Convergence and stability analysis of stochastic optimization algorithms

Williamson, Måns

2023

Document Version:
Publisher's PDF, also known as Version of record

– CENTRUM SCIENTIARUM MATHEMATICARUM –

Convergence and Stability Analysis of Stochastic Optimization Algorithms

MÅNS WILLIAMSON

Lund University
Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

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 Licentiate
Thesis advisors: Dr. M. Eisenmann, Prof. E. Hansen, Dr. T. Stillfjord
Faculty opponent: Prof. K. C. Zygalakis

To be presented, with the permission of the LTH, Faculty of Engineering of Lund University, for public criticism at Matematikcentrum, MH:333 on Tuesday, the 14th of March 2023 at 13:15.

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

Document name
LICENTIATE DISSERTATION

Date of disputation
2023-03-14

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 [1], 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 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
numerical analysis, optimization, stochastic optimization, machine learning

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title 1404-0034

ISBN
978-91-8039-558-8 (print)
978-91-8039-559-5 (pdf)

Recipient's notes

Number of pages
104

Price

Security classification

Convergence and Stability Analysis of Stochastic Optimization Algorithms.

by Måns Williamson

LUND
UNIVERSITY

Funding information: The thesis work was financially supported by Wallenberg AI, Autonomous Systems and Software program.

LTH, Faculty of Engineering, Centre for Mathematical Sciences

ISBN: 978-91-8039-558-8 (print)
ISBN: 978-91-8039-559-5 (pdf)

Printed in Sweden by Media-Tryck, Lund University, Lund 2023

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

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.**
E. 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

Acknowledgements

I would like to express my deep gratitude to my supervisors for their guidance and for patiently reading this thesis over and over again and coming with valuable feedback. I would also like to thank my family and my girlfriend Sana for always supporting me.

Popular summary in English

This thesis is concerned with stochastic optimization methods. In the fields of Artificial intelligence and Machine learning, it is common that one wants to estimate parameters in a statistical model in order to make predictions. One then uses a cost function that penalizes predictions that are far away from the true value, and minimize this with respect to the statistical parameters. An example is artificial neural networks that often have a very complicated structure and are difficult to minimize efficiently. Usual optimization methods are not suitable for these as they are too computationally demanding and one then uses stochastic optimization methods, as these are less costly and faster to use. Although they have proven to work well for such problems, they are often sensitive to the choice of step size/learning rate. If it is chosen too small it will take too long for it to find a good minimum, and if it is chosen too large it may blow up. In this thesis, we investigate different methods for stabilizing stochastic optimization schemes. More precisely, we look at methods for solving differential equations numerically, that have been shown to have good stability properties, and make use of them in the context of stochastic optimization algorithms.

Populärvetenskaplig sammanfattning på svenska

Den här uppsatsen handlar om stokastiska optimerings metoder. Inom områdena Artificiell intelligens och Maskininlärning är det vanligt att man vill skatta parametrar i en statistisk modell för att kunna göra prediktioner. Man använder sig då av en kostnadsfunktion som bestraffar prediktioner som ligger långt ifrån det riktiga värdet och minimerar denna med avseende på de statistiska parametrarna. Ett exempel på detta är artificiella neurala nätverk, som ofta är väldigt komplexa och svåra att minimera effektivt. Vanliga optimeringsmetoder är då inte lämpliga eftersom de är för beräkningsmässigt krävande och man använder sig istället av så kallade stokastiska optimerings metoder, som är mindre kostsamma och snabbare att använda. Även om dessa har visat sig fungera bra för dylika problem, är de ofta känsliga för valet av steglängd. Väljs den för liten tar det en evighet att hitta minimat till kostnadsfunktionen och väljs den för stor kan algoritmen explodera. I den här uppsatsen undersöks olika metoder för att stabilisera stokastiska optimeringsalgoritmer. Mer exakt tittar vi på metoder för att lösa differentialekvationer numeriskt, som har visat sig ha väldigt bra stabilitetsegenskaper och omformulerar dessa som stokastiska optimeringsalgoritmer.

Convergence and Stability Analysis of Stochastic Optimization Algorithms.

Chapter 1

Introduction

The research presented in this thesis is concerned with stochastic optimization methods. Stochastic optimization methods are often used in the field of machine learning in order to estimate the parameters in a statistical model, e.g., a regression- or a classification model. A typical example is the *stochastic gradient descent* method (SGD), which is a randomized version of the gradient descent algorithm. Using the SGD has shown to have several advantages; it is less computationally costly compared to the traditional algorithms such as the 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 [4, 5]. 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, compare [6]. 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 do not add much to the information in the gradient update. Here, stochastic algorithms that only make use of a subset of the data tend to use information more efficiently, see for example Sec. 3.3 in [7].

Despite its advantages, the step size 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 enter the picture. In the field of numerical analysis for differential equations, methods that allow for larger step sizes have long been used. Let $F : \mathbb{R}^d \to \mathbb{R}$. Then the gradient descent algorithm can be viewed as an explicit

Euler discretization of the gradient flow equation

$$\begin{aligned} w' &= -\nabla F(w), \\ w(0) &= w_0. \end{aligned} \tag{1.1}$$

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_*, which by definition satisfies

$$\nabla F(w_*) = 0,$$

i.e, a stationary point. There is a severe step size restriction on the explicit Euler scheme, and this can be remedied by using other schemes with larger stability region. In 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 thesis is arranged as follows; the second chapter gives an introduction to 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. 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 what their advantages are. We also treat stochastic optimization methods, and mention some of the most common results and their proof strategies.

In Chapter 5, we summarize the research done so far in the project and its conclusions, and consider some possible paths for future research.

Chapter 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 expected value 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 background to 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 predict the label y_i for a given feature x_i 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. 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 [8, Definition 5.1.1]). If we for example were to classify the images of the famous MNIST dataset [9], where each feature is a 28×28-pixel image of a handwritten digit between 0 and 9, we could have

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

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 could 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 is a Borel set, compare [10, 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

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

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 [11, 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 [8, 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 [12, 4.4] and the discussion on the outer expectation in [13]) and set-valued random variables (see [14, 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 [12], 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

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

in probability or almost surely (depending on whether 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 [11]. 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 [11, Thm. 3.5]. In machine learning, one often considers classes of prediction functions $\mathcal{H}$ with finite *VC-dimension*. 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 [15]. 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 [15, Ch. 3.7], but this is not the focus of this thesis.

Chapter 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 *explicit Euler's method*. We start with choosing a grid of time points $\{t_k\}_{k=0}^K$ defined by $t_{k+1} = t_k + K \cdot h$, with $h = \frac{T-t_0}{K}$, where h is the *step size*. Using the knowledge that $w(t_0) = w_0$, we define a sequence of approximations $\{w_k\}_{k=0}^K$ 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 to $w(t_{k+1})$, if we make use of (3.2). Assuming that the solution w is twice continuously differentiable, we get by Taylor expansion that

$$\begin{aligned} w(t_{k+1}) &= w(t_k + h) = w(t_k) + h \cdot w'(t_k) + \frac{h^2}{2} \cdot w''(\theta_k) \\ &= w(t_k) + h \cdot f(t_k, w(t_k)) + \frac{h^2}{2} \cdot w''(\theta_k), \quad \theta_k \in [t_k, t_{k+1}]. \end{aligned} \tag{3.3}$$

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 add and subtract w_{k+1} to both sides which yields the equation

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

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

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

It can be shown by induction, along with the fact that $1 + x \leq e^x$ for $x \geq 0$, [16, p. 6] that the global error satisfies

$$\|e_k\| \leq \frac{c}{L} \left(e^{(T - t_0)L} - 1 \right) h. \tag{3.4}$$

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 (3.4) is not very good for practical purposes; it is however possible to obtain sharper error bounds, see [16].

3.2 Implicit Euler

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

$$w_{k+1} = w_k + hf(t_{k+1}, w_{k+1}). \tag{3.5}$$

This can be rewritten on the form

$$w_{k+1} = R_f w_k, \tag{3.6}$$

where $R_f = (I + hf)^{-1}$ 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 [16, 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

$$r_{k+1} = w(t_{k+1}) - w(t_k) - hf(t_{k+1}, w(t_{k+1})) = -\frac{h^2}{2}w''(t_k) + \mathcal{O}(h^3). \tag{3.7}$$

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 (3.4). See e.g. [16, 17]. 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

$$w(t_{k+1}) = w(t_k) + \int_{t_k}^{t_{k+1}} f(s, w(s))\mathrm{d}s. \tag{3.8}$$

Given an approximation $w_k \approx w(t_k)$, we can in order to obtain an approximation for the function value at t_{k+1}, use a quadrature formula and approximate the integral on the right-hand side of (3.8), i.e.

$$w_{k+1} = w_k + h\sum_{i=0}^{s} b_i f(t_{k,i}, w_{k,i}). \tag{3.9}$$

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

$$w_{k,i} = w_k + h\sum_{j=1}^{s} a_{i,j} f(t + hc_j, w_{k,j}). \tag{3.10}$$

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 (3.9) and (3.10) 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 [18]. 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 [19].

3.4 Stability

Consider the initial value problem

$$\begin{aligned} w'(t) &= f(w(t)),\ t \geq t_0, \\ w(t_0) &= w_0, \end{aligned} \tag{3.11}$$

where $f : \mathbb{R}^d \to \mathbb{R}$. For simplicity, we consider autonomous systems in this section. We say that w_* is an *equilibrium solution* if $f(w_*) = 0$, 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 holds that addition $\lim_{t\to\infty}\|w(t) - w_*\| = 0$, the solution is said to be *asymptotically stable*.

It is possible to show that w_* is an asymptotically stable equilibrium solution if and only if all the eigenvalues of the Jacobian of f at w_* have negative real part, see [20, Thm. 1.2.5]. If we assume that the Jacobian at w_* is diagonalizable, then the linearized system

$$\begin{aligned} w'(t) &= J_f(w_*)w(t),\ t \geq t_0, \\ w(t_0) &= w_0, \end{aligned} \tag{3.12}$$

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

$$\begin{aligned} x'(t) &= \Lambda x(t),\ t \geq t_0, \\ x(0) &= x_0, \end{aligned}$$

where Λ is a diagonal matrix, with the eigenvalues of $J_f(w_*)$ on the diagonal. We thus have d linear equations, all of the form

$$\begin{aligned} y'(t) &= \lambda y(t),\ \lambda \in \mathbb{C},\ t \geq t_0, \\ y(0) &= y_0. \end{aligned} \tag{3.13}$$

Equation 3.13 is known as the *linear test equation*. Since we have $|y(t)| = e^{\mathrm{Re}(\lambda)t}|y_0|$ for equation 3.13, we see that $y_* = 0$ is asymptotically stable if and only if $\mathrm{Re}(\lambda) < 0$.

For a numerical method that produces a sequence of approximations $\{y_k\}_{k\geq 0}$ to the solution to (3.13), it would be desirable that it mimicked this behavior; i.e. it should satisfy

$$\lim_{k\to\infty} y_k = 0, \tag{3.14}$$

when applied to equation 3.13 with $\mathrm{Re}(\lambda) < 0$.

If we apply the *explicit Euler method* from Section 3.1 to (3.13), 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), (3.14) holds as we have that $|y_{n+1}| < |R(z)||y_n|$. 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), the step size restriction becomes $h < -\frac{2}{\lambda}$.

For the implicit Euler method from Section 3.2, the stability function is given by $R(z) = (1 - z)^{-1}$. The stability region 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 region. A method that satisfies this, is said to be *A-stable*, see [21, Def. 3.3]. In particular, the negative real line $\mathbb{R}_-$, is included in the stability region of an A-stable method.

For the Runge–Kutta methods introduced in Section 3.3, applying (3.9) and (3.10) to (3.13), gives the update $y_{n+1} = R(z)y_n$, where $R(z) = 1 + zb^t(I - zA)^{-1} \cdot \mathbb{1}$, where $\mathbb{1} = (1, \ldots, 1) \in \mathbb{R}^s$ 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.

The larger part of the negative real axis the stability domain contains, the larger step size it allows for. Following [19], 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 domain. For any explicit s-stage Runge–Kutta method, it holds that $\beta_R \leq 2s^2$, compare [19, 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

$$R_s(z) = T_s\left(1 + \frac{z}{s^2}\right), \tag{3.15}$$

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

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

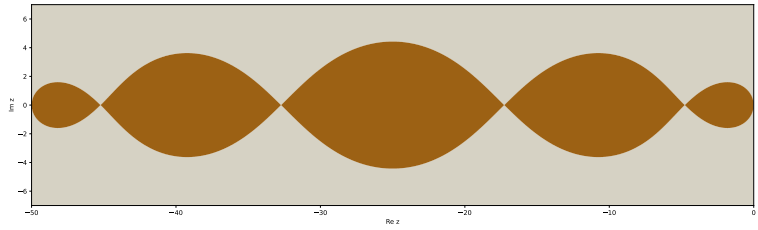

Figure 3.1: Stability region of a RKC method with 5 stages. We see that there are points z on the negative real axis for which $|R(z)| = 1$.

In Figure 3.1, 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 [19, V.1], is to introduce a damping factor. Instead of using the stability polynomials in (3.15), one uses a damped version 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 [19]. For small $\varepsilon > 0$ it is a slight reduction compared to that of the un-damped method, but instead we gain some margin around the critical points. See Figure 3.1 for an illustration of the stability region of a damped RKC-method with 5 stages.

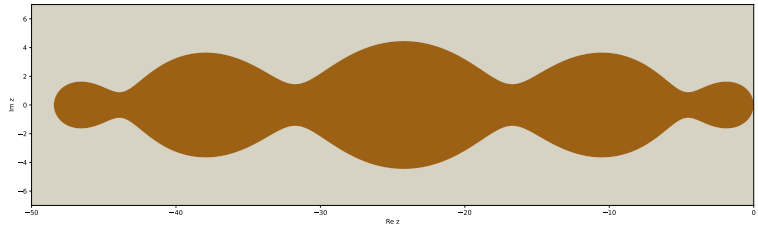

Figure 3.2: Stability region of a damped RKC method with 5 stages. The damping parameter $\varepsilon = 0.05$.

Chapter 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 $\alpha_k > 0$. 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 [22]. This is sometimes referred to as *L-smoothness*. 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 function value $F(w_k)$ decreases with each iteration. 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 [22, 1.2.3], 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 [22, Cor. 1.2.2, Thm. 1.2.4] that if $\|w_k - w_*\|_2 < \frac{2l}{M}$, then

$$\|I - \alpha_k G_k\|_2 < 1.$$

From this, and the fact that $\|w_{k+1} - w_*\|_2 \leq \|I - \alpha_k G_k\|_2 \|w_k - w_*\|_2$, we see that the sequence $\{w_k\}_{k \geq 0}$ converges to w_*. One can prove [22, Thm. 1.2.4] that 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

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

This result is tells us that even for a non-convex function, as long as it is sufficiently smooth, and we start close enough to a local minimum, we will converge to that minimum linearly, given that the step size is chosen according to the theorem. See Thm. 1.2.4. in [22] for details. One 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 the right-hand side of (4.6) and ensure that the convergence is linear. 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). Therefore, it is clear that this result, although interesting and informative in its own right, is purely theoretical.

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 (1.1). 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

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

and is known as the *proximal point method*. 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, once we have obtained the 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 $\nabla f(w_k, \xi_k)$ and uses this in the update:

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

Here $\{\xi_k\}_{k\geq 1}$ is a sequence of i.i.d random variables 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 expectation taken with respect to ξ_k given 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 $\{\xi_k\}_{k\geq 1}$ is independent.

There are several strategies for showing convergence of the stochastic gradient descent method. In this section, we will closely follow the approach in [7]. 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. Then we can use the bound (4.3) from Section 4.1. We can then insert

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

into inequality (4.3), to 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,

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

A common assumption on stochastic optimization algorithms is that they satisfy $\mathbb{E}_{\xi_k}\left[\|\nabla f(w_k,\xi_k)\|_2^2\right] \leq M$, for some constant $M > 0$. It was however shown in [23], that this is not satisfied for strongly convex functions on unbounded domains. Hence, as in [7], we make the following, weaker 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 $M, M_G > 0$. If we now insert (4.15) into (4.14), we get

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

Here we see that if we impose the step size restriction $\alpha_k \leq \frac{1}{LM_G}$, the term $\left(1 - \frac{\alpha_k L M_G}{2}\right) > \frac{1}{2}$. Thus, for a step size that satisfies this, the previous bound becomes

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

We now take the expectation with respect to the joint distribution of all the variables ξ_j, i.e. $\mathbb{E}[\cdot] = \mathbb{E}_{\xi_1}\mathbb{E}_{\xi_2}\ldots\mathbb{E}_{\xi_k}[\cdot]$, for $j \leq k$,

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

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

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

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,

$$\min_{1\leq k\leq K}\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right]\sum_{k=1}^{K}\alpha_k \leq \sum_{k=1}^{K}\alpha_k\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right]. \tag{4.17}$$

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

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

where we have used (4.12) in order to bound $-\mathbb{E}\left[F(w_{K+1})\right]$ by $-F_*$. From (4.18) we see that the sequence

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

converges to 0, if we require that

$$\sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty. \tag{4.20}$$

With the additional regularity assumptions that $\|\nabla F(w)\|_2^2$ is differentiable, we can also say that

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

although not with a rate, compare Cor. 4.12 in [7].

We now look at a proof strategy for the convex case. A common assumption is that the objective function is strongly convex with convexity constant $c > 0$, i.e.

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

It is proved in [7, (4.12)] that

$$2c\left(F(w) - F(w_*)\right) \leq \|\nabla F(w)\|_2^2, \tag{4.21}$$

where w_* is the unique global minimum of F. The fact that such a minimum exists follows from (4.12), the continuity of F along with the strong convexity [24, Cor. 11.17]. Inserting (4.21) into inequality (4.16), 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

$$\mathbb{E}\left[F(w_{k+1})\right] - F(w_*) \leq (1 - c\alpha_k)\left(\mathbb{E}\left[F(w_k)\right] - F(w_*)\right) + \frac{LM}{2}\alpha_k^2. \tag{4.22}$$

Using an induction argument as in [7, Thm. 4.7], we can use (4.22) to show that with $\alpha_k = \frac{\beta}{k+\gamma}$, where $\beta > \frac{1}{c}$ and $\gamma > 0$, we have

$$\mathbb{E}\left[F(w_k) - F(w_*)\right] \leq \frac{\nu}{k+\gamma}, \tag{4.23}$$

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 [7].

The decreasing step size in (4.20) and (4.23) is needed for convergence. If we use a fixed step size, the bound (4.18) becomes

$$\min_{1\leq k\leq K} \mathbb{E}\left[\|\nabla F(w_k)\|_2^2\|\right] \leq \frac{2\left(F(w_1) - F_*\right)}{\alpha K} + \frac{LM}{2}\alpha. \tag{4.24}$$

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 (4.19) stays bounded, but it does not converge to 0. This is sometimes referred to as a *noise-ball* around a stationary point. Similarly, we can use (4.22) 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 [7, Thm. 4.6]. Indeed, by subtracting $\frac{LM\alpha}{2c}$ from both sides of (4.22), we get the bound

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

If the constant step size $\alpha < \frac{2}{c}$ this will be a contraction, and we find that

$$\mathbb{E}\left[F(w_{k+1}) - F(w_*)\right] \leq \frac{LM\alpha}{2c} + (1 - c\alpha)^k\left(\mathbb{E}\left[F(w_1) - F(w_*) - \frac{LM\alpha}{2c}\right]\right), \tag{4.25}$$

from which we conclude that $\mathbb{E}\left[F(w_{k+1}) - F(w_*)\right]$ is bounded by $\frac{LM\alpha}{2c}$ 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 (4.25) and (4.24), 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 (4.24), we see that we will have achieved an error of size $\mathcal{O}(1/\sqrt{K})$ after K iterations in the non-convex case.

Chapter 5

Research and Outlook

In this chapter, we summarize the results from Paper I and Paper II 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.

5.1 Summary and Conclusions

In Section 4.2 of Chapter 4, we introduced the proximal point method and noted that it can be viewed as the implicit Euler scheme from Section 3.2 in Chapter 3, applied to the gradient flow equation.

$$\begin{aligned} \dot{w}(t) &= -\nabla F(w(t)),\ t \geq 0, \\ w(0) &= w_0. \end{aligned}$$

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. Suppose that H is a real Hilbert space and $F : H \to \mathbb{R}$ a strongly convex function. We are interested in finding the unique solution w_* to the problem

$$w_* = \arg\min_{w \in H} F(w). \tag{5.1}$$

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 (5.1) by producing a sequence of iterates $\{w_k\}_{k\geq 1}$ according to the update rule

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

where $\{\alpha_k\}_{k\geq 1}$ is a step size sequence, i.e. $\alpha_k > 0$ for every $k \geq 1$. 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 few exceptions to this are [25], where a weak type of convergence for maximal monotone operators is proved and [26], where norm convergence at a rate, but with a rather strong global Lipschitz condition on the objective function is proved. Under the assumption that the gradient of $f(\cdot, \xi)$ satisfy 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.

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

for some constant C and where w_* is defined by (5.1). The research in Paper I, generalizes that in [27] and extends it to an infinite dimensional setting. This result is new in that it provides a convergence rate for the scheme (5.2) in the infinite dimensional setting. In many cases, a closed form solution of (5.2), to obtain w_{k+1}, can be found, and then the stochastic proximal method provides a more stable alternative to the SGD, at essentially the same computational cost, see [2, Sec. 5].

Although the proximal point method has very good stability properties, it can be computationally costly to compute the implicit update (5.1) 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 Section 3.4 of Chapter 3, 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 [28], 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 Chebshev methods introduced in Section 3.4 of Chapter 3

for approximating the solution to (5.1). 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 [29] by for example making use of Lemma 2.1 in [30] or Lemma 2.1 in [31], under suitable assumptions. We obtain convergence guarantees in expectation at a sublinear rate, see Thm. 2.6 in Paper II. Under slightly stricter regularity assumptions, we obtain convergence in expectation to a stationary point, see Thm. 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.

$$\min_{1\leq k\leq K} \mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] = \mathcal{O}\left(\frac{1}{\log(K)}\right), \tag{5.3}$$

in the non-convex case. This follows from (4.17) in Section 4.3 in Chapter 4, (2.10) in Thm. 2.8 in Paper II, along with the fact that

$$A_K = \sum_{k=1}^{K} \frac{\beta}{k+\gamma} \geq \int_1^{K+1} \frac{\beta}{x+\gamma} \mathrm{d}x = \beta\left(\log(K+1+\gamma) - \log(1+\gamma)\right). \tag{5.4}$$

The argument is essentially the same as that in Section 4.3 in Chapter 4 and is therefore omitted here.

Something else worth noting is that although we prove convergence in expectation in Thm. 2.1 and and Thm. 2.6 in Paper II, a standard result in probability theory states that this implies convergence in probability, compare [32, Prop. 3.1.5]. Thus, we can for example use (5.3), to say that

$$\mathbb{P}\Big(\Big\{\omega : \min_{1\leq k\leq K} \|\nabla F(w_k)\|_2^2 > \varepsilon\Big\}\Big) = \mathcal{O}\left(\frac{1}{\log(K)}\right). \tag{5.5}$$

Note however that the error constant inversely proportional to ε, see [32, Prop. 3.1.5]. From (5.5) we can also conclude that the sequence

$$\min_{1\leq k\leq K} \|\nabla F(w_k)\|_2^2 \tag{5.6}$$

converges almost surely to 0, i.e. the set where (5.6) fails to converge has measure 0. To see this, we note that (5.6) is a decreasing sequence in K almost surely. Thus, we can appeal to [33, Thm. 1 Sec. 2.10.3], which states that a sequence of random variables $\{\zeta_k\}_{k\geq 1}$ converges almost surely to a random variable ζ, if and only if

$$\lim_{n\to\infty} \mathbb{P}\Big(\Big\{\omega \in \Omega : \sup_{k\geq n} |\zeta_k(\omega) - \zeta(\omega)| \geq \varepsilon\Big\}\Big) = 0,$$

for every $\varepsilon > 0$, to conclude that (5.6) converges almost surely.

It is relatively common to show convergence of the type (5.3), see e.g. [7, Thm. 4.10] (and keep in mind (4.17) in Section 4.3 in Chapter 4). Here, we see that it follows that (5.6) actually converges almost surely, although not with a convergence rate. It is a weaker result than that in e.g. [34], [35] or [36], where it is shown that every converging subsequence of $\{w_k\}_{k\geq 1}$ converges almost surely to a point in the set $\{w : \nabla F(w) = 0\}$. The later can likely be shown for the algorithms in Paper I and Paper II as well, but then a different proof strategy is probably needed.

Methods with large stability region are particularly useful for *stiff* problems. See [37] for a discussion on this. For a convex quadratic optimization problem, this essentially corresponds to 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.

5.2 Outlook

In this section, we discuss possible paths for future research, based on the conclusions from the papers in the thesis.

5.2.1 Other stabilized schemes

The stochastic proximal iteration in Paper I and the SRKCD in Paper II, are two examples of schemes that are "stabilized" versions of the SGD. Another option for stabilizing schemes, is to take into account information about the gradient when choosing the step size. Consider the stochastic gradient update

$$w_{k+1} = w_k - \alpha_k \nabla f(w_k, \xi), \tag{5.7}$$

which was discussed in Section 4.3. If the stochastic gradient $\nabla f(w_k, \xi)$ is very large, then we will take a large step in that direction. A large step in a "bad" direction could make us end up at an iterate very far from a local minimum which could delay the convergence considerably. A way to remedy this is to rescale the step size in order to compensate for having a large gradient in the update. This is sometimes referred to as *gradient clipping*, compare [6, Sec. 10.11.1]. Recently, a stochastic optimization scheme based on the *tamed Euler scheme*

was proposed in [30]. Instead of (5.7), one considers the update

$$w_{k+1} = w_k - \frac{\alpha_k \nabla f(w_k, \xi)}{1 + \alpha_k \|\nabla f(w_k, \xi)\|}. \tag{5.8}$$

Large gradients are compensated for, by the gradient in the denominator. With a decreasing step size, such as $\alpha_k = \frac{\beta}{k}$, for some $\beta > 0$, the longer we let the algorithm run, the smaller α_k becomes and the smaller the rescaling becomes. The tamed Euler scheme (5.8) has shown to work very well, see [30]. Another possibility is to apply the rescaling term component-wise, compare Sec. 10.11.1 in [6] and Sec. 3.3 in [38]. The idea of element-wise rescaling was made use of in the popular *Adam* algorithm, proposed in [39]. The component-wise version of (5.8) would take the form

$$(w_{k+1})_i = (w_k)_i - \frac{\alpha_k \frac{\partial f(w_k, \xi_k)}{\partial x_i}}{1 + \alpha_k \left| \frac{\partial f(w_k, \xi_k)}{\partial x_i} \right|}. \tag{5.9}$$

Here $(w)_i$ denotes the i:th component of the vector w. It can be shown that the tamed Euler scheme is a second order perturbation of the SGD, see [30]. This means that we can write (5.9) as

$$(w_{k+1})_i = (w_k)_i - \alpha_k \frac{\partial f(w_k, \xi)}{\partial x_i} + \alpha_k^2 h_k \left(\frac{\partial f(w_k, \xi_k)}{\partial x_i} \right), \tag{5.10}$$

where $h_k(x) = \frac{x|x|}{1+\alpha_k|x|}$, i.e.

$$x - \alpha_k h_k(x) = \frac{x}{1 + \alpha_k |x|}. \tag{5.11}$$

An ansatz for further studies is to consider the scheme (5.10) for more general functions h_k. This would allow for using a broad class of stabilizing methods, from which the most optimal could be chosen for the given optimization problem that one is presented with.

We will now prove convergence of a generalized version of (5.10). We will go through the proof, and impose assumptions on the functions $h_k(x)$, "as we go" by looking at the properties of h_k for the tamed Euler scheme and generalizing these. The assumptions that we make on the objective function F is that it has a Lipschitz continuous gradient and that it is bounded below, i.e. there is a $F_* \in \mathbb{R}$ such that

$$F(w) \geq F_*,$$

for all $w \in \mathbb{R}^d$.

The full update (5.10) can be written on the form

$$w_{k+1} = w_k - \alpha_k \nabla f(w_k, \xi_k) + \alpha_k^2 H_k(\nabla f(w_k, \xi_k)), \tag{5.12}$$

where $H_k : \mathbb{R}^d \to \mathbb{R}^d$ applies h_k element-wise to the gradient and $\{\xi_k\}_{k\geq 1}$ is a sequence of i.i.d. random variables.

Assuming that ∇F is Lipschitz continuous, we can plug (5.12) into (4.3) from Section 4.1 in Chapter 4. This yields

$$\begin{aligned} F(w_{k+1}) - F(w_k) \leq & -\alpha_k \langle \nabla F(w_k), \nabla f(w_k, \xi_k) \rangle \\ & + \alpha_k^2 \langle \nabla F(w_k), H_k\left(\nabla f(w_k, \xi_k)\right) \rangle \\ & + \frac{L\alpha_k^2}{2} \| \nabla f(w_k, \xi_k) - \alpha_k H_k\left(\nabla f(w_k, \xi_k)\right) \|_2^2. \end{aligned} \tag{5.13}$$

Assuming that $\nabla f(w, \xi)$ is an unbiased estimator of $\nabla F(w)$, we find that

$$\begin{aligned} \mathbb{E}_{\xi_k}\left[F(w_{k+1})\right] - F(w_k) \leq & -\alpha_k \|\nabla F(w_k)\|_2^2 \\ & + \alpha_k^2 \langle \nabla F(w_k), \mathbb{E}_{\xi_k}\left[H_k\left(\nabla f(w_k, \xi_k)\right)\right] \rangle \\ & + \frac{L\alpha_k^2}{2} \mathbb{E}_{\xi_k}\left[\| \nabla f(w_k, \xi_k) - \alpha_k H_k\left(\nabla f(w_k, \xi_k)\right) \|_2^2 \right], \end{aligned} \tag{5.14}$$

where we have used the fact that $\mathbb{E}_{\xi_k}\left[\nabla f(w_k, \xi_k)\right] = \nabla F(w_k)$, on the first term on the right-hand side. For the component-wise tamed Euler scheme, it holds for the last term on the right-hand side of (5.14) that

$$\mathbb{E}_{\xi_k}\left[\| \nabla f(w_k, \xi_k) - \alpha_k H_k\left(\nabla f(w_k, \xi_k)\right) \|_2^2 \right] \leq \mathbb{E}_{\xi_k}\left[\| \nabla f(w_k, \xi_k) \|_2^2 \right], \tag{5.15}$$

by (5.11) and since $\left(\frac{x}{1+\alpha_k|x|}\right)^2 \leq x^2$. Hence, we impose this as a condition on the function H_k, i.e. $(x - \alpha_k h_k(x))^2 \leq x^2$ for every k.

We now turn our attention to the second term on the right-hand side of (5.14),

$$\alpha_k^2 \langle \mathbb{E}_{\xi_k}\left[\nabla f(w_k, \xi_k)\right], \mathbb{E}_{\xi_k}\left[H_k\left(\nabla f(w_k, \xi_k)\right)\right] \rangle. \tag{5.16}$$

We can write this out in terms of the partial derivatives:

$$\alpha_k^2 \sum_{i=1}^{d} \mathbb{E}_{\xi_k}\left[X_i\right] \cdot \mathbb{E}_{\xi_k}\left[h_k\left(X_i\right)\right],$$

where $X_i := \frac{\partial f(w_k,\xi_k)}{\partial x_i}$. We can at this point make use of Chebyshev's second inequality, see e.g. [40] or [8, Thm. 4.7.9], which states that if ϕ and ψ are two increasing functions, it holds that

$$\mathbb{E}_{\xi_k}\left[\phi\left(X_i\right)\right]\mathbb{E}_{\xi_k}\left[\psi\left(X_i\right)\right] \leq \mathbb{E}_{\xi_k}\left[\phi\left(X_i\right)\psi\left(X_i\right)\right]. \tag{5.17}$$

If we thus require that $h_k(x)$ is increasing for every $k \geq 1$ (as it is for the component-wise tamed Euler scheme (5.9)), we can use (5.17), with $\phi(x) = x$ and $\psi(x) = h_k(x)$, to obtain that

$$\mathbb{E}_{\xi_k}\left[X_i\right]\mathbb{E}_{\xi_k}\left[h_k\left(X_i\right)\right] \leq \mathbb{E}_{\xi_k}\left[X_i h_k(X_i)\right].$$

With $h_k(x) = \frac{x|x|}{1+\alpha_k|x|}$, we have that $xh_k(x) \leq |x|^3$. If we enforce this as a condition on the functions $h_k(x)$, it thus holds that

$$\alpha_k^2\langle\mathbb{E}_{\xi_k}\left[\nabla f(w_k,\xi_k)\right],\mathbb{E}_{\xi_k}\left[H_k\left(\nabla f(w_k,\xi_k)\right)\right]\rangle \leq \alpha_k^2\mathbb{E}_{\xi_k}\left[\|\nabla f(w_k,\xi_k)\|_2^3\right]. \tag{5.18}$$

By inserting (5.15) and (5.18) into (5.14), we thus get

$$\begin{aligned}\mathbb{E}_{\xi_k}\left[F(w_{k+1})\right] - F(w_k) \leq &-\alpha_k\|\nabla F(w_k)\|_2^2 + \alpha_k^2\mathbb{E}_{\xi_k}\left[\|\nabla f(w_k,\xi_k)\|_2^3\right] \\ &+ \frac{L\alpha_k^2}{2}\mathbb{E}_{\xi_k}\left[\|\nabla f(w_k,\xi_k)\|_2^2\right].\end{aligned} \tag{5.19}$$

If we further assume that we can bound the second- and third moments of the stochastic gradient [1]

$$\begin{aligned}&\mathbb{E}_{\xi_k}\left[\|\nabla f(w_k,\xi_k)\|_2^2\right] \leq \sigma^2, \\ &\mathbb{E}_{\xi_k}\left[\|\nabla f(w_k,\xi_k)\|_2^3\right] \leq \kappa,\ \kappa > 0,\end{aligned} \tag{5.20}$$

we get by inserting (5.20) into (5.19), the bound

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

Taking the full expectation with respect to all $\{\xi_k\}_{k\geq 1}$, and summing up from 1 to K as in Chapter 4, Section 4.3, we get

$$\sum_{k=1}^{K}\alpha_k\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] \leq F(w_1) - F_* + \sum_{k=1}^{K}\alpha_k^2\left(\kappa + \frac{L\sigma^2}{2}\right). \tag{5.21}$$

[1] We can compare this to the bound 4.15 in Section 4.3 of Chapter 4. As we are considering non-convex objective functions in the discussion here, it is not a contradictory assumption in this case.

Here we have used that $\mathbb{E}\left[F(w_1)\right] = F(w_1)$ and that $\mathbb{E}\left[F(w_{K+1})\right] \geq F_*$. We can now use the tricks introduced in Section 4.3 Chapter 4 on (5.21). With a decreasing step size such as $\alpha_k = \frac{\beta}{k}$, for some $\beta > 0$, we get that

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

where we have used an argument similar to (5.4) along with (4.17). With a fixed step size, we get a bound similar to that of (4.24). If we let the step size depend on the number of iterations, i.e. $\alpha = \frac{1}{\sqrt{K}}$, we get that

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

Determining which functions h_k are good candidates in (5.10) is a topic of future research. Some examples of candidate functions are $h_k(x) = x - g(x)$, where $g(x)$ is a function with sigmoid shape satisfying $g(x) = 0$, such as $g(x) = \arctan(x)$ or $g(x) = \frac{1}{1+e^{-x}} - \frac{1}{2}$. As noted above, the assumption (5.20) is slightly stronger than the usual assumptions. It is the bound (5.18), that makes this necessary. An alternative could be to note that for (5.9), we have for $\alpha_k > \frac{1}{B'}$ where B' is some constant strictly greater than 0, that

$$xh_k(x) = \frac{|x|^3}{1+\alpha_k|x|} \leq \frac{|x|^3}{\alpha_k|x|} \leq B'x^2.$$

On the other hand, for α_k small enough, the scheme (5.10), behaves essentially like SGD since $\alpha_k h_k(x)$ will be small. Based on these observations, one could thus imagine splitting the analysis into two cases, depending on whether $\alpha_k > \frac{1}{B'}$ or not.

A different approach could be to analyze the scheme presented above, in the framework of monotone operators in the infinite dimension, as in [30] and Paper I. This would likely require an a priori bound similar to Lemma 2 in Paper I and [30, Lemma 4.1] in order to show that the difference $\|w_{k+1} - w_*\|$ is bounded by something like $C_k\|w_k - w_*\| + D\alpha_k^2$, where $0 < C_k < 1$ and D is some constant.

5.2.2 Almost sure convergence

Another option that could be investigated further is the almost sure convergence of the schemes analyzed in Paper I and Paper II. Almost sure convergence of the classical SGD scheme has been done in for example [35, 36, 41, 42] and more

recently [34]. Very roughly, one can say that the main idea is to construct a stochastic process which is an interpolation of the sequence $\{w_k\}_{k\geq 1}$, shift this in time and then make use of the Arzela–Ascoli theorem along with martingale techniques to conclude that the sequence $\{w_k\}_{k\geq 1}$ converges to a stationary point of the gradient flow, see [35, 41, 43]. The assumptions and strategies vary, but one of the main assumptions is that $\sup_k \|w_k\| < \infty$, which does not always hold, see [44]. Recently, [34] showed that it does hold for SGD under some additional assumptions, such as Lipschitz continuity of the objective function and that the gradient is bounded.

5.2.3 Avoidance analysis

Convergence to an equilibrium point does not guarantee that the point is a local minimum, but in [34] it is shown that the probability that the stochastic gradient descent scheme converges to a saddle-point manifold is 0. A possible path for further research is to perform such an analysis for the schemes in Paper 1 and Paper 2.

5.2.4 Extension to the online framework

In Section 4.3 of Chapter 4, we noted that the objective function in machine learning problems often take the form

$$F(w) = \frac{1}{K}\sum_{i=1}^{K} \ell(h(x_i, w), y_i), \tag{5.22}$$

where ℓ is a loss function, $h(\cdot, w)$ a prediction function and $\{x_i, y_i\}_{i=1}^{K}$ are samples. The minimization of the function (5.22) is based on the assumption that we have the whole data set at hand when we start the optimization procedure. There are cases when one might want to perform the optimization procedure and update the prediction based on the parameter w, as new data is obtained. This gives rise to a class of algorithms known as *online algorithms*. It is common to phrase the problem in the following way; at time k we receive a sample x_k and based on this we choose an iterate w_k with which we predict $h(x_k, w_k)$. The true label y_k of x_k is then revealed, and we suffer the loss $\ell(h(w_k, x_k), y_k)$, i.e. $\ell(h(w_k, x_k), y_k)$ gives a measurement on how far the prediction $h(w_k, x_k)$ was from the label y_k. The goal of the algorithm is to incur a low average loss

$$\frac{1}{K}\sum_{k=1}^{K} \ell(h(w_k, x_k), y_k) \tag{5.23}$$

and we can evaluate its performance by considering the *average regret*

$$\frac{R_K}{K} = \frac{1}{K}\sum_{k=1}^{K} \ell(h(w_k, x_k), y_k) - \frac{1}{K}\sum_{k=1}^{K} \ell(h(w_*, x_k), y_k), \tag{5.24}$$

where w_* is the parameter that minimizes (5.22). Essentially, this measures how much the average loss differs from the loss, if we stop after K iterations and then find the argument that minimizes (5.22). Suppose that our sequence of predictions $\{w_k\}_{k\geq 1}$ is defined by

$$w_{k+1} = w_k - \alpha_k \nabla \ell(h(w_k, x_k), y_k), \tag{5.25}$$

where the gradient is taken with respect to w. We then have the following result due to Zinkevic [45]. Assume that the gradients are differentiable and uniformly bounded in the sense that $\|\nabla \ell(h(w, x_k), y_k)\|_2^2 \leq B_1$ for some constant $B_1 > 0$, and that the iterates satisfy $\|w_k - w_*\|_2^2 \leq B_2$, where $B_2 > 0$. If we choose the step size $\alpha_k = k^{-1/2}$, we have the following bound for the average regret,

$$\frac{R_K}{K} \leq \frac{B_1}{2\sqrt{K}} + \left(\frac{1}{\sqrt{K}} - \frac{1}{2K}\right) B_2. \tag{5.26}$$

What (5.26) is saying is that if the average loss (5.23) is larger than

$$\min_w \frac{1}{K}\sum_{k=1}^{K} \ell(h(w_*, x_k), y_k), \tag{5.27}$$

after K iterations, then we have a bound on the difference, and we know that as we increase the number of iterations - or data points - this gap will decrease. In practice, the optimal parameter w_* that minimizes (5.27) is usually not known, and the performance of the online algorithm can be evaluated by looking at the behavior of (5.23).

Extending the SRKCD algorithm to the online-framework could also be a direction for future studies. Finding a strategy for choosing the number of stages in the algorithm, that is optimal with respect to the computational cost and efficiency, is a potential course for forthcoming research as well.

5.3 References

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

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

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

[4] 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.

[5] 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.

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

[7] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173.

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

[9] Yann LeCun, Corinna Cortes, and Christopher Burges. MNIST handwritten digit database. Available at `http://yann.lecun.com/exdb/mnist`, 2010.

[10] O. Kallenberg. *Foundations of Modern Probability, Third Edition*. Springer Switzerland, 2021. doi: https://doi.org/10.1007/978-3-030-61871-1.

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

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

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

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

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

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

[17] I. Faragó. *Note on the Convergence of the Implicit Euler Method.*, volume 8236. Lecture Notes in Computer Science, vol 8236. Springer, Berlin, Heidelberg. doi: https://doi.org/10.1007/978-3-642-41515-9_1.

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

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

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

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

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

[23] L. Nguyen, P. H. Nguyen, M. van Dijk, P. Richtarik, K. Scheinberg, and M. Takac. SGD and hogwild! Convergence without the bounded gradients assumption. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3750–3758. PMLR, 2018.

[24] H. H. Bauschke and P.L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* Springer, 2 edition, 2017. doi: 10.1007/978-3-319-48311-5.

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

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

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

[28] 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. doi: https://doi.org/10.1007/s10543-020-00819-y.

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

[30] M. Eisenmann and T. Stillfjord. Sub-linear convergence of a tamed stochastic method in Hilbert space. *SIAM Journal on Optimization*, 32, 2022.

[31] M.V. Balashov. Maximization of a function with Lipschtiz continuous gradient. *Journal of Mathematical Sciences.*, 209(1), 2015.

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

[33] A.N. Shiryaev. *Probability-1.* Graduate Texts in Mathematics. Springer New York, 2016.

[34] 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.

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

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

[37] G. Söderlind, L. Jay, and M. Calvo. Stiffness 1952–2012: Sixty years in search of a definition. *BIT Numerical Mathematics*, 55:531–558, 2015. doi: https://doi.org/10.1007/s10543-014-0503-3.

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

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

[40] A. M. Fink and Jr. M. Jodeit. On Chebyshev's other inequality. In Y.L. Tong, editor, *Inequalities in Statistics and Probability: Proceedings of the Symposium on Inequalities in Statistics and Probability, October 27-30, 1982, Lincoln, Nebraska*, volume 5, pages 115–120, 1984.

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

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

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

[44] P. Wang, Y. Lei, Y. Ying, and H. Zhang. Differentially private SGD with non-smooth losses. *Applied and Computational Harmonic Analysis*, 56, 2022. doi: https://doi.org/10.1016/j.acha.2021.09.001.

[45] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In Tom Fawcett and Nina Mishra, editors, *ICML*, pages 928–936. AAAI Press, 2003.

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.

Paper I

M. Williamson, M. Eisenmann and T. Stillfjord
Sub-linear convergence of a stochastic proximal iteration. method in Hilbert space
Computational Optimization and Applications, 2022, vol. 83, issue 1.

Computational Optimization and Applications (2022) 83:181–210
https://doi.org/10.1007/s10589-022-00380-0

Sub-linear convergence of a stochastic proximal iteration method in Hilbert space

Monika Eisenmann[1] · **Tony Stillfjord**[1] · **Måns Williamson**[1]

Received: 24 September 2021 / Accepted: 19 May 2022 / Published online: 20 June 2022

Abstract
We consider a stochastic version of the proximal point algorithm for convex optimization problems posed on a Hilbert space. A typical application of this is supervised learning. While the method is not new, it has not been extensively analyzed in this form. Indeed, most related results are confined to the finite-dimensional setting, where error bounds could depend on the dimension of the space. On the other hand, the few existing results in the infinite-dimensional setting only prove very weak types of convergence, owing to weak assumptions on the problem. In particular, there are no results that show strong convergence with a rate. In this article, we bridge these two worlds by assuming more regularity of the optimization problem, which allows us to prove convergence with an (optimal) sub-linear rate also in an infinite-dimensional setting. In particular, we assume that the objective function is the expected value of a family of convex differentiable functions. While we require that the full objective function is strongly convex, we do not assume that its constituent parts are so. Further, we require that the gradient satisfies a weak local Lipschitz continuity property, where the Lipschitz constant may grow polynomially given certain guarantees on the variance and higher moments near the minimum. We illustrate these results by discretizing a concrete infinite-dimensional classification problem with varying degrees of accuracy.

Keywords Stochastic proximal point · Convergence analysis · Convergence rate · Infinite-dimensional · Hilbert space

Mathematics Subject Classification 46N10 · 65K10 · 90C15

✉ Monika Eisenmann
monika.eisenmann@math.lth.se

Tony Stillfjord
tony.stillfjord@math.lth.se

Måns Williamson
mans.williamson@math.lth.se

[1] Centre for Mathematical Sciences, Lund University, P.O. Box 118, 22100 Lund, Sweden

1 Introduction

We consider convex optimization problems of the form

$$w^* = \operatorname{argmin}_{w \in H} F(w), \tag{1}$$

where H is a real Hilbert space and

$$F(w) = \mathbf{E}_\xi[f(w, \xi)].$$

The main applications we have in mind are supervised learning tasks. In such a problem, a set of data samples $\{x_j\}_{j=1}^n$ with corresponding labels $\{y_j\}_{j=1}^n$ is given, as well as a classifier h depending on the parameters w. The goal is to find w such that $h(w, x_j) \approx y_j$ for all $j \in \{1, \dots, n\}$. This is done by minimizing

$$F(w) = \frac{1}{n} \sum_{j=1}^{n} \ell(h(w, x_j), y_j), \tag{2}$$

where ℓ is a given loss function. We refer to, e.g., Bottou et al. [9] for an overview. In order to reduce the computational costs, it has been proved to be useful to split F into a collection of functions f of the type

$$f(w, \xi) = \frac{1}{|B_\xi|} \sum_{j \in B_\xi} \ell(h(w, x_j), y_j),$$

where B_ξ is a random subset of $\{1, \dots, n\}$, referred to as a batch. In particular, the case of $|B_\xi| = 1$ is interesting for applications, as it corresponds to a separation of the data into single samples.

A commonly used method for such problems is the stochastic gradient method (SGD), given by the iteration

$$w^{k+1} = w^k - \alpha_k \nabla f(w^k, \xi^k),$$

where $\alpha_k > 0$ denotes a step size, $\{\xi^k\}_{k \in \mathbb{N}}$ is a family of jointly independent random variables and ∇ denotes the Gâteaux derivative with respect to the first variable. The idea is that in each step we choose a random part $f(\cdot, \xi)$ of F and go in the direction of the negative gradient of this function. SGD corresponds to a stochastic version of the explicit (forward) Euler scheme applied to the gradient flow

$$\dot{w} = -\nabla F(w).$$

This differential equation is frequently stiff, which means that the method often suffers from stability issues.

The restatement of the problem as a gradient flow suggests that we could avoid such stability problems by instead considering a stochastic version of implicit (backward) Euler, given by

$$w^{k+1} = w^k - \alpha_k \nabla f(w^{k+1}, \xi^k).$$

In the deterministic setting, this method has a long history under the name *proximal point method*, because it is equivalent to

$$w^{k+1} = \operatorname{argmin}_{w \in H} \left\{ \alpha F(w) + \frac{1}{2}\|w - w^k\|^2 \right\} = \operatorname{prox}_{\alpha F}(w^k),$$

where

$$\operatorname{prox}_{\alpha F}(w^k) = (I + \alpha \nabla F)^{-1} w^k.$$

The proximal point method has been studied extensively in the infinite dimensional but deterministic case, beginning with the work of Rockafellar [28]. Several convergence results and connections to other methods such as the Douglas–Rachford splitting are collected in Eckstein and Bertsekas [13], see also Güler [17]. In the strongly convex case, the main convergence analysis idea is to observe that the gradient is strongly monotone. Then the resolvent $(I + \alpha \nabla F)^{-1}$ is a strict contraction, and the Banach fixed point theorem shows that $\{w^k\}_{k \in \mathbb{N}}$ converges to w^* in norm.

Following Ryu and Boyd [32], we will refer to the stochastic version as *stochastic proximal iteration* (SPI). We note that the computational cost of one SPI step is in general much higher than for SGD, and indeed often infeasible. However, in many special cases a clever reformulation can result in very similar costs. If so, then SPI should be preferred over SGD, as it will converge more reliably. We provide such an example in Sect. 5.

The main goal of this paper is to prove sub-linear convergence of the type

$$\mathbf{E}\left[\|w^k - w^*\|^2\right] \leq \frac{C}{k}$$

in an infinite-dimensional setting, i.e. where $\{w^k\}_{k \in \mathbb{N}}$ and w^* are elements in a Hilbert space H. As shown in e.g. [1, 26], this is optimal in the sense that we cannot expect a better asymptotic rate even in the finite-dimensional case.

Most previous convergence results in this setting only provide guarantees for convergence, without an explicit error bound. The convergence is usually also in a rather weak norm. This is mainly due to weak assumptions on the involved functions and operators. Overall, little work has been done to consider SPI in an infinite dimensional space. A few exceptions are given by Bianchi [7], where maximal monotone operators $\nabla F : H \to 2^H$ are considered and weak ergodic convergence and norm convergence is proved. In Rosasco et al. [30], the authors work with an infinite dimensional setting and an implicit-explicit splitting where ∇F is decomposed in a regular and an irregular part. The regular part is considered explicitly but with a stochastic approximation while the irregular part is used in a deterministic proximal step. They prove both $\nabla F(w^k) \to \nabla F(w^*)$ and $w^k \to w^*$ in H as $k \to \infty$. Without further assumptions, neither of these approaches yield convergence rates.

In the finite-dimensional case, stronger assumptions are typically made, with better convergence guarantees as a result. Nevertheless, for the SPI scheme in particular, we are only aware of the unpublished manuscript [32], which suggests $1/k$ convergence in $\mathbb{R}^d$. Based on [32], the implicit method has also been considered in a few other works: In Patrascu and Necoara [24], a SPI method with additional constraints

on the domain was studied. A slightly more general setting that includes the SPI has been considered in Davis and Drusvyatskiy [12]. Toulis and Airoldi and Toulis et al. studied such an implicit scheme in [35–37]. Finally, very recently and during the preparation of this work, [20] was published, wherein both SGD and proximal methods for composite problems are analyzed in a common framework based on bounded gradients. This is a generalization of the basic setting in a different direction than our work.

Whenever using an implicit scheme, it is essential to solve the appearing implicit equation effectively. This can be impeded by large batches for the stochastic approximation of F. On the other hand, a larger batch improves the accuracy of the approximation of the function. In Toulis et al. [39, 40] and Ryu and Yin [33], a compromise was found by solving several implicit problems on small batches and taking the average of these results. This corresponds to a sum splitting. Furthermore, implicit-explicit splittings can be found in Patrascu and Irofti [23], Ryu and Yin [33], Salim et al. [34], Bianchi and Hachem [8] and Bertsekas [6]. A few more related schemes have been considered in Asi and Duchi [2, 3] and Toulis et al. [38]. More information about the complexity of solving these kinds of implicit equations and the corresponding implementation can be found in Fagan and Iyengar [16] and Tran et al. in [40].

Our aim is to bridge the gap between the "strong finite-dimensional" and "weak infinite-dimensional" settings, by extending the approach of [32] to the infinite-dimensional case. We also further extend the results by allowing for more general Lipschitz conditions on $\nabla f(\cdot, \xi)$, provided that sufficient guarantees can be made on the integrability near the minimum w^*. In particular, we make the less restrictive assumption that for every function $f(\cdot, \xi)$ and every ball of radius $R > 0$ around the origin there is a Lipschitz constant $L_\xi(R)$ that grows polynomially with R. We also weaken the standard assumption of strong convexity and only demand that the functions are strongly convex for some realizations.

We note that if F is only convex then there might be multiple local minima, and proving convergence in norm is in general not possible. On the other hand, if every $f(\cdot, \xi)$ is strongly convex then parts of the analysis can be simplified. The assumptions made in this article are thus situated between these two extremes, where it is still possible to prove convergence results similar to the strongly convex case but under milder assumptions.

These strong convergence results can then be applied to, e.g., the setting where there is an original infinite-dimensional optimization problem which is subsequently discretized into a series of finite-dimensional problems. Given a reasonable discretization, each of those problems will then satisfy the same convergence guarantees.

Our analysis closely follows the finite-dimensional approach [32]. However, several arguments no longer work in the infinite-dimensional case (such as the unit ball being compact, or a linear operator having a minimal eigenvalue) and we fix those. Additionally, we simplify several of the remaining arguments, provide many omitted, but critical, details and extend the results to more general operators.

A brief outline of the paper is as follows. The main assumptions that we make are stated in Sect. 2, as well as the main theorem. Then we prove a number of preliminary results in Sect. 3, before we can tackle the main proof in Sect. 4. In Sect. 5 we

describe a numerical experiment that illustrates our results, and then we summarize our findings in Sect. 6.

2 Assumptions and main theorem

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a complete probability space and let $\{\xi^k\}_{k\in\mathbb{N}}$ be a family of jointly independent random variables on Ω. Each realization of ξ^k corresponds to a different batch. Let $(H, (\cdot,\cdot), \|\cdot\|)$ be a real Hilbert space and $(H^*, (\cdot,\cdot)_{H^*}, \|\cdot\|_{H^*})$ its dual. Since H is a Hilbert space, there exists an isometric isomorphism $\imath : H^* \to H$ such that $\imath^{-1} : H \to H^*$ with $\imath^{-1} : u \mapsto (u, \cdot)$. Furthermore, the dual pairing is denoted by $\langle u', u\rangle = u'(u)$ for $u' \in H^*$ and $u \in H$. It satisfies

$$\langle \imath^{-1}u, v\rangle = (u, v) \quad \text{and} \quad \langle u', v\rangle = (\imath u', v), \quad u, v \in H, u' \in H^*.$$

We denote the space of linear bounded operators mapping H into H by $\mathcal{L}(H)$. For a symmetric operator S, we say that it is positive if $(Su, u) \geq 0$ for all $u \in H$. It is called strictly positive if $(Su, u) > 0$ for all $u \in H$ such that $u \neq 0$.

For the function $f(\cdot, \xi) : H \times \Omega \to (-\infty, \infty]$, we use ∇, as in $\nabla f(u, \xi)$, to denote differentiation with respect to the first variable. When we present an argument that holds almost surely, we will frequently omit ξ from the notation and simply write $f(u)$ rather than $f(u, \xi)$. Given a random variable X on Ω, we denote the expectation with respect to $\mathbf{P}$ by $\mathbf{E}[X]$. We use sub-indices, such as in $\mathbf{E}_\xi[\cdot]$, to denote expectations with respect to the probability distribution of the random variable ξ.

We consider the stochastic proximal iteration (SPI) scheme given by

$$w^{k+1} = w^k - \alpha_k \imath \nabla f(w^{k+1}, \xi^k) \quad \text{in } H, \qquad w^1 = w_1 \quad \text{in } H, \tag{3}$$

for minimizing

$$F(w) = \mathbf{E}_\xi[f(w, \xi)],$$

where f and F fulfill the following assumption.

For the family of jointly independent random variables $\{\xi^k\}_{k\in\mathbb{N}}$, we are interested in the total expectation

$$\mathbf{E}_k\left[\|X\|^2\right] := \mathbf{E}_{\xi^1}\left[\mathbf{E}_{\xi^2}\left[\cdots \mathbf{E}_{\xi^k}\left[\|X\|^2\right]\cdots\right]\right].$$

Since the random variables $\{\xi^k\}_{k\in\mathbb{N}}$ are jointly independent, and w^k only depends on ξ^j, $j \leq k-1$, this expectation coincides with the expectation with respect to the joint probability distribution of $\xi^1, \ldots, \xi^{k-1}$. In the rest of the paper, it often occurs that a statement does not involve an expectation but contains a random variable. Where it does not cause any confusion, such a statement is assumed to hold almost surely even if this is not explicitly stated.

Assumption 1 For a random variable ξ on Ω, let the function $f(\cdot, \xi) : \Omega \times H \to (-\infty, \infty]$ be given such that $\omega \mapsto f(v, \xi(\omega))$ is measurable for

every $v \in H$ and such that $f(\cdot, \xi)$ is convex, lower semi-continuous and proper almost surely. Additionally, $f(\cdot, \xi)$ fulfills the following conditions:

- The expectation $\mathbf{E}_\xi\big[f(\cdot, \xi)\big] =: F(\cdot)$ is lower semi-continuous and proper.
- The function $f(\cdot, \xi)$ is Gâteaux differentiable almost surely on a non-empty common domain $\mathcal{D}(\nabla f) \subseteq H$, i.e. for all for all $v, w \in \mathcal{D}(\nabla f)$ the inequality $\langle \nabla f(v, \xi), w\rangle = \lim_{h\to 0} \frac{f(v+hw,\xi)-f(v,\xi)}{h}$ is fulfilled almost surely.
- There exists $m \in \mathbb{N}$ such that $\left(\mathbf{E}_\xi\left[\|\nabla f(w^*, \xi)\|_{H^*}^{2^m}\right]\right)^{2^{-m}} =: \sigma < \infty$.
- For every $R > 0$ there exists $L_\xi(R) : \Omega \to \mathbb{R}$ such that

$$\|\nabla f(u, \xi) - \nabla f(v, \xi)\|_{H^*} \leq L_\xi(R)\|u - v\|$$

  almost surely for all $u, v \in \mathcal{D}(\nabla f)$ with $\|u\|, \|v\| \leq R$. Furthermore, there exists a polynomial $P : \mathbb{R} \to \mathbb{R}$ of degree $2^m - 2$ such that $L_\xi(R) \leq P(R)$ almost surely.
- There exist a random variable $M_\xi : \Omega \to \mathcal{L}(H)$ such that the image is symmetric and a random variable $\mu_\xi : \Omega \to [0, \infty)$ such that $\mathbf{E}_\xi[\mu_\xi] = \mu > 0$ and $\mathbf{E}_\xi[\mu_\xi^2] = \nu^2 < \infty$. Moreover,

$$\langle \nabla f(u, \xi) - \nabla f(v, \xi), u - v\rangle \geq (M_\xi(u - v), u - v) \geq \mu_\xi\|u - v\|^2$$

  is fulfilled almost surely for all $u, v \in \mathcal{D}(\nabla f)$.

An immediate result of Assumption 1, is that the gradient $\nabla f(\cdot, \xi)$ is maximal monotone almost surely, see [27, Theorem A]. As a consequence, the resolvent (proximal operator)

$$T_{f,\xi} = (I + \nabla f(\cdot, \xi))^{-1}$$

is well-defined almost surely, see Lemma 1 for more details. Further, each resolvent maps into $\mathcal{D}(\nabla f)$, and as a consequence every iterate $w^k \in \mathcal{D}(\nabla f)$. Finally, we may interchange expectation and differentiation so that $\nabla F(w) = \mathbf{E}_\xi[\nabla f(\xi, w)]$. Note that this means that the approximation $\nabla f(\cdot, \xi)$ is an *unbiased* estimate of the full gradient ∇F. In our case, this property can be shown via a straightforward argument based on dominated convergence similar to [32, Lemma 6], but we note that it also holds in more general settings [21, 29].

Remark 1 The idea behind the operators M_ξ is that each $f(\cdot, \xi)$ is is allowed to be only convex rather than strongly convex. However, they should be strongly convex for *some* realizations, such that $f(\cdot, \xi)$ is strongly convex *in expectation*. By assumption, F is lower semi-continuous, proper and strongly convex, so there is a minimum w^* of (1) (c.f. [4, Proposition 1.4]) which is unique due to the strong convexity.

Remark 2 Note that the local Lipschitz constant of Assumption 1 is a generalization compared to [32] and other existing literature. Instead of asking for one Lipschitz constant L_ξ that is valid on the entire domain, we only ask for a Lipschitz constant $L_\xi(R)$ that depends on the norm of the input elements $u, v \in \mathcal{D}(\nabla f)$. This means in

particular that $L_\xi(R)$ may tend to infinity as $R \to \infty$. In the coming analysis we handle this by applying an a priori bound (Lemma 2) that shows that the solution is bounded and thus R is bounded too.

While the properness of F needs to be verified by application-specific means, the lower semi-continuity can be guaranteed on a more general level in different ways. If, e.g., it is additionally known that $\mathbf{E}_\xi\left[\inf_{u\in H} f(u,\xi)\right] > -\infty$ then one can employ Fatou's lemma ([22, Theorem 2.3.6]) as in [32, Lemma 5], or slightly modify [5, Corollary 9.4].

We note that from a function analytic point of view, we are dealing with bounded rather than unbounded operators ∇F. However, also operators that are traditionally seen as unbounded fit into the framework, given that the space H is chosen properly. For example, the functional $F(w) = \frac{1}{2}\int \|\nabla w\|^2$ corresponding to $\nabla F = -\Delta$, the negative Laplacian, is unbounded on $H = L^2$. But if we instead choose $H = H_0^1$, then $H^* = H^{-1}$ and ∇F is bounded and Lipschitz continuous. In this case, the splitting of $F(w)$ into $f(w,\xi^k)$ is less obvious than in our main application, but e.g. (randomized) domain decomposition as in [25] is a natural idea. In each step, an elliptic problem then has to be solved (to apply $\imath$), but this can often be done very efficiently.

Our main theorem states that we have sub-linear convergence of the iterates w^k to w^* in expectation:

Theorem 1 *Let Assumption 1 be fulfilled and let $\{\xi^k\}_{k\in\mathbb{N}}$ be a family of jointly independent random variables on Ω. Then the scheme (3) converges sub-linearly if the step sizes fulfill $\alpha_k = \frac{\eta}{k}$ with $\eta > \frac{1}{\mu}$. In particular, the error bound*

$$\mathbf{E}_{k-1}\left[\|w^k - w^*\|^2\right] \le \frac{C}{k}$$

is fulfilled, where C depends on $\|w_1 - w^*\|$, μ, ν, σ, η and m.

When $m = 1$, there is a L such that $L_\xi(R) \le L$ almost surely for all R and we have the explicit bound

$$C = \left(\|w^1 - w^*\|^2 + \frac{2^{\mu\eta}\eta^2}{\mu\eta - 1}\left(\sigma^2 + 2L\sigma\left(\|w^1 - w^*\|^2 + \sigma^2\sum_{j=1}^{k-1}\alpha_j^2\right)^{\frac{1}{2}}\right)\right)\exp\left(\frac{\nu^2\eta^2\pi^2}{4}\right).$$

For details on the error constant when $m > 1$, we refer the reader to the proof, which is given in Sect. 4. We note that there is no upper bound on the step size α_k, as would be the case for an explicit method like SGD. There is still a lower bound, but this is not as critical. Similarly to the finite-dimensional case (see e.g. [32, Theorem 15]), the method still converges if the assumption $\eta > \frac{1}{\mu}$ is not fulfilled, albeit at a slower rate $\mathcal{O}(1/k^\gamma)$ with $\gamma < 1$. This follows from a straightforward extension of Lemma 10 and the above theorem, but we omit these details for brevity. Moreover, we note that the exponential terms in the error constant are an artifact of the proof. They are not observed in practice and could likely be removed by the use of more refined algebraic inequalities.

The main idea of the proof is to acquire a contraction property of the form

$$\mathbf{E}_{k-1}\left[\|w^k - w^*\|^2\right] \leq C_k \mathbf{E}_{k-2}\left[\|w^{k-1} - w^*\|^2\right] + \alpha_k^2 D,$$

where $C_k < 1$ and D are certain constants depending on the data. Inevitably, $C_k \to 1$ as $k \to \infty$, but because of the chosen step size sequence this happens slowly enough to still guarantee the optimal rate. To reach this point, we first show two things: First, an a priori bound of the form $\mathbf{E}_{k-1}\left[\|w^k - w^*\|^2\right] \leq C$, i.e. unlike the SGD, the SPI is always stable regardless of how large the step size is. Secondly, that the resolvents $T_{f,\xi}$ are contractive with

$$\mathbf{E}_{\xi}\left[\|T_{f,\xi}u - T_{f,\xi}v\|^2\right] \leq C_k \|u - v\|^2.$$

Similarly to [32], we do the latter by approximating the functions $f(\cdot, \xi)$ by convex quadratic functions $\tilde{f}(\cdot, \xi)$ for which the property is easier to verify, and then establishing a relation between the approximated and the true contraction factors. The series of lemmas in the next section is devoted to this preparatory work.

3 Preliminaries

First, let us show that the scheme is in fact well-defined, in the sense that every iterate is measurable if the random variables $\{\xi^k\}_{k\in\mathbb{N}}$ are.

Lemma 1 *Let Assumption 1 be fulfilled. Further, let $\{\xi^k\}_{k\in\mathbb{N}}$ be a family of jointly independent random variables. Then for every $k \in \mathbb{N}$ there exists a unique mapping $w^{k+1} : \Omega \to \mathcal{D}(\nabla f)$ that fulfills (3) and is measurable with respect to the σ-algebra generated by $\xi^1, \ldots, \xi^k$.*

Proof We define the mapping

$$h : \mathcal{D}(\nabla f) \times \Omega \to H, \quad (u, \omega) \mapsto w^k - (I + \alpha_k \iota \nabla f(\cdot, \xi^k(\omega)))u.$$

For almost all $\omega \in \Omega$, the mapping $f(\cdot, \xi^k(\omega))$ is lower semi-continuous, proper and convex. Thus, by [27, Theorem A] $\nabla f(\cdot, \xi^k(\omega))$ is maximal monotone. By [4, Theorem 2.2], this shows that the operator $\iota^{-1} + \alpha_k \nabla f(\cdot, \xi^k(\omega)) : \mathcal{D}(\nabla f) \to H^*$ is surjective. Note that the two previously cited results are stated for multi-valued operators. As we are in a more regular setting, the sub-differential of $f(\cdot, \xi^k(\omega))$ only consists of a single element at each point. Therefore, it is possible to apply these multi-valued results also in our setting and interpret the appearing operators as single-valued. Furthermore, due to the monotonicity of $\nabla f(\cdot, \xi^k(\omega))$ it follows that for $u, v \in \mathcal{D}(\nabla f)$

$$\langle \big(\imath^{-1} + \alpha_k \nabla f(\cdot, \xi^k(\omega))\big)u - \big(\imath^{-1} + \alpha_k \nabla f(\cdot, \xi^k(\omega))\big)v, u - v\rangle \geq \|u - v\|^2$$

which implies

$$\Big\|\big(\imath^{-1} + \alpha_k \nabla f(\cdot, \xi^k(\omega))\big)u - \big(\imath^{-1} + \alpha_k \nabla f(\cdot, \xi^k(\omega))\big)v\Big\| \geq \|u - v\|.$$

This verifies that $I + \alpha_k \imath \nabla f(\cdot, \xi^k(\omega))$ is injective. As we have proved that the operator is both injective and surjective, it is, in particular, bijective. Therefore, there exists a unique element $w^{k+1}(\omega)$ such that

$$h(w^{k+1}(\omega), \omega) = w^k - (I + \alpha_k \imath \nabla f(\cdot, \xi^k(\omega)))w^{k+1}(\omega) = 0.$$

We can now apply [14, Lemma 2.1.4] or [15, Lemma 4.3] and obtain that $\omega \mapsto w^{k+1}(\omega)$ is measurable. □

Proving that the scheme is always stable is relatively straightforward, as shown in the next lemma. With some extra effort, we also get stability in stronger norms, i.e. we can bound not only $\mathbf{E}_k\big[\|w^{k+1} - w^*\|^2\big]$ but also higher moments $\mathbf{E}_k\big[\|w^{k+1} - w^*\|^{2^m}\big]$, $m \in \mathbb{N}$. This will be important since we only have the weaker local Lipschitz continuity stated in Assumption 1 rather than global Lipschitz continuity. The idea of the proof stems from a standard technique mostly applied in the field of evolution equations in a variational framework, compare for example [31, Lemma 8.6]. The main difficulty is to incorporate the stochastic gradient in the presentation.

Lemma 2 *Let Assumption 1 be fulfilled, and suppose that* $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$. *Then there exists a constant* $D \geq 0$ *depending only on* $\|w_1 - w^*\|$, $\sum_{k=1}^{\infty} \alpha_k^2$ *and* σ, *such that*

$$\mathbf{E}_k\big[\|w^{k+1} - w^*\|^{2^m}\big] \leq D$$

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

Proof Within the proof, we abbreviate the function $f(\cdot, \xi^k)$ by f_k, $k \in \mathbb{N}$. First, we consider the case $m = 1$. Recall the identity $(a - b, a) = \frac{1}{2}\big(\|a\|^2 - \|b\|^2 + \|a - b\|^2\big)$, $a, b \in H$. We write the scheme as

$$w^{k+1} - w^k + \alpha_k \imath \nabla f_k(w^{k+1}) = 0,$$

subtract $\alpha_k \imath \nabla f_k(w^*)$ from both sides, multiply by two and test it with $w^{k+1} - w^*$ to obtain

$$\begin{aligned}&\|w^{k+1} - w^*\|^2 - \|w^k - w^*\|^2 + \|w^{k+1} - w^k\|^2\\ &\quad + 2\alpha_k(\imath\nabla f_k(w^{k+1}) - \imath\nabla f_k(w^*), w^{k+1} - w^*)\\ &= -2\alpha_k(\imath\nabla f_k(w^*), w^{k+1} - w^*).\end{aligned}$$

For the right-hand side, we have by Young's inequality that

$$\begin{aligned}&-2\alpha_k(\imath\nabla f_k(w^*), w^{k+1} - w^*)\\ &\quad = -2\alpha_k\langle\nabla f_k(w^*), w^{k+1} - w^k\rangle - 2\alpha_k\langle\nabla f_k(w^*), w^k - w^*\rangle\\ &\quad \le 2\alpha_k\|\nabla f_k(w^*)\|_{H^*}\|w^{k+1} - w^k\| - 2\alpha_k\langle\nabla f_k(w^*), w^k - w^*\rangle\\ &\quad \le \alpha_k^2\|\nabla f_k(w^*)\|_{H^*}^2 + \|w^{k+1} - w^k\|^2 - 2\alpha_k\langle\nabla f_k(w^*), w^k - w^*\rangle.\end{aligned}$$

Together with the monotonicity condition, it then follows that

$$\|w^{k+1} - w^*\|^2 - \|w^k - w^*\|^2 \le \alpha_k^2\|\nabla f_k(w^*)\|_{H^*}^2 - 2\alpha_k\langle\nabla f_k(w^*), w^k - w^*\rangle. \tag{4}$$

Since $w^k - w^*$ is independent of ξ^k and $\mathbf{E}_{\xi^k}[\nabla f_k(w^*)] = \nabla F(w^*) = 0$, taking the expectation $\mathbf{E}_{\xi^k}$ thus leads to the following bound:

$$\mathbf{E}_{\xi^k}\left[\|w^{k+1} - w^*\|^2\right] \le \|w^k - w^*\|^2 + \alpha_k^2\sigma^2.$$

Repeating this argument, we obtain that

$$\mathbf{E}_k\left[\|w^{k+1} - w^*\|^2\right] \le \|w_1 - w^*\|^2 + \sigma^2\sum_{j=1}^{k}\alpha_j^2. \tag{5}$$

In order to find the higher moment bound, we recall (4). We then follow a similar idea as in [10, Lemma 3.1], where we multiply this inequality with $\|w^{k+1} - w^*\|^2$ and use the identity $(a-b)a = \frac{1}{2}\left(|a|^2 - |b|^2 + |a-b|^2\right)$ for $a, b \in \mathbb{R}$. It then follows that

$$\begin{aligned}&\|w^{k+1} - w^*\|^4 - \|w^k - w^*\|^4 + \left|\|w^{k+1} - w^*\|^2 - \|w^k - w^*\|^2\right|^2\\ &\quad \le \left(\alpha_k^2\|\nabla f_k(w^*)\|_{H^*}^2 - 2\alpha_k\langle\nabla f_k(w^*), w^k - w^*\rangle\right)\|w^{k+1} - w^*\|^2\\ &\quad \le \left(\alpha_k^2\|\nabla f_k(w^*)\|_{H^*}^2 - 2\alpha_k\langle\nabla f_k(w^*), w^k - w^*\rangle\right)\\ &\qquad \times \left(\|w^k - w^*\|^2 + \alpha_k^2\|\nabla f_k(w^*)\|_{H^*}^2 - 2\alpha_k\langle\nabla f_k(w^*), w^k - w^*\rangle\right)\\ &\quad \le \alpha_k^2\|w^k - w^*\|^2\|\nabla f_k(w^*)\|_{H^*}^2 - 2\alpha_k\|w^k - w^*\|^2\langle\nabla f_k(w^*), w^k - w^*\rangle\\ &\qquad + \alpha_k^4\|\nabla f_k(w^*)\|_{H^*}^4 - 4\alpha_k^3\|\nabla f_k(w^*)\|_{H^*}^2\langle\nabla f_k(w^*), w^k - w^*\rangle\\ &\qquad + 4\alpha_k^2\left(\langle\nabla f_k(w^*), w^k - w^*\rangle\right)^2.\end{aligned}$$

Applying Young's inequality to the first and fourth term of the previous row then implies that

$$
\begin{aligned}
&\|w^{k+1}-w^*\|^4-\|w^k-w^*\|^4\\
&\quad\le \frac{\alpha_k^2}{2}\|w^k-w^*\|^4-2\alpha_k\|w^k-w^*\|^2\langle\nabla f_k(w^*),w^k-w^*\rangle\\
&\qquad+\left(3\alpha_k^4+\frac{\alpha_k^2}{2}\right)\|\nabla f_k(w^*)\|_{H^*}^4+6\alpha_k^2\|\nabla f_k(w^*)\|_{H^*}^2\|w^k-w^*\|^2\\
&\quad\le \frac{\alpha_k^2}{2}\|w^k-w^*\|^4-2\alpha_k\|w^k-w^*\|^2\langle\nabla f_k(w^*),w^k-w^*\rangle\\
&\qquad+\left(3\alpha_k^4+\frac{\alpha_k^2}{2}\right)\|\nabla f_k(w^*)\|_{H^*}^4+3\alpha_k^2\|\nabla f_k(w^*)\|_{H^*}^4+3\alpha_k^2\|w^k-w^*\|^4\\
&\quad\le \frac{7\alpha_k^2}{2}\|w^k-w^*\|^4-2\alpha_k\|w^k-w^*\|^2\langle\nabla f_k(w^*),w^k-w^*\rangle\\
&\qquad+\left(3\alpha_k^4+\frac{7\alpha_k^2}{2}\right)\|\nabla f_k(w^*)\|_{H^*}^4.
\end{aligned}
$$

Summing up from $j=1$ to k and taking the expectation $\mathbf{E}_k$, yields

$$
\begin{aligned}
&\mathbf{E}_k\left[\|w^{k+1}-w^*\|^4\right]\\
&\quad\le \|w_1-w^*\|^4+\sum_{j=1}^k\frac{7\alpha_j^2}{2}\mathbf{E}_{j-1}\left[\|w^j-w^*\|^4\right]+\sigma^4\sum_{j=1}^k\left(3\alpha_j^4+\frac{7\alpha_j^2}{2}\right).
\end{aligned}
$$

We then apply the discrete Grönwall inequality for sums (see, e.g., [11]) which shows that

$$
\mathbf{E}_k\left[\|w^{k+1}-w^*\|^4\right]\le\left(\|w_1-w^*\|^4+\sigma^4\sum_{j=1}^k\left(3\alpha_j^4+\frac{7\alpha_j^2}{2}\right)\right)\exp\left(\frac{7}{2}\sum_{j=1}^k\alpha_j^2\right).
$$

For the next higher bound $\mathbf{E}_k\left[\|w^{k+1}-w^*\|^8\right]$, we recall that

$$
\begin{aligned}
&\|w^{k+1}-w^*\|^4-\|w^k-w^*\|^4\\
&\quad\le \frac{7\alpha_k^2}{2}\|w^k-w^*\|^4-2\alpha_k\|w^k-w^*\|^2\langle\nabla f_k(w^*),w^k-w^*\rangle\\
&\qquad+\left(3\alpha_k^4+\frac{7\alpha_k^2}{2}\right)\|\nabla f_k(w^*)\|_{H^*}^4,
\end{aligned}
$$

which we can multiply with $\|w^{k+1}-w^*\|^4$ in order to follow the same strategy as before. Following this approach, we find bounds for $\mathbf{E}_k\left[\|w^{k+1}-w^*\|^{2^m}\right]$ recursively for all $m\in\mathbb{N}$. □

Remark 3 In particular, Lemma 2 implies that there exists a constant D depending on $\|w_1-w^*\|$, $\sum_{k=1}^\infty\alpha_k^2$ and σ such that

$$\mathbf{E}_k\left[\|w^{k+1} - w^*\|^p\right] \leq D$$

for all $p \leq 2^m$ and $k \in \mathbb{N}$. Further, comparing (5)

$$\mathbf{E}_k\left[\|w^{k+1} - w^*\|^2\right] \leq \|w_1 - w^*\| + \sum_{i=1}^{k} \alpha_i^2 \mathbf{E}_{\xi^i}\left[\|\nabla f(w^*, \xi^i)\|^2\right],$$

to the corresponding bound for the SGD

$$\mathbf{E}_k\left[\|w^{k+1} - w^*\|^2\right] \leq \|w_1 - w^*\| + \sum_{i=1}^{k} \alpha_i^2 \mathbf{E}_i\left[\|\nabla f(w^i, \xi^i)\|^2\right],$$

indicates that the SPI has a smaller a priori bound than the SGD. This bound plays a crucial part in the error constant in the convergence proof of Theorem 1. In practice one would expect the terms $\mathbf{E}_{\xi^i}\left[\|\nabla f(w^*, \xi^i)\|^2\right]$ to be significantly smaller than $\mathbf{E}_i\left[\|\nabla f_i(w^i, \xi^i)\|^2\right]$ if the variance of $\nabla f(\cdot, \xi^i)$ is small. Note that since we assume that we have an unbiased estimate, the variance is given by $\mathbf{E}_{\xi^i}\left[\|\nabla f(w, \xi^i)\|^2\right] - \|\mathbf{E}_{\xi^i}\left[\nabla f(w, \xi^i)\right]\|^2 = \mathbf{E}_{\xi^i}\left[\|\nabla f(w, \xi^i)\|^2\right]$.

Following Ryu and Boyd [32], we now introduce the function $\tilde{f}(\cdot, \xi) : H \times \Omega \to (-\infty, \infty]$ given by

$$\tilde{f}(u, \xi) = f(u_0, \xi) + \langle \nabla f(u_0, \xi), u - u_0 \rangle + \frac{1}{2}(M_\xi(u - u_0), u - u_0), \tag{6}$$

where $u_0 \in \mathcal{D}(\nabla f)$ is a fixed parameter. This mapping is a convex approximation of f. Furthermore, we define the function $\tilde{r}(\cdot, \xi) : H \times \Omega \to (-\infty, \infty]$ given by

$$\tilde{r}(u, \xi) = f(u, \xi) - \tilde{f}(u, \xi). \tag{7}$$

Their gradients $\nabla\tilde{f}(\cdot, \xi) : H \times \Omega \to H^*$ and $\nabla\tilde{r}(\cdot, \xi) : \mathcal{D}(\nabla f) \times \Omega \to H^*$ can be stated as

$$\begin{aligned}
\nabla\tilde{f}(u, \xi) &= \nabla f(u_0, \xi) + (M_\xi(u - u_0), \cdot), \quad u \in H, \\
\nabla\tilde{r}(u, \xi) &= \nabla f(u, \xi) - \nabla f(u_0, \xi) - (M_\xi(u - u_0), \cdot), \quad u \in \mathcal{D}(\nabla f)
\end{aligned}$$

almost surely. In the following lemma, we collect some standard properties of these operators.

Lemma 3 *The function $\tilde{r}(\cdot, \xi)$ defined in (7) is convex almost surely, i.e., it fulfills $\tilde{r}(u, \xi) \geq \tilde{r}(v, \xi) + \langle \nabla\tilde{r}(v, \xi), u - v \rangle$ for all $u, v \in \mathcal{D}(\nabla f)$ almost surely. As a consequence, the gradient $\nabla\tilde{r}(\cdot, \xi)$ is monotone almost surely.*

Proof In the following proof, let us omit ξ for simplicity and let $u, v \in \mathcal{D}(\nabla f)$ be given. Due to the monotonicity property of ∇f stated in Assumption 1, it follows that

$$f(u) \geq f(v) + \langle \nabla f(v), u - v \rangle + \frac{1}{2}(M(u-v), u-v).$$

For the function $\tilde{f}$ we can write

$$\begin{aligned} \tilde{f}(u) &= f(u_0) + \langle \nabla f(u_0), u - u_0 \rangle + \frac{1}{2}(M(u-u_0), u-u_0), \\ \nabla \tilde{f}(u) &= \nabla f(u_0) + (M(u-u_0), \cdot) \quad \text{and} \quad \nabla^2 \tilde{f}(u) = M. \end{aligned}$$

All further derivatives are zero. Thus, we can use a Taylor expansion around v to write

$$\tilde{f}(u) = \tilde{f}(v) + \langle \nabla \tilde{f}(v), u - v \rangle + \frac{1}{2}(M(u-v), u-v).$$

It then follows that

$$\begin{aligned} \tilde{r}(u) &\geq f(v) + \langle \nabla f(v), u - v \rangle + \frac{1}{2}(M(u-v), u-v) \\ &\quad - \left(\tilde{f}(v) + \langle \nabla \tilde{f}(v), u - v \rangle + \frac{1}{2}(M(u-v), u-v) \right) \\ &= \tilde{r}(v) + \langle \nabla \tilde{r}(v), u - v \rangle. \end{aligned}$$

By [41, Proposition 25.10], it follows that $\nabla \tilde{r}$ is monotone. □

The following lemma demonstrates that the resolvents $T_{\tilde{f},\xi}$ and certain perturbations of them are well-defined. Furthermore, we will provide a more explicit formula for such resolvents. A comparable result is mentioned in [32, page 10], we include a proof for the sake of completeness.

Lemma 4 *Let Assumption 1 be fulfilled and let $\tilde{f}(\cdot, \xi)$ be defined as in (6). Then the operator*

$$T_{\tilde{f},\xi} = (I + \imath \nabla \tilde{f}(\cdot, \xi))^{-1} : H \times \Omega \to H$$

is well-defined. If a function $r(\cdot, \xi) : H \times \Omega \to (-\infty, \infty]$ is Gâteaux differentiable with the common domain $\mathcal{D}(\nabla r) = \mathcal{D}(\nabla f)$, lower semi-continuous, convex and proper almost surely, then

$$T_{\tilde{f}+r,\xi} = (I + \imath \nabla \tilde{f}(\cdot, \xi) + \imath \nabla r(\cdot, \xi))^{-1} : H \times \Omega \to \mathcal{D}(\nabla f)$$

is well-defined.

If there exist $Q_\xi : \mathcal{D}(\nabla f) \times \Omega \to H^*$ and $z_\xi : \Omega \to H^*$ such that $\nabla r(u, \xi) = Q_\xi u + z_\xi$ then the resolvent can be represented by

$$T_{\tilde{f}+r,\xi} u = (I + M_\xi + \imath Q_\xi)^{-1} \left(u - \imath \nabla f(u_0, \xi) + M_\xi u_0 - \imath z_\xi \right).$$

Proof For simplicity, let us omit ξ again. In order to prove that $T_{\tilde{f}}$ and $T_{\tilde{f}+r}$ are well-defined, we can apply [27, Theorem A] and [4, Theorem 2.2] analogously to the argumentation in the proof of Lemma 1.

Assuming that $\nabla r(u) = Qu + z$, we find an explicit representation for $T_{\tilde{f}+r}$. To this end, for $v \in H$, consider

$$(I + \imath\nabla\tilde{f} + \imath\nabla r)^{-1}v = T_{\tilde{f}+r}v =: u \in \mathcal{D}(\nabla f).$$

Then it follows that

$$v = (I + \imath\nabla\tilde{f} + \imath\nabla r)u = (I + M + \imath Q)u + \imath\nabla f(u_0) - Mu_0 + \imath z.$$

Rearranging the terms, yields

$$T_{\tilde{f}+r}v = (I + M + \imath Q)^{-1}\big(v - \imath\nabla f(u_0) + Mu_0 - \imath z\big).$$

□

Next, we will show that the contraction factors of $T_{f,\xi}$ and $T_{\tilde{f},\xi}$ are related. For this, we need the following basic identities and some stronger inequalities that hold for symmetric positive operators on H. These results are fairly standard and similar statements can be found in [32, Lemma 9 and Lemma 10]. For the sake of completeness, we provide an alternative proof that is better adapted to our notation.

Lemma 5 *Let Assumption 1 be satisfied and let $\tilde{f}(\cdot,\xi)$ and $\tilde{r}(\cdot,\xi)$ be given as in (6) and (7), respectively. Then the identities*

$$\imath\nabla f(T_{f,\xi},\xi) = I - T_{f,\xi} \quad \text{and} \quad \imath\nabla\tilde{f}(T_{f,\xi},\xi) + T_{f,\xi} - I = -\imath\nabla\tilde{r}(T_{f,\xi},\xi)$$

are fulfilled almost surely.

Proof By the definition of $T_{f,\xi}$, we have that

$$T_{f,\xi} + \imath\nabla f(T_{f,\xi},\xi) = (I + \imath\nabla f(\cdot,\xi))T_{f,\xi} = I,$$

from which the first claim follows immediately. The second identity then follows from

$$\imath\nabla\tilde{f}(T_{f,\xi},\xi) + T_{f,\xi} - I = \imath\nabla\tilde{f}(T_{f,\xi},\xi) - \imath\nabla f(T_{f,\xi},\xi) = -\imath\nabla\tilde{r}(T_{f,\xi},\xi).$$

□

As a consequence of Lemma 5 we have the following basic inequalities:

Lemma 6 *Let Assumption 1 be satisfied. It then follows that*

$$\|T_{f,\xi}u - u\| \le \|\nabla f(u,\xi)\|_{H^*}$$

almost surely for every $u \in \mathcal{D}(\nabla f)$. *Additionally, if for* $R > 0$ *the bound* $\|u\| + \|\nabla f(u, \xi)\| \leq R$ *holds true almost surely, then*

$$\|\iota^{-1}(T_{f,\xi}u - u) + \nabla f(u, \xi)\|_{H^*} \leq L_\xi(R)\|\nabla f(u, \xi)\|_{H^*}$$

is fulfilled almost surely.

Proof In order to shorten the notation, we omit the ξ in the following proof and let u be in $\mathcal{D}(\nabla f)$. For the first inequality, we note that since ∇f is monotone, we have

$$\langle \nabla f(T_f u) - \nabla f(u), T_f u - u \rangle \geq 0.$$

Thus, by the first identity in Lemma 5, it follows that

$$\begin{aligned}\langle -\nabla f(u), T_f u - u \rangle &= \langle \nabla f(T_f u) - \nabla f(u), T_f u - u \rangle - \langle \nabla f(T_f u), T_f u - u \rangle \\ &\geq \langle \iota^{-1}(T_f u - u), T_f u - u \rangle \\ &= (T_f u - u, T_f u - u) = \|T_f u - u\|^2.\end{aligned}$$

But by the Cauchy-Schwarz inequality, we also have

$$\langle -\nabla f(u), T_f u - u \rangle \leq \|\nabla f(u)\|_{H^*}\|T_f u - u\|,$$

which in combination with the previous inequality proves the first claim.

The second inequality follows from the first part of this lemma. Because

$$\|T_f u\| \leq \|T_f u - u\| + \|u\| \leq \|\nabla f(u)\|_{H^*} + \|u\|,$$

both u and $T_f u$ are in a ball of radius R. Thus, we obtain

$$\begin{aligned}\|\iota^{-1}(T_f u - u) + \nabla f(u)\|_{H^*} &= \|\nabla f(u) - \nabla f(T_f u)\|_{H^*} \\ &\leq L(R)\|u - T_f u\| \leq L(R)\|\nabla f(u)\|_{H^*}.\end{aligned}$$

Lemma 7 *Let* $Q, S \in \mathcal{L}(H)$ *be symmetric operators. Then the following holds*:

- If Q is invertible and S and Q^{-1} are strictly positive, then $(Q + S)^{-1} < Q^{-1}$. If S is only positive, then $(Q + S)^{-1} \leq Q^{-1}$.
- If Q is a positive and contractive operator, i.e. $\|Qu\| \leq \|u\|$ for all $u \in H$, then it follows that $\|Qu\|^2 \leq (Qu, u)$ for all $u \in H$.
- If Q is a strongly positive invertible operator, such that there exists $\beta > 0$ with $(Qu, u) \geq \beta\|u\|^2$ for all $u \in H$, then $\|Qu\| \geq \beta\|u\|$ for all $u \in H$ and $\|Q^{-1}\|_{\mathcal{L}(H)} \leq \frac{1}{\beta}$.

Proof We start by expressing $(Q + S)^{-1}$ in terms of Q^{-1} and S, similar to the Sherman-Morrison-Woodbury formula for matrices [18]. First observe that the operator $(I + Q^{-1}S)^{-1} \in \mathcal{L}(H)$ by e.g. [19, Lemma 2A.1]. Then, since

$$\begin{aligned}&\left(Q^{-1} - Q^{-1}S\left(I + Q^{-1}S\right)^{-1}Q^{-1}\right)(Q+S)\\ &\quad = I + Q^{-1}S - Q^{-1}S\left(I + Q^{-1}S\right)^{-1}\left(I + Q^{-1}S\right) = I\end{aligned}$$

and

$$\begin{aligned}&(Q+S)\left(Q^{-1} - Q^{-1}S\left(I + Q^{-1}S\right)^{-1}Q^{-1}\right)\\ &\quad = I + SQ^{-1} - S\left(I + Q^{-1}S\right)\left(I + Q^{-1}S\right)^{-1}Q^{-1} = I,\end{aligned}$$

we find that

$$(Q+S)^{-1} = Q^{-1} - Q^{-1}S\left(I + Q^{-1}S\right)^{-1}Q^{-1}.$$

Since Q^{-1} is symmetric, we see that $(Q+S)^{-1} < Q^{-1}$ if and only if $S\left(I + Q^{-1}S\right)^{-1}$ is strictly positive. But this is true, as we see from the change of variables $z = (I + Q^{-1}S)^{-1}u$. Because then

$$\left(S\left(I + Q^{-1}S\right)^{-1}u, u\right) = \left(Sz, z + Q^{-1}Sz\right) = (Sz, z) + \left(Q^{-1}Sz, Sz\right) > 0$$

for any $u \in H$, $u \neq 0$, since S and Q^{-1} are strictly positive. If S is only positive, it follows analogously that $\left(S\left(I + Q^{-1}S\right)^{-1}u, u\right) \geq 0$.

In order to prove the second statement, we use the fact that there exists a unique symmetric and positive square root $Q^{1/2} \in \mathcal{L}(H)$ such that $Q = Q^{1/2}Q^{1/2}$. Since $\|Q\| = \sup_{x\in H}(Qx, x) = \sup_{x\in H}(Q^{\frac{1}{2}}x, Q^{\frac{1}{2}}x) = \|Q^{1/2}\|^2$, also $Q^{1/2}$ is contractive. Thus, it follows that

$$\|Qu\|^2 = \|Q^{1/2}Q^{1/2}u\|^2 \leq \|Q^{1/2}u\|^2 = (Q^{1/2}u, Q^{1/2}u) = (Qu, u).$$

Now, we prove the third statement. First we notice that $(Qu, u) \geq \beta\|u\|^2$ and $(Qu, u) \leq \|Qu\|\|u\|$ imply that $\|Qu\| \geq \beta\|u\|$ for all $u \in H$. Substituting $v = Q^{-1}u$, then shows $\|v\| \geq \beta\|Q^{-1}v\|$, which proves the final claim. □

The previous lemma now allows us to extend [32, Theorem 10], which we have reformulated and restructured to match our setting. It relates the contraction factors of the true and approximated operators.

Lemma 8 *Let Assumption 1 be fulfilled and let $\tilde{f}(\cdot, \xi)$ be given as in (6). Then*

$$\mathbf{E}_\xi\left[\frac{\|T_{f,\xi}u - T_{f,\xi}v\|^2}{\|u - v\|^2}\right] \leq \left(\mathbf{E}_\xi\left[\frac{\|T_{\tilde{f},\xi}u - T_{\tilde{f},\xi}v\|^2}{\|u - v\|^2}\right]\right)^{1/2}$$

holds for every $u, v \in H$.

Proof For better readability, we once again omit ξ where there is no risk of confusion. For $u, v \in \mathcal{D}(\nabla f)$ with $u \neq v$ and $\varepsilon > 0$, we approximate the function $\tilde{r}(\cdot, \xi)$ defined in (7) by

$$\tilde{r}_\varepsilon(\cdot, \xi) : H \times \Omega \to (-\infty, \infty], \quad \tilde{r}_\varepsilon(z, \xi) = \langle \nabla \tilde{r}(T_f u, \xi), z \rangle + \frac{\big(\langle v_\varepsilon, z - T_f u \rangle\big)^2}{2a_\varepsilon},$$

where

$$v_\varepsilon = -\nabla \tilde{r}(T_f u) + \nabla \tilde{r}(T_f v) + \varepsilon \imath^{-1}(T_f v - T_f u) \in H \quad \text{and} \quad a_\varepsilon = \langle v_\varepsilon, T_f v - T_f u \rangle.$$

As we can write

$$\begin{aligned} a_\varepsilon &= \langle -\nabla \tilde{r}(T_f u) + \nabla \tilde{r}(T_f v) + \varepsilon \imath^{-1}(T_f v - T_f u), T_f v - T_f u \rangle \\ &= \langle \nabla \tilde{r}(T_f u) - \nabla \tilde{r}(T_f v), T_f u - T_f v \rangle + \varepsilon (T_f v - T_f u, T_f v - T_f u) \\ &\geq \varepsilon \|T_f v - T_f u\|^2 > 0, \end{aligned}$$

$\tilde{r}_\varepsilon$ is well-defined. The derivative is given by $\nabla \tilde{r}_\varepsilon(\cdot, \xi) : H \times \Omega \to H^*$,

$$\nabla \tilde{r}_\varepsilon(z) = \nabla \tilde{r}(T_f u) + \frac{\langle v_\varepsilon, z - T_f u \rangle}{a_\varepsilon} v_\varepsilon = \frac{\langle v_\varepsilon, z \rangle}{a_\varepsilon} v_\varepsilon + \nabla \tilde{r}(T_f u) - \frac{\langle v_\varepsilon, T_f u \rangle}{a_\varepsilon} v_\varepsilon.$$

This function $\nabla \tilde{r}_\varepsilon$ is an interpolation between the points

$$\begin{aligned} \nabla \tilde{r}_\varepsilon(T_f u) &= \nabla \tilde{r}(T_f u) \quad \text{and} \\ \nabla \tilde{r}_\varepsilon(T_f v) &= \nabla \tilde{r}(T_f u) + \frac{\langle v_\varepsilon, T_f v - T_f u \rangle}{a_\varepsilon} v_\varepsilon \\ &= \nabla \tilde{r}(T_f u) + \frac{\langle v_\varepsilon, T_f v - T_f u \rangle}{\langle v_\varepsilon, T_f v - T_f u \rangle} v_\varepsilon \\ &= \nabla \tilde{r}(T_f u) - \nabla \tilde{r}(T_f u) + \nabla \tilde{r}(T_f v) + \varepsilon \imath^{-1}(T_f v - T_f u) \\ &= \nabla \tilde{r}(T_f v) + \varepsilon \imath^{-1}(T_f v - T_f u). \end{aligned}$$

Furthermore, since $T_{\tilde{f}+\tilde{r}_\varepsilon} = (I + \imath \nabla \tilde{f} + \imath \nabla \tilde{r}_\varepsilon)^{-1}$, it follows that

$$\begin{aligned} (I + \imath \nabla \tilde{f} + \imath \nabla \tilde{r}_\varepsilon) T_f u &= T_f u + \imath \nabla \tilde{f}(T_f u) + \imath \nabla \tilde{r}(T_f u) \\ &= T_f u + \imath \nabla f(T_f u) = (I + \imath \nabla f) T_f u = u, \end{aligned}$$

and therefore

$$T_f u = (I + \imath \nabla \tilde{f} + \imath \nabla \tilde{r}_\varepsilon)^{-1} u = T_{\tilde{f}+\tilde{r}_\varepsilon} u.$$

Applying Lemma 5, we find that

$$
\begin{aligned}
&(I + \imath\nabla\tilde{f} + \imath\nabla\tilde{r}_\varepsilon)T_f v \\
&\quad = T_f v + \imath\nabla\tilde{f}(T_f v) + \imath\nabla\tilde{r}(T_f v) + \varepsilon(T_f v - T_f u) \\
&\quad = T_f v + \imath\nabla f(T_f v) + \varepsilon(T_f v - T_f u) = v + \varepsilon(T_f v - T_f u).
\end{aligned}
$$

This shows that

$$
T_f v = (I + \imath\nabla\tilde{f} + \imath\nabla\tilde{r}_\varepsilon)^{-1}(v + \varepsilon(T_f v - T_f u)) = T_{\tilde{f}+\tilde{r}_\varepsilon}(v + \varepsilon(T_f v - T_f u)). \tag{8}
$$

Using the explicit representation of $T_{\tilde{f}+\tilde{r}_\varepsilon}$ from Lemma 4, it follows that

$$
T_{\tilde{f}+\tilde{r}_\varepsilon} z = \left(I + M + \imath\left(\frac{\langle v_\varepsilon, \cdot\rangle}{a_\varepsilon} v_\varepsilon\right)\right)^{-1}\left(z - \imath\nabla f(u_0) + Mu_0 - \imath\left(\nabla\tilde{r}(T_f u) - \frac{\langle v_\varepsilon, T_f u\rangle}{a_\varepsilon} v_\varepsilon\right)\right).
$$

Therefore, we have

$$
\begin{aligned}
&\|T_{\tilde{f}+\tilde{r}_\varepsilon} v - T_{\tilde{f}+\tilde{r}_\varepsilon}(v + \varepsilon(T_f v - T_f u))\| \\
&\quad \le \left\| \left(I + M + \imath\left(\frac{\langle v_\varepsilon, \cdot\rangle}{a_\varepsilon} v_\varepsilon\right)\right)^{-1} \right\|_{\mathcal{L}(H)} \|v - v - \varepsilon(T_f v - T_f u)\| \\
&\quad \le \varepsilon\|T_f v - T_f u\| \to 0 \quad \text{as } \varepsilon \to 0,
\end{aligned}
$$

since

$$
\left(\left(I + M + \imath\left(\frac{\langle v_\varepsilon, \cdot\rangle}{a_\varepsilon} v_\varepsilon\right)\right)u, u\right) \ge \|u\|^2
$$

means that we can apply Lemma 7. Thus, this shows that $T_f u = T_{f+\tilde{r}_\varepsilon} u$ and $T_f v = \lim_{\varepsilon\to 0} T_{\tilde{f}+\tilde{r}_\varepsilon} v$. Further, we can state an explicit representation for $T_{\tilde{f}}$ using Lemma 4 given by

$$
T_{\tilde{f}} z = (I + \imath\nabla\tilde{f})^{-1} z = (I + M)^{-1}\big(z - \imath\nabla f(u_0) + Mu_0\big).
$$

For $n = \frac{u-v}{\|u-v\|}$ with $\|n\| = 1$, we obtain using Lemma 7

$$
\begin{aligned}
\frac{\|T_{\tilde{f}} u - T_{\tilde{f}} v\|}{\|u - v\|} &= \|(I + M)^{-1} n\| \\
&\ge ((I + M)^{-1} n, n) \\
&\ge \left(\left(I + M + \imath\left(\frac{\langle v_\varepsilon, \cdot\rangle}{a_\varepsilon} v_\varepsilon\right)\right)^{-1} n, n\right) \\
&\ge \left\| \left(I + M + \imath\left(\frac{\langle v_\varepsilon, \cdot\rangle}{a_\varepsilon} v_\varepsilon\right)\right)^{-1} n \right\|^2 \\
&= \frac{\|T_{\tilde{f}+\tilde{r}_\varepsilon} u - T_{\tilde{f}+\tilde{r}_\varepsilon} v\|^2}{\|u - v\|^2} \to \frac{\|T_f u - T_f v\|^2}{\|u - v\|^2} \quad \text{as } \varepsilon \to 0.
\end{aligned}
$$

Finally, as $\mathbf{E}_\xi\left[\frac{\|T_{\tilde{f}}u-T_{\tilde{f}}v\|}{\|u-v\|}\right]$ is finite, we can apply the dominated convergence theorem to obtain that

$$\mathbf{E}_\xi\left[\frac{\|T_f u - T_f v\|^2}{\|u-v\|^2}\right] \leq \mathbf{E}_\xi\left[\frac{\|T_{\tilde{f}} u - T_{\tilde{f}} v\|}{\|u-v\|}\right] \leq \left(\mathbf{E}_\xi\left[\frac{\|T_{\tilde{f}} u - T_{\tilde{f}} v\|^2}{\|u-v\|^2}\right]\right)^{\frac{1}{2}}.$$

□

After having established a connection between the contraction properties of $T_{f,\xi}$ and $T_{\tilde{f},\xi}$, the next step is to provide a concrete result for the contraction factor of $T_{\tilde{f},\xi}$. Applying Lemma 4, we can express this resolvent in terms of M_ξ, which is easier to handle due to its linearity. The following lemma extends [32, Theorem 11]. As we are in an infinite dimensional setting, we can no longer argue using the smallest eigenvalue of an operator. This proof instead uses the convexity parameters directly. Moreover, we provide an explicit, non-asymptotic, bound for the contraction constant.

Lemma 9 *Let Assumption 1 be satisfied and let $\tilde{f}(\cdot,\xi)$ be given as in (6). Then for $u, v \in H$ and $\alpha > 0$,*

$$\mathbf{E}_\xi\left[\|T_{\alpha\tilde{f},\xi}u - T_{\alpha\tilde{f},\xi}v\|^2\right] < \mathbf{E}_\xi\left[\|(I+\alpha M_\xi)^{-1}\|^2_{\mathcal{L}(H)}\right]\|u-v\|^2$$

is fulfilled. Furthermore, it follows that

$$\mathbf{E}_\xi\left[\|(I+\alpha M_\xi)^{-1}\|^2_{\mathcal{L}(H)}\right] < 1 - 2\mu\alpha + 3\nu^2\alpha^2.$$

Proof Due to the explicit representation of $T_{\alpha\tilde{f},\xi}$ stated in Lemma 4, we find that

$$T_{\alpha\tilde{f},\xi}u - T_{\alpha\tilde{f},\xi}v = (I+\alpha M_\xi)^{-1}(u-v)$$

for $u, v \in H$. As $u - v$ does not depend on Ω, it follows that

$$\mathbf{E}_\xi\left[\|(I+\alpha M_\xi)^{-1}(u-v)\|^2\right] \leq \mathbf{E}_\xi\left[\|(I+\alpha M_\xi)^{-1}\|^2_{\mathcal{L}(H)}\right]\|u-v\|^2.$$

Thus, we have reduced the problem to a question about "how contractive" the resolvent of M_ξ is in expectation. We note that for any $u \in H$, we have

$$((I+\alpha M_\xi)u, u) \geq (1+\mu_\xi\alpha)\|u\|^2.$$

Due to Lemma 7 it follows that

$$\|(I+\alpha M_\xi)^{-1}\|^2_{\mathcal{L}(H)} \leq (1+\mu_\xi\alpha)^{-2}.$$

The right-hand-side bound is a $C^2(-\frac{1}{\mu_\xi}, \infty)$-function with respect to α or even a $C^2(\mathbb{R})$-function if $\mu_\xi = 0$. By a second-order expansion in a Taylor series we can therefore conclude that

$$\|(I + \alpha M_\xi)^{-1}\|^2_{\mathcal{L}(H)} \leq 1 - 2\mu_\xi \alpha + 3\mu_\xi^2 \alpha^2.$$

Combining these results, we obtain

$$\mathbf{E}_\xi \Big[\|(I + \alpha M_\xi)^{-1}\|^2_{\mathcal{L}(H)} \Big] \leq \mathbf{E}_\xi \Big[1 - 2\mu_\xi \alpha + 3\mu_\xi^2 \alpha^2 \Big] = 1 - 2\mu\alpha + 3\nu^2 \alpha^2.$$

□

Finally, the proof of the main theorem relies on iterating the step-wise bounds arising from the contraction properties of the resolvents which we just established. This leads to certain products of the contraction factors. The following algebraic inequalities show that these are bounded in the desired way. While this type of result has been stated previously for first-order polynomials in $1/j$ (see e.g. [24, Theorem 14]), we prove here a particular version for second-order polynomials that matches the approximation of the contraction factor stated in Lemma 9.

Lemma 10 *Let $C_1, C_2 > 0$, $p > 0$ and $r \geq 0$ satisfy $C_1 p > r$ and $4C_2 \geq C_1^2$. Then the following inequalities are satisfied*:

(i) $\prod_{j=1}^k \Big(1 - \frac{C_1}{j} + \frac{C_2}{j^2}\Big)^p \leq \exp\Big(\frac{C_2 p \pi^2}{6}\Big)(k+1)^{-C_1 p}$,

(ii) $\sum_{j=1}^k \frac{1}{j^{1+r}} \prod_{i=j+1}^k \Big(1 - \frac{C_1}{i} + \frac{C_2}{i^2}\Big)^p \leq 2^{C_1 p} \exp\Big(\frac{C_2 p \pi^2}{6}\Big) \frac{1}{C_1 p - r}(k+1)^{-r}$.

Proof The proof relies on the trivial inequality $1 + u \leq \mathrm{e}^u$ for $u \geq -1$ and the following two basic inequalities involving (generalized) harmonic numbers

$$\ln(k+1) - \ln(m) \leq \sum_{i=m}^k \frac{1}{i} \quad \text{and} \quad \sum_{i=1}^k i^{C-1} \leq \frac{1}{C}(k+1)^C.$$

The first one follows quickly by treating the sum as a lower Riemann sum approximating the integral $\int_m^{k+1} u^{-1}\,\mathrm{d}u$. The second one can be proved analogously by approximating the integral $\int_0^{k+1} u^{C-1}\,\mathrm{d}u$ with an upper ($C < 1$) or lower ($C > 1$) Riemann sum.

The condition $4C_2 \geq C_1^2$ implies that all the factors in the product *(i)* are positive. We therefore have that $0 \leq 1 - \frac{C_1}{j} + \frac{C_2}{j^2} \leq \exp\big(-\frac{C_1}{j}\big)\exp\big(\frac{C_2}{j^2}\big)$. Thus, it follows that

$$\prod_{j=1}^{k}\left(1-\frac{C_1}{j}+\frac{C_2}{j^2}\right)^p \le \exp\left(-C_1p\sum_{j=1}^{k}\frac{1}{j}\right)\exp\left(C_2p\sum_{j=1}^{k}\frac{1}{j^2}\right)$$
$$\le \exp\big(-C_1p\ln(k+1)\big)\exp\left(\frac{C_2p\pi^2}{6}\right),$$

from which the first claim follows directly. For the second claim, we similarly have

$$\sum_{j=1}^{k}\frac{1}{j^{1+r}}\prod_{i=j+1}^{k}\left(1-\frac{C_1}{i}+\frac{C_2}{i^2}\right)^p \le \exp\left(\frac{C_2p\pi^2}{6}\right)\sum_{j=1}^{k}\frac{1}{j^{1+r}}\exp\left(-C_1p\sum_{i=j+1}^{k}\frac{1}{i}\right),$$

where the latter sum can be bounded by

$$\sum_{j=1}^{k}\frac{1}{j^{1+r}}\exp\left(-C_1p\sum_{i=j+1}^{k}\frac{1}{i}\right) \le \sum_{j=1}^{k}\frac{1}{j^{1+r}}\exp\left(-C_1p\ln\left(\frac{k+1}{j+1}\right)\right)$$
$$\le \sum_{j=1}^{k}\frac{1}{j^{1+r}}\left(\frac{k+1}{j+1}\right)^{-C_1p}$$
$$= (k+1)^{-C_1p}\sum_{j=1}^{k}j^{C_1p-r-1}\cdot\left(\frac{j+1}{j}\right)^{C_1p}$$
$$\le \frac{2^{C_1p}}{C_1p-r}(k+1)^{-r}.$$

The final inequality is where we needed $C_1p > r$, in order to have something better than j^{-1} in the sum. □

4 Proof of main theorem

Using the lemmas presented in the previous section, we are now in a position to prove Theorem 1. Compared to the earlier results in the literature, we can provide a more general result with respect to the Lipschitz condition. More precisely, with the help of our a priori bound from Lemma 2, we can exchange the global Lipschitz condition by a local Lipschitz condition.

Proof of Theorem 1 Given the sequence of mutually independent random variables ξ^k, we abbreviate the random functions $f_k = f(\cdot,\xi^k)$ and $T_k = T_{\alpha_k f,\xi^k}$, $k\in\mathbb{N}$. Then the scheme can be written as $w^{k+1} = T_kw^k$. If $T_kw^* = w^*$, we would essentially only have to invoke Lemmas 8 and 9 to finish the proof. But due to the stochasticity, this does not hold, so we need to be more careful.

We begin by adding and subtracting the term T_kw^* and find that

$$\|w^{k+1} - w^*\|^2 = \|T_k w^k - T_k w^*\|^2 + 2(T_k w^k - T_k w^*, T_k w^* - w^*) + \|T_k w^* - w^*\|^2.$$

By Lemmas 8 and 9 the expectation $\mathbf{E}_{\xi^k}$ of the first term on the right-hand side is bounded by $(1 - 2\mu\alpha_k + 3v^2\alpha_k^2)^{1/2}\|w^k - w^*\|^2$ while by Lemma 6 the last term is bounded in expectation by $\alpha_k^2\sigma^2$. The second term is the problematic one. We add and subtract both w^k and w^* in order to find terms that we can control:

$$\begin{aligned}
&(T_k w^k - T_k w^*, T_k w^* - w^*) \\
&\quad = \big((T_k - I)w^k - (T_k - I)w^*, (T_k - I)w^*\big) + \big(w^k - w^*, (T_k - I)w^*\big) \\
&\quad =: I_1 + I_2.
\end{aligned}$$

In order to bound I_1 and I_2, we first need to apply the a priori bound from Lemma 2. This will also enable us to utilize the local Lipschitz condition. First, we notice that due to Lemma 6, we find that

$$\big(\mathbf{E}_{\xi^k}\big[\|T_k w^*\|^j\big]\big)^{\frac{1}{j}} \le \|w^*\| + \Big(\mathbf{E}_{\xi^k}\Big[\|\nabla f_k(w^*)\|_{H^*}^j\Big]\Big)^{\frac{1}{j}} \le \|w^*\| + \sigma$$

is bounded for $j \le 2^m$. As T_k is a contraction, we also obtain

$$\begin{aligned}
\big(\mathbf{E}_k\big[\|T_k w^k\|^j\big]\big)^{\frac{1}{j}} &\le \big(\mathbf{E}_k\big[\|T_k w^k - T_k w^*\|^j\big]\big)^{\frac{1}{j}} + \big(\mathbf{E}_{\xi^k}\big[\|T_k w^*\|^j\big]\big)^{\frac{1}{j}} \\
&\le \big(\mathbf{E}_k\big[\|w^k - w^*\|^j\big]\big)^{\frac{1}{j}} + \|w^*\| + \sigma.
\end{aligned}$$

Thus, there exists a random variable R_1 such that

$$\max\big(\|T_k w^k\|, \|T_k w^*\|\big) \le R_1,$$

and $\mathbf{E}_k[R_1^j]$ is bounded for $j \le 2^m$. For I_1, we then obtain that

$$\begin{aligned}
I_1 &\le \big((T_k - I)w^k - (T_k - I)w^*, (T_k - I)w^*\big) \\
&\le \|\alpha_k \nabla f_k(T_k w^k) - \alpha_k \nabla f_k(T_k w^*)\|_{H^*} \|\alpha_k \nabla f_k(w^*)\|_{H^*} \\
&\le \alpha_k^2 L_{\xi^k}(R_1)\|T_k w^k - T_k w^*\| \|\nabla f_k(w^*)\|_{H^*} \\
&\le \alpha_k^2 L_{\xi^k}(R_1)\|w^k - w^*\| \|\nabla f_k(w^*)\|_{H^*},
\end{aligned}$$

where we used the fact that T_k is contractive in the last step. Taking the expectation, we then have by Hölder's inequality that

$$\begin{aligned}
\mathbf{E}_k[I_1] &\le \alpha_k^2 \mathbf{E}_k\big[L_{\xi^k}(R_1)\|w^k - w^*\| \|\nabla f_k(w^*)\|_{H^*}\big] \\
&\le \alpha_k^2 \tilde{L}_1 \big(\mathbf{E}_{k-1}\big[\|w^k - w^*\|^{2^m}\big]\big)^{2^{-m}} \big(\mathbf{E}_{\xi^k}\big[\|\nabla f_k(w^*)\|_{H^*}^{2^m}\big]\big)^{2^{-m}},
\end{aligned}$$

where

$$\tilde{L}_1 = \begin{cases} \left(\mathbf{E}_k\left[P(R_1)^{\frac{2^m}{2^m-2}}\right]\right)^{\frac{2^m-2}{2^m}}, & m > 1, \\ \sup|P(R_1)|, & m = 1. \end{cases}$$

As P is a polynomial of at most order $2^m - 2$, the expression only contains terms R_1^j where the exponent j is at most $\left(\frac{2^m}{2^m-2}\right)(2^m - 2) = 2^m$. Hence $\tilde{L}_1$ is bounded, and in view of Lemma 2 we get that

$$\mathbf{E}_k[I_1] \leq D_1\alpha_k^2,$$

where $D_1 \geq 0$ is a constant depending only on $\|w^*\|$, $\|w_1 - w^*\|$, σ and η. For I_2, we add and subtract $\alpha_k \imath \nabla f_k(w^*)$ to get

$$\begin{aligned} I_2 &= \left(w^k - w^*, (T_k - I)w^*\right) \\ &= \left(w^k - w^*, (T_k - I)w^* + \alpha_k \imath \nabla f_k(w^*)\right) - \left(w^k - w^*, \alpha_k \imath \nabla f_k(w^*)\right). \end{aligned}$$

Since $w^k - w^*$ is independent of $\alpha_k \nabla f_k(w^*)$, it follows that

$$\mathbf{E}_{\xi^k}[\left(w^k - w^*, \alpha_k \imath \nabla f_k(w^*)\right)] = \left(w^k - w^*, \mathbf{E}_{\xi^k}[\alpha_k \imath \nabla f_k(w^*)]\right) = 0.$$

Using the Cauchy-Schwarz inequality and Lemma 6, we find that

$$\begin{aligned} \mathbf{E}_k[I_2] &\leq \mathbf{E}_k\left[\|w^k - w^*\| \|\imath^{-1}(T_k - I)w^* + \alpha_k \nabla f_k(w^*)\|_{H^*}\right] \\ &\leq \mathbf{E}_k\left[L_{\xi^k}(R_2)\alpha_k^2 \|w^k - w^*\| \|\nabla f_k(w^*)\|_{H^*}\right] \\ &\leq \alpha_k^2 \tilde{L}_2 \left(\mathbf{E}_{k-1}\left[\|w^k - w^*\|^{2^m}\right]\right)^{2^{-m}} \left(\mathbf{E}_{\xi^k}\left[\|\nabla f_k(w^*)\|_{H^*}^{2^m}\right]\right)^{2^{-m}}, \end{aligned}$$

where $R_2 = \max(\|w^*\|, \|\nabla f_k(w^*)\|_{H^*})$ and

$$\tilde{L}_2 = \begin{cases} \left(\mathbf{E}_k\left[P(R_2)^{\frac{2^m}{2^m-2}}\right]\right)^{\frac{2^m-2}{2^m}}, & m > 1, \\ \sup|P(R_2)|, & m = 1. \end{cases}$$

Just as for I_1, we therefore get by Lemma 2 that

$$\mathbf{E}_k[I_2] \leq D_2\alpha_k^2,$$

where $D_2 \geq 0$ is a constant depending only on $\|w^*\|$, $\|w_1 - w^*\|$, σ and η.

Summarising, we now have

$$\mathbf{E}_k\left[\|w^{k+1} - w^*\|^2\right] \leq \tilde{C}_k \mathbf{E}_{k-1}\left[\|w^k - w^*\|^2\right] + \alpha_k^2 D$$

with $\tilde{C}_k = \left(1 - 2\mu\alpha_k + 3\nu^2\alpha_k^2\right)^{1/2}$ and $D = \sigma^2 + D_1 + D_2$. Recursively applying the above bound yields

$$\mathbf{E}_k\left[\|w^{k+1} - w^*\|^2\right] \leq \prod_{j=1}^{k} \tilde{C}_j \|w_1 - w^*\|^2 + D \sum_{j=1}^{k} \alpha_j^2 \prod_{i=j+1}^{k} \tilde{C}_i.$$

Applying Lemma 10 (i) and (ii) with $p = 1/2$, $r = 1$, $C_1 = 2\mu\eta$ and $C_2 = 3\nu^2\eta^2$ then shows that

$$\prod_{j=1}^{k} \tilde{C}_j \leq \exp\left(\frac{\nu^2\eta^2\pi^2}{4}\right)(k+1)^{-\mu\eta}$$

and

$$\sum_{j=1}^{k} \alpha_j^2 \prod_{i=j+1}^{k} \tilde{C}_i \leq \eta^2 2^{\mu\eta} \exp\left(\frac{\nu^2\eta^2\pi^2}{4}\right) \frac{1}{\mu\eta - 1}(k+1)^{-1}.$$

Thus, we finally arrive at

$$\mathbf{E}_k\left[\|w^{k+1} - w^*\|^2\right] \leq \frac{C}{k+1},$$

where C depends on $\|w^*\|$, $\|w_1 - w^*\|$, μ, σ and η. □

Remark 4 The above proof is complicated mainly due to the stochasticity and due to the lack of strong convexity. We consider briefly the simpler, deterministic, full-batch, case with

$$w^{k+1} = w^k - \alpha_k \nabla F(w^{k+1}),$$

where F is strongly convex with convexity constant μ. Then it can easily be shown that

$$(\nabla F(v) - \nabla F(w), v - w) \geq \mu\|v - w\|^2.$$

This means that

$$\|(I + \alpha\nabla F)^{-1}(v) - (I + \alpha\nabla F)^{-1}(w)\| \leq (1 + \alpha\mu)^{-1}\|v - w\|,$$

i.e. the resolvent is a strict contraction. Since $\nabla F(w^*) = 0$, it follows that $(I + \alpha\nabla F)^{-1}w^* = w^*$ so a simple iterative argument shows that

$$\|w^{k+1} - w^*\|^2 \leq \prod_{j=1}^{k} \left(1 + \alpha_j\mu\right)^{-1} \|w_1 - w^*\|^2.$$

Using $(1 + \alpha\mu)^{-1} \leq 1 - \mu\alpha + \mu^2\alpha^2$, choosing $\alpha_k = \eta/k$ and applying Lemma 10 then shows that

$$\|w^{k+1} - w^*\|^2 \leq C(k+1)^{-1}$$

for appropriately chosen η. In particular, these arguments do not require the Lipschitz continuity of ∇F, which is needed in the stochastic case to handle the terms arising due to $\nabla f(w^*, \xi) \neq 0$.

5 Numerical experiments

In order to illustrate our results, we set up a numerical experiment along the lines given in the introduction. In the following, let $H = L^2(0,1)$ be the Lebesgue space of square integrable functions equipped with the usual inner product and norm. Further, let $x_j^i \in H$ for $i = 1,\ j = 1, \ldots, \left\lfloor \frac{n}{2} \right\rfloor$ and $i = 2,\ j = \left\lfloor \frac{n}{2} \right\rfloor + 1, \ldots, n$ be elements from two different classes within the space H. In particular, we choose each x_j^1 to be a polynomial of degree 4 and each x_j^2 to be a trigonometric function with bounded frequency for $j = 1, \ldots, n$. The polynomial coefficients and the frequencies were randomly chosen.

We want to classify these functions as either polynomial or trigonometric. To do this, we set up an affine (SVM-like) classifier by choosing the loss function $\ell(h, y) = \ln(1 + \mathrm{e}^{-hy})$ and the prediction function $h([w, \overline{w}], x) = (w, x) + \overline{w}$ with $[w, \overline{w}] \in L^2(0,1) \times \mathbb{R}$. Without $\overline{w}$, this would be linear, but by including $\overline{w}$ we can allow for a constant bias term and thereby make it affine. We also add a regularization term $\frac{\lambda}{2}\|w\|^2$ (not including the bias), such that the minimization objective is

$$F([w, \overline{w}], \xi) = \frac{1}{n} \sum_{j=1}^{n} \ell(h([w, \overline{w}], x_j), y_j) + \frac{\lambda}{2}\|w\|^2,$$

where $[x_j, y_j] = [x_j^1, -1]$ if $j \leq \left\lfloor \frac{n}{2} \right\rfloor$ and $[x_j, y_j] = [x_j^2, 1]$ if $j > \left\lfloor \frac{n}{2} \right\rfloor$, similar to Eq. (2). In one step of SPI, we use the function

$$f([w, \overline{w}], \xi) = \ell(h([w, \overline{w}], x_\xi), y_\xi) + \frac{\lambda}{2}\|w\|^2,$$

with a random variable $\xi : \Omega \to \{1, \ldots, n\}$. Since we cannot do computations directly in the infinite-dimensional space, we discretize all the functions using N equidistant points in $[0, 1]$, omitting the endpoints. For each N, this gives us an optimization problem on $\mathbb{R}^N$, which approximates the problem on H.

For the implementation, we make use of the following computational idea, which makes SPI essentially as fast as SGD. Differentiating the chosen ℓ and h shows that the scheme is given by the iteration

$$[w, \overline{w}]^{k+1} = [w, \overline{w}]^k + c_k[x_k, 1] - \lambda\alpha_k[w, 0]^{k+1},$$

where $c_k = \frac{\alpha_k y_k}{1+\exp((w^{k+1}, x_k)y_k + \overline{w}^{k+1} y_k)}$. This is equivalent to

$$w^{k+1} = \frac{1}{1+\alpha_k\lambda}\left(w^k + c_k x_k\right) \quad \text{and} \quad \overline{w}^{k+1} = \overline{w}^k + c_k.$$

Inserting the expression for $[w, \overline{w}]^{k+1}$ in the definition of c_k, we obtain that

$$c_k = \frac{\alpha_k y_k}{1 + \exp\left(\frac{1}{1+\alpha_k\lambda}(w^k + c_k x_k, x_k)y_k + (\overline{w}^k + c_k)y_k\right)}.$$

We thus only need to solve one scalar-valued equation. This is at most twice as expensive as SGD, since the equation solving is essentially free and the only additional costly term is (x_k, x_k) (the term (w^k, x_k) of course has to be computed also in SGD). By storing the scalar result, the extra cost will be essentially zero if the same sample is revisited. We note that extending this approach to larger batch-sizes is straightforward. If the batch size is B, then one has to solve a B-dimensional equation.

Using this idea, we implemented the method in Python and tested it on a series of different discretizations. We took $n = 1000$, i.e. 500 functions of each type, $M = 10{,}000$ time steps and discretization parameters $N = 100 \cdot 2^i$ for $i = 1, \ldots, 11$ to approximate the infinite dimensional space $L^2(0, 1)$. We used $\lambda = 10^{-3}$ and the initial step size $\eta = 2/\lambda$, since in this case it can be shown that $\mu \geq \lambda$. There is no closed-form expression for the exact minimum w^*, so instead we ran SPI with $10M$ time steps and used the resulting reference solution as an approximation to w^*. Further, we approximated the expectation $\mathbf{E}_k$ by running the experiment 100 times and averaging the resulting errors. In order to compensate for the vectors becoming longer as

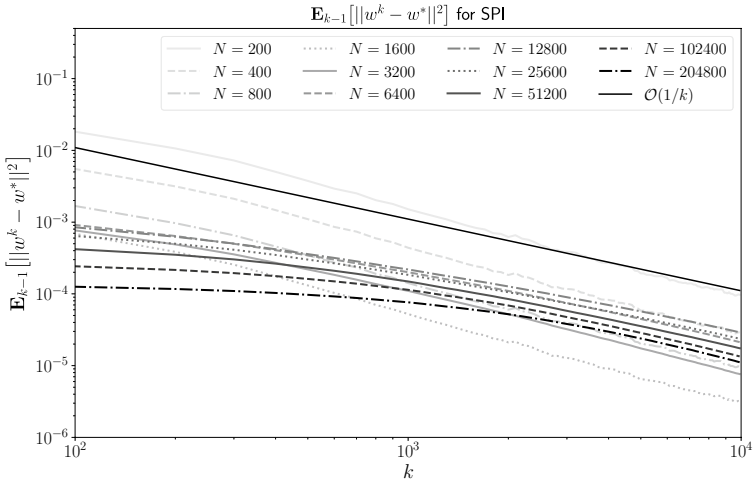

Fig. 1 Approximated errors $\mathbf{E}_{k-1}[\|w^k - w^*\|_N^2]$ for the SPI method, measured in RMS-norm, for discretizations with varying number of grid points N. Statistics were only computed at every 100 time steps, this is why the plot starts at $k = 100$. The $1/k$-convergence is clearly seen by comparing to the uppermost solid black reference line

N increases, we measure the errors in the RMS-norm $\|\cdot\|_N = \|\cdot\|_{\mathbb{R}^N}/\sqrt{N+1}$. As $N \to \infty$, this tends to the L^2 norm.

Figure 1 shows the resulting approximated errors $\mathbf{E}_{k-1}[\|w^k - w^*\|_N^2]$. As expected, we observe convergence proportional to $1/k$ for all N. The error constants do vary to a certain extent, but they are reasonably similar. As the problem approaches the infinite-dimensional case, they vary less. In order to decrease the computational requirements, we only compute statistics at every 100 time steps, this is why the plot starts at $k = 100$.

In contrast, redoing the same experiment but with the explicit SGD method instead results in Fig. 2. We note that except for $N = 200$ and $N = 400$, the method seemingly does not converge at all. This is partially explained by the fact that the Lipschitz constant grows with N (at least for the coarsest discretizations, for which we could estimate it), such that we get closer to the stability boundary. The main reason, however, is because of rare "bad" paths. In those, the method initially takes a large step in the wrong direction. Theoretically, it will eventually recover from this. In practice, it does not, due to the finite computational budget. Even when such bad paths are omitted from the results and $\mathcal{O}(1/k)$–convergence is observed, the errors are much larger than in Fig. 1. Many more steps would be necessary to reach the same accuracy as SPI. Since our implementations are certainly not optimal in any sense, we do not show a comparison of computational times here. They are, however, very similar, meaning that SPI is more efficient than SGD for this problem.

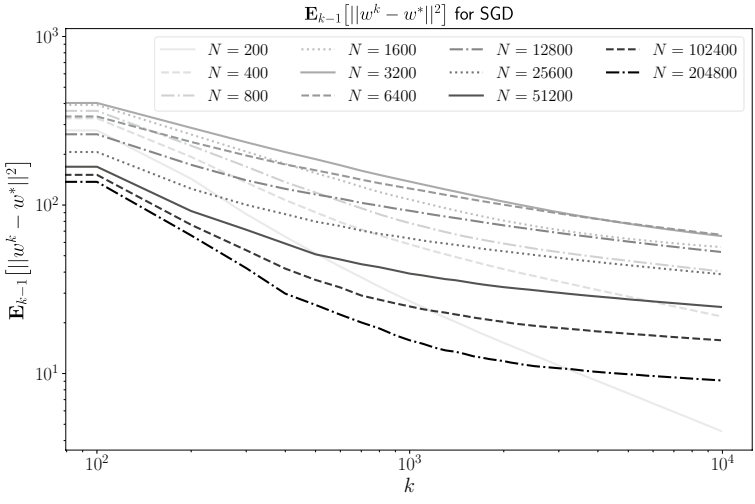

Fig. 2 Approximated errors $\mathbf{E}_{k-1}[\|w^k - w^*\|_N^2]$ for the SGD method, measured in RMS-norm, for discretizations with varying number of grid points N. Statistics were only computed at every 100 time steps, this is why the plot starts at $k = 100$. Except for $N = 200$ and $N = 400$, the method does not converge at all. Even when it does, the errors are much larger than in Fig. 1

6 Conclusions

We have rigorously proved convergence with an optimal rate for the stochastic proximal iteration method in a general Hilbert space. This improves the analysis situation in two ways. Firstly, by providing an extension of similar results in a finite-dimensional setting to the infinite-dimensional case, as well as extending these to more general operators. Secondly, by improving on similar infinite-dimensional results that only achieve convergence, without any error bounds. The latter improvement comes at the cost of stronger assumptions on the cost functional. Global Lipschitz continuity of the gradient is, admittedly, a rather strong assumption. However, as we have demonstrated, this can be replaced by local Lipschitz continuity where the maximal growth of the Lipschitz constant is determined by higher moments of the gradient applied to the minimum. This is a weaker condition. Finally, we have seen that the theoretical results are applicable also in practice, as demonstrated by the numerical results in the previous section.

Acknowledgements The authors would like to thank the anonymous referee and Eskil Hansen for valuable feedback.

Author Contributions All authors contributed to all parts of the article.

Funding Open access funding provided by Lund University. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

Availability of data and materials Not applicable.

Code availability The code used for the numerical experiments is available on request from the authors.

Declarations

Computations The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at LUNARC partially funded by the Swedish Research Council through grant agreement no. 2018–05973.

Conflict of interest/Competing interests The authors have no conflicts of interest to declare that are relevant to the content of this article.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Agarwal, A., Bartlett, P.L., Ravikumar, P., Wainwright, M.J.: Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. IEEE Trans. Inf. Theory **58**(5), 3235–3249 (2012). https://doi.org/10.1109/TIT.2011.2182178
2. Asi, H., Duchi, J.C.: Modeling simple structures and geometry for better stochastic optimization algorithms. In: Chaudhuri, K., Sugiyama, M. (eds.) Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 89, pp. 2425–2434. PMLR, Naha, Japan (2019). https://proceedings.mlr.press/v89/asi19a.html
3. Asi, H., Duchi, J.C.: Stochastic (approximate) proximal point methods: convergence, optimality, and adaptivity. SIAM J. Optim. **29**(3), 2257–2290 (2019). https://doi.org/10.1137/18M1230323
4. Barbu, V.: Nonlinear Differential Equations of Monotone Types in Banach Spaces, p. 272. Springer, New York (2010). https://doi.org/10.1007/978-1-4419-5542-5
5. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 2nd edn. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, p. 619. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-48311-5
6. Bertsekas, D.P.: Incremental proximal methods for large scale convex optimization. Math. Program. **129**(2, Ser. B), 163–195 (2011). https://doi.org/10.1007/s10107-011-0472-0
7. Bianchi, P.: Ergodic convergence of a stochastic proximal point algorithm. SIAM J. Optim. **26**(4), 2235–2260 (2016). https://doi.org/10.1137/15M1017909
8. Bianchi, P., Hachem, W.: Dynamical behavior of a stochastic forward–backward algorithm using random monotone operators. J. Optim. Theory Appl. **171**(1), 90–120 (2016). https://doi.org/10.1007/s10957-016-0978-y
9. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Rev. **60**(2), 223–311 (2018). https://doi.org/10.1137/16M1080173
10. Brzeźniak, Z., Carelli, E., Prohl, A.: Finite-element-based discretizations of the incompressible Navier–Stokes equations with multiplicative random forcing. IMA J. Numer. Anal. **33**(3), 771–824 (2013). https://doi.org/10.1093/imanum/drs032
11. Clark, D.S.: Short proof of a discrete Gronwall inequality. Discrete Appl. Math. **16**(3), 279–281 (1987). https://doi.org/10.1016/0166-218X(87)90064-3
12. Davis, D., Drusvyatskiy, D.: Stochastic model-based minimization of weakly convex functions. SIAM J. Optim. **29**(1), 207–239 (2019). https://doi.org/10.1137/18M1178244
13. Eckstein, J., Bertsekas, D.P.: On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**(3, Ser. A), 293–318 (1992). https://doi.org/10.1007/BF01581204
14. Eisenmann, M.: Methods for the temporal approximation of nonlinear, nonautonomous evolution equations. PhD thesis, TU Berlin (2019)
15. Eisenmann, M., Kovács, M., Kruse, R., Larsson, S.: On a randomized backward Euler method for nonlinear evolution equations with time-irregular coefficients. Found. Comput. Math. **19**(6), 1387–1430 (2019). https://doi.org/10.1007/s10208-018-09412-w
16. Fagan, F., Iyengar, G.: Unbiased scalable softmax optimization. ArXiv Preprint, arXiv:1803.08577 (2018)
17. Güler, O.: On the convergence of the proximal point algorithm for convex minimization. SIAM J. Control Optim. **29**(2), 403–419 (1991). https://doi.org/10.1137/0329022
18. Hager, W.W.: Updating the inverse of a matrix. SIAM Rev. **31**(2), 221–239 (1989). https://doi.org/10.1137/1031049
19. Lasiecka, I., Triggiani, R.: Control Theory for Partial Differential Equations: Continuous and Approximation Theories. I. Encyclopedia of Mathematics and its Applications, vol. 74, p. 644. Cambridge University Press, Cambridge (2000). https://doi.org/10.1017/CBO9781107340848 . (**Abstract parabolic systems**)

20. Necoara, I.: General convergence analysis of stochastic first-order methods for composite optimization. J. Optim. Theory Appl. **189**(1), 66–95 (2021). https://doi.org/10.1007/s10957-021-01821-2
21. Papageorgiou, N.S.: Convex integral functionals. Trans. Am. Math. Soc. **349**(4), 1421–1436 (1997). https://doi.org/10.1090/S0002-9947-97-01478-5
22. Papageorgiou, N.S., Winkert, P.: Applied Nonlinear Functional Analysis. An Introduction, p. 612. De Gruyter, Berlin (2018). https://doi.org/10.1515/9783110532982
23. Patrascu, A., Irofti, P.: Stochastic proximal splitting algorithm for composite minimization. ArXiv Preprint, arXiv:1912.02039v2 (2020)
24. Patrascu, A., Necoara, I.: Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. J. Mach. Learn. Res. **18**(198), 1–42 (2018)
25. Quarteroni, A., Valli, A.: Domain Decomposition Methods for Partial Differential Equations Numerical Mathematics and Scientific Computation, p. 360. Oxford University Press, New York (1999)
26. Raginsky, M., Rakhlin, A.: Information-based complexity, feedback and dynamics in convex programming. IEEE Trans. Inf. Theory **57**(10), 7036–7056 (2011). https://doi.org/10.1109/TIT.2011.2154375
27. Rockafellar, R.T.: On the maximal monotonicity of subdifferential mappings. Pac. J. Math. **33**, 209–216 (1970)
28. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control Optim. **14**(5), 877–898 (1976). https://doi.org/10.1137/0314056
29. Rockafellar, R.T., Wets, R.J.-B.: On the interchange of subdifferentiation and conditional expectations for convex functionals. Stochastics **7**(3), 173–182 (1982). https://doi.org/10.1080/17442508208833217
30. Rosasco, L., Villa, S., Vũ, B.C.: Convergence of stochastic proximal gradient algorithm. Appl. Math. Optim. **82**(3), 891–917 (2020). https://doi.org/10.1007/s00245-019-09617-7
31. Roubíček, T.: Nonlinear Partial Differential Equations with Applications, 2nd edn., p. 476. Springer, Basel (2013). https://doi.org/10.1007/978-3-0348-0513-1
32. Ryu, E., Boyd, S.: Stochastic proximal iteration: a non-asymptotic improvement upon stochastic gradient descent. www.math.ucla.edu/eryu/papers/spi.pdf (2016). Accessed 20 March 2020
33. Ryu, E.K., Yin, W.: Proximal–proximal–gradient method. J. Comput. Math. **37**(6), 778–812 (2019). https://doi.org/10.4208/jcm.1906-m2018-0282
34. Salim, A., Bianchi, P., Hachem, W.: Snake: a stochastic proximal gradient algorithm for regularized problems over large graphs. IEEE Trans. Automat. Control **64**(5), 1832–1847 (2019). https://doi.org/10.1109/tac.2019.2890888
35. Toulis, P., Airoldi, E.M.: Scalable estimation strategies based on stochastic approximations: classical results and new insights. Stat. Comput. **25**(4), 781–795 (2015). https://doi.org/10.1007/s11222-015-9560-y
36. Toulis, P., Airoldi, E.M.: Asymptotic and finite-sample properties of estimators based on stochastic gradients. Ann. Stat. **45**(4), 1694–1727 (2017). https://doi.org/10.1214/16-AOS1506
37. Toulis, P., Airoldi, E., Rennie, J.: Statistical analysis of stochastic gradient methods for generalized linear models. In: Xing, E.P., Jebara, T. (eds.) Proceedings of the 31st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 32, pp. 667–675. PMLR, Bejing, China (2014). https://proceedings.mlr.press/v32/toulis14.html
38. Toulis, P., Horel, T., Airoldi, E.M.: The proximal Robbins–Monro method. ArXiv Preprint, arXiv:1510.00967v4 (2020)
39. Toulis, P., Tran, D., Airoldi, E.: Towards stability and optimality in stochastic gradient descent. In: Gretton, A., Robert, C.C. (eds.) Proceedings of the 19th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 51, pp. 1290–1298. PMLR, Cadiz, Spain (2016). http://proceedings.mlr.press/v51/toulis16.html
40. Tran, D., Toulis, P., Airoldi, E.M.: Stochastic gradient descent methods for estimation with large data sets. ArXiv Preprint, arXiv:1509.06459 (2015)
41. Zeidler, E.: Nonlinear Functional Analysis and Its Applications. II/B, pp. 469–1202. Springer, New York (1990). https://doi.org/10.1007/978-1-4612-0985-0. Nonlinear monotone operators

Paper II

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

Journal of Computational and Applied Mathematics 417 (2023) 114575

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

SRKCD: A stabilized Runge–Kutta method for stochastic optimization

Tony Stillfjord *, Måns Williamson

Centre for Mathematical Sciences, Lund University, P.O. Box 118, 221 00 Lund, Sweden

ARTICLE INFO

Article history:
Received 29 January 2022
Received in revised form 22 June 2022

MSC:
90C15
65K05
65L20

Keywords:
Stochastic optimization
Convergence analysis
Runge–Kutta–Chebyshev
Stability

ABSTRACT

We introduce a family of stochastic optimization methods based on the Runge–Kutta–Chebyshev (RKC) schemes. The RKC methods are explicit methods originally designed for solving stiff ordinary differential equations by ensuring that their stability regions are of maximal size. In the optimization context, this allows for larger step sizes (learning rates) and better robustness compared to e.g. the popular stochastic gradient descent method. Our main contribution is a convergence proof for essentially all stochastic Runge–Kutta optimization methods. This shows convergence in expectation with an optimal sublinear rate under standard assumptions of strong convexity and Lipschitz-continuous gradients. For non-convex objectives, we get convergence to zero in expectation of the gradients. The proof requires certain natural conditions on the Runge–Kutta coefficients, and we further demonstrate that the RKC schemes satisfy these. Finally, we illustrate the improved stability properties of the methods in practice by performing numerical experiments on both a small-scale test example and on a problem arising from an image classification application in machine learning.

1. Introduction

In this article we consider the optimization problem

$$\min_w F(w)$$

where F is differentiable. Such problems frequently arise in many contexts, e.g. for training neural networks in the currently popular subject of machine learning. We focus on the large-scale case where computing $\nabla F(w)$ is expensive, and assume that cheap stochastic approximations to $\nabla F(w)$ are available.

At a (local) minimum w_*, it holds that $\nabla F(w_*) = 0$, and such a stationary point of the gradient may be found by evolving the gradient flow

$$\dot{w}(t) = -\nabla F(w(t))$$

over the pseudo-time $t \in [0, \infty)$. The benefit of this reformulation is that many optimization methods for the original problem may now be stated as time-stepping methods for the gradient flow. We recognize e.g. the explicit Euler method with varying step sizes α_k

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

* Corresponding author.
E-mail addresses: tony.stillfjord@math.lth.se (T. Stillfjord), mans.williamson@math.lth.se (M. Williamson).

https://doi.org/10.1016/j.cam.2022.114575

as the gradient descent (GD) method. The popular *stochastic* gradient descent (SGD) [1] method uses the same formula but with the approximation $g(\xi_k, w_k)$ instead of $\nabla F(w_k)$, where ξ_k is a random variable that determines which parts of F to use. SGD is therefore a perturbed version of explicit Euler.

As was observed already in [2], the gradient flows arising from neural networks tend to be stiff. As a consequence, explicit methods suffer from severe step size restrictions. This is particularly inconvenient when one wants to reach a stationary state, which typically requires evolving the system for a long time. While it is difficult to quantify exactly how the stochasticity introduced in methods like SGD affects this, they suffer from similar step size restrictions.

A way to avoid such step size restrictions would be to instead use methods with better stability properties, such as A-stable methods. This, however, requires that the method is implicit. One such method would be implicit Euler, which, when applied to the gradient flow is equivalent to the proximal point method in the context of optimization [3,4]. While this can be applied in certain cases when F has a specific structure that allows the arising nonlinear equation systems to be solved efficiently, in general (usually) this is not feasible.

An alternative, which to our knowledge has only been considered to a very small extent in the optimization community, is the use of explicit stabilized schemes. These are constructed such that their stability regions are maximized. Thus, there will still be a step size restriction, but of a more benign type. A large class of such methods are the Runge–Kutta–Chebyshev methods [5], see also [6] for an overview and further references. They are explicit Runge–Kutta methods, i.e. of the form

$$w_{k,i} = w_k - \alpha_k \sum_{j=1}^{i-1} a_{i,j} \nabla F(w_{k,j}), \quad i = 1, \dots, s,$$

$$w_{k+1} = w_k - \alpha_k \sum_{i=1}^{s} b_i \nabla F(w_{k,i}),$$

where the coefficients $a_{i,j}$ and b_i have been chosen in a very specific way such that the stability region extends as far into the left half-plane as possible. The tradeoff compared to GD is that such a scheme with s stages requires s times as many gradient evaluations. However, it still pays off, because the stability region grows as s^2. An optimization method called the Runge–Kutta–Chebyshev descent (RKCD) based on this idea has recently been investigated in [7]. However, only for the case where ∇F can be computed exactly and for a rather restrictive class of problems. In this article, we propose a stochastic version of such a scheme which we call the stochastic Runge–Kutta–Chebyshev descent (SRKCD). Compared to e.g. SGD, it has superior stability properties.

There are of course other advanced methods that can be applied to the problem, and there is a rather large number of papers on the subject. We refer to [8] for a general overview. Here, we mention for example accelerated gradient-type methods such as the SGD with momentum [9,10], the stochastic heavy ball method [11] and Nesterov's accelerated gradient method [12]. These do not use only the approximate gradient at the current iteration w_k but modify this gradient using other gradient information acquired in previous steps. A different class of methods are the adaptive learning rate methods, containing e.g. AdaGrad [13], AdaDelta [14], Adam [15], RMSprop [16] and AdaMax [15]. These are typically formulated as adapting the step size α_k based on a constantly updated model of the local cost landscape, acquired from gradient information computed in previous iterations. However, since most of them adjust the step size for each component of w_k separately, they could in a certain sense be seen as instead modifying the approximation $g(\xi_k, w_k)$ like the accelerated gradient methods.

In contrast to this, the method we propose simply uses the available gradient information without modifications and allows each step to be longer. Just like SGD may be extended to e.g. SGD with momentum, one might also consider SRKCD with momentum, provided that further analysis on the properties of this combined method is performed.

The main contribution of this article is a rigorous proof of convergence for a general Runge–Kutta method, under weak assumptions on its coefficients and standard assumptions on the optimization problem and the approximations $g(\xi, w)$. We emphasize that while the proof applies to SRKCD, it is more widely applicable. We consider two settings. First, the usual strongly convex setting, wherein we can prove optimal convergence orders of the type $\mathcal{O}(1/k)$. Secondly, the fully non-convex setting where we show that the squared norm of $\nabla F(w_k)$ goes to zero in expectation. This is also essentially optimal. In both cases, the results are direct extensions of similar results for SGD.

We note that nonlinear stability analysis is a very complex topic with few generally applicable results, and that the stability region of a method only refers to the setting of linear problems. For these reasons, it is not possible to use the available information on the RKC stability regions to tailor the general convergence proof further for SRKCD. The benefits of the improved stability properties in SRKCD are therefore not directly illustrated by the convergence proof. For this reason, we also perform numerical experiments which demonstrate that in practice they are present also in the stochastic non-linear and non-convex setting.

The outline of the paper is as follows. Section 2 contains the main error analysis for the general Runge–Kutta methods. It begins by formalizing the notation and assumptions on the problem, then presents preliminary results in Sections 2.1 and 2.2. The actual convergence proofs are presented in Sections 2.3 (convex case) and 2.4 (nonconvex case). Then we study the SRKCD method specifically in Section 3 and discuss its properties. The numerical experiments follow in Section 4 and we sum up some conclusions in Section 5. Finally, the Appendix contains a few results on Chebyshev polynomials which are needed but which are otherwise not of interest here.

2. General Runge–Kutta error analysis

Let us first fix the notation and specify our assumptions on the underlying problem. We denote by $\|\cdot\|$ the usual Euclidean norm on $\mathbb{R}^d$ and by $\langle\cdot,\cdot\rangle$ the corresponding inner product $\langle u, v\rangle = v^T u$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a complete probability space. For a random variable ξ on Ω, we consider the functions $f(\xi, \cdot) : \Omega \times \mathbb{R}^d \to \mathbb{R}$ and the main objective function $F : \mathbb{R}^d \to \mathbb{R}$,

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

Here, $\mathbb{E}_\xi[\cdot]$ denotes the expectation with respect to the probability distribution of ξ. We note that we have not specified the target space of the random variable ξ, because its properties does not matter for our analysis. However, if $\omega \in \Omega$ then $\xi(\omega)$ should be interpreted as a specific selection of the problem data, in machine learning terminology known as a batch. A typical situation would be to have an objective function of the form

$$F(w) = \frac{1}{N}\sum_{j=1}^{N} f(j, w).$$

Then a specific realization of $\xi(\omega)$ could be a single j, corresponding to a single data sample. Alternatively, in the common mini-batch setting, a realization of $\xi(\omega)$ could be a $B_\xi \subset \{1, \dots, N\}$, corresponding to a small subset of the data.

We approximate $\nabla F(w)$ by $g(\xi, w)$, where $g(\xi(\cdot), \cdot) : \Omega \times \mathbb{R}^d \to \mathbb{R}^d$ is integrable. In the above typical situation, we would usually have either $g(\xi, w) = \nabla f(\xi, w)$, where $\xi(\omega) \in \{1, \dots, N\}$ (single sample) or $g(\xi, w) = \frac{1}{|B_\xi|}\sum_{j\in B_\xi} \nabla f(j, w)$ with $B_{\xi(\omega)} \subset \{1, \dots, N\}$ (mini-batch).

In general, we consider a sequence of jointly independent random variables $\{\xi_k\}_{k=1}^\infty$ on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with the idea that step k of the method will depend on a realization of ξ_k. For such a sequence we define the total expectation $\mathbb{E}_k[X]$ of a random variable X by

$$\mathbb{E}_k[X] = \mathbb{E}_{\xi_1}\left[\mathbb{E}_{\xi_2}\left[\dots \mathbb{E}_{\xi_{k-1}}[X]\right]\right].$$

As the variables ξ_k are jointly independent, this coincides with the expectation of X with respect to the joint probability distribution of $(\xi_1, \dots, \xi_k)$.

The following assumptions on the full problem are standard:

Assumption 1. $F : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and ∇F is Lipschitz continuous with Lipschitz constant $L > 0$:

$$\|\nabla F(u) - \nabla F(v)\| \le L\|u - v\|, \ \forall u, v \in \mathbb{R}^d.$$

Assumption 2. F is strongly convex with convexity constant $c > 0$. That is,

$$F(u) \ge F(v) + \langle \nabla F(v), v - u\rangle + \frac{c}{2}\|v - u\|^2, \ \forall u, v \in \mathbb{R}^d.$$

We also make standard assumptions on the approximation g. The first is that it is Lipschitz-continuous with respect to the second argument:

Assumption 3. The function g is Lipschitz continuous with respect to the second argument with (for simplicity) the same Lipschitz constant $L > 0$ as ∇F:

$$\|g(\xi, u) - g(\xi, v)\| \le L\|u - v\|, \ a.s. \ \forall u, v \in \mathbb{R}^d.$$

Next, we assume that g is a reasonable approximation to ∇F in the following sense, following [8]:

Assumption 4. There exist scalars $\mu_G \ge \mu > 0$, $M \ge 0$ and $M_G \ge \mu^2$ such that the gradient ∇F and its approximation g satisfy the following conditions for all $w \in \mathbb{R}^d$:

(i) $\langle \nabla F(w), \mathbb{E}_\xi[g(\xi, w)]\rangle \ge \mu\|\nabla F(w)\|^2$,
(ii) $\|\mathbb{E}_\xi[g(\xi, w)]\| \le \mu_G\|\nabla F(w)\|$ and
(iii) $\mathbb{E}_\xi\left[\|g(\xi, w)\|^2\right] \le M + M_G\|\nabla F(w)\|^2$.

Assumption 4(i) and (ii) are fulfilled by assumption with $\mu = \mu_G = 1$ if we are considering (e.g.) the previously described single sample case. The third item puts a weak limit on the variance, which means that the approximation to the gradient is not too noisy.

Remark 2.1. We note that the statements "for all $w \in \mathbb{R}^d$" in the above assumptions could be replaced by "for all w_k", where w_k are the method iterates, i.e. the assumptions only need to hold where the method is actually evaluated. However, this is not helpful in practice, since the iterates are not known a priori.

Finally, we make a general assumption on the numerical optimization method. As shown in the previous section, this will be satisfied in particular for the SRKCD method.

Assumption 5. Given a sequence of step sizes $\{\alpha_k\}_{k\in\mathbb{N}}$, such that $\alpha_k > 0$ for all k and an initial condition $w_1 \in \mathbb{R}^d$, the optimization method is of the form

$$w_{k,i} = w_k - \alpha_k \sum_{j=1}^{i-1} a_{i,j} g(\xi_k, w_{k,j}), \quad i = 1, \dots, s,$$
$$w_{k+1} = w_k - \alpha_k \sum_{i=1}^{s} b_i g(\xi_k, w_{k,i}).$$

For brevity, denote $a_{s+1,i} \coloneqq b_i$, $i = 1, \dots, s$, and set $w_{k,s+1} \coloneqq w_{k+1}$. With this notation, the coefficients $a_{i,j}$ satisfy

(i) $\sum_{i=1}^{s} a_{s+1,i} = 1$,

(ii) $\sum_{j=1}^{i-1} |a_{i,j}| \leq 1, \quad i = 1, \dots, s+1.$

We note that item *(i)* would be satisfied for any Runge–Kutta method which is of order 1 when applied to $\dot{w} = -\nabla F(w)$. Further, we note that $a_{s+1,i} = b_i$ and $w_{k,s+1} = w_{k+1}$ means that the stage update formula coincides with the updating formula for w_{k+1}. This makes the following induction proofs less cumbersome.

2.1. Preliminary results

In the following lemma, we list some consequences of the basic assumptions.

Lemma 2.1. *Under Assumptions 1 and 2, there exists a unique $w_* \in \mathbb{R}^d$ such that*

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

and $\nabla F(w_*) = 0$. Further, it follows that

$$F(u) - F(v) \leq \langle \nabla F(v), u - v \rangle + \frac{L}{2}\|u - v\|^2 \tag{1}$$

for all $u, v \in \mathbb{R}^d$. Finally, the difference $F(w) - F(w_*)$ is bounded by

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

Proof. The existence of a unique minimizer in this benign situation is well-known, see e.g. [17, Corollary 11.17]. The first inequality follows directly from a first-order expansion in Taylor series and Assumption 1. For the final inequality, see e.g. [8, Appendix B]. □

2.2. Bound on $\|w_{k+1} - w_k\|$

First, we consider what the method does in one step and bound $\|w_{k+1} - w_k\| = \|w_{k,s+1} - w_{k,1}\|$. To this end, we now define a sequence of polynomials $P_n(\alpha)$, $n = 0, \dots, s$, by

$$P_0(\alpha) = 0, \quad P_1(\alpha) = \alpha,$$
$$P_n(\alpha) = \alpha + \alpha L \sum_{i=1}^{n} |a_{n+1,i}| P_{i-1}(\alpha), \text{ where } 2 \leq n \leq s.$$

Note that the sequence depends on s, but for brevity we do not add an extra index to indicate this.

Lemma 2.2. *Let Assumptions 3 and 5 be satisfied. Then for a fixed s, it holds that $\|w_{k,n+1} - w_k\| \leq P_n(\alpha_k)\|g(\xi_k, w_k)\|$ for all $n \leq s$.*

Proof. We prove the statement by induction over n. In the case $n = 1$ it follows immediately from the definition that

$$\|w_{k,2} - w_k\| = \|w_{k,2} - w_{k,1}\| = |a_{2,1}|\alpha_k \|g(\xi_k, w_k)\|.$$

Since $|a_{2,1}| \leq 1$ by Assumption 5(ii), the base case is satisfied. Assume that the claim holds for all $i \leq n$ with $n < s$. Then, using Assumption 3 and the induction assumption

$$\begin{aligned}
&\|w_{k,n+1} - w_k\| \\
&= \left\| -\alpha_k \sum_{i=1}^{n} a_{n+1,i} g(\xi_k, w_k) - \alpha_k \sum_{i=1}^{n} a_{n+1,i} \big(g(\xi_k, w_{k,i}) - g(\xi_k, w_k)\big) \right\| \\
&\leq \alpha_k \sum_{i=1}^{n} |a_{n+1,i}| \|g(\xi_k, w_k)\| + \alpha_k \sum_{i=1}^{n} |a_{n+1,i}| \|g(\xi_k, w_{k,i}) - g(\xi_k, w_k)\| \\
&\leq \alpha_k \sum_{i=1}^{n} |a_{n+1,i}| \|g(\xi_k, w_k)\| + \alpha_k L \sum_{i=1}^{n} |a_{n+1,i}| \|w_{k,i} - w_k\| \\
&\leq \alpha_k \sum_{i=1}^{n} |a_{n+1,i}| \|g(\xi_k, w_k)\| + \alpha_k L \sum_{i=1}^{n} |a_{n+1,i}| P_{i-1}(\alpha_k) \|g(\xi_k, w_k)\| \\
&\leq P_n(\alpha_k) \|g(\xi_k, w_k)\|,
\end{aligned}$$

where we used Assumption 5(ii) in the last step. This concludes the inductive step. □

Lemma 2.3. *Under Assumption 5, it holds for $2 \leq n \leq s$ that*

$$P_n(\alpha) = \alpha + \alpha \sum_{i=1}^{n-1} (\alpha L)^i c_{n,i}$$

where the $c_{n,i}$ are constants not depending on α or L. Further, $c_{n,i} \leq 1$ for $2 \leq n \leq s$ and $1 \leq i \leq n-1$.

Proof. Once again, we employ induction. For $n = 2$, we have

$$P_2(\alpha) = \alpha + \alpha L(|a_{3,1}|\alpha),$$

which is on the stated form with $c_{2,1} = |a_{3,1}|$, and by Assumption 5(ii), $c_{2,1} \leq 1$. That is, the claim is valid for $n = 2$. Assume that P_m can be written on the stated form for all $m \leq n$ and that all the constants $c_{m,i}$ are bounded by 1. Then inserting this in the definition of P_{n+1} shows that

$$\begin{aligned}
P_{n+1} = \alpha &+ \alpha^2 L \sum_{i=2}^{n+1} |a_{n+2,i}| + \alpha^3 L^2 \sum_{i=3}^{n+1} |a_{n+2,i}| c_{i-1,1} \\
&+ \alpha^4 L^3 \sum_{i=4}^{n+1} |a_{n+2,i}| c_{i-1,2} + \cdots + \alpha^{n+1} L^n |a_{n+2,n+1}| c_{n,n-1}.
\end{aligned}$$

That is, we can write P_{n+1} on the desired form by taking $c_{n+1,1} = \sum_{i=2}^{n+1} |a_{n+2,i}|$ and $c_{n+1,j} = \sum_{i=j+1}^{n+1} |a_{n+2,i}| c_{i-1,j-1}$ for $j = 2, \ldots, n$. By Assumption 5(ii),

$$c_{n+1,1} = \sum_{i=2}^{n+1} |a_{n+2,i}| \leq \sum_{i=1}^{n+1} |a_{n+2,i}| \leq 1.$$

Similarly, since all the $c_{i-i,j-1}$ are bounded by 1 by the induction assumption,

$$c_{n+1,j} = \sum_{i=j+1}^{n+1} |a_{n+2,i}| c_{i-1,j-1} \leq \sum_{i=1}^{n+1} |a_{n+2,i}| \leq 1.$$

for $j = 2, \ldots, n$. This concludes the induction step. □

We can now bound the difference $F(w_{k+1}) - F(w_k)$ by using (1) from Lemma 2.1 to write

$$F(w_{k+1}) - F(w_k) \leq \langle \nabla F(w_k), w_{k+1} - w_k \rangle + \frac{L}{2} \|w_{k+1} - w_k\|^2. \tag{3}$$

For the first term on the right-hand side, we add and subtract terms to get

$$\begin{aligned}
&\langle \nabla F(w_k), w_{k+1} - w_k \rangle \\
&\quad = \left\langle \nabla F(w_k), -\alpha_k \sum_{i=1}^{s} b_i g(\xi_k, w_k) - \alpha_k \sum_{i=1}^{s} b_i (g(\xi_k, w_{k,i}) - g(\xi_k, w_k)) \right\rangle \\
&\quad \leq -\alpha_k \sum_{i=1}^{s} b_i \langle \nabla F(w_k), g(\xi_k, w_k) \rangle + \alpha_k L \sum_{i=1}^{s} |b_i| \|\nabla F(w_k)\| \|w_{k,i} - w_k\|.
\end{aligned}$$

We now use Lemma 2.2 and Young's inequality $ab \leq \frac{a^2}{4} + b^2$ with $a = \alpha_k \sqrt{L} \|\nabla F(w_k)\|$ and $b = \sqrt{L} \|w_{k,i} - w_k\|$ to bound the last sum in the previous expression

$$\begin{aligned}
&\sum_{i=1}^{s} \alpha_k L |b_i| \|\nabla F(w_k)\| \|w_{k,i} - w_k\| \\
&\leq \frac{\alpha_k^2 L}{4} \sum_{i=1}^{s} |b_i| \|\nabla F(w_k)\|^2 + L \sum_{i=1}^{s} |b_i| \|w_{k,i} - w_k\|^2 \\
&\leq \frac{\alpha_k^2 L}{4} \sum_{i=1}^{s} |b_i| \|\nabla F(w_k)\|^2 + L \sum_{i=1}^{s} |b_i| P_{i-1}(\alpha_k)^2 \|g(\xi_k, w_k)\|^2.
\end{aligned}$$

In total, by using Lemma 2.2 also on the last term of (3), we get

$$\begin{aligned}
&F(w_{k+1}) - F(w_k) \\
&\quad \leq -\alpha_k \sum_{i=1}^{s} b_i \langle \nabla F(w_k), g(\xi_k, w_k) \rangle + \frac{\alpha_k^2 L}{4} \sum_{i=1}^{s} |b_i| \|\nabla F(w_k)\|^2 \\
&\qquad + L \sum_{i=1}^{s} |b_i| P_{i-1}(\alpha_k)^2 \|g(\xi_k, w_k)\|^2 + \frac{L}{2} P_s(\alpha_k)^2 \|g(\xi_k, w_k)\|^2.
\end{aligned}$$

Taking expectations with respect to the distribution of ξ_k (recall that w_k does not depend on ξ_k) leads to

$$\begin{aligned}
&\mathbb{E}_{\xi_k}[F(w_{k+1}) - F(w_k)] \\
&\quad \leq -\alpha_k \sum_{i=1}^{s} b_i \langle \nabla F(w_k), \mathbb{E}_{\xi_k}[g(\xi_k, w_k)] \rangle + \frac{\alpha_k^2 L}{4} \sum_{i=1}^{s} |b_i| \|\nabla F(w_k)\|^2 \\
&\qquad + L \left(\sum_{i=1}^{s} |b_i| P_{i-1}(\alpha_k)^2 + \frac{1}{2} P_s(\alpha_k)^2 \right) \mathbb{E}_{\xi_k}\left[\|g(\xi_k, w_k)\|^2 \right].
\end{aligned} \tag{4}$$

By Assumption 4 we have that

$$\mathbb{E}_{\xi_k}\left[\|g(\xi_k, w_k)\|^2 \right] \leq M + M_G \|\nabla F(w_k)\|^2,$$

and applying this to the last term of (4) gives

$$\begin{aligned}
&\mathbb{E}_{\xi_k}[F(w_{k+1}) - F(w_k)] \\
&\quad \leq -\alpha_k \mu \|\nabla F(w_k)\|^2 + \frac{\alpha_k^2 L}{4} \sum_{i=1}^{s} |b_i| \|\nabla F(w_k)\|^2 \\
&\qquad + L \left(\sum_{i=1}^{s} |b_i| P_{i-1}(\alpha_k)^2 + \frac{1}{2} P_s(\alpha_k)^2 \right) \left(M + M_G \|\nabla F(w_k)\|^2 \right).
\end{aligned} \tag{5}$$

Here we have also used Assumption 4(i) and Assumption 5(i) on the first term on the right-hand side of (4) to obtain the $-\alpha_k \mu \|\nabla F(w_k)\|^2$-term in (5). Reordering the terms, we find

$$\begin{aligned}
&\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \\
&\quad \leq Q(\alpha_k) \|\nabla F(w_k)\|^2 + LM \left(\sum_{i=1}^{s} |b_i| P_{i-1}(\alpha_k)^2 + \frac{1}{2} P_s(\alpha_k)^2 \right).
\end{aligned} \tag{6}$$

with

$$Q(\alpha_k) = -\alpha_k \mu + LM_G \sum_{i=1}^{s} |b_i| P_{i-1}(\alpha_k)^2 + \frac{LM_G}{2} P_s(\alpha_k)^2 + \frac{1}{4} \alpha_k^2 L \sum_{i=1}^{s} |b_i|.$$

Since $P_0(\alpha_k) = 0$ and the smallest power of α_k in $P_i(\alpha_k)^2$ for $i = 1, \dots, s$ is α_k^2, we can choose $\alpha_k > 0$ small enough that

$$Q(\alpha_k) < -\frac{\alpha_k \mu}{2}. \tag{7}$$

This means that the first term in (6) is negative, and we can estimate it by using the strong convexity property

$$-\|\nabla F(w_k)\|^2 \leq -2c\big(F(w_k) - F(w_*)\big)$$

from (2) in Lemma 2.1. Adding and subtracting $F(w_*)$, rearranging and taking total expectations on both sides thus leads to

$$\begin{aligned}\mathbb{E}_k[F(w_{k+1}) - F(w_*)] \leq (1 - \alpha_k \mu c)\, \mathbb{E}_k[F(w_k) - F(w_*)] \\ + LM\Big(\sum_{i=1}^{s} |b_i| P_{i-1}(\alpha_k)^2 + \frac{1}{2} P_s(\alpha_k)^2\Big).\end{aligned} \tag{8}$$

This means that the next error is the previous error multiplied by a factor which is strictly less than one, plus two terms that are small. Hence it will tend to zero as $k \to \infty$, as we show formally in the next section.

Remark 2.2. Let us elaborate on the choice of α_k in (7). We can make the choice because the negative term is multiplied with α_k while the positive terms are all multiplied with higher powers of α_k, meaning that for a sufficiently small α_k the negative term will dominate. To make this more concrete, suppose that $\alpha_k \leq \frac{1}{Lm}$ for an integer $m \geq 2$. Then by Lemma 2.3,

$$P_i(\alpha_k)^2 \leq \alpha_k^2 \Big(1 + \frac{1}{m} + \frac{1}{m^2} + \cdots + \frac{1}{m^{s-1}}\Big)^2 = \alpha_k^2 \frac{m^2}{(m-1)^2} \leq 4\alpha_k^2$$

for every $i = 1, \dots, s$. Thus, since $\sum_{i=1}^{s} |b_i| \leq 1$ by Assumption 5,

$$\begin{aligned} Q(\alpha_k) &\leq -\alpha_k \mu + \alpha_k^2 \Big(4LM_G + 4\frac{LM_G}{2} + \frac{L}{4}\Big) \\ &\leq -\alpha_k \mu + L\alpha_k^2 (6M_G + \frac{1}{4}) \\ &\leq -\alpha_k \mu + \alpha_k \frac{6M_G + \frac{1}{4}}{m}. \end{aligned}$$

This is bounded by $-\frac{\alpha_k \mu}{2}$ and thereby satisfies (7) if

$$m \geq \frac{12M_G + \frac{1}{2}}{\mu}.$$

We can guarantee this by choosing m large enough, and a moderately small m is sufficient unless the estimator of the gradient is very bad (small μ) or the variance of the data is very large (large M_G). In a typical situation, both of these constants can be set to 1, which leads to a step size restriction of $\alpha_k \leq \frac{2}{25L}$. We note that this argument could be further refined to improve the bound, since the current estimations of $P_i(\alpha_k)^2$ are quite crude. For example, clearly $P_1(\alpha_k)^2 = \alpha_k^2$.

2.3. Convergence proof

Theorem 2.1. *Let Assumptions 1–5 be satisfied. Further assume that the scheme is run with the step size $\alpha_k = \frac{\beta}{k+\gamma}$, where $\gamma > 0$, $\beta > \frac{1}{c\mu}$ and α_1 satisfies (7). Then with*

$$\nu = \max\left\{ \frac{LM\Big(\sum_{i=1}^{s} |b_i| P_{i-1}(\beta)^2 + \frac{1}{2} P_s(\beta)^2\Big)}{\beta\mu c - 1}, (\gamma + 1)\,(F(w_1) - F(w_*)) \right\},$$

it holds that

$$\mathbb{E}_k[F(w_k) - F(w_*)] \leq \frac{\nu}{k + \gamma}, \tag{9}$$

for $k = 1, 2, \dots$.

Remark 2.3. The error constant ν can be bounded by a constant which is independent of s by using Assumption 5(ii). However, for some methods b_i decreases rapidly with increasing i (such as the SRKCD methods). In that case, such an estimation would be rather crude. We therefore keep these terms in the statement and leave it to the reader to insert their specific coefficients.

Proof of Theorem 2.1. We prove this using induction, inspired by [8, Theorem 4.7]. Let us abbreviate $\hat{k} = k + \gamma$. For the base case we note that it follows from the definition of ν that

$$\mathbb{E}_k[F(w_1) - F(w_*)] = (\gamma + 1)\,\frac{F(w_1) - F(w_*)}{\gamma + 1} \leq \frac{\nu}{\gamma + 1},$$

since w_1 is not chosen randomly. For the induction step we assume that (9) holds for some k. Using (8) we then have

$$\begin{aligned}\mathbb{E}_k[F(w_{k+1}) - F(w_*)] \leq (1 - \alpha_k \mu c)\,\frac{\nu}{\hat{k}} \\ + LM\Big(\sum_{i=1}^{s} |b_i| P_{i-1}(\alpha_k)^2 + \frac{1}{2} P_s(\alpha_k)^2\Big).\end{aligned} \tag{10}$$

Using that $\alpha_k = \frac{\beta}{\hat{k}}$ and adding and subtracting $\frac{\nu}{\hat{k}^2}$, we find that the right-hand side of (10) equals $S_1 + S_2$ where

$$S_1 = \left(\frac{\hat{k} - 1}{\hat{k}^2}\right)\nu \quad \text{and} \quad S_2 = -\left(\frac{\beta\mu c - 1}{\hat{k}^2}\right)\nu + LM\left(\sum_{i=1}^{s} |b_i| P_{i-1}\Big(\frac{\beta}{\hat{k}}\Big)^2 + \frac{1}{2} P_s\Big(\frac{\beta}{\hat{k}}\Big)^2\right).$$

By the inequality $\hat{k}^2 \geq \big(\hat{k} - 1\big)\big(\hat{k} + 1\big)$ we directly have that

$$S_1 \leq \frac{\nu}{\hat{k} + 1}.$$

To bound S_2, we first note that the polynomials $\frac{P_i(\alpha)}{\alpha}$ are increasing on the positive real axis since all the coefficients of $P_i(\alpha)$ are non-negative. It thus holds that

$$\hat{k} P_i\Big(\frac{\beta}{\hat{k}}\Big) \leq P_i(\beta).$$

By the definition of ν, this yields

$$LM\left(\sum_{i=1}^{s} |b_i| P_{i-1}\Big(\frac{\beta}{\hat{k}}\Big)^2 + \frac{1}{2} P_s\Big(\frac{\beta}{\hat{k}}\Big)^2\right) \leq \left(\frac{\beta\mu c - 1}{\hat{k}^2}\right)\nu.$$

Thus $S_2 \leq 0$. In conclusion, $S_1 + S_2 \leq \frac{\nu}{\hat{k}+1}$, so the bound (9) holds for all $k \geq 1$. □

2.4. Nonconvex setting

Without any convexity assumption, it is typically impossible to prove convergence with a certain speed. But we may still prove convergence. The following section is an adaptation of similar arguments in [8] to the Runge–Kutta setting. Since we do not know a priori that there is a unique minimum w_* or even a lower bound on F, we make the following assumption:

Assumption 6. The sequence of iterates $\{w_k\}_{k\in\mathbb{N}}$ is contained in an open set over which F is bounded from below by F_{inf}.

Theorem 2.2. *Let Assumption 1 and Assumptions 3–6 be satisfied. Further, let the step-size sequence $\{\alpha_k\}_{k\geq 1}$ satisfy $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$ and $\sum_{k=1}^{\infty} \alpha_k = \infty$, where $\alpha_k > 0$ for all k and α_1 satisfies (7). Then the following bound holds:*

$$\lim_{K\to\infty} \frac{1}{A_K} \sum_{k=1}^{K} \alpha_k \mathbb{E}_k\big[\|\nabla F(w_k)\|^2\big] = 0,$$

where $A_K = \sum_{k=1}^{K} \alpha_k$.

Remark 2.4. This means that $\liminf_{k\to\infty} \mathbb{E}_k\big[\|\nabla F(w_k)\|^2\big] = 0$, i.e. w_k tends to a (local) minimum of F in a weak sense. But we do not get any further information on how fast this convergence is.

Proof of Theorem 2.2. If α_1 satisfies (7) then so does every α_k, $k \geq 1$, and by taking total expectations in (6) we find that

$$\begin{aligned}\mathbb{E}_k[F(w_{k+1})] - \mathbb{E}_k[F(w_k)] \leq -\frac{1}{2}\alpha_k \mu \mathbb{E}_k\big[\|\nabla F(w_k)\|^2\big] \\ + LM\Big(\sum_{i=1}^{s} |b_i| P_{i-1}(\alpha_k)^2 + \frac{1}{2} P_s(\alpha_k)^2\Big).\end{aligned}$$

By the independence of the $\{\xi_k\}_{k=1}^{\infty}$ and the fact that w_k is independent of ξ_K for $K > k$ we have that $\mathbb{E}_K[F(w_k)] = \mathbb{E}_k[F(w_k)]$ for $K \geq k$. Using this, we obtain a telescopic sum on the left-hand side when we sum over K terms. Along with the fact that

$$F_{\text{inf}} - \mathbb{E}_K[F(w_1)] \leq \mathbb{E}_K[F(w_{K+1})] - \mathbb{E}_K[F(w_1)]$$

and rearranging the terms we thus get

$$\begin{aligned}\frac{1}{2}\mu \sum_{k=1}^{K} \alpha_k \mathbb{E}_K\left[\|\nabla F(w_k)\|^2\right] \leq \mathbb{E}_K[F(w_1)] - F_{\text{inf}} \\ + \sum_{k=1}^{K}\Big(\sum_{i=1}^{s} |b_i| P_{i-1}(\alpha_k)^2 + \frac{L}{2} P_s(\alpha_k)^2\Big) M.\end{aligned} \tag{11}$$

By assumption, we have $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$, which means that also $\sum_{k=1}^{\infty} \alpha_k^i < \infty$ for any integer $i > 2$. But $P_j(\alpha)$ is a polynomial in α of degree j without a constant term, see e.g. Lemma 2.3. Hence

$$P_j(\alpha_k)^2 = \sum_{i=2}^{2j} C_i \alpha_k^i,$$

where C_i are certain constants. This immediately shows that the terms on the second line of (11) are finite, and thus we can conclude that

$$\lim_{K\to\infty} \sum_{k=1}^{K} \alpha_k \mathbb{E}_k\left[\|\nabla F(w_k)\|^2\right] < \infty.$$

By assumption, $\sum_{k=1}^{\infty} \alpha_k = \infty$, and (recalling $A_K = \sum_{k=1}^{K} \alpha_k$) hence

$$\lim_{K\to\infty} \frac{1}{A_K} \sum_{k=1}^{K} \alpha_k \mathbb{E}_k\left[\|\nabla F(w_k)\|^2\right] = 0. \quad \square$$

We may replace the lim inf in Remark 2.4 by a strong limit, if we also assume that F is twice differentiable. We state this result for completeness, but omit the proof since it is very similar to that of [8, Corollary 4.12].

Theorem 2.3. *Let Assumption 1 and Assumptions 3–6 be satisfied, and let $\{\alpha_k\}_{k\geq 1}$ be a step-size sequence as in Theorem 2.2. If we also assume that F is twice differentiable it holds that*

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

3. Specific SRKCD analysis

The first-order RKC method with s stages applied to the gradient flow $\dot{w} = -\nabla F(w)$ with constant time step α is defined by

$$\begin{aligned} w_{k,1} &= w_k, \\ w_{k,2} &= w_k - \tilde{\mu}_1 \alpha \nabla F(w_k), \\ w_{k,j+1} &= (1-\nu_j) w_{k,j} + \nu_j w_{k,j-1} - \tilde{\mu}_j \alpha \nabla F(w_{k,j}), \quad j = 2, \dots, s, \\ w_{k+1} &= w_{k,s+1}, \end{aligned} \tag{12}$$

see e.g. [6, Section V.1]. Here, $w_{k,j}$ denotes the jth internal stage, and $\nabla F(w_{k,j})$ is the corresponding stage derivative. The scalars $\tilde{\mu}_j$ and ν_j are the method-specific coefficients. They are defined via Chebyshev polynomials T_j as

$$\tilde{\mu}_1 = \frac{\omega_1}{T_1(\omega_0)}, \quad \tilde{\mu}_j = \frac{2\omega_1 T_{j-1}(\omega_0)}{T_j(\omega_0)} \text{ and } \nu_j = -\frac{T_{j-2}(\omega_0)}{T_j(\omega_0)}$$

where $\omega_0 = 1 + \frac{\epsilon}{s^2}$ and $\omega_1 = \frac{T_s(\omega_0)}{T_s'(\omega_0)}$. There is thus a single design parameter, ω_0, which is given in terms of ϵ. Setting $\epsilon = 0$ results in the original, un-damped, RKC methods. Instead setting $\epsilon > 0$ introduces extra numerical damping and makes sure that the stability region never degenerates into a single point on the negative real axis. In our numerical experiments, we use the value $\epsilon = 0.01$. We note that we write $\tilde{\mu}_j$ rather than simply μ_j to be consistent with [6], where μ_j would be the quantity $1 - \nu_j$ and an extra term $(1 - \mu_j - \nu_j) w_k$ appears. In our first-order setting, $\mu_j + \nu_j = 1$, and this term cancels. Similarly, the variables ω_0 and ω_1 indicate scalars and should not be confused with elements of the probability space Ω.

Approximating the gradient $\nabla F(w_k)$ by $g(\xi_k, w_k)$ in step k and using the step size α_k now gives us the method we call SRKCD:

$$\begin{aligned} w_{k,1} &= w_k, \\ w_{k,2} &= w_k - \tilde{\mu}_1 \alpha_k g(\xi_k, w_k), \\ w_{k,j+1} &= (1-\nu_j) w_{k,j} + \nu_j w_{k,j-1} - \tilde{\mu}_j \alpha_k g(\xi_k, w_{k,j}), \quad j = 2, \dots, s, \\ w_{k+1} &= w_{k,s+1}. \end{aligned} \tag{13}$$

The method is formulated as a three-term recursion in order to preserve its stability properties under round-off error perturbations. This is similar to how computing the Chebyshev polynomials directly in a naive way quickly leads to a complete loss of precision, whereas evaluating them via a three-term recursion is backwards stable. In order to apply the analysis in the previous section, however, we need to state the method on the standard Runge–Kutta form. This, and verifying Assumption 5, is what the rest of the section is concerned with. Since the SRKCD method has precisely the same coefficients as the RKC method for the full problem $\dot{w} = -\nabla F(w)$, we will consider the RKC formulation for brevity. We will also dispense with the subscript k in α_k, since the varying step size does not matter for the reformulation.

We start by noting that by Lemmas A.1 and A.2 (in the appendix), both $T_s(\omega_0)$ and $T_s'(\omega_0)$ are positive for $s \geq 1$. Hence, $\omega_1 > 0$. Lemma A.1 also shows that $T_j(\omega_0) \geq 1$ for any j, which directly implies that $\tilde{\mu}_1 > 0$, $\tilde{\mu}_j > 0$ and $\nu_j < 0$ for every $j \in \mathbb{N}$. We collect these inequalities in a lemma for later reference:

Lemma 3.1. *With $\omega_0 = 1 + \frac{\epsilon}{s^2}$ chosen as above with $\epsilon \geq 0$, it holds for every $j \in \mathbb{N}$ that $\tilde{\mu}_1 > 0$, $\tilde{\mu}_j > 0$ and $\nu_j < 0$.*

3.1. One-stage update

We first derive an alternative expression for the update $w_{k,j+1} - w_{k,j}$, i.e. what happens from one stage to the next.

Lemma 3.2. *The iterates defined by (12) satisfy*

$$w_{k,j+1} - w_{k,j} = -\alpha \sum_{i=1}^{j} (-1)^{j+i} \left(\prod_{\ell=i+1}^{j} \nu_\ell \right) \tilde{\mu}_i \nabla F(w_{k,i}) \tag{14}$$

for $j = 2, \ldots, s$.

Proof. The proof is by induction. For the base case $j = 1$, we have using (14) that

$$w_{k,2} - w_{k,1} = -\alpha \tilde{\mu}_1 \nabla F(w_{k,1}),$$

which corresponds to the first update of (12). Assume that the identity holds for some j with $2 \leq j \leq s - 1$. According to (12), we then have

$$w_{k,j+2} - w_{k,j+1} = -\nu_{j+1}(w_{k,j+1} - w_{k,j}) - \tilde{\mu}_{j+1}\alpha \nabla F(w_{k,j+1}).$$

We plug in (14) instead of $w_{k,j+1} - w_{k,j}$ and find that the right-hand-side equals

$$-\nu_{j+1} \left(-\alpha \sum_{i=1}^{j} (-1)^{j+i} \left(\prod_{\ell=i+1}^{j} \nu_\ell \right) \tilde{\mu}_i \nabla F(w_{k,i}) \right) - \tilde{\mu}_{j+1}\alpha \nabla F(w_{k,j+1}).$$

Because the product does not depend on i, we can move the ν_{j+1} into it. We can also extend the sum to incorporate the final gradient term, since $i = j + 1$ makes the product equal 1. This leaves us with

$$\begin{aligned} w_{k,j+2} - w_{k,j+1} &= -\alpha \sum_{i=1}^{j} (-1)^{i+j+1} \left(\prod_{\ell=i+1}^{j+1} \nu_\ell \right) \tilde{\mu}_i \nabla F(w_{k,i}) - \tilde{\mu}_{j+1}\alpha \nabla F(w_{k,j+1}) \\ &= -\alpha \sum_{i=1}^{j+1} (-1)^{i+j+1} \left(\prod_{\ell=i+1}^{j+1} \nu_\ell \right) \tilde{\mu}_i \nabla F(w_{k,i}). \end{aligned}$$

The identity (14) thus holds also for $j + 1$ and the proof is complete. □

3.2. Full update

Next, we consider the "full" stage updates $w_{k+1} - w_k$.

Lemma 3.3. *For $1 \leq i \leq s + 1$, the iterates of the RKC method (12) satisfy*

$$w_{k,i} = w_k - \alpha \sum_{j=1}^{i-1} a_{i,j} \nabla F(w_{k,j}),$$

where

$$a_{i,j} = \sum_{n=j}^{i-1} (-1)^{n+j} \left(\prod_{\ell=j+1}^{n} \nu_\ell \right) \tilde{\mu}_j. \tag{15}$$

In particular,

$$w_{k+1} = w_k - \alpha \sum_{i=1}^{s} b_i \nabla F(w_{k,i}),$$

where $b_i = a_{s+1,i}$. *Additionally, every* $a_{i,j} > 0$.

Proof. The particular form of $w_{k,i}$ follows from (14) in the preceding section since $w_k = w_{k,1}$ implies that

$$w_{k,i} - w_k = \sum_{n=1}^{i-1} w_{k,n+1} - w_{k,n} = -\alpha \sum_{n=1}^{i-1} \sum_{j=1}^{n} (-1)^{n+j} \left(\prod_{\ell=j+1}^{n} \nu_\ell \right) \tilde{\mu}_j \nabla F(w_{k,j-1}).$$

Interchanging the order of summation gives

$$w_{k,i} - w_k = -\alpha \sum_{j=1}^{i-1} \left(\sum_{n=j}^{i-1} (-1)^{n+j} \left(\prod_{\ell=j+1}^{n} \nu_\ell \right) \tilde{\mu}_j \right) \nabla F(w_{k,j-1}),$$

where we recognize the coefficients $a_{i,j}$. The expression for w_{k+1} follows by setting $i = s + 1$.

For the final assertion, we note that each of the terms

$$(-1)^{n+j} \left(\prod_{\ell=j+1}^{n} \nu_\ell \right) \tilde{\mu}_j$$

in the sum (15) is positive, since they are the product of $2n$ negative factors: $n+j$ from $(-1)^{n+j}$ and $n-j$ from the product. Since it is a sum of positive terms, the coefficient $a_{i,j}$ is therefore also positive. □

3.3. Convergence

We can now transfer these properties to the SRKCD method and prove that it converges.

Lemma 3.4. *The SRKCD method* (13) *satisfies Assumption* 5.

Proof. The methods (12) and (13) share the same coefficients. By recalling that $w_{k,1} = w_k$ and replacing ∇F with $g(\xi_k, \cdot)$, Lemma 3.3 proves that the method is given on the desired form.

One of the basic Runge–Kutta order conditions requires that $\sum_{i=1}^{s} b_i = 1$. This can be easily verified by inserting the exact solution into the scheme and expanding in Taylor series, see e.g. [18, Section II.1]. Since the corresponding RKC methods are designed to be of order 1 regardless of which s is chosen, part *(i)* of Assumption 5 is fulfilled.

For part *(ii)*, we note that by (15) in Lemma 3.3 we have

$$\begin{aligned}
\sum_{i=1}^{n} a_{n+1,i} &= \sum_{i=1}^{n} \sum_{j=i}^{n} (-1)^{j+i} \left(\prod_{\ell=i+1}^{j} \nu_\ell \right) \tilde{\mu}_i \\
&= \sum_{i=1}^{n-1} \sum_{j=i}^{n} (-1)^{j+i} \left(\prod_{\ell=i+1}^{j} \nu_\ell \right) \tilde{\mu}_i + \tilde{\mu}_n \\
&= \sum_{i=1}^{n-1} a_{n,i} + \sum_{i=1}^{n-1} (-1)^{n+i} \left(\prod_{\ell=i+1}^{n} \nu_l \right) \tilde{\mu}_i + \tilde{\mu}_n.
\end{aligned}$$

By Lemma 3.1, the $\tilde{\mu}_i$-terms are positive, while the ν_l-terms are negative. Each of the terms in the middle sum is thus the product of an even number of negative factors and is therefore positive. From this fact, we conclude that $\sum_{i=1}^{n} a_{n+1,i} > \sum_{i=1}^{n-1} a_{n,i}$. Since the coefficients $a_{n,i}$ are positive by Lemma 3.3 we immediately get also $\sum_{i=1}^{n} |a_{n+1,i}| > \sum_{i=1}^{n-1} |a_{n,i}|$. The sum $\sum_{i=1}^{n-1} |a_{n,i}|$ is thus strictly increasing with n, and bounded from above by $\sum_{i=1}^{s} |a_{s+1,i}| = \sum_{i=1}^{s} a_{s+1,i} = 1$. □

Corollary 3.1. *If Assumptions* 1–4 *are satisfied, then SRKCD converges as stated in Theorem* 2.1. *If instead Assumptions* 1, 3, 4 *and* 6 *are satisfied, SRKCD converges as stated in Theorem* 2.2.

Proof. By Lemma 3.4, Assumption 5 is satisfied. We can therefore apply either Theorem 2.1 or Theorem 2.2. □

3.4. Linearization

We note that Corollary 3.1 does not use the properties of the scheme that makes it an RKC-type method. This is both because we apply it to a nonlinear problem, and because of the stochastic modification. In the rest of this subsection, we will elaborate on this matter.

Consider the full, nonlinear problem $\dot{w} = -\nabla F(w)$ and suppose that F is twice continuously differentiable. Let $z(t)$ be a second, arbitrary solution with $\dot{z} = -\nabla F(z)$, such that $w(t) = z(t) + y(t)$. A linearization around z is then

$$\dot{y} = -\nabla^2 F(z(t))y, \tag{16}$$

where $\nabla^2 F(z(t))$ is the Hessian at $z(t)$. If we further take an equilibrium solution $z(t) \equiv w_*$, we get an autonomous linear initial value problem $\dot{y} = Ay = -\nabla^2 F(w_*)y$. Under Assumption 2, the matrix A has negative eigenvalues, which means that the exact solution $y(t)$ tends to zero as t grows.

If we now apply a Runge–Kutta method and approximate $y(t_k)$ by y_k, then the stability of the scheme is governed by the eigenvalues of A. This is easily seen by diagonalizing A and doing a change of variables. In particular, if R is the stability function of the Runge–Kutta scheme and α_k is the temporal step size, then

$$|R(\alpha_k \lambda_j)| \leq 1$$

should hold for every eigenvalue λ_j of A. With strict inequality, we do not only have stability but that y_k tends to zero just like the exact solution. By considering the situation in somewhat more detail, one can prove that in fact

$$G(y_{k+1}) - G(0) \leq \max_j R(\alpha_k \lambda_j)^2 \big(G(y_k) - G(0)\big),$$

where $G(y) = y^T \nabla^2 F(w_*)y$ with the minimum $y_* = 0$. This is [7, Proposition 1], which considers the (slightly) more general situation $G(y) = y^T A y - b^T y$ with a constant vector b.

We can now utilize information on the stability functions R. For gradient descent, corresponding to the explicit Euler method, stability is guaranteed for step sizes α_k such that $|1 + \alpha_k \lambda_j| \leq 1$ for all j, which implies that $\alpha_k \leq \min_j \frac{-2}{\lambda_j}$. The RKC methods, on the other hand, are constructed such that their stability regions $\{z \in \mathbb{C} \mid |R(z)| \leq 1\}$ cover as much as possible of the negative real line. With s stages, the stability limit will instead be roughly[1] $\alpha_k \leq \min_j \frac{-2s^2}{\lambda_j}$, which allows much larger steps than for normal gradient descent. If the linearized system (16) is a reasonably good approximation of the full nonlinear problem $\dot{w} = -\nabla F(w)$, then we can expect the same behaviour when applying the methods to the full problem.

If we instead apply SGD to the linearized system, we get the iteration

$$\begin{aligned} y_{k+1} &= y_k - \alpha_k \nabla g(\xi_k, w_*) y_k \\ &= \prod_{i=1}^{k} \Big(I - \alpha_k \nabla g(\xi_i, w_*)\Big) y_1. \end{aligned}$$

This indicates that the scheme would be stable if $\|I - \alpha_k \nabla g(\xi_i, w_*)\| \leq 1$ for every i, i.e. $|1 + \alpha_k \lambda_j^i| \leq 1$ for all i and j, where λ_j^i now denotes the eigenvalues of the matrix $\nabla g(\xi_i, w_*)$. Similarly, for SRKCD we get the stability condition $|R(\alpha_k \lambda_j^i)| \leq 1$ for all i and j, which allows a step size which is roughly s^2 larger.

However, in practice this condition is likely both too restrictive and impractical. It is too restrictive because the maximal eigenvalues $\lambda_{\max}^i = \max_j \lambda_j^i$ typically vary significantly with i, see Fig. 1 for an example. The likelihood that the corresponding "worst" $\nabla g(\xi_i, w_*)$ is chosen often enough to be the dominating factor in terms of stability is very small. That is, with high probability, many of the steps could be significantly larger without issue. It is impractical, because there is no clear relation between the eigenvalues of $\nabla g(\xi_i, w_*)$ and those of $\nabla^2 F(w_*)$, meaning that any known overall statistics about the data cannot be used. Further, there is no way to a priori find out which $g(\xi_i, \cdot)$ will be chosen such that the above issue could be alleviated.

For these reasons, we find it unlikely that one could find a proof of convergence of SRKCD with a stability condition that is reasonably sharp and illustrates the benefit of the scheme. Nevertheless, since the RKC methods have stability regions that are roughly s^2 times larger than that of the explicit Euler method, we expect to be able to take roughly s^2 times larger steps with SRKCD instead of SGD.

4. Numerical experiments

In order to investigate the behaviour of the SRKCD method in practice, we have performed numerical experiments on a simple academic test example and on a more complex optimization problem arising in a supervised learning application. The different setups are described in the following subsections.

[1] The exact value depends on the damping parameter ϵ. For small ϵ it is approximately $\min_j \frac{-(2-4/3\epsilon)s^2}{\lambda_j}$, see [6, Section V.1].

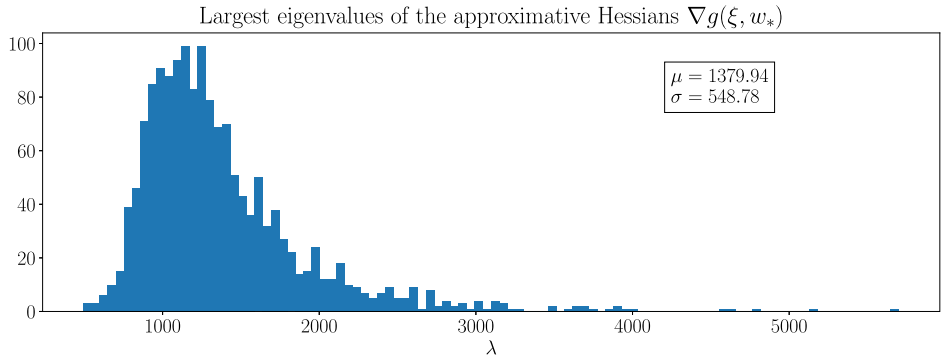

Fig. 1. Here we see the distribution of the largest eigenvalues of $\nabla g(\xi_i, w_*)$ for an optimization problem arising from using a convolutional neural network for image classification. The data set with 60000 images is split into non-overlapping batches of 32 images each, and each ξ_i corresponds to one such batch. Each bar indicates how many such batches have a maximal eigenvalue in the specific range. The mean is $\mu = 1379.94$ and the standard deviation $\sigma = 548.78$.

We have implemented the method in Tensorflow with Keras by observing that (13) can be alternatively expressed as SGD with a very specific momentum term that changes with each stage, and where the same batch of data is used in s consecutive steps. The same idea could equally well be applied in other common machine learning frameworks such as PyTorch. However, we note that it is only valid for relatively small values of s; for large s the three-term recursion (13) is needed to avoid catastrophic round-off error accumulation. We write the momentum equations as

$$\begin{aligned} u_{k,j+1} &= \eta_j u_{k,j} - \ell_j g(\xi_k, w_{k,j}), \\ w_{k,j+1} &= w_{k,j} + u_{k,j+1}, \end{aligned} \tag{17}$$

i.e. $w_{k,j+1} - w_{k,j} = u_{k,j+1}$. But according to (13) we have

$$w_{k,j+1} - w_{k,j} = -\nu_j(w_{k,j} - w_{k,j-1}) - \tilde{\mu}_j \alpha_k g(\xi_k, w_{k,j}),$$

so we see that the two formulations (17) and (13) are equivalent if we set

$$\eta_j = \begin{cases} -\nu_j, & 2 \le s, \\ 0, & j = 1, \end{cases} \qquad \text{and} \qquad \ell_j = \tilde{\mu}_j \alpha_k.$$

The main expected benefit of SRKCD is improved stability properties, and this is what we will focus on in the numerical experiments. First, however, we illustrate Corollary 3.1.

4.1. Convergence with small-scale quadratic convex problem

Let us consider the cost functional

$$F(w) = \frac{1}{N}\sum_{i=1}^{N} f(x^i, w) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d} \frac{(x_j^i)^2 w_j^2}{d},$$

where $d \in \mathbb{N}$ and $w \in \mathbb{R}^d$ are the optimization parameters and each $x^i \in \mathbb{R}^d$ is a known data vector. We take $N = 1000$ and $d = 50$. The vectors x^i were sampled randomly from normal distributions with standard deviation 1 and means $1 + \frac{10i}{d}$. This means that

$$\nabla F(w) = Aw,$$

where A is a diagonal matrix with the diagonal entries

$$\lambda_j = A_{j,j} = \frac{2}{Nd}\sum_{i=1}^{N} (x_j^i)^2.$$

We note that $\{\lambda_j\}_{j=1}^d$ are also the eigenvalues of A. The system is diagonal by design for simplicity, but any system $\dot{w} = Aw$ with a diagonalizable matrix A can be transformed into this form with the eigenvalues preserved. Thus this choice implies no loss of generality.

Since ∇F is linear, we see that Assumptions 1 and 2 are satisfied, with $c = \lambda_{\min} := \min_i \lambda_i$ and $L = \lambda_{\max} := \max_i \lambda_i$. With our particular choice of data, one realization resulted in $\lambda_{\min} = 0.0791$ and $\lambda_{\max} = 4.704$.

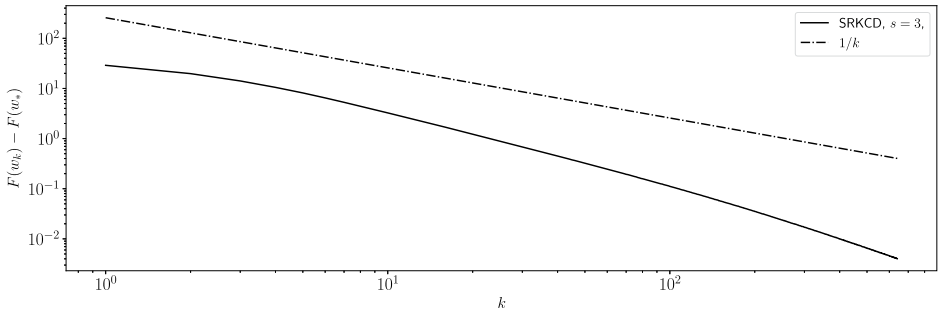

Fig. 2. Loglog plot of the loss for the SRKCD-scheme with 3 stages, run for 20 epochs with a batch-size of 32. Here all the assumptions in Theorem 2.1 are satisfied and we see that we get sublinear convergence.

Further, we approximate ∇F using a batch size of 32, i.e.

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

where $B_\xi \subset \{1, \dots, N\}$ with $|B_\xi| = 32$. Similarly to ∇F, this means that we can write the approximation as

$$g(\xi, w) = \tilde{A}(\xi) w,$$

where $\tilde{A}(\xi)$ is a diagonal matrix with the diagonal entries

$$\tilde{\lambda}_j(\xi) = \tilde{A}(\xi)_{j,j} = \frac{2}{|B_\xi| d} \sum_{i \in B_\xi} (x_j^i)^2.$$

Thus $g(\xi, \cdot)$ is Lipschitz continuous, and since $\tilde{\lambda}_j(\xi) \leq \frac{N}{|B_\xi|} \lambda_j$, Assumption 3 is satisfied. This inequality also shows that Assumption 4(iii) is satisfied with $M = 0$ and $M_G \leq N^2/32^2$. Finally, a straightforward calculation shows that Assumption 4(i) and (ii) holds with $\mu = \mu_G = 1$.

We now choose $\beta = \frac{1}{c} \approx 12.65$ according to the condition in Theorem 2.1. The above bound for M_G is based on estimating the partial sums with the full sums, which in general leads to a large overestimation. By stochastic sampling we determined that $M_G = 2$ constitutes a more representative, sharper, bound, and we therefore use this value below for determining the initial step size α_1. This step size must satisfy (7), so we solve the equation

$$Q\left(\frac{\beta}{\gamma + 1}\right) + \frac{\beta}{2(\gamma + 1)} = 0,$$

which gives us $\gamma \approx 178$ and $\alpha_1 \approx 0.07$.

We apply the SRKCD with $s = 3$ stages to this problem with the parameters stated above and run it for 20 epochs, corresponding to 640 iterations. The loss $F(w_k) - F(w_*) = F(w_k)$ is plotted in Fig. 2, and as expected we observe (at least) sublinear convergence.

The assumptions on the step size given in Theorem 2.1 are sufficient but not necessary. In particular, we expect that the good stability properties of RKC will allow us to use larger step sizes. In Fig. 3 we see the result of the same experiment but where the chosen parameters do not satisfy the assumptions in Theorem 2.1. Here we have chosen $\gamma = 1$ and $\beta = 4$, which results in $\alpha_1 = 2$. This is significantly larger than in the previous experiment, but we see that we still get similar sublinear convergence.

4.2. Stability with small-scale quadratic convex problem

As seen in the previous section, often the variance-related quantity M_G is not explicitly known. Thus, even in the case of SGD where the condition (7) reduces to $\alpha \leq \frac{\mu}{LM_G}$, it is not obvious how to best choose the initial step size; some trial-and-error and parameter sweeps is required. If a step size is chosen too large during this process, there will be issues with stability. A method with good stability properties such as SRKCD will be less affected by this, and could work well also in the case of badly estimated parameters.

We continue to investigate the setting described in the previous section and use the same data. To better illustrate the stability properties of the methods, from now on we only use fixed step sizes α rather than the previous sequences α_k. Stability is only an issue for large step sizes, so with a decreasing step size we will always have stability eventually.

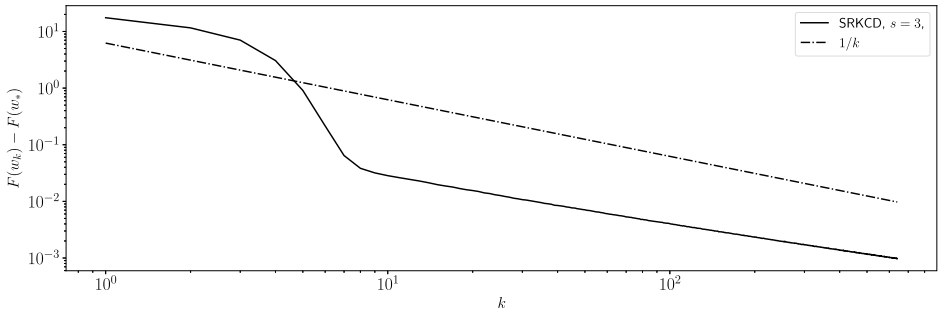

Fig. 3. Loglog plot of the loss for the SRKCD-scheme with 3 stages, run for 20 epochs with a batch-size of 32. Here the initial step size is larger than allowed for in Theorem 2.1 but we still get sublinear convergence. Here $\beta = 4$ and $\gamma = 1$.

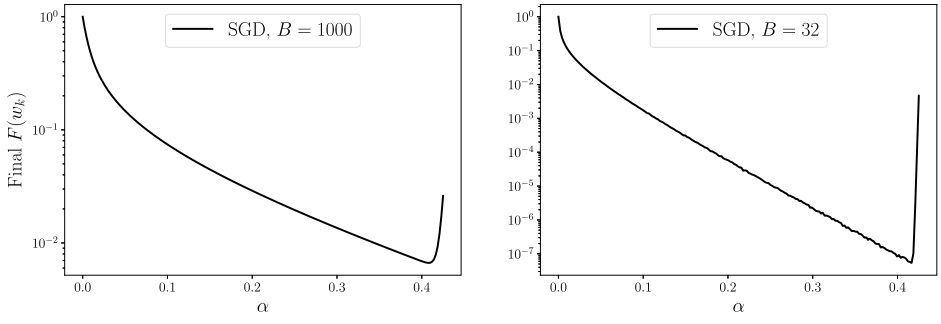

Fig. 4. SGD with batch size 1000, i.e. GD, (left) and SGD with batch size 32 (right) when applied to the problem described in Section 4.1. Note the different scales on the y-axes and that different number of iterations were used.

However, typically even a small number of unstable steps in the initial phase will lead to an extremely large $F(w_k)$ that will take an unfeasibly large number of steps to recover from.

Since the system is diagonal, stability is determined by the eigenvalues as discussed in Section 3.4. In particular, if we use ∇F instead of stochastic approximations $g(\xi, \cdot)$, then for stability we must have $\alpha_k \leq \frac{2}{\lambda_{\max}} = \frac{2}{L}$. Further, the optimal step size which minimizes $\max_j |R(h\lambda_j)|$ is $\alpha = \frac{2}{\lambda_{\min}+\lambda_{\max}}$, see e.g. [7].

We ran 15 iterations of GD and 3 epochs of SGD with a batch size of 32 and with different step sizes between 0 and $2/L = 0.4251$. We use more iterations for SGD simply because the GD iterations are more expensive. The final values $F(w)$ are plotted in Fig. 4. For GD, we can clearly observe the optimal step size choice $\frac{2}{\lambda_{\min}+\lambda_{\max}} = 0.4181$. Closer to $\alpha = 2/L$, the values start to increase again and larger step sizes will lead to instability and divergence. Interestingly, the picture is very similar for SGD. In this case, the step size limit is very slightly smaller than $\alpha = 2/L$ and we can observe some wiggles in the curve due to the stochastic approximations. But the optimal step size choice stays at almost the same position.

In Fig. 5, we repeat the experiment with a batch size 32 but now with the SRKCD methods with different s. For each s, we try step sizes $\alpha \in (0, \frac{b_R}{L})$ where b_R is the maximal value such that $(-b_R, 0)$ is included in the stability region for the corresponding RKC method. It can be shown that $b_R = \frac{2\omega_0 T_s'(\omega_0)}{T_s(\omega_0)}$ [6, p. 425].

The first thing to note is that as expected, the stability regions are much larger than for SGD. For larger s, they do not quite reach b_R/L in this stochastic setting, but the differences are extremely small. Secondly, we note that all the methods exhibit a characteristic "dip" at a relatively small step size. This is similar to the optimal step size dip at $\frac{2}{\lambda_{\min}+\lambda_{\max}}$ for SGD. However, since the stability function of the corresponding RKC method has s zeroes instead of only one, there are also many other choices of larger α which yield comparable performance. Indeed, while SGD performs quite well in the interval $\alpha \in (0.3, 0.42)$, SRKCD with $s = 5$ performs roughly equally well for all $\alpha \in (0.5, 10.2)$.

We note that these plots cannot be used for efficiency comparisons, since the latter method has used 5 times as many evaluations of $g(\xi, \cdot)$ as SGD. Nevertheless, it is clear that the improved RKC stability properties makes SRKCD more robust. If, e.g. the values of $\lambda_{\min}$ and $\lambda_{\max}$ were not known, then selecting a good step size for SGD is difficult. For SRKCD, the choice almost does not matter.

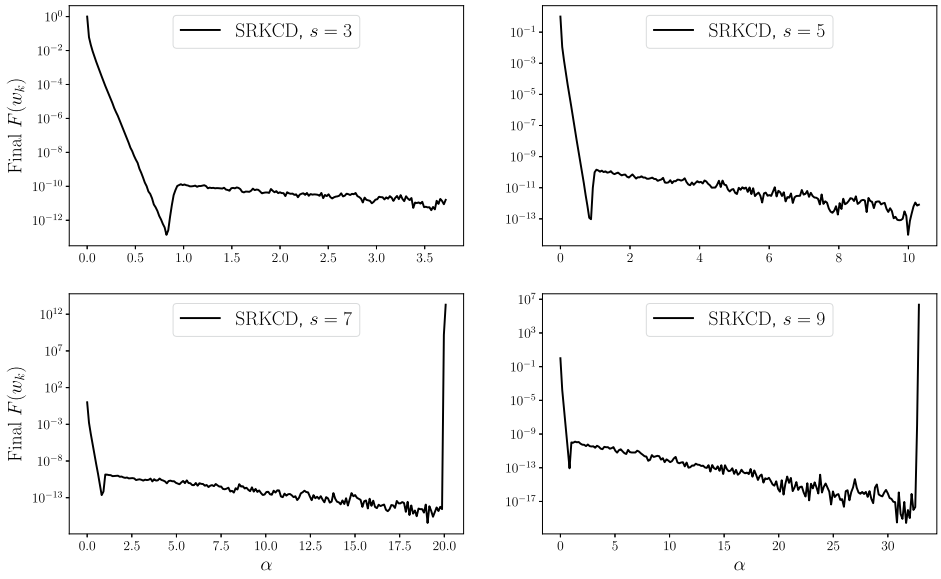

Fig. 5. SRKCD with batch size 32 for various values of s when applied to the problem described in Section 4.1. In each case, 3 epochs were run.

4.3. Convolutional neural network

Next, we consider also an example arising from a real-world problem, namely the classification of images by convolutional neural networks. Such a problem can also be stated on the form $\min_w F(w)$, where F now depends on the collection of images, the network structure, and the loss function used to penalize mis-classifications. We refer to e.g. [8] for details. For this particular experiment, we set up a simple convolutional neural network consisting of one convolutional layer with a kernel size of 32×32 upon which we stack two fully connected dense layers with 128 and 10 neurons each. The activation function is ReLu for the first dense layer and softmax for the output layer and we use a crossentropy loss function. We train this network on the MNIST dataset [19] using both the SGD and the SRKCD algorithm with various step sizes and number of stages s.

While a single training sequence is not so expensive, repeating it many times like in the previous section quickly becomes very time-consuming. Instead of illustrating the behaviour of the methods over a whole interval $\alpha \in (0, a)$ for some a, we therefore settle for trying to pin down the practical stability boundary. We recall that since this problem is nonlinear, we cannot expect the stability properties to behave as nicely as in the previous experiment. This problem is also larger, but we still use a batch size of 32. As a consequence, the variance is larger than in the previous experiment, i.e. every realization is noisier. To alleviate this, we run each step size 5 times and take the average.

Fig. 6 shows the final averaged loss values $F(w_k)$ after 1000 iterations for SGD and SRKCD with $s = 3, 4, 5$, for 10 step sizes close to the stability limit. The loss function F saturates around 2.4 which means that for such values the methods are unstable. Smaller values do not rule out that the methods could diverge in further iterations, but typically it rather indicates that we simply did not yet use enough iterations to decrease the loss further. Thus we can observe that for SGD, the practical stability limit is at around $\alpha = 0.35$. For SRKCD with $s = 3$, we instead estimate it to about $\alpha = 1.9$. For $s = 4$ and $s = 5$, we get about $\alpha = 2.8$ and $\alpha = 3.9$. Clearly these are very rough estimates, but as expected the stability properties of SRKCD are superior also in the nonlinear case. We note that e.g. $1.9 < 3^2 \cdot 0.35 = 3.15$, i.e. the s^2-scaling of the stability regions is not preserved for nonlinear problems. However, this is just one example and other types of problems might behave differently. Fully understanding the general nonlinear setting is a significant research undertaking.

5. Conclusions

We have introduced and analysed the stochastic Runge–Kutta–Chebyshev descent (SRKCD) method by showing convergence in expectation to a unique minimum for a strongly convex objective function, and to a stationary point under certain regularity assumptions in the nonconvex case. While we have focused on the SRKCD methods because they exhibit the particular stability properties that were our original motivation, the proof is more general and applies

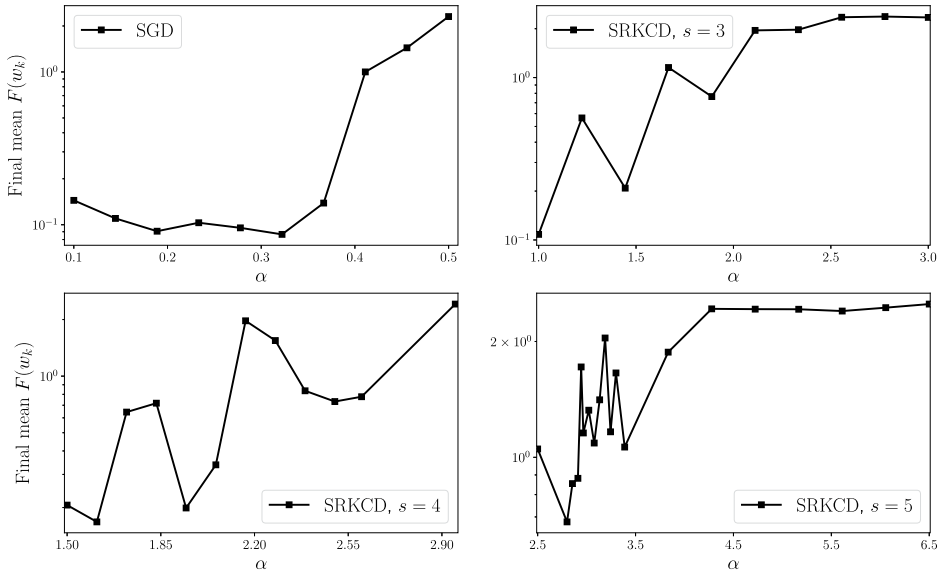

Fig. 6. SGD and SRKCD with batch size 32 for various values of s when applied to the problem described in Section 4.2. In each case, 1000 iterations were run and the average final value of $F(w_k)$ over 5 paths is plotted versus the step size α.

to essentially any Runge–Kutta method. Other such methods may have properties that are of interest in this setting, this remains an open interesting research question.

As we have seen from the numerical experiments, the stability properties of the SRKCD methods are superior to SGD. This remains true also for nonlinear and nonconvex problems. We aim to investigate the efficiency of SRKCD in more detail, and also to compare it more extensively to other popular optimization methods. A key point to take into account here is of course that one iteration of SRKCD requires s approximative gradient evaluations, while most similar methods such as SGD require only one. In the usual setting of stiff ODEs, this is outweighed by being able to take much longer steps. In the current optimization context where it is not necessarily ideal to take the largest possible step, it is no longer as clear. We have, nevertheless, seen from the numerical experiment in Section 4.2 that we can expect the SRKCD methods to be more robust in the sense that more step size choices give reasonable results in the absence of good model parameter estimates.

Finally, we note that in this stochastic setting one must use a decreasing step size sequence to actually reach a local minimum. With a fixed step size, we will only reach a neighbourhood of the minimum, whose size depends on the step size and on the variance of the approximative gradients. But with a very small step size, the better stability properties of SRKCD are irrelevant. These methods are therefore best employed in the initial phase where larger step sizes can and should be used, and where the convergence towards the minimum is rapid. We think that a hybrid method which utilizes SRKCD with decreasing values of s, eventually becoming SGD at $s = 1$, could be ideal.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, Sweden.

Appendix. Auxiliary results

In this appendix, we collect a few results that are important to our analysis but which are not of great interest on their own.

A.1. Chebyshev polynomials

The Chebyshev polynomials are given by

$$T_0(x) = 1, \quad T_1(x) = x,$$
$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x), \quad n \geq 2.$$

Lemma A.1. *For fixed $x \geq 1$ it holds that $T_n(x) \geq T_{n-1}(x)$ for $n \geq 1$. As a consequence, $T_n(x) \geq 1$ for all $n \geq 0$ if $x \geq 1$.*

Proof. We prove the lemma by induction. The statement is clearly true for $n = 1$. Assume that it is true for $n = k$, i.e. $T_k(x) - T_{k-1}(x) \geq 0$ for $x \geq 1$. Then

$$\begin{aligned} T_{k+1}(x) &= 2xT_k(x) - T_{k-1}(x) \\ &\geq 2T_k(x) - T_{k-1}(x) \\ &= T_k(x) + (T_k(x) - T_{k-1}(x)) \geq T_k(x). \end{aligned}$$

The fact that $T_n(x) \geq 1$ then follows directly from $T_0(x) = 1$. □

The RKC-update also depends on the derivatives of the Chebyshev polynomials so we also prove the same result for these:

Lemma A.2. *For fixed $x \geq 1$ it holds that $T_n'(x) \geq T_{n-1}'(x)$ for $n \geq 1$. Further, $T_n'(x) \geq 4$ for $n \geq 2$ if $x \geq 1$.*

Proof. From the definition of T_n, we find the following recursive formula for the derivatives $T_n'(x)$:

$$\begin{aligned} &T_0'(x) = 0, \ T_1'(x) = 1, \\ &T_n'(x) = 2T_{n-1}(x) + 2xT_{n-1}'(x) - T_{n-2}'(x), \quad n \geq 2. \end{aligned}$$

Now we can use induction again like in the previous Lemma. We clearly have $T_1'(x) \geq T_0'(x)$. Assuming that $T_n'(x) \geq T_{n-1}'(x)$ holds we get

$$\begin{aligned} T_{n+1}'(x) &= 2T_n(x) + 2xT_n'(x) - T_{n-1}'(x) \\ &\geq T_n'(x) + \left(T_n'(x) - T_{n-1}'(x)\right) \geq T_n'(x), \end{aligned}$$

where we used $T_n(x) \geq 1$ from Lemma A.1 in the first inequality. The final statement follows directly from the fact that $T_2'(x) = 4x$. □

References

[1] H. Robbins, S. Monro, A stochastic approximation method, Ann. Math. Statist. 22 (1951) 400–407, http://dx.doi.org/10.1214/aoms/1177729586.
[2] A.J. Owens, D.L. Filkin, Efficient training of the backpropagation network by solving a system of stiff ordinary differential equations, in: International 1989 Joint Conference on Neural Networks, Vol. 2, 1989, pp. 381–386, http://dx.doi.org/10.1109/IJCNN.1989.118726.
[3] P. Bianchi, Ergodic convergence of a stochastic proximal point algorithm, SIAM J. Optim. 26 (2016) 2235–2260, http://dx.doi.org/10.1137/15M1017909.
[4] M. Eisenmann, T. Stillfjord, M. Williamson, Sub-linear convergence of a stochastic proximal iteration method in Hilbert space, 2020, arXiv e-prints http://arxiv.org/abs/2010.12348 [arXiv:2010.12348].
[5] P.J. van der Houwen, B.P. Sommeijer, On the internal stability of explicit, m-stage runge–kutta methods for large m-values, Z. Angew. Math. Mech. 60 (1980) 479–485, http://dx.doi.org/10.1002/zamm.19800601005.
[6] W. Hundsdorfer, J.G. Verwer, Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations, Springer, Berlin, Heidelberg, 2003, http://dx.doi.org/10.1007/978-3-662-09017-6.
[7] A. Eftekhari, B. Vandereycken, G. Vilmart, K.C. Zygalakis, Explicit stabilised gradient descent for faster strongly convex optimisation, BIT Numer. Math. 61 (2021) 119–139, http://dx.doi.org/10.1007/s10543-020-00819-y.
[8] L. Bottou, F.E. Curtis, J. Nocedal, Optimization methods for large-scale machine learning, SIAM Rev. 60 (2018) 223–311, http://dx.doi.org/10.1137/16M1080173.
[9] B. Polyak, Some methods of speeding up the convergence of iteration methods, USSR Comput. Math. Math. Phys. 4 (1964) 1–17, http://dx.doi.org/10.1016/0041-5553(64)90137-5.
[10] I. Sutskever, J. Martens, G. Dahl, G. Hinton, On the importance of initialization and momentum in deep learning, in: S. Dasgupta, D. McAllester (Eds.), Proceedings of the 30th International Conference on Machine Learning, 28 of Proceedings of Machine Learning Research, PMLR, Atlanta, Georgia, USA, 2013, pp. 1139–1147, URL https://proceedings.mlr.press/v28/sutskever13.html.
[11] S. Gadat, F. Panloup, S. Saadane, Stochastic heavy ball, Electron. J. Stat. 12 (2018) 461–529, http://dx.doi.org/10.1214/18-EJS1395.
[12] Y.E. Nesterov, A method for solving the convex programming problem with convergence rate $O(1/k^2)$, Sov. Math. Dokl. 27 (1983) 372–376; Translation from Dokl. Akad. Nauk SSSR 269 (3) (1983) 543-547.
[13] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, J. Mach. Learn. Res. 12 (2011) 2121–2159, URL http://jmlr.org/papers/v12/duchi11a.html.
[14] M.D. Zeiler, ADADELTA: An adaptive learning rate method, 2012, arXiv e-prints http://arxiv.org/abs/1212.5701 [arXiv:1212.5701].
[15] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Published As a Conference Paper At the 3rd International Conference for Learning Representations, San Diego, 2015, 2017, arXiv e-prints http://arxiv.org/abs/1412.6980 [arXiv:1412.6980].
[16] G. Hinton, Coursera Neural Networks for Machine Learning Lecture 6, 2018.
[17] H.H. Bauschke, P.L. Combettes, Convex analysis and monotone operator theory in Hilbert spaces, in: CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, second ed., Springer, Cham, 2017, http://dx.doi.org/10.1007/978-3-319-48311-5.
[18] E. Hairer, S.P. Nörsett, G. Wanner, Solving ordinary differential equations, in: I, 8 of Springer Series in Computational Mathematics, Second revised edition, Springer, Berlin, 2009, http://dx.doi.org/10.1007/978-3-540-78862-1, nonstiff problems, paperback.
[19] Y. LeCun, C. Cortes, C. Burges, MNIST handwritten digit database, 2010, Available at http://yann.lecun.com/exdb/mnist.

Printed by Media-Tryck, Lund 2023 NORDIC SWAN ECOLABEL 3041 0903

Licentiate Thesis in Mathematical Sciences 2023:1

ISBN 978-91-8039-558-8
ISSN 1404-0034