Convergence and Stability Analysis of Stochastic Optimization Algorithms

Williamson, Måns

2025

– CENTRUM SCIENTIARUM MATHEMATICARUM –

Convergence and Stability Analysis of Stochastic Optimization Algorithms

MÅNS WILLIAMSON

Lund University
Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

Printed by Media-Tryck, Lund 2025 NORDIC SWAN ECOLABEL 3041 0903

Doctoral Theses in Mathematical Sciences 2025:???????
ISBN 978-91-8104-331-0
ISSN 1404-0034

Convergence and Stability Analysis of Stochastic Optimization Algorithms

Convergence and Stability Analysis of Stochastic Optimization Algorithms

by Måns Williamson

LUND
UNIVERSITY

Thesis for the degree of Doctorate
Thesis advisors: Dr. T. Stillfjord, Dr. M. Eisenmann, Prof. E. Hansen
Faculty opponent: Prof. K. C. Zygalakis

To be presented, with the permission of the Faculty of Engineering of Lund University, for public defense at Matematikcentrum, MH:Hörmander on Friday, the 31st of January 2025 at 13:00.

Organization
LUND UNIVERSITY

Centre for Mathematical Sciences
Box 118
SE–221 00 LUND
Sweden

Document name
DOCTORAL DISSERTATION

Date of disputation
2025-01-31

Sponsoring organization
Wallenberg AI, Autonomous Systems and Software Program

Author(s)
Måns Williamson

Title and subtitle
Convergence and Stability Analysis of Stochastic Optimization Algorithms

Abstract

This thesis is concerned with stochastic optimization methods. The pioneering work in the field is the article "A stochastic approximation algorithm" by Robbins and Monro (1951), in which they proposed the *stochastic gradient descent*; a stochastic version of the classical gradient descent algorithm. Since then, many improvements and extensions of the theory have been published, as well as new versions of the original algorithm. Despite this, a problem that many stochastic algorithms still share, is the sensitivity to the choice of the step size/learning rate. One can view the stochastic gradient descent algorithm as a stochastic version of the *explicit Euler scheme* applied to the corresponding gradient flow equation. There are other schemes for solving differential equations numerically that allow for larger step sizes. In this thesis, we investigate the properties of some of these methods, and how they perform, when applied to stochastic optimization problems.

Key words
stochastic optimization, optimization, machine learning, numerical analysis

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title
1404-0034

ISBN
978-91-8104-331-0 (print)
978-91-8104-332-7 (pdf)

Recipient's notes

Number of pages
209

Price

Security classification

Convergence and Stability Analysis of Stochastic Optimization Algorithms

by Måns Williamson

LUND
UNIVERSITY

Faculty of Engineering, Centre for Mathematical Sciences

ISBN: 978-91-8104-331-0 (print)
ISBN: 978-91-8104-332-7 (pdf)

Printed in Sweden by Media-Tryck, Lund University, Lund 2025

Media-Tryck is a Nordic Swan Ecolabel
certified provider of printed material.
Read more about our environmental
work at www.mediatryck.lu.se

MADE IN SWEDEN

Contents

List of publications . . . iii
Acknowledgements . . . iv
Popular summary in English . . . v
Populärvetenskaplig sammanfattning på svenska . . . vii

1 Introduction **1**

2 Supervised learning and risk minimization **5**
2.1 Supervised learning . . . 5
2.2 Empirical risk minimization . . . 6
2.3 Generalization error . . . 7

3 Time integration **9**
3.1 Explicit Euler . . . 9
3.2 Implicit Euler . . . 11
3.3 Runge–Kutta methods . . . 12
3.4 Stability . . . 13

4 Optimization **17**
4.1 Gradient descent . . . 17
4.2 Proximal point method . . . 20
4.3 Stochastic gradient descent . . . 20

5 ODE method **27**
5.1 ODE method . . . 28
5.2 The Robbins–Siegmund theorem . . . 35
5.3 Asymptotic pseudo-trajectories and chain recurrence . . . 36

6 Research **45**
6.1 Paper I . . . 46
6.2 Paper II . . . 47
6.3 Paper III . . . 48
6.4 Paper IV . . . 50
6.5 Outlook . . . 53

References **55**

Scientific publications **63**
Author contributions 63
Paper I: Sub-linear convergence of a stochastic proximal iteration method in Hilbert space 65
Paper II: SRKCD: A stabilized Runge–Kutta method for stochastic optimization 97
Paper III: Analysis of a class of stochastic component-wise soft-clipping schemes 117
Paper IV: Almost sure convergence of stochastic Hamiltonian descent methods 148

List of publications

This thesis is based on the following publications, referred to by their Roman numerals:

I **Sub-linear convergence of a stochastic proximal iteration method in Hilbert space**
M. Eisenmann, T. Stillfjord, **M. Williamson**
Computational Optimization and Applications, 83(1) (2022), p. 181-210

II **SRKCD: A stabilized Runge–Kutta method for stochastic optimization**
T. Stillfjord, **M. Williamson**
Journal of Computational and Applied Mathematics, 417(2023), 114575

III **Analysis of a class of stochastic component-wise soft-clipping schemes**
M. Eisenmann, T. Stillfjord, **M. Williamson**
ArXiv Preprint, arXiv:2406.16640, 2024

IV **Almost sure convergence of stochastic Hamiltonian descent methods**
T. Stillfjord, **M. Williamson**
ArXiv Preprint: arXiv:2406.16649, 2024

Acknowledgements

This thesis would not have been possible without the funding from WASP, Wallenberg AI, Autonomous Systems and Software Program. I would also like to express my deep gratitude to my supervisors for patiently reading this thesis over and over again and coming with valuable feedback. A huge thanks to Tony who has been my main supervisor for these 5 years. It's been a pleasure to pursue a PhD under your guidance. I would also like to thank my family for their support.

Popular summary in English

A statistical model is a type of function that is fitted to a dataset in order to make predictions about the future. An example of such a function could be a program that, given an X-ray image, predicts whether a patient has cancer or not. Another example is a computer program that generates text. Given a new word, the next word in the sequence is determined by the program predicting which word is most likely. This is possible because the program has previously been exposed to a dataset consisting of large amounts of text.

In 1958, Rosenblatt published his paper 'The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain', in which he proposed and analyzed the so-called perceptron. This is a statistical model inspired by how the human brain processes information and can be considered the ancestor of today's artificial neural networks, which have become very large in artificial intelligence.

When fitting a statistical model to a dataset, this is often done by introducing a so-called cost function. This is a function that applies a penalty if the model's prediction is far from the true value. One then tries to find model parameters that result in the smallest penalty possible. In artificial intelligence, this process is referred to as 'training' the model. An analogy for this procedure could be descending a mountainous landscape. At each stage, you take a step downward. The penalty we receive from the cost function is our position's altitude, and the goal is to find the lowest point in a valley. The size of the step we take is called the step size or learning rate. It is common in artificial intelligence for cost functions to be very complicated and computationally intensive to minimize. The landscape in the analogy above is very rocky, and it is very difficult to take each step. In such cases, traditional optimization methods are less suitable. Instead, one uses methods for which each step is cheaper to take - so-called stochastic optimization methods. The most famous of these is stochastic gradient descent, whose precursor was introduced in 1951 by Robbins & Monro in the paper 'A Stochastic Approximation Method'. To simplify, these methods can be described as follows: instead of always taking a step downward, one moves somewhat randomly, but in such a way that, on average, one moves downward. A problem with these methods is that the size of the steps must be chosen carefully. If they are too large, one may move farther and farther from the desired valley. If they are too small, it might take an eternity to reach the lowest point. The focus of this thesis is exploring different ways to make stochastic optimization methods more robust and less sensitive to the step size. This is important from a sustainability perspective, as artificial intelligence is

becoming an increasingly large part of our daily lives and is claiming more and more of our total energy consumption. More robust methods ensure that less computational power and energy are required to train the models.

Populärvetenskaplig sammanfattning på svenska

En statistisk modell är en slags funktion som man anpassar till en datamängd för att kunna göra förutsägelser om framtiden. Ett exempel på en sådan funktion skulle kunna vara ett program som givet en röntgenbild förutsäger om en patient har cancer eller inte. Ett annat är ett datorprogram som genererar text. Givet ett nytt ord, bestäms nästa ord i ordföljden genom att programmet förutspår vilket ord som är mest troligt. Detta är möjligt tack vare att programmet tidigare har fått se en datamängd som består av stora mängder text.

1958 publicerade Rosenblatt sin artikel "The perceptron: a probabilistic model for information storage and organization in the brain", i vilken han föreslog och analyserar den så kallade perceptronen. Detta är en statistisk modell som inspirerats av hur den mänskliga hjärnan hanterar information och kan sägas vara förfadern till de artificiella neuronnätverk som har kommit att bli väldigt stora inom artificiell intelligens.

När man anpassar en statistisk modell till en datamängd gör man ofta det genom att introducera en så kallad kostnadsfunktion. Detta är en funktion som ger ett "straff" om modellens förutsägning ligger långt ifrån det sanna värdet. Man försöker sedan hitta parametrar till modellen som ger ett så litet straff som möjligt. Inom artificiell intelligens kallas detta för att man "tränar" modellen. En analogi för den här proceduren skulle kunna vara en nedstigning genom ett bergigt landskap. Vid varje etapp tar man ett steg nedåt. Bestraffningen vi får från kostnadsfunktionen är vår positions altitud och målet är att hitta den lägsta punkten i en dal. Storleken på steget vi tar kallas för steglängd eller inlärningshastighet (från eng. learning rate). Det är vanligt inom artificiell intelligens att kostnadsfuntionerna är väldigt komplicerade och beräkningsmässigt krävande att minimera. Landskapet i analogin ovan är väldigt klippigt och det är väldigt jobbigt att ta varje enskilt steg. Då är traditionella metoder för att utföra mindre lämpliga att använda. I stället använder man metoder där varje steg är "billigare" att ta - så kallade *stokastiska optimeringsmetoder*. Den kanske mest kända av dessa är *stochastic gradient descent*, vars förlaga introducerades 1951 av Robbins & Monro med artikeln "A stochastic approximation method". Förenklat skulle dessa metoder kunna beskrivas på följande vis: i stället för att alltid ta ett steg som tar en nedåt, går man lite på måfå, men på ett sådant sätt att man i genomsnitt rör sig nedåt. Ett problem med dessa metoder är att man måste välja stegen man tar på ett lämpligt sätt. Väljs de för stora kanske man rör sig längre och längre ifrån den eftersökta dalen. Väljs de för små kan de ta en evighet att komma till den lägsta punkten. Fokuset i den här avhandlingen är att undersöka olika sätt att göra stokastiska optimeringsmetoder mer

robusta och mindre känsliga för steglängden som tas. Detta är viktigt ur ett hållbarhetsperspektiv då artificiell intelligens är på väg att bli en större och större del av vår vardag och gör anspråk på mer och mer av vår totala energiåtgång. Mer robusta metoder gör att det går åt mindre beräkningskapacitet och energi för att träna modellerna.

1. Introduction

For a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$, we consider the problem of solving

$$w_* \in \arg\min_{w \in \mathbb{R}^d} F(w). \tag{1.1}$$

Such problems are frequently encountered in the field of machine learning when one seeks to estimate the parameters of a statistical model. A classical approach for iteratively approximating a solution to (1.1) is to make use of the *gradient descent* method:

$$w_{k+1} = w_k - \alpha_k \nabla F(w_k), \tag{1.2}$$

where α_k is the *step size* or *learning rate*. Gradient descent is an example of a deterministic optimization algorithm. In machine learning, the objective function F frequently takes the form

$$F(w) = \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, h(x_i, w)), \tag{1.3}$$

where $\{x_i, y_i\}_{i=1}^N$ is a dataset of feature-label pairs and $h(\cdot, w)$ is a model with model parameters w, e.g. a regression- or a classification model. It is common that the size of the dataset N is very large and in this case it may become very computationally expensive to compute the gradient in (1.2). A cheaper alternative is to make use of *stochastic optimization methods*. This is what the research presented in this thesis is concerned with. The most classical example is the *stochastic gradient descent* method (SGD):

$$w_{k+1} = w_k - \alpha_k \nabla f(w_k, \xi_k),$$

which is a randomized version of (1.2). Here, $\nabla f(\xi_k, w_k)$ is a stochastic approximation to the gradient $\nabla F(w_k)$. A typical choice is to take

$$\nabla f(w, \xi_k) = \frac{1}{|B_{\xi_k}|} \sum_{i \in B_{\xi_k}} \nabla \ell(y_i, h(x_i, w)), \tag{1.4}$$

where $B_{\xi_k} \subset \{1, \dots, N\}$ is a *mini-batch*, chosen such that $|B_{\xi_k}| \ll N$. Using SGD has been shown to have several advantages; it is less computationally costly compared to the traditional algorithms such as gradient descent or Newton's method; another advantage is that the randomness allows the iterates to escape local saddle points in the non-convex case, see [7, 21, 26]. The latter is an important property, as many machine learning problems are indeed non-convex. Perhaps most notable are deep neural networks, for which evidence suggest that saddle points at which the value of the cost function is high, appear more frequently than shallow local minima, [33]. Yet another benefit is the following: the objective function used in machine learning problems is typically based on the sample data set. In practice, the latter often contains data that is similar and does not add much information to the gradient update. Here, stochastic algorithms that only make use of a subset of the data tend to use information more efficiently, see e.g. Section 3.3 in [14] or 8.1.3 in [33].

Despite the advantages of SGD, the step size α_k often needs to be carefully tuned; if it is chosen too small, it can take a long time before an acceptable value of the objective function is reached; if it is chosen too large, the method may blow up. Here, the need for stabilized methods that are less sensitive to the choice of step size enters the picture. In the field of numerical analysis for differential equations, methods that allow for larger step sizes have long been used. Let $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ be a continuous function and consider the initial value problem

$$\begin{aligned} y' &= f(t, y), \\ y(t_0) &= y_0, \ t_0 \in \mathbb{R}. \end{aligned} \tag{1.5}$$

The *explicit Euler method* is an iterative method for approximating the solution to (1.5) on an interval $[t_0, T]$:

$$y_{k+1} = y_k + hf(t_k, y_k),$$

for $k = 0, \dots, n$. Here $n \in \mathbb{N}$, $t_k = t_0 + hk$ and $h = \frac{T - t_0}{n}$ is the step size. SGD can be viewed as a stochastic explicit Euler discretization of the gradient flow equation

$$\begin{aligned} w' &= -\nabla F(w), \\ w(0) &= w_0. \end{aligned} \tag{1.6}$$

When solving an initial value problem we want to find the solution over a certain time period, while in the optimization case we want to solve it over an infinite time interval to find an equilibrium solution $w(t) = w_*$, which by definition

satisfies

$$w'(t) = 0.$$

By (1.6), an equilibrium solution to the gradient flow is also a stationary point of F. There is a severe step size restriction on the explicit Euler scheme, and thus it can be good to use other schemes with larger stability regions. An example of such a scheme are the *Runge–Kutta–Chebyshev* methods that we analyze in Paper II of this thesis.

In Paper I and II of this thesis, we investigate how methods with good stability properties that have proven to work well for solving differential equations numerically, work when they are applied in the context of stochastic optimization problems. The central theme is how the concept of stability of numerical methods for ODEs translates to the stochastic optimization setting. In Paper III and IV, we deviate from this perspective and investigate so-called clipping algorithms. These are methods that make use of gradient information to rescale the vector field or step size to ensure stability. In Paper IV we also consider *momentum* algorithms. A popular version of SGD is *SGD with momentum*:

$$\begin{aligned} p_{k+1} &= \beta p_k - \alpha \nabla f(w_k, \xi_k), \\ q_{k+1} &= q_k + p_{k+1}, \end{aligned} \tag{1.7}$$

which makes use of a weighted average of all the past gradients: By an iterative argument, it holds that

$$p_{k+1} = -\sum_{i=0}^{k} \left(\prod_{j=i+1}^{k} \beta_j \right) \alpha_i \nabla f(q_i, \xi_i).$$

The algorithm determined by (1.7) can be viewed as a discretization of ODE

$$p'(t) = \beta p(t) - \nabla F(w(t)) \tag{1.8}$$

$$q'(t) = p(t). \tag{1.9}$$

Similar to (1.6), the equilibrium solutions of (1.8) are also the stationary points of the objective function F. In Paper IV, we explore how other ODEs (more precisely dissipative Hamiltonian systems) can be used to study clipped momentum algorithms.

The thesis is arranged as follows; Chapters 2 to 5 are intended to serve as an introduction to the concepts encountered in Papers I– IV. The second chapter gives an overview of supervised learning and risk minimization in general. Although the research presented in the papers of the thesis is not mainly concerned

with this, it is important to have an understanding of the underlying problems that the presented optimization algorithms aim to solve.

Chapter 3 gives a brief introduction to time stepping methods and stability of numerical methods. First, we look at the explicit Euler method and discuss its properties. Next, we discuss the implicit Euler method, which was the inspiration for Paper I. The concept of stability of a numerical method is one of the core concepts of the thesis. In connection with this, we also give a short introduction to Runge–Kutta–Chebyshev methods, with which the second paper in the thesis is concerned. In Chapter 4, we go through some of the most common optimization methods and explain what their advantages and disadvantages are. This thesis places a strong emphasis on optimization for non-convex functions, and therefore we start with discussing what can be said in the deterministic case. We also discuss stochastic optimization methods, and mention some of the most common types of results that one encounters and their proof strategies. Chapter 5 is dedicated to the ODE-method, on which the analysis in Paper IV is based. This is an approach for proving almost sure convergence of certain stochastic algorithms that can be viewed as discretizations of ODEs. The analysis is more involved than in the previous chapters, but it also allows us to demonstrate that the entire sequence of iterates generated by the algorithm converges almost surely to a stationary point in the non-convex case. In Chapter 6, we summarize the research of the project and its conclusions, and consider some possible paths for future research.

2. Supervised learning and risk minimization

One of the main applications of stochastic optimization methods is to minimize an objective function F that takes the form of a sum

$$F(w) = \frac{1}{N} \sum_{i=1}^{N} F_i(w),$$

in order to estimate a statistical parameter w. The hope is that the objective function is a good approximation of an expectation that one does not have at hand. In this chapter, we discuss when this is the case, and under what conditions. Although the research presented in this thesis is not mainly concerned with this topic, it plays an important role in the theory of machine learning problems, and is important for understanding the problems that the thesis is concerned with.

2.1 Supervised learning

In a supervised learning problem, we have some measurements $\{(x_i, y_i)\}_{i=1}^{n}$, where x_i are called *features* and y_i *labels*. The task is to estimate the label y for new inputs x by finding a prediction function h such that $h(x)$ is not too far from y for any feature-label pair (x, y) that could be produced. The precise meaning of "not too far" will be made clear later on. In image classification, each x_i could correspond to an image and the y_i to the class of that particular image. In linear regression, x_i would be the independent variable and y_i the dependent variable. Regardless of what the underlying problem is or from where the data emanates, before we set out and gather the data for the experiment, we do not know what the actual value of either the features or the corresponding labels will be.

Thus, it is not unreasonable to think of the feature-label pairs as independent, identically distributed random vectors $\{(X_i, Y_i)\}_{i=1}^N$, defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where X takes values in the feature space and Y in the space of all labels. In the field of statistical inference, one would refer to $\{(X_i, Y_i)\}_{i=1}^N$ as a *random sample* (compare with [16, Definition 5.1.1]). If we for example were to classify the images of the famous MNIST dataset [45], where each feature is a 28×28-pixel image of a handwritten digit between 0 and 9, we could

$$X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}^{28\times 28},$$
$$Y : (\Omega, \mathcal{A}, \mathbb{P}) \to \{0, \dots, 9\}.$$

2.2 Empirical risk minimization

The question now is how to determine a good prediction function h. Suppose that we have some class of measurable functions $\mathcal{H} = \{h(\cdot, w)\}_{w\in\Theta}$, that we restrict ourselves to. Here w is a parameter and Θ the parameter space. The set of functions $\mathcal{H}$ could for example be all functions of the form $h(x, w) = ax+b$, with $w = (a, b)$, or all convolutional neural networks with a certain structure. Our question in the following will be, how do we know if a certain function $h(\cdot, w)$ from the chosen class $\mathcal{H}$ is a good candidate. The common way to measure this is to introduce a loss function ℓ that gives us a penalty if $h(x, w)$ is not equal to the true value of y - the farther away, the larger the penalty. In a linear regression problem we would for example use the square loss function $\ell(y, h(x, w)) = (y - h(x, w))^2$, where $h(x, w) = ax + b$ as above. We then seek to minimize the risk functional

$$R(w) = \int_\Omega \ell\left(h(X(\omega), w), Y(\omega)\right) \mathbb{P}(d\omega). \tag{2.1}$$

Rather than working with the integral in the abstract probability space, it is often more convenient to work with the measure $\mathbb{P}_{(X,Y)}$ induced by $\mathbb{P}$ in the feature-label space, i.e. $\mathbb{P}_{(X,Y)}(A) = \mathbb{P}\left(\{\omega : (X(\omega), Y(\omega)) \in A\}\right)$, where $A \in \mathcal{B}\left(\mathbb{R}^d\right)$, the Borel $\sigma-$algebra on $\mathbb{R}^d$, compare [41, p. 10]. This allows us to talk about the joint-, marginal- and conditional distributions of (X, Y). In the MNIST example, (2.1) would become

$$R(w) = \int_{\mathbb{R}^{28\times 28}\times\{0,\dots,9\}} \ell\left(h(x, w), y\right) \mathbb{P}_{X,Y}(dx \times dy).$$

where $\mathbb{P}_{X,Y}$ is the joint distribution of (X, Y) defined by

$$\mathbb{P}_{X,Y}\left(A\right) = \mathbb{P}\left(\{\omega : (X(\omega), Y(\omega)) \in A\}\right).$$

In many cases the conditional distribution of Y given X can be modelled as deterministic. Using the MNIST dataset as an example again, it is natural to put $\mathbb{P}(Y = 3) = 1$ given that X is an image of a 3, and so on. The rationale for this procedure is that choosing a function $h(\cdot, w) \in \mathcal{H}$ that gives a low value for the risk functional will give us a low loss $\ell(h(x, w), y)$ on average. The problem is that in most cases, the joint distribution $\mathbb{P}_{X,Y}$ is unknown to us. We can however obtain a random sample $\{(X_i, Y_i)\}_{i=1}^N$ and hence what we can minimize is the *empirical risk functional*

$$R_N(w) = \frac{1}{N} \sum_{i=1}^{N} \ell\left(h(X_i, w), Y_i\right). \tag{2.2}$$

Minimizing (2.2) rather than (2.1) is sometimes referred to as *the principle of empirical risk minimization*, see [64, p. 32]. We note that the minimizer w_* of (2.2) is an *estimator*, i.e. a function of the random sample $\{(X_i, Y_i)\}_{i=1}^N$ (compare with [16, Definition 7.1.1] and the discussion that follows). In general, w_* could be non-measurable and/or set-valued. In this discussion, we for simplicity assume that it is a random variable, i.e. single-valued and measurable. There are several ways to deal with non-measurability (see for example [66, 4.4] and the discussion on the outer expectation in [63]) and set-valued random variables (see [1, 14.91]), but this is outside the scope of this thesis.

2.3 Generalization error

An important concept in machine learning is that of *generalization*. Assume that there is w_0 in the parameter space Θ such that $R(w_0) = \inf_{w \in \Theta} R(w)$ and let w_* be a minimizer of (2.2). Suppose that we want to estimate w_0 by finding w_*. Can we guarantee that $R(w_*)$ will get closer to $R(w_0)$ in some sense -either in probability or almost surely- if we increase the number of samples?

Closely following [66], the difference $R(w_*) - R(w_0)$ can be split up as follows

$$\begin{aligned} R(w_*) - R(w_0) = &\underbrace{R(w_*) - R_N(w_*)}_{T_1} + \underbrace{R_N(w_*) - R_N(w_0)}_{T_2} + \\ &\underbrace{R_N(w_0) - R(w_0)}_{T_3}. \end{aligned}$$

The second term T_2 is less than or equal to 0 since w_* is a minimizer of $R_N(w)$.

According to the *law of large numbers* we have for a fixed w that

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \ell\left(h(X_i, w), Y_i\right) = \int_{\Omega} \ell\left(h(X(\omega), w)\right), Y(\omega)\right) d\mathbb{P}(\omega),$$

in probability or almost surely (depending on wether we use the strong- or weak law of large numbers). The parameter w_0, being the minimizer of $R(w)$, is independent of the random sample and is thus a deterministic quantity. Thus we can conclude that T_3 converges to 0. (Again, at this point we keep the discussion general, so we are not specifying the mode of convergence).

We now turn our attention to the first term T_1. The problem is that w_* is not fixed as it depends on the random variables $\{(X_i, Y_i)\}_{i=1}^{N}$. Therefore, we need a uniform bound on the difference $R(w) - R_N(w)$ so that we can guarantee beforehand that the difference will not be too large, independent of what sample we get and what distribution they have. The common approach to ensure this is to restrict the functions that we consider to various function classes for which uniform convergence holds, see [64]. Under certain conditions on the class $\mathcal{H}$ it is for example possible to say that R_N converges uniformly, almost surely, to R, i.e.

$$\mathbb{P}\Big(\Big\{\omega : \lim_{N\to\infty} \sup_{w\in\Theta} |R(w) - R_N(w)| \neq 0\Big\}\Big) = 0,$$

compare [64, Thm. 3.5]. In machine learning, one often considers classes of prediction functions $\mathcal{H}$ with finite *VC-dimension*. In the expression (2.2), we could for example have $\mathcal{H} = \{h(\cdot, w)\}_{w\in\mathbb{R}^d}$. Intuitively, one can say that these are classes of functions that do not overfit the data. Suppose that the functions are also bounded, in the sense that there are constants A and B such that $A \leq h(x) \leq B$ holds for all $h \in \mathcal{H}$. Then it holds for a class $\mathcal{H}$ of finite VC-dimension v that

$$\mathbb{P}\Big(\Big\{\omega : \sup_{w\in\Theta} |R(w) - R_N(w)| > \epsilon\Big\}\Big) \leq 4\exp\left\{N\left(\frac{v\left(\ln(\frac{2N}{v}) + 1\right)}{N} - \frac{\epsilon^2}{B - A}\right)\right\}, \tag{2.3}$$

when $N > \frac{v}{2}$, compare (3.10) and Thm. 3.3 in [65]. That is, R_N converges uniformly in probability to R. If we look at how the constant on the right-hand side of (2.3) behaves, we see that for a fixed VC-dimension v, we have uniform convergence in w as the number of samples N is increased. We also see that for a fixed number of samples, N, the gap between $R(w)$ and $R_N(w)$ can increase if we use a function class with larger VC-dimension v. For other classes of functions, such as the set of unbounded, non-negative functions, there are similar bounds on the gap between the empirical risk and the risk functional, compare [65, Ch. 3.7], but this is not the focus of this thesis.

3. Time integration

The optimization methods that are proposed in the papers of this thesis are based on numerical schemes for time-integration. In this chapter we therefore give a brief overview of the corresponding time-integration schemes, as well as some of the most relevant concepts from the field.

3.1 Explicit Euler

The goal of time integration methods is to approximate the solution to the problem

$$\begin{aligned} w'(t) &= f(t, w(t)), \quad t \in (t_0, T], \\ w(t_0) &= w_0, \end{aligned} \tag{3.1}$$

where $f : [t_0, T] \times \mathbb{R}^d \to \mathbb{R}$.

The most simple method is perhaps the *explicit Euler method*. We start with choosing a grid of time points $\{t_k\}_{k=0}^N$ defined by $t_{k+1} = t_k + h$, with $h = \frac{T-t_0}{N}$, where h is the *step size*. Using the knowledge that $w(t_0) = w_0$, we define a sequence of approximations $\{w_k\}_{k=0}^N$ iteratively, where each $w_k \approx w(t_k)$, by approximating the left-hand side of (3.1) with a forward difference approximation

$$\frac{w(t+h) - w(t)}{h} \approx f(t, w(t))$$

which gives the recursion

$$w_{k+1} = w_k + hf(t_k, w_k). \tag{3.2}$$

Suppose that we at time point t_k actually have the exact value of $w(t_k)$ at hand. An important question that arises is how far the approximation w_{k+1} will be

from $w(t_{k+1})$, if we make use of (3.2). Assuming that f is twice continuously differentiable, we get by Taylor expansion that

$$w(t_{k+1}) = w(t_k + h) = w(t_k) + h \cdot w'(t_k) + \frac{h^2}{2} \cdot w''(\theta_k) \tag{3.3}$$

$$= w(t_k) + h \cdot f(t_k, w(t_k)) + \frac{h^2}{2} \cdot w''(\theta_k), \quad \theta_k \in [t_k, t_{k+1}]. \tag{3.4}$$

Hence we see that if $\sup_{\theta_k \in [t_k, t_{k+1}]} \|w''(\theta_k)\|$ is bounded, the *local error* defined by $r_k = w(t_{k+1}) - (w(t_k) + h \cdot f(t_k, w(t_k)))$ satisfies

$$\|r_k\| \leq C \cdot h^2, \ C > 0,$$

which tends to 0 as h tends to 0. A method that satisfies this property is referred to as a *consistent* method. In this case, as the local error is $\mathcal{O}(h^2)$, we say that the order of consistency of the method is 1.

Another quantity of interest is the *global error*, given by $e_k = w(t_k) - w_k$. While the local error measures the error made in one step, the global error measures the accumulated error at time t_k. With starting point in the local error (3.3), we subtract w_{k+1} from both sides which yields the equation

$$w(t_{k+1}) - w_{k+1} = w(t_k) - w_k + h \cdot (f(t_k, w(t_k)) - f(t_k, w_k)) + w''(\theta_k)\frac{h^2}{2}.$$

If we assume that f is Lipschitz continuous with Lipschitz constant L, we get the following bound on the global error

$$\|e_{k+1}\| \leq (1 + hL) \|e_k\| + C \cdot h^2,$$

where $C = \frac{w''(\theta_k)}{2}$. It can be shown by induction, along with the fact that $1 + x \leq e^x$ for $x \geq 0$, [40, p. 6] that the global error satisfies

$$\|e_k\| \leq \frac{c}{L} \left(e^{(T - t_0)L} - 1 \right) h. \tag{3.5}$$

This also means that the explicit Euler method is *convergent*; the maximum error tends to 0 as the step size tends to 0, and this holds for any initial value problem (3.1) for which the function f on the left-hand side is Lipschitz-continuous and whose solution is twice continuously differentiable, with bounded second derivative. The error constant in (3.5) is not very good for practical purposes; it is however possible to obtain sharper error bounds, see [40].

3.2 Implicit Euler

Instead of evaluating the function on the right-hand side of (3.1) at (t_{k+1}, w_{k+1}) one obtains the implicit Euler update

$$w_{k+1} = w_k + hf(t_{k+1}, w_{k+1}). \tag{3.6}$$

This can be rewritten on the form

$$w_{k+1} = R_k w_k, \tag{3.7}$$

where $R_k = (I + hf(t_{k+1}, \cdot))^{-1}$ is the *resolvent* of f. In order to investigate the order of consistency, we consider the difference

$$r_{k+1} = w(t_{k+1}) - w(t_k) - hf(t_{k+1}, w(t_{k+1})).$$

Following [40, Chap. 1.4], we expand the first term in Taylor series around t_k and exchange the last for $w'(t_{k+1})$. This yields

$$r_{k+1} = w(t_k) + hw'(t_k) + \frac{h^2}{2}w''(t_k) + \mathcal{O}(h^3) - w(t_k) - hw'(t_{k+1}).$$

We proceed with expanding the last term in Taylor series around t_k which gives

$$\begin{aligned} r_{k+1} =& w(t_k) + w'(t_k)h + \frac{h^2}{2}w''(t_k) + \mathcal{O}(h^3) - w(t_k) - \\ & h\Big\{ w'(t_k) + hw''(t_k) + \mathcal{O}(h^2) \Big\}. \end{aligned}$$

From this we see that

$$r_{k+1} = w(t_{k+1}) - w(t_k) - hf(t_{k+1}, w(t_{k+1})) = -\frac{h^2}{2}w''(t_k) + \mathcal{O}(h^3), \tag{3.8}$$

We see that the local error is $\mathcal{O}(h^2)$, and hence the implicit Euler scheme is consistent of order 1.

As in the explicit Euler case, it is possible to show that the global error $e_k = w(t_k) - w_k$ satisfies a bound similar to (3.5). See e.g. [27, 40]. The advantage of using the implicit Euler method over the explicit Euler method is that it is more stable and allows for larger step sizes. It is however more computationally costly in general compared to explicit methods, as one needs to solve an implicit equation in order to obtain the next iterate in each step.

3.3 Runge–Kutta methods

The starting point of Runge–Kutta methods, is the observation that the problem (3.1) equivalently can be written as an integral equation

$$w(t) = w_0 + \int_0^t f(s, w(s))\mathrm{d}s.$$

The relation between the solution to (3.1) at time t_k and t_{k+1} can thus be expressed as

$$w(t_{k+1}) = w(t_k) + \int_{t_k}^{t_{k+1}} f(s, w(s))\mathrm{d}s. \tag{3.9}$$

Given an approximation $w_k \approx w(t_k)$, we can obtain an approximation to $w(t_{k+1})$ by using a quadrature formula to approximate the integral on the right-hand side of (3.9), i.e.

$$w_{k+1} = w_k + h\sum_{i=0}^{s} b_i f(t_{k,i}, w_{k,i}). \tag{3.10}$$

Here $t_{k,i} \in [t_k, t_{k+1}]$ and the coefficients b_i are weights from the quadrature rule. As we do not have the function $w(t)$ at hand, we need approximations $w_{k,i}$ to the points $w(t_{k,i})$. In Runge–Kutta methods, the *intermediate stages* $w_{k,i}$ are computed in a recursive fashion according to the rule

$$w_{k,i} = w_k + h\sum_{j=1}^{s} a_{i,j} f(t + hc_j, w_{k,j}). \tag{3.11}$$

If $a_{i,j} = 0$ for $j \geq i$, the method is *explicit*, otherwise *implicit*. The coefficients $a_{i,j}, b_i, c_j$, in (3.10) and (3.11) are chosen such that the local and global error satisfies certain order conditions. A common assumption is that $c_i = \sum_{j=1}^{s} a_{i,j}$, see [35]. For consistency of order 1, which is used in Paper II, we need to impose the condition that $\sum_{i=1}^{s} b_i = 1$, see Section II.1.1 of [38].

The advantages of Runge–Kutta methods is that they can have larger stability regions than the explicit Euler method and allow for larger step sizes. This comes with the price of a higher computational cost in each step.

3.4 Stability

Consider the initial value problem

$$\begin{aligned} w'(t) &= f(w(t)),\ t > t_0, \\ w(t_0) &= w_0, \end{aligned} \tag{3.12}$$

where $f : \mathbb{R}^d \to \mathbb{R}$. For simplicity, we consider autonomous systems in this section. We say that $w(t) = w_*$ is an *equilibrium solution* if $w'(t) = 0$, i.e. it is constant in time. It is said to be a *stable equilibrium solution*, if for any $\varepsilon > 0$, there is a $\delta > 0$ such that $\|w(t_0) - w_*\| < \delta$ implies that $\|w(t) - w_*\| < \varepsilon$, for all $t \geq t_0$. That is, any small perturbation of the equilibrium solution will remain in an $\varepsilon-$neighborhood of w_* at any time $t \geq t_0$. If it also holds that $\lim_{t\to\infty}\|w(t) - w_*\| = 0$, the solution is said to be *asymptotically stable*.

It is possible to show that w_* is an asymptotically stable equilibrium solution if all the eigenvalues of the Jacobian of f at w_* have negative real part, see Theorem 1.2.5 in [68]. If we assume that the Jacobian at w_* is diagonalizable, then the linearized system

$$\begin{aligned} w_l'(t) &= J_f(w_*)w_l(t),\ t > t_0, \\ w_l(t_0) &= w_0, \end{aligned} \tag{3.13}$$

is equivalent to a $d-$dimensional system of equations

$$\begin{aligned} x'(t) &= \Lambda x(t),\ t > t_0, \\ x(t_0) &= x_0- \end{aligned}$$

Here Λ is a diagonal matrix such that $J_f(w_*) = Q^{-1}\Lambda Q$, for some invertible matrix Q, and whose diagonal elements are the eigenvalues of $J_f(w_*)$. We have also made the change of variable $x(t) = Qw_l(t)$. We thus have d linear equations, all of the form

$$\begin{aligned} y'(t) &= \lambda y(t),\ \lambda \in \mathbb{C},\ t > t_0, \\ y(0) &= y_0. \end{aligned} \tag{3.14}$$

Equation (3.14) is known as the *linear test equation*. Since we have $|y(t)| = e^{\mathrm{Re}(\lambda)t}|y_0|$ for Equation (3.14), we see that $y_* = 0$ is stable if and only if $\mathrm{Re}(\lambda) \leq 0$. For $\mathrm{Re}(\lambda) < 0$ it is asymptotically stable.

For a numerical method that produces a sequence of approximations $\{y_k\}_{k\geq 0}$ to the solution to (3.14), it would be desirable that it mimicked this behavior; i.e. it should satisfy

$$\lim_{k\to\infty} y_k = 0, \tag{3.15}$$

when applied to equation 3.14 with $\mathrm{Re}(\lambda) < 0$.

If we apply the *explicit Euler method* from Section 3.1 to (3.14), we obtain the difference equation

$$y_{n+1} = R(z)y_n,$$

where $R(z) = 1 + z$ and $z = h\lambda$. The function $R(z)$ is referred to as the *stability function* of the method. For the values $z \in \mathbb{C}$ such that $|R(z)| < 1$ (the *stability domain* of the method), (3.15) holds as we have that $|y_{n+1}| < |y_n|$. In the case of the explicit Euler method, we require that $|1+z| < 1$. If $\lambda \in \mathbb{R}_-$ (the negative real line including 0), this results in a step size restriction $h < -\frac{2}{\lambda}$.

For the implicit Euler method from Section 3.2, the stability function is given by $R(z) = (1-z)^{-1}$. The stability domain is thus $\{z \in \mathbb{C} : |1-z| > 1\}$. For the implicit Euler method it holds that $\mathbb{C}_- = \{z \in \mathbb{C} : \mathrm{Re}(z) < 0\}$ is contained in the stability domain. A method that satisfies this, is said to be *A-stable*, see Definition 3.3 in [36]. In particular, the negative real line $\mathbb{R}_-$, is included in the stability domain of an A-stable method. This means that there is no restriction on the step size for $\lambda \in \mathbb{C}_-$. However, A-stable Runge–Kutta methods are always implicit, compare Lemma 4.2 in [40] and its corollary. This means that it is in generally necessary to solve a non-linear equation to obtain the update in each step. A-stable Runge–Kutta methods are therefore generally more computationally demanding than explicit Runge–Kutta methods.

For the Runge–Kutta methods introduced in Section 3.3, applying (3.10) and (3.11) to (3.14), gives the update $y_{n+1} = R(z)y_n$, where $R(z) = 1 + zb^t(I - zA)^{-1}\mathbb{1}$, where $\mathbb{1} = (1, \ldots, 1)^T \in \mathbb{R}^s$, b is a vector containing the b_i coefficients of (3.10) and

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,s} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,s} \\ \vdots & \vdots & \ddots & \vdots \\ a_{s,1} & a_{s,2} & \ldots & a_{s,s} \end{pmatrix}.$$

Hence the stability domain $\mathcal{S} = \{z \in \mathbb{C} : |R(z)| < 1\}$ of a Runge–Kutta method depends on the coefficient matrix A and the vector b.

One thing to note is that the larger part of the negative real axis the stability region $\mathcal{S}$ of a method contains, the larger step size it allows for. Following [38], we define the *real stability boundary* of a method, $\beta_R > 0$, as the largest number such that $[-\beta_R, 0] \subset \bar{\mathcal{S}}$. Here $\bar{\mathcal{S}}$ denotes the closure of the stability region. For any explicit s-stage Runge–Kutta method, it holds that $\beta_R \leq 2s^2$,

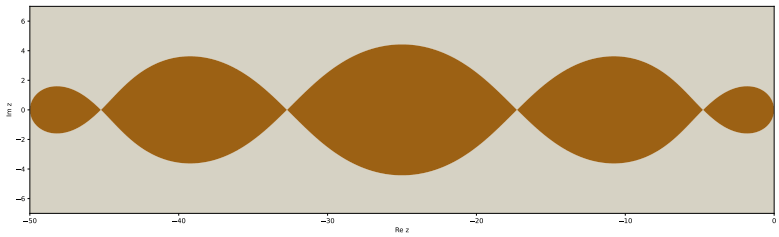

Figure 3.1: Stability region of a RKC method with 5 stages. We see that there are points z on the negative real axis for which $|R(z)| = 1$.

compare [38, Thm. 1.1]. There is a class of Runge–Kutta methods whose real stability boundary satisfies $\beta_R = 2s^2$. These are knows as *Runge–Kutta–Chebyshev methods*. For brevity, we will refer to these as RKC methods. The stability function of such a method is given by

$$R_s(z) = T_s\left(1 + \frac{z}{s^2}\right), \tag{3.16}$$

where s is the number of stages of the method and T_s is the s-th Chebyshev polynomial, defined by the recurrence relation

$$\begin{aligned} T_0(x) &= 1, \\ T_1(x) &= x, \\ T_{n+1}(x) &= 2xT_n(x) - T_{n-1}(x). \end{aligned}$$

In Figure 3.2, we see the stability region for a RKC method with $s = 5$ stages. A problem with RKC methods, is that there will be points $z \in (-\beta_R, 0)$ such that $|R(z)| = 1$. This means that a small error due to numerical inaccuracy could cause the iterates to end up outside of the stability domain. A remedy for this, see [38, V.1], is to introduce a damping factor. Instead of using the stability polynomials in (3.16), one uses damped versions of these;

$$R_s(z) = \frac{T_s\left(\omega_0 + \omega_1 z\right)}{T_s\left(\omega_0\right)}, \ \omega_1 = \frac{T_s\left(\omega_0\right)}{T_s'\left(\omega_0\right)},$$

where $\omega_0 > 1$ is a parameter. With $\omega_0 = 1 + \frac{\varepsilon}{s^2}$, for $\varepsilon > 0$, the real stability boundary then becomes $\beta_R = \frac{2\omega_0 T_s'(\omega_0)}{T_s(\omega_0)} \approx \left(2 - \frac{4}{3}\varepsilon\right) s^2$, see [38]. For small $\varepsilon > 0$ it is a slight reduction compared to that of the undamped method, but instead we gain some margin around the critical points. See Figure 3.2 for an illustration of the stability region of a damped RKC-method with 5 stages.

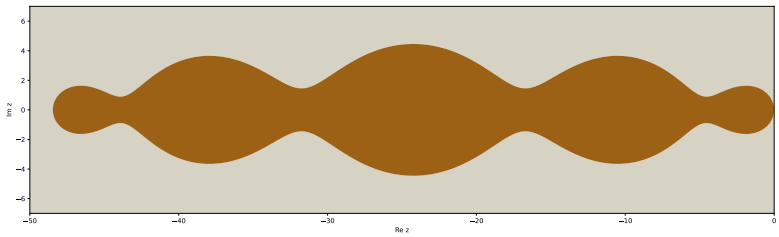

Figure 3.2: Stability region of a damped RKC method with 5 stages. The damping parameter $\varepsilon = 0.05$.

4. Optimization

The principle of empirical risk minimization, described in Chapter 2, tells us that we can minimize the empirical risk functional (2.2), instead of the risk functional (2.1). Thus, we have transformed the problem from that of finding the minimum of the unknown function (2.1), to an unconstrained optimization problem with the empirical risk functional (2.2) as the objective function. In this chapter, we will describe various common optimization methods for approximating the solution to such problems.

4.1 Gradient descent

Let $F : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function such that its derivative is Lipschitz-continuous with Lipschitz constant L. Further, assume that F is bounded below by some number F_*. Suppose that we want to find a solution to the problem

$$w_* = \arg\min_{w \in \mathbb{R}^d} F(w). \tag{4.1}$$

A common algorithm for approximating the solution w_*, is the *gradient descent method*. We start by choosing an initial iterate w_1. A sequence of approximations $\{w_k\}_{k \geq 1}$ is then produced by letting

$$w_{k+1} = w_k - \alpha_k \nabla F(w_k), \tag{4.2}$$

for a step size/learning rate $\alpha_k > 0$. We note that (4.2) corresponds to the explicit Euler method in Chapter 3. By the Lipschitz continuity of the gradient of F, it holds that

$$F(w_{k+1}) \leq F(w_k) + \langle \nabla F(w_k), w_{k+1} - w_k \rangle + \frac{L}{2} \|w_{k+1} - w_k\|_2^2, \tag{4.3}$$

compare Lemma 1.2.3 in [52]. If we use (4.2) in this expression, we obtain

$$F(w_{k+1}) \leq F(w_k) - \alpha_k \left(1 - \frac{L\alpha_k}{2}\right) \|\nabla F(w_k)\|_2^2. \tag{4.4}$$

Assuming that $\alpha_k < \frac{2}{L}$, the term $1 - \frac{L\alpha_k}{2}$ is positive and hence we see that the sequence $\{F(w_k)\}_{k\geq 1}$ is a decreasing sequence. By differentiating $\varphi(\alpha) = -\alpha + \frac{L\alpha^2}{2}$, we find that the maximum decrease we can achieve in an iteration is when we take $\alpha_k = \frac{1}{L}$. Let us now suppose for simplicity that $\alpha_k = \frac{1}{L}$. Then (4.4) becomes

$$\frac{1}{2L}\|\nabla F(w_k)\|_2^2 \leq F(w_k) - F(w_{k+1}).$$

By summing up from 1 to K we see that

$$\frac{1}{2L}\sum_{k=0}^{K}\|\nabla F(w_k)\|_2^2 \leq F(w_0) - F(w_{K+1}) \leq F(w_0) - F_*,$$

where F_* is the lower bound of (4.1). If we let K tend to ∞ in the sum above, we see that the sum it is finite, since the right-hand side is independent of K. Thus, we can conclude that

$$\lim_{k\to\infty} \|\nabla F(w_k)\|_2 = 0,$$

i.e. we reach a stationary point of F in the limit. It turns out that we can say more about the local convergence under further assumptions. Closely following Section 1.2.3 in [52], we assume that

1. The Hessian $\nabla^2 F$ of F is Lipschitz continuous with Lipschitz constant M.

2. There is a local minimum w_* at which the Hessian is positive definite, with the smallest eigenvalue $l > 0$ and largest eigenvalue $L > 0$.

3. The initial iterate w_0 is close enough to w_* in the sense that

$$\|w_0 - w_*\|_2 < \frac{2l}{M}. \tag{4.5}$$

Then we can ensure that $\|w_{k+1} - w_*\|_2 < \|w_k - w_*\|_2$. To see this, we start with noting that

$$\begin{aligned}\nabla F(w_k) &= \nabla F(w_k) - \nabla F(w_*) = \int_0^1 \nabla^2 F(w_* + t(w_k - w_*))(w_k - w_*)dt \\ &=: G_k(w_k - w_*).\end{aligned}$$

By adding subtracting w_* from both sides of (4.2) we get the recurrence relation

$$w_{k+1} - w_* = (I - \alpha_k G_k)(w_k - w_*).$$

Using the Lipschitz continuity of $\nabla^2 F$, it is possible to show that if $\|w_k - w_*\|_2 < \frac{2l}{M}$, then

$$\|I - \alpha_k G_k\| < 1,$$

compare Corollary 1.2.2 and Theorem 1.2.4 in [52]. In the expression above, $\|\cdot\|$ denotes the 2-norm for matrices. From this, and the fact that $\|w_{k+1} - w_*\|_2 \leq \|I - \alpha_k G_k\| \|w_k - w_*\|_2$, we see that the sequence $\{w_k\}_{k \geq 0}$ converges to w_*. For the optimal choice of step size

$$\alpha_k = \frac{2}{l + L}, \forall k \geq 1, \tag{4.6}$$

one obtains a linear convergence rate, in the sense that

$$\|w_k - w_*\| \leq \frac{2lL\|w_0 - w_*\|}{2l - L\|w_0 - w_*\|} \left(1 - \frac{2l}{L + 3l}\right)^k.$$

For a derivation of this result, see Theorem 1.2.4 in [52]. As an example of a function that fulfills the assumptions above, we can take $F(w) = \sin(w)$ for $w \in \mathbb{R}$. Given that we start close enough to $w_* = \frac{3\pi}{2}$, we will converge to w_* if the step size is small enough.

The previous convergence guarantee holds for non-convex functions that are sufficiently smooth. As long as we start close enough to a local minimum, we will converge to that minimum linearly. This is under the premise that the step size is chosen according to (4.6). It essentially tells us that the non-convex case behaves like the convex case if we are close enough to a local minimum. The issue with this is that it might be hard in practice to estimate the constants l, L and M. Therefore, it is difficult to estimate (4.6) and ensure that the convergence is linear. Similarly, it is in most cases not possible to find the constant on the right-hand side of (4.5). Furthermore, the local minimum w_* is not known beforehand, and it is not feasible to choose the initial iterate w_0 according to (4.5).

4.2 Proximal point method

In the previous section, we noted that the update (4.2) could be seen as an *explicit Euler discretization* of the gradient flow equation (1.6). Another common option is to instead use the *implicit Euler scheme* as discretization; instead of evaluating ∇F at w_k, we choose to evaluate it at w_{k+1}, which gives the update

$$w_{k+1} = w_k - \alpha_k \nabla F(w_{k+1}). \tag{4.7}$$

In the optimization setting, the update is often seen in the form

$$w_{k+1} = \text{prox}_{F,\alpha_k}(w_k) = \underset{w\in\mathbb{R}^d}{\arg\min} \big\{ F(w) + \frac{1}{2\alpha_k}\|w - w_k\|_2^2 \big\}. \tag{4.8}$$

For differentiable F, the equivalence of (4.7) and (4.8) can be seen by differentiating the expression $F(w) + \frac{1}{2\alpha_k}\|w - w_k\|_2^2$. Another way to look at (4.8), at least for convex functions, is as a generalization of orthogonal projection. If we let C be a convex set, then the indicator function

$$I_C(w) = \begin{cases} 0 \ , & w \in C, \\ \infty, & w \notin C, \end{cases}$$

is a convex function and we have that

$$\text{prox}_{I_C,\alpha_k}(w_k) = \underset{w\in\mathbb{R}^d}{\arg\min} \big\{ I_C(w) + \frac{1}{2\alpha_k}\|w - w_k\|_2^2 \big\},$$

which is the orthogonal projection of w_k onto C.

4.3 Stochastic gradient descent

For machine learning problems, the function F in (4.1) is often of the form

$$F(w) = \frac{1}{N}\sum_{i=1}^{N} \ell(h(x_i, w), y_i), \tag{4.9}$$

where ℓ is a loss function, $h(\cdot, w)$ is a prediction function and $\{(x_i, y_i)\}_{i=1}^N$ is a sample of feature-label pairs. In Chapter 2, we adopted the point of view that the objective function depended on a random sample. Now, we are instead concerned with the problem of minimizing (4.9) with respect to w, for a known sample $\{(x_i, y_i)\}_{i=1}^N$. Hence, the objective function (4.9), is a deterministic function. For each of the functions in the sum of (4.9), we need to evaluate

the gradient if we want to compute $\nabla F(w)$. Hence, the gradient update (4.2) can be very computationally expensive for a large number of samples N. A solution to this is the stochastic gradient descent method which, instead of computing the full gradient at each iteration, computes an approximation $\nabla f(w_k, \xi_k)$ which is used instead:

$$w_{k+1} = w_k - \alpha_k \nabla f(w_k, \xi_k). \tag{4.10}$$

Here $\{\xi_k\}_{k\geq 1}$ is a sequence of i.i.d random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and could for example denote the act of choosing a *batch*, i.e. a random subset of indices $B_k \subset \{1, \dots, N\}$. In this case, we would get

$$\nabla f(w_k, \xi_k) = \frac{1}{|B_k|} \sum_{i \in B_k} \ell(h(x_i, w), y_i). \tag{4.11}$$

In the following, we will by $\mathbb{E}_{\xi_k}[\cdot]$ denote the conditional expectation taken with respect to $\sigma-$algebra generated by the sequence $\xi_{k-1}, \dots, \xi_1$. Note that as w_k only depends on ξ_j for $j < k$, w_k is independent of ξ_k, by the assumption that the sequence $\{\xi_k\}_{k\geq 1}$ are mutually independent. There are several strategies for showing convergence of the stochastic gradient descent method. In this section, we will closely follow the approach in [14]. We start with looking at the results in the non-convex case, and we assume that there exists some global lower bound F_* such that

$$F_* \leq F(w), \ \forall w \in \mathbb{R}^d. \tag{4.12}$$

Another common assumption, which we will also make, is that F has Lipschitz continuous gradients. By (4.3) and the fact that

$$w_{k+1} - w_k = -\alpha_k \nabla f(w_k, \xi_k),$$

we get that

$$F(w_{k+1}) - F(w_k) \leq -\alpha_k \langle \nabla F(w_k), \nabla f(w_k, \xi_k) \rangle + \frac{L\alpha_k^2}{2} \|\nabla f(w_k, \xi_k)\|_2^2.$$

If the stochastic gradient is an unbiased estimate of $\nabla F(w)$, i.e.

$$\mathbb{E}_\xi[\nabla f(w, \xi)] = \nabla F(w), \tag{4.13}$$

we get, after taking the expectation w.r.t. ξ_k and using that w_k is independent of ξ_k,

$$\mathbb{E}_{\xi_k}\left[F(w_{k+1})\right] - F(w_k) \leq -\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{L\alpha_k^2}{2} \mathbb{E}_{\xi_k}\left[\|\nabla f(w_k, \xi_k)\|_2^2\right]. \tag{4.14}$$

We now introduce the assumption that

$$\mathbb{E}_{\xi_k}\left[\|\nabla f(w_k,\xi_k)\|_2^2\right] \leq M + M_G\|\nabla F(w_k)\|_2^2,\ \forall k \geq 1, \tag{4.15}$$

for some constants $M, M_G > 0$. This essentially means that the variance is allowed to grow if the gradient grows and is bounded at stationary points. If we insert (4.15) into (4.14), we get

$$\mathbb{E}_{\xi_k}\left[F(w_{k+1})\right] - F(w_k) \leq -\alpha_k\left(1 - \frac{\alpha_k LM_G}{2}\right)\|\nabla F(w_k)\|_2^2 + \frac{LM\alpha_k^2}{2}.$$

Here we see that if we impose the step size restriction $\alpha_k \leq \frac{1}{LM_G}$, the term $1 - \frac{\alpha_k LM_G}{2} > \frac{1}{2}$. Thus, the previous bound becomes

$$\mathbb{E}_{\xi_k}\left[F(w_{k+1})\right] - F(w_k) \leq -\frac{\alpha_k}{2}\|\nabla F(w_k)\|_2^2 + \frac{LM\alpha_k^2}{2}. \tag{4.16}$$

We now take the expectation of the previous expression:

$$\mathbb{E}\left[F(w_{k+1})\right] - \mathbb{E}\left[F(w_k)\right] \leq -\frac{\alpha_k}{2}\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] + \frac{LM}{2}\alpha_k^2. \tag{4.17}$$

If we rearrange the terms and sum from 1 to K, we arrive at the inequality

$$\sum_{k=1}^{K}\alpha_k\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] \leq 2\left(F(w_1) - \mathbb{E}\left[F(w_{K+1})\right]\right) + LM\sum_{k=1}^{K}\alpha_k^2. \tag{4.18}$$

Here we have used the fact that $\mathbb{E}\left[F(w_1)\right] = F(w_1)$ since w_1 is deterministic. The left-hand side of the previous inequality can be bounded from below by as follows

$$\min_{1\leq k\leq K}\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right]\sum_{k=1}^{K}\alpha_k \leq \sum_{k=1}^{K}\alpha_k\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right]. \tag{4.19}$$

After dividing both sides by $\sum_{k=1}^{K}\alpha_k$ we then get

$$\min_{1\leq k\leq K}\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] \leq \frac{2\left(F(w_1) - F_*\right) + LM\sum_{k=1}^{K}\alpha_k^2}{\sum_{k=1}^{K}\alpha_k}, \tag{4.20}$$

where we have used (4.12) in order to bound $-\mathbb{E}\left[F(w_{K+1})\right]$ by $-F_*$. From (4.20) we see that the sequence

$$\left\{\min_{1\leq k\leq K}\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right]\right\}_{K\geq 1} \tag{4.21}$$

converges to 0, if we require that

$$\sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty. \tag{4.22}$$

With the additional regularity assumptions that $\|\nabla F(w)\|_2^2$ is differentiable, it is possible to show

$$\lim_{k\to\infty} \mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] = 0,$$

although not with a rate, compare Corollary 4.12 in [14].

We will now turn our attention to the convex case. A common assumption is that the objective function is strongly convex with convexity constant $c > 0$, i.e.

$$F(w') - F(w) \geq \langle \nabla F(w), w' - w \rangle + \frac{c}{2}\|w' - w\|_2^2, \quad w, w' \in \mathbb{R}^d.$$

Strongly convex (and differentiable) functions satisfy the inequality

$$2c\,(F(w) - F(w_*)) \leq \|\nabla F(w)\|_2^2, \tag{4.23}$$

where w_* is the unique global minimum of F, see Appendix B in [14]. The fact that such a minimum exists follows from (4.12), the continuity of F along with the strong convexity, compare Corollary 11.17 in [5]. Inserting (4.23) into inequality (4.17), we get

$$\mathbb{E}\left[F(w_{k+1})\right] - F(w_k) \leq -c\alpha_k \left(\mathbb{E}\left[F(w_k)\right] - F(w_*)\right) + \frac{LM}{2}\alpha_k^2.$$

Here we can subtract $F(w_*)$ and add $F(w_k)$ from both sides, which yields the recurrence inequality

$$\mathbb{E}\left[F(w_{k+1})\right] - F(w_*) \leq (1 - c\alpha_k)\left(\mathbb{E}\left[F(w_k)\right] - F(w_*)\right) + LM\alpha_k^2. \tag{4.24}$$

Using an induction argument as in Theorem 4.7 in [14], we can use (4.24) to show that with $\alpha_k = \frac{\beta}{k+\gamma}$, where $\beta > \frac{1}{c}$ and $\gamma > 0$, we have

$$\mathbb{E}\left[F(w_k) - F(w_*)\right] \leq \frac{\nu}{k+\gamma}, \tag{4.25}$$

and where

$$\nu = \max\left\{ \frac{LM\beta^2}{2(c\beta - 1)}, (1+\gamma)\left(F(w_1) - F(w_*)\right) \right\}.$$

The constant ν is chosen such that we can perform the base- and induction step of the proof, as in [14]. The decreasing step size in (4.22) and (4.25) is needed for convergence. If we use a fixed step size, i.e. $\alpha_k = \alpha$ for all $k \in \mathbb{N}$, the bound (4.20) becomes

$$\min_{1 \leq k \leq K} \mathbb{E}\left[\|\nabla F(w_k)\|_2^2\|\right] \leq \frac{2\left(F(w_1) - F_*\right)}{\alpha K} + LM\alpha. \tag{4.26}$$

Letting the number of iterations K tend to infinity, the first term on the right-hand side tends to 0, while the second is unaffected. Thus, the sequence (4.21) stays bounded, but it does not necessarily converge to 0. This is sometimes referred to as a *noise-ball* around a stationary point. Similarly, we can use (4.24) with a fixed step size, to show that the sequence $\{F(w_k)\}_{k \geq 1}$ converges to a bounded region around the minimum $F(w_*)$, see [14, Thm. 4.6]. Indeed, by subtracting $\frac{L\alpha}{c}$ from both sides of (4.24), we get the bound

$$\mathbb{E}\left[F(w_{k+1})\right] - F(w_*) - \frac{L\alpha}{c} \leq (1 - 2c\alpha)\left(\mathbb{E}\left[F(w_k)\right] - F(w_*) - \frac{L\alpha}{c}\right).$$

If $\alpha < \frac{1}{c}$ this will be a contraction, and we find that

$$\mathbb{E}\left[F(w_{k+1})\right] - F(w_*) \leq \frac{L\sigma^2\alpha}{4c} + (1 - 2c\alpha)^k\left(F(w_1) - F(w_*) - \frac{L\sigma^2\alpha}{4c}\right), \tag{4.27}$$

from which we conclude that $\mathbb{E}\left[F(w_{k+1})\right] - F(w_*)$ is bounded by $\frac{L\sigma^2\alpha}{4c}$ as k tends to infinity. A potential strategy is to start a scheme with a constant step size until we are close to the bounded region around the stationary point, and then use a decreasing step size to obtain convergence. It is possible to control the size of the bound in (4.27) and (4.26), by choosing the constant step size α small enough. A classical choice is to take $\alpha = 1/\sqrt{K}$, so that the step size depends on the number of iterations. If we plug this value of α into (4.26), we see that we will have achieved an error of size $\mathcal{O}(1/\sqrt{K})$ after K iterations in the non-convex case.

Another result that one sometimes encounters is the following; consider (4.18) and divide both sides by $\sum_{j=1}^K \alpha_j$:

$$\sum_{k=1}^K \frac{\alpha_k}{\sum_{j=1}^K \alpha_j} \mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] \leq \frac{2\left(F(w_1) - F_*\right) + LM\sum_{k=1}^K \alpha_k^2}{\sum_{k=1}^K \alpha_k}.$$

If we for a fixed number of total iterations K introduce a random variable R such that

$$\mathbb{P}\left(R = k\right) = \frac{\alpha_k}{\sum_{k=1}^K \alpha_k}, \quad \text{for } k = 1, \ldots, K. \tag{4.28}$$

Then we can rewrite the previous inequality as

$$\mathbb{E}\left[\|\nabla F(w_R)\|_2^2\right] \leq \frac{2\left(F(w_1) - F_*\right) + LM\sum_{k=1}^{K}\alpha_k^2}{\sum_{k=1}^{K}\alpha_k}.$$

With a constant step size of $\alpha_k = \frac{1}{\sqrt{K}}$ we get

$$\mathbb{E}\left[\|\nabla F(w_R)\|_2^2\right] \leq \frac{2\left(F(w_1) - F_*\right) + LM}{\sqrt{K}}.$$

This means that an algorithm whose output is a randomly selected iterate according to (4.28), will on average have "small" gradients. How small the gradients are on average depends on the total number of iterations K. This view point is for instance adopted in [32]. The rationale is that such an algorithm does not require any additional computational effort to estimate $\min_{0\leq k\leq K}\|\nabla F(w_k)\|_2$. We can obtain similar results for the algorithms in Paper II and Paper III as well, although this is not explicitly stated.

5. ODE method

The results in Chapter 4.4.3 allows us to conclude that a subsequence of SGD converges to a stationary point of the objective function for *non-convex* functions. This can for instance be shown by making use of (4.21) in the previous chapter, along with Chebyshev's inequality. See e.g. Corollary 7 of Paper III. We can however not conclude that the entire sequence converges to a stationary point. In this chapter, we will discuss the *ODE method*, which provides a framework for demonstrating such convergence. The analysis presented in Paper IV is based on this method. For the sake of illustrating the method, we will here for simplicity perform the analysis for the gradient flow

$$w'(t) = -\nabla F(w(t)), \tag{5.1}$$

and SGD-update

$$w_{k+1} = w_k - \alpha_k \nabla f(w_k, \xi_k), \tag{5.2}$$

but the main ideas translates to the analysis of the Hamiltonian system (7) and update (9) in Paper IV. The method originates from [47] and the particular strategy that we will employ here is due to [43, 44]. It can be broadly summarized in four steps:

Step 1 Introduce a pseudo-time $t_k = \sum_{i=0}^{k-1} \alpha_i$ and construct a piecewise constant interpolation $W_0(t)$ of $\{w_k\}_{k\geq 0}$ generated by (5.2).

Step 2 Show that the time shifted process $W_k(t) = W_0(t_k + t)$ asymptotically satisfies (5.1).

Step 3 Demonstrate that every subsequence of $\{W_k\}_{k\geq 0}$ has a further subsequence converging to a solution to (5.1).

Step 4 At last, make use of the underlying dynamics of (5.1) to conclude that $\{w_k\}_{k\geq 0}$ converges almost surely to a stationary point of F. We split this up into two steps:

i) Show that the sublevel sets of F, $\{w : F(w) \leq c\}$ are *locally asymptotically stable*.

ii) Make use of the *Kushner & Clark theorem* (Theorem 5.2 in [44]) to show that the sequence $\{w_k\}_{k\geq 0}$ generated by (5.2) converges to a stationary point of F.

We are interested in extracting converging subsequences of $\{W_k\}_{k\geq 0}$ whose limit satisfies (5.1). This will allow us to argue by contradiction to conclude that $\{w_k\}_{k\geq 0}$ converges to a stationary point. **Step 2** above implies that *if* a subsequence of $\{W_k\}_{k\geq 0}$ converges, *then* it converges to a solution to (5.1). To show **Step 3**, we first demonstrate that $\{W_k\}_{k\geq 0}$ satisfies an extended form of equicontinuity. We can then appeal to a version of the Arzelà–Ascoli theorem, compare [44].

We will divide the outline of the proof in the next section according to four above-mentioned steps.

One of the key assumptions for the approach to work is that

$$\sup_{k\in\mathbb{N}} \|w_k\| < \infty \tag{5.3}$$

almost surely. This is relatively strong, but as we shall see, it does hold for (5.2) under the same assumptions on the noise that we made in the previous chapter if the objective function is coercive. In this chapter, we thus make the same assumptions on the objective function F as in Chapter 4.4.1 and the stochastic setting is the same as in Chapter 4.4.3.

5.1 ODE method

In this section, we follow the approach of [44] adapted to Equation (5.1), while filling in many details that are omitted in [44].

Step 1

Consider the sequence $\{w_k\}_{k\geq 0}$ defined by the stochastic gradient update (5.2). We can write it as

$$w_{k+1} = w_k - \alpha_k \nabla F(w_k) + \alpha_k \delta M_k, \tag{5.4}$$

where $\delta M_k = \nabla F(w_k) - \nabla f(w_k, \xi_k)$. We start by defining a sequence of pseudo-time points $\{t_k\}_{k\geq 0}$ by $t_0 = 0$ and $t_k = \sum_{i=0}^{k-1} \alpha_i$. Let W_0 be given by

$$W_0(t) = \begin{cases} w_k, \ t_k \leq t < t_{k+1} \\ w_0, \ t < t_0 \end{cases},$$

or equivalently

$$W_0(t) = w_0 \cdot I_{(-\infty, t_0)}(t) + \sum_{i=0}^{\infty} w_i \cdot I_{[t_i, t_{i+1})}(t), \tag{5.5}$$

where $I_{[t_k, t_{k+1})}(t)$ denotes the indicator function of the interval $[t_k, t_{k+1})$. We note that $W_0(t)$ is a stochastic process defined on the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, depending on $\omega \in \Omega$ through the random variables $\{w_k\}_{k\geq 0}$. By introducing the function $m(t)$, defined by

$$m(t) = k, \ t_k \leq t < t_{k+1}, \tag{5.6}$$

we can write

$$W_0(t) = w_0 - \sum_{i=0}^{m(t)-1} \alpha_i \nabla F(w_i) + M_0(t), \tag{5.7}$$

where $M_0(t) = \sum_{i=0}^{m(t)-1} \alpha_i \delta M_i$. At this point, we introduce the shifted sequences $\{W_k(t)\}_{k\geq 0}$ and $\{M_k(t)\}_{k\geq 0}$:

$$W_k(t) = W_0(t_k + t) \text{ and } M_k(t) = M_0(t_k + t) - M_0(t_k). \tag{5.8}$$

We note that we can write

$$M_k(t) = \sum_{i=k}^{m(t_k+t)-1} \alpha_i \delta M_i. \tag{5.9}$$

From (5.7) and the fact that $W_k(0) = w_k$ we have that

$$W_k(t) = W_k(0) - \sum_{i=k}^{m(t_k+t)-1} \alpha_i \nabla F(w_i) + M_k(t). \tag{5.10}$$

Step 2

We will now show that $\{W_k\}_{k\geq 0}$ can be written as the integral equation corresponding to (5.1), except for terms that converge to 0 uniformly on compact sets

as $k \to \infty$. For this sake, consider the integral

$$I_k = -\int_0^t \nabla F(W_k(s))\mathrm{d}s.$$

Assume that $t \geq 0$ to begin with. From (5.5) and (5.8) we have that

$$\begin{aligned} I_k &= -\int_0^t \nabla F(W_0(t_k + s))\mathrm{d}s \\ &= -\int_0^t \nabla F\left(w_0 \cdot I_{(-\infty,0)}(t) + \sum_{i=0}^{\infty} w_i \cdot I_{[t_i,t_{i+1})}(t)\right)\mathrm{d}s. \end{aligned}$$

Since $t_k + s$ belongs to a single interval $[t_i, t_{i+1})$ and $[t_i, t_{i+1}) \cap [t_j, t_{j+1}) = \emptyset$, for $i \neq j$, we can further rewrite this as

$$I_k = -\int_0^t \nabla F\left(w_0\right) \cdot I_{(-\infty,0)}(t_k + s) + \sum_{i=0}^{\infty} \nabla F\left(w_i\right) \cdot I_{[t_i,t_{i+1})}(t_k + s)\mathrm{d}s.$$

Since $t_k + s \geq 0$ and $t \geq 0$ (by assumption), we see that $(-\infty, 0)$ is not in the interval of integration. Hence, first term inside the parenthesis does not contribute to the integral. Similarly, since $t_i - t_k \leq 0$ for $i \leq k$, any interval $[t_i, t_{i+1})$ for $i < k$ does not contribute to the integral. Thus we can ignore the first term in the integral and start the sum of the second term at $i = k$. Consequently,

$$I_k = -\int_0^t \sum_{i=k}^{\infty} \nabla F\left(w_i\right) \cdot I_{[t_i,t_{i+1})}(t_k + s)\mathrm{d}s.$$

Furthermore, by (5.8), it holds that $t_{m(t_k+t)} \leq t_k + t < t_{m(t_k+t)+1}$. This means that the terms of the sum of index $i \geq m(t_k + t) + 1$ will be outside the region of integration as well. Therefore

$$I_k = -\int_0^t \sum_{i=k}^{m(t_k+t)} \nabla F\left(w_i\right) \cdot I_{[t_i,t_{i+1})}(t_k + s)\mathrm{d}s.$$

At last, since $\alpha_i = \int_{t_i}^{t_{i+1}} I_{[t_i,t_{i+1})}(t)\mathrm{d}t$, we see that the previous expression can be written as

$$-\sum_{i=k}^{m(t_k+t)-1} \alpha_i \nabla F(w_i) + \rho_k(t),$$

where

$$\rho_k(t) = -\nabla F(w_{m(t_k+t)})(t - t_{m(t_k+t)} + t_k).$$

and accounts for the fact that we are not (necessarily) integrating over the entire interval $[t_{m(t_k+t)}, t_{m(t_k+t)+1})$. It follows that

$$-\sum_{i=k}^{m(t_k+t)-1} \alpha_i \nabla F(w_i) = -\int_0^t \nabla F(W_k(s))\mathrm{d}s + \rho_k(t). \tag{5.11}$$

A similar manipulation of the integral gives an equivalent expression when $t < 0$.

Under the assumption that (5.3) holds almost surely, we have by the continuity of ∇F that

$$\|\rho_k(t)\|_2 \leq \|\nabla F(w_{m(t_k+t)})\|_2(t - t_{m(t_k+t)} + t_k) \leq \sup_{k\in\mathbb{N}} \|\nabla F(w_k)\| \alpha_{m(t_k+t)}.$$

This, along with the fact that $\lim_{k\to\infty} \alpha_k = 0$ tells us that

$$\lim_{k\to\infty} \sup_{t\in[0,T]} \|\rho_k(t)\| = 0,$$

for every compact set $T > 0$, $\mathbb{P}-$almost surely. Thus, the sequence $\{\rho_k\}_{k\geq 0}$ tends to 0 uniformly on compact sets in t, except on a set of probability 0 in ω. Combining (5.10) and (5.11) we see that

$$W_k(t) = W_k(0) - \int_0^t \nabla F(W_k(s))\mathrm{d}s - \rho_k(t) + M_k(t). \tag{5.12}$$

One can show that for any $T \geq 0$, it holds that

$$\lim_{k\to\infty} \sup_{t\in[0,T]} \|M_k(t)\|_2 = 0,$$

$\mathbb{P}-$almost surely, see Paper IV or [44] for a proof of this. The main idea is essentially to use the bounded variance, $\mathbb{E}\left[\|\delta M_k\|_2^2\right] < \infty$, to construct a martingale which bounds $\sup_{t\in[0,T]} \|M_k(t)\|_2$, and then apply Doob's submartingale inequality, see e.g. [44, 69]. This means that any limit function W (with respect to uniform convergence) of $\{W_k\}_{k\geq 0}$ satisfies

$$W(t) = W(0) - \int_0^t \nabla F(W(s))\mathrm{d}s. \tag{5.13}$$

The argument is essentially same as that in the proof of Lemma 4.5 in Paper IV.

Step 3

Our next task is to establish that any subsequence of $\{W_k\}_{k\geq 0}$ has a further subsequence that converges to a solution to (5.13). We will do this by showing that $\{W_k\}_{k\geq 0}$ satisfies a generalized form of equicontinuity and then appeal to a version of the Arzelà–Ascoli theorem:

Definition 5.1 (Extended equicontinuity)**.** *A sequence of functions $\{W_k\}_{k\geq 0}$, where $W_k : \mathbb{R}^d \to \mathbb{R}$ for each k, is said to be* equicontinuous in the extended sense *if $\sup_k \|W_k(0)\| < \infty$ and for every T and $\epsilon > 0$ there is $\delta > 0$ such that*

$$\limsup_{k\to\infty} \sup_{0<|t-s|\leq\delta,\ t,s\in[0,T]} |W_k(t) - W_k(s)| \leq \epsilon. \tag{5.14}$$

Theorem 5.2 (Arzelà–Ascoli)**.** *Let $\{W_k\}_{k\geq 0}$ be a sequence of functions $W_k : \mathbb{R}^d \to \mathbb{R}$ that are equicontinuous in the extended sense. Then there is a subsequence $\{W_{n_k}\}_{k\geq 0}$ of $\{W_k\}_{k\geq 0}$ that converges to a continuous function W.*

The limit functions W that we can extract are not necessarily unique and may depend on the subsequence. The proof can be found in e.g. [22] and [12].

We now show that the process $\{W_k\}_{k\geq 0}$ is equicontinuous in the extended sense, $\mathbb{P}$-almost surely. This means that any limit function of $\{W_k\}_{k\geq 0}$ (i.e. any limit of a subsequence of $\{W_k\}_{k\geq 0}$) satisfies (5.13). Since W is continuous by the previous theorem, this implies that W is in fact differentiable and a solution to (5.1). By the construction of the process $\{W_k\}_{k\geq 0}$ as a piecewise constant interpolation of $\{w_k\}_{k\geq 0}$, it also follows that the limit function W takes values in the set of limit points of $\{w_k\}_{k\geq 0}$, compare Proposition 1.b of [28]. We will use this fact later on.

Lemma 5.3 (Equicontinuous in the extended sense)**.** *Consider the sequence $\{W_k\}_{k\geq 0}$ defined by* (5.8)*. Suppose that $\{w_k\}_{k\geq 0}$ is given by* (5.4)*, that the assumptions on the noise and the objective function satisfies those in Chapter 4.4.3 and that $\sup_{k\in\mathbb{N}}\|w_k\| < \infty$, $\mathbb{P}$-almost surely. Then $\{W_k\}_{k\geq 0}$ is equicontinuous in the extended sense, $\mathbb{P}$-almost surely.*

The proof is very similar to that of Lemma 4.3 in paper IV and is therefore omitted here. Roughly, the idea is to make use of the Lipschitz continuity of ∇F to bound the difference $\|W_k(t) - W_k(s)\|_2$ by $C(\omega)\cdot\alpha_k$, where $C(\omega)$ is finite almost surely. Then the fact that α_k tends to 0 as $k \to \infty$ yields the desired result.

Step 4.i)

We now have a connection between the algorithm determined by (5.2) and the solutions to (5.1) through the interpolation sequence $\{W_k\}_{k\geq 0}$. Our next goal is to determine the limit behavior of the solutions to (5.1). But first, we will introduce the concept of a *locally asymptotically stable set* [13, 44]:

Definition 5.4 (Locally asymptotically stable set)**.** *A set A is said to be* Lyapunov stable *if for any $\epsilon > 0$, there exists a $\delta > 0$ such that every trajectory initiated in the δ-neighborhood of A (which we denote by $N_\delta(A)$) remains in its ϵ-neighborhood, $N_\epsilon(A)$. A set A is said to be* locally asymptotically stable *if every such path ultimately goes to A. The largest open set $\mathcal{O}$ such that all paths initiated in $\mathcal{O}$ stays in $\mathcal{O}$ and converges to A is called the* domain of attraction *of A.*

Now, we show that the sublevel sets of F are locally asymptotically stable:

Lemma 5.5. *Suppose that the objective function F is differentiable and coercive. Then the sub-level sets $\{w : F(w) \leq c\}$ are locally asymptotically stable sets for* (5.1)*.*

Demonstrating Lyapunov stability requires a bit of caution; we need to relate the sublevel sets of F to the $\delta-$neighborhoods of $A = \{w : F(w) \leq c\}$ for a given c. For more details, see the proof of Lemma 4.12 in paper IV.

Since F is coercive and continuous, the sublevel sets are compact, compare Proposition 11.12 in [5]. Furthermore, F is decreasing along the paths of (5.1) since for any solution W to (5.1) which is not an equilibrium solution, it holds that

$$\frac{d}{dt}F(W(t)) = \langle \nabla F(W(t)), \dot{W}(t)\rangle = -\|\nabla F(W(t))\|_2^2 < 0,$$

Hence any solution tends to a stationary point of F. However, to find a δ-neighborhood of the set $\{w : F(w) \leq c\}$ such that all paths starting within it tend toward $\{w : F(w) \leq c\}$, we need to ensure that this neighborhood contains no other stationary points of F. Hence, we make the following assumption:

Assumption 5.6. *Let Λ be the set of stationary points of F. For every compact set $K \subset \mathbb{R}$, the cardinality of $K \cap F(\Lambda)$ is finite.*

Under this assumption, we can show that any solution to (5.1) that starts close enough to a given sub-level set, will ultimately end up in it. Hence, we have shown that the sublevel sets of the objective function are asymptotically stable. The details can be found in Lemma 5.16 of Paper IV.

Step 4.ii)

We have seen that we can relate the algorithm (5.2) to the dynamical system (5.1) through the sequence of interpolations $\{W^k\}$, and we have shown that the sublevel sets of the objective function are locally asymptotically stable for the solutions to (5.1). We will now see how this translates to the limit behavior of the algorithm.

Let A be a locally asymptotically stable set and assume that $\{w_k\}_{k\geq 0}$ enters a compact set K in the domain of attraction of A infinitely often (i.e. there exists a subsequence $\{w_{n_k}\}_{k\geq 0}$ of $\{w_k\}_{k\geq 0}$ which is contained in K). The following lemma states that this implies that for every δ-neighborhood $N_\delta(A)$ of A, there exists a subsequence of $\{w_k\}_{k\geq 0}$ that lies in $N_\delta(A)$:

Lemma 5.7. *Let the sequence $\{w_k\}_{k\geq 0}$ be given by* (5.4) *and F be a differentiable and coercive function which is bounded below. Further, assume that* (5.3) *holds. Then for every $\delta > 0$ there exists a subsequence $\{w_{m_k}\}_{k\geq 0}$ of $\{w_k\}_{k\geq 0}$ that lies in $N_\delta(A)$.*

For a proof, see Lemma A.6 in Paper IV. The previous lemma is then used to show that for any $\delta > 0$, the sequence $\{w_k\}_{k\geq 0}$ cannot escape the δ-neighborhood of A infinitely often.

Theorem 5.8 (Kushner & Clark)**.** *Let the assumptions of Lemma 5.7 hold. Further, assume that A is a locally asymptotically stable set for* (5.1) *and that $\{w_k\}_{k\geq 0}$ enters a compact set in the domain of attraction of A infinitely often. Then $\{w_k\}_{k\geq 0}$ converges to the set A almost surely. That is, there is a set U such that $\mathbb{P}(U) = 1$, and for any $\omega \in U$ it holds that*

$$\lim_{k\to\infty} \inf_{a\in A} \|w_k(\omega) - a\|_2 = 0.$$

The proof can be carried out by a contradiction argument: assuming that there do exist a $\epsilon > 0$ and a subsequence $\{w_{n_k}\}_{k\geq 0}$ outside $N_\epsilon(A)$, we can by the Arzelà–Ascoli theorem construct solutions $w(t)$ to (5.1) that starts in $N_\delta(A)$ as in Definition 5.4, and either leaves the ϵ-neighborhood of A or never reaches A. This contradicts the local asymptotic stability of A. For details see the proof of Theorem 5.16 in Paper IV.

The next step is to show that $\{w_k\}_{k\geq 0}$ actually enters a compact set in the domain of attraction of $A = \{w : F(w) \leq c\}$ infinitely often, where $c = \liminf_{k\to\infty} F(w_k)$. Theorem 5.8 then implies that $\{w_k\}_{k\geq 0}$ converges to A.

Moreover, it holds that

$$\lim_{k\to\infty} F(w_k) = c, \tag{5.15}$$

since otherwise we would have some subsequence of $\{w_k\}_{k\geq 0}$ that converges to a lower functional value which is impossible by the choice of c.

From this it follows that $\{w_k\}_{k\geq 0}$ converges to the set of stationary points: Otherwise we can construct a solution $\tilde{W}(\cdot)$ to (5.1) such that $F(\tilde{W}(0)) = c$. This solution decreases along the paths of (5.1); therefore we can find a $t' > 0$ such that $F(\tilde{W}(0)) > F(\tilde{W}(t'))$. But $\tilde{W}$ takes values in the set of limit points of $\{w_k\}_{k\geq 0}$. Hence, we can find a subsequence $\{w_{m_k}\}_{k\geq 0}$ that converges to $F(\tilde{W}(t'))$, which contradicts (5.15).

If we further know that the equilibria of F are isolated, we can conclude that $\{w_k\}_{k\geq 0}$ converges to a unique equilibrium point of (5.1): From the assumption that $\sup_k \|w_k\|_2 < \infty$, along with the continuity of the gradient ∇F, we have that $\sup_k \|\nabla F(w_k)\|_2 < \infty$, almost surely. Thus

$$\|w_{k+1} - w_k\|_2 \leq \alpha_k \cdot \sup_k \|\nabla F(w_k)\|_2,$$

which tends to 0 for $\omega \in U$ where U is as in Theorem 5.8. From the assumption that $\sup_k \|w_k\|_2 < \infty$, it follows that for fixed $\omega \in U$, the limit set

$$L(\{w_k\}) = \{w : \exists \{w_{n_k}\}_{k\geq 0} \subset \{w_k\}_{k\geq 0} : \lim_{n_k\to\infty} w_{n_k} = w\}$$

is a connected set, compare [3]. But we know from the previous section that $L(\{w_k\}) \subset \{w : \nabla F(w) = 0\}$. Hence, if $\{w : \nabla F(w) = 0\}$ consists of isolated points, we must have that $L(\{w_k\}) = \{w_*\}$ for some equilibrium point w_*.

Remark. *Note that in the case that the assumptions of Theorem 5.9 in Section 5.2 below are fulfilled, we can shorten the proof to conclude that* (5.15) *holds, without making use of Theorem 5.8, since* (5.15) *follows from* (5.17) *in Theorem 5.9. This is for instance the case for* (5.2) *when the step size satisfies* (4.22)*, F is $L-$smooth, coercive and bounded below and the noise satisfies the assumptions in Chapter 4.4.3.*

5.2 The Robbins–Siegmund theorem

Several of the arguments in the previous section relies on the assumption that $\sup_k \|w_k\|_2 < \infty$ almost surely. To demonstrate this, one can make use of the Robbins–Siegmund theorem:

Theorem 5.9 (Robbins–Siegmund [58]). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$ be a sequence of sub-σ-algebras of $\mathcal{F}$. For each $k = 1, 2, \dots$ let V_k, β_k, X_k and Y_k be non-negative $\mathcal{F}_k$-measurable random variables such that*

$$\mathbb{E}\left[V_{k+1}|\mathcal{F}_k\right] \leq V_k(1+\beta_k) + X_k - Y_k. \tag{5.16}$$

Then

$$V = \lim_{k\to\infty} V_k \tag{5.17}$$

exists and is finite and $\sum_k Y_k < \infty$ on the set

$$\left\{\omega : \sum_k \beta_k < \infty, \sum_k X_k < \infty\right\}.$$

We can apply the theorem to the inequality (4.16) in Chapter 4.4.3:

$$\mathbb{E}_{\xi_k}\left[F(w_{k+1})\right] - F(w_k) \leq -\frac{\alpha_k}{2}\|\nabla F(w_k)\|_2^2 + \frac{LM\alpha_k^2}{2}. \tag{5.18}$$

taking $V_k = F(w_k) - F_*, \beta_k = 0, Y_k = \frac{\alpha_k}{2}\|\nabla F(w_k)\|_2^2$ and $X_k = \frac{LM\alpha_k^2}{2}$. Since by assumption $\sum_{k\geq 0}\alpha_k^2 < \infty$, we have that $\sum_{k\geq 0} X_k < \infty$ everywhere. We can thus conclude that $\lim_{k\to\infty} F(w_k)$ exists and satisfies

$$\lim_{k\to\infty} F(w_k) < \infty,$$

almost surely. If the objective function F is coercive, then since the sub-level sets are bounded, it holds that $\sup_{k\geq 1}\|w_k\|_2 < \infty$, almost surely. In paper IV, we derive a similar bound for the random variables $V_k = H(p_k, q_k) - F_* - \phi_*$ in order to apply Theorem 5.9.

5.3 Asymptotic pseudo-trajectories and chain recurrence

In this section we will look at an alternative approach to the one in the previous section. The approach in this section is due to Benaïm [6, 7] and requires the introduction of more technical terms, but also provides a more general form of convergence to a so-called *internally chain recurrent set*. However, we will see that under similar assumptions as in Section 5.1 it allows us to draw the same conclusions.

We will structure this section similar to Section 5.1 and divide it into three steps:

Step 1 Introduce the notion of *asymptotic pseudo-trajectories*.

An asymptotic pseudo-trajectory for an ODE is a function that asymptotically approximates solutions to the ODE; for any finite time interval and any tolerance level, there is a solution to the ODE which it tracks within that tolerance. We will see that we can extend the functions (5.10) to an asymptotic pseudo-trajectory of (5.1).

Step 2 Present the concept of *chain recurrence*.

Roughly, chain recurrence is a generalization of periodic points, but for which a certain margin of error is tolerated. Chain recurrence is intimately linked to the limit behavior of stochastic algorithms, and the next step is to

Step 3 Characterize the limits of stochastic algorithms in terms of chain recurrent sets.

We will look at a result due to Benaïm [6, 7], which states that certain stochastic algorithms converge to a so-called *internally chain recurrent set*. We will also touch upon a converse result from [8], which asserts that for any compact, connected *internally chain recurrent set*, there is a stochastic algorithm that has that set as its limit set. This means that it is not possible in general to say more than that the stochastic algorithms we are interested in, converge to a compact, connected, *internally chain recurrent set*. However, if there exists a Lyapunov function for the ODE, then we will see that one can obtain convergence to a stationary point.

Step 1

Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a Lipschitz-continuous function. The semiflow of f is the family of mappings $\phi : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$, defined by

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi(t, x_0) = f(\phi(t, x_0)),\ x_0 \in \mathbb{R}^d. \tag{5.19}$$

The mapping ϕ is continuous and satisfies the group property $\phi(0, x_0) = x_0$ and $\phi(s, \phi(t, x_0)) = \phi(s + t, x_0)$. We now introduce the notion of an *asymptotic pseudo-trajectory*:

Definition 5.10. *An asymptotic pseudo-trajectory for* (5.19) *is a function* $X : \mathbb{R} \to \mathbb{R}$ *such that*

$$\lim_{s\to\infty} \sup_{t\in[0,T]} \|X(t+s) - \phi(t, X(s))\| = 0 \tag{5.20}$$

holds for any $T > 0$.

The intuition behind Definition 5.10 is the following: given an interval $[0, T]$ we can make X track a solution to (5.19) on that interval which starts at $X(s)$ with arbitrary precision, if we choose s large enough.

Letting ϕ be the semiflow of (5.1), we see that for $T > 0$ and $t \in [0, T]$, we have by (5.1) and (5.12) that

$$\begin{aligned}\|W_k(t) - \phi(t, w_k)\|_2 &= \left\| \int_0^t \nabla F(W_k(s)) - \nabla F(\phi(s, w_k))\mathrm{d}s + \rho_k(t) + M_k(t) \right\|_2 \\ &\leq L \int_0^t \|W_k(s) - \phi(s, w_k)\|_2 \mathrm{d}s + \sup_{t\in[0,T]} \left(\|\rho_k(t)\|_2 + \|M_k(t)\|_2\right),\end{aligned}$$

where we have used the assumption that ∇F is Lipschitz continuous. Hence, making use of Grönwall's inequality (see e.g. Theorem 5.1 in [25]) we obtain

$$\|W_k(t) - \phi(t, w_k)\|_2 \leq \sup_{t\in[0,T]} \left(\|\rho_k(t)\|_2 + \|M_k(t)\|_2\right) e^{LT}.$$

Now,the fact that the functions ρ_k and M_k in (5.12) tends to 0 uniformly on compact intervals as k tends to infinity implies that

$$\lim_{k\to\infty} \sup_{t\in[0,T]} \|W_k(t) - \phi(t, w_k)\|_2 = 0. \tag{5.21}$$

This means that the entire sequence $\{W_k\}_{k\geq 0}$ converges to the set of solutions to (5.1) on every interval $[0, T]$. It also follows that W_0 is an asymptotic pseudo-trajectory to ϕ. This can be seen from the fact that

$$\begin{aligned}\sup_{t\in[0,T]} \|W_0(s+t) - \phi(t, W_0(s))\|_2 \leq &\sup_{t\in[0,T]} \|W_0(s+t) - W_0(t_{m(s)} + t)\|_2 \\ &+ \sup_{t\in[0,T]} \|W_0(t_{m(s)} + t) - \phi_t(W_0(t_{m(s)}))\|_2 \\ &+ \sup_{t\in[0,T]} \|\phi(t, W_0(t_{m(s)})) - \phi(t, W_0(s))\|_2.\end{aligned}$$

The second term on the right-hand side of the previous inequality tends to 0 as a consequence of (5.21). The first term also converges to 0, as a consequence of the

extended equicontinuity of $\{W_k\}_{k\geq 0}$ (making use of the fact that $W_0(t_{m(s)}+t) = W_{m(s)}(t)$) and that $s - t_{m(s)} \leq \alpha_{m(s)}$. For the same reason, the third term can be made arbitrarily small where we also make use of the continuity of ϕ.

The next theorem is an adaption of Lemma 4.2 in [6]. (See also Proposition 4.1 in [7]). It tells us that under certain conditions on the stochastic algorithm, its interpolated process is an asymptotic pseudo-trajectory:

Lemma 5.11 (Benaïm [6])**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be such that* (5.19) *has a unique solution for each initial condition x_0. Let $\{w_k\}_{k\geq 0}$ be a solution to the recursion*

$$w_{k+1} = w_k + \alpha_k f(w_k) + u_k + b_k. \tag{5.22}$$

Assume that

i) $\sup_k \|w_k\|_2 < \infty$,

ii) $\lim_{k\to\infty} b_k = 0$, *and*

iii) for each $T > 0$, it holds that

$$\lim_{k\to\infty} \left\| \sum_{i=k}^{m(t_k+T)-1} \alpha_i u_i \right\|_2 = 0.$$

Let $t_k = \sum_{i=0}^{k-1} \alpha_i$ and $t_0 = 0$. Then, the interpolated process

$$W_0(t) = \begin{cases} w_k, & t_k \leq t < t_{k+1}, \\ w_0, & t < t_0, \end{cases} \tag{5.23}$$

is an asymptotic pseudo-trajectory for (5.19) *whose image has compact closure.*

Remark. *Note that the statement of Lemma 5.11 is deterministic but it also translates to the stochastic setting: If Assumptions i) to iii) holds almost surely, then for almost all $\omega \in \Omega$, $W_0(t)$ is an asymptotic pseudo-trajectory for* (5.19)*.*

Remark. *Assumptions i) to iii) of Lemma 5.11 are known as the* Kushner & Clark *assumptions, compare [6]. In the setting of the previous section when $\{w_k\}_{k\geq 0}$ is generated by* (5.2)*, we would have $b_k = 0$ and $u_k = \nabla F(w_k) - \nabla f(w_k, \xi_k)$ for all k. Assumption iii) corresponds to the sequence $\{M_k\}_{k\geq 0}$ given by* (5.9) *converging uniformly on compact sets to* 0*.*

The proof can be found in [6, 7] for piecewise linear interpolation processes. For piecewise constant interpolations, we essentially perform the same calculations as we did in Section 5.1 and then apply Grönwall's lemma as we did above. From the fact that $\sup_{k\in\mathbb{N}} \|w_k\|_2 < \infty$, it follows that the image of W_0 has compact closure.

Step 2

The limit set of the interpolated process W_0 from (5.7) is given by

$$L(W_0) = \{x \in \mathbb{R}^d : \exists\ \{a_k\}_{k\geq 0},\ \text{s.t. } a_k \to \infty \text{ and } \lim_{k\to\infty} W_0(a_k) = x\}. \quad (5.24)$$

In the case that W_0 is a piecewise constant interpolation of a sequence $\{w_k\}_{k\geq 0}$ as above, it holds that $L(W_0) = L(\{w_k\})$, where $L(\{w_k\})$ is the set of limit points of $\{w_k\}_{k\geq 0}$.

Therefore, our next goal is to characterize $L(W_0)$ for asymptotic pseudotrajectories.

In order to do this, we will now introduce the concept of *chain recurrence*; an idea that was originally introduced in [19]. One can think of a chain recurrent point as one that would be taken to be periodic if one allows for arbitrarily small measurement errors:

Definition 5.12. *An (ϵ, T)-pseudo orbit from x to y is a sequence of points $\{x_i\}_{i=0}^n$ together with time points $t_i \geq T$ such that $x_0 = x$, $x_n = y$ and*

$$\|\phi(t_i, x_i) - x_{i+1}\| < \epsilon.$$

If there for every $\epsilon > 0$ and $T > 0$ exists a (ϵ, T)-pseudo orbit from x to itself, x is said to be chain recurrent *for ϕ. The set of chain recurrent points for ϕ is denoted by $R(\phi)$. A non-empty, compact invariant set Λ is said to be* internally chain recurrent *if the semiflow ϕ restricted to Λ satisfies $R(\phi|_\Lambda) = \Lambda$.*

Remark. *The difference between a chain recurrent set and an internally chain recurrent set is that the chains (i.e. the points $\{x_i\}_{i=0}^n$ in Definition 5.12) has to lie inside the set itself for an internally chain recurrent set.*

The set of $\omega-$limit points of ϕ, denoted $L(\phi)$, is defined by

$$L(\phi) = \{y \in \Gamma : \exists x_0, \{t_n\}_{n\geq 0} : t_n \to \infty, \phi(t_n, x_0) \to y\}.$$

If $\{\phi(t, x_0) : t \geq 0\}$, for $x_0 \in \mathbb{R}^d$, has compact closure[1], it holds that

$$L(\phi) \subset R(\phi), \quad (5.25)$$

compare e.g. Proposition 1.5 in [51]. By Theorem 15.0.3 [68] the set $L(\phi)$ then consists of equilibria for equation (5.1).[2] Likewise, all periodic points are chain recurrent since these are contained in $L(\phi)$, see [6]. However, in Figure 5.2 we see an example of a chain recurrent set containing points that are not $\omega-$limit points. (As well as a chain recurrent set which is not internally chain recurrent).

[1] This is true for the solutions to (5.1) if F is coercive. See e.g. Theorem 2 in [72].

[2] This also holds for the equation considered in Paper IV since the Hamiltonian (6) is decreasing along the paths of (7).

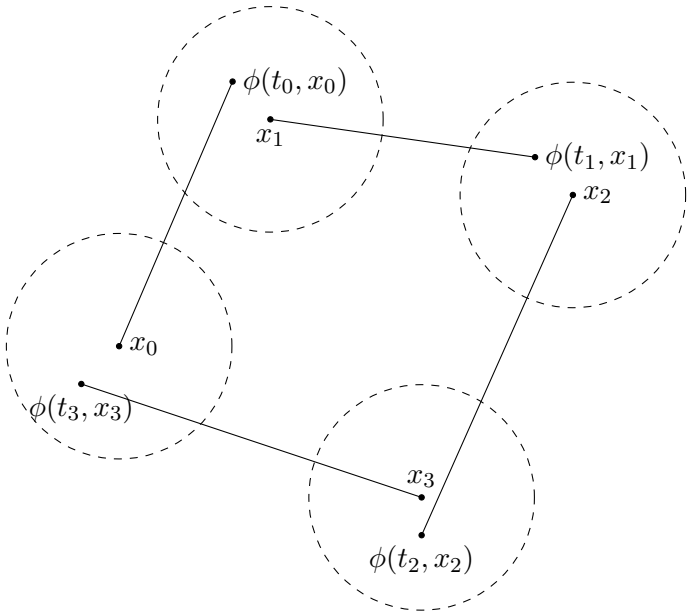

Figure 5.1: Illustration of an (ϵ, T)-pseudo orbit with 4 points: We can find points $\{x_i\}_{i=0}^4$ where $x_4 = x_0$ and time points $\{t_i\}_{i=0}^3$ with $t_i \geq T, i = 0, \ldots, 3$ such that $\|\phi(t_i, x_i) - x_{i+1}\| < \epsilon$ for $i = 0, \ldots, 3$.

Step 3

The next theorem is an adaption of (a part of) Corollary 4.3 in [6]. It characterizes the limit set of an asymptotic pseudo-trajectory in terms of internally chain recurrent sets:

Lemma 5.13 (Benaïm [6])**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be such that* (5.19) *has a unique solution for each initial condition x_0, and let ϕ be the semiflow generated by f. If X is an asymptotic pseudo-trajectory for ϕ whose image has compact closure, then $L(X)$ is a compact and internally chain recurrent set for ϕ.*

Using Lemma 5.11 and 5.13 in tandem we get that:

Theorem 5.14. *Let $\{w_k\}_{k\geq 0}$ be a sequence that satisfies assumptions i) – iii) of Lemma 5.11. Then its limit set $L(\{w_k\}_{k\geq 0})$ is a compact, connected and internally chain recurrent set.*

As noted in the end of Section 5.1, connectedness follows from Assumption i)–iii) of Lemma 5.11. For a proof see Corollary 4.3 in [6]. A converse result to Lemma 5.11 and Theorem 5.13 is the following:

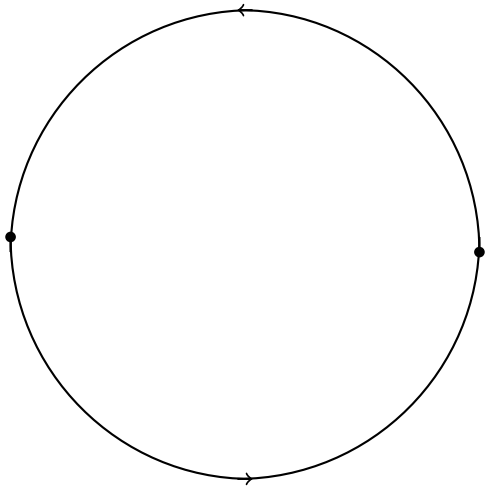

Figure 5.2: Consider $\dot{\theta} = \sin^2(\theta)$, for $\theta \in S^1 = \mathbb{R}\backslash 2\pi\mathbb{Z}$. The points 0 and π are fixed points and periodic points. The entire circle is chain recurrent for the semiflow generated by the equation and the sets $\{0\}, \{\pi\}$ and S^1 are internally chain recurrent. The set $[0, \pi]$ consists of chain recurrent points but is not an internally chain recurrent set since the chains are not contained in it. Compare e.g. [2, 6].

Theorem 5.15 (Benaim & Hirsch [8])**.** *Let L be a compact, connected set which is internally chain recurrent for the semiflow induced by f (where f is as in Theorem 5.11). Then there exists sequences $\{w_k\}_{k\geq 0}$, $\{b_k\}_{k\geq 0}$ and $\{u_k\}_{k\geq 0}$ such that $\{w_k\}_{k\geq 0}$ is a solution to* (5.22) *of Theorem 5.11, and $\{w_k\}_{k\geq 0}$, $\{b_k\}_{k\geq 0}$ and $\{u_k\}_{k\geq 0}$ satisfies i), ii) and iii) respectively and the set of limit points of $\{w_k\}_{k\geq 0}$ is equal to L.*

The previous theorem implies that the result in Theorem 5.14 is tight in the following sense: under the Kushner & Clark assumptions, one cannot in general be more specific on the set of limit points of the sequence, than that it is an internally chain recurrent, compact and connected set.

The conclusion of Theorem 5.13 is very general in that it holds for all algorithms satisfying the Kushner & Clark assumptions. If more structure is imposed on the system, we can however be more specific regarding the set of limit points of the algorithm:

Definition 5.16 (Lyapunov function)**.** *Let ϕ be a semiflow on a compact metric space X and let $\Lambda \subset X$ be an invariant set. A function $V : X \to \mathbb{R}$ is called a Lyapunov function for Λ if the function $t \in \mathbb{R}_+ \mapsto V(\phi_t(x))$ is constant for $x \in \Lambda$ and strictly decreasing for $x \notin \Lambda$. If Λ equals the set of equilibria, V is called a* strict Lyapunov function*.*

For (5.1), the objective function F is a strict Lyapunov function for $\Lambda = \{w :$

$\nabla F(w) = 0\}$. For the algorithm in Paper IV, the Hamiltonian $H(p, q) = F(q) + \varphi(p)$ is a strict Lyapunov function.

We can now appeal to the two following results from [6]:

Lemma 5.17 (Proposition 3.2 [6]). *Let ϕ be a semiflow on a compact metric space X and let $\Lambda \subset X$ be a compact invariant set. Let $V : X \to \mathbb{R}$ be a strict Lyapunov function for Λ and let the cardinality of $V(\Lambda)$ be finite. Then*

$$\mathcal{R}(\phi) \subset \Lambda.$$

A more general version of this theorem appears in [7] (Proposition 6.4) in which the assumption that $V(\Lambda)$ is finite is relaxed to $V(\Lambda)$ having empty interior.

Corollary 5.18 (Corollary 3.3 [6]). *Let Λ denote the set of equilibria of* (5.1) *and suppose that these are isolated. Let $\{w_k\}_{k\geq 0}$ satisfy the assumptions of Lemma 5.11. Then $\{w_k\}_{k\geq 0}$ converges towards an equilibrium.*

As noted above, it holds that $L(\{w_k\}_{k\geq 0}) = L(W_0)$. Let E denote the set of equilibria of ϕ. Under the assumptions of Lemma 5.11, we have by Theorem 5.14 that $L(W_0)$ is a compact, connected and internally chain recurrent set. We can then apply Lemma 5.17 to the semiflow restricted to $L(W_0)$. Since $L(W_0)$ is internally chain recurrent it holds that

$$L(W_0) = R(\phi|_{L(W_0)}) \supset L(\phi|_{L(W_0)})$$

by (5.25). As noted above, the set $L(\phi|_{L(W_0)})$ consists of equilibria (by e.g. Theorem 15.0.3 in [68]). This implies that the set $E \cap L(W_0)$ is non empty. Thus, since $L(W_0)$ is compact, we can apply Theorem 5.14 to $\phi|_{L(W_0)}$ with $\Lambda = E \cap L(W_0)$ and conclude that $\Lambda \supset R(\phi|_{L(W_0)}) = L(W_0)$. Since $L(W_0)$ is connected (from the fact that $L(W_0) = L(\{w_k\})$) and Λ consists of isolated points, $L(W_0)$ (and thus $L(\{w_k\}_{k\geq 0})$) is an equilibrium.

In this section, we have seen that the approach with asymptotic-pseudo trajectories provides an alternative way to demonstrate convergence of stochastic optimization algorithms. The convergence to an internally chain recurrent set that can be deduced from Theorem 5.14 is weaker (but also under weaker assumptions) than the convergence that we obtained in the previous section. However, in the case that a strict Lyapunov function exists, we saw that we could draw the same conclusion. We have also seen that Theorem 5.15 justifies the notion of chain recurrence for characterizing the limit sets of stochastic algorithms.

6. Research

In this chapter, we summarize the results from Paper I–IV and link them to the concepts introduced in the previous chapters of the thesis. We discuss the implications of the research and touch on some possible paths for future studies.

The inspiration for the project was the idea that the gradient descent algorithm

$$w_{k+1} = w_k - \alpha_k \nabla F(w_k),$$

can be viewed as an explicit Euler discretization of the gradient flow equation

$$w'(t) = -\nabla F(w(t)). \tag{6.1}$$

As in Section 3.4, this equation can be linearized around an equilibrium solution, and we can consider the system

$$w'(t) = -H_F(w_*)w(t).$$

From the discussion in Section 3.4, we saw that it is reasonable to expect algorithms that have good stability properties for the linear test equation to perform well on the original problem, at least for strongly convex functions, whose Hessians are positive definite or around a local minimum where it has positive eigenvalues.

This is the view-point that is adopted in Paper I and Paper II, where we investigate the behavior of classical time-stepping methods in the context of stochastic optimization. In Paper III we analyze a generalization of *tamed-Euler method*, which is used to integrate stochastic differential equations. Unlike the so-called *clipping*-methods, which are frequently used to stabilize optimization algorithms within the machine learning community. Another popular practice in the field is the custom to use *momentum*: the introduction of an additional momentum variable, in which the average of the past gradients accumulates. This also tends to have a stabilizing effect on the algorithm. In Paper IV, we study a particular form of clipping methods with momentum, and show that it can be viewed as a discretization of a certain Hamiltonian system.

6.1 Paper I

In Chapter 4.4.2, we introduce the proximal point method and note that it can be viewed as the implicit Euler scheme from Section 3.2 in Chapter 3, applied to the equation (6.1). In Paper I, we show convergence for a stochastic proximal point method for convex functions. The analysis in Chapter 3 and 4 was done on $\mathbb{R}^d$ for simplicity, but in Paper I, the analysis is performed in a general Hilbert space setting. Let H be a real Hilbert space and $F : H \to \mathbb{R}$ a strongly convex function. We are then interested in finding the unique solution w_* to the problem

$$w_* = \arg\min_{w \in H} F(w). \tag{6.2}$$

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{\xi_k\}_{k\geq 1}$ be a sequence of jointly independent random variables on Ω.

The stochastic proximal point method seeks to approximate the solution to the problem (6.2) by producing a sequence of iterates $\{w_k\}_{k\geq 1}$ according to the update rule

$$w_{k+1} = w_k - \alpha_k \nabla f(w_{k+1}, \xi_k), \tag{6.3}$$

where $\{\alpha_k\}_{k\geq 1}$ is a step size sequence, i.e. $\alpha_k > 0$ for every $k \geq 0$. In Paper I we assume that the random functions $f(\cdot, \xi)$ are unbiased estimates of $F(\cdot)$, i.e. that $\mathbb{E}_\xi [f(w, \xi)] = F(w)$. Although the stochastic proximal point algorithm is not new, it has not been analyzed in the infinite-dimensional framework to a large degree before. A notable exception to this is [10] where a weak type of convergence for maximal monotone operators is proved in a general setting. Another example is [59] where the authors demonstrate norm convergence at a rate, albeit with a rather strong global Lipschitz condition on the objective function. Under the assumption that the gradient of $f(\cdot, \xi)$ satisfies a local Lipschitz condition, and that it is μ_ξ-strongly convex for a positive random variable μ_ξ (see Paper I for details), we get sublinear convergence in expectation to the solution, i.e.

$$\mathbb{E}\left[\|w_k - w_*\|_H^2\right] \leq \frac{C}{k},$$

for some constant C and where w_* is defined by (6.2). The research in Paper I generalizes that in [60] and extends it to an infinite-dimensional setting. In several cases, a closed-form solution of (6.3), to obtain w_{k+1}, can be found, and then the stochastic proximal method provides a more stable alternative to SGD, at essentially the same computational cost, see [24, Sec. 5].

6.2 Paper II

Although the proximal point method has very good stability properties, it can be computationally costly to compute the implicit update (6.2) in the cases when there is no closed-form solution at hand. An alternative in these cases is to use explicit methods with larger stability regions. As noted in Chapter 3.3.4, an example class of methods that are optimal in the sense that they maximize the *real stability boundary*, are Runge–Kutta–Chebyshev methods. Although well-known in the time-stepping community, the utility of these methods for solving optimization problems have not been extensively studied. A notable exception is [23], in which a deterministic optimization method that is based on Runge–Kutta–Chebyshev methods is proposed.

In Paper II, we propose a stochastic optimization algorithm –the *Stochastic Runge–Kutta–Chebyshev descent* method (abbreviated as *SRKCD*)– based on the Runge–Kutta Chebyshev methods introduced in Section 3.4 of Chapter 3 for approximating the solution to (6.2). The analysis is performed in a finite-dimensional setting on $\mathbb{R}^d$ for simplicity. It can likely be extended to the infinite-dimensional setting in the framework of monotone operators as in [37]. We obtain convergence guarantees in expectation at a sublinear rate, see Theorem 2.6 in Paper II. Under slightly stricter regularity assumptions, we obtain convergence in expectation to a stationary point, see Theorem 2.10 in Paper II. Although not explicitly stated in the article, we obtain convergence at a rate for the sequence $\{\min_{1\leq k\leq K} \mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right]\}_{K\geq 1}$, i.e.

$$\min_{1\leq k\leq K} \mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] = \mathcal{O}\left(\frac{1}{\log(K)}\right), \tag{6.4}$$

in the non-convex case similar to Theorem 4 of paper III. This follows from (4.19) in Section 4.3, (2.10) in Theorem 2.8 in Paper II, along with the fact that

$$A_K = \sum_{k=1}^{K} \frac{\beta}{k+\gamma} \geq \int_1^{K+1} \frac{\beta}{x+\gamma} dx = \beta\left(\log(K+1+\gamma) - \log(1+\gamma)\right). \tag{6.5}$$

The argument is essentially the same as that in Chapter 4.4.3 or in the proof of Theorem 4 in paper III and is therefore omitted here.

It is also worth remarking that although we prove convergence in expectation in Theorem 2.1 and Theorem 2.6 in Paper II, a standard result in probability theory states that this implies convergence in probability, compare [18, Prop. 3.1.5]. Thus, we can for example use (6.4), to say that

$$\mathbb{P}\Big(\Big\{\omega : \min_{1\leq k\leq K} \|\nabla F(w_k)\|_2^2 > \varepsilon\Big\}\Big) = \mathcal{O}\left(\frac{1}{\log(K)}\right). \tag{6.6}$$

Note however, that the error constant inversely proportional to ε, see [18, Prop. 3.1.5]. From (6.6) we can also conclude that the sequence

$$\{\min_{1\leq k\leq K}\|\nabla F(w_k)\|_2^2\}_{K\geq 1} \tag{6.7}$$

converges almost surely to 0. The argument is the same as in the proof of Corollary 7 of Paper III. This result is less strong than the one obtained in Paper IV, where we obtain almost sure convergence of the entire sequence $\{w_k\}_{k\geq 0}$.

Methods with large stability region are particularly useful for *stiff* problems. See [62] for a discussion of this. For a convex quadratic optimization problem, this essentially corresponds to the Hessian having one very large eigenvalue, that puts a severe step-size restriction on the gradient update. In Paper II, we saw that the use of SRKCD allowed for a much larger step size than SGD for such problems.

6.3 Paper III

The main inspiration for Paper III was the *tamed Euler* method, introduced in [39]. This is a scheme that is used to obtain convergence for certain stochastic differential equations where the explicit Euler scheme is known to fail. It is similar to the so-called *clipping*-methods: if the norm of an iterate exceeds a predefined threshold, the iterate is rescaled to prevent the method from exploding. This concept was studied as early as in 1967 by Poljak, see [55], but was made popular in the machine learning community by Mikolov in 2013 in the context of large language models, compare [50].

Another widespread practice is the usage of *component-wise rescaling* of the gradient, which was introduced in the Adam paper [42]. The idea in Paper III was to investigate a component-wise version of the tamed-Euler scheme as a stochastic optimization method. During preliminary investigations for the paper, we realized that the rescaling function for the tamed-Euler scheme exhibits similar behavior to the arctan function. This spurred the idea of considering other nonlinear gradient clipping functions. One could loosely define a "clipping-function" as a function that behaves like the identity close to 0, but is "sufficiently" bounded below and above at $\pm\infty$. This is the behavior that we aimed to capture with Assumption 1 in Paper III. The analysis also allows for clipping functions with "slow" growth at $\pm\infty$ or functions that are bounded. See Figure 6.1 for an illustration.

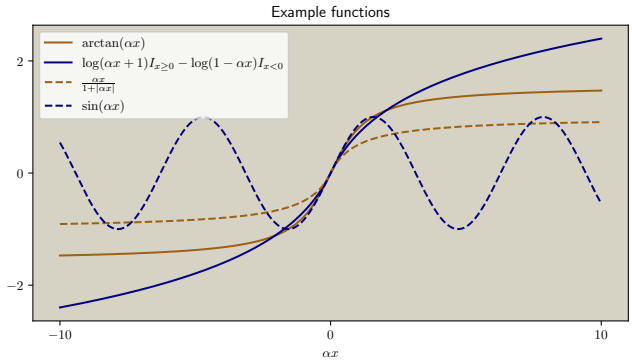

Figure 6.1: Examples of some clipping functions that are covered by the analysis in Paper III.

The scheme analyzed in Paper III is given by

$$w_{k+1} = w_k - \alpha_k G(\nabla f(w_k, \xi_k), \alpha_k) \tag{6.8}$$

where G is an operator that applies a clipping function component-wise to the stochastic gradient $\nabla f(w_k, \xi_k)$. The assumptions about the noise in Paper III differ somewhat from those typically encountered. Assumption 4 reads: *There exists* $w_* \in \arg\min F(w)$ *such that*

$$\mathbb{E}\left[\|\nabla f(w_*, \xi)\|^2\right] \leq \sigma^2.$$

The assumption is that the variance is bounded at *some* stationary point of F. In addition to this we posit that there exists $w_* \in \arg\min F(w)$ such that

$$\mathbb{E}\left[\|w_k - w_*\|_2^3\right] \leq M, \ \forall k. \tag{6.9}$$

This stability assumption is relatively strong, but it is in some sense analogous to an assumption of having bounded stochastic gradients or a step-size restriction and was necessary to deal with the nonlinearity of the clipping functions. This is discussed in more detail in Appendix A of Paper III. If a (rather restrictive) step-size restriction is imposed, it is also possible to show convergence of the scheme (6.8) under the assumption that

$$\mathbb{E}\left[\|\nabla f(w, \xi)\|_2^3\right] \leq M_1 + M_2 \|\nabla F(w)\|_2^3, \tag{6.10}$$

where $M_1, M_2 \geq 1$.

We also note that the $\|\cdot\|_2^3-$terms that appear in (6.9) and (6.10) are due to the fact that (6.8) can be rewritten as a second-order perturbation of SGD, where the the perturbation term is of the order $\mathcal{O}\left(\|\nabla f(w_k, \xi_k)\|_2^3\right)$.

Another thing to touch upon is Assumption 5 of Paper III, which in the literature is known as an *interpolation assumption*; it states that there exists a minimum w_* of F, which is simultaneously a minimum of all the stochastic functions $f(\cdot, \xi)$, almost surely. Over-parametrized machine learning models frequently have the capacity to interpolate the data and achieve 0 loss on training data. Empirical evidence suggests that the interpolation assumption is satisfied for such models, compare [48, 74]. In Paper III, we make use of Assumption 5 in order to obtain a smaller error constant in the bound in Theorem 4.

In section 6 of Paper III, the algorithms are tested on some classical machine learning tasks such as image recognition and text prediction. We see that most of the schemes exhibit performance on par with state-of-the-art algorithms such as Adam and SGD with momentum. The analysis in Paper III opens up for the usage of a large class of non-linear clipping functions for stochastic optimization algorithms.

6.4 Paper IV

In Paper IV, we continue to explore the realm of clipped stochastic optimization algorithms. We now also consider clipped momentum algorithms. Momentum-based optimization algorithms were first considered in the deterministic case in [56]. The main idea is to make use of a weighted average of all the past gradients instead of just the gradient. There are many different formulations of SGD with momentum and the particular formulation that we consider is on the form

$$\begin{aligned} p_{k+1} &= p_k - \alpha_k \nabla f(q_k, \xi_k) - \alpha_k \gamma p_k \\ q_{k+1} &= q_k + \alpha_k p_{k+1}. \end{aligned} \tag{6.11}$$

One can view this as a stochastic implicit-explicit discretization of the ODE

$$\begin{aligned} \dot{p} &= -\nabla F(q) - \gamma p, \\ \dot{q} &= p. \end{aligned}$$

Introducing a separable Hamiltonian $H(p, q) = F(q) + \frac{\|p\|_2^2}{2}$, one finds that the previous system is a dissipative Hamiltonian system:

$$\begin{aligned} \dot{p} &= -\nabla_q H(p, q) - \gamma \nabla_p H(p, q), \\ \dot{q} &= \nabla_p H(p, q). \end{aligned} \tag{6.12}$$

If we now generalize this to $H(p, q) = F(q) + \varphi(p)$, we can get other schemes that are interesting. A case of special interest is

$$\varphi(p) = \sqrt{\epsilon + \|p\|_2^2},$$

where $\epsilon > 0$. Since $\nabla\varphi(p) = \frac{p}{\sqrt{\epsilon+\|p\|_2^2}}$, the corresponding discretization gives us a version of normalized SGD with momentum for small ϵ. For larger values of ϵ (e.g. $\epsilon = 1$), the scheme behaves like a soft-clipping algorithm. Similar formulations have been previously considered in the deterministic case in e.g. [29, 30]. The algorithm we consider in Paper IV is therefore

$$\begin{aligned} p_{k+1} &= p_k - \alpha_k \nabla f(q_k, \xi_k) - \alpha_k \gamma \nabla\varphi(p_k) \\ q_{k+1} &= q_k + \alpha_k \nabla\varphi(p_{k+1}). \end{aligned} \tag{6.13}$$

The analysis is based on the ODE method which is illustrated in Section 5.1. The analysis of the algorithm in Paper IV is different than that in [44] since the algorithm in Paper IV is an explicit-implicit discretization of (6.12) and this compels us to demonstrate that the difference between the explicit and the implicit discretization

$$\kappa_k(t) = \int_0^t \nabla\varphi(P_{k+1}(s)) - \nabla\varphi(P_k(s))\mathrm{d}s$$

converges to 0 uniformly on compact intervals. The assumption on the noise in [44] is also different; translated to the algorithm in Paper IV, it would correspond to requiring that

$$\sup_k \mathbb{E}\left[\|\nabla f(q_k, \xi_k)\|_2^2\right] < \infty.$$

We also show that the iterates generated by (6.13) are finite almost surely in three different settings that are adapted to the stochastic optimization setting. Another merit of Paper IV is its rigorous proof, within the specific setting, of the results cited in [44], where details are scarce.

We first show that the algorithm converges to a stationary point under the assumption that the iterates $\{p_k\}_{k\geq 0}$ and $\{q_k\}_{k\geq 0}$ are almost surely bounded. This is Theorem 4.14 in Paper IV. We then demonstrate that the latter holds if the objective function is coercive. This is Theorem 4.15.

The analysis in Paper IV is done under three different sets of assumptions. In the first, we assume that the gradient is Lipschitz continuous and that the stochastic gradients satisfy

$$\mathbb{E}\left[\|\nabla f(q, \xi) - \nabla F(q)\|_2^2\right] \leq \sigma^2 + \kappa\left(F(q) - F_*\right) + \tau\|\nabla F(q)\|_2^2.$$

Since the objective function is assumed to be coercive, this allows for infinite variance in the case that the iterates escape to infinity or when the gradient

blows up. In the second setting, we weaken the regularity assumption on the gradient to the so-called (L_0, L_1)-smoothness assumption introduced in [75]:

$$\|\nabla F(x) - \nabla F(y)\|_2 \leq (L_0 + L_1 \|\nabla F(y)\|_2) \, \|x - y\|_2,$$

whenever $\|x - y\|_2 \leq \frac{1}{L_1}$. In this setting, we assume that the noise satisfies

$$\mathbb{E}\left[\|\nabla f(q, \xi) - \nabla F(q)\|_2^2\right] \leq \sigma^2.$$

Empirical evidence suggests that the stochastic gradient noise for SGD may be more heavy-tailed in some cases, compare [34]. To account for such noise, we also analyze the algorithm defined by (6.13) in the *empirical risk minimization* setting. More precisely, we assume that the objective function is of the form

$$F(q) = \frac{1}{N} \sum_{i=1}^{N} f_i(q),$$

for some loss functions f_i that are bounded below and that the stochastic gradients can be written as

$$\nabla f(q, \xi) = \frac{1}{|B_\xi|} \sum_{i \in B_\xi} \nabla f_i(q),$$

where $B_\xi \subset \{1, \ldots, N\}$ and $|B_\xi|$ is the cardinality of B_ξ. We further assume that the stochastic functions $f(\cdot, \xi)$ are $(L_0, L_1)-$smooth. In this setting, we can prove convergence for noise that is merely bounded in expectation:

$$\mathbb{E}\left[\|\nabla f(q, \xi) - \nabla F(q)\|_2\right] \leq \sigma.$$

The assumptions on the noise are relatively weak compared to other results analyzing stochastic algorithms in the (L_0, L_1)-smoothness setting. A common assumption is that the noise is almost surely bounded, see [20, 46, 73, 75]. An exception to this is [67], which analyzes AdaGrad for (L_0, L_1)-smooth functions under the *affine variance* assumption (Assumption 4.iii of Paper II). The assumption we pose on the noise in third setting of Paper IV is strictly weaker than the affine variance assumption, since it allows for heavy-tailed noise. Some studies suggest that this may be a problem for certain machine learning models, compare [34, 76] as well as [53].

At last, a remark on the existence of solutions to (6.12) is in order. To use the ODE method we need to know that the solutions to (6.12) exists for all future time. This is less obvious for an objective function that is $(L_0, L_1)-$smooth, but it follows from the fact that there exists a Lyapunov function which is coercive:

Take $z = (p, q)$ and put $V(z) = H(p, q) - F_* - \varphi_*$. Since V is decreasing along the paths of V we have that

$$V(z(t)) \leq V(z(0)).$$

Suppose that there exists a $T > 0$ such that $\lim_{t \to T^-} \|z(t)\| = \infty$. Then the fact that V is coercive implies that $\lim_{t \to T^-} V(z(t)) = \infty$, which is a contradiction. Hence, the solutions to (6.12) exist for all future time and are bounded. Compare with Theorem 2 in [72].

6.5 Outlook

The non-convex results in Paper II– IV only guarantees convergence to a stationary point of the objective function F. Without more knowledge, this could be a saddle point or a maximum. Under the assumption that

$$\mathbb{E}\left[\langle \nabla F(w) - \nabla f(w, \xi), v \rangle_+\right] \geq c, \tag{6.14}$$

for some constant $c > 0$ and all unit vectors v, one can show that SGD and Adam, under suitable assumptions converges almost surely to a local minimum, compare [4, 31, 49]. Here $(\cdot)_+ = \max\{x, 0\}$. This type of analysis dates back to [54] and [15] (see also [7, 9]). The assumption (6.14) essentially means that the algorithm has noise in all directions and will get "pushed out" of a stable manifold, see [7, 54]. The following example from [54] illustrates what happens if (6.14) fails to hold: Consider the system

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

The origin is an unstable equilibrium for this equation, but for any initial condition $(x_0, 0)$, the solutions will converge to the origin along the stable manifold. If we take X to be a standard Gaussian random variable and consider a stochastic algorithm which has the noise $\alpha_k X$ in the $x-$direction (but no noise in the $y-$direction) and starts at $(x_0, 0)$, this algorithm will converge to the origin as well.

A similar convergence result for a version of SGD with momentum was obtained in [31], and this can likely be extended to the algorithms in Paper IV, as they behave similarly. An analysis of this kind could be complemented by a convergence rate analysis in the region around local minima, as in [49].

Another possible direction for future research is to extend the results in Paper IV to demonstrate convergence in the constant step size case. Consider the algorithm

$$w_{k+1}^{\alpha} = w_k^{\alpha} - \alpha \nabla f(w_k^{\alpha}, \xi_k),$$

where the superscript α is to denote the dependence on the constant step-size α. Introducing the piecewise constant interpolation process

$$W_0^{\alpha}(t) = \begin{cases} w_k^{\alpha}, \ k\alpha \leq t < (k+1)\alpha, \\ w_0^{\alpha}, \ t \leq 0. \end{cases}$$

we may as in Section 5.1, study the time-shifted process $W_0^{\alpha}(k_\alpha \alpha + t)$, where $\{k_\alpha\}$ is a sequence of integers that tends to ∞ as $\alpha \to 0$. Theorem 2.1 of Section 8.2 in [44] provides an analogue of Theorem 5.8 of Section 5.1: Loosely speaking, it holds under suitable assumptions that every subsequence of $\{W_0^{\alpha}(k_\alpha \alpha + t)\}_{k_\alpha}$ has a further subsequence that converges in distribution to a solution to the ODE (5.1), as $\alpha \to 0$. It is also possible to obtain convergence guarantees in probability of the type

$$\limsup_{k \to \infty} \mathbb{P}\left(d(w_k^{\alpha}, S) > \epsilon\right) = 0, \text{ as } \alpha \to 0,$$

see [11] in which a projected version of SGD is analyzed.

Other conceivable paths for future research are to extend the algorithms in this thesis to account for correlated noise and functions with non-differentiable gradients. For non-convex functions, the *generalized subgradient* defined by

$$\partial f(x) = \{y : f^0(x; v) \geq \langle y, v \rangle \text{ for all } v \in \mathbb{R}^d\}, \tag{6.15}$$

where $f^0(x; v)$ is

$$f^0(x; v) = \limsup_{y \to x, t \to 0^+} \frac{f(y + tv) - f(y)}{t},$$

the generalized directional derivative of f at x in the direction of v, compare [17]. This was for instance made use of in [11] to analyze SGD for non-differentiable functions.

References

[1] C. Aliprantis and K. Border. *Infinite dimensional analysis; a hitchhiker's guide*. Springer Berlin, 1994.

[2] J. M. Alongi and G. S. Nelson. *Recurrence and Topology*, volume 85. American Mathematical Society, 2007.

[3] M. D. Asic and D. D. Adamovic. Limit points of sequences in metric spaces. *Am. Math. Mon.*, 77(6):613–616, 1970.

[4] A. Barakat, P. Bianchi, W. Hachem, and S. Schechtman. Stochastic optimization with momentum: Convergence, fluctuations, and traps avoidance. *Electronic Journal of Statistics*, 15(2), 2021.

[5] H. H. Bauschke and P.L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2 edition, 2017.

[6] M. Benaïm. A dynamical system approach to stochastic approximations. 34(2), 1996.

[7] M. Benaïm. Dynamics of stochastic approximation algorithms. Séminaire de probabilités de Strasbourg. 33, 1999.

[8] M. Benaïm and M. Hirsch. Asymptotic pseudotrajectories and chain recurrent flows, with applications. *Journal of Dynamics and Differential Equations*, 8, 12 1996.

[9] M. Benaïm and M. W. Hirsch. Dynamics of morse-smale urn processes. *Ergodic Theory and Dynamical Systems*, 15(6), 1995.

[10] P. Bianchi. Ergodic convergence of a stochastic proximal point algorithm. *SIAM J. Optim.*, 26(4):2235–2260, 2016.

[11] P. Bianchi, W. Hachem, and S. Schechtman. Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. *Set-Valued and Variational Analysis*, 30(3), 2022.

[12] P. Billingsley. *Convergence of Probability Measures*. Wiley, 2 edition, 1968.

[13] V. S. Borkar. *Stochastic Approximation; A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[14] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[15] O. Brandière and M. Duflo. Les algorithmes stochastiques contournent-ils les pièges? *Annales de l'I.H.P. Probabilités et statistiques*, 32(3), 1996.

[16] G. Casella and R. Berger. *Statistical Inference*. Duxbury Resource Center, 2001.

[17] F.H. Clarke, Yu.S. Ledyaev, R.J. Stern, and P.R. Wolenski. *Nonsmooth Analysis and Control Theory*. Graduate texts in mathematics. Springer, 1991.

[18] D.L. Cohn. *Measure Theory, Second Edition*. Springer, 2013.

[19] C.C. Conley. *Isolated Invariant Sets and the Morse Index*. Conference Board of the Mathematical Sciences, 1978.

[20] M. Crawshaw, M. Liu, F. Orabona, W. Zhang, and Z. Zhuang. Robustness to unbounded smoothness of generalized SignSGD. *Advances in neural information processing systems*, 35, 2022.

[21] H. Daneshmand, J. Kohler, A. Lucchi, and T. Hofmann. Escaping saddles with stochastic gradients. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1155–1164. PMLR, 2018.

[22] J. Droniou and R. Eymard. Uniform-in-time convergence of numerical methods for non-linear degenerate parabolic equations. *Numerische Mathematik*, 132, 2016.

[23] A. Eftekhari, B. Vandereycken, G. Vilmart, and K. C. Zygalakis. Explicit stabilised gradient descent for faster strongly convex optimisation. *BIT Numerical Mathematics*, 61:119–139, 2021.

[24] M. Eisenmann, T. Stillfjord, and M. Williamson. Sub-linear convergence of a stochastic proximal iteration in Hilbert space. *Computational optimization and applications*, 83, 2022.

[25] S.N. Ethier and T.G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley Series in Probability and Statistics. Wiley.

[26] C. Fang, Z. Lin, and T. Zhang. Sharp analysis for nonconvex SGD escaping from saddle points. *Proceedings of Machine Learning Research*, 99:1192–1234, 2019.

[27] I. Faragó. *Note on the Convergence of the Implicit Euler Method.*, volume 8236. Lecture Notes in Computer Science, vol 8236. Springer, Berlin, Heidelberg.

[28] J.-C. Fort and G. Pagès. Convergence of stochastic algorithms: from the Kushner-Clark theorem to the Lyapounov functional method. *Adv. in Appl. Probab.*, 28(4):1072–1094, 1996.

[29] G. Franca, J. Sulam, D. Robinson, and R. Vidal. Conformal symplectic and relativistic optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020.

[30] G. Franca, M. I. Jordan, and R. Vidal. On dissipative symplectic integration with applications to gradient-based optimization. *J. Stat. Mech.: Theory Exp.*, 2021.

[31] S. Gadat, F. Panloup, and F. Saadane. Stochastic heavy ball. *Electron. J. Stat.*, 12(1), 2018.

[32] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23 (4):2341–2368, 2013.

[33] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[34] M. Gurbuzbalaban, U. Simsekli, and L. Zhu. The heavy-tail phenomenon in SGD. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021.

[35] E. Hairer, S.P. Norsett, and G. Wanner. *Solving ordinary differential equations I.* Springer Berlin, 1987.

[36] E. Hairer, S.P. Norsett, and G. Wanner. *Solving ordinary differential equations II.* Springer Berlin, 1991.

[37] E. Hansen. Runge–Kutta time discretizations of nonlinear dissipative evolution equations. *Mathematics of Computations*, 75(254), 2005.

[38] W. Hundsdorfer and J. Verwer. *Numerical solution of time-dependent advection-diffusion-reaction equations.* Springer Berlin, 2003.

[39] M. Hutzenthaler, A. Jentzen, and P. Kloeden. Strong convergence of an explicit numerical method for sdes with nonglobally lipschitz continuous coefficients. *The Annals of Applied Probability*, 22(4), 2012.

[40] A Iserles. *A first course in numerical analysis of differential equations.* Cambridge university press, 2009.

[41] O. Kallenberg. *Foundations of Modern Probability, Third Edition.* Springer Switzerland, 2021.

[42] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014. Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[43] H. J. Kushner and D.S Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* Springer New York, 1978.

[44] H. J. Kushner and D.S Clark. *Stochastic Approximation and Recursive Algorithms and Applications.* Springer New York, 2003.

[45] Y. LeCun, C. Cortes, and C. Burges. MNIST handwritten digit database. Available at `http://yann.lecun.com/exdb/mnist`, 2010.

[46] H. Li, A. Rakhlin, and A. Jadbabaie. Convergence of Adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36, 2024.

[47] L. Ljung. Analysis of Recursive Stochastic Algorithms. Technical report, Department of Automatic Control, Lund Institute of Technology (LTH)., 1976.

[48] S. Ma, R. Bassily, and M. Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[49] P. Mertikopoulos, N. Hallak, A. Kavis, and V. Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1117–1128. Curran Associates, Inc., 2020.

[50] T. Mikolov. *Statistical language models based on neural networks.* PhD thesis, Brno University of Technology, 2013.

[51] K. Mischaikow, H. Smith, and H. Thieme. Asymptotically autonomous semiflows: Chain recurrence and lyapunov functions. *Transactions of the American Mathematical Society*, 347:1669–1685, 1995.

[52] Y. Nesterov. *Introductory Lectures on Convex Optimization.* Springer New York, 2004.

[53] A. Panigrahi, R. Somani, N. Goyal, and P. Netrapalli. Non-gaussianity of stochastic gradient noise, 2019. URL `https://arxiv.org/abs/1910.09626`.

[54] R. Pemantle. Nonconvergence to Unstable Points in Urn Models and Stochastic Approximations. *The Annals of Probability*, 18(2), 1990.

[55] B.T. Poljak. A general method for solving extremum problems. *Sov. Math., Dokl.*, 8(3), 1967.

[56] B. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Comp. Math. Math. Phys.*, 4, 1964.

[57] H. Robbins and S. Monro. A stochastic approximation algorithm. *Ann. Math. Statist.*, 22, 1951.

[58] H. Robbins and D. Siegmund. A convergence theorem for nonnegative almost supermartingales and some applications. In J. S. Rustagi, editor, *Optimizing Methods in Statistics.* Academic Press, 1971.

[59] L. Rosasco, S. Villa, and B.C Vu. Convergence of stochastic proximal gradient algorithm. *Applied Mathematics & Optimization.*, 82(3):891–917, 2020.

[60] E.K. Ryu and S. Boyd. Stochastic proximal iteration: A non-asymptotic improvement upon stochastic gradient descent.

[61] T. Stillfjord and M. Williamson. SRKCD: A stabilized Runge–Kutta method for stochastic optimization. *Journal of computational and applied mathematics*, 417:114575, 2023.

[62] G. Söderlind, L. Jay, and M. Calvo. Stiffness 1952–2012: Sixty years in search of a definition. *BIT Numerical Mathematics*, 55:531–558, 2015.

[63] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer New York, NY, 1996.

[64] V. N. Vapnik. *Statistical learning theory*. Wiley, 1998.

[65] V. N. Vapnik. *The nature of statistical learning*. Springer New York, 2000.

[66] Martin J. Wainwright. *High-dimensional statistics: a non-asymptotic viewpoint*. Cambridge university press, 2019.

[67] B. Wang, H. Zhang, Z. Ma, and W. Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In G. Neu and L. Rosasco, editors, *COLT*, volume 195 of *Proceedings of Machine Learning Research*. PMLR, 2023.

[68] S. Wiggins. *Introduction to applied non-linear dynamics and chaos.* Springer New York, 2003.

[69] D. Williams. *Probability with Martingales.* Cambridge University Press, 1991.

[70] M. Williamson and T. Stillfjord. Almost sure convergence of stochastic hamiltonian descent methods. *ArXiv Preprint: arXiv:2406.16649*, 2024. URL `https://arxiv.org/abs/2406.16649`.

[71] M. Williamson, M. Eisenmann, and T. Stillfjord. Analysis of a class of stochastic component-wise soft-clipping schemes. *ArXiv Preprint, arXiv:2406.16640*, 2024. URL `https://arxiv.org/abs/2406.16640`.

[72] T. Yoshizawa. Liapunov's function and boundedness of solutions. *Funkcialaj Ekvacioj*, (2), 1958.

[73] B. Zhang, J. Jin, C. Fang, and L. Wang. Improved analysis of clipping algorithms for non-convex optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15511–15521. Curran Associates, Inc., 2020.

[74] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64 (3), 2021.

[75] J. Zhang, T. He, S. Sra, and A. Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

[76] J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. Reddi, S. Kumar, and S. Sra. Why are adaptive methods good for attention models? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020.

Scientific publications

Author contributions

In Paper I, I contributed to the analysis and the numerical experiments. I participated in proofreading and revising the article.

In Paper II, I proved the majority of the results. I and my co-author contributed equally to writing the article and implementing the numerical experiments.

In Paper III, I proved several of the results and implemented most of the numerical experiments.

In Paper IV, I proved the majority of the results. I wrote the larger part of the article and implemented most of the numerical experiments.