterates and (possibly noisy) oracles called at the previous iterates. In addition, the next iterate can either be given directly or through a probability distribution parameterized through the previously acquired information. The following results hold for methods of the former type, where the next step is given directly, i.e. only depends on the previously acquired information and the randomness therein.

If  $N(\nu)$  is known for the problem class of interest, we can immediately judge, whether a method is minimax-optimal for this class, based on how many steps the method maximally needs to reach accuracy  $\nu$ . This makes calculating (or at least bounding)  $N(\nu)$  an interesting task. As pointed out by Nemirovski [1994], having a method that requires maximally  $n_0(\nu)$  oracle calls to reach accuracy  $\nu$  on any problem in the considered class immediately gives an upper bound  $n_0(\nu) \geq N(\nu)$ . Given a theoretic lower bound  $n_1(\nu)$ , it is, of course, desirable to construct a method, which (at least up to constants) achieves  $n_0(\nu) = n_1(\nu)$ , as this implies that not only the lower bound  $n_1(\nu)$  on  $N(\nu)$  is tight, but also that the method is essentially unimprovable with respect to the worst case.

## 2. Background

### Complexity results for exact first order oracles

The following results hold under the assumption that the used method has access to exact first order oracles. This means that while we may not have knowledge of the entire target function  $F$ , we can at every point  $t \in \Theta$  calculate  $F(t)$  and  $\nabla F(t)$ . As a result, no randomness is present in the approximation methods considered in the following. For now,  $t_k$ ,  $k \geq 1$ , denotes the deterministic sequence of approximate minimizers achieved through  $k$  steps of a generic method starting with  $t_0$ .

Nemirovski [1994] presents for the class of unconstrained,  $L$ -smooth, convex problems on  $\mathbb{R}^d$  (for which it is assumed that at least one minimizer  $t_*$  exists) that  $N(\nu)$  has lower and upper bounds of the form  $O(\sqrt{L/\nu}\|t_0 - t_*\|_2)$  for sufficiently large dimension  $d$ . Accordingly, Nesterov [2004] gives under the additional assumption that the target function is infinitely often continuously differentiable and that  $1 \leq k \leq (d-1)/2$  the lower bound

$$F(t_k) - F(t_*) \geq \frac{3L\|t_0 - t_*\|_2^2}{32(k+1)^2}.$$

As explained by Nemirovski [1994], gradient descent cannot achieve this. For gradient descent, the optimal performance is  $O(1/k)$ . The presented lower bound of the type  $O(1/k^2)$  is achieved by an algorithm called *Nesterov's method*. Interestingly, Nesterov's method achieves this convergence rate independently of the dimension [Nemirovski, 1994]. Since the lower bound holds for sufficiently large  $d$ , Nesterov's algorithm is optimal in the high-dimensional setting.

As for the class of unconstrained,  $L$ -smooth,  $\mu$ -strongly convex problems, Nemirovski [1994] presents the bounds

$$c_1 \cdot \min \left\{ d, \sqrt{\kappa} \log \left( \frac{1}{2\nu} \right) \right\} \leq N(\nu) \leq c_2 \cdot \sqrt{\kappa} \log \left( \frac{2}{\nu} \right),$$

where  $c_1, c_2 > 0$  are constant. As they point out, if  $d \geq \sqrt{\kappa}$ , the lower bound can be modified to

$$c_3 \cdot \frac{\sqrt{\kappa}}{\log(\kappa)} \log \left( \frac{1}{2\nu} \right),$$

where  $c_3 > 0$  is again a constant. This refinement is helpful for all  $\nu \in (0, 1)$ , other than the first lower bound, which is only helpful for sufficiently large  $\nu$  but possibly (depending on  $\kappa$ ) tighter. These results yield that the best possible convergence rate of the accuracy we can hope for is linear in  $k$ .

Interestingly, these results are less helpful in lower dimensions. Nemirovski [1994] compares the bounds achieved in lower dimensions with those for non-smooth problems and come to the conclusion that a high dimension is necessary for the smoothness to improve the problem

complexity. As a result, for  $L$ -smooth,  $\mu$ -strongly convex functions on an infinitely dimensional domain, Nesterov [2004] presents the lower bound

$$F(t_k) - F(t_*) \geq \frac{\mu}{2} \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{2k} \|t_0 - t_*\|_2^2$$

on the achievable accuracy, which then holds for all  $k \in \mathbb{N}$ .

### Complexity results for noisy first order oracles

Here, the results are extended to methods which only have access to noisy first order oracles. As opposed to the previous passage, we can therefore at every point  $t \in \Theta$  only sample an estimator for  $F(t)$  and  $\nabla F(t)$ , which is typically assumed to be unbiased. Here, the trajectory of the solution path  $(\tilde{\theta}_k)_{k \geq 0}$  is consequently random, with the oracle noise causing the randomness.

Nemirovski [1994] make the central assumption that the gradient noise has finite second moment. Under that assumption, they argue that given noisy oracles of a certain variance, a convex, compact domain and an optimization procedure consisting of  $k$  steps based on a stochastic oracle, there will be a convex function  $F$  and an oracle, such that the expected accuracy  $\mathbb{E}[F(\tilde{\theta}_k) - F(t_*)]$  is lower bounded by a quantity of the form  $O(1/\sqrt{k})$ .

As argued by Nemirovski [1994], this lower bound no longer holds, when we restrict  $F$  to the class of strongly convex functions, which are upper and lower bounded by two parabolas, coinciding at the optimum value  $F(t_*)$ . This is exactly what the expression

$$\forall t \in \Theta : \quad F(t_*) + \frac{\mu}{2} \|t - t_*\|_2^2 \leq F(t) \leq F(t_*) + \frac{L}{2} \|t - t_*\|_2^2$$

describes, which is an implication of  $\mu$ -strong convexity together with  $L$ -smoothness, as explained in Section 2.4. For example, SGD with properly tuned step sizes can achieve a rate of  $O(1/k)$  in this case.

Agarwal et al. [2009] extend the work by Nemirovsky and Yudin [1983], Nemirovski [1994] and Nesterov [2004]. Specifically, under the assumption of access to unbiased stochastic oracles of first order with finite, fixed variance, the authors derive the minimax oracle complexity. Under a sufficiently restrictive class of strongly convex functions, they show that given the minimax optimal estimator for a problem, i.e. the method which decreases the expected excess risk the most for the worst possible function in the class, will lead to a convergence rate no faster than  $O(1/k)$  if the worst allowed unbiased oracle with bounded variance is sampled, i.e. we receive the most problematic allowed input. As pointed out by Bach [2014], results from this line of work imply that for the entire class of  $\mu$ -strongly convex functions, a constant  $\mu$  in the denominator is unavoidable, i.e. the optimal convergence in  $k$  is  $O(1/(k\mu))$ .

## 2. Background

### Summary of the presented complexity results

Firstly, it should be reiterated that these results only relate to the number of oracle calls, not to the computational complexity of the same.

For smooth optimization with exact first order oracles, the optimal rate of convergence is of the type  $O(1/k^2)$  for convex functions if the dimension is sufficiently high and of the type  $O(\rho^k)$ , with  $\rho \in [0, 1)$ , for strongly-convex functions. Note that the value of  $\rho$  depends on the condition number  $\kappa = L/\mu$  of the problem.

For numeric optimization with unbiased stochastic first order oracles of fixed variance, the optimal rate of convergence is  $O(1/\sqrt{k})$  for convex functions. For sufficiently regular, strongly convex functions, the optimal rate improves to  $O(1/k)$ .

### 2.9.2. Optimality in empirical risk minimization and M-estimation

As discussed in the introduction, estimating the parameter value  $\theta_* \in \Theta$ , which minimizes the risk, by calculating the empirical risk minimizer  $\hat{\theta}_n$ , i.e. the minimum of  $\hat{f}_n(\theta)$ , is part of the well-studied M-estimation framework in statistics.

In many applications, such as maximum likelihood estimation,  $\hat{\theta}_n$  is under some regularity conditions efficient, but might be computationally expensive to calculate. Nonetheless, the estimation performance of  $\hat{\theta}$  is a natural benchmark. The closer a computationally fast, iterative algorithm comes to its performance, the more inclined one should be to use this algorithm instead of  $\hat{\theta}_n$ .

As for the excess risk of the ERM, note that under suitable conditions on the loss and the distribution of the data, it holds that

$$\lim_{n \rightarrow \infty} n \cdot \mathbb{E} \left[ f(\hat{\theta}_n) - f(\theta_*) \right] = \tau^2,$$

for

$$\tau^2 = \frac{1}{2} \mathbb{E} \left[ (\nabla_{\theta} l(x_1, \theta_*))^T [\nabla^2 f(\theta_*)]^{-1} \nabla_{\theta} l(x_1, \theta_*) \right],$$

as argued by Frostig et al. [2015].

Thus, the aim of stochastic approximation methods will typically be to reduce the excess risk with a dependence of  $O(1/n)$  on the number of data points  $n$  used, while having a multiplicative constant not much larger than  $\tau^2$ . In online SGD, exactly one data point is used at each iteration, meaning that here  $n$  also refers to the number of iterations.

As noted by several authors, the rate of  $O(\tau^2/n)$  is in accordance with the Cramér–Rao lower bound (see for example Frostig et al. [2015]).

In this thesis, we are especially interested in the case of least squares regression, which will be discussed in the following.

### 2.9.3. The ERM for least squares regression

For the remainder of this section, Example 1 of random design least squares regression will be continued.

In this context, the ERM, i.e. the minimizer of

$$\hat{f}_n(\theta) = \frac{1}{2n} \sum_{i=1}^n (y_i - \theta^T x_i)^2$$

over  $\Theta = \mathbb{R}^d$  is, of course, simply the ordinary least squares estimator (OLS-estimator), given through

$$\hat{\theta}_n = \left( \frac{1}{n} \sum_{i=1}^n x_i x_i^T \right)^{-1} \frac{1}{n} \sum_{i=1}^n y_i x_i$$

for  $n \geq d$ , if the matrix in parenthesis is invertible.<sup>2</sup>

For the considered case of least squares regression with i.i.d. residuals  $\varepsilon_i = y_i - \theta_*^T x_i$ ,  $i \geq 1$ , we have  $\nabla_{\theta} l(y_i, \theta_*^T x_i) = -x_i \varepsilon_i$  and  $\nabla^2 f(\theta_*) = H$ . If we further assume that a  $\sigma > 0$  exists satisfying  $\mathbb{E} [\varepsilon_1^2 x_1 x_1^T] \preceq \sigma^2 H$ , we have the following asymptotic variance constant:

$$\begin{aligned} \tau^2 &= \frac{1}{2} \mathbb{E} [\varepsilon_1^2 x_1^T H^{-1} x_1] \\ &= \frac{1}{2} \mathbb{E} \left[ \text{tr} \left( H^{-1} \varepsilon_1^2 x_1 x_1^T \right) \right] \\ &= \frac{1}{2} \text{tr} \left( H^{-1} \mathbb{E} [\varepsilon_1^2 x_1 x_1^T] \right) \\ &\leq \frac{1}{2} \text{tr} (\sigma^2 I_d) \\ &= \frac{1}{2} \sigma^2 d. \end{aligned} \tag{2.7}$$

Note that if the variables  $\varepsilon_i$  and  $x_i$  are independent and  $\mathbb{E} [\varepsilon_1^2] = \sigma^2$ , (2.7) is an equality. In the fixed design model (i.e. deterministic  $x_i$ ) under standard assumptions and i.i.d. errors, it is easy to see that the corresponding result holds non-asymptotically.

### Computational cost of OLS

Calculation of the ERM (or OLS-estimator) comes at the cost of computational complexity. More specifically, for the common implementation of OLS through QR decomposition, the complexity is given by  $O(nd^2)$  for  $n \geq d$ , when using Householder QR (see Golub and Van Loan [2013, Subsection 5.3.3]).

As a result, SGD which only has a complexity of  $O(d)$  per iteration and consequently  $O(nd)$  for

---

<sup>2</sup>If the matrix in parenthesis is singular (which is automatically the case if  $n < d$ ), the ERM is not unique.

## 2. Background

$n$  steps, while still keeping the memory requirements low at  $O(d)$ , can be used to speed up the estimation process, especially when  $d$  is large.

Another advantage of SGD is that it can be used for online learning with the same runtime and memory complexity, due to its iterative nature with respect to the data points. More specifically, the complexity of performing SGD for  $n$  data points and then updating it for additional  $N - n$  (for some  $N > n$ ) data points, is the same as performing SGD for  $N$  data points (i.e.  $O(Nd)$ ). For OLS, the estimator can also be updated to account for the  $N - n$  new data points in a manner that is more efficient than calculating the estimator based on the  $N$  data points from scratch (see Golub and Van Loan [2013, Section 6.5]). Specifically, updating a QR-decomposition for another observation costs  $O(d^2)$  (see Golub and Van Loan [2013, Subsections 5.2.6, 6.5.3]). Subsequent calculation of the OLS-estimator given the QR-decomposition costs  $O((n+1)d)$  (see Golub and Van Loan [2013, Subsection 5.3.3]). As a result, given  $n \geq d$ , we have  $\max\{d^2, (n+1)d\} = (n+1)d$  and therefore updating the estimator for one new observation is of runtime complexity  $O(nd)$ , which is worse than SGD, for which we have  $O(d)$  for one new observation.

### Summary of the complexity of random design least squares regression

The given asymptotic behavior of the type  $\sigma^2 d/n$  of the excess risk is known to be unimprovable for least squares regression [Bach and Moulines, 2013]. However, the cost of calculation does not grow linearly in the number of variables, which might be an issue in large-scale learning. Thus, if  $n$  or  $d$  are large, in practice the computational cost of  $O(nd)$  of SGD might be preferable over the cost of  $O(nd^2)$  for OLS-estimation, even though an optimization error is added.

Even more application oriented is the argument for this by Nemirovski [1994]. As he puts it, non-iterative “exact” solutions are helpful, when you can pay the entire computational cost. However, if you cannot pay the computational cost (e.g. time runs out), you do not have any solution to the optimization problem at all. This is an inherent advantage of iterative methods in practice; you can stop them at any time and will get a solution. The later you stop the iteration, the better your approximation will roughly be. This also touches upon the basic principles of large-scale learning outlined by Bottou and Bousquet [2008], who view the number of used data points  $n$  as a variable, which has to be chosen such that the available time does not run out. For OLS-estimation, one might have to choose a substantially smaller subset of the data in order to guarantee that a solution is reached, as opposed to SGD.

## 2.10. Concluding remark

It is worth briefly looking back at the convergence results of GD for smooth functions, which were given in Subsection 2.6.1. As Bubeck [2015] points out, neither in the regular convex case, nor in the strongly convex can GD achieve the optimal number of oracle calls (i.e. steps) discussed in Subsection 2.9.1. Given the bounds of Subsection 2.6.1, we need  $O(1/\nu)$  oracle calls in order to achieve an accuracy of  $\nu$  in the case of general convexity, while we need  $O(\kappa \log(1/\nu))$

in the case of strong convexity. As discussed, the information theoretic minimum number of first order oracle calls needed to reduce the accuracy to below  $\nu$  are actually  $\Omega(1/\sqrt{\nu})$  for the general convex case (if the dimension is high enough) and  $\Omega(\sqrt{\kappa} \log(1/\nu))$  in the strongly convex case. The discrepancy for the strongly convex case is particularly telling, as it shows that GD depends more on the condition number of the problem than what would be optimal. This means in particular that for problems with large condition number  $\kappa = L/\mu$ , resulting from weak  $L$ -smoothness (large  $L$ ) or weak  $\mu$ -strong convexity (small  $\mu$ ), GD will need suboptimally many steps.

As Bach and Moulines [2013] point out, a similar effect can be observed in the context of SGD. For general convex functions, SGD can achieve a convergence rate of  $O(1/\sqrt{k})$ . As will be discussed in the literature review in Chapter 4, iterate averaging (see Subsection 2.7.3) has practical benefits. However, prior to Bach and Moulines [2013], averaged SGD for strongly convex functions was only shown to achieve a convergence rate of  $O(1/(k\mu))$ . This  $\mu$  in the denominator means again that if the strong convexity is very weak (small  $\mu$ ), averaged SGD will result in slow convergence, which might, in non-asymptotic settings, be in fact worse than a rate of  $O(1/\sqrt{k})$  independent of  $\mu$ , which we get for generally convex functions. This problem lies at the heart of the next chapter.



### 3. Non-asymptotic analysis of least squares regression

One of the applications of stochastic gradient descent is least squares regression. Besides logistic regression, this is a problem considered by Bach and Moulines [2013].

In the first section of this chapter, the problem of least squares regression as addressed in Example 1 will be examined in the context of stochastic gradient descent. After that, the assumptions and result of Theorem 1 in Bach and Moulines [2013] will be given and subsequently discussed. Lastly, a short proof outline is given. An in extensive detail described version of the proof by Bach and Moulines [2013] is provided in Appendix B.

#### 3.1. Revisiting the example of least squares regression

**Problem setup.** In Example 1, the problem of least squares regression was presented. Recall that in this case for some  $d \in \mathbb{N}$ , we have  $\Theta = \mathbb{R}^d$ ,  $x \in \mathbb{R}^d$ ,  $y \in \mathbb{R}$  and  $\mathcal{G} = \{g_\theta : \mathbb{R}^d \rightarrow \mathbb{R} \mid g_\theta(\cdot) = \langle \theta, \cdot \rangle, \theta \in \mathbb{R}^d\}$ . The quadratic loss function is given by  $l(y, y') = \frac{1}{2}(y - y')^2$  for  $y, y' \in \mathbb{R}$ . Let the risk function be denoted by  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ . Here, it is given by the quadratic function

$$f(\theta) = \mathbb{E} [l(y, g_\theta(x))] = \mathbb{E} \left[ \frac{1}{2}(y - \theta^T x)^2 \right] = \frac{1}{2} \mathbb{E} [y^2] - \theta^T \mathbb{E} [xy] + \frac{1}{2} \theta^T \mathbb{E} [xx^T] \theta.$$

It immediately follows that in order to evaluate the risk,  $x$  and  $y$  have to exhibit finite second moments. Specifically, it has to hold that  $\mathbb{E} [y^2] < \infty$  and  $\mathbb{E} [\|x\|_2^2] < \infty$ . Define  $H = \mathbb{E} [xx^T]$ , which is consequently a real, positive semi-definite matrix. A slightly more compact form of  $f$  is thus

$$f(\theta) = \frac{1}{2} \theta^T H \theta - \theta^T \mathbb{E} [xy] + \frac{1}{2} \mathbb{E} [y^2]. \quad (3.1)$$

**Global minimum.** Since  $f(\theta)$  is a continuous, quadratic function on  $\mathbb{R}^d$  and  $H$  is positive semi-definite, a (not necessarily unique) global minimizer of  $f$  exists. Such a global minimizer  $\theta_*$  has to satisfy the first and second order conditions, as  $f$  is twice continuously differentiable.

### 3. Non-asymptotic analysis of least squares regression

The first order condition gives that  $\theta_*$  has to satisfy

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta^T} f(\theta)|_{\theta_*} \\ \iff 0 &= H\theta_* - \mathbb{E}[x_1 y_1] \\ \iff H\theta_* &= \mathbb{E}[x_1 y_1]. \end{aligned}$$

If in addition  $H$  is invertible (and thus positive definite), there is exactly one point  $\theta_*$  satisfying the first order condition, which is given by

$$\theta_* = H^{-1} \mathbb{E}[x_1 y_1].$$

As for the second order condition, note that by our assumption that  $H$  is invertible,  $\frac{\partial^2}{\partial \theta \partial \theta^T} f(\theta) = H$  is positive definite, implying that  $\theta_*$  is indeed a global minimum.

**Linear model.** One typical motivation for the least squares setup is the estimation of a parameter  $\theta_{**}$ , which governs the joint distribution of the explanatory variables  $x_i$  and the response variable  $y_i$  through the model formula  $y_i = \theta_{**}^T x_i + u_i$ , where  $u_i \in \mathbb{R}$  is a sequence of random zero-mean errors of finite variance, such that the sequence  $((x_i, u_i))_{i \in \mathbb{N}}$  is i.i.d. Note that the model might be misspecified and that  $x_i$  and  $u_i$  might not be independent. In particular, we might have that  $\mathbb{E}[x_1 u_1] \neq 0$  and consequently  $\mathbb{E}[u_1 | x_1] \neq 0$ .

Linear models with random explanatory variables, like we have here, are oftentimes referred to as *random design* linear models. This distinguishes the setting from so-called *fixed design* linear models in which the values of the explanatory variables  $x_i$  are fixed and the randomness in the response variable  $y_i$  consequently only stems from the error  $u_i$ . The condition that  $H$  is invertible, which was imposed in the previous paragraph, is analogous to the condition of full-rank for design matrices in fixed design models.

The connection to the global minimizer  $\theta_*$  discussed in the previous paragraph is as follows. We have  $\theta_* = \theta_{**}$  if and only if  $\mathbb{E}[x_1 u_1] = 0$ , since firstly if  $\mathbb{E}[x_1 u_1] = 0$

$$\theta_* = H^{-1} \mathbb{E}[x_1 y_1] = H^{-1} \mathbb{E}[x_1 x_1^T \theta_{**} + x_1 u_1] = H^{-1} H \theta_{**} + H^{-1} \mathbb{E}[x_1 u_1] = \theta_{**}$$

and conversely if  $\theta_* = \theta_{**}$

$$\mathbb{E}[x_1 u_1] = \mathbb{E}[x_1 y_1 - x_1 x_1^T \theta_{**}] = \mathbb{E}[x_1 y_1] - H \theta_* = \mathbb{E}[x_1 y_1] - H H^{-1} \mathbb{E}[x_1 y_1] = 0.$$

In summary, under model misspecification there is a discrepancy between the parameter of the model  $\theta_{**}$  and the risk minimizer  $\theta_*$ . The discrepancy vanishes if  $\theta_* = \theta_{**}$ , which was shown to be equivalent to weak exogeneity, i.e.  $\mathbb{E}[x_1 u_1] = 0$ . As shown in the former of the two preceding displays, one direction of the implication follows from the equation  $\theta_* = \theta_{**} + H^{-1} \mathbb{E}[x_1 u_1]$ , meaning that the discrepancy between the true model parameter and the risk minimizer is

### 3.2. The non-asymptotic bound by Bach and Moulines [2013]

continuous in  $\mathbb{E}[x_1 u_1]$ . The behavior is analogous to that of the ordinary least squares estimator in misspecified models.

In their paper, Bach and Moulines [2013] introduce the i.i.d. residuals  $\varepsilon_i = y_i - \theta_*^T x_i$ ,  $i \in \mathbb{N}$ . Since the sequence  $((x_i, u_i))_{i \in \mathbb{N}}$  is assumed i.i.d. and  $\varepsilon_i = (\theta_{**} - \theta_*)^T x_i + u_i$ ,  $i \in \mathbb{N}$ , the sequence  $((x_i, \varepsilon_i))_{i \in \mathbb{N}}$  is also i.i.d. By construction, we have  $\mathbb{E}[x_1 \varepsilon_1] = 0$  which can be verified by following the steps of the last preceding display. Still,  $\mathbb{E}[\varepsilon_1 | x_1] = 0$  does not need to hold, unless  $\mathbb{E}[u_1 | x_1] = 0$ .

For the remainder of this thesis, the residuals  $\varepsilon_i$  will be of much higher importance than the errors  $u_i$ .

**Stochastic Gradient Descent.** In general, a popular choice for the gradient estimator  $\phi(y, x, \theta)$  for SGD is the loss gradient  $\nabla_{\theta} l(y, g_{\theta}(x))$ . In the case of least squares regression, the recursion of SGD then follows easily. The gradient of the loss function with respect to  $\theta$  is given by

$$\nabla_{\theta} l(y, \theta^T x) = (y - \theta^T x)(-x) = x x^T \theta - x y.$$

Given that the gradient of the risk is

$$\nabla f(\theta) = H\theta - \mathbb{E}[x_1 y_1],$$

where  $H$  was defined through  $H = \mathbb{E}[x_1 x_1^T]$ , it is immediately obvious that for any fixed  $\theta \in \Theta$  and for any of the i.i.d. observations  $(x_i, y_i)$ ,

$$\nabla_{\theta} l(y_i, \theta^T x_i) = x_i x_i^T \theta - x_i y_i$$

is an unbiased estimator for  $\nabla f(\theta)$ .

Stochastic gradient descent with starting point  $\theta_0 \in \mathbb{R}^d$  and step size sequence  $\gamma_n > 0$ ,  $n \geq \mathbb{N}$ , is therefore given through the iteration

$$\begin{aligned} \theta_n &= \theta_{n-1} - \gamma_n (x_n x_n^T \theta_{n-1} - x_n y_n) \\ &= \theta_{n-1} - \gamma_n x_n x_n^T \theta_{n-1} + \gamma_n x_n y_n \\ &= (I - \gamma_n x_n x_n^T) \theta_{n-1} + \gamma_n x_n y_n \end{aligned}$$

over  $n \in \mathbb{N}$ .

## 3.2. The non-asymptotic bound by Bach and Moulines [2013]

Theorem 1 by Bach and Moulines [2013] gives an upper bound on the expected excess risk that averaged stochastic gradient descent achieves in  $n$  steps for an arbitrary starting point  $\theta_0 \in \mathbb{R}^d$  for the problem of least squares regression.

### 3. Non-asymptotic analysis of least squares regression

In Bach and Moulines [2013], the authors present their results in a slightly more general setting by considering a  $d$ -dimensional response variable instead of a scalar response variable. While the more general case is of high relevance in the derivation of subsequent results in their paper, the simplification of the result for scalar response variables is interesting in its own right. Since their general version bears no importance for the remainder of this thesis, the following description is restricted to scalar response variables. For the sake of completeness, the general case described by Bach and Moulines [2013] is considered and proved in Appendix B.

Following Bach and Moulines [2013], we make a series of assumptions concerning the distribution of the observations and the stochastic gradient descent procedure. Most of these assumptions were already informally introduced or motivated in the previous sections and are restated here for clarity. Let  $d \in \mathbb{N}$ .

- (A1)  $(y_i, x_i) \in \mathbb{R} \times \mathbb{R}^d$ ,  $i \in \mathbb{N}$ , are independent and identically distributed random variables.
- (A2) The variables have finite second moments, i.e.  $\mathbb{E}[\|x_1\|_2^2] < \infty$  and  $\mathbb{E}[y_1^2] < \infty$ . We write  $H = \mathbb{E}[x_1 x_1^T]$  for the second moment matrix, which is assumed to be invertible.
- (A3) The function  $f(\theta) = \frac{1}{2} \mathbb{E}[(y_1 - \theta^T x_1)^2]$  is minimal at  $\theta_* \in \mathbb{R}^d$ . Let  $\varepsilon_i = y_i - \theta_*^T x_i$  for  $i \in \mathbb{N}$  denote the residuals.
- (A4) There exist  $R > 0$  and  $\sigma > 0$  such that  $\mathbb{E}[\varepsilon_1^2 x_1 x_1^T] \preceq \sigma^2 H$  and  $\mathbb{E}[\|x_1\|_2^2 x_1 x_1^T] \preceq R^2 H = R^2 \mathbb{E}[x_1 x_1^T]$ .
- (A5) The step size sequence is chosen constant with  $\gamma_n = \gamma > 0$  for all  $n \in \mathbb{N}$  such that  $\gamma R^2 < 1$ .
- (A6) The stochastic gradient descent recursion is defined as

$$\theta_n = (I - \gamma x_n x_n^T) \theta_{n-1} + \gamma x_n y_n \quad (3.2)$$

for  $n \in \mathbb{N}$  with  $\theta_0 \in \mathbb{R}^d$  arbitrary and non-random. The averaged SGD estimator is

$$\bar{\theta}_n = \frac{1}{n+1} \sum_{k=0}^n \theta_k$$

for  $n \in \mathbb{N} \cup \{0\}$ .

Assumption (A3) is not restrictive but rather provides us with the necessary definitions, as long as the other conditions hold. Assumptions (A1), (A2) and (A4) impose regularity conditions on the distribution of the observations. Lastly, assumptions (A5) and (A6) describe the specifics of the chosen SGD variant.

Under the assumptions (A1)-(A6) Bach and Moulines [2013] derive an upper bound for the expected excess risk of  $\bar{\theta}_n$  for arbitrary finite  $n \in \mathbb{N}$ . This is the subject of the following theorem.

**Theorem 1.** (Bach and Moulines [2013]) Under the assumptions **(A1)**–**(A6)**

$$\mathbb{E} [f(\bar{\theta}_{n-1}) - f(\theta_*)] \leq \frac{1}{2n} \left[ \frac{\sigma\sqrt{d}}{1 - \sqrt{\gamma R^2}} + R\|\theta_0 - \theta_*\|_2 \frac{1}{\sqrt{\gamma R^2}} \right]^2$$

holds for all  $n \in \mathbb{N}$ . For  $\gamma = 1/(4R^2)$  this simplifies to

$$\mathbb{E} [f(\bar{\theta}_{n-1}) - f(\theta_*)] \leq \frac{2}{n} \left[ \sigma\sqrt{d} + R\|\theta_0 - \theta_*\|_2 \right]^2.$$

Some aspects of this result are discussed in the next section.

### 3.3. Discussion

The discussion is split up into remarks about the underlying assumptions and comments regarding the bound derived by Bach and Moulines [2013]. Some more context is then provided through the literature review in Chapter 4.

#### 3.3.1. Discussion of the assumptions

**Convexity and smoothness.** Firstly, it should be noted that the risk function satisfies the definitions of strong convexity and smoothness as defined in Section 2.4, since assumption **(A2)** implies that  $H$  is positive definite. The strong convexity parameter is given through the smallest eigenvalue of  $H$ , i.e.  $\mu = \lambda_{\min}(H)$ , and the smoothness parameter corresponds to the largest eigenvalue of  $H$ , i.e.  $L = \lambda_{\max}(H)$ .

While assumption **(A2)** implies that the eigenvalues of  $H$  are strictly positive, no lower bound is imposed on them. As a consequence,  $\mu$  can be an arbitrarily small positive quantity, implying an arbitrarily weak strong convexity. Therefore, the condition number  $\kappa = L/\mu$  is allowed to be arbitrarily large.

**Fourth-moment constant  $R^2$ .** If such a value  $R$  as defined in assumption **(A4)** exists, it is not unique, since the defining inequality would then also hold for all  $\tilde{R} > R$ . For the sake of this discussion, define  $R_0$  as the infimum of all values satisfying the inequality in assumption **(A4)**. Lower bounds on  $R_0$  translate to lower bounds on the values for  $R$ .

Since  $R_0$  depends on the distribution of the explanatory variables  $x_i$ , it (and all other values  $R$ ) will typically be unknown in practice. However, according to assumption **(A5)**,  $\gamma > 0$  has to be chosen such that it satisfies  $\gamma < 1/R^2 < 1/R_0^2$ , with  $R^2 > R_0^2$ . To guarantee this, upper bounds (no matter how loose) on  $R_0^2$  are needed.

Furthermore, the relationship of such  $R^2$  to other quantities, such as  $H$ ,  $\mu$ ,  $L$  and  $d$  is not immediately obvious.

First, any  $R^2$  satisfies  $\text{tr}(H) \leq R^2$ , since we have from the definition of  $R^2$  in assumption **(A4)**

### 3. Non-asymptotic analysis of least squares regression

that

$$\begin{aligned} & \mathbb{E} \left[ \|x_1\|_2^2 x_1 x_1^T \right] \preceq R^2 H \\ \implies & \operatorname{tr} \left( \mathbb{E} \left[ \|x_1\|_2^2 x_1 x_1^T \right] \right) \leq R^2 \operatorname{tr}(H) \\ \iff & \mathbb{E} \left[ \|x_1\|_2^4 \right] \leq R^2 \operatorname{tr}(H). \end{aligned}$$

Using either the Lyapunov or the Cauchy–Schwarz inequality gives us  $\left( \mathbb{E} \left[ \|x_1\|_2^2 \right] \right)^2 \leq \mathbb{E} \left[ \|x_1\|_2^4 \right]$ .

Furthermore,  $\mathbb{E} \left[ \|x_1\|_2^2 \right] = \operatorname{tr} \left( \mathbb{E} \left[ x_1 x_1^T \right] \right) = \operatorname{tr}(H)$ . The last inequality therefore implies

$$\left( \operatorname{tr}(H) \right)^2 \leq R^2 \operatorname{tr}(H)$$

and thus

$$\operatorname{tr}(H) \leq R^2.$$

This gives a necessary condition on  $R^2$ . The inequality directly implies that

$$L + \mu \leq R^2$$

and of course also

$$\mu \leq \frac{1}{d} \operatorname{tr}(H) \leq L \leq \operatorname{tr}(H) \leq R^2.$$

The necessary condition further implies that

$$\mu d \leq R^2.$$

An immediately obvious sufficient condition for  $R^2$  is that the data is bounded, such that  $\|x_i\|_2^2 \leq R^2$  holds almost surely for all  $i \in \mathbb{N}$ . The necessary condition  $\operatorname{tr}(H) \leq R^2$  is equivalent to  $\mathbb{E} \left[ \|x_1\|_2^2 \right] \leq R^2$ , yielding in summary that

$$\|x_1\|_2^2 \leq R^2 \text{ almost surely} \implies R^2 \text{ satisfies } \mathbf{(A4)} \implies \mathbb{E} \left[ \|x_1\|_2^2 \right] \leq R^2$$

which gives some characterization of  $R^2$ .

For bounded data, this immediately yields a suitable value of  $R^2$  for which assumption **(A4)** is satisfied.

To give a specific example, for  $x_i \sim \mathcal{N}_d(0, I_d)$ , it is easy to show that  $R_0^2 = d + 2$ , showing that the presented necessary conditions are of the correct order  $\Theta(d)$  for this case.

**Upper bound on the step size.** Recall from assumptions **(A5)** and **(A6)** that we perform SGD with  $\gamma < 1/R^2$  as given by the iteration

$$\theta_n = (I - \gamma x_n x_n^T) \theta_{n-1} + \gamma x_n y_n$$

for  $n \in \mathbb{N}$  with some initial point  $\theta_0 \in \mathbb{R}^d$ . As pointed out by Bottou et al. [2018], for a suitable step size, we deal with a contraction. This idea lies at the heart of this and other proofs of convergence and will therefore be explained in a little more detail.

Recall that we have shown that  $R^2$  needs to satisfy  $\text{tr}(H) \leq R^2$ . As will be shown in Appendix B.2, this implies that  $0 \preceq \gamma H \preceq I$ , with the first inequality being immediately obvious. As a result  $0 \preceq I - \gamma H \preceq I$ , where in turn the second inequality is immediately obvious. This of course also implies that  $-I \preceq I - \gamma H \preceq I$ , which would also hold if we only required  $\gamma < 2/R^2$  instead of  $\gamma < 1/R^2$ . Either way, if assumption **(A5)** was ignored and  $\gamma$  was chosen arbitrarily large, the first inequality would be violated, and we might even have  $I - \gamma H \preceq -cI$  for some  $c > 1$ . This is important for the following reason.

Consider the deterministic sequence  $(t_n)_{n \geq 0}$  in  $\mathbb{R}^d$  defined by some starting point  $t_0 \in \mathbb{R}^d$  and the recursion

$$t_n = (I - \gamma H) t_{n-1}$$

for  $n \in \mathbb{N}$ . For the sequence of the norms of the  $t_n$ , this means that through spectral decomposition of  $I - \gamma H$  one can show that

$$\begin{aligned} \|t_n\|_2 &= \|(I - \gamma H) t_{n-1}\|_2 = \sqrt{t_{n-1}^T (I - \gamma H)^2 t_{n-1}} \\ &\leq \|t_{n-1}\|_2 \sqrt{\max\{(1 - \gamma \lambda_{\min}(H))^2, (1 - \gamma \lambda_{\max}(H))^2\}} \\ &= \|t_{n-1}\|_2 \max\{|1 - \gamma \mu|, |1 - \gamma L|\}. \end{aligned} \tag{3.3}$$

Note that the previously derived inequalities  $-I \preceq I - \gamma H \preceq I$  are equivalent to  $-1 \leq 1 - \gamma \lambda_i(H) \leq 1$  for all  $i \in \{1, \dots, d\}$ , which is in turn equivalent to  $-1 \leq 1 - \gamma L \leq 1 - \gamma \mu \leq 1$ . As long as  $\gamma < 2/L$  (and thus  $\gamma < 2/\mu$ ), it holds that  $\gamma L \in (0, 2)$  and  $\gamma \mu \in (0, 2)$ . As a result, we have that  $-1 < 1 - \gamma L \leq 1 - \gamma \mu < 1$ , which yields  $\max\{|1 - \gamma \mu|, |1 - \gamma L|\} \in (0, 1)$ .

Assumption **(A5)** is even stricter, requiring that  $\gamma < 1/R^2 < 1/L \leq 1/\mu$  and thus implying  $\gamma L \in (0, 1)$  as well as  $\gamma \mu \in (0, 1)$ . As a result we even have  $0 < 1 - \gamma L \leq 1 - \gamma \mu < 1$ , which yields  $\max\{|1 - \gamma \mu|, |1 - \gamma L|\} = 1 - \gamma \mu \in (0, 1)$ . This shows that the choice of step size in assumption **(A5)** guarantees that the sequence  $(t_n)_{n \geq 0}$  converges towards 0 for  $n \rightarrow \infty$ . Even more, (3.3) even gives linear convergence  $O(\rho^n)$  of  $\|t_n\|_2$ , where  $\rho = \max\{|1 - \gamma \mu|, |1 - \gamma L|\}$ , as long as  $\gamma < 2/L$ .

To see why this bears relevance for the sequence  $(\theta_n)_{n \geq 0}$ , note the following reformulation of the

### 3. Non-asymptotic analysis of least squares regression

iteration formula (3.2):

$$\begin{aligned}
& \theta_n = (I - \gamma x_n x_n^T) \theta_{n-1} + \gamma x_n y_n \\
\iff & \theta_n - \theta_* = (I - \gamma x_n x_n^T) \theta_{n-1} - \theta_* + \gamma x_n x_n^T \theta_* - \gamma x_n x_n^T \theta_* + \gamma x_n y_n \\
\iff & \theta_n - \theta_* = (I - \gamma x_n x_n^T) (\theta_{n-1} - \theta_*) + \gamma x_n (y_n - x_n^T \theta_*) \\
\iff & \theta_n - \theta_* = (I - \gamma H + \underbrace{\gamma(x_n x_n^T - H)}_{=: E_n}) (\theta_{n-1} - \theta_*) + \underbrace{\gamma x_n \varepsilon_n}_{=: e_n}.
\end{aligned}$$

This shows that the iteration sequence for the error  $\theta_n - \theta_*$  is quite similar to the one of  $t_n$  examined above. The difference lies in the zero-mean errors  $E_n$  and  $e_n$ , where  $E_n$  has a multiplicative influence on the iterate, whereas  $e_n$  has an additive influence. Note that  $\mathbb{E}[E_n] = 0$  is obvious and  $\mathbb{E}[e_n] = 0$  was discussed in Section 3.1 in the introduction of  $\varepsilon_n$ . The second moments of these quantities are exactly the objects of interest in assumption **(A4)**, where upper bounds are defined through the definition of  $\sigma^2$  and  $R^2$ .

Note further that  $\theta_{n-1}$  and  $x_n$  are independent for all  $n \in \mathbb{N}$ , which by iteration means that

$$\mathbb{E}[\theta_n - \theta_*] = (I - \gamma H) \mathbb{E}[\theta_{n-1} - \theta_*] = (I - \gamma H)^n (\theta_0 - \theta_*).$$

As a result, the bias of the estimator  $\theta_n$  depends entirely on the choice of  $\theta_0$  and will be larger if  $\gamma$  is chosen small, as is reflected in the bound by Bach and Moulines [2013], which will be discussed later.

If  $E_n$  and  $e_n$  were uniformly equal to 0, convergence would follow through  $t_n = \theta_n - \theta_*$  and would be of linear rate. The difficulty lies in analyzing the influence of  $E_n$  and  $e_n$ . Some technicalities on how Bach and Moulines [2013] tackle this problem will be presented after the discussion in Section 3.4 with the full proof being available in Appendix B. The distinction between multiplicative and additive errors will reappear in the literature review in Chapter 4.

**Comparison to the step size of GD.** Compare the step size to the given results for GD (see Subsection 2.6.1). As Dieuleveut et al. [2017] point out,  $L$  is the spectral radius of  $H$  (i.e. the largest eigenvalue), while  $R^2$  is of the order  $\text{tr}(H)$  (i.e. the sum of all eigenvalues). We get the following inequalities for  $1/R^2$ ,  $1/L$  and  $2/(L + \mu)$ :

$$\gamma < \frac{1}{R^2} \leq \frac{1}{\text{tr}(H)} = \frac{1}{\sum_{i=1}^d \lambda_i(H)} < \frac{1}{L} \leq \frac{2}{L + \mu} < \frac{2}{L},$$

where  $\lambda_i(H)$ ,  $i \in \{1, \dots, d\}$ , are the eigenvalues of  $H$ .

This shows the relationship between the step sizes considered here and in Subsection 2.6.1. The fact that the step size of SGD is chosen smaller than GD is not that surprising, as  $\gamma$  appears as a multiplicative factor in the errors  $E_n$  and  $e_n$ , meaning that larger values of  $\gamma$  lead to more pronounced effects of the errors  $E_n$  and  $e_n$ .

Recall further that we have also derived that  $R^2$  necessarily satisfies  $R^2 > d\mu$ , and thus  $\gamma <$

$1/(d\mu)$  follows from assumption **(A5)**. This is interesting as it means that *ceteris paribus*, a higher dimension leads to smaller step sizes.

**Iterate averaging.** In assumption **(A6)**, iterate averaging (which was already introduced as an algorithmic primitive in Subsection 2.7.3) was redefined and the bound holds indeed for averaged SGD. The effects of iterate averaging are twofold. Firstly, as will be discussed in more detail in the literature review in Chapter 4, it allows for a more robust choice of the step size. Secondly, SGD with a constant step size without averaging will in general not converge (see for example Pflug [1986] or Dieuleveut et al. [2020]). Thus, some adaption of the algorithm like iterate averaging is necessary in order to ensure convergence if constant step sizes are employed. Interestingly, if averaged SGD with constant step size converges even though the risk function is not quadratic, the point of convergence will not be the minimizer  $\theta_*$ . This is discussed by Dieuleveut et al. [2020] and will be explained in the literature review in Chapter 4.

### 3.3.2. Discussion of the bound

After having discussed the importance and interpretation of some of the assumptions, we will now look at an interpretation of the non-asymptotic expected excess risk bound shown by Bach and Moulines [2013] and given in Theorem 1. For the convenience of the reader, the result is restated here. For all  $n \in \mathbb{N}$ , it holds under assumptions **(A1)**-**(A6)** that

$$\mathbb{E} [f(\bar{\theta}_{n-1}) - f(\theta_*)] \leq \underbrace{\frac{1}{2n}}_{=:(*)} \left[ \underbrace{\frac{\sigma\sqrt{d}}{1 - \sqrt{\gamma R^2}}}_{=:(**)} + R\|\theta_0 - \theta_*\|_2 \underbrace{\frac{1}{\sqrt{\gamma R^2}}}_{=:(***)} \right]^2.$$

In the following, the three parts of the bound  $(*)$ ,  $(**)$  and  $(***)$  will be discussed. The structure and interpretation of the two components  $(**)$  and  $(***)$  is of special interest and a particular strength of this result by Bach and Moulines [2013].

**Convergence rate.** With  $(*) = 1/(2n)$ , the result provides a sublinear convergence rate of order  $1/n$ , which is unimprovable as the leading order of convergence [Bach and Moulines, 2013].

**Variance.** The term  $(**)$  is known as the *variance term*, as it quantifies the random variation of the residuals  $\varepsilon_n$  respectively the influence of the additive errors  $e_n = \varepsilon_n x_n$ . As Bach and Moulines [2013] point out, no upper bound can be better than  $\sigma^2 d/n$  up to constants. Recall that this is the asymptotically optimal result in the context of M-estimation, as discussed in Section 2.9.

Note further the term  $\sqrt{\gamma R^2}$  in the denominator. Since  $\gamma R^2 \in (0, 1)$ , this quantity serves as a trade-off between the sizes of  $(**)$  and  $(***)$ . Large values of  $\gamma R^2$  (i.e. comparatively large step sizes) lead to a large influence of the additive error  $e_n$  or the residuals  $\varepsilon_n$ .

### 3. Non-asymptotic analysis of least squares regression

**Bias.** The term (\*\*\*) is known as the *bias term*, as it quantifies the importance of the choice of the starting point  $\theta_0$ . Unsurprisingly, the larger the distance from the optimum, the larger the bias term and consequently the upper bound will be. Conversely to the influence of the step size on the variance term, large values of  $\gamma R^2$  lead to a small influence of the choice of the starting point. Intuitively this trade-off is clear. While large step sizes can lead to propagation of errors and random noise, they help to forget the starting point more quickly.

Note that the bias term contains some redundancy, as the  $R$  in the numerator and the  $\sqrt{R^2}$  in the denominator cancel. However, since  $\gamma$  has to be chosen in accordance with  $R^2$  through assumption **(A5)** and since the discussed trade-off presents itself, it makes sense to consider  $\gamma R^2$  as a unit. In other words, one could rephrase assumption **(A5)** to require the choice of a number  $c \in (0, 1)$  controlling the trade-off, with the step size then being defined through  $\gamma = c/R^2$ . This point of view then shows that  $R$  does indeed appear as an important factor in the bias term, thus leading to a larger influence of  $R$  if one tries to balance the trade-off towards less influence of the variance term. Given the necessary condition  $R^2 \geq d\mu$  derived above, this shows that not only the variance term, but also the bias term increases at least as quickly as  $\sqrt{d}$ . We therefore recover a bound of the common structure  $O(d/n)$ .

While the variance term is optimal, several authors seek to improve the bias term in order to reduce the effect the choice of  $\theta_0$  (which might be largely uninformed or even random) has. This will be commented on in further detail in the literature review in Chapter 4.

**Influence of  $L$  and  $\mu$ .** Note that through the necessary condition  $L < R^2$ , less smooth functions, i.e. functions with large  $L$ , will require larger values of  $R^2$ . This automatically leads to an, at least, linear increase in the bias term with growing  $L$ .

Very small values of  $\mu$  do not seem to have any adverse effects, which Bach and Moulines [2013] give as the central merit of this bound. Note, however, that the result gives an upper bound on the expected excess risk  $\mathbb{E}[f(\bar{\theta}_{n-1}) - f(\theta_*)]$ , not the mean squared error  $\mathbb{E}[\|\bar{\theta}_{n-1} - \theta_*\|_2^2]$ . The relationship between these two quantities is

$$\frac{\mu}{2} \mathbb{E}[\|\bar{\theta}_{n-1} - \theta_*\|_2^2] \leq \mathbb{E}[f(\bar{\theta}_{n-1}) - f(\theta_*)]$$

(see the inequalities in (2.4)) such that the bound by Bach and Moulines [2013] for the mean squared error increases as  $1/\mu$  if  $\mu$  decreases. Such an increase is unavoidable and can also be observed in OLS-estimation, where higher correlation of the covariates leads to a higher mean squared error. If, however, the upper bound for the expected excess risk already contained  $\mu$  in the denominator, such a result would imply a growth of the upper bound of the mean squared error as  $1/\mu^2$ .

This shows that results on SGD for general strongly convex, smooth functions, in which case the expected excess risk will be no better than  $O(1/(n\mu))$  lead to a convergence of  $O(1/(n\mu^2))$  for the mean squared error, meaning that SGD is too sensitive with respect to the strong convexity constant. The bound by Bach and Moulines [2013] shows that this can be improved in the case

of least squares regression to match the dependence on  $\mu$  of the empirical risk minimizer.

### 3.4. Proof outline

The proof will be presented in extensive detail in Appendix B. The present section will give a largely non-technical explanation of the proof technique employed by Bach and Moulines [2013]. A graphical representation of their proof is given in Figure 3.1.

First, recall the risk function for least squares regression, which was derived in equation (3.1) and is restated here for convenience:

$$f(\theta) = \frac{1}{2}\theta^T H \theta - \theta^T \mathbb{E}[x_1 y_1] + \frac{1}{2}\mathbb{E}[y_1^2].$$

Using that  $\mathbb{E}[x_1 y_1] = H\theta_*$ , it follows directly from the preceding equality that the excess risk  $f(\cdot) - f(\theta_*)$  can at every point  $\theta \in \mathbb{R}^d$  be rewritten as an inner product, i.e.

$$f(\theta) - f(\theta_*) = \frac{1}{2}(\theta - \theta_*)^T H(\theta - \theta_*).$$

The expected excess risk for our estimator  $\bar{\theta}_{n-1}$  therefore satisfies

$$\mathbb{E}[f(\bar{\theta}_{n-1}) - f(\theta_*)] = \frac{1}{2}\mathbb{E}[(\bar{\theta}_{n-1} - \theta_*)^T H(\bar{\theta}_{n-1} - \theta_*)].$$

For convenience, define the norm on random variables  $\|\cdot\|_H$  through  $\|X\|_H := \sqrt{\mathbb{E}[X^T H X]}$  for any  $\mathbb{R}^d$ -valued random variable  $X$  with finite second moments. Define further the error  $\bar{\eta}_n := \bar{\theta}_n - \theta_*$  for all  $n \geq 0$ . We then get

$$\sqrt{2\mathbb{E}[f(\bar{\theta}_{n-1}) - f(\theta_*)]} = \|\bar{\eta}_n\|_H,$$

which justifies the derivation of an upper bound on  $\|\bar{\eta}_n\|_H$  in order to get an upper bound on the expected excess risk.

As summarized in Figure 3.1, the proof then continues as follows. Through a straight-forward derivation of a formula for  $\bar{\eta}_n$  based on  $\bar{\eta}_0$ ,  $(x_i)_{i=1}^n$  and  $(\varepsilon_i)_{i=1}^n$  for any  $n \in \mathbb{N}$ , one can split up  $\bar{\eta}_n$  into a term  $a_n$ , quantifying the influence of  $\bar{\eta}_0$  combined with  $(x_i)_{i=1}^n$ , and a second term  $b_n$ , quantifying the influence of  $(x_i)_{i=1}^n$  and  $(\varepsilon_i)_{i=1}^n$ . Through the triangle inequality, the norm of  $\bar{\eta}_n$  can be upper bounded by the norms on  $a_n$  and  $b_n$ .

The first term  $a_n$  contains the dependence on the bias and can be conveniently written as an iteration. Through upper bounding the change in its norm at each step of the iteration,  $\|a_n\|_H$  can be upper bounded for every  $n \in \mathbb{N}$  through the usage of  $a_0$  or, equivalently,  $\bar{\eta}_0$ . This already yields the bias term for which no more calculations are necessary.

### 3. Non-asymptotic analysis of least squares regression

The calculation of the variance term is slightly more complicated. Roughly speaking, an integer-valued approximation parameter  $r \geq 0$  is introduced and a sequence  $(c_{n,r})_{n,r \geq 0}$  is constructed. Through the triangle inequality, it is easy to see that we can derive separate upper bounds on the error of this approximation  $\|b_n - c_{n,r}\|_H$ , which will be denoted by  $A_{n,r}$ , and the size of this approximation  $\|c_{n,r}\|_H$ , which we will call  $B_{n,r}$ . For  $r \rightarrow \infty$ , the upper bound  $A_{n,r}$  for the error of the approximation vanishes. At the same time, the upper bound on the size of the approximation  $B_{n,r}$  converges to the variance term in the result of Theorem 1. This gives the variance term as an upper bound on  $\|b_n\|_H$ , as  $\|b_n\|_H$  does not depend on  $r$ . Algebraically, the derivations of  $A_{n,r}$  and  $B_{n,r}$  are largely based on the analysis of recursion equations. Combining the bias and the variance term finally gives the bound derived by Bach and Moulines [2013].

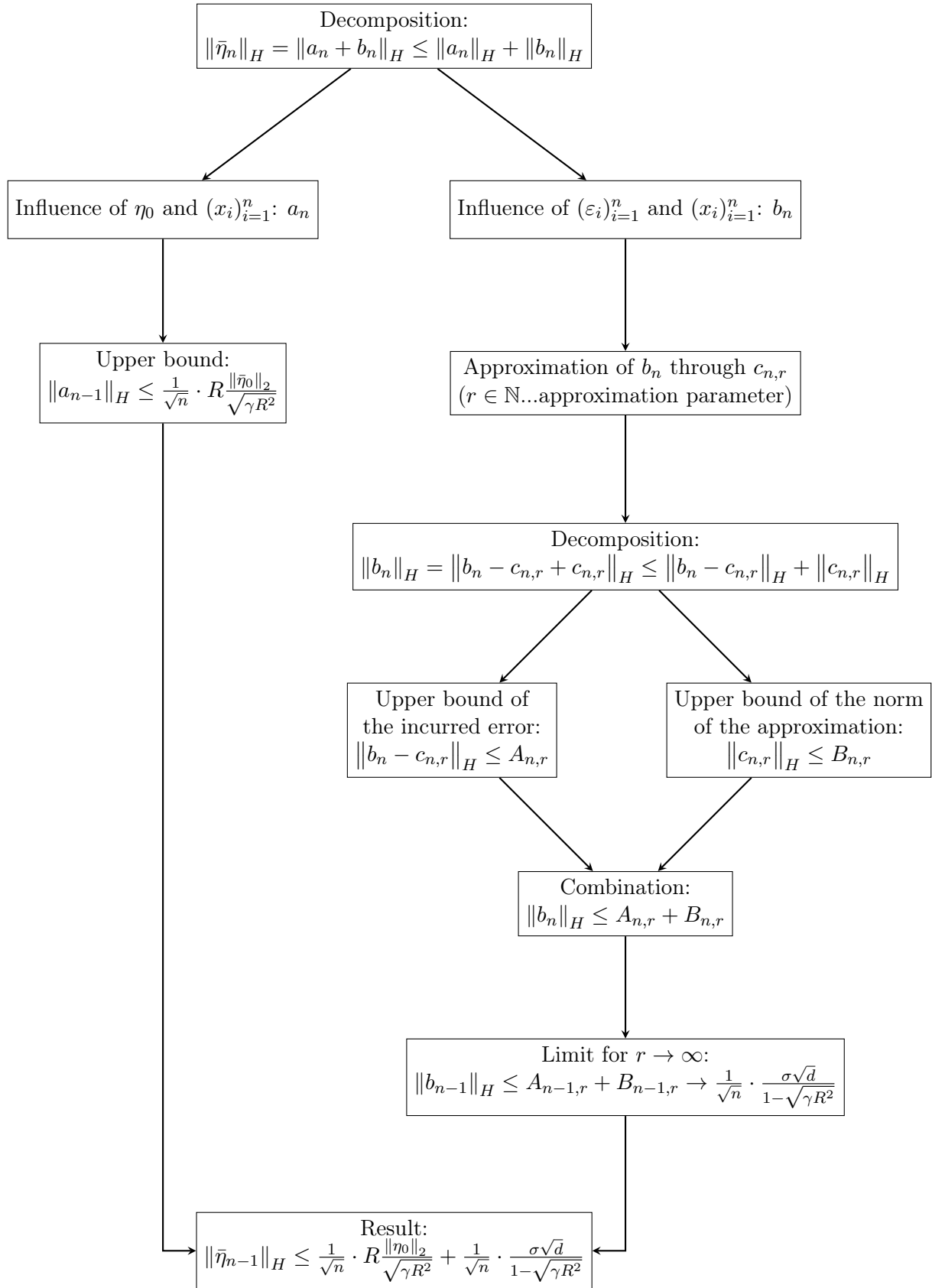


Figure 3.1.: The diagram visualizes the proof used by Bach and Moulines [2013] in order to prove Theorem 1. For a short description of the appearing terms, see Section 3.4. The full proof is presented in detail in Appendix B.



## 4. Related results and literature review

The main focus of this thesis is the work by Bach and Moulines [2013]. While the advances and improvements made after their publication are certainly of interest, it is also worth looking at prior results in order to understand the significance of the innovations by Bach and Moulines. This is why this chapter will be chronologically split into the period before and after their publication. After that, one more section will give a brief overview over the different choices for the step size in the literature and will discuss two related lines of work.

First, however, a very brief discussion of the common goal of the presented literature will be given.

### 4.1. Introduction and common goal

While the literature discussed over the course of this chapter considers different settings, different assumptions and different algorithms with varying performance, the common goal is to find a value  $\theta_*$  in a predefined parameter space  $\Theta$  through an iterative procedure using as little computational resources as possible. Typically,  $\theta_*$  will be an optimizer, though it might also be the solution to a root-finding problem, as will be discussed shortly.

However, this characterization is still too broad and a discussion of all the publications that fit this description is undoubtedly far beyond the scope of this thesis. Instead, the discussion will be restricted to methods based on stochastic first-order oracles, i.e. variants of stochastic gradient descent.

The common goal of the presented literature is to find an algorithm within this framework that is usable in practice (i.e. does not require the practitioner to have prior information that is typically unavailable) and allows for some performance guarantees, such as upper bounds on the expected excess risk after  $n$  oracle calls. Corresponding information-theoretic lower bounds on the expected excess risk serve as a benchmark to evaluate whether a specific algorithm is optimal. By decomposing the behavior of the algorithm into a deterministic and a stochastic component, it is argued in the literature that we can find lower bounds by lending from optimization theory and statistical theory separately (see for example Dieuleveut et al. [2017]). The relevant results were discussed in Section 2.9. To give a very brief summary, assuming smoothness, we can hope for a linear reduction  $O(\rho^n)$  in the case of strong convexity and a sublinear convergence of order  $O(1/n^2)$  in the case of general convexity if the oracles are deterministic. When dealing with noisy oracles, these rates decrease to  $O(1/n)$  for the strongly convex case and  $O(1/\sqrt{n})$  for the

## 4. Related results and literature review

general convex case. For the case of least squares regression, we will certainly always have a term of the type  $\frac{\sigma^2 d}{n}$  which cannot be improved. The results on deterministic oracles are of interest insofar as they give us guidance on the possible convergence of the bias term, i.e. how fast the initial condition  $\theta_0$  of the iterative algorithm can be forgotten.

## 4.2. Preceding literature

### 4.2.1. Origins

The origin of gradient descent is commonly attributed to Cauchy [1847]. Stochastic gradient descent emerged much later in the work by Robbins and Monro [1951]. Stochastic gradient descent in its most general form is thus also known as the *Robbins–Monro procedure* in the literature. Their motivation lies in root-finding, i.e. finding the value  $\theta_* \in \mathbb{R}$  such that the function  $M_{RM} : \mathbb{R} \rightarrow \mathbb{R}$  satisfies  $M_{RM}(\theta_*) = \beta$ , for some given  $\beta \in \mathbb{R}$  (e.g.  $\beta = 0$ ).

Robbins and Monro [1951] introduce the following case. Let  $Z = Z(\theta)$  have distribution  $P_\theta$  and let  $M_{RM}(\theta) = \int Z dP_\theta$  denote the mean of  $Z$  for a given  $\theta$ . The authors further assume that the response  $Z$  is uniformly (in  $\theta$ )  $P_\theta$ -a.s. bounded. Regarding the shape of  $M_{RM}$ , they assume  $M_{RM}(\theta) \geq \beta$  if  $\theta > \theta_*$  and  $M_{RM}(\theta) \leq \beta$  if  $\theta < \theta_*$ . For an arbitrary starting point  $\theta_0$ , they then propose the iteration

$$\theta_n = \theta_{n-1} + \gamma_n(\beta - Z_n(\theta_{n-1})) \quad (4.1)$$

for all  $n \in \mathbb{N}$  and suitable  $\gamma_n > 0$ , where  $Z_n$  is a sequence of random variables with distribution depending on  $\theta_{n-1}$  through  $Z_n(\theta) \sim P_\theta$ , i.e.  $M_{RM}(\theta_{n-1}) = \mathbb{E}[Z_n(\theta_{n-1})|\theta_{n-1}]$ . The interpretation of the algorithm for the given root-finding problem is quite straightforward; if for some  $\theta_{n-1}$  the estimator  $Z_n(\theta_{n-1})$  for  $M_{RM}(\theta_{n-1})$  lies above  $\beta$ ,  $\theta_n$  is chosen smaller than  $\theta_{n-1}$  and conversely if  $Z_n(\theta_{n-1})$  lies below  $\beta$ . Indeed, Robbins and Monro [1951] show for certain classes of functions  $M_{RM}$  that  $\theta_n$  converges in quadratic mean (and thus also in probability) towards  $\theta_*$  for  $n \rightarrow \infty$  if  $\gamma_n = c/n$  for some constant  $c > 0$ . For example, one such class would be all non-decreasing functions  $M_{RM}$ , with  $M_{RM}(\theta_*) = \beta$  and  $M'_{RM}(\theta_*) > 0$ . However, Robbins and Monro [1951] point out themselves that their assumptions are quite restrictive and can probably be weakened.

The extension on their work by Kiefer and Wolfowitz [1952] comes already closer to what we now refer to as *stochastic gradient methods*. Specifically, they investigate a function  $M_{KW}$  of the same definition as  $M_{RM}$ , but assume that  $M_{KW}$  is strictly increasing on  $(-\infty, \theta_*)$  and strictly decreasing on  $(\theta_*, \infty)$ . The goal is then to find the maximizer  $\theta_*$ , through the sequentially sampled observations  $z_n(\theta)$ ,  $n \geq 1$ , at the points of interest  $\theta = \theta_{n-1}$ . Instead of boundedness of these observations, Kiefer and Wolfowitz [1952] only require finite second moments conditional

on  $\theta$ . To solve this maximization problem, they propose introducing a suitable sequence  $\delta_n > 0$ ,  $n \geq 1$ , and using the iteration

$$\theta_n = \theta_{n-1} + \gamma_n \frac{z_n^{(1)}(\theta_{n-1} + \delta_n) - z_n^{(2)}(\theta_{n-1} - \delta_n)}{\delta_n}$$

for  $n \in \mathbb{N}$  and an arbitrary  $\theta_0 \in \mathbb{R}$ . Here,  $z_n^{(1)}(\theta_{n-1} + \delta_n)$  and  $z_n^{(2)}(\theta_{n-1} - \delta_n)$  are independent and are both sampled from  $P_\theta$  with  $\theta = \theta_{n-1} + \delta_n$  and  $\theta = \theta_{n-1} - \delta_n$  respectively. The updates are thus defined by a very simple gradient estimator. Under some regularity conditions on  $M_{KW}$ , such as Lipschitz-continuity in the neighborhood of the optimum, the authors show convergence in probability of the sequence  $\theta_n$  towards  $\theta_*$  for  $n \rightarrow \infty$ .

Both Robbins and Monro [1951] and Kiefer and Wolfowitz [1952] require that the learning rate  $\gamma_n$  is chosen in such a way that  $\sum_{n=1}^{\infty} \gamma_n = \infty$  and  $\sum_{n=1}^{\infty} \gamma_n^2 < \infty$ . If  $\gamma_n = c/n^\alpha$  for some constant  $c > 0$  and some real value  $\alpha$ , this is satisfied for  $\alpha \in (1/2, 1]$ .

An early extension of the formulated problems is given by Blum [1954b]. His straight-forward generalization to the  $d$ -dimensional case includes the function  $M_{RM} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ , with  $M_{RM}(\theta) = \mathbb{E}[Z_n(\theta)|\theta]$  for the Robbins–Monro procedure with  $Z_n(\theta) \in \mathbb{R}^d$  and  $\theta \in \mathbb{R}^d$ . For the Kiefer–Wolfowitz procedure, they define analogously  $M_{KW} : \mathbb{R}^d \rightarrow \mathbb{R}$  through  $M_{KW}(\theta) = \mathbb{E}[z_n(\theta)|\theta]$ , where  $z_n(\theta) \in \mathbb{R}$  and  $\theta \in \mathbb{R}^d$ .

In this subsection, the Robbins–Monro procedure and the Kiefer–Wolfowitz procedure were both described in their original form. For the latter, the relationship to SGD is quite obvious. Specifically, in the notation used throughout the preceding chapters of this thesis, we could think of  $M_{KW}$  as the negative risk function  $-f$  and  $z_n(\theta_{n-1})$  as (for example) the loss incurred by a data point by choosing  $\theta = \theta_{n-1}$ .<sup>1</sup> For the Robbins–Monro procedure, set  $\beta = 0$ . The recurrence (4.1) thus simplifies to

$$\theta_n = \theta_{n-1} - \gamma_n Z_n(\theta_{n-1})$$

for  $n \in \mathbb{N}$  and arbitrary  $\theta_0 \in \mathbb{R}^d$  in the multivariate extension by Blum [1954b]. We can then think of  $M_{RM}(\theta)$  as the gradient of the risk  $\nabla f(\theta)$  and of the observation  $Z_n(\theta_{n-1})$  as the gradient estimator based on the  $n$ -th data point at the location  $\theta_{n-1}$ . As a consequence, results derived for the Robbins–Monro procedure translate directly to SGD, with assumptions on  $M_{RM}$  translating to assumptions on the gradient of  $f$ .

---

<sup>1</sup>Since at every point  $\theta_{n-1}$  we need two independent realizations of  $z_n$ , namely  $z_n^{(1)}$  and  $z_n^{(2)}$ , a sample of size  $n$  actually only allows us to carry out  $\lfloor n/2 \rfloor$  steps of the algorithm.

### 4.2.2. Subsequent asymptotic analysis

Practical advantages of the Robbins–Monro and the Kiefer–Wolfowitz procedure were soon proved. For example, Blum [1954a] shows convergence with probability 1 for both procedures in the univariate case. To that end, the author considers quite similar assumptions as Kiefer and Wolfowitz [1952], i.e. finite second moments of the sequentially sampled estimators  $Z_n$  (resp.  $z_n$ ) of  $M_{RM}$  (resp.  $M_{KW}$ ) and regularity conditions on  $M_{RM}$  (resp.  $M_{KW}$ ). Shortly after, in Blum [1954b], the author extends the results to the multivariate case introduced above.

Around the same time, back in the univariate case  $d = 1$ , Chung [1954] derives conditions under which  $n^{\alpha/2}(\theta_n - \theta_*)$  (with step size of the type  $\gamma_n = c/n^\alpha$  and suitable  $\alpha \in (1/2, 1]$ ) converges in distribution to a normal distribution with mean 0 in the univariate Robbins–Monro procedure, by calculating the (scaled) limits of all moments of  $\theta_n - \theta_*$ . Ruppert [1988] points out that the optimal choice for the result by Chung [1954] is  $\alpha = 1$ , in which case the step size sequence has to be chosen as  $\gamma_n = c/n$  with  $c > 1/(2M'_{RM}(\theta_*))$ . In this case, the variance of the limiting distribution turns out to be  $c^2\sigma^2/(2M'_{RM}(\theta_*)c - 1)$ . Here,  $\sigma^2$  is the variance of the response which is assumed homoscedastic by Chung [1954]. Ruppert [1988] further notes that the optimal choice of  $c$  is  $c_* = 1/M'_{RM}(\theta_*)$ , which is easy to see through analysis of the variance of the limiting distribution. However, in practice  $c_*$  will typically be unknown, as  $M'_{RM}(\theta_*)$  is unknown.

As mentioned by Ruppert [1988], in the 1960s so-called *adaptive* methods, which define the learning rate as  $\gamma_n = c_n/n$ , where  $c_n$  is a consistent estimator for  $c_* = 1/M'_{RM}(\theta_*)$ , were introduced to remedy this issue.

### 4.2.3. Iterate averaging in stochastic approximation

Iterate averaging is a different approach to tackle the problem of not knowing in practice which values of  $c$  can be chosen. In the context of risk minimization, recall that  $M_{RM}$  corresponds to the risk gradient  $\nabla f(\theta)$ . As a result, the derivative of  $M_{RM}$  evaluated at the optimum  $\nabla M_{RM}(\theta_*)$  corresponds to the Hessian of the risk evaluated at the optimum  $\nabla^2 f(\theta_*)$ . Polyak and Juditsky [1992] argue that while subsequent research on stochastic approximation was able to achieve optimal rates of convergence, the requirement to have knowledge of unknown quantities such as the matrix  $\nabla^2 f(\theta_*)$  is problematic in practice.<sup>2</sup>

In the papers by Ruppert [1988], Polyak [1990] and Polyak and Juditsky [1992], the authors use iterate averaging in order to reduce the amount of prior information needed. Polyak and Juditsky [1992] point out that the idea of using averaging is of course not new, but the given line of work achieved substantial improvements in the effects of iterate averaging by using learning rates that decrease more slowly than  $\gamma_n = c/n$ , as for example  $\gamma_n = c/n^\alpha$  with  $\alpha \in (1/2, 1)$ . Due to the

---

<sup>2</sup>As we will see, this translates to knowledge about the strong convexity constant  $\mu$ . Recall that for the least squares regression problem, we simply have  $H = \nabla^2 f(\theta_*)$  with  $\mu = \lambda_{\min}(H)$ .

importance of the works by Ruppert and Polyak, averaging the iterates of the Robbins–Monro process is often times called *Polyak(–Ruppert) averaging*.

In Ruppert [1988], the author considers the univariate Robbins–Monro procedure. The author shows that for  $\gamma_n = c/n^\alpha$  with arbitrary  $c > 0$  and  $\alpha \in (1/2, 1)$  the term  $\sqrt{n}(\bar{\theta}_n - \theta_*)$  converges in distribution to a normal distribution with mean 0 and minimal variance. Consequently, he manages to show (for the univariate case) that while the non-averaged Robbins–Monro procedure does not achieve the optimal rate of convergence for  $\alpha \in (1/2, 1)$ , the rate of convergence becomes optimal when the procedure is extended through iterate averaging. The benefit of this insight is that here, the choice of the step sizes can be carried out without any prior information on the unknown function  $M_{RM}$  (i.e.  $\nabla f$ ).

Polyak [1990] uses the same idea of combining iterate averaging with slowly decaying step sizes in multivariate Robbins–Monro procedures and shows convergence in the quadratic mean with asymptotic rate  $O(1/n)$  for several problems, including parameter estimation for least squares regression under certain assumptions, such as finite fourth moments and independent errors.

In the well-known paper by Polyak and Juditsky [1992], the authors extend these results in several settings. Under certain assumptions on the target function (such as smoothness and strong convexity), as well as regularity conditions on the errors, they derive convergence with probability 1 of the averaged iterates. Furthermore, they show that the scaled estimation errors  $\sqrt{n}(\bar{\theta}_n - \theta_*)$  converge in distribution to a multivariate normal distribution with mean 0. The results include the case of linear regression, where they achieve a minimal variance-covariance matrix, thus showing asymptotic efficiency of the averaged Robbins–Monro procedure with slowly decaying step sizes in the multivariate case.

#### 4.2.4. Constant step sizes for least squares regression

In gradient descent, the use of constant step sizes is quite common (see Subsection 2.6.1).

In stochastic gradient descent, however, the process will in general not converge for constant step sizes (see Pflug [1986]) and even when averaging is employed, the issue is not that simple. An enlightening analysis of this problem using Markov process theory and extending the analysis by Pflug [1986] is given by Dieuleveut et al. [2020], which will be discussed later. For now, a discussion of the case of least squares regression will be given.

In the case of least squares regression, Györfi and Walk [1996] point out that the previously presented optimality results by Polyak [1990] and Polyak and Juditsky [1992] do not translate to constant step size averaged SGD if we do not make the assumption  $x_n x_n^T = H$ , i.e. require the multiplicative error  $E_n$  to be uniformly equal to 0.<sup>3</sup> More specifically, in the case of no

---

<sup>3</sup>Note that this directly relates to the presented result by Bach and Moulines [2013], who show that optimal rates of convergence can indeed be recovered with a constant step size, even under the presence of multiplicative errors  $E_n$ .

#### 4. Related results and literature review

multiplicative errors (i.e. known  $H$ ) the least squares regression problem can be written as the linear problem presented in Polyak [1990, Section 2], for which an asymptotic rate of  $O(1/n)$  is achieved for a sufficiently small constant step size. For the results in Polyak and Juditsky [1992], the situation is quite similar. It is worth pointing out that in the presence of multiplicative errors  $E_n$ , Polyak [1990] does already recover an asymptotic rate of  $O(1/n)$  of the mean squared error for a constant step size, but the appearing constant is improvable, meaning that they cannot show asymptotic efficiency of the estimator.

As will be discussed later when the results by Dieuleveut et al. [2020] are introduced, averaged SGD with constant step size can give asymptotically biased estimators when dealing with more general problems than least squares regression. In the simpler case of least squares regression, this can also occur when the observations are not i.i.d., which is discussed by Györfi and Walk [1996].

##### 4.2.5. Non-asymptotic viewpoint

So far, we have discussed that, under certain conditions, averaged stochastic gradient descent with appropriate step size choice can be argued to be asymptotically optimal. However, asymptotic results have the substantial disadvantage that they might not be informative for finite sample applications. As a consequence, there has been some interest in analyzing the behavior of (averaged) SGD for all  $n \in \mathbb{N}$  (or at least for all  $n \geq n_0$  with some characterized  $n_0 \in \mathbb{N}$ ).

As a motivation, Spall [2003] argues why the promising asymptotic results achieved through averaging do not necessarily translate to finite sample sizes. In short, in Spall [2003, Chapter 4.5.3] the author argues that averaging improves stochastic gradient descent typically only when  $\theta_n$  is already “close” to the optimum  $\theta_*$  and oscillates around it. This oscillation can appear for example in situations with high noise or if the learning rate is chosen constant (for a simple illustrative example for the latter situation, see for example Bottou et al. [2018, Example 3.1]). As long as  $\theta_n$  is not close to the optimum  $\theta_*$ , numerical results have shown that averaging often times worsens the performance (Spall [2003, Chapter 4.5.3]). One possible explanation for these results is that while the reduction in variability of the estimator through averaging is helpful in the oscillation phase around  $\theta_*$ , the higher variability of non-averaged stochastic gradient descent can drive the estimator faster to the optimal value. As a result, even though asymptotic optimality of averaged stochastic gradient descent is established under certain sensible conditions, optimality in a finite setting is an entirely different issue.

In the previously presented literature, non-asymptotic performance was sometimes discussed (see for example Polyak [1990], where the author comments on the time of onset of the derived asymptotic behavior), but was not the focus of their publications. An analysis of the non-asymptotic convergence rate for (tail-averaged) stochastic gradient descent was provided in the well-known article by Nemirovski et al. [2009]. In this paper, the authors assume that the

parameter space  $\Theta \subset \mathbb{R}^d$  is convex, compact and non-empty. As a consequence, they have to modify the stochastic gradient descent algorithm by including projection steps, which were introduced in Subsection 2.7.3. It should be pointed out that this non-asymptotic analysis certainly was not the first, but their result was one of the motivations for the work by Bach and Moulines [2013].

First, Nemirovski et al. [2009] derive a non-asymptotic upper bound for the expected excess risk of the type  $C/n$  (for some constant  $C > 0$ ) for non-averaged SGD under the assumption of  $\mu$ -strong convexity, which will be presented shortly. To that end, they require a step size sequence of the form  $\gamma_n = c/n$ , where  $c > 1/(2\mu)$ . On the first glance, the achieved bound looks promising, as a rate of  $1/n$  is unimprovable. However, as the authors point out, the step size used to achieve this convergence rate depends on the strong-convexity parameter  $\mu$ . Since  $\mu$  is typically unknown, this is exactly the problem that Polyak and Juditsky [1992] saw in the preceding optimality results of non-averaged SGD. Nemirovski et al. [2009] make this concern more obvious by showing in a simple example with a quadratic function that a bad choice of  $c$ , due to a misconception of the value of  $\mu$ , can lead to extremely slow convergence. To illustrate the graveness of this issue, a slightly more general version of the example will be stated here.

**Example 9.** *Let  $d = 1$ ,  $\Theta = [-1, 1]$  and  $F(t) = \frac{\mu}{2}t^2$ . This function  $F : \Theta \rightarrow \mathbb{R}$  is univariate and quadratic and thus the most simple strongly convex, smooth function one can imagine. The minimizer is  $t_* = 0$ . Suppose the even simpler case that our gradient estimators  $\phi$  are identical to the gradient, i.e. we have access to deterministic first order oracles.*

*The iteration is therefore effectively given through*

$$t_n = t_{n-1} - \gamma_n F'(t_{n-1}) = (1 - \gamma_n \mu) t_{n-1}$$

*for  $n \in \mathbb{N}$  with some  $t_0 \in [-1, 1]$ .*

*The considered case is obviously quite trivial and optimally tuned gradient descent (i.e.  $\gamma_n = 1/L = 1/\mu$ ) would give immediate convergence, no matter the choice of  $t_0$ . However, this cannot be used as the benchmark if we impose the assumption that the first order oracle might be noisy, as it is then only reasonable to pick the step size accordingly. If we therefore choose  $\gamma_n = c/n$  with  $c > 1/(2\mu)$ , we will have a convergence of order  $1/n$  as we satisfy the assumptions by Nemirovski et al. [2009].*

*However, say we start in  $t_0 = 1$ . Following the same steps as Nemirovski et al. [2009] in a little bit more generality, one can construct the following lower bound for step sizes which are misspecified through the choice  $c < 1/(2\mu)$ :*

$$F(t_n) - F(t_*) > \frac{\mu}{4e} \cdot \frac{1}{n^{2c\mu}}.$$

*Again, recall that  $\mu$  is unknown, meaning that such a misspecification might happen in practice. The result is of course catastrophic, as the convergence rate gets arbitrarily low, since  $2c\mu$  goes*

#### 4. Related results and literature review

towards 0 with decreasing  $c$ . Note further that while  $\mu$  appears as a multiplicative factor, the dependence on  $n$  also gets worse when  $\mu$  decreases, if  $c$  is not adapted accordingly, as might happen when  $\mu$  is unknown.

Another issue of the result by Nemirovski et al. [2009] is the bound itself. The authors achieve for SGD with projection steps and  $\gamma_n = c/n$ , where  $c > 1/(2\mu)$ , a bound of the form

$$\mathbb{E} [f(\theta_{n-1}) - f(\theta_*)] \leq \frac{L}{2n} \cdot \max \left\{ \frac{c^2 M^2}{2\mu c - 1}, \|\theta_0 - \theta_*\|_2^2 \right\},$$

for all  $n \in \mathbb{N}$ . Here,  $M > 0$  is assumed to exist such that  $\mathbb{E} [\|\phi(y_1, x_1, \theta)\|_2^2] \leq M^2$  for any  $\theta \in \Theta$ . In other words, the variance of the gradient estimator is uniformly bounded on the domain. Note that this condition is not too restrictive, as  $\Theta$  is assumed compact. The first argument in the maximum-operator is minimal in  $c$  for  $c_* = 1/\mu$ . This gives the following lower bound of the upper bound under optimal tuning of the step size:

$$\frac{L}{2n} \cdot \max \left\{ \frac{c^2 M^2}{2\mu c - 1}, \|\theta_0 - \theta_*\|_2^2 \right\} \geq \frac{L}{2n} \cdot \frac{c^2 M^2}{2\mu c - 1} \geq \frac{L}{2n} \cdot \frac{(1/\mu)^2 M^2}{2\mu(1/\mu) - 1} = \frac{LM^2}{2n\mu^2} \geq \frac{M^2}{2n\mu},$$

using  $L \geq \mu$  in the last inequality. This shows that even under perfect tuning of the step size, i.e. choosing  $\gamma_n = 1/(n\mu)$ , the upper bound becomes uninformative if  $\mu$  decreases towards 0.<sup>4</sup>

Nemirovski et al. [2009] further investigate convex functions. Through averaged SGD with slowly decreasing step sizes  $\gamma_n = c/\sqrt{n}$  and decreasing averaging weights proportional to  $\gamma_n$ , the authors achieve a non-asymptotic expected excess risk bound of rate  $C/\sqrt{n}$  (for some constant  $C > 0$ ). As argued by the authors, this rate is optimal for the class of generally convex problems. It should however be pointed out that Nemirovski et al. [2009] assume a compact, convex, non-empty  $\Theta$  and  $c$  depends in a certain sense on the size of  $\Theta$  (and the choice of  $\theta_0 \in \Theta$ ), making the result not immediately informative if  $\Theta = \mathbb{R}^d$ .

These results give a good, non-asymptotic illustration of the problems SGD yields when trying to leverage strong convexity. Firstly, the choice of the step size becomes more difficult and secondly, the bounds quickly become less informative if  $\mu$  decreases.

A thorough examination of non-asymptotic bounds for the mean squared error and the expected excess risk of an estimator sequence is given by Bach and Moulines [2011], who consider different combinations of averaging, strong convexity, smoothness and boundedness of the gradient estimators. Note that the results for the mean squared error oftentimes directly translate to results

---

<sup>4</sup>Regarding the effect of  $\mu$  on  $M^2$ , note that for unbiased gradient estimators  $\mathbb{E} [\|\phi(y_1, x_1, \theta)\|_2^2] \geq \|\mathbb{E} [\phi(y_1, x_1, \theta)]\|_2^2 = \|\nabla f(\theta)\|_2^2$ . For a fixed domain  $\Theta$ , this lower bound on possible values of  $M^2$  can only be guaranteed to decrease towards 0 if  $L$  converged to 0, but not if merely  $\mu$  converges to 0.

on the expected excess risk, as explained in Section 2.4. Bach and Moulines [2011] main take-away is that averaging with slowly decaying step sizes can yield optimal rates for the strongly convex setting and close to optimal rates for the non-strongly convex setting, even with respect to their non-asymptotic behavior.

### 4.3. The result by Bach and Moulines [2013]

When looking at the specific case of least squares regression, we find ourselves in a very benign setting. Nevertheless, the literature presented so far gives unsatisfactory non-asymptotic bounds, as will be explained in the following.

Recall that as discussed in Chapter 3, the target function is  $L$ -smooth and, under the absence of perfect multicollinearity of the explanatory variables  $x_i$ ,  $\mu$ -strongly convex. Here,  $L$  coincides with the largest eigenvalue of  $H = \mathbb{E} [x_1 x_1^T]$ , and  $\mu$  with the smallest. Typically,  $\mu$  will be unknown and can therefore not be used for deciding the step size sequence. This makes Polyak–Ruppert averaging an appealing choice, as Bach and Moulines [2011] have derived non-asymptotic bounds for averaged SGD with optimal rates for strongly-convex functions.

Bach and Moulines [2013] identify the following problem: upper bounds on the expected excess risk grow with a rate of  $\Omega(1/\mu)$  for  $\mu \rightarrow 0$ , thus becoming rapidly less informative for a decreasing strong convexity parameter  $\mu$ . This can, for example, be observed in the results by Nemirovski et al. [2009] just presented. At the same time, for non-strongly convex functions we can only achieve a rate of  $O(1/\sqrt{n})$ . As Bach and Moulines [2013] point out, in a non-asymptotic setting this leads to the paradoxical situation that for very weak strong convexity (i.e.  $\mu$  close to 0) bounds derived without the assumption of strong convexity are in fact more informative than those with this additional assumption.

This is of course unsatisfactory and Bach and Moulines [2013] manage to show that for the case of least squares regression, it is possible to achieve an upper bound on the expected excess risk that decreases with a rate of  $C/n$  (for some constant  $C > 0$ ) for all  $n \in \mathbb{N}$  and does not grow with decreasing  $\mu$ . They point out that this is highly relevant in practice, as, especially in high-dimensional settings, explanatory variables might be highly correlated, thus leading to very small values of  $\mu$ . They achieve this bound by using constant step sizes in combination with Polyak–Ruppert averaging. Details of their result are discussed in Chapter 3.

### 4.4. Subsequent literature

As discussed, the variance term in the result by Bach and Moulines [2013] is unimprovable. As a consequence, the subsequent literature mainly deals with either improving the convergence of the bias term or generalizing the results.

For instance, with respect to the bias term, Bach and Moulines [2013] point out that in their earlier work (see Bach and Moulines [2011]), they have shown that averaged stochastic gradient

#### 4. Related results and literature review

descent can achieve a convergence of the bias term of order  $O(1/n^2)$ . While this specific result in Bach and Moulines [2011] does not give an optimal rate of convergence overall, the convergence of the bias term as  $O(1/n^2)$  is better than their results in Bach and Moulines [2013], where the bias term vanishes with  $O(1/n)$ .

As argued by Dieuleveut et al. [2017], the optimal rate of convergence of the bias term is known from the deterministic results in oracle based optimization, which were discussed in Section 2.9. As a consequence, under smoothness, a convergence of the sublinear rate  $O(1/n^2)$  of the bias term is optimal for general convexity and a linear rate of  $O(\rho^n)$  for some  $\rho \in [0, 1)$  is optimal for strongly convex functions.

##### 4.4.1. Improving the sublinear convergence of the bias

A natural extension of the work by Bach and Moulines [2013] is given by Defossez and Bach [2015]. Here, the authors also use a constant step size with averaging for the problem of least squares regression. Instead of upper bounds on the expected excess risk, Defossez and Bach [2015] derive information on the expected excess risk itself. Again, the step size has to be sufficiently small, i.e.  $\gamma \in (0, \gamma_{\max})$ , with  $\gamma_{\max} > 0$  depending on the distribution of the explanatory variables  $(x_i)_{i \in \mathbb{N}}$  in a slightly more involved way. As the authors point out, the allowed range is less strict than the one in Bach and Moulines [2013]. A sufficient condition is therefore that  $\gamma < 1/R^2$  and as a necessary condition the authors give  $\gamma < 2/\text{tr}(H) < 2/L$ .<sup>5</sup> Therefore,  $\gamma_{\max} \in (1/R^2, 2/\text{tr}(H))$ , which gives an idea of the extend of the relaxation. Using a more refined proof technique than Bach and Moulines [2013], Defossez and Bach [2015] show a convergence of order  $O(1/n^2)$  of the bias term. However, their bias term has several shortcomings. Most importantly, the leading order of  $O(1/n^2)$  of the bias term comes at the cost of  $\gamma^2$  in the denominator of this bound. As a result, as noted by the authors, small step sizes can lead to an explosion of the bias term. The authors note that the variance term - for which the authors recover the unimprovable rate  $O(1/n)$  - does not exhibit the same problem. As pointed out by Jain et al. [2018b], this unimprovable rate is, however, not achieved for finite  $n$  and a fixed step size  $\gamma \in (0, \gamma_{\max})$ . Instead, the asymptotic (in  $n$ ) behavior of the variance term is equivalent to  $O(1/n)$  in the limit of the step size  $\gamma \rightarrow 0$ . The proof technique for this result is similar to the one presented in Bach and Moulines [2013] with the difference that after the introduction of the necessary algebraic structures, the second moment matrix  $\mathbb{E}[\bar{\eta}_n \bar{\eta}_n^T]$  of the errors  $\bar{\eta}_n = \bar{\theta}_n - \theta_*$  is investigated instead of merely a norm of  $\bar{\eta}_n$ . This technique was adopted by several authors, such as Jain et al. [2018a,b].

Dieuleveut et al. [2017] also extend the work by Bach and Moulines [2013] without leaving the non-asymptotic setting with a fixed step size  $\gamma > 0$ . Here, the authors apply averaged SGD with constant step size to the  $l_2$ -regularized problem (i.e. if  $f$  denotes the least squares risk,

---

<sup>5</sup>Recall that  $\text{tr}(H) \leq R^2$ , i.e.  $2/\text{tr}(H) \geq 2/R^2 > 1/R^2$ .

$\theta \mapsto f(\theta) + \frac{c}{2}\|\theta - \theta_0\|_2^2$  is minimized over  $\mathbb{R}^d$  for some  $c > 0$ ).<sup>6</sup> They note that several of their results stay valid for  $c = 0$ , which is the unpenalized case considered by Bach and Moulines [2013]. For example, they achieve a non-asymptotic bound with the bias term decreasing as  $C/n^2$  (for some constant  $C > 0$ ) while retaining the optimal rate for the variance under mostly the same assumptions as Bach and Moulines [2013]. More specifically, they reach the following bound for small constant step sizes  $\gamma \in (0, 1/(2R^2))$ :

$$\mathbb{E} [f(\bar{\theta}_n) - f(\theta_*)] \leq \frac{3(1+d)}{\gamma^2 n^2} \left\| H^{-1/2}(\theta_0 - \theta_*) \right\|_2^2 + \frac{6\sigma^2 d}{n+1}.$$

An important drawback of this result is the appearance of  $H^{-1/2}$  in the norm in the bias term. Note that since  $H$  is real and symmetric, it can be diagonalized through  $H = U\Lambda U^T$  where  $U$  is orthogonal and  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$  the diagonal matrix containing the eigenvalues of  $H$ . Define  $v = U(\theta_0 - \theta_*)$  and let  $\lambda_{\min} = \min_{i \in \{1, \dots, d\}} \lambda_i$ . It then holds that

$$\left\| H^{-1/2}(\theta_0 - \theta_*) \right\|_2^2 = v^T \Lambda^{-1} v = \sum_{i=1}^d \frac{(v^{(i)})^2}{\lambda_i} \geq \frac{1}{\lambda_{\min}} \|v\|_2^2 = \frac{1}{\lambda_{\min}} \|\theta_0 - \theta_*\|_2^2$$

where  $v^{(i)}$  denotes the  $i$ -th element of the vector  $v$ . Recall that the strong convexity constant  $\mu$  is equal to  $\lambda_{\min}$ . The above result therefore shows that the bias term increases at least as fast as  $1/\mu$  if  $\mu$  decreases.

Note that the authors improve this by using the following algorithm, which performs a two-step iteration for some  $\theta_0 = \nu_0 \in \mathbb{R}^d$  using a momentum coefficient  $\delta \in \mathbb{R}$  and a gradient estimator  $\phi(y_n, x_n, \theta)$  at step  $n \in \mathbb{N}$ :

$$\begin{aligned} \theta_n &= \nu_{n-1} - \gamma \phi(y_n, x_n, \nu_{n-1}) \\ \nu_n &= \theta_n + \delta(\theta_n - \theta_{n-1}), \end{aligned}$$

for  $n \in \mathbb{N}$ . The first line can be seen as the regular SGD update step, while the second line gives an acceleration through the addition of a momentum term. For the averaged iterates of this *accelerated stochastic gradient descent* method, Dieuleveut et al. [2017] achieve the bound

$$\mathbb{E} [f(\bar{\theta}_n) - f(\theta_*)] \leq 36 \frac{\|\theta_0 - \theta_*\|_2^2}{\gamma(n+1)^2} + 8 \frac{\sigma^2 d}{n+1}$$

for all  $n \in \mathbb{N}$  with  $\delta = 1$  and  $\gamma > 0$  such that  $\gamma H \preceq I$  (e.g.  $\gamma \leq 1/\text{tr}(H)$  and thus in particular if  $\gamma \leq 1/R^2$ ). However, this result is only achieved in the case that the matrix  $H$  is known and

---

<sup>6</sup>Note that also for other risk functions,  $l_2$ -penalization can be helpful for optimization. Specifically, it has the advantage that even if  $f$  was only generally convex instead of strongly convex, the regularized function is automatically  $c$ -strongly convex. As a result, especially for GD (say for minimizing the empirical risk based on non-strongly convex losses), adding  $l_2$ -penalization can lead to much faster convergence.

#### 4. Related results and literature review

used in the construction of the gradient estimate (i.e.  $E_n = 0$  for all  $n \geq 1$ ). More specifically, they achieve this result by using  $\phi(y_n, x_n, \theta) = H\theta - y_n x_n$ . Recall from Bach and Moulines [2013] that we have there  $\phi(y_n, x_n, \theta) = x_n x_n^T \theta - y_n x_n$ , which is a more realistic case for random predictors  $x_n$ .

##### 4.4.2. Linear convergence of the bias

As Jain et al. [2018b] point out, averaging regulates the trade-off between the rate of convergence of the bias term and the one of the variance term. For regular SGD under strong convexity and  $\gamma_n = c/n^\alpha$  with optimally chosen  $c > 0$ , the bias term converges linearly, while the variance term converges as  $O(1/n^\alpha)$  for  $\alpha \in (0, 1)$ , which is suboptimal (see Bach and Moulines [2011, Theorem 1]). For averaged SGD, Jain et al. [2018b] argue that the convergence of the variance is sped up to its optimal rate of  $O(1/n)$  at the cost of a slower convergence of the bias.

In Jain et al. [2018b] the authors make this explicit by analyzing the effects of tail-averaging. They further generalize their results by simultaneously allowing for mini-batching (see Subsection 2.7.3). More specifically, they provide an upper bound for the generalization error given that the direction at each iteration of SGD is based on the average of  $b \in \mathbb{N}$  gradient estimators, each of which are calculated using one data point.<sup>7</sup> As a result, only  $\lfloor n/b \rfloor$  iterations are possible with  $n$  oracle calls. Averaging is then not carried out with respect to all  $\lfloor n/b \rfloor$  iterates, but only with respect to the later iterates, ignoring the first  $s \in \mathbb{N}$  iterates. Oftentimes, a moving window approach is chosen, where  $s = \Theta(n)$ . For example, Nemirovski et al. [2009] and Jain et al. [2018b] suggest  $s = r \cdot n$  for some constant  $r \in (0, 1)$ .<sup>8</sup> For batch size  $b \geq 1$ , let the constant step size  $\gamma > 0$  be sufficiently small such that it satisfies  $\gamma \leq \gamma_{b,\max}/2$ , where  $\gamma_{b,\max} = 2b/(R^2 \rho_m + (b-1)L)$  and  $\rho_m \geq 1$  is a problem-specific constant introduced to quantify the extend to which the model is wrongly specified. If the model is correctly specified with i.i.d. errors, then  $\rho_m = 1$ . Note that the maximally allowed step size decreases with increasing model misspecification  $\rho_m$  and increases with increasing batch size  $b$ . Both dependencies seem intuitively sensible. In the later case, better gradient estimates mean more informed steps and thus allow for larger updates, while in the former case, under higher model misspecification, more cautious updates are recommendable. For the case of correct model specification (i.e.  $\rho_m = 1$ ) and no mini-batching (i.e.  $b = 1$ ), we simply have  $\gamma \in (0, 1/R^2]$ , which is almost identical to the step size requirement by Bach and Moulines [2013]. If we would - for the sake of the argument - let the batch size  $b$  tend to infinity, we would need  $\gamma$  to be smaller than  $1/L$ , which was the step size required to reach optimal rate of convergence for deterministic gradient descent (see Subsection 2.6.1). Under these considerations, for correctly specified models, the step size by Jain et al. [2018b] seems like a sensible interpolation over  $b$  between highly informative gradient estimates (i.e. large batches or exact first order oracles) and less informative gradient estimates

<sup>7</sup>Note that mini-batching SGD with batch sizes  $b = 1$  is identical to regular SGD.

<sup>8</sup> $r = 0$  is also possible and leads to regular Polyak–Ruppert averaging.

(i.e. no batches or regular SGD).

Let the estimator resulting from mini-batching tail-averaged SGD be denoted by

$$\bar{\theta}_{s+1}^{\lfloor n/b \rfloor} = \frac{1}{\lfloor n/b \rfloor - s} \sum_{s < i \leq \lfloor n/b \rfloor} \theta_i.$$

The authors then achieve the following non-asymptotic bound for all  $n \in \mathbb{N}$ ,  $b \in \mathbb{N}$ ,  $s < \lfloor n/b \rfloor$ :

$$\mathbb{E} \left[ f \left( \bar{\theta}_{s+1}^{\lfloor n/b \rfloor} \right) - f(\theta_*) \right] \leq \frac{2}{\gamma^2 \mu^2} \cdot \frac{(1 - \gamma \mu)^s}{(\frac{n}{b} - s)^2} \cdot (f(\theta_0) - f(\theta_*)) + \frac{8\tau^2}{b \cdot (\frac{n}{b} - s)}$$

for

$$\tau^2 = \frac{1}{2} \text{tr} \left( H^{-1} \mathbb{E} \left[ \varepsilon_1^2 x_1 x_1^T \right] \right).$$

Note that by choosing  $s = r \cdot n$  for  $r \in (0, 1)$  the expressions  $s$  and  $\frac{n}{b} - s$  grow linearly in  $n$ . The first summand is the bias term which decreases linearly towards 0 with increasing  $s$ . By setting  $\rho = (1 - \gamma \mu)^r$ , the influence of the bias term (first summand) decays as  $O(\rho^n/n^2)$ , which is linear, since  $0 < \gamma \mu \leq 1$  for  $\gamma \leq \gamma_{b,\max}/2$ , as the last inequality implies  $\gamma < 1/L$ . At the same time, the variance term (second summand) decays as  $O(1/n)$ , which is optimal up to constants depending on  $b$  and  $r$ . It should be noted that the bias term of this and the following bounds have again the deficiency Bach and Moulines [2013] tried to remedy, namely the problematic dependence on  $\mu$ .

Jain et al. [2018b] further show that their bound can be stated for the case of  $\gamma = 1/R^2$  without mini-batching (i.e.  $b = 1$ ) in a correctly specified model with homoscedastic, additive errors, i.e. when  $\tau^2 = d\sigma^2/2$ . In that case, they derive the following upper bound:

$$\mathbb{E} [f(\bar{\theta}_{s+1}^n) - f(\theta_*)] \leq \frac{2R^4}{(n-s)^2 \mu^2} e^{-s\mu/R^2} (f(\theta_0) - f(\theta_*)) + 4 \frac{d\sigma^2}{n-s}. \quad (4.2)$$

This result is very telling, as it shows the effect of tail-averaging as opposed to complete averaging. For  $s = 0$  (i.e. complete averaging), the variance term is optimal up to constants, but the bias only decreases sublinearly. However, if one chooses  $s$  such that it increases linearly in  $n$ , the bias term converges linearly at the cost of increasing the variance term by a constant factor depending on the exact relationship of  $s$  and  $n$ .

As Jain et al. [2018b] point out, this last result for  $s = 0$  achieves the convergence rate of  $O(1/n^2)$  in the bias, while maintaining minimax-optimality of the variance term, thus giving the same rates of convergence as Defossez and Bach [2015] even with a fixed step size.

The proof technique used by Jain et al. [2018b] to achieve the given bounds is based on the same framework as Bach and Moulines [2013], meaning that it uses bias-variance-decomposition and subsequent separate analysis of the two components.

While the presented bounds are quite informative, they have some obvious drawbacks. Most

#### 4. Related results and literature review

importantly, dependence on the strong convexity parameter  $\mu$  is reintroduced in such a way that the bias term becomes very large for small values of  $\mu$ . As a consequence, it loses one of the prime benefits of the bound by Bach and Moulines [2013]. Furthermore, while the interdependence of the upper bound for the step size  $\gamma$  and the batch size  $b$  has an intuitive interpretation, in practice it depends on the unknown quantities  $R^2$ ,  $L$  and  $\rho_m$ .

Finally, one should point out the following caveat. Taking  $s = r \cdot n$  with  $r \in (0, 1)$ , as commonly proposed, leads to the following reformulation of the preceding bound (4.2):

$$\mathbb{E} [f(\bar{\theta}_{s+1}^n) - f(\theta_*)] \leq \frac{2R^4}{n^2(1-r)^2\mu^2} e^{-rn\mu/R^2} (f(\theta_0) - f(\theta_*)) + 4 \frac{d\sigma^2}{n(1-r)}.$$

One might now be tempted to make the statement that we have reached a bound that decreases the bias term linearly and has an (up to constants) optimal variance term. While this may be correct, for small sample sizes the optimal choice of  $r$  might be highly relevant, but quite difficult.

#### The influence of model misspecification and mini-batching on the step size

The constant learning rate  $\gamma$  employed by Jain et al. [2018b] has to be in the range  $(0, \gamma_{b,\max}/2)$ , with  $\gamma_{b,\max} = 2b/(R^2\rho_m + (b-1)L)$ . The quantity  $\rho_m \geq 1$  describes the extend of the misspecification and takes the value  $\rho_m = 1$  if the model is correctly specified with i.i.d. errors (for details see Jain et al. [2018b]). Note that the influence of  $\rho_m$  on the step size becomes insignificant for very large  $b$ . Conversely, for  $b = 1$ , we have  $\gamma_{1,\max} = \frac{2}{R^2\rho_m}$  as Jain et al. [2018b] point out. This result illustrates that mini-batching can be used to remedy the restricting effect model misspecification has on the allowed step size.

Furthermore, Jain et al. [2018b] point out that their maximum step size without mini-batching is thus  $\gamma_{1,\max}/2 = 1/(R^2\rho_m)$ . For the correctly specified case, this is obviously similar to Bach and Moulines [2013].

The two preceding arguments illustrate that mini-batching can keep the learning rate large if the model is misspecified, while the absence of mini-batching can lead to the need for small learning rates under model misspecification.

#### 4.4.3. Geometrically decaying averaging weights

A variety of literature considers a penalized risk function, achieved for example through  $l_2$ -regularization. In Subsection 4.4.1, this was already mentioned and results by Dieuleveut et al. [2017], which belong to this line of work, were presented. However, tuning of the regularization parameter can be costly, as argued by Neu and Rosasco [2018]. Interestingly, for the case of least squares regression, where  $l_2$ -regularization amounts to ridge regression, Neu and Rosasco [2018] build on the work by Dieuleveut et al. [2017] and show that the effect of regularization with parameter  $c > 0$  can be recovered through an appropriate averaging scheme, where  $c$  enters through the chosen weights. To that end, they propose *geometric Polyak–Ruppert averaging*,

which can be carried out instead of applying SGD to the regularized risk. The proposed scheme is given through

$$\tilde{\theta}_n = \frac{1}{\sum_{i=0}^n (1 + \gamma c)^{-i}} \sum_{j=0}^n (1 + \gamma c)^{-j} \theta_j,$$

where  $(\theta_j)_{j \geq 0}$  is the result of constant step size SGD for the non-penalized risk function with sufficiently small learning rate  $\gamma > 0$  and  $c > 0$  is the penalization parameter. We evidently have  $1 + \gamma c > 1$ , meaning that the weights  $(1 + \gamma c)^{-j}$  decrease geometrically in  $j$ . For  $c = 0$  (i.e. no penalization) we recover equal weights, which Bach and Moulines [2013] have shown to be a good choice for the non-penalized risk. Neu and Rosasco [2018] show that this approach recovers the ridge regression estimator through SGD, while allowing to set the appropriate regularization parameter  $c$  after performing the costly calculation of the SGD iterates.

Using this construction, they show that for sufficiently small step sizes  $\gamma > 0$  (e.g.  $\gamma < 1/R^2$  is sufficient), the expected value of this estimator converges to the solution of the  $l_2$ -regularized risk with parameter  $c$ . Furthermore, for  $\gamma \leq 1/(2R^2)$  and  $c < 1/\gamma$ , they give a non-asymptotic bound on the expected excess risk. Neu and Rosasco [2018] point out that their results are similar to the ones achieved for SGD for the regularized risk by Dieuleveut et al. [2017].

Lastly, as discussed by Neu and Rosasco [2018], their result is quite the opposite approach to tail-averaging. Instead of omitting the initial iterates, geometric Polyak–Ruppert averaging puts a lot of weight on them. This can be interpreted as regularization through a probabilistic early stopping method, as noted by the authors. Specifically, the time at which SGD is to be stopped is chosen according to a geometrically distributed random variable with mean  $(1 + \gamma c)/(\gamma c)$  conditional on running no more than  $n$  iterations and deciding the stopping time independently of the iterates  $\theta_i$ ,  $i \geq 1$ .

As a last note, Nemirovski et al. [2009] also employ decreasing weights, but there the weight for the  $i$ -th iterate is proportional to  $1/\sqrt{i}$ , meaning that the weights decrease much more slowly.

#### 4.4.4. Characterization through Markov chains

Over the course of this thesis, it was discussed how stochastic gradient descent with constant step size and averaging can achieve optimal results for least squares regression. One obvious question is, whether these properties translate to other problems and if not, why. One particularly useful tool to answer this question is the framework of Markov processes. Such an approach was, for example, chosen by Pflug [1986], Pflug [1990] and more recently by Dieuleveut et al. [2020]. Bach and Moulines [2013] also explain the behavior of averaged SGD with constant learning rate for least squares regression using similar ideas as Dieuleveut et al. [2020].

As shown in Dieuleveut et al. [2020], under certain assumptions including strong convexity of  $f$  and uniform boundedness of the observations  $x_n$  (which can be weakened for some results), the iterates of SGD  $(\theta_n)_{n \geq 0}$  have a unique stationary distribution, which does not depend on the starting point (or possibly even starting distribution) of  $\theta_0$ . This stationary distribution does,

#### 4. Related results and literature review

however, depend on the learning rate  $\gamma$  and the corresponding invariant probability measure will therefore be denoted by  $\pi_\gamma$ . The convergence to a stationary distribution and the dependence of the same on  $\gamma$  were already noted by Pflug [1986], who shows that constant step size SGD converges to a stationary distribution  $\pi_\gamma$ . Under suitable conditions, the author shows that a random variable  $\vartheta \sim \pi_\gamma$  converges in distribution to a normal distribution for  $\gamma \rightarrow 0$  if it is adequately scaled through  $\gamma^{-1/2}(\vartheta - \theta_*)$ . Dieuleveut et al. [2020] note that this implies an average oscillation of the order  $\sqrt{\gamma}$  of the non-averaged iterates of SGD in the limit of  $n \rightarrow \infty$  around the optimum  $\theta_*$  and investigate the effect of averaging. Let  $\bar{\theta}_\gamma$  denote the mean of the distribution  $\pi_\gamma$ . Bach and Moulines [2013] and Dieuleveut et al. [2020] show that while  $\theta_* = \bar{\theta}_\gamma$  for quadratic risk  $f$ , this equality does not need to hold for a general risk function  $f$ . As a result, while one might hope that the result of averaged SGD  $\bar{\theta}_n$  converges to  $\bar{\theta}_\gamma$  in some sense (see for example Pflug [1990, Lemma 5]), there will be an error incurred by the so-called *drift*  $\bar{\theta}_\gamma - \theta_*$ . The special case of the drift being 0 for a quadratic risk  $f$  stems from the fact that the derivative  $\nabla f$  is consequently a linear function (see Bach and Moulines [2013] or Dieuleveut et al. [2020]). The results by Dieuleveut et al. [2020] provide linear convergence of the distribution of the iterates of SGD  $\theta_n$  towards the stationary distribution in terms of the Wasserstein distance (of order 2) with non-asymptotic bounds.<sup>9</sup> Such a linear convergence has already been noted by Pflug [1990]. Furthermore, Dieuleveut et al. [2020] show that for general strongly convex, smooth  $f$ , the drift  $\bar{\theta}_\gamma - \theta_*$  is of the order  $\gamma$  (plus a term decaying as  $O(\gamma^2)$ ). Furthermore, a random vector  $\vartheta$  following the stationary distribution  $\pi_\gamma$  has a root mean squared error for  $\theta_*$  of the order  $\sqrt{\gamma}$  (plus a term decaying as  $O(\gamma)$ ), or as Dieuleveut et al. [2020] put it through the mean squared error matrix

$$\mathbb{E} \left[ (\vartheta - \theta_*)(\vartheta - \theta_*)^T \right] = \Theta(\gamma) + O(\gamma^2).$$

As a result, for small step sizes  $\gamma$  and in the limit  $n \rightarrow \infty$  where the iterates converge to stationarity, Dieuleveut et al. [2020] point out that the final iterate of SGD fluctuates around  $\theta_*$  in a radius of about  $\sqrt{\gamma}$  with the expectation  $\bar{\theta}_\gamma$  of the stationary iterates having a distance of  $\gamma$  from  $\theta_*$ . As  $\sqrt{\gamma} > \gamma$  for  $\gamma < 1$ , this formalizes and illustrates the advantage of averaging the iterates of SGD if the learning rate is to be kept constant and small, even in the non-quadratic case. As discussed, for the quadratic case, we even have  $\bar{\theta}_\gamma = \theta_*$ , making the advantage of using the average sequence  $\bar{\theta}_n$  to approximate  $\bar{\theta}_\gamma$  obvious.

### 4.5. Further related results

In this section, some further results will be presented, which are relevant to the topic, but can neither be classified as subsequent results building on Bach and Moulines [2013] nor as prior results leading towards Bach and Moulines [2013]. The first subsection will give more of an

---

<sup>9</sup>It should be noted that the base  $\rho \in [0, 1)$  of the linear convergence rate  $O(\rho^n)$  increases towards 1 if  $\mu \rightarrow 0$ .

overview regarding different choices of the step size. Afterwards, a practitioner-oriented comment on the oscillation phase of constant step size SGD will be given.

#### 4.5.1. Choice of the step size sequence

The most common choices of the step size sequence are of the type  $\gamma_n = c/n^\alpha$ , with  $\alpha \geq 0$  and  $c > 0$  (or even  $\gamma_n = c/(n+C)^\alpha$  for some constant  $C \geq 0$ ). In the literature, theoretic results, like worst-case excess risk bounds and convergence rates, are based on a specific choice of  $\alpha$  under the problematic assumption of appropriate (typically problem-specific) choice of  $c$ . This can include a dependence on quantities such as  $\mu$ ,  $L$  or  $R^2$ . One illustrative example of the problem of  $\gamma$  depending on  $\mu$  for the case  $\alpha = 1$  was discussed in Example 9.

One body of literature deals with the constant step size policy  $\gamma_n = \gamma$ , which is achieved by choosing  $\alpha = 0$  (and  $c = \gamma$ ). As argued by Dieuleveut et al. [2020], this is particularly appealing in practice, as it requires the appropriate selection of only one parameter (namely  $\gamma$ ) instead of several jointly (namely  $c$ ,  $\alpha$  and possibly  $C$ ). In the context of averaged SGD, see, for example, Polyak and Juditsky [1992], Györfi and Walk [1996], Bach and Moulines [2013], Defossez and Bach [2015], Dieuleveut et al. [2017], Jain et al. [2018b]. It should be noted that Polyak and Juditsky [1992] also consider decreasing step sizes, as explained below. By viewing the sequence of iterates resulting from SGD as a Markov chain, Dieuleveut et al. [2020] provide an insightful analysis of the drawbacks of choosing a constant step size for non-quadratic target functions. This mainly concerns the convergence of the iterates to a non-degenerate stationary distribution and the convergence of averaged SGD to the mean of this stationary distribution, which in general does not coincide with the minimum of the target function. Details are provided in Subsection 4.4.4.

At its inception, the learning rate for SGD was chosen with  $\alpha = 1$ , stemming from the requirements

$$\sum_{n=1}^{\infty} \gamma_n = \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \gamma_n^2 < \infty, \quad (4.3)$$

as given in Robbins and Monro [1951]. Intuitively these conditions seem sensible as the former guarantees that the learning rate decreases slowly enough to avoid a significant slow down before the optimum is reached (i.e. before the gradient estimator is typically small), while the latter guarantees that the learning rate converges sufficiently quickly, in order to avoid or reduce the oscillating behavior of SGD observed for example for constant step size policies. Interestingly, as pointed out by Pflug [1990], for well-behaved target functions (e.g. smooth and strongly convex), the former of the two conditions in (4.3) together with convergence of  $\gamma_n$  towards 0 is sufficient for convergence of the iterates of SGD in probability. The later condition in (4.3) is then needed to also guarantee almost sure convergence. The conditions in (4.3) are satisfied for the slower learning rates achieved through  $\alpha \in (1/2, 1)$ , which are in certain settings asymptotically

#### 4. Related results and literature review

efficient for averaged SGD (see Ruppert [1988], Polyak and Juditsky [1992] and Spall [2003, Subsection 4.5.3]). Bach and Moulines [2011] provide a useful overview over non-asymptotic convergence rates with and without iterate averaging under a variety of assumptions based on different values of  $\alpha \in (0, 1)$  in the setting  $\gamma_n = c/n^\alpha$ .

An example for an application of a 3-parameter learning rate sequence  $\gamma_n = c/(n+C)^\alpha$  satisfying (4.3) is given by Bottou et al. [2018, Theorem 4.7]. There, the authors incorporate properties of the gradient estimator, such as its bias and its variance, into the choice of the step size sequence. More specifically, they use  $\gamma_n = c/(n+C)$ , where  $C > 0$  is a constant that has to be chosen sufficiently large in order to account for the mentioned properties of the gradient estimator. However, in practice, these might not be known.

##### 4.5.2. Detection of the stationary oscillation phase in practice

As discussed in Subsection 4.4.4, Dieuleveut et al. [2020] derive a linear convergence rate (in terms of the second order Wasserstein distance) towards a stationary distribution characterizing the oscillation of the iterates of SGD with constant step size  $\gamma > 0$  around some point  $\bar{\theta}_\gamma$  which is equal to the optimum  $\theta_*$  if the target function is quadratic or else lies in a sphere with radius of order  $\sqrt{\gamma}$  around the optimum  $\theta_*$ .

In practice, it might not be easy to judge whether we have reached this stationary oscillation phase or whether we are still in the transient phase in which the starting point has not yet been forgotten. A practitioner-oriented avenue to answer this question is pursued in the recent paper by Chee and Toulis [2018], who analyze a method which they name *Pflug diagnostic*, given that it is largely based on ideas by Pflug [1990]. The underlying idea is to investigate whether subsequent improvement directions tend to have a positive or a negative inner product. While a positive inner product indicates movement of the iterates in roughly the same direction, negative results signify direction changes which typically happen frequently in the oscillation phase. Pflug [1990] proposes a test statistic based on this idea, for which such inner products are calculated and the running mean over these products is recorded. By looking at the time at which this running mean first turns negative (after a suitable burn-in time), the author receives a stopping time. Note that a negative running mean indicates that the negative inner products induced through the oscillation phase outweigh the positive ones from the transient phase. Through terminating SGD at this stopping time and calculating the average of the iterates, Pflug [1990] proposes a confidence region for the optimum  $\theta_*$  which is uniform in the choice of the initialization  $\theta_0$ .

Chee and Toulis [2018] investigate a simplification and calculate the inner products of successive loss gradients  $\nabla_{\theta} l(y_n, \theta^T x_n)|_{\theta=\theta_{n-1}}$  and  $\nabla_{\theta} l(y_{n+1}, \theta^T x_{n+1})|_{\theta=\theta_n}$ , which are commonly used as gradient estimators. For quadratic loss functions, the authors show that the region in which the inner product of subsequent gradient estimates is expected to be negative is an ellipsis around the optimum  $\theta_*$ .

The analyses by Pflug [1990] and Chee and Toulis [2018] are not purely of theoretical interest. As Bottou et al. [2018] point out, it is quite common in practice to use a constant step size and

reduce it once it seems that no further improvement is made. Given the results by Dieuleveut et al. [2020] regarding the oscillation in a radius of order  $\sqrt{\gamma}$  of the stationary distribution around the optimum  $\theta_*$ , such an approach makes sense. Pflug [1990] and Chee and Toulis [2018] propose to use their measures of stationarity in order to decide, when the step size should be reduced. More specifically, they propose running SGD with step size  $\gamma$  until the running mean of inner products is negative and then restarting SGD at the last point with step size  $\gamma/2$ . Repeating this idea iteratively will then lead to convergence of SGD with constant step size even without averaging the iterates, as the radius of oscillation is successively shrunk.

These ideas are of course not new and are related to the concept of acceleration mentioned in Subsection 2.7.3.

### 4.5.3. Finite horizon

Many authors consider a setting in which the total number of iterations  $N$  is known in advance. In a way, the iterations are still performed online, as each sample is called into memory exactly once and then discarded, but the steps are not carried out indefinitely. Assuming such a finite horizon, one can incorporate  $N$  in the design of the algorithm. For instance, the step size  $\gamma_n$  at iteration  $n \in \{1, \dots, N\}$  may depend on  $N$ . Considering the insights by Pflug [1990] and Dieuleveut et al. [2020] (see Subsection 4.4.4) that for certain target functions the iterates of SGD with constant step size  $\gamma$  converge to a stationary distribution with a mean distance of order  $O(\sqrt{\gamma})$  to the optimum  $\theta_*$  justifies the idea of choosing the step size constant in  $n$ , but decreasing in  $N$ . Such a choice of the step size sequence then allows the claim that the iterates resulting from constant step size SGD converge towards the optimum for  $N \rightarrow \infty$ . While these claims critically depend on the assumption of having a finite horizon of known size  $N$ , Bach [2014] points out that such results can still be of interest for fully online schemes by restarting the process with a new batch of data points and a reduced constant step size once the first horizon  $N$  is reached.

One example for a finite horizon setting is presented in Nemirovski et al. [2009], where the authors recover the optimal rate  $O(1/\sqrt{N})$  for generally convex functions on a compact domain  $\Theta$ . This is reached by using averaged SGD with constant step size  $\gamma \propto 1/\sqrt{N}$ .

Another example is the result by Bach [2014]. Here, the author considers self-concordant target functions, i.e. functions where the third derivative is effectively bounded by the second derivative. Further, he considers Lipschitz-continuity and convexity of  $f : \Theta \rightarrow \mathbb{R}$ , but notably no strong convexity. This makes the results applicable to logistic regression (and other generalized linear models with bounded explanatory variables), even without restricting  $\Theta$  to a compact set and instead also allowing  $\Theta = \mathbb{R}^d$ . Using a constant step size  $\gamma \propto 1/\sqrt{N}$ , he achieves a non-asymptotic upper bound on the expected excess risk for averaged SGD of the type  $C/(N\mu_*)$  (for some constant  $C > 0$ ), where  $\mu_*$  is the *local strong convexity constant*, given by the smallest eigenvalue of the Hessian at the optimum  $\theta_*$ , which is assumed to exist. It should be noted that the step size is chosen independently of  $\mu_*$ .

#### 4. Related results and literature review

This is further improved upon by Bach and Moulines [2013, Theorem 3], where the authors achieve a rate of  $O(1/N)$  for logistic regression with bounded explanatory variables without the assumption of any local strong convexity of the target function. The result is of a non-asymptotic nature, but requires  $N$  to be sufficiently large with respect to, on one hand, the radius in which the explanatory variables lie and, on the other hand, the distance of  $\theta_0$  to the optimum  $\theta_*$ . Their bound is achieved through performing  $\lfloor N/2 \rfloor$  steps of averaged SGD on the target function of logistic regression with a constant step size  $\gamma \propto 1/\sqrt{N}$ , before then approximating the target function in the reached point through a quadratic function and performing averaged SGD with constant step size with the remaining data points on the approximating quadratic function.

## 5. Discussion

Over the course of this thesis, the very basics of optimization with (noisy) first-order oracles were discussed by introduction of the necessary terminology, assumptions and algorithmic variants. It was illustrated how gradient descent achieves close-to-optimal performance when constant step sizes are used, but it was pointed out that gradient descent is too dependent on the condition number  $\kappa = L/\mu$  of the problem for  $L$ -smooth,  $\mu$ -strongly convex functions. For stochastic gradient descent for strongly convex functions, dependence on  $\mu$  is a problem as well.

In the context of stochastic gradient descent, we face a trade-off governed by the choice of step size. In the case of smooth, strongly convex target functions, mirroring the behavior of gradient descent, constant step sizes can be used to forget the initial conditions exponentially fast, but the iterates will then not converge. Using the historically first choice of step sizes  $\gamma_n \propto 1/n$  will lead to optimal convergence rates, but require the practitioner to have knowledge about  $\mu$  in order to properly tune the step size. While this was remedied by Ruppert [1988] and Polyak and Juditsky [1992] through averaging with slowly decaying step sizes  $\gamma_n \propto 1/n^\alpha$  for  $\alpha \in (1/2, 1)$ , for general classes of smooth, strongly convex target functions,  $\mu$  will linger in the denominator of the bound with optimal rate  $O(1/n)$ . This makes results sensitive to small values of  $\mu$ .

For more specific problem classes, such as least squares regression and logistic regression, Bach and Moulines [2013] have shown how this can be avoided by combining constant step sizes and averaging.

While such a method recovers for the case of least squares regression for any finite sample size  $n \in \mathbb{N}$  the optimal dependence on the variability in the data, the dependence on the starting point  $\theta_0$  is good, but suboptimal. Furthermore, the choice of the step size depends on a quantity that is typically unknown, unless specific assumptions regarding the distribution of the explanatory variables (e.g. boundedness with known radius) are made.

Regarding the dependence on the starting point, open questions still linger. Firstly, without leveraging any strong convexity properties, Dieuleveut et al. [2017] have shown the optimal rate of  $O(1/n^2)$  for generally convex problems. However, this comes at the cost of too restrictive assumptions regarding the knowledge about the distribution of the explanatory variables. Jain et al. [2018b] have shown that by leveraging strong convexity, one can retain the linear convergence known from gradient descent by performing tail-averaging. However, this comes at the cost of reintroducing the factor  $1/\mu$  in the bound. As far as the author of this thesis is aware, no variants of SGD exist to date, for which the initial condition can be forgotten at a rate of  $O(1/n^2)$  without requiring too much knowledge about the distribution of the explanatory variables, or a

## 5. *Discussion*

linear rate where  $\mu$  enters in a way similar to what would be optimal for optimization procedures based on deterministic first-order oracles.

# References

- A. Agarwal, M. J. Wainwright, P. Bartlett, and P. Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, volume 22, pages 1–9. Curran Associates, Inc., 2009. URL <https://proceedings.neurips.cc/paper/2009/file/2387337ba1e0b0249ba90f55b2ba2521-Paper.pdf>.
- F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(19):595–627, 2014. URL <http://jmlr.org/papers/v15/bach14a.html>.
- F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, pages 451–459. Curran Associates, Inc., 2011. URL <https://proceedings.neurips.cc/paper/2011/file/40008b9a5380fcacce3976bf7c08af5b-Paper.pdf>.
- F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate  $O(1/n)$ . In *Neural Information Processing Systems (NIPS)*, 2013. URL <https://hal.archives-ouvertes.fr/hal-00831977>.
- J. R. Blum. Approximation Methods which Converge with Probability one. *The Annals of Mathematical Statistics*, 25(2):382–386, 1954a. doi:10.1214/aoms/1177728794.
- J. R. Blum. Multidimensional Stochastic Approximation Methods. *The Annals of Mathematical Statistics*, 25(4):737–744, 1954b. doi:10.1214/aoms/1177728659.
- L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pages 177–187. Springer, 2010. URL <http://leon.bottou.org/papers/bottou-2010>.
- L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, volume 20, pages 161–168. Curran Associates, Inc., 2008. URL <https://proceedings.neurips.cc/paper/2007/file/0d3180d672e08b4c5312dcdafdf6ef36-Paper.pdf>.
- L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi:10.1137/16M1080173.

## References

- S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015. ISSN 1935-8237. doi:10.1561/22000000050.
- A. Cauchy. Méthode générale pour la résolution des systèmes d’équations simultanées. *Comptes rendus hebdomadaires des séances de l’Académie des sciences*, 25:536–538, 1847. URL <https://cs.uwaterloo.ca/~y328yu/classics/cauchy-en.pdf>. Translated by Richard J. Pulskamp.
- J. Chee and P. Toulis. Convergence diagnostics for stochastic gradient descent with constant learning rate. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1476–1485. PMLR, 2018. URL <http://proceedings.mlr.press/v84/chee18a/chee18a.pdf>.
- K. L. Chung. On a Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 25(3):463–483, 1954. doi:10.1214/aoms/1177728716.
- A. Defossez and F. Bach. Averaged Least-Mean-Squares: Bias-Variance Trade-offs and Optimal Sampling Distributions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 205–213. PMLR, 2015. URL <http://proceedings.mlr.press/v38/defossez15.pdf>.
- A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016. URL <https://projecteuclid.org/euclid.aos/1467894702>.
- A. Dieuleveut, N. Flammarion, and F. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18(101):1–51, 2017. URL <http://jmlr.org/papers/v18/16-335.html>.
- A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. *The Annals of Statistics*, 48(3):1348–1382, 2020. doi:10.1214/19-AOS1850. Accessed through <https://arxiv.org/pdf/1707.06386> on the 30<sup>th</sup> of November 2020.
- R. Frostig, R. Ge, S. M. Kakade, and A. Sidford. Competing with the empirical risk minimizer in a single pass. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 728–763. PMLR, 2015. URL <http://proceedings.mlr.press/v40/Frostig15.pdf>.
- G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 2013.
- L. Györfi and H. Walk. On the averaged stochastic approximation for linear regression. *SIAM Journal on Control and Optimization*, 34(1):31–61, 1996. doi:10.1137/S0363012992226661.

- P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating stochastic gradient descent for least squares regression. In *COLT*, 2018a. URL <http://proceedings.mlr.press/v75/jain18a/jain18a.pdf>.
- P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18(223):1–42, 2018b. URL <http://jmlr.org/papers/v18/16-595.html>.
- J. Kiefer and J. Wolfowitz. Stochastic Estimation of the Maximum of a Regression Function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952. doi:10.1214/aoms/1177729392.
- H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*. Springer, 2003.
- A. Nemirovski. Efficient methods in convex programming, 1994. URL <https://www2.isye.gatech.edu/~nemirovs/>. Lecture Notes.
- A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. doi:10.1137/070704277.
- A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. John Wiley & Sons, 1983.
- Y. Nesterov. *Introductory Lectures on Convex Optimization - A basic course*. Springer, 2004.
- G. Neu and L. Rosasco. Iterate averaging as regularization for stochastic gradient descent. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 3222–3242. PMLR, 2018.
- J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2006.
- G. C. Pflug. Stochastic minimization with constant step-size: Asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986. doi:10.1137/0324039.
- G. C. Pflug. Non-asymptotic confidence bounds for stochastic approximation algorithms with constant step size. *Monatshefte für Mathematik*, 110:297–314, 1990. doi:10.1007/BF01301683.
- B. Polyak. New method of stochastic approximation type. *Automation and Remote Control*, 51: 937–946, 1990.
- B. Polyak and A. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30:838–855, 1992. doi:10.1137/0330046.

## References

- H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi:10.1214/aoms/1177729586.
- N. Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems*, volume 25, pages 2663–2671. Curran Associates, Inc., 2012. URL <https://proceedings.neurips.cc/paper/2012/file/905056c1ac1dad141560467e0a99e1cf-Paper.pdf>.
- D. Ruppert. Efficient estimators from a slowly convergent robbins-monro procedure. *School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, Technical Report*, 781, 1988. URL <https://ecommons.cornell.edu/bitstream/handle/1813/8664/TR000781.pdf?sequence=1>.
- M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, 2017. doi:10.1007/s10107-016-1030-6.
- R. Schwabe and H. Walk. On a stochastic approximation procedure based on averaging. *Metrika*, 44:165–180, 1996. doi:10.1007/BF02614063.
- J. C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. John Wiley & Sons, 2003.
- P. Toulis and E. M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017. doi:10.1214/16-AOS1506.
- A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

## List of Algorithms

|    |                                                                       |    |
|----|-----------------------------------------------------------------------|----|
| 1. | Stochastic gradient descent for empirical risk minimization . . . . . | 17 |
| 2. | Stochastic gradient descent for online learning . . . . .             | 18 |

## List of Figures

|      |                                                              |    |
|------|--------------------------------------------------------------|----|
| 3.1. | Schematic representation of the proof of Theorem 1 . . . . . | 47 |
|------|--------------------------------------------------------------|----|



# Appendices



## A. An additional result

### A.1. Asymptotic equalities in Bottou [2010]

**Claim 1.** *If the asymptotic ( $n \rightarrow \infty$ ) equalities  $(\log(n)/n)^\alpha \sim \delta_n \sim \varepsilon_n$  hold for sequences  $(\varepsilon_n)_{n \geq 1}$  and  $(\delta_n)_{n \geq 1}$  for some  $\alpha \in [1/2, 1]$ , then*

$$n \log \left( \frac{1}{\delta_n} \right) \sim \frac{1}{\alpha \varepsilon_n^{1/\alpha}} \log^2 \left( \frac{1}{\varepsilon_n} \right) \quad (\text{A.1})$$

*holds.*

*Proof.* Recall that for two sequences  $(a_n)_{n \geq 1}$  and  $(b_n)_{n \geq 1}$ , we write  $a_n \sim b_n$  if and only if  $\lim_{n \rightarrow \infty} a_n/b_n = r \in (0, \infty)$ . For positive sequences  $(a_n)_{n \geq 1}$ ,  $(b_n)_{n \geq 1}$ ,  $(c_n)_{n \geq 1}$  and  $(d_n)_{n \geq 1}$ , note the following properties of asymptotic equality

- (i) If  $a_n \sim b_n$ , then  $\forall p \in \mathbb{R} : a_n^p \sim b_n^p$ .
- (ii) If  $a_n \sim b_n$  and  $c_n \sim d_n$ , then  $a_n c_n \sim b_n d_n$ .
- (iii) If  $b_n \rightarrow b \in \{0, \infty\}$  for  $n \rightarrow \infty$  (and thus  $\log(b_n) \neq 0$  for sufficiently large  $n$ ), it holds that  $\log(a_n) \sim \log(b_n)$ .

The first two results follow immediately from the definition and basic properties of the limit. Note that the first property in particular holds for  $p = -1$ . For the third implication, note that

$$\frac{\log(a_n)}{\log(b_n)} = \frac{\log \left( b_n \cdot \frac{a_n}{b_n} \right)}{\log(b_n)} = 1 + \frac{\log \left( \frac{a_n}{b_n} \right)}{\log(b_n)}$$

holds, where all quantities are well-defined for sufficiently large  $n$ , as then  $b_n > 0$  and  $\log(b_n) \neq 0$ . Noting that  $\log(a_n/b_n)$  converges to a constant  $\log(r)$  and  $\log(b_n)$  converges to  $\pm\infty$ , we have

$$\lim_{n \rightarrow \infty} \frac{\log(a_n)}{\log(b_n)} = 1,$$

i.e. the quotient of the logarithms converges to a constant in  $(0, \infty)$ , thus showing the third property.

The claim now follows from these three properties. For simplicity, the subscripts of  $\varepsilon_n$  and  $\delta_n$  are omitted and dependence on  $n$  is implied. Note that  $(\log(n)/n)^\alpha \sim \delta \sim \varepsilon$  implies that all

A. An additional result

three quantities converge to 0 for  $n \rightarrow \infty$ , so that their inverses go to  $\infty$ . Property (i) tells us for  $p = -1/\alpha$  that

$$\left(\frac{\log(n)}{n}\right)^\alpha \sim \varepsilon \implies \frac{n}{\log(n)} \sim \frac{1}{\varepsilon^{1/\alpha}}. \quad (\text{A.2})$$

Applying property (iii) to the right hand side of (A.2) then gives us

$$\frac{1}{\alpha} \log\left(\frac{1}{\varepsilon}\right) \sim \log(n) - \log(\log(n)). \quad (\text{A.3})$$

The combination of property (i) for  $p = -1$  and property (iii) further yields that

$$\varepsilon \sim \delta \implies \log\left(\frac{1}{\varepsilon}\right) \sim \log\left(\frac{1}{\delta}\right). \quad (\text{A.4})$$

Combining (A.2), (A.3) and (A.4) by applying property (ii) gives us for the right hand side of (A.1) that

$$\begin{aligned} \frac{1}{\alpha \varepsilon^{1/\alpha}} \log^2\left(\frac{1}{\varepsilon}\right) &= \frac{1}{\varepsilon^{1/\alpha}} \cdot \frac{1}{\alpha} \log\left(\frac{1}{\varepsilon}\right) \cdot \log\left(\frac{1}{\varepsilon}\right) \\ &\sim \frac{n}{\log(n)} \cdot (\log(n) - \log(\log(n))) \cdot \log\left(\frac{1}{\delta}\right) \\ &= n \log\left(\frac{1}{\delta}\right) \cdot \left(1 - \frac{\log(\log(n))}{\log(n)}\right) \\ &\sim n \log\left(\frac{1}{\delta}\right). \end{aligned}$$

The last step immediately follows from the definition of asymptotic equality and the fact that  $\log(\log(n))/\log(n) \rightarrow 0$  for  $n \rightarrow \infty$ . This proves the claim. □

## B. The proof of Theorem 1 by Bach and Moulines [2013]

The given theorem and proof were taken from Bach and Moulines [2013] and are in no way original work by the author of this thesis. In this appendix, their work will be explained, commented and possibly extended.

Prior to the actual proof of the theorem, three lemmata are given by Bach and Moulines [2013] which will be needed later.

In order to simplify the notation a little bit, we will write  $\|\cdot\|$  instead of  $\|\cdot\|_2$ , i.e. the norm without subscript is the Euclidean norm.

### B.1. Lemmata

**Lemma 1.** *Let  $(\mathcal{F}_n)_{n \in \mathbb{N} \cup \{0\}}$  be a sequence of increasing  $\sigma$ -fields. Let  $(x_n, \zeta_n) \in \mathbb{R}^d \times \mathbb{R}^d$  be  $\mathcal{F}_n$ -measurable for all  $n \in \mathbb{N}$ . Furthermore, let  $\mathbb{E}[\zeta_n | \mathcal{F}_{n-1}] = 0$ ,  $\mathbb{E}[\|\zeta_n\|^2] < \infty$  and  $\mathbb{E}[\|x_n\|^2 x_n x_n^T | \mathcal{F}_{n-1}] \preceq R^2 H$  a.s. with  $H = \mathbb{E}[x_n x_n^T | \mathcal{F}_{n-1}]$  invertible a.s. and some  $R > 0$ . Consider the sequence of random vectors  $(\alpha_n)_{n \in \mathbb{N}}$  with  $\alpha_n = (I - \gamma x_n x_n^T) \alpha_{n-1} + \gamma \zeta_n$  for  $n \geq 1$  with some non-random  $\gamma > 0$  such that  $\gamma R^2 \leq 1$  and some arbitrary, non-random  $\alpha_0$ . Define for all  $n \geq 0$ :*

$$\bar{\alpha}_n = \frac{1}{n+1} \sum_{i=0}^n \alpha_i.$$

*It then holds that*

$$(1 - \gamma R^2) \mathbb{E}[\langle \bar{\alpha}_{n-1}, H \bar{\alpha}_{n-1} \rangle] + \frac{1}{2n\gamma} \mathbb{E}[\|\alpha_n\|^2] \leq \frac{1}{2n\gamma} \|\alpha_0\|^2 + \frac{\gamma}{n} \sum_{k=1}^n \mathbb{E}[\|\zeta_k\|^2]. \quad (\text{B.1})$$

*Remark.* Note that (B.1) can be reformulated to the following inequality if  $\gamma R^2 \neq 1$  which will be more convenient later on:

$$\mathbb{E}[\langle \bar{\alpha}_{n-1}, H \bar{\alpha}_{n-1} \rangle] \leq \frac{1}{1 - \gamma R^2} \frac{\gamma}{n} \left( \frac{1}{2\gamma^2} \|\alpha_0\|^2 - \frac{1}{2\gamma^2} \mathbb{E}[\|\alpha_n\|^2] + \sum_{k=1}^n \mathbb{E}[\|\zeta_k\|^2] \right). \quad (\text{B.2})$$

*Proof.* The proof technique is rather simple. We will first derive an upper bound for  $\|\alpha_n\|^2$  and then calculate its expected value by first taking the conditional expected value with respect to

*B. The proof of Theorem 1 by Bach and Moulines [2013]*

$\mathcal{F}_{n-1}$ . Lastly we will use this to derive the bound on the expected value of  $\langle \bar{\alpha}_{n-1}, H\bar{\alpha}_{n-1} \rangle$  by using the convexity of  $x \mapsto \langle x, Hx \rangle$ .

In this spirit we first make use of the recursive definition of  $\alpha_n$ . Let  $n \in \mathbb{N}$  be arbitrary. It holds that

$$\begin{aligned}
\|\alpha_n\|^2 &= \left\| (I - \gamma x_n x_n^T) \alpha_{n-1} + \gamma \zeta_n \right\|^2 \\
&= \left\| \alpha_{n-1} - \gamma x_n x_n^T \alpha_{n-1} + \gamma \zeta_n \right\|^2 \\
&= \left\| \alpha_{n-1} + \gamma (\zeta_n - x_n x_n^T \alpha_{n-1}) \right\|^2 \\
&= \|\alpha_{n-1}\|^2 + 2\gamma \langle \alpha_{n-1}, \zeta_n - x_n x_n^T \alpha_{n-1} \rangle + \underbrace{\gamma^2 \left\| \zeta_n - x_n x_n^T \alpha_{n-1} \right\|^2}_{(*)}. \tag{B.3}
\end{aligned}$$

Next, we will use an extension of the well known statement

$$\forall a_0, b_0 \in \mathbb{R} : \quad (a_0 - b_0)^2 \leq 2a_0^2 + 2b_0^2$$

to higher dimensions in order to find an upper bound for the term  $(*)$ . More specifically, we will use that for any  $k \in \mathbb{N}$ :

$$\forall a, b \in \mathbb{R}^k : \|a - b\|^2 \leq 2\|a\|^2 + 2\|b\|^2.$$

This is true, since for all  $a, b \in \mathbb{R}^k$  it holds that

$$\begin{aligned}
\|a - b\|^2 &= \|a\|^2 - 2a^T b + \|b\|^2 \\
\iff \|a - b\|^2 + \|a + b\|^2 &= 2\|a\|^2 + 2\|b\|^2 \\
\implies \|a - b\|^2 &\leq 2\|a\|^2 + 2\|b\|^2.
\end{aligned}$$

Applying this to  $(*)$  for  $k = d$ ,  $a = \zeta_n$  and  $b = x_n x_n^T \alpha_{n-1}$  yields together with (B.3) that

$$\|\alpha_n\|^2 \leq \|\alpha_{n-1}\|^2 + 2\gamma \langle \alpha_{n-1}, \zeta_n - x_n x_n^T \alpha_{n-1} \rangle + 2\gamma^2 \|\zeta_n\|^2 + 2\gamma^2 \left\| x_n x_n^T \alpha_{n-1} \right\|^2.$$

Note that  $\alpha_n$  is  $\mathcal{F}_n$ -measurable for all  $n \in \mathbb{N}$ . This can be shown by noting that for any  $n \in \mathbb{N}$ ,  $\alpha_n$  is  $\mathcal{F}_n$ -measurable if it holds that  $x_n$  and  $\zeta_n$  are  $\mathcal{F}_n$ -measurable and that  $\alpha_{n-1}$  is  $\mathcal{F}_{n-1}$ -measurable. Since the measurability of  $x_n$  and  $\zeta_n$  is given by the corresponding assumption, it is sufficient to show that  $\alpha_{n-1}$  is  $\mathcal{F}_{n-1}$ -measurable. Since  $\alpha_0$  is non-random and therefore measurable with respect to any choice of  $\mathcal{F}_0$ , the measurability of  $\alpha_n$  with respect to  $\mathcal{F}_n$  for any  $n \in \mathbb{N}$  follows by induction.

Therefore, by taking the conditional expected value we get

$$\begin{aligned} \mathbb{E} \left[ \|\alpha_n\|^2 \mid \mathcal{F}_{n-1} \right] &\leq \|\alpha_{n-1}\|^2 + 2\gamma \underbrace{\mathbb{E} \left[ \langle \alpha_{n-1}, \zeta_n - x_n x_n^T \alpha_{n-1} \rangle \mid \mathcal{F}_{n-1} \right]}_A \\ &\quad + 2\gamma^2 \mathbb{E} \left[ \|\zeta_n\|^2 \mid \mathcal{F}_{n-1} \right] + 2\gamma^2 \underbrace{\mathbb{E} \left[ \|x_n x_n^T \alpha_{n-1}\|^2 \mid \mathcal{F}_{n-1} \right]}_B. \end{aligned} \quad (\text{B.4})$$

We can reformulate  $A$  as follows:

$$\begin{aligned} A &= \mathbb{E} \left[ \alpha_{n-1}^T \zeta_n \mid \mathcal{F}_{n-1} \right] - \mathbb{E} \left[ \alpha_{n-1}^T x_n x_n^T \alpha_{n-1} \mid \mathcal{F}_{n-1} \right] \\ &= \alpha_{n-1}^T \underbrace{\mathbb{E} [\zeta_n \mid \mathcal{F}_{n-1}]}_{=0} - \alpha_{n-1}^T \underbrace{\mathbb{E} [x_n x_n^T \mid \mathcal{F}_{n-1}]}_{=H} \alpha_{n-1} \\ &= -\langle \alpha_{n-1}, H \alpha_{n-1} \rangle. \end{aligned} \quad (\text{B.5})$$

As for  $B$ , note that

$$\begin{aligned} B &= \mathbb{E} \left[ \langle x_n x_n^T \alpha_{n-1}, x_n x_n^T \alpha_{n-1} \rangle \mid \mathcal{F}_{n-1} \right] \\ &= \mathbb{E} \left[ \alpha_{n-1}^T x_n x_n^T x_n x_n^T \alpha_{n-1} \mid \mathcal{F}_{n-1} \right] \\ &= \alpha_{n-1}^T \underbrace{\mathbb{E} \left[ \|x_n\|^2 x_n x_n^T \mid \mathcal{F}_{n-1} \right]}_{\preccurlyeq R^2 H} \alpha_{n-1} \\ &\leq R^2 \langle \alpha_{n-1}, H \alpha_{n-1} \rangle. \end{aligned} \quad (\text{B.6})$$

Plugging the results for  $A$  and  $B$  from (B.5) and (B.6) back into (B.4) gives us the upper bound

$$\begin{aligned} \mathbb{E} \left[ \|\alpha_n\|^2 \mid \mathcal{F}_{n-1} \right] &\leq \|\alpha_{n-1}\|^2 - 2\gamma \langle \alpha_{n-1}, H \alpha_{n-1} \rangle + 2\gamma^2 \mathbb{E} \left[ \|\zeta_n\|^2 \mid \mathcal{F}_{n-1} \right] + 2\gamma^2 R^2 \langle \alpha_{n-1}, H \alpha_{n-1} \rangle \\ &= \|\alpha_{n-1}\|^2 - 2\gamma(1 - \gamma R^2) \langle \alpha_{n-1}, H \alpha_{n-1} \rangle + 2\gamma^2 \mathbb{E} \left[ \|\zeta_n\|^2 \mid \mathcal{F}_{n-1} \right]. \end{aligned}$$

By taking the expected value on both sides, we get

$$\mathbb{E} \left[ \|\alpha_n\|^2 \right] \leq \mathbb{E} \left[ \|\alpha_{n-1}\|^2 \right] - 2\gamma(1 - \gamma R^2) \mathbb{E} \left[ \langle \alpha_{n-1}, H \alpha_{n-1} \rangle \right] + 2\gamma^2 \mathbb{E} \left[ \|\zeta_n\|^2 \right]$$

or equivalently

$$(1 - \gamma R^2) \mathbb{E} \left[ \langle \alpha_{n-1}, H \alpha_{n-1} \rangle \right] + \frac{1}{2\gamma} \mathbb{E} \left[ \|\alpha_n\|^2 \right] \leq \frac{1}{2\gamma} \mathbb{E} \left[ \|\alpha_{n-1}\|^2 \right] + \gamma \mathbb{E} \left[ \|\zeta_n\|^2 \right]. \quad (\text{B.7})$$

Note that (B.7) was derived for all  $n \in \mathbb{N}$ . By taking the sum over all  $k \in \{1, \dots, n\}$  on both

*B. The proof of Theorem 1 by Bach and Moulines [2013]*

sides and dividing by  $n$  we get

$$(1 - \gamma R^2) \frac{1}{n} \sum_{k=1}^n \mathbb{E} [\langle \alpha_{k-1}, H \alpha_{k-1} \rangle] + \frac{1}{2n\gamma} \sum_{k=1}^n \mathbb{E} [\|\alpha_k\|^2] \leq \frac{1}{2n\gamma} \sum_{k=1}^n \mathbb{E} [\|\alpha_{k-1}\|^2] + \frac{\gamma}{n} \sum_{k=1}^n \mathbb{E} [\|\zeta_k\|^2]$$

which is by a simple index shift in the sums equivalent to the result

$$\begin{aligned} (1 - \gamma R^2) \underbrace{\frac{1}{n} \sum_{k=0}^{n-1} \mathbb{E} [\langle \alpha_k, H \alpha_k \rangle]}_{(**)} + \frac{1}{2n\gamma} \sum_{k=0}^{n-1} \mathbb{E} [\|\alpha_{k+1}\|^2] &\leq \\ &\leq \frac{1}{2n\gamma} \sum_{k=0}^{n-1} \mathbb{E} [\|\alpha_k\|^2] + \frac{\gamma}{n} \sum_{k=0}^{n-1} \mathbb{E} [\|\zeta_{k+1}\|^2]. \end{aligned} \quad (\text{B.8})$$

Note that since  $H$  is positive definite,  $x \mapsto \langle x, Hx \rangle$  is convex. Therefore the following inequality holds:

$$\frac{1}{n} \sum_{k=0}^{n-1} \langle \alpha_k, H \alpha_k \rangle \geq \left\langle \frac{1}{n} \sum_{k=0}^{n-1} \alpha_k, H \frac{1}{n} \sum_{k=0}^{n-1} \alpha_k \right\rangle = \langle \bar{\alpha}_{n-1}, H \bar{\alpha}_{n-1} \rangle.$$

The equality follows directly from the definition of  $\bar{\alpha}_{n-1}$ . Using this and the fact that  $1 - \gamma R^2 \geq 0$  in (B.8) gives us the inequality

$$\begin{aligned} (1 - \gamma R^2) \mathbb{E} [\langle \bar{\alpha}_{n-1}, H \bar{\alpha}_{n-1} \rangle] + \frac{1}{2n\gamma} \sum_{k=0}^{n-1} \mathbb{E} [\|\alpha_{k+1}\|^2] &\leq \\ &\leq \frac{1}{2n\gamma} \sum_{k=0}^{n-1} \mathbb{E} [\|\alpha_k\|^2] + \frac{\gamma}{n} \sum_{k=0}^{n-1} \mathbb{E} [\|\zeta_{k+1}\|^2]. \end{aligned} \quad (\text{B.9})$$

Lastly, using the simple result that

$$\frac{1}{2n\gamma} \sum_{k=0}^{n-1} \mathbb{E} [\|\alpha_{k+1}\|^2] - \frac{1}{2n\gamma} \sum_{k=0}^{n-1} \mathbb{E} [\|\alpha_k\|^2] = \frac{1}{2n\gamma} \mathbb{E} [\|\alpha_n\|^2] - \frac{1}{2n\gamma} \underbrace{\mathbb{E} [\|\alpha_0\|^2]}_{=\|\alpha_0\|^2}$$

completes the proof as (B.9) is then simplified to

$$(1 - \gamma R^2) \mathbb{E} [\langle \bar{\alpha}_{n-1}, H \bar{\alpha}_{n-1} \rangle] + \frac{1}{2n\gamma} \mathbb{E} [\|\alpha_n\|^2] \leq \frac{1}{2n\gamma} \|\alpha_0\|^2 + \frac{\gamma}{n} \sum_{k=0}^{n-1} \mathbb{E} [\|\zeta_{k+1}\|^2].$$

□

**Lemma 2.** *Let  $(\mathcal{F}_n)_{n \in \mathbb{N} \cup \{0\}}$  be a sequence of increasing  $\sigma$ -fields. Let  $\zeta_n$ ,  $n \in \mathbb{N}$ , be a sequence of random vectors, such that  $\zeta_n$  is  $\mathcal{F}_n$ -measurable for all  $n \in \mathbb{N}$ . Let furthermore for all  $n \in \mathbb{N}$ ,*

$\mathbb{E}[\zeta_n | \mathcal{F}_{n-1}] = 0$  a.s. and  $\mathbb{E}[\zeta_n \zeta_n^T] \preceq C$ . We are interested in the recursion  $\alpha_n = (I - \gamma H)\alpha_{n-1} + \gamma \zeta_n$  for  $n \in \mathbb{N}$  with  $\gamma > 0$ , arbitrary, non-random  $\alpha_0$  and positive definite, orthogonally diagonalizable  $H$ . Lastly, we assume that  $\gamma H \preceq I$ .

Then it holds for all  $n \in \mathbb{N}$  that

$$\mathbb{E}[\alpha_n \alpha_n^T] \preceq (I - \gamma H)^n \alpha_0 \alpha_0^T (I - \gamma H)^n + \gamma^2 \sum_{k=1}^n (I - \gamma H)^{n-k} C (I - \gamma H)^{n-k} \quad (\text{B.10})$$

and

$$\mathbb{E}[\langle \bar{\alpha}_{n-1}, H \bar{\alpha}_{n-1} \rangle] \leq \frac{1}{n\gamma} \|\alpha_0\|^2 + \frac{\text{tr}(CH^{-1})}{n}, \quad (\text{B.11})$$

where for all  $n \in \mathbb{N}$

$$\bar{\alpha}_{n-1} := \frac{1}{n} \sum_{i=0}^{n-1} \alpha_i.$$

*Remark.* In their paper, Bach and Moulines [2013] (whose proof this is) wrongly claim that the result in (B.10) is an equality.

*Proof.* First, note that by iteratively plugging in the recursion equation we get

$$\begin{aligned} \alpha_n &= (I - \gamma H)\alpha_{n-1} + \gamma \zeta_n \\ &= (I - \gamma H)^2 \alpha_{n-2} + (I - \gamma H)\gamma \zeta_{n-1} + \gamma \zeta_n \\ &= (I - \gamma H)^3 \alpha_{n-3} + (I - \gamma H)^2 \gamma \zeta_{n-2} + (I - \gamma H)\gamma \zeta_{n-1} + \gamma \zeta_n \\ &= \dots \\ &= (I - \gamma H)^n \alpha_0 + \gamma \sum_{j=0}^{n-1} (I - \gamma H)^j \zeta_{n-j}. \end{aligned}$$

Changing the index of the sum then yields for all  $n \in \mathbb{N}$ :

$$\alpha_n = (I - \gamma H)^n \alpha_0 + \gamma \sum_{j=1}^n (I - \gamma H)^{n-j} \zeta_j. \quad (\text{B.12})$$

*B. The proof of Theorem 1 by Bach and Moulines [2013]*

We can now use this to verify the first result as given in (B.10):

$$\begin{aligned}
\mathbb{E} \left[ \alpha_n \alpha_n^T \right] &= (I - \gamma H)^n \alpha_0 \alpha_0^T (I - \gamma H)^n + (I - \gamma H)^n \alpha_0 \gamma \sum_{j=1}^n \mathbb{E} \left[ \zeta_j^T \right] (I - \gamma H)^{n-j} \\
&\quad + \gamma \sum_{j=1}^n (I - \gamma H)^{n-j} \mathbb{E} \left[ \zeta_j \right] \alpha_0^T (I - \gamma H)^n \\
&\quad + \gamma^2 \sum_{j=1}^n \sum_{k=1}^n (I - \gamma H)^{n-j} \mathbb{E} \left[ \zeta_j \zeta_k^T \right] (I - \gamma H)^{n-k}. \tag{B.13}
\end{aligned}$$

Note that  $\mathbb{E} [\zeta_j] = \mathbb{E} [\mathbb{E} [\zeta_j | \mathcal{F}_{j-1}]] = \mathbb{E} [0] = 0$  for all  $j \in \mathbb{N}$ .

Consider the expression  $\mathbb{E}[\zeta_j \zeta_k^T]$  in the case  $k \neq j$ . Let without loss of generality  $k < j$  (and therefore  $k \leq j - 1$ ) hold. It then follows that

$$\begin{aligned}
\mathbb{E} \left[ \zeta_j \zeta_k^T \right] &= \mathbb{E} \left[ \mathbb{E} \left[ \zeta_j \zeta_k^T | \mathcal{F}_{j-1} \right] \right] \\
&= \mathbb{E} \left[ \mathbb{E} \left[ \zeta_j | \mathcal{F}_{j-1} \right] \zeta_k^T \right] \\
&= \mathbb{E} \left[ 0 \cdot \zeta_k^T \right] \\
&= 0.
\end{aligned}$$

Consequently

$$\mathbb{E} \left[ \zeta_j \zeta_k^T \right] = \begin{cases} 0 & \text{if } k \neq j \\ \mathbb{E} \left[ \zeta_k \zeta_k^T \right] & \text{if } k = j \end{cases} \tag{B.14}$$

for all  $k, j \in \mathbb{N}$ . By using these results on  $\mathbb{E} [\zeta_j]$  and  $\mathbb{E} [\zeta_j \zeta_k^T]$ , equation (B.13) is simplified to

$$\mathbb{E} \left[ \alpha_n \alpha_n^T \right] = (I - \gamma H)^n \alpha_0 \alpha_0^T (I - \gamma H)^n + \gamma^2 \sum_{k=1}^n (I - \gamma H)^{n-k} \mathbb{E} \left[ \zeta_k \zeta_k^T \right] (I - \gamma H)^{n-k}.$$

Finally, we get the first result as given in (B.10) by using the assumption that  $\mathbb{E} \left[ \zeta_k \zeta_k^T \right] \preceq C$  for all  $k \in \mathbb{N}$ .

Now on to the second result which is given in (B.11).

Since  $\bar{\alpha}_{n-1} = \frac{1}{n} \sum_{k=0}^{n-1} \alpha_k$ , we can use (B.12) to get

$$\begin{aligned}
\bar{\alpha}_{n-1} &= \frac{1}{n} \sum_{k=0}^{n-1} (I - \gamma H)^k \alpha_0 + \frac{\gamma}{n} \sum_{k=1}^{n-1} \sum_{j=1}^k (I - \gamma H)^{k-j} \zeta_j \\
&= \frac{1}{n} (I - (I - \gamma H)^n) (I - (I - \gamma H))^{-1} \alpha_0 + \frac{\gamma}{n} \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} (I - \gamma H)^{k-j} \zeta_j \\
&= \frac{1}{n} (I - (I - \gamma H)^n) (\gamma H)^{-1} \alpha_0 + \frac{\gamma}{n} \sum_{j=1}^{n-1} \left( \sum_{k=j}^{n-1} (I - \gamma H)^{k-j} \right) \zeta_j \\
&= \frac{1}{n} (I - (I - \gamma H)^n) (\gamma H)^{-1} \alpha_0 + \frac{\gamma}{n} \sum_{j=1}^{n-1} \left( \sum_{k=0}^{n-1-j} (I - \gamma H)^k \right) \zeta_j \\
&= \frac{1}{n} (I - (I - \gamma H)^n) (\gamma H)^{-1} \alpha_0 + \frac{\gamma}{n} \sum_{j=1}^{n-1} (I - (I - \gamma H)^{n-j}) (\gamma H)^{-1} \zeta_j. \tag{B.15}
\end{aligned}$$

By plugging (B.15) into  $\langle \bar{\alpha}_{n-1}, H \bar{\alpha}_{n-1} \rangle = \bar{\alpha}_{n-1}^T H \bar{\alpha}_{n-1}$ , we get

$$\begin{aligned}
\langle \bar{\alpha}_{n-1}, H \bar{\alpha}_{n-1} \rangle &= \\
&= \underbrace{\frac{1}{n^2} \alpha_0^T (\gamma H)^{-1} (I - (I - \gamma H)^n) H (I - (I - \gamma H)^n) (\gamma H)^{-1} \alpha_0}_{A_1} \\
&\quad + \underbrace{2 \frac{\gamma}{n^2} \sum_{j=1}^{n-1} \zeta_j^T (\gamma H)^{-1} (I - (I - \gamma H)^{n-j}) H (I - (I - \gamma H)^n) (\gamma H)^{-1} \alpha_0}_{A_2} \\
&\quad + \underbrace{\frac{\gamma^2}{n^2} \sum_{j=1}^{n-1} \sum_{k=1}^{n-1} \zeta_j^T (\gamma H)^{-1} (I - (I - \gamma H)^{n-j}) H (I - (I - \gamma H)^{n-k}) (\gamma H)^{-1} \zeta_k}_{A_3}. \tag{B.16}
\end{aligned}$$

Note that we are actually interested in  $\mathbb{E}[\langle \bar{\alpha}_{n-1}, H \bar{\alpha}_{n-1} \rangle]$  and will therefore, in the hope of better readability, first examine the expected values of  $A_1$ ,  $A_2$  and  $A_3$  in (B.16).

Obviously,  $\mathbb{E}[A_1] = A_1$ . However, we can simplify  $A_1$ , since  $H$  and  $I - (I - \gamma H)^n$  commute, as can easily be seen by observing that  $H$  and  $I - \gamma H$  commute. Therefore

$$\begin{aligned}
\mathbb{E}[A_1] &= A_1 = \frac{1}{n^2 \gamma^2} \alpha_0^T (I - (I - \gamma H)^n)^2 H^{-1} \alpha_0 \\
&= \frac{1}{n \gamma} \alpha_0^T (I - (I - \gamma H)^n)^2 (n \gamma H)^{-1} \alpha_0.
\end{aligned}$$

We will now apply Lemma 3, which is stated and proved below, to  $(I - (I - \gamma H)^n)^2$ . Since Lemma 3 is stated for  $u \in [0, 1]$  and not for matrices like  $\gamma H$ , we will have to justify this step. Let  $U \Lambda U^T = \gamma H$  be the diagonalization of  $\gamma H$  where  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$  is the diagonal

B. The proof of Theorem 1 by Bach and Moulines [2013]

matrix of the eigenvalues  $\lambda_1, \dots, \lambda_d$  of  $\gamma H$  and  $U$  is an orthogonal matrix of according dimension. Existence of such matrices  $\Lambda$  and  $U$  is guaranteed by the assumption that  $H$  can be orthogonally diagonalized.

Note that for all  $i = 1, \dots, d$  it holds that  $\lambda_i \geq 0$ , since  $\gamma H$  is positive semi-definite, and  $\lambda_i \leq 1$ , since  $\gamma H \preceq I$ .

We can now rewrite  $(I - (I - \gamma H)^n)^2$  by following these steps:

$$\begin{aligned}
& \gamma H = U \Lambda U^T \\
\iff & I - \gamma H = U(I - \Lambda)U^T \\
\iff & (I - \gamma H)^n = U(I - \Lambda)^n U^T \\
\iff & I - (I - \gamma H)^n = U(I - (I - \Lambda)^n)U^T \\
\iff & (I - (I - \gamma H)^n)^2 = U(I - (I - \Lambda)^n)^2 U^T.
\end{aligned}$$

The matrix  $(I - (I - \Lambda)^n)^2$  is diagonal with the  $i$ -th entry being  $(1 - (1 - \lambda_i)^n)^2$ . We can now apply Lemma 3 and get  $(1 - (1 - \lambda_i)^n)^2 \leq n\lambda_i$ . Consequently,  $(I - (I - \Lambda)^n)^2 \preceq n\Lambda$  which leads to

$$\begin{aligned}
(I - (I - \gamma H)^n)^2 &= U(I - (I - \Lambda)^n)^2 U^T \\
&\preceq nU\Lambda U^T \\
&= n\gamma H.
\end{aligned} \tag{B.17}$$

By now using (B.17), we get

$$\begin{aligned}
A_1 &\leq \frac{1}{n\gamma} \alpha_0^T (n\gamma H) (n\gamma H)^{-1} \alpha_0 \\
&= \frac{1}{n\gamma} \|\alpha_0\|^2.
\end{aligned}$$

Now we will examine the term  $A_2$ . Since we have already established that under the assumptions it holds that  $\mathbb{E}[\zeta_j] = 0$ , it follows immediately that  $\mathbb{E}[A_2] = 0$ .

As for  $A_3$ , we will use properties of the trace of a matrix:

$$\begin{aligned}
\mathbb{E}[A_3] &= \frac{\gamma^2}{n^2} \sum_{j=1}^{n-1} \sum_{k=1}^{n-1} \mathbb{E} \left[ \zeta_j^T (\gamma H)^{-1} (I - (I - \gamma H)^{n-j}) H (I - (I - \gamma H)^{n-k}) (\gamma H)^{-1} \zeta_k \right] \\
&= \frac{\gamma^2}{n^2} \sum_{j=1}^{n-1} \sum_{k=1}^{n-1} \text{tr} \left( \mathbb{E} \left[ \zeta_k \zeta_j^T \right] (\gamma H)^{-1} (I - (I - \gamma H)^{n-j}) H (I - (I - \gamma H)^{n-k}) (\gamma H)^{-1} \right) \\
&= \frac{1}{n^2} \sum_{j=1}^{n-1} \sum_{k=1}^{n-1} \text{tr} \left( \mathbb{E} \left[ \zeta_k \zeta_j^T \right] H^{-1} (I - (I - \gamma H)^{n-j}) H (I - (I - \gamma H)^{n-k}) H^{-1} \right).
\end{aligned}$$

By recalling (B.14) and using that a matrix  $H^{1/2}$  with  $H^{1/2}H^{1/2} = H$  exists, which in addition commutes with  $I - (I - \gamma H)^i$  for all  $i \in \mathbb{N}$ , we get that

$$\begin{aligned}
\mathbb{E}[A_3] &= \frac{1}{n^2} \sum_{j=1}^{n-1} \sum_{k=1}^{n-1} \text{tr} \left( \mathbb{E} \left[ \zeta_k \zeta_j^T \right] H^{-1} (I - (I - \gamma H)^{n-j}) H (I - (I - \gamma H)^{n-k}) H^{-1} \right) \\
&= \frac{1}{n^2} \sum_{k=1}^{n-1} \text{tr} \left( \mathbb{E} \left[ \zeta_k \zeta_k^T \right] H^{-1} (I - (I - \gamma H)^{n-k}) H (I - (I - \gamma H)^{n-k}) H^{-1} \right) \\
&= \frac{1}{n^2} \sum_{k=1}^{n-1} \text{tr} \left( \mathbb{E} \left[ \zeta_k \zeta_k^T \right] H^{-1} (I - (I - \gamma H)^{n-k}) H^{1/2} H^{1/2} (I - (I - \gamma H)^{n-k}) H^{-1} \right) \\
&= \frac{1}{n^2} \sum_{k=1}^{n-1} \text{tr} \left( \left( \mathbb{E} \left[ \zeta_k \zeta_k^T \right] \right)^{1/2} H^{-1} H^{1/2} (I - (I - \gamma H)^{n-k})^2 H^{1/2} H^{-1} \left( \mathbb{E} \left[ \zeta_k \zeta_k^T \right] \right)^{1/2} \right).
\end{aligned}$$

Since  $0 \preccurlyeq I - \gamma H$  by assumption, it holds that  $(I - (I - \gamma H)^{n-k})^2 \preccurlyeq I$ . Using the previous result, this implies

$$\begin{aligned}
\mathbb{E}[A_3] &\leq \frac{1}{n^2} \sum_{k=1}^{n-1} \text{tr} \left( \left( \mathbb{E} \left[ \zeta_k \zeta_k^T \right] \right)^{1/2} H^{-1/2} \cdot I \cdot H^{-1/2} \left( \mathbb{E} \left[ \zeta_k \zeta_k^T \right] \right)^{1/2} \right) \\
&= \frac{1}{n^2} \sum_{k=1}^{n-1} \text{tr} \left( H^{-1/2} \cdot \mathbb{E} \left[ \zeta_k \zeta_k^T \right] \cdot H^{-1/2} \right).
\end{aligned}$$

Lastly, using the assumption  $\mathbb{E} \left[ \zeta_k \zeta_k^T \right] \preccurlyeq C$ , we get

$$\begin{aligned}
\mathbb{E}[A_3] &\leq \frac{1}{n^2} \sum_{k=1}^{n-1} \text{tr} \left( H^{-1/2} \cdot C \cdot H^{-1/2} \right) \\
&= \frac{n-1}{n^2} \text{tr} \left( C H^{-1} \right) \\
&\leq \frac{1}{n} \text{tr} \left( C H^{-1} \right).
\end{aligned}$$

By now combining the upper bounds we got for  $\mathbb{E}[A_1]$  and  $\mathbb{E}[A_3]$  and by using the fact that  $\mathbb{E}[A_2] = 0$ , we get

$$\mathbb{E} \left[ \langle \bar{\alpha}_{n-1}, H \bar{\alpha}_{n-1} \rangle \right] \leq \frac{1}{n\gamma} \|\alpha_0\|^2 + \frac{1}{n} \text{tr} \left( C H^{-1} \right),$$

which proves the second result of the lemma as given in (B.11).  $\square$

**Lemma 3.** *It holds that*

$$\forall u \in [0, 1] \quad \forall n \in \mathbb{N} : \quad (1 - (1 - u)^n)^2 \leq nu$$

*B. The proof of Theorem 1 by Bach and Moulines [2013]*

*Proof.* First note that since  $u \in [0, 1]$  and  $n \geq 1$ ,

$$\begin{aligned}
& 1 - u \in [0, 1] \\
\iff & (1 - u)^n \in [0, 1] \\
\iff & 1 - (1 - u)^n \in [0, 1] \\
\implies & 1 - (1 - u)^n \leq 1.
\end{aligned} \tag{B.18}$$

Furthermore, it holds that

$$\begin{aligned}
& 1 - u \in [0, 1] \\
\iff & (1 - u)^{n-1} \in [0, 1] \\
\implies & n(1 - u)^{n-1} \leq n.
\end{aligned} \tag{B.19}$$

Integration of both sides of (B.19) from 0 to  $u$  yields  $nu$  on the right-hand side and the following result on the left-hand side:

$$\begin{aligned}
\int_0^u n(1 - x)^{n-1} dx &= [-(1 - x)^n]_0^u \\
&= -(1 - u)^n + 1.
\end{aligned}$$

Consequently, (B.19) implies

$$1 - (1 - u)^n \leq nu. \tag{B.20}$$

Multiplying the inequalities (B.18) and (B.20) gives

$$(1 - (1 - u)^n)^2 \leq nu,$$

since all terms are non-negative. This proves the lemma.  $\square$

## B.2. Assumptions

The proof by Bach and Moulines [2013] considers a slightly more general setting than was presented in Chapter 3. In particular, instead of a scalar response  $y_i \in \mathbb{R}$ ,  $i \geq 1$ , a vector-valued response  $z_i \in \mathbb{R}^d$ ,  $i \geq 1$ , is considered. As a result, instead of the scalar residuals  $\varepsilon_i = y_i - \theta_*^T x_i$ , we have vector-valued residuals  $\xi_i = z_i - (\theta_*^T x_i)x_i$  for all  $i \geq 1$ . Note that this is indeed a generalization, as all results that will now be presented can be translated into the notation of Chapter 3 through setting  $z_i = y_i x_i$  and  $\xi_i = \varepsilon_i x_i$  for all  $i \geq 1$ . If one wishes, one can therefore also view  $z_i$  and  $\xi_i$  as merely more concise notation for the frequently appearing products  $y_i x_i$

and  $\varepsilon_i x_i$ .

Restating the assumptions made in Chapter 3 with this notation gives the following assumptions, which are essentially identical to the ones made by Bach and Moulines [2013]. Let  $d \in \mathbb{N}$ .

- (A1')  $(z_i, x_i) \in \mathbb{R}^d \times \mathbb{R}^d$ ,  $i \in \mathbb{N}$ , are independent and identically distributed random variables.
- (A2') The random variables have finite second moments, i.e.  $\mathbb{E}[\|x_1\|^2] < \infty$  and  $\mathbb{E}[\|z_1\|^2] < \infty$ . We write  $H = \mathbb{E}[x_1 x_1^T]$  for the second moment matrix of the explanatory variables, which is assumed to be invertible.
- (A3') The function  $f(\theta) = \frac{1}{2} \mathbb{E}[\langle \theta, x_1 \rangle^2 - 2\langle \theta, z_1 \rangle]$  is minimal at  $\theta_* \in \mathbb{R}^d$ . Let  $\xi_i = z_i - \langle \theta_*, x_i \rangle x_i$  for  $i \in \mathbb{N}$  denote the residuals.
- (A4') There exist  $R > 0$  and  $\sigma > 0$  such that  $\mathbb{E}[\xi_1 \xi_1^T] \preceq \sigma^2 H$  and  $\mathbb{E}[\|x_1\|^2 x_1 x_1^T] \preceq R^2 H = R^2 \mathbb{E}[x_1 x_1^T]$ .
- (A5') The step size sequence is chosen constant with  $\gamma_n = \gamma > 0$  for all  $n \in \mathbb{N}$  such that  $\gamma R^2 < 1$ .
- (A6') The stochastic gradient descent (SGD) recursion is defined as

$$\theta_n = (I - \gamma x_n x_n^T) \theta_{n-1} + \gamma z_n$$

for  $n \in \mathbb{N}$  with  $\theta_0 \in \mathbb{R}^d$  arbitrary and non-random. The averaged SGD estimator is

$$\bar{\theta}_n = \frac{1}{n+1} \sum_{k=0}^n \theta_k$$

for  $n \in \mathbb{N} \cup \{0\}$ .

Note that one can easily show that for  $z_i = y_i x_i$ , the function in assumption (A3') only differs by an additive constant from the function in assumption (A3). Note that the residuals  $\xi_i$ ,  $i \geq 1$ , defined in assumption (A3'), are i.i.d. with  $\mathbb{E}[\xi_1] = 0$ .

*Remark.* Note that the quantities  $H$ ,  $R^2$  and  $\gamma$  remain unaffected by the presented reformulation of the assumptions, as they only depend on properties of the explanatory variables  $x_i$ ,  $i \geq 1$ . In Subsection 3.3.1, we have shown that  $R^2$  needs to satisfy  $R^2 \geq \text{tr}(H)$ . We can use this in order to show that  $\gamma R^2 < 1$  implies that  $\gamma H \preceq I$ , as mentioned in Subsection 3.3.1. To that end, we have to show that  $H \preceq \text{tr}(H) I$ . Given this inequality, the proof immediately follows from

$$H \preceq \text{tr}(H) I \preceq R^2 I \preceq \frac{1}{\gamma} I,$$

where  $\gamma R^2 < 1$  was used in the last inequality.

The fact that  $H \preceq \text{tr}(H) I$  holds can be shown as follows. Since  $H$  is symmetric, there exists an orthogonal matrix  $U$  of the same dimension such that  $H = U \Lambda U^T$  with  $\Lambda$  denoting the diagonal

### B. The proof of Theorem 1 by Bach and Moulines [2013]

matrix containing the eigenvalues of  $H$  on its diagonal. Let  $\lambda_1, \dots, \lambda_d$  denote the eigenvalues of  $H$ . We want to show that  $\text{tr}(H)I - H$  is positive semidefinite. Note that

$$\text{tr}(H)I - H = \text{tr}(H)UU^T - U\Lambda U^T = U(\text{tr}(H)I - \Lambda)U^T = U(\text{tr}(\Lambda)I - \Lambda)U^T.$$

The fact that  $H$  is assumed positive definite implies  $\text{tr}(\Lambda) > \lambda_i$  for all  $i = 1, \dots, d$ . Since this means that  $\text{tr}(\Lambda)I - \Lambda$  is a diagonal matrix with positive diagonal elements,  $\text{tr}(H)I - H$  is positive definite.

## B.3. The theorem of interest

For the convenience of the reader, the theorem is restated here.

**Theorem 1.** *Under the assumptions (A1')-(A6')*

$$\mathbb{E}[f(\bar{\theta}_{n-1}) - f(\theta_*)] \leq \frac{1}{2n} \left[ \frac{\sigma\sqrt{d}}{1 - \sqrt{\gamma R^2}} + R\|\theta_0 - \theta_*\| \frac{1}{\sqrt{\gamma R^2}} \right]^2 \quad (\text{B.21})$$

holds for all  $n \in \mathbb{N}$ . For  $\gamma = 1/(4R^2)$  this simplifies to

$$\mathbb{E}[f(\bar{\theta}_{n-1}) - f(\theta_*)] \leq \frac{2}{n} \left[ \sigma\sqrt{d} + R\|\theta_0 - \theta_*\| \right]^2.$$

Over the course of the proof, we will mostly consider the deviation of  $\theta_n$  from  $\theta_*$ , which we will denote by  $\eta_n := \theta_n - \theta_*$ , or its averaged version  $\bar{\eta}_n = \frac{1}{n+1} \sum_{k=0}^n \eta_k = \bar{\theta}_n - \theta_*$ , for  $n \in \mathbb{N} \cup \{0\}$ . Before getting to the actual derivation of the upper bound, we will examine some of the properties of  $\eta_n$  and  $\bar{\eta}_n$  in the following section.

## B.4. Properties of $\eta_n = \theta_n - \theta_*$ and $\bar{\eta}_n = \bar{\theta}_n - \theta_*$

The reason for considering properties of this term is that the expression of interest, namely  $\mathbb{E}[f(\bar{\theta}_{n-1}) - f(\theta_*)]$ , can interestingly be rewritten as the square of a norm of  $\bar{\eta}_{n-1}$ . This will be established in the following claim.

**Claim 2.** *We can rewrite the LHS of (B.21) as follows for all  $n \in \mathbb{N}$ :*

$$\mathbb{E}[f(\bar{\theta}_{n-1}) - f(\theta_*)] = \frac{1}{2} \mathbb{E}[\langle \bar{\eta}_{n-1}, H\bar{\eta}_{n-1} \rangle]. \quad (\text{B.22})$$

*Proof.* To show this, consider a non-random  $\theta \in \mathbb{R}^d$  and  $\theta_*$  as in assumption (A3'). Let  $\eta :=$

B.4. Properties of  $\eta_n = \theta_n - \theta_*$  and  $\bar{\eta}_n = \bar{\theta}_n - \theta_*$

$\theta - \theta_*$ . It then holds that

$$\begin{aligned}
f(\theta) - f(\theta_*) &= \frac{1}{2} \mathbb{E} \left[ \langle \theta, x_n \rangle^2 - 2 \langle \theta, z_n \rangle \right] - \frac{1}{2} \mathbb{E} \left[ \langle \theta_*, x_n \rangle^2 - 2 \langle \theta_*, z_n \rangle \right] \\
&= \frac{1}{2} \mathbb{E} \left[ \langle \theta, x_n \rangle^2 - 2 \langle \theta, z_n \rangle - \langle \theta_*, x_n \rangle^2 + 2 \langle \theta_*, z_n \rangle \right] \\
&= \frac{1}{2} \mathbb{E} \left[ (\langle \theta, x_n \rangle - \langle \theta_*, x_n \rangle)(\langle \theta, x_n \rangle + \langle \theta_*, x_n \rangle) - 2 \langle \theta - \theta_*, z_n \rangle \right] \\
&= \frac{1}{2} \mathbb{E} \left[ \underbrace{\langle \theta - \theta_*, x_n \rangle}_{=\eta} \underbrace{\langle \theta + \theta_*, x_n \rangle}_{=\eta + 2\theta_*} - 2 \underbrace{\langle \theta - \theta_*, z_n \rangle}_{=\eta} \right] \\
&= \frac{1}{2} \mathbb{E} \left[ \eta^T x_n x_n^T (\eta + 2\theta_*) - 2\eta^T z_n \right] \\
&= \frac{1}{2} \left( \eta^T \mathbb{E} \left[ x_n x_n^T \right] (\eta + 2\theta_*) - 2\eta^T \mathbb{E} [z_n] \right). \tag{B.23}
\end{aligned}$$

Note that by assumption **(A2')**  $\mathbb{E} [x_n x_n^T] = H$  for all  $n \in \mathbb{N}$ . As for  $\mathbb{E} [z_n]$ , we can use

$$\begin{aligned}
\mathbb{E} [z_n] &= \mathbb{E} [\xi_n + \langle \theta_*, x_n \rangle x_n] \\
&= \underbrace{\mathbb{E} [\xi_n]}_{=0} + \mathbb{E} [x_n \langle \theta_*, x_n \rangle] \\
&= \mathbb{E} [x_n x_n^T \theta_*] \\
&= H \theta_*.
\end{aligned}$$

Using this in (B.23), we get

$$\begin{aligned}
f(\theta) - f(\theta_*) &= \frac{1}{2} \left( \eta^T H (\eta + 2\theta_*) - 2\eta^T H \theta_* \right) \\
&= \frac{1}{2} \left( \eta^T H \eta + 2\eta^T H \theta_* - 2\eta^T H \theta_* \right) \\
&= \frac{1}{2} \eta^T H \eta. \tag{B.24}
\end{aligned}$$

Since this holds for any  $\theta \in \mathbb{R}^d$  and  $\eta = \theta - \theta_*$ , it also holds almost surely for the random estimator  $\bar{\theta}_{n-1}$  achieved here with averaged stochastic gradient descent and its corresponding error  $\bar{\eta}_{n-1} = \bar{\theta}_{n-1} - \theta_*$ .

Consequently, with (B.24) the equation

$$f(\bar{\theta}_{n-1}) - f(\theta_*) = \frac{1}{2} \bar{\eta}_{n-1}^T H \bar{\eta}_{n-1}$$

holds almost surely, which implies

$$\mathbb{E} [f(\bar{\theta}_{n-1}) - f(\theta_*)] = \frac{1}{2} \mathbb{E} [\bar{\eta}_{n-1}^T H \bar{\eta}_{n-1}].$$

B. The proof of Theorem 1 by Bach and Moulines [2013]

This proves the claim made in (B.22). □

For the remainder of the proof of the theorem, we will work on creating an upper bound on the RHS of the previous claim, as this will naturally give us an upper bound on the LHS.

**Claim 3.** *The recursion on  $\theta_n$  in assumption (A6') implies the following recursion for  $\eta_n = \theta_n - \theta_*$  for all  $n \in \mathbb{N}$ :*

$$\eta_n = (I - \gamma x_n x_n^T) \eta_{n-1} + \gamma \xi_n.$$

*Proof.* It follows directly from the recursion  $\theta_n = (I - \gamma x_n x_n^T) \theta_{n-1} + \gamma z_n$  for  $n \in \mathbb{N}$  with  $\theta_0$  arbitrary that

$$\begin{aligned} \eta_n &= \theta_n - \theta_* \\ &= (I - \gamma x_n x_n^T) \theta_{n-1} - \theta_* + \gamma z_n \\ &= (I - \gamma x_n x_n^T) \theta_{n-1} - \theta_* + \gamma x_n x_n^T \theta_* - \gamma x_n x_n^T \theta_* + \gamma z_n \\ &= (I - \gamma x_n x_n^T) (\theta_{n-1} - \theta_*) + \gamma (z_n - x_n x_n^T \theta_*) \\ &= (I - \gamma x_n x_n^T) \underbrace{(\theta_{n-1} - \theta_*)}_{=\eta_{n-1}} + \gamma \underbrace{(z_n - \langle x_n, \theta_* \rangle x_n)}_{=\xi_n} \\ &= (I - \gamma x_n x_n^T) \eta_{n-1} + \gamma \xi_n, \end{aligned}$$

thus immediately completing the proof of the claim. □

Due to this recursion,  $\eta_n$  can be written as a function of the starting point  $\eta_0$ , the errors  $\xi_k$ ,  $k \in \{1, \dots, n\}$ , and the observations  $x_k$ ,  $k \in \{1, \dots, n\}$ . This will be shown in the next claim.

**Claim 4.** *For  $i, j \in \mathbb{N}$  with  $j \geq i$ , let*

$$M_i^j(x_i, x_{i+1}, \dots, x_{j-1}, x_j) = M_i^j := (I - \gamma x_j x_j^T) \cdot (I - \gamma x_{j-1} x_{j-1}^T) \cdot \dots \cdot (I - \gamma x_i x_i^T)$$

and let

$$M_i^{i-1} = I.$$

We can write

$$\eta_n = M_1^n \eta_0 + \gamma \sum_{k=1}^n M_{k+1}^n \xi_k \tag{B.25}$$

B.4. Properties of  $\eta_n = \theta_n - \theta_*$  and  $\bar{\eta}_n = \bar{\theta}_n - \theta_*$

and consequently also

$$\bar{\eta}_{n-1} = \frac{1}{n} \sum_{k=0}^{n-1} M_1^k \eta_0 + \frac{\gamma}{n} \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} M_{j+1}^k \xi_j \quad (\text{B.26})$$

for all  $n \in \mathbb{N}$ .

*Proof.* The first result follows from simply applying the recursion equation from the previous claim  $\eta_n = (I - \gamma x_n x_n^T) \eta_{n-1} + \gamma \xi_n$  for  $n \in \mathbb{N}$  iteratively such that

$$\begin{aligned} \eta_n &= (I - \gamma x_n x_n^T) \eta_{n-1} + \gamma \xi_n \\ &= (I - \gamma x_n x_n^T) (I - \gamma x_{n-1} x_{n-1}^T) \eta_{n-2} + (I - \gamma x_n x_n^T) \gamma \xi_{n-1} + \gamma \xi_n \\ &= (I - \gamma x_n x_n^T) (I - \gamma x_{n-1} x_{n-1}^T) (I - \gamma x_{n-2} x_{n-2}^T) \eta_{n-3} \\ &\quad + (I - \gamma x_n x_n^T) (I - \gamma x_{n-1} x_{n-1}^T) \gamma \xi_{n-2} + (I - \gamma x_n x_n^T) \gamma \xi_{n-1} + \gamma \xi_n \\ &= \dots \\ &= (I - \gamma x_n x_n^T) \cdot \dots \cdot (I - \gamma x_1 x_1^T) \cdot \eta_0 + \gamma \sum_{k=1}^{n-1} (I - \gamma x_n x_n^T) \cdot \dots \cdot (I - \gamma x_{k+1} x_{k+1}^T) \cdot \xi_k + \gamma I \xi_n \\ &= M_1^n \eta_0 + \gamma \sum_{k=1}^{n-1} M_{k+1}^n \xi_k + \gamma M_{n+1}^n \xi_n \\ &= M_1^n \eta_0 + \gamma \sum_{k=1}^n M_{k+1}^n \xi_k \end{aligned}$$

which proves (B.25).

For the second result as given in equation (B.26), we will use the first result, but rename the summation index as  $j$  instead of  $k$ . The second result follows immediately by using the definition of  $\bar{\eta}_{n-1}$  and a change in the summation indices of the double sum.

B. The proof of Theorem 1 by Bach and Moulines [2013]

More specifically, it holds that

$$\begin{aligned}
\bar{\eta}_{n-1} &= \frac{1}{n} \sum_{k=0}^{n-1} \eta_k \\
&= \frac{1}{n} \eta_0 + \frac{1}{n} \sum_{k=1}^{n-1} \eta_k \\
&= \frac{1}{n} \eta_0 + \frac{1}{n} \sum_{k=1}^{n-1} \left[ M_1^k \eta_0 + \gamma \sum_{j=1}^k M_{j+1}^k \xi_j \right] \\
&= \frac{1}{n} \eta_0 + \frac{1}{n} \sum_{k=1}^{n-1} M_1^k \eta_0 + \frac{\gamma}{n} \sum_{k=1}^{n-1} \sum_{j=1}^k M_{j+1}^k \xi_j \\
&= \frac{1}{n} M_1^0 \eta_0 + \frac{1}{n} \sum_{k=1}^{n-1} M_1^k \eta_0 + \frac{\gamma}{n} \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} M_{j+1}^k \xi_j \\
&= \frac{1}{n} \sum_{k=0}^{n-1} M_1^k \eta_0 + \frac{\gamma}{n} \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} M_{j+1}^k \xi_j.
\end{aligned}$$

This completes the proof of the claim.  $\square$

## B.5. Construction of the upper bound on the expected excess risk

We will now put the results from the claims of the previous section to use. This part of the proof is crucial as it derives the structure of the upper bound given in the theorem. Recall the following definition, which was already introduced in Section 2.1 and used in Section 3.4.

**Definition.** For ease of notation, we define the norm  $\|\cdot\|_H$  by

$$\|X\|_H := \sqrt{\mathbb{E} [\langle X, HX \rangle]}$$

for all square integrable random variables  $X \in \mathbb{R}^d$ .

*Remark.* The fact that this expression truly satisfies the triangle inequality follows directly from the Minkowski inequality and the fact that we could also write  $\|X\|_H = \sqrt{\mathbb{E} [\langle H^{1/2} X, H^{1/2} X \rangle]}$  since  $H$  is symmetric and positive semi-definite.

By making use of the triangle inequality and the claims shown in the previous section (more

### B.5. Construction of the upper bound on the expected excess risk

specifically (B.22) and (B.26)), we can write

$$\begin{aligned} \sqrt{2\mathbb{E} [f(\bar{\theta}_{n-1}) - f(\theta_*)]} &= \|\bar{\eta}_{n-1}\|_H \\ &= \left\| \frac{1}{n} \sum_{k=0}^{n-1} M_1^k \eta_0 + \frac{\gamma}{n} \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} M_{j+1}^k \xi_j \right\|_H \\ &\leq \left\| \frac{1}{n} \sum_{k=0}^{n-1} M_1^k \eta_0 \right\|_H + \left\| \frac{\gamma}{n} \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} M_{j+1}^k \xi_j \right\|_H. \end{aligned}$$

This will directly lead to the structure of the final bound in (B.21) by reformulation to

$$\mathbb{E} [f(\bar{\theta}_{n-1}) - f(\theta_*)] \leq \frac{1}{2} \left( \left\| \frac{1}{n} \sum_{k=0}^{n-1} M_1^k \eta_0 \right\|_H + \left\| \frac{\gamma}{n} \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} M_{j+1}^k \xi_j \right\|_H \right)^2. \quad (\text{B.27})$$

Motivated by (B.27), our goal will now be to find upper bounds on

$$\left\| \frac{1}{n} \sum_{k=0}^{n-1} M_1^k \eta_0 \right\|_H$$

and

$$\left\| \frac{\gamma}{n} \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} M_{j+1}^k \xi_j \right\|_H.$$

The former term can be seen as the influence of the starting point  $\theta_0$ , while the errors  $\xi_k$ ,  $k \in \{1, \dots, n-1\}$ , do not have an influence. The latter term can be described as the noise process that results from having the starting point  $\theta_0 = \theta_*$ . Thus, the former term will be upper bounded by the bias term in the bound by Bach and Moulines [2013], while the latter will be upper bounded by their variance term.

In the following, we will assume **(A1')**-(**A6'**). Furthermore,  $\mathcal{F}_n$  will denote the  $\sigma$ -field generated by  $\{(x_1, z_1), \dots, (x_n, z_n)\}$  for all  $n \in \mathbb{N}$ . The  $\sigma$ -field  $\mathcal{F}_0$  will be arbitrary with the sole condition that  $\mathcal{F}_0 \subseteq \mathcal{F}_1$ . This directly implies  $\mathcal{F}_n$ -measurability of  $\xi_n = z_n - \langle \theta_*, x_n \rangle x_n$  for all  $n \in \mathbb{N}$ .

#### B.5.1. Construction of the bias term

**Claim 5.** *With  $\eta_n = \theta_n - \theta_*$  for  $n \in \mathbb{N} \cup \{0\}$  like defined above, it holds that*

$$\left\| \frac{1}{n} \sum_{k=0}^{n-1} M_1^k \eta_0 \right\|_H \leq \frac{\|\theta_0 - \theta_*\|}{\sqrt{n\gamma}}. \quad (\text{B.28})$$

*Proof.* Consider the sequence of random variables  $(\nu_n)_{n \in \mathbb{N} \cup \{0\}}$  which satisfies  $\nu_n = (I - \gamma x_n x_n^T) \nu_{n-1}$

B. The proof of Theorem 1 by Bach and Moulines [2013]

for  $n \in \mathbb{N}$  with  $\nu_0 = \eta_0$ . Note that the only difference between the recursion of  $\eta_n$  and  $\nu_n$  is that in the latter the term " $+\gamma\xi_n$ " is missing. Since omitting this term in the recursion equation of  $\eta_n$  would be equal to assuming errors equal to zero, i.e.  $\xi_n = 0$  for all  $n \in \mathbb{N}$ , the second result from Claim 4 (see (B.26)) gives us for  $\bar{\nu}_{n-1} = \frac{1}{n} \sum_{k=0}^{n-1} \nu_k$ :

$$\bar{\nu}_{n-1} = \frac{1}{n} \sum_{k=0}^{n-1} M_1^k \nu_0 = \frac{1}{n} \sum_{k=0}^{n-1} M_1^k \eta_0$$

and therefore

$$\left\| \frac{1}{n} \sum_{k=0}^{n-1} M_1^k \eta_0 \right\|_H = \|\bar{\nu}_{n-1}\|_H.$$

We will therefore prove the claim by deriving an upper bound for  $\|\bar{\nu}_{n-1}\|_H$ .

Note first that we can write for  $\|\nu_n\|^2$  with arbitrary  $n \in \mathbb{N}$ :

$$\begin{aligned} \|\nu_n\|^2 &= \langle (I - \gamma x_n x_n^T) \nu_{n-1}, (I - \gamma x_n x_n^T) \nu_{n-1} \rangle \\ &= \langle \nu_{n-1} - \gamma x_n x_n^T \nu_{n-1}, \nu_{n-1} - \gamma x_n x_n^T \nu_{n-1} \rangle \\ &= \|\nu_{n-1}\|^2 - 2\gamma \langle \nu_{n-1}, x_n x_n^T \nu_{n-1} \rangle + \gamma^2 \langle x_n x_n^T \nu_{n-1}, x_n x_n^T \nu_{n-1} \rangle \\ &= \|\nu_{n-1}\|^2 - 2\gamma \langle \nu_{n-1}, x_n x_n^T \nu_{n-1} \rangle + \gamma^2 \nu_{n-1}^T x_n \underbrace{x_n^T x_n}_{=\|x_n\|^2} x_n^T \nu_{n-1} \\ &= \|\nu_{n-1}\|^2 - 2\gamma \langle \nu_{n-1}, x_n x_n^T \nu_{n-1} \rangle + \gamma^2 \langle \nu_{n-1}, \|x_n\|^2 x_n x_n^T \nu_{n-1} \rangle. \end{aligned}$$

By taking the expected value conditional on  $\mathcal{F}_{n-1}$  and using that for any  $n \in \mathbb{N}$  it holds that  $x_n$  and  $\mathcal{F}_{n-1}$  are independent, this yields

$$\begin{aligned} \mathbb{E} [\|\nu_n\|^2 | \mathcal{F}_{n-1}] &= \|\nu_{n-1}\|^2 - 2\gamma \underbrace{\langle \nu_{n-1}, \mathbb{E} [x_n x_n^T | \mathcal{F}_{n-1}] \nu_{n-1} \rangle}_{=\mathbb{E}[x_n x_n^T]=H} + \gamma^2 \underbrace{\langle \nu_{n-1}, \mathbb{E} [\|x_n\|^2 x_n x_n^T | \mathcal{F}_{n-1}] \nu_{n-1} \rangle}_{=\mathbb{E}[\|x_n\|^2 x_n x_n^T] \preceq R^2 H} \\ &\leq \|\nu_{n-1}\|^2 - 2\gamma \langle \nu_{n-1}, H \nu_{n-1} \rangle + \gamma^2 R^2 \langle \nu_{n-1}, H \nu_{n-1} \rangle \\ &= \|\nu_{n-1}\|^2 + \gamma \underbrace{(\gamma R^2 - 2)}_{\leq -1, \text{ since } \gamma R^2 \leq 1} \langle \nu_{n-1}, H \nu_{n-1} \rangle \\ &\leq \|\nu_{n-1}\|^2 - \gamma \langle \nu_{n-1}, H \nu_{n-1} \rangle. \end{aligned}$$

By now taking the unconditional expected value, we get

$$\mathbb{E} [\|\nu_n\|^2] \leq \mathbb{E} [\|\nu_{n-1}\|^2] - \gamma \mathbb{E} [\langle \nu_{n-1}, H \nu_{n-1} \rangle]$$

### B.5. Construction of the upper bound on the expected excess risk

or equivalently

$$\mathbb{E} \left[ \|\nu_{n-1}\|^2 \right] - \mathbb{E} \left[ \|\nu_n\|^2 \right] \geq \gamma \mathbb{E} \left[ \langle \nu_{n-1}, H \nu_{n-1} \rangle \right].$$

Since this holds for any  $n \in \mathbb{N}$ , summing up on both sides leads to

$$\frac{1}{n} \sum_{k=1}^n \left( \mathbb{E} \left[ \|\nu_{k-1}\|^2 \right] - \mathbb{E} \left[ \|\nu_k\|^2 \right] \right) \geq \frac{\gamma}{n} \sum_{k=1}^n \mathbb{E} \left[ \langle \nu_{k-1}, H \nu_{k-1} \rangle \right].$$

Using the linearity of the expected value on the RHS and the fact that the LHS is a telescopic sum, gives us

$$\frac{1}{n} \left( \mathbb{E} \left[ \|\nu_0\|^2 \right] - \mathbb{E} \left[ \|\nu_n\|^2 \right] \right) \geq \gamma \mathbb{E} \left[ \frac{1}{n} \sum_{k=0}^{n-1} \langle \nu_k, H \nu_k \rangle \right]. \quad (\text{B.29})$$

Note that since  $x \mapsto \langle x, Hx \rangle$  is convex, it follows for any  $a_1, \dots, a_n \in \mathbb{R}^d$  with  $n \in \mathbb{N}$ :

$$\frac{1}{n} \sum_{k=1}^n \langle a_k, H a_k \rangle \geq \left\langle \left( \frac{1}{n} \sum_{k=1}^n a_k \right), H \left( \frac{1}{n} \sum_{k=1}^n a_k \right) \right\rangle.$$

Together with the fact that  $\nu_0$  is assumed non-random and that we have  $\bar{\nu}_{n-1} = \frac{1}{n} \sum_{k=0}^{n-1} \nu_k$ , (B.29) implies

$$\frac{1}{n} \left( \|\nu_0\|^2 - \mathbb{E} \left[ \|\nu_n\|^2 \right] \right) \geq \gamma \mathbb{E} \left[ \langle \bar{\nu}_{n-1}, H \bar{\nu}_{n-1} \rangle \right] = \gamma \|\bar{\nu}_{n-1}\|_H^2.$$

We thus arrive at the following upper bound for  $\|\bar{\nu}_{n-1}\|_H^2$ :

$$\|\bar{\nu}_{n-1}\|_H^2 \leq \frac{\|\nu_0\|^2}{n\gamma} - \frac{\mathbb{E} \left[ \|\nu_n\|^2 \right]}{n\gamma} \leq \frac{\|\nu_0\|^2}{n\gamma}.$$

Using that  $\|\bar{\nu}_{n-1}\|_H^2 = \left\| \frac{1}{n} \sum_{k=0}^{n-1} M_1^k \eta_0 \right\|_H^2$  and that  $\nu_0 = \eta_0 = \theta_0 - \theta_*$  gives us

$$\left\| \frac{1}{n} \sum_{k=0}^{n-1} M_1^k \eta_0 \right\|_H^2 \leq \frac{\|\theta_0 - \theta_*\|^2}{n\gamma}.$$

Since taking the square root on both sides gives us that

$$\left\| \frac{1}{n} \sum_{k=0}^{n-1} M_1^k \eta_0 \right\|_H \leq \frac{\|\theta_0 - \theta_*\|}{\sqrt{n\gamma}},$$

this completes the proof.  $\square$

### B.5.2. Construction of the variance term

The derivation of this bound is far more complicated than the one of the last bound. It will therefore be structured as several claims, which are then combined.

In the end, we will have shown that

$$\left\| \frac{\gamma}{n} \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} M_{j+1}^k \xi_j \right\|_H \leq \frac{\sigma \sqrt{d}}{\sqrt{n}} \frac{1}{1 - \sqrt{\gamma R^2}}. \quad (\text{B.30})$$

**Proof Principle.** The proof is complementary to the one presented in the previous subsection. Let us once again define a sequence of random variables  $(\nu_n)_{n \in \mathbb{N} \cup \{0\}}$  through the recursion equation  $\nu_n = (I - \gamma x_n x_n^T) \nu_{n-1} + \gamma \xi_n$ , which now starts with  $\nu_0 = 0$ . Here, the difference between the sequences  $(\eta_n)_{n \in \mathbb{N} \cup \{0\}}$  and  $(\nu_n)_{n \in \mathbb{N} \cup \{0\}}$  lies in the starting point  $\eta_0$  or  $\nu_0$  respectively. Therefore, the second result from Claim 4 (see (B.26)) gives for  $\bar{\nu}_{n-1} = \frac{1}{n} \sum_{k=0}^{n-1} \nu_k$ :

$$\bar{\nu}_{n-1} = \frac{\gamma}{n} \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} M_{j+1}^k \xi_j.$$

We can thus find an upper bound on

$$\left\| \frac{\gamma}{n} \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} M_{j+1}^k \xi_j \right\|_H$$

by deriving an upper bound on  $\|\bar{\nu}_{n-1}\|_H$ .

To that end, we will define further sequences:

$$\begin{aligned} \nu_n^0 &= (I - \gamma H) \nu_{n-1}^0 + \gamma \xi_n & \text{for } n \in \mathbb{N} \text{ with } \nu_0^0 = 0, \\ \text{for } r \in \mathbb{N} : \nu_n^r &= (I - \gamma H) \nu_{n-1}^r + \gamma (H - x_n x_n^T) \nu_{n-1}^{r-1} & \text{for } n \in \mathbb{N} \text{ with } \nu_0^r = 0. \end{aligned}$$

In accordance with the notation so far, we define for  $r \in \mathbb{N} \cup \{0\}$ :  $\bar{\nu}_{n-1}^r = \frac{1}{n} \sum_{k=0}^{n-1} \nu_k^r$ .

The proof of the upper bound in (B.30) is then based on the following inequality, for which we

### B.5. Construction of the upper bound on the expected excess risk

need to apply the triangle inequality twice:

$$\begin{aligned}\|\bar{\nu}_{n-1}\|_H &= \left\| \bar{\nu}_{n-1} - \sum_{i=0}^r \bar{\nu}_{n-1}^i + \sum_{i=0}^r \bar{\nu}_{n-1}^i \right\|_H \\ &\leq \left\| \bar{\nu}_{n-1} - \sum_{i=0}^r \bar{\nu}_{n-1}^i \right\|_H + \left\| \sum_{i=0}^r \bar{\nu}_{n-1}^i \right\|_H \\ &\leq \left\| \bar{\nu}_{n-1} - \sum_{i=0}^r \bar{\nu}_{n-1}^i \right\|_H + \sum_{i=0}^r \left\| \bar{\nu}_{n-1}^i \right\|_H.\end{aligned}$$

We will then derive an upper bound

$$A_r \geq \left\| \bar{\nu}_{n-1} - \sum_{i=0}^r \bar{\nu}_{n-1}^i \right\|_H$$

and an upper bound

$$B_r \geq \sum_{i=0}^r \left\| \bar{\nu}_{n-1}^i \right\|_H$$

and we will show that

$$\lim_{r \rightarrow \infty} (A_r + B_r) = \frac{\sigma\sqrt{d}}{\sqrt{n}} \frac{1}{1 - \sqrt{\gamma R^2}},$$

which will complete the proof of the claim, as  $\|\bar{\nu}_{n-1}\|_H$  does not depend on  $r$ .

#### Construction of the upper bound $A_r$

**Claim 6.** For  $\nu_n, \nu_n^r$  with  $r, n \in \mathbb{N} \cup \{0\}$ , defined in the previous paragraph, the following recursion equation holds for all  $n \in \mathbb{N}$  and  $r \in \mathbb{N} \cup \{0\}$ :

$$\nu_n - \sum_{i=0}^r \nu_n^i = (I - \gamma x_n x_n^T) \left( \nu_{n-1} - \sum_{i=0}^r \nu_{n-1}^i \right) + \gamma (H - x_n x_n^T) \nu_{n-1}^r.$$

*Proof.* For  $n = 0$  the result is trivial for all  $r \in \mathbb{N} \cup \{0\}$ . We will show by induction over  $r$  that for all  $r \in \mathbb{N} \cup \{0\}$  the claim holds for all  $n \in \mathbb{N}$ .

B. The proof of Theorem 1 by Bach and Moulines [2013]

$r = 0$ : By using the recursion equations for the sequences we get

$$\begin{aligned}
\nu_n - \nu_n^0 &= (I - \gamma x_n x_n^T) \nu_{n-1} + \gamma \xi_n - (I - \gamma H) \nu_{n-1}^0 - \gamma \xi_n \\
&= (I - \gamma x_n x_n^T) \nu_{n-1} - (I - \gamma H) \nu_{n-1}^0 \\
&= (I - \gamma x_n x_n^T) \nu_{n-1} - (I - \gamma x_n x_n^T + \gamma x_n x_n^T - \gamma H) \nu_{n-1}^0 \\
&= (I - \gamma x_n x_n^T) \nu_{n-1} - (I - \gamma x_n x_n^T) \nu_{n-1}^0 - \gamma (x_n x_n^T - H) \nu_{n-1}^0 \\
&= (I - \gamma x_n x_n^T) (\nu_{n-1} - \nu_{n-1}^0) + \gamma (H - x_n x_n^T) \nu_{n-1}^0
\end{aligned}$$

which shows that the claim is true for all  $n \in \mathbb{N}$  for the induction start  $r = 0$ .

$r \mapsto r + 1$ : This is again rather straight forward by using the induction assumption and the recursion equation for  $\nu_n^{r+1}$ :

$$\begin{aligned}
\nu_n - \sum_{i=0}^{r+1} \nu_n^i &= \nu_n - \sum_{i=0}^r \nu_n^i - \nu_n^{r+1} \\
&= (I - \gamma x_n x_n^T) \left( \nu_{n-1} - \sum_{i=0}^r \nu_{n-1}^i \right) + \gamma (H - x_n x_n^T) \nu_{n-1}^r \\
&\quad - (I - \gamma H) \nu_{n-1}^{r+1} - \gamma (H - x_n x_n^T) \nu_{n-1}^r \\
&= (I - \gamma x_n x_n^T) \left( \nu_{n-1} - \sum_{i=0}^r \nu_{n-1}^i \right) - (I - \gamma x_n x_n^T + \gamma x_n x_n^T - \gamma H) \nu_{n-1}^{r+1} \\
&= (I - \gamma x_n x_n^T) \left( \nu_{n-1} - \sum_{i=0}^r \nu_{n-1}^i \right) - (I - \gamma x_n x_n^T) \nu_{n-1}^{r+1} - \gamma (x_n x_n^T - H) \nu_{n-1}^{r+1} \\
&= (I - \gamma x_n x_n^T) \left( \nu_{n-1} - \sum_{i=0}^r \nu_{n-1}^i - \nu_{n-1}^{r+1} \right) + \gamma (H - x_n x_n^T) \nu_{n-1}^{r+1} \\
&= (I - \gamma x_n x_n^T) \left( \nu_{n-1} - \sum_{i=0}^{r+1} \nu_{n-1}^i \right) + \gamma (H - x_n x_n^T) \nu_{n-1}^{r+1}
\end{aligned}$$

thus completing the proof of the claim, since the arguments above hold for all  $n \in \mathbb{N}$ .  $\square$

Before we get to the upper bound for  $\|\bar{\nu}_{n-1} - \sum_{i=0}^r \bar{\nu}_{n-1}^i\|_H$ , we need to derive an upper bound on  $\mathbb{E} \left[ \nu_n^r (\nu_n^r)^T \right]$ , the importance of which is not obvious just yet. This motivates the next claim.

**Claim 7.** For all  $r \in \mathbb{N} \cup \{0\}$  and  $n \in \mathbb{N} \cup \{0\}$  it holds that

$$\mathbb{E} \left[ \nu_n^r (\nu_n^r)^T \right] \preceq \gamma^{r+1} R^{2r} \sigma^2 I.$$

*Proof.* The proof is trivial for  $n = 0$ . Once again we will use a proof by induction over  $r \in \mathbb{N} \cup \{0\}$ .

### B.5. Construction of the upper bound on the expected excess risk

$r = 0$ : Here, we will use the recursion equation of  $\nu_n^0$ , the assumption  $\mathbb{E} [\xi_n \xi_n^T] \preceq \sigma^2 H$  for all  $n \in \mathbb{N}$  and the fact that  $\mathbb{E} [\nu_{n-1}^0 \xi_n^T] = \mathbb{E} [\nu_{n-1}^0 \mathbb{E} [\xi_n^T | \mathcal{F}_{n-1}]] = \mathbb{E} [\nu_{n-1}^0 \cdot 0] = 0$ . These arguments give us for any  $n \in \mathbb{N}$ :

$$\begin{aligned} \mathbb{E} [\nu_n^0 (\nu_n^0)^T] &= \mathbb{E} [((I - \gamma H) \nu_{n-1}^0 + \gamma \xi_n)((I - \gamma H) \nu_{n-1}^0 + \gamma \xi_n)^T] \\ &= (I - \gamma H) \mathbb{E} [\nu_{n-1}^0 (\nu_{n-1}^0)^T] (I - \gamma H) + \gamma \mathbb{E} [\xi_n (\nu_{n-1}^0)^T] (I - \gamma H) \\ &\quad + \gamma (I - \gamma H) \mathbb{E} [\nu_{n-1}^0 \xi_n^T] + \gamma^2 \mathbb{E} [\xi_n \xi_n^T] \\ &\preceq (I - \gamma H) \mathbb{E} [\nu_{n-1}^0 (\nu_{n-1}^0)^T] (I - \gamma H) + \gamma^2 \sigma^2 H. \end{aligned}$$

Applying this inequality recursively gives us

$$\begin{aligned} \mathbb{E} [\nu_n^0 (\nu_n^0)^T] &\preceq (I - \gamma H) \mathbb{E} [\nu_{n-1}^0 (\nu_{n-1}^0)^T] (I - \gamma H) + \gamma^2 \sigma^2 H \\ &\preceq (I - \gamma H)^2 \mathbb{E} [\nu_{n-2}^0 (\nu_{n-2}^0)^T] (I - \gamma H)^2 + \gamma^2 \sigma^2 (I - \gamma H) H (I - \gamma H) + \gamma^2 \sigma^2 H \\ &\preceq \dots \\ &\preceq (I - \gamma H)^n \underbrace{\mathbb{E} [\nu_0^0 (\nu_0^0)^T]}_{=0} (I - \gamma H)^n + \gamma^2 \sigma^2 \sum_{k=0}^{n-1} (I - \gamma H)^k H (I - \gamma H)^k \\ &= \gamma^2 \sigma^2 \sum_{k=0}^{n-1} (I - \gamma H)^k H (I - \gamma H)^k. \end{aligned}$$

Since  $I - \gamma H$  and  $H$  commute, it is obvious that  $(I - \gamma H)^k$  and  $H$  commute for all  $k \in \mathbb{N} \cup \{0\}$ . Therefore, we can simplify the previous inequality to get

$$\mathbb{E} [\nu_n^0 (\nu_n^0)^T] \preceq \gamma^2 \sigma^2 \sum_{k=0}^{n-1} (I - \gamma H)^{2k} H. \quad (\text{B.31})$$

The RHS can be further simplified by noting that it includes a geometric sum for which it is known that

$$\begin{aligned} \sum_{k=0}^{n-1} (I - \gamma H)^{2k} &= \sum_{k=0}^{n-1} [(I - \gamma H)^2]^k \\ &= \left( I - (I - \gamma H)^{2n} \right) \cdot \left( I - (I - \gamma H)^2 \right)^{-1}. \end{aligned}$$

Recall that as shown in Section B.2, choosing  $\gamma > 0$  such that  $\gamma R^2 \leq 1$  implies  $\gamma H \preceq I$ . Note that  $\gamma H$  is symmetric and therefore orthogonally diagonalizable. Furthermore, since  $0 \prec \gamma H \preceq I$ , its eigenvalues lie in the range  $(0, 1]$ . Let  $\lambda_i$ ,  $i \in \{1, \dots, d\}$ , denote the eigenvalues of  $\gamma H$  and let  $\gamma H = U \Lambda U^T$  with  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$  and  $U$  orthogonal. The fact that it holds for all

B. The proof of Theorem 1 by Bach and Moulines [2013]

$i \in \{1, \dots, d\}$  that  $\lambda_i \in (0, 1]$  gives us that  $1 - \lambda_i \in [0, 1)$  and thus

$$1 - \lambda_i \geq (1 - \lambda_i)^2 \implies 1 - (1 - \lambda_i)^2 \geq \lambda_i \implies (1 - (1 - \lambda_i)^2)^{-1} \leq \lambda_i^{-1}.$$

The previous equality can therefore be restated as

$$\begin{aligned} \sum_{k=0}^{n-1} (I - \gamma H)^{2k} &= U \left( I - (I - \Lambda)^{2n} \right) \cdot \left( I - (I - \Lambda)^2 \right)^{-1} U^T \\ &= U \text{diag} \left( \underbrace{(1 - (1 - \lambda_1)^{2n})}_{\leq 1} \underbrace{(1 - (1 - \lambda_1)^2)^{-1}}_{\leq \lambda_1^{-1}}, \dots, \underbrace{(1 - (1 - \lambda_d)^{2n})}_{\leq 1} \underbrace{(1 - (1 - \lambda_d)^2)^{-1}}_{\leq \lambda_d^{-1}} \right) U^T \\ &\preceq U \text{diag} \left( \lambda_1^{-1}, \dots, \lambda_d^{-1} \right) U^T \\ &= U \Lambda^{-1} U^T \\ &= (\gamma H)^{-1}. \end{aligned}$$

We therefore have the following upper bound on the geometric sum:

$$\sum_{k=0}^{n-1} (I - \gamma H)^{2k} \preceq \frac{1}{\gamma} H^{-1}, \quad (\text{B.32})$$

which we can use in (B.31) to get

$$\begin{aligned} \mathbb{E} \left[ \nu_n^0 (\nu_n^0)^T \right] &\preceq \gamma^2 \sigma^2 \sum_{k=0}^{n-1} (I - \gamma H)^{2k} H \\ &= \gamma^2 \sigma^2 H^{1/2} \sum_{k=0}^{n-1} (I - \gamma H)^{2k} H^{1/2} \\ &\preceq \gamma^2 \sigma^2 H^{1/2} \frac{1}{\gamma} H^{-1} H^{1/2} \\ &= \gamma \sigma^2 I, \end{aligned}$$

thus completing the induction start. Note that the reformulation in the second line is indeed necessary before we can use (B.32), as arbitrary quadratic matrices  $M_1, M_2, M_3 \succcurlyeq 0$  of equal dimension with  $M_1 \preceq M_2$  in general do not need to satisfy  $M_1 M_3 \preceq M_2 M_3$ . However, for such matrices the inequality  $M_3 M_1 M_3 \preceq M_3 M_2 M_3$  does hold.

$r \mapsto r + 1$ : For the induction step we will need the first result of Lemma 2 as given in (B.10).

We want to show that  $\mathbb{E} \left[ \nu_n^{r+1} (\nu_n^{r+1})^T \right] \preceq \gamma^{r+2} R^{2(r+1)} \sigma^2 I$  for all  $r \in \mathbb{N} \cup \{0\}$ ,  $n \in \mathbb{N}$ .

We will apply Lemma 2 for  $\alpha_n := \nu_n^{r+1}$  and  $\zeta_n := (H - x_n x_n^T) \nu_{n-1}^r$  for all  $n \in \mathbb{N}$  and arbitrary  $r \in \mathbb{N} \cup \{0\}$ . We define  $\alpha_0 = \nu_0^{r+1} = 0$ . We will first check the assumptions made in Lemma 2.

- Recursion equation: The chosen definition for  $\alpha_n$  and  $\zeta_n$  satisfies  $\alpha_n = (I - \gamma H) \alpha_{n-1} + \gamma \zeta_n$

### B.5. Construction of the upper bound on the expected excess risk

for all  $r \in \mathbb{N} \cup \{0\}$  and all  $n \in \mathbb{N}$ . The matrix  $H$  is by assumption invertible and thus positive definite.  $H$  is furthermore orthogonally diagonalizable, due to being symmetric.

- Measurability: Note that we have that  $\zeta_n = (H - x_n x_n^T) \nu_{n-1}^r$  is  $\mathcal{F}_n$ -measurable if  $\nu_{n-1}^r$  is at least  $\mathcal{F}_n$ -measurable, since  $x_n$  is  $\mathcal{F}_n$ -measurable by construction of  $\mathcal{F}_n$ . Note that  $\nu_{n-1}^r = (I - \gamma H) \nu_{n-2}^r + \gamma (H - x_{n-1} x_{n-1}^T) \nu_{n-2}^{r-1}$  is  $\mathcal{F}_{n-1}$ -measurable if  $\nu_{n-2}^r$  and  $\nu_{n-2}^{r-1}$  are  $\mathcal{F}_{n-2}$ -measurable. This leads to a recursive argument that leads to the conclusion that  $\nu_{n-1}^r$  is indeed  $\mathcal{F}_{n-1}$ -measurable, since for all  $r \in \mathbb{N} \cup \{0\}$  the starting point  $\nu_0^r$  is non-random and thus measurable with respect to all  $\sigma$ -fields and at the same time  $\nu_n^0 = (I - \gamma H) \nu_{n-1}^0 + \gamma \xi_n$  is measurable with respect to  $\mathcal{F}_n$  for all  $n \in \mathbb{N}$ .
- Conditionally centered: We have to check that  $\mathbb{E} [\zeta_n | \mathcal{F}_{n-1}] = 0$  for all  $n \in \mathbb{N}$ . Note that we have

$$\mathbb{E} [\zeta_n | \mathcal{F}_{n-1}] = \mathbb{E} \left[ (H - x_n x_n^T) \nu_{n-1}^r | \mathcal{F}_{n-1} \right] = \left( H - \mathbb{E} [x_n x_n^T | \mathcal{F}_{n-1}] \right) \nu_{n-1}^r = (H - H) \nu_{n-1}^r = 0$$

by using the independence assumption in **(A1')** and the assumption **(A2')**.

- Bound on second moments: We need to find a matrix  $C$  such that  $\mathbb{E} [\zeta_n \zeta_n^T] \preceq C$  for all  $n \in \mathbb{N}$ . Using the independence of  $x_n$  and  $\nu_{n-1}^r$ , the fact that  $\mathbb{E} [\|x_n\| x_n x_n^T] \preceq R^2 H$  due to assumption **(A4')**, that  $\mathbb{E} [x_n x_n^T] = H$  as given in assumption **(A2')**, and the induction assumption, we get

$$\begin{aligned} \mathbb{E} [\zeta_n \zeta_n^T] &= \mathbb{E} \left[ (H - x_n x_n^T) \nu_{n-1}^r (\nu_{n-1}^r)^T (H - x_n x_n^T) \right] \\ &= \mathbb{E} \left[ (H - x_n x_n^T) \mathbb{E} [\nu_{n-1}^r (\nu_{n-1}^r)^T] (H - x_n x_n^T) \right] \\ &\preceq \mathbb{E} \left[ (H - x_n x_n^T) \gamma^{r+1} R^{2r} \sigma^2 I (H - x_n x_n^T) \right] \\ &= \gamma^{r+1} R^{2r} \sigma^2 \mathbb{E} \left[ (H - x_n x_n^T)^2 \right] \\ &= \gamma^{r+1} R^{2r} \sigma^2 \mathbb{E} \left[ H^2 - H x_n x_n^T - x_n x_n^T H + x_n x_n^T x_n x_n^T \right] \\ &= \gamma^{r+1} R^{2r} \sigma^2 \left( H^2 - H \mathbb{E} [x_n x_n^T] - \mathbb{E} [x_n x_n^T] H + \mathbb{E} [\|x_n\| x_n x_n^T] \right) \\ &\preceq \gamma^{r+1} R^{2r} \sigma^2 (H^2 - H^2 - H^2 + R^2 H) \\ &= \gamma^{r+1} R^{2r} \sigma^2 (-H^2 + R^2 H) \\ &\preceq \gamma^{r+1} R^{2r+2} \sigma^2 H \end{aligned}$$

and we can therefore choose  $C = \gamma^{r+1} R^{2r+2} \sigma^2 H$ .

- Choice of  $\gamma$ :  $\gamma H \preceq I$  is implied by  $\gamma R^2 \leq 1$  as argued in Section B.2.

B. The proof of Theorem 1 by Bach and Moulines [2013]

Now that we have checked all the assumptions, we can apply Lemma 2, which gives us

$$\mathbb{E}[\alpha_n \alpha_n^T] \preceq (I - \gamma H)^n \alpha_0 \alpha_0^T (I - \gamma H)^n + \gamma^2 \sum_{k=1}^n (I - \gamma H)^{n-k} C (I - \gamma H)^{n-k}.$$

Since  $\alpha_0 = 0$  for all  $r \in \mathbb{N} \cup \{0\}$ , this gives us

$$\begin{aligned} \mathbb{E}[\nu_n^{r+1} (\nu_n^{r+1})^T] &\preceq \gamma^2 \sum_{k=1}^n (I - \gamma H)^{n-k} \gamma^{r+1} R^{2r+2} \sigma^2 H (I - \gamma H)^{n-k} \\ &= \gamma^{r+3} R^{2r+2} \sigma^2 \sum_{k=1}^n \left( (I - \gamma H)^2 \right)^{n-k} H. \end{aligned}$$

Note that we have

$$\begin{aligned} \sum_{k=1}^n \left( (I - \gamma H)^2 \right)^{n-k} &= \sum_{k=0}^{n-1} \left( (I - \gamma H)^2 \right)^k \\ &\preceq \frac{1}{\gamma} H^{-1} \end{aligned}$$

by making use of the same upper bound as in (B.32).

This simplifies the previous result to

$$\mathbb{E} \left[ \nu_n^{r+1} (\nu_n^{r+1})^T \right] \preceq \gamma^{r+2} R^{2r+2} \sigma^2 I,$$

which completes the proof of the claim.  $\square$

*Remark.* Before we go on, it is worth noting that over the course of this proof, we have found an upper bound  $\mathbb{E} \left[ \left( (H - x_n x_n^T) \nu_{n-1}^r \right) \left( (H - x_n x_n^T) \nu_{n-1}^r \right)^T \right] \preceq \gamma^{r+1} R^{2r+2} \sigma^2 H$  for all  $n \in \mathbb{N}$  and  $r \in \mathbb{N} \cup \{0\}$ .

We now have everything we need to find the upper bound  $A_r$  on  $\left\| \bar{\nu}_{n-1} - \sum_{i=0}^r \bar{\nu}_{n-1}^i \right\|_H$ . This will be the subject of the next claim.

**Claim 8.** For all  $r \in \mathbb{N} \cup \{0\}$  and  $n \in \mathbb{N}$  it holds that

$$\left\| \bar{\nu}_{n-1} - \sum_{i=0}^r \bar{\nu}_{n-1}^i \right\|_H^2 \leq \frac{1}{1 - \gamma R^2} \gamma^{r+1} R^{2r+2} \sigma^2 d.$$

*Remark.* Note that this is not the upper bound given by Bach and Moulines [2013]. While they have achieved a better upper bound in this step, both bounds are sufficiently low in order to prove the final result as given in Theorem 1. The discrepancy results from the upper bounds

### B.5. Construction of the upper bound on the expected excess risk

in (B.33) and (B.34). It is unclear how the smaller upper bound in Bach and Moulines [2013, Appendix A.4] is achieved and, in particular, how the explicit factor  $d$  is avoided.

*Proof.* We will first use the recursion equation achieved in Claim 6 in order to apply Lemma 1. We therefore set  $\alpha_n := \nu_n - \sum_{i=0}^r \nu_n^i$  and  $\zeta_n := (H - x_n x_n^T) \nu_{n-1}^r$  for all  $r \in \mathbb{N} \cup \{0\}$  and  $n \in \mathbb{N}$ . We define  $\alpha_0 = \nu_0 - \sum_{i=0}^r \nu_0^i = 0 - 0 = 0$ . We will now check the assumptions of Lemma 1 for those choices of  $\alpha_n$  and  $\zeta_n$ . First, note that here  $\zeta_n$  is defined just like it was defined in the proof of the previous claim.

- Recursion equation: For the chosen  $\alpha_n$  and  $\zeta_n$ , the recursion equation  $\alpha_n = (I - \gamma x_n x_n^T) \alpha_{n-1} + \gamma \zeta_n$  holds for  $n \in \mathbb{N}$ , as shown in Claim 6.
- Measurability: We know that  $x_n$  is  $\mathcal{F}_n$ -measurable for all  $n \in \mathbb{N}$  by construction of  $\mathcal{F}_n$ .  $\zeta_n$  is  $\mathcal{F}_n$  is measurable as argued in the proof of the previous claim.
- Conditionally centered: Once again, we already know that  $\mathbb{E} [\zeta_n | \mathcal{F}_{n-1}] = 0$  for all  $n \in \mathbb{N}$ , as this was shown in the proof of the previous claim.
- Finite second moment of the norm: Recall from the remark right before this claim that for all  $n \in \mathbb{N}$  it holds that  $\mathbb{E} [\zeta_n \zeta_n^T] \preceq \gamma^{r+1} R^{2r+2} \sigma^2 H$ . This implies that

$$\begin{aligned} \mathbb{E} [\|\zeta_n\|^2] &= \mathbb{E} \left[ \text{tr} (\|\zeta_n\|^2) \right] = \mathbb{E} \left[ \text{tr} (\zeta_n \zeta_n^T) \right] = \text{tr} \left( \mathbb{E} [\zeta_n \zeta_n^T] \right) \\ &\leq \text{tr} \left( \gamma^{r+1} R^{2r+2} \sigma^2 H \right) = \gamma^{r+1} R^{2r+2} \sigma^2 \text{tr} (H). \end{aligned} \quad (\text{B.33})$$

Even more, since we know that  $\gamma H \preceq I$  as shown in the discussion of the assumptions, we know  $H \preceq \frac{1}{\gamma} I$  and therefore

$$\begin{aligned} \mathbb{E} [\|\zeta_n\|^2] &\leq \gamma^{r+1} R^{2r+2} \sigma^2 \text{tr} \left( \frac{1}{\gamma} I \right) \\ &\leq \gamma^r R^{2r+2} \sigma^2 d \\ &< \infty. \end{aligned} \quad (\text{B.34})$$

- Bound on higher moments: The condition that  $\mathbb{E} [\|x_n\|^2 x_n x_n^T | \mathcal{F}_{n-1}] \preceq R^2 H$  with  $H = \mathbb{E} [x_n x_n^T | \mathcal{F}_{n-1}]$  invertible for all  $n \in \mathbb{N}$  is satisfied, due to the assumptions **(A2')** and **(A4')** and the independence of the observations  $(x_n, z_n)$ ,  $n \in \mathbb{N}$ , as assumed in **(A1')**.
- Choice of  $\gamma$ :  $\gamma R^2 \leq 1$  is obviously satisfied due to assumption **(A5')**.

Note that we can now apply Lemma 1. Using the reformulation provided in the remark after

B. The proof of Theorem 1 by Bach and Moulines [2013]

Lemma 1 (see (B.2)) with  $\gamma R^2 < 1$ , this gives us for all  $n \in \mathbb{N}$ :

$$\|\bar{\alpha}_{n-1}\|_H^2 = \mathbb{E} [\langle \bar{\alpha}_{n-1}, H \bar{\alpha}_{n-1} \rangle] \leq \frac{1}{1 - \gamma R^2} \frac{\gamma}{n} \left( \frac{1}{2\gamma^2} \|\alpha_0\|^2 - \frac{1}{2\gamma^2} \mathbb{E} [\|\alpha_n\|^2] + \sum_{k=1}^n \mathbb{E} [\|\zeta_k\|^2] \right).$$

Since  $\bar{\alpha}_{n-1} = \bar{\nu}_{n-1} - \sum_{i=0}^r \bar{\nu}_{n-1}^i$  and  $\alpha_0 = \nu_0 - \sum_{i=0}^r \nu_0^i = 0 - \sum_{i=0}^r 0 = 0$ , we get

$$\begin{aligned} \left\| \bar{\nu}_{n-1} - \sum_{i=0}^r \bar{\nu}_{n-1}^i \right\|_H^2 &\leq \frac{1}{1 - \gamma R^2} \frac{\gamma}{n} \left( \frac{1}{2\gamma^2} \underbrace{\|\alpha_0\|^2}_{=0} - \frac{1}{2\gamma^2} \underbrace{\mathbb{E} [\|\alpha_n\|^2]}_{\geq 0} + \sum_{k=1}^n \mathbb{E} [\|\zeta_k\|^2] \right) \\ &\leq \frac{1}{1 - \gamma R^2} \frac{\gamma}{n} \sum_{k=1}^n \mathbb{E} [\|\zeta_k\|^2]. \end{aligned}$$

Recall that  $\mathbb{E} [\|\zeta_k\|^2] \leq \gamma^r R^{2r+2} \sigma^2 d$  as we have just discussed in the assumptions. This yields

$$\begin{aligned} \left\| \bar{\nu}_{n-1} - \sum_{i=0}^r \bar{\nu}_{n-1}^i \right\|_H^2 &\leq \frac{1}{1 - \gamma R^2} \frac{\gamma}{n} \sum_{k=1}^n \gamma^r R^{2r+2} \sigma^2 d \\ &= \frac{1}{1 - \gamma R^2} \gamma^{r+1} R^{2r+2} \sigma^2 d, \end{aligned}$$

which completes the proof of the claim. □

We have therefore found the upper bound

$$A_r = \sqrt{\frac{1}{1 - \gamma R^2} \gamma^{r+1} R^{2r+2} \sigma^2 d}. \quad (\text{B.35})$$

**Construction of the upper bound  $B_r$**

Here, it will be our goal to find an upper bound  $B_r$  such that  $B_r \geq \sum_{i=0}^r \|\bar{\nu}_{n-1}^i\|_H$  for all  $n \in \mathbb{N}$  for any  $r \in \mathbb{N} \cup \{0\}$ . To that end, we will first derive an upper bound on  $\|\bar{\nu}_{n-1}^r\|_H^2$  for all  $r \in \mathbb{N} \cup \{0\}$  and all  $n \in \mathbb{N}$ :

**Claim 9.** *For all  $r \in \mathbb{N} \cup \{0\}$  and all  $n \in \mathbb{N}$  it holds that*

$$\|\bar{\nu}_{n-1}^r\|_H^2 \leq \frac{1}{n} \gamma^r R^{2r} \sigma^2 d.$$

*Proof.* Note that the second result of Lemma 2 (see (B.11)) can be used to derive this result for  $\alpha_n = \nu_n^r$  and  $\zeta_n = (H - x_n x_n^T) \nu_{n-1}^{r-1}$ . In the case  $r \neq 0$ , this is merely an index shift in  $r$  by 1 compared to the definition of  $\alpha_n$  and  $\zeta_n$  in the induction step of the proof of Claim 6. Therefore, we already know that the assumptions of Lemma 2 are met, and that  $\mathbb{E} [\zeta_n \zeta_n^T] \preccurlyeq \gamma^r R^{2r} \sigma^2 H =: C$

### B.5. Construction of the upper bound on the expected excess risk

if  $r \neq 0$ . Note, furthermore, that  $\alpha_0 = 0$  for all  $r \in \mathbb{N} \cup \{0\}$ .

Therefore, Lemma 2 gives us for  $r \in \mathbb{N}, n \in \mathbb{N}$ :

$$\begin{aligned} \|\bar{\nu}_{n-1}^r\|_H^2 &\leq \frac{1}{n\gamma} \|0\|^2 + \frac{\text{tr}(\gamma^r R^{2r} \sigma^2 H H^{-1})}{n} \\ &\leq \frac{1}{n} \text{tr}(\gamma^r R^{2r} \sigma^2 I) \\ &= \frac{1}{n} \gamma^r R^{2r} \sigma^2 d, \end{aligned}$$

which completes the proof for  $r \neq 0$ .

As for the case  $r = 0$ , we can still apply Lemma 2 for  $\alpha_n = \nu_n^0$  for all  $n \in \mathbb{N} \cup \{0\}$  and  $\zeta_n = \xi_n$  for all  $n \in \mathbb{N}$ . The assumptions of the lemma are again met.

- Recursion equation: The recursion equation  $\alpha_n = (I - \gamma H)\alpha_{n-1} + \gamma \zeta_n$  holds for  $n \in \mathbb{N}$  by definition of the sequence  $(\nu_n^0)_{n \in \mathbb{N} \cup \{0\}}$ . Furthermore,  $H$  is assumed invertible in **(A2')** and is therefore positive definite.  $H$  is symmetric and therefore orthogonally diagonalizable.
- Measurability: By construction of  $\mathcal{F}_n$ ,  $\zeta_n = \xi_n$  is  $\mathcal{F}_n$ -measurable for all  $n \in \mathbb{N}$ .
- Conditionally centered: Due to the independence assumption in **(A1')** and the fact that the random vectors  $\xi_n$  are centered for all  $n \in \mathbb{N}$  as argued in Section B.2, it holds that  $\mathbb{E}[\zeta_n | \mathcal{F}_{n-1}] = \mathbb{E}[\xi_n | \mathcal{F}_{n-1}] = \mathbb{E}[\xi_n] = 0$ .
- Bound on second moments: By assumption **(A4')**, it holds for all  $n \in \mathbb{N}$  that  $\mathbb{E}[\zeta_n \zeta_n^T] = \mathbb{E}[\xi_n \xi_n^T] \preceq \sigma^2 H$ .
- Choice of  $\gamma$ : Due to the assumption  $\gamma R^2 < 1$  made in **(A5')**, it holds that  $\gamma H \preceq I$  as argued in Section B.2.

Note that once again  $\alpha_0 = 0$ .

We therefore get from Lemma 2 (see (B.11)):

$$\begin{aligned} \|\bar{\nu}_{n-1}^0\|_H^2 &\leq \frac{1}{n\gamma} \|0\|^2 + \frac{\text{tr}(\sigma^2 H H^{-1})}{n} \\ &\leq \frac{1}{n} \text{tr}(\sigma^2 I) \\ &= \frac{1}{n} \sigma^2 d, \end{aligned}$$

which completes the proof for  $r = 0$  and thus for all  $r \in \mathbb{N} \cup \{0\}$ . □

This directly yields the upper bound

$$B_r = \sum_{i=0}^r \sqrt{\frac{1}{n} \gamma^i R^{2i} \sigma^2 d} = \frac{\sigma \sqrt{d}}{\sqrt{n}} \sum_{i=0}^r \left( \sqrt{\gamma R^2} \right)^i. \quad (\text{B.36})$$

B. The proof of Theorem 1 by Bach and Moulines [2013]

**Finalizing the upper bound on**  $\left\| \frac{\gamma}{n} \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} M_{j+1}^k \xi_j \right\|_H$

As we have argued in the beginning of this subsection, it holds for all  $r \in \mathbb{N} \cup \{0\}$  that

$$\left\| \frac{\gamma}{n} \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} M_{j+1}^k \xi_j \right\|_H \leq A_r + B_r.$$

Since the LHS does not depend on  $r$ , it also holds true that

$$\left\| \frac{\gamma}{n} \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} M_{j+1}^k \xi_j \right\|_H \leq \lim_{r \rightarrow \infty} (A_r + B_r).$$

Recall the bound  $A_r$  from (B.35) to see that

$$\begin{aligned} \lim_{r \rightarrow \infty} A_r &= \lim_{r \rightarrow \infty} \sqrt{\frac{1}{1 - \gamma R^2} \gamma^{r+1} R^{2r+2} \sigma^2 d} \\ &= \sqrt{\frac{\gamma R^2 \sigma^2 d}{1 - \gamma R^2}} \cdot \lim_{r \rightarrow \infty} \sqrt{(\gamma R^2)^r} \\ &= 0, \end{aligned}$$

since  $0 < \gamma R^2 < 1$ .

As for  $B_r$ , recall the bound from (B.36) to see that we have a geometric sum:

$$\begin{aligned} \lim_{r \rightarrow \infty} B_r &= \lim_{r \rightarrow \infty} \frac{\sigma \sqrt{d}}{\sqrt{n}} \sum_{i=0}^r \left( \sqrt{\gamma R^2} \right)^i \\ &= \frac{\sigma \sqrt{d}}{\sqrt{n}} \frac{1}{1 - \sqrt{\gamma R^2}}, \end{aligned}$$

since  $0 < \gamma R^2 < 1$ .

We consequently get

$$\begin{aligned} \left\| \frac{\gamma}{n} \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} M_{j+1}^k \xi_j \right\|_H &\leq 0 + \frac{\sigma \sqrt{d}}{\sqrt{n}} \frac{1}{1 - \sqrt{\gamma R^2}} \\ &= \frac{\sigma \sqrt{d}}{\sqrt{n}} \frac{1}{1 - \sqrt{\gamma R^2}}, \end{aligned}$$

which at last proves the inequality from (B.30).

### B.5.3. Completion of the upper bound on the expected excess risk

Recall from (B.27) that

$$\mathbb{E} [f(\bar{\theta}_{n-1}) - f(\theta_*)] \leq \frac{1}{2} \left( \left\| \frac{1}{n} \sum_{k=0}^{n-1} M_1^k \eta_0 \right\|_H + \left\| \frac{\gamma}{n} \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} M_{j+1}^k \xi_j \right\|_H \right)^2.$$

Recall further from (B.28) that

$$\left\| \frac{1}{n} \sum_{k=0}^{n-1} M_1^k \eta_0 \right\|_H \leq \frac{\|\theta_0 - \theta_*\|}{\sqrt{n\gamma}}$$

and lastly that we have just shown (B.30):

$$\left\| \frac{\gamma}{n} \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} M_{j+1}^k \xi_j \right\|_H \leq \frac{\sigma\sqrt{d}}{\sqrt{n}} \frac{1}{1 - \sqrt{\gamma R^2}}.$$

Combining these three results, immediately gives the final upper bound

$$\begin{aligned} \mathbb{E} [f(\bar{\theta}_{n-1}) - f(\theta_*)] &\leq \frac{1}{2} \left( \frac{\|\theta_0 - \theta_*\|}{\sqrt{n\gamma}} + \frac{\sigma\sqrt{d}}{\sqrt{n}} \frac{1}{1 - \sqrt{\gamma R^2}} \right)^2 \\ &= \frac{1}{2n} \left( \frac{\sigma\sqrt{d}}{1 - \sqrt{\gamma R^2}} + R \frac{\|\theta_0 - \theta_*\|}{\sqrt{\gamma R^2}} \right)^2, \end{aligned}$$

thus completing the proof of the theorem.

## B.6. Improvement of the bound

In this section, a slight improvement of the bound  $B_r$  by Bach and Moulines [2013] will be derived using some of their proof technique. In the previous section, we have defined the following sequences:

$$\begin{aligned} \nu_n &= (I - \gamma x_n x_n^T) \nu_{n-1} + \gamma \xi_n && \text{for } n \in \mathbb{N} \text{ with } \nu_0 = 0, \\ \nu_n^0 &= (I - \gamma H) \nu_{n-1}^0 + \gamma \xi_n && \text{for } n \in \mathbb{N} \text{ with } \nu_0^0 = 0, \\ \text{for } r \in \mathbb{N} : \nu_n^r &= (I - \gamma H) \nu_{n-1}^r + \gamma (H - x_n x_n^T) \nu_{n-1}^{r-1} && \text{for } n \in \mathbb{N} \text{ with } \nu_0^r = 0. \end{aligned}$$

Note that the third recursion equation implies that for any  $r, n \in \mathbb{N}$  it holds that  $\nu_n^r = 0$  if  $\nu_{n-1}^r = 0$  and  $\nu_{n-1}^{r-1} = 0$ . Since we can apply this argument iteratively, and  $\nu_0^0 = \nu_0^r = 0$  for all  $r \in \mathbb{N}$ , it leads to the idea that  $\nu_n^r = 0$  in certain cases. More specifically,  $\nu_n^r = 0$  as long as  $r \geq n$ , as we will prove in the following claim.

B. The proof of Theorem 1 by Bach and Moulines [2013]

**Claim 10.** *With the previously defined sequences, it holds that*

$$\forall n \in \mathbb{N} \cup \{0\} \quad \forall r \geq n : \quad \nu_n^r = 0.$$

*Proof.* The proof by induction over  $n$  is very short.

$n = 0$ :  $\nu_0^r = 0$  holds for all  $r \geq n = 0$  by definition of the sequences.

$n \mapsto n + 1$ : Let  $r \geq n + 1$  be arbitrary. Note that this implies that  $r \geq 1$ ,  $r - 1 \geq n$  and  $r \geq n$ . This means that under the induction assumption, we have  $\nu_n^{r-1} = 0$  and  $\nu_n^r = 0$ . We therefore have

$$\nu_{n+1}^r = (I - \gamma H) \underbrace{\nu_n^r}_{=0} + \gamma (H - x_{n+1} x_{n+1}^T) \underbrace{\nu_n^{r-1}}_{=0} = 0,$$

thus showing that for all  $r \geq n + 1$  it holds that  $\nu_{n+1}^r = 0$  and completing the proof.  $\square$

The important implication of this result is that  $\|\bar{\nu}_{n-1}^r\|_H^2$  is not only bounded by the term  $\frac{1}{n} \gamma^r R^{2r} \sigma^2 d$  for all  $r \in \mathbb{N} \cup \{0\}$  as shown in Claim 9, but actually constant for all  $r \geq r_0$  with appropriate  $r_0 \in \mathbb{N}$ . This is the motivation for the next claim. Note that this will lead to an improvement of the bound  $B_r$  which directly translates to the final result.

**Claim 11.** *With the previously defined sequences, it holds for  $\bar{\nu}_{n-1}^r = \frac{1}{n} \sum_{k=0}^{n-1} \nu_k^r$  for all  $r \in \mathbb{N} \cup \{0\}, n \in \mathbb{N}$ :*

$$\|\bar{\nu}_{n-1}^r\|_H \leq \begin{cases} \sqrt{\frac{n-r}{n^2} \gamma^r R^{2r} \sigma^2 d} & \text{for } r < n - 1, \\ 0 & \text{for } r \geq n - 1. \end{cases}$$

*Proof.* For the proof, we will first look at the very easy case that  $r \geq n - 1$ , before then considering the case  $r < n - 1$ .

Case  $r \geq n - 1$ : Using the previous claim, the proof is very simple. Let  $n \in \mathbb{N}$  and  $r \geq n - 1$  be arbitrary. It holds that

$$\bar{\nu}_{n-1}^r = \frac{1}{n} \sum_{k=0}^{n-1} \underbrace{\nu_k^r}_{=0} = 0.$$

Case  $r < n - 1$ : Note that the case  $r = 0$  has already been shown, since the upper bound presented here is then simplified to the one by Bach and Moulines [2013]. Therefore, let us only consider cases in which  $1 \leq r < n - 1$ . First, we will make use of the fact that using the previous claim, we can write

$$\bar{\nu}_{n-1}^r = \frac{1}{n} \sum_{k=0}^{n-1} \nu_k^r = \frac{1}{n} \left( \sum_{k=0}^r \underbrace{\nu_k^r}_{=0} + \sum_{k=r+1}^{n-1} \nu_k^r \right) = \frac{1}{n} \sum_{k=r+1}^{n-1} \nu_k^r = \frac{1}{n} \sum_{k=r}^{n-1} \nu_k^r$$

with the last equality being true since  $\nu_r^r = 0$ .

Let us then consider the sequence  $(\alpha_j)_{j \geq 0}$  with  $\alpha_j = \nu_{r+j}^r$  for  $j \in \mathbb{N} \cup \{0\}$ . We then have  $\alpha_0 = \nu_r^r = 0$  and the familiar recursion equation  $\alpha_n = (I - \gamma H)\alpha_{n-1} + \gamma \zeta_n$  for  $\zeta_n = (H - x_{r+n} x_{r+n}^T) \nu_{r+n-1}^{r-1}$ . As we have already shown in the proofs of Claim 7 and Claim 9, we can apply Lemma 2 for the accordingly shifted sequence of  $\sigma$ -fields  $(\mathcal{F}_{r+n})_{n \in \mathbb{N} \cup \{0\}}$  and with  $C = \gamma^r R^{2r} \sigma^2 H$ . Recall that (B.11) with a slight change of notation is

$$\|\bar{\alpha}_{m-1}\|_H^2 = \mathbb{E}[\langle \bar{\alpha}_{m-1}, H \bar{\alpha}_{m-1} \rangle] \leq \frac{1}{m\gamma} \|\alpha_0\|^2 + \frac{\text{tr}(CH^{-1})}{m} \quad (\text{B.37})$$

for all  $m \in \mathbb{N}$ . Note that

$$\bar{\nu}_{n-1}^r = \frac{1}{n} \sum_{k=r}^{n-1} \nu_k^r = \frac{1}{n} \sum_{j=0}^{n-r-1} \alpha_j^r = \frac{n-r}{n} \frac{1}{n-r} \sum_{j=0}^{n-r-1} \alpha_j^r = \frac{n-r}{n} \bar{\alpha}_{n-r-1}^r$$

and consequently

$$\|\bar{\nu}_{n-1}^r\|_H = \frac{n-r}{n} \|\bar{\alpha}_{n-r-1}^r\|_H$$

or equivalently

$$\frac{n^2}{(n-r)^2} \|\bar{\nu}_{n-1}^r\|_H^2 = \|\bar{\alpha}_{n-r-1}^r\|_H^2.$$

Using furthermore that  $\alpha_0 = 0$ , (B.37) gives us for  $m = n - r$  together with the last equation that

$$\frac{n^2}{(n-r)^2} \|\bar{\nu}_{n-1}^r\|_H^2 = \|\bar{\alpha}_{n-r-1}^r\|_H^2 \leq \frac{\text{tr}(\gamma^r R^{2r} \sigma^2 H H^{-1})}{n-r}$$

and consequently

$$\|\bar{\nu}_{n-1}^r\|_H^2 \leq \frac{n-r}{n^2} \gamma^r R^{2r} \sigma^2 d,$$

which immediately proves the claim.  $\square$

We can now use this result in order to find a slightly better upper bound  $B_r^*$ , such that

$$B_r \geq B_r^* \geq \sum_{i=0}^r \|\bar{\nu}_{n-1}^i\|_H$$

holds for all  $r \in \mathbb{N} \cup \{0\}$  and  $n \in \mathbb{N}$ . This will be dealt with in the next claim.

B. The proof of Theorem 1 by Bach and Moulines [2013]

**Claim 12.** For all  $n \in \mathbb{N}$  and  $r \geq n - 1$  it holds that

$$\sum_{i=0}^r \left\| \bar{\nu}_{n-1}^i \right\|_H \leq \sigma \sqrt{d} \sum_{i=0}^{n-2} \frac{\sqrt{n-i}}{n} \left( \sqrt{\gamma R^2} \right)^i \quad (\text{B.38})$$

or equivalently

$$\sum_{i=0}^r \left\| \bar{\nu}_{n-1}^i \right\|_H \leq \frac{\sigma \sqrt{d}}{n} \left( \sqrt{\gamma R^2} \right)^n \sum_{i=2}^n \sqrt{i} \left( \frac{1}{\sqrt{\gamma R^2}} \right)^i.$$

*Proof.* Let  $n \in \mathbb{N}$  be arbitrary. Let  $r \geq n - 1$  be arbitrary. Using the result of the preceding claim we can then write

$$\begin{aligned} \sum_{i=0}^r \left\| \bar{\nu}_{n-1}^i \right\|_H &= \sum_{i=0}^{n-2} \underbrace{\left\| \bar{\nu}_{n-1}^i \right\|_H}_{\leq \sqrt{\frac{n-i}{n^2} (\gamma R^2)^i \sigma^2 d}} + \sum_{i=n-1}^r \underbrace{\left\| \bar{\nu}_{n-1}^i \right\|_H}_{=0} \\ &\leq \sigma \sqrt{d} \sum_{i=0}^{n-2} \frac{\sqrt{n-i}}{n} \left( \sqrt{\gamma R^2} \right)^i, \end{aligned}$$

which is the first result. The second, equivalent result follows immediately:

$$\begin{aligned} \sigma \sqrt{d} \sum_{i=0}^{n-2} \frac{\sqrt{n-i}}{n} \left( \sqrt{\gamma R^2} \right)^i &= \frac{\sigma \sqrt{d}}{n} \sum_{i=0}^{n-2} \sqrt{(n-i)(\gamma R^2)^i} \\ &= \frac{\sigma \sqrt{d}}{n} \sum_{i=2}^n \sqrt{i(\gamma R^2)^{n-i}} \\ &= \frac{\sigma \sqrt{d}}{n} \left( \sqrt{\gamma R^2} \right)^n \sum_{i=2}^n \sqrt{i} \left( \frac{1}{\sqrt{\gamma R^2}} \right)^i. \end{aligned}$$

□

We can therefore set

$$B_r^* := \begin{cases} B_r & \text{for } r < n - 1, \\ \sigma \sqrt{d} \sum_{i=0}^{n-2} \frac{\sqrt{n-i}}{n} \left( \sqrt{\gamma R^2} \right)^i & \text{for } r \geq n - 1. \end{cases}$$

*Remark.* Note that we could easily improve the bound for  $r < n - 1$  as well by using Claim 11 directly in order to get

$$\sum_{i=0}^r \left\| \bar{\nu}_{n-1}^i \right\|_H \leq \sum_{i=0}^r \sqrt{\frac{n-i}{n^2} (\gamma R^2)^i \sigma^2 d},$$

but we are only interested in the asymptotic behaviour of the bounds  $B_r^*$  and  $B_r$  for  $r \rightarrow \infty$ .

*Remark.* We can achieve an upper bound with a much simpler form than  $B_r^*$  that is still an improvement compared to  $B_r$ .

Since  $\frac{\sqrt{n-i}}{n} \leq \frac{\sqrt{n}}{n} = \frac{1}{\sqrt{n}}$  for all  $0 \leq i \leq n$ , (B.38) gives us the slightly weaker bound

$$\begin{aligned} \sigma\sqrt{d} \sum_{i=0}^{n-2} \frac{\sqrt{n-i}}{n} \left(\sqrt{\gamma R^2}\right)^i &\leq \frac{\sigma\sqrt{d}}{\sqrt{n}} \sum_{i=0}^{n-2} \left(\sqrt{\gamma R^2}\right)^i \\ &= \frac{\sigma\sqrt{d}}{\sqrt{n}} \frac{1 - \left(\sqrt{\gamma R^2}\right)^{n-1}}{1 - \sqrt{\gamma R^2}}. \end{aligned}$$

This would be another alternative for  $B_r$  for large  $r$ , i.e. we could define

$$B_r^{**} := \begin{cases} B_r & \text{for } r < n-1, \\ \frac{\sigma\sqrt{d}}{\sqrt{n}} \frac{1 - \left(\sqrt{\gamma R^2}\right)^{n-1}}{1 - \sqrt{\gamma R^2}} & \text{for } r \geq n-1. \end{cases}$$

However, note that  $\lim_{r \rightarrow \infty} B_r = \frac{\sigma\sqrt{d}}{\sqrt{n}} \frac{1}{1 - \sqrt{\gamma R^2}}$ . As a result, the improvement we reach by making use of  $B_r^{**}$  is

$$\begin{aligned} \frac{\lim_{r \rightarrow \infty} B_r^{**}}{\lim_{r \rightarrow \infty} B_r} &= \frac{\frac{\sigma\sqrt{d}}{\sqrt{n}} \frac{1 - \left(\sqrt{\gamma R^2}\right)^{n-1}}{1 - \sqrt{\gamma R^2}}}{\frac{\sigma\sqrt{d}}{\sqrt{n}} \frac{1}{1 - \sqrt{\gamma R^2}}} \\ &= 1 - \left(\sqrt{\gamma R^2}\right)^{n-1}. \end{aligned}$$

This means that the improvement vanishes exponentially fast with increasing sample size  $n$ . Nevertheless, this remark directly implies that we could improve the final bound to

$$\mathbb{E} [f(\bar{\theta}_{n-1}) - f(\theta_*)] \leq \frac{1}{2n} \left( \sigma\sqrt{d} \frac{1 - \left(\sqrt{\gamma R^2}\right)^{n-1}}{1 - \sqrt{\gamma R^2}} + R \frac{\|\theta_0 - \theta_*\|}{\sqrt{\gamma R^2}} \right)^2.$$