Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Seminar for Applied Mathematics SAM

Lower error bounds for the stochastic gradient descent optimization algorithm: Sharp convergence rates for slowly and fast decaying learning rates

A. Jentzen and Ph. von Wurstemberger

Research Report No. 2018-12
April 2018

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

Lower error bounds for the stochastic gradient descent optimization algorithm: Sharp convergence rates for slowly and fast decaying learning rates

Arnulf Jentzen[1] and Philippe von Wurstemberger[2]

[1]Department of Mathematics, ETH Zurich,
e-mail: arnulf.jentzen@sam.math.ethz.ch

[2]Department of Mathematics, ETH Zurich,
e-mail: vwurstep@student.ethz.ch

April 6, 2018

Abstract

The stochastic gradient descent (SGD) optimization algorithm plays a central role in a series of machine learning applications. The scientific literature provides a vast amount of upper error bounds for the SGD method. Much less attention as been paid to proving lower error bounds for the SGD method. It is the key contribution of this paper to make a step in this direction. More precisely, in this article we establish for every $\gamma, \nu \in (0, \infty)$ essentially matching lower and upper bounds for the mean square error of the SGD process with learning rates $(\frac{\gamma}{n^\nu})_{n\in\mathbb{N}}$ associated to a simple quadratic stochastic optimization problem. This allows us to precisely quantify the mean square convergence rate of the SGD method in dependence on the asymptotic behavior of the learning rates.

Contents

1 Introduction

The stochastic gradient descent (SGD) optimization algorithm plays a central role in machine learning and, in particular, deep learning applications such as image analysis and speech recognition (cf., e.g., [12, 13, 16, 23]). It is therefore important to analyze and quantify the convergence speed of the SGD method. There is a vast amount of scientific literature investigating and providing upper bounds for the SGD method and modifications of it (cf., e.g., [3, 4, 5, 6, 7, 8, 9, 10, 11, 18, 20, 21, 24] and cf., e.g., [14] for a more comprehensive review of the literature). Much less attention has been paid to proving lower error bounds for the SGD method, that is, to quantifying the best possible speed of convergence which the SGD method can achieve (cf., e.g., [2, 17, 19, 22, 25]). It is the key contribution of this paper to make a step in this direction.

To be more specific, in this paper we precisely quantify the speed of convergence of the SGD process in the case of a simple quadratic stochastic optimization problem (cf. item (i) in Theorem 1.1 below) for both slowly as well as fast decaying learning rates. In particular, in Theorem 1.1 below we provide for every $\gamma, \nu \in (0, \infty)$ essentially matching upper and lower bounds for the root mean square distance between the global minimum of the considered stochastic optimization problem and the SGD process with the learning rates $(\frac{\gamma}{n^\nu})_{n\in\mathbb{N}}$.

Theorem 1.1. *Let $d \in \mathbb{N}$, $\alpha, \gamma, \nu \in (0, \infty)$, $\xi \in \mathbb{R}^d$, let $\langle \cdot, \cdot \rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the d-dimensional Euclidean scalar product, let $\|\cdot\| : \mathbb{R}^d \to [0, \infty)$ be the d-dimensional Euclidean norm, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X_n \colon \Omega \to \mathbb{R}^d$, $n \in \mathbb{N}$, be i.i.d. random variables with $\mathbb{E}[\|X_1\|^2] < \infty$ and $\mathbb{P}(X_1 = \mathbb{E}[X_1]) < 1$, let $(r_{\varepsilon,i})_{\varepsilon\in(0,\infty),i\in\{0,1\}} \subseteq \mathbb{R}$ satisfy for all $\varepsilon \in (0, \infty)$, $i \in \{0, 1\}$ that*

$$r_{\varepsilon,i} = \begin{cases} \nu/2 & : \nu < 1 \\ \min\{1/2, \gamma\alpha + (-1)^i\varepsilon\} & : \nu = 1 \\ 0 & : \nu > 1, \end{cases} \tag{1}$$

let $F = (F(\theta, x))_{(\theta,x)\in\mathbb{R}^d\times\mathbb{R}^d} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $f \colon \mathbb{R}^d \to \mathbb{R}$ be the functions which satisfy for all $\theta, x \in \mathbb{R}^d$ that

$$F(\theta, x) = \tfrac{\alpha}{2}\|\theta - x\|^2 \qquad \text{and} \qquad f(\theta) = \mathbb{E}\big[F(\theta, X_1)\big], \tag{2}$$

and let $\Theta \colon \mathbb{N}_0 \times \Omega \to \mathbb{R}^d$ be the stochastic process which satisfies for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \tfrac{\gamma}{n^\nu}(\nabla_\theta F)(\Theta_{n-1}, X_n). \tag{3}$$

Then

(i) *there exists a unique* $\vartheta \in \mathbb{R}^d$ *such that* $\{\theta \in \mathbb{R}^d \colon f(\theta) = \inf_{w \in \mathbb{R}^d} f(w)\} = \{\vartheta\}$,

(ii) *for every* $\varepsilon \in (0, \infty)$ *there exist* $c_0, c_1 \in (0, \infty)$ *such that for all* $n \in \mathbb{N}$ *it holds that*

$$c_0 n^{-r_{\varepsilon,0}} \leq \left(\mathbb{E}\left[\|\Theta_n - \vartheta\|^2\right]\right)^{1/2} \leq c_1 n^{-r_{\varepsilon,1}}, \tag{4}$$

and

(iii) *for every* $\varepsilon \in (0, \infty)$ *there exist* $C_0, C_1 \in (0, \infty)$ *such that for all* $n \in \mathbb{N}$ *it holds that*

$$C_0 n^{-2r_{\varepsilon,0}} \leq \mathbb{E}[f(\Theta_n)] - f(\vartheta) \leq C_1 n^{-2r_{\varepsilon,1}}. \tag{5}$$

Theorem 1.1 is an immediate consequence of Theorem 4.13 below, which is the main result of this article. We now roughly describe the dependence, exhibited in Theorem 1.1, of the root mean square convergence rate of the SGD process on the learning rates. In the case of slowly decaying learning rates (corresponding to the case $\nu < 1$ in Theorem 1.1) the convergence rate of the SGD process does not depend on the size of the learning rates (corresponding to the parameter γ in Theorem 1.1). In this case we note that faster decay of the learning rates (corresponding to larger ν in Theorem 1.1) results in a higher convergence rate of the SGD process. In the case of fast decaying but large learning rates (corresponding to the case $\nu = 1$ and $\gamma > 1/(2\alpha)$ in Theorem 1.1) the SGD process attains the optimal convergence rate of $1/2$. In the case of fast decaying and small learning rates (corresponding to the case $\nu = 1$ and $\gamma \leq 1/(2\alpha)$ in Theorem 1.1) the convergence rate of the SGD process depends on the size of the learning rates. In this case we observe that the smaller the learning rates are (corresponding to smaller γ in Theorem 1.1) the lower is the resulting convergence rate. Note that this is contrary to the effect observed above in the case of slowly decaying learning rates. The phenomenon that the convergence rate increases as ν increases (so that the learning rates get smaller) but also increases as γ increases (so that the learning rates get larger), roughly speaking, arises from the interplay of two sources of errors: The error due to the randomness in the SGD method (which gets smaller when the learning rates get smaller) and the error due to the fact that the deterministic gradient method does not reach in finite time the whole infinite time interval of the underlying gradient flow (which gets smaller when the learning rates get larger). Finally, in the case of very fast decaying learning rates (corresponding to the case $\nu > 1$ in Theorem 1.1) the SGD process fails to converge to the global minimum of the objective function.

The remainder of this paper is organized as follows. In Section 2 we introduce the setting of the stochastic optimization problem considered in this paper and we

establish a few basic properties for the objective function, the loss function, and the SGD process. In Section 3 we derive upper bounds for the root mean square error of the SGD process. In Section 4 we first establish in Subsections 4.1–4.3 lower bounds for the root mean square error of the SGD process which essentially match the upper bounds of Section 3. Then, in Subsection 4.4, we combine the upper and lower bounds of this article in Theorem 4.13 and thereby obtain a sharp convergence rate of the SGD process in dependence of the learning rates. Theorem 1.1 above is an immediate consequence of Theorem 4.13.

2 Basic properties for the stochastic gradient descent (SGD) optimization method

In Section 3 and Section 4 below we provide a detailed error analysis for the SGD optimization method in the case of a simple quadratic loss function; cf., particularly, Theorem 4.13 below. In this section we introduce the setting of the considered optimization problem (see Setting 2.1 in Subsection 2.1 below) and we establish some elementary properties for the optimization problem under consideration (see Lemma 2.4 below) and the associated SGD process (see Proposition 2.6 below). These elementary properties will be repeatedly used in the convergence rate proofs in our detailed error analysis in Section 3 and Section 4 below.

2.1 Setting

Throughout this article the following setting is frequently used.

Setting 2.1. *Let $d \in \mathbb{N}$, $\alpha, \gamma, \nu \in (0, \infty)$, $\xi \in \mathbb{R}^d$, let $\langle \cdot, \cdot \rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the d-dimensional Euclidean scalar product, let $\|\cdot\| : \mathbb{R}^d \to [0, \infty)$ be the d-dimensional Euclidean norm, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X_n \colon \Omega \to \mathbb{R}^d$, $n \in \mathbb{N}$, be i.i.d. random variables with $\mathbb{E}[\|X_1\|^2] < \infty$, let $F = (F(\theta, x))_{(\theta,x)\in\mathbb{R}^d\times\mathbb{R}^d} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $f \colon \mathbb{R}^d \to \mathbb{R}$ be the functions which satisfy for all $\theta, x \in \mathbb{R}^d$ that*

$$F(\theta, x) = \tfrac{\alpha}{2}\|\theta - x\|^2 \qquad \text{and} \qquad f(\theta) = \mathbb{E}\big[F(\theta, X_1)\big], \tag{6}$$

and let $\Theta \colon \mathbb{N}_0 \times \Omega \to \mathbb{R}^d$ be the stochastic process which satisfies for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \tfrac{\gamma}{n^\nu}(\nabla_\theta F)(\Theta_{n-1}, X_n). \tag{7}$$

2.2 Basic properties of the objective and the loss function

In this subsection we establish in Lemma 2.4 below some basic properties for the objective and the loss function of the optimization problem under consideration (cf. Setting 2.1 above). Our proof of Lemma 2.4 employs the elementary and well-known results in Lemma 2.2 and Lemma 2.3. For completeness we also provide the proofs of Lemma 2.2 and Lemma 2.3 here.

2.2.1 Bias-variance decomposition of the mean square error

Lemma 2.2. *Let $d \in \mathbb{N}$, $\vartheta \in \mathbb{R}^d$, let $\langle \cdot, \cdot \rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a scalar product, let $\|\cdot\| : \mathbb{R}^d \to [0, \infty)$ be the function which satisfies for all $v \in \mathbb{R}^d$ that $\|v\| = \sqrt{\langle v, v \rangle}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $Z \colon \Omega \to \mathbb{R}^d$ be a random variable with $\mathbb{E}[\|Z\|] < \infty$. Then*

$$\mathbb{E}\big[\|Z - \vartheta\|^2\big] = \mathbb{E}\big[\|Z - \mathbb{E}[Z]\|^2\big] + \|\mathbb{E}[Z] - \vartheta\|^2. \tag{8}$$

Proof of Lemma 2.2. Observe that the hypothesis that $\mathbb{E}[\|Z\|] < \infty$ and the Cauchy-Schwarz inequality ensure that

$$\begin{aligned} \mathbb{E}\big[|\langle Z - \mathbb{E}[Z], \mathbb{E}[Z] - \vartheta \rangle|\big] &\leq \mathbb{E}\big[\|Z - \mathbb{E}[Z]\| \|\mathbb{E}[Z] - \vartheta\|\big] \\ &\leq (\mathbb{E}[\|Z\|] + \|\mathbb{E}[Z]\|) \|\mathbb{E}[Z] - \vartheta\| < \infty. \end{aligned} \tag{9}$$

The linearity of the expectation hence shows that

$$\begin{aligned} &\mathbb{E}\big[\|Z - \vartheta\|^2\big] = \mathbb{E}\big[\|(Z - \mathbb{E}[Z]) + (\mathbb{E}[Z] - \vartheta)\|^2\big] \\ &= \mathbb{E}\big[\|Z - \mathbb{E}[Z]\|^2 + 2\langle Z - \mathbb{E}[Z], \mathbb{E}[Z] - \vartheta \rangle + \|\mathbb{E}[Z] - \vartheta\|^2\big] \\ &= \mathbb{E}\big[\|Z - \mathbb{E}[Z]\|^2\big] + 2\langle \mathbb{E}[Z] - \mathbb{E}[Z], \mathbb{E}[Z] - \vartheta \rangle + \|\mathbb{E}[Z] - \vartheta\|^2 \\ &= \mathbb{E}\big[\|Z - \mathbb{E}[Z]\|^2\big] + \|\mathbb{E}[Z] - \vartheta\|^2. \end{aligned} \tag{10}$$

The proof of Lemma 2.2 is thus completed. □

2.2.2 On the derivative of the Euclidean norm

Lemma 2.3 (Derivative of the Euclidean norm)**.** *Let $d \in \mathbb{N}$, $\vartheta \in \mathbb{R}^d$, let $\|\cdot\| : \mathbb{R}^d \to [0, \infty)$ be the d-dimensional Euclidean norm, and let $f \colon \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that*

$$f(\theta) = \|\theta - \vartheta\|^2. \tag{11}$$

Then it holds for all $\theta \in \mathbb{R}^d$ that $f \in C^\infty(\mathbb{R}^d, \mathbb{R})$ and

$$(\nabla f)(\theta) = 2(\theta - \vartheta). \tag{12}$$

Proof of Lemma 2.3. Throughout this proof let $\vartheta_1, \ldots, \vartheta_d \in \mathbb{R}$ satisfy that $\vartheta = (\vartheta_1, \ldots, \vartheta_d)$. Note that the fact that for all $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ it holds that

$$f(\theta) = \sum_{i=1}^{d} \big[\theta_i - \vartheta_i\big]^2 \tag{13}$$

implies that for all $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ it holds that $f \in C^\infty(\mathbb{R}^d, \mathbb{R})$ and

$$(\nabla f)(\theta) = \begin{pmatrix} (\frac{\partial f}{\partial \theta_1})(\theta) \\ \vdots \\ (\frac{\partial f}{\partial \theta_d})(\theta) \end{pmatrix} = \begin{pmatrix} 2(\theta_1 - \vartheta_1) \\ \vdots \\ 2(\theta_d - \vartheta_d) \end{pmatrix} = 2(\theta - \vartheta). \tag{14}$$

The proof of Lemma 2.3 is thus completed. □

2.2.3 Basic properties of the objective and the loss function

Lemma 2.4. *Assume Setting 2.1. Then*

(i) it holds for all $\theta \in \mathbb{R}^d$ *that* $f(\theta) = \frac{\alpha}{2}\|\theta - \mathbb{E}[X_1]\|^2 + \frac{\alpha}{2}\,\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big]$,

(ii) it holds that $\{\theta \in \mathbb{R}^d \colon f(\theta) = \inf_{w \in \mathbb{R}^d} f(w)\} = \{\mathbb{E}[X_1]\}$,

(iii) it holds for all $\theta, x \in \mathbb{R}^d$ *that* $(\nabla_\theta F)(\theta, x) = \alpha(\theta - x)$,

(iv) it holds for all $\theta \in \mathbb{R}^d$ *that* $(\nabla f)(\theta) = \mathbb{E}[(\nabla_\theta F)(\theta, X_1)] = \alpha(\theta - \mathbb{E}[X_1])$,

(v) it holds for all $\theta \in \mathbb{R}^d$ *that* $\langle \theta - \mathbb{E}[X_1], (\nabla f)(\theta)\rangle = \alpha\|\theta - \mathbb{E}[X_1]\|^2$,

(vi) it holds for all $\theta \in \mathbb{R}^d$ *that* $\|(\nabla f)(\theta)\| = \alpha\|\theta - \mathbb{E}[X_1]\|$, *and*

(vii) it holds for all $\theta \in \mathbb{R}^d$ *that*

$$\mathbb{E}\big[\|(\nabla_\theta F)(\theta, X_1) - (\nabla f)(\theta)\|^2\big] = \alpha^2\,\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big]. \tag{15}$$

Proof of Lemma 2.4. First, note that the hypothesis that $\mathbb{E}[\|X_1\|^2] < \infty$ and Lemma 2.2 (with $\vartheta = \theta$, $Z = X_1$ in the notation of Lemma 2.2) ensure that for all $\theta \in \mathbb{R}^d$ it holds that

$$\begin{aligned} f(\theta) &= \mathbb{E}\big[F(\theta, X_1)\big] = \tfrac{\alpha}{2}\,\mathbb{E}\big[\|X_1 - \theta\|^2\big] \\ &= \tfrac{\alpha}{2}\left(\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|\big]^2 + \|\theta - \mathbb{E}[X_1]\|^2\right). \end{aligned} \tag{16}$$

This establishes item (i). Next observe that item (i) proves item (ii). In addition, note that Lemma 2.3 proves that for all $\theta, x \in \mathbb{R}^d$ it holds that

$$(\nabla_\theta F)(\theta, x) = \tfrac{\alpha}{2}(2(\theta - x)) = \alpha(\theta - x). \tag{17}$$

This establishes item (iii). Moreover, observe that Lemma 2.3, item (i), and item (iii) ensure that for all $\theta \in \mathbb{R}^d$ it holds that

$$\begin{aligned}(\nabla f)(\theta) &= \tfrac{\alpha}{2}(2(\theta - \mathbb{E}[X_1])) = \alpha(\theta - \mathbb{E}[X_1]) \\ &= \mathbb{E}[\alpha(\theta - X_1)] = \mathbb{E}[(\nabla_\theta F)(\theta, X_1)]\,.\end{aligned} \tag{18}$$

This proves item (iv). Next note that item (iv) implies items (v)–(vi). Moreover, note that item (iii) and item (iv) demonstrate that for all $\theta \in \mathbb{R}^d$ it holds that

$$\begin{aligned}\mathbb{E}\big[\|(\nabla_\theta F)(\theta, X_1) - (\nabla f)(\theta)\|^2\big] &= \mathbb{E}\big[\|\alpha(\theta - X_1) - \alpha(\theta - \mathbb{E}[X_1])\|^2\big] \\ &= \alpha^2\,\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big].\end{aligned} \tag{19}$$

This establishes item (vii). The proof of Lemma 2.4 it thus completed. □

2.3 On explicit formulas for the SGD process

In this subsection we establish in Proposition 2.6 below a few explicit formulas for the SGD process in (7). Our proof of Proposition 2.6 employs the elementary and well-known result for affine recursions in Lemma 2.5 below. For completeness we also present the proof of Lemma 2.5 here.

2.3.1 On a recursive equality

Lemma 2.5. *Let $d \in \mathbb{N}$, $(\alpha_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}$, $(\beta_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}^d$, $(e_n)_{n\in\mathbb{N}_0} \subseteq \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}$ that*

$$e_n = \alpha_n e_{n-1} + \beta_n. \tag{20}$$

Then it holds for all $n \in \mathbb{N}_0$ that

$$e_n = \left[\prod_{l=1}^{n} \alpha_l\right] e_0 + \sum_{k=1}^{n} \left(\left[\prod_{l=k+1}^{n} \alpha_l\right] \beta_k\right). \tag{21}$$

Proof of Lemma 2.5. We prove (21) by induction on $n \in \mathbb{N}_0$. For the base case $n = 0$ observe that

$$\left[\prod_{l=1}^{0} \alpha_l\right] e_0 + \sum_{k=1}^{0} \left(\left[\prod_{l=k+1}^{0} \alpha_l\right] \beta_k\right) = 1 \cdot e_0 + 0 = e_0. \tag{22}$$

This establishes (21) in the case case $n = 0$. For the induction step $\mathbb{N}_0 \ni (n-1) \to n \in \mathbb{N}$ note that (20) implies that for all $n \in \mathbb{N}$ with $e_{n-1} = \left[\prod_{l=1}^{n-1} \alpha_l\right] e_0 + \sum_{k=1}^{n-1} \left(\left[\prod_{l=k+1}^{n-1} \alpha_l\right] \beta_k\right)$ it holds that

$$\begin{aligned} e_n &= \alpha_n e_{n-1} + \beta_n \\ &= \alpha_n \left(\left[\prod_{l=1}^{n-1} \alpha_l \right] e_0 + \sum_{k=1}^{n-1} \left(\left[\prod_{l=k+1}^{n-1} \alpha_l \right] \beta_k \right) \right) + \beta_n \\ &= \left[\prod_{l=1}^{n} \alpha_l \right] e_0 + \sum_{k=1}^{n-1} \left(\left[\prod_{l=k+1}^{n} \alpha_l \right] \beta_k \right) + \left[\prod_{l=n+1}^{n} \alpha_l \right] \beta_n \\ &= \left[\prod_{l=1}^{n} \alpha_l \right] e_0 + \sum_{k=1}^{n} \left(\left[\prod_{l=k+1}^{n} \alpha_l \right] \beta_k \right). \end{aligned} \tag{23}$$

Induction thus establishes (21). The proof of Lemma 2.5 is thus completed. □

2.3.2 Explicit formulas for the SGD process

Proposition 2.6. *Assume Setting 2.1. Then*

(i) it holds for all $n \in \mathbb{N}$ that $\Theta_n = (1 - \frac{\gamma\alpha}{n^\nu})\Theta_{n-1} + \frac{\gamma\alpha}{n^\nu} X_n$,

(ii) it holds for all $n \in \mathbb{N}_0$ that

$$\Theta_n = \left[\prod_{l=1}^{n} (1 - \tfrac{\gamma\alpha}{l^\nu}) \right] \xi + \sum_{k=1}^{n} \left(\tfrac{\gamma\alpha}{k^\nu} \left[\prod_{l=k+1}^{n} (1 - \tfrac{\gamma\alpha}{l^\nu}) \right] X_k \right), \tag{24}$$

(iii) it holds for all $n \in \mathbb{N}$ that

$$\begin{aligned} &\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] \\ &= (1 - \tfrac{\gamma\alpha}{n^\nu})^2 \, \mathbb{E}\big[\|\Theta_{n-1} - \mathbb{E}[X_1]\|^2\big] + (\tfrac{\gamma\alpha}{n^\nu})^2 \, \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big], \end{aligned} \tag{25}$$

and

(iv) it holds for all $n \in \mathbb{N}_0$ that

$$\begin{aligned} \infty > \ \mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] &= \left[\prod_{l=1}^{n} (1 - \tfrac{\gamma\alpha}{l^\nu}) \right]^2 \|\xi - \mathbb{E}[X_1]\|^2 \\ &\quad + \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] \left[\sum_{k=1}^{n} \left[\tfrac{\gamma\alpha}{k^\nu} \left(\prod_{l=k+1}^{n} (1 - \tfrac{\gamma\alpha}{l^\nu}) \right) \right]^2 \right]. \end{aligned} \tag{26}$$

Proof of Proposition 2.6. First of all, observe that item (iii) in Lemma 2.4 assures that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}\Theta_n &= \Theta_{n-1} - \tfrac{\gamma}{n^\nu}(\nabla_\theta F)(\Theta_{n-1}, X_n)\\ &= \Theta_{n-1} - \tfrac{\gamma\alpha}{n^\nu}(\Theta_{n-1} - X_n)\\ &= (1 - \tfrac{\gamma\alpha}{n^\nu})\Theta_{n-1} + \tfrac{\gamma\alpha}{n^\nu} X_n.\end{aligned} \tag{27}$$

This establishes item (i). Next note that Lemma 2.5 (with $d = d$, $(\alpha_n)_{n\in\mathbb{N}} = (1 - \frac{\gamma\alpha}{n^\nu})_{n\in\mathbb{N}}$, $(\beta_n)_{n\in\mathbb{N}} = (\frac{\gamma\alpha}{n^\nu} X_n(\omega))_{n\in\mathbb{N}}$, $(e_n)_{n\in\mathbb{N}_0} = (\Theta_n(\omega))_{n\in\mathbb{N}_0}$ for $\omega \in \Omega$ in the notation of Lemma 2.5) and item (i) demonstrate that for all $n \in \mathbb{N}_0$, $\omega \in \Omega$ it holds that

$$\begin{aligned}\Theta_n(\omega) &= \left[\prod_{l=1}^{n}(1 - \tfrac{\gamma\alpha}{l^\nu})\right]\Theta_0(\omega) + \sum_{k=1}^{n}\left(\left[\prod_{l=k+1}^{n}(1 - \tfrac{\gamma\alpha}{l^\nu})\right]\left(\tfrac{\gamma\alpha}{k^\nu} X_k(\omega)\right)\right)\\ &= \left[\prod_{l=1}^{n}(1 - \tfrac{\gamma\alpha}{l^\nu})\right]\xi + \sum_{k=1}^{n}\left(\tfrac{\gamma\alpha}{k^\nu}\left[\prod_{l=k+1}^{n}(1 - \tfrac{\gamma\alpha}{l^\nu})\right] X_k(\omega)\right).\end{aligned} \tag{28}$$

This proves item (ii). Furthermore, note that item (ii) and the fact that $\forall\, k \in \mathbb{N}\colon \mathbb{E}[\|X_k\|^2] = \mathbb{E}[\|X_1\|^2] < \infty$ assure that for all $n \in \mathbb{N}_0$ it holds that $\mathbb{E}[\|\Theta_n\|^2] < \infty$. This ensures that for all $n \in \mathbb{N}_0$ it holds that

$$\mathbb{E}[\|\Theta_n - \mathbb{E}[X_1]\|^2] < \infty. \tag{29}$$

The fact that $\forall\, k \in \mathbb{N}\colon \mathbb{E}[\|X_k\|^2] = \mathbb{E}[\|X_1\|^2] < \infty$ and item (i) therefore imply that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}&\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big]\\ &= \mathbb{E}\big[\|(1 - \tfrac{\gamma\alpha}{n^\nu})\Theta_{n-1} + \tfrac{\gamma\alpha}{n^\nu} X_n - \mathbb{E}[X_1]\|^2\big]\\ &= \mathbb{E}\big[\|(1 - \tfrac{\gamma\alpha}{n^\nu})(\Theta_{n-1} - \mathbb{E}[X_1]) + \tfrac{\gamma\alpha}{n^\nu}(X_n - \mathbb{E}[X_1])\|^2\big]\\ &= \mathbb{E}\Big[\|(1 - \tfrac{\gamma\alpha}{n^\nu})(\Theta_{n-1} - \mathbb{E}[X_1])\|^2\\ &\quad + 2\,\langle(1 - \tfrac{\gamma\alpha}{n^\nu})(\Theta_{n-1} - \mathbb{E}[X_1]), \tfrac{\gamma\alpha}{n^\nu}(X_n - \mathbb{E}[X_1])\rangle + \|\tfrac{\gamma\alpha}{n^\nu}(X_n - \mathbb{E}[X_1])\|^2\Big]\\ &= (1 - \tfrac{\gamma\alpha}{n^\nu})^2\,\mathbb{E}\big[\|\Theta_{n-1} - \mathbb{E}[X_1]\|^2\big] + (\tfrac{\gamma\alpha}{n^\nu})^2\,\mathbb{E}\big[\|X_n - \mathbb{E}[X_1]\|^2\big]\\ &\quad + 2(1 - \tfrac{\gamma\alpha}{n^\nu})(\tfrac{\gamma\alpha}{n^\nu})\mathbb{E}\big[\langle\Theta_{n-1} - \mathbb{E}[X_1], X_n - \mathbb{E}[X_1]\rangle\big].\end{aligned} \tag{30}$$

In addition, note that the fact that for all independent random variables $Y, Z\colon \Omega \to \mathbb{R}$ with $\mathbb{E}[|Y| + |Z|] < \infty$ it holds that $\mathbb{E}[|YZ|] < \infty$ and $\mathbb{E}[YZ] = \mathbb{E}[Y]\,\mathbb{E}[Z]$ (cf., e.g., Klenke [15, Theorem 5.4]), the fact that for all $n \in \mathbb{N}$ it holds that Θ_{n-1} and X_n are

independent, and the fact that for all $n \in \mathbb{N}$ it holds that $\mathbb{E}[\|\Theta_{n-1}\| + \|X_n\|] < \infty$ assure that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}\mathbb{E}\big[\langle \Theta_{n-1} - \mathbb{E}[X_1], X_n - \mathbb{E}[X_1]\rangle\big] &= \big\langle \mathbb{E}\big[\Theta_{n-1} - \mathbb{E}[X_1]\big], \mathbb{E}\big[X_1 - \mathbb{E}[X_1]\big]\big\rangle \\ &= \big\langle \mathbb{E}[\Theta_{n-1}] - \mathbb{E}[X_1], \mathbb{E}[X_1] - \mathbb{E}[X_1]\big\rangle = 0.\end{aligned} \tag{31}$$

This, (30), and the fact that $(X_n)_{n\in\mathbb{N}}$ are i.i.d random variables demonstrate that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}&\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] \\ &= (1 - \tfrac{\gamma\alpha}{n^\nu})^2\, \mathbb{E}\big[\|\Theta_{n-1} - \mathbb{E}[X_1]\|^2\big] + (\tfrac{\gamma\alpha}{n^\nu})^2\, \mathbb{E}\big[\|X_n - \mathbb{E}[X_1]\|^2\big] \\ &= (1 - \tfrac{\gamma\alpha}{n^\nu})^2\, \mathbb{E}\big[\|\Theta_{n-1} - \mathbb{E}[X_1]\|^2\big] + (\tfrac{\gamma\alpha}{n^\nu})^2\, \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big]\,.\end{aligned} \tag{32}$$

This proves item (iii). Combining Lemma 2.5 (with $d = 1$, $(\alpha_n)_{n\in\mathbb{N}} = ((1 - \frac{\gamma\alpha}{n^\nu})^2)_{n\in\mathbb{N}}$, $(\beta_n)_{n\in\mathbb{N}} = ((\frac{\gamma\alpha}{n^\nu})^2\, \mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|^2])_{n\in\mathbb{N}}$, $(e_n)_{n\in\mathbb{N}_0} = (\mathbb{E}[\|\Theta_n - \mathbb{E}[X_1]\|^2])_{n\in\mathbb{N}_0}$ in the notation of Lemma 2.5) with item (iii) and (29) demonstrates that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{aligned}\infty > \ &\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] \\ &= \left[\prod_{l=1}^{n} (1 - \tfrac{\gamma\alpha}{l^\nu})^2\right] \mathbb{E}\big[\|\Theta_0 - \mathbb{E}[X_1]\|^2\big] \\ &\quad + \sum_{k=1}^{n} \left[\left(\prod_{l=k+1}^{n} (1 - \tfrac{\gamma\alpha}{l^\nu})^2\right) (\tfrac{\gamma\alpha}{k^\nu})^2\, \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big]\right] \\ &= \left[\prod_{l=1}^{n} (1 - \tfrac{\gamma\alpha}{l^\nu})\right]^2 \|\xi - \mathbb{E}[X_1]\|^2 \\ &\quad + \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] \left[\sum_{k=1}^{n} \left[\tfrac{\gamma\alpha}{k^\nu} \left(\prod_{l=k+1}^{n} (1 - \tfrac{\gamma\alpha}{l^\nu})\right)\right]^2\right].\end{aligned} \tag{33}$$

This establishes item (iv). The proof of Proposition 2.6 it thus completed. □

3 Upper error estimates for the SGD optimization method

In this section we establish in Proposition 3.3 and Corollary 3.7 below upper bounds for the root mean square distance between the SGD process in (7) and the global

minimum of the considered optimization problem (cf. item (ii) in Lemma 2.4). In our analysis we distinguish between the case of slowly decaying learning rates (see Subsection 3.1 below), the case of fast decaying learning rates (see Subsection 3.2 and Subsection 3.3 below), and the case of very fast decaying learning rates (see Subsection 3.4 below).

3.1 Upper errors estimates in the case of slowly decaying learning rates

In this subsection we establish in Proposition 3.3 below an upper bound for the root mean square error of the SGD process in (7) in the case of slowly decaying learning rates (corresponding to the case $\nu < 1$ in Setting 2.1). In our proof of Proposition 3.3 we employ the auxiliary and elementary results in Lemma 3.1 and Lemma 3.2 below. A result similar to Lemma 3.1 can, e.g., be found in [14, Corollary 2.18] and a result similar to Lemma 3.2 can, e.g., be found in [14, Lemma 4.1].

3.1.1 On a recursive inequality and an a priori estimate

Lemma 3.1. *Let $\kappa \in [0, \infty)$, $(e_n)_{n\in\mathbb{N}_0} \subseteq [0, \infty)$, $(\gamma_n)_{n\in\mathbb{N}} \subseteq (0, \infty)$ satisfy for all $n \in \mathbb{N}$ that*

$$e_n \leq (1 - \gamma_n)^2 e_{n-1} + \kappa(\gamma_n)^2 \qquad \text{and} \tag{34}$$

$$\liminf_{l\to\infty} \left[\frac{\gamma_l - \gamma_{l-1}}{(\gamma_l)^2} + \frac{2\gamma_{l-1}}{\gamma_l} - \gamma_{l-1} \right] > 0. \tag{35}$$

Then there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}$ it holds that

$$e_n \leq C\gamma_n. \tag{36}$$

Proof of Lemma 3.1. Throughout this proof let $m \in \mathbb{N} \cap (1, \infty)$, $\mathcal{I} \in (0, \infty)$ satisfy that

$$\mathcal{I} = \inf_{l\in\mathbb{N}\cap(m,\infty)} \left[\frac{\gamma_l - \gamma_{l-1}}{(\gamma_l)^2} + \frac{2\gamma_{l-1}}{\gamma_l} - \gamma_{l-1} \right] > 0 \tag{37}$$

(cf. (35)) and let $C \in [0, \infty)$ be given by

$$C = \max\left\{ \frac{e_m}{\gamma_m}, \frac{\kappa}{\mathcal{I}} \right\}. \tag{38}$$

We claim that for all $n \in \{m, m + 1, \ldots\}$ it holds that

$$e_n \leq C\gamma_n. \tag{39}$$

We now prove (39) by induction on $n \in \{m, m+1, \ldots\}$. For the base case $n = m$ note that

$$e_m = \left[\frac{e_m}{\gamma_m}\right] \gamma_m \leq C\gamma_m. \tag{40}$$

This establishes (39) in the base case $n = m$. For the induction step $\{m, m+1, \ldots\} \ni (n-1) \to n \in \mathbb{N} \cap (m, \infty)$ note that (34) assures that for all $n \in \mathbb{N} \cap (m, \infty)$ with $e_{n-1} \leq C\gamma_{n-1}$ it holds that

$$\begin{aligned} e_n &\leq (1-\gamma_n)^2 e_{n-1} + \kappa(\gamma_n)^2 \\ &\leq (1 - 2\gamma_n + (\gamma_n)^2) C\gamma_{n-1} + \kappa(\gamma_n)^2 - C\gamma_n + C\gamma_n \\ &= C\big(\gamma_{n-1} - 2\gamma_n\gamma_{n-1} + (\gamma_n)^2\gamma_{n-1} - \gamma_n\big) + \kappa(\gamma_n)^2 + C\gamma_n \\ &= (\gamma_n)^2 \left[C\left(\frac{\gamma_{n-1} - \gamma_n}{(\gamma_n)^2} - \frac{2\gamma_{n-1}}{\gamma_n} + \gamma_{n-1} \right) + \kappa \right] + C\gamma_n \\ &= C\gamma_n - (\gamma_n)^2 \left[C\left(\frac{\gamma_n - \gamma_{n-1}}{(\gamma_n)^2} + \frac{2\gamma_{n-1}}{\gamma_n} - \gamma_{n-1} \right) - \kappa \right]. \end{aligned} \tag{41}$$

This, (37), and the fact that $C \geq \frac{\kappa}{\mathcal{I}}$ demonstrate that for all $n \in \mathbb{N} \cap (m, \infty)$ with $e_{n-1} \leq C\gamma_{n-1}$ it holds that

$$\begin{aligned} e_n &\leq C\gamma_n - (\gamma_n)^2 \left[C\mathcal{I} - \kappa \right] \\ &\leq C\gamma_n - (\gamma_n)^2 \left[\tfrac{\kappa}{\mathcal{I}} \mathcal{I} - \kappa \right] \\ &= C\gamma_n - (\gamma_n)^2 \left[\kappa - \kappa \right] \\ &= C\gamma_n. \end{aligned} \tag{42}$$

Induction thus establishes (39). Next observe that (39) implies that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} e_n &\leq \left[\max\left\{ \frac{e_1}{\gamma_1}, \frac{e_2}{\gamma_2}, \ldots, \frac{e_{m-1}}{\gamma_{m-1}}, C \right\} \right] \gamma_n \\ &= \left[\max\left\{ \frac{e_1}{\gamma_1}, \frac{e_2}{\gamma_2}, \ldots, \frac{e_m}{\gamma_m}, \frac{\kappa}{\mathcal{I}} \right\} \right] \gamma_n. \end{aligned} \tag{43}$$

This completes the proof of Lemma 3.1. □

3.1.2 On an asymptotic property of the learning rates

Lemma 3.2. *Let $\beta \in (0, \infty)$, $\nu \in (0, 1)$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq (0, \infty)$ satisfy for all $n \in \mathbb{N}$ that $\gamma_n = \beta n^{-\nu}$. Then*

$$\liminf_{l \to \infty} \left[\frac{\gamma_l - \gamma_{l-1}}{(\gamma_l)^2} + \frac{2\gamma_{l-1}}{\gamma_l} - \gamma_{l-1} \right] \geq 2 > 0. \tag{44}$$

Proof of Lemma 3.2. First, note that for all $l \in \{2, 3, \ldots\}$ it holds that

$$\begin{aligned}\gamma_l - \gamma_{l-1} &= \beta\left(l^{-\nu} - (l-1)^{-\nu}\right) = \beta\left[x^{-\nu}\right]_{x=l-1}^{x=l} = \beta\left[\int_{l-1}^{l} (-\nu)x^{-\nu-1}\,\mathrm{d}x\right] \\ &= -\beta\nu \underbrace{\left[\int_{l-1}^{l} \frac{1}{x^{1+\nu}}\,\mathrm{d}x\right]}_{\leq \frac{1}{(l-1)^{1+\nu}}} \geq \frac{-\beta\nu}{(l-1)^{1+\nu}}.\end{aligned} \tag{45}$$

The fact that $1 + \nu - 2\nu = 1 - \nu > 0$ hence demonstrates that

$$\begin{aligned}\liminf_{l\to\infty}\left[\frac{\gamma_l - \gamma_{l-1}}{(\gamma_l)^2}\right] &\geq \liminf_{l\to\infty}\left[\frac{\left(\frac{-\beta\nu}{(l-1)^{1+\nu}}\right)}{\left(\frac{\beta}{l^\nu}\right)^2}\right] = -\tfrac{\nu}{\beta}\limsup_{l\to\infty}\left[\tfrac{l^{2\nu}}{(l-1)^{1+\nu}}\right] \\ &= -\tfrac{\nu}{\beta}\limsup_{l\to\infty}\left[\tfrac{1}{(l-1)^{1+\nu-2\nu}}\left(\tfrac{l}{l-1}\right)^{2\nu}\right] = 0.\end{aligned} \tag{46}$$

Therefore, we obtain that

$$\begin{aligned}&\liminf_{l\to\infty}\left[\frac{\gamma_l - \gamma_{l-1}}{(\gamma_l)^2} + \frac{2\gamma_{l-1}}{\gamma_l} - \gamma_{l-1}\right] \\ &\geq \liminf_{l\to\infty}\left[\frac{\gamma_l - \gamma_{l-1}}{(\gamma_l)^2}\right] + 2\liminf_{l\to\infty}\left[\frac{\gamma_{l-1}}{\gamma_l}\right] - \limsup_{l\to\infty}\gamma_{l-1} \\ &\geq 2\liminf_{l\to\infty}\left[\frac{\left(\frac{\beta}{(l-1)^\nu}\right)}{\left(\frac{\beta}{l^\nu}\right)}\right] - \limsup_{l\to\infty}\left[\tfrac{\beta}{(l-1)^\nu}\right] \\ &= 2\liminf_{l\to\infty}\left[\left(\tfrac{l}{l-1}\right)^\nu\right] = 2.\end{aligned} \tag{47}$$

The proof of Lemma 3.2 is thus completed. □

3.1.3 Upper error estimates

Proposition 3.3. *Assume Setting 2.1 and assume that $\nu < 1$. Then there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}$ it holds that*

$$\left(\mathbb{E}\left[\|\Theta_n - \mathbb{E}[X_1]\|^2\right]\right)^{1/2} \leq Cn^{-\nu/2}. \tag{48}$$

Proof of Proposition 3.3. Note that items (iii)–(iv) in Proposition 2.6 assure that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}&\mathbb{E}\left[\|\Theta_n - \mathbb{E}[X_1]\|^2\right] \\ &= (1 - \tfrac{\gamma\alpha}{n^\nu})^2\,\mathbb{E}\left[\|\Theta_{n-1} - \mathbb{E}[X_1]\|^2\right] + (\tfrac{\gamma\alpha}{n^\nu})^2\,\mathbb{E}\left[\|X_1 - \mathbb{E}[X_1]\|^2\right] < \infty.\end{aligned} \tag{49}$$

Moreover, observe that Lemma 3.2 (with $\beta = \gamma\alpha$, $\nu = \nu$ in the notation of Lemma 3.2) ensures that

$$\liminf_{l\to\infty}\left[\frac{\left(\frac{\gamma\alpha}{l^\nu}\right)-\left(\frac{\gamma\alpha}{(l-1)^\nu}\right)}{\left(\frac{\gamma\alpha}{l^\nu}\right)^2}+\frac{2\left(\frac{\gamma\alpha}{(l-1)^\nu}\right)}{\left(\frac{\gamma\alpha}{l^\nu}\right)}-\left(\frac{\gamma\alpha}{(l-1)^\nu}\right)\right]>0. \tag{50}$$

Combining this, (49) and Lemma 3.1 (with $\kappa = \mathbb{E}[\|X_1-\mathbb{E}[X_1]\|^2]$, $(e_n)_{n\in\mathbb{N}_0} = (\mathbb{E}[\|\Theta_n-\mathbb{E}[X_1]\|^2])_{n\in\mathbb{N}_0}$, $(\gamma_n)_{n\in\mathbb{N}} = (\frac{\gamma\alpha}{n^\nu})_{n\in\mathbb{N}}$ in the notation of Lemma 3.1) establishes that there exists $C\in(0,\infty)$ such that for all $n\in\mathbb{N}$ it holds that

$$\mathbb{E}\big[\|\Theta_n-\mathbb{E}[X_1]\|^2\big]\leq C(\tfrac{\gamma\alpha}{n^\nu}) = [C\gamma\alpha]\,n^{-\nu}. \tag{51}$$

Therefore, we obtain for all $n\in\mathbb{N}$ that

$$\big(\mathbb{E}\big[\|\Theta_n-\mathbb{E}[X_1]\|^2\big]\big)^{1/2}\leq [C\gamma\alpha]^{1/2}\,n^{-\nu/2}. \tag{52}$$

The proof of Proposition 3.3 is thus completed. □

3.2 Upper errors estimates in the case of fast decaying learning rates

In this subsection we establish in Proposition 3.4 below an upper bound for the root mean square error of the SGD process in (7) in the case of fast decaying learning rates (corresponding to the case $\nu = 1$ in Setting 2.1).

Proposition 3.4. *Assume Setting 2.1 and assume that $\nu = 1$. Then for every $\varepsilon\in(0,\infty)$ there exists $C\in(0,\infty)$ such that for all $n\in\mathbb{N}$ it holds that*

$$\big(\mathbb{E}\big[\|\Theta_n-\mathbb{E}[X_1]\|^2\big]\big)^{1/2}\leq Cn^{-(\min\{1/2,\gamma\alpha\}-\varepsilon)}. \tag{53}$$

Proof of Proposition 3.4. Throughout this proof let $\varepsilon\in(0,\min\{1/2,\gamma\alpha\})$, let $\beta\in(0,1/2)$ be given by

$$\beta = \min\{1/2,\gamma\alpha\}-\varepsilon, \tag{54}$$

and let $c=(c_n)_{n\in\mathbb{N}}\colon\mathbb{N}\to[0,\infty]$ be the function which satisfies for all $n\in\mathbb{N}$ that

$$c_n = \max\Big\{\Big\{\tfrac{\mathbb{E}[\|\Theta_k-\mathbb{E}[X_1]\|^2]}{k^{-2\beta}}\colon k\in\{1,2,\ldots,n\}\Big\}\cup\big\{(\gamma\alpha)^2\,\mathbb{E}\big[\|X_1-\mathbb{E}[X_1]\|^2\big]\big\}\Big\}. \tag{55}$$

Note that item (iv) in Proposition 2.6 implies that for all $n\in\mathbb{N}$ it holds that

$$\mathbb{E}\big[\|\Theta_n-\mathbb{E}[X_1]\|^2\big]<\infty. \tag{56}$$

Hence, we obtain that for all $n \in \mathbb{N}$ it holds that

$$c_n < \infty. \tag{57}$$

Next observe that

$$\begin{aligned}
&\limsup_{n\to\infty} \left[n^2\big((1-\tfrac{\gamma\alpha}{n})^2(n-1)^{-2\beta} - n^{-2\beta}\big)\right] \\
&= \limsup_{n\to\infty} \left[n^2\Big(\Big[1 - \tfrac{2\gamma\alpha}{n} + (\tfrac{\gamma\alpha}{n})^2\Big](n-1)^{-2\beta} - n^{-2\beta}\Big)\right] \\
&= \limsup_{n\to\infty} \left[n^2((n-1)^{-2\beta} - n^{-2\beta}) - n^2(\tfrac{2\gamma\alpha}{n})(n-1)^{-2\beta} + n^2(\tfrac{\gamma\alpha}{n})^2(n-1)^{-2\beta}\right] \\
&= \limsup_{n\to\infty} \left[n^2((n-1)^{-2\beta} - n^{-2\beta}) - (2\gamma\alpha)n(n-1)^{-2\beta} + \tfrac{(\gamma\alpha)^2}{(n-1)^{2\beta}}\right] \\
&= \limsup_{n\to\infty} \left[n^2((n-1)^{-2\beta} - n^{-2\beta}) - (2\gamma\alpha)n(n-1)^{-2\beta}\right].
\end{aligned} \tag{58}$$

The fact that for all $n \in \{2, 3, \ldots\}$ it holds that

$$\begin{aligned}
(n-1)^{-2\beta} - n^{-2\beta} &= -\left[x^{-2\beta}\right]_{x=n-1}^{x=n} = -\int_{n-1}^{n} (-2\beta)x^{-2\beta-1}\,\mathrm{d}x \\
&= 2\beta\left[\int_{n-1}^{n} x^{-2\beta-1}\,\mathrm{d}x\right] \leq 2\beta(n-1)^{-2\beta-1}
\end{aligned} \tag{59}$$

hence proves that

$$\begin{aligned}
&\limsup_{n\to\infty} \left[n^2\big((1-\tfrac{\gamma\alpha}{n})^2(n-1)^{-2\beta} - n^{-2\beta}\big)\right] \\
&\leq \limsup_{n\to\infty} \left[n^2 2\beta(n-1)^{-2\beta-1} - (2\gamma\alpha)n(n-1)^{-2\beta}\right] \\
&= \limsup_{n\to\infty} \left[n(n-1)^{-2\beta}(2\beta n(n-1)^{-1} - 2\gamma\alpha)\right].
\end{aligned} \tag{60}$$

Moreover, note that the fact that $2\beta = \min\{1, 2\gamma\alpha\} - 2\varepsilon \leq 2\gamma\alpha - 2\varepsilon < 2\gamma\alpha$ ensures that

$$\limsup_{n\to\infty} \left[2\beta(\tfrac{n+1}{n}) - 2\gamma\alpha\right] = \liminf_{n\to\infty} \left[2\beta(\tfrac{n+1}{n}) - 2\gamma\alpha\right] = \ 2\beta - 2\gamma\alpha < 0. \tag{61}$$

The fact that $2\beta < 2\gamma\alpha$, the fact that $2\beta < 1$, and (60) hence establish that

$$\begin{aligned}
&\limsup_{n\to\infty} \left[n^2\big((1-\tfrac{\gamma\alpha}{n})^2(n-1)^{-2\beta} - n^{-2\beta}\big)\right] \\
&\leq \limsup_{n\to\infty} \left[\tfrac{n}{(n-1)^{2\beta}}\left(2\beta(\tfrac{n}{n-1}) - 2\gamma\alpha\right)\right] \\
&= \limsup_{n\to\infty} \left[\left[\tfrac{(n+1)}{n^{2\beta}}\right]\left[2\beta(\tfrac{n+1}{n}) - 2\gamma\alpha\right]\right] \\
&= \lim_{n\to\infty} \left[\left[\tfrac{(n+1)}{n^{2\beta}}\right]\left[2\beta(\tfrac{n+1}{n}) - 2\gamma\alpha\right]\right] \\
&= \left[\lim_{n\to\infty}\left[\tfrac{(n+1)}{n^{2\beta}}\right]\right]\left[\lim_{n\to\infty}\left[2\beta(\tfrac{n+1}{n}) - 2\gamma\alpha\right]\right] \\
&= \left[\lim_{n\to\infty}\left[n^{1-2\beta} + n^{-2\beta}\right]\right][2\beta - 2\gamma\alpha] \\
&= \left[\lim_{n\to\infty} n^{1-2\beta}\right][2\beta - 2\gamma\alpha] = -\infty.
\end{aligned} \tag{62}$$

This implies that there exists $m \in \mathbb{N}$ such that for all $n \in \mathbb{N} \cap (m, \infty)$ it holds that

$$n^2\big((1-\tfrac{\gamma\alpha}{n})^2(n-1)^{-2\beta} - n^{-2\beta}\big) \leq -1. \tag{63}$$

Next note that (57) establishes that $c_m < \infty$. Moreover, observe that (55) ensures that for all $n \in \{1, 2, \ldots, m\}$ it holds that

$$\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] = \left[\tfrac{\mathbb{E}[\|\Theta_n - \mathbb{E}[X_1]\|^2]}{n^{-2\beta}}\right] n^{-2\beta} \leq c_m n^{-2\beta}. \tag{64}$$

Next we claim that for all $n \in \{m, m+1, \ldots\}$ it holds that

$$\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] \leq c_m n^{-2\beta}. \tag{65}$$

Note that (64) establishes (65) in the base case $n = m$. For the induction step $\{m, m+1, \ldots\} \ni (n-1) \to n \in \mathbb{N} \cap (m, \infty)$ note that item (iii) in Proposition 2.6 assures that for all $n \in \{2, 3, \ldots\}$ with $\mathbb{E}\big[\|\Theta_{n-1} - \mathbb{E}[X_1]\|^2\big] \leq c_m(n-1)^{-2\beta}$ it holds that

$$\begin{aligned}
&\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] \\
&= (1-\tfrac{\gamma\alpha}{n})^2\,\mathbb{E}\big[\|\Theta_{n-1} - \mathbb{E}[X_1]\|^2\big] + (\tfrac{\gamma\alpha}{n})^2\,\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] \\
&\leq (1-\tfrac{\gamma\alpha}{n})^2\, c_m(n-1)^{-2\beta} + (\tfrac{\gamma\alpha}{n})^2\,\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] \\
&= (1-\tfrac{\gamma\alpha}{n})^2\, c_m(n-1)^{-2\beta} - c_m n^{-2\beta} + (\tfrac{\gamma\alpha}{n})^2\,\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] + c_m n^{-2\beta} \\
&= \tfrac{1}{n^2}\Big[c_m\left[n^2\big((1-\tfrac{\gamma\alpha}{n})^2(n-1)^{-2\beta} - n^{-2\beta}\big)\right] + (\gamma\alpha)^2\,\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big]\Big] \\
&\quad + c_m n^{-2\beta}.
\end{aligned} \tag{66}$$

This, (55), and (63) demonstrate that for all $n \in \mathbb{N} \cap (m, \infty)$ with $\mathbb{E}\big[\|\Theta_{n-1} - \mathbb{E}[X_1]\|^2\big] \leq c_m (n-1)^{-2\beta}$ it holds that

$$\begin{aligned}
&\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] \\
&\leq \tfrac{1}{n^2}\Big[- c_m + (\gamma\alpha)^2 \, \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] \Big] + c_m n^{-2\beta} \\
&\leq \tfrac{1}{n^2}\Big[- (\gamma\alpha)^2 \, \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] + (\gamma\alpha)^2 \, \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] \Big] + c_m n^{-2\beta} \\
&= c_m n^{-2\beta}.
\end{aligned} \tag{67}$$

Induction thus proves (65). Combining (64) and (65) establishes (53). The proof of Proposition 3.4 is thus completed. □

3.3 Refined upper errors estimates in the case of fast decaying learning rates

3.3.1 A characterization of an asymptotic property for fast decaying learning rates

Lemma 3.5. *Let $\beta \in (0, \infty)$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq (0, \infty)$ satisfy for all $n \in \mathbb{N}$ that $\gamma_n = \beta / n$. Then the following two statements are equivalent:*

(i) It holds that

$$\liminf_{l \to \infty} \left[\frac{\gamma_l - \gamma_{l-1}}{(\gamma_l)^2} + \frac{2\gamma_{l-1}}{\gamma_l} - \gamma_{l-1} \right] > 0. \tag{68}$$

(ii) It holds that $\beta > 1/2$.

Proof of Lemma 3.5. First, observe that

$$\begin{aligned}
&\liminf_{l\to\infty}\left[\frac{\gamma_l-\gamma_{l-1}}{(\gamma_l)^2}+\frac{2\gamma_{l-1}}{\gamma_l}-\gamma_{l-1}\right]\\
&=\liminf_{l\to\infty}\left[\frac{\left(\frac{\beta}{l}\right)-\left(\frac{\beta}{l-1}\right)}{\left(\frac{\beta}{l}\right)^2}+\frac{2\left(\frac{\beta}{l-1}\right)}{\left(\frac{\beta}{l}\right)}-\left(\tfrac{\beta}{l-1}\right)\right]\\
&=\liminf_{l\to\infty}\left[\frac{1}{\beta}\left[\frac{\left(\frac{l-1-l}{l(l-1)}\right)}{\left(\frac{1}{l}\right)^2}\right]+2\left(\tfrac{l}{l-1}\right)-\beta\left(\tfrac{1}{l-1}\right)\right]\\
&=\liminf_{l\to\infty}\left[-\tfrac{1}{\beta}\left(\tfrac{l^2}{l(l-1)}\right)+2\left(\tfrac{l}{l-1}\right)-\beta\left(\tfrac{1}{l-1}\right)\right]\\
&=\liminf_{l\to\infty}\left[-\tfrac{1}{\beta}\left(\tfrac{l}{l-1}\right)+2\left(\tfrac{l}{l-1}\right)-\beta\left(\tfrac{1}{l-1}\right)\right]\\
&=\lim_{l\to\infty}\left[-\tfrac{1}{\beta}\left(\tfrac{l}{l-1}\right)+2\left(\tfrac{l}{l-1}\right)-\beta\left(\tfrac{1}{l-1}\right)\right]\\
&=-\tfrac{1}{\beta}+2=2-\tfrac{1}{\beta}.
\end{aligned}\tag{69}$$

Therefore, we obtain that

$$\begin{aligned}
&\left(\liminf_{l\to\infty}\left[\frac{\gamma_l-\gamma_{l-1}}{(\gamma_l)^2}+\frac{2\gamma_{l-1}}{\gamma_l}-\gamma_{l-1}\right]>0\right)\\
\Leftrightarrow\quad&\left(2-\tfrac{1}{\beta}>0\right)\\
\Leftrightarrow\quad&\left(2>\tfrac{1}{\beta}\right)\\
\Leftrightarrow\quad&\left(\tfrac{1}{2}<\beta\right).
\end{aligned}\tag{70}$$

The proof of Lemma 3.5 is thus completed. □

3.3.2 Improved upper error estimates in the case of large but fast decaying learning rates

Proposition 3.6. *Assume Setting 2.1 and assume that $\nu = 1$ and $\gamma\alpha > 1/2$. Then there exists $C \in (0,\infty)$ such that for all $n \in \mathbb{N}$ it holds that*

$$\left(\mathbb{E}\big[\|\Theta_n-\mathbb{E}[X_1]\|^2\big]\right)^{1/2}\leq Cn^{-1/2}.\tag{71}$$

Proof of Proposition 3.6. Observe that items (iii)–(iv) in Proposition 2.6 imply that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
&\mathbb{E}\big[\|\Theta_n-\mathbb{E}[X_1]\|^2\big]\\
&=(1-\tfrac{\gamma\alpha}{n})^2\,\mathbb{E}\big[\|\Theta_{n-1}-\mathbb{E}[X_1]\|^2\big]+(\tfrac{\gamma\alpha}{n})^2\,\mathbb{E}\big[\|X_1-\mathbb{E}[X_1]\|^2\big]<\infty.
\end{aligned}\tag{72}$$

Moreover, note that the hypothesis that $\gamma\alpha > 1/2$ and Lemma 3.5 (with $\beta = \gamma\alpha$ in the notation of Lemma 3.5) ensures that

$$\liminf_{l\to\infty}\left[\frac{\left(\frac{\gamma\alpha}{l}\right)-\left(\frac{\gamma\alpha}{(l-1)}\right)}{\left(\frac{\gamma\alpha}{l}\right)^2}+\frac{2\left(\frac{\gamma\alpha}{(l-1)}\right)}{\left(\frac{\gamma\alpha}{l}\right)}-\left(\frac{\gamma\alpha}{(l-1)}\right)\right]>0. \tag{73}$$

Combining this and (72) with Lemma 3.1 (with $\kappa = \mathbb{E}[\|X_1-\mathbb{E}[X_1]\|^2]$, $(e_n)_{n\in\mathbb{N}_0} = (\mathbb{E}[\|\Theta_n-\mathbb{E}[X_1]\|^2])_{n\in\mathbb{N}_0}$, $(\gamma_n)_{n\in\mathbb{N}} = (\frac{\gamma\alpha}{n})_{n\in\mathbb{N}}$ in the notation of Lemma 3.1) demonstrates that there exists $C\in(0,\infty)$ such that for all $n\in\mathbb{N}$ it holds that

$$\mathbb{E}\big[\|\Theta_n-\mathbb{E}[X_1]\|^2\big]\leq C(\tfrac{\gamma\alpha}{n}) = [C\gamma\alpha]\,n^{-1}. \tag{74}$$

Therefore, we obtain for all $n\in\mathbb{N}$ that

$$\big(\mathbb{E}\big[\|\Theta_n-\mathbb{E}[X_1]\|^2\big]\big)^{1/2}\leq [C\gamma\alpha]^{1/2}\,n^{-1/2}. \tag{75}$$

The proof of Proposition 3.6 is thus completed. □

3.3.3 Refined upper error estimates in the case of fast decaying learning rates

The next result, Corollary 3.7 below, combines the upper error bounds for fast decaying learning rates obtained in Proposition 3.4 and Proposition 3.6 above.

Corollary 3.7. *Assume Setting 2.1 and assume that $\nu = 1$. Then for every $\varepsilon\in(0,\infty)$ there exists $C\in(0,\infty)$ such that for all $n\in\mathbb{N}$ it holds that*

$$\big(\mathbb{E}\big[\|\Theta_n-\mathbb{E}[X_1]\|^2\big]\big)^{1/2}\leq Cn^{-\min\{1/2,\gamma\alpha-\varepsilon\}}. \tag{76}$$

Proof of Corollary 3.7. To prove (76) we distinguish between the cases $\gamma\alpha > 1/2$ and $\gamma\alpha\leq 1/2$. We first consider the case $\gamma\alpha > 1/2$. In this case observe that Proposition 3.6 proves that there exists $C\in(0,\infty)$ such that for all $\varepsilon\in(0,\infty)$, $n\in\mathbb{N}$ it holds that

$$\begin{aligned}\big(\mathbb{E}\big[\|\Theta_n-\mathbb{E}[X_1]\|^2\big]\big)^{1/2} &\leq Cn^{-1/2}\\ &= Cn^{[\min\{1/2,\gamma\alpha-\varepsilon\}-1/2]}n^{-\min\{1/2,\gamma\alpha-\varepsilon\}}\\ &= Cn^{\min\{0,\gamma\alpha-\varepsilon-1/2\}}n^{-\min\{1/2,\gamma\alpha-\varepsilon\}}\\ &\leq Cn^{-\min\{1/2,\gamma\alpha-\varepsilon\}}.\end{aligned} \tag{77}$$

Hence, we obtain that for every $\varepsilon \in (0, \infty)$ there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}$ it holds that

$$\left(\mathbb{E}\left[\|\Theta_n - \mathbb{E}[X_1]\|^2\right]\right)^{1/2} \leq C n^{-\min\{1/2, \gamma\alpha - \varepsilon\}}. \tag{78}$$

This establishes (76) in the case $\gamma\alpha > 1/2$. Next we consider the case $\gamma\alpha \leq 1/2$. In this case we observe that Proposition 3.4 proves that for every $\varepsilon \in (0, \infty)$ there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}$ it holds that

$$\left(\mathbb{E}\left[\|\Theta_n - \mathbb{E}[X_1]\|^2\right]\right)^{1/2} \leq C n^{-(\min\{1/2, \gamma\alpha\} - \varepsilon)} = C n^{-(\gamma\alpha - \varepsilon)} = C n^{-\min\{1/2, \gamma\alpha - \varepsilon\}}. \tag{79}$$

This proves (76) in the case $\gamma\alpha \leq 1/2$. Combining (78) and (79) establishes (76). The proof of Corollary 3.7 is thus completed. □

3.4 Upper error estimates in the case of very fast decaying learning rates

In this subsection we establish in Lemma 3.8 below that the root mean square error of the SGD process in (7) is bounded from above in the case of very fast decaying learning rates (corresponding to the case $\nu > 1$ in Setting 2.1).

Lemma 3.8. *Assume Setting 2.1 and assume that $\nu > 1$. Then there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}$ it holds that*

$$\left(\mathbb{E}\left[\|\Theta_n - \mathbb{E}[X_1]\|^2\right]\right)^{1/2} \leq C. \tag{80}$$

Proof of Lemma 3.8. Throughout this proof let $m \in \mathbb{N} \cap (\gamma\alpha, \infty)$ and let $C \in (0, \infty)$ be given by

$$C = \max\left(\left[\bigcup_{\substack{k, r \in \mathbb{N}, \\ r \leq k \leq m}} \left\{\prod_{l=r}^{k} \left|1 - \tfrac{\gamma\alpha}{l^\nu}\right|^2\right\}\right] \cup \{1\}\right). \tag{81}$$

Observe that item (iv) in Proposition 2.6 ensures that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
&\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] \\
&= \|\xi - \mathbb{E}[X_1]\|^2 \left[\prod_{l=1}^{n}(1 - \tfrac{\gamma\alpha}{l^\nu})\right]^2 \\
&\quad + \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] \left[\sum_{k=1}^{n}\left[\tfrac{\gamma\alpha}{k^\nu}\left(\prod_{l=k+1}^{n}(1 - \tfrac{\gamma\alpha}{l^\nu})\right)\right]^2\right] \\
&= \|\xi - \mathbb{E}[X_1]\|^2 \left[\prod_{l=1}^{\min\{m,n\}}\big|1 - \tfrac{\gamma\alpha}{l^\nu}\big|^2\right]\left[\prod_{l=\min\{m,n\}+1}^{n}\big|1 - \tfrac{\gamma\alpha}{l^\nu}\big|^2\right] \\
&\quad + (\gamma\alpha)^2\,\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] \\
&\quad \cdot \left[\sum_{k=1}^{n}\left(\frac{1}{k^{2\nu}}\left[\prod_{l=k+1}^{\min\{m,n\}}\big|1 - \tfrac{\gamma\alpha}{l^\nu}\big|^2\right]\left[\prod_{l=\max\{k,\min\{m,n\}\}+1}^{n}\big|1 - \tfrac{\gamma\alpha}{l^\nu}\big|^2\right]\right)\right]
\end{aligned} \tag{82}$$

This and (81) establish that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
&\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] \\
&\leq C\|\xi - \mathbb{E}[X_1]\|^2 \left[\prod_{l=\min\{m,n\}+1}^{n}\big|1 - \tfrac{\gamma\alpha}{l^\nu}\big|^2\right] \\
&\quad + (\gamma\alpha)^2\,\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big]\left[\sum_{k=1}^{n}\left(\frac{C}{k^{2\nu}}\left[\prod_{l=\max\{k,\min\{m,n\}\}+1}^{n}\big|1 - \tfrac{\gamma\alpha}{l^\nu}\big|^2\right]\right)\right].
\end{aligned} \tag{83}$$

Next note that the fact that $m > \gamma\alpha$ assures that for all $l \in \mathbb{N} \cap (m, \infty)$ it holds that

$$0 = 1 - 1 < 1 - \tfrac{\gamma\alpha}{m} \leq 1 - \tfrac{\gamma\alpha}{m^\nu} \leq 1 - \tfrac{\gamma\alpha}{l^\nu} < 1 - 0 = 1. \tag{84}$$

Hence, we obtain that for all $l \in \mathbb{N} \cap (m, \infty)$ it holds that

$$\big(1 - \tfrac{\gamma\alpha}{l^\nu}\big) \in (0, 1). \tag{85}$$

This implies that for all $n, l \in \mathbb{N}$ with $\min\{m, n\} < l \leq n$ it holds that

$$\big|1 - \tfrac{\gamma\alpha}{l^\nu}\big|^2 = \big(1 - \tfrac{\gamma\alpha}{l^\nu}\big)^2 \in (0, 1). \tag{86}$$

Combining this with (83) demonstrates that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} &\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] \\ &\leq C\|\xi - \mathbb{E}[X_1]\|^2 + (\gamma\alpha)^2 C\,\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big]\left[\sum_{k=1}^{n} \frac{1}{k^{2\nu}}\right]. \end{aligned} \tag{87}$$

Moreover, note that the hypothesis that $\nu > 1$ ensures that

$$\begin{aligned} \sum_{k=1}^{\infty} \frac{1}{k^{2\nu}} &= 1 + \sum_{k=2}^{\infty}\left[\int_{k-1}^{k} \frac{1}{k^{2\nu}}\,\mathrm{d}x\right] \leq 1 + \sum_{k=2}^{\infty}\left[\int_{k-1}^{k} \frac{1}{x^{2\nu}}\,\mathrm{d}x\right] = 1 + \int_{1}^{\infty} \frac{1}{x^{2\nu}}\,\mathrm{d}x \\ &= 1 + \left[\left(\tfrac{1}{(1-2\nu)}\right) x^{1-2\nu}\right]_{x=1}^{x=\infty} = 1 - \left(\tfrac{1}{(1-2\nu)}\right) = 1 + \tfrac{1}{(2\nu-1)} < \infty. \end{aligned} \tag{88}$$

This and (87) establish (80). The proof of Lemma 3.8 is thus completed. □

4 Lower error estimates for the SGD optimization method

In this section we establish in Proposition 4.5 and Proposition 4.11 below lower bounds for the root mean square distance between the SGD process in (7) and the global minimum of the considered optimization problem (cf. item (ii) in Lemma 2.4). These results show that the upper error bounds obtained in Section 3 (see Proposition 3.3 and Corollary 3.7 above) can essentially not be improved. Moreover, in Subsection 4.3 below we demonstrate that the SGD process fails to converge to the global minimum of the objective function in the case of very fast decaying learning rates (see Lemma 4.12 below for details). Finally, in Subsection 4.4 below we present Theorem 4.13 which combines the main findings of this article.

4.1 Lower errors estimates in the case of slowly and fast decaying learning rates

In this subsection we establish in Proposition 4.5 below a lower bound for the root mean square error of the SGD process in (7) in the case of slowly and fast decaying learning rates (corresponding to the case $\nu \leq 1$ in Setting 2.1). Our proof of Proposition 4.5 employs the elementary result in Lemma 4.1 and the elementary and well-known result in Lemma 4.3. For completeness we also provide the proofs of Lemma 4.1 and Lemma 4.3 here.

4.1.1 On the strict positivity of the mean square errors

Lemma 4.1. *Assume Setting 2.1 and assume that* $\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] > 0$. *Then it holds for all* $n \in \mathbb{N}$ *that*

$$\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] > 0. \tag{89}$$

Proof of Lemma 4.1. Observe that item (iv) in Proposition 2.6 and the assumption that $\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] > 0$ assure that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
&\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] \\
&= \left[\prod_{l=1}^{n}(1 - \tfrac{\gamma\alpha}{l^\nu})\right]^2 \|\xi - \mathbb{E}[X_1]\|^2 \\
&\quad + \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] \left[\sum_{k=1}^{n}\left[\tfrac{\gamma\alpha}{k^\nu}\left(\prod_{l=k+1}^{n}(1 - \tfrac{\gamma\alpha}{l^\nu})\right)\right]^2\right] \\
&\geq \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big]\,\big(\tfrac{\gamma\alpha}{n^\nu}\big) > 0.
\end{aligned} \tag{90}$$

The proof of Lemma 4.1 is thus completed. □

4.1.2 Approximations of the exponential function

Lemma 4.2. *Let* $(a_l)_{l\in\mathbb{N}} \subseteq \mathbb{R}$, $(n_l)_{l\in\mathbb{N}} \subseteq \mathbb{N}$ *satisfy that* $\liminf_{l\to\infty} a_l = \limsup_{l\to\infty} a_l$ *and* $\liminf_{l\to\infty} n_l = \infty$. *Then*

$$\limsup_{l\to\infty}\left|\left[1 + \tfrac{a_l}{n_l}\right]^{n_l} - \exp\Big(\lim_{l\to\infty} a_l\Big)\right| = 0. \tag{91}$$

Proof of Lemma 4.2. Throughout this proof let $f_l\colon \mathbb{N}_0 \to \mathbb{R}$, $l \in \mathbb{N}$, be the functions which satisfy for all $l \in \mathbb{N}$, $k \in \mathbb{N}_0$ that

$$f_l(k) = \begin{cases} \left[\prod_{r=0}^{k-1}(n_l - r)\right]\frac{(a_l)^k}{(n_l)^k k!} & : k \leq n_l \\ 0 & : k > n_l, \end{cases} \tag{92}$$

let $F\colon \mathbb{N}_0 \to \mathbb{R}$ be the function which satisfies for all $k \in \mathbb{N}_0$ that

$$F(k) = \frac{[\sup_{l\in\mathbb{N}}|a_l|]^k}{k!}, \tag{93}$$

and let $\#\colon \mathcal{P}(\mathbb{N}_0) \to [0,\infty]$ be the counting measure on $\mathbb{N}_0$. Observe that the binomial theorem proves that for all $l \in \mathbb{N}$ it holds that

$$\left[1 + \tfrac{a_l}{n_l}\right]^{n_l} = \sum_{k=0}^{n_l} \binom{n_l}{k} \left[\tfrac{a_l}{n_l}\right]^k = \sum_{k=0}^{\infty} f_l(k) = \int_{\mathbb{N}_0} f_l(k)\, \#(\mathrm{d}k). \tag{94}$$

Moreover, note the hypothesis that $\liminf_{l\to\infty} n_l = \infty$ ensures that for all $k \in \mathbb{N}_0$ it holds that

$$\lim_{l\to\infty} f_l(k) = \lim_{l\to\infty} \left[\left[\prod_{r=0}^{k-1} (1 - \tfrac{r}{n_l}) \right] \tfrac{(a_l)^k}{k!} \right] = \frac{\left[\lim_{l\to\infty} a_l\right]^k}{k!}. \tag{95}$$

In addition, note that for all $k \in \mathbb{N}_0$ it holds that

$$\sup_{l\in\mathbb{N}} |f_l(k)| \leq F(k). \tag{96}$$

The fact that

$$\int_{\mathbb{N}_0} F(k)\, \#(\mathrm{d}k) = \sum_{k=0}^{\infty} \left[\frac{\left[\sup_{l\in\mathbb{N}} |a_l|\right]^k}{k!} \right] = \exp\left(\sup_{l\in\mathbb{N}} |a_l| \right) < \infty, \tag{97}$$

Lebesgue's theorem of dominated convergence, and (95) hence demonstrate that

$$\begin{aligned} &\lim_{l\to\infty} \left[\int_{\mathbb{N}_0} f_l(k)\, \#(\mathrm{d}k) \right] = \int_{\mathbb{N}_0} \left[\lim_{l\to\infty} f_l(k) \right] \#(\mathrm{d}k) \\ &= \int_{\mathbb{N}_0} \frac{\left[\lim_{l\to\infty} a_l\right]^k}{k!}\, \#(\mathrm{d}k) = \sum_{k=0}^{\infty} \frac{\left[\lim_{l\to\infty} a_l\right]^k}{k!} = \exp\left(\lim_{l\to\infty} a_l \right). \end{aligned} \tag{98}$$

This and (94) ensure that

$$\lim_{l\to\infty} \left[\left[1 + \tfrac{a_l}{n_l}\right]^{n_l} \right] = \exp\left(\lim_{l\to\infty} a_l \right). \tag{99}$$

The proof of Lemma 4.2 is thus completed. □

Lemma 4.3. *Let $(a_l)_{l\in\mathbb{N}} \subseteq \mathbb{R}$, $(n_l)_{l\in\mathbb{N}} \subseteq \mathbb{R}$ satisfy that $\liminf_{l\to\infty} a_l = \limsup_{l\to\infty} a_l$ and $\liminf_{l\to\infty} n_l = \infty$.*

$$\limsup_{l\to\infty} \left| \left[1 + \tfrac{a_l}{n_l}\right]^{n_l} - \exp\left(\lim_{l\to\infty} a_l \right) \right| = 0. \tag{100}$$

Proof of Lemma 4.3. Throughout this proof let $\lfloor \cdot \rfloor \colon \mathbb{R} \to \mathbb{Z}$ be the function which satisfies for all $x \in \mathbb{R}$ that $\lfloor x \rfloor = \max((-\infty, x] \cap \mathbb{Z})$. Observe that the hypothesis that $\liminf_{l\to\infty} n_l = \infty$ ensures that

$$1 \geq \limsup_{l\to\infty} \left[\tfrac{\lfloor n_l \rfloor}{n_l}\right] \geq \liminf_{l\to\infty} \left[\tfrac{\lfloor n_l \rfloor}{n_l}\right] \geq \liminf_{l\to\infty} \left[\tfrac{(n_l-1)}{n_l}\right] = \liminf_{l\to\infty} \left[1 - \tfrac{1}{n_l}\right] = 1. \quad (101)$$

This implies that $\lim_{l\to\infty} \left[\frac{\lfloor n_l \rfloor}{n_l}\right] = 1$. The hypothesis that $\liminf_{l\to\infty} a_l = \limsup_{l\to\infty} a_l$ therefore assures that

$$\lim_{l\to\infty} \left[\tfrac{a_l \lfloor n_l \rfloor}{n_l}\right] = \left[\lim_{l\to\infty} a_l\right] \left[\lim_{l\to\infty} \tfrac{\lfloor n_l \rfloor}{n_l}\right] = \lim_{l\to\infty} a_l. \quad (102)$$

This, the fact that $\liminf_{l\to\infty} \lfloor n_l \rfloor = \infty$, and Lemma 4.2 (with $(a_l)_{l\in\mathbb{N}} = (\frac{a_l \lfloor n_l \rfloor}{n_l})_{l\in\mathbb{N}}$, $(n_l)_{l\in\mathbb{N}} = (\lfloor n_l \rfloor)_{l\in\mathbb{N}}$ in the notation of Lemma 4.2) proves that

$$\lim_{l\to\infty} \left(\left[1 + \frac{\left(\frac{a_l \lfloor n_l \rfloor}{n_l}\right)}{\lfloor n_l \rfloor} \right]^{\lfloor n_l \rfloor} \right) = \exp\left(\lim_{l\to\infty} \left[\tfrac{a_l \lfloor n_l \rfloor}{n_l} \right] \right) = \exp\left(\lim_{l\to\infty} a_l \right). \quad (103)$$

Next note that the fact that for all $\alpha \in [0,1], r \in (0,1]$ it holds that $r \leq r^\alpha \leq 1$ and the fact that for all $\alpha \in [0,1], r \in [1,\infty)$ it holds that $1 \leq r^\alpha \leq r$ show that for all $\alpha \in [0,1], r \in (0,\infty)$ it holds that

$$|1 - r^\alpha| \leq |1 - r|. \quad (104)$$

Combining this and the fact that for all $l \in \mathbb{N}$ it holds that $n_l - \lfloor n_l \rfloor \in [0,1]$ with the hypothesis that $\liminf_{l\to\infty} n_l = \infty$ and the fact that $\sup_{l\in\mathbb{N}} |a_l| < \infty$ demonstrates that

$$\limsup_{l\to\infty} \left| 1 - \left[1 + \tfrac{a_l}{n_l}\right]^{n_l - \lfloor n_l \rfloor} \right| \leq \limsup_{l\to\infty} \left| 1 - \left[1 + \tfrac{a_l}{n_l}\right] \right| = \limsup_{l\to\infty} \left| \tfrac{a_l}{n_l} \right| = 0. \quad (105)$$

This and (103) establish that

$$\begin{aligned}
\lim_{l\to\infty} \left[\left[1 + \tfrac{a_l}{n_l}\right]^{n_l} \right] &= \lim_{l\to\infty} \left[\left[1 + \tfrac{a_l}{n_l}\right]^{\lfloor n_l \rfloor} \left[1 + \tfrac{a_l}{n_l}\right]^{n_l - \lfloor n_l \rfloor} \right] \\
&= \lim_{l\to\infty} \left[\left[1 + \frac{\left(\frac{a_l \lfloor n_l \rfloor}{n_l}\right)}{\lfloor n_l \rfloor} \right]^{\lfloor n_l \rfloor} \left[1 + \tfrac{a_l}{n_l}\right]^{n_l - \lfloor n_l \rfloor} \right] \\
&= \left[\lim_{l\to\infty} \left[1 + \frac{\left(\frac{a_l \lfloor n_l \rfloor}{n_l}\right)}{\lfloor n_l \rfloor} \right]^{\lfloor n_l \rfloor} \right] \left[\lim_{l\to\infty} \left[1 + \tfrac{a_l}{n_l}\right]^{n_l - \lfloor n_l \rfloor} \right] \\
&= \exp\left(\lim_{l\to\infty} a_l \right).
\end{aligned} \quad (106)$$

The proof of Lemma 4.3 is thus completed. □

4.1.3 Lower error estimates

Lemma 4.4. *Let $\beta \in (0,\infty)$, $\nu \in (0,1]$. Then*

$$\liminf_{n\to\infty}\left[n^{\nu}\left(\sum_{k=1}^{n}\left[\tfrac{\beta}{k^{\nu}}\left(\prod_{l=k+1}^{n}(1-\tfrac{\beta}{l^{\nu}})\right)\right]^{2}\right)\right] \geq \frac{\beta^{2}\exp(-2^{\nu}\beta)}{2}. \tag{107}$$

Proof of Lemma 4.4. Throughout this proof let $\lceil\cdot\rceil\colon \mathbb{R}\to\mathbb{Z}$ be the function which satisfies for all $x\in\mathbb{R}$ that $\lceil x\rceil = \min([x,\infty)\cap\mathbb{Z})$. Observe that the fact that for all $n\in\mathbb{N}$ it holds that $\lceil n-\frac{n^{\nu}}{2}\rceil \geq n-\frac{n^{\nu}}{2}\geq n-\frac{n}{2}=\frac{n}{2}>0$ ensures that

$$\begin{aligned}
&\liminf_{n\to\infty}\left[n^{\nu}\left(\sum_{k=1}^{n}\left[\tfrac{\beta}{k^{\nu}}\left(\prod_{l=k+1}^{n}(1-\tfrac{\beta}{l^{\nu}})\right)\right]^{2}\right)\right]\\
&\geq \liminf_{n\to\infty}\left[n^{\nu}\left(\sum_{k=\lceil n-\frac{n^{\nu}}{2}\rceil}^{n}\left[\tfrac{\beta}{k^{\nu}}\left(\prod_{l=k+1}^{n}(1-\tfrac{\beta}{l^{\nu}})\right)\right]^{2}\right)\right]\\
&\geq \liminf_{n\to\infty}\left[n^{\nu}\left(\sum_{k=\lceil n-\frac{n^{\nu}}{2}\rceil}^{n}\left[\tfrac{\beta}{n^{\nu}}\left(\prod_{l=\lceil n-\frac{n^{\nu}}{2}\rceil+1}^{n}(1-\tfrac{\beta}{l^{\nu}})\right)\right]^{2}\right)\right]\\
&\geq \liminf_{n\to\infty}\left[n^{\nu}\left[\tfrac{\beta}{n^{\nu}}\right]^{2}\left(\sum_{k=\lceil n-\frac{n^{\nu}}{2}\rceil}^{n}\left[\left[1-\frac{\beta}{(\lceil n-\frac{n^{\nu}}{2}\rceil)^{\nu}}\right]^{n-\lceil n-\frac{n^{\nu}}{2}\rceil}\right]^{2}\right)\right]\\
&= \liminf_{n\to\infty}\left[\tfrac{\beta^{2}}{n^{\nu}}\left(n-\lceil n-\tfrac{n^{\nu}}{2}\rceil+1\right)\left[\left[1-\frac{\beta}{(\lceil n-\frac{n^{\nu}}{2}\rceil)^{\nu}}\right]^{n-\lceil n-\frac{n^{\nu}}{2}\rceil}\right]^{2}\right].
\end{aligned} \tag{108}$$

The fact that for all $n\in\mathbb{N}$ it holds that $\lceil n-\frac{n^{\nu}}{2}\rceil \leq n-\frac{n^{\nu}}{2}+1$ and the fact that for

all $n \in \mathbb{N}$ it holds that $\lceil n - \frac{n^\nu}{2} \rceil \geq n - \frac{n^\nu}{2} \geq n - \frac{n}{2} = \frac{n}{2}$ hence demonstrate that

$$\begin{aligned}
&\liminf_{n\to\infty} \left[n^\nu \left(\sum_{k=1}^{n} \left[\tfrac{\beta}{k^\nu} \left(\prod_{l=k+1}^{n} (1 - \tfrac{\beta}{l^\nu}) \right) \right]^2 \right) \right] \\
&\geq \liminf_{n\to\infty} \left[\tfrac{\beta^2}{n^\nu} \left(n - (n - \tfrac{n^\nu}{2} + 1) + 1 \right) \left[\left[1 - \tfrac{\beta}{(\frac{n}{2})^\nu} \right]^{n - \left(n - \frac{n^\nu}{2}\right)} \right]^2 \right] \\
&= \liminf_{n\to\infty} \left[\tfrac{\beta^2}{n^\nu} \left(\tfrac{n^\nu}{2} \left[1 - \tfrac{2^\nu \beta}{n^\nu} \right]^{n^\nu} \right) \right] = \liminf_{n\to\infty} \left(\tfrac{\beta^2}{2} \left[1 - \tfrac{2^\nu \beta}{n^\nu} \right]^{n^\nu} \right) \\
&= \tfrac{\beta^2}{2} \left[\liminf_{n\to\infty} \left(\left[1 - \tfrac{2^\nu \beta}{n^\nu} \right]^{n^\nu} \right) \right].
\end{aligned} \tag{109}$$

Combining this with Lemma 4.3 (with $a_l = -2^\nu \beta$, $n_l = l^\nu$ for $l \in \mathbb{N}$ in the notation of Lemma 4.3) establishes that

$$\liminf_{n\to\infty} \left[n^\nu \left(\sum_{k=1}^{n} \left[\tfrac{\beta}{k^\nu} \left(\prod_{l=k+1}^{n} (1 - \tfrac{\beta}{l^\nu}) \right) \right]^2 \right) \right] \geq \frac{\beta^2 \exp(-2^\nu \beta)}{2}. \tag{110}$$

The proof of Lemma 4.4 is thus completed. □

Proposition 4.5. *Assume Setting 2.1 and assume that $\nu \leq 1$ and $\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] > 0$. Then there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}$ it holds that*

$$\big(\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2 \big] \big)^{1/2} \geq C n^{-\nu/2}. \tag{111}$$

Proof of Proposition 4.5. First, observe that item (iv) in Proposition 2.6 ensures that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2 \big] &= \left[\prod_{l=1}^{n} (1 - \tfrac{\gamma\alpha}{l^\nu}) \right]^2 \|\xi - \mathbb{E}[X_1]\|^2 \\
&\quad + \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2 \big] \left[\sum_{k=1}^{n} \left[\tfrac{\gamma\alpha}{k^\nu} \left(\prod_{l=k+1}^{n} (1 - \tfrac{\gamma\alpha}{l^\nu}) \right) \right]^2 \right] \\
&\geq \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2 \big] \left[\sum_{k=1}^{n} \left[\tfrac{\gamma\alpha}{k^\nu} \left(\prod_{l=k+1}^{n} (1 - \tfrac{\gamma\alpha}{l^\nu}) \right) \right]^2 \right].
\end{aligned} \tag{112}$$

Moreover, note that Lemma 4.4 (with $\beta = \gamma\alpha$, $\nu = \nu$ in the notation of Lemma 4.4) implies that there exists $m \in \mathbb{N}$ such that for all $n \in \{m, m+1, \ldots\}$ it holds that

$$n^{\nu}\left(\sum_{k=1}^{n}\left[\tfrac{\gamma\alpha}{k^{\nu}}\left(\prod_{l=k+1}^{n}(1-\tfrac{\gamma\alpha}{l^{\nu}})\right)\right]^2\right) \geq \tfrac{1}{2}\big(\tfrac{(\gamma\alpha)^2\exp(-2^{\nu}\gamma\alpha)}{2}\big) = \tfrac{(\gamma\alpha)^2\exp(-2^{\nu}\gamma\alpha)}{4}. \tag{113}$$

Therefore, we obtain for all $n \in \{m, m+1, \ldots\}$ that

$$\sum_{k=1}^{n}\left[\tfrac{\gamma\alpha}{k^{\nu}}\left(\prod_{l=k+1}^{n}(1-\tfrac{\gamma\alpha}{l^{\nu}})\right)\right]^2 \geq \left[\tfrac{(\gamma\alpha)^2\exp(-2^{\nu}\gamma\alpha)}{4}\right] n^{-\nu}. \tag{114}$$

This and (112) demonstrate that for all $n \in \{m, m+1, \ldots\}$ it holds that

$$\begin{aligned}\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] &\geq \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big]\left[\sum_{k=1}^{n}\left[\tfrac{\gamma\alpha}{k^{\nu}}\left(\prod_{l=k+1}^{n}(1-\tfrac{\gamma\alpha}{l^{\nu}})\right)\right]^2\right] \\ &\geq \left[\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big]\Big(\tfrac{(\gamma\alpha)^2\exp(-2^{\nu}\gamma\alpha)}{4}\Big)\right] n^{-\nu}.\end{aligned} \tag{115}$$

Furthermore, observe that Lemma 4.1 and the hypothesis that $\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] > 0$ prove that for all $n \in \mathbb{N} \cap (0, m)$ it holds that

$$\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] > 0. \tag{116}$$

Hence, we obtain for all $n \in \mathbb{N} \cap (0, m)$ that

$$\begin{aligned}\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] &= \left[\tfrac{\mathbb{E}[\|\Theta_n - \mathbb{E}[X_1]\|^2]}{n^{-\nu}}\right] n^{-\nu} \\ &\geq \left[\min\left\{\tfrac{\mathbb{E}[\|\Theta_k - \mathbb{E}[X_1]\|^2]}{k^{-\nu}} \colon k \in \mathbb{N} \cap (0, m)\right\}\right] n^{-\nu} > 0.\end{aligned} \tag{117}$$

Combining this, (115), and the hypothesis that $\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] > 0$ assures that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}&\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] \\ &\geq \left[\min\left(\left\{\tfrac{\mathbb{E}[\|\Theta_k - \mathbb{E}[X_1]\|^2]}{k^{-\nu}} \colon k \in \mathbb{N} \cap (0, m)\right\}\right.\right. \\ &\quad \left.\left. \cup \left\{\left[\tfrac{(\gamma\alpha)^2\exp(-2^{\nu}\gamma\alpha)}{4}\right]\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big]\right\}\right)\right] n^{-\nu} > 0.\end{aligned} \tag{118}$$

Therefore, we obtain for all $n \in \mathbb{N}$ that

$$\begin{aligned}&\big(\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big]\big)^{1/2}\\ &\geq n^{-\nu/2}\Bigg[\min\Bigg(\Big\{\tfrac{\mathbb{E}[\|\Theta_k - \mathbb{E}[X_1]\|^2]}{k^{-\nu}} \colon k \in \mathbb{N} \cap (0, m)\Big\}\\ &\quad \cup \Big\{\Big[\tfrac{(\gamma\alpha)^2 \exp(-2^\nu \gamma\alpha)}{4}\Big]\, \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big]\Big\}\Bigg)\Bigg]^{1/2} > 0.\end{aligned} \tag{119}$$

The proof of Proposition 4.5 is thus completed. □

4.2 Refined lower errors estimates in the case of fast decaying learning rates

In this subsection we establish in Lemma 4.9 below a lower error bound for the SGD process in (7) in the case of fast decaying learning rates (corresponding to the case $\nu = 1$ in Setting 2.1). Combining this lower bound with the lower bound from Proposition 4.5 (see Lemma 4.10 below) allows us to establish the refined lower error bound in Proposition 4.11 below.

4.2.1 An estimate for the natural logarithm

In Lemma 4.6 below we recall an elementary and well-known property of the natural logarithm (see, e.g., [1]). Lemma 4.6 will be employed in our proof of Lemma 4.7 which, in turn, will be used to prove Lemma 4.9. For completeness we provide the proof of Lemma 4.6 here.

Lemma 4.6. *It holds for all $x \in (0, \infty)$ that*

$$\ln(x) \geq \frac{(x-1)}{x}. \tag{120}$$

Proof of Lemma 4.6. Throughout this proof let $f \colon (0, \infty) \to \mathbb{R}$ be the function which satisfies for all $x \in (0, \infty)$ that

$$f(x) = \ln(x) - \frac{(x-1)}{x} = \ln(x) - 1 + \frac{1}{x} = \ln(x) - 1 + x^{-1}. \tag{121}$$

Note that

$$f(1) = \ln(1) - \frac{(1-1)}{1} = \ln(1) = 0. \tag{122}$$

Moreover, observe that for all $x \in (0, \infty)$ it holds that

$$f'(x) = \frac{1}{x} - \frac{1}{x^2} = \frac{(x-1)}{x^2}. \tag{123}$$

This ensures that for all $x \in [1, \infty)$ it holds that $f'(x) \geq 0$. The fundamental theorem of calculus and (122) hence imply that for all $x \in [1, \infty)$ it holds that

$$f(x) = f(1) + \int_1^x f'(t) \, \mathrm{d}t \geq f(1) = 0. \tag{124}$$

Moreover, note that (123) assures that for all $x \in (0, 1]$ it holds that $f'(x) \leq 0$. The fundamental theorem of calculus and (122) therefore ensure that for all $x \in (0, 1]$ it holds that

$$0 = f(1) = f(x) + \int_x^1 f'(t) \, \mathrm{d}t \leq f(x). \tag{125}$$

Combining this with (124) proves that for all $x \in (0, \infty)$ it holds that

$$\ln(x) - \frac{(x-1)}{x} = f(x) \geq 0. \tag{126}$$

Therefore, we obtain for all $x \in (0, \infty)$ that

$$\ln(x) \geq \frac{(x-1)}{x}. \tag{127}$$

The proof of Lemma 4.6 is thus completed. □

4.2.2 Errors due to the deterministic gradient descent dynamic

Lemma 4.7. *Let $m \in \mathbb{N}$, $\beta \in (0, \infty) \backslash \{m, m+1, m+2, \ldots\}$. Then for every $\varepsilon \in (0, \infty)$ there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N} \cap [m, \infty)$ it holds that*

$$\left[\prod_{l=m}^{n} \left| 1 - \tfrac{\beta}{l} \right| \right] \geq C n^{-(\beta+\varepsilon)}. \tag{128}$$

Proof of Lemma 4.7. Throughout this proof let $\varepsilon \in (0, \infty)$, let $L \in \mathbb{N} \cap (\max\{m, \beta\}, \infty)$ satisfy that

$$\frac{\beta}{(1 - \frac{\beta}{L})} \leq (\beta + \varepsilon), \tag{129}$$

and let $C \in [0,\infty)$ be given by

$$C = \min\left(\left\{ \tfrac{\left[\prod_{l=m}^{k} |1-\frac{\beta}{l}|\right]}{k^{-(\beta+\varepsilon)}} \colon k \in \mathbb{N} \cap [m, L] \right\} \cup \left\{ \prod_{l=m}^{L} \left|1 - \tfrac{\beta}{l}\right| \right\}\right). \tag{130}$$

Note that the fact that $\beta \notin \{m, m+1, m+2, \ldots\} = \mathbb{N} \cap [m, \infty)$ ensures that for all $l \in \mathbb{N} \cap [m, \infty)$ it holds that

$$\left|1 - \tfrac{\beta}{l}\right| > 0. \tag{131}$$

This and (130) establish that $C > 0$. Moreover, observe that the fact that $L > \beta$ assures that for all $l \in \mathbb{N} \cap [L, \infty)$ it holds that

$$0 = 1 - 1 < 1 - \tfrac{\beta}{L} \leq 1 - \tfrac{\beta}{l} < 1 - 0 = 1. \tag{132}$$

Hence, we obtain that for all $l \in \mathbb{N} \cap [L, \infty)$ it holds that

$$\left(1 - \tfrac{\beta}{l}\right) \in (0, 1). \tag{133}$$

Lemma 4.6 and (129) therefore assure that for all $n \in \mathbb{N} \cap (L, \infty)$ it holds that

$$\begin{aligned}
&\ln\left(\prod_{l=L+1}^{n} \left|1 - \tfrac{\beta}{l}\right|\right) = \sum_{l=L+1}^{n} \ln\left(1 - \tfrac{\beta}{l}\right) \\
&\geq \sum_{l=L+1}^{n} \left(\frac{\left(1 - \frac{\beta}{l}\right) - 1}{\left(1 - \frac{\beta}{l}\right)}\right) = -\left[\sum_{l=L+1}^{n} \left(\frac{1}{l}\left[\frac{\beta}{\left(1 - \frac{\beta}{l}\right)}\right]\right)\right] \\
&\geq -\left[\sum_{l=L+1}^{n} \left(\frac{1}{l}\left[\frac{\beta}{\left(1 - \frac{\beta}{L}\right)}\right]\right)\right] \geq -\left[\sum_{l=L+1}^{n} \frac{(\beta+\varepsilon)}{l}\right] \\
&\geq -(\beta+\varepsilon)\left[\sum_{l=2}^{n} \frac{1}{l}\right].
\end{aligned} \tag{134}$$

The fact that for all $n \in \mathbb{N}$ it holds that

$$\sum_{l=2}^{n} \frac{1}{l} = \sum_{l=2}^{n} \left[\int_{l-1}^{l} \frac{1}{l}\,\mathrm{d}x\right] \leq \sum_{l=2}^{n} \left[\int_{l-1}^{l} \frac{1}{x}\,\mathrm{d}x\right] = \int_{1}^{n} \frac{1}{x}\,\mathrm{d}x = \ln(n) \tag{135}$$

hence ensures that for all $n \in \mathbb{N} \cap (L, \infty)$ it holds that

$$\begin{aligned}
\prod_{l=L+1}^{n} \left|1 - \tfrac{\beta}{l}\right| &= \exp\left(\ln\left(\prod_{l=L+1}^{n} \left|1 - \tfrac{\beta}{l}\right|\right)\right) \\
&\geq \exp\left(-(\beta+\varepsilon)\left[\sum_{l=2}^{n} \frac{1}{l}\right]\right) \\
&\geq \exp\big(-(\beta+\varepsilon)\ln(n)\big) = n^{-(\beta+\varepsilon)}.
\end{aligned} \tag{136}$$

This and (130) demonstrate that for all $n \in \mathbb{N} \cap (L, \infty)$ it holds that

$$\begin{aligned} \prod_{l=m}^{n} \left|1 - \tfrac{\beta}{l}\right| &= \left[\prod_{l=m}^{L} \left|1 - \tfrac{\beta}{l}\right|\right] \left[\prod_{l=L+1}^{n} \left|1 - \tfrac{\beta}{l}\right|\right] \\ &\geq \left[\prod_{l=m}^{L} \left|1 - \tfrac{\beta}{l}\right|\right] n^{-(\beta+\varepsilon)} \geq C n^{-(\beta+\varepsilon)}. \end{aligned} \tag{137}$$

Moreover, note that (130) implies that for all $n \in \mathbb{N} \cap [m, L]$ it holds that

$$\prod_{l=m}^{n} \left|1 - \tfrac{\beta}{l}\right| = \left[\frac{\left[\prod_{l=m}^{n} \left|1 - \frac{\beta}{l}\right|\right]}{n^{-(\beta+\varepsilon)}}\right] n^{-(\beta+\varepsilon)} \geq C n^{-(\beta+\varepsilon)}. \tag{138}$$

Combining this and (137) establishes that for all $n \in \mathbb{N} \cap [m, \infty)$ it holds that

$$\prod_{l=m}^{n} \left|1 - \tfrac{\beta}{l}\right| \geq C n^{-(\beta+\varepsilon)}. \tag{139}$$

The fact that $C > 0$ therefore establishes (128). The proof of Lemma 4.7 is thus completed. □

Lemma 4.8 (Lower bound for deterministic gradient descent)**.** *Let $d \in \mathbb{N}$, $\kappa \in \mathbb{R}$, $\vartheta \in \mathbb{R}^d$, $\xi \in \mathbb{R}^d \backslash \{\vartheta\}$, $\alpha \in (0, \infty)$, $\gamma \in (0, \infty) \backslash \{\frac{1}{\alpha}, \frac{2}{\alpha}, \frac{3}{\alpha}, \ldots\}$, let $\|\cdot\| : \mathbb{R}^d \to [0, \infty)$ be the d-dimensional Euclidean norm, let $f \colon \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that*

$$f(\theta) = \tfrac{\alpha}{2} \|\theta - \vartheta\|^2 + \kappa, \tag{140}$$

and let $\Theta \colon \mathbb{N}_0 \times \Omega \to \mathbb{R}^d$ be the function which satisfies for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \tfrac{\gamma}{n} (\nabla f)(\Theta_{n-1}). \tag{141}$$

Then

(i) it holds that $\{\theta \in \mathbb{R}^d \colon f(\theta) = \inf_{w \in \mathbb{R}^d} f(w)\} = \{\vartheta\}$ and

(ii) for every $\varepsilon \in (0, \infty)$ there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}$ it holds that

$$\|\Theta_n - \vartheta\| \geq C n^{-(\gamma\alpha+\varepsilon)}. \tag{142}$$

Proof of Lemma 4.8. Throughout this proof let $\varepsilon \in (0,\infty)$. Observe that (140) proves item (i). It thus remains to prove item (ii). For this note that Lemma 2.3 and (140) ensure that for all $\theta \in \mathbb{R}^d$ it holds that

$$(\nabla f)(\theta) = \tfrac{\alpha}{2}(2(\theta-\vartheta)) = \alpha(\theta-\vartheta). \tag{143}$$

Therefore, we obtain for all $n \in \mathbb{N}$ that

$$\begin{aligned} \Theta_n - \vartheta &= \Theta_{n-1} - \tfrac{\gamma}{n}(\nabla f)(\Theta_{n-1}) - \vartheta \\ &= \Theta_{n-1} - \vartheta - \tfrac{\gamma\alpha}{n}(\Theta_{n-1}-\vartheta) \\ &= (1-\tfrac{\gamma\alpha}{n})(\Theta_{n-1}-\vartheta). \end{aligned} \tag{144}$$

Induction hence proves that for all $n \in \mathbb{N}$ it holds that

$$\Theta_n - \vartheta = \left[\prod_{l=1}^{n}(1-\tfrac{\gamma\alpha}{l})\right](\Theta_0-\vartheta) = \left[\prod_{l=1}^{n}(1-\tfrac{\gamma\alpha}{l})\right](\xi-\vartheta). \tag{145}$$

This assures that for all $n \in \mathbb{N}$ it holds that

$$\|\Theta_n - \vartheta\| = \left[\prod_{l=1}^{n}\left|1-\tfrac{\gamma\alpha}{l}\right|\right]\|\xi-\vartheta\|. \tag{146}$$

Next observe that Lemma 4.7 (with $m=1$, $\beta=\gamma\alpha$ in the notation of Lemma 4.7) and the fact that $\gamma\alpha \notin \mathbb{N}$ imply that there exists $C \in (0,\infty)$ such that for all $n \in \mathbb{N}$ it holds that

$$\left[\prod_{l=1}^{n}\left|1-\tfrac{\gamma\alpha}{l}\right|\right] \geq Cn^{-(\gamma\alpha+\varepsilon)}. \tag{147}$$

Combining this with (146) demonstrates that for all $n \in \mathbb{N}$ it holds that

$$\|\Theta_n - \vartheta\| \geq \left[Cn^{-(\gamma\alpha+\varepsilon)}\right]\|\xi-\vartheta\| = \left[C\|\xi-\vartheta\|\right]n^{-(\gamma\alpha+\varepsilon)}. \tag{148}$$

The hypothesis that $\xi \neq \vartheta$ hence establishes item (ii). The proof of Lemma 4.8 is thus completed. □

Lemma 4.9. *Assume Setting 2.1 and assume that $\nu = 1$ and $\mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|^2] > 0$. Then for every $\varepsilon \in (0,\infty)$ there exists $C \in (0,\infty)$ such that for all $n \in \mathbb{N}$ it holds that*

$$\left(\mathbb{E}\left[\|\Theta_n - \mathbb{E}[X_1]\|^2\right]\right)^{1/2} \geq Cn^{-(\gamma\alpha+\varepsilon)}. \tag{149}$$

Proof of Lemma 4.9. Throughout this proof let $\varepsilon \in (0,\infty)$, let $m \in \mathbb{N} \cap (\gamma\alpha - 1, \infty)$, and let $\mathcal{M} \in [0,\infty)$ be given by

$$\mathcal{M} = \min\left\{ \frac{\left(\mathbb{E}[\|\Theta_k - \mathbb{E}[X_1]\|^2]\right)^{1/2}}{k^{-(\gamma\alpha+\varepsilon)}} \colon k \in \{1,2,\ldots,m\} \right\}. \tag{150}$$

Observe that Lemma 4.1 and the hypothesis that $\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] > 0$ assure that for all $n \in \{1,2,\ldots,m\}$ it holds that

$$\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] > 0. \tag{151}$$

This ensures that $\mathcal{M} > 0$. Next note that item (iv) in Proposition 2.6 assures that for all $n \in \mathbb{N} \cap (m,\infty)$ it holds that

$$\begin{aligned}
&\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] \\
&= \left[\prod_{l=1}^{n}(1-\tfrac{\gamma\alpha}{l})\right]^2 \|\xi - \mathbb{E}[X_1]\|^2 \\
&\quad + \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] \left[\sum_{k=1}^{n}\left[\tfrac{\gamma\alpha}{k}\left(\prod_{l=k+1}^{n}(1-\tfrac{\gamma\alpha}{l})\right)\right]^2\right] \\
&\geq \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] \left[\tfrac{\gamma\alpha}{m}\left(\prod_{l=m+1}^{n}(1-\tfrac{\gamma\alpha}{l})\right)\right]^2 \\
&= (\tfrac{\gamma\alpha}{m})^2\, \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] \left[\prod_{l=m+1}^{n}\left|1-\tfrac{\gamma\alpha}{l}\right|\right]^2.
\end{aligned} \tag{152}$$

Moreover, observe that the fact that $m+1 > \gamma\alpha$ ensures that $\gamma\alpha \notin \{m+1, m+2, \ldots\}$. Lemma 4.7 (with $m = m+1$, $\beta = \gamma\alpha$ in the notation of Lemma 4.7) therefore demonstrates that there exists $C \in (0,\infty)$ such that for all $n \in \mathbb{N} \cap [m+1,\infty) = \mathbb{N} \cap (m,\infty)$ it holds that

$$\left[\prod_{l=m+1}^{n}\left|1-\tfrac{\gamma\alpha}{l}\right|\right] \geq Cn^{-(\gamma\alpha+\varepsilon)}. \tag{153}$$

Combining this with (152) proves that for all $n \in \mathbb{N} \cap (m,\infty)$ it holds that

$$\begin{aligned}
\left(\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big]\right)^{1/2} &\geq \left[\frac{\gamma\alpha\left(\mathbb{E}[\|X_1-\mathbb{E}[X_1]\|^2]\right)^{1/2}}{m}\right]\left[\prod_{l=m+1}^{n}\left|1-\tfrac{\gamma\alpha}{l}\right|\right] \\
&\geq \left[\frac{\gamma\alpha C\left(\mathbb{E}[\|X_1-\mathbb{E}[X_1]\|^2]\right)^{1/2}}{m}\right] n^{-(\gamma\alpha+\varepsilon)}.
\end{aligned} \tag{154}$$

In addition, note that for all $n \in \{1, 2, \ldots, m\}$ it holds that

$$\left(\mathbb{E}\left[\|\Theta_n - \mathbb{E}[X_1]\|^2\right]\right)^{1/2} = \left[\frac{\left(\mathbb{E}[\|\Theta_n - \mathbb{E}[X_1]\|^2]\right)^{1/2}}{n^{-(\gamma\alpha+\varepsilon)}}\right] n^{-(\gamma\alpha+\varepsilon)} \geq \mathcal{M} n^{-(\gamma\alpha+\varepsilon)}. \tag{155}$$

This and (154) establish that for all $n \in \mathbb{N}$ it holds that

$$\left(\mathbb{E}\left[\|\Theta_n - \mathbb{E}[X_1]\|^2\right]\right)^{1/2} \geq \left[\min\left\{\mathcal{M}, \left[\frac{\gamma\alpha C\left(\mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|^2]\right)^{1/2}}{m}\right]\right\}\right] n^{-(\gamma\alpha+\varepsilon)}. \tag{156}$$

The hypothesis that $\mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|^2] > 0$ and the fact that $\mathcal{M} > 0$ therefore establish (149). The proof of Lemma 4.9 is thus completed. □

4.2.3 Errors due to the randomness in the SGD method

Lemma 4.10. *Assume Setting 2.1 and assume that $\nu = 1$ and $\mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|^2] > 0$. Then there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}$ it holds that*

$$\left(\mathbb{E}\left[\|\Theta_n - \mathbb{E}[X_1]\|^2\right]\right)^{1/2} \geq C n^{-1/2}. \tag{157}$$

Proof of Lemma 4.10. Note that Proposition 4.5 and the hypothesis that $\nu = 1$ demonstrate that there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}$ it holds that

$$\left(\mathbb{E}\left[\|\Theta_n - \mathbb{E}[X_1]\|^2\right]\right)^{1/2} \geq C n^{-\nu/2} = C n^{-1/2}. \tag{158}$$

The proof of Lemma 4.10 is thus completed. □

4.2.4 Composition of the errors

Proposition 4.11. *Assume Setting 2.1 and assume that $\nu = 1$ and $\mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|^2] > 0$. Then for every $\varepsilon \in (0, \infty)$ there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}$ it holds that*

$$\left(\mathbb{E}\left[\|\Theta_n - \mathbb{E}[X_1]\|^2\right]\right)^{1/2} \geq C n^{-\min\{1/2, \gamma\alpha+\varepsilon\}}. \tag{159}$$

Proof of Proposition 4.11. Throughout this proof let $\varepsilon \in (0, \infty)$. Note that Lemma 4.10 demonstrates that there exists $c \in (0, \infty)$ such that for all $n \in \mathbb{N}$ it holds that

$$\left(\mathbb{E}\left[\|\Theta_n - \mathbb{E}[X_1]\|^2\right]\right)^{1/2} \geq c n^{-1/2}. \tag{160}$$

Moreover, observe that Lemma 4.9 assures that there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}$ it holds that

$$\left(\mathbb{E}\left[\|\Theta_n - \mathbb{E}[X_1]\|^2\right]\right)^{1/2} \geq C n^{-(\gamma\alpha+\varepsilon)}. \tag{161}$$

Combining this and (160) ensures that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}\big(\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big]\big)^{1/2} &\geq \max\big\{cn^{-1/2}, Cn^{-(\gamma\alpha+\varepsilon)}\big\} \\ &\geq \min\{c, C\}\max\big\{n^{-1/2}, n^{-(\gamma\alpha+\varepsilon)}\big\} \\ &= \big[\min\{c, C\}\big] n^{\max\{-1/2, -(\gamma\alpha+\varepsilon)\}} \\ &= \big[\min\{c, C\}\big] n^{-\min\{1/2, \gamma\alpha+\varepsilon\}}.\end{aligned} \tag{162}$$

The proof of Proposition 4.11 is thus completed. □

4.3 Lower errors estimates in the case of very fast decaying learning rates

In this subsection we establish in Lemma 4.12 below that the SGD process in (7) fails to converge to the global minimum of the objective function in the case of very fast decaying learning rates (corresponding to the case $\nu > 1$ in Setting 2.1).

Lemma 4.12. *Assume Setting 2.1 and assume that $\nu > 1$ and $\mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|^2] > 0$. Then there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}$ it holds that*

$$\big(\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big]\big)^{1/2} \geq C. \tag{163}$$

Proof of Lemma 4.12. Throughout this proof let $m \in \mathbb{N} \cap (\gamma\alpha, \infty)$. Observe that Lemma 4.1 and the hypothesis that $\mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|^2] > 0$ ensure that for all $n \in \mathbb{N}$ it holds that

$$\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] > 0. \tag{164}$$

Next note that item (iv) in Proposition 2.6 demonstrates that for all $n \in \mathbb{N} \cap (m, \infty)$ it holds that

$$\begin{aligned}\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] &= \left[\prod_{l=1}^{n}(1 - \tfrac{\gamma\alpha}{l^\nu})\right]^2 \|\xi - \mathbb{E}[X_1]\|^2 \\ &\quad + \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] \left[\sum_{k=1}^{n}\left[\tfrac{\gamma\alpha}{k^\nu}\left(\prod_{l=k+1}^{n}(1 - \tfrac{\gamma\alpha}{l^\nu})\right)\right]^2\right] \\ &\geq \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big] \left[\tfrac{\gamma\alpha}{m^\nu}\left(\prod_{l=m+1}^{n}(1 - \tfrac{\gamma\alpha}{l^\nu})\right)\right]^2 \\ &= \left[\tfrac{(\gamma\alpha)^2\mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|^2]}{m^{2\nu}}\right]\left[\prod_{l=m+1}^{n}\big|1 - \tfrac{\gamma\alpha}{l^\nu}\big|\right]^2.\end{aligned} \tag{165}$$

Moreover, note that the fact that $m > \gamma\alpha$ ensures that for all $l \in \mathbb{N} \cap [m, \infty)$ it holds that

$$0 < 1 - \tfrac{\gamma\alpha}{m} \leq 1 - \tfrac{\gamma\alpha}{l} \leq 1 - \tfrac{\gamma\alpha}{l^\nu} < 1. \tag{166}$$

Therefore, we obtain that for all $l \in \mathbb{N} \cap [m, \infty)$ it holds that

$$\left(1 - \tfrac{\gamma\alpha}{l^\nu}\right) \in (0, 1). \tag{167}$$

Lemma 4.6 hence assures that for all $n \in \mathbb{N} \cap (m, \infty)$ it holds that

$$\begin{aligned} &\ln\left(\prod_{l=m+1}^{n} \left|1 - \tfrac{\gamma\alpha}{l^\nu}\right|\right) = \sum_{l=m+1}^{n} \ln\left(1 - \tfrac{\gamma\alpha}{l^\nu}\right) \\ &\geq \sum_{l=m+1}^{n} \left(\frac{\left(1 - \frac{\gamma\alpha}{l^\nu}\right) - 1}{\left(1 - \frac{\gamma\alpha}{l^\nu}\right)}\right) = -\left[\sum_{l=m+1}^{n} \left(\frac{\gamma\alpha}{l^\nu \left(1 - \frac{\gamma\alpha}{l^\nu}\right)}\right)\right] \\ &\geq -\left[\frac{\gamma\alpha}{\left(1 - \frac{\gamma\alpha}{m^\nu}\right)}\right] \left[\sum_{l=m+1}^{n} \frac{1}{l^\nu}\right] \geq -\left[\frac{\gamma\alpha}{\left(1 - \frac{\gamma\alpha}{m^\nu}\right)}\right] \left[\sum_{l=2}^{\infty} \frac{1}{l^\nu}\right]. \end{aligned} \tag{168}$$

In addition, note that the hypothesis that $\nu > 1$ implies that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \sum_{l=2}^{\infty} \frac{1}{l^\nu} &= \sum_{l=2}^{\infty} \left[\int_{l-1}^{l} \frac{1}{l^\nu} \, \mathrm{d}x\right] \leq \sum_{l=2}^{\infty} \left[\int_{l-1}^{l} \frac{1}{x^\nu} \, \mathrm{d}x\right] \\ &= \int_1^\infty x^{-\nu} \, \mathrm{d}x = \left[\left(\tfrac{1}{(1-\nu)}\right) x^{1-\nu}\right]_{x=1}^{x=\infty} = -\tfrac{1}{(1-\nu)} = \tfrac{1}{(\nu-1)}. \end{aligned} \tag{169}$$

This and (168) prove that for all $n \in \mathbb{N} \cap (m, \infty)$ it holds that

$$\ln\left(\prod_{l=m+1}^{n} \left|1 - \tfrac{\gamma\alpha}{l^\nu}\right|\right) \geq -\left[\frac{\gamma\alpha}{\left(1 - \frac{\gamma\alpha}{m^\nu}\right)}\right] \frac{1}{(\nu - 1)} = \frac{-\gamma\alpha}{\left(1 - \frac{\gamma\alpha}{m^\nu}\right)(\nu - 1)}. \tag{170}$$

Combining this and (167) with (165) demonstrates that for all $n \in \mathbb{N} \cap (m, \infty)$ it holds that

$$\begin{aligned} \mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] &\geq \left[\tfrac{(\gamma\alpha)^2 \mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|^2]}{m^{2\nu}}\right] \exp\left(\ln\left(\left[\prod_{l=m+1}^{n} \left|1 - \tfrac{\gamma\alpha}{l^\nu}\right|\right]^2\right)\right) \\ &= \left[\tfrac{(\gamma\alpha)^2 \mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|^2]}{m^{2\nu}}\right] \exp\left(2 \ln\left(\prod_{l=m+1}^{n} \left|1 - \tfrac{\gamma\alpha}{l^\nu}\right|\right)\right) \\ &\geq \left[\tfrac{(\gamma\alpha)^2 \mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|^2]}{m^{2\nu}}\right] \exp\left(\tfrac{-2\gamma\alpha}{\left(1 - \frac{\gamma\alpha}{m^\nu}\right)(\nu-1)}\right). \end{aligned} \tag{171}$$

The hypothesis that $\mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|^2] > 0$, (164), and (167) hence establish that for all $n \in \mathbb{N}$ it holds that

$$\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big] \geq \min\Bigg(\Big\{\mathbb{E}\big[\|\Theta_k - \mathbb{E}[X_1]\|^2\big] : k \in \{1, 2, \ldots, m\}\Big\} \cup \left\{\left[\tfrac{(\gamma\alpha)^2 \mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|^2]}{m^{2\nu}}\right] \exp\left(\tfrac{-2\gamma\alpha}{(1 - \frac{\gamma\alpha}{m^\nu})(\nu - 1)}\right)\right\}\Bigg) > 0. \quad (172)$$

The proof of Lemma 4.12 is thus completed. □

4.4 Main result of this article

The following theorem summarizes the main findings of this article.

Theorem 4.13. *Let $d \in \mathbb{N}$, $\alpha, \gamma, \nu \in (0, \infty)$, $\xi \in \mathbb{R}^d$, let $\langle \cdot, \cdot \rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the d-dimensional Euclidean scalar product, let $\|\cdot\| : \mathbb{R}^d \to [0, \infty)$ be the d-dimensional Euclidean norm, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X_n \colon \Omega \to \mathbb{R}^d$, $n \in \mathbb{N}$, be i.i.d. random variables with $\mathbb{E}[\|X_1\|^2] < \infty$ and $\mathbb{P}(X_1 = \mathbb{E}[X_1]) < 1$, let $(r_{\varepsilon,i})_{\varepsilon \in (0,\infty), i \in \{0,1\}} \subseteq \mathbb{R}$ satisfy for all $\varepsilon \in (0, \infty)$, $i \in \{0, 1\}$ that*

$$r_{\varepsilon,i} = \begin{cases} \nu/2 & : \nu < 1 \\ \min\{1/2, \gamma\alpha + (-1)^i \varepsilon\} & : \nu = 1 \\ 0 & : \nu > 1, \end{cases} \quad (173)$$

let $F = (F(\theta, x))_{(\theta,x) \in \mathbb{R}^d \times \mathbb{R}^d} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $f \colon \mathbb{R}^d \to \mathbb{R}$ be the functions which satisfy for all $\theta, x \in \mathbb{R}^d$ that

$$F(\theta, x) = \tfrac{\alpha}{2}\|\theta - x\|^2 \qquad \text{and} \qquad f(\theta) = \mathbb{E}\big[F(\theta, X_1)\big], \quad (174)$$

and let $\Theta \colon \mathbb{N}_0 \times \Omega \to \mathbb{R}^d$ be the stochastic process which satisfies for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \tfrac{\gamma}{n^\nu}(\nabla_\theta F)(\Theta_{n-1}, X_n). \quad (175)$$

Then

(i) it holds for all $\theta \in \mathbb{R}^d$ that $f(\theta) = \frac{\alpha}{2}\|\theta - \mathbb{E}[X_1]\|^2 + \frac{\alpha}{2}\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big]$,

(ii) it holds that $\{\theta \in \mathbb{R}^d \colon f(\theta) = \inf_{w \in \mathbb{R}^d} f(w)\} = \{\mathbb{E}[X_1]\}$,

(iii) it holds for all $\theta \in \mathbb{R}^d$ that $\langle \theta - \mathbb{E}[X_1], (\nabla f)(\theta) \rangle = \alpha\|\theta - \mathbb{E}[X_1]\|^2$,

(iv) *it holds for all* $\theta \in \mathbb{R}^d$ *that* $\|(\nabla f)(\theta)\| = \alpha\|\theta - \mathbb{E}[X_1]\|$,

(v) *it holds for all* $\theta \in \mathbb{R}^d$ *that*

$$\mathbb{E}\big[\|(\nabla_\theta F)(\theta, X_1) - (\nabla f)(\theta)\|^2\big] = \alpha^2\,\mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big], \tag{176}$$

(vi) *for every* $\varepsilon \in (0,\infty)$ *there exist* $C_0, C_1 \in (0,\infty)$ *such that for all* $n \in \mathbb{N}$ *it holds that*

$$C_0 n^{-r_{\varepsilon,0}} \leq \big(\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big]\big)^{1/2} \leq C_1 n^{-r_{\varepsilon,1}}, \tag{177}$$

and

(vii) *for every* $\varepsilon \in (0,\infty)$ *there exist* $C_0, C_1 \in (0,\infty)$ *such that for all* $n \in \mathbb{N}$ *it holds that*

$$C_0 n^{-2r_{\varepsilon,0}} \leq \mathbb{E}[f(\Theta_n)] - f(\mathbb{E}[X_1]) \leq C_1 n^{-2r_{\varepsilon,1}}. \tag{178}$$

Proof of Theorem 4.13. First, note that items (i), (ii), (v), (vi), and (vii) in Lemma 2.4 establish items (i)–(v). In addition, observe that the hypothesis that $\mathbb{P}(X_1 = \mathbb{E}[X_1]) < 1$ ensures that $\mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|^2] > 0$. Proposition 3.3, Corollary 3.7, Lemma 3.8, Proposition 4.5, Proposition 4.11, and Lemma 4.12 therefore prove item (vi). Moreover, note that item (i) ensures that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}\mathbb{E}[f(\Theta_n)] - f(\mathbb{E}[X_1]) &= \mathbb{E}\big[\tfrac{\alpha}{2}\big(\|\Theta_n - \mathbb{E}[X_1]\|^2 + \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big]\big)\big] \\ &\quad - \tfrac{\alpha}{2}\big(\|\mathbb{E}[X_1] - \mathbb{E}[X_1]\|^2 + \mathbb{E}\big[\|X_1 - \mathbb{E}[X_1]\|^2\big]\big) \\ &= \tfrac{\alpha}{2}\,\mathbb{E}\big[\|\Theta_n - \mathbb{E}[X_1]\|^2\big]\,.\end{aligned} \tag{179}$$

Combining this and item (vi) establishes item (vii). The proof of Theorem 4.13 is thus completed. □

References

[1] Lower bound of natural logarithm - pr∞fwiki. `https://proofwiki.org/wiki/Lower_Bound_of_Natural_Logarithm`. [Accessed 08-March-2018].

[2] Agarwal, A., and Bottou, L. A lower bound for the optimization of finite sums. *arXiv:1410.0723* (2014), 19 pages.

[3] Bach, F., and Moulines, E. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems (NIPS)* (2011).

[4] Bach, F. R., and Moulines, E. Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). *arXiv:1306.2119* (2013), 42 pages.

[5] Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Physica-Verlag/Springer, Heidelberg, 2010, pp. 177–186.

[6] Bottou, L., and Bousquet, O. The tradeoffs of large scale learning. *Optimization for Machine Learning, MIT Press* (2011), 351–368.

[7] Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *arXiv:1606.04838* (2016), 95 pages.

[8] Bottou, L., and LeCun, Y. Large scale online learning. *In Thrun, Sebastian, Saul, Lawrence and Schölkopf, Bernhard (eds.), Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA* (2004).

[9] Chau, H. N., Kumar, C., Rásonyi, M., and Sabanis, S. On fixed gain recursive estimators with discontinuity in the parameters. *preprint, arXiv:1609.05166* (2017).

[10] Dereich, S., and Mueller-Gronbach, T. General multilevel adaptations for stochastic approximation algorithms. *arXiv:1506.05482* (2017), 33 pages.

[11] Dieuleveut, A., Durmus, A., and Bach, F. Bridging the gap between constant step size stochastic gradient descent and markov chains. *preprint, hal-01565514* (2017).

[12] Graves, A., Mohamed, A.-r., and Hinton, G. Speech recognition with deep recurrent neural networks. *Acoustics, Speech and Signal Processing (ICASSP)* (2013), 6645–6649.

[13] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., and Sainath, T. N. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE 29*, 6 (2012), 82–97.

[14] Jentzen, A., Kuckuck, B., Neufeld, A., and von Wurstemberger, P. Strong error analysis for stochastic gradient descent optimization algorithms. *arXiv:1801.09324* (2018), 75 pages.

[15] Klenke, A. *Probabilitly Theory*, 2 ed. Universitext. Springer-Verlag London Ltd., 2014.

[16] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[17] Lan, G., and Zhou, Y. An optimal randomized incremental gradient method. *Mathematical programming* (2017), 1–49.

[18] Li, Q., Tai, C., and E, W. Dynamics of stochastic gradient algorithms. *arXiv:1511.06251* (2015), 29 pages.

[19] Müller-Gronbach, T., and Ritter, K. Minimal errors for strong and weak approximation of stochastic differential equations. In *Monte Carlo and Quasi-Monte Carlo Methods 2006*. Springer, 2008, pp. 53–82.

[20] Murata, N. A statistical study of on-line learning. *Online Learning and Neural Networks. Cambridge University Press, Cambridge, UK* (1998), 63–92.

[21] Nguyen, L. M., Nguyen, N. H., Phan, D. T., Kalagnanam, J. R., and Scheinberg, K. When does stochastic gradient algorithm work well? *arXiv:1801.06159* (2018), 22 pages.

[22] Rakhlin, A., Shamir, O., Sridharan, K., et al. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML* (2012), Citeseer.

[23] Ruder, S. An overview of gradient descent optimization algorithms. *arXiv:1609.04747* (2016), 12 pages.

[24] Tang, C., and Monteleoni, C. On the convergence rate of stochastic gradient descent for strongly convex functions. In *Regularization, optimization, kernels, and support vector machines*, Chapman & Hall/CRC Mach. Learn. Pattern Recogn. Ser. CRC Press, Boca Raton, FL, 2015, pp. 159–175.

[25] Woodworth, B. E., and Srebro, N. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 3639–3647.