SUB-LINEAR CONVERGENCE OF A STOCHASTIC PROXIMAL ITERATION METHOD IN HILBERT SPACE

MONIKA EISENMANN, TONY STILLFJORD, AND MÅNS WILLIAMSON

Abstract. We consider a stochastic version of the proximal point algorithm for 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 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. We illustrate these results by discretizing a concrete infinite-dimensional classification problem with varying degrees of accuracy.

1. Introduction

We consider convex optimization problems of the form

$$w^* = \arg\min_{w \in H} F(w), \tag{1.1}$$

where

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

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

$$F(w) = \frac{1}{n} \sum_{j=1}^{n} \ell(h(w, x_j), y_j), \tag{1.2}$$

where ℓ is a given loss function. We refer to, e.g., Bottou, Curtis & Nocedal [1] 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

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

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

E-mail addresses: monika.eisenmann@math.lth.se, tony.stillfjord@math.lth.se, mans.williamson@math.lth.se.

2020 *Mathematics Subject Classification.* 46N10 and 65K10 and 90C15.

Key words and phrases. stochastic proximal point and convergence analysis and convergence rate and infinite-dimensional and Hilbert space.

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The authors would like to thank Eskil Hansen for valuable feedback.

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, $\{\xi^k\}_{k\in\mathbb{N}}$ are 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

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

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 [2]. Several convergence results and connections to other methods such as the Douglas–Rachford splitting are collected in Eckstein & Bertsekas [3]

Following Ryu & Boyd [4], 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 Section 5.

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

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

in an infinite-dimensional setting, i.e. where $\{w^k\}_{k\in\mathbb{N}}$ and w^* live in a Hilbert space H. As shown in e.g. [5, 6], 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 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 $\nabla F \colon H \to 2^H$ are considered and weak ergodic convergence is proved. In Rosasco et al. [8], 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 $k \to \infty$.

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 [4], which suggests $1/k$ convergence in $\mathbb{R}^d$. Based on [4], the implicit method has also been considered in a few other works: In Patrascu & Necoara [9], 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 & Drusvyatskiy [10]. Toulis & Airoldi and Toulis et al. studied such an implicit scheme in [11, 12, 13].

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, Tran & Airoldi [14, 15] and Ryu & Yin [16], 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 & Irofti [17], Ryu & Yin [16], Salim et al. [18], Bianchi & Hachem [19] and Bertsekas [20]. A few more related schemes have been considered in Asi & Duchi [21, 22] and Toulis, Horel & Airoldi [23]. More information about the complexity of solving these kinds of implicit equations and the corresponding implementation can be found in Fagan & Iyengar [24] and Tran, Toulis & Airoldi in [15].

Our aim is to bridge the gap between the "strong finite-dimensional" and "weak infinite-dimensional" settings, by extending the approach of [4] 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^*. 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. We will follow [4] closely, because their approach is sound. 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 less bounded operators.

A brief outline of the paper is as follows. The main assumptions that we make are stated in Section 2, as well as the main theorem. Then we prove a number of preliminary results in Section 3, before we can tackle the main proof in Section 4. In Section 5 we describe a numerical experiment that illustrates our results, and then we summarize our findings in Section 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 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 $\iota\colon H^* \to H$ such that $\iota^{-1}\colon H \to H^*$ with $\iota^{-1} : u \mapsto (u, \cdot)$. Furthermore, the dual pairing is denoted by $\langle u', u\rangle = u'(u)$ for $u' \in H^*$ and $u \in H$. It satisfies

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

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 $f(\cdot, \xi)\colon H \times \Omega \to \mathbb{R}$, 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 a certain random variable.

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

$$w^{k+1} = w^k - \alpha_k \iota \nabla f(w^{k+1}, \xi^k) \quad \text{in } H, \qquad w^1 = w_1 \quad \text{in } H, \tag{2.1}$$

for minimizing

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

where f and F fulfil the following assumption.

Assumption 1. *For a random variable ξ, the function $f(\cdot, \xi)\colon H \times \Omega \to \mathbb{R}$ is lower semi-continuous, convex and proper as well as Gâteaux differentiable on H almost surely. Additionally, f fulfills the following conditions:*

- *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 $L_R\colon \Omega \to \mathbb{R}$ such that*

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

  almost surely for all $u, v \in H$ with $\|u\|, \|v\| \leq R$. Furthermore, there exists a polynomial $P\colon \mathbb{R} \to \mathbb{R}$ of degree $2^m - 2$ such that $L_R \leq P(R)$ almost surely.
- *There exists a random variable $M\colon \Omega \to \mathcal{L}(H)$ such that the image is symmetric and for all $u, v \in H$*

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

  is fulfilled almost surely. Further, there exists a random variable $\bar{\mu}\colon \Omega \to [0, \infty)$ such that $(Mu, u) \geq \bar{\mu}\|u\|^2$ for all $u \in H$ with $\mathbf{E}[\bar{\mu}] = \mu > 0$ and $\mathbf{E}[\bar{\mu}^2] = \nu^2 < \infty$.
- *The situation is not degenerate, in the sense that there exists a $u \in H$ such that $F(u) < \infty$ and*

  $$\mathbf{E}_\xi\Big[\inf_{u \in H} f(u, \xi)\Big] > -\infty.$$

An immediate result of Assumption 1, is that the gradient $\nabla f(\cdot, \xi)$ is maximal monotone almost surely, see [25, Theorem A]. As a consequence, the resolvent $T_f = (I + \nabla f(\cdot, \xi))^{-1}$ is well-defined almost surely, see Lemma 3.1 for more details. Furthermore, it is straightforward to show that F is Gâteaux differentiable, lower semi-continuous, strictly convex and proper by employing dominated convergence and Fatou's lemma ([26, Theorem 2.3.6]). See, e.g., [4] for the main ideas. As a consequence, there is a unique minimum w^* of (1.1).

Remark 2.1. The idea behind the operators M is that each f is is allowed to be only convex rather than strongly convex. However, they should be strongly convex in *some* directions, such that f is strongly convex *in expectation*.

Remark 2.2. 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 [27] is a natural idea. In each step, an elliptic problem then has to be solved (to apply ι), but this can often be done very efficiently.

For a random variable X, we let $\mathbf{E}_{\xi^k}[X]$ denote the expectation with respect to ξ^k given w^{k-1}. We are interested in the total expectation

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

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 $\xi^1, \dots, \xi^{k-1}$. Our main theorem states that we have sub-linear convergence of the iterates w^k to w^* measured in this expectation:

Theorem 2.1. *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 (2.1) converges sub-linearly if the step sizes fulfill $\alpha_k = \frac{\eta}{k}$ with $\eta > \frac{1}{\mu}$. In particular, the error bound*

$$\mathbf{E}_{k-1}\big[\|w^k - w^*\|^2\big] \leq \frac{C}{k}$$

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

The proof of this theorem is given in Section 4. In order to prepare for this, however, we first need a series of preliminary results.

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 3.1. *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 $w^{k+1}\colon \Omega \to H$ that fulfills (2.1) and is measurable with respect to the joint probability distribution of $\xi^1, \dots, \xi^k$.*

Proof. We define the mapping

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

For almost all $\omega \in \Omega$, the mapping $f(\cdot, \xi^k(\omega))$ is lower semi-continuous, proper and convex. Thus, by [25, Theorem A] $\nabla f(\cdot, \xi^k(\omega))$ is maximal monotone. By [28, Theorem 2.2], this shows that the operator $\iota^{-1} + \alpha_k \nabla f(\cdot, \xi^k(\omega))\colon H \to H^*$ is surjective. Furthermore, due to the monotonicity of $\nabla f(\cdot, \xi^k(\omega))$ it follows that

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

which implies

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

This verifies that $I + \alpha_k \iota \nabla f(\cdot, \xi^k(\omega))$ is injective and, in particular, bijective. Therefore, there exists a unique element $w^{k+1}(\omega)$ such that

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

We can now apply [29, Lemma 2.1.4] or [30, Lemma 4.3] and obtain that $\omega \mapsto w^{k+1}(\omega)$ is measurable. □

For the further analysis, we now introduce the function $\tilde{f}(\cdot,\xi)\colon H\times\Omega\to\mathbb{R}$ given by

$$\tilde{f}(u,\xi) = f(u_0,\xi) + \langle\nabla f(u_0,\xi), u-u_0\rangle + \frac{1}{2}(M(u-u_0),u-u_0), \tag{3.1}$$

where $u_0\in H$ is a fixed parameter. This mapping is a convex approximation of f. Furthermore, we define the function $\tilde{r}(\cdot,\xi)\colon H\times\Omega\to\mathbb{R}$ given by

$$\tilde{r}(u,\xi) = f(u,\xi) - \tilde{f}(u,\xi). \tag{3.2}$$

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

$$\begin{aligned}\nabla\tilde{f}(u,\xi) &= \nabla f(u_0,\xi) + (M(u-u_0),\cdot)\\ \nabla\tilde{r}(u,\xi) &= \nabla f(u,\xi) - \nabla f(u_0,\xi) - (M(u-u_0),\cdot)\end{aligned}$$

for $u\in H$ almost surely.

Lemma 3.2. *The function $\tilde{r}$ defined in (3.2) 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 $u,v\in H$ almost surely. The gradient $\nabla\tilde{r}(\cdot,\xi)$ is monotone almost surely.*

Proof. In the following proof, let us omit ξ for simplicity and let $u,v\in H$. Since f is M-convex almost surely, it follows that

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

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

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

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

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

It then follows that

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

As $\tilde{r}$ is convex almost surely this implies that $\nabla\tilde{r}$ is monotone almost surely. □

Lemma 3.3. *Let Assumption 1 be fulfilled and let $\tilde{f}$ be defined as in (3.1). Then the operator*

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

is well-defined. If a function $r\colon H\times\Omega\to\mathbb{R}$ is Gâteaux differentiable, lower semi-continuous, convex and proper almost surely then

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

is well-defined.

If there exist $Q\colon H\times\Omega\to H^*$ and $z\colon\Omega\to H^*$ such that $\nabla r(u,\xi)=Qu+z$ then the resolvent can be represented by

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

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 [25, Theorem A] and [28, Theorem 2.2] analogously to the argumentation in the proof of Lemma 3.1.

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

$$(I + \iota\nabla\tilde{f} + \iota\nabla r)^{-1}v = T_{\tilde{f}+r}v =: u \quad \text{almost surely.}$$

Then it follows that

$$\begin{aligned} v &= (I + \iota\nabla\tilde{f} + \iota\nabla r)u \\ &= (I + M + \iota Q)u + \iota\nabla f(u_0) - Mu_0 + \iota z \quad \text{almost surely.} \end{aligned}$$

Rearranging the terms, yields

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

□

Next, we will show that the contraction factors of T_f and $T_{\tilde{f}}$ are related. For this, we need the following basic identities and inequalities for operators on H, as well as some stronger inequalities that hold for symmetric positive operators on H.

Lemma 3.4. *Let Assumption 1 be satisfied and let $\tilde{f}$ and $\tilde{r}$ be given as in (3.1) and (3.2), respectively. Then the identities*

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

are fulfilled almost surely.

Proof. By the definition of T_f, we have that

$$T_f + \iota\nabla f(T_f, \xi) = (I + \iota\nabla f(\cdot, \xi))T_f = I,$$

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

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

□

As a consequence of Lemma 3.4 we have the following basic inequalities:

Lemma 3.5. *Let Assumption 1 be satisfied. It then follows that*

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

almost surely for every $u \in H$. Additionally, if for $R > 0$ the bound $\|u\| + \|\nabla f(u, \xi)\| \leq R$ holds true almost surely, then the second-order estimate

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

is fulfilled almost surely.

Proof. In order to shorten the notation, we omit the ξ in the following proof. For the first inequality, we note that since ∇f is monotone, we have

$$\langle \nabla f(T_f u) - \nabla f(u), T_f u - u \rangle \geq 0 \quad \text{almost surely.}$$

Thus, by the first identity in Lemma 3.4,

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

follows almost surely. But by Cauchy-Schwarz, we also have

$$(-\nabla f(u), T_f u - u) \leq \|\nabla f(u)\|_{H^*} \|T_f u - u\| \quad \text{almost surely,}$$

which in combination with the previous inequality proves the 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 almost surely. Thus, we obtain

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

almost surely. □

Lemma 3.6. *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 [31]. First observe that the operator $(I + Q^{-1}S)^{-1} \in \mathcal{L}(H)$ by e.g. [32, Lemma 2A.1]. Then, since

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

and

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

we find that

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

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

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

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 $\big(S\big(I + Q^{-1}S\big)^{-1}u, u\big) \geq 0$.

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

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

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. □

Lemma 3.7. *Let Assumption 1 be fulfilled and let $\tilde{f}$ be given as in (3.1). Then*

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

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

Proof. For better readability, we once again omit ξ where there is no risk of confusion. For $u, v \in H$ and $\varepsilon > 0$, we approximate the function $\tilde{r}$ defined in (3.2) by

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

where

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

almost surely. As we can write

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

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

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

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

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

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

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

and therefore

$$T_f u = (I + \iota \nabla \tilde{f} + \iota \nabla \tilde{r}_\varepsilon)^{-1} u = T_{\tilde{f}+\tilde{r}_\varepsilon} u \quad \text{almost surely.}$$

Applying Lemma 3.4, we find that

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

almost surely. This shows that

$$T_f v = (I + \iota \nabla \tilde{f} + \iota \nabla \tilde{r}_\varepsilon)^{-1}(v + \varepsilon(T_f v - T_f u)) = T_{\tilde{f}+\tilde{r}_\varepsilon}(v + \varepsilon(T_f v - T_f u)) \tag{3.3}$$

almost surely. Using the explicit representation of $T_{\tilde{f}+\tilde{r}_\varepsilon}$ from Lemma 3.3, it follows that

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

almost surely. Therefore, we have

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

almost surely, since

$$\Big(\Big(I + M + \iota\Big(\frac{\langle v_\varepsilon, \cdot\rangle}{a_\varepsilon} v_\varepsilon\Big)\Big)u, u\Big) \ge \|u\|^2 \quad \text{almost surely}$$

means that we can apply Lemma 3.6. Thus, this shows that $T_f u = T_{f+\tilde{r}_\varepsilon} u$ and $T_f v = \lim_{\varepsilon\to 0} T_{\tilde{f}+\tilde{r}_\varepsilon} v$ almost surely. Further, we can state an explicit representation for $T_{\tilde{f}}$ using Lemma 3.3 given by

$$T_{\tilde{f}} z = (I + \iota\nabla\tilde{f})^{-1} z = (I + M)^{-1}\big(z - \iota\nabla f(u_0) + Mu_0\big) \quad \text{almost surely.}$$

For $n = \frac{u-v}{\|u-v\|}$ with $\|n\| = 1$, we obtain using Lemma 3.6

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

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

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

□

Lemma 3.8. *Let Assumption 1 be satisfied and let $\tilde{f}$ be given as in (3.1). Then for $u, v \in H$ and $\alpha > 0$,*

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

is fulfilled. Furthermore, it follows that

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

Proof. Due to the explicit representation of $T_{\alpha\tilde{f}}$ stated in Lemma 3.3, we find that

$$T_{\alpha\tilde{f}} u - T_{\alpha\tilde{f}} v = (I + \alpha M)^{-1}(u - v) \quad \text{almost surely}$$

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

$$\mathbf{E}_\xi\big[\|(I + \alpha M)^{-1}(u - v)\|^2\big] \le \mathbf{E}_\xi\big[\|(I + \alpha M)^{-1}\|^2_{\mathcal{L}(H)}\big]\|u - v\|^2.$$

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

$$((I + \alpha M)u, u) \geq (1 + \overline{\mu}\alpha)\|u\|^2 \quad \text{almost surely.}$$

Due to Lemma 3.6 it follows that

$$\|(I + \alpha M)^{-1}\|^2_{\mathcal{L}(H)} \leq (1 + \overline{\mu}\alpha)^{-2} \quad \text{almost surely.}$$

The right-hand-side bound is a C^2-function with respect to α (in fact, C^∞). By a second-order expansion in a Taylor series we can therefore conclude that

$$\|(I + \alpha M)^{-1}\|^2_{\mathcal{L}(H)} \leq 1 - 2\overline{\mu}\alpha + 3\overline{\mu}^2\alpha^2 \quad \text{almost surely.}$$

Combining these results, we obtain

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

□

For the proof of the main theorem, we will also make use of the following algebraic inequalities.

Lemma 3.9. *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) $\sum_{j=1}^k \frac{1}{j^{1+r}} \prod_{i=j}^k \big(1 - \frac{C_1}{i} + \frac{C_2}{i^2}\big)^p \leq \exp\Big(\frac{C_2 p \pi^2}{6} + C_1 p\Big) \frac{1}{C_1 p - r}(k+1)^{-r}$.

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 $0 \leq 1 - \frac{C_1}{j} + \frac{C_2}{j^2} \leq \exp(-\frac{C_1}{j})\exp(\frac{C_2}{j^2})$. Thus, it follows that

$$\begin{aligned}\prod_{j=1}^k \Big(1 - \frac{C_1}{j} + \frac{C_2}{j^2}\Big)^p &\leq \exp\Big(-C_1 p \sum_{j=1}^k \frac{1}{j}\Big)\exp\Big(C_2 p \sum_{j=1}^k \frac{1}{j^2}\Big) \\ &\leq \exp\Big(-C_1 p \ln(k+1)\Big)\exp\Big(\frac{C_2 p \pi^2}{6}\Big),\end{aligned}$$

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

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

where the latter sum can be bounded by

$$\begin{aligned}
&\sum_{j=1}^{k} \frac{1}{j^{1+r}} \exp\Big(- C_1 p \sum_{i=j}^{k} \frac{1}{i} \Big) \\
&\leq \sum_{j=1}^{k} \frac{1}{j^{1+r}} \exp\Big(- C_1 p \ln\Big(\frac{k+1}{j} \Big) \Big) \\
&\leq \sum_{j=1}^{k} \frac{1}{j^{1+r}} \Big(\frac{k+1}{j} \Big)^{-C_1 p} \\
&\leq (k+1)^{-C_1 p} \sum_{j=1}^{k} j^{C_1 p - r - 1} \leq \exp(C_1 p) \frac{1}{C_1 p - r} (k+1)^{-r}.
\end{aligned}$$

The final inequality is where we needed $C_1 p > r$, in order to have something better than j^{-1} in the sum. □

As a final step before showing that the errors converge sub-linearly to zero, we now prove the following a priori bound which shows that they are uniformly bounded.

Lemma 3.10. *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* $\|w_1 - w^*\|$ *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, $k \in \mathbb{N}$. We consider first 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

$$w^{k+1} - w^k + \alpha_k \iota \nabla f_k(w^{k+1}) = 0 \quad \text{almost surely,}$$

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

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

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

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

almost surely. Together with the monotonicity condition, it then follows that

$$\|w^{k+1} - w^*\|^2 - \|w^k - w^*\|^2 \leq \alpha_k^2 \|\nabla f_k(w^*)\|_{H^*}^2 - 2\alpha_k \langle \nabla f_k(w^*), w^k - w^* \rangle \tag{3.4}$$

almost surely. Taking the expectation $\mathbf{E}_{\xi^k}$, we find that

$$\mathbf{E}_{\xi^k} \big[\|w^{k+1} - w^*\|^2 \big] \leq \|w^k - w^*\|^2 + \alpha_k^2 \sigma^2,$$

since $\mathbf{E}_{\xi^k}[\nabla f_k(w^*)] = 0$ and $w^k - w^*$ is independent of ξ^k. Repeating this argument, we obtain that

$$\mathbf{E}_k \big[\|w^{k+1} - w^*\|^2 \big] \leq \|w_1 - w^*\|^2 + \sigma^2 \sum_{j=1}^{k} \alpha_j^2.$$

In order to find the higher moment bound, we recall (3.4). We then follow a similar idea as in [33, Lemma 3.1], where we multiply this inequality with $\|w^{k+1}-w^*\|^2$ and use the identity $(a-b)a = \frac{1}{2}\big(|a|^2-|b|^2+|a-b|^2\big)$ for $a,b\in\mathbb{R}$. It then follows that

$$\begin{aligned}
&\|w^{k+1}-w^*\|^4 - \|w^k-w^*\|^4 + \big|\|w^{k+1}-w^*\|^2 - \|w^k-w^*\|^2\big|^2\\
&\leq \Big(\alpha_k^2\|\nabla f_k(w^*)\|_{H^*}^2 - 2\alpha_k\langle\nabla f_k(w^*), w^k-w^*\rangle\Big)\|w^{k+1}-w^*\|^2\\
&\leq \Big(\alpha_k^2\|\nabla f_k(w^*)\|_{H^*}^2 - 2\alpha_k\langle\nabla f_k(w^*), w^k-w^*\rangle\Big)\\
&\quad\times\Big(\|w^k-w^*\|^2 + \alpha_k^2\|\nabla f_k(w^*)\|_{H^*}^2 - 2\alpha_k\langle\nabla f_k(w^*), w^k-w^*\rangle\Big)\\
&\leq \alpha_k^2\|w^k-w^*\|^2\|\nabla f_k(w^*)\|_{H^*}^2 - 2\alpha_k\|w^k-w^*\|^2\langle\nabla f_k(w^*), w^k-w^*\rangle\\
&\quad+\alpha_k^4\|\nabla f_k(w^*)\|_{H^*}^4 - 4\alpha_k\|\nabla f_k(w^*)\|_{H^*}^2\alpha_k^2\langle\nabla f_k(w^*), w^k-w^*\rangle\\
&\quad+4\alpha_k^2\big(\langle\nabla f_k(w^*), w^k-w^*\rangle\big)^2
\end{aligned}$$

almost surely. Applying Young's inequality to the first and fourth term then implies that

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

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

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

We then apply the discrete Grönwall inequality for sums (see, e.g., [34]) which shows that

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

For the next higher bound, we recall that

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

almost surely, which we can multiply with $\|w^{k+1}-w^*\|^4$ in order to follow the same strategy as before. □

Remark 3.1. In particular, Lemma 3.10 implies that there exists a constant D depending on $\|w_1 - w^*\|$ and σ such that

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

for all $p \leq 2^m$ and $k \in \mathbb{N}$.

4. Proof of main theorem

We are now in a position to prove Theorem 2.1.

Proof. Given the sequence of independent random variables ξ^k, we define the random functions $f_k = f(\cdot, \xi^k)$, $k \in \mathbb{N}$. Then the scheme can be written as $w^{k+1} = T_{\alpha_k f_k} w^k$. If $T_{\alpha_k f_k} w^* = w^*$, we would essentially only have to invoke Lemma 3.7 and Lemma 3.8 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 $T_{\alpha_k f_k} w^*$ and find that

$$\begin{aligned}\|w^{k+1} - w^*\|^2 = \|T_{\alpha_k f_k} w^k - T_{\alpha_k f_k} w^*\|^2 \\ + 2(T_{\alpha_k f_k} w^k - T_{\alpha_k f_k} w^*, T_{\alpha_k f_k} w^* - w^*) + \|T_{\alpha_k f_k} w^* - w^*\|^2\end{aligned}$$

almost surely. By Lemma 3.7 and Lemma 3.8 the expectation $\mathbf{E}_{\xi^k}$ of the first term on the right-hand side is bounded by $(1 - 2\mu\alpha_k + 3\nu^2\alpha_k^2)^{1/2}\|w^k - w^*\|^2$ while by Lemma 3.5 the last term is bounded 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:

$$\begin{aligned}&(T_{\alpha_k f_k} w^k - T_{\alpha_k f_k} w^*, T_{\alpha_k f_k} w^* - w^*) \\ &= \big((T_{\alpha_k f_k} - I)w^k - (T_{\alpha_k f_k} - I)w^*, (T_{\alpha_k f_k} - I)w^*\big) \\ &\quad + \big(w^k - w^*, (T_{\alpha_k f_k} - I)w^*\big) \\ &=: I_1 + I_2 \quad \text{almost surely.}\end{aligned}$$

In order to bound I_1 and I_2, we first need to apply the a priori bound from Lemma 3.10. This will also enable us to utilize the local Lipschitz condition. First, we notice that due to Lemma 3.5, we find that

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

is bounded for $j \leq 2^m$. As $T_{\alpha_k f_k}$ is a contraction, we also obtain

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

Thus, there exists a random variable R_1 such that

$$\max\Big(\max_{j\in\{1,\dots,k\}} \|w^j\|, \max_{j\in\{1,\dots,k\}} \|T_{\alpha_j f_j} w^j\|, \|w^*\|, \max_{j\in\{1,\dots,k\}} \|T_{\alpha_j f_j} w^*\|\Big) \leq R_1,$$

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

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

where we used the fact that $T_{\alpha_k f_k}$ is non-expansive in the last step. Taking the expectation, we then have by Hölder's inequality that

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

where

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

As P is a polynomial of at most order $2^m - 2$, the exponent for P is bounded by $\big(\frac{2^{m-1}}{2^{m-1}-1}\big)\big(2^m - 2\big) = 2^m$. Hence $\tilde{L}_1$ is bounded, and in view of Lemma 3.10 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 $\|w^*\|$, $\|w_1 - w^*\|$ and σ. For I_2, we add and subtract $\alpha_k \nabla f_k w^*$ to get

$$\begin{aligned}&\big(w^k - w^*, (T_{\alpha_k f_k} - I)w^*\big) \\ &= \big(w^k - w^*, (T_{\alpha_k f_k} - I)w^* + \alpha_k \iota \nabla f_k(w^*)\big) - \big(w^k - w^*, \alpha_k \iota \nabla f_k(w^*)\big).\end{aligned}$$

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

$$\mathbf{E}_{\xi_k}[\big(w^k - w^*, \alpha_k \iota \nabla f_k(w^*)\big)] = \big(w^k - w^*, \mathbf{E}_{\xi_k}[\alpha_k \iota \nabla f_k(w^*)]\big) = 0.$$

Thus, we only have to consider the first term. Using Cauchy-Schwarz and Lemma 3.5, we find that

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

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

$$\tilde{L}_2 = \begin{cases} \big(\mathbf{E}_k\big[P(R_2)^{\frac{2^{m-1}}{2^{m-1}-1}}\big]\big)^{\frac{2^{m-1}-1}{2^{m-1}}}, & m > 1, \\ \inf_{R_2 \in \mathbb{R}}|P(R_2)|, & m = 1. \end{cases}$$

Just as for I_1, we therefore get by Lemma 3.10 that

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

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

Summarising, we now have

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

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

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

Applying Lemma 3.9 (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

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

and

$$\sum_{j=1}^{k} \alpha_j^2 \prod_{i=j+1}^{k} \tilde{C}_i \leq \frac{1}{\tilde{C}_1} \sum_{j=1}^{k} \alpha_j^2 \prod_{i=j}^{k} \tilde{C}_i \leq \frac{\eta^2}{\tilde{C}_1} \exp\Big(\frac{\nu^2\eta^2\pi^2}{4} + \mu\eta\Big) \frac{1}{\mu\eta - 1} (k+1)^{-1},$$

since $\nu \geq \mu$. Thus, we finally arrive at

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

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

Remark 4.1. 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 λ. Then it can easily be shown that

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

This means that

$$\|\big(I + \alpha\nabla F\big)^{-1}(u) - \big(I + \alpha\nabla F\big)^{-1}(v)\| \leq (1 + \alpha\lambda)^{-1}\|u - v\|,$$

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

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

Using $(1 + \alpha\lambda)^{-1} \leq 1 - \lambda\alpha + \lambda^2\alpha^2$, choosing $\alpha_k = \eta/k$ and applying Lemma 3.9 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 $i = 1, 2$ and $j = 1, \ldots, n$ 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 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 $j \leq n/2$ and $[x_j, y_j] = [x_j^2, 1]$ if $j > n/2$, similar to Equation (1.2). 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 $\xi \colon \Omega \to \{1, \dots, n\}$. 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

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

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

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

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, $M = 10000$ time steps and $N = 100 \cdot 2^i$ for $i = 1, \dots, 11$. We used $\lambda = 10^{-3}$ and the initial step size $\eta = \frac{2}{\lambda}$, 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 10 times and averaging the resulting errors. This may seem like a small number of paths but using more (or less) such paths does not seem to influence the results much, indicating that the convergence is likely actually almost surely rather than only in expectation. In order to compensate for the vectors becoming longer as 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 $\mathbf{E}_k[\|w^{k+1} - w^*\|_N^2]$. 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 Figure 2. We note that except for $N = 200$ and $N = 400$, the method does not converge at all, likely because as N grows the problem also becomes more stiff. Even when it does converge, the errors are much larger than in Figure 1. Many more steps would be necessary to reach the same accuracy as

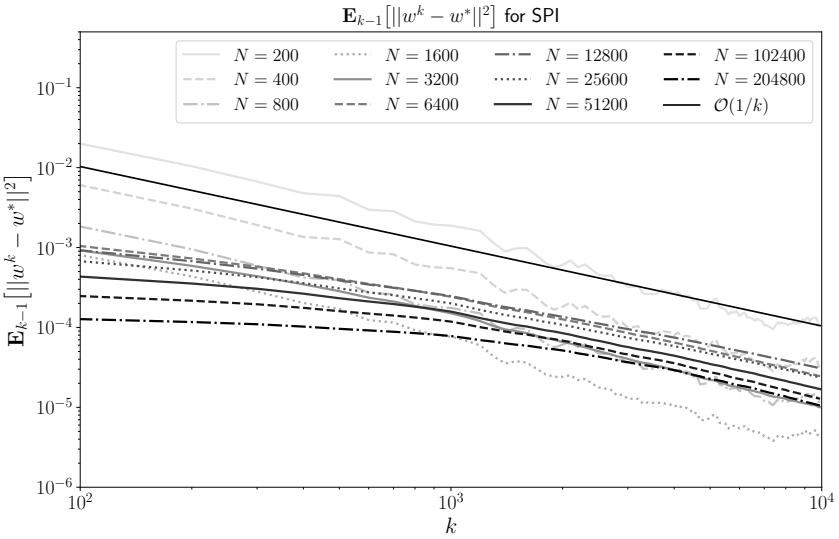

FIGURE 1. Approximated errors $\mathbf{E}_k[\|w^{k+1} - w^*\|_N^2]$ for the SPI method, measured in RMS-norm, for discretizations with varying number of grid points N. Statistics were only computed at every 100 time steps, this is why the plot starts at $k = 100$. The $1/k$-convergence is clearly seen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

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.

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 less bounded 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.

References

[1] Bottou, L., Curtis, F., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Rev. **60**(2), 223–311 (2018). URL https://doi.org/10.1137/16M1080173

[2] Rockafellar, R.: Monotone operators and the proximal point algorithm. SIAM J. Control Optimization **14**(5), 877–898 (1976). URL https://doi.org/10.1137/0314056

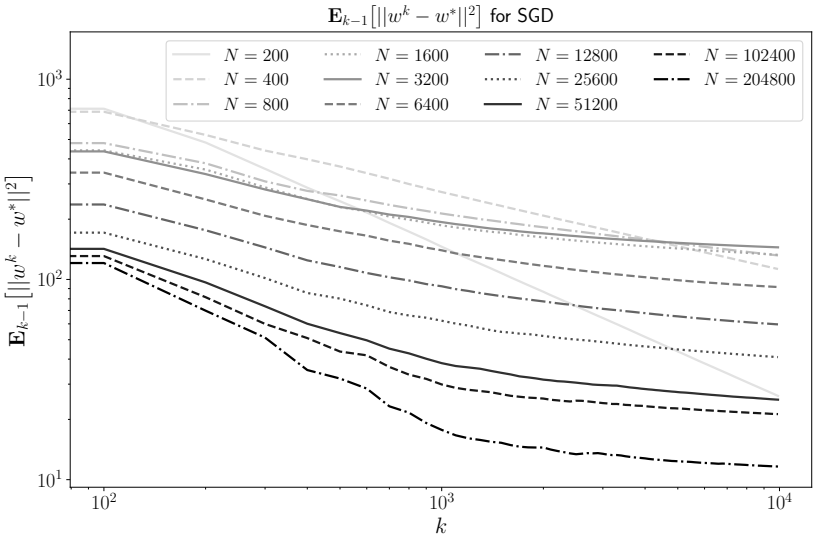

FIGURE 2. Approximated errors $\mathbf{E}_k[\|w^{k+1} - w^*\|_N^2]$ for the SGD method, measured in RMS-norm, for discretizations with varying number of grid points N. Statistics were only computed at every 100 time steps, this is why the plot starts at $k = 100$. Except for $N = 200$ and $N = 400$,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

[3] Eckstein, J., Bertsekas, D.: On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Programming **55**(3, Ser. A), 293–318 (1992). URL https://doi.org/10.1007/BF01581204

[4] Ryu, E., Boyd, S.: Stochastic proximal iteration: A non-asymptotic improvement upon stochastic gradient descent. www.math.ucla.edu/eryu/papers/spi.pdf (2016)

[5] Agarwal, A., Bartlett, P.L., Ravikumar, P., Wainwright, M.J.: Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. IEEE Trans. Inform. Theory **58**(5), 3235–3249 (2012). URL https://doi.org/10.1109/TIT.2011.2182178

[6] Raginsky, M., Rakhlin, A.: Information-based complexity, feedback and dynamics in convex programming. IEEE Trans. Inform. Theory **57**(10), 7036–7056 (2011). URL https://doi.org/10.1109/TIT.2011.2154375

[7] Bianchi, P.: Ergodic convergence of a stochastic proximal point algorithm. SIAM J. Optim. **26**(4), 2235–2260 (2016). URL https://doi.org/10.1137/15M1017909

[8] Rosasco, L., Villa, S., Vũ, B.: Convergence of stochastic proximal gradient algorithm. Appl Math Optim (2019)

[9] Patrascu, A., Necoara, I.: Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. J. Mach. Learn. Res. **18**, Paper No. 198, 42 (2017)

[10] Davis, D., Drusvyatskiy, D.: Stochastic model-based minimization of weakly convex functions. SIAM J. Optim. **29**(1), 207–239 (2019). URL https://doi.org/10.1137/18M1178244

[11] Toulis, P., Airoldi, E.M.: Scalable estimation strategies based on stochastic approximations: classical results and new insights. Stat. Comput. **25**(4), 781–795 (2015). URL https://doi.org/10.1007/s11222-015-9560-y

[12] Toulis, P., Airoldi, E.M.: Asymptotic and finite-sample properties of estimators based on stochastic gradients. Ann. Statist. **45**(4), 1694–1727 (2017). URL https://doi.org/10.1214/16-AOS1506

[13] Toulis, P., Rennie, J., Airoldi, E.M.: Statistical analysis of stochastic gradient methods for generalized linear models. Proceedings of the 31st International Conference on Machine Learning (2014)

[14] Toulis, P., Tran, D., Airoldi, E.: Towards stability and optimality in stochastic gradient descent. In: A. Gretton, C.C. Robert (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). URL http://proceedings.mlr.press/v51/toulis16.html

[15] Tran, D., Toulis, P., Airoldi, E.M.: Stochastic gradient descent methods for estimation with large data sets. ArXiv Preprint, arXiv:1509.06459 (2015)

[16] Ryu, E.K., Yin, W.: Proximal-proximal-gradient method. J. Comput. Math. **37**(6), 778–812 (2019). URL https://doi.org/10.4208/jcm.1906-m2018-0282

[17] Patrascu, A., Irofti, P.: Stochastic proximal splitting algorithm for composite minimization. ArXiv Preprint, arXiv:1912.02039v2 (2020)

[18] 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). URL https://doi.org/10.1109/tac.2019.2890888

[19] 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). URL https://doi.org/10.1007/s10957-016-0978-y

[20] Bertsekas, D.P.: Incremental proximal methods for large scale convex optimization. Math. Program. **129**(2, Ser. B), 163–195 (2011). URL https://doi.org/10.1007/s10107-011-0472-0

[21] Asi, H., Duchi, J.C.: Modeling simple structures and geometry for better stochastic optimization algorithms. Proceedings of the 22 International Conference on Artificial Intelligence and Statistics (2019)

[22] Asi, H., Duchi, J.C.: Stochastic (approximate) proximal point methods: convergence, optimality, and adaptivity. SIAM J. Optim. **29**(3), 2257–2290 (2019). URL https://doi.org/10.1137/18M1230323

[23] Toulis, P., Horel, T., Airoldi, E.M.: The proximal robbins–monro method. ArXiv Preprint, arXiv:1510.00967v4 (2020)

[24] Fagan, F., Iyengar, G.: Unbiased scalable softmax optimization. ArXiv Preprint, arXiv:1803.08577 (2018)

[25] Rockafellar, R.T.: On the maximal monotonicity of subdifferential mappings. Pacific J. Math. **33**, 209–216 (1970). URL http://projecteuclid.org/euclid.pjm/1102977253

[26] Papageorgiou, N., Winkert, P.: Applied Nonlinear Functional Analysis. An Introduction. De Gruyter, Berlin (2018)

[27] Quarteroni, A., Valli, A.: Domain decomposition methods for partial differential equations. Numerical Mathematics and Scientific Computation. The Clarendon Press, Oxford University Press, New York (1999). Oxford Science Publications

[28] Barbu, V.: Nonlinear Differential Equations of Monotone Types in Banach Spaces. Springer-Verlag, New York (2010). URL https://doi.org/10.1007/978-1-4419-5542-5

[29] Eisenmann, M.: Methods for the temporal approximation of nonlinear, nonautonomous evolution equations. PhD thesis, TU Berlin (2019)

[30] 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). URL https://doi.org/10.1007/s10208-018-09412-w

[31] Hager, W.W.: Updating the inverse of a matrix. SIAM Rev. **31**(2), 221–239 (1989). URL https://doi.org/10.1137/1031049

[32] Lasiecka, I., Triggiani, R.: Control theory for partial differential equations: continuous and approximation theories. I, *Encyclopedia of Mathematics and its Applications*, vol. 74. Cambridge University Press, Cambridge (2000). Abstract parabolic systems

[33] 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). URL https://doi.org/10.1093/imanum/drs032

[34] Clark, D.: Short proof of a discrete Gronwall inequality. Discrete Appl. Math. **16**(3), 279–281 (1987). URL http://dx.doi.org/10.1016/0166-218X(87)90064-3