SUB-LINEAR CONVERGENCE OF A TAMED STOCHASTIC GRADIENT DESCENT METHOD IN HILBERT SPACE

MONIKA EISENMANN AND TONY STILLFJORD

Abstract. In this paper, we introduce the tamed stochastic gradient descent method (TSGD) for optimization problems. Inspired by the tamed Euler scheme, which is a commonly used method within the context of stochastic differential equations, TSGD is an explicit scheme that exhibits stability properties similar to those of implicit schemes. As its computational cost is essentially equivalent to that of the well-known stochastic gradient descent method (SGD), it constitutes a very competitive alternative to such methods.

We rigorously prove (optimal) sub-linear convergence of the scheme for strongly convex objective functions on an abstract Hilbert space. The analysis only requires very mild step size restrictions, which illustrates the good stability properties. The analysis is based on a priori estimates more frequently encountered in a time integration context than in optimization, and this alternative approach provides a different perspective also on the convergence of SGD. Finally, we demonstrate the usability of the scheme on a problem arising in a context of supervised learning.

1. Introduction

We consider the gradient flow

$$w' = -\nabla F(w), \quad w(0) = w_1,$$

on the interval $t \in [0, \infty)$ in order to approximate its steady state w^* which satisfies $\nabla F(w^*) = 0$. We are interested in this problem because for a suitable F its solution solves the minimization problem

$$w^* = \arg\min_w F(w).$$

Standard optimization methods may thereby be formulated as time-stepping methods for an evolution equation, which provides an alternative viewpoint on their behaviour and on how to analyze them.

We are mainly interested in the case where $F = \frac{1}{N}\sum_{i=1}^N f_i$ is a sum of many functions f_i of the same type. This setting occurs in, e.g., supervised learning applications, where each f_i corresponds to either a single data point or to a small

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.

2010 *Mathematics Subject Classification.* 46N10; 65K10; 90C15.

Key words and phrases. stochastic optimization; tamed Euler scheme; convergence analysis; convergence rate; 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 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.

subset (batch) of the data. In order to cover also the infinite data case, we assume more generally that

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

where ξ is a random variable and $\mathbf{E}_\xi$ denotes the corresponding expectation. Then a realization of ξ corresponds to a specific batch. In supervised learning applications, the amount of data is frequently very large, and computing the full gradient ∇F is not feasible. Instead, one typically applies *stochastic* methods where instead of ∇F the gradient $\nabla f(\xi, \cdot)$ is used, see [6] for a general overview.

A popular method is stochastic gradient descent (SGD), given by

$$w^{n+1} = w^n - \alpha_n \nabla f(\xi_n, w^n), \quad w^1 = w_1,$$

where $\{\xi_n\}_{n\in\mathbb{N}}$ denotes a sequence of jointly independent random variables and $\{\alpha_n\}_{n\in\mathbb{N}}$ is a sequence of step sizes (learning rates). In essence, we apply the standard gradient descent method but in each step only utilize a randomly chosen (small) part of ∇F. More advanced methods such as Adam [20] exist as well, but most are still based on the underlying SGD idea.

Viewed as a time-stepping method, SGD is equivalent to an inexact version of the explicit (forward) Euler method and thereby suffers from the same stability issues. In particular, if the problem is at all stiff, then there is a severe limit on the step sizes α_n, $n \in \mathbb{N}$, where the iterates quickly explode in size if it is violated. It has been observed that neural networks do indeed tend to give rise to such stiff gradient flows, see e.g. [23] for an early concrete example. On the other hand, for optimal performance, we want to choose the step sizes as large as possible, and thus as close to this limit as possible. Since the limit depends on properties of F that are not always known, like its Lipschitz constant, this is difficult.

Ideally, one would like to instead use an implicit scheme which is unconditionally stable. This would remove the step size restrictions altogether. In certain cases, such a method can be implemented very efficiently and is then the best choice. See, e.g. [4, 8, 11, 26, 28, 34, 35, 36] for analyses of this setting. In general, however, it means that we have to solve an unfeasibly large system of nonlinear equations in each step.

The situation is similar for certain stochastic differential equations (SDEs), where it can be shown that the explicit (forward) Euler-Maruyama method does neither converge in strong mean-square sense nor in the numerically weak sense to the exact solution at a finite time point, compare [17] and also [19] for a generalized result. At the same time the implicit (backward) Euler-Maruyama scheme might be too expensive. In this context, the *tamed Euler* scheme provides a fully explicit alternative, with better stability properties. This scheme was introduced for SDEs in [18] and has been studied further in, e.g., [15, 16, 30, 31]. Very recently, the taming idea has also been extended to a setting similar to ours involving stochastic gradient Langevin dynamics [22], which generalizes the deterministic setting from [7, 32].

We propose to use a method of this type also in the current context, which we call the tamed stochastic gradient descent (TSGD). It is defined by

$$w^{n+1} = w^n - \frac{\alpha_n \nabla f(\xi_n, w^n)}{1 + \alpha_n \|\nabla f(\xi_n, w^n)\|}, \quad w^1 = w_1.$$

We note that it is a fully explicit scheme. Further, as the step sizes or the gradients tend to zero, the method tends to the SGD. In fact, it is straightforward to show that TSGD is a second-order perturbation of SGD. However, due to the specific rescaling of the gradient, its stability properties are much better and large step sizes do not cause issues.

The main contribution of this paper is a rigorous error analysis of TSGD in a strongly convex setting, which demonstrates that it converges as $\mathcal{O}(\frac{1}{n})$. This is the optimal rate which can be expected in this stochastic setting. Notably, we require very weak or no bounds on the initial step size, and its size only affects the error constants in a mild manner. Another feature of our analysis is that we consider the problem in a (possibly) infinite-dimensional Hilbert space, which means that the error bounds are applicable not only to optimization of $\mathbb{R}^d$-valued data, but to, e.g. classification of functions. We also directly prove convergence of $\{w^n\}_{n\in\mathbb{N}}$ towards w^* rather than of $\{F(w^n)\}_{n\in\mathbb{N}}$ towards $F(w^*)$. While these types of convergence are equivalent in the current setting, our approach provides better error constants for the first type of convergence than using this equivalence together with more standard arguments.

We refer to [6] for a general overview of optimization methods for our problem setting. This includes a general proof of convergence for first-order explicit methods in which many similar methods fit. We note that verifying the required assumptions for the method suggested here is non-trivial. Furthermore, applying such a general result would not highlight the benefit of the scheme. We also note that our analysis is based on a different idea which relies on a priori estimates. The same ideas can be applied also to, e.g., SGD, which similarly shows convergence without a strict step size restriction. The limitation instead shows up in the error constant, which becomes infeasibly large. For the proposed method, the error constant is instead of a moderate size. Our analysis thus provides a different viewpoint on the convergence of these kinds of methods, which does not rely on prescribed step size limitations.

There are other related methods which might be useful in the given context, such as implicit-explicit schemes [3, 5, 25, 29, 33], where only part of the problem is considered in an implicit way, and sum-splitting methods [29, 37, 38] where the problem is decomposed into many small subproblems and each is considered in an implicit way. Both of these approaches rely on there being such easily identifiable splittings, which is typically not the case in the general setting. More closely related to our proposed method are the stabilized Runge-Kutta schemes proposed in [1, 39] for parabolic problems rather than optimization. See also e.g. [12, 40] and [14, Section V] for an overview. Recently, they were adapted to solve a special class of deterministic optimization problems in [10].

While our proofs of convergence require rather strong assumptions, such as strong convexity, we hasten to add that the method performs well also in more general settings, such as that of general neural networks. This is demonstrated by our numerical experiments in Section 6. It is therefore likely that our assumptions can be much weakened while still guaranteeing, e.g., local convergence to a local minimum. Such considerations would, however, add a considerable amount of technical details that would obscure the general idea, and we thus choose to limit ourselves to this setting.

The paper is organized as follows. In Section 2 we fix the notation and state the basic assumptions on the optimization problem. Then we formally introduce the method in Section 3. As stated above, our main proof relies on a priori estimates, and we prove these in Section 4. These are then used in the main error analysis in Section 5. In Section 6, we provide several numerical experiments that illustrate our claims, both in a setting satisfying our basic assumptions and in a more general setting. Section 7 summarises our conclusions. Finally, we collect some generally applicable results that are critical for our analysis, but whose proofs are overly technical and do not contribute to an understanding of the main ideas in Appendix A.

2. Preliminaries

In the following, we denote by $\mathbb{N}$ the natural numbers, not including 0. Let $(H, \langle\cdot,\cdot\rangle, \|\cdot\|)$ be a real Hilbert space (e.g. $H = \mathbb{R}^d$ for $d \in \mathbb{N}$). Its dual space is denoted by $(H^*, \langle\cdot,\cdot\rangle_{H^*}, \|\cdot\|_{H^*})$. 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 $\iota^{-1} : v \mapsto \langle v, \cdot\rangle$.

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a complete probability space and let $\{\xi_n\}_{n\in\mathbb{N}}$ be a family of jointly independent random variables on Ω. For a random variable $X\colon \Omega \to H$, let $\mathbf{E}_\xi[X]$ denote the expectation with respect to the probability distribution of ξ. We are mainly interested in the total expectation

$$\mathbf{E}_n\big[\|X\|^2\big] = \mathbf{E}_{\xi_1}\big[\mathbf{E}_{\xi_2}\big[\cdots \mathbf{E}_{\xi_n}\big[\|X\|^2\big]\cdots\big]\big].$$

Since the random variables $\{\xi_n\}_{n\in\mathbb{N}}$ are jointly independent, this expectation coincides with the expectation with respect to the joint probability distribution of $\xi_1, \dots, \xi_n$. We also note here that if one of the following statements does not involve an expectation but does contain a random variable, then it is assumed to hold almost surely (a.s.) even if this is not explicitly stated.

For a measurable space $(E, \mathcal{E})$, let $\xi\colon \Omega \to E$ be a random variable and $f\colon E \times H \to \mathbb{R}$ be a function. We then consider the composition function $f(\xi,\cdot)\colon \Omega \times H \to \mathbb{R}$ which we assume fulfils

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

and aim to find

$$w^* = \arg\min_w F(w).$$

The existence of such a minimum will be guaranteed by a strong convexity assumption below. We note that this means that $\nabla F(w^*) = 0$.

In the following theory, we only consider the composition function $f(\xi,\cdot)\colon \Omega \times H \to \mathbb{R}$ in detail instead of $f\colon E \times H \to \mathbb{R}$ and therefore do not need to state the measurable space $(E, \mathcal{E})$ explicitly. For a fixed ω, the element $\xi(\omega) \in E$ can, for example, represent the batch chosen to approximate F as explained in the introduction.

Below, we collect all the assumptions that will be used throughout the paper. Each lemma and theorem specifies which particular assumptions are in effect at that point. The first assumption concerns the properties of the functions $f(\xi,\cdot)$, which will be used as stochastic approximations to F.

Assumption 1. *Let $f(\xi,\cdot)\colon \Omega \times H \to \mathbb{R}$ be given such that*

- $\langle \iota\nabla f(\xi, v), w\rangle = \lim_{h\to 0} \frac{f(\xi, v+hw) - f(\xi,v)}{h}$ *a.s. for all $v, w \in H$, i.e. $f(\xi,\cdot)$ is Gâteaux differentiable a.s.;*

- *there exists a random variable* $\mu_\xi \colon \Omega \to [0,\infty)$ *with* $\mathbf{E}_\xi[\mu_\xi] =: \mu \in (0,\infty)$ *such that*

$$\langle \iota\nabla f(\xi,v) - \iota\nabla f(\xi,w), v-w\rangle \geq \mu_\xi \|v-w\|^2 \quad \text{a.s. for all } v,w \in H;$$

- *there exists a random variable* $L_\xi \colon \Omega \to [0,\infty)$ *with* $\left(\mathbf{E}_\xi\left[L_\xi^2\right]\right)^{\frac{1}{2}} =: L \in (0,\infty)$ *such that*

$$\|\iota\nabla f(\xi,v) - \iota\nabla f(\xi,w)\| \leq L_\xi \|v-w\| \quad \text{a.s. for all } v,w \in H;$$

- *for* $w^* \in H$ *with* $\nabla F(w^*) = 0$*, there exists a finite value* $\sigma \in [0,\infty)$ *such that* $\left(\mathbf{E}_\xi\left[\|\iota\nabla f(\xi,w^*)\|^2\right]\right)^{\frac{1}{2}} = \sigma$.

The above assumption is enough to prove convergence with a sub-optimal rate and the optimal rate in some cases. To guarantee the optimal rate in all cases, we additionlly make the following assumption on certain higher moments.

Assumption 2. *Let* f *be given such that Assumption 1 is fulfilled. Further, assume that for all* $v,w \in H$

- $\langle \iota\nabla f(\xi,v) - \iota\nabla f(\xi,w), v-w\rangle \geq \mu_\xi \|v-w\|^2$ *with* $\left(\mathbf{E}_\xi\left[\|\mu_\xi^2\|\right]\right)^{\frac{1}{2}} =: \mu_2 \in (0,\infty)$.
- $\|\iota\nabla f(\xi,v) - \iota\nabla f(\xi,w)\| \leq L_\xi \|v-w\|$ *with* $\left(\mathbf{E}_\xi\left[L_\xi^4\right]\right)^{\frac{1}{4}} =: L_4 \in (0,\infty)$;
- *for* $w^* \in H$ *with* $\nabla F(w^*) = 0$*, there exists a finite value* $\sigma_4 \in [0,\infty)$ *such that* $\left(\mathbf{E}_\xi\left[\|\iota\nabla f(\xi,w^*)\|^4\right]\right)^{\frac{1}{4}} =: \sigma_4$.

Finally, in the case that the gradient is also globally bounded, the convergence result can be further improved. For technical reasons we also need to ensure that at points away from the minimum, the stochastic gradients are not significantly smaller than they are at the minimum of F. This is the content of the next assumption.

Assumption 3. *Let* f *be given such that Assumption 1 is fulfilled, and such that there exists* $B \in (0,\infty)$ *with* $\|\iota\nabla f(\xi,w)\| \leq B$ *a.s. for all* $w \in H$*. Further, for* $w^* \in H$ *such that* $\nabla F(w^*) = 0$ *there exists* $D \in [0,\infty)$ *such that*

$$\left(\mathbf{E}_\xi\left[\chi_{\|\iota\nabla f(\xi,w)\|>0}\frac{\|\iota\nabla f(\xi,w^*)\|^2}{\|\iota\nabla f(\xi,w)\|^2}\right]\right)^{\frac{1}{2}} \leq D$$

is fulfilled for all $w \in H$.

As shown in the auxiliary Lemma A.3, Assumption 1 means that F is also Gâteaux differentiable and $\nabla F = \mathbf{E}_\xi\left[\nabla f(\xi,\cdot)\right]$. The following lemma summarises a few further consequences of the above assumptions.

Lemma 2.1. *Let Assumption 1 be fulfilled. Then* F *is strongly convex with convexity constant* μ *and* ∇F *is Lipschitz continuous with Lipschitz constant* $L_F \leq L$*, i.e. for all* $v,w \in H$ *it holds that*

$$\|\nabla F(v) - \nabla F(w)\|_{H^*} = \|\iota\nabla F(v) - \iota\nabla F(w)\| \leq L_F\|v-w\| \leq L\|v-w\| \quad \text{and}$$
$$F(v) \geq F(w) + \langle \iota\nabla F(w), v-w\rangle + \frac{\mu}{2}\|v-w\|^2.$$

Further, the first inequality implies that

$$F(v) \leq F(w) + \langle \iota\nabla F(w), v-w\rangle + \frac{L}{2}\|v-w\|^2.$$

Finally, there exists a unique $w^* \in H$ *such that* $F(w^*) = \min_{w\in H} F(w)$.

Proof. Using $\nabla F(w) = \mathbf{E}_\xi\big[\nabla f(\xi, w)\big]$, we obtain

$$\langle \iota\nabla F(v) - \iota\nabla F(w), v - w\rangle \geq \mu\|v - w\|^2 \quad \text{for all } v, w \in H.$$

Thus, the function $v \mapsto \nabla F(v) - \mu\iota^{-1}v$ is monotone. Applying [42, Proposition 25.10], it follows that $v \mapsto F(v) - \frac{\mu}{2}\|v\|^2$ is convex such that F is strongly convex and the variational inequality stated in the lemma is fulfilled. The Lipschitz continuity of $\iota\nabla F$ similarly follows from the Lipschitz continuity of $\iota\nabla f(\xi, \cdot)$ by the identification provided in Lemma A.3. The final inequality follows by expanding F in a zeroth-order Taylor expansion around w and using the Lipschitz continuity. See e.g. [6, Appendix B] for more details. Since F is strongly convex, it is coercive. Combined with the Gâteaux differentiability, this guarantees the existence of a unique global minimum, see e.g. [42, Theorem 25.D, Proposition 25.20 and Corollary 25.15]. □

3. The stochastic tamed Euler scheme

Throughout the paper, we will assume that $\{\xi_n\}_{n\in\mathbb{N}}$ is a given a family of jointly independent random variables and we will abbreviate $f_n(w) = f(\xi_n, w)$ for $n \in \mathbb{N}$. The ξ_n typically correspond to what batches we choose in each iteration, i.e. on which part of the data we evaluate the gradient. Let $\{\alpha_n\}_{n\in\mathbb{N}}$ be a sequence of of positive real numbers. We then consider the stochastic tamed Euler scheme

$$w^{n+1} = w^n - \frac{\alpha_n \iota\nabla f_n(w^n)}{1 + \alpha_n\|\iota\nabla f_n(w^n)\|} \quad \text{for } n \in \mathbb{N}, \quad w^1 = w_1. \tag{3.1}$$

Note that it is also possible to choose a random initial value w^1, and our convergence statements can be extended to this setting in a straightforward way. For simplicity, we restrict ourselves to a fixed initial value $w_1 \in H$ in the following.

We note that the computational effort of the scheme is essentially the same as that of SGD, since once $\nabla f_n(w^n)$ has been found it is cheap to compute its norm. We also note that TSGD can be interpreted as a second order perturbation of SGD, since

$$\frac{\alpha_n \iota\nabla f_n(w^n)}{1 + \alpha_n\|\iota\nabla f_n(w^n)\|} = \alpha_n \iota\nabla f_n(w^n) - \frac{\alpha_n^2\|\nabla f_n(w^n)\|\iota\nabla f_n(w^n)}{1 + \alpha_n\|\iota\nabla f_n(w^n)\|}.$$

This second order perturbation mainly offers advantages if $\alpha_n\|\iota\nabla f_n(w^n)\|$ is large. In this case we make use of the fact that

$$\frac{1}{2}\min\big\{1, \alpha_n\|\iota\nabla f_n(w^n)\|\big\} \leq \frac{\alpha_n\|\iota\nabla f_n(w^n)\|}{1 + \alpha_n\|\iota\nabla f_n(w^n)\|} \leq \min\big\{1, \alpha_n\|\iota\nabla f_n(w^n)\|\big\}.$$

Thus, the growth of $w^{n+1} - w^n = \frac{-\alpha_n \iota\nabla f_n(w^n)}{1+\alpha_n\|\iota\nabla f_n(w^n)\|}$ is always bounded.

4. A priori bounds

Our main results will show that $\mathbf{E}_n\big[\|w^{n+1} - w^*\|^2\big]$ tends to zero as $\frac{1}{n}$ in the strongly convex case. In this section, we prepare for the proofs of this by first showing that the errors are bounded. We note that the main argument here only requires convexity rather than strong convexity, and the w^* in the following three lemmas could therefore equally well be any $w^* \in H$ that satifies $\nabla F(w^*) = 0$.

Lemma 4.1. *Let Assumption 1 be fulfilled and let $\{\alpha_n\}_{n\in\mathbb{N}}$ be a sequence of positive real numbers such that $\sum_{n=1}^{\infty}\alpha_n^2<\infty$. For $\Phi\in[0,\infty)$ the a priori bound*

$$\begin{aligned}
&\mathbf{E}_n\big[\|w^{n+1}-w^*\|^2\big]\\
&\le \|w_1-w^*\|^2\exp\Big(\sum_{i=1}^{\infty}\big(2\sigma^2\min\big\{\Phi^{-2},\alpha_i^2\big\}+4\alpha_i^2L^2m_i\big)\Big)\\
&\quad+\sum_{i=1}^{\infty}\big(\Phi^2\min\big\{\mathbf{E}_i\big[\|\iota\nabla f_i(w^i)\|^{-2}\big],\alpha_i^2\big\}+2(1-m_i)+2\min\big\{1,2\alpha_i^2\sigma^2\big\}\big)\\
&\qquad\times\exp\Big(\sum_{j=i+1}^{\infty}\big(2\sigma^2\min\big\{\Phi^{-2},\alpha_j^2\big\}+4\alpha_j^2L^2m_j\big)\Big)=:M_2
\end{aligned}$$

is fulfilled, where $\min\big\{\Phi^{-2},x\big\}=x$ for $\Phi=0$ and every $x\in\mathbb{R}$ and

$$m_i=\begin{cases}1, & \text{if } 2\alpha_i^2L^2\mathbf{E}_{i-1}\big[\|w^i-w^*\|^2\big]\le 1,\\ 0, & \text{otherwise.}\end{cases}$$

Furthermore, there exists $n_0\in\mathbb{N}$ such that $m_i=1$ for all $i\ge n_0$.

Remark 4.2. The advantage of this particular a priori bound is that the bound does not grow very much when the initial step size is increased. The corresponding proof for the SGD method looks very similar, but does not have the factors m_n or $\min\{\dots,\alpha_n^2\}$, $n\in\mathbb{N}$. This means that the first few terms in the products become very large, even for moderately sized Lipschitz constants, reflecting the fact that a too large step size can lead to instability. In our case, these large terms are multiplied by 0 or cut off by the min-function. The constant Φ can be used to tune the error bound further in case σ or $\mathbf{E}_n\big[\|\iota\nabla f_n(w^n)\|^{-2}\big]$, $n\in\mathbb{N}$, is large.

Proof of Lemma 4.1. We test the scheme defined by (3.1) with w^n-w^*, in order to obtain that

$$\begin{aligned}
&\langle w^{n+1}-w^*-(w^n-w^*),w^n-w^*\rangle\\
&\quad+\frac{\alpha_n\langle\iota\nabla f_n(w^n)-\iota\nabla f_n(w^*),w^n-w^*\rangle}{1+\alpha_n\|\iota\nabla f_n(w^n)\|}=-\frac{\alpha_n\langle\iota\nabla f_n(w^*),w^n-w^*\rangle}{1+\alpha_n\|\iota\nabla f_n(w^n)\|}.
\end{aligned}\tag{4.1}$$

Using the identity $\langle u-v,u\rangle=\frac{1}{2}(\|u\|^2-\|v\|^2+\|u-v\|^2)$, $u,v\in H$, the first summand on the left-hand side can be written as

$$\begin{aligned}
&-\langle w^n-w^*-(w^{n+1}-w^*),w^n-w^*\rangle\\
&=-\frac{1}{2}\big(\|w^n-w^*\|^2-\|w^{n+1}-w^*\|^2+\|w^{n+1}-w^n\|^2\big)\\
&=\frac{1}{2}\Big(\|w^{n+1}-w^*\|^2-\|w^n-w^*\|^2-\frac{\alpha_n^2\|\iota\nabla f_n(w^n)\|^2}{(1+\alpha_n\|\iota\nabla f_n(w^n)\|)^2}\Big),
\end{aligned}$$

where we inserted the scheme in the last step. Thus, inserting the monotonicity condition for f_n into (4.1) and multiplying the inequality with the factor two, it follows that

$$\begin{aligned}
&\|w^{n+1}-w^*\|^2-\|w^n-w^*\|^2\\
&\le-\frac{2\alpha_n\langle\iota\nabla f_n(w^*),w^n-w^*\rangle}{1+\alpha_n\|\iota\nabla f_n(w^n)\|}+\frac{\alpha_n^2\|\iota\nabla f_n(w^n)\|^2}{(1+\alpha_n\|\iota\nabla f_n(w^n)\|)^2}=:I_1+I_2.
\end{aligned}$$

Since the tamed Euler scheme is the forward Euler scheme with a second order perturbation, it follows that

$$\frac{\alpha_n \iota \nabla f_n(w^*)}{1+\alpha_n\|\iota\nabla f_n(w^n)\|} = \frac{\alpha_n \iota \nabla f_n(w^*)}{1+\alpha_n\Phi} + \frac{\alpha_n^2\big(\Phi - \|\iota\nabla f_n(w^n)\|\big)\iota\nabla f_n(w^*)}{(1+\alpha_n\|\iota\nabla f_n(w^n)\|)(1+\alpha_n\Phi)}$$

for $\Phi \in [0,\infty)$. Note that we have w^* in the numerator of the left-hand-side but w^n in the denominator. We insert this equality into I_1 and use the Cauchy–Schwarz inequality and Young's inequality for products in order to obtain

$$\begin{aligned}
I_1 &= -\frac{\alpha_n\langle 2\iota\nabla f_n(w^*), w^n - w^*\rangle}{1+\alpha_n\|\iota\nabla f_n(w^n)\|} \\
&\leq -\frac{2\alpha_n\langle \iota\nabla f_n(w^*), w^n - w^*\rangle}{1+\alpha_n\Phi} + \frac{2\alpha_n^2\Phi\|\iota\nabla f_n(w^*)\|\|w^n-w^*\|}{(1+\alpha_n\|\iota\nabla f_n(w^n)\|)(1+\alpha_n\Phi)} \\
&\quad + \frac{2\alpha_n^2\|\iota\nabla f_n(w^n)\|\|\iota\nabla f_n(w^*)\|\|w^n-w^*\|}{(1+\alpha_n\|\iota\nabla f_n(w^n)\|)(1+\alpha_n\Phi)} \\
&\leq -\frac{2\alpha_n\langle \iota\nabla f_n(w^*), w^n - w^*\rangle}{1+\alpha_n\Phi} + \frac{\alpha_n^2\big(\Phi^2 + \|\iota\nabla f_n(w^n)\|^2\big)}{(1+\alpha_n\|\iota\nabla f_n(w^n)\|)^2} \\
&\quad + \frac{2\alpha_n^2\|\iota\nabla f_n(w^*)\|^2\|w^n-w^*\|^2}{(1+\alpha_n\Phi)^2} =: I_{1,1} + I_{1,2} + I_{1,3}.
\end{aligned}$$

For $I_{1,1} = -\frac{2\alpha_n\langle \iota\nabla f_n(w^*), w^n-w^*\rangle}{1+\alpha_n\Phi}$, we notice that $\mathbf{E}_{\xi_n}[I_{1,1}] = 0$. Moreover, for $I_{1,2} = \frac{\alpha_n^2(\Phi^2+\|\iota\nabla f_n(w^n)\|^2)}{(1+\alpha_n\|\iota\nabla f_n(w^n)\|)^2}$, we get

$$\begin{aligned}
I_{1,2} &\leq \Phi^2 \min\big\{\|\iota\nabla f_n(w^n)\|^{-2}, \alpha_n^2\big\} + \min\big\{1, \alpha_n^2\|\iota\nabla f_n(w^n)\|^2\big\} \\
&\leq \Phi^2 \min\big\{\|\iota\nabla f_n(w^n)\|^{-2}, \alpha_n^2\big\} + \min\big\{1, 2\alpha_n^2 L_{\xi_n}^2\|w^n-w^*\|^2\big\} \\
&\quad + \min\big\{1, 2\alpha_n^2\|\iota\nabla f_n(w^*)\|^2\big\}
\end{aligned}$$

and

$$I_{1,3} = \frac{2\alpha_n^2\|\iota\nabla f_n(w^*)\|^2\|w^n-w^*\|^2}{(1+\alpha_n\Phi)^2} \leq 2\|\iota\nabla f_n(w^*)\|^2 \min\big\{\Phi^{-2}, \alpha_n^2\big\}\|w^n-w^*\|^2.$$

A bound for I_2 is given by

$$\begin{aligned}
I_2 &= \frac{\alpha_n^2\|\iota\nabla f_n(w^n)\|^2}{(1+\alpha_n\|\iota\nabla f_n(w^n)\|)^2} \leq \min\big\{1, \alpha_n^2\|\iota\nabla f_n(w^n)\|^2\big\} \\
&\leq \min\big\{1, 2\alpha_n^2 L_{\xi_n}^2\|w^n-w^*\|^2\big\} + \min\big\{1, 2\alpha_n^2\|\iota\nabla f_n(w^*)\|^2\big\}.
\end{aligned}$$

Then it follows

$$\begin{aligned}
&\|w^{n+1}-w^*\|^2 - \|w^n-w^*\|^2 \leq I_1 + I_2 \\
&\leq -\frac{2\alpha_n\langle \iota\nabla f_n(w^*), w^n - w^*\rangle}{1+\alpha_n\Phi} + \Phi^2 \min\big\{\|\iota\nabla f_n(w^n)\|^{-2}, \alpha_n^2\big\} \\
&\quad + 2\min\big\{1, 2\alpha_n^2 L_{\xi_n}^2\|w^n-w^*\|^2\big\} + 2\min\big\{1, 2\alpha_n^2\|\iota\nabla f_n(w^*)\|^2\big\} \\
&\quad + 2\|\iota\nabla f_n(w^*)\|^2 \min\big\{\Phi^{-2}, \alpha_n^2\big\}\|w^n-w^*\|^2.
\end{aligned} \tag{4.2}$$

Taking the $\mathbf{E}_n$-expectation, we then obtain

$$\begin{aligned}
&\mathbf{E}_n\big[\|w^{n+1}-w^*\|^2\big]\\
&\quad\le \big(1+2\sigma^2\min\big\{\Phi^{-2},\alpha_n^2\big\}\big)\mathbf{E}_{n-1}\big[\|w^n-w^*\|^2\big]\\
&\qquad+\Phi^2\min\big\{\mathbf{E}_n\big[\|\iota\nabla f_n(w^n)\|^{-2}\big],\alpha_n^2\big\}\\
&\qquad+2\min\big\{1,2\alpha_n^2L^2\mathbf{E}_{n-1}\big[\|w^n-w^*\|^2\big]\big\}+2\min\big\{1,2\alpha_n^2\sigma^2\big\}\\
&\quad=\big(1+2\min\big\{\Phi^{-2},\alpha_n^2\big\}\sigma^2+4\alpha_n^2L^2m_n\big)\mathbf{E}_{n-1}\big[\|w^n-w^*\|^2\big]\\
&\qquad+\Phi^2\min\big\{\mathbf{E}_n\big[\|\iota\nabla f_n(w^n)\|^{-2}\big],\alpha_n^2\big\}+2(1-m_n)+2\min\big\{1,2\alpha_n^2\sigma^2\big\},
\end{aligned}\tag{4.3}$$

with m_n defined as in the lemma statement. Reinserting the bound repeatedly thus yields

$$\begin{aligned}
&\mathbf{E}_n\big[\|w^{n+1}-w^*\|^2\big]\\
&\le\|w_1-w^*\|^2\prod_{i=1}^n\big(1+2\sigma^2\min\big\{\Phi^{-2},\alpha_i^2\big\}+4\alpha_i^2L^2m_i\big)\\
&\quad+\sum_{i=1}^n\big(\Phi^2\min\big\{\mathbf{E}_i\big[\|\iota\nabla f_i(w^i)\|^{-2}\big],\alpha_i^2\big\}+2(1-m_i)+2\min\big\{1,2\alpha_i^2\sigma^2\big\}\big)\\
&\qquad\times\prod_{j=i+1}^n\big(1+2\sigma^2\min\big\{\Phi^{-2},\alpha_j^2\big\}+4\alpha_j^2L^2m_j\big).
\end{aligned}$$

Finally, we apply the inequality $1+x\le\exp(x)$, $x\in\mathbb{R}$, and make the bound independent of n by bounding the final sums by the corresponding infinite sums, in order to obtain

$$\begin{aligned}
&\mathbf{E}_n\big[\|w^{n+1}-w^*\|^2\big]\\
&\le\|w_1-w^*\|^2\exp\Big(\sum_{i=1}^\infty\big(2\sigma^2\min\big\{\Phi^{-2},\alpha_i^2\big\}+4\alpha_i^2L^2m_i\big)\Big)\\
&\quad+\sum_{i=1}^\infty\big(\Phi^2\min\big\{\mathbf{E}_i\big[\|\iota\nabla f_i(w^i)\|^{-2}\big],\alpha_i^2\big\}+2(1-m_i)+2\min\big\{1,2\alpha_i^2\sigma^2\big\}\big)\\
&\qquad\times\exp\Big(\sum_{j=i+1}^\infty\big(2\sigma^2\min\big\{\Phi^{-2},\alpha_j^2\big\}+4\alpha_j^2L^2m_j\big)\Big).
\end{aligned}$$

It remains to verify, that there exists $n_0\in\mathbb{N}$ such that $m_n=1$ for all $n\ge n_0$. This can be done by estimating (4.3) and following a similar line of argumentation as before. First, we can write

$$\mathbf{E}_n\big[\|w^{n+1}-w^*\|^2\big]\le\big(1+2\alpha_n^2\sigma^2+4\alpha_n^2L^2\big)\mathbf{E}_{n-1}\big[\|w^n-w^*\|^2\big]+\Phi^2\alpha_n^2+4\alpha_n^2\sigma^2.$$

Reinserting the inequality $n-1$ times, it follows that

$$\begin{aligned}
&\mathbf{E}_n\big[\|w^{n+1}-w^*\|^2\big]\\
&\le\|w_1-w^*\|^2\prod_{i=1}^n\big(1+2\sigma^2\alpha_i^2+4\alpha_i^2L^2\big)+\sum_{i=1}^n4\alpha_i^2\sigma^2\prod_{j=i+1}^n\big(1+2\sigma^2\alpha_j^2+4\alpha_j^2L^2\big)\\
&\le\|w_1-w^*\|^2\exp\Big(\big(2\sigma^2+4L^2\big)\sum_{i=1}^\infty\alpha_i^2\Big)+\sum_{i=1}^\infty4\alpha_i^2\sigma^2\exp\Big(\big(2\sigma^2+4L^2\big)\sum_{j=i+1}^\infty\alpha_j^2\Big).
\end{aligned}$$

Since $\sum_{n=1}^\infty\alpha_n^2<\infty$ there is thus a $n_0\in\mathbb{N}$ such that $2\alpha_n^2L^2\mathbf{E}_{n-1}\big[\|w^n-w^*\|^2\big]\le1$ for all $n\ge n_0$. □

Lemma 4.3. *Let Assumption 2 be fulfilled and let $\{\alpha_n\}_{n\in\mathbb{N}}$ be a sequence of positive real numbers such that $\sum_{n=1}^{\infty}\alpha_n^2<\infty$. Then the a priori bound*

$$\begin{aligned}\mathbf{E}_n\big[\|w^{n+1}-w^*\|^4\big] &\leq \|w^1-w^*\|^4\exp\Big(\sum_{i=1}^{\infty}c_1^i(\alpha_i)\Big)\\ &\quad+\sum_{i=1}^{\infty}\big(c_2^i(\alpha_i)M_2+c_3^i(\alpha_i)\big)\exp\Big(\sum_{j=i+1}^{\infty}c_1^j(\alpha_j)\Big)=:M_4\end{aligned}$$

is fulfilled for $c_1^i,c_2^i,c_3^i\colon(0,\infty)\to(0,\infty)$ such that there exist $C_1^k,C_2^k,C_3^k,C_4^k\in(0,\infty)$ with $c_k^i(\alpha)\leq C_1^i\min\{C_2^i,\alpha^4\}+C_3^i\min\{C_4^i,\alpha^2\}$ for all $\alpha\in(0,\infty)$, $k\in\{1,2,3\}$ and $i\in\mathbb{N}$.

Proof. Within the proof of Lemma 4.1, we verified the inequality (4.2). Starting from this point, we find

$$\|w^{n+1}-w^*\|^2-\|w^n-w^*\|^2\leq A_n+B_n, \tag{4.4}$$

where

$$\begin{aligned}A_n&=-\frac{2\alpha_n\langle\iota\nabla f_n(w^*),w^n-w^*\rangle}{1+\alpha_n\Phi}\qquad\text{and}\\ B_n&=\Phi^2\min\big\{\|\iota\nabla f_n(w^n)\|^{-2},\alpha_n^2\big\}+2\min\big\{1,2\alpha_n^2L_{\xi_n}^2\|w^n-w^*\|^2\big\}\\ &\quad+2\min\big\{1,2\alpha_n^2\|\iota\nabla f_n(w^*)\|^2\big\}+2\|\iota\nabla f_n(w^*)\|^2\min\big\{\Phi^{-2},\alpha_n^2\big\}\|w^n-w^*\|^2\end{aligned}$$

for $\Phi\in[0,\infty)$. We note that the parameter Φ can be chosen such that $\mathbf{E}_n[B_n]$ is as small as possible. From this it follows that

$$\begin{aligned}&\mathbf{E}_{\xi_n}[A_n]=0,\\ &\mathbf{E}_{\xi_n}[B_n]\leq\Phi^2\min\big\{\mathbf{E}_{\xi_n}\big[\|\iota\nabla f_n(w^n)\|^{-2}\big],\alpha_n^2\big\}+2\min\big\{1,2\alpha_n^2L^2\|w^n-w^*\|^2\big\}\\ &\qquad+2\min\big\{1,2\alpha_n^2\sigma^2\big\}+2\sigma^2\min\big\{\Phi^{-2},\alpha_n^2\big\}\|w^n-w^*\|^2,\\ &\mathbf{E}_{\xi_n}[A_n^2]\leq4\sigma^2\min\{\Phi^{-2},\alpha_n^2\}\|w^n-w^*\|^2\qquad\text{and}\\ &\mathbf{E}_{\xi_n}[B_n^2]\leq4\Phi^4\min\big\{\big(\mathbf{E}_{\xi_n}\big[\|\iota\nabla f_n(w^n)\|^{-2}\big]\big)^2,\alpha_n^4\big\}+16\min\big\{1,4\alpha_n^4L_4^4\|w^n-w^*\|^4\big\}\\ &\qquad+16\min\big\{1,4\alpha_n^4\sigma_4^4\big\}+16\sigma_4^4\min\big\{\Phi^{-4},\alpha_n^4\big\}\|w^n-w^*\|^4.\end{aligned}$$

We note that for $a,b\in\mathbb{R}$ we have the identity $(a-b)a=\frac{1}{2}\big(|a|^2-|b|^2+|a-b|^2\big)$, and thus $|a|^2-|b|^2\leq2(a-b)a$. By multiplying the inequality from (4.4) with the factor $2\|w^{n+1}-w^*\|^2$, we therefore obtain

$$\begin{aligned}\|w^{n+1}-w^*\|^4-\|w^n-w^*\|^4&\leq2\big(A_n+B_n\big)\|w^{n+1}-w^*\|^2\\ &\leq2\big(A_n+B_n\big)\big(\|w^n-w^*\|^2+A_n+B_n\big)\\ &\leq2(A_n+B_n)\|w^n-w^*\|^2+4A_n^2+4B_n^2\end{aligned}$$

and in $\mathbf{E}_{\xi_n}$-expectation

$$\begin{aligned}
&\mathbf{E}_{\xi_n}\big[\|w^{n+1}-w^*\|^4\big]-\|w^n-w^*\|^4\\
&\le 4\Big(2L^2\min\big\{\tfrac{1}{2}L^{-2}\|w^n-w^*\|^{-2},\alpha_n^2\big\}+\sigma^2\min\big\{\Phi^{-2},\alpha_n^2\big\}\\
&\qquad+64L_4^4\min\big\{\tfrac{1}{4}L_4^{-4}\|w^n-w^*\|^{-4},\alpha_n^4\big\}+16\sigma_4^4\min\big\{\Phi^{-4},\alpha_n^4\big\}\Big)\|w^n-w^*\|^4\\
&\quad+2\Big(\Phi^2\min\big\{\mathbf{E}_{\xi_n}\big[\|\iota\nabla f_n(w^n)\|^{-2}\big],\alpha_n^2\big\}+4\sigma^2\min\big\{\tfrac{1}{2}\sigma^{-2},\alpha_n^2\big\}\\
&\qquad+8\sigma^2\min\{\Phi^{-2},\alpha_n^2\}\Big)\|w^n-w^*\|^2\\
&\quad+16\Phi^4\min\big\{\big(\mathbf{E}_{\xi_n}\big[\|\iota\nabla f_n(w^n)\|^{-2}\big]\big)^2,\alpha_n^4\big\}+256\sigma_4^4\min\big\{\tfrac{1}{4}\sigma_4^{-4},\alpha_n^4\big\}\\
&=: c_1^n(\alpha_n)\|w^n-w^*\|^4+c_2^n(\alpha_n)\|w^n-w^*\|^2+c_3^n(\alpha_n).
\end{aligned}$$

Adding $\|w^n-w^*\|^4$ to both sides of the inequality, taking the $\mathbf{E}_{n-1}$-expectation and reinserting the bound, we obtain

$$\begin{aligned}
&\mathbf{E}_n\big[\|w^{n+1}-w^*\|^4\big]\\
&\le\big(1+c_1^n(\alpha_n)\big)\mathbf{E}_{n-1}\big[\|w^n-w^*\|^4\big]+c_2^n(\alpha_n)M_2+c_3^n(\alpha_n)\\
&\le\|w^1-w^*\|^4\prod_{i=1}^n\big(1+c_1^i(\alpha_i)\big)+\sum_{i=1}^n\big(c_2^i(\alpha_i)M_2+c_3^i(\alpha_i)\big)\prod_{j=i+1}^n\big(1+c_1^j(\alpha_j)\big)\\
&\le\|w^1-w^*\|^4\exp\Big(\sum_{i=1}^\infty c_1^i(\alpha_i)\Big)+\sum_{i=1}^\infty\big(c_2^i(\alpha_i)M_2+c_3^i(\alpha_i)\big)\exp\Big(\sum_{j=i+1}^\infty c_1^j(\alpha_j)\Big).
\end{aligned}$$

Finally, this is finite due to the assumption $\sum_{n=1}^\infty\alpha_n^2<\infty$. □

It is much easier to show the following *pathwise* a priori bound, which provides the intuition for why the scheme is good; in n steps, we can only make the error worse by n in the worst case. This is a marked improvement over the situation for other explicit methods such as SGD, where the error may grow without bound. It is in fact similar to what one would get from an implicit scheme such as the implicit Euler, corresponding to the proximal point method in the context of optimization.

Lemma 4.4. *Let $f(\xi,\cdot)\colon\Omega\times H\to\mathbb{R}$ be Gâteaux differentiable a.s. and let $\{\alpha_n\}_{n\in\mathbb{N}}$ be a sequence of positive real numbers. Then the a priori bound*

$$\|w^{n+1}-w^*\|\le\|w_1-w^*\|+\sum_{i=1}^n\min\{1,\alpha_n\|\iota\nabla f_i(w^i)\|\}\le\|w_1-w^*\|+n$$

is fulfilled.

Proof. We recall the TSGD scheme from (3.1) and obtain that

$$\begin{aligned}
\|w^{n+1}-w^*\|&\le\|w^{n+1}-w^n\|+\|w^n-w^*\|\\
&=\frac{\alpha_n\|\iota\nabla f_n(w^n)\|}{1+\alpha_n\|\iota\nabla f_n(w^n)\|}+\|w^n-w^*\|\\
&\le\min\{1,\alpha_n\|\iota\nabla f_n(w^n)\|\}+\|w^n-w^*\|.
\end{aligned}$$

Reinserting this inequality shows that $\|w^{n+1}-w^*\|\le\|w_1-w^*\|+n$ holds. □

5. Error analysis

Given $z \in H$ and $\alpha > 0$, we define $T_{\alpha f_n, z}(w) \colon \Omega \times H \to H$ by

$$T_{\alpha f_n, z}(w) = w - \frac{\alpha \iota \nabla f_n(w)}{1 + \alpha \|\iota \nabla f_n(z)\|}.$$

This implies that the next iterate w^{n+1} is given by $T_{\alpha_n f_n, w^n}(w^n)$.

Lemma 5.1. *Let Assumption 1 be fulfilled and let $z \in H$, $\alpha \in (0, \infty)$ and $\Xi \in [0, \infty)$ be given. It then follows that*

$$\Big\| T_{\alpha f_n, z}(w) - w + \frac{\alpha \iota \nabla f_n(w)}{1 + \alpha \Xi} \Big\| = \frac{\alpha^2 \big| \Xi - \|\iota \nabla f_n(z)\| \big| \|\iota \nabla f_n(w)\|}{(1 + \alpha \|\iota \nabla f_n(z)\|)(1 + \alpha \Xi)}$$

for all $w \in H$.

Proof. Inserting the definition of $T_{\alpha f_n, z}$, it follows that

$$\begin{aligned}
&\Big\| T_{\alpha f_n, z}(w) - w + \frac{\alpha \iota \nabla f_n(w)}{1 + \alpha \Xi} \Big\| = \Big\| \frac{\alpha \iota \nabla f_n(w)}{1 + \alpha \|\iota \nabla f_n(z)\|} - \frac{\alpha \iota \nabla f_n(w)}{1 + \alpha \Xi} \Big\| \\
&= \Big\| \frac{\alpha \iota \nabla f_n(w) + \alpha^2 \Xi \iota \nabla f_n(w)}{(1 + \alpha \|\iota \nabla f_n(z)\|)(1 + \alpha \Xi)} - \frac{\alpha \iota \nabla f_n(w) + \alpha^2 \|\iota \nabla f_n(z)\| \iota \nabla f_n(w)}{(1 + \alpha \|\iota \nabla f_n(z)\|)(1 + \alpha \Xi)} \Big\| \\
&= \frac{\alpha^2 \big| \Xi - \|\iota \nabla f_n(z)\| \big| \|\iota \nabla f_n(w)\|}{(1 + \alpha \|\iota \nabla f_n(z)\|)(1 + \alpha \Xi)},
\end{aligned}$$

for all $w \in H$, which proves the claim. □

Lemma 5.2. *Let Assumption 1 be fulfilled and let $\{\alpha_n\}_{n \in \mathbb{N}}$ be a sequence of positive real numbers. For any $\Xi \in [0, \infty)$ it then follows that*

$$\begin{aligned}
\mathbf{E}_{\xi_n}\big[\|w^{n+1} - w^*\|^2\big] \leq \Big(1 - \mathbf{E}_{\xi_n}\Big[\frac{2\alpha_n \mu_{\xi_n}}{1 + \alpha_n \|\iota \nabla f_n(w^n)\|}\Big]\Big) \|w^n - w^*\|^2 \\
+ 2\mathbf{E}_{\xi_n}\Big[\frac{\alpha_n^2 \|\iota \nabla f_n(w^n) - \iota \nabla f_n(w^*)\|^2}{(1 + \alpha_n \|\iota \nabla f_n(w^n)\|)^2}\Big] \\
+ 2\mathbf{E}_{\xi_n}\Big[\frac{\alpha_n^2 \|\iota \nabla f_n(w^*)\|^2}{(1 + \alpha_n \|\iota \nabla f_n(w^n)\|)^2}\Big] \\
+ 2\mathbf{E}_{\xi_n}\Big[\frac{\alpha_n^2 \big| \Xi - \|\iota \nabla f_n(w^n)\| \big| \|\iota \nabla f_n(w^*)\|}{(1 + \alpha_n \|\iota \nabla f_n(w^n)\|)(1 + \alpha_n \Xi)}\Big] \|w^n - w^*\|
\end{aligned}$$

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

Proof. From $w^{n+1} = T_{\alpha_n f_n, w^n}(w^n)$ we obtain that

$$\begin{aligned}
\|w^{n+1} - w^*\|^2 &= \|T_{\alpha_n f_n, w^n}(w^n) - T_{\alpha_n f_n, w^n}(w^*) + T_{\alpha_n f_n, w^n}(w^*) - w^*\|^2 \\
&= \|T_{\alpha_n f_n, w^n}(w^n) - T_{\alpha_n f_n, w^n}(w^*)\|^2 + \|T_{\alpha_n f_n, w^n}(w^*) - w^*\|^2 \\
&\quad + 2\langle T_{\alpha_n f_n, w^n}(w^n) - T_{\alpha_n f_n, w^n}(w^*), T_{\alpha_n f_n, w^n}(w^*) - w^* \rangle \\
&=: I_1 + I_2 + 2I_3.
\end{aligned}$$

For I_1, we can write

$$\begin{aligned}
I_1 &= \|T_{\alpha_n f_n, w^n}(w^n) - T_{\alpha_n f_n, w^n}(w^*)\|^2 \\
&= \big\| w^n - w^* + (T_{\alpha_n f_n, w^n} - I)(w^n) - (T_{\alpha_n f_n, w^n} - I)(w^*) \big\|^2 \\
&= \|w^n - w^*\|^2 + 2\big\langle w^n - w^*, (T_{\alpha_n f_n, w^n} - I)(w^n) - (T_{\alpha_n f_n, w^n} - I)(w^*) \big\rangle \\
&\quad + \|(T_{\alpha_n f_n, w^n} - I)(w^n) - (T_{\alpha_n f_n, w^n} - I)(w^*)\|^2 =: I_{1,1} + 2I_{1,2} + I_{1,3}.
\end{aligned}$$

The term $I_{1,2}$ can be estimated by applying the Cauchy–Schwarz inequality:

$$\begin{aligned} I_{1,2} &= \big\langle w^n - w^*, (T_{\alpha_n f_n, w^n} - I)(w^n) - (T_{\alpha_n f_n, w^n} - I)(w^*) \big\rangle \\ &= -\Big\langle w^n - w^*, \frac{\alpha_n \iota \nabla f_n(w^n)}{1 + \alpha_n \|\iota \nabla f_n(w^n)\|} - \frac{\alpha_n \iota \nabla f_n(w^*)}{1 + \alpha_n \|\iota \nabla f_n(w^n)\|} \Big\rangle \\ &\leq -\frac{\alpha_n \mu_{\xi_n}}{1 + \alpha_n \|\iota \nabla f_n(w^n)\|} \|w^n - w^*\|^2. \end{aligned}$$

For $I_{1,3}$, we insert the definition of $T_{\alpha_n f_n, w^n}$ and find

$$\begin{aligned} I_{1,3} &= \|(T_{\alpha_n f_n, w^n} - I)(w^n) - (T_{\alpha_n f_n, w^n} - I)(w^*)\|^2 \\ &= \frac{\alpha_n^2 \|\iota \nabla f_n(w^n) - \iota \nabla f_n(w^*)\|^2}{(1 + \alpha \|\iota \nabla f_n(w^n)\|)^2}. \end{aligned}$$

Thus, for I_1, we have

$$I_1 \leq \Big(1 - \frac{2\alpha_n \mu_{\xi_n}}{1 + \alpha_n \|\iota \nabla f_n(w^n)\|}\Big) \|w^n - w^*\|^2 + \frac{\alpha_n^2 \|\iota \nabla f_n(w^n) - \iota \nabla f_n(w^*)\|^2}{(1 + \alpha \|\iota \nabla f_n(w^n)\|)^2}.$$

Further, I_2 can be written as

$$I_2 = \|T_{\alpha_n f_n, w^n}(w^*) - w^*\|^2 = \frac{\alpha_n^2 \|\iota \nabla f_n(w^*)\|^2}{(1 + \alpha_n \|\iota \nabla f_n(w^n)\|)^2}.$$

Finally, I_3 can be rewritten as

$$\begin{aligned} I_3 &= \langle T_{\alpha_n f_n, w^n}(w^n) - T_{\alpha_n f_n, w^n}(w^*), T_{\alpha_n f_n, w^n}(w^*) - w^* \rangle \\ &= \langle (T_{\alpha_n f_n, w^n} - I)(w^n) - (T_{\alpha_n f_n, w^n} - I)(w^*), (T_{\alpha_n f_n, w^n} - I)(w^*) \rangle \\ &\quad + \langle w^n - w^*, (T_{\alpha_n f_n, w^n} - I)(w^*) \rangle =: I_{3,1} + I_{3,2}. \end{aligned}$$

Then for $I_{3,1}$, we insert the definition of $T_{\alpha_n f_n, w^n}$ and obtain

$$\begin{aligned} I_{3,1} &= \langle (T_{\alpha_n f_n, w^n} - I)(w^n) - (T_{\alpha_n f_n, w^n} - I)(w^*), (T_{\alpha_n f_n, w^n} - I)(w^*) \rangle \\ &\leq \frac{\alpha_n \|\iota \nabla f_n(w^n) - \iota \nabla f_n(w^*)\|}{1 + \alpha_n \|\iota \nabla f_n(w^n)\|} \cdot \frac{\alpha_n \|\iota \nabla f_n(w^*)\|}{1 + \alpha_n \|\iota \nabla f_n(w^n)\|} \\ &\leq \frac{1}{2} \frac{\alpha_n^2 \|\iota \nabla f_n(w^n) - \iota \nabla f_n(w^*)\|^2}{(1 + \alpha_n \|\iota \nabla f_n(w^n)\|)^2} + \frac{1}{2} \frac{\alpha_n^2 \|\iota \nabla f_n(w^*)\|^2}{(1 + \alpha_n \|\iota \nabla f_n(w^n)\|)^2}, \end{aligned}$$

where we applied the Cauchy–Schwarz inequality and Young's inequality. To estimate $I_{3,2}$, we add and subtract an additional summand, so that

$$\begin{aligned} I_{3,2} &= \langle w^n - w^*, (T_{\alpha_n f_n, w^n} - I)(w^*) \rangle \\ &= \Big\langle w^n - w^*, (T_{\alpha_n f_n, w^n} - I)(w^*) + \frac{\alpha_n \iota \nabla f_n(w^*)}{1 + \alpha_n \Xi} \Big\rangle - \Big\langle w^n - w^*, \frac{\alpha_n \iota \nabla f_n(w^*)}{1 + \alpha_n \Xi} \Big\rangle, \end{aligned}$$

where $\mathbf{E}_{\xi_n}\big[\big\langle w^n - w^*, \frac{\alpha_n \iota \nabla f_n(w^*)}{1 + \alpha_n \Xi} \big\rangle\big] = 0$ is fulfilled. Moreover, applying Lemma 5.1 and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} &\Big\langle w^n - w^*, (T_{\alpha_n f_n, w^n} - I)(w^*) + \frac{\alpha_n \iota \nabla f_n(w^*)}{1 + \alpha_n \Xi} \Big\rangle \\ &\leq \frac{\alpha_n^2 \big|\Xi - \|\iota \nabla f_n(w^n)\|\big| \|\iota \nabla f_n(w^*)\|}{(1 + \alpha_n \|\iota \nabla f_n(w^n)\|)(1 + \alpha_n \Xi)} \|w^n - w^*\|. \end{aligned}$$

Thus, the expectation of $I_3 = I_{3,1} + I_{3,2}$ can be bounded by

$$\begin{aligned}\mathbf{E}_{\xi_n}[I_3] \leq \frac{1}{2}\mathbf{E}_{\xi_n}\Big[\frac{\alpha_n^2\|\iota\nabla f_n(w^n) - \iota\nabla f_n(w^*)\|^2}{\big(1+\alpha_n\|\iota\nabla f_n(w^n)\|\big)^2} + \frac{\alpha_n^2\|\iota\nabla f_n(w^*)\|^2}{(1+\alpha_n\|\iota\nabla f_n(w^n)\|)^2}\Big] \\ + \mathbf{E}_{\xi_n}\Big[\frac{\alpha_n^2\big|\Xi - \|\iota\nabla f_n(w^n)\|\big|\|\iota\nabla f_n(w^*)\|}{(1+\alpha_n\|\iota\nabla f_n(w^n)\|)(1+\alpha_n\Xi)}\Big]\|w^n - w^*\|.\end{aligned}$$

Inserting the bounds for I_1, I_2 and I_3 into $\mathbf{E}_{\xi_n}\big[\|w^{n+1} - w^*\|^2\big] = \mathbf{E}_{\xi_n}\big[I_1 + I_2 + 2I_3\big]$ finishes the proof. □

Theorem 5.3. *Let Assumption 2 be fulfilled. For $\alpha_n = \frac{\vartheta}{n+\gamma}$, $n \in \mathbb{N}$, with $\gamma \in [0,\infty)$, $\vartheta \in (0, \frac{1+\gamma}{2\mu}]$ and*

$$K = 2\vartheta^2 L\mu_2 M_4^{\frac{3}{4}} + \big(2\vartheta^2L^2 + 2\vartheta^2 L\sigma + 2\vartheta^2\mu_2\sigma\big)M_2 + 2\vartheta^2\sigma^2 M_2^{\frac{1}{2}} + 2\vartheta^2\sigma^2,$$

it follows that

$$\begin{aligned}&\mathbf{E}_n\big[\|w^{n+1} - w^*\|^2\big] \\ &\leq \|w_1 - w^*\|^2(1+\gamma)^{2\vartheta\mu}(n+1+\gamma)^{-2\vartheta\mu} \\ &\quad + \exp\Big(\frac{2\vartheta\mu}{1+\gamma}\Big)K\begin{cases}(n+1+\gamma)^{-1}\frac{1}{2\vartheta\mu-1}, & 2\vartheta\mu \in (1,\infty), \\ (n+1+\gamma)^{-1}(1+\ln(n+\gamma)), & 2\vartheta\mu = 1, \\ (n+1+\gamma)^{-2\vartheta\mu}\frac{(1+\gamma)^{2\vartheta\mu-2}(2\vartheta\mu-2-\gamma)}{2\vartheta\mu-1}, & 2\vartheta\mu \in [0,1),\end{cases}\end{aligned}$$

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

Remark 5.4. By choosing $\vartheta \in (\frac{1}{2\mu}, \infty)$ we obtain the optimal convergence rate. A value of ϑ much larger than $\frac{1}{2\mu}$ does not improve the overall rate further, but does affect the exponent in the first term of the bound that involves the initial error. We note that since $\frac{1+\gamma}{2\mu} \geq \frac{1}{2\mu}$, it is possible to make this choice for any γ. We could in fact instead have analyzed the simpler step size sequence with $\alpha_n = \frac{\vartheta}{n}$, but chose to present the results in this form in order to match our other results and comparable results for e.g. SGD [6, Theorem 4.7]. The results for the simpler sequence are recovered by simply setting $\gamma = 0$.

Proof of Theorem 5.3. The main idea of the proof is to apply the bound from Lemma 5.2 with $\Xi = 0$ and bound the denominators of the last three summands

from below by one. We then get

$$
\begin{aligned}
&\mathbf{E}_{\xi_n}\big[\|w^{n+1}-w^*\|^2\big]\\
&\le \Big(1-\mathbf{E}_{\xi_n}\Big[\frac{2\alpha_n\mu_{\xi_n}}{1+\alpha_n\|\iota\nabla f_n(w^n)\|}\Big]\Big)\|w^n-w^*\|^2+2\alpha_n^2L^2\|w^n-w^*\|^2+2\alpha_n^2\sigma^2\\
&\quad+2\alpha_n^2\big(\mathbf{E}_{\xi_n}\big[\|\iota\nabla f_n(w^n)\|^2\big]\big)^{\frac12}\sigma\|w^n-w^*\|\\
&\le \Big(1-\mathbf{E}_{\xi_n}\Big[\frac{2\alpha_n\mu_{\xi_n}}{1+\alpha_nL_{\xi_n}\|w^n-w^n\|+\alpha_n\|\iota\nabla f_n(w^*)\|}\Big]\Big)\|w^n-w^*\|^2\\
&\quad+\big(2\alpha_n^2L^2+2\alpha_n^2L\sigma\big)\|w^n-w^*\|^2+2\alpha_n^2\sigma^2\|w^n-w^*\|+2\alpha_n^2\sigma^2\\
&\le \Big(1-\mathbf{E}_{\xi_n}\Big[\frac{2\alpha_n\mu_{\xi_n}}{1+\alpha_n\|\iota\nabla f_n(w^*)\|}+\frac{2\alpha_n^2L_{\xi_n}\mu_{\xi_n}}{(1+\alpha_n\|\iota\nabla f_n(w^*)\|)^2}\|w^n-w^*\|\Big]\Big)\|w^n-w^*\|^2\\
&\quad+\alpha_n^2\big(\big(2L^2+2L\sigma\big)\|w^n-w^*\|^2+2\sigma^2\|w^n-w^*\|+2\sigma^2\big)\\
&\le \Big(1-\mathbf{E}_{\xi_n}\Big[\frac{2\alpha_n\mu_{\xi_n}}{1+\alpha_n\|\iota\nabla f_n(w^*)\|}\Big]\Big)\|w^n-w^*\|^2\\
&\quad+\alpha_n^2\big(2L\mu_2\|w^n-w^*\|^3+\big(2L^2+2L\sigma\big)\|w^n-w^*\|^2+2\sigma^2\|w^n-w^*\|+2\sigma^2\big),
\end{aligned}
$$

where we added and subtracted $\iota\nabla f_n(w^*)$ and applied Minkowski's inequality in the second step and Lemma A.2 in the third. For the first summand of the previous inequality, we apply Lemma A.2 once more and find

$$
\begin{aligned}
&\Big(1-\mathbf{E}_{\xi_n}\Big[\frac{2\alpha_n\mu_{\xi_n}}{1+\alpha_n\|\iota\nabla f_n(w^*)\|}\Big]\Big)\|w^n-w^*\|^2\\
&=\Big(1-\mathbf{E}_{\xi_n}\Big[\frac{2\vartheta\mu_{\xi_n}}{n+\gamma+\vartheta\|\iota\nabla f_n(w^*)\|}\Big]\Big)\|w^n-w^*\|^2\\
&\le \Big(1-\mathbf{E}_{\xi_n}\Big[\frac{2\vartheta\mu_{\xi_n}}{n+\gamma}\Big]+\mathbf{E}_{\xi_n}\Big[\frac{2\vartheta\mu_{\xi_n}}{(n+\gamma)^2}\vartheta\|\iota\nabla f_n(w^*)\|\Big]\Big)\|w^n-w^*\|^2\\
&\le \Big(1-\frac{2\vartheta\mu}{n+\gamma}\Big)\|w^n-w^*\|^2+2\alpha_n^2\mu_2\sigma\|w^n-w^*\|^2.
\end{aligned}
$$

Taking the $\mathbf{E}_{n-1}$-expectation and applying the a priori bounds from Lemma 4.1 and Lemma 4.3, we find that

$$
\begin{aligned}
&\mathbf{E}_n\big[\|w^{n+1}-w^*\|^2\big]\\
&\le \Big(1-\frac{2\vartheta\mu}{n+\gamma}\Big)\mathbf{E}_{n-1}\big[\|w^n-w^*\|^2\big]+2\alpha_n^2L\mu_2\mathbf{E}_{n-1}\big[\|w^n-w^*\|^3\big]\\
&\quad+\big(2\alpha_n^2L^2+2\alpha_n^2L\sigma+2\alpha_n^2\mu_2\sigma\big)\mathbf{E}_{n-1}\big[\|w^n-w^*\|^2\big]\\
&\quad+2\alpha_n^2\sigma^2\mathbf{E}_{n-1}\big[\|w^n-w^*\|\big]+2\alpha_n^2\sigma^2\\
&\le \Big(1-\frac{2\vartheta\mu}{n+\gamma}\Big)\mathbf{E}_{n-1}\big[\|w^n-w^*\|^2\big]+2\alpha_n^2L\mu_2M_4^{\frac34}\\
&\quad+\big(2\alpha_n^2L^2+2\alpha_n^2L\sigma+2\alpha_n^2\mu_2\sigma\big)M_2+2\alpha_n^2\sigma^2M_2^{\frac12}+2\alpha_n^2\sigma^2\\
&=\Big(1-\frac{2\vartheta\mu}{n+\gamma}\Big)\mathbf{E}_{n-1}\big[\|w^n-w^*\|^2\big]+\frac{K}{(n+\gamma)^2}.
\end{aligned}
$$

Reinserting the inequality $n-1$ times yields

$$
\begin{aligned}
&\mathbf{E}_n\big[\|w^{n+1}-w^*\|^2\big]\\
&\le \|w_1-w^*\|^2\prod_{i=1}^n\Big(1-\frac{2\vartheta\mu}{\gamma+i}\Big)+K\sum_{i=1}^n\frac{1}{(\gamma+i)^2}\prod_{j=i+1}^n\Big(1-\frac{2\vartheta\mu}{\gamma+j}\Big).
\end{aligned}
$$

Due to the assumption $\vartheta \in (0, \frac{1+\gamma}{2\mu}]$, we can apply Lemma A.1 with $x = 2\vartheta\mu$ and $y = \gamma$ and obtain the claimed error bound. □

In comparison to convergence results regarding SGD, e.g. [6, Theorem 4.7], the higher-moment bounds in Assumption 2 are not necessary. These are in fact not needed to prove convergence of TSGD, as the following theorem demonstrates. The drawback is that the contraction parameter is given by $1 - \frac{2\alpha_n\mu_{\xi_n}}{1+\alpha_n\|\iota\nabla f_n(w^n)\|}$, where we cannot verify that $\|\iota\nabla f_n(w^n)\|$ is bounded. Thus, it is not necessarily possible to prove the optimal rate of convergence in all cases. We note that the step size sequence involving γ is important here, as it allows us to choose a large ϑ for which the parameter C defined in the theorem below becomes as large as possible, leading to the best possible rate. A larger γ also increases the error term arising from the initial error, but as argued in [6, p. 251] the influence of this term can be minimized by precomputing a better w_1 using e.g. TSGD with a constant step size. We note that the condition $C \leq 1 + \gamma$ is mostly technical, will likely not be an issue in practice, and can always be satisfied by choosing $\gamma \geq 1$.

Theorem 5.5. *Let Assumption 1 be fulfilled. For $\alpha_n = \frac{\vartheta}{n+\gamma}$, $n \in \mathbb{N}$, with $\gamma \in [0, \infty)$, $\vartheta \in (0, \infty)$ such that*

$$C = \mathbf{E}_\xi\Big[\frac{2\gamma\vartheta\mu_\xi}{\gamma + \vartheta L_\xi\|w_1 - w^*\| + \gamma\vartheta L_\xi + \vartheta\|\iota\nabla f(\xi, w^*)\|}\Big],$$
$$K = 2\vartheta^2\big((L^2 + L\sigma)M_2 + \sigma^2 + \sigma^2 M_2^{\frac{1}{2}}\big)$$

and $C \in (0, 1 + \gamma]$ it follows that

$$\begin{aligned}
&\mathbf{E}_n\big[\|w^{n+1} - w^*\|^2\big] \\
&\leq \|w_1 - w^*\|^2(1+\gamma)^C(n+1+\gamma)^{-C} \\
&\quad + \exp\Big(\frac{C}{1+\gamma}\Big)K \begin{cases} (n+1+\gamma)^{-1}\frac{1}{C-1}, & C \in (1, \infty), \\ (n+1+\gamma)^{-1}(1 + \ln(n+\gamma)), & C = 1, \\ (n+1+\gamma)^{-C}\frac{(1+\gamma)^{C-2}(C-2-\gamma)}{C-1}, & C \in [0, 1), \end{cases}
\end{aligned}$$

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

Proof. As in the proof of Theorem 5.3, the main idea of the proof is to apply the bound from Lemma 5.2 with $\Xi = 0$, where we also bound the denominators of the last three summands from below by one. We then get

$$\begin{aligned}
&\mathbf{E}_{\xi_n}\big[\|w^{n+1} - w^*\|^2\big] \\
&\leq \Big(1 - \mathbf{E}_{\xi_n}\Big[\frac{2\alpha_n\mu_{\xi_n}}{1+\alpha_n\|\iota\nabla f_n(w^n)\|}\Big]\Big)\|w^n - w^*\|^2 + 2\alpha_n^2L^2\|w^n - w^*\|^2 + 2\alpha_n^2\sigma^2 \\
&\quad + 2\alpha_n^2\big(\mathbf{E}_{\xi_n}\big[\|\iota\nabla f_n(w^n)\|^2\big]\big)^{\frac{1}{2}}\sigma\|w^n - w^*\| \\
&\leq \Big(1 - \mathbf{E}_{\xi_n}\Big[\frac{2\alpha_n\mu_{\xi_n}}{1+\alpha_n\|\iota\nabla f_n(w^n)\|}\Big]\Big)\|w^n - w^*\|^2 + 2\alpha_n^2(L^2 + L\sigma)\|w^n - w^*\|^2 \\
&\quad + 2\alpha_n^2\sigma^2 + 2\alpha_n^2\sigma^2\|w^n - w^*\| \\
&=: \Big(1 - \mathbf{E}_{\xi_n}\Big[\frac{2\alpha_n\mu_{\xi_n}}{1+\alpha_n\|\iota\nabla f_n(w^n)\|}\Big]\Big)\|w^n - w^*\|^2 + \alpha_n^2 I,
\end{aligned}$$

where we added and subtracted $\iota\nabla f_n(w^*)$ and applied Minkowski's inequality in the second step. Using the a priori bound from Lemma 4.1, we find that

$$\begin{aligned}\alpha_n^2\mathbf{E}_{n-1}[I] &= 2\alpha_n^2\big((L^2+L\sigma)\mathbf{E}_{n-1}\big[\|w^n-w^*\|^2\big]+\sigma^2+\sigma^2\mathbf{E}_{n-1}\big[\|w^n-w^*\|\big]\big)\\ &\leq \frac{2\vartheta^2}{(n+\gamma)^2}\big((L^2+L\sigma)M_2+\sigma^2+\sigma^2M_2^{\frac{1}{2}}\big)=\frac{K}{(n+\gamma)^2}.\end{aligned}$$

Applying the pathwise a priori bound from Lemma 4.4, it follows that

$$\begin{aligned}\alpha_n\|\iota\nabla f_n(w^n)\| &\leq \alpha_nL_{\xi_n}\|w^n-w^*\|+\alpha_n\|\iota\nabla f_n(w^*)\|\\ &\leq \frac{\vartheta}{n+\gamma}L_{\xi_n}\|w_1-w^*\|+\frac{\vartheta n}{\gamma+n}L_{\xi_n}+\frac{\vartheta}{n+\gamma}\|\iota\nabla f_n(w^*)\|\\ &\leq \frac{\vartheta}{\gamma}L_{\xi_n}\|w_1-w^*\|+\vartheta L_{\xi_n}+\frac{\vartheta}{\gamma}\|\iota\nabla f_n(w^*)\|.\end{aligned}$$

Thus, we find

$$\begin{aligned}&1-\mathbf{E}_{\xi_n}\Big[\frac{2\alpha_n\mu_{\xi_n}}{1+\alpha_n\|\iota\nabla f_n(w^n)\|}\Big]\\ &\leq 1-\frac{1}{n+\gamma}\cdot\mathbf{E}_{\xi}\Big[\frac{2\gamma\vartheta\mu_{\xi}}{\gamma+\vartheta L_{\xi}\|w_1-w^*\|+\gamma\vartheta L_{\xi}+\vartheta\|\iota\nabla f(\xi,w^*)\|}\Big]=1-\frac{C}{n+\gamma}.\end{aligned}$$

Altogether, this implies that

$$\begin{aligned}&\mathbf{E}_n\big[\|w^{n+1}-w^*\|^2\big]\\ &\leq \|w_1-w^*\|^2\prod_{i=1}^n\Big(1-\frac{C}{\gamma+i}\Big)+K\sum_{i=1}^n\frac{1}{(\gamma+i)^2}\prod_{j=i+1}^n\Big(1-\frac{C}{\gamma+j}\Big).\end{aligned}$$

is fulfilled. As $C\in(0,1+\gamma]$ is fulfilled by assumption, the claim of the theorem can be verified by an application of Lemma A.1 with $x=C$ and $y=\gamma$. □

In the penultimate theorem, we prove a convergence result under the additional assumption that the gradient is bounded. Note that the error bound does not increase uncontrollably with growing γ or ϑ. The terms γ and ϑB always appear with a positive exponent in one factor and with the same, but negative, exponent in another factor. This verifies that TSGD is very stable with respect to large initial step sizes.

Theorem 5.6. *Let Assumption 3 be fulfilled. For $\alpha_n=\frac{\vartheta}{n+\gamma}$, $n\in\mathbb{N}$ with $\gamma\in[0,\infty)$, $1+\gamma\geq\vartheta\big(2\mu-B\big)$ and $K=\big(4+6D^2\big)B^2+2\big(B^2D+\sigma B\big)M_2^{\frac{1}{2}}$ it follows that*

$$\begin{aligned}&\mathbf{E}_n\big[\|w^{n+1}-w^*\|^2\big]\\ &\leq \|w_1-w^*\|^2(1+\gamma+\vartheta B)^{2\vartheta\mu}(n+1+\gamma+\vartheta B)^{-2\vartheta\mu}\\ &\quad+\exp\Big(\frac{2\mu\vartheta}{1+\gamma+\vartheta B}\Big)\vartheta^2K\times\\ &\quad\begin{cases}(n+1+\gamma+\vartheta B)^{-1}\frac{1}{2\vartheta\mu-1}, & 2\vartheta\mu\in(1,\infty),\\ (n+1+\gamma+\vartheta B)^{-1}\big(1+\ln\left(n+\gamma+\vartheta B\right)\big), & 2\vartheta\mu=1,\\ (n+1+\gamma+\vartheta B)^{-2\vartheta\mu}\frac{(1+\gamma+\vartheta B)^{2\vartheta\mu-2}(2\vartheta\mu-2-\gamma+\vartheta B)}{2\vartheta\mu-1}, & 2\vartheta\mu\in[0,1)\end{cases}\end{aligned}$$

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

Proof. Again, we apply the bound from Lemma 5.2 but this time with $\Xi = B$ to acquire

$$\begin{aligned}
&\mathbf{E}_{\xi_n}\big[\|w^{n+1} - w^*\|^2\big] \\
&\leq \Big(1 - \mathbf{E}_{\xi_n}\Big[\frac{2\alpha_n\mu_{\xi_n}}{1+\alpha_n\|\iota\nabla f_n(w^n)\|}\Big]\Big)\|w^n - w^*\|^2 \\
&\quad + 4\mathbf{E}_{\xi_n}\Big[\frac{\alpha_n^2\|\iota\nabla f_n(w^n)\|^2}{(1+\alpha_n\|\iota\nabla f_n(w^n)\|)^2}\Big] + 6\mathbf{E}_{\xi_n}\Big[\frac{\alpha_n^2\|\iota\nabla f_n(w^*)\|^2}{(1+\alpha_n\|\iota\nabla f_n(w^n)\|)^2}\Big] \\
&\quad + 2\mathbf{E}_{\xi_n}\Big[\frac{\alpha_n^2\big(B + \|\iota\nabla f_n(w^n)\|\big)\|\iota\nabla f_n(w^*)\|}{(1+\alpha_n\|\iota\nabla f_n(w^n)\|)(1+\alpha_n B)}\Big]\|w^n - w^*\| \\
&\leq \Big(1 - \mathbf{E}_{\xi_n}\Big[\frac{2\alpha_n\mu_{\xi_n}}{1+\alpha_n B}\Big]\Big)\|w^n - w^*\|^2 + \frac{4\alpha_n^2 B^2}{(1+\alpha_n B)^2} \\
&\quad + 6\mathbf{E}_{\xi_n}\Big[\frac{\alpha_n^2\|\iota\nabla f_n(w^*)\|^2}{(1+\alpha_n\|\iota\nabla f_n(w^n)\|)^2}\Big] \\
&\quad + 2\Big(\frac{\alpha_n B}{1+\alpha_n B}\mathbf{E}_{\xi_n}\Big[\frac{\alpha_n\|\iota\nabla f_n(w^*)\|}{1+\alpha_n\|\iota\nabla f_n(w^n)\|}\Big] + \frac{\alpha_n^2\sigma B}{(1+\alpha_n B)^2}\Big)\|w^n - w^*\|,
\end{aligned}$$

where we used in the last step that the function $x \mapsto \frac{x}{1+x}$ is monotonically increasing for $x \in [0,\infty)$. We also have

$$\begin{aligned}
\mathbf{E}_{\xi_n}\Big[\frac{\alpha_n^2\|\iota\nabla f_n(w^*)\|^2}{(1+\alpha_n\|\iota\nabla f_n(w^n)\|)^2}\Big] &\leq \mathbf{E}_{\xi_n}\Big[\frac{\|\iota\nabla f_n(w^*)\|^2}{\|\iota\nabla f_n(w^n)\|^2}\cdot\frac{\alpha_n^2\|\iota\nabla f_n(w^n)\|^2}{(1+\alpha_n\|\iota\nabla f_n(w^n)\|)^2}\Big] \\
&\leq \frac{\alpha_n^2 B^2 D^2}{(1+\alpha_n B)^2},
\end{aligned}$$

and analogously $\mathbf{E}_{\xi_n}\big[\frac{\alpha_n\|\iota\nabla f_n(w^*)\|}{1+\alpha_n\|\iota\nabla f_n(w^n)\|}\big] \leq \frac{\alpha_n BD}{1+\alpha_n B}$. It then follows that

$$\begin{aligned}
\mathbf{E}_{\xi_n}\big[\|w^{n+1} - w^*\|^2\big] &\leq \Big(1 - \frac{2\alpha_n\mu}{1+\alpha_n B}\Big)\|w^n - w^*\|^2 + \frac{\alpha_n^2 B^2\big(4+6D^2\big)}{(1+\alpha_n B)^2} \\
&\quad + \frac{2\alpha_n^2 B^2 D + 2\alpha_n^2\sigma B}{(1+\alpha_n B)^2}\|w^n - w^*\|,
\end{aligned}$$

and taking the $\mathbf{E}_{n-1}$-expectation, we find that

$$\begin{aligned}
&\mathbf{E}_n\big[\|w^{n+1} - w^*\|^2\big] \\
&\leq \Big(1 - \frac{2\alpha_n\mu}{1+\alpha_n B}\Big)\mathbf{E}_{n-1}\big[\|w^n - w^*\|^2\big] + \frac{\alpha_n^2 B^2\big(4+6D^2\big)}{(1+\alpha_n B)^2} \\
&\quad + \frac{2\alpha_n^2 B^2 D + 2\alpha_n^2\sigma B}{(1+\alpha_n B)^2}\big(\mathbf{E}_{n-1}\big[\|w^n - w^*\|^2\big]\big)^{\frac{1}{2}} \\
&\leq \Big(1 - \frac{2\alpha_n\mu}{1+\alpha_n B}\Big)\mathbf{E}_{n-1}\big[\|w^n - w^*\|^2\big] + \frac{\alpha_n^2 B^2\big(4+6D^2\big)}{(1+\alpha_n B)^2} + \frac{2\alpha_n^2\big(B^2 D + \sigma B\big)M_2^{\frac{1}{2}}}{(1+\alpha_n B)^2} \\
&= \Big(1 - \frac{2\vartheta\mu}{n+\gamma+\vartheta B}\Big)\mathbf{E}_{n-1}\big[\|w^n - w^*\|^2\big] + \frac{\vartheta^2 K}{(n+\gamma+\vartheta B)^2}.
\end{aligned}$$

Reinserting the bound $n-1$ times, it follows that

$$\begin{aligned}
\mathbf{E}_n\big[\|w^{n+1} - w^*\|^2\big] &\leq \|w_1 - w^*\|^2\prod_{i=1}^{n}\Big(1 - \frac{2\vartheta\mu}{i+\gamma+\vartheta B}\Big) \\
&\quad + \vartheta^2 K\sum_{i=1}^{n}\frac{1}{(i+\gamma+\vartheta B)^2}\prod_{j=i+1}^{n}\Big(1 - \frac{2\vartheta\mu}{j+\gamma+\vartheta B}\Big).
\end{aligned}$$

Due to the restriction $n + \gamma \geq \vartheta(2\mu - B)$, we can now apply Lemma A.1 with $x = 2\vartheta\mu$ and $y = \gamma + \vartheta B$ in order to finish the proof of the theorem. □

We note that convergence results in this area are often stated in the form $F(w^n) - F(w^*) \leq \frac{C}{n}$. The above theorems are a slightly stronger version, in that they prove convergence of the iterates themselves. However, in our setting these types of convergence are actually equivalent, as the following theorem shows. We note that it is possible to use an approach similar to the one above to directly prove the convergence of $\{F(w^n)\}_{n\in\mathbb{N}}$. The error constants thereby acquired are similar to those acquired from a combination of one of the theorems above and Theorem 5.7 below. However, the reverse approach of proving convergence of $\{w^n\}_{n\in\mathbb{N}}$ by using convergence of $\{F(w^n)\}_{n\in\mathbb{N}}$ results in an additional factor $\frac{2}{\mu}$ which is typically very large.

Theorem 5.7. *Let Assumption 1 be fulfilled. Then* $\{\mathbf{E}_n[\|w^{n+1} - w^*\|^2]\}_{n\in\mathbb{N}}$ *behaves asymptotically the same as* $\{\mathbf{E}_n[F(w^{n+1})] - F(w^*)\}_{n\in\mathbb{N}}$. *More precisely,*

$$\mathbf{E}_n[F(w^{n+1})] - F(w^*) \leq \frac{L}{2}\mathbf{E}_n[\|w^{n+1} - w^*\|^2], \quad \text{and}$$
$$\mathbf{E}_n[\|w^{n+1} - w^*\|^2] \leq \frac{2}{\mu}\mathbf{E}_n[F(w^{n+1})] - F(w^*)$$

are fulfilled for every $n \in \mathbb{N}$.

Proof. By applying Lemma 2.1, with $v = w^{n+1}$ and $w = w^*$ we find that

$$F(w^{n+1}) \leq F(w^*) + \langle \iota\nabla F(w^*), w^{n+1} - w^*\rangle + \frac{L}{2}\|w^{n+1} - w^*\|^2.$$

But since $\nabla F(w^*) = 0$, this directly implies

$$\mathbf{E}_n[F(w^{n+1})] - F(w^*) \leq \frac{L}{2}\mathbf{E}_n[\|w^{n+1} - w^*\|^2].$$

As F is strongly convex, it follows that

$$\begin{aligned}\mathbf{E}_n[F(w^{n+1})] - F(w^*) &\geq \mathbf{E}_n[\langle \iota\nabla F(w^*), w^{n+1} - w^*\rangle] + \frac{\mu}{2}\mathbf{E}_n[\|w^{n+1} - w^*\|^2] \\ &= \frac{\mu}{2}\mathbf{E}_n[\|w^{n+1} - w^*\|^2],\end{aligned}$$

since $\nabla F(w^*) = 0$ which verifies the second inequality. □

6. Numerical experiments

In this section, we illustrate our theoretical results and the advantages of the TSGD method by performing a few numerical experiments. We consider binary classification, which means that we have $N \in \mathbb{N}$ given data samples $x_i \in \mathbb{R}^d$ and corresponding labels $y_i \in \{0, 1\}$, $i \in \{1, \dots, N\}$. Each x_i belongs to one of two classes; to the first one if $y_i = 0$ and to the second if $y_i = 1$. The goal is to find a prediction function $h_w \colon \mathbb{R}^d \to \mathbb{R}$ such that $h_w(x_i) \approx y_i$ for every $i \in \{1, \dots, N\}$. The prediction function depends on the parameters $w \in \mathbb{R}^{n_w}$ and is of a specific, given type. Here, we consider two different types; the first is a support vector machine (SVM) where $h_w(x) = \langle \hat{w}, x\rangle + b$ for $w = (\hat{w}, b)$. This is an affine classifier, which fits into our analysis. The second type is a general fully connected neural network [13] where $h_w(x)$ depends nonlinearly on the parameters w. This type of classifier does not fit directly into our analysis, but the TSGD method still performs well.

To measure the performance of the classifier we use the log loss function $\ell\colon \mathbb{R}^2 \to \mathbb{R}$ given by $\ell(h_w(x), y) = \ln(1 + \exp(-h_w(x)y))$. We also add a regularization term $\frac{\lambda}{2}\|w\|^2$ with $\lambda \in (0, \infty)$, which makes the problem strongly convex in the SVM case. The overall problem is then to minimize the empirical risk $F(w)$, where

$$F(w) = \frac{1}{N}\sum_{i=1}^{N} \ell(h_w(x_i), y_i) + \frac{\lambda}{2}\|w\|^2.$$

We choose two different data sets from the LIBSVM collection*, namely the *mushroom* data set (originally from the UCI Machine Learning Repository [9]†) and the *rcv1.binary* data set [21]‡. The former has $N = 8124$ samples with $d = 112$ features while the latter contains $N = 20242$ samples with $d = 47236$ features each.

The regularization parameter λ corresponds to the convexity parameter μ from Assumption 1. The other parameters L and σ appearing in our theory are more difficult to state explicitly. As our theoretical results show that the choice of TSGD step size does not depend on these parameters, this is not an issue for TSGD. In comparison, for the step size sequence $\{\alpha_n\}_{n\in\mathbb{N}}$ with $\alpha_n = \frac{\vartheta}{n+\gamma}$, the optimal choice for the parameter γ for SGD can depend on $\frac{1}{L}$, see [6, Theorem 4.7]. This makes it more difficult to find a suitable initial step size.

We have implemented both the SGD and TSGD methods in Python. Due to the low complexity of the methods, this is fairly straightforward, and the main issue is how to compute the gradients $\nabla f_n(w^n)$, $n \in \mathbb{N}$. In the SVM case, this is also straightforward, and we can directly write down a closed-form expression that depends on the data x_i, $i \in \{1, \dots, N\}$. In the neural network case, we rely on the scikit-learn library [27] and its backpropagation implementation. In both cases, we deviate slightly from the presented analysis in that we do not choose the batches completely randomly. Instead, we follow the conventional procedure of splitting the data set into a number of batches and picking from these without replacement. When there are none left, the data is reshuffled and new batches are created. One such sequence is referred to as an epoch.

In the examples, we compare the TSGD method with the classical SGD method. We plot the errors $\mathbf{E}_n\big[F(w^{n+1})\big] - F(w^*)$, which according to Theorem 5.7 behave similarly to the errors $\|w^{n+1} - w^*\|^2$ for the number of steps $n \in \mathbb{N}$. As we only prove the convergence in expectation, we use 100 sample paths and plot the average error for them. As the mushroom data set is comparably small, we can compute a reference solution $F(w^*)$ by using the nonlinear equation solver provided by the package `scipy.optimize` [41] in the SVM setting. This enables us to show the exact values $F(w^n) - F(w^*)$. In all the other examples, we compute a reference solution $F(w^*)$ by simply running the TSGD scheme for more steps and choosing the lowest value obtained during all steps. The $F(w^*)$ thereby acquired is not the exact minimum but a very good approximation thereof. We choose TSGD to obtain the reference solution $F(w^*)$ as smaller values $F(w^n)$, $n \in \mathbb{N}$, are obtained using this method and therefore the obtained value $F(w^*)$ is as small as possible.

In the following two sub-sections we further describe parameter choices and the results of the different settings.

*Hosted at https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.

†Available at https://archive.ics.uci.edu/ml/datasets/mushroom.

‡Available at https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#rcv1.binary.

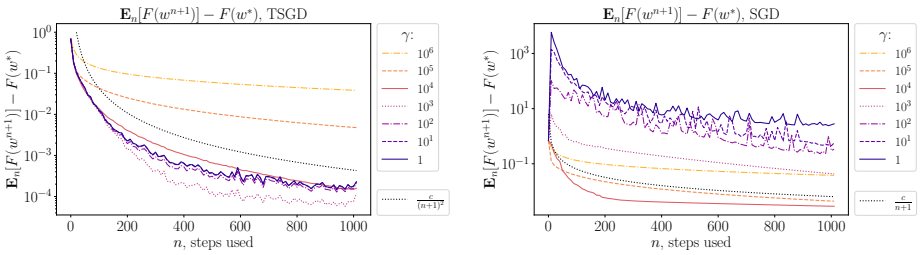

FIGURE 1. TSGD (left) versus SGD (right), different step sizes, SVM, mushrooms

6.1. **Support vector machine.** We used a batch size of 1% of the amount of samples for both data sets. The regularization parameter was chosen as $\lambda = 10^{-5}$. We ran the example for 10 epochs but only stored every tenth value $F(w^n)$ in order to save computational costs. For the step size $\alpha_n = \frac{\vartheta}{n+\gamma}$ we chose $\vartheta = 2 \cdot 10^5 = 2\lambda^{-1}$ in order to ensure the optimal speed of convergence of TSGD from Theorems 5.3 and 5.6 and to fit the restriction for SGD from [6, Theorem 4.7]. Further, we varied $\gamma = 10^m$ for $m \in \{0, \dots, 6\}$ to investigate how larger initial step sizes effect the errors. Note that in [6, Theorem 4.7] there is also a lower bound for γ. This restriction cannot be stated easily as it depends for example on the Lipschitz constant of ∇f_n. The optimal rate can be observed for γ large enough in the SVM examples.

In Figure 1, we observe very well how larger initial step sizes change the outcome. For the TSGD method, we see how the error decreases while decreasing γ within $\{10^3, 10^4, 10^5, 10^6\}$ and thereby increasing the initial step size. When γ is chosen within $\{1, 10, 10^2, 10^3\}$ the error stops decreasing but remains within the same ballpark. This behavior cannot be observed for SGD. While increasing the initial step size has a positive effect for γ between $\{10^4, 10^5, 10^6\}$, it leads to large errors within the first few steps for $\gamma = 10^m$ for $m \in \{0, 1, 2, 3\}$ that can no longer be compensated for at later points. We note that we observe a faster asymptotic convergence for TSGD than suggested by our bounds (although we acknowledge that choosing a representative reference curve is a non-trivial task, given the number of different results in the plots). A possible explanation could be that the error in Theorems 5.3, 5.5 and 5.6 consists of two parts where the exponent in the second summand cannot be smaller than -1. In our case it could be the first error part that is dominating the total error. Here, the error can decrease faster than $\sim n^{-1}$ for large ϑ. The second part of the error corresponds to the question how well the operator $T_{\alpha f_n, z}(w) = w - \frac{\alpha \iota \nabla f_n(w)}{1+\alpha \|\iota \nabla f_n(z)\|}$ preserves the optimum w^*. In the deterministic case, i.e. $f_n = F$, it follows that $T_{\alpha F, z}(w^*) = w^*$. Thus, the stochastic approximation of F by f_n could be better than expected in our examples.

In Figure 2, we observe similar results for the second data set.

6.2. **Neural network.** We used a fully connected neural network with one hidden layer containing 100 neurons. The activation function was $f(x) = \max\{0, x\}$ on the hidden layers and $f(x) = 1/(1 + \exp(-x))$ on the output layer. The regularization parameter was again $\lambda = 10^{-5}$. We allowed a maximum of 10 epochs, and used a batch size which was 1% of the amount of samples. We stored every tenth value $F(w^n)$ in order to save computational costs. For the steps size sequence $\{\alpha_n\}_{n \in \mathbb{N}}$

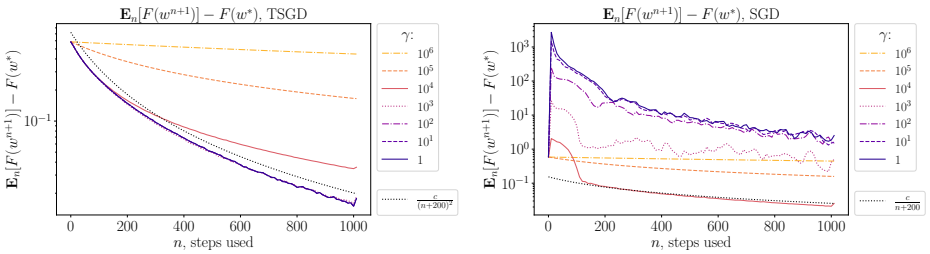

FIGURE 2. TSGD (left) versus SGD (right), different step sizes, SVM, rcv1.binary

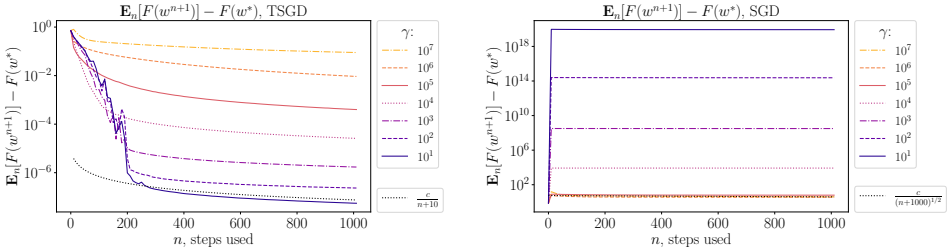

FIGURE 3. TSGD (left) versus SGD (right), different step sizes, neural network, mushrooms

with $\alpha_n = \frac{\vartheta}{n+\gamma}$, $n \in \mathbb{N}$, we chose $\vartheta = 10^5 = \lambda^{-1}$ and we varied $\gamma = 10^m$ for $m \in \{1, \dots, 7\}$.

The positive effects of TSGD are showing even more clearly in this example. In Figure 3, we observe that for growing initial step sizes TSGD improves, while SGD becomes worse. For the second example in Figure 4 we still observe that TSGD is much more stable than SGD even though the best result is achieved with $\gamma = 10^4$ and it becomes worse after. Compared to SGD, the speed of convergence is faster.

Altogether, we note that *if* we choose the initial step size optimally, we do achieve the optimal rate also for SGD. This is, however, difficult to do in a real large-scale application, and the method is very sensitive to this choice. In contrast, TSGD performs similarly well for many different parameter choices, and is thus not sensitive at all. Further, in these examples, the TSGD decay is usually also faster than the best SGD decay. Using a different step size sequence for SGD which decreases faster initially and slower later might change this result, but it is unclear how to choose this optimally. Providing such an automatically tuned step size sequence is also, in fact, essentially what TSGD does.

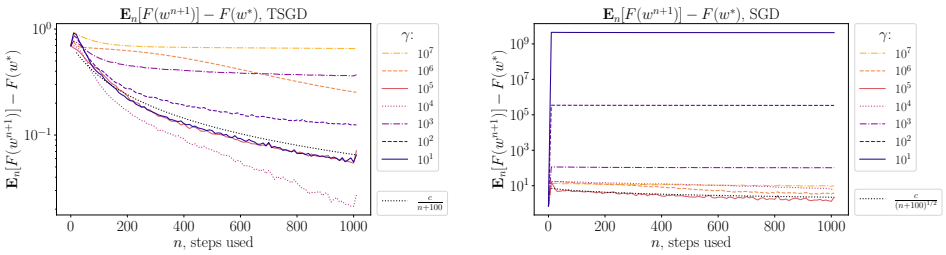

FIGURE 4. TSGD (left) versus SGD (right), different step sizes, neural network, rcv1.binary

7. Conclusions

We have introduced the TSGD method as an alternative to the well-known SGD method. While being comparably inexpensive, TSGD still offers better stability properties in comparison to this standard method. We have provided a general convergence analysis in an infinite dimensional framework for TSGD. While the infinite dimensional setting ensures that the error constants are independent of the underlying dimension of the problem, our analysis also shows that they are only mildly affected by large step sizes. This is in contrast to SGD, where large step sizes can lead to extremely large error constants. In practice, this means that larger step sizes can be used for TSGD which may lead to fast convergence results. We have also observed that TSGD is much less sensitive to the choice of parameters, in that similar convergence behaviour is often achieved for very different initial step sizes.

The advantages of TSGD were demonstrated in a numerical experiment involving a classification problem. We applied both an affine classifier (SVM) and a nonlinear classifier (neural network). The affine setting fits into our theory and illustrated the theoretical results, while the good performance in the nonlinear framework suggested that there is a wider range of applications of the TSGD scheme than those covered by our assumptions.

Appendix A. Auxiliary results

This section contains three results that are required for our main theory, but which are more generally applicable. The first lemma provides the main algebraic inequalities which we base our convergence analysis on:

Lemma A.1. *Let $x, y \in (0,\infty)$ and $n, m \in \mathbb{N}$ be given such that $\frac{x}{1+y} \leq 1$. Then the following inequalities are satisfied:*

(i) $\prod_{i=m}^{n} \left(1 - \frac{x}{i+y}\right) \leq \left(\frac{n+1+y}{m+y}\right)^{-x}$,

(ii) $\sum_{i=1}^{n} \frac{1}{(i+y)^2} \prod_{j=i+1}^{n} \left(1 - \frac{x}{j+y}\right)$

$$\leq \exp\left(\tfrac{x}{1+y}\right) \begin{cases} (n+1+y)^{-1} \frac{1}{x-1}, & x \in (1,\infty), \\ (n+1+y)^{-1}\big(1 + \ln(n+y)\big), & x = 1, \\ (n+1+y)^{-x} \frac{(1+y)^{x-2}(x-2-y)}{x-1}, & x \in [0,1). \end{cases}$$

Proof. In this proof, we apply the following basic inequalities involving (generalized) harmonic numbers

$$\sum_{i=m}^{n}(i+y)^{-1} \geq \ln(n+1+y) - \ln(m+y), \quad m \in \{1,\dots,n\},$$

$$\sum_{i=1}^{n}(i+y)^{p} \leq \begin{cases} \frac{(n+1+y)^{p+1}}{p+1}, & p \in [0,\infty), \\ \frac{(n+y)^{p+1}}{p+1}, & p \in (-1,0), \\ 1+\ln(n+y), & p=-1, \\ \frac{(1+y)^{p}(p-y)}{p+1}, & p \in (-\infty,-1), \end{cases}$$

for $y \in (0,\infty)$. These inequalities follow by treating the sums as a lower or upper Riemann sums approximating the integral $\int (u+y)^p \,\mathrm{d}u$ over the intervals $[0,n]$, $[1,n]$ or $[0,n+1]$.

Using the inequality $1+u \leq \mathrm{e}^u$ for $u \in [-1,\infty)$, it follows that $0 \leq 1-\frac{x}{i+y} \leq \exp(-\frac{x}{i+y})$ is fulfilled for every $i \in \mathbb{N}$ since $\frac{x}{1+y} \leq 1$. It then follows that

$$\begin{aligned}\prod_{i=m}^{n}\Big(1-\frac{x}{i+y}\Big) &\leq \exp\Big(-x\sum_{i=m}^{n}(i+y)^{-1}\Big) \\ &\leq \exp\big(-x\big(\ln(n+1+y)-\ln(m+y)\big)\big) \\ &= \exp\Big(-x\ln\Big(\frac{n+1+y}{m+y}\Big)\Big) = \Big(\frac{n+1+y}{m+y}\Big)^{-x}\end{aligned}$$

from which the first claim follows directly. For the second claim, we use the fact that $\frac{i+1+y}{i+y} = 1+\frac{1}{i+y} \leq 1+\frac{1}{1+y} \leq \exp(\frac{1}{1+y})$ for all $i \in \mathbb{N}$ and find that

$$\begin{aligned}&\sum_{i=1}^{n}\frac{1}{(i+y)^2}\prod_{j=i+1}^{n}\Big(1-\frac{x}{j+y}\Big) \leq \sum_{i=1}^{n}\frac{1}{(i+y)^2}\Big(\frac{n+1+y}{i+1+y}\Big)^{-x} \\ &\leq (n+1+y)^{-x}\sum_{i=1}^{n}\Big(\frac{i+1+y}{i+y}\Big)^{x}(i+y)^{x-2} \\ &\leq \exp\Big(\frac{x}{1+y}\Big)(n+1+y)^{-x}\sum_{i=1}^{n}(i+y)^{x-2} \\ &\leq \exp\Big(\frac{x}{1+y}\Big)\begin{cases}(n+1+y)^{-1}\frac{1}{x-1}, & x \in (1,\infty), \\ (n+1+y)^{-1}\big(1+\ln(n+y)\big), & x=1, \\ (n+1+y)^{-x}\frac{(1+y)^{x-2}(x-2-y)}{x-1}, & x \in [0,1),\end{cases}\end{aligned}$$

where we applied the basic inequalities from the beginning of the proof. □

Lemma A.2. *Given $a,b \in (0,\infty)$, $\frac{-1}{ax+b} \leq -\frac{1}{b}+\frac{a}{b^2}x$ for every $x \in (0,\infty)$.*

Proof. We consider the function $f\colon [0,\infty) \to \mathbb{R}$ with $f(x) = \frac{-1}{ax+b}$. Then the first and second derivative of f are given by $f'(x) = \frac{a}{(ax+b)^2}$ and $f''(x) = \frac{-2a^2}{(ax+b)^3}$. Using a first-order Taylor expansion of f then shows that

$$f(x) = -\frac{1}{b}+\frac{a}{b^2}x - \frac{a^2}{(a\xi+b)^3}x^2 \leq -\frac{1}{b}+\frac{a}{b^2}x,$$

where $\xi \in (0,x)$. □

The final lemma shows that ∇F does in fact exist and equals $\mathbf{E}_\xi\big[\nabla f(\xi,\cdot)\big]$. The proof is essentially the same as in [28, Lemma 6] for the finite-dimensional case but

we include it for completeness. Similar results can be proved also in more general setting, see e.g. [24].

Lemma A.3. *Let Assumption 1 be fulfilled. Then* $F = \mathbf{E}_\xi[f(\xi,\cdot)]$ *is Gâteaux differentiable and its derivative is given by*

$$\langle \iota\nabla F(v), w\rangle = \mathbf{E}_\xi[\langle \iota\nabla f(\xi, v), w\rangle] \quad v, w \in H.$$

Proof. For $v, w \in H$, we see that by the definition of the Gâteaux derivative,

$$\begin{aligned}
\langle \iota\nabla F(v), w\rangle &= \lim_{h\to 0} \frac{F(v+hw) - F(v)}{h} \\
&= \lim_{h\to 0} \frac{\mathbf{E}_\xi[f(\xi, v+hw)] - \mathbf{E}[f(\xi, v)]}{h} \\
&= \lim_{h\to 0} \mathbf{E}_\xi\Big[\frac{f(\xi, v+hw) - f(\xi, v)}{h}\Big] \\
&= \lim_{h\to 0} \mathbf{E}_\xi\Big[\frac{1}{h}\int_0^h \langle \iota\nabla f(\xi, v+sw), w\rangle \,\mathrm{d}s\Big].
\end{aligned}$$

In order to apply the dominated convergence theorem, we bound the integral as

$$\begin{aligned}
&\Big|\frac{1}{h}\int_0^h \langle \iota\nabla f(\xi, v+sw), w\rangle \,\mathrm{d}s\Big| \\
&\quad\le \frac{1}{h}\int_0^h \|\iota\nabla f(\xi, v+sw)\|\|w\| \,\mathrm{d}s \\
&\quad\le \frac{1}{h}\int_0^h \big(\|\iota\nabla f(\xi, v+sw) - \iota\nabla f(\xi, w^*)\| + \|\iota\nabla f(\xi, w^*)\|\big)\|w\| \,\mathrm{d}s \\
&\quad\le \frac{1}{h}\int_0^h \big(L_\xi\|v+sw-w^*\| + \|\iota\nabla f(\xi, w^*)\|\big)\|w\| \,\mathrm{d}s \\
&\quad\le \sup_{s\in(0,h)} L_\xi\|v+sw-w^*\|\|w\| + \|\iota\nabla f(\xi, w^*)\|\|w\| \\
&\quad\le L_\xi h\|w\|^2 + L_\xi\|v-w^*\|\|w\| + \|\iota\nabla f(\xi, w^*)\|\|w\|,
\end{aligned}$$

where the last term is integrable on Ω. This implies that

$$\begin{aligned}
\langle \iota\nabla F(v), w\rangle &= \lim_{h\to 0} \mathbf{E}_\xi\Big[\frac{1}{h}\int_0^h \langle \iota\nabla f(\xi, v+sw), w\rangle \,\mathrm{d}s\Big] \\
&= \mathbf{E}_\xi\Big[\lim_{h\to 0}\frac{1}{h}\int_0^h \langle \iota\nabla f(\xi, v+sw), w\rangle \,\mathrm{d}s\Big] = \mathbf{E}_\xi[\langle \iota\nabla f(\xi, v), w\rangle].
\end{aligned}$$

□

References

[1] A. Abdulle and A. Medovikov, *Second order Chebyshev methods based on orthogonal polynomials*, Numerische Mathematik, 90 (2001), pp. 1–18.

[2] V. Barbu, *Nonlinear Differential Equations of Monotone Types in Banach Spaces*, Springer Monographs in Mathematics, Springer, New York, 2010, https://doi.org/10.1007/978-1-4419-5542-5, https://doi.org/10.1007/978-1-4419-5542-5.

[3] D. Bertsekas, *Incremental proximal methods for large scale convex optimization*, Math. Program., 129 (2011), pp. 163–195.

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

[5] P. Bianchi and W. Hachem, *Dynamical behavior of a stochastic forward-backward algorithm using random monotone operators*, J. Optim. Theory Appl., 171 (2016), pp. 90–120.

[6] L. Bottou, F. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, SIAM Rev., 60 (2018), pp. 223–311.
[7] N. Brosse, A. Durmus, E. Moulines, and S. Sabanis, *The tamed unadjusted Langevin algorithm*, Stochastic Process. Appl., 129 (2019), pp. 3638–3663.
[8] D. Davis and D. Drusvyatskiy, *Stochastic model-based minimization of weakly convex functions*, SIAM J. Optim., 29 (2019), pp. 207–239.
[9] D. Dua and C. Graff, *UCI machine learning repository*, 2017, http://archive.ics.uci.edu/ml.
[10] A. Eftekhari, B. Vandereycken, G. Vilmart, and K. Zygalakis, *Explicit stabilised gradient descent for faster strongly convex optimisation*, BIT Numerical Mathematics, 61 (2021), pp. 119–139.
[11] M. Eisenmann, T. Stillfjord, and M. Williamson, *Sub-linear convergence of a stochastic proximal iteration method in Hilbert space*, ArXiv Preprint, arXiv:2010.12348, (2020).
[12] E. Hairer and G. Wanner, *Solving ordinary differential equations. II*, vol. 14 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 2010. Stiff and differential-algebraic problems, Second revised edition, paperback.
[13] C. Higham and D. Higham, *Deep learning: An introduction for applied mathematicians*, SIAM Rev., 61 (2019), pp. 860–891.
[14] W. Hundsdorfer and J. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer, Berlin, Heidelberg, 2003.
[15] M. Hutzenthaler and A. Jentzen, *Numerical approximations of stochastic differential equations with non-globally Lipschitz continuous coefficients*, Mem. Amer. Math. Soc., 236 (2015), pp. v+99, https://doi.org/10.1090/memo/1112, https://doi.org/10.1090/memo/1112.
[16] M. Hutzenthaler and A. Jentzen, *On a perturbation theory and on strong convergence rates for stochastic ordinary and partial differential equations with nonglobally monotone coefficients*, Ann. Probab., 48 (2020), pp. 53–93, https://doi.org/10.1214/19-AOP1345, https://doi.org/10.1214/19-AOP1345.
[17] M. Hutzenthaler, A. Jentzen, and P. Kloeden, *Strong and weak divergence in finite time of Euler's method for stochastic differential equations with non-globally Lipschitz continuous coefficients*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 467 (2011), pp. 1563–1576, https://doi.org/10.1098/rspa.2010.0348, https://doi.org/10.1098/rspa.2010.0348.
[18] M. Hutzenthaler, A. Jentzen, and P. Kloeden, *Strong convergence of an explicit numerical method for SDEs with nonglobally Lipschitz continuous coefficients*, Ann. Appl. Probab., 22 (2012), pp. 1611–1641.
[19] M. Hutzenthaler, A. Jentzen, and P. Kloeden, *Divergence of the multilevel Monte Carlo Euler method for nonlinear stochastic differential equations*, Ann. Appl. Probab., 23 (2013), pp. 1913–1966.
[20] D. Kingma and J. Ba, *Adam: A method for stochastic optimization*, arXiv e-prints, (2014), arXiv:1412.6980, pp. 1–15. Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.
[21] D. Lewis, Y. Yang, T. Rose, and F. Li, *Rcv1: A new benchmark collection for text categorization research*, J. Mach. Learn. Res., 5 (2004), p. 361–397.
[22] A. Lovas, I. Lytras, M. Rásonyi, and S. Sabanis, *Taming neural networks with TUSLA: Non-convex learning via adaptive stochastic gradient Langevin algorithms*, ArXiv Preprint, arXiv:2006.14514, (2020).
[23] Owens and 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, https://doi.org/10.1109/IJCNN.1989.118726.
[24] N. S. Papageorgiou, *Convex integral functionals*, Trans. Amer. Math. Soc., 349 (1997), pp. 1421–1436, https://doi.org/10.1090/S0002-9947-97-01478-5.
[25] A. Patrascu and P. Irofti, *Stochastic proximal splitting algorithm for composite minimization*, ArXiv Preprint, arXiv:1912.02039v2, (2020).
[26] A. Patrascu and I. Necoara, *Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization*, J. Mach. Learn. Res., 18 (2017), pp. Paper No. 198, 42.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, 12 (2011), pp. 2825–2830.

[28] E. Ryu and S. Boyd, *Stochastic proximal iteration: A non-asymptotic improvement upon stochastic gradient descent*, https://stanford.edu/~boyd/papers/pdf/spi.pdf, (2016).

[29] E. Ryu and W. Yin, *Proximal-proximal-gradient method*, J. Comput. Math., 37 (2019), pp. 778–812.

[30] S. Sabanis, *A note on tamed Euler approximations*, Electron. Commun. Probab., 18 (2013), pp. no. 47, 10.

[31] S. Sabanis, *Euler approximations with varying coefficients: the case of superlinearly growing diffusion coefficients*, Ann. Appl. Probab., 26 (2016), pp. 2083–2105.

[32] S. Sabanis and Y. Zhang, *Higher order Langevin Monte Carlo algorithm*, Electron. J. Stat., 13 (2019), pp. 3805–3850.

[33] A. Salim, P. Bianchi, and W. Hachem, *Snake: a stochastic proximal gradient algorithm for regularized problems over large graphs*, IEEE Trans. Automat. Control, 64 (2019), pp. 1832–1847.

[34] P. Toulis and E. Airoldi, *Scalable estimation strategies based on stochastic approximations: classical results and new insights*, Stat. Comput., 25 (2015), pp. 781–795.

[35] P. Toulis and E. Airoldi, *Asymptotic and finite-sample properties of estimators based on stochastic gradients*, Ann. Statist., 45 (2017), pp. 1694–1727.

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

[37] P. Toulis, D. Tran, and E. Airoldi, *Towards stability and optimality in stochastic gradient descent*, in Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, A. Gretton and C. C. Robert, eds., vol. 51 of Proceedings of Machine Learning Research, Cadiz, Spain, 09–11 May 2016, PMLR, pp. 1290–1298.

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

[39] P. J. van Der Houwen and B. Sommeijer, *On the internal stability of explicit, m-stage Runge-Kutta methods for large m-values*, ZAMM, 60 (1980), pp. 479–485.

[40] J. Verwer, *Explicit Runge-Kutta methods for parabolic partial differential equations*, Appl. Numer. Math., 22 (1996), pp. 359–379. Special issue celebrating the centenary of Runge-Kutta methods.

[41] P. Virtanen, R. Gommers, T. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. van der Walt, M. Brett, J. Wilson, K. Millman, N. Mayorov, A. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. Quintero, C. Harris, A. Archibald, A. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python*, Nature Methods, 17 (2020), pp. 261–272.

[42] E. Zeidler, *Nonlinear Functional Analysis and its Applications. II/B*, Springer-Verlag, New York, 1990. Nonlinear monotone operators, Translated from the German by the author and Leo F. Boron.