Optimization in Machine Learning

Simon Weissmann
University of Mannheim

Abstract
This lecture course covers Optimization methods applied in the research area of machine learning. In the first part of the lecture, we will start with a detailed overview of first order gradient methods. After a brief recap of the basics of unconstrained optimization problems, we will study (accelerated) gradient descent methods for various conditions on the corresponding cost function. In the second part of this lecture course, we will study stochastic variants of gradient descent. We will provide a rigorous introduction of stochastic gradient estimations and standard convergence results. Due to the stochastic approximations we discuss the incorporation of variance reduction methods.

Contents

1

Introduction

These lecture notes were originally prepared for the course *Optimization in Machine Learning* offered during the Spring semester of 2023 at the University of Mannheim. The current version includes updates and modifications for the Fall 2024 iteration of the same course. Several of the presented results are taken from books and monographs, and the corresponding references are provided in place. I would like to thank Marc Schäfer for carefully proofreading this manuscript.

In this course, we will delve into optimization methods applied in the research field of *machine learning*. In Chapter 2, we provide a comprehensive introduction to unconstrained optimization, with a particular emphasis on gradient descent methods. We will present a series of convergence results for various assumptions regarding the cost function, such as differentiability, (strong) convexity, and smoothness. Moreover, we discuss the non-smooth scenario introducing the notion of sub-differential. The incorporation of momentum leading to accelerated gradient descent methods will be discussed in Chapter 3. This discussion will include both Polyak's heavy ball method and Nesterov's acceleration method. In Chapter 4, we study the role of empirical approximations in the area of expected risk minimization. Building upon this foundation, we will introduce stochastic variants of gradient descent methods, which play a crucial role in solving empirical and expected risk minimization problems. We will discuss different types of convergence results and introduce schemes for variance reduction.

Before we start discussing the important aspects of optimization methods, we will provide a brief overview of the research field machine learning and position the role of this lecture within this field. The focus of this course lies in the theoretical analysis and practical implementation of (stochastic) optimization methods which are typically applied in the *training* (or *learning*) task of machine learning models.
The field of machine learning is often divided into the following three classes:

1. *Supervised learning:* Given pairs of *input* and *output* vectors, the aim in supervised learning is to learn/ describe connections between the input and output. This includes the task of *classification* (discrete output state) and *regression* (continuous output state).

2. *Unsupervised learning:* In this class, the data set is given by input vectors without specified labels. The aim is for example to separate the data set into cluster (*clustering*) or to learn a probability distribution describing the data set (*density estimation*)

3. *Reinforcement learning:* In this class, an *agent* aims to find an optimal strategy of actions in order to maximize some *reward* resulting from the action. However, there is no access to any pre-observed data set, and the optimal strategy needs to be learned via trial and error of different applied actions.

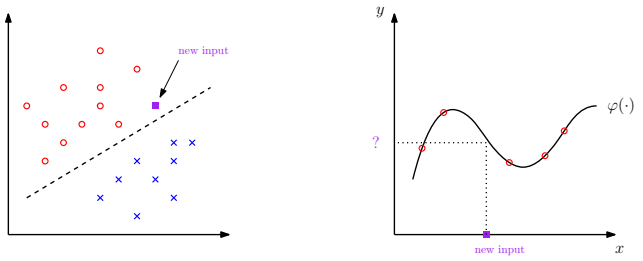

Figure 1.1: Illustration of classification and regression problems.

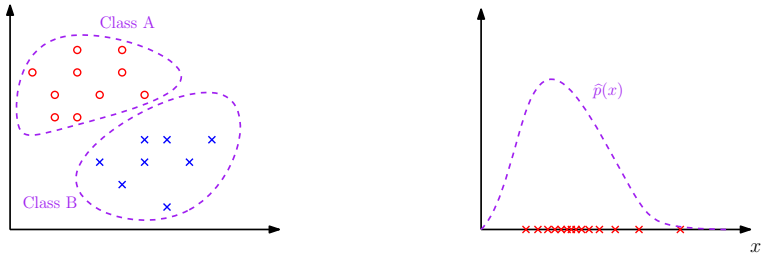

Figure 1.2: Illustration of clustering and density estimation problems.

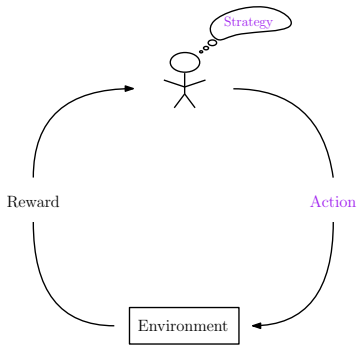

Figure 1.3: Illustration of reinforcement learning.

The presented lecture course focuses mainly on optimization methods in the area of supervised learning. In order to introduce the learning task in supervised learning, we will introduce the following notation:

- Input set: $\mathcal{Z} \subset \mathbb{R}^{d_z}$ (e.g. images, observation points, features)
- Output/target set: $\mathcal{Y} \subset \mathbb{R}^{d_y}$ (e.g. label $\{0,1\}$ or function evaluation in $\mathbb{R}^{d_y}$
- Training data: $S = \{(z^{(1)}, y^{(1)}), \dots, (z^{(m)}, y^{(m)})\}$ with $(z^{(i)}, y^{(i)}) \in \mathcal{Z} \times \mathcal{Y}$, $i = 1, \dots, m$.

The goal in supervised learning is to approximate the unknown model

$$\varphi : \mathcal{Z} \to \mathcal{Y}, \quad z \mapsto \varphi(z) = y$$

The *task of the learner* is to construct a prediction or approximation $g : \mathcal{Z} \to \mathcal{Y}$, which is actually a task of function approximation. Typically, the learner aims to find a (finite dimensional) parameter $\theta \in \Theta$ and compute the approximation $g_\theta : \mathcal{Z} \to \mathcal{Y}$. Here, we will call both $\mathcal{G}$ and Θ the *learning class* of possible candidates for $g \in \mathcal{G}$ or $\theta \in \Theta$ respectively. We assume that the learning class $\mathcal{G}$ is a Hilbert space.
In the following, we fix an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The *data model* is usually described by the *training data set* as a family of random variables. For example, one may assume that $Z^{(i)} \sim \mu_Z$ can be generated independently and pushed through the "true" model φ:

$$Y^{(i)} := \varphi(Z^{(i)}) + \xi^{(i)}, \quad i = 1, \dots, m,$$

where $(\xi^{(i)})_{i=1}^m$ denotes possible noise. Hence, we will model the in- and output as jointly varying random variable $(Z, Y) \sim \mu_{(Z,Y)}$ with joint unknown distribution $\mu_{(Z,Y)}$. Later, we will usually assume that we are able to generate iid. sample of the data $\{(Z^{(i)}, Y^{(i)})\}_{i=1,\dots,m}, m \in \mathbb{N}$, with $(Z^{(1)}, Y^{(1)}) \sim \mu_{(Z,Y)}$. In the first step, we want to define a measure of success for the prediction of the learner. Let $f : \mathcal{G} \times \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$ be measurable (wrt. $\mathcal{B}(\mathcal{G}) \otimes \mathcal{B}(\mathcal{Z}) \otimes \mathcal{B}(\mathcal{Y})/\mathcal{B}(\mathbb{R})$) and let $(Z, Y) \sim \mu_{(Z,Y)}$ with $\mathbb{E}[|f(g, Z, Y)|] < +\infty$ for all $g \in \mathcal{G}$.

(i) We define the *expected risk* $F : \mathcal{G} \to \mathbb{R}$ by

$$F(g) := \mathbb{E}_{\mu_{(Z,Y)}}[f(g, Z, Y)] := \int_{\mathcal{Z}\times\mathcal{Y}} f(g, z, y)\, \mu_{(Z,Y)}(d(z, y)).$$

(ii) Let $(Z^{(1)}, Y^{(1)}), \dots, (Z^{(m)}, Y^{(m)})$ be iid. random variables with $(Z^{(1)}, Y^{(1)}) \sim \mu_{(Z,Y)}$. We define the *empirical risk* $F_m : \mathcal{G} \to \mathbb{R}$ by

$$F_m(g) := \frac{1}{m} \sum_{i=1}^m f(g, Z^{(i)}, Y^{(i)})$$

We call f the *risk function*, and the tasks

$$\min_{g\in\mathcal{G}} F(g) \quad (\min_{g\in\mathcal{G}} F_m(g))$$

the *expected (empirical) risk minimization problem.*

Example 1.0.1 (Classification problem)**.** *Let $\mathcal{Y} = \{0,1\} \subset \mathbb{R}$ and assume that the training data is generated without noise as $Y^{(i)} = \varphi(Z^{(i)})\,,\ i = 1,\dots,m$. In a classification problem one usually aims to classify between a finite number of classes, here for simplicity between two classes. The first class corresponds to 0 and the second one to 1. Hence, the true model φ maps to the set $\{0,1\}$ and states whether the input belongs to 0 or 1.*
A common choice for a risk function is an indicator function over the predicted classification through $g \in \mathcal{G}$. We receive a penalty 1 if the prediction is wrong, whereas 0 if the prediction is correct. Mathematically written, the risk function can be defined by

$$f(g,z,y) := \mathbb{1}_{\{g(z)\neq y\}} = \mathbb{1}_{\{g(z)\neq\varphi(z)\}}$$

and the corresponding expected risk correspond to the probability of giving a wrong prediction

$$F(g) = \mathbb{E}_{\mu_{(Z,Y)}}[\mathbb{1}_{\{g(Z)\neq Y\}}] = \mathbb{P}_{\mu_{(Z,Y)}}(g(Z)\neq Y)\,.$$

When having access to a training data set $\{(Z^{(i)},Y^{(i)})\}_{i=1}^m$ the empirical risk counts the relative number of failures in predicting the correct label

$$F_m(g) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}_{\{g(Z^{(i)})\neq Y^{(i)}\}} = \frac{|\{i\in\{1,\dots,m\} \mid g(Z^{(i)})\neq Y^{(i)}\}|}{m}\,.$$

Example 1.0.2 (Regression problem)**.** *The second classical example of supervised learning tasks are regression problems. As an example we consider the task to approximate a function $\varphi : \mathbb{R}^{d_z} \to \mathbb{R}^{d_y}$ based on the risk function*

$$f(g,z,y) := \frac{1}{2}\|g(z)-y\|^2_{\mathbb{R}^{d_y}}\,.$$

The corresponding expected risk measures the expected squared distance of the prediction g

$$F(g) = \mathbb{E}_{\mu_{(Z,Y)}}[\frac{1}{2}\|g(Z)-Y\|^2_{\mathbb{R}^{d_y}}]\,,$$

and the empirical risk computes the averaged squared distance of the prediction

$$F_m(g) = \frac{1}{m}\sum_{i=1}^m \frac{1}{2}\|g(Z^{(i)})-Y^{(i)}\|^2_{\mathbb{R}^{d_y}}$$

within a training data set $\{(Z^{(i)},Y^{(i)})\}_{i=1}^m$.

Based on the learners chosen objective function, three central questions have accumulated significant attention in the machine learning research field:

1. *Approximation (expressive power):* In this research area, the concern revolves around the choice of the function class used to approximate the true model. The question is whether a suitable function class even exists for approximating the true model and how large one should choose the class to approximate the true model up to a certain accuracy.

2. *Training/ Learning task:* Once we have decided on a specific class of functions in which we aim to approximate the underlying true model, the next step is to find the best possible representation within this chosen class $\mathcal{G}$ or Θ. For example, assuming a parametrized representation $g_\theta : \mathbb{R}^{d_z} \to \mathbb{R}^{d_y}$ of the function approximation and given a training data set $\{(z^{(i)}, y^{(i)})\}_{i=1}^N$ constructed through the true model, i.e. $y^{(i)} = \varphi(z^{(i)})$, $i = 1, \dots, N$, we aim to find $\theta \in \Theta$ such that we obtain the best possible approximation $y^{(i)} \approx g_\theta(z^{(i)})$. As described above, the task in training/learning of the model is then to solve an optimization problem of the form
$$\min_{\theta\in\Theta} \; f_N(\theta, \{(z^{(i)}, y^{(i)})\}_{i=1}^N),$$
where $f_N : \Theta \times \left(\times_{i=1}^N (\mathbb{R}^{d_z} \times \mathbb{R}^{d_y})\right) \to \mathbb{R}$ is a suitable cost function. As we have seen in the example of regression, a typical example of cost function is
$$f_N(\theta, \{z^{(i)}, y^{(i)}\}_{i=1}^N) = \frac{1}{N}\sum_{i=1}^N \|g_\theta(z^{(i)}) - y^{(i)}\|^2 + \mathcal{R}(\theta),$$
where $\mathcal{R} : \Theta \to \mathbb{R}$ is a regularization function in order to avoid so-called *over fitting*.

3. *Generalization:* Once we learned/trained an optimal parameter θ_* approximating the true model, the natural question arises of how good does this approximation generalizes. We seek to evaluate the quality of the approximation
$$g_{\theta_*}(z) \approx \varphi(z)\,,$$
when applied to data points $(z, \varphi(z))$ which have not been used in the training task.

For better illustration, we consider the following example:

Example 1.0.3 (Polynomial regression)**.** *We assume that the underlying true model is described through*
$$y = \varphi(z) := \sin(2\pi z), \quad z \in [0, 1].$$
Given a training data set $\{(z^{(i)}, y^{(i)} = \varphi(z^{(i)}))\}_{i=1}^N$ *we want to construct a polynomial of* M *degrees*
$$g_\theta(z) = \theta_0 + \theta_1 z + \dots, \theta_M z^M,$$
in order to approximate φ*. The coefficients* $\theta = (\theta_0, \dots, \theta_M)^\top \in \mathbb{R}^{M+1}$ *are the parameters to be*

learned. Therefore, we firstly minimize the cost function

$$f^M(\theta) = \frac{1}{N}\sum_{i=1}^{N} |g_\theta(z^{(i)}) - y^{(i)}|^2,$$

also known as the data misfit functional. The resulting approximations are shown in Figure 1.4.

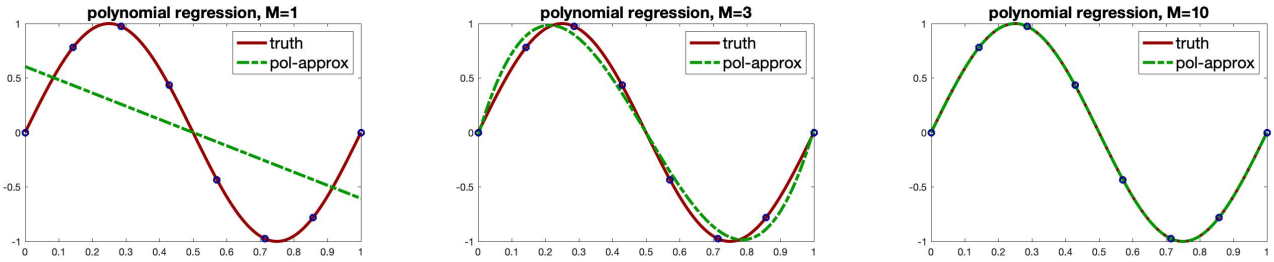

Figure 1.4: Polynomial approximation of φ for minimizing the data misfit functional for unperturbed data $y^{(i)} = \varphi(z^{(i)})$, $i = 1, \dots, N$.

In the case where the training dataset $\{(z^{(i)}, y^{(i)} = \varphi(z^{(i)}) + \eta^{(i)})\}_{i=1}^N$ *is perturbed by some noise* $\eta^{(i)} \in \mathbb{R}$ *the situation changes. The resulting optimization over the data misfit functional is ill-posed and regularization is needed. Therefore, we minimize the regularized cost function*

$$f^M(\theta) = \frac{1}{N}\sum_{i=1}^{N} |g_\theta(z^{(i)}) - y^{(i)}|^2 + \mathcal{R}_\beta(\theta),$$

where $\mathcal{R}_\beta : \mathbb{R}^{M+1} \to \mathbb{R}$ *denotes a regularization function with parameter* $\beta > 0$*, in our case we choose Tikhonov Regularization leading to penalization through* $\mathcal{R}(\theta) := \beta(\theta_0^2 + \cdots + \theta_{M+1}^2)$ *and* $\beta = 0.01$*. The resulting approximation with and without regularization are shown in Figure 1.5.*

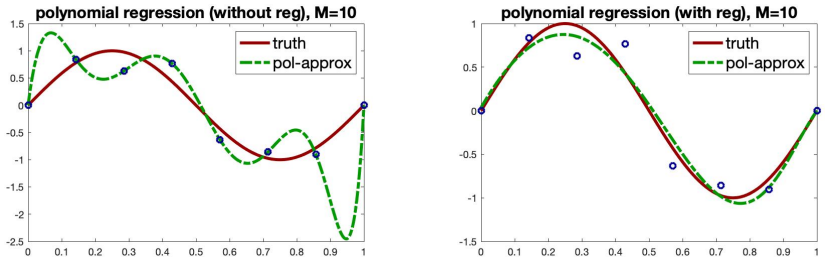

Figure 1.5: Resulting polynomial approximation of φ for minimizing the data misfit functional without (left) and with (right) regularization for perturbed data $y^{(i)} = \varphi(z^{(i)}) + \eta^{(i)}$, $i = 1, \dots, N$.

2

Unconstrained Optimization methods

In this chapter, we will provide a brief overview of basic concepts in unconstrained optimization. The concepts presented in this chapter are based on [1, 17, 15]. We will introduce fundamental definitions and properties around optimality conditions. Building upon this foundation, we will explore a series of gradient-based optimization methods to iteratively solve the problems. The convergence analysis of the studied schemes is presented for different classes of cost functions ((non-)convex, (non-)differentiable, (non-)smooth).
Let us consider the following problem formulation, which will be the central focus of this lecture.

Problem 2.0.1. Let $f : X \to \mathbb{R}$ be a (continuous) function with domain $X \subset \mathbb{R}^d$. For which value(s) $x \in X$ is the function evaluation $f(x)$ minimal?

We will refer to Problem 2.0.1 to *optimization* or *minimization problem* and will write shortly

$$\min_{x \in X} \; f(x)\,. \tag{2.1}$$

The corresponding function f is called *cost function* or sometimes also *objective function*.

Remark 2.0.2.

- In this lecture our special focus will lie in optimization problems without constraints, in which we assume $X = \mathbb{R}^d$. For the case $X \subsetneq \mathbb{R}^d$ we refer to (2.1) as optimization problem under the constrain $x \in X$, where we also write

$$\min_{x \in \mathbb{R}^d} \; f(x), \quad \text{s.t.} \quad x \in X,$$

  in order to highlight the constrain. Note that "s.t." stands for "subjected to".

- For simplicity we will consider Minimization problems, since each Maximization problem can be equivalently rewritten as Minimization problem:

$$\max_{x \in X} \; f(x) \quad \widehat{=} \quad \min_{x \in X} \; -f(x)\,.$$

- This lecture focuses on so-called continuous optimization problems, where the feasible set is infinite (uncountable). In case the feasible set is countable or finite, the optimization problem is called discrete.

- Typically, the cost function f is nonlinear, such that (2.1) is often called nonlinear program.

In the following, we introduce the notion of local and global solutions for (2.1).

Definition 2.0.3. Let $f : X \to \mathbb{R}$ for some $X \subset \mathbb{R}^d$. The point $x_* \in X$ is called

a) *local Minimum* of f over X, if there exists $\varepsilon > 0$ such that $f(x_*) \leq f(x)$ for all $x \in X$ with $\|x - x_*\| < \varepsilon$. If it even holds true that $f(x_*) < f(x)$ for all $x \in X \setminus \{x_*\}$ with $\|x - x_*\| < \varepsilon$, we call x_* *strict local Minimum*.

b) *global Minimum* of f over X, if $f(x_*) \leq f(x)$ for all $x \in X$. If it even holds true that $f(x_*) < f(x)$ for all $x \in X \setminus \{x_*\}$, we call x_* *strict global Minimum*.

Remark 2.0.4. A local Minimum $x_* \in X$ minimizes the cost function f only in a local neighborhood $\mathcal{B}_\varepsilon(x_*) = \{x \in \mathbb{R}^d \mid \|x - x_*\| < \varepsilon\}$, whereas a global Minimum minimizes the cost function over the whole domain/ feasible set X. We note that this definition includes constrained optimization problems. Furthermore, every global Minimum is also a local Minimum. The other way around does not hold.

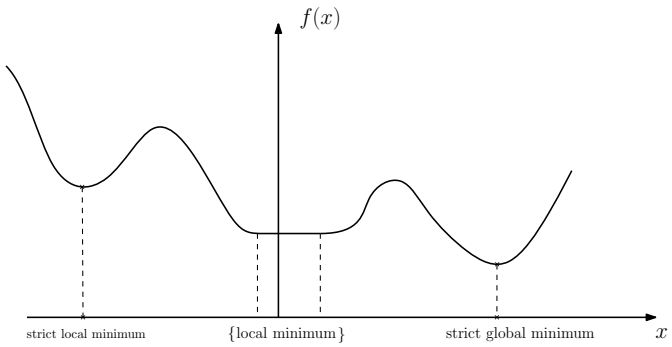

Figure 2.1: Illustration of different types of minima of the cost function f based on Figure 1.1.1 in [1].

2.1 Optimality conditions

We will now discuss necessary as well as sufficient conditions characterizing local and global minima. We start with the necessary optimality conditions of first order, which only needs differentiability of the cost function.

Theorem 2.1.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable over the open set $S \subset \mathbb{R}^d$ and suppose $x_* \in S$ is a local minimum of f. Then it holds true that $\nabla f(x_*) = 0$.*

Proof. Let $x_* \in S$ be a local minimum of f and consider the corresponding ball $\mathcal{B}_\varepsilon(x_*)$ such that $f(x_*) \leq f(x)$ for all $x \in \mathcal{B}_\varepsilon(x_*)$. Let $z \in \mathbb{R}^d$ be arbitrary but fixed, such that for $\bar{\alpha} > 0$ small enough $x_* + \alpha z \in \mathcal{B}_\varepsilon(x_*)$ and hence $f(x_* + \alpha z) \geq f(x_*)$ for all $\alpha \in [0, \bar{\alpha})$. We define $\alpha \mapsto g(\alpha) = f(x_* + \alpha z)$, $\alpha > 0$ and observe

$$\frac{\mathrm{d}g(0)}{d\alpha} = \lim_{\alpha \to 0} \frac{f(x_* + \alpha z) - f(x_*)}{\alpha} \geq 0.$$

On the other side by chain rule we also have

$$\frac{\mathrm{d}g(0)}{\mathrm{d}\alpha} = z^\top \nabla f(x_* + \alpha z).$$

In particular, this implies $0 \leq z^\top \nabla f(x_*)$. Since $z \in \mathbb{R}^d$ is arbitrary, with $y = -z \in \mathbb{R}^d$ we also obtain $0 \leq y^\top \nabla f(x_*) = -z^\top \nabla f(x_*)$. Therefore, for all $z \in \mathbb{R}^d$ we have $z^\top \nabla f(x_*) = 0$, which yields $\nabla f(x_*) = 0$. □

Under the additional assumption that f is twice continuously differentiable there is also a necessary optimality condition of second order.

Theorem 2.1.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable over the open set $S \subset \mathbb{R}^d$ and suppose $x_* \in S$ is a local minimum of f. Then it holds true that $\nabla^2 f(x_*)$ is positive semi-definite.*

Proof. Let again $x_* \in S$ be the local minimum of f with the corresponding ball $\mathcal{B}_\varepsilon(x_*)$ such that $f(x_*) \leq f(x)$ for all $x \in \mathcal{B}_\varepsilon(x_*)$ and consider an arbitrary $z \in \mathbb{R}^d$. With the Taylor expansion it holds

$$f(x_* + \alpha z) - f(x_*) = \alpha \underbrace{\nabla f(x_*)^\top}_{=0} z + \frac{\alpha^2}{2} z^\top \nabla^2 f(x_*) z + o(\alpha^2) = \frac{\alpha^2}{2} z^\top \nabla^2 f(x_*) z + o(\alpha^2)\,.$$

Let $\bar{\alpha} > 0$ be small enough such that $f(x_* + \alpha z) \geq f(x_*)$ for all $\alpha \in [0, \bar{\alpha})$. With the above equation we obtain

$$0 \leq \frac{f(x_* + \alpha z) - f(x_*)}{\alpha^2} = \frac{1}{2} z^\top \nabla^2 f(x_*) z + \frac{o(\alpha^2)}{\alpha^2} \to \frac{1}{2} z^\top \nabla^2 f(x_*) z$$

for $\alpha \to 0$. Since $z \in \mathbb{R}^d$ is chosen arbitrary, this leads to $z^\top \nabla^2 f(x_*) z \geq 0$ for all $z \in \mathbb{R}^d$ and the assertion follows. □

Remark 2.1.3. Theorem 2.1.1 and 2.1.2 characterize necessary, but not sufficient, conditions for optimality. Firstly, consider for example $f(x) = -x^2$ with stationary point $x_* = 0$, which is no

local minimum. Secondly, consider $f(x) = x_1^2 - x_2^4$, $x = (x_1, x_2)^\top \in \mathbb{R}^2$ with stationary point $x_* = (0,0)^\top$ and positive semi-definite Hessian $\nabla^2 f(x_*) = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}$, which is no local minimum.

In the following theorem, we formulate second order sufficient conditions for optimality.

Theorem 2.1.4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable over the open set $S \subset \mathbb{R}^d$. Let $x_* \in S$ with*

1. *$\nabla f(x_*) = 0$*
2. *$\nabla^2 f(x_*)$ (strictly) positive definite.*

Then x_* is a strict local minimum of f and there exist $\gamma > 0$, $\varepsilon > 0$ such that

$$f(x) \geq f(x_*) + \frac{\gamma}{2}\|x - x_*\|^2$$

for all $x \in \mathcal{B}_\varepsilon(x_*)$.

Before proving Theorem 2.1.4, we will need to prove the following auxiliary result.

Lemma 2.1.5. Let $A \in \mathbb{R}^{d\times d}$ be a symmetric matrix with real valued eigenvalues $\lambda_1 \leq \cdots \leq \lambda_d$ and corresponding eigenvectors $v_1, \dots, v_d$. Then it holds true that:

1. $\lambda_1\|z\|^2 \leq z^\top A z \leq \lambda_d \|z\|^2$ for all $z \in \mathbb{R}^d$.
2. The matrix A is (strict) positive definite if and only if all eigenvalues are (strictly) positive.

Proof. We start with the first assertion. Let $z \in \mathbb{R}^d$ and write

$$z = \sum_{i=1}^d \xi_i v_i$$

with coefficients $\xi_i \in \mathbb{R}$. Then we can write

$$z^\top A z = \sum_{i=1}^d \xi_i^2 \langle v_i, \underbrace{Av_i}_{=\lambda_i v_i}\rangle = \sum_{i=1}^d \lambda_i \xi_i^2 \|v_i\|^2 \begin{cases} \geq & \lambda_1 \sum_{i=1}^d \xi_i^2 \|v_i\|^2 = \lambda_i \|z\|^2, \\ \leq & \lambda_d \sum_{i=1}^d \xi_i^2 \|v_i\|^2 = \lambda_d \|z\|^2., \end{cases}$$

which finishes the proof of the first claim. For the second assertion we start with "$\Rightarrow$". Let (λ_i, v_i) be eigenvalue and corresponding eigenvector of the (strictly) positive definite matrix A. By definition of a positive definite matrix we have

$$0 \leq (<) v_i^\top A v_i = v_i^\top (\lambda_i v_i) = \lambda_i \|v_i\|^2.$$

Since $\|v_i\|^2 > 0$ for $v_i \neq 0$, we obtain $\lambda_i \geq (>)0$ for all $i = 1, \dots, d$. For the other way around "$\Leftarrow$", we assume that $0 \leq (<)\lambda_1 \leq \dots \leq \lambda_d$ are eigenvalues of A with corresponding eigenvectors $v_1 \dots, v_d$. Then it follows with the first assertion

$$z^\top A z \geq \lambda_1 \|z\|^2 \geq (>)0$$

for all $z \in \mathbb{R}^d$ with $z \neq 0$. □

We are now ready to prove Theorem 2.1.4.

Proof of Theorem 2.1.4. Let $\lambda > 0$ be the smallest eigenvalue of the positive definite matrix $\nabla^2 f(x_*)$. By Lemma 2.1.5 it holds true that

$$z^\top \nabla^2 f(x_*) z \geq \lambda \|z\|^2.$$

Application of Taylor's expansion around x_* together with $\nabla f(x_*) = 0$ yields

$$\begin{aligned} f(x_* + d) - f(x_*) &= \nabla f(x_*)^\top d + \frac{1}{2} d^\top \nabla^2 f(x_*) d + o(\|d\|^2) \\ &= \frac{1}{2} d^\top \nabla^2 f(x_*) d + o(\|d\|^2) \\ &\geq \frac{\lambda}{2} \|d\|^2 + o(\|d\|^2) \\ &= \left(\frac{\lambda}{2} + \frac{o(\|d\|^2)}{\|d\|^2} \right) \|d\|^2. \end{aligned}$$

Let $\varepsilon > 0$ be sufficiently small such that it holds $\frac{o(\|d\|^2)}{\|d\|^2} \in (-\frac{\lambda}{4}, \frac{\lambda}{4})$ for $\|d\|^2 < \varepsilon$. For $x \in \mathbb{R}^d$ with $\|x - x_*\| < \varepsilon$ it follows

$$f(x) \geq f(x_*) + \left(\frac{\lambda}{2} + \frac{o(\|x - x_*\|^2)}{\|x - x_*\|^2} \right) \|x - x_*\|^2 \geq f(x_*) + \underbrace{(\frac{\lambda}{2} - \frac{\lambda}{4})}_{=: \frac{\gamma}{2}} \|x - x_*\|^2 > f(x_*).$$

□

Under the additional assumption of convex cost function f, we can further characterize sufficient optimality conditions.

Proposition 2.1.6. Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and convex over the convex set $S \subset \mathbb{R}^d$. Then it holds true that:

1. Every local minimum of f over S is also a global minimum over S.
2. If f is even strictly convex, then there exists at most one global minimum.

3. Let S be an open set. Then the condition $\nabla f(x_*) = 0$ is a sufficient and necessary condition for $x_* \in S$ to be a global minimum of f over S.

Exercise 2.1.1. Prove Proposition 2.1.6.

2.2 Optimization methods based on descent directions

Recall that we are interested in solving $\min_{x\in\mathbb{R}^d} f(x)$. In practical applications, it is often challenging and, most of the time, even impossible to compute solutions of this problem analytically. Therefore, we will introduce iterative methods for solving the minimization task numerically. The focus will lie in so-called *descent methods* based on descent directions.

Definition 2.2.1. Let $f : \mathbb{R}^d \to \mathbb{R}$ be the cost function. We call a vector $d \in \mathbb{R}^d$ *descent direction* of f in $x \in \mathbb{R}^d$, if there exists $\bar{\alpha} > 0$ such that

$$f(x + \alpha d) < f(x) \quad \text{for all } \alpha \in (0, \bar{\alpha}].$$

Our aim is to construct an iterative scheme $x_0, x_1, x_2, \dots$ initialized with $x_0 \in \mathbb{R}^d$, such that $f(x_{k+1}) < f(x_k)$ for $k = 0, 1, 2, \dots$. The descent direction will be the key to construct this scheme.

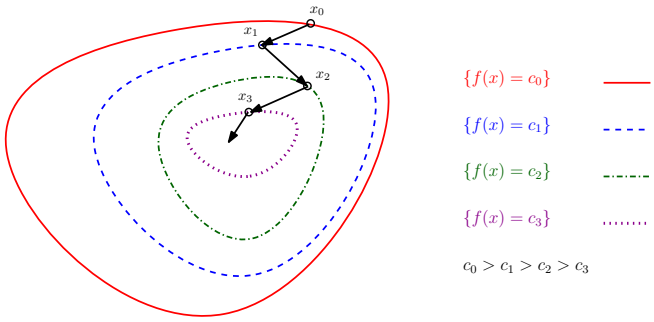

Figure 2.2: Illustration of descent methods based on Figure 1.2.1 in [1].

Lemma 2.2.2. Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable in $x \in \mathbb{R}^d$. Then the condition

$$\nabla f(x)^\top d < 0 \tag{2.2}$$

is sufficient for $d \in \mathbb{R}^d$ being a descent direction of f in x.

Proof. We fix $x \in \mathbb{R}^d$ and $d \in \mathbb{R}^d$ satisfying (2.2) and define $\varphi(\alpha) = f(x+\alpha d)$. By Taylor expansion it follows

$$\varphi(\alpha) = \varphi(0) + \alpha\varphi'(0) + o(\alpha).$$

Note that $\varphi(0) = f(x)$ and $\varphi'(0) = \nabla f(x)^\top d$ such that

$$\frac{\varphi(\alpha) - \varphi(0)}{\alpha} = \underbrace{\nabla f(x)^\top d}_{<0} + \underbrace{\frac{o(\alpha)}{\alpha}}_{\to 0,\ \alpha \to 0}.$$

This implies that there exists $\bar{\alpha} > 0$ such that

$$\frac{\varphi(\alpha) - \varphi(0)}{\alpha} < 0$$

for all $\alpha \in (0, \bar{\alpha}]$. □

Remark 2.2.3. The condition (2.2) is no necessary condition for $d \in \mathbb{R}^d$ being a descent direction. See Exercise 2.2.1 for more details.

Exercise 2.2.1. Let $x_* \in \mathbb{R}^d$ be a strict local maximum of $f : \mathbb{R}^d \to \mathbb{R}$. Prove that every $d \in \mathbb{R}^d$ is a descent direction of f in x_*.

Example 2.2.4. *The following two choices are classical examples of descent directions.*

- *Given $x \in \mathbb{R}^d$, the choice $d = -\nabla f(x)$ is a descent direction of f in x. This direction is also called steepest descent direction.*
- *Given $x \in \mathbb{R}^d$ and positive definite matrix $M \in \mathbb{R}^{d \times d}$, the choice $d = -M\nabla f(x)$ is a descent direction of f in x. We also call it preconditioned gradient-based descent direction.*

The resulting iterative descent method is formulated in the following algorithm.

Algorithm 1 Descent method

1: **Input:**
   - cost function $f : \mathbb{R}^d \to \mathbb{R}$
   - initial $x_0 \in \mathbb{R}^d$
2: set $k = 0$
3: **while** "convergence/stopping criterion not met" **do**
4: find a descent direction $d_k \in \mathbb{R}^d$ of f in x_k
5: determine a step size $\alpha_k > 0$ such that $f(x_k + \alpha_k d_k) < f(x_k)$
6: set $x_{k+1} = x_k + \alpha_k d_k$, $k \mapsto k + 1$
7: **end while**

Remark 2.2.5.
- The "convergence/stopping criterion" is of practical relevance. We will suppress this criterion in our theoretical analysis and study the various types of algorithms in its long time behavior for number of iterations $k \to \infty$.

- The values $\alpha_k > 0$ are called *step size* in iteration $k \in \mathbb{N}$. In the research area of machine learning these step sizes are called *learning rate*. In the literature of optimization the choice of the step size/ learning rate is often based on line-search. We will give more details on this in Section 2.2.2.

2.2.1 Examples of descent directions

We consider a row of examples for descent directions of the unified form

$$d_k = -D_k \nabla f(x_k), \tag{2.3}$$

where $D_k \in \mathbb{R}^{d\times d}$ is a positive definite matrix.

Definition 2.2.6. We define iterative schemes of the form

$$x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k), \quad x_0 \in \mathbb{R}^d,\ k \geq 0, \tag{2.4}$$

with $\alpha_k > 0$ and $D_k \in \mathbb{R}^{d\times d}$ positive definite as *gradient methods*.

Remark 2.2.7. The particular choice $d_k = -D_k \nabla f(x_k)$ describes a descent direction due to

$$\nabla f(x_k)^\top d_k = -\nabla f(x_k)^\top D_k \nabla f(x_k) < 0.$$

We consider the following examples of gradient methods.

a) *Method of steepest descent:* This scheme is described by the simplified choice $D_k = \mathrm{Id}$, i.e.

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \tag{2.5}$$

and is also known under the name *gradient descent method* (GD). The name "steepest descent" can be motivated by the normalized descent direction

$$d_k = -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}.$$

More details will follow later.

b) *Newton method:* We consider the quadratic approximation (second order Taylor approximation) of the cost function f. Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable, $x_k \in \mathbb{R}^d$ be the current iteration and approximate

$$f(x) \approx f_q(x) := f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{1}{2}(x_k - x)^\top \nabla^2 f(x_k)(x - x_k).$$

In order to find a stationary point $x_* \in \mathbb{R}^d$ of f_q we need to solve

$$\nabla f_q(x) = \nabla f(x_k) + \nabla^2 f(x_k)(x - x_k) = 0$$

which yields

$$x_* = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k) \tag{2.6}$$

assuming that $\nabla^2 f(x_k)$ is positive definite. This iteration corresponds to the *Newton iteration*. The Newton method is a gradient method with the particular choice $D_k = (\nabla^2 f(x_k))^{-1}$ provided that $\nabla^2 f(x_k)$ is regular, i.e.

$$x_{k+1} = x_k - \alpha_k (\nabla^2 f(x_k))^{-1} \nabla f(x_k).$$

The Newton method is a second-order gradient method, since derivatives of the second order are used to formulate the iterative scheme. However, our focus in this lecture will be on first-order gradient methods.

c) *Quasi-Newton method:* In case the Hessian $\nabla^2 f(x_k)$ is not invertible or only invertible with high afford, in practical applications one often applies numerical schemes in order to solves (2.6). This leads to the class of *quasi-Newton methods*.

2.2.2 Selection of the step size/ learning rate

After we have considered descent direction we will now take a look into the choice of the step size / learning rate α_k.

a) *Constant step size:* The most straightforward choice of step size is the constant step size $\alpha_k = s$ for all $k \geq 0$ and some $s > 0$ sufficiently small. However, convergence to a stationary point might be slow for too small s. In contrast, for a too large choice of s the resulting scheme might diverge. We will see examples for which we can derive specific upper bounds on s in order to ensure convergence of the scheme.

b) *Diminishing step size:* Another popular choice, in particular for stochastic optimization schemes, are diminishing step sizes $\alpha_k \to 0$ for $k \to \infty$. Again for too large choices of α_k the resulting scheme violates the monotonic descent along the iteration. Furthermore, it can easily happen that α_k degenerates too fast such α_k is too small in order to make progress towards a stationary point, although it might be still far away. A popular condition for choosing α_k is

$$\sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

c) *Armijo step size rule:* In the best possible way we would like to choose α_k such that the gain along the chosen descent direction is maximized, i.e. such that the cost function is minimized. One corresponding choice would be

$$\alpha_k \in \arg\min_{\alpha\in[0,s]} f(x_k + \alpha d_k)$$

for some pre-specified $s > 0$. In most scenarios it is impossible to solve this minimization task exactly, and algorithmic schemes are applied to solve the line-search. The line-search based on the Armijo rule is one of the most popular schemes. Here, the step size is successively decreased until it leads to a decrease in the evaluation of the cost function.The Armijo rule is given by the condition

$$f(x + \alpha d) \leq f(x) + \sigma\alpha\nabla f(x)^\top d, \tag{2.7}$$

where x denotes the current iteration, d is the chosen descent direction and $\sigma > 0$. Assume that $d \in \mathbb{R}^d$ is chosen such that $\nabla f(x)^\top d < 0$ (e.g. $d = -\nabla f(x)$), then condition (2.7) leads to an decrease of the cost function. Furthermore, there exists some $\alpha > 0$ satisfying condition (2.7). In order to choose a suitable step size, we can apply a so-called *backtracking* line search. Given a certain initial step size $\alpha^{(0)} = s_0 > 0$ we will reduce the step size sequentially until condition (2.7) is satisfied. The Armijo step size rule is summarized in Algorithm 2.

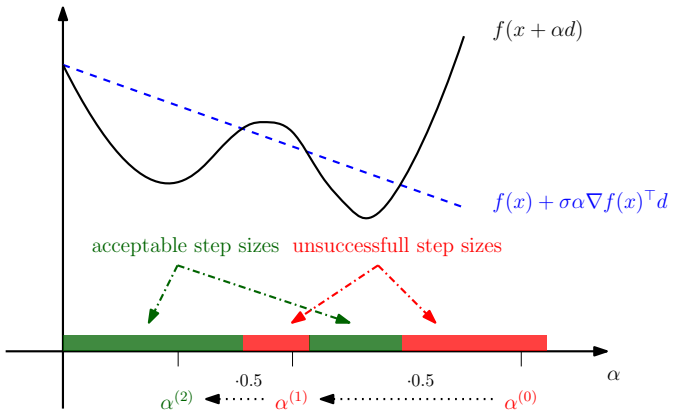

Figure 2.3: Illustration of the Armijo step size rule based on Figure 1.2.7 in [1]. In the green area condition (2.7) is satisfied such that α is accepted.

We refer interested readers to [1] for more information on alternative step size rules.

Exercise 2.2.2. Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and $(x_k)_{k\in\mathbb{N}}$ be defined by

$$x_{k+1} = x_k - \alpha_k\nabla f(x_k), \quad x_0 \in \mathbb{R}^d,$$

with diminishing step size $\alpha_k > 0$ such that $\sum_{k=1}^\infty \alpha_k = \infty$. Suppose that $(x_k)_{k\in\mathbb{N}}$ converges to

Algorithm 2 Armijo step size rule

Input:

- current iteration x_k and descent direction d_k
- parameter $\sigma \in (0,1)$, $\rho \in (0,1)$
- initial step size $s_0 > 0$

set $\ell = 0$, $\alpha^{(0)} = s_0$
while $f(x_k + \alpha^{(\ell)} d_k) > f(x) + \sigma \alpha^{(\ell)} \nabla f(x_k)^\top d_k$ **do**
    set $\alpha^{(\ell+1)} = \rho \cdot \alpha^{(\ell)}$
    set $\ell \mapsto \ell + 1$
end while
set $\alpha_k = \alpha^{(\ell)}$

some $x_* \in \mathbb{R}^d$. Prove that x_* is a stationary point of f, i.e. $\nabla f(x_*) = 0$.

2.2.3 Discussion about convergence behavior

Our aim in this course is to analyze the convergence of various optimization algorithms. The question is which behavior of convergence can we expect? In an optimal scenario we wish that the optimization scheme should converge from any initial state to a global minimum of the cost function. Unfortunately, typically this scenario is way too optimistic.

We consider a gradient descent scheme (2.5), where the state in each iteration moves into direction of steepest descent. This means the iteration always moves downhill independent of the global structure of the cost function. On the one side the iteration gets attracted from local minimums, but on the other side gets stuck in any stationary point. Without in the case of convex cost function we can only hope for convergence to stationary points. In general it is not clear if there exists an accumulation or even limit point of the sequence $(x_k)_{k\in\mathbb{N}}$ constructed by the gradient descent method (2.5).

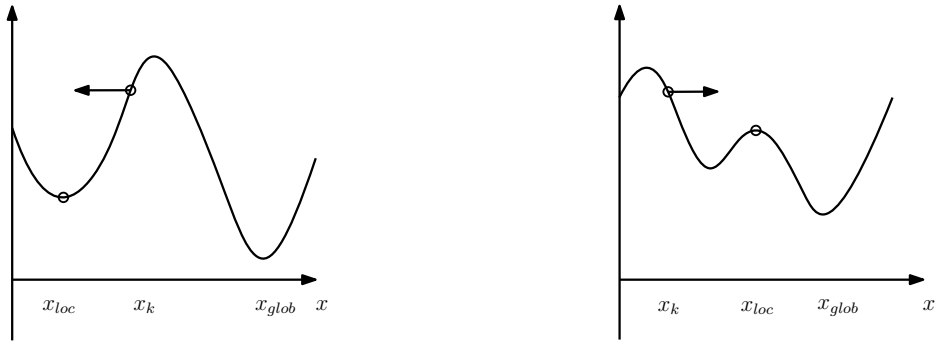

Figure 2.4: Illustration of possible terminations of the gradient descent method.

2.3 Gradient descent method

Before going into more details of the convergence analysis for gradient descent methods, we will motivate description of the method as *steepest descent method*. We have seen in Lemma 2.2.2 that $\nabla f(x)^\top d < 0$ characterizes the strength of the descent direction $d \in \mathbb{R}^d$ (keyword: Taylor expansion).
Let us consider $x \in \mathbb{R}^d$ and choose a normalized descent direction $d \in \mathbb{R}^d$ such that

$$\min_{d\in\mathbb{R}^d} \nabla f(x)^\top d, \quad \text{s.t.} \|d\| = 1. \tag{2.8}$$

We consider normalized descent directions, since it only determines the direction, whereas the length is scaled by the step size α_k after the descent direction has been chosen.
By the Cauchy-Schwarz inequality we firstly observe that for $\|d\| = 1$

$$0 \leq |\nabla f(x)^\top d| \leq \|\nabla f(x)\|\|d\| = \|\nabla f(x)\|.$$

Then it holds also true that

$$\nabla f(x)^\top d \geq -|\nabla f(x)^\top d| \geq -\|\nabla f(x)\|.$$

Since the choice $d_* = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$ leads to $\nabla f(x)^\top d = -\|\nabla f(x)\|$, d_* is a solution of (2.8). This means that the negative gradient $-\nabla f(x)$ is the steepest descent direction over all possible directions in the sense of minimizing the descent condition (2.2).
In order to do characterize limit points of the steepest descent scheme, we will need more assumptions on our underlying cost function f. One important property considered in this lecture course is *smoothness*. This property guarantees a descent of the objective function along the trajectories of the gradient descent method as long as the step size is sufficiently small.

Definition 2.3.1. We call a function $f : \mathbb{R}^d \to \mathbb{R}$ L-smooth for $L > 0$, if f is differentiable and the corresponding gradient ∇f is L-Lipschitz continuous, i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \text{for all } x, y \in \mathbb{R}^d$$

If f is twice continuously differentiable, then a sufficient condition of L-smoothness is given by $\sup_{x\in\mathbb{R}^d} \|\nabla^2 f(x)\|^2 \leq L$. Assuming that our cost function f is L-smooth allows us to apply the following descent Lemma.

Lemma 2.3.2 (Descent lemma)**.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and consider

$x, y \in \mathbb{R}^d$ with

$$\|\nabla f(x+ty) - \nabla f(x)\| \leq L\|ty\|, \tag{2.9}$$

for some $L > 0$ and all $t \in [0,1]$. Then it holds true that

$$f(x+y) \leq f(x) + y^\top \nabla f(x) + \frac{L}{2}\|y\|^2.$$

Proof. We define $\phi(t) = f(x+ty)$ and apply chain rule in order to derive

$$\phi'(t) = y^\top \nabla f(x+ty), \quad t \in [0,1].$$

By the fundamental theorem of calculus it follows

$$\begin{aligned} f(x+y) - f(x) = \phi(1) - \phi(0) &= \int_0^1 \phi'(t)\,\mathrm{d}t = \int_0^1 y^\top \nabla f(x+ty)\,\mathrm{d}t \\ &= \int_0^1 y^\top \nabla f(x)\,\mathrm{d}t + \int_0^1 y^\top (\nabla f(x+ty) - \nabla f(x))\,\mathrm{d}t \\ &\leq y^\top \nabla f(x) + \int_0^1 \|y\| \cdot \|\nabla f(x+ty) - \nabla f(x)\|\,\mathrm{d}t \\ &\leq y^\top \nabla f(x) + \|y\| \int_0^1 Lt \cdot \|y\|\,\mathrm{d}t \\ &= y^\top \nabla f(x) + \frac{L}{2}\|y\|^2, \end{aligned}$$

where we have applied Cauchy-Schwarz followed by the assumption (2.9). □

Following the descent lemma we obtain for L-smooth functions f the upper bound

$$f(y) \leq f(x) + \nabla f(x)^\top (y-x) + \frac{L}{2}\|x-y\|^2, \quad \text{for all } x, y \in \mathbb{R}^d. \tag{2.10}$$

In comparison for μ-strongly convex functions f we have a lower bound

$$f(y) \geq f(x) + \nabla f(x)^\top (y-x) + \frac{\mu}{2}\|x-y\|^2, \quad \text{for all } x, y \in \mathbb{R}^d. \tag{2.11}$$

If we put both together we obtain for L-smooth and μ-strongly convex functions the following characterization

$$\frac{\mu}{2}\|x-y\|^2 \leq f(y) - f(x) - \nabla f(x)^\top (y-x) \leq \frac{L}{2}\|x-y\|^2.$$

This motivates the definition of the Bregman divergence.

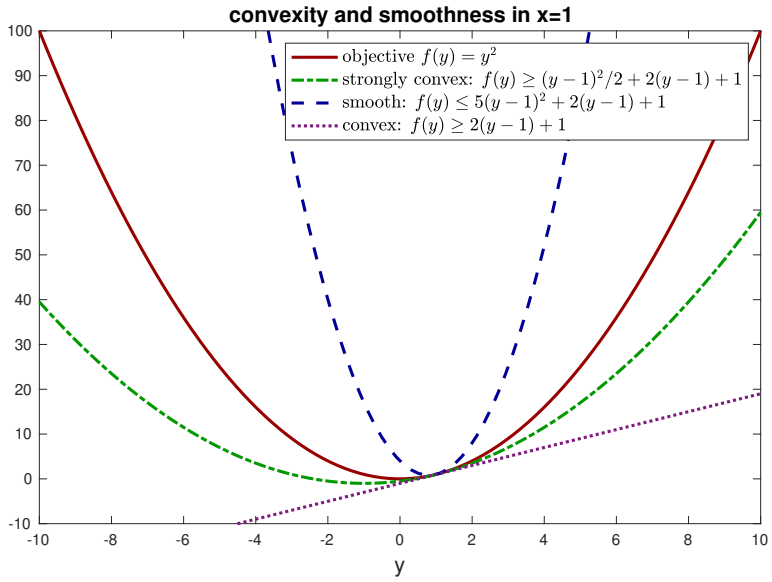

Figure 2.5: Illustration of L-smoothness and μ-strong convexity. We consider the function $f(x) = x^2$ such that f is L-smooth and μ-strongly convex with $1 := \mu < f''(x) = 2 < L =: 10$ for all $x \in \mathbb{R}$. This plot illustrates convexity and smoothness of f in $x = 1$ using the upper and lower bound (2.10) and (2.11) on $f(y)$, $y \in \mathbb{R}$.

Definition 2.3.3. Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable, then the Bregman divergence $D_f^{(B)} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is defined as

$$D_f^{(B)}(y, x) = f(y) - f(x) - \nabla f(x)^\top (y - x), \quad y, x \in \mathbb{R}^d.$$

If we assume L-smoothness and convexity of f (without μ-strong convexity), we are able to obtain an additional bound on the differences of the gradient evaluation.

Lemma 2.3.4. Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and L-smooth, then

$$\frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2 \leq D_f^{(B)}(y, x)$$

is satisfied for all $x, y \in \mathbb{R}^d$.

Remark 2.3.5. In general, for a convex function it only holds true that

$$D_f^{(B)}(y, x) = f(y) - f(x) - \nabla f(x)^\top (y - x) \geq 0.$$

With the additional assumption of L-smoothness we obtain the stronger characterization

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2.$$

Exercise 2.3.1. Prove Lemma 2.3.4.

In the next subsection, we will begin the convergence analysis of the gradient descent method in a non-convex setting.

2.3.1 Convergence for non-convex cost function

We start the convergence analysis of gradient descent scheme in the most general setting, where the cost function is assumed to be continuously differentiable without any other assumption. In this Subsection we follow very closely to Section 1.2.2. in [1].
As discussed in Section 2.2.3, we do not expect to prove convergence to a unique global minimum in this scenario. However, assuming that there exists an accumulation point of the sequence generated by gradient descent, we are able to characterize this point as stationary point.

Theorem 2.3.6 (GD with Armijo rule)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and $(x_k)_{k\in\mathbb{N}}$ be generated by*

$$x_{k+1} = x_k + \alpha_k d_k, \quad d_k = -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|},$$

where $\alpha_k > 0$ is chosen by the Armijo step size rule Algorithm 2. Then it holds true that every accumulation point $\bar{x} \in \mathbb{R}^d$ of the sequence $(x_k)_{k\in\mathbb{N}}$ is a stationary point of f, i.e. $\nabla f(\bar{x}) = 0$.

Proof. We will prove the assertion via contradiction. Firstly, suppose that $\bar{x} \in \mathbb{R}^d$ is an accumulating point of the sequence $(x_k)_{k\in\mathbb{N}}$ satisfying $\nabla f(\bar{x}) \neq 0$.
We consider the corresponding sub-sequence $(x_{k_n})_{n\in\mathbb{N}}$ converging to $\bar{x}$. Since the cost function f is continuous, we also have that $(f(x_{k_n}))_{n\in\mathbb{N}}$ converges to $f(\bar{x})$.
Moreover, through the Armijo step size rule it follows that the sequence $(f(x_k))_{k\in\mathbb{N}}$ is monotonically decreasing, i.e. $f(x_{k+1}) < f(x_k)$, and hence, the whole sequence $(f(x_k))_{k\in\mathbb{N}}$ converges to $f(\bar{x})$.
In particular, it follows that $(f(x_k))_{k\in\mathbb{N}}$ is a Cauchy-sequence and therefore

$$\lim_{k\to\infty} (f(x_{k+1}) - f(x_k)) = 0.$$

Due to the Armijo step size rule it holds true that

$$f(x_k) - f(x_{k+1}) \geq \sigma\alpha_k \|\nabla f(x_k)\|$$

implying that

$$\lim_{k\to\infty} \alpha_k \|\nabla f(x_k)\| = 0.$$

Since we have assumed that $\nabla f(\bar{x}) \neq 0$ it follows by continuity of ∇f that $\lim_{n\to\infty} \|\nabla f(x_{k_n})\| \neq 0$. Hence, it follows that $\lim_{n\to\infty} \alpha_{k_n} = 0$. By construction of the Armijo step size rule, we can write $\alpha_{k_n} = \rho^{\ell_n} \cdot s_0$, where $\ell_n \in \mathbb{N}$ denotes the first iteration such that condition (2.7) is satisfied. Since $\lim_{n\to\infty} \alpha_{k_n} = 0$, for $n \in \mathbb{N}$ large enough there exists $\ell_n > 0$ such that $f(x_{k_n} + \rho^{\ell_n - 1} s_0 d_{k_n}) > f(x_{k_n}) + \sigma \rho^{\ell_n - 1} s_0 \nabla f(x_{k_n})^\top d_{k_n}$, which means that condition (2.7) is at least once not satisfied during the application of Algorithm 2. Thus, we have

$$\frac{f(x_{k_n} + \rho^{\ell_n - 1} s_0 d_{k_n}) - f(x_{k_n})}{\rho^{\ell_n - 1} s_0} > \sigma \nabla f(x_{k_n})^\top d_{k_n}. \tag{2.12}$$

With the mean-value theorem there exists some $r_n \in [0, \rho^{\ell_n - 1} \cdot s_0]$ such that

$$\begin{aligned} f(x_{k_n} + \rho^{\ell_n - 1} s_0 d_{k_n}) - f(x_{k_n}) &= \nabla f(x_{k_n} + r_n d_{k_n})^\top (x_{k_n} + \rho^{\ell_n - 1} s_0 d_{k_n} - x_{k_n}) \\ &= \rho^{\ell_n - 1} s_0 \nabla f(x_{k_n} + r_n d_{k_n})^\top d_{k_n}, \end{aligned}$$

and with (2.12) it follows

$$\nabla f(x_{k_n} + r_n d_{k_n})^\top d_{k_n} > \sigma \nabla f(x_{k_n})^\top d_{k_n}. \tag{2.13}$$

Due to continuity of ∇f and the assumption that $\lim_{n\to\infty} x_{k_n} = \bar{x}$, we obtain

$$\lim_{n\to\infty} d_{k_n} = \lim_{n\to\infty} -\frac{\nabla f(x_{k_n})}{\|\nabla f(x_{k_n})\|} = -\frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|},$$

and since $\lim_{n\to\infty} \alpha_{k_n} = 0$ we obtain

$$\lim_{n\to\infty} r_n = \rho^{\ell_n - 1} s_0 = \frac{\rho^{\ell_n}}{\rho} s_0 = \frac{1}{\rho} \alpha_{k_n} = 0.$$

Both limits together imply that

$$\lim_{n\to\infty} \nabla f(x_{k_n} + r_n d_{k_n})^\top d_{k_n} = -\frac{\nabla f(\bar{x})^\top \nabla f(\bar{x})}{\|\nabla f(\bar{x})\|} = -\|\nabla f(\bar{x})\|.$$

and similarly,

$$\lim_{n\to\infty} \sigma \nabla f(x_{k_n})^\top d_{k_n} = -\sigma \|\nabla f(\bar{x})\|.$$

Finally, taking the limit $n \to \infty$ in equation (2.13) it follows

$$-\|\nabla f(\bar{x})\| \geq -\sigma \|\nabla f(\bar{x})\|,$$

which contradicts the assumption $\sigma \in (0, 1)$ for $\nabla f(\bar{x}) \neq 0$. Hence, we have proved $\nabla f(\bar{x}) = 0$. □

We continue with the convergence analysis of gradient descent for L-smooth cost functions f with-

out further assumptions such as convexity. The following Theorem then quantifies accumulation points of gradient descent with fixed but sufficiently small step size.

Theorem 2.3.7 (GD with constant step size)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and $(x_k)_{k\in\mathbb{N}}$ be generated by*

$$x_{k+1} = x_k - \bar{\alpha}\nabla f(x_K),$$

where $\bar{\alpha} \in [\varepsilon, \frac{2-\varepsilon}{L}]$ for some $\varepsilon > 0$ with $\varepsilon \leq \frac{2}{L+1}$. Then it holds true that every accumulation point $\bar{x} \in \mathbb{R}^d$ of the sequence $(x_k)_{k\in\mathbb{N}}$ is a stationary point of f, i.e. $\nabla f(\bar{x}) = 0$.

Proof. Since f is assumed to be L-smooth we can apply the descent Lemma 2.3.2 (with the choice $y \equiv -\bar{\alpha}\nabla f(x_k)$ and $x \equiv x_k$),

$$f(x_k - \bar{\alpha}\nabla f(x_k)) - f(x_k) \leq (-\bar{\alpha}\nabla f(x_k))^\top \nabla f(x_k) + \frac{L}{2}\|\bar{\alpha}\nabla f(x_k)\|^2 = \bar{\alpha}\|\nabla f(x_k)\|^2(\frac{\bar{\alpha}L}{2} - 1). \quad (2.14)$$

Due to our choice of $\bar{\alpha} \leq \frac{2-\varepsilon}{L}$ we can bound

$$\frac{\bar{\alpha}L}{2} - 1 \leq -\frac{\varepsilon}{2} < 0.$$

We reformulate the inequality (2.14) and obtain

$$f(x_k) - f(x_k - \bar{\alpha}\nabla f(x_k)) \geq \frac{\varepsilon}{2}\bar{\alpha}\|\nabla f(x_k)\|^2 \geq \frac{\varepsilon^2}{2}\|\nabla f(x_k)\|^2. \quad (2.15)$$

Similar to the proof of Theorem 2.3.6, we firstly assume that $(x_{k_n})_{n\in\mathbb{N}}$ is a sub-sequence with limit point $\bar{x} \in \mathbb{R}^d$ and $\nabla f(\bar{x}) \neq 0$. With (2.15) we can imply that the sequence $(f(x_k))_{k\in\mathbb{N}}$ is monotonically decreasing and therefore, it converges to $f(\bar{x})$ using continuity of f. In particular, we have

$$\lim_{k\to\infty}(f(x_{k+1}) - f(x_k)) = 0.$$

However, with (2.15) it then follows that

$$\lim_{k\to\infty}\frac{\varepsilon^2}{2}\|\nabla f(x_k)\|^2 = 0,$$

which is in contradiction to $\nabla f(\bar{x}) \neq 0$. □

We can formulate a similar Theorem for gradient descent with diminishing step size choice. In order to quantify accumulation points as stationary points, we need to force α_k to degenerate but not too fast.

Theorem 2.3.8 (GD with diminishing step size)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and $(x_k)_{k\in\mathbb{N}}$ be generated by*

$$x_{k+1} = x_k - \alpha_k \nabla f(x_K),$$

with

$$\lim_{k\to\infty} \alpha_k = 0 \quad \text{and} \quad \sum_{k=0}^{\infty} \alpha_k = \infty.$$

Then for the sequence $(f(x_k))_{k\in\mathbb{N}}$ it holds true that either

$$\lim_{k\to\infty} f(x_k) = -\infty \quad \text{or} \quad \lim_{k\to\infty} \nabla f(x_k) = 0.$$

Moreover, every accumulation point $\bar{x} \in \mathbb{R}^d$ of the sequence $(x_k)_{k\in\mathbb{N}}$ is a stationary point of f, i.e. $\nabla f(\bar{x}) = 0$.

Proof. Similar to the proof of Theorem 2.3.7 we apply Lemma 2.3.2 to derive

$$f(x_{k+1}) = f(x_k - \alpha_k \nabla f(x_k)) \leq f(x_k) - (\alpha_k - \frac{L\alpha_k^2}{2})\|\nabla f(x_k)\|^2 = f(x_k) - \alpha_k(1 - \frac{L\alpha_k}{2})\|\nabla f(x_k)\|^2.$$

Since we have assumed $\lim_{k\to\infty} \alpha_k = 0$, there exists $k_0 \geq 0$ such that

$$f(x_{k+1}) \leq f(x_k) - \alpha_k c \|\nabla f(x_k)\|^2$$

for some $c > 0$ and all $k \geq k_0$. Hence, the sequence $(f(x_k))_{k\geq k_0}$ is decreasing and it either holds true that $\lim_{k\to\infty} f(x_k) = -\infty$ or $\lim_{k\to\infty} f(x_k) = M$ for some $M < \infty$.
Suppose that we are in the case where $\lim_{k\to\infty} f(x_k) = M$. It follows

$$\sum_{k=k_0}^{K} \alpha_k c \|\nabla f(x_k)\|^2 \leq \sum_{k=k_0}^{K} \{f(x_k) - f(x_{k+1})\}$$

for $K > k_0$. Using that the rhs is a telescoping sum, we obtain

$$\sum_{k=k_0}^{K} \{f(x_k) - f(x_{k+1})\} = f(x_{k_0}) - f(x_K)$$

and for $K \to \infty$ we even imply

$$c \sum_{k=k_0}^{\infty} \alpha_k \|\nabla f(x_k)\|^2 \leq f(x_{k_0}) - M < \infty. \tag{2.16}$$

Since $\sum_{k=k_0}^{\infty} \alpha_k = \infty$, there can not exist any $\varepsilon > 0$ such that $\|\nabla f(x_k)\|^2 > \varepsilon$ for all $k \geq \hat{k} \geq 0$.

However, this only means that

$$\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0.$$

In order to prove $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$ we will use (2.16) to prove that $\limsup_{k\to\infty} \|\nabla f(x_k)\| = 0$. Suppose that $\limsup_{k\to\infty} \|\nabla f(x_k)\| \geq \varepsilon$ for some $\varepsilon > 0$ and consider two sub-sequences $(m_j)_{j\in\mathbb{N}}$, $(n_j)_{j\in\mathbb{N}}$, $n_j, m_j \in \mathbb{N}$, with $m_j < n_j < m_{j+1}$ such that

$$\frac{\varepsilon}{3} < \|\nabla f(x_k)\|, \quad \text{for } m_j \leq k < n_j$$

and

$$\|\nabla f(x_k)\| \leq \frac{\varepsilon}{3}, \quad \text{for } n_j \leq k < m_{j+1}.$$

Moreover, let $\bar{j} \in \mathbb{N}$ be sufficiently large such that

$$\sum_{k=m_{\bar{j}}}^{\infty} \alpha_k c \|\nabla f(x_k)\|^2 \leq \frac{\varepsilon^2}{9L}.$$

Using L-smoothness for $j \geq \bar{j}$ and $m_j \leq m \leq n_j - 1$ it holds true that

$$\begin{aligned}
\|\nabla f(x_{n_j}) - \nabla f(x_m)\| \leq \sum_{k=m}^{n_j-1} \|\nabla f(x_{k+1}) - \nabla f(x_k)\| &\leq L \sum_{k=m}^{n_j-1} \|x_{k+1} - x_k\| \\
&= \frac{3\varepsilon}{3\varepsilon} L \sum_{k=m}^{n_j-1} \alpha_k \|\nabla f(x_k)\| \\
&\leq L \frac{3}{\varepsilon} \sum_{k=m}^{n_j-1} \alpha_k \|\nabla f(x_k)\|^2 \\
&\leq L \frac{3}{\varepsilon} \frac{\varepsilon^2}{9L} = \frac{\varepsilon}{3},
\end{aligned}$$

where we have used that $\|\nabla f(x_k)\| > \frac{\varepsilon}{3}$ for $m_j \leq k \leq n_j - 1$. This implies that

$$\|\nabla f(x_m)\| \leq \|\nabla f(x_{n_j})\| + \|\nabla f(x_{n_j}) - \nabla f(x_m)\| \leq \|\nabla f(x_{n_j})\| + \frac{\varepsilon}{3} \leq \frac{2\varepsilon}{3}$$

and therefore $\|\nabla f(x_m)\| \leq \frac{2\varepsilon}{3}$ for all $m \geq m_{\bar{j}}$. This is in contradiction to $\limsup_{k\to\infty} \|\nabla f(x_k)\| \geq \varepsilon$ and we have proved that

$$\limsup_{k\to\infty} \|\nabla f(x_k)\| = \liminf_{k\to\infty} \|\nabla f(x_k)\| = \lim_{k\to\infty} \|\nabla f(x_k)\| = 0.$$

Finally, let $\bar{x} \in \mathbb{R}^d$ be an accumulating point of $(x_k)_{k\in\mathbb{N}}$. Since $(f(x_k))_{k\geq k_0}$ is decreasing, it follows by continuity that

$$\lim_{k\to\infty} f(x_k) = f(\bar{x}) < \infty$$

and then also

$$\nabla f(\bar{x}) = \lim_{k\to\infty} \nabla f(x_k) = 0.$$

□

Remark 2.3.9. Under the conditions of Theorem 2.3.8 there is no guarantee of decrease in the cost function $f(x_k)$ along the initial iterations $k \leq k_0$, which we have only verified for k_0 sufficiently large. The decrease can be forced under the additional assumption $\alpha_k \leq \frac{2-\varepsilon}{L}$ for $\varepsilon < \frac{2}{L+1}$ which was also used in Theorem 2.3.7 for an upper bound on the constant step size $\bar{\alpha}$.

Although in the previous theorems we have characterized accumulation points of the gradient descent method, we are not able to characterize the existence of limit points or even the convergence rate. However, under the additional assumption of lower bounded objective functions and following the proofs of Theorem 2.3.7 and Theorem 2.3.8, we can quantify the speed of degeneration of the gradients along the iterations of gradient descent.

Corollary 2.3.10. Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and $(x_k)_{k\in\mathbb{N}}$ be generated by

$$x_{k+1} = x_k - \alpha_k \nabla f(x_K), \quad \alpha_k > 0.$$

We define $g_N^* := \min_{k\in\mathbb{N}, 0\leq k\leq N} \|\nabla f(x_k)\|^2$ and $\bar{g}_N := \frac{1}{N}\sum_{k=0}^{N} \|\nabla f(x_k)\|^2$ for $N \in \mathbb{N}$.

- For the choice of a constant step size $\bar{\alpha} \in [\varepsilon, \frac{2-\varepsilon}{L}]$ with $\varepsilon < \frac{2}{L+1}$ it holds true that

$$g_N^*, \bar{g}_N \in \mathcal{O}(\frac{1}{N}).$$

- Consider a diminishing step size $\alpha_k = \frac{c}{\sqrt{k+1}}$, $k \geq 0$, for $c > 0$. Suppose that $\lim_{k\to\infty} f(x_k) = M \in (-\infty, \infty)$, then it holds true that

$$g_N^*, \bar{g}_N \in \mathcal{O}(\frac{1}{\sqrt{N}}).$$

Exercise 2.3.2. Prove Corollary 2.3.10.

We will close this section with the property of gradient descent which guarantees that the iteration gets captured by isolated local minimums. This means, that once the iteration moves into a sufficiently small neighborhood around an isolated local minimum, it will remain within this neighborhood and even converge to the corresponding local minimum.

Theorem 2.3.11 (Capture Theorem). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and the sequence $(x_k)_{k\in\mathbb{N}}$ be generated by gradient descent*

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),$$

with bounded step size $\alpha_k \leq s$, $s > 0$. Moreover, assume that the sequence $(f(x_k))_{k\in\mathbb{N}}$ is decreasing and every accumulation point of $(x_k)_{k\in\mathbb{N}}$ is a stationary point. Let $x_* \in \mathbb{R}^d$ be an isolated local minimum of f, i.e. there exists an open neighborhood $U \subset \mathbb{R}^d$ of x_* such that x_* is the only stationary point within the set U. Then there exists an open set $S \subset \mathbb{R}^d$ with $x_* \in S$ with the following property: If $x_{\bar{k}} \in S$ for some $\bar{k} \in \mathbb{N}$, then $x_k \in S$ for all $k \geq \bar{k}$ and in particular the sequence $(x_k)_{k\in\mathbb{N}}$ converges to x_*.

Proof. Since $x_* \in \mathbb{R}^d$ is assumed to be an isolated local minimum, we can find some $r > 0$ such that within a closed Ball $\bar{\mathcal{B}}_r(x_*) := \{x \in \mathbb{R}^d \mid \|x - x_*\| \leq r\}$ with radius r around x_* it holds true that

$$f(x_*) < f(x), \quad x \in \bar{\mathcal{B}}_r(x_*)$$

while there exists no stationary point $x' \in \bar{\mathcal{B}}_r(x_*) \setminus \{x_*\}$. It directly follows that

$$\min_{x\in\bar{\mathcal{B}}_r(x_*)\setminus\mathcal{B}_t(x_*)} f(x) > f(x_*)$$

for any open ball $\mathcal{B}_t(x_*) := \{x \in \mathbb{R}^d \mid \|x - x_*\| < t\}$ with radius $t \in (0, r]$ and we can the function

$$\Phi(t) = \min_{x\in\bar{\mathcal{B}}_r(x_*)\setminus\mathcal{B}_t(x_*)} f(x) - f(x_*),$$

which is decreasing for decreasing t. Since the gradient ∇f is assumed to be continuous and $\nabla f(x_*) = 0$, for a fixed but arbitrary $\varepsilon > 0$ we can find $q \in (0, \varepsilon]$ such that

$$\|x - x_*\| + s\|\nabla f(x)\| < \varepsilon \quad \text{for all } x \in \mathcal{B}_q(x_*). \tag{2.17}$$

We will use (2.17) in order to prove that the open set S defined by

$$S := \{x \in \mathbb{R}^d \mid \|x - x_*\| < \varepsilon,\ f(x) < f(x_*) + \Phi(q)\}$$

captures the iteration of gradient descent, i.e. if $x_k \in S$ then it also holds true that $x_{k+1} \in S$. Let us suppose $x_k \in S$, then for $\tau = \|x_k - x_*\|$ we have that $\Phi(\tau) \leq f(x_k) - f(x_*) < \Phi(q)$ by definition of S. Due to decreasing behavior of Φ we obtain $\|x_k - x_*\| = \tau < q$ and therefore $x_k \in \mathcal{B}_q(x_*)$. By (2.17) we obtain

$$\|x_k - x_*\| + s\|\nabla f(x_k)\| < \varepsilon.$$

Since the step size satisfies $\alpha_k \leq s$, applying triangular inequality we can imply that

$$\|x_{k+1} - x_*\| \leq \|x_k - x_*\| + \alpha_k \|\nabla f(x_k)\| \leq \|x_k - x_*\| + s\|\nabla f(x_k)\| < \varepsilon.$$

By assumption $(f(x_k))_{k\in\mathbb{N}}$ is decreasing such that

$$f(x_{k+1}) - f(x_*) < f(x_k) - f(x_*) < \Phi(q),$$

where we have used $x_k \in S$, and therefore, we obtain $x_{k+1} \in S$ as well. Once we find $\bar{k} \in \mathbb{N}$ such that $x_{\bar{k}} \in S$, we can imply $x_k \in S$ for all $k \geq \bar{k}$.
It is left to argue that in this case $\lim_{k\to\infty} x_k = x_*$. Consider the closure $\bar{S}$ of S which is a compact set. Then there exists at least one accumulation point $\bar{x}$ of $(x_k)_{k\in\mathbb{N}}$, which by assumption is a stationary point. By construction $\bar{S} \subset \mathcal{B}_r(x_*)$ such that $x_* \in \bar{S}$ is the unique stationary point of f in $\bar{S}$ implying that $\lim_{k\to\infty} x_k = x_*$. □

2.3.2 Convergence for convex and smooth cost function

In the following, we will study the convergence behavior of gradient descent under the additional assumption of (strong) convex cost functions. While in the previous section we have quantified possible accumulation points, we will now consider the description of convergence through some error function. Let $(x_k)_{k\in\mathbb{N}}$ be the sequence generated through some optimization scheme for a cost function f. We consider an error function $e : \mathbb{R}^d \to \mathbb{R}$ with the property $e(x) \geq 0$ for all $x \in \mathbb{R}^d$ and, $e(x_*) = 0$ for some $x_* \in \mathbb{R}^d$, e.g. $x_* \in \arg\min_{x\in\mathbb{R}} f(x)$ assuming it exists. Typical examples include

$$e(x) = \|x - x_*\| \quad \text{or} \quad e(x) = |f(x) - f(x_*)|.$$

We define the following type of convergence behavior.

Definition 2.3.12. We say that the sequence of errors $(e(x_k))_{k\in\mathbb{N}}$ converges linearly, if there exists $c \in (0, 1)$ such that

$$e(x_{k+1}) \leq ce(x_k)$$

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

Since the focus of this lecture course lies in first order methods, we do not expect faster convergence than linear such as super-linear or quadratic convergence. We will now study the convergence of gradient descent for general convex and L-smooth cost functions. Since for general convex functions there is no guarantee for existence of a unique global minimum, we do only expect convergence of the error function $e(x_k) = f(x_k) - f(x_*)$ for some global minimum $x_* \in \mathbb{R}^d$. The convergence is slower than linear, sometimes also referred to sub-linear convergence.

Theorem 2.3.13 (GD for convex and smooth cost function)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and L-smooth, and let $(x_k)_{k\in\mathbb{N}}$ be generated by*

$$x_{k+1} = x_k - \bar{\alpha}\nabla f(x_k)$$

with $\bar{\alpha} \leq \frac{1}{L}$. Moreover, we assume that the set of all global minimums of f is non-empty. Then the sequence $(x_k)_{k\in\mathbb{N}}$ converges in the sense that

$$e(x_k) := f(x_k) - f_* \leq \frac{c}{k}, \quad k \geq 1$$

for some constant $c > 0$ and $f_* = \min_{x\in\mathbb{R}^d} f(x)$.

Proof. We again apply the descent Lemma 2.3.2 (with $t = 1$, $y = (x_{k+1} - x_k)$ and $x = x_k$) to derive

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x)^\top (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2.$$

With $x_{k+1} - x_k = -\bar{\alpha}\nabla f(x_k)$ we obtain

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \bar{\alpha}\|\nabla f(x_k)\|^2 + \frac{L}{2}\bar{\alpha}^2\|\nabla f(x_k)\|^2 \\ &= f(x_k) + (\frac{L}{2}\bar{\alpha} - 1)\bar{\alpha}\|\nabla f(x_k)\|^2. \end{aligned}$$

Since $\bar{\alpha} \leq \frac{1}{L}$, we have $(\frac{L}{2}\bar{\alpha} - 1) \leq -\frac{1}{2} < 0$ and therefore, the sequence $(f(x_k))_{k\in\mathbb{N}}$ is decreasing. Now, let $x_* \in \mathbb{R}^d$ be some global minimum of f such that due to convexity it holds true that

$$f(x_k) + (x_* - x_k)^\top \nabla f(x_k) \leq f(x_*).$$

We plug this in into the above inequality to imply

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \bar{\alpha}(\frac{L}{2}\bar{\alpha} - 1)\|\nabla f(x_k)\|^2 \\ &\leq f(x_*) - \frac{\bar{\alpha}}{\bar{\alpha}}(x_* - x_k)^\top \nabla f(x_k) + \bar{\alpha}(\frac{L}{2}\bar{\alpha} - 1)\|\nabla f(x_k)\|^2 \\ &= f(x_*) + \frac{1}{\bar{\alpha}}\left\{\frac{1}{2}\|x_* - x_k\|^2 + \frac{\bar{\alpha}^2}{2}\|\nabla f(x_k)\|^2 - \frac{1}{2}\|(x_* - x_k) + \bar{\alpha}\nabla f(x_k)\|^2\right\} \\ &\quad + \bar{\alpha}(\frac{L}{2}\bar{\alpha} - 1)\|\nabla f(x_k)\|^2, \end{aligned}$$

where we have used $-a^\top b = \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2 - \frac{1}{2}\|a+b\|^2$ for $a, b \in \mathbb{R}^d$. Rearranging the rhs leads to

$$\begin{aligned} f(x_{k+1}) &\leq f(x_*) + \frac{1}{2\bar{\alpha}}\left(\|x_* - x_k\|^2 - \|x_* - x_{k+1}\|^2\right) + \bar{\alpha}(\frac{L}{2}\bar{\alpha} - \frac{1}{2})\|\nabla f(x_k)\|^2 \\ &\leq f(x_*) + \frac{1}{2\bar{\alpha}}\left(\|x_* - x_k\|^2 - \|x_* - x_{k+1}\|^2\right), \end{aligned}$$

where we have used again that $\bar{\alpha} \leq \frac{1}{L}$. Taking the sum over all iterations gives

$$\sum_{k=0}^{N}\{f(x_{k+1}) - f(x_*)\} \leq \frac{1}{2\bar{\alpha}}\sum_{k=0}^{N}\{\|x_* - x_k\|^2 - \|x_* - x_{k+1}\|^2\} \leq \frac{1}{2\bar{\alpha}}\{\|x_* - x_0\|^2 - \|x_* - x_{N+1}\|^2\},$$

where we have applied a telescoping sum. With the decrease of $(f(x_k))_{k\in\mathbb{N}}$ it holds true that

$$\sum_{k=0}^{N} f(x_{k+1}) \geq (N+1) f(x_{N+1}),$$

and therefore, the assertion follows with

$$f(x_{N+1}) - f(x_*) \leq \frac{1}{N+1}\frac{1}{2\bar{\alpha}}\|x_* - x_0\|^2 =: \frac{c}{N+1}.$$

□

2.3.3 Convergence for strongly convex and smooth cost function

Under the additional assumption that the cost function f is μ-strongly convex for some $\mu > 0$, we can even prove linear convergence of gradient descent with sufficiently small constant step size. The convergence holds even for the error function $e(x_k) = \|x_k - x_*\|$, where $x_* \in \mathbb{R}^d$ is the unique global minimum of f.

Theorem 2.3.14 (GD for strongly convex and smooth cost function)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex and L-smooth, and let $(x_k)_{k\in\mathbb{N}}$ be generated by*

$$x_{k+1} = x_k - \bar{\alpha}\nabla f(x_k)$$

with $\bar{\alpha} \leq \frac{1}{L}$. Then the sequence $(x_k)_{k\in\mathbb{N}}$ converges linearly in the sense that there exists $c \in (0,1)$ such that

$$e(x_k) := \|x_k - x_*\| \leq c^k\|x_0 - x_*\|, \quad k \in \mathbb{N}$$

where $x_* \in \mathbb{R}^d$ is the unique global minimum of f with $f(x_*) = \min_{x\in\mathbb{R}^d} f(x)$.

Proof. Let $x_* \in \mathbb{R}^d$ be the unique global minimum of f with $\nabla f(x_*) = 0$. Since f is assumed to

be μ-strongly convex, by Definition A.1.10 it holds true that

$$\frac{\mu}{2}\|x_{k+1}-x_*\|^2 = \nabla f(x_*)^\top (x_{k+1}-x_*) + \frac{\mu}{2}\|x_{k+1}-x_*\|^2 \leq f(x_{k+1}) - f(x_*).$$

On the other side in the previous proof of Theorem 2.3.13 we have derived that

$$f(x_{k+1}) - f(x_*) \leq \frac{1}{2\bar{\alpha}}\{\|x_k - x_*\|^2 - \|x_{k+1}-x_*\|^2\}$$

and together we obtain

$$(\frac{\mu}{2} + \frac{1}{2\bar{\alpha}})\|x_{k+1}-x_*\|^2 \leq \frac{1}{2\bar{\alpha}}\|x_k - x_*\|^2.$$

The assertion follows via induction using the inequality

$$\|x_{k+1}-x_*\|^2 \leq \frac{\frac{1}{\bar{\alpha}}}{\mu + \frac{1}{\bar{\alpha}}}\|x_k - x_*\|^2 = \frac{1}{1+\mu\bar{\alpha}}\|x_k - x_*\|^2 =: c^2\|x_k - x_*\|^2,$$

where $c = \sqrt{\frac{1}{1+\mu\bar{\alpha}}} \in (0,1)$. □

Remark 2.3.15. For the choice $\bar{\alpha} = \frac{1}{L}$ the upper bound of gradient descent is given by

$$\|x_k - x_*\| \leq \left(\sqrt{\frac{L}{L+\mu}}\right)^k \|x_0 - x_*\|.$$

This means that the speed of convergence is determined through the ratio of smoothness L and strong convexity μ:

$$c = \sqrt{\frac{1}{1+\frac{\mu}{L}}},$$

which decreases for decreasing L and increasing μ.

The proof of the previous Theorem 2.3.14 for convergence under strong convexity builds directly up on the proof of Theorem 2.3.13 under convexity. However, we can even improve the rate of convergence if we do not go the direct way from convex to strongly convex. Therefore, we consider the following convergence result.

Theorem 2.3.16. *Assume that the same conditions as in Theorem 2.3.14 are satisfied, and assume additionally that* $\bar{\alpha} \leq \min(\frac{2}{\mu+L}, \frac{\mu+L}{2\mu L})$. *Then the sequence* $(x_k)_{k\in\mathbb{N}}$ *generated by*

$$x_{k+1} = x_k - \bar{\alpha}\nabla f(x_k)$$

converges linearly in the sense that there exists $c \in (0,1)$ such that

$$e(x_k) := \|x_k - x_*\| \leq c^k \|x_0 - x_*\|, \quad k \in \mathbb{N}$$

where $x_* \in \mathbb{R}^d$ is the unique global minimum of f with $f(x_*) = \min_{x \in \mathbb{R}^d} f(x)$.

Proof. Let $x_* \in \mathbb{R}^d$ be the unique global minimizer of f, i.e. $\nabla f(x_*) = 0$.For $k \in \mathbb{N}$ and $x_k \in \mathbb{R}^d$ it holds true that

$$\|x_{k+1} - x_*\|^2 = \|x_k - \bar{\alpha}\nabla f(x_k) - x_*\|^2 = \|x_k - x_*\|^2 - 2\langle x_k - x_*, \bar{\alpha}\nabla f(x_k)\rangle + \bar{\alpha}^2\|\nabla f(x_k)\|^2.$$

We will make use of the inequality

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{\mu L}{\mu + L}\|x - y\|^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|^2$$

for any $x, y \in \mathbb{R}^d$, which is left as exercise in Lemma 2.3.17. It then follows that

$$\langle \nabla f(x_k), x_k - x_*\rangle \geq \frac{\mu L}{\mu + L}\|x_k - x_*\|^2 + \frac{1}{\mu + L}\|\nabla f(x_k)\|^2,$$

since $\nabla f(x_*) = 0$. We obtain the bound

$$\begin{aligned}
\|x_{k+1} - x_k\|^2 &\leq \|x_k - x_*\|^2 - 2\bar{\alpha}\left(\frac{\mu L}{\mu + L}\|x_k - x_*\|^2 + \frac{1}{\mu + L}\|\nabla f(x_k)\|^2\right) + \bar{\alpha}^2\|\nabla f(x_k)\|^2 \\
&= \left(1 - 2\bar{\alpha}\frac{\mu L}{\mu + L}\right)\|x_k - x_*\|^2 + \bar{\alpha}\left(\bar{\alpha} - \frac{2}{\mu + L}\right)\|\nabla f(x_k)\|^2 \\
&\leq \left(1 - 2\bar{\alpha}\frac{\mu L}{\mu + L}\right)\|x_k - x_*\|^2,
\end{aligned}$$

where we have used that $(\bar{\alpha} - \frac{2}{\mu+L}) \leq 0$. Since we have assumed that $\bar{\alpha} \leq \frac{\mu+L}{2\mu L}$, we finally obtain linear convergence with $c = \sqrt{1 - 2\bar{\alpha}\frac{\mu L}{\mu+L}} \in (0,1)$ in the sense that

$$\|x_k - x_*\| \leq c^k \|x_0 - x_*\|.$$

□

In the previous proof we have used the following bound for smooth and strong convex cost functions.

Lemma 2.3.17. Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and μ-strongly convex. Then it holds true that

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x - y\|^2$$

and

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|^2.$$

Exercise 2.3.3. Prove Lemma 2.3.17.

Remark 2.3.18. The motivation behind the alternative convergence result for gradient descent under μ-strong convexity and L-smoothness is the following. The optimal convergence rate is in the sense of maximizing $c^2(\bar{\alpha}) = (1 - 2\bar{\alpha}\frac{\mu L}{\mu+L}) \in (0,1)$ is obtained for the choice $\bar{\alpha} = \frac{2}{\mu+L}$. The upper bound of the error of gradient descent is then given by

$$\|x_k - x_*\| \leq c^k \|x_0 - x_*\|$$

with

$$c = \sqrt{\frac{(\mu + L)^2}{(\mu + L)^2} - \frac{4\mu L}{(\mu + L)^2}} = \frac{L - \mu}{L + \mu} = \frac{\kappa - 1}{\kappa + 1},$$

where $\kappa := \frac{L}{\mu}$ denotes the ratio between smoothness and strong convexity. We sometimes also refer to κ as the condition number of f. Usually, we have $L \geq \mu$, such that $\kappa \in (0,1)$. Finally, we can rewrite c through

$$c = \frac{\kappa - 1}{\kappa + 1} = 1 - \frac{2}{\kappa + 1},$$

which again decreases for decreasing L and increasing μ.

2.3.4 Convergence under PL-condition and smooth cost function

In this section, we will consider additional properties on the cost function f under which gradient descent converge with given rate. We will assume that the cost function is L-smooth and consider two settings, where the function evaluation of f satisfies a regularity condition related to its gradient norm. In [12] the authors consider a linear convergence analysis under the so-called Polyak-Łojasiewicz (PL) condition.

Theorem 2.3.19. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and satisfies the PL condition*

$$\|\nabla f(x)\|^2 \geq 2r(f(x) - f_*) \tag{2.18}$$

for some $r \in (0, L)$ and all $x \in \mathbb{R}^d$ with $f_* = \min_{x \in \mathbb{R}^d} f(x) > -\infty$. Then the sequence $(x_k)_{k \in \mathbb{N}}$ generated by

$$x_{k+1} = x_k - \bar{\alpha} \nabla f(x_k)$$

with $\bar{\alpha} = \frac{1}{L}$ converges linearly in the sense that there exists $c = 1 - \frac{r}{L} \in (0,1)$ such that

$$e(x_k) := f(x_k) - f_* \leq c^k (f(x_0) - f_*).$$

Exercise 2.3.4. 1. Prove Theorem 2.3.19.

2. Prove that μ-strong convexity and L-smoothness imply the PL condition (2.18).
3. Use a graphing calculator to find r such that $f(x) = x^2 + 3\sin^2(x)$ satisfies the PL condition (2.18) (argue why $x \to \infty$ is not a problem) and prove that f is not convex.

We can even assume a slightly weaker condition than PL (2.18) and remain convergence of gradient descent. However, the weaker condition is not enough to guarantee linear convergence.

Theorem 2.3.20. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and satisfies the "weak" PL condition*

$$\|\nabla f(x)\| \geq 2r(f(x) - f_*) \tag{2.19}$$

for some $r \in (0, L)$ and all $x \in \mathbb{R}^d$ with $f_* = \min_{x \in \mathbb{R}^d} f(x) > -\infty$. Then the sequence $(x_k)_{k \in \mathbb{N}}$ generated by

$$x_{k+1} = x_k - \bar{\alpha} \nabla f(x_k)$$

with $\bar{\alpha} = \frac{1}{L}$ converges with

$$e(x_k) := f(x_k) - f_* \leq \frac{L}{2r^2(k+1)}.$$

Exercise 2.3.5. Prove Theorem 2.3.20.

2.3.5 Convergence for non-smooth and convex cost function

As last class of cost functions for which we will study the convergence behavior of gradient descent, we will consider non-differentiable cost functions. We have to reformulate the scheme of gradient descent in a way such that we are avoiding the computation of the gradient ∇f. In the following section, we will formulate the so-called *sub-gradient descent method*. We begin with the definition of *sub-gradients* and the corresponding *sub-differential*.

Definition 2.3.21. We call $g_x \in \mathbb{R}^d$ a *sub-gradient* of $f : \mathbb{R}^d \to \mathbb{R}$ in $x \in \mathbb{R}^d$ if for all $y \in \mathbb{R}^d$ it holds true that

$$f(y) \geq f(x) + g_x^\top (y - x). \tag{2.20}$$

We call the set of all sub-gradients of f in x *sub-differential* of f denoted by $\partial f(x)$.

Note that (2.20) is closely related to convexity which can be formulated in case of differentiable functions as condition

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x).$$

This suggests that there is a close connection between sub-gradients and convexity as also stated in the next proposition.

Proposition 2.3.22. Let $C \subset \mathbb{R}^d$ be a convex set and $f : \mathbb{R}^d \to \mathbb{R}$.

1. If $\partial f(x) \neq \emptyset$ for all $x \in C$, then f is convex over C.
2. If f is convex, then $\partial f(x) \neq \emptyset$ for all x in the interior of C.

Proof. We will only prove the first assertion, for the second one we refer to Proposition 2.5 in [15]. Suppose that $\partial f(x) \neq \emptyset$ and let $z_\lambda = \lambda x + (1-\lambda)y \in C$ for $\lambda \in (0,1)$ and $x, y \in C$. Moreover, consider an arbitrary sub-gradient $g_{z_\lambda} \in \partial f(z_\lambda)$, then by definition of the sub-gradient it holds true that

$$f(y) \geq f(z_\lambda) + g_{z_\lambda}^\top (y - z_\lambda) = f(\lambda x + (1-\lambda)y) + \lambda g_{z_\lambda}^\top (y - x)$$

and similarly

$$f(x) \geq f(z_\lambda) + g_{z_\lambda}^\top (x - z_\lambda) = f(\lambda x + (1-\lambda)y) + (1-\lambda) g_{z_\lambda}^\top (x - y) -$$

We combine both inequalities to obtain

$$\begin{aligned}(1-\lambda)f(y) + \lambda f(x) &\geq (1-\lambda) f(\lambda x + (1-\lambda)y) + (1-\lambda)\lambda g_{z_\lambda}^\top (y - x) \\ &\quad + \lambda f(\lambda x + (1-\lambda)y) + \lambda(1-\lambda) g_{z_\lambda}^\top (x - y) \\ &= f((1-\lambda)y + \lambda x)\end{aligned}$$

which proves convexity of f over C. □

Example 2.3.23. *Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be a family of convex and differentiable functions $i = 1, \dots, N$ for some $N \in \mathbb{N}$. We define $F(x) = \max_{i=1,\dots,N} f_i(x)$ and for given $x \in \mathbb{R}^d$ we consider $j \in \arg\max_{i=1,\dots,N} f_i(x)$. Then we can compute a sub-gradient of F in x through $\nabla f_j(x)$, i.e. it holds true that $\nabla f_j(x) \in \partial F(x)$. To prove this, we observe that by convexity we have*

$$f_j(y) \geq f_j(x) + \nabla f_j(x)^\top (y - x)$$

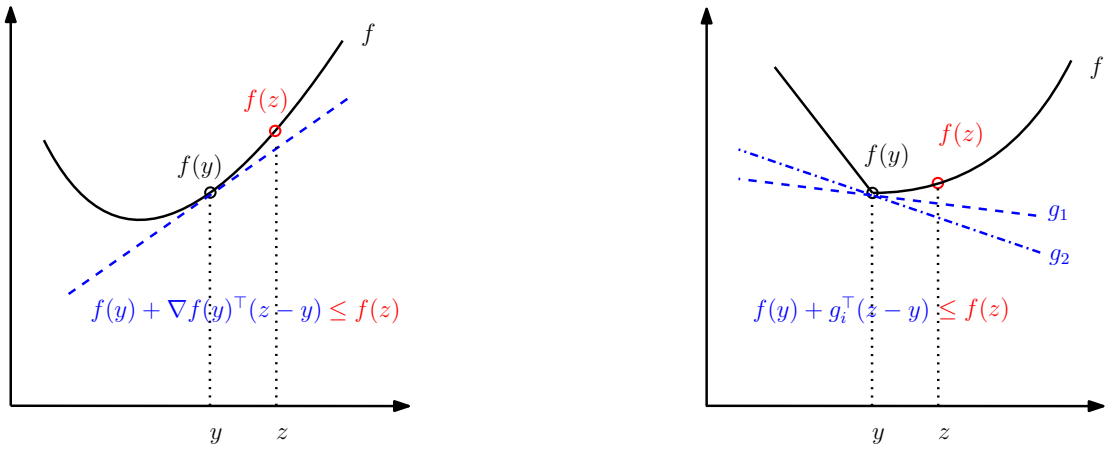

Figure 2.6: Illustration of sub-gradients for differentiable convex (left) and non-differentiable convex (right) functions. For continuously differentiable and convex f, the sub-gradient is unique, i.e. $\partial f(x) = \{\nabla f(x)\}$.

for all $y \in \mathbb{R}^d$. *This implies*

$$F(y) \geq f_j(y) \geq f_j(x) + \nabla f_j(x)^\top (y - x) = F(x) + \nabla f_j(x)^\top (y - x),$$

where we have used that $F(x) = f_j(x)$ *and* $F(y) \geq f_j(y)$ *for* $y \neq x$. *This proves that* $\nabla f_j(x) \in \partial F(x)$.

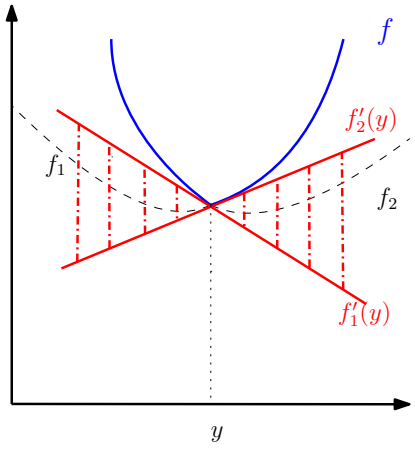

Figure 2.7: Illustration of Example 2.3.23. We consider $N = 2$ and define $f(x) = \max(f_1(x), f_2(x))$ for two convex and differentiable functions f_1, f_2. In the above Situation, we find unique sub-gradients for $f_1(x) > f_2(x)$ given by $\partial f(x) = \{f_1'(x)\}$, and similarly for $f_1(x) < f_2(x)$ given by $\partial f(x) = \{f_2'(x)\}$. In the case of $f_1(x) = f_2(x)$, the sub-differential is given by $\partial f(x) = [f_2'(x), f_1'(x)]$.

There are similar rules for the computation of sub-gradients, which are left as an exercise:

Exercise 2.3.6. Let $f, f_1, f_2 : \mathbb{R}^d \to \mathbb{R}$ be a convex functions. Then the following holds true:

- Prove that $\partial f(x)$ is a convex set for all $x \in \mathbb{R}^d$.
- Prove for $a > 0$ that $\partial(af)(x) = a\partial f(x)$.
- Prove that $\partial f_1(x) + \partial f_2(x) \subset \partial(f_1 + f_2)(x)$ for any $x \in \mathbb{R}^d$.
- Let $h(x) = f(Ax + b)$ for $A \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^d$. Prove that $A^\top \partial f(Ax + b) \subset \partial h(x)$. Prove equality for invertible A.

In case of continuously differentiable functions, the sub-gradient is unique and corresponds to the gradient.

Proposition 2.3.24. Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and convex in $x \in \mathbb{R}^d$. Then the sub-differential is a one-point set $\partial f(x) = \{\nabla f(x)\}$.

Proof. Firstly, it is obvious to see that $\nabla f(x) \in \partial f(x)$, since f is convex and it holds

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x)$$

for all $y \in \mathbb{R}^d$. Let us consider any $g_x \in \partial f(x)$. We will prove that it necessarily follows that $g_x = \nabla f(x)$. Let $y = x + \lambda z$ for $\lambda > 0$, such that

$$f(x + \lambda z) \geq f(x) + g_x^\top (\lambda z)$$

or rewritten

$$\frac{f(x + \lambda z) - f(x)}{\lambda} \geq g_x^\top z.$$

Since f is assumed to be continuously differentiable, we obtain taking the limit $\lambda \to 0$

$$\lim_{\lambda\to 0} \frac{f(x + \lambda z) - f(x)}{\lambda} = \nabla f(x)^\top z \geq g_x^\top z,$$

which implies that

$$(\nabla f(x) - g_x)^\top z \geq 0.$$

Since $z \in \mathbb{R}^d$ is arbitrary, we can choose $z = -(\nabla f(x) - g_x)$ in order to prove that

$$-(\nabla f(x) - g_x)^\top (\nabla f(x) - g_x) = -\|\nabla f(x) - g_x\|^2 \geq 0$$

which proves that $\nabla f(x) = g_x$. □

We now formulate an additional optimality condition for non-differentiable but convex cost functions, which can be characterized through the sub-differential.

Proposition 2.3.25. Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and continuous. Then $x_* \in \mathbb{R}^d$ is a global minimum of f if and only if $0 \in \partial f(x_*)$.

Proof. We start with $x_* \in \mathbb{R}^d$ being a global minimum of f, i.e. for all $y \in \mathbb{R}^d$ we have $f(x_*) \leq f(y)$. Therefore, we directly obtain $0 \in \partial f(x)$, since

$$f(y) \geq f(x_*) = f(x_*) + 0^\top (y - x_*).$$

Now let $0 \in \partial f(x_*)$ for some $x_* \in \mathbb{R}^d$, then by definition of the sub-differential we have

$$f(y) \geq f(x_*) + 0^\top (y - x_*) = f(x_*)$$

for all $y \in \mathbb{R}^d$ and it follows that x_* is a global minimum of f. □

We can apply the previous stated optimality condition for solving the next exercise:

Exercise 2.3.7. Let

$$f(x) = \frac{1}{2}\|x - y\|^2 + \lambda \|x\|_1, \quad x \in \mathbb{R}^d,$$

be the Lagrangian form of the least squares Lasso method. Note that $\|\cdot\|_1$ denotes the 1-norm defined as $\|x\|_1 = \sum_{i=1}^d |x_i|$ for $x = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$.

1. Compute a sub-gradient of f.
2. Prove that f is a convex function.
3. Apply Proposition 2.3.25 to find a global minimum of f.

We now come to the formulation of the gradient descent "like" method for non-differentiable cost functions - also called sub-gradient descent method. Instead of moving into direction of the steepest descent given as the negative gradient, the iterative scheme moves into direction of some arbitrary negative sub-gradient, see Algorithm 3. Note that in general the sub-gradient descent method is not necessarily a descent method as defined in the beginning of this section.
Since we do not assume differentiability of f and in particular, we do not assume L-smoothness of f, we are not able to directly apply the descent lemma 2.3.2.
Before going into details of the proof of convergence for sub-gradient descent methods, we derive the following useful property.

Algorithm 3 Sub-gradient descent method

1: **Input:**
- cost function $f : \mathbb{R}^d \to \mathbb{R}$
- initial $x_0 \in \mathbb{R}^d$
- sequence of step sizes $(\alpha_k)_{k\in\mathbb{N}}$, $\alpha_k > 0$

2: set $k = 0$
3: **while** "convergence/stopping criterion not met" **do**
4: find a sub-gradient $g_{x_k} \in \partial f(x_k)$
5: set $x_{k+1} = x_k - \alpha_k g_{x_k}$, $k \mapsto k+1$
6: **end while**

Lemma 2.3.26. Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and M-Lipschitz continuous, i.e.

$$|f(x) - f(y)| \leq M\|x - y\|$$

for all $x, y \in \mathbb{R}^d$. Then every sub-gradient $g_x \in \partial f(x)$ for all $x \in \mathbb{R}^d$ remains uniformly bounded by

$$\|g_x\| \leq M.$$

Proof. Let $g_x \in \partial f(x)$ for any $x \in \mathbb{R}^d$. Then by definition of the sub-gradient it follows that

$$f(x + z) \geq f(x) + g_x^\top z$$

for any $z \in \mathbb{R}^d$. We can reformulate the inequality such that

$$g_x^\top z \leq f(x + z) - f(x) \leq |f(x + z) - f(x)| \leq M\|z\|.$$

Since $z \in \mathbb{R}^d$ is arbitrary, we set $z = g_x$ implying that

$$g_x^\top g_x = \|g_x\|^2 \leq M\|g_x\|$$

and therefore $\|g_x\| \leq M$. □

We are now ready to formulate the convergence of the sub-gradient descent method.

Theorem 2.3.27 (Convergence sub-gradient descent method)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and M-Lipschitz continuous, i.e.*

$$|f(x) - f(y)| \leq M\|x - y\|$$

for all $x, y \in \mathbb{R}^d$. Moreover, let $(x_k)_{k\in\mathbb{N}}$ be generated by

$$x_{k+1} = x_k - \alpha_k g_{x_k},$$

with $\alpha_k > 0$ and arbitrary sub-gradient $g_{x_k} \in \partial f(x_k)$. Then, assuming existence of a global minimum $x_* \in \mathbb{R}^d$, it holds true that

$$e(x_k) = f(\bar{x}_N) - f(x_*) \leq \frac{\|x_0 - x_*\| + M^2 \sum_{k=0}^N \alpha_k^2}{2\sum_{k=0}^N \alpha_k},$$

where $\bar{x}_N := \sum_{k=0}^N w_k x_k$ is a weighted average over all iterations with weights

$$w_k = \frac{\alpha_k}{\sum_{s=0}^N \alpha_s}, \quad k = 1, \dots, N.$$

Proof. Following the iteration of $(x_k)_{k\in\mathbb{N}}$, we have

$$\begin{aligned}\|x_{k+1} - x_*\|^2 = \|x_k - \alpha_k g_{x_k} - x_*\|^2 &= \|x_k - x_*\|^2 - 2\alpha_k \langle g_{x_k}, x_k - x_*\rangle + \alpha_k^2 \|g_{x_k}\|^2 \\ &\leq \|x_k - x_*\|^2 - 2\alpha_k (f(x_k) - f(x_*)) + \alpha_k^2 M^2,\end{aligned}$$

where we have used that g_{x_k} is a sub-gradient of f and $\|g_{x_k}\|^2 \leq M^2$ by Lemma 2.3.26. We reformulate the abover inequality and proceed with summing over all iterations to obtain

$$\begin{aligned}2\sum_{k=0}^N \alpha_k (f(x_k) - f(x_*)) &\leq \sum_{k=0}^N \left(\|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2 + \alpha_k^2 M^2\right) \\ &= \|x_0 - x_*\|^2 - \|x_{N+1} - x_*\|^2 + M^2 \sum_{k=0}^N \alpha_k^2 \\ &\leq \|x_0 - x_*\|^2 + M^2 \sum_{k=0}^N \alpha_k^2,\end{aligned}$$

where we have used that the first two terms are a telescoping sum. We apply Jensen's inequality, Proposition A.1.6, to imply

$$f(\bar{x}_N) - f(x_*) \leq \sum_{k=0}^N w_k \left(f(x_k) - f(x_*)\right) \leq \frac{\|x_0 - x_*\|^2 + M^2 \sum_{k=0}^N \alpha_k^2}{2\sum_{k=0}^N \alpha_k},$$

with $w_k = \frac{\alpha_k}{\sum_{s=0}^N \alpha_k} \in (0,1)$ and $\sum_{k=0}^N w_k = 1$. □

Remark 2.3.28. Assuming that $\lim_{N\to\infty} \sum_{k=0}^N \alpha_k^2 < \infty$ and $\lim_{N\to\infty} \sum_{k=0}^N \alpha_k = \infty$ implies convergence of the sub-gradient descent method through the upper bound in Theorem 2.3.27. It is left as an exercise to quantify the speed of convergence for different choices of step sizes $(\alpha_k)_{k\in\mathbb{N}}$.

3

Accelerated gradient descent methods (Momentum)

We consider a class of first order optimization schemes devoted to accelerate the convergence behavior of gradient descent methods. The idea is to incorporate information of the previous iterations - the so-called momentum - into the iterative update scheme instead of just moving into direction of the current steepest descent direction. We will motivate the effects of momentum through the example of minimizing a quadratic cost function presented in [19]. As we have seen in Section 2.3.3, in particular in Theorem 2.3.16, the convergence rate for the error $e(x_k) = \|x_k - x_*\|$ of gradient descent, given by

$$c = \frac{\kappa - 1}{\kappa + 1} = 1 - \frac{2}{\kappa + 1},$$

scales poorly when the condition number $\kappa = L/\mu$ is large, where L denotes the smoothness parameter and μ the strong convexity parameter. In case of quadratic cost functions of the form $f(x) = \frac{1}{2}x^\top Qx$ with positive definite matrix $Q \in \mathbb{R}^{d\times d}$, the ratio κ corresponds to the condition number of Q. We make this more precise in the following example.

Example 3.0.1 (Quadratic cost function)**.** *Let $Q \in \mathbb{R}^{d\times d}$ be a positive definite matrix with eigenvalues $\lambda_{\max} = \lambda_1 \geq \cdots \geq \lambda_d = \lambda_{\min} > 0$. We aim to solve*

$$\min_{x\in\mathbb{R}^d} f(x), \quad f(x) = \frac{1}{2}x^\top Qx$$

using the gradient descent method. Let us consider an eigendecomposition of Q such that $Q = UDU^\top$ where $U \in \mathbb{R}^{d\times d}$ is an orthogonal matrix and $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^{d\times d}$ a diagonal matrix with the eigenvalues along the diagonal. The inner product scaled by Q can be rewritten as

$$\frac{1}{2}x^\top Qx = \frac{1}{2}x^\top(UDU^\top)x = \frac{1}{2}(U^\top x)^\top D(U^\top x) = \frac{1}{2}z^\top Dz$$

with $z = U^\top x$. We can then consider the equivalent optimization problem (since all eigenvalues

are positive) of the form

$$\min_{x\in\mathbb{R}^d} f(x), \quad f(x) = \frac{1}{2}x^\top Dx.$$

The gradient and Hessian compute as

$$\nabla f(x) = Dx \quad \textit{and} \quad \nabla^2 f(x) = D$$

and the unique global minimum is given by $x_* = 0 \in \mathbb{R}^d$. *Let* $x_0 \in \mathbb{R}^d$, $x_0 \neq 0$ *be the initialization and* $\alpha_k = \alpha \in (0, \frac{2}{\lambda_{\max}})$ *(smoothness parameter* $L = \lambda_{\max}$*) a fixed step size. Recall that the gradient descent method is written as*

$$x_{k+1} = x_k - \alpha\nabla f(x_k) = x_k - \alpha D x_k.$$

The distance to the global minimum is given by

$$\|x_{k+1} - x_*\| = \|x_{k+1}\| = \|x_k - \alpha D x_k\| = \|(I - \alpha D)x_k\| \leq \max(|1 - \alpha\lambda_{\min}|, |1 - \alpha\lambda_{\max}|)\|x_k\|.$$

We can now compute the step size such that the derived upper bound is minimized in the sense that

$$\min_{\alpha\in(0,\frac{2}{\lambda_{\max}})} \max(|1 - \alpha\lambda_{\min}|, |1 - \alpha\lambda_{\max}|).$$

It is an exercise to prove that the resulting optimal step size is given by $\alpha_* = \frac{2}{\lambda_{\min}+\lambda_{\max}}$ *(note that this step size coincides with the one derived for general* μ*-strongly convex and* L*-smooth cost functions in Remark 2.3.18). Let* $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$ *be the condition number of* Q *and* D *respectively. Then we have*

$$\begin{aligned}\max(|1 - \alpha_*\lambda_{\min}|, |1 - \alpha_*\lambda_{\max}|) &= |1 - \alpha_*\lambda_{\min}| = |1 - \alpha_*\lambda_{\max}| \\ &= \left|1 - \frac{2\lambda_{\max}}{\lambda_{\min} + \lambda_{\max}}\right| = \frac{\kappa - 1}{\kappa + 1} = 1 - \frac{2}{\kappa + 1}.\end{aligned}$$

Therefore, the gradient descent method with fixed step size α_* *converges linearly with*

$$\|x_k - x_*\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x_0 - x_*\|.$$

In order to achieve an error of tolerance $\varepsilon > 0$, *we need to iterate a certain amount of steps:*

$$\begin{aligned}\left(\frac{\kappa - 1}{\kappa + 1}\right)^k < \varepsilon \quad &\Leftrightarrow \quad \log\left(\frac{1}{\varepsilon}\right) < k \log\left(\frac{\kappa + 1}{\kappa - 1}\right) \\ &\Leftrightarrow \quad k > \log\left(\frac{1}{\varepsilon}\right) \log\left(1 + \frac{2}{\kappa - 1}\right)^{-1},\end{aligned}$$

which increases with increasing condition number κ. Hence, gradient descent may perform poorly for quadratic functions with high condition number κ.

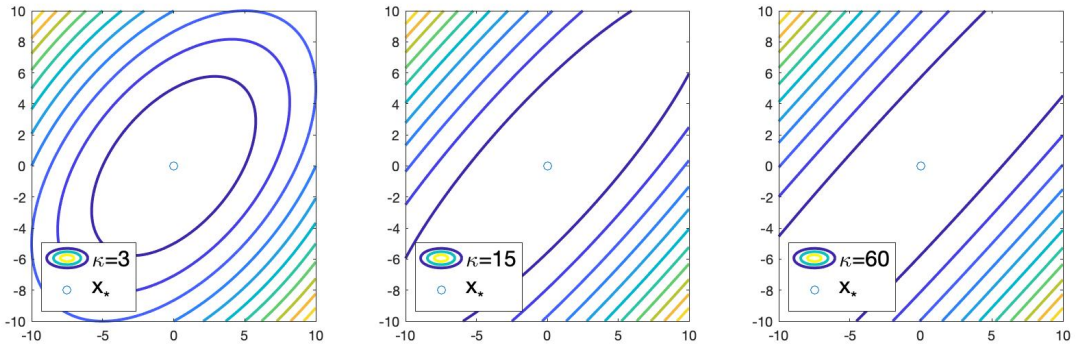

Figure 3.1: Contour lines of a quadratic function for increasing condition number κ.

3.1 Polyak's heavy ball method

The idea of incorporating momentum into iterative optimization schemes goes back to Polyak's so-called "heavy-ball" method (1964) [18]. As the name suggest the motivation behind the method is a heavy ball rolling down the hill into direction of a valley.
While a "light" ball is significantly influenced by tight curvatures, such as in a ravine, and loses in velocity through high oscillation, a heavy ball is accelerated due to low influence through curvatures. From a mathematical point of view, the momentum is incorporated as form of damping of the descent direction.
We formulate Polyak's heavy ball method (HBM) in the following algorithm:

Algorithm 4 Heavy ball method

Input:

- cost function $f : \mathbb{R}^d \to \mathbb{R}$
- initial $x_0 \in \mathbb{R}^d$
- sequence of step sizes $(\alpha_k)_{k\in\mathbb{N}}$, $\alpha_k > 0$, and sequence of momentum parameters $(\beta_k)_{k\in\mathbb{N}}$, $\beta_k \geq 0$.

set $x_1 = x_0 - \alpha_0 \nabla f(x_0)$, and $k = 1$
while "convergence/stopping criterion not met" **do**
    set $x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \beta_k(x_k - x_{k-1})$, $k \mapsto k+1$
end while

Let us take a closer look into the iterative update of HBM:

$$x_{k+1} \quad = \quad \underbrace{x_k - \alpha_k \nabla f(x_k)}_{\text{gradient descent}} \quad + \quad \underbrace{\beta_k(x_k - x_{k-1})}_{\text{Heavy ball momentum}} .$$

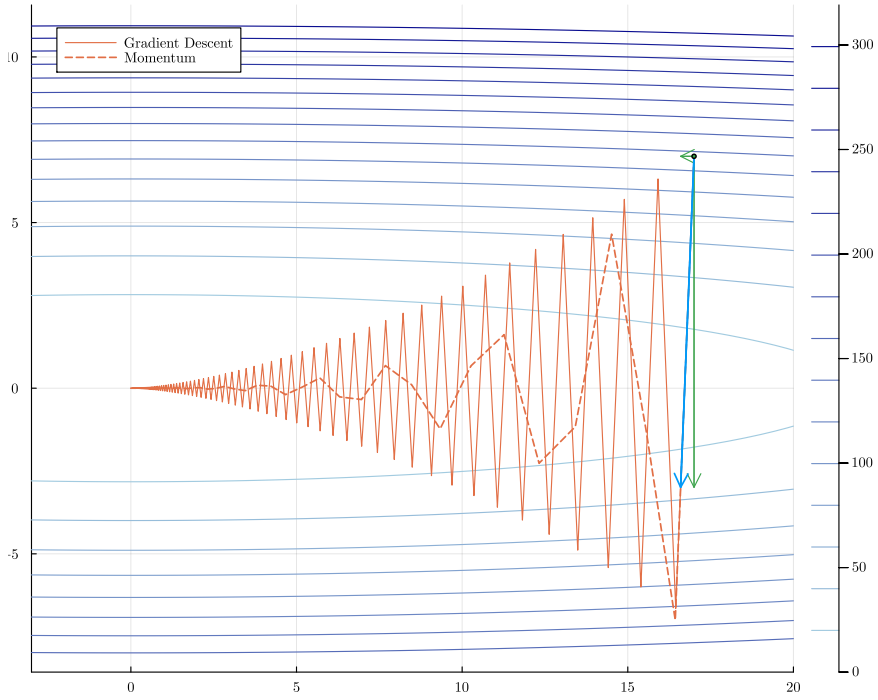

Figure 3.2: Illustration of the damping effect through momentum.

This means, we incorporate the update from the previous iteration through

$$x_k - x_{k-1} = -\alpha_{k-1}\nabla f(x_{k-1}) + \beta_{k-1}(x_{k-1} - x_{k-2})$$

into the next iteration. For example, in the second iteration, the HBM update is given by

$$x_2 = x_1 - \alpha_1 \nabla f(x_1) - \beta_1 \alpha_0 \nabla f(x_0).$$

The hyperparameter $\beta_k \geq 0$ controls the strength of the influence of momentum and can be seen as damping parameter. Moreover, we can choose $\beta_k = 0$ to recover the gradient descent method (without momentum). Of course, the performance of the HBM method highly depends on a good choice the parameter β_k.

Remark 3.1.1. In general HBM is no descent method and therefore, we do not expect a monotonic decrease of the cost function along the iterations.

We will continue with Example 3.0.1 and derive an optimal choice of step size α_k and momentum parameter β_k for quadratic cost function.

Example 3.1.2 (Continuation of Example 3.0.1)**.** *Let us come back to minimizing our quadratic cost function* $f(x) = \frac{1}{2}x^\top Dx$*. We consider HBM with fixed step size* $\alpha_k = \alpha > 0$ *and fixed*

momentum parameter $\beta_k = \beta > 0$. The iterative scheme is given by

$$x_{k+1} = x_k - \alpha D x_k + \beta(x_k - x_{k-1}).$$

In this case, we consider the joint update of the vector

$$\begin{pmatrix} x_{k+1} - x_* \\ x_k - x_* \end{pmatrix} = \begin{pmatrix} x_{k+1} \\ x_k \end{pmatrix} \in \mathbb{R}^{2d}.$$

The iteration can be written as

$$\begin{aligned} \begin{pmatrix} x_{k+1} \\ x_k \end{pmatrix} &= \begin{pmatrix} x_k - \alpha D x_k + \beta(x_k - x_{k-1}) \\ x_k \end{pmatrix} \\ &= \begin{pmatrix} (1+\beta) I x_k - \alpha D x_k - \beta I x_{k-1} \\ I x_k \end{pmatrix} \\ &= \begin{pmatrix} (1+\beta) I - \alpha D & -\beta I \\ I & 0 \end{pmatrix} \begin{pmatrix} x_k \\ x_{k-1} \end{pmatrix} =: T \begin{pmatrix} x_k \\ x_{k-1} \end{pmatrix}. \end{aligned}$$

Since the matrix $T \in \mathbb{R}^{2d \times 2d}$ is independent of the iteration $k \in \mathbb{N}$, we obtain

$$\begin{pmatrix} x_{k+1} \\ x_k \end{pmatrix} = T^k \begin{pmatrix} x_1 \\ x_0 \end{pmatrix},$$

and therefore, the error can be written as

$$\left\| \begin{pmatrix} x_{k+1} - x_* \\ x_k - x_* \end{pmatrix} \right\| = \left\| T^k \begin{pmatrix} x_1 - x_* \\ x_0 - x_* \end{pmatrix} \right\| \leq \|T^k\| \left\| \begin{pmatrix} x_1 - x_* \\ x_0 - x_* \end{pmatrix} \right\|,$$

for some matrix norm which is consistent with the euclidean 2-norm. Let $\rho(T) = \max_{i=1,\dots,2d} |\lambda_i(T)|$ be the largest eigenvalue $\lambda_i(T)$ (in absolute value) of T, also called the spectral radius of T. We will make use of the Gelfand formula which states that $\rho(T) = \lim_{k\to\infty} \|A^k\|^{1/k}$ for any matrix norm, and in particular there exists a sequence $(\varepsilon_k)_{k\in\mathbb{N}}$ converging to 0, such that

$$\|T^k\| \leq (\rho(T) + \varepsilon_k)^k.$$

Before going into details, we transform T to a block diagonal matrix without changing the corresponding eigenvalues (note that the eigenvectors change). From Exercise 3.1.1 we observe that T

is self-similar to a block diagonal matrix

$$\widehat{T} = \begin{pmatrix} T_1 & & \\ & \ddots & \\ & & T_d \end{pmatrix}$$

with blocks

$$T_i = \begin{pmatrix} 1+\beta-\alpha\lambda_i & -\beta \\ 1 & 0 \end{pmatrix} \in \mathbb{R}^{2\times 2}.$$

Let μ_i be an eigenvalue of $\widehat{T}$ with corresponding eigenvector $v_i \in \mathbb{R}^{2d}$, then we can compute

$$\widehat{T}v_i = \mu_i v_i = S^{-1}TSv_i \quad \Leftrightarrow \quad \mu_i Sv_i = TSv_i,$$

which means that μ_i is also an eigenvalue of T with corresponding eigenvector Sv_i. Without loss of generality we will compute the eigenvalues of $\widehat{T}$ instead of T. Due to the block structure, we can deduce the computation of the eigenvalues from the computation of the eigenvalues of each block $T_i \in \mathbb{R}^{2\times 2}$. To do so, we aim to find $\mu_i \in \mathbb{C}$ satisfying

$$\det(T_i - \mu_i I) = \det\left(\begin{pmatrix} 1+\beta-\alpha\lambda_i-\mu_i & -\beta \\ 1 & -\mu_i \end{pmatrix}\right) = \mu_i^2 - \mu_i(1+\beta-\alpha\lambda_i) + \beta \overset{!}{=} 0.$$

For each block we obtain two eigenvalues given as

$$\mu_i^{(1)} = \frac{1+\beta-\alpha\lambda_i}{2} - \sqrt{\left(\frac{1+\beta-\alpha\lambda_i}{2}\right)^2 - \beta}$$
$$\mu_i^{(2)} = \frac{1+\beta-\alpha\lambda_i}{2} + \sqrt{\left(\frac{1+\beta-\alpha\lambda_i}{2}\right)^2 - \beta}.$$

We restrict our self to $\beta > 0$ such that $\left(\frac{1+\beta-\alpha\lambda_i}{2}\right)^2 - \beta \leq 0$ and therefore, the eigenvalues are complex-valued. It then holds true that the absolute value is given by

$$\begin{aligned} |\mu_i^{(1)}| = |\mu_i^{(2)}| &= \frac{1}{2}\sqrt{(1-\alpha\lambda_i+\beta)^2 + |\underbrace{(1-\alpha\lambda_i+\beta)^2 - 4\beta}_{\leq 0}|} \\ &= \frac{1}{2}\sqrt{(1-\alpha\lambda_i+\beta)^2 - (1-\alpha\lambda_i+\beta)^2 + 4\beta} = \sqrt{\beta}. \end{aligned}$$

Motivated by this observation, we aim to satisfy $\left(\frac{1+\beta-\alpha\lambda_i}{2}\right)^2 - \beta \leq 0$ for all i, such that the spectral

radius is given by $\rho(T)=\sqrt{\beta}$. *Let us consider*

$$\Delta(\beta):=\left(\frac{1+\beta-\alpha\lambda_i}{2}\right)^2-\beta=\frac{1}{4}(\beta^2-2(1+\alpha\lambda_i)\beta+(1-\alpha\lambda_i)^2).$$

The mapping $\beta\mapsto\Delta(\beta)$, $\beta\in\mathbb{R}$, *describes a parabola, such that* $\Delta(\beta)<0$ *between two points* $\beta^{(1)},\beta^{(2)}$ *satisfying* $\Delta(\beta^{(1)})=\Delta(\beta^{(2)})=0$ *(if two exist). Therefore, we solve* $\Delta(\beta)\overset{!}{=}0$. *We focus on the cases where* $\beta<1$ *and* $\alpha<\frac{4}{\lambda_{\max}}$ *such that it holds*

$$|1-\sqrt{\alpha\lambda_{\min}}|,|1-\sqrt{\alpha\lambda_{\max}}|\in(0,1).$$

Moreover, we observe that

$$(1-\sqrt{\alpha\lambda_i})^2\leq\max\left((1-\sqrt{\alpha\lambda_{\min}})^2,(1-\sqrt{\alpha\lambda_{\max}})^2\right)$$

such that it is sufficient to choose

$$1>\beta\geq\max\left((1-\sqrt{\alpha\lambda_{\min}})^2,(1-\sqrt{\alpha\lambda_{\max}})^2\right).$$

in order to force $\rho(T)=\sqrt{\beta}$. *With the specific choice*

$$\alpha=\frac{4}{(\sqrt{\lambda_{\min}}+\sqrt{\lambda_{\max}})^2}\quad\textit{and}\quad\beta=\max\left((1-\sqrt{\alpha\lambda_{\min}})^2,(1-\sqrt{\alpha\lambda_{\max}})^2\right)$$

we deduce that

$$\beta=(1-\sqrt{\alpha\lambda_{\max}})^2=(1-\sqrt{\alpha\lambda_{\min}})^2=\left(\frac{\sqrt{\lambda_{\max}}-\sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}}+\sqrt{\lambda_{\min}}}\right)^2=\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$$

and therefore, $\rho(T)=\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$. *Finally, using the Gelfand formula we obtain an improved upper bound (for sufficiently large* k*) on the error compared to gradient descent method*

$$\left\|\begin{pmatrix}x_{k+1}-x_*\\x_k-x_*\end{pmatrix}\right\|\leq\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}+\varepsilon_k\right)^k\left\|\begin{pmatrix}x_1-x_*\\x_0-x_*\end{pmatrix}\right\|.$$

Similarly as before, in order to achieve an error of tolerance $\varepsilon>0$, *we need to iterate a certain amount of steps:*

$$\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k<\varepsilon\Leftrightarrow k>\log\left(\frac{1}{\varepsilon}\right)\log\left(1+\frac{2}{\sqrt{\kappa}-1}\right)^{-1},$$

where $\log\left(1+\frac{2}{\sqrt{\kappa}-1}\right)^{-1}\leq\log\left(1+\frac{2}{\kappa-1}\right)^{-1}$, *since* $\kappa\geq1$ *and therefore* $\sqrt{\kappa}\leq\kappa$.

Exercise 3.1.1. Let $T \in \mathbb{R}^{2d\times 2d}$ be defined as

$$T = \begin{pmatrix} (1+\beta)I - \alpha D & -\beta I \\ I & 0 \end{pmatrix},$$

with diagonal matrix $D = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$, $\alpha, \beta > 0$. Prove that there exists a regular matrix $S \in \mathbb{R}^{2d\times 2d}$ such that

$$S^{-1}TS = \widehat{T} = \begin{pmatrix} T_1 & & \\ & \ddots & \\ & & T_d \end{pmatrix},$$

where $\widehat{T}$ is a block diagonal matrix with

$$T_i = \begin{pmatrix} 1+\beta-\alpha\lambda_i & -\beta \\ 1 & 0 \end{pmatrix} \in \mathbb{R}^{2\times 2}.$$

We observed that in the case of quadratic cost function we are able to improve the rate of convergence compared to gradient descent. This suggests that the convergence behavior of gradient descent, as derived in Section 2.3.2 and Section 2.3.3, might be sub-optimal.

3.2 Discussion about optimality of the convergence behavior

For quadratic cost functions we have seen that the rate of convergence of the gradient descent method can be improved through the incorporation of momentum in form of HBM. This raises the question about optimality of gradient descent as a first order method. Or the other way around, what is the best possible convergence behavior we can expect using only first order information. We consider the following class of first order iterative methods.

Assumption 3.2.1. The sequence $(x_k)_{k\in\mathbb{N}}$ (generated by some iterative scheme) satisfies the condition

$$x_k \in x_0 + \mathrm{span}\{\nabla f(x_0), \dots, \nabla f(x_{k-1})\}$$

for all $k \geq 1$.

Assumption 3.2.1 means that each iteration x_k can be expressed as a linear combination of the initialization x_0 and all previous gradients $\nabla f(x_0), \dots, \nabla f(x_{k-1})$. Both gradient descent and HBM are examples satisfying Assumption 3.2.1.

We recall, that for cost functions which are μ-strongly convex and L-smooth, we obtain linear convergence of gradient descent with fixed step size $\bar{\alpha} = \frac{2}{\mu+L}$ of the form

$$\|x_k - x_*\|^2 \leq \left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \|x_0 - x_*\|^2,$$

see Section 2.3.3. For the specific case of quadratic cost function, the upper bound can be improved through HBM to

$$\|x_k - x_*\|^2 \leq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \|x_0 - x_*\|^2.$$

The following lower bound from Nesterov, see e.g. [17], shows that we can not expect more than this improvement as long as we do not include more than first order information, i.e. as long as our iterative scheme satisfies Assumption 3.2.1. However, the lower bound is derived for a specifically constructed function with high- or even infinite-dimensional domain, to be more precise, for a function $f : \ell^2(\mathbb{R}) \to \mathbb{R}$, where

$$\ell^2(\mathbb{R}) := \{(z_i)_{i\in\mathbb{N}} \mid z_i \in \mathbb{R}, \ \sum_{i=1}^{\infty} |z_i|^2 < \infty\}.$$

Note that $\ell^2(\mathbb{R})$ can be equipped with a norm as well as an inner product, such that it forms a Banach and even a Hilbert space.

Theorem 3.2.2 (Lower bound strong convex and smooth, Theorem 2.1.13 in[17]). *For each $x_0 \in \ell^2(\mathbb{R})$, $\mu, L > 0$ with $\kappa = \frac{L}{\mu} > 1$, there exists a μ-strongly convex and L-smooth function $f : \ell^2(\mathbb{R}) \to \mathbb{R}$ such that every iterative scheme $(x_k)_{k\in\mathbb{N}}$ satisfying Assumption 3.2.1 satisfies a lower bound on the error given by*

$$e(x_k) := \|x_k - x_*\|^2 \geq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \|x_0 - x_*\|^2,$$

where $x_* \in \ell^2(\mathbb{R})$ denotes the unique global minimum of f.

Remark 3.2.3. The proof of Theorem 3.2.2 in [17] is constructive, i.e. one can construct a certain μ-strongly convex and L-smooth function $f : \ell^2(\mathbb{R}) \to \mathbb{R}$ satisfying the lower bound. We will take a close look on this as part of the exercises.

Remark 3.2.4. We note that the lower bound in Theorem 3.2.2 covers a more general setting than the one consider in this lecture course so far. Up to now, we have only considered cost functions $f : \mathbb{R}^d \to \mathbb{R}$ with finite dimensional domain $\mathbb{R}^d$. Strictly speaking, we would need to go back to the beginning of this course in order to move from finite dimensional to infinite dimensional domains. One needs to re-define derivatives and consider more general optimality conditions. This would

lead to the so-called Fréchet derivative, which are needed to formulate gradient descent methods in Hilbert spaces. However, this is beyond the scope of this lecture course.

We now return to the setting of Section 2.3.2, where we have assumed general convex and L-smooth functions (without a strong convexity assumption). Under these properties, gradient descent with a fixed step size $\bar{\alpha} \leq \frac{1}{L}$ converges with upper bound of the form

$$f(x_k) - f_* \leq \frac{c}{k+1}, \quad k \in \mathbb{N}, c > 0,$$

where $f_* = \min_{x \in \mathbb{R}^d} f(x)$. Indeed, also in this scenario, it is possible to derive a lower bound on iterative schemes satisfying Assumption 3.2.1, which suggest a gap between upper and lower bound.

Theorem 3.2.5 (Lower bound convex and smooth, Theorem 2.1.7 in[17]). *For every $k \in \mathbb{N}$ with $1 \leq k \leq \frac{1}{2}(d-1)$, $L > 0$ and every $x_0 \in \mathbb{R}^d$ (d denotes the dimension of the domain), there exists a convex and L-smooth function $f : \mathbb{R}^d \to \mathbb{R}$ such that every iterative scheme $(x_k)_{k \in \mathbb{N}}$ satisfying Assumption 3.2.1 satisfies a lower bound on the error given by*

$$e(x_k) := f(x_k) - f_* \geq \frac{3L\|x_0 - x_*\|^2}{32(k+1)^2},$$

where $f_* = \min_{x \in \mathbb{R}^d} f(x) > -\infty$ exists.

Remark 3.2.6. The considered lower bound in Theorem 3.2.5 is only satisfied for $k \leq \frac{1}{2}(d-1)$, which again, particularly for high dimensional ($d \gg 1$) optimization tasks, suggests a gap between lower and upper bound. The proof in [17] is again via construction, and will be considered in more detail as part of the exercises.

We ask our self if we can improve the upper bounds derived for gradient descent methods (both for convex and strong-convex setting) through momentum methods. This will be part of the next section.

3.3 Nesterov's acceleration method

Recall that the iteration of HBM is given by

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \beta_k (x_k - x_{k-1}).$$

We have seen that we can obtain linear convergence for quadratic cost function $f(x) = \frac{1}{2}x^\top Qx$ with rate $c = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$. To do so, we have derived

$$\alpha_k = \bar{\alpha} = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2} \quad \text{and} \quad \beta_k = \bar{\beta} = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2, \tag{3.1}$$

where $L > 0$ denotes the largest eigenvalue and $\mu > 0$ the smallest eigenvalue of Q, and $\kappa = \frac{L}{\mu}$ is the condition number. We wonder if it is possible to extend this result to general L-smooth and μ-strongly convex functions. If we similarly set α, β fixed as in (3.1), do we still obtain linear convergence with rate $c = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$? Unfortunately, this is not true. We consider the following counter example presented in [16], where it turns out that one can construct a one-dimensional L-smooth and μ-strongly convex function, for which HBM runs into a circle and does not converge. This example is formulated as exercise:

Exercise 3.3.1. 1. Find a continuous function $f : \mathbb{R} \to \mathbb{R}$ such that

$$f'(x) = \begin{cases} 25x, & x \leq 1 \\ x + 24, & 1 < x < 2 \\ 25x - 24, & 2 \leq x \end{cases}.$$

Prove that f is μ-strongly convex with $\mu = 1$, L-smooth with $L = 25$ and has a unique global minimum in $x_* = 0$.

2. Implement HBM with the optimal step size α and momentum parameter β following (3.1).

3. Prove that the application of HBM on f with the parameters in (3.1) result in the recursion

$$x_{k+1} = \frac{13}{9}x_k - \frac{4}{9}x_{k-1} - \frac{1}{9}\nabla f(x_k).$$

4. Find a cycle of points $p \to q \to r \to p$, such that for $x_0 = p$ we have

$$x_{3k} = p, \; x_{3k+1} = q, \; x_{3k+2} = r$$

for all $k \in \mathbb{N}$. To do so, assume $p, q < 1$ and $r > 2$, apply the heavy ball recursion to create a linear equation for p, q, r and solve it. What does it mean for the convergence behavior?

Motivation: We will try to motivate a different approach of incorporating momentum - Nesterov's acceleration method (NAM). Recall that the iteration of HBM firstly computes the gradient at the

current location $\nabla f(x_k)$ and then moves into direction of a weighted sum of all previous gradients

$$d_k = -\alpha_k \nabla f(x_k) + \beta_k(x_k - x_{k-1}) = -\alpha_k \nabla f(x_k) - \beta_k \alpha_{k-1} \nabla f(x_{k-1}) + \beta_k \beta_{k-1}(x_{k-1} - x_{k-2}) = \dots .$$

In NAM this step is split into two sub-steps, where we firstly move into direction of the iterated previous gradients and then compute the next gradient correcting the first move. We can describe this method through a coupled system of two vectors $[p_k, q_k] \in \mathbb{R}^d \times \mathbb{R}^d$:

1. Assume that we are in location $p_k \in \mathbb{R}^d$, such that we can obtain information from the previous iterations through the computation

$$q_k = p_k + \beta(p_k - p_{k-1}),$$

for some momentum parameter $\beta > 0$.

2. In location q_k we compute the next gradient information in order to correct the previously gained information

$$p_{k+1} = q_k - \alpha \nabla f(q_k),$$

where $\alpha > 0$ denotes a step size/ learning rate.

3. Compute the iterated weighted information for the next iteration

$$q_{k+1} = p_{k+1} + \beta(p_{k+1} - p_k).$$

We summarize NAM in Algorithm 5.

Algorithm 5 Nesterov's accelerated gradient descent method

Input:

- cost function $f : \mathbb{R}^d \to \mathbb{R}$
- initial $q_0, p_0 \in \mathbb{R}^d$
- sequence of step sizes $(\alpha_k)_{k\in\mathbb{N}}$, $\alpha_k > 0$, and sequence of momentum parameters $(\beta_k)_{k\in\mathbb{N}}$, $\beta_k \geq 0$.

set $p_1 = q_0 - \alpha_0 \nabla f(q_0)$
set $q_1 = p_1 + \beta_0(p_1 - p_0)$
set $k = 1$
while "convergence/stopping criterion not met" **do**
    set $p_{k+1} = q_k - \alpha_k \nabla f(q_k)$
    set $q_{k+1} = p_{k+1} + \beta_k(p_{k+1} - p_k)$
    set $k \mapsto k + 1$
end while

3.3.1 Convergence for convex and smooth functions

We will start the convergence analysis of NAM with the assumption of convex and L-smooth cost functions. In order to analyze NAM, we will consider a slightly more general way of writing the iterative scheme described through three variable $(x_k, y_k, z_k)_{k\in\mathbb{N}}$. The convergence analysis is based on Lyapunov methods for optimization schemes as it has been analyzed in [25] for NAM. The analysis presented in [25] covers a wide range of acceleration algorithms in continuous and discrete time setting. In terms of the scope of this lecture course, we will focus on a very simplified setting, where we write NAM as system of the form

$$\begin{aligned} x_k &= \tau_k z_k + (1-\tau_k) y_k, \\ y_{k+1} &= x_k - \alpha_k \nabla f(x_k), \\ z_{k+1} &= z_k - \gamma_k \nabla f(x_k), \end{aligned} \tag{3.2}$$

with parameters $\alpha_k, \gamma_k > 0$ and $\tau_k \in (0,1)$. The variable y_k represents the current gradient step, whereas z_k iterates momentum in form of memorizing all the previous gradient steps. The variable x_k then combines both steps. Consider the following example to see that this system can be seen as NAM.

Example 3.3.1. *We ask ourselves if we can transform the system (3.2) back to the form of Algorithm 5. Indeed, one can choose the parameters $\gamma_k, \tau_k > 0$ such that the system reduces into the form of Algorithm 5. Therefore, we rewrite the update*

$$\begin{aligned} z_{k+1} = z_k + \frac{1}{\tau_k}(x_k - x_k) - \gamma_k \nabla f(x_k) &= y_k + \frac{1}{\tau_k}(x_k - y_k) - \gamma_k \nabla f(x_k) \\ &= y_k + \frac{1}{\tau_k}(x_k - \gamma_k \tau_k \nabla f(x_k) - y_k) \\ &= y_k + \frac{1}{\tau_k}(y_{k+1} - y_k), \end{aligned}$$

where we have chosen $(\alpha_k, \gamma_k, \tau_k)$ such that $\gamma_k \tau_k = \alpha_k$. We can plug this into the update of x_{k+1} in order to eliminate z_{k+1} from (3.2):

$$x_{k+1} = \tau_{k+1} z_{k+1} + (1-\tau_{k+1}) y_{k+1} = y_{k+1} + \frac{\tau_{k+1}(1-\tau_k)}{\tau_k}(y_{k+1} - y_k).$$

Finally, we have written the system (3.2) as update of two variables $(x_k, y_k)_{k\in\mathbb{N}}$ described through

$$\begin{aligned} y_{k+1} &= x_k - \alpha_k \nabla f(x_k), \\ x_{k+1} &= y_{k+1} + \beta_k (y_{k+1} - y_k), \end{aligned}$$

where $\beta_k := \frac{\tau_{k+1}(1-\tau_k)}{\tau_k} > 0$.

As mentioned earlier, we will consider a simplified analysis based on Lyapunov methods as presented in [25]. There are many different ways of applying Lyapunov methods for analyzing optimization methods. See also Appendix A.2 for a brief motivation of Lyapunov methods in optimization. We will follow a specific strategy for proving convergence of an iterative scheme $(x_k)_{k\in\mathbb{N}}$ based on Lyapunov theory:

1. We firstly choose an error function $e : \mathbb{R}^d \to \mathbb{R}_+$ for which we want to prove convergence towards 0, i.e. $\lim_{k\to\infty} e(x_k) = 0$.

2. Construct a *Lyapunov function* of the form

$$E_k := E(x_k) = r(x_k) + A_k e(x_k),$$

where $r : \mathbb{R}^d \to \mathbb{R}_+$ is some auxiliary function, and $(A_k)_{k\in\mathbb{N}}$ is a monotonically increasing sequence with $A_0 \geq 0$ devoted to describe the speed of convergence.

3. We aim to bound the increments of the sequence $(E_k)_{k\in\mathbb{N}}$ by

$$E_{k+1} - E_k \leq \varepsilon_{k+1},$$

where $(\varepsilon_k)_{k\in\mathbb{N}}$ is a real-valued sequence with $\limsup_{k\to\infty} \varepsilon_k < +\infty$. In our specific case, we aim to prove that $\varepsilon_{k+1} \leq 0$ for all $k \in \mathbb{N}$ such that $(E_k)_{k\in\mathbb{N}}$ is non-increasing and particularly bounded by $E_k \leq E_0$. It then follows, that

$$A_k e(x_k) \leq r(x_k) + A_k e(x_k) = E_k \leq E_0$$

and therefore, we obtain

$$e(x_k) \leq \frac{E_0}{A_k}$$

which illustrates why $(A_k)_{k\in\mathbb{N}}$ describes the speed of convergence.

Motivated by [25], in order to analyze the convergence of the system (3.2), we will construct the Lyapunov function of the form

$$E_k = \frac{1}{2}\|z_k - x_*\|^2 + A_k(f(y_k) - f_*), \tag{3.3}$$

where f is assumed to satisfy $f_* = \min_{x\in\mathbb{R}^d} > -\infty$, $x_* \in \mathbb{R}^d$ is some global minimum of f and $(A_k)_{k\in\mathbb{N}}$ is a monotonically increasing sequence with $A_0 > 0$. We will firstly derive the following upper bound on the increments of $(E_k)_{k\in\mathbb{N}}$.

Lemma 3.3.2. Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and convex with $\min_{x\in\mathbb{R}^d} f > -\infty$, and assume there exists at least one global minimum $x_* \in \mathbb{R}^d$ of f. Moreover, let $(x_k, y_k, z_k)_{k\in\mathbb{N}}$ be generated by (3.2) with parameters $\alpha_k = \frac{1}{L}$, $\gamma_k = A_{k+1} - A_k$ and $\tau_k = \frac{\gamma_k}{A_{k+1}} = \frac{A_{k+1}-A_k}{A_{k+1}} \in (0, 1)$. i.e.

$$\begin{aligned} x_k &= y_k + \frac{A_{k+1} - A_k}{A_{k+1}}(z_k - y_k), \\ y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k), \\ z_{k+1} &= z_k - (A_{k+1} - A_k)\nabla f(x_k), \end{aligned}$$

initialized with $(y_0, z_0) \in \mathbb{R}^d \times \mathbb{R}^d$. Then the increments of the sequence $(E_k)_{k\in\mathbb{N}}$ defined in (3.3) satisfy

$$E_{k+1} - E_k \leq \varepsilon_{k+1} := \frac{1}{2}(A_{k+1} - A_k)^2\|\nabla f(x_k)\|^2 + A_{k+1}(f(y_{k+1}) - f(x_k))$$

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

Remark 3.3.3. In order to show monotonic behavior of the error $(E_k)_{k\in\mathbb{N}}$, i.e. $\varepsilon_{k+1} \leq 0$, we will later apply convexity and L-smoothness of f to derive

$$f(y_{k+1}) - f(x_k) \propto -\|\nabla f(x_k)\|^2.$$

It will turn out, that the choice of $(A_k)_{k\in\mathbb{N}}$ is the key to prove convergence of the error $e_k = f(y_k) - f_*$.

Proof of Lemma 3.3.2. Firstly, we write down the increments of $(E_k)_{k\in\mathbb{N}}$

$$E_{k+1} - E_k = \frac{1}{2}\|z_{k+1} - x_*\|^2 - \frac{1}{2}\|z_k - x_*\|^2 + A_{k+1}(f(y_{k+1}) - f_*) - A_k(f(y_k) - f_*)$$

and observe that

$$\begin{aligned} \frac{1}{2}\|z_{k+1} - x_*\|^2 - \frac{1}{2}\|z_k - x_*\|^2 &= \frac{1}{2}\|z_{k+1} - x_*\|^2 - \frac{1}{2}\|(z_k - z_{k+1}) + (z_{k+1} - x_*)\|^2 \\ &= \frac{1}{2}\|z_{k+1} - x_*\|^2 - \frac{1}{2}\|z_k - z_{k+1}\|^2 \\ &\quad - \langle z_k - z_{k+1}, z_{k+1} - x_*\rangle - \frac{1}{2}\|z_{k+1} - x_*\|^2 \\ &= -\frac{1}{2}\|z_k - z_{k+1}\|^2 - \langle z_k - z_{k+1}, z_{k+1} - x_*\rangle \\ &= -\frac{1}{2}\|z_k - z_{k+1}\|^2 + \langle x_* - z_{k+1}, (A_{k+1} - A_k)\nabla f(x_k)\rangle. \end{aligned}$$

This leads to

$$\begin{aligned}E_{k+1}-E_k &= \langle x_*-z_{k+1},(A_{k+1}-A_k)\nabla f(x_k)\rangle - \frac{1}{2}\|z_k-z_{k+1}\|^2\\ &\quad + A_{k+1}(f(y_{k+1})-f_*) - A_k(f(y_k)-f_*)\\ &= \langle x_*-z_k,(A_{k+1}-A_k)\nabla f(x_k)\rangle + \langle z_k-z_{k+1},(A_{k+1}-A_k)\nabla f(x_k)\rangle - \frac{1}{2}\|z_k-z_{k+1}\|^2\\ &\quad + A_{k+1}(f(y_{k+1})-f_*) - A_k(f(y_k)-f_*).\end{aligned}$$

Since $0 \leq \frac{1}{2}\|a-b\|^2 = \frac{1}{2}\|a\|^2 - \langle a,b\rangle + \frac{1}{2}\|b\|^2$ for $a,b\in\mathbb{R}^d$, we can apply the inequality $\langle a,b\rangle - \frac{1}{2}\|a\|^2 \leq \frac{1}{2}\|b\|^2$ to derive

$$\langle z_k-z_{k+1},(A_{k+1}-A_k)\nabla f(x_k)\rangle - \frac{1}{2}\|z_k-z_{k+1}\|^2 \leq \frac{1}{2}\|(A_{k+1}-A_k)\nabla f(x_k)\|^2$$

such that

$$\begin{aligned}E_{k+1}-E_k &= \frac{1}{2}(A_{k+1}-A_k)^2\|\nabla f(x_k)\|^2 + \langle x_*-z_k,(A_{k+1}-A_k)\nabla f(x_k)\rangle\\ &\quad + A_{k+1}(f(y_{k+1})-f_*) - A_k(f(y_k)-f_*).\end{aligned}$$

Moreover, we observe that

$$\begin{aligned}A_{k+1}(f(y_{k+1})-f_*) - A_k(f(y_k)-f_*) &= (A_{k+1}-A_k)(f(x_k)-f_*) + A_k(f(x_k)-f(y_k))\\ &\quad + A_{k+1}(f(y_{k+1}-f(x_k)).\end{aligned}$$

With $\varepsilon_{k+1} := \frac{1}{2}(A_{k+1}-A_k)^2\|\nabla f(x_k)\|^2 + A_{k+1}(f(y_{k+1})-f(x_k))$ it follows that

$$\begin{aligned}E_{k+1}-E_k &\leq \varepsilon_{k+1} + \langle x_*-z_k,(A_{k+1}-A_k)\nabla f(x_k)\rangle\\ &\quad + (A_{k+1}-A_k)(f(x_k)-f_*) + A_k(f(x_k)-f(y_k)),\end{aligned}$$

and it is left to prove that

$$\mathcal{R} = \langle x_*-z_k,(A_{k+1}-A_k)\nabla f(x_k)\rangle + (A_{k+1}-A_k)(f(x_k)-f_*) + A_k(f(x_k)-f(y_k)) \leq 0.$$

We add $-y_k+y_k=0$ to derive

$$\begin{aligned}\mathcal{R} &= \langle x_*-y_k,(A_{k+1}-A_k)\nabla f(x_k)\rangle + \langle y_k-z_k,(A_{k+1}-A_k)\nabla f(x_k)\rangle\\ &\quad + (A_{k+1}-A_k)(f(x_k)-f_*) + A_k(f(x_k)-f(y_k)),\end{aligned}$$

where we now aim to apply convexity of f in form of

$$f(z) - f(y) + \langle y - z, \nabla f(z)\rangle \leq 0.$$

Therefore, using $y_k - z_k = \frac{A_{k+1}}{A_{k+1}-A_k}(y_k - x_k)$ we again rewrite $\mathcal{R}$ in form of

$$\begin{aligned}
\mathcal{R} &= \langle x_* - y_k, (A_{k+1} - A_k)\nabla f(x_k)\rangle + A_{k+1}\langle y_k - x_k, \nabla f(x_k)\rangle \\
&\quad + (A_{k+1} - A_k)(f(x_k) - f_*) + A_k(f(x_k) - f(y_k)) \\
&= \langle x_* - y_k, (A_{k+1} - A_k)\nabla f(x_k)\rangle + A_{k+1}\langle y_k - x_k, \nabla f(x_k)\rangle \\
&\quad - A_k\langle x_k, \nabla f(x_k)\rangle + A_k\langle x_k, \nabla f(x_k)\rangle \\
&\quad + (A_{k+1} - A_k)(f(x_k) - f_*) + A_k(f(x_k) - f(y_k)) \\
&= (A_{k+1} - A_k)\langle x_* - x_k, \nabla f(x_k)\rangle + A_k\langle y_k - x_k, \nabla f(x_k)\rangle \\
&\quad + (A_{k+1} - A_k)(f(x_k) - f_*) + A_k(f(x_k) - f(y_k)) \\
&= (A_{k+1} - A_k)\left\{f(x_k) - f(x_*) + \langle x_* - x_k, \nabla f(x_k)\rangle\right\} \\
&\quad + A_k\left\{f(x_k) - f(y_k) + \langle y_k - x_k, \nabla f(x_k)\rangle\right\} \\
&\leq 0
\end{aligned}$$

by convexity of f. Finally, we have proved that $E_{k+1} - E_k \leq \varepsilon_{k+1}$. □

The previous Lemma proved an upper bound on the increments of the form

$$E_{k+1} - E_k \leq \frac{1}{2}(A_{k+1} - A_k)^2\|\nabla f(x_k)\|^2 + A_{k+1}(f(y_{k+1}) - f(x_k)).$$

Next, we want to apply L-smoothness in order to derive

$$f(y_{k+1}) - f(x_k) \propto -\|\nabla f(x_k)\|^2$$

and therefore, to imply the decrease of $(E_k)_{k\in\mathbb{N}}$. In particular, we will then obtain convergence of the error $e_k = f(y_k) - f_*$ of the order $\mathcal{O}(\frac{1}{(k+1)k})$. Note that for gradient descent under similar assumptions on f we did only prove convergence of order $\mathcal{O}(\frac{1}{k+1})$.

Theorem 3.3.4 (NAM for convex and smooth cost function)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and convex with $\min_{x\in\mathbb{R}^d} f > -\infty$, and assume there exists at least one global minimum $x_* \in \mathbb{R}^d$ of f. Moreover, let $(x_k, y_k, z_k)_{k\in\mathbb{N}}$ be generated by (3.2) with parameters $\alpha_k = \frac{1}{L}$, $\gamma_k = A_{k+1} - A_k$*

and $\tau_k = \frac{\gamma_k}{A_{k+1}} = \frac{A_{k+1}-A_k}{A_{k+1}} \in (0,1)$. *i.e.*

$$\begin{aligned} x_k &= y_k + \frac{A_{k+1}-A_k}{A_{k+1}}(z_k - y_k), \\ y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k), \\ z_{k+1} &= z_k - (A_{k+1}-A_k)\nabla f(x_k), \end{aligned}$$

initialized with $(y_0, z_0) \in \mathbb{R}^d \times \mathbb{R}^d$. *Then the increments of the sequence* $(E_k)_{k\in\mathbb{N}}$ *defined in* (3.3) *satisfy*

$$E_{k+1} - E_k \leq \left(\frac{1}{2}(A_{k+1}-A_k)^2 - \frac{1}{2L}A_{k+1}\right)\|\nabla f(x_k)\|^2$$

for all $k \in \mathbb{N}$. *For the particular choice* $A_k = \frac{1}{4L}(k+1)k$, $k \geq 1$, *and* $A_0 = A_1$, *we obtain*

$$e_k = f(y_k) - f_* \leq \frac{4LE_0}{(k+1)k}, \quad k \geq 1.$$

Proof. We define $G_L(x_k) = x_k - \frac{1}{L}\nabla f(x_k)$ and apply L-smoothness of f to deduce

$$\begin{aligned} f(G_L(x_k)) &\leq f(x_k) + \langle \nabla f(x_k), G_L(x_k) - x_k\rangle + \frac{L}{2}\|G_L(x_k) - x_k\|^2 \\ &= f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2, \end{aligned}$$

implying that

$$f(y_{k+1}) - f(x_k) \leq -\frac{1}{2L}\|\nabla f(x_k)\|^2.$$

With the previous Lemma 3.3.2 we obtain

$$E_{k+1} - E_k \leq \left(\frac{1}{2}(A_{k+1}-A_k)^2 - \frac{1}{2L}A_{k+1}\right)\|\nabla f(x_k)\|^2.$$

With $A_k = \frac{1}{4L}(k+1)k$ it follows that

$$\frac{1}{2}(A_{k+1}-A_k)^2 - \frac{1}{2L}A_{k+1} \leq 0,$$

since

$$\frac{(A_{k+1}-A_k)^2}{A_{k+1}} = \frac{1}{L}\frac{(k+1)^2}{(k+2)(k+1)} \leq \frac{1}{L}.$$

It follows that $E_{k+1} - E_k \leq 0$ and therefore, we have

$$\frac{1}{2}\|z_k - x_*\|^2 + A_k(f(y_k) - f_*) \leq E_0$$

which implies convergence

$$f(y_k) - f_* \leq \frac{4LE_0}{(k+1)k}, \quad k \geq 1.$$

□

Remark 3.3.5. To draw the connection to Algorithm 5 we have to choose

$$\beta_k = \frac{\tau_{k+1}(1-\tau_k)}{\tau_k},$$

where

$$\tau_k = \frac{A_{k+1} - A_k}{A_{k+1}} = \frac{2k+2}{(k+1)(k+2)} = \frac{2}{k+2}$$

and hence,

$$\beta_k = \frac{\frac{2}{k+3}(1-\frac{2}{k+2})}{\frac{2}{k+2}} = \frac{k}{k+3}.$$

Note that

$$\gamma_k\tau_k = \frac{(A_{k+1}-A_k)^2}{A_{k+1}} = \frac{1}{4L}\frac{(2k+2)^2}{(k+1)(k+2)} \to \frac{1}{L},$$

which is in minor contrast to the choice $\gamma_k\tau_k = \alpha_k = \frac{1}{L}$ in Example 3.3.1.

3.3.2 Convergence for strongly convex and smooth function

We have seen that NAM leads to an improvement of the convergence for convex and smooth functions. We now want to consider the strongly convex and smooth setting, where we again show improvement compared to the optimal convergence behavior of the gradient descent method discussed in Remark 2.3.18. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a μ-strongly convex and L-smooth cost function. Motivated from [25] we again consider a slightly different system for NAM of three variable $(x_k, y_k, z_k)_{k\in\mathbb{N}}$ generated by

$$\begin{aligned} x_k &= \frac{\tau}{1+\tau}z_k + \frac{1}{1+\tau}y_k \\ y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ z_{k+1} &= z_k + \tau(x_k - z_k) - \frac{\tau}{\mu}\nabla f(x_k) \end{aligned} \tag{3.4}$$

with $\tau > 0$. Similarly as before, the authors in [25] consider a more general family of accelerated gradient methods, but due to the scope of this lecture course we again consider only the simplified formulation. In the following example, we make the connection to NAM formulated in Algorithm 5. We show that the system (3.4) with fixed choice $\tau = \sqrt{\frac{\mu}{L}}$ can be viewed as special case of Algorithm 5 with fixed $\alpha = \frac{1}{L}$ and $\beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$.

Example 3.3.6. *We formulate system (3.4) as update of two variables $(x_k, y_k)_{k\in\mathbb{N}}$ written as*

$$\begin{aligned} y_{k+1} &= x_k - \alpha_k \nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \beta_k (y_{k+1} - y_k) \end{aligned}.$$

Firstly, we observe that

$$(1-\tau)z_k = (1-\tau)\left(\frac{1+\tau}{\tau}x_k - \frac{1}{\tau}y_k\right) = \left(\frac{1}{\tau} - \tau\right)x_k + \frac{\tau-1}{\tau}y_k,$$

such that we can write the update of z_{k+1} through

$$\begin{aligned} z_{k+1} &= z_k + \tau(x_k - z_k) - \frac{\tau}{\mu}\nabla f(x_k) \\ &= (1-\tau)z_k + \tau x_k - \frac{\tau}{\mu}\nabla f(x_k) \\ &= \left(\frac{1}{\tau}x_k - \tau + \tau\right)x_k + \left(1 - \frac{1}{\tau}\right)y_k - \frac{\tau}{\tau}\frac{\tau}{\mu}\nabla f(x_k). \end{aligned}$$

We set $\tau = \sqrt{\frac{\mu}{L}}$ such that $\frac{\tau^2}{\mu} = \frac{1}{L}$ and therefore,

$$\begin{aligned} z_{k+1} = \frac{1}{\tau}x_k - \frac{1}{\tau}\frac{1}{L}\nabla f(x_k) + \left(1 - \frac{1}{\tau}\right)y_k &= y_k + \frac{1}{\tau}\left(x_k - \frac{1}{L}\nabla f(x_k) - y_k\right) \\ &= y_k + \frac{1}{\tau}(y_{k+1} - y_k). \end{aligned}$$

We plug this into the update of x_{k+1} to obtain

$$\begin{aligned} x_{k+1} = \frac{\tau}{\tau+1}z_{k+1} + \frac{1}{\tau+1}y_{k+1} &= \frac{\tau}{\tau+1}\left(y_k + \frac{1}{\tau}(y_{k+1} - y_k)\right) + \frac{1}{\tau+1}y_{k+1} \\ &= y_{k+1} + \frac{\tau-1}{\tau+1}y_k + \frac{1-\tau}{1+\tau} \\ &= y_{k+1} + \frac{1-\tau}{1+\tau}(y_{k+1} - y_k). \end{aligned}$$

Finally, with the choice $\tau = \sqrt{\frac{\mu}{L}}$ we can write the iterative update scheme as

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}(y_{k+1} - y_k), \end{aligned}$$

such that we recover Algorithm 5 with fixed $\alpha_k = \alpha = \frac{1}{L}$ and $\beta_k = \beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$.

We are now ready to prove linear convergence of Algorithm 5 through system (3.4) in the strongly

convex and smooth setting.

Theorem 3.3.7 (NAM for strongly convex and smooth cost function)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex and L-smooth with $L > \mu$, and let $x_* \in \mathbb{R}^d$ be the corresponding unique global minimum of f. Moreover, let $(x_k, y_k, z_k)_{k\in\mathbb{N}}$ be generated by (3.4) with $\tau = \sqrt{\frac{\mu}{L}} \in (0,1)$ and initialized by $(y_0, z_0) \in \mathbb{R}^d \times \mathbb{R}^d$. Then NAM converges linearly in the sense that*

$$e_k := f(y_k) - f(x_*) + \frac{\mu}{2}\|z_k - x_*\|^2 \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \left(f(y_0) - f(x_*) + \frac{\mu}{2}\|z_0 - x_*\|^2\right).$$

Proof. We define $e_k := f(y_k) - f(x_*) + \frac{\mu}{2}\|z_k - x_*\|^2$ and aim to prove

$$e_{k+1} \leq (1-\tau)e_k.$$

Firstly, applying L-smoothness of f gives

$$f(y_{k+1}) \leq f(x_k) + \langle \nabla f(x_x), y_{k+1} - x_k\rangle + \frac{L}{2}\|x_k - y_{k+1}\|^2 = f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2. \tag{3.5}$$

Next, we consider the update of $e_k^{(1)} = \|z_k - x_*\|^2$:

$$\begin{aligned}
e_{k+1}^{(1)} &= \|z_k + \tau(x_k - z_k) - \frac{\tau}{\mu}\nabla f(x_k) - x_*\|^2 \\
&= \|z_k - x_*\|^2 + 2\langle \tau(x_k - z_k) - \frac{\tau}{\mu}\nabla f(x_k), z_k - x_*\rangle + \| \underbrace{\tau(x_k - z_k) - \frac{\tau}{\mu}\nabla f(x_k)}_{=z_{k+1}-z_k} \|^2 \\
&= e_k^{(1)} + 2\tau\langle x_k - z_k, \underbrace{z_k - x_*}_{=z_k - x_k + x_k - x_*}\rangle - 2\frac{\tau}{\mu}\langle \nabla f(x_k), \underbrace{z_k - x_*}_{=z_k - x_k + x_k - x_*}\rangle + \|z_{k+1} - z_k\|^2 \\
&= e_k^{(1)} + \|z_{k+1} - z_k\|^2 + 2\tau\langle x_k - z_k, z_k - x_k\rangle - 2\frac{\tau}{\mu}\langle \nabla f(x_k), z_k - x_k\rangle \\
&\quad + 2\tau\langle x_k - z_k, x_k - x_*\rangle - 2\frac{\tau}{\mu}\langle \nabla f(x_k), x_k - x_*\rangle
\end{aligned}$$

Recall that by strong convexity it holds true that

$$f(x_*) - f(x_k) \geq \langle x_* - x_k, \nabla f(x_k)\rangle + \frac{\mu}{2}\|x_k - x_*\|^2,$$

such that

$$\begin{aligned}
e_{k+1}^{(1)} \leq e_k^{(1)} &+ \|z_{k+1} - z_k\|^2 + 2\tau\langle x_k - z_k, z_k - x_k\rangle - 2\frac{\tau}{\mu}\langle \nabla f(x_k), z_k - x_k\rangle \\
&+ 2\frac{\tau}{\mu}(f(x_*) - f(x_k) - \frac{\mu}{2}\|x_k - x_*\|^2) + 2\tau\langle x_k - z_k, x_k - x_*\rangle.
\end{aligned}$$

We observe that

$$\tau(x_k - z_k) = \tau x_k - ((1+\tau)x_k - y_k) = y_k - x_k$$

and obtain

$$\begin{aligned} e_{k+1}^{(1)} \leq e_k^{(1)} + \|z_{k+1} - z_k\|^2 - 2\frac{\tau}{\mu}(f(x_k) - f(x_*)) + \frac{2}{\mu}\langle \nabla f(x_k), y_k - x_k\rangle \\ + 2\tau\langle x_k - z_k, x_k - x_*\rangle - 2\tau\|x_k - z_k\|^2 - \tau\|x_k - x_*\|^2. \end{aligned}$$

Note that

$$\begin{aligned} 2\tau\langle x_k - z_k, x_k - x_*\rangle - 2\tau\|x_k - z_k\|^2 - \tau\|x_k - x_*\|^2 &= -\tau\|x_k - x_* - (x_k - z_k)\|^2 - \tau\|x_k - z_k\|^2 \\ &= -\tau e_k^{(1)} - \tau\|x_k - z_k\|^2, \end{aligned}$$

which yields

$$e_{k+1}^{(1)} \leq (1-\tau)e_k^{(1)} + \|z_{k+1} - z_k\|^2 - 2\frac{\tau}{\mu}(f(x_k) - f(x_*)) + \frac{2}{\mu}\langle \nabla f(x_k), y_k - x_k\rangle - \tau\|x_k - z_k\|^2.$$

Finally, with $e_k^{(2)} = f(y_k) - f(x_*)$ we consider the evolution of e_k through

$$\begin{aligned} e_{k+1} &= \frac{\mu}{2}e_{k+1}^{(1)} + e_{k+1}^{(2)} \leq (1-\tau)\frac{\mu}{2}e_k^{(1)} + \frac{\mu}{2}\|z_{k+1} - z_k\|^2 - \tau(f(x_k) - f(x_*)) + \langle \nabla f(x_k), y_k - x_k\rangle \\ &\quad - \tau\frac{\mu}{2}\|x_k - z_k\|^2 + f(y_{k+1}) - f(x_*) \\ &\leq (1-\tau)\frac{\mu}{2}e_k^{(1)} + \frac{\mu}{2}\|z_{k+1} - z_k\|^2 - \tau(f(x_k) - f(x_*)) + \langle \nabla f(x_k), y_k - x_k\rangle \\ &\quad - \tau\frac{\mu}{2}\|x_k - z_k\|^2 + f(x_k) - f(x_*) - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \{(1-\tau)e_k^{(2)} - (1-\tau)e_k^{(2)}\}\} \\ &= (1-\tau)e_k - (1-\tau)(f(y_k) - f(x_*)) - \tau(f(x_k) - f(x_*)) + f(x_k) - f(x_*) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \\ &\quad + \frac{\mu}{2}\|z_{k+1} - z_k\|^2 - \tau\frac{\mu}{2}\|x_k - z_k\|^2 + \langle \nabla f(x_k), y_k - x_k\rangle \\ &=: (1-\tau)e_k + \mathcal{R}, \end{aligned}$$

where we have used (3.5) in the first inequality and added a zero $(1-\tau)(f(y_k) - f(x_*)) - (1-\tau)(f(y_k) - f(x_*)) = 0$. It is left to prove that $\mathcal{R} \leq 0$. First note, that $\mathcal{R}$ simplifies to

$$\mathcal{R} = (1-\tau)(f(x_k) - f(y_k)) - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{\mu}{2}\|z_{k+1} - z_k\|^2 - \tau\frac{\mu}{2}\|x_k - z_k\|^2 + (1-\tau+\tau)\langle \nabla f(x_k), y_k - x_k\rangle.$$

We apply strong convexity in form of

$$\langle \nabla f(x_k), y_k - x_k\rangle \leq f(y_k) - f(x_k) - \frac{\mu}{2}\|y_k - x_k\|^2$$

such that

$$\begin{aligned}\mathcal{R} &\leq (1-\tau)(f(x_k)-f(y_k))+(1-\tau)(f(y_k)-f(x_k))-\frac{\mu}{2}\|y_k-z_k\|^2\\ &\quad-\frac{1}{2L}\|\nabla f(x_k)\|^2+\frac{\mu}{2}\|z_{k+1}-z_k\|^2-\tau\frac{\mu}{2}\|x_k-z_k\|^2+\tau\langle\nabla f(x_k),y_k-x_k\rangle\\ &=\tau\langle\nabla f(x_k),y_k-x_k\rangle-(1-\tau)\frac{\mu}{2}\|y_k-x_k\|^2-\frac{1}{2L}\|\nabla f(x_k)\|^2\\ &\quad+\frac{\mu}{2}\|z_{k+1}-z_k\|^2-\tau\frac{\mu}{2}\|x_k-z_k\|^2.\end{aligned}$$

Recall that we have used $z_{k+1}-z_k=\tau(x_k-z_k)-\frac{\tau}{\mu}\nabla f(x_k)$, which with $\tau(x_k-z_k)=y_k-x_k$ gives

$$\frac{\mu}{2}\|z_{k+1}-z_k\|^2=\frac{\mu}{2}\|y_k-x_k\|^2-\tau\langle\nabla f(x_k),y_k-x_k\rangle+\frac{\tau^2}{2\mu}\|\nabla f(x_k)\|^2,$$

and therefore, we obtain

$$\begin{aligned}\mathcal{R} &\leq -(1-\tau)\frac{\mu}{2}\|y_k-x_k\|^2+\frac{\mu}{2}\|y_k-x_k\|^2+\tau\langle\nabla f(x_k),y_k-x_k\rangle-\tau\langle\nabla f(x_k),y_k-x_k\rangle\\ &\quad+\frac{\tau^2}{2\mu}\|\nabla f(x_k)\|^2-\frac{1}{2L}\|\nabla f(x_k)\|^2-\tau\frac{\mu}{2}\frac{\tau}{\tau}\|x_k-z_k\|^2\\ &=\tau\frac{\mu}{2}\|y_k-x_k\|^2-\frac{\mu}{2\tau}\|\tau(x_k-z_k)\|^2+\left(\frac{\tau^2}{2\mu}-\frac{1}{2L}\right)\|\nabla f(x_k)\|^2\\ &=\left(\frac{\tau\mu}{2}-\frac{\mu}{2\tau}\right)\|y_k-x_k\|^2+\left(\frac{\tau^2}{2\mu}-\frac{1}{2L}\right)\|\nabla f(x_k)\|^2,\end{aligned}$$

where we have used again that $\tau(x_k-z_k)=y_k-x_k$. With the choice $\tau=\sqrt{\frac{\mu}{L}}\in(0,1)$ we observe that

$$\frac{\tau^2}{2\mu}-\frac{1}{2L}=\frac{1}{2L}-\frac{1}{2L}=0$$

and

$$\frac{\tau\mu}{2}-\frac{\mu}{2\tau}=\frac{\mu}{2}(\tau-\frac{1}{\tau})\leq 0.$$

This proves that $\mathcal{R}\leq 0$ and the assertion follows:

$$e_{k+1}\leq(1-\tau)e_k=\left(1-\sqrt{\frac{\mu}{L}}\right)e_k.$$

□

Remark 3.3.8. We can rewrite the error bound for NAM of the previous theorem through

$$e_k=\frac{\mu}{2}\|z_k-x_*\|^2+f(y_k)-f(x_*)\leq(1-\tau)^k e_0=\left(1-\frac{1}{\sqrt{\kappa}}\right)^k e_0=\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}}\right)^k e_0.$$

The corresponding optimal convergence rate of the simple gradient descent scheme was given by

$$e_k^{\mathrm{GD}} := \|x_k - x_*\|^2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} e_0^{\mathrm{GD}}.$$

In Figure 3.3 we compare both rates of convergence for increasing condition number $\kappa = \frac{L}{\mu}$ illustrating the improvement through NAM.

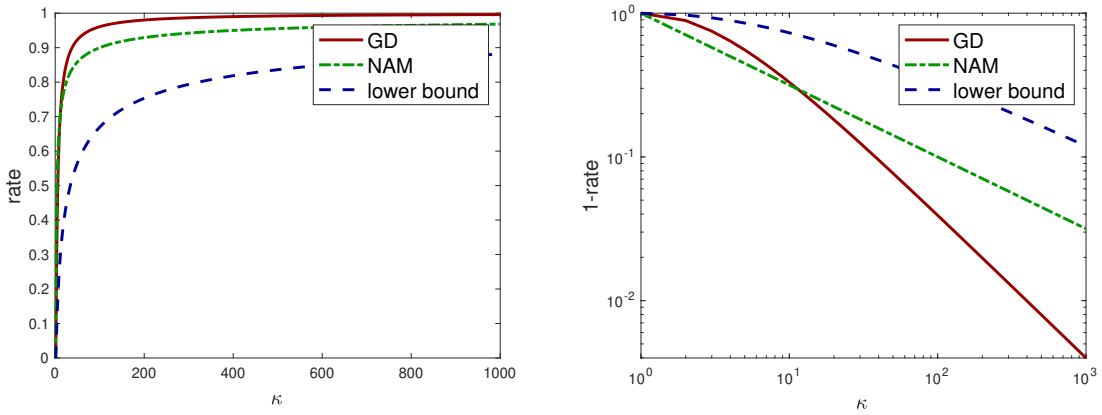

Figure 3.3: Illustration of the linear convergence rate depending on the condition number $\kappa = \frac{\mu}{L}$ for GD and NAM. The left plot shows the convergence rate $c^{\mathrm{GD}}(\kappa) = \left(\frac{\kappa-1}{\kappa+1}\right)^2$ and $c^{\mathrm{NAM}}(\kappa) = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}}\right)$, whereas the right plot shows the difference to 1, i.e. $1 - c(\kappa)$, in logarithmic scale.

4

Stochastic approximation in Optimization

We will start with a motivating example to introduce the problem of minimizing expected and empirical risk. This will motivate the consideration of stochastic variants of gradient descent methods.

Example 4.0.1. *We revisit the regression problem discussed in Chapter 1, which arises in supervised learning. Recall that we aim to approximate an unknown model*

$$z \mapsto \varphi(z) = y, \quad \varphi : \mathbb{R}^{d_z} \to \mathbb{R}^{d_y}$$

through a parametrized family of functions $g_\theta : \mathbb{R}^{d_z} \to \mathbb{R}^{d_y}$, $\theta \in \Theta$. Given a training data set $\{(z^{(i)}, y^{(i)})\}_{i=1}^N$, we have described the training task as optimization problem

$$\min_{\theta \in \Theta} \; f_N(\theta, \{(z^{(i)}, y^{(i)})\}_{i=1}^N),$$

where $f_N : \Theta \times (\times_{i=1}^N (\mathbb{R}^{d_z} \times \mathbb{R}^{d_y}) \to \mathbb{R}$ denotes the to cost function. In the example of regression, we have considered

$$f_N(\theta, \{(z^{(i)}, y^{(i)})\}_{i=1}^N) = \frac{1}{N} \sum_{i=1}^N \|g_\theta(z^{(i)}) - y^{(i)}\|^2 + \mathcal{R}(\theta), \tag{4.1}$$

where $\mathcal{R} : \Theta \to \mathbb{R}$ is some regularization function. We aim to incorporate a probabilistic framework in order to introduce (empirical) risk minimization. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be our underlying probability space. We model the input and output variable as jointly distributed random variables (Z, Y) on $(\Omega, \mathcal{A}, \mathbb{P})$ with state space $(\mathbb{R}^{d_z} \times \mathbb{R}^{d_y}, \mathcal{B}(\mathbb{R}^{d_z}) \otimes \mathcal{B}(\mathbb{R}^{d_y}))$. The goal is to find $\theta \in \Theta$ such that g_θ represents the stochastic model

$$Y = \varphi(Z)[+\xi], \quad Z \sim \mu_Z,$$

where ξ denotes possible noise.

Challenge: *We assume that distribution μ_Z of Z and the joint distribution $\mu_{(Z,Y)}$ respectively are unknown. Instead, we assume that we are able to generate (arbitrarily many) i.i.d. sample*

$(Z^{(i)}, Y^{(i)}) \sim \mu_{(Z,Y)}$.

We are now interested in how to choose an optimal approximation g_θ. The natural extension of the cost function (4.1) to the probabilistic setting is the task of minimizing the cost function

$$F(\theta) = \mathbb{E}_{(Z,Y)\sim\mu_{(Z,Y)}}[\|g_\theta(Z) - Y\|^2] + \mathcal{R}(\theta),$$

where $\mathbb{E}_{(Z,Y)\sim\mu_{(Z,Y)}}$ denotes the expectation w.r.t. (Z,Y). Given a training data set of i.i.d. random variables $\{(Z^{(i)}, Y^{(i)})\}_{i=1}^N$ distributed according to $(Z^{(1)}, Y^{(1)}) \sim \mu_{(Z,Y)}$, we can apply a Monte Carlo approximation of the above expectation

$$\mathbb{E}_{(Z,Y)\sim\mu_{(Z,Y)}}[\|g_\theta(Z) - Y\|^2] \approx \frac{1}{N}\sum_{i=1}^N \|g_\theta(Z^{(i)}) - Y^{(i)}\|^2$$

to construct the empirical cost function

$$F_N(\theta) = \frac{1}{N}\sum_{i=1}^N \|g_\theta(Z^{(i)}) - Y^{(i)}\|^2 + \mathcal{R}(\theta),$$

which coincides with (4.1).

Motivated by this example we introduce the definition of (empirical) risk minimization problems.

Definition 4.0.2. Let $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ be $\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}^p)/\mathcal{B}(\mathbb{R})$ measurable and $Z : \Omega \to \mathbb{R}^p$ a random variable with distribution μ such that $\mathbb{E}[|f(x,Z)|] < \infty$ for all $x \in \mathbb{R}^d$.

1. We define the *expected risk* $F : \mathbb{R}^d \to \mathbb{R}$ as

$$F(x) = \mathbb{E}_{Z\sim\mu}[f(x,Z)] =: \int_{\mathbb{R}^p} f(x,z)\,\mu(\mathrm{d}z), \quad x \in \mathbb{R}^d.$$

   We call the minimization problem

$$\min_{x\in\mathbb{R}^d} F(x), \quad F(x) = \mathbb{E}_{Z\sim\mu}[f(x,Z)]$$

   risk minimization problem.

2. Let $Z_1, \dots, Z_N$ be i.i.d. random variables with $Z_1 \sim \mu$. We define the *empirical risk* $F_N : \mathbb{R}^d \to \mathbb{R}$ by

$$F_N(x) = \frac{1}{N}\sum_{i=1}^N f(x, Z^{(i)}).$$

   The empirical risk is sometimes also called *population risk*. We call the minimization

problem

$$\min_{x\in\mathbb{R}^d} F_N(x), \quad F_N(x) = \frac{1}{N}\sum_{i=1}^{N} f(x, Z^{(i)})$$

empirical risk minimization problem.

This chapter will focus on analyzing stochastic optimization methods for solving (empirical and expected) risk minimization problems. It's important to note that this discussion only addresses a subset of the typical challenges that arise in machine/supervised learning.

Remark 4.0.3 (Statistical learning perspective)**.** For example, in the area of statistical learning among other questions one is interested in the consistency of solutions of the empirical risk minimization problem. Let $\widehat{X}_N$ be the minimizer of the empirical risk and x_* be the minimizer of the corresponding expected risk (provided both exist), then we can decompose the error

$$F(\widehat{X}_N)-F(x_*) = F(\widehat{X}_N)-F_N(\widehat{X}_N)+F_N(\widehat{X}_N)-F_N(x_*)+F_N(x_*)-F(x_*) \leq 2\sup_{x\in\mathbb{R}^d}|F_N(x)-F(x)|.$$

We emphasize that F_N as function depends on the random variables $Z^{(1)},\dots,Z^{(N)}$ and is therefore random. Hence, the minimizer $\widehat{X}_N$ itself is a random variable. In statistical learning theory one concerns about questions such as the consistency of $\widehat{X}_N$ for number of data points N approaching infinity.

Remark 4.0.4 (Inverse problem perspective)**.** For a fixed number of data points $N \in \mathbb{N}$ the empirical risk minimization problem is typically ill-posed and it is necessary to include regularization. This is the topic of the lecture course *inverse problems*. As motivation, we will treat the training task of supervised learning as an inverse problem. Recall, that we are interested to approximate a model $\varphi : \mathbb{R}^{d_z} \to \mathbb{R}^{d_y}$ by a parametrized family of functions $g_\theta : \mathbb{R}^{d_z} \to \mathbb{R}^{d_y}$, $\theta \in \Theta$. Given a training data set $\{(Z^{(i)}, Y^{(i)})\}$ we want to minimize the empirical risk

$$\min_{\theta\in\Theta} \frac{1}{N}\|g_\theta(Z^{(i)}) - Y^{(i)}\|^2.$$

An alternative persepective/interpretation is the following. Define a *forward map* $H : \Theta \to \mathbb{R}^{N\cdot d_y}$,

$$H(\theta) := (g_\theta(Z^{(1)},\dots,g_\theta(Z^{(N)}))^\top \in \mathbb{R}^{N\cdot d_y}.$$

With observations $\widehat{Y} = (Y^{(1)},\dots,Y^{(N)})^\top \in \mathbb{R}^{N\cdot d_y}$, we aim to solve the inverse problem of recovering the parameter $\theta \in \Theta$ such that

$$\widehat{Y} = H(\theta). \tag{4.2}$$

This problem is typically ill-posed (in the sense of a well-posed problem following Hadamard [8]), due to the following reasons:

1. There might not exist any $\theta \in \Theta$ solving (4.2).

2. The solution of (4.2) is not necessarily unique, i.e. there might be $\theta_1, \theta_2 \in \Theta$ with $H(\theta_1) = H(\theta_2) = \widehat{Y}$.

3. The solution of (4.2) might be instable w.r.t. changes in $\widehat{Y}$ (e.g. due to measurement noise).

Therefore, it is not the best idea to simply solve $\min_{\theta\in\Theta} \|H(\theta) - \widehat{Y}\|^2$. We illustrate the resulting issues in Figure 4.1–4.3. In inverse problems a large focus lies in the study of regularization methods, which can, for example, be incorporated as a penalty function. Instead of simply minimizing the data misfit functional, one considers solving the regularized optimization problem

$$\min_{\theta\in\Theta} \|H(\theta) - \widehat{Y}\|^2 + \mathcal{R}(\theta),$$

where $\mathcal{R} : \Theta \to \mathbb{R}$ is a regularization function.

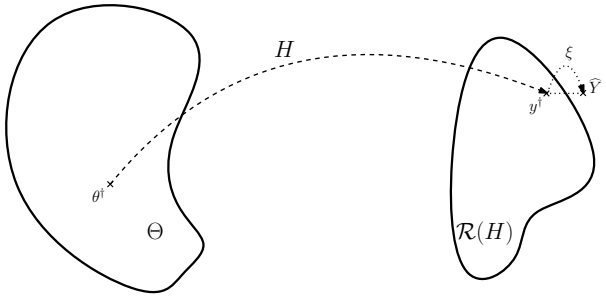

Figure 4.1: Illustration of ill-posedness through observational noise. The occurrence of noise might shift the observed data outside of the range of the forward map H.

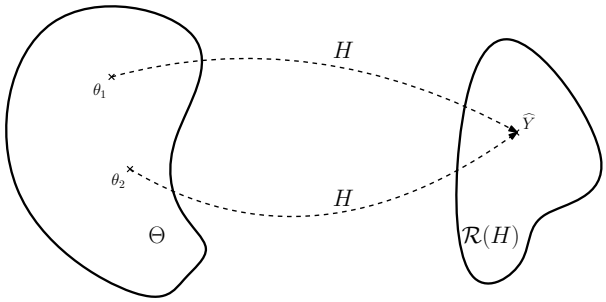

Figure 4.2: Illustration of ill-posedness through multiple solutions. Two different parameters $\theta_1, \theta_2 \in \Theta$ might map onto the observed data $\widehat{Y}$.

In the following chapter we are going to study optimization methods for solving the expected and empirical risk minimization problem. We do not consider questions around generalization and regularization, which are beyond the scope of this lecture course.

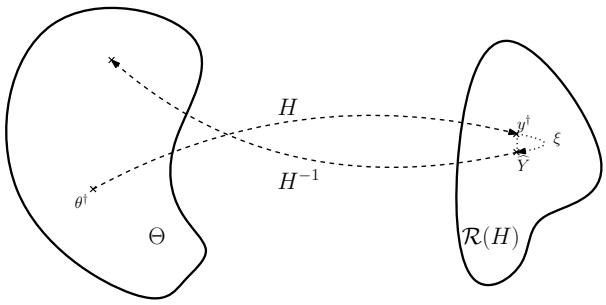

Figure 4.3: Illustration of ill-posedness through discontinuity. Even if H is invertible, the instability may occur in the solution of the inverse problem resulting from possible discontinuity of the inverse operator.

4.1 Stochastic gradient descent method (SGD)

We want to introduce and analyze a stochastic variant of the gradient descent method for solving the expected and empirical risk minimization problem. For a comprehensive overview of stochastic gradient methods, the interested reader may refer to [4, 7, 22].
In the following, let $(\Omega, \mathcal{A}, \mathbb{P})$ be the underlying probability space, $Z : \Omega \to \mathbb{R}^p$ be a random variable on $(\Omega, \mathcal{A}, \mathbb{P})$ with distribution μ_Z. We are interested in solving the optimization problem

$$\min_{x \in \mathbb{R}^d} F(x),$$

where the cost function $F : \mathbb{R}^d \to \mathbb{R}$ is defined as the expectation function in form

$$F(x) = \mathbb{E}_{Z \sim \mu_Z}[f(x, Z)] = \int_{\mathbb{R}^p} f(x, z) \mu_Z(\mathrm{d}z), \quad x \in \mathbb{R}^d,$$

for a function $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$. Throughout this lecture course we make the following assumption:

Assumption 4.1.1. 1. The function $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ is $\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}^p)/\mathcal{B}(\mathbb{R})$-measurable.

2. For every $z \in \mathbb{R}^p$ the function $x \mapsto f(x, z)$ is continuously differentiable.

3. For every $x \in \mathbb{R}^d$ we have

$$\mathbb{E}[|f(x, Z)| + \|\nabla_x f(x, Z)\|] < \infty$$

and

$$\mathbb{E}[\|\nabla_x f(x, Z) - \mathbb{E}[\nabla_x f(x, Z)]\|^q \leq b(1 + \|x\|^q)$$

for some $q \geq 1$ and $b > 0$.

In order to apply the gradient methods introduced in the previous sections, we run into the

question: How do we compute the derivative of F? More fundamentally, we can ask under which conditions F is differentiable?
The following result from [9] addresses the question about differentiability of F and illustrates that we can use the random variable Z to construct an unbiased estimator of $\nabla_x F(x)$ for every fixed $x \in \mathbb{R}^d$. The result provides confirmation that under Assumption 4.1.1 we are allowed to interchange derivative and expectation for the computation of $\nabla_x F(x)$.

Lemma 4.1.2 (Lemma 4.8 in [9])**.** Suppose Assumption 4.1.1 is satisfied, then it holds true that

1. the function $F(x) = \mathbb{E}[f(x, Z)]$ is continuously differentiable,
2. $\nabla_x f(x, Z)$ is an unbiased estimator of $\nabla_x F(x)$ for every $x \in \mathbb{R}^d$, i.e. it holds true that

$$\nabla_x F(x) = \mathbb{E}[\nabla_x f(x, Z)].$$

The stochastic gradient descent (SGD) method serves as an approximation to the gradient descent method, where each update guides the current iteration in the direction of a (stochastic) approximation of the negative gradient. Since we observed that $\nabla_x f(x, Z)$ can be viewed as unbiased estimator of $\nabla_x F(x)$, we expect that the scheme will perform well in average. As part of this lecture course we will verify this expectation. The algorithm is formulated in Algorithm 6. Throughout this chapter, we assume the following scenario.

Assumption 4.1.3. We assume that we can generate arbitrarily many i.i.d. samples according to μ_Z. This is, we assume that we have access to a sequence of i.i.d. random variables $(Z_k)_{k\in\mathbb{N}}$, where $Z_1 \sim \mu_Z$.

Remark 4.1.4. In Algorithm 6, for each $k \in \mathbb{N}$, the random variable Z_{k+1} is independent of X_m, $0 \leq m \leq k$. In the definition of G_k we have used the index $k+1$ for Z_{k+1} such that X_k remains measurable w.r.t. $\sigma(Z_m, m \leq k)$ for all $k \geq 1$. Indeed, we can then consider the natural filtration $\mathcal{F}_k^X = \sigma(X_m, 0 \leq m \leq k) = \sigma(X_0, Z_m, m \leq k)$.

Remark 4.1.5. We note that Algorithm 6 in practice is often used to minimize cost functions in form of empirical risks,

$$F_n(x) = \frac{1}{N}\sum_{i=1}^{N} f(x, z^{(i)}) = \mathbb{E}_{Z\sim\widehat{\mu}_N}[f(x, Z)],$$

where $\widehat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{z^{(i)}}$ denotes the empirical measure over the fixed data set $\{z^{(i)}, i = 1, \dots, N\}$. For the application of SGD in each iteration a random index $\mathfrak{i}_k \sim \mathcal{U}(\{1, \dots, N\})$ (or even a random

Algorithm 6 Stochastic gradient descent method (SGD)

Input:

- cost function $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$
- initial random variable $X_0 : \Omega \to \mathbb{R}^d$
- sequence of step sizes $(\alpha_k)_{k\in\mathbb{N}}$, $\alpha_k > 0$ (deterministic or $\mathcal{F}$-adapted)
- sequence of i.i.d. random variables $(Z_k)_{k\in\mathbb{N}}$ with $Z_1 \sim \mu_Z$.

set $k = 0$
while "convergence/stopping criterion not met" **do**
    approximate the gradient $\nabla_x F(X_k)$ through

$$G_k = \nabla_x f(X_k, Z_{k+1})$$

    set $X_{k+1} = X_k - \alpha_k G_k$, $k \mapsto k+1$
end while

index set $\mathfrak{I}_k \subset \{1, \dots, N\}$) is generated independently, and $\nabla_x F(x)$ is approximated by

$$G_k = \nabla_x f(x, z^{(\mathfrak{i}_k)}).$$

We describe this scheme in Algorithm 7.

Algorithm 7 SGD with finite data

Input:

- cost function $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$
- initial random variable $X_0 : \Omega \to \mathbb{R}^d$
- sequence of step sizes $(\alpha_k)_{k\in\mathbb{N}}$, $\alpha_k > 0$ (deterministic or $\mathcal{F}$-adapted)
- fixed realization of fixed deterministic data set $\{z^{(i)}\}_{i=1}^N$ with $z^{(i)} \in \mathbb{R}^p$.

set $k = 0$
while "convergence/stopping criterion not met" **do**
    generate independently $\mathfrak{i}_{k+1} \sim \mathcal{U}(\{1, \dots, N\})$
    approximate the gradient $\nabla_x F_N(X_k)$ through

$$G_k = \nabla_x f(X_k, z^{\mathfrak{i}_{k+1}})$$

    set $X_{k+1} = X_k - \alpha_k G_k$, $k \mapsto k+1$
end while

While Algorithm 6 generates a new independent realization of Z in each iteration, Algorithm 7 first fixes the number of realized independent samples of Z and then randomly iterates through this data set during SGD. The randomness in Algorithm 7 occurs through the realization of the random indices $(\mathfrak{i}_k)_{k\in\mathbb{N}}$. For both algorithms, the resulting iteration $(X_k)_{k\in\mathbb{N}}$ is a stochastic process,

which is path-wise constructed via $(Z_k)_{k\in\mathbb{N}}$ (Algorithm 6) and $(\mathfrak{i}_k)_{k\in\mathbb{N}}$ (Algorithm 7) respectively. In both cases, we can view the stochastic process as an adapted process with respect to the natural filtration

$$\mathcal{F}_k^X = \sigma(X_m, m \leq k) = \sigma(X_0,\ Z_m, m \leq k)$$

and

$$\mathcal{F}_k^X = \sigma(X_m, m \leq k) = \sigma(X_0,\ \mathfrak{i}_m, m \leq k).$$

This filtration will be relevant when analyzing the convergence behavior of SGD, where we take the expectation conditioned on the information from the past.
We will first discuss SGD from an stochastic approximation perspective motivated by the Robbins & Monro algorithm [21].

Outlook 1. (Robbins & Monro Algorithm) We consider a brief outlook to the original stochastic approximation method introduced in [21] aiming to root-finding. The authors considered the following question: Given a family of real-valued random variables $(Y_x)_{x\in\mathbb{R}}$ one can define the expectation function (in x)

$$M(x) = \mathbb{E}[Y_x] = \int_R y\, \mu(dy; x)$$

where $\mu(\cdot; x)$ denotes the distribution of Y_x. Given $z \in \mathbb{R}$, the aim is construct an algorithm to find the (unique) solution of the equation

$$M(x) = z.$$

The challenging aspect in this question is the unknown expectation function M. However, the authors assume that one is able to sample independently from $(Y_x)_{x\in\mathbb{R}}$ (or from $\mu(\cdot; x)$ respectively). The Robbins & Monro algorithm in its original from iterates through

$$X_{k+1} = X_k + \alpha_k(z - Y_k),$$

where Y_k, $k \in \mathbb{N}$ are independent random variables with distribution $\mu(\cdot; X_k)$ and $(\alpha_k)_{k\in\mathbb{N}}$ is a sequence of step sizes $\alpha_k > 0$. Robbins & Monro proved convergence of $(X_k)_{k\in\mathbb{N}}$ in L^2 towards z under certain assumptions on M and $(\mu(\cdot; x))_{x\in\mathbb{R}}$. Slightly later Blum [3] (1954) proved almost sure convergence under additional condition on the sequence of step sizes

$$\sum_{k=0}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty.$$

We will apply the Robbins & Siegmund Theorem based on Doob's supermartingale convergence theorem in order to derive an almost sure convergence result for SGD. However, this theorem

only leads to an asymptotic convergence result. In principle, we can view the approximation G_k in Algorithm 6 and 7 as a form of Monte Carlo approximation, which quantifies an error to the (exact/full) gradient descent method. A smaller variance of the estimator G_k suggests a better convergence behavior of SGD. Therefore, we will later consider a variance reduced version of the SGD method.
In order to analyze SGD, we rewrite the iterative scheme as follows

$$\begin{aligned} X_{k+1} = X_k - \alpha_k \nabla_x f(X_k, Z_{k+1}) &= X_k - \alpha_k \nabla_x F(X_k) + \alpha_k \left(\nabla_x F(X_k) - \nabla_x f(X_k, Z_{k+1})\right) \\ &=: X_k - \alpha_k \nabla_x F(X_k) + \alpha_k M_{k+1}. \end{aligned}$$

Recall, that we consider $(X_k)_{k\in\mathbb{N}}$ as an adapted process with respect to its natural filtration $\mathcal{F}_k = \sigma(X_0,\ Z_m, m \leq k)$. In the next section, it will turn out, that the process $(M_k)_{k\in\mathbb{N}}$ satisfies

$$\mathbb{E}[M_{k+1} \mid \mathcal{F}_k] = \mathbb{E}[\nabla_x F(X_k) - \nabla_x f(X_k, Z_{k+1}) \mid \mathcal{F}_k] = 0, \tag{4.3}$$

since we always assume that we are able to apply Lemma 4.1.2 to derive $\nabla_x F(x) = \mathbb{E}[\nabla_x f(x, Z)]$. Note that this property holds for fixed $x \in \mathbb{R}^d$ and we will extend this behavior to the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}]$, where it will be particularly relevant that $(X_k)_{k\in\mathbb{N}}$ is $\mathcal{F}$-adapted and Z_{k+1} is independent of $\mathcal{F}_k$.

4.1.1 Technical detail: Factorization of conditional expectation

In this section, we will discuss one important property which is needed for the verification of $\nabla_x f(X_k, Z_{k+1})$ as an unbiased estimator of $\nabla_x F(X_k)$ conditioned on the iterations of SGD represented through the natural filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$. In the literature of SGD it is often claimed that $\mathbb{E}[\nabla_x F(X_k) - \mathbb{E}[\nabla_x f(X_k, Z_{k+1}) \mid \mathcal{F}_k] = 0$, since it is assumed that $\nabla_x f(x, Z_{k+1})$ is an unbiased estimator of $\nabla_x F(x)$ for every $x \in \mathbb{R}^d$. However, in general this implication is non-trivial and we will need to investigate some more work. It turns out that the verification will require technical tools from measure theory such as the monotone class theorem in order to derive some form of factorization of the conditional expectation.
We will prove the following Lemma which can be found in [5, Proposition 1.12] and [9, Corollary 2.9].

Lemma 4.1.6. Let

- $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{F} \subset \mathcal{A}$ be some sub-σ-algebra of $\mathcal{A}$ on Ω,
- $(\mathbb{X}, \mathcal{X})$ and $(\mathbb{Y}, \mathcal{Y})$ be measurable spaces, $X : \Omega \to \mathbb{X}$ be $\mathcal{A}/\mathcal{X}$-measurable and independent of $\mathcal{F}$, and $Y : \Omega \to \mathbb{Y}$ be $\mathcal{F}/\mathcal{Y}$-measurable.

- $\Phi : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ be $(\mathcal{X} \otimes \mathcal{Y})/\mathcal{B}(\mathbb{R})$-measurable with $\mathbb{E}[|\Phi(X,Y)|] < \infty$, and $\mathbb{E}[|\Phi(X,y)|] < \infty$ for all $y \in \mathbb{Y}$.

With $\varphi : \mathbb{Y} \to \mathbb{R}$ defined by $\varphi(y) = \mathbb{E}[\Phi(X,y)]$, $y \in \mathbb{Y}$ we have

1. φ is $\mathcal{Y}/\mathcal{B}(\mathbb{R})$-measurable,
2. for all $A \in \mathcal{F}$ it holds true that

$$\mathbb{E}[\Phi(X,Y)\mathbb{1}_A] = \mathbb{E}[\varphi(Y)\mathbb{1}_A].$$

Remark 4.1.7. We consider Lemma 4.1.6 as a generalization of the following rule for conditional expectation. Let X, Y be real-valued random variables on $(\Omega, \mathcal{A}, \mathbb{P})$, $\mathcal{F} \subset \mathcal{A}$ be some sub-σ-algebra and X be independent of $\mathcal{F}$. Then we can compute the conditional expectation

$$\mathbb{E}[X \cdot Y \mid \mathcal{F}] = \mathbb{E}[X] \cdot \mathbb{E}[Y \mid \mathcal{F}].$$

For Y being $\mathcal{F}$-measurable, we deduce the assertion of Lemma 4.1.6 with $\Phi(x,y) = x \cdot y$.

Caution: We have some minor conflict of notation. When applying Lemma 4.1.6 to SGD, the random variables Z_k will take the role of X and the random variables X_k will take the role of Y.

Proof of Lemma 4.1.6. Firstly, note that it follows from Fubini's theorem that φ is $\mathcal{Y}/\mathcal{B}(\mathbb{R})$-measurable. Let us start with a brief outline of the proof for the second assertion:

Step 1 We prove that the assertion holds for $\Phi(x,y) = \mathbb{1}_B(x,y)$ with arbitrary $B \in \mathcal{X} \otimes \mathcal{Y}$.

Step 2 We use **step 1** in order to prove the assertion for step functions $\Phi_N(x,y) = \sum_{k=1}^N d_k \mathbb{1}_{D_k}(x,y)$ for $d_k \geq 0$ and $D_k \in \mathcal{X} \otimes \mathcal{Y}$.

Step 3 We prove the assertion for positive functions Φ, which can be expressed as limit of monotonically increasing step functions.

Step 4 We finish the proof by splitting Φ into positive and negative part.

We go through all of the steps.
Step 1: Let $B \in \mathcal{X} \otimes \mathcal{Y}$ be arbitrary and consider $\Phi(x,y) = \mathbb{1}_B(x,y)$ as well as $\varphi(y) = \mathbb{E}[\mathbb{1}_B(X,y)]$. We will apply the strategy discussed in Remark A.3.6 in order to verify that the property

$$\mathbb{E}[\mathbb{1}_B(X,Y)\mathbb{1}_A] = \mathbb{E}[\varphi(Y)\mathbb{1}_A], \quad \text{for all } A \in \mathcal{F}$$

is satisfied for any $B \in \mathcal{X} \otimes \mathcal{Y}$. We define the set

$$\mathcal{M} = \{D \in \mathcal{X} \otimes \mathcal{Y} \mid \mathbb{E}[\mathbb{1}_D(X,Y)\mathbb{1}_A] = \mathbb{E}[(\mathbb{E}[\mathbb{1}_D(X,y)])\,|_{y=Y}\mathbb{1}_A], \quad \text{for all } A \in \mathcal{F}\},$$

which will be the candidate for the Dynkin system. Moreover, we consider the $\cap$-stable generator

$$\mathcal{E} = \{S \in \mathcal{X} \otimes \mathcal{Y} \mid S = E_1 \times E_2,\ E_1 \in \mathcal{X},\ E_2 \in \mathcal{Y}\}$$

with $\sigma(\mathcal{E}) = \mathcal{X} \otimes \mathcal{Y}$. We need to prove that $\mathcal{M}$ is a Dynkin system, and $\mathcal{E} \subset \mathcal{M}$. We begin with the latter property. Let $E_1 \in \mathcal{X}$ and $E_2 \in \mathcal{Y}$, then we have for all $A \in \mathcal{F}$ that

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_{E_1 \times E_2}(X,Y)\mathbb{1}_A] &= \mathbb{P}(\{X \in E_1\} \cap \{Y \in E_2\} \cap A) \\
&\overset{X \text{ indep. } \mathcal{F}}{=} \mathbb{P}(\{X \in E_1\})\mathbb{P}(\{Y \in E_2\} \cap A) \\
&= \mathbb{E}[\mathbb{E}[\mathbb{1}_{E_1}(X)]\mathbb{1}_{E_2}(Y)\mathbb{1}_A] \\
&= \mathbb{E}[(\mathbb{E}[\mathbb{1}_{E_1}(X)\mathbb{1}_{E_2}(y)])\,|_{y=Y}\mathbb{1}_A] \\
&= \mathbb{E}[(\mathbb{E}[\mathbb{1}_{E_1 \times E_2}(X,y)])\,|_{y=Y}\mathbb{1}_A],
\end{aligned}$$

which proves that $E_1 \times E_2 \in \mathcal{M}$, i.e. $\mathcal{E} \subset \mathcal{M}$. Next, it is easy to verify, that $\mathcal{M}$ is a Dynkin system (we will skip the details here). Therefore, using Theorem A.3.5, we imply that

$$\mathcal{X} \otimes \mathcal{Y} = \sigma(\mathcal{E}) \overset{\mathcal{E}\ \cap\text{-stable}}{=} d(\mathcal{E}) \overset{\mathcal{E} \subset \mathcal{M}}{\subset} d(\mathcal{M}) \overset{\mathcal{M}\ \text{Dynkin system}}{=} \mathcal{M} \subset \mathcal{X} \otimes \mathcal{Y},$$

which means that for all $B \in \mathcal{X} \otimes \mathcal{Y}$ we have

$$\mathbb{E}[\mathbb{1}_B(X,Y)\mathbb{1}_A] = \mathbb{E}[\varphi(Y)\mathbb{1}_A], \quad \text{for all } A \in \mathcal{F},$$

which finishes **step 1**.

Step 2: Let $\Phi(x,y) = \sum_{k=1}^N d_k \mathbb{1}_{D_k}(x,y)$ for $d_k \geq 0$ and $D_k \in \mathcal{X} \otimes \mathcal{Y}$. We apply linearity of the expectation and **step 1** to obtain

$$\begin{aligned}
\mathbb{E}[\Phi(X,Y)\mathbb{1}_A] &= \sum_{k=1}^N d_k \mathbb{E}[\mathbb{1}_{D_k}(X,Y)\mathbb{1}_A] \\
&\overset{\textbf{step 1}}{=} \sum_{k=1}^N d_k \mathbb{E}[(\mathbb{E}[\mathbb{1}_{D_k}(X,y))\,|_{y=Y}\mathbb{1}_A] \\
&= \mathbb{E}[\left(\mathbb{E}[\sum_{k=1}^N d_k \mathbb{1}_{D_k}(X,y)\right)|_{y=Y}\mathbb{1}_A] = \mathbb{E}[(\mathbb{E}[\Phi(X,y)])\,|_{y=Y}\mathbb{1}_A] = \mathbb{E}[\varphi(Y)\mathbb{1}_A].
\end{aligned}$$

Step 3: Let $\Phi : \mathbb{X} \times \mathbb{Y} \to [0,\infty)$ be $(\mathcal{X} \otimes \mathcal{Y})/\mathcal{B}([0,\infty))$-measurable, then we can find a monotonically increasing sequence of step functions $(\Phi_n)_{n \in \mathbb{N}}$,

$$\Phi_n(x,y) = \sum_{k=1}^{N_n} d_k^{(n)} \mathbb{1}_{D_k^{(n)}}(x,y),$$

such that $\lim_{n\to\infty}\Phi_n(x,y)=\Phi(x,y)$ point-wise. Monotonicity is to understand in the sense of $\Phi_n(x,y)\leq\Phi_{n+1}(x,y)$ for all $(x,y)\in\mathbb{X}\times\mathbb{Y}$ and all $n\in\mathbb{N}$. We apply monotone convergence and the findings of **step 2** to imply

$$\begin{aligned}\mathbb{E}[\Phi(X,Y)\mathbb{1}_A]\mathbb{E}[\lim_{n\to\infty}\Phi_n(X,Y)\mathbb{1}_A] &= \lim_{n\to\infty}\mathbb{E}[\Phi_n(X,Y)\mathbb{1}_A]\\ &= \lim_{n\to\infty}\mathbb{E}[(\mathbb{E}[\Phi_n(X,y)])\,|_{y=Y}\mathbb{1}_A]\\ &= \mathbb{E}[(\mathbb{E}[\Phi(X,y)])\,|_{y=Y}\mathbb{1}_A]=\mathbb{E}[\varphi(Y)\mathbb{1}_A],\end{aligned}$$

for all $A\in\mathcal{F}$.

Step 4: Let $\Phi:\mathbb{X}\times\mathbb{Y}\to\mathbb{R}$ be $(\mathcal{X}\otimes\mathcal{Y})/\mathcal{B}(\mathbb{R})$-measurable and consider the decomposition

$$\Phi(x,y)=\Phi^+(x,y)-\Phi^-(x,y)$$

for positive and $(\mathcal{X}\otimes\mathcal{Y})/\mathcal{B}(\mathbb{R})$-measurable functions Φ^+,Φ^-. Moreover, we define

$$\varphi^+(y)=\mathbb{E}[\Phi^+(X,y)]\leq\mathbb{E}[|\Phi(X,y)|]<\infty,\quad \varphi^-(y)=\mathbb{E}[\Phi^-(X,y)]\leq\mathbb{E}[|\Phi(X,y)|]<\infty$$

where we can write

$$\varphi(y)=\mathbb{E}[\Phi(X,y)]=\mathbb{E}[\Phi^+(X,y)-\Phi^-(X,y)]=\varphi^+(y)-\varphi^-(y).$$

We apply linearity of the expectation and the findings of **step 3** to deduce

$$\mathbb{E}[\Phi(X,Y)\mathbb{1}_A]=\mathbb{E}[\Phi^+(X,Y)\mathbb{1}_A]-\mathbb{E}[\Phi^-(X,Y)\mathbb{1}_A]=\mathbb{E}[\varphi^+(Y)\mathbb{1}_A]-\mathbb{E}[\varphi^-(Y)\mathbb{1}_A]=\mathbb{E}[\varphi(Y)\mathbb{1}_A].$$

□

Remark 4.1.8. We can apply Lemma 4.1.2 and Lemma 4.1.6 to verify (4.3).

4.1.2 Almost sure convergence for non-convex cost function

In the following section, we will analyze the almost sure convergence behavior of SGD. We will make use of an almost sure convergence theorem of Robbins & Siegmund [20] which is based on Doob's supermartingale convergence theorem. We refer to Appendix A.4 for a brief summary/recall on martingales.

Theorem 4.1.9 (Robbins & Siegmund)**.** *Let $(\Omega,\mathcal{A},\mathcal{F},\mathbb{P})$ be a filtered probability space, $(Z_k)_{k\in\mathbb{N}}$,*

$(A_k)_{k\in\mathbb{N}}$, $(B_k)_{k\in\mathbb{N}}$ and $(C_k)_{k\in\mathbb{N}}$ be non-negative and $\mathcal{F}$-adapted stochastic processes, such that

$$\sum_{k=0}^{\infty} A_k < \infty \quad \text{and} \quad \sum_{k=0}^{\infty} B_k < \infty$$

almost surely. Moreover, suppose

$$\mathbb{E}[Z_{k+1} \mid \mathcal{F}_k] \leq Z_k(1+A_k) + B_k - C_k.$$

Then

1. *there exists an almost surely finite random variable Z_∞ such that $Z_k \to Z_\infty$ almost surely for $k \to \infty$,*
2. *it holds true that $\sum_{k=0}^{\infty} C_k < \infty$ almost surely.*

Proof. We want to apply Doob's martingale convergence theorem, Theorem A.4.2, in order to prove the assertion. Therefore, we are going to construct a supermartingale based on the stated stochastic processes.

Step 1 (construction of a supermartingale): We define the auxiliary random variables

$$\widehat{Z}_k = \frac{Z_k}{\prod_{i=0}^{k-1}(1+A_i)}, \quad \widehat{B}_k = \frac{B_k}{\prod_{i=0}^{k}(1+A_i)}, \quad \widehat{C}_k = \frac{C_k}{\prod_{i=0}^{k}(1+A_i)}$$

and observe that

$$\begin{aligned}\mathbb{E}[\widehat{Z}_{k+1} \mid \mathcal{F}_k] = \left(\prod_{i=0}^{k}(1+A_i)^{-1}\right)\mathbb{E}[Z_{k+1} \mid \mathcal{F}_k] &\leq \left(\prod_{i=0}^{k}(1+A_i)^{-1}\right)(Z_k(1+A_k)+B_k-C_k) \\ &= \widehat{Z}_k + \widehat{B}_k - \widehat{C}_k.\end{aligned} \tag{4.4}$$

Our candidate for the supermartingale is

$$M_k = \widehat{Z}_k - \sum_{i=0}^{k-1}(\widehat{B}_i - \widehat{C}_i),$$

for which we observe

$$\begin{aligned}\mathbb{E}[M_{k+1} \mid \mathcal{F}_k] = \mathbb{E}[\widehat{Z}_{k+1} \mid \mathcal{F}_k] - \sum_{i=0}^{k}\Big(\mathbb{E}[\widehat{B}_i \mid \mathcal{F}_k] - \mathbb{E}[\widehat{C}_i \mid \mathcal{F}_k]\Big) &\leq \widehat{Z}_k + \widehat{B}_k - \widehat{C}_k - \sum_{i=0}^{k}(\widehat{B}_i - \widehat{C}_i) \\ &= \widehat{Z}_k - \sum_{i=0}^{k-1}(\widehat{B}_i - \widehat{C}_i) = M_k,\end{aligned}$$

where we have used (4.4) and that $\widehat{B}_i$, $\widehat{C}_i$ are $\mathcal{F}_k$-measurable for $i \leq k$. In order to apply Doob's

martingale convergence theorem, we need to verify $\sup_{k\in\mathbb{N}} \mathbb{E}[M_k^-] < \infty$. Since in general, it is not obvious that this property will hold, we introduce a localization

Step 2 (localization): We define the stopping time $\tau_\varepsilon = \inf\{k \geq 1 : \sum_{i=0}^k \widehat{B}_i > \varepsilon\}$ for $\varepsilon > 0$. Since $(B_k)_{k\in\mathbb{N}}$ is $\mathcal{F}$-adapted, τ_ε is a stopping time with respect to $\mathcal{F}$. Moreover, $(M_{k\wedge\tau_\varepsilon})_{k\in\mathbb{N}}$ is still a supermartingale, and additionally satisfies

$$M_{k\wedge\tau_\varepsilon} = \widehat{Z}_{k\wedge\tau_\varepsilon} - \sum_{i=0}^{(k\wedge\tau_\varepsilon)-1} \widehat{B}_i + \sum_{i=0}^{(k\wedge\tau_\varepsilon)-1} \widehat{C}_i \geq - \sum_{i=0}^{(k\wedge\tau_\varepsilon)-1} \widehat{B}_i \geq -\varepsilon,$$

since $\sum_{i=0}^{(k\wedge\tau_\varepsilon)-1} \widehat{B}_i \leq \varepsilon$ by construction of the stopping time τ_ε. Since $(M_{k\wedge\tau_\varepsilon})_{k\in\mathbb{N}}$ is uniformly bounded from below (and due to the monotonic decrease of the expectation for supermartingales) we obtain

$$\sup_{k\in\mathbb{N}} \mathbb{E}[|M_{k\wedge\tau_\varepsilon}|] < \infty.$$

We are now ready to apply Theorem A.4.2 to find a integrable random variable M_∞^ε with $\lim_{k\to\infty} M_{k\wedge\tau_\varepsilon} = M_\infty^\varepsilon$ almost surely. Next, we have to remove the stopping time.

Step 3 (remove localization): Let $(\varepsilon_n)_{n\in\mathbb{N}}$ be an increasing sequence with $\lim_{n\to\infty} \varepsilon_n = \infty$. First note, that for each $n \in \mathbb{N}$ we have

$$\lim_{k\to\infty} M_{k\wedge\tau_{\varepsilon_n}}(\omega) = M_\infty^{\varepsilon_n}(\omega)$$

for almost all $\omega \in \Omega$. We observe that for each $\omega \in \Omega$ with $\sum_{i=0}^\infty \widehat{B}_i(\omega) < \infty$ there exists $N \in \mathbb{N}$ such that $\omega \in \{\tau_{\varepsilon_N} = \infty\}$, i.e. for this ω it holds

$$M_{k\wedge\tau_{\varepsilon_N}}(\omega) = M_k(\omega)$$

for all $k \in \mathbb{N}$, but similarly

$$\lim_{k\to\infty} M_k(\omega) = \lim_{k\to\infty} M_{k\wedge\tau_{\varepsilon_N}}(\omega) = M_\infty^{\tau_N}(\omega) < \infty,$$

where the last inequality $< \infty$ holds since $\mathbb{E}[|M_\infty^{\tau_N}|] < \infty$.

Step 4 (conclusion): Finally, we move back to the assertion regarding $(Z_k)_{k\in\mathbb{N}}$ and $(C_k)_{k\in\mathbb{N}}$. Observe that

$$-\infty < -\sum_{i=0}^\infty \widehat{B}_i(\omega) \leq \lim_{k\to\infty} M_k(\omega) = \lim_{k\to\infty} \widehat{Z}_k(\omega) - \sum_{i=0}^{k-1} (\widehat{B}_i(\omega) - \widehat{C}_i(\omega)) < \infty,$$

where $\widehat{Z}_k(\omega), \widehat{B}_i(\omega), \widehat{C}_i(\omega) \geq 0$ implying that

$$\lim_{k\to\infty} \widehat{Z}_k(\omega) < \infty \quad \text{and} \quad \sum_{i=0}^{\infty} \widehat{C}_i(\omega) < \infty$$

for almost all $\omega \in \Omega$. Moreover, it holds true that

$$Z_k(\omega) = \widehat{Z}_k(\omega) \prod_{i=0}^{k-1} (1 + A_i(\omega)),$$

where both $\widehat{Z}_k(\omega)$ and $\prod_{i=0}^{k-1}(1 + A_i(\omega))$ converge for almost all $\omega \in \Omega$. The latter one follows by monotonicity and

$$0 \leq \prod_{i=0}^{k-1}(1 + A_i(\omega)) \leq \exp(\sum_{i=0}^{k-1} A_i(\omega)),$$

where the upper bound converges by assumption. Therefore, $\lim_{k\to\infty} Z_k(\omega) = Z_\infty(\omega)$ exists for almost all $\omega \in \Omega$. Similarly, we have

$$\sum_{i=0}^{k} C_i(\omega) = \sum_{i=0}^{k} \widehat{C}_i(\omega) \prod_{j=0}^{i}(1 + A_j(\omega)) \leq \left(\prod_{j=0}^{\infty}(1 + A_j(\omega))\right) \sum_{i=0}^{k} \widehat{C}_i(\omega)$$

which implies

$$\sum_{i=0}^{\infty} C_i(\omega) < \infty$$

for almost all $\omega \in \Omega$. □

The following corollary is an easy, but very useful, extension of Theorem 4.1.9.

Corollary 4.1.10. Let $(\Omega, \mathcal{A}, \mathcal{F}, \mathbb{P})$ be a filtered probability space, $(Z_k)_{k\in\mathbb{N}}$, $(A_k)_{k\in\mathbb{N}}$, $(B_k)_{k\in\mathbb{N}}$ and $(D_k)_{k\in\mathbb{N}}$ be non-negative and $\mathcal{F}$-adapted stochastic processes, such that

$$\sum_{k=0}^{\infty} A_k < \infty, \quad \sum_{k=0}^{\infty} B_k < \infty \quad \text{and} \quad \sum_{k=0}^{\infty} D_k = \infty$$

almost surely. Moreover, suppose

$$\mathbb{E}[Z_{k+1} \mid \mathcal{F}_k] \leq Z_k(1 + A_k - D_k) + B_k.$$

Then Z_k converges almost surely to 0 for $k \to \infty$.

Exercise 4.1.1. Prove Corollary 4.1.10.

We are now ready to prove convergence of SGD in the non-convex setting. We will assume that the cost function F is L-smooth and lower bounded, i.e. $\inf_{x\in\mathbb{R}^d} F(x) > -\infty$. Similar to the case of gradient descent, we do not expect more than convergence to stationary points. In particular, we are able to extend Theorem 2.3.8 to the stochastic version.

Theorem 4.1.11 (SGD almost sure convergence)**.** *Let $F:\mathbb{R}^d\to\mathbb{R}$ be L-smooth and bounded from below by $F_* = \inf_{x\in\mathbb{R}^d} F(x) > -\infty$, let $(\alpha_k)_{k\in\mathbb{N}}$ (deterministic or $\mathcal{F}$-adapted) satisfy*

$$\alpha_k > 0, \quad \sum_{k=0}^{\infty}\alpha_k = \infty \quad \text{and} \quad \sum_{k=0}^{\infty}\alpha_k^2 < \infty$$

(almost surely). We assume that the Assumptions of Lemma 4.1.2 are satisfied, and

$$\mathbb{E}[\|\nabla_x f(x,Z) - \mathbb{E}[\nabla_x f(x,Z)]\|^2] \leq c(1+(F(x)-F_*))$$

for some constant $c>0$ and all $x\in\mathbb{R}^d$. Moreover, let X_0 be a random variable such that $\mathbb{E}[F(X_0)] < \infty$ and $(X_k)_{k\in\mathbb{N}}$ be the sequence of random variables generated by Algorithm 6. Then it holds true that the sequence of random variables $(F(X_k))_{k\in\mathbb{N}}$ converges almost surely to some random variable F_∞, almost surely finite, and

$$\lim_{k\to\infty}\|\nabla_x F(X_k)\|^2 = 0$$

almost surely.

Proof. We define the natural filtration $\mathcal{F} = (\mathcal{F}_k)_{k\in\mathbb{N}}$ through $\mathcal{F}_k = \sigma(X_m, m\leq k) = \sigma(X_0,\ Z_m, m\leq k)$ and note that $(\alpha_k)_{k\in\mathbb{N}}$ is $\mathcal{F}$-adapted per construction. Using the L-smoothness of F we obtain (path-wise) that

$$\begin{aligned}
F(X_{k+1}) &= F(X_k - \alpha_k\nabla_x f(X_k, Z_{k+1})) \\
&\leq F(X_k) - \alpha_k\langle\nabla_x F(X_k), \nabla_x f(X_k,Z_{k+1})\rangle + \alpha_k^2\frac{L}{2}\|\nabla_x f(X_k,Z_{k+1})\|^2 \\
&= F(X_k) - \alpha_k\|\nabla_x F(X_k)\|^2 + \alpha_k\langle\nabla_x F(X_k), M_{k+1}\rangle \\
&\quad + \alpha_k^2\frac{L}{2}(\|\nabla_x F(X_k)\|^2 - 2\langle\nabla_x F(X_k), M_{k+1}\rangle + \|M_{k+1}\|^2,
\end{aligned}$$

where $M_{k+1} := \nabla_x F(X_k) - \nabla_x f(X_k, Z_{k+1})$. By Lemma 4.1.2 and Lemma 4.1.6 we obtain

$$\mathbb{E}[M_{k+1}\mid\mathcal{F}_k] = 0$$

and

$$\mathbb{E}[\|M_{k+1}\|^2 \mid \mathcal{F}_k] \leq c(1 + (F(X_k) - F_*)).$$

This yields

$$\begin{aligned}\mathbb{E}[F(X_{k+1}) - F_* \mid \mathcal{F}_k] &\leq (F(X_k) - F_*) + \frac{L}{2}(\alpha_k^2 - \alpha_k)\|\nabla_x F(X_k)\|^2 + \frac{L}{2}\alpha_k^2 c(1 + (F(X_k) - F_*)) \\ &= (1 + c\frac{L}{2}\alpha_k^2)(F(X_k) - F_*) + c\frac{L}{2}\alpha_k^2 - \alpha_k(1 - \frac{L}{2}\alpha_k)\|\nabla_x F(X_k)\|^2.\end{aligned}$$

W.l.o.g. we assume that $\alpha_k \leq (1-\varepsilon)\frac{2}{L}$ for some $\varepsilon \in (0,1)$ (else let k be sufficiently large), such that $(1 - \frac{L}{2}\alpha_k) \geq \varepsilon > 0$. We can now apply Theorem 4.1.9 to imply that $\lim_{k\to\infty} F(X_k) - F_*$ exists almost surely and is finite, as well as

$$\varepsilon \sum_{k=0}^{\infty} \alpha_k \|\nabla_x F(X_k)\|^2 \leq \sum_{k=0}^{\infty} \alpha_k (1 - \frac{L}{2}\alpha_k)\|\nabla_x F(X_k)\|^2 < \infty$$

almost surely. Since we have assumed $\sum_{k=0}^{\infty} \alpha_k = \infty$ almost surely, using the same argument as in the proof of Theorem 2.3.8 path-wise we obtain

$$\lim_{k\to\infty} \|\nabla_x F(X_k)\|^2 = 0$$

almost surely. □

Remark 4.1.12. We note that the condition

$$\mathbb{E}[\|\nabla_x f(x, Z) - \mathbb{E}[\nabla_x f(x, Z)]\|^2] \leq c(1 + (F(x) - F_*))$$

in Theorem 4.1.11 can also be replaced by

$$\mathbb{E}[\|\nabla_x f(x, Z) - \mathbb{E}[\nabla_x f(x, Z)]\|^2] \leq c(1 + \|\nabla_x F(x)\|^2),$$

see for example [2]. Both conditions are relaxations of an uniform (in $x \in \mathbb{R}^d$) variance bound

$$\mathbb{E}[\|\nabla_x f(x, Z) - \mathbb{E}[\nabla_x f(x, Z)]\|^2] \leq c.$$

Before delving into the derivation of convergence rates for SGD, we formulate the following Corollary which states almost sure convergence under same assumptions of Theorem 4.1.11, but with the additional property of strong convexity of F.

Corollary 4.1.13. Suppose that the assumptions of Theorem 4.1.11 are satisfied and additionally, assume that F is μ-strongly convex. Then the sequence of random variables $(X_k)_{k\in\mathbb{N}}$

converges almost surely to the unique global minimum $x_* \in \mathbb{R}^d$ of F.

Exercise 4.1.2. Prove Corollary 4.1.13.

Similar as in the setting of the deterministic gradient descent scheme, the previous result states convergence to a stationary point without explicit rate of convergence, but under rather mild assumptions on the cost function F. In order to obtain a speed of convergence, additional properties such as convexity must be assumed for F. We will observe that the convergence behavior is worse than in the setting of deterministic gradient descent methods. However, in order to make a fair comparison between deterministic and stochastic gradient descent schemes, one must consider the complexity of both algorithms, including the computational cost of implementation.

4.1.3 Convergence for convex and smooth cost function

In the following, we will prove convergence of SGD under the assumption that the cost function is convex. Compared to the convergence result for gradient descent schemes, we will prove a slower convergence behavior. The following Theorem presents the resulting error bound in expectation.

Theorem 4.1.14 (SGD for convex and smooth cost function)**.** *Let $F : \mathbb{R}^d \to \mathbb{R}$ be convex and L-smooth, and assume that the set of global minima of F is non-empty. We assume that the assumptions of Lemma 4.1.2 are satisfied and that there exists $c > 0$ such that*

$$\mathbb{E}[\|\nabla_x f(x, Z) - \mathbb{E}[\nabla_x f(x, Z)]\|^2] \leq c$$

for all $x \in \mathbb{R}^d$. Let X_0 be a random variable such that $\mathbb{E}[|F(X_0)| + \|X_0 - x_*\|^2] < \infty$ for some $x_* \in \arg\min_{x\in\mathbb{R}^d} F(x)$. Moreover, let $(X_k)_{k\in\mathbb{N}}$ be generated by Algorithm 6 with deterministic and decreasing sequence of step sizes $(\alpha_k)_{k\in\mathbb{N}}$ such that $\alpha_k \in (0, \frac{1}{L}]$. Then for

$$\bar{X}_N := \sum_{k=0}^{N-1} w_k^N X_{k+1}, \quad w_k^N := \frac{\alpha_k}{\sum_{j=0}^{N-1} \alpha_j}, \quad N \geq 2,$$

it holds true that

$$\mathbb{E}[F(\bar{X}_N) - F(x_*)] \leq \frac{\mathbb{E}[\|X_0 - x_*\|^2]}{2\sum_{j=0}^{N-1} \alpha_j} + \frac{c(1 + \alpha_0 L)\sum_{k=0}^{N-1} \alpha_k^2}{2\sum_{j=0}^{N-1} \alpha_j}.$$

Proof. Let $x_* \in \arg\min_{x\in\mathbb{R}^d} F(x)$ and $\mathcal{F}_k = \sigma(X_m, m \leq k)$ be the natural filtration. Similar as in the proof of Theorem 4.1.11 we have

$$\mathbb{E}[F(X_{k+1})] = \mathbb{E}[\mathbb{E}[F(X_{k+1}) \mid \mathcal{F}_k]] \leq \mathbb{E}[F(X_k)] - \alpha_k(1 - \frac{L\alpha_k}{2})\mathbb{E}[\|\nabla_x F(X_k)\|^2] + c\frac{L}{2}\alpha_k^2.$$

We can also derive the following,

$$\begin{aligned}\|X_{k+1}-x_*\|^2 &= \|X_k-x_*\|^2 - 2\alpha_k\langle\nabla_x f(X_k,Z_{k+1}),X_k-x_*\rangle + \alpha_k^2\|\nabla_x f(X_k,Z_{k+1})\|^2\\ &= \|X_k-x_*\|^2 - 2\alpha_k\langle\nabla_x F(X_k),X_k-x_*\rangle + \alpha_k^2\|\nabla_x F(X_k)\|^2\\ &\quad + 2\alpha_k\langle M_{k+1},X_k-x_*\rangle + \alpha_k^2\|M_{k+1}\|^2 + 2\alpha_k^2\langle M_{k+1},\nabla_x F(X_k)\rangle,\end{aligned}$$

where again $M_{k+1} = \nabla_x F(X_k) - \nabla_x f(X_k,Z_{k+1})$. Taking the conditional expectation wrt. $\mathcal{F}_k$ results in the bound

$$\begin{aligned}\mathbb{E}[\|X_{k+1}-x_*\|^2 \mid \mathcal{F}_k] &= \|X_k-x_*\|^2 - 2\alpha_k\langle\nabla_x F(X_k),X_k-x_*\rangle + \alpha_k^2\|\nabla_x F(X_k)\|^2 + \alpha_k^2\mathbb{E}[\|M_{k+1}\|^2 \mid \mathcal{F}_k]\\ &\leq \|X_k-x_*\|^2 - 2\alpha_k\langle\nabla_x F(X_k),X_k-x_*\rangle + \alpha_k^2\|\nabla_x F(X_k)\|^2 + \alpha_k^2 c\,.\end{aligned}$$

We take again expectation and rewrite the derived inequality in form

$$2\alpha_k\mathbb{E}[\langle\nabla_x F(X_k),X_k-x_*\rangle] \leq \mathbb{E}[\|X_k-x_*\|^2] - \mathbb{E}[\|X_{k+1}-x_*\|^2] + \alpha_k^2\mathbb{E}[\|\nabla_x F(X_k)\|^2] + \alpha_k^2 c\,.$$

By convexity of F we have almost surely that

$$F(X_k) \leq F(x_*) + \langle X_k - x_*, \nabla_x F(X_k)\rangle,$$

such that

$$\begin{aligned}\mathbb{E}[F(X_{k+1})] &\leq F(x_*) + \mathbb{E}[\langle X_k-x_*,\nabla_x F(X_k)\rangle] - \alpha_k(1-\frac{L\alpha_k}{2})\mathbb{E}[\|\nabla_x F(X_k)\|^2] + c\frac{L}{2}\alpha_k^2\\ &\leq F(x_*) + \frac{1}{2\alpha_k}\left(\mathbb{E}[\|X_k-x_*\|^2] - \mathbb{E}[\|X_{k+1}-x_*\|^2]\right)\\ &\quad - \alpha_k\left(\frac{1}{2}-\frac{L\alpha_k}{2}\right)\mathbb{E}[\|\nabla_x F(X_k)\|^2] + (\frac{\alpha_k}{2}+\frac{L\alpha_k^2}{2})c\\ &\leq F(x_*) + \frac{1}{2\alpha_k}\left(\mathbb{E}[\|X_k-x_*\|^2] - \mathbb{E}[\|X_{k+1}-x_*\|^2]\right) + \left(\frac{\alpha_k}{2}+\frac{L\alpha_k^2}{2}\right)c,\end{aligned}$$

where we have used that $(\frac{1}{2}-\frac{L\alpha_k}{2}) \geq 0$. Note that $\sum_{k=0}^{N-1} w_k^N = 1$, such that by Jensen's inequality

it follows that

$$\begin{aligned}\mathbb{E}[F(\bar{X}_N) - F(x_*)] &\leq \frac{1}{\sum_{j=0}^{N-1}\alpha_j}\sum_{k=0}^{N-1}\alpha_k\mathbb{E}[F(X_{k+1}) - F(x_*)] \\ &\leq \frac{1}{2\sum_{j=0}^{N-1}\alpha_j}\sum_{k=0}^{N-1}\left(\mathbb{E}[\|X_k - x_*\|^2] - \mathbb{E}[\|X_{k+1} - x_*\|^2]\right) \\ &\quad + \frac{1}{\sum_{j=0}^{N-1}\alpha_j}\sum_{k=0}^{N-1}\left(\frac{\alpha_k^2}{2} + \frac{\alpha_k^3 L}{2}\right)c \\ &\leq \frac{\mathbb{E}[\|X_0 - x_*\|^2]}{2\sum_{j=0}^{N-1}\alpha_j} + \frac{c(1+\alpha_0 L)\sum_{k=0}^{N-1}\alpha_k^2}{2\sum_{j=0}^{N-1}\alpha_j},\end{aligned}$$

where we have used that α_k is decreasing and therefore, $\alpha_k \leq \alpha_0$. □

From the derived upper bound we obtain convergence under the sufficient condition that

$$\sum_{k=0}^{\infty}\alpha_j = \infty \quad \text{and} \quad \sum_{k=0}^{\infty}\alpha_k^2 < \infty.$$

In order to quantify the speed of convergence, we provide a specific choice of the step sizes. Note that in the case below, we have $\sum_{k=0}^{\infty}\alpha_k^2 = \infty$ but

$$\frac{\sum_{k=0}^{N}\alpha_k^2}{\sum_{k=0}^{N}\alpha_k} \to 0\,.$$

Corollary 4.1.15. Suppose that the same conditions of Theorem 4.1.14 are satisfied. Moreover, let $\alpha_k := \frac{1}{L\sqrt{k+1}}$. Then it holds true that

$$\mathbb{E}[F(\bar{X}_N) - F(x_*)] \in \mathcal{O}\left(\frac{\log(N)}{\sqrt{N}}\right).$$

Proof. Firstly, we observe that

$$\sum_{j=0}^{N-1}\alpha_j = \frac{1}{L}\sum_{j=0}^{N-1}\frac{1}{\sqrt{j+1}} \geq \frac{1}{L}\int_1^{N-1}\frac{1}{\sqrt{t+1}}\,\mathrm{d}t = \frac{2}{L}(\sqrt{N} - \sqrt{2}) \geq \frac{1}{L}\sqrt{N},$$

for sufficiently large N ($N \geq 8$). On the other side, we have

$$\sum_{j=0}^{N-1}\alpha_j^2 = \frac{1}{L^2}\sum_{j=0}^{N-1}\frac{1}{j+1} \leq \frac{1}{L^2}\left(1 + \int_0^{N-1}\frac{1}{t+1}\,\mathrm{d}t\right) = \frac{1}{L^2}(1 + \log(N)).$$

By the upper bound derived in Theorem 4.1.14, we obtain

$$\mathbb{E}[F(\bar{X}_N) - F(x_*)] \leq \frac{L\mathbb{E}[\|X_0 - x_*\|^2] + \frac{c}{2L}(1+\alpha_0 L)}{\sqrt{N}} + \frac{c}{2L}(1+\alpha_0 L)\frac{\log(N)}{\sqrt{N}} \in \mathcal{O}\left(\frac{\log(N)}{\sqrt{N}}\right).$$

□

4.1.4 Convergence for strongly convex and smooth cost function

If we additionally assume strong convexity, we can further improve the derived upper bound. Moreover, we can even prove convergence to a unique global minimum of F, as we also did for the deterministic gradient descent scheme. However, due to the stochastic approximation of the gradient, we lose the behavior of linear convergence.

Theorem 4.1.16 (SGD for strong convex and smooth cost function)**.** *Let $F : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex and L-smooth. We assume that the assumptions of Lemma 4.1.2 are satisfied and that there exists $c > 0$ such that*

$$\mathbb{E}[\|\nabla_x f(x,Z) - \mathbb{E}[\nabla_x f(x,Z)]\|^2] \leq c$$

for all $x \in \mathbb{R}^d$. Let X_0 be a random variable such that $\mathbb{E}[|F(X_0)| + \|X_0 - x_*\|^2] < \infty$, where $x_* \in \mathbb{R}^d$ is the unique global minimum of F. Moreover, let $(X_k)_{k\in\mathbb{N}}$ be generated by Algorithm 6 with deterministic and decreasing sequence of step sizes $(\alpha_k)_{k\in\mathbb{N}}$ such that $\alpha_k \in (0, \frac{1}{L}]$. Then it holds true that

$$\mathbb{E}[\|X_{k+1} - x_*\|^2] \leq (1 - \alpha_k\mu)\mathbb{E}[\|X_k - x_*\|^2] + c\alpha_k^2$$

for all $k \geq 0$.

Proof. Let $\mathcal{F}_k = \sigma(X_m, m \leq k)$ be again the natural filtration and recall that we have derived in the proof of Theorem 4.1.14 that

$$\mathbb{E}[\|X_{k+1} - x_*\|^2] \leq \mathbb{E}[\|X_k - x_*\|^2] - 2\alpha_k\mathbb{E}[\langle \nabla_x F(X_k), X_k - x_*\rangle] + \alpha_k^2\mathbb{E}[\|\nabla_x F(X_k)\|^2] + \alpha_k^2 c.$$

By μ-strong convexity we have that

$$F(x_*) - F(X_k) \geq \langle x_* - X_k, \nabla_x F(X_k)\rangle + \frac{\mu}{2}\|X_k - x_*\|^2,$$

which can be rewritten as

$$-\langle X_k - x_*, \nabla_x F(X_k)\rangle \leq -\left(F(X_k) - F(x_*)\right) - \frac{\mu}{2}\|X_k - x_*\|^2 \quad \text{almost surely}.$$

Combining both inequalities, we obtain that

$$\mathbb{E}[\|X_{k+1} - x_*\|^2] \leq (1 - \alpha_k\mu)\mathbb{E}[\|X_k - x_*\|^2] - 2\alpha_k\mathbb{E}[F(X_k) - F(x_*)] + \alpha_k^2\mathbb{E}[\|\nabla_x F(X_k)\|^2] + \alpha_k^2 c\,.$$

The assumption of L-smoothness implies that

$$-(F(X_k) - F(x_*)) \leq -\frac{1}{2L}\|\nabla_x F(X_k)\|^2 \quad \text{almost surely,}$$

such that

$$\begin{aligned}\mathbb{E}[\|X_{k+1} - x_*\|^2] &\leq (1 - \alpha_k\mu)\mathbb{E}[\|X_k - x_*\|^2] + \alpha_k(\alpha_k - \frac{1}{L})\mathbb{E}[\|\nabla_x F(X_k)\|^2] + \alpha_k^2 c \\ &\leq (1 - \alpha_k\mu)\mathbb{E}[\|X_k - x_*\|^2] + \alpha_k^2 c,\end{aligned}$$

where we have used that $\alpha_k \leq \frac{1}{L}$. □

From the above derived error bound we observe that the iterated error decomposes into the error arising due to the optimization error from the exact gradient descent scheme applied to strongly convex and smooth cost functions, and into an error arising from the variance of the stochastic approximation of the gradients. In order to obtain a convergence rate along the iterations, we need to balance both errors by either decreasing the step size α_k sufficiently or by decreasing the variance term. The latter one will be the topic of Section 4.2, where we consider methods of variance reduction. In the following, we will present the former approach of decreasing the step size to 0.

Corollary 4.1.17. Suppose that the same conditions as in Theorem 4.1.16 are satisfied. Moreover, let $\alpha_k = \frac{\tau}{\mu(k+s)}$ for some $\tau \geq 2$ and $s \geq \kappa\tau = \frac{L}{\mu}\tau$. Then it holds true that $\alpha_0 \leq \frac{1}{L}$ and there exists $\gamma \geq (s+1)2\max(\mathbb{E}[\|X_0 - x_*\|^2], \frac{\tau^2 c}{s\mu^2})$ such that

$$\mathbb{E}[\|X_k - x_*\|^2] \leq \frac{\gamma}{k+s}$$

for all $k \geq 1$.

Proof. Firstly, we observe that by definition it holds true that

$$\alpha_0 = \frac{\tau}{\mu \cdot s} \leq \frac{\tau\mu}{\mu L\tau} = \frac{1}{L}\,,$$

such that $\alpha_k \leq \frac{1}{L}$ for all $k \geq 0$. We define $\Delta_k = \mathbb{E}[\|X_k - x_*\|^2]$ and prove the second claim via

induction. For $k = 1$ it holds true that

$$\Delta_1 \leq (1-\alpha_0\mu)\Delta_0 + \alpha_0^2 c \leq (1-\frac{1}{\kappa})\Delta_0 + \frac{\tau^2 c}{\mu^2 s} \\ \leq \frac{(s+1)2\max(\Delta_0, \frac{\tau^2 c}{\mu^2 s})}{s+1} \leq \frac{\gamma}{s+1}.$$

Now, suppose that the upper bound $\Delta_k \leq \frac{\gamma}{k+s}$ is satisfied for some $k \geq 1$, then it follows that

$$\begin{aligned}\Delta_{k+1} \leq (1-\alpha_k\mu)\Delta_k + \alpha_k^2 c &\leq \left(1-\frac{\tau}{k+s}\right)\frac{\gamma}{k+s} + \frac{\tau^2 c}{\mu^2}\frac{1}{(k+s)^2} \\ &= \frac{\gamma}{k+1+s} + \frac{\gamma}{(k+s)(k+1+s)} - \frac{\gamma\tau}{(k+s)^2} + \frac{\tau^2 c}{\mu^2}\frac{1}{(k+s)^2} \\ &\leq \frac{\gamma}{k+1+s} + \frac{\gamma - \tau\gamma + \frac{\tau^2 c}{\mu^2}}{(k+s)^2} \\ &\leq \frac{\gamma}{k+1+s},\end{aligned}$$

where we have used that $\gamma(1-\tau) \leq -\gamma$ and $\gamma \geq (s+1)\max(\Delta_0, \frac{\tau^2 c}{\mu^2}) \geq \frac{\tau^2 c}{\mu^2}$. □

4.1.5 Convergence under PL-condition and smooth cost function

Similar to the deterministic gradient descent method, we are able to prove convergence under the PL-condition. We obtain the same type of convergence behavior as in the strong convex setting, with the main difference being the error discrepancy in the cost function evaluation.

Theorem 4.1.18 (SGD under PL-condition)**.** *Let $F : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and assume that F satisfies the PL-condition*

$$\|\nabla_x F(x)\|^2 \geq 2r(F(x) - F_*)$$

for some $r \in (0, L)$ and all $x \in \mathbb{R}^d$, where $F_* = \inf_{x\in\mathbb{R}^d} F(x) > -\infty$. We assume that the assumptions of Lemma 4.1.2 are satisfied and that there exists $c > 0$ such that

$$\mathbb{E}[\|\nabla_x f(x, Z) - \mathbb{E}[\nabla_x f(x, Z)]\|^2] \leq c$$

for all $x \in \mathbb{R}^d$. Let X_0 be a random variable such that $\mathbb{E}[|F(X_0)| + \|X_0\|^2] < \infty$, and let $(X_k)_{k\in\mathbb{N}}$ be generated by Algorithm 6 with deterministic and decreasing sequence of step sizes

$(\alpha_k)_{k\in\mathbb{N}}$ *such that* $\alpha_k \in (0, \frac{1}{L}]$. *Then it holds true that*

$$\mathbb{E}[F(X_k) - F_*] \leq (1 - \alpha_k r)\mathbb{E}[F(X_k) - F_*] + c\frac{L}{2}\alpha_k^2 .$$

Proof. We have already seen that under L-smoothness we obtain the following bound

$$\mathbb{E}[F(X_{k+1}) - F_*] \leq \mathbb{E}[F(X_k) - F_*] - \alpha_k(1 - \frac{L}{2}\alpha_k)\mathbb{E}[\|\nabla_x F(X_k)\|^2] + c\frac{L}{2}\alpha_k^2 .$$

Under the PL-condition and the fact that $1 - \frac{L}{2}\alpha_k \geq \frac{1}{2} > 0$, we improve the upper bound to

$$\begin{aligned}\mathbb{E}[F(X_{k+1}) - F_*] &\leq \mathbb{E}[F(X_k) - F_*] - \alpha_k(1 - \frac{L}{2}\alpha_k)2r\mathbb{E}[F(X_k) - F_*] + c\frac{L}{2}\alpha_k^2 \\ &= (1 - \alpha_k(1 - \frac{L}{2}\alpha_k)2r)\mathbb{E}[F(X_k) - F_*] + c\frac{L}{2}\alpha_k^2 \\ &\leq (1 - \alpha_k r)\mathbb{E}[F(X_k) - F_*] + c\frac{L}{2}\alpha_k^2 .\end{aligned}$$

□

We can apply the same step size strategy as in the strongly convex setting to derive convergence.

Corollary 4.1.19. Suppose that the same conditions as in Theorem 4.1.18 are satisfied. Moreover, let $\alpha_k = \frac{\tau}{r(k+s)}$ for some $\tau \geq 2$ and $s = \frac{L}{r}\tau$. Then it holds true that $\alpha_0 \leq \frac{1}{L}$ and there exists $\gamma \geq (s+1)2\max(\mathbb{E}[F(X_0) - F_*], \frac{\tau^2 c}{r^2})$ such that

$$\mathbb{E}[F(X_k) - F_*] \leq \frac{\gamma}{k+s}$$

for all $k \geq 1$.

Proof. The proof proceeds line by line as the proof of Corollary 4.1.17. □

4.1.6 Discussion about the complexity of SGD

In the previous sections, we derived convergence rates of SGD under convexity, strong convexity and the PL-condition. We obtained similar results for the deterministic GD scheme. Comparing both GD and SGD, we observe that the derived results for SGD are significantly worse.

	convex	strongly convex	PL
GD	$\mathcal{O}\left(\frac{1}{k+1}\right)$	$\mathcal{O}(\rho^k),\ \rho \in (0,1)$	$\mathcal{O}(\rho^k),\ \rho \in (0,1)$
SGD	$\mathcal{O}\left(\frac{\log(k)}{\sqrt{k+1}}\right)$	$\mathcal{O}(\frac{1}{k+1})$	$\mathcal{O}(\frac{1}{k+1})$

However, the implementation of GD might be expensive (e.g. for large data sets in empirical risk minimization) or even impossible. In this case, it is infeasible to implement GD and there is no other choice to work with SGD.

The comparison of GD and SGD gets more involved in cases where we are able to compute the exact gradient. Let us consider the empirical risk minimization problem

$$\min_{x\in\mathbb{R}^d} F_N(x), \quad F_n(x) = \frac{1}{N}\sum_{i=1}^{N} f(x, z^{(i)}),\ z^{(i)} \in \mathbb{R}^p,\ i = 1, \dots, N,$$

where we assume that $x \mapsto f(x, z^{(i)})$ are μ-strongly convex and L-smooth for all i. We have seen that GD with a fixed step size $\bar{\alpha} \leq \frac{1}{L}$ converges linearly with rate $\rho \in (0, 1)$ such that

$$\|x_{k,N}^{\mathrm{GD}} - x_N\|^2 \leq \rho^k \|x_0 - x_N\|^2,$$

where $(x_{k,N}^{\mathrm{GD}})_{k\in\mathbb{N}}$ denotes the iteration generated by GD and $x_N = \arg\min_{x\in\mathbb{R}^d} F_N(x)$. In comparison, SGD (with decreasing step size) converges sub-linear with

$$\mathbb{E}[\|X_{k,N}^{\mathrm{SGD}} - x_N\|^2] \leq \frac{\gamma}{k+s}.$$

In order to achieve an error of a certain tolerance $\varepsilon > 0$ we need to iterate

(i) $k^{\mathrm{GD}} \geq \mathcal{O}(\log(\varepsilon^{-1}))$, such that $\|x_{k,N}^{\mathrm{GD}} - x_N\|^2 \leq \varepsilon$,

(ii) $k^{\mathrm{SGD}} \geq \mathcal{O}(\varepsilon^{-1})$, such that $\mathbb{E}[\|X_{k,N}^{\mathrm{SGD}} - x_N\|^2] \leq \varepsilon$.

The first guess is, that as long we are able to compute the full gradient, there is no reason to implement SGD over GD. However, this train of thought is too naive. The reason is, that up to now we have ignored the empirical error which occurs through solving $x_N = \arg\min\ F_N(x)$. Indeed, x_N should be treated as random variable X_N, which depends on $Z^{(1)}, \dots, Z^{(N)}$. We want to quantify the error of both GD and SGD to $x_* = \arg\min_{x\in\mathbb{R}^d} F(x)$, where $F(x) = \mathbb{E}[f(x, Z)]$ is the expected risk. For SGD, implemented through Algorithm 6, we again obtain convergence (with decreasing step size) in form of

$$\mathbb{E}[\|X_{k,N}^{\mathrm{SGD}} - x_*\|^2] \leq \frac{\gamma}{k+s}.$$

However, the situation changes for GD. Assuming that we are not able to compute the exact gradient of F, we firstly have to approximate F through F_N and then apply GD to find $X_N = \arg\min_{x\in\mathbb{R}^d} F_N(x)$. The final error decomposes to

$$\frac{1}{2}\mathbb{E}[\|X_{k,N}^{\mathrm{GD}} - x_*\|^2] \leq \mathbb{E}[\|X_{k,N}^{\mathrm{GD}} - X_N\|^2] + \mathbb{E}[\|X_N - x_*\|^2] \leq \rho^k \mathbb{E}[\|X_{0,N}^{\mathrm{GD}} - X_N\|^2] + \mathbb{E}[\|X_N - x_*\|^2],$$

where X_N denotes the random minimum of F_N given $Z^{(1)}, \dots, Z^{(N)}$. In the following, we study the error $\mathbb{E}[\|X_N - x_*\|^2]$ for strongly convex cost functions depending on the number of data points.

Theorem 4.1.20. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex and let $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ such that $x \mapsto f(x,z)$ is μ-strongly convex for all $z \in \mathbb{R}^p$. Moreover, let $x_* = \arg\min_{x\in\mathbb{R}^d} F(x)$, and assume that $\mathbb{E}[\nabla_x f(x_*, Z)] = \nabla_x F(x_*)$ and*

$$\mathbb{E}[\|\nabla_x f(x_*, Z) - \mathbb{E}[\nabla_x f(x_*, Z)]\|^2] \leq B$$

for some $B > 0$. Let $Z^{(1)}, \dots, Z^{(N)}$ be a family of iid. random variables with distribution μ_Z, then it holds true that

$$\mathbb{E}[\|X_n - x_*\|^2] \leq \frac{B}{\mu^2}\frac{1}{N},$$

where $X_N = \arg\min_{x\in\mathbb{R}^d} F_N(x)$.

Proof. Let

$$x_* = \arg\min_{x\in\mathbb{R}^d} F(x), \quad F(x) = \mathbb{E}[f(x,Z)]$$
$$X_N = \arg\min_{x\in\mathbb{R}^d} F_N(x), \quad F_N(x) = \frac{1}{N}\sum_{i=1}^N f(x, Z^{(i)}),$$

which means that $\nabla_x F(x_*) = 0$ and $\nabla_x F_N(X_N) = 0$ almost surely. By strong convexity we can apply Lemma 2.3.17 to imply

$$\begin{aligned}\|X_N - x_*\|^2 \leq \frac{1}{\mu}\langle X_n - x_*, \nabla_x F_N(X_N) - \nabla_x F_N(x_*)\rangle &= \frac{1}{\mu}\langle X_n - x_*, \nabla_x F(x_*) - \nabla_x F_N(x_*)\rangle \\ &\leq \frac{1}{\mu}\|X_N - x_*\|\|\nabla_x F(x_*) - \nabla_x F_N(x_*)\|,\end{aligned}$$

almost surely, where we have used Cauchy-Schwarz inequality in the last line. Reordering the above inequality leads to

$$\|X_N - x_*\| \leq \frac{1}{\mu}\|\nabla_x F(x_*) - \nabla_x F_N(x_*)\| \tag{4.5}$$

almost surely. Note that

$$\mathbb{E}[\nabla_x F_N(x_*)] = \mathbb{E}[\frac{1}{N}\sum_{i=1}^N \nabla_x f(x_*, Z^{(i)})] = \nabla_x F(x_*)$$

such that we obtain

$$\begin{aligned}\mathbb{E}[\|\nabla_x F_N(x_*) - \nabla_x F(x_*)\|^2] &= \mathbb{E}[\|\frac{1}{N}\sum_{i=1}^N \left(\nabla_x f(x_*, Z^{(i)}) - \mathbb{E}[\nabla_x f(x_*, Z^{(i)})]\right)\|^2] \\ &= \frac{1}{N^2}\sum_{i,j=1}^N \mathbb{E}[\langle \nabla_x f(x_*, Z^{(i)}) - \mathbb{E}[\nabla_x f(x_*, Z^{(i)})], \\ &\qquad \nabla_x f(x_*, Z^{(j)}) - \mathbb{E}[\nabla_x f(x_*, Z^{(j)})]\rangle] \\ &= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}[\|\nabla_x f(x_*, Z^{(i)}) - \mathbb{E}[\nabla_x f(x_*, Z^{(i)})]\|^2] \leq \frac{B}{N},\end{aligned}$$

where we have used that the $Z^{(1)}, \dots, Z^{(N)}$ are iid. random variables. Together with inequality (4.5) we close the proof with

$$\mathbb{E}[\|X_N - x_*\|^2] \leq \frac{B}{\mu^2}\frac{1}{N}.$$

□

The overall error of GD is then given by

$$\frac{1}{2}\mathbb{E}[\|X_{k,N}^{\mathrm{GD}} - x_*\|^2] \leq c(\frac{1}{N} + \rho^k)$$

for some constant $c > 0$. Therefore, it is sufficient to choose $N \geq \mathcal{O}(\varepsilon^{-1})$ and $k \geq \mathcal{O}(\log(\varepsilon^{-1})$, such that the computational cost of GD are given by

$$\mathrm{cost}_{\mathrm{GD}}(\varepsilon) = N \cdot k \simeq \varepsilon^{-1}\log(\varepsilon^{-1}),$$

whereas the computational cost of SGD are given by

$$\mathrm{cost}_{\mathrm{SGD}}(\varepsilon) = 1 \cdot k \simeq \varepsilon^{-1}.$$

4.1.7 Lower bound of SGD

In the following, we will observe that we do not expect to improve the derived upper bound on SGD in the strongly convex setting. Therefore, we consider the example of a simple quadratic function. The example to be considered has been studied in detail in [10].

Example 4.1.21 (SGD lower error bound)**.** *Let $(Z_k)_{k\in\mathbb{N}}$ be a sequence of iid. random variables in $\mathbb{R}^d$ with distribution μ_Z and $\mathbb{E}[\|Z_1\|^2] < \infty$. We define*

$$f(x, z) = \frac{1}{2}\|x - z\|^2, \quad x, z \in \mathbb{R}^d$$

and the corresponding expectation function

$$F(x) = \mathbb{E}[f(x,Z)], \quad x \in \mathbb{R}^d, \quad Z \sim \mu_Z \,.$$

Firstly, observe that this expectation computes as

$$F(x) = \frac{1}{2}\mathbb{E}[\|x-Z\|^2] = \frac{1}{2}\mathbb{E}[\|x-\mathbb{E}[Z]\|^2] + \frac{1}{2}\mathbb{E}[\|Z-\mathbb{E}[Z]\|^2]$$

such that we identify $x_* = \mathbb{E}[Z] \in \mathbb{R}^d$ as the global minimum of F. In this formulation we are also able to compute exact derivatives of F and obtain

$$\mathbb{E}[\|\nabla_x f(x,Z) - \nabla_x F(x)\|^2] = \mathbb{E}[\|(x-Z)-(x-\mathbb{E}[Z])\|^2] = \mathbb{E}[\|Z-\mathbb{E}[Z]\|^2] =: \sigma^2 \,.$$

Next, consider the iteration generated by SGD with sequence of step sizes $\alpha_k = \frac{\tau}{k^\nu}$, $k \in \mathbb{N}$ for some $\nu > 0$ and $\tau > 0$, which can be written as

$$X_{k+1} = X_k - \frac{\tau}{k^\nu}(X_k - Z_{k+1}) = (1-\frac{\tau}{k^\nu})X_k + \frac{\tau}{k^\nu}Z_{k+1} \,.$$

In this specific example, we are then able to compute the iterated error analytically given by

$$\begin{aligned}
\mathbb{E}[\|X_{k+1}-x_*\|^2] &= \mathbb{E}[\|X_{k+1}-\mathbb{E}[Z]\|^2] \\
&= \left(1-\frac{\tau}{k^\nu}\right)^2 \mathbb{E}[\|X_k-\mathbb{E}[Z]\|^2] + 2\left(1-\frac{\tau}{k^\nu}\right)\frac{\tau}{k^\nu}\mathbb{E}[\langle X_k-\mathbb{E}[Z], Z_{k+1}-\mathbb{E}[Z]\rangle] \\
&\quad + \left(\frac{\tau}{k^\nu}\right)^2 \mathbb{E}[\|Z_{k+1}-\mathbb{E}[Z]\|^2] \\
&= \left(1-\frac{\tau}{k^\nu}\right)^2 \mathbb{E}[\|X_k-\mathbb{E}[Z]\|^2] + \left(\frac{\tau}{k^\nu}\right)^2 \sigma^2 \,.
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{E}[\|X_k-\mathbb{E}[Z]\|^2] &= \prod_{j=0}^{k-1}\left(1-\frac{\tau}{j^\nu}\right)^2 \mathbb{E}[\|X_0-\mathbb{E}[Z]\|^2] + \sigma^2 \sum_{j=0}^{k-1}\left(\frac{\tau}{k^\nu}\right)^2 \prod_{i=j+1}^{k-1}\left(1-\frac{\tau}{i^\nu}\right)^2 \\
&\geq \sigma^2 \sum_{j=0}^{k-1}\left(\frac{\tau}{k^\nu}\right)^2 \prod_{i=j+1}^{k-1}\left(1-\frac{\tau}{i^\nu}\right)^2 ,
\end{aligned}$$

where the lower bound behaves as $Ck^{-\nu}$ for sufficiently large $k \in \mathbb{N}$, see [10] for more details. In particular, for the choice $\nu = 1$ the lower bound matches our derived upper bound in Section 4.1.4.

4.2 Variance reduction

In the derived convergence results for SGD we have always assumed that $\nabla_x f(x, Z)$ is an unbiased estimator of $\nabla_x F(x)$ with uniformly bounded variance

$$\mathbb{E}[\|\nabla_x f(x, Z) - \nabla_x F(x)\|^2] \leq \text{var}\,.$$

This upper bound $\text{var} > 0$ occurs in all derived error bounds in a similar way:

Assumption	error bound
convex	$\frac{C_1}{\sqrt{k}} + \text{var} \cdot \frac{\log(k)}{\sqrt{k}}$
strong convex	$(1 - \alpha_k \mu) e_k + \text{var} \cdot \alpha_k^2$
PL-condition	$(1 - \alpha_k r) e_k + \text{var} \cdot \alpha_k^2$

In order to push the total error towards zero, we had to choose $\alpha_k \to 0$. In the specific cases of strong convexity or under PL-condition, we lose the behavior of linear convergence which we obtained for the exact (deterministic) gradient descent scheme. In the deterministic setting, we were able to choose $\alpha_k = \bar{\alpha} > 0$ such that

$$e_{k+1} \leq \rho(\bar{\alpha}) e_k, \quad \rho(\bar{\alpha}) \in (0, 1)$$

By controlling the variance error term through $\alpha_k \to 0$, we obtain an error bound for SGD of the form

$$e_{k+1} \leq \rho_k e_k + \text{var}_k,$$

where $\rho_k \to 1$ for $k \to \infty$. This behavior makes the analysis challenging and in particular, we have seen that SGD does not converge linearly.
In the following section, we consider different types of variance reduction methods, which control the variance error term in a different way. We are no longer forced to consider $\alpha_k \to 0$ and will choose a fixed step size $\alpha_k = \bar{\alpha} > 0$ for all $k \in \mathbb{N}$.

4.2.1 Dynamic Sampling

Our first method to be considered is called *Dynamical Sampling*, where we control the variance error term through a dynamical batch-sampling strategy. The unbiased estimator of the descent direction $\nabla_x F(X_k)$ is estimated through a batch of samples $(Z_k^{(m)})_{k \in \mathbb{N},\ m=1,\dots,B_{k-1}}$, where the random variables are assumed to be independent in k and in m with an identical distribution μ_Z.
In the following, we will analyze SGD with dynamical batch-sampling. The focus will be placed on the strongly convex setting and the method will be compared to a fixed batch-size $\bar{B} > 0$ across all iterations. For both schemes, we will consider a fixed step size $\bar{\alpha} > 0$.

Algorithm 8 SGD with dynamical sampling

1: **Input:**
- cost function $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$
- initial random variable $X_0 : \Omega \to \mathbb{R}^d$
- sequence of step sizes $(\alpha_k)_{k\in\mathbb{N}}$, $\alpha_k > 0$ (deterministic or $\mathcal{F}$-adapted)
- sequence of batch sizes $(B_k)_{k\in\mathbb{N}}$
- sequence of i.i.d. random variables $(Z_k^{(m)})_{k\in\mathbb{N},\ m=1,\dots,B_{k-1}}$ with $Z_1^{(1)} \sim \mu_Z$.

2: set $k = 0$
3: **while** "convergence/stopping criterion not met" **do**
4: approximate the gradient $\nabla_x F(X_k)$ through

$$G_k = \frac{1}{B_k}\sum_{m=1}^{B_k} \nabla_x f(X_k, Z_{k+1}^{(m)})$$

5: set $X_{k+1} = X_k - \alpha_k G_k$, $k \mapsto k+1$
6: **end while**

Firstly, we discuss how the assumed uniform upper bound on the variance is effected through the incorporation of batch-sampling.

Lemma 4.2.1. Let the assumptions of Lemma 4.1.2 be satisfied and assume that

$$\mathbb{E}[\|\nabla_x f(x, Z) - \nabla_x F(x)\|^2] \leq c$$

for some $c > 0$ and all $x \in \mathbb{R}^d$. Moreover, let $Z^{(1)}, \dots, Z^{(B)}$ be iid. random variables with distribution μ_Z. Then for $G := \frac{1}{B}\sum_{m=1}^{B} \nabla_x f(x, Z^{(m)})$ it holds true that

$$\mathbb{E}[\|G - \nabla_x F(x)\|^2 \leq \frac{c}{B}$$

for all $x \in \mathbb{R}^d$.

Proof. By $\mathbb{E}[\nabla_x f(x, Z^{(1)})] = \nabla_x F(x)$ and the independence of $Z^{(1)}, \dots, Z^{(m)}$ we have

$$\begin{aligned}\mathbb{E}[\|G - \nabla_x F(x)\|^2] &= \frac{1}{B^2}\sum_{m,n=1}^{B} \mathbb{E}[\langle \nabla_x f(x, Z^{(m)}) - \nabla_x F(x), \nabla_x f(x, Z^{(m)}) - \nabla_x F(x)\rangle] \\ &= \frac{1}{B^2}\sum_{m=1}^{B} \mathbb{E}[\|\nabla_x f(x, Z^{(m)}) - \nabla_x F(x)\|^2] \leq \frac{c}{B}\,.\end{aligned}$$

□

With the previous observations we are now able to extend Theorem 4.1.16 to dynamical batch-

sampling.

Theorem 4.2.2 (SGD with dynamical sampling)**.** *Let $F : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex and L-smooth. We assume that the assumptions of Lemma 4.1.2 are satisfied and that there exists $c > 0$ such that*

$$\mathbb{E}[\|\nabla_x f(x, Z) - \mathbb{E}[\nabla_x f(x, Z)]\|^2] \leq c$$

for all $x \in \mathbb{R}^d$. Let X_0 be a random variable such that $\mathbb{E}[|F(X_0)| + \|X_0 - x_*\|^2] < \infty$, where $x_* \in \mathbb{R}^d$ is the unique global minimum of F. Moreover, let $(X_k)_{k\in\mathbb{N}}$ be generated by Algorithm 8 with sequence of batch-sizes $(B_k)_{k\in\mathbb{N}}$, $B_k \geq 1$ and deterministic, decreasing sequence of step sizes $(\alpha_k)_{k\in\mathbb{N}}$ such that $\alpha_k \in (0, \frac{1}{L}]$. Then for the error $e_k := \mathbb{E}[\|X_k - x_*\|^2]$ it holds true that

$$e_{k+1} \leq (1 - \alpha_k \mu) e_k + \frac{c\alpha_k^2}{B_k}$$

for all $k \geq 0$. Furthermore, for a fixed step size $\alpha_k = \bar{\alpha} = \frac{\tau}{\mu}$ with $\tau \leq \frac{1}{\kappa} = \frac{\mu}{L}$, it holds true that

$$e_{k+1} \leq \rho e_k + \frac{c\bar{\alpha}^2}{B_k} \tag{4.6}$$

where $\rho = (1 - \tau) \in (0, 1)$.

Proof. The proof follows by Theorem 4.1.16 combined with Lemma 4.2.1. □

In the following, we will keep the step size $\bar{\alpha} = \frac{\tau}{\mu} \leq \frac{1}{L}$ fixed and aim to derive an optimal sequence of batch-sizes. Firstly, we need to formulate what we mean by an *optimal* batch-size. Therefore, we will assume that the computation of each iteration of SGD occurs with cost which are determined through the generation of the samples $(Z_k^{(m)})_{m=1,\dots,B_{k-1}}$. We will assume that these cost are normalized.

Assumption 4.2.3. The generation of the state X_k by Algorithm 8 with sequence of batch-sizes $(B_k)_{k\in\mathbb{N}}$ occurs with computational cost

$$\text{cost}(X_k) = \sum_{j=0}^{k-1} B_j \,.$$

We want to choose the batch-sizes such that we minimize the computational cost under the constraint that the final error bound is below a specified tolerance level $\varepsilon > 0$, i.e. we want to solve the constrained minimization problem

$$\min_{B_0,\dots,B_{K-1}} \sum_{j=0}^{K-1} B_j \quad \text{s.t. } e_K \leq \varepsilon \,.$$

We refer the interested reader to [24], where this approach has been considered in a more general framework. The total error after iteration $K \geq 1$ can be upper bounded by iterating the error bound (4.6):

$$e_K \leq \rho^K e_0 + c\bar{\alpha}^2 \sum_{j=0}^{K-1} \rho^{K-1-j} \frac{1}{B_j}\,.$$

For simplicity, we assume that $K \geq \lceil \log(\rho^{-1}) \log(\frac{\varepsilon}{2} e_0^{-1}) \rceil$ such that

$$e_K \leq \frac{\varepsilon}{2}.$$

For this given K we want to determine $B_0, \dots, B_{K-1}$ under the constrain

$$c\bar{\alpha}^2 \sum_{j=0}^{K-1} \rho^{K-1-j} \frac{1}{B_j} \leq \frac{\varepsilon}{2}\,.$$

We start with the following auxiliary result:

Lemma 4.2.4. Let $\varepsilon > 0$, $\gamma > 0$ and $a_j > 0$, $j \in \{0, \dots, K-1\}$. Then the choice

$$B_j = C(\varepsilon, K) \cdot a_j^{\frac{1}{1+\gamma}}, \quad C(\varepsilon, K) = \varepsilon^{-\frac{1}{\gamma}} \left(\sum_{s=0}^{K-1} a_s^{\frac{1}{1+\gamma}} \right)^{\frac{1}{\gamma}}$$

solves the constrained optimization problem

$$\min_{B_0,\dots,B_{K-1}} \sum_{j=0}^{K-1} B_j, \quad \text{s.t.} \sum_{j=0}^{K-1} a_j B_j^{-\gamma} \leq \varepsilon\,.$$

Proof. We only derive a stationary point of the considered constrained optimization problem. The Lagrange function is given by

$$\mathcal{L}(B_0, \dots, B_{K-1}, \lambda) = \sum_{j=0}^{K-1} B_j + \lambda \left(\sum_{j=0}^{K-1} a_j B_j^{-\gamma} - \varepsilon \right)$$

and the corresponding optimality conditions are

$$\begin{aligned} &\text{(I)} \quad 1 - \lambda - \lambda\gamma a_j B_j^{-(1+\gamma)} = 0, \ j = 0, \dots, K-1, \\ &\text{(II)} \quad \sum_{j=0}^{K-1} a_j B_j^{-\gamma} - \varepsilon = 0\,. \end{aligned}$$

We solve (I) to derive

$$B_j = (\lambda\gamma)^{\frac{1}{1+\gamma}} a_j^{\frac{1}{1+\gamma}},$$

which together with (II) gives

$$(\lambda\gamma)^{-\frac{\gamma}{1+\gamma}} \sum_{j=0}^{K-1} a_j \cdot a_j^{-\frac{\gamma}{\gamma+1}} = (\lambda\gamma)^{-\frac{\gamma}{1+\gamma}} \sum_{j=0}^{K-1} a_j^{\frac{1}{\gamma+1}} = \varepsilon$$

and therefore,

$$\lambda\gamma = \varepsilon^{-\frac{1+\gamma}{\gamma}} \left(\sum_{j=0}^{K-1} a_j^{\frac{1}{1+\gamma}} \right)^{\frac{1+\gamma}{\gamma}}.$$

This results in

$$B_j = C(\varepsilon, K) \cdot a_j^{\frac{1}{1+\gamma}}, \quad C(\varepsilon, K) = \varepsilon^{-\frac{1}{\gamma}} \left(\sum_{s=0}^{K-1} a_s^{\frac{1}{1+\gamma}} \right)^{\frac{1}{\gamma}}.$$

□

We are now ready to choose the optimal batch-size for Algorithm 8 under strong convexity assumption:

$$\min_{B_0,\dots,B_{K-1}} \sum_{j=0}^{K-1} B_j, \quad \text{s.t.} \sum_{j=0}^{K-1} c\bar{\alpha}^2 \rho^{K-1-j} B_j^{-1} \leq \varepsilon/2\,,$$

which by Lemma 4.2.4 leads to

$$C(\varepsilon, K) = 2\varepsilon^{-1}\sqrt{c}\bar{\alpha} \sum_{j=0}^{K-1} \rho^{\frac{K-1-j}{2}} = 2\varepsilon^{-1}\sqrt{c}\bar{\alpha} \sum_{j=0}^{K-1} \rho^{\frac{j}{2}} = 2\varepsilon^{-1}\sqrt{c}\bar{\alpha} \left(\frac{1-\rho^{\frac{K}{2}}}{1-\rho^{\frac{1}{2}}} \right)$$

and therefore, to an optimal dynamical batch-size

$$B_j = 2\varepsilon^{-1} c\bar{\alpha}^2 \left(\frac{1-\rho^{\frac{K}{2}}}{1-\rho^{\frac{1}{2}}} \right) \rho^{\frac{K-1-j}{2}}. \tag{4.7}$$

The corresponding computational cost are given by

$$\sum_{j=0}^{K-1} B_j = 2\varepsilon^{-1} c\bar{\alpha}^2 \left(\frac{1-\rho^{\frac{K}{2}}}{1-\rho^{\frac{1}{2}}} \right) \sum_{j=0}^{K-1} \rho^{\frac{K-1-j}{2}} = 2\varepsilon^{-1} c\bar{\alpha}^2 \left(\frac{1-\rho^{\frac{K}{2}}}{1-\rho^{\frac{1}{2}}} \right)^2 \simeq \varepsilon^{-1},$$

where $\left(\frac{1-\rho^{\frac{K}{2}}}{1-\rho^{\frac{1}{2}}} \right)^2 \in \left(1, \left(\frac{1}{1-\sqrt{\rho}} \right)^2 \right)$, independent of K. We compare the derived dynamical batch-sampling strategy to a fixed batch size $\bar{B} \geq 1$ for all $k = 0, \dots, K-1$. This fixed batch-size has again to be chosen such that $e_K \leq \varepsilon$. For simplicity, let again $K \geq \lceil \log(\rho^{-1}) \log(\frac{\varepsilon}{2} e_0^{-1}) \rceil$ such that $\rho^K e_0 \leq \frac{\varepsilon}{2}$ and therefore,

$$e_K \leq \frac{\varepsilon}{2} + c\bar{\alpha}^2 \frac{1}{\bar{B}} \sum_{j=0}^{K-1} \rho^{K-1-j} = \frac{\varepsilon}{2} + c\bar{\alpha}^2 \frac{1-\rho^K}{1-\rho} \frac{1}{\bar{B}},$$

where $\frac{1-\rho^K}{1-\rho} \leq \frac{1}{1-\rho}$. The fixed batch-size $\bar{B}$ needs to be chosen such that

$$\bar{B} \geq 2\varepsilon^{-1} c\bar{\alpha}^2 (1-\rho)^{-1} \simeq \varepsilon^{-1} \tag{4.8}$$

and the corresponding computational cost are given by

$$\sum_{j=0}^{K-1} B_j = K \cdot \bar{B} \simeq |\log(\varepsilon^{-1})|\varepsilon^{-1}\,.$$

We summarize the derived batch-sampling strategies in the following theorem.

Theorem 4.2.5. *Suppose that the conditions of Theorem 4.2.2 are satisfied and define* $e_0 = \mathbb{E}[\|X_0 - x_*\|^2]$. *For* $\varepsilon > 0$ *and* $K \geq \lceil \log(\rho^{-1}) \log(\frac{\varepsilon}{2} e_0^{-1}) \rceil$, *let* $(X_k^{\mathrm{DS}})_{k=0,\dots,K}$ *be generated by Algorithm 8 with* $(B_k)_{k=0,\dots,K-1}$ *defined in (4.7). Moreover let* $(X_k^{\mathrm{FB}})_{k=0,\dots,K}$ *be generated by Algorithm 8 with fixed batch-size* $B_k = \bar{B}$ *given in (4.8) for* $k = 0, \dots, K-1$. *Then it holds true that*

$$e_K^{\mathrm{DS}} := \mathbb{E}[\|X_K^{\mathrm{DS}} - x_*\|^2] \leq \varepsilon$$
$$e_K^{\mathrm{FB}} := \mathbb{E}[\|X_K^{\mathrm{FB}} - x_*\|^2] \leq \varepsilon$$

while the computational cost are given by

$$\mathrm{cost}^{\mathrm{DS}} := \mathrm{cost}(X_K^{\mathrm{DS}}) = \sum_{j=0}^{K-1} B_j \simeq \varepsilon^{-1},$$
$$\mathrm{cost}^{\mathrm{FB}} := \mathrm{cost}(X_K^{\mathrm{FB}}) = K \cdot \bar{B} \simeq |\log(\varepsilon^{-1})|\varepsilon^{-1}\,.$$

4.2.2 Stochastic average gradient method (SAG)

In the following three sections, we consider different variance reduction methods for solving the empirical risk minimization problem. We fix the realization of the data set $\{z^{(1)}, \dots, z^{(N)}\}$ and ignore the error of the empirical approximation. For the next three algorithms, let $F_N : \mathbb{R}^d \to \mathbb{R}$ be a cost function of finite sum form

$$F_N(x) = \frac{1}{N} \sum_{i=1}^{N} f(x, z^{(i)}) =: \frac{1}{N} \sum_{i=1}^{N} f_i(x), \quad x \in \mathbb{R}^d,$$

for fixed $z^{(1)}, \dots, z^{(N)} \in \mathbb{R}^p$ and $f_i(x) := f(x, z^{(i)})$.

Assuming that F_N is μ-strongly convex and L-smooth, the deterministic gradient descent method applied to F_N converges linearly to the global optimum of F_N. However, in each iteration of the

algorithm we need to evaluate the gradients

$$\nabla_x f_i(x_k), \quad i = 1, \dots, N\,.$$

across the entire data set. In comparison, when applying SGD with finite data, Algorithm 7, in each iteration, we need to evaluate only one gradient $\nabla_x f_{\mathfrak{i}_{k+1}}(X_k)$. However, in this case, we lose the property of linear convergence to the global minimum.
We consider a class of algorithms, which apply variance reduction techniques, in order to obtain linear convergence for modified algorithms of SGD. In each iteration, the goal is to evaluate only one new gradient across the family of functions $\{f_i\}_{i=1,\dots,N}$ such as it is the case in Algorithm 7. Note that all of these algorithms essentially assume that the cost function is in form of a finite sum F_N.
In the following it is clear out of context, that the gradient $\nabla_x f_i(\cdot)$ is computed wrt. x, such that from now on we will omit the dependence on x and simply write $\nabla f_i(\cdot)$.

Motivation: Suppose that we want to estimate an unknown parameter $\theta \in \mathbb{R}$ and G be an unbiased estimator of θ, i.e. $\mathbb{E}[G] = \theta$. Moreover, let ξ be a random variable with mean close to zero, $\mathbb{E}[\xi] \approx 0$, such that the modified random variable $G_\xi := G - \xi$ is *nearly* unbiased, i.e. $\mathbb{E}[G_\xi] = \mathbb{E}[G] - \mathbb{E}[\xi] \approx \theta$. (In case $\mathbb{E}[\xi] = 0$, G_ξ even remains unbiased). The modification becomes interesting when considering the resulting variance:

$$\mathbb{V}(G_\xi) = \mathbb{V}(G - \xi) = \mathbb{V}(G) + \mathbb{V}(\xi) - 2\mathrm{Cov}(G, \xi)\,.$$

So in case we find a high (positive) correlation between G and ξ, we hope for a significant reduction of the variance without introducing a large bias. This concept can be viewed as motivation for the following three algorithms to be considered.

We consider the first algorithm which forms the basis for introducing variance reduction in SGD. In [23] the authors propose the *stochastic average gradient* (SAG) method, which achieves linear convergence for strongly convex cost functions while having same complexity characteristics as SGD. The idea is to reuse the gradient information obtained from the past.
The algorithm stores all gradient approximations across the entire data set $i = 1, \dots, N$, and in each iteration it updates the approximation of the gradient $f_{\mathfrak{i}}$ for only one randomly picked index $\mathfrak{i}$. Similarly to SGD in form of Algorithm 7, only one new gradient needs to be evaluated per iteration. However, one needs to have capacity for storing the gradient approximation for each index $i = 1, \dots, N$, which can be seen as the main disadvantage of SAG. As mentioned above, the algorithm achieves linear convergence toward the global optimum of F_N as shown in [23].

Algorithm 9 Stochastic average gradient method (SAG)

1: **Input:**
- cost function $F_N : \mathbb{R}^d \to \mathbb{R}$, $F_N(x) = \frac{1}{N}\sum_{i=1}^N f_i(x)$
- initial random variable $X_0 : \Omega \to \mathbb{R}^d$
- sequence of step sizes $(\alpha_k)_{k\in\mathbb{N}}$, $\alpha_k > 0$ (deterministic or $\mathcal{F}$-adapted)

2: set $k = 0$, initialize $G_0^{(i)} = 0$, $i = 1, \dots, N$
3: compute $\bar{G}_0 = \frac{1}{N}\sum_{i=1}^N G_0^{(i)}$
4: **while** "convergence/stopping criterion not met" **do**
5: generate independently $\mathfrak{i}_{k+1} \sim \mathcal{U}(\{1, \dots, N\})$
6: set $G_k^{(i)} = \begin{cases} \nabla f_i(X_k), & i = \mathfrak{i}_{k+1} \\ G_{k-1}^{(i)}, & i \neq \mathfrak{i}_{k+1} \end{cases}$
7: approximate the gradient $\nabla F_N(X_k)$ through

$$\bar{G}_k = \frac{1}{N}\sum_{i=1}^N G_k^{(i)} = \bar{G}_{k-1} - \frac{1}{N}G_{k-1}^{(\mathfrak{i}_{k+1})} + \frac{1}{N}\nabla f_{\mathfrak{i}_{k+1}}(X_k)$$

8: set $X_{k+1} = X_k - \alpha_k \bar{G}_k$, $k \mapsto k+1$
9: **end while**

Theorem 4.2.6 (Theorem 1 in [23]). *Let F_N, $f_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1, \dots, N$, be μ-strongly convex and L-smooth. Moreover, let $(X_k)_{k\in\mathbb{N}}$ be generated by Algorithm 9 with fixed step size $\alpha_k = \bar{\alpha} = \frac{1}{16L}$. Then it holds true that*

$$\frac{\mu}{2}\mathbb{E}[\|X_k - x_*\|^2] \leq \mathbb{E}[F_N(X_k) - F_N(x_*)] \leq \left(1 - \min\left\{\frac{\mu}{16L}, \frac{1}{8N}\right\}\right)^k C_0,$$

where $C_0 = \mathbb{E}[F_N(X_0) - F_N(x_*)] + \frac{4L}{N}\mathbb{E}[\|X_0 - x_*\|^2] + \frac{\sigma^2}{16L}$ with $\sigma^2 = \frac{1}{N}\sum_{i=1}^N \|\nabla f_i(x_*)\|^2$.

Remark 4.2.7. Ignoring the cost of storing $\{G_k^{(i)}\}_{i=1,\dots,N}$ we can run N iterations of Algorithm 9 to achieve a similar complexity as the deterministic full gradient descent scheme. To compare both SAG and deterministic GD we can view SAG as linearly converging scheme with rate

$$\rho^{\text{SAG}} = \left(1 - \min\left\{\frac{\mu}{16L}, \frac{1}{8N}\right\}\right)^N.$$

For small N ($N \leq \frac{2L}{\mu}$) the rate is dominated by

$$\rho^{\text{SAG}} = \left(1 - \frac{\mu}{16L}\right)^N$$

which corresponds to N steps of GD with step size $\bar{\alpha} = \frac{1}{16L}$. In comparison, for large N ($N \geq$

$\frac{2L}{\mu} = 2\kappa$) the rate is dominated by

$$\rho^{\text{SAG}} = \left(1 - \frac{1}{8N}\right)^N \leq \exp(-\frac{1}{8}),$$

such that the rate can be uniformly controlled in N.

4.2.3 SAGA

The convergence analysis of SAG is challenging due to the biased estimation of the gradients $\nabla F_N(X_k)$ in each iteration:

$$\mathbb{E}[\bar{G}_{k+1} \mid \mathcal{F}_k] = \underbrace{\mathbb{E}[\bar{G}_k - \frac{1}{N} G_k^{(\mathfrak{i}_{k+1})}]}_{\neq 0} + \frac{1}{N}\nabla F_N(X_k) = (1 - \frac{1}{N})\bar{G}_{k-1} + \frac{1}{N}\nabla F_N(X_k).$$

This problem can be solved by replacing the approximation of $\nabla F_N(X_k)$ through an unbiased estimator of form

$$\bar{G}_k = \frac{1}{N}\sum_{i=1}^N G_{k-1}^{(i)} - G_{k-1}^{(i)} + \nabla f_{\mathfrak{i}_{k+1}}(X_k).$$

This estimator is unbiased since

$$\mathbb{E}[\frac{1}{N}\sum_{i=1}^N G_{k-1}^{(i)} - G_{k-1}^{(i)} \mid \mathcal{F}_k] = 0 \quad \text{and} \quad \mathbb{E}[\nabla f_{\mathfrak{i}_{k+1}}(X_k) \mid \mathcal{F}_k] = \nabla F_N(X_k).$$

Note that $\bar{G}_k$ does not correspond to the mean over all stored gradients $\bar{G}_k \neq \frac{1}{N}\sum_{i=1}^N G_k^{(i)}$ anymore. This observation led to a modified algorithm called *SAGA* which has been introduced in [6].

We follow the proof of linear convergence under strong convexity for Algorithm 10 presented in [6]. Let us assume that F_N, f_i, $i = 1, \dots, N$ are μ-strongly convex and L-smooth. For $X_0(\omega) = x_0 \in \mathbb{R}^d$ we define point-wise

$$\phi_0^{(i)}(\omega) = x_0, \quad \phi_{k+1}^{(i)}(\omega) = \begin{cases} \phi_k^{(i)}(\omega), & i \neq \mathfrak{i}_{k+1}(\omega) \\ X_k(\omega), & i = \mathfrak{i}_{k+1}(\omega) \end{cases}$$

and consider the error function of form

$$E_k = \underbrace{c\|X_k - x_*\|^2}_{=:E_k^{(2)}} + \underbrace{\frac{1}{N}\sum_{i=1}^N \left(f_i(\phi_k^{(i)}) - f_i(x_*) - \langle \nabla f_i(x_*), \phi_k^{(i)} - x_*\rangle\right)}_{=:E_k^{(1)}} \geq c\|X_k - x_*\|^2,$$

Algorithm 10 SAGA

Input:

- cost function $F_N : \mathbb{R}^d \to \mathbb{R}$, $F_N(x) = \frac{1}{N}\sum_{i=1}^N f_i(x)$
- initial random variable $X_0 : \Omega \to \mathbb{R}^d$
- sequence of step sizes $(\alpha_k)_{k\in\mathbb{N}}$, $\alpha_k > 0$ (deterministic or $\mathcal{F}$-adapted)

set $k = 0$, initialize $G_0^{(i)} = 0$, $i = 1, \dots, N$
compute $\bar{G}_0 = \frac{1}{N}\sum_{i=1}^N G_0^{(i)}$
while "convergence/stopping criterion not met" **do**
    generate independently $\mathfrak{i}_{k+1} \sim \mathcal{U}(\{1, \dots, N\})$
    set $G_k^{(i)} = \begin{cases} \nabla f_i(X_k), & i = \mathfrak{i}_{k+1} \\ G_{k-1}^{(i)}, & i \neq \mathfrak{i}_{k+1} \end{cases}$
    approximate the gradient $\nabla F_N(X_k)$ through

$$\bar{G}_k = \bar{G}_{k-1} - G_{k-1}^{(\mathfrak{i}_{k+1})} + \nabla f_{\mathfrak{i}_{k+1}}(X_k) = \frac{1}{N}\sum_{i=1}^N G_k^{(i)}$$

    set $X_{k+1} = X_k - \alpha_k \bar{G}_k$, $k \mapsto k+1$
end while

where $x_* = \arg\min_{x\in\mathbb{R}^d} F_N(x)$. Note that $E_k^{(1)} \geq 0$ by convexity of f_i. We observe that

$$\|X_{k+1} - x_*\|^2 = \|X_k - x_*\|^2 - 2\bar{\alpha}\langle X_k - x_*, \bar{G}_k\rangle + \bar{\alpha}^2\|\bar{G}_k\|^2$$

and with $\mathcal{F}_k := \sigma(X_0, \mathfrak{i}_m, m \leq k)$ and $\mathbb{E}[\bar{G}_k \mid \mathcal{F}_k] = \nabla F_N(X_k)$ we have that

$$\mathbb{E}[\|X_{k+1} - x_*\|^2 \mid \mathcal{F}_k] = \|X_k - x_*\|^2 - 2\bar{\alpha}\langle X_k - x_*, \nabla F_N(X_k)\rangle + \bar{\alpha}^2\mathbb{E}[\|\bar{G}_k\|^2 \mid \mathcal{F}_k]\,.$$

In order to obtain an improved convergence result compared to SGD, we need to control $\mathbb{E}[\|\bar{G}_k\|^2 \mid \mathcal{F}_k]$ sufficiently well.

Lemma 4.2.8. Let $F_N,\ f_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1, \dots, N$, be μ-strongly convex and L-smooth. Then for any $\beta > 0$ it holds true that

$$\begin{aligned}\mathbb{E}[\|\bar{G}_k\|^2 \mid \mathcal{F}_k] \leq\ & (1+\beta^{-1})\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(\phi_k^{(i)}) - \nabla f_i(x_*)\|^2 \\ & + (1+\beta)\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(X_k) - \nabla f_i(x_*)\|^2 - \beta\|\nabla F_N(X_k)\|^2\,,\end{aligned}$$

where $x_* = \arg\min_{x\in\mathbb{R}^d} F_N(x)$.

Proof. We will apply multiple times the following equality

$$\mathbb{E}[\|Q-\mathbb{E}[Q]\|^2]=\mathbb{E}[\|Q\|^2]-\|\mathbb{E}[Q]\|^2 \tag{4.9}$$

for any random vector Q with $\mathbb{E}[\|Q\|^2]<\infty$. By construction of $\bar{G}_k$ we can write

$$\begin{aligned}\mathbb{E}[\|\bar{G}_k\|^2\mid\mathcal{F}_k]&=\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^N\nabla f_i(\phi_k^{(i)})-\nabla f_{\mathrm{i}_{k+1}}(\phi_k^{(\mathrm{i}_{k+1})})+\nabla f_{\mathrm{i}_{k+1}}(X_k)\|^2\mid\mathcal{F}_k]\\&=:\mathbb{E}[\|Q\|^2\mid\mathcal{F}_k]=\mathbb{E}[\|Q-\mathbb{E}[Q\mid\mathcal{F}_k]\|^2\mid\mathcal{F}_k]+\|\mathbb{E}[Q\mid\mathcal{F}_k]\|^2\\&=\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^N\nabla f_i(\phi_k^{(i)})-\nabla f_{\mathrm{i}_{k+1}}(\phi_k^{(\mathrm{i}_{k+1})})+\nabla f_{\mathrm{i}_{k+1}}(X_k)-\nabla F_N(X_k)\|^2\mid\mathcal{F}_k]\\&\quad+\|\nabla F_N(X_K)\|^2\end{aligned}$$

Let us consider the first term

$$\begin{aligned}&\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^N\nabla f_i(\phi_k^{(i)})-\nabla f_{\mathrm{i}_{k+1}}(\phi_k^{(\mathrm{i}_{k+1})})+\nabla f_{\mathrm{i}_{k+1}}(X_k)-\nabla F_N(X_k)\|^2\mid\mathcal{F}_k]\\&=\mathbb{E}[\|\left\{\nabla f_{\mathrm{i}_{k+1}}(\phi_k^{(\mathrm{i}_{k+1})})-\nabla f_{\mathrm{i}_{k+1}}(x_*)-\frac{1}{N}\sum_{i=1}^N\nabla f_i(\phi_k^{(i)})\right\}\\&\qquad-\left\{\nabla f_{\mathrm{i}_{k+1}}(X_k)-\nabla f_{\mathrm{i}_{k+1}}(x_*)-\nabla F_N(X_k)\right\}\|^2\mid\mathcal{F}_k]\\&\le(1+\beta^{-1})\mathbb{E}[\|\nabla f_{\mathrm{i}_{k+1}}(\phi_k^{(\mathrm{i}_{k+1})})-\nabla f_{\mathrm{i}_{k+1}}(x_*)-\frac{1}{N}\sum_{i=1}^N\nabla f_i(\phi_k^{(i)})\|^2\mid\mathcal{F}_k]\\&\quad+(1+\beta)\mathbb{E}[\|\nabla f_{\mathrm{i}_{k+1}}(X_k)-\nabla f_{\mathrm{i}_{k+1}}(x_*)-\nabla F_N(X_k)\|^2\mid\mathcal{F}_k]\,,\end{aligned}$$

where we have used that $\|x+y\|^2\le(1+\beta^{-1})\|x\|^2+(1+\beta)\|y\|^2$ for any $\beta>0$. We define $Q_1=\nabla f_{\mathrm{i}_{k+1}}(\phi_k^{(\mathrm{i}_{k+1})})-\nabla f_{\mathrm{i}_{k+1}}(x_*)$ with

$$\mathbb{E}[Q_1\mid\mathcal{F}_k]=\frac{1}{N}\sum_{i=1}^N\nabla f_i(\phi_k^{(i)})-\nabla F_N(x_*)=\frac{1}{N}\sum_{i=1}^N\nabla f_i(\phi_k^{(i)})\,,$$

and similarly

$$Q_2=\nabla f_{\mathrm{i}_{k+1}}(X_k)-\nabla f_{\mathrm{i}_{k+1}}(x_*)\quad\text{with}\quad\mathbb{E}[Q_2\mid\mathcal{F}_k]=\nabla F_N(X_k)\,.$$

Finally, we obtain

$$\begin{aligned}
&(1+\beta^{-1})\mathbb{E}[\|\nabla f_{i_{k+1}}(\phi_k^{(i_{k+1})}) - \nabla f_{i_{k+1}}(x_*) - \frac{1}{N}\sum_{i=1}^N \nabla f_i(\phi_k^{(i)})\|^2 \mid \mathcal{F}_k] \\
&\quad + (1+\beta)\mathbb{E}[\|\nabla f_{i_{k+1}}(X_k) - \nabla f_{i_{k+1}}(x_*) - \nabla F_N(X_k)\|^2 \mid \mathcal{F}_k] \\
&\leq (1+\beta^{-1})\left\{\mathbb{E}[\|\nabla f_{i_{k+1}}(\phi_k^{(i_{k+1})}) - \nabla f_{i_{k+1}}(x_*)\|^2 \mid \mathcal{F}_k] - \|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\phi_k^{(i)})\|^2\right\} \\
&\quad + (1+\beta)\left\{\mathbb{E}[\|\nabla f_{i_{k+1}}(X_k) - \nabla f_{i_{k+1}}(x_*)\|^2 \mid \mathcal{F}_k] - \|\nabla F_N(X_k)\|^2\right\}
\end{aligned}$$

and all together

$$\begin{aligned}
\mathbb{E}[\|\bar{G}_k\|^2 \mid \mathcal{F}_k] \leq (1+\beta^{-1})\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(\phi_k^{(i)}) - \nabla f_i(x_*)\|^2 \\
+ (1+\beta)\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(X_k) - \nabla f_i(x_*)\|^2 - \beta\|\nabla F_N(X_k)\|^2 .
\end{aligned}$$

□

Applying the upper bound in Lemma 4.2.8 results in

$$\begin{aligned}
\mathbb{E}[\|X_{k+1} - x_*\|^2 \mid \mathcal{F}_k] \leq \|X_k - x_*\|^2 - 2\bar{\alpha}\langle X_k - x_*, \nabla F_N(X_k)\rangle - \bar{\alpha}^2\beta\|\nabla F_N(X_k)\|^2 \\
+ \bar{\alpha}^2(1+\beta^{-1})\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(\phi_k^{(i)}) - \nabla f_i(x_*)\|^2 \\
+ \bar{\alpha}^2(1+\beta)\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(X_k) - \nabla f_i(x_*)\|^2 .
\end{aligned} \tag{4.10}$$

With the following Lemma, we are able to control $\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(X_k) - \nabla f_i(x_*)\|^2$ by setting it in relation to $-\langle X_k - x_*, \nabla F_N(X_k)\rangle$.

Lemma 4.2.9. Let $F_N,\ f_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1, \dots, N$, be μ-strongly convex and L-smooth. Then for all $x \in \mathbb{R}^d$ it holds true that

$$\begin{aligned}
\langle \nabla F_N(x), x_* - x\rangle \leq -\frac{L-\mu}{L}(F_N(x) - F_N(x_*)) - \frac{\mu}{2}\|x - x_*\|^2 \\
- \frac{1}{2L}\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(x) - \nabla f_i(x_*)\|^2 ,
\end{aligned}$$

where $x_* = \arg\min_{x\in\mathbb{R}^d}\ F_N(x)$.

The proof of Lemma 4.2.9 is left as an exercise for the interested reader. We can now combine the

upper bound in Lemma 4.2.9 with (4.10) to obtain

$$\begin{aligned}\mathbb{E}[\|X_{k+1}-x_*\|^2 \mid \mathcal{F}_k] \leq (1-\bar{\alpha}\mu)\|X_k-x_*\|^2 - 2\bar{\alpha}\frac{L-\mu}{L}\left(F_N(X_k)-F_N(x_*)\right) - \bar{\alpha}^2\beta\|\nabla F_N(X_k)\|^2 \\ + (\bar{\alpha}^2(1+\beta)-\frac{\bar{\alpha}}{L})\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(X_k)-\nabla f_i(x_*)\|^2 \\ + \bar{\alpha}^2(1+\beta^{-1})\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(\phi_k^{(i)})-\nabla f_i(x_*)\|^2\end{aligned}$$

It remains to control

$$\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(\phi_k^{(i)})-\nabla f_i(x_*)\|^2 .$$

Since all f_i, $i=1,\dots,N$, are assumed to be L-smooth with the same $L>0$, we have that

$$\|\nabla f_i(\phi_k^{(i)})-\nabla f_i(x_*)\|^2 \leq 2L(f_i(\phi_k^{(i)})-f_i(x_*)-\langle \nabla f_i(x_*), \phi_k^{(i)}-x_*\rangle)$$

(see also Lemma 4.2.14 below) implying that

$$\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(\phi_k^{(i)})-\nabla f_i(x_*)\|^2 \leq \frac{1}{N}\sum_{i=1}^N f_i(\phi_k^{(i)}) - F_N(x_*) - \frac{1}{N}\sum_{i=1}^N \langle \nabla f_i(x_*), \phi_k^{(i)}-x_*\rangle = E_k^{(1)},$$

which also explains the origin of the error function $E_k = E_k^{(2)} + E_k^{(1)}$. We are now ready to prove the linear convergence of SAGA.

Theorem 4.2.10 (Theorem 1 in [6]). *Let F_N, $f_i : \mathbb{R}^d \to \mathbb{R}$, $i=1,\dots,N$, be μ-strongly convex and L-smooth, and let X_0 be a random variable such that $\mathbb{E}[E_0] = c\mathbb{E}[\|X_0-x_*\|^2] + \mathbb{E}[F_N(X_0)-F_N(x_*)] < \infty$. Moreover, let $(X_k)_{k\in\mathbb{N}}$ be generated by Algorithm 10 with fixed step size $\alpha_k = \bar{\alpha} = \frac{1}{2(\mu N+L)}$. Then for $c = \frac{1}{2\bar{\alpha}(1-\bar{\alpha}\mu)N}$ it holds true that*

$$\mathbb{E}[E_{k+1}] \leq (1-\bar{\alpha}\mu)\mathbb{E}[E_k] .$$

Proof. Firstly, we observe that each $\phi_{k+1}^{(i)}$ given $\mathcal{F}_k$ is distributed according to

$$\phi_{k+1}^{(i)} \mid \mathcal{F}_k \sim \frac{1}{N}\delta_{X_k} + (1-\frac{1}{N})\delta_{\phi_k^{(i)}} ,$$

such that

$$\mathbb{E}[\frac{1}{N}\sum_{i=1}^N f_i(\phi_{k+1}^{(i)}) \mid \mathcal{F}_k] = \frac{1}{N}\sum_{i=1}^N \mathbb{E}[f_i(\phi_{k+1}^{(i)}) \mid \mathcal{F}_k] = \frac{1}{N}\sum_{i=1}^N \left(\frac{1}{N} f_i(X_k) + (1-\frac{1}{N}) f_i(\phi_k^{(i)})\right)$$
$$= \frac{1}{N} F_N(X_k) + (1-\frac{1}{N})\frac{1}{N}\sum_{i=1}^N f_i(\phi_k^{(i)})\,.$$

With a similar computation we also obtain

$$\mathbb{E}[-\frac{1}{N}\sum_{i=1}^N \langle \nabla f_i(x_*), \phi_{k+1}^{(i)} - x_*\rangle] = -\frac{1}{N}\langle \nabla F_N(x_*), X_k - x_*\rangle - (1-\frac{1}{N})\sum_{i=1}^N \langle \nabla f_i(x_*), \phi_k^{(i)} - x_*\rangle$$
$$= -(1-\frac{1}{N})\sum_{i=1}^N \langle \nabla f_i(x_*), \phi_k^{(i)} - x_*\rangle\,.$$

Finally, we obtain the iterative error bound

$$\begin{aligned}
\mathbb{E}[E_{k+1} \mid \mathcal{F}_k] &\leq \left(\frac{1}{N} - c2\bar{\alpha}\frac{L-\mu}{L}\right)(F_N(X_k) - F_N(x_*)) - c\bar{\alpha}^2\beta\|\nabla F_N(X_K)\|^2 \\
&\quad + (1-\bar{\alpha}\mu)\, c\|X_k - x_*\|^2 \\
&\quad + \left(1 - \frac{1}{N} + 2c(1+\beta^{-1})\bar{\alpha}^2 L - \bar{\alpha}\mu + \bar{\alpha}\mu\right) E_k^{(1)} \\
&\quad + \left(c\bar{\alpha}(\bar{\alpha}(1+\beta) - \frac{1}{L})\right)\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(X_k) - \nabla f_i(x_*)\|^2 \\
&\leq (1-\bar{\alpha}\mu)E_k \\
&\quad + \left(\frac{1}{N} - 2c\bar{\alpha}\frac{L-\mu}{L} - 2\mu\bar{\alpha}^2\beta c\right)\cdot(F_N(X_k) - F_N(x_*)) \\
&\quad + \left(\bar{\alpha}\mu - \frac{1}{N} + 2c(1+\beta^{-1})\bar{\alpha}^2 L\right) E_k^{(1)} \\
&\quad + c\bar{\alpha}(\bar{\alpha}(1+\beta) - \frac{1}{L})\frac{1}{N}\sum_{i=1}^N \|\nabla f_i(X_k) - \nabla f_i(x_*)\|^2\,,
\end{aligned}$$

where we have used $-\|\nabla F_N(X_K)\|^2 \leq -2\mu(F_N(X_k) - F_N(x_*))$ by μ-strong convexity of F_N. With the choice $\bar{\alpha} = \frac{1}{2(\mu N + L)}$, $\beta = \frac{2\mu N + L}{L}$ and $c = \frac{1}{2\bar{\alpha}(1-\bar{\alpha}\mu)N}$ one can verify that

- $\bar{\alpha}(1+\beta) - \frac{1}{L} = 0$,
- $\bar{\alpha}\mu - \frac{1}{N} + 2c(1+\beta^{-1})\bar{\alpha}^2 L \leq 0$,
- $\frac{1}{N} - 2c\bar{\alpha}\frac{L-\mu}{L} - 2\mu\bar{\alpha}^2\beta c = 0$,

such that we conclude the proof with

$$\mathbb{E}[E_{k+1}] = \mathbb{E}[\mathbb{E}[E_{k+1} \mid \mathcal{F}_k]] \leq (1 - \bar{\alpha}\mu)\mathbb{E}[E_k]\,.$$

□

We observe that $c\mathbb{E}[\|X_k - x_*\|^2] \leq E_k$ such that we obtain the following corollary:

Corollary 4.2.11 (Corollary 1 in [6])**.** Under the same assumptions as in Theorem 4.2.10 it holds true that

$$\mathbb{E}[\|X_k - x_*\|^2] \leq \left(1 - \frac{\mu}{2(\mu N + L)}\right)^k \left(\mathbb{E}[\|X_0 - x_*\|^2 + \frac{N}{\mu N + L}\mathbb{E}[F_N(X_k) - F_N(x_*)]\right),$$

where $x_* = \arg\min_{x\in\mathbb{R}^d} \; F_N(x)$.

Remark 4.2.12. We emphasize again, that for both SAG and SAGA it is necessary to store $\{G_k^{(i)}\}_{i=1,\dots,N}$ which may occur with additional cost.

In the following, we take a look at Table 1 of [23] (extended by the values for SAGA), where the theoretically derived rates of convergence of SAG (and SAGA) are compared to various deterministic first order methods, which in each iteration need to evaluate the gradients across the entire data set $i = 1, \dots, N$. We observe a significant improvement through SAG and SAGA.

Algorithm	$\bar{\alpha}$	rate	$\mu = 0.01$	$\mu = 0.0001$
GD	$\frac{1}{L}$	$(1 - \frac{\mu}{L})^2$	~ 0.9998	~ 1
GD	$\frac{2}{\mu+L}$	$(1 - \frac{2\mu}{L+\mu})^2$	~ 0.9996	~ 1
NAM	$\frac{1}{L}$	$(1 - \sqrt{\frac{\mu}{L}})$	~ 0.99	~ 0.999
lower bound	–	$(1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}})^2$	~ 0.9608	~ 0.996
SAG	$\frac{1}{16L}$	$(1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^N$	~ 0.8825	~ 0.9938
SAGA	$\frac{1}{2(\mu N+L)}$	$(1 - \frac{\mu}{2(\mu N+L)})^N$	~ 0.635	~ 0.956

Table 4.1: For both scenarios we assume that $L = 100$ and $N = 10^5$.

4.2.4 Stochastic variance reduced gradient (SVRG)

In the following, we consider a variance reduction method for SGD that avoids storing gradients across the entire dataset. The *stochastic variance reduced gradient* (SVRG) method has been introduced in [11]. The algorithm operates cyclic by a loop of SGD followed by an exact gradient update. Every M iterations, the scheme updates the gradient information across the entire data set. For each cycle of SGD followed by the exact gradient iteration, we need to evaluate $2M + N$ gradients. In order to compare the algorithm to GD, SAG and SAGA, we will need to pay attention to this observation and rescale the effective rate of convergence accordingly.

Algorithm 11 Stochastic variance reduced gradient method (SVRG)

Input:

- cost function $F_N : \mathbb{R}^d \to \mathbb{R}$, $F_N(x) = \frac{1}{N}\sum_{i=1}^N f_i(x)$
- initial random variable $X_0 : \Omega \to \mathbb{R}^d$
- length of cycle $M \geq 1$
- sequence of step sizes $(\alpha_k^{(m)})_{k\in\mathbb{N},\ m=0,\dots,M-1}$, $\alpha_k^{(m)} > 0$ (deterministic or $\mathcal{F}$-adapted)

set $k = 0$
set $\tilde{X}_0 = X_0^{(0)}$, $\tilde{G}_0 = \nabla F_N(\tilde{X}_0)$
set $X_0^{(0)} = X_0$ $\bar{G}_0^{(0)} = \tilde{G}_0$
while "convergence/stopping criterion not met" **do**
  for $m = 0, \dots, M-1$ **do**
    generate independently $\mathfrak{i}_{k+1}^{(m+1)} \sim \mathcal{U}(\{1, \dots, N\})$
    approximate the gradient $\nabla F_N(X_k^{(m)})$ through

$$\bar{G}_k^{(m)} = \nabla f_{\mathfrak{i}_{k+1}^{(m+1)}}(X_k^{(m)}) - \nabla f_{\mathfrak{i}_{k+1}^{(m+1)}}(\tilde{X}_k) + \tilde{G}_k$$

    set $X_k^{(m+1)} = X_k^{(m)} - \alpha_k^{(m)} \bar{G}_k^{(m)}$,
  end for
  generate independently $\mathfrak{m}_{k+1} \sim \mathcal{U}(\{0, \dots, M-1\})$,
  set $\tilde{X}_{k+1} = X_k^{(\mathfrak{m}_{k+1})}$,
  set $\tilde{G}_{k+1} = \nabla F_N(\tilde{X}_{k+1}) = \frac{1}{N}\sum_{i=1}^N \nabla f_i(\tilde{X}_{k+1})$
  set $X_{k+1}^{(0)} = \tilde{X}_{k+1}$, $\bar{G}_{k+1}^{(0)} = \tilde{G}_{k+1}$
  set $k \mapsto k+1$
end while

Remark 4.2.13. As mentioned above, the advantage of SVRG compared to SAG and SAGA is the avoidance of storing all gradients for $i = 1, \dots, N$.

We will follow the proof of linear convergence of SVRG presented in [11]. Let us start with the following auxiliary bound on the gradients ∇f_i, $i = 1, \dots, N$. Note that this result can be viewed

as an extension of the inequality

$$\frac{1}{2L}\|\nabla F_N(x) - \nabla F_N(x_*)\|^2 \leq F_N(x) - F_N(x_*),$$

which we have derived under L-smoothness and convexity of F_N.

Lemma 4.2.14. Let $F_N,\ f_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1, \dots, N$, be convex and L-smooth. Then for all $x \in \mathbb{R}^d$ it holds true that

$$\frac{1}{N}\sum_{i=1}^{N} \|\nabla f_i(x) - \nabla f_i(x_*)\|^2 \leq 2L\left(F_N(x) - F_N(x_*)\right),$$

where $x_* = \arg\min_{x\in\mathbb{R}^d}\ F_N(x)$.

Proof. For each $i = 1, \dots, N$, define the function

$$\varphi_i(x) = f_i(x) - f_i(x_*) - \langle \nabla f_i(x_*), x - x_*\rangle \geq 0$$

such that $\nabla\varphi_i(x) = \nabla f_i(x) - \nabla f_i(x_*)$ and $\nabla\varphi_i(x_*) = 0$. Since $\varphi_i(x_*) = 0$ and $\varphi_i(x) \geq 0$ by convexity of f_i, we observe that $\varphi_i(x_*) = \min_{x\in\mathbb{R}^d}\ \varphi_i(x)$. For any $x \in \mathbb{R}^d$ it holds that

$$\begin{aligned} 0 = \varphi_i(x_*) \leq \min_{\alpha\geq 0}\ \varphi_i(x - \alpha\nabla f_i(x)) &\leq \min_{\alpha\geq 0}\ \left(\varphi_i(x) - \alpha(1 - \frac{L}{2}\alpha)\|\nabla\varphi_i(x)\|^2\right) \\ &= \varphi_i(x) - \frac{1}{2L}\|\nabla\varphi_i(x)\|^2, \end{aligned}$$

where we have used that φ_i is L-smooth (since $\nabla\varphi_i(x) = \nabla f_i(x) - \nabla f_i(x_*)$ remains L-Lipschitz continuous) and that $\max_\alpha \alpha(1 - \frac{L}{2}\alpha) = \frac{1}{2L}$. Hence, it follows that

$$\|\nabla\varphi_i(x)\|^2 = \|\nabla f_i(x) - \nabla f_i(x_*)\|^2 \leq 2L\varphi_i(x) = 2L\left(f_i(x) - f_i(x_*) - \langle\nabla f_i(x_*), x - x_*\rangle\right).$$

Taking the average over all $i = 1, \dots, N$, finishes the proof

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} \|\nabla f_i(x) - \nabla f_i(x_*)\|^2 &\leq 2L\left(\frac{1}{N}\sum_{i=1}^{N} f_i(x) - \frac{1}{N}\sum_{i=1}^{N} f_i(x_*) - \langle\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_*), x - x_*\rangle\right) \\ &= 2L\left(F_N(x) - F_N(x_*)\right). \end{aligned}$$

□

Let $(X_k^{(m)})_{k\in\mathbb{N},\ m=0,\dots,M}$ be generated by Algorithm 11 with fixed step size $\bar{\alpha} > 0$ and consider the

following natural filtration

$$\mathcal{F}_k^{(m)} = \begin{cases} \sigma(\{X_0\}, \{\mathfrak{i}_\ell^{(s)}, \ell \leq k, s \leq M\}, \{\mathfrak{i}_k^{(s)}, s \leq m\}, \{\mathfrak{m}_\ell, \ell \leq k\}), & k \geq 1 \\ \sigma(\{X_0\}, \{\mathfrak{i}_0^{(s)}, s \leq m\}), & k = 0 \end{cases}.$$

Similarly as in the case of SAGA, we have that

$$\mathbb{E}[\|X_k^{(m+1)} - x_*\|^2 \mid \mathcal{F}_k^{(m)}] = \|X_k^{(m)} - x_*\|^2 - 2\bar{\alpha}\langle X_k^{(m)} - x_*, \nabla F(X_k^{(m)})\rangle + \bar{\alpha}^2 \mathbb{E}[\|\bar{G}_k^{(m)}\|^2 \mid \mathcal{F}_k^{(m)}],$$

since $\bar{G}_k^{(m)}$ is an unbiased estimator of $\nabla F_N(X_k^{(m)})$:

$$\mathbb{E}[\bar{G}_k^{(m)} \mid \mathcal{F}_k^{(m)}] = \mathbb{E}[\nabla f_{\mathfrak{i}_{k+1}^{(m+1)}}(X_k^{(m)}) \mid \mathcal{F}_k^{(m)}] - \underbrace{\mathbb{E}[\nabla f_{\mathfrak{i}_{k+1}^{(m+1)}}(\tilde{X}_k) - \tilde{G}_k \mid \mathcal{F}_k^{(m)}]}_{=0} = \nabla F_N(X_k^{(m)}).$$

We again aim to control $\mathbb{E}[\|\bar{G}_k^{(m)}\|^2 \mid \mathcal{F}_k^{(m)}]$ in order to achieve significant variance reduction.

Lemma 4.2.15. Let F_N, $f_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1, \dots, N$, be μ-strongly convex and L-smooth, and let X_0 be a random variable such that $\mathbb{E}[\|X_0\|^2] + \mathbb{E}[|F_N(X_0)|] < \infty$. Moreover, let $(X_k^{(m)})_{k\in\mathbb{N}, m=0,\dots,M}$ be generated by Algorithm 11 with fixed step size $\alpha_k^{(m)} = \bar{\alpha} > 0$. Then for all $k \geq 0$ and $m \geq 1$ it holds true that

$$\mathbb{E}[\|\bar{G}_k^{(m)}\|^2 \mid \mathcal{F}_k^{(m)}] \leq 4L \left(F_N(X_k^{(m)}) - F_N(x_*) + F_N(\tilde{X}_k) - F_N(x_*) \right),$$

where $x_* = \arg\min_{x\in\mathbb{R}^d} F_N(x)$.

Proof. The proof follows by similar argumentation as the proof of Lemma 4.2.8. We again apply (4.9) together with $\|x + y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$:

$$\begin{aligned}
\mathbb{E}[\|\bar{G}_k^{(m)}\|^2 \mid \mathcal{F}_k^{(m)}] &\leq 2\mathbb{E}[\|\nabla f_{\mathfrak{i}_{k+1}^{(m+1)}}(X_k^{(m)}) - \nabla f_{\mathfrak{i}_{k+1}^{(m+1)}}(x_*)\|^2 \mid \mathcal{F}_k^{(m)}] \\
&\quad + 2\mathbb{E}[\|\nabla f_{\mathfrak{i}_{k+1}^{(m+1)}}(\tilde{X}_k) - \nabla f_{\mathfrak{i}_{k+1}^{(m+1)}}(x_*) - \nabla F_N(\tilde{X}_k)\|^2 \mid \mathcal{F}_k^{(m)}] \\
&= 2\mathbb{E}[\|\nabla f_{\mathfrak{i}_{k+1}^{(m+1)}}(X_k^{(m)}) - \nabla f_{\mathfrak{i}_{k+1}^{(m+1)}}(x_*)\|^2 \mid \mathcal{F}_k^{(m)}] \\
&\quad + 2\mathbb{E}[\|\{\nabla f_{\mathfrak{i}_{k+1}^{(m+1)}}(\tilde{X}_k) - \nabla f_{\mathfrak{i}_{k+1}^{(m+1)}}(x_*)\} - \{\nabla F_N(\tilde{X}_k) - \nabla F_N(x_*)\}\|^2 \mid \mathcal{F}_k^{(m)}] \\
&\leq 2\mathbb{E}[\|\nabla f_{\mathfrak{i}_{k+1}^{(m+1)}}(X_k^{(m)}) - \nabla f_{\mathfrak{i}_{k+1}^{(m+1)}}(x_*)\|^2 \mid \mathcal{F}_k^{(m)}] \\
&\quad + 2\mathbb{E}[\|\nabla f_{\mathfrak{i}_{k+1}^{(m+1)}}(\tilde{X}_k) - \nabla f_{\mathfrak{i}_{k+1}^{(m+1)}}(x_*)\|^2 \mid \mathcal{F}_k^{(m)}] \\
&= \frac{2}{N}\sum_{i=1}^N \left(\|\nabla f_i(X_k^{(m)}) - \nabla f_i(x_*)\|^2 + \|\nabla f_i(\tilde{X}_k) - \nabla f_i(x_*)\|^2 \right).
\end{aligned}$$

The assertion follows by application of Lemma 4.2.14. □

We are now ready to prove linear convergence of SVRG under strong convexity.

Theorem 4.2.16 (Theorem 1 in [11]). *Let F_N, $f_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1, \dots, N$, be μ-strongly convex and L-smooth, and let X_0 be a random variable such that $\mathbb{E}[\|X_0\|^2] + \mathbb{E}[|F_N(X_0)|] < \infty$. Moreover, let $(X_k^{(m)})_{k\in\mathbb{N},\ m=0,\dots,M}$ be generated by Algorithm 11 with fixed step size $\alpha_k^{(m)} = \bar{\alpha} > 0$ and $M \geq 1$ such that*

$$0 < \rho := \frac{1}{\mu(1-2\bar{\alpha}L)\bar{\alpha}M} + \frac{2\bar{\alpha}L}{1-2\bar{\alpha}L} < 1\,.$$

Then it holds true that

$$\mathbb{E}[F_N(\tilde{X}_k) - F_N(x_*)] \leq \rho^k \mathbb{E}[F_N(X_0) - F_N(x_*)]\,.$$

Proof. By convexity of F_N we have

$$\begin{aligned}
\mathbb{E}[\|X_k^{(m+1)} - x_*\|^2 \mid \mathcal{F}_k^{(m)}] &\leq \|X_k^{(m)} - x_*\|^2 - 2\bar{\alpha}\left(F_N(X_k^{(m)}) - F_N(x_*)\right) + \bar{\alpha}^2\mathbb{E}[\|\bar{G}_k^{(m)}\|^2 \mid \mathcal{F}_k^{(m)}] \\
&\leq \|X_k^{(m)} - x_*\|^2 - 2\bar{\alpha}\left(F_N(X_k^{(m)}) - F_N(x_*)\right) \\
&\quad + 4L\bar{\alpha}^2\left(F_N(X_k^{(m)}) - F_N(x_*)\right) + 4L\bar{\alpha}^2\left(F_N(\tilde{X}_k) - F_N(x_*)\right) \\
&= \|X_k^{(m)} - x_*\|^2 - 2\bar{\alpha}(1-2L\bar{\alpha})\left(F_N(X_k^{(m)}) - F_N(x_*)\right) \\
&\quad + 4L\bar{\alpha}^2\left(F_N(\tilde{X}_k) - F_N(x_*)\right)
\end{aligned}$$

where we have applied Lemma 4.2.15. By construction of the Algorithm 11 we have

$$\mathbb{E}[F_N(\tilde{X}_{k+1}) - F_N(x_*) \mid \mathcal{F}_k^{(M)}] = \frac{1}{M}\sum_{m=0}^{M-1}\left(F_N(X_k^{(m)}) - F_N(x_*)\right),$$

such that

$$\begin{aligned}
&\mathbb{E}[\|X_k^{(M)} - x_*\|^2] + 2\bar{\alpha}(1-2L\bar{\alpha})M\mathbb{E}[F_N(\tilde{X}_{k+1}) - F_N(x_*)] \\
&\leq \mathbb{E}[\|X_k^{(0)} - x_*\|^2] - 2\bar{\alpha}(1-2L\bar{\alpha})\sum_{m=0}^{M-1}\mathbb{E}[F_N(X_k^{(m)}) - F_N(x_*)] \\
&\quad + 2\bar{\alpha}(1-2L\bar{\alpha})M\frac{1}{M}\sum_{m=0}^{M-1}\mathbb{E}[F_N(X_k^{(m)}) - F_N(x_*)] \\
&\quad + 4L\bar{\alpha}^2 M\mathbb{E}[F_N(\tilde{X}_k) - F_N(x_*)] \\
&\leq \frac{2}{\mu}\mathbb{E}[F_N(X_k^{(0)}) - F_N(x_*)] + 4L\bar{\alpha}^2 M\mathbb{E}[F_N(\tilde{X}_k) - F_N(x_*)] \\
&= 2(\frac{1}{\mu} + 2L\bar{\alpha}^2 M)\mathbb{E}[F_N(\tilde{X}_k) - F_N(x_*)]\,,
\end{aligned}$$

where we have used that $\frac{\mu}{2}\|x - x_*\|^2 \leq F_N(x) - F_N(x_*)$ by strong convexity of F_N. Finally, we

have

$$\mathbb{E}[F_N(\tilde{X}_{k+1}) - F_N(x_*)] \leq \left(\frac{1}{\mu(1-2\bar{\alpha}L)\bar{\alpha}M} + \frac{2\bar{\alpha}L}{1-2\bar{\alpha}L}\right)\mathbb{E}[F_N(\tilde{X}_k) - F_N(x_*)]\,.$$

□

Remark 4.2.17. Returning to the setting of Table 4.1. For $L = 100$, $\mu = 0.01$ and $N = 10^5 = 10\kappa$, we want to choose $\bar{\alpha} = \frac{\tau}{L}$, $\tau \leq \frac{1}{2}$ such that ρ defined in Theorem 4.2.16 is gieven by

$$\rho = \frac{1}{1-2\tau}\left(\frac{\kappa}{\tau M} + 2\tau\right) = \frac{1}{1-2\tau}\left(\frac{N}{10\tau M} + 2\tau\right)\,.$$

We set $\tau = \frac{1}{10}$, $M = 4N$ such that $\rho = \frac{9}{16}$. The number of gradient evaluations of one cycle in SVRG is $2M + N$, such that the effective rate of convergence compared to full GD is

$$\rho^{\text{SVRG}} = \rho^{\frac{N}{2M+N}} = \left(\frac{9}{16}\right)^{\frac{1}{9}} \approx 0.94,$$

which still achieves a better convergence rate than GD. Note that the above choice of step size has been selected by trial and might be improved significantly. Moreover, we emphasize again that there are no additional storage cost when applying Algorithm 11, such as it was the case for Algorithm 9 and Algorithm 10.

Bibliography

[1] Dimitri P. Bertsekas. *Nonlinear programming*. Athena Scientific, 2nd edition, September 2008.

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

[3] Julius R. Blum. Approximation Methods which Converge with Probability one. *The Annals of Mathematical Statistics*, 25(2):382 – 386, 1954.

[4] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[5] G. Da Prato and J. Zabczyk. *Stochastic Equations in Infinite Dimensions*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1992.

[6] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[7] Guillaume Garrigos and Robert M. Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv-Preprint*, 2023.

[8] J. Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, pages 49–52, 1902.

[9] Arnulf Jentzen, Benno Kuckuck, Ariel Neufeld, and Philippe von Wurstemberger. Strong error analysis for stochastic gradient descent optimization algorithms. *IMA Journal of Numerical Analysis*, 41(1):455–492, 05 2020.

[10] Arnulf Jentzen and Philippe von Wurstemberger. Lower error bounds for the stochastic gradient descent optimization algorithm: Sharp convergence rates for slowly and fast decaying learning rates. *Journal of Complexity*, 57:101438, 2020.

[11] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[12] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. pages 795–811, 2016.

[13] H.K. Khalil. *Nonlinear Systems*. Pearson Education. Prentice Hall, 2002.

[14] Achim Klenke. *Wahrscheinlichkeitstheorie*. Springer, 2006.

[15] Guanghui Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer Cham, 1st edition, 2021.

[16] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[17] Yurii Nesterov. *Lectures on Convex Optimization*. Springer Cham, 2nd edition, 2018.

[18] B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[19] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999.

[20] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In Jagdish S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257. Academic Press, 1971.

[21] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951.

[22] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv-Preprint*, 2017.

[23] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, 2017.

[24] Simon Weissmann, Ashia Wilson, and Jakob Zech. Multilevel optimization for inverse problems. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 5489–5524. PMLR, 02–05 Jul 2022.

[25] Ashia C. Wilson, Ben Recht, and Michael I. Jordan. A Lyapunov analysis of accelerated methods in Optimization. *J. Mach. Learn. Res.*, 22(1), jan 2021.

A

Appendix

A.1 Convex sets and functions

We will give a brief overview of convex sets and functions. We start with the following definition.

Definition A.1.1 (convex set)**.** A subset $C \subset \mathbb{R}^d$ is called *convex*, if

$$\lambda x + (1 - \lambda)y \in C$$

for all $x, y \in C$ and $\lambda \in [0, 1]$.

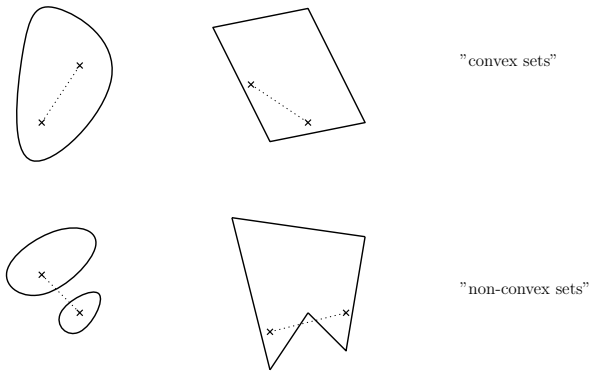

Figure A.1: Examples of convex and non-convex sets.

Similarly, we can define convex functions.

Definition A.1.2 (convex function)**.** Let $C \subset \mathbb{R}^d$ be a convex set. A function $f : C \to \mathbb{R}$ is called *convex*, if

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$$

for all $x, y \in C$ and $\lambda \in [0, 1]$. A function f is called concave, if $-f$ is convex. Moreover, a

function f is called strictly convex, if

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$$

for all $x, y \in C$, $x \neq y$ and $\lambda \in (0,1)$.

Example A.1.3. *Let's consider the following examples.*

1. *Let $\{C_i,\ i \in I\}$ be a family of convex sets. Then $\cap_{i\in I} C_i$ is a convex set.*
2. *Let $C_1, C_2 \subset \mathbb{R}^d$ be two convex sets, then the set*

$$C = \{x \in \mathbb{R}^d \mid x = x_1 + x_2,\ x_1 \in C_1, x_2 \in C_2\}$$

   is a convex set.

3. *The image of a convex set under linear transformation is again a convex set.*
4. *Let $C \subset \mathbb{R}^d$ be a convex set and let $f : C \to \mathbb{R}$ be a convex function. Then the level sets*

$$A_\alpha = \{x \in C \mid f(x) \leq \alpha\} \quad \textit{and} \quad B_\alpha = \{x \in C \mid f(x) < \alpha\}$$

   are convex sets for all $\alpha \in \mathbb{R}$.

Definition A.1.4. Let $f : C \to \mathbb{R}$ be a function and $C \subset \mathbb{R}^d$ be a convex set. We define the *epigraph* of f as

$$\mathcal{E}(f) := \{(x, w) \in C \times \mathbb{R} \mid f(x) \leq w\},$$

the set of all points above the *graph* of f defined as

$$\mathcal{G}(f) := \{(x, w) \in C \times \mathbb{R} \mid f(x) = w\}.$$

We can characterize convex functions via convexity of its epigraph.

Proposition A.1.5 (Fact)**.** Let $f : C \to \mathbb{R}$ be a function and $C \subset \mathbb{R}^d$ be a convex set. The f is convex if and only if its epigraph $\mathcal{E}(f)$ is a convex set.

Next, we describe an inequality for convex functions which will be used often times in this lecture.

Proposition A.1.6 (Jensen's inequality)**.** Let $f : C \to \mathbb{R}$ be a convex function, $C \subset \mathbb{R}^d$ be a

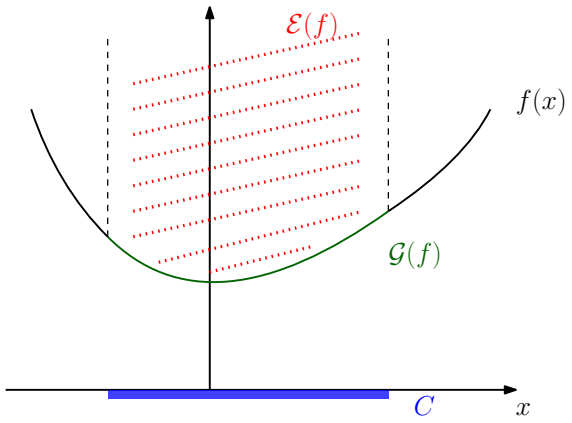

Figure A.2: Illustration of the connection between the graph $\mathcal{G}(f)$, the epigraph $\mathcal{E}(f)$ and the convexity of f.

convex set and $\lambda_1, \dots, \lambda_n \in (0,1)$ with $\sum_{i=1}^n \lambda_i = 1$. Then it holds true that

$$f(\sum_{i=1}^n \lambda_i x_i) \leq \sum_{i=1}^n \lambda_i f(x_i).$$

Exercise A.1.1. Prove Proposition A.1.6.

One typical example for the application of Jensen's inequality is

$$\left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2 \leq \frac{1}{n}\sum_{i=1}^n x_i^2.$$

Example A.1.7. *In the following we present a row of useful properties in the context of convex functions.*

1. *Every linear function is convex.*
2. *Every norm in $\mathbb{R}^d$ is convex.*
3. *Let $f_1, \dots, f_n$ be convex functions and $c_1, \dots, c_n \geq 0$.*
   a) *Then $x \mapsto \sum_{i=1}^n f_i(x)$ is a convex function.*
   b) *Then $x \mapsto \sup_{i=1,\dots,n} f_i(x)$ is a convex function.*

Exercise A.1.2. Prove the properties presented in Example A.1.7.

Proposition A.1.8. Let $C \subset \mathbb{R}^d$ be a convex set and let $f : C \to \mathbb{R}$ be differentiable over C. Then the following holds true:

1. The function f is convex over C if and only if
$$f(z) \geq f(x) + (z-x)^\top \nabla f(x),$$
for all $x, z \in C$.
2. If
$$f(z) > f(x) + (z-x)^\top \nabla f(x), \quad x \neq z,$$
for all $x, z \in C$, then f is strictly convex.

Proof. We only prove the first assertion, the second one follows by similar argumentation. Firstly, assume that f is convex and let $x, z \in C$. Since C is a convex set, $x + a(z-x) \in C$ for all $a \in (0,1)$. The convexity of f implies that

$$f(x + a(z-x)) \leq af(z) + (1-a)f(x),$$

which we can rewrite to
$$\frac{f(x + a(z-x)) - f(x)}{a} \leq f(z) - f(x).$$
On the other side, since f is differentiable, we have

$$(z-a)^\top \nabla f(x) = \lim_{a \searrow 0} \frac{f(x + a(z-x)) - f(x)}{a} \leq f(z) - f(x).$$

This proves the first direction " $\Rightarrow$ ". Secondly, for the direction " $\Leftarrow$ " suppose that

$$f(z) \geq f(x) + (z-x)^\top \nabla f(x) \tag{A.1}$$

for all $x, z \in C$. We fix arbitrary $x, y \in C$ and $\lambda \in (0,1)$, and define $z := \lambda x + (1-\lambda)y$. Then by (A.1) we have both

$$f(x) \geq f(z) + (x-z)^\top \nabla f(z)$$
$$f(y) \geq f(z) + (y-z)^\top \nabla f(z).$$

Combining both inequalities yields

$$\begin{aligned}\lambda f(x) + (1-\lambda) f(y) &\geq \lambda f(z) + \lambda (x-z)^\top \nabla f(z) + (1-\lambda) f(z) + (1-\lambda)(y-z)^\top \nabla f(z) \\ &= f(z) + (\lambda x + (1-\lambda) y - z) \nabla f(z) \\ &= f(z) = f(\lambda x + (1-\lambda) y).\end{aligned}$$

Since $x, y \in C$ are arbitrary, we have proven convexity of f. □

Proposition A.1.9. Let $C \subset \mathbb{R}^d$ be a convex set, $f : \mathbb{R}^d \to \mathbb{R}$ twice continuous differentiable over C and let $Q \in \mathbb{R}^{d \times d}$ be a symmetric matrix.

1. If $\nabla^2 f(x)$ is positive semi-definite for all $x \in C$, then f is convex over C.
2. If $\nabla^2 f(x)$ is positive definite for all $x \in C$, then f is strictly convex over C.
3. If $C = \mathbb{R}^d$ and f is convex, then $\nabla^2 f(x)$ is positive semi-definite for all $x \in C$.
4. The quadratic function $x \mapsto f(x) = x^\top Q x$ is convex if and only if Q is positive semi-definite. Moreover, f is strictly convex if and only if Q is positive definite.

Proof. 1. Let $x, y \in C$, then by the mean-value theorem there exists $\alpha \in [0,1]$ such that

$$\begin{aligned}f(y) &= f(x) + (y-x)^\top \nabla f(x) + \frac{1}{2}(y-x)^\top \nabla^2 f(x + \alpha(y-x))(y-x) \\ &\geq f(x) + (y-x)^\top f(x),\end{aligned}$$

where we have used that $\nabla^2 f(x + \alpha(y-x))$ is positive semi-definite, since $x + \alpha(y-x) \in C$. Then f is convex by Proposition A.1.8.

2. Follows similarly as the proof of the first claim.

3. Let f be convex and suppose that there exists $x \in \mathbb{R}^d$ such that $z^\top \nabla^2 f(x) z < 0$ for some $z \in \mathbb{R}^d$. By assumption we have that $x \mapsto \nabla^2 f(x)$ is continuous, such that for $\bar{\alpha} > 0$ small enough it holds true that

$$z^\top \nabla^2 f(x) z < 0, \quad \text{for } \alpha \in [0, \bar{\alpha}).$$

For an arbitrary $\tilde{\alpha} \in [0, \bar{\alpha})$ we set $\tilde{z} = \tilde{\alpha} z$. By the mean-value theorem there exists $\beta \in [0,1]$

such that

$$\begin{aligned} f(x+\tilde{z}) &= f(x) + \tilde{z}^\top \nabla f(x) + \frac{1}{2}\tilde{z}^\top \nabla^2 f(x+\beta\tilde{z})\tilde{z} \\ &= f(x) + \tilde{z}^\top \nabla f(x) + \frac{1}{2}\tilde{\alpha}^2 z^\top \nabla f(x + \underbrace{\beta\tilde{\alpha}}_{<\tilde{\alpha}} z)z \\ &< f(x) + \tilde{z}^\top \nabla f(x), \end{aligned}$$

which is in contradiction to the convexity of f (see Proposition A.1.8).

4. The hessian of f is given by $\nabla^2 f(x) = 2Q$ for all $x \in \mathbb{R}^d$. With 1. and 3. it follows that f is convex if and only if Q is positive semi-definite. If Q is positive definite, then convexity of f follows by the 3. assertion. It is left to prove that strict convexity of f implies that Q is positive definite. Assume that f is strictly convex, then by 3. we know that Q is positive semi-definite. If a matrix $A \in \mathbb{R}^{d\times d}$ is positive semi-definite but not positive definite, then there exists at least one $x \in \mathbb{R}^d$, $x \neq 0$ such that $x^\top Ax = 0$, i.e. $Ax = 0 = 0 \cdot x$. Hence, to prove that Q is positive definite we will show that $\lambda = 0$ is no eigenvalue of Q. Suppose $\lambda = 0$ is an eigenvalue of Q, then there exists $x \neq 0$ such that $Qx = \lambda x = 0$. This implies that
$$\frac{1}{2}f(x) + \frac{1}{2}f(-x) = \frac{1}{2}(x^\top Qx + (-x)^\top Q(-x) = x^\top Qx = 0 = f(\frac{1}{2}x - \frac{1}{2}x),$$
which is in contradiction to strict convexity of f. Hence, Q is positive definite.

□

We will often consider the class of strong convex functions.

Definition A.1.10 (strongly convex function)**.** Let $C \subset \mathbb{R}^d$ be a convex set. A function $f : C \to \mathbb{R}$ is called *μ-strongly convex* over C, if

$$f(y) \geq f(x) + (y-x)^\top \nabla f(x) + \frac{\mu}{2}\|y-x\|^2$$

for all $x, y \in C$.

Proposition A.1.11. 1. Every μ-strongly convex function f is also strictly convex.

2. A function $f : C \to \mathbb{R}$ is μ-strongly convex if and only if its hessian $\nabla^2 f(x)$ is uniformly positive definite with
$$z^\top \nabla^2 f(x) z \geq \mu\|z\|^2$$
for all $x \in C$ and all $z \in \mathbb{R}^d$.

Exercise A.1.3. Prove Proposition A.1.11.

A.2 Lyapunov methods for optimization

We will motivate the application of Lyapunov theory to optimization methods based on stability analysis of ordinary differential equations (ODEs). Lyapunov theory can be applied to analyze the behavior of dynamical systems without solving them analytically. Let

$$\frac{\mathrm{d}z(t)}{\mathrm{d}t} = g(z(t)), \quad z(0) = z_0 \in \mathbb{R}^d \tag{A.2}$$

be a continuous-time dynamical system described as ODE with $g : \mathbb{R}^d \to \mathbb{R}^d$ locally Lipschitz-continuous.

Definition A.2.1. We call a point $\bar{z} \in \mathbb{R}^d$ *equilibrium point* of (A.2) if $g(\bar{z}) = 0$.

Lyapunov methods can be applied to describe stability of equilibrium points of (A.2). Without loss of generality we assume that $\bar{z} = 0 \in \mathbb{R}^d$ is an equilibrium point of (A.2).

Definition A.2.2. The equilibrium point $\bar{z} = 0$ is called

1. *stable*, if for all $\varepsilon > 0$ there exists a $\delta = \delta(\varepsilon) > 0$ such that:
$$\|z(0)\| < \delta \quad \Longrightarrow \quad \|z(t)\| < \varepsilon \quad \text{for all } t \geq 0.$$
2. *unstable*, if $\bar{z}$ is not stable.
3. *locally asymptotically stable*, if $\bar{z}$ is stable and $\delta > 0$ can be chosen such that:
$$\|z(0)\| < \delta \quad \Longrightarrow \quad \lim_{t\to\infty} z(t) = 0.$$
4. *globally stable*, if $\bar{z}$ is stable and $\lim_{\varepsilon\to\infty} \delta(\varepsilon) = \infty$.
5. *globally asymptotically stable*, if $\lim_{t\to\infty} z(t) = 0$ for all $z_0 \in \mathbb{R}^d$.

We consider a continuously differentiable function $V : \mathbb{R}^d \to \mathbb{R}$ - in the following called candidate of a Lyapunov function - which will be used to be described along the trajectories. With the help of chain rule, we can describe the dynamical behavior of V along the trajectory of $z(t)$:

$$\frac{\mathrm{d}V(z(t))}{\mathrm{d}t} = \langle \nabla_z V(z(t)), \frac{\mathrm{d}z(t)}{\mathrm{d}t} \rangle = \langle \nabla_z V(z(t)), g(z(t)) \rangle.$$

Under specific assumptions on the function V one can verify global stability of $\bar{z}$.

Theorem A.2.3 (Theorem 3.2 in [13]). *Let $\bar{z} = 0$ be an equilibrium point of (A.2). Moreover, let $V : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable with*

1. *$V(\bar{z}) = 0$ and $V(z) > 0$ for all $z \in \mathbb{R}^d \setminus \{\bar{z}\}$,*
2. *$V(z) \to \infty$ for $\|z\| \to \infty$,*
3. *$\frac{\mathrm{d}V(z(t))}{\mathrm{d}t} = \langle \nabla_z V(z(t)), g(z(t)) \rangle \leq -W(z(t))$, for some continuous function $W : \mathbb{R}^d \to \mathbb{R}$.*

If $W(z) \geq 0$ for all $z \in \mathbb{R}^d$, then $\bar{z}$ is globally stable. Moreover, if $W(z) > 0$ for all $z \in \mathbb{R}^d \setminus \{\bar{z}\}$, then $\bar{z}$ is even globally asymptotically stable.

A function satisfying the above conditions is sometimes also referred to a *Lyapunov function*.

We want to apply Lyapunov theory as tool for analyzing convergence of optimization methods. Let us start with an motivating example in continuous-time.

Example A.2.4. *The gradient descent method with fixed step size $\bar{\alpha} > 0$ is written as*

$$x_{k+1} = x_k - \bar{\alpha} \nabla f(x_k),$$

which for $\bar{\alpha} \to 0$ can be interpreted as Euler-scheme of the ODE

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -\nabla f(x(t)). \tag{A.3}$$

In general, optimization schemes are often described and/or analysed through its continuous-time formulation. The system is sometimes also called gradient flow and intuitively speaking, it describes gradient descent with degenerated step size. It can be used as indicator of how gradient descent may perform with sufficiently small step size $\bar{\alpha} > 0$. We have analysed gradient descent methods under various settings such as (strong) convexity and smoothness. The convergence analysis of the gradient flow (A.3) can be done in a very similar way using Lyapunov theory. We want to construct a function describing the convergence behavior of the dynamical system (A.3) without solving it explicitly. The most straightforward analysis can be done for the error function $V(x(t)) = \mathcal{E}(x(t)) = f(x(t)) - f_*$, where $f_* = \min_{x \in \mathbb{R}^d} f(x) > -\infty$. Similar as before, we can describe the dynamical behavior of V through the ODE

$$\frac{\mathrm{d}V(x(t))}{\mathrm{d}t} = \langle \nabla_x V(x(t)), \frac{\mathrm{d}x(t)}{\mathrm{d}t} \rangle = -\langle \nabla f(x(t)), \nabla f(x(t)) \rangle = -\|\nabla f(x(t))\|^2.$$

Under suitable conditions on f one can now apply Theorem A.2.3 in order to quantify stability of a (unique) stationary point $x_* \in \mathbb{R}^d$ of f as equilibrium point of the gradient flow (A.3).

Let us consider a optimization scheme in continuous-time of the form

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = g(x(t)), \tag{A.4}$$

and a corresponding error function $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}$ to be analyzed. We aim to quantify the convergence of (A.4) through the error function along the trajectory

$$\frac{\mathrm{d}\mathcal{E}(x(t))}{\mathrm{d}t} = \langle \nabla_x \mathcal{E}(x(t)), g(x(t)) \rangle \begin{cases} \leq 0 \\ < 0 \end{cases} .$$

This can be used to obtain results such as monotonically decreasing error (≤ 0) or even convergence of the error (< 0). However, these properties are not sufficient to describe the speed of convergence. In order to say something about the convergence speed, we can define a time-dependent error function $\widehat{\mathcal{E}} : [0, \infty) \times \mathbb{R}^d \to \mathbb{R}$ of the form

$$\widehat{\mathcal{E}}(t, x) = \gamma(t)\mathcal{E}(x),$$

where $\gamma : [0, \infty) \to \mathbb{R}_+$ (smooth and continuously differentiable) is devoted to describe the speed of convergence. Let us illustrate the strategy of proving convergence with the help of this error function throught the following example.

Example A.2.5. *We want to minimize f using (A.4) and prove convergence of the error function $\mathcal{E}(x) = f(x) - f_*$. One possible strategy is to construct the time dependent error function*

$$\widehat{\mathcal{E}}(t, x(t)) = \gamma(t)\mathcal{E}(x(t)) + r(x(t)),$$

where $\gamma : [0, \infty)$ with $\frac{\mathrm{d}\gamma(t)}{\mathrm{d}t} > 0$ describes our guess of convergence rate and $r : \mathbb{R}^d \to \mathbb{R}$ is an auxiliary function with $r(x) \geq 0$. Suppose we are able to prove monotonicity of the error in form of

$$\frac{\mathrm{d}\widehat{\mathcal{E}}(t, x(t))}{\mathrm{d}t} \leq 0$$

(using again chain rule), we can directly imply by definition of $\widehat{\mathcal{E}}$ that

$$\widehat{\mathcal{E}}(t, x(t)) = \gamma(t)(f(x(t)) - f_*) + r(x(t)) \leq \widehat{\mathcal{E}}(0, x(0)),$$

and therefore, we can imply convergence of the error $\mathcal{E}$ in form of

$$\mathcal{E}(x(t)) = f(x(t)) - f_* \leq \frac{\widehat{\mathcal{E}}(0, x(0))}{\gamma(t)}.$$

A similar strategy for discrete-time optimization schemes is described in Section 3.3.1.

A.3 Measure theoretical background

In the following section, we will briefly recall the definition of Dynkin systems and σ-algebras, and a useful tool which allows to prove measure theoretical properties on a $\cap$-stable generator of a σ-algebra. For example, if we want to prove that two measures μ_1 and μ_2 on the same measurable space $(\Omega, \mathcal{A})$ are equal, it is sufficient to verify the equality on an $\cap$-stable generator $\mathcal{E}$ of $\mathcal{A}$ (i.e. $\sigma(\mathcal{E}) = \mathcal{A}$). For more details we refer to [14, Section 1].

Definition A.3.1. Let $\mathcal{A} \subset \mathcal{P}(\Omega)$ be a non-empty family of subsets of Ω. We call $\mathcal{A}$ a *σ-algebra* if

1. $\Omega \in \mathcal{A}$,
2. $A \in \mathcal{A} \implies A^{\complement} \in \mathcal{A}$,
3. $A_1, A_2, \dots \in \mathcal{A} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$

Definition A.3.2. Let $\mathcal{D} \subset \mathcal{P}(\Omega)$ be a non-empty family of subsets of Ω. We call $\mathcal{D}$ a *Dynkin system* if

1. $\Omega \in \mathcal{D}$,
2. $A \in \mathcal{D} \implies A^{\complement} \in \mathcal{D}$,
3. $A_1, A_2, \dots \in \mathcal{D}$ pairwise disjoint (i.e. $A_i \cap A_j = \emptyset$, $i \neq j$) $\implies \biguplus_{i=1}^{\infty} A_i \in \mathcal{D}$

Proposition A.3.3. Let $\mathcal{D}$ be a Dynkin system, then the following is equivalent:

- $\mathcal{D}$ is a σ-algebra,
- $\mathcal{D}$ is $\cap$-stable, i.e. for $A, B \in \mathcal{D}$ it follows $A \cap B \in \mathcal{D}$.

Definition A.3.4. Let $\mathcal{E} \subset \mathcal{P}(\Omega)$ be a non-empty family of subsets of Ω. Then we define

$$\sigma(\mathcal{E}) = \bigcap_{\mathcal{E} \subset B,\ B\ \sigma-\text{algebra}} B$$

as the σ-Algebra generated by $\mathcal{E}$. Similarly we define

$$d(\mathcal{E}) = \bigcap_{\mathcal{E}\subset B,\ B \text{ Dynkin system}} B$$

as the Dynkin system generated by $\mathcal{E}$.

Theorem A.3.5. *Let $\mathcal{E} \subset \mathcal{P}(\Omega)$ be a non-empty family of subsets of Ω which is $\cap$-stable, then $d(\mathcal{E}) = \sigma(\mathcal{E})$.*

Remark A.3.6. In order to prove that some condition $\oplus$ is satisfied for all $A \in \mathcal{A}$, where $\mathcal{A}$ is a σ-algebra over Ω, we can follow a certain strategy:

1. define the set $\mathcal{M} = \{A \in \mathcal{A} \mid \text{condition } \oplus \text{ is satisfied for } A\} \subset \mathcal{A}$,
2. prove that $\mathcal{M}$ is a Dynkin system,
3. find a $\cap$-stable generator $\mathcal{E}$ of $\mathcal{A}$ such that $\mathcal{E} \subset \mathcal{M}$,
4. imply that
$$\mathcal{A} = \sigma(\mathcal{E}) = d(\mathcal{E}) \overset{\mathcal{E}\subset\mathcal{M}}{\subset} d(\mathcal{M}) \overset{\mathcal{M} \text{ Dynkin system}}{=} \mathcal{M} \subset \mathcal{A},$$
which yields that $\mathcal{M} = \mathcal{A}$, i.e. condition $\oplus$ is satisfied for all $A \in \mathcal{A}$.

A.4 Martingales

In this section, we briefly recall Doob's martingale convergence theorem, in particular for supermartingales. We refer to [14, Section 9–11] for more details.

Definition A.4.1. Let $(\Omega, \mathcal{A}, \mathcal{F}, \mathbb{P})$ be a filtered probability space. We call stochastic process $X = (X_k)_{k\in\mathbb{N}}$ *martingale* with respect to the filtration $\mathcal{F}$, if

1. X is $\mathcal{F}$-adapted,
2. $\mathbb{E}[|X_k|] < \infty$ for all $k \in \mathbb{N}$,
3. $\mathbb{E}[X_k \mid \mathcal{F}_l] = X_l$ for all $k, l \in \mathbb{N}$ with $l \leq k$.

If 3. holds with $\leq$ ($\geq$), then we call X a *supermartingale* (*submartingale*).

Theorem A.4.2 (Doob's supermartingale convergence)**.** *Let $X = (X_k)_{k\in\mathbb{N}}$ be a supermartingale or submartingale with*

$$\sup_{k\in\mathbb{N}} \mathbb{E}[|X_k|] < \infty,$$

then $(X_k)_{k\in\mathbb{N}}$ converges almost surely to an $\mathcal{F}_\infty$-measurable and integrable random variable X_∞.

Remark A.4.3. For a martingale we have the property of constant expectation $\mathbb{E}[X_k] = \mathbb{E}[X_0]$ for all $k \in \mathbb{N}$. For a supermartingale in contrast, we have decreasing expectation $\mathbb{E}[X_k] \leq \mathbb{E}[X_l] \leq \mathbb{E}[X_0]$ for $k \geq l$. Therefore, it is sufficient to replace the condition $\sup_{k\in\mathbb{N}} \mathbb{E}[|X_k|] < \infty$ of Theorem A.4.2 with the expectation of the negative part $X_k^- = \max(0, -X_k)$, since we have

$$\mathbb{E}[|X_k|] = \mathbb{E}[X_k^+ + X_k^-] = \mathbb{E}[X_k^+ - X_k^- + X_k^- + X_k^-] = \mathbb{E}[X_k] + 2\mathbb{E}[X_k^-] \leq \mathbb{E}[X_0] + 2\mathbb{E}[X_k^-].$$

In case that X_k is bounded from below it follows that $\mathbb{E}[X_k^-] < \infty$. This means it is sufficient to prove a uniform lower bound on $(X_k)_{k\in\mathbb{N}}$ in order to apply Theorem A.4.2.