Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Seminar for
Applied
Mathematics
SAM

Strong error analysis for stochastic gradient descent optimization algorithms

A. Jentzen and B. Kuckuck and P. von Wurstemberger

Research Report No. 2018-03
January 2018

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

Strong error analysis for stochastic gradient descent optimization algorithms

Arnulf Jentzen[1], Benno Kuckuck[2],
Ariel Neufeld[3], and Philippe von Wurstemberger[4]

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

[2]Department of Mathematics, Universität Düsseldorf,
e-mail: kuckuck@math.uni-duesseldorf.de

[3]Department of Mathematics, ETH Zurich,
e-mail: ariel.neufeld@math.ethz.ch

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

January 31, 2018

Abstract

Stochastic gradient descent (SGD) optimization algorithms are key ingredients in a series of machine learning applications. In this article we perform a rigorous strong error analysis for SGD optimization algorithms. In particular, we prove for every arbitrarily small $\varepsilon \in (0, \infty)$ and every arbitrarily large $p \in (0, \infty)$ that the considered SGD optimization algorithm converges in the strong L^p-sense with order $1/2 - \varepsilon$ to the global minimum of the objective function of the considered stochastic approximation problem under standard convexity-type assumptions on the objective function and relaxed assumptions on the moments of the stochastic errors appearing in the employed SGD optimization algorithm. The key ideas in our convergence proof are, first, to employ techniques from the theory of Lyapunov-type functions for dynamical systems to develop a general convergence machinery for SGD optimization algorithms based on such functions, then, to apply this general machinery to concrete Lyapunov-type functions with polynomial structures, and, thereafter, to perform an induction argument along the powers appearing in the Lyapunov-type functions in order to achieve for every arbitrarily large $p \in (0, \infty)$ strong L^p-convergence rates. This article also contains an extensive review of results on SGD optimization algorithms in the scientific literature.

Contents

1 Introduction

Stochastic gradient descent (SGD) type optimization algorithms are fundamental tools in many machine and deep learning applications such as object and speech recognition or image analysis (cf., for example, Ruder [92]). To ensure the performance of such algorithms it is important to analyze their approximation errors and, in particular, to investigate their speeds of convergence. A very common approach to study SGD type optimization algorithms is to formulate them as so-called stochastic approximation algorithms (SAAs). SAAs were first introduced in Robbins & Monro [91] and SAAs and SGD type optimization algorithms, respectively, have been widely studied in the scientific literature; cf., for example, [2, 19, 39, 63, 64, 71, 74, 75, 77, 83, 93, 99, 102, 107] and the references mentioned therein for the derivation and the proposal of SAAs, cf., for example, [10, 26, 27, 29, 35, 36, 49, 60, 65, 72, 79, 85, 86, 87, 98, 100, 101, 103, 103, 109, 110] and the references mentioned therein for the derivation and the proposal of SGD type optimization algorithms, cf., for example, [1, 12, 17, 18, 20, 21, 23, 31, 33, 34, 40, 41, 48, 51, 52, 55, 56, 59, 61, 62, 68, 70, 76, 78, 81, 82, 94] and the references mentioned therein for numerical simulations and convergence rates proofs for SAAs, cf., for example, [3, 4, 5, 13, 14, 16, 25, 32, 46, 67, 73, 80, 84, 88, 89, 90, 104, 105, 106, 108, 111] and the references mentioned therein for numerical simulations and convergence rates proofs for SGD type optimization algorithms, cf., for example, [6, 7, 9, 11, 22, 37, 38, 54, 57, 58, 69, 95, 97] and the references mentioned therein for overview articles and monographs on SAAs, cf., for example, [8, 15, 92] and the references mentioned therein for overview articles on SGD type optimization algorithms, and cf., for example, [28, 30, 42, 43, 44, 45, 53, 66, 96] and the references mentioned therein for applications involving neural networks and SGD type optimization algorithms.

In this paper we develop a rigorous strong error analysis for SAAs and SGD optimization algorithms. In particular, we prove for every arbitrarily small $\varepsilon \in (0, \infty)$ and every arbitrarily large $p \in (0, \infty)$ that the considered SGD optimization algorithm converges in the strong L^p-sense with order ${}^1\!/{}_2 - \varepsilon$ to the global minimum of the objective function of the considered stochastic approximation problem under standard convexity-type assumptions on the objective function (cf. (2) in Theorem 1.1 below, (214) in Theorem 3.7 in Subsection 3.4 below, and, e.g., Dereich & Mueller-Gronbach [31, Assumption A.1]) and relaxed assumptions on the moments of the stochastic errors appearing in the employed SGD optimization algorithm (cf. (3) in Theorem 1.1 below and (217) in Theorem 3.7 in Subsection 3.4 below). To illustrate the findings of this article, we now present in the following theorem a special case of our strong error analysis for SAAs and SGD optimization algorithms (cf. Theorem 3.7 in Subsection 3.4 below and Corollary 4.9 in Subsection 4.2 below).

Theorem 1.1. *Let $d \in \mathbb{N}$, $p, \alpha, \kappa, c \in (0, \infty)$, $\nu \in (0,1)$, $q = \min(\{2, 4, 6, \dots\} \cap [p, \infty))$, $\xi, \vartheta \in \mathbb{R}^d$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, let $X_n \colon \Omega \to S$, $n \in \mathbb{N}$, be i.i.d. random variables, let $F = (F(\theta, x))_{\theta \in \mathbb{R}^d, x \in S} \colon \mathbb{R}^d \times S \to \mathbb{R}$ be $(\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{S})/\mathcal{B}(\mathbb{R})$-measurable, assume for all $x \in S$ that $(\mathbb{R}^d \ni \theta \mapsto F(\theta, x) \in \mathbb{R}) \in C^1(\mathbb{R}^d, \mathbb{R})$, assume for all $\theta \in \mathbb{R}^d$ that*

$$\mathbb{E}\big[|F(\theta, X_1)| + \|(\nabla_\theta F)(\theta, X_1)\|_{\mathbb{R}^d}\big] < \infty, \tag{1}$$

$$\langle \theta - \vartheta, \mathbb{E}[(\nabla_\theta F)(\theta, X_1)] \rangle_{\mathbb{R}^d} \geq c \max\big\{\|\theta - \vartheta\|^2_{\mathbb{R}^d}, \|\mathbb{E}[(\nabla_\theta F)(\theta, X_1)]\|^2_{\mathbb{R}^d}\big\}, \tag{2}$$

$$\mathbb{E}\big[\|(\nabla_\theta F)(\theta, X_1) - \mathbb{E}[(\nabla_\theta F)(\theta, X_1)]\|^q_{\mathbb{R}^d}\big] \leq \kappa\big(1 + \|\theta\|^q_{\mathbb{R}^d}\big), \tag{3}$$

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

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \tfrac{\alpha}{n^\nu}(\nabla_\theta F)(\Theta_{n-1}, X_n). \tag{4}$$

Then

(i) it holds that $\big\{\theta \in \mathbb{R}^d \colon \big(\mathbb{E}[F(\theta, X_1)] = \inf_{v \in \mathbb{R}^d} \mathbb{E}[F(v, X_1)]\big)\big\} = \{\vartheta\}$ and

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

$$\big(\mathbb{E}\big[\|\Theta_n - \vartheta\|^p_{\mathbb{R}^d}\big]\big)^{1/p} \leq C n^{-\nu/2}. \tag{5}$$

Theorem 1.1 is an immediate consequence of Jensen's inequality and Corollary 4.9 in Subsection 4.2 below. Corollary 4.9, in turn, follows from Theorem 3.7 in Subsection 3.4 below, which is the main result of this article. A strong convergence result related to Theorems 1.1 and 3.7 in this article has been obtained in Dereich & Mueller-Gronbach [31, Theorem 2.4] (cf. also [31, Proposition 2.2]). A key difference between Theorem 2.4 in [31] and Theorems 1.1 and 3.7 in this article is the hypothesis on the moments of the stochastic errors appearing in the employed SAA (cf. item (ii) in Assumption A.2 in [31] with (3) in Theorem 1.1 and (217) in Theorem 3.7 in this article). More formally, in Theorem 2.4 in [31] the stochastic errors appearing in the employed SAA are assumed to be bounded in the state space variable $\theta \in \mathbb{R}^d$ (cf. item (ii) in Assumption A.2 in [31]) while Theorems 1.1 and 3.7 in this article allow the L^p-norm of the stochastic errors to grow linearly in the state space variable $\theta \in \mathbb{R}^d$ (cf. (3) in Theorem 1.1 and (217) in Theorem 3.7 in this article). This relaxed hypothesis enables us to achieve for every arbitrarily small $\varepsilon \in (0, \infty)$ and every arbitrarily large $p \in (0, \infty)$ the essentially sharp strong L^p-convergence rate $1/2 - \varepsilon$ in the case of very natural stochastic optimization examples with quadratically growing loss functions such as in the case of linear regression; see Corollary 4.11 in Subsection 4.3 below for details. The key ideas in our proofs of

Theorem 1.1 and Theorem 3.7, respectively, are, first, to employ techniques from the theory of Lyapunov-type functions for dynamical systems to develop a general convergence result for SAAs and SGD optimization algorithms based on such functions (see Proposition 3.2 and Corollary 3.3 in Subsection 3.2 below for details), then, to apply this general convergence result to concrete Lyapunov-type functions of the form $\mathbb{R}^d \ni \theta \mapsto V_q(\theta) = \|\theta - \vartheta\|_{\mathbb{R}^d}^q \in [0, \infty)$ for $q \in \{2, 4, 6, 8, \dots\}$ (see (168) in the proof of Proposition 3.4 in Subsection 3.3 as well as (212) in the proof of Proposition 3.6 in Subsection 3.4 below for details), and, thereafter, to perform an induction argument on $q \in \{2, 4, 6, 8, \dots\} \cap [0, p]$ in order to establish for every arbitrarily large $p \in (0, \infty)$ strong L^p-convergence rates. In previous error analysis results for SAAs and SGD optimization algorithms in the literature induction arguments have been frequently employed along the time variable (cf., e.g., also Lemma 2.17 in Subsection 2.6 below as well as (139) in the proof of Proposition 3.2 in Subsection 3.2 below). A key idea in this work is to perform an induction argument along the powers $q \in \{2, 4, 6, 8, \dots\}$ appearing in the Lyapunov-type functions $\mathbb{R}^d \ni \theta \mapsto V_q(\theta) = \|\theta - \vartheta\|_{\mathbb{R}^d}^q \in [0, \infty)$.

The remainder of this article is organized as follows. In Section 2 we present several auxiliary results which we employ in our strong L^p-error analysis. In Section 3 we develop our strong L^p-error analysis for general SAAs. In particular, in Subsection 3.4 of Section 3 we present and prove Theorem 3.7, which is the main result of this article. In Section 4 we specialize the abstract findings of Section 3 to SGD optimization algorithms. In particular, in Corollary 4.9 in Subsection 4.2 we establish for every arbitrarily large $p \in (0, \infty)$ strong L^p-convergence rates for SGD optimization algorithms. Theorem 1.1 above is in immediate consequence of Jensen's inequality and Corollary 4.9 in Subsection 4.2 below. In Subsection 4.3 we also illustrate the statement of Corollary 4.9 by means of a simple example.

2 Auxiliary Results

2.1 Norms on Euclidean spaces

In this subsection we establish in Lemmas 2.1–2.4 below some elementary and essentially well-known results for norms in Euclidean spaces. Lemmas 2.1, 2.3, and 2.4 are used in our strong error analysis for SGD methods in Proposition 3.6 in Subsection 3.4 below. Lemma 2.2, in turn, is employed in the proof of Lemma 2.3.

Lemma 2.1 (Convexity of powers of the norm)**.** *Let $d \in \mathbb{N}$, $p \in [1, \infty)$, $v, w \in \mathbb{R}^d$*

and let $\|\cdot\| \colon \mathbb{R}^d \to [0,\infty)$ *be a norm. Then*

$$\begin{split}\|v+w\|^p &\leq \left[\sup_{x,y\in(0,\infty)} \frac{(x+y)^p}{(x^p+y^p)}\right] (\|v\|^p + \|w\|^p) \\ &= 2^{p-1}(\|v\|^p + \|w\|^p) \leq 2^p(\|v\|^p + \|w\|^p).\end{split} \tag{6}$$

Proof of Lemma 2.1. Throughout this proof assume w.l.o.g. that $p > 1$ and let $f \colon (0,\infty) \to \mathbb{R}$ be the function which satisfies for all $t \in (0,\infty)$ that

$$f(t) = \frac{(1+t)^p}{1+t^p}. \tag{7}$$

Note that

$$\sup_{x,y\in(0,\infty)} \left[\frac{(x+y)^p}{x^p+y^p}\right] = \sup_{x,t\in(0,\infty)} \left[\frac{(x+tx)^p}{x^p+(tx)^p}\right] = \sup_{t\in(0,\infty)} \left[\frac{(1+t)^p}{1+t^p}\right] = \sup_{t\in(0,\infty)} f(t). \tag{8}$$

Next observe that for all $t \in (0,\infty)$ it holds that

$$\begin{split}f'(t) &= \frac{p(1+t)^{p-1}(1+t^p) - p(1+t)^p t^{p-1}}{(1+t^p)^2} \\ &= \frac{p(1+t)^{p-1}(1+t^p-(1+t)t^{p-1})}{(1+t^p)^2} \\ &= \frac{p(1+t)^{p-1}(1-t^{p-1})}{(1+t^p)^2}.\end{split} \tag{9}$$

This implies that

$$\{t \in (0,\infty) \colon f'(t) = 0\} = \{1\}. \tag{10}$$

Moreover, observe that the fact that for all $t \in (0,\infty)$ it holds that

$$f(t) = \frac{(1+t)^p}{1+t^p} = \frac{(1+{}^1\!/_t)^p}{1+{}^1\!/_{t^p}} \tag{11}$$

assures that $\lim_{t\searrow 0} f(t) = \lim_{t\to\infty} f(t) = 1$. The fact that $f(1) = 2^{p-1} > 1$, (8), and (10) hence ensure that

$$\sup_{x,y\in(0,\infty)} \left[\frac{(x+y)^p}{x^p+y^p}\right] = \sup_{t\in(0,\infty)} f(t) = f(1) = 2^{p-1}. \tag{12}$$

Therefore, we obtain that

$$\begin{aligned} \|v+w\|^p &\leq \big(\|v\|+\|w\|\big)^p \\ &\leq \left[\sup_{x,y\in(0,\infty)} \frac{(x+y)^p}{x^p+y^p}\right]\big(\|v\|^p+\|w\|^p\big) \\ &= 2^{p-1}\big(\|v\|^p+\|w\|^p\big). \end{aligned} \tag{13}$$

The proof of Lemma 2.1 is thus completed. □

Lemma 2.2. *Let $p \in \mathbb{N}$. Then it holds for all $x, y \in [0,\infty)$ that*

$$|x^p - y^p| \leq 2^p|x-y|\big(\min\{x^{p-1}, y^{p-1}\} + |x-y|^{p-1}\big). \tag{14}$$

Proof of Lemma 2.2. First, observe that for all $x, y \in [0,\infty)$ with $x \geq y$ it holds that

$$\begin{aligned} |x^p - y^p| &= x^p - y^p = (y + (x-y))^p - y^p \\ &= \left[\sum_{k=0}^{p} \binom{p}{k} y^{p-k}(x-y)^k\right] - y^p \\ &= \left[\sum_{k=1}^{p} \binom{p}{k} y^{p-k}(x-y)^k\right] \\ &= (x-y)\left[\sum_{k=1}^{p} \binom{p}{k} y^{p-k}(x-y)^{k-1}\right]. \end{aligned} \tag{15}$$

This demonstrates that for all $x, y \in [0,\infty)$ with $x \geq y$ it holds that

$$\begin{aligned} |x^p - y^p| &\leq (x-y)\left[\sum_{k=1}^{p} \binom{p}{k} \big[\max\{y, x-y\}\big]^{p-1}\right] \\ &\leq (x-y)\max\big\{y^{p-1}, (x-y)^{p-1}\big\}\left[\sum_{k=0}^{p} \binom{p}{k}\right] \\ &= 2^p(x-y)\max\big\{y^{p-1}, (x-y)^{p-1}\big\} \\ &\leq 2^p(x-y)(y^{p-1} + (x-y)^{p-1}) \\ &= 2^p|x-y|(y^{p-1} + |x-y|^{p-1}). \end{aligned} \tag{16}$$

Hence, we obtain that for all $x, y \in [0,\infty)$ with $x \leq y$ it holds that

$$|x^p - y^p| = |y^p - x^p| \leq 2^p|y-x|(x^{p-1} + |y-x|^{p-1}) \tag{17}$$

This and (16) establish (14). The proof of Lemma 2.2 is thus completed. □

Lemma 2.3. *Let $d, p \in \mathbb{N}$, $v, w \in \mathbb{R}^d$ and let $\|\cdot\| \colon \mathbb{R}^d \to [0,\infty)$ be a norm. Then*

$$\begin{aligned}\big|\|v\|^p - \|w\|^p\big| &\leq 2^p \|v-w\| \big(\min\{\|v\|^{p-1}, \|w\|^{p-1}\} + \|v-w\|^{p-1}\big) \\ &\leq 2^p \|v-w\| \big(\|w\|^{p-1} + \|v-w\|^{p-1}\big).\end{aligned} \tag{18}$$

Proof of Lemma 2.3. Observe that Lemma 2.2 ensures that

$$\begin{aligned}\big|\|v\|^p - \|w\|^p\big| &\leq 2^p \big|\|v\| - \|w\|\big| \big(\min\{\|v\|^{p-1}, \|w\|^{p-1}\} + \big|\|v\| - \|w\|\big|^{p-1}\big) \\ &\leq 2^p \|v-w\| \big(\min\{\|v\|^{p-1}, \|w\|^{p-1}\} + \|v-w\|^{p-1}\big).\end{aligned} \tag{19}$$

The proof of Lemma 2.3 is thus completed. □

Lemma 2.4 (Derivative of the norm)**.** *Let $d \in \mathbb{N}$, $p \in \{2, 3, \ldots\}$, $\vartheta \in \mathbb{R}^d$, let $\langle\cdot,\cdot\rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a scalar product, let $\|\cdot\| \colon \mathbb{R}^d \to [0,\infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that $\|\theta\| = \sqrt{\langle\theta,\theta\rangle}$, and let $V \colon \mathbb{R}^d \to [0,\infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that $V(\theta) = \|\theta + \vartheta\|^p$. Then*

(i) it holds that $V \in C^1(\mathbb{R}^d, [0,\infty))$ and

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

$$V'(\theta)(v) = p\|\theta+\vartheta\|^{p-2}\langle\theta+\vartheta, v\rangle. \tag{20}$$

Proof of Lemma 2.4. Throughout this proof assume w.l.o.g. that $p \geq 3$ and let $f \colon \mathbb{R}^d \to [0,\infty)$ and $g \colon \mathbb{R} \to [0,\infty)$ be the functions which satisfy for all $\theta \in \mathbb{R}^d$, $x \in \mathbb{R}$ that

$$f(\theta) = \|\theta+\vartheta\|^2 \qquad \text{and} \qquad g(x) = |x|^{p/2}. \tag{21}$$

Note that for all $x \in \mathbb{R}$ it holds that $g \in C^1(\mathbb{R}, [0,\infty))$ and

$$g'(x) = \begin{cases} \frac{p}{2}\,|x|^{p/2-1} & : x \geq 0 \\ -\frac{p}{2}\,|x|^{p/2-1} & : x < 0. \end{cases} \tag{22}$$

The chain rule hence implies that for all $\theta, v \in \mathbb{R}^d$ it holds that $g \circ f \in C^1(\mathbb{R}^d, [0,\infty))$ and

$$\begin{aligned}\big((g\circ f)'(\theta)\big)(v) &= \tfrac{p}{2}\,\big|\|\theta+\vartheta\|^2\big|^{p/2-1}\big(2\langle\theta+\vartheta, v\rangle\big) \\ &= p\|\theta+\vartheta\|^{p-2}\langle\theta+\vartheta, v\rangle.\end{aligned} \tag{23}$$

Combining this with the fact that $V = g \circ f$ completes the proof of Lemma 2.4. □

2.2 Conditional expectation

In this subsection we present in Lemma 2.5 a well-known property associated to conditional expectations (cf., e.g., Klenke [50, Theorem 8.14]), which we employ in our strong error analyses in Propositions 3.4 and 3.6 below. For completeness we also provide the proof of Lemma 2.5 in this subsection.

Lemma 2.5. *Let $d \in \mathbb{N}$, let $\langle \cdot, \cdot \rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a scalar product, let $\|\cdot\| \colon \mathbb{R}^d \to [0, \infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that $\|\theta\| = \sqrt{\langle \theta, \theta \rangle}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{G} \subseteq \mathcal{F}$ be a sigma-algebra on Ω, let $X \colon \Omega \to \mathbb{R}^d$ be $\mathcal{F}/\mathcal{B}(\mathbb{R}^d)$-measurable, let $Y \colon \Omega \to \mathbb{R}^d$ be $\mathcal{G}/\mathcal{B}(\mathbb{R}^d)$-measurable, and assume for all $A \in \mathcal{G}$ that*

$$\mathbb{E}\big[\|X\| + \|Y\| + \|X\|\|Y\|\big] < \infty \qquad \text{and} \qquad \mathbb{E}\big[X \mathbb{1}_A\big] = 0. \tag{24}$$

Then it holds for all $A \in \mathcal{G}$ that

$$\mathbb{E}\big[|\langle X, Y \rangle|\big] < \infty \qquad \text{and} \qquad \mathbb{E}\big[\langle X, Y \rangle \mathbb{1}_A\big] = 0. \tag{25}$$

Proof of Lemma 2.5. Throughout this proof let $c \in (0, \infty)$ satisfy

$$c = \sup_{\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d \backslash \{0\}} \left(\frac{\big(\sum_{i=1}^d |\theta_i| \big)}{\|\theta\|} \right), \tag{26}$$

let $X_i \colon \Omega \to \mathbb{R}$, $i \in \{1, 2, \dots, d\}$, and $Y_i \colon \Omega \to \mathbb{R}$, $i \in \{1, 2, \dots, d\}$, be the functions which satisfy that

$$X = (X_1, X_2, \dots, X_d) \qquad \text{and} \qquad Y = (Y_1, Y_2, \dots, Y_d), \tag{27}$$

let $e_1 = (1, 0, \dots, 0)$, $e_2 = (0, 1, 0, \dots, 0)$, $\dots$, $e_d = (0, \dots, 0, 1) \in \mathbb{R}^d$, and let $M = (M_{i,j})_{(i,j) \in \{1,2,\dots,d\}^2} \in \mathbb{R}^{d \times d}$ be the $(d \times d)$-matrix which satisfies for all $i, j \in \{1, \dots, d\}$ that

$$M_{i,j} = \langle e_i, e_j \rangle. \tag{28}$$

Observe that the Cauchy-Schwarz inequality and the hypothesis that $\mathbb{E}\big[\|X\|\|Y\|\big] < \infty$ imply that

$$\mathbb{E}\big[|\langle X, Y \rangle|\big] \le \mathbb{E}\big[\|X\|\|Y\|\big] < \infty. \tag{29}$$

Next note that (26) and the hypothesis that $\mathbb{E}\big[\|X\|\|Y\|\big] < \infty$ ensure that for all $i, j \in \{1, 2, \dots, d\}$ it holds that

$$\mathbb{E}\big[|X_i Y_j|\big] = \mathbb{E}\big[|X_i||Y_j|\big] \le \mathbb{E}\big[\big(\textstyle\sum_{k=1}^d |X_k|\big)\big(\sum_{k=1}^d |Y_k|\big)\big] \le c^2 \mathbb{E}\big[\|X\|\|Y\|\big] < \infty. \tag{30}$$

Item (iii) in Theorem 8.14 in Klenke [50], the hypothesis that the function Y is $\mathcal{G}/\mathcal{B}(\mathbb{R}^d)$-measurable, (24), and (29) hence ensure that for all $A \in \mathcal{G}$ it holds that

$$\begin{aligned}
\mathbb{E}\big[\langle X, Y\rangle \mathbb{1}_A\big] &= \mathbb{E}\left[\Big\langle \sum_{i=1}^{d} X_i e_i, \sum_{j=1}^{d} Y_j e_j \Big\rangle \mathbb{1}_A\right] \\
&= \mathbb{E}\left[\bigg(\sum_{i,j=1}^{d} X_i Y_j \, \langle e_i, e_j\rangle \bigg) \mathbb{1}_A\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\bigg(\sum_{i,j=1}^{d} X_i Y_j M_{i,j} \bigg) \mathbb{1}_A \middle| \mathcal{G}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{i,j=1}^{d} X_i Y_j M_{i,j} \middle| \mathcal{G}\right] \mathbb{1}_A\right] \\
&= \mathbb{E}\left[\bigg(\sum_{i,j=1}^{d} \mathbb{E}\big[X_i Y_j \big| \mathcal{G}\big] M_{i,j} \bigg) \mathbb{1}_A\right] \\
&= \mathbb{E}\left[\bigg(\sum_{i,j=1}^{d} \mathbb{E}\big[X_i \big| \mathcal{G}\big] Y_j M_{i,j} \bigg) \mathbb{1}_A\right] \\
&= \mathbb{E}\Big[\big\langle \mathbb{E}\big[X \big| \mathcal{G}\big], Y\big\rangle \mathbb{1}_A\Big] = 0.
\end{aligned} \tag{31}$$

This and (29) establish (25). The proof of Lemma 2.5 is thus completed. □

2.3 Factorization lemma for conditional expectations

In this subsection we recall the statement and the proof of the well-known factorization lemma for conditional expectations from the literature (cf., e.g., Da Prato & Zabczyk [24, Proposition 1.12] and Pusnik & Jentzen [47, Subsection 2.1]).

Lemma 2.6. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{G} \subseteq \mathcal{F}$ be a sigma-algebra on Ω, let $(\mathbb{X}, \mathcal{X})$ and $(\mathbb{Y}, \mathcal{Y})$ be measurable spaces, let $X \colon \Omega \to \mathbb{X}$ be $\mathcal{F}/\mathcal{X}$-measurable, assume that X is independent of $\mathcal{G}$, let $Y \colon \Omega \to \mathbb{Y}$ be $\mathcal{G}/\mathcal{Y}$-measurable, let $B \in (\mathcal{X} \otimes \mathcal{Y})$, and let $\phi \colon \mathbb{Y} \to [0, \infty)$ be the function which satisfies for all $y \in \mathbb{Y}$ that $\phi(y) = \mathbb{E}\big[\mathbb{1}_B(X, y)\big]$. Then*

(i) it holds that the function ϕ is $\mathcal{Y}/\mathcal{B}([0,\infty))$-measurable and

(ii) it holds for all $A \in \mathcal{G}$ that

$$\mathbb{E}\big[\mathbb{1}_B(X, Y)\mathbb{1}_A\big] = \mathbb{E}\big[\phi(Y)\mathbb{1}_A\big]. \tag{32}$$

Proof of Lemma 2.6. Throughout this proof for every set S and every subset $\mathcal{S} \subseteq \mathcal{P}(S)$ of the power set $\mathcal{P}(S)$ of S let $\delta_S(\mathcal{S})$ be the set given by

$$\delta_S(\mathcal{S}) = \bigcap\nolimits_{\mathcal{B} \in \left\{ \substack{\mathcal{C} \text{ is a Dynkin system} \\ \text{on } S \text{ with } \mathcal{C} \supseteq \mathcal{S}} \right\}} \mathcal{B}, \tag{33}$$

for every set S and every subset $\mathcal{S} \subseteq \mathcal{P}(S)$ of the power set $\mathcal{P}(S)$ of S let $\sigma_S(\mathcal{S})$ be the set given by

$$\sigma_S(\mathcal{S}) = \bigcap\nolimits_{\mathcal{B} \in \left\{ \substack{\mathcal{C} \text{ is a sigma-algebra} \\ \text{on } S \text{ with } \mathcal{C} \supseteq \mathcal{S}} \right\}} \mathcal{B}, \tag{34}$$

let $\mathcal{E} \subseteq (\mathcal{X} \otimes \mathcal{Y})$ be the set given by

$$\mathcal{E} = \big\{ S \in (\mathcal{X} \otimes \mathcal{Y}) \colon (\exists\, E_1 \in \mathcal{X}, E_2 \in \mathcal{Y} \colon S = E_1 \times E_2) \big\}, \tag{35}$$

and let $\mathcal{D} \subseteq (\mathcal{X} \otimes \mathcal{Y})$ be the set given by

$$\mathcal{D} = \left\{ \begin{array}{l} D \in (\mathcal{X} \otimes \mathcal{Y}) \colon \\ \left[\begin{array}{l} \big(\mathbb{Y} \ni y \mapsto \mathbb{E}[\mathbb{1}_D(X,y)] \in [0,\infty)\big) \text{ is } \mathcal{Y}/\mathcal{B}([0,\infty))\text{-measurable} \\ \text{and } \big(\forall\, A \in \mathcal{G} \colon \mathbb{E}\big[\mathbb{1}_D(X,Y)\mathbb{1}_A\big] = \mathbb{E}\big[(\mathbb{E}[\mathbb{1}_D(X,y)])|_{y=Y}\mathbb{1}_A\big]\big) \end{array} \right] \end{array} \right\} \tag{36}$$

Note that Fubini's theorem (cf., e.g., Klenke [50, (14.6) in Theorem 14.16]) and the assumption that the function $X \colon \Omega \to \mathbb{X}$ is $\mathcal{F}/\mathcal{X}$-measurable demonstrate that for all $D \in (\mathcal{X} \otimes \mathcal{Y})$ it holds that the function

$$\mathbb{Y} \ni y \mapsto \mathbb{E}\big[\mathbb{1}_D(X,y)\big] = \int_\Omega \mathbb{1}_D(X(\omega),y)\, \mathbb{P}(\mathrm{d}\omega) \in [0,\infty) \tag{37}$$

is $\mathcal{Y}/\mathcal{B}([0,\infty))$-measurable. Hence, we obtain that

$$\begin{split} &\mathcal{D} = \\ &\big\{ D \in (\mathcal{X} \otimes \mathcal{Y}) \colon \big(\forall\, A \in \mathcal{G} \colon \mathbb{E}\big[\mathbb{1}_D(X,Y)\mathbb{1}_A\big] = \mathbb{E}\big[(\mathbb{E}[\mathbb{1}_D(X,y)])|_{y=Y}\mathbb{1}_A\big]\big) \big\}. \end{split} \tag{38}$$

Next observe that the hypothesis that the function X is independent of $\mathcal{G}$ and the hypothesis that the function Y is $\mathcal{G}/\mathcal{Y}$-measurable ensure that for all $E_1 \in \mathcal{X}, E_2 \in \mathcal{Y}, A \in \mathcal{G}$ it holds that

$$\begin{split} &\mathbb{E}\Big[\big(\mathbb{E}\big[\mathbb{1}_{E_1 \times E_2}(X,y)\big]\big)\big|_{y=Y}\mathbb{1}_A\Big] = \mathbb{E}\Big[\big(\mathbb{E}\big[\mathbb{1}_{E_1}(X)\mathbb{1}_{E_2}(y)\big]\big)\big|_{y=Y}\mathbb{1}_A\Big] \\ &= \mathbb{E}\big[\mathbb{P}(X \in E_1)\mathbb{1}_{E_2}(Y)\mathbb{1}_A\big] = \mathbb{P}(X \in E_1)\,\mathbb{P}(\{Y \in E_2\} \cap A) \\ &= \mathbb{P}(\{X \in E_1\} \cap \{Y \in E_2\} \cap A) = \mathbb{E}\big[\mathbb{1}_{E_1 \times E_2}(X,Y)\mathbb{1}_A\big]. \end{split} \tag{39}$$

Therefore, we obtain that $\mathcal{E} \subseteq \mathcal{D}$. Next observe that for all $D \in \mathcal{D}$, $A \in \mathcal{G}$ it holds that

$$\begin{aligned} &\mathbb{E}\big[\mathbb{1}_{((\mathbb{X}\times\mathbb{Y})\backslash D)}(X,Y)\mathbb{1}_A\big] = \mathbb{E}\big[(1-\mathbb{1}_D(X,Y))\mathbb{1}_A\big] \\ &= \mathbb{E}\big[\mathbb{1}_A\big] - \mathbb{E}\big[\mathbb{1}_D(X,Y)\mathbb{1}_A\big] = \mathbb{E}\big[\mathbb{1}_A\big] - \mathbb{E}\Big[\big(\mathbb{E}\big[\mathbb{1}_D(X,y)\big]\big)\big|_{y=Y}\mathbb{1}_A\Big] \\ &= \mathbb{E}\Big[\big(1-\big(\mathbb{E}\big[\mathbb{1}_D(X,y)\big]\big)\big|_{y=Y}\big)\mathbb{1}_A\Big] = \mathbb{E}\Big[\big(\mathbb{E}\big[1-\mathbb{1}_D(X,y)\big]\big)\big|_{y=Y}\mathbb{1}_A\Big] \\ &= \mathbb{E}\Big[\big(\mathbb{E}\big[\mathbb{1}_{((\mathbb{X}\times\mathbb{Y})\backslash D)}(X,y)\big]\big)\big|_{y=Y}\mathbb{1}_A\Big]. \end{aligned} \tag{40}$$

Moreover, note that the monotone convergence theorem implies that for all $A \in \mathcal{G}$, $(D_k)_{k\in\mathbb{N}} \subseteq \mathcal{D}$ with $\forall\, i \in \mathbb{N}, j \in \mathbb{N}\backslash\{i\}\colon D_i \cap D_j = \emptyset$ it holds that

$$\begin{aligned} \mathbb{E}\Big[\mathbb{1}_{(\cup_{k=1}^{\infty} D_k)}(X,Y)\,\mathbb{1}_A\Big] &= \mathbb{E}\Big[\lim_{n\to\infty}\big[\mathbb{1}_{(\cup_{k=1}^{n} D_k)}(X,Y)\,\mathbb{1}_A\big]\Big] \\ &= \mathbb{E}\Big[\lim_{n\to\infty}\big[\textstyle\sum_{k=1}^{n}\mathbb{1}_{D_k}(X,Y)\,\mathbb{1}_A\big]\Big] \\ &= \lim_{n\to\infty}\mathbb{E}\Big[\textstyle\sum_{k=1}^{n}\mathbb{1}_{D_k}(X,Y)\,\mathbb{1}_A\Big] \\ &= \lim_{n\to\infty}\left[\sum_{k=1}^{n}\mathbb{E}\big[\mathbb{1}_{D_k}(X,Y)\,\mathbb{1}_A\big]\right] \\ &= \lim_{n\to\infty}\left[\sum_{k=1}^{n}\mathbb{E}\Big[\big(\mathbb{E}\big[\mathbb{1}_{D_k}(X,y)\big]\big)\big|_{y=Y}\,\mathbb{1}_A\Big]\right]. \end{aligned} \tag{41}$$

Again the monotone convergence theorem hence implies that for all $A \in \mathcal{G}$, $(D_k)_{k\in\mathbb{N}} \subseteq \mathcal{D}$ with $\forall\, i \in \mathbb{N}, j \in \mathbb{N}\backslash\{i\}\colon D_i \cap D_j = \emptyset$ it holds that

$$\begin{aligned} \mathbb{E}\Big[\mathbb{1}_{(\cup_{k=1}^{\infty} D_k)}(X,Y)\,\mathbb{1}_A\Big] &= \mathbb{E}\left[\lim_{n\to\infty}\Big(\textstyle\sum_{k=1}^{n}\mathbb{E}\big[\mathbb{1}_{D_k}(X,y)\big]\Big)\Big|_{y=Y}\,\mathbb{1}_A\right] \\ &= \mathbb{E}\left[\Big(\lim_{n\to\infty}\textstyle\sum_{k=1}^{n}\mathbb{E}\big[\mathbb{1}_{D_k}(X,y)\big]\Big)\Big|_{y=Y}\,\mathbb{1}_A\right] \\ &= \mathbb{E}\left[\Big(\lim_{n\to\infty}\mathbb{E}\big[\textstyle\sum_{k=1}^{n}\mathbb{1}_{D_k}(X,y)\big]\Big)\Big|_{y=Y}\,\mathbb{1}_A\right] \\ &= \mathbb{E}\Big[\big(\mathbb{E}\big[\mathbb{1}_{(\cup_{k=1}^{\infty} D_k)}(X,y)\big]\big)\big|_{y=Y}\,\mathbb{1}_A\Big]. \end{aligned} \tag{42}$$

This, (40), and the fact that $(\mathbb{X}\times\mathbb{Y}) \in \mathcal{D}$ show that $\mathcal{D}$ is a Dynkin-system. The fact that $\mathcal{E}$ is $\cap$-stable, the fact that $\mathcal{E} \subseteq \mathcal{D}$, and Dynkin's π-λ-Theorem (see, e.g., [47, Theorem 2.5]) therefore demonstrate that

$$(\mathcal{X}\otimes\mathcal{Y}) = \sigma_{\mathbb{X}\times\mathbb{Y}}(\mathcal{E}) = \delta_{\mathbb{X}\times\mathbb{Y}}(\mathcal{E}) \subseteq \mathcal{D} \subseteq (\mathcal{X}\otimes\mathcal{Y}). \tag{43}$$

Hence, we obtain that $\mathcal{D} = \mathcal{X} \otimes \mathcal{Y}$. The assumption that $B \in (\mathcal{X} \otimes \mathcal{Y})$ hence assures that $B \in \mathcal{D}$. This completes the proof of Lemma 2.6. □

Lemma 2.7. *Let $N \in \mathbb{N}$, $c_1, \ldots, c_N \in [0, \infty)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{G} \subseteq \mathcal{F}$ be a sigma-algebra on Ω, let $(\mathbb{X}, \mathcal{X})$ and $(\mathbb{Y}, \mathcal{Y})$ be measurable spaces, let $D_1, \ldots, D_N \in (\mathcal{X} \otimes \mathcal{Y})$, let $X \colon \Omega \to \mathbb{X}$ be $\mathcal{F}/\mathcal{X}$-measurable, assume that X is independent of $\mathcal{G}$, let $Y \colon \Omega \to \mathbb{Y}$ be $\mathcal{G}/\mathcal{Y}$-measurable, let $\Phi \colon \mathbb{X} \times \mathbb{Y} \to [0, \infty)$ be $(\mathcal{X} \otimes \mathcal{Y})/\mathcal{B}([0, \infty))$-measurable, assume for all $x \in \mathbb{X}$, $y \in \mathbb{Y}$ that*

$$\Phi(x, y) = \sum_{k=1}^{N} c_k \mathbb{1}_{D_k}(x, y), \tag{44}$$

and let $\phi \colon \mathbb{Y} \to [0, \infty)$ be the function which satisfies for all $y \in \mathbb{Y}$ that $\phi(y) = \mathbb{E}\big[\Phi(X, y)\big]$. Then

(i) it holds that the function ϕ is $\mathcal{Y}/\mathcal{B}([0, \infty))$-measurable and

(ii) it holds for all $A \in \mathcal{G}$ that

$$\mathbb{E}\big[\Phi(X, Y)\mathbb{1}_A\big] = \mathbb{E}\big[\phi(Y)\mathbb{1}_A\big]. \tag{45}$$

Proof of Lemma 2.7. First, note that for all $y \in \mathbb{Y}$ it holds that

$$\phi(y) = \mathbb{E}\big[\Phi(X, y)\big] = \sum_{k=1}^{N} c_k \mathbb{E}[\mathbb{1}_{D_k}(X, y)]\,. \tag{46}$$

Item (i) in Lemma 2.6 therefore ensures that the function ϕ is $\mathcal{Y}/\mathcal{B}([0, \infty))$-measurable. This establishes item (i). In addition, observe that Lemma 2.6 implies that

$$\begin{aligned} \mathbb{E}\big[\Phi(X, Y)\mathbb{1}_A\big] &= \mathbb{E}\big[\textstyle\sum_{k=1}^{N} c_k \mathbb{1}_{D_k}(X, Y)\mathbb{1}_A\big] \\ &= \sum_{k=1}^{N} c_k \mathbb{E}\big[\mathbb{1}_{D_k}(X, Y)\mathbb{1}_A\big] \\ &= \sum_{k=1}^{N} c_k \mathbb{E}\big[(\mathbb{E}[\mathbb{1}_{D_k}(X, y)])|_{y=Y}\mathbb{1}_A\big] \\ &= \mathbb{E}\big[\big(\mathbb{E}\big[\textstyle\sum_{k=1}^{N} c_k \mathbb{1}_{D_k}(X, y)\big]\big)\big|_{y=Y}\mathbb{1}_A\big] \\ &= \mathbb{E}\big[(\mathbb{E}[\Phi(X, y)])|_{y=Y}\mathbb{1}_A\big] \\ &= \mathbb{E}\big[\phi(Y)\mathbb{1}_A\big]. \end{aligned} \tag{47}$$

This establishes item (ii). The proof of Lemma 2.7 is thus completed. □

Lemma 2.8. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{G} \subseteq \mathcal{F}$ be a sigma-algebra on Ω, let $(\mathbb{X}, \mathcal{X})$ and $(\mathbb{Y}, \mathcal{Y})$ be measurable spaces, let $X \colon \Omega \to \mathbb{X}$ be $\mathcal{F}/\mathcal{X}$-measurable, assume that X is independent of $\mathcal{G}$, let $Y : \Omega \to \mathbb{Y}$ be $\mathcal{G}/\mathcal{Y}$-measurable, let $\Phi \colon \mathbb{X} \times \mathbb{Y} \to [0, \infty]$ be $(\mathcal{X} \otimes \mathcal{Y})/\mathcal{B}([0, \infty])$-measurable, and let $\phi \colon \mathbb{Y} \to [0, \infty]$ be the function which satisfies for all $y \in \mathbb{Y}$ that $\phi(y) = \mathbb{E}\big[\Phi(X, y)\big]$. Then*

(i) it holds that the function ϕ is $\mathcal{Y}/\mathcal{B}([0, \infty])$-measurable and

(ii) it holds for all $A \in \mathcal{G}$ that

$$\mathbb{E}\big[\Phi(X, Y)\mathbb{1}_A\big] = \mathbb{E}\big[\phi(Y)\mathbb{1}_A\big]. \tag{48}$$

Proof of Lemma 2.8. First, note that Fubini's theorem (cf., e.g., Klenke [50, (14.6) in Theorem 14.16]), the assumption that the function $X : \Omega \to \mathbb{X}$ is $\mathcal{F}/\mathcal{X}$-measurable, and the assumption that the function $\Phi \colon \mathbb{X} \times \mathbb{Y} \to [0, \infty]$ is $(\mathcal{X} \otimes \mathcal{Y})/\mathcal{B}([0, \infty])$-measurable demonstrate that the function

$$\mathbb{Y} \ni y \mapsto \phi(y) = \mathbb{E}\big[\Phi(X, y)\big] = \int_\Omega \Phi(X(\omega), y)\, \mathbb{P}(\mathrm{d}\omega) \in [0, \infty] \tag{49}$$

is $\mathcal{Y}/\mathcal{B}([0, \infty])$-measurable. This establishes item (i). It thus remains to prove item (ii). For this let $\Phi_n \colon \mathbb{X} \times \mathbb{Y} \to [0, \infty)$, $n \in \mathbb{N}$, be the functions which satisfy for all $n \in \mathbb{N}$, $x \in \mathbb{X}$, $y \in \mathbb{Y}$ that

$$\begin{split}&\Phi_n(x, y) = \\ &2^n\, \mathbb{1}_{\{(v,w) \in \mathbb{X}\times\mathbb{Y} : \Phi(v,w) \geq 2^n\}}(x, y) + \sum_{k=0}^{2^{2n}-1} \left[\frac{k}{2^n}\, \mathbb{1}_{\{(v,w)\in\mathbb{X}\times\mathbb{Y} \colon k2^{-n} \leq \Phi(v,w) < (k+1)2^{-n}\}}(x, y)\right].\end{split} \tag{50}$$

Observe that the hypothesis that the function $\Phi \colon \mathbb{X}\times\mathbb{Y} \to [0, \infty]$ is $(\mathcal{X}\otimes\mathcal{Y})/\mathcal{B}([0, \infty])$-measurable assures that for all $n \in \mathbb{N}$, $k \in \{0, 1, \ldots, 2^{2n} - 1\}$ it holds that

$$\big\{(v, w) \in \mathbb{X} \times \mathbb{Y} \colon \Phi(v, w) \geq k2^n\big\} \in (\mathcal{X} \otimes \mathcal{Y}) \tag{51}$$

and

$$\big\{(v, w) \in \mathbb{X} \times \mathbb{Y} \colon k2^{-n} \leq \Phi(v, w) < (k + 1)2^{-n}\big\} \in (\mathcal{X} \otimes \mathcal{Y}). \tag{52}$$

This and Lemma 2.7 ensure that for all $n \in \mathbb{N}$, $A \in \mathcal{G}$ it holds that

$$\mathbb{E}\big[\Phi_n(X, Y)\mathbb{1}_A\big] = \mathbb{E}\Big[\big(\mathbb{E}\big[\Phi_n(X, y)\big]\big)\big|_{y=Y}\mathbb{1}_A\Big]. \tag{53}$$

The fact that $\forall\,(x,y)\in\mathbb{X}\times\mathbb{Y}, n\in\mathbb{N}\colon \Phi_n(x,y)\leq\Phi_{n+1}(x,y)$, the fact that $\forall\,(x,y)\in\mathbb{X}\times\mathbb{Y}\colon \lim_{n\to\infty}\Phi_n(x,y)=\Phi(x,y)$, and the monotone convergence theorem therefore imply that for all $A\in\mathcal{G}$ it holds that

$$\begin{split}\mathbb{E}\big[\Phi(X,Y)\mathbb{1}_A\big] &= \lim_{n\to\infty}\mathbb{E}\big[\Phi_n(X,Y)\mathbb{1}_A\big] = \lim_{n\to\infty}\mathbb{E}\Big[\big(\mathbb{E}\big[\Phi_n(X,y)\big]\big)\big|_{y=Y}\mathbb{1}_A\Big]\\ &= \mathbb{E}\Big[\big(\lim_{n\to\infty}\mathbb{E}\big[\Phi_n(X,y)\big]\big)\big|_{y=Y}\mathbb{1}_A\Big]\\ &= \mathbb{E}\Big[\big(\mathbb{E}\big[\Phi(X,y)\big]\big)\big|_{y=Y}\mathbb{1}_A\Big] = \mathbb{E}\big[\phi(Y)\mathbb{1}_A\big].\end{split} \tag{54}$$

This establishes item (ii). The proof of Lemma 2.8 is thus completed. □

Corollary 2.9. *Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, let $\mathcal{G}\subseteq\mathcal{F}$ be a sigma-algebra on Ω, let $(\mathbb{X},\mathcal{X})$ and $(\mathbb{Y},\mathcal{Y})$ be measurable spaces, let $X\colon\Omega\to\mathbb{X}$ be $\mathcal{F}/\mathcal{X}$-measurable, assume that X is independent of $\mathcal{G}$, let $Y\colon\Omega\to\mathbb{Y}$ be $\mathcal{G}/\mathcal{Y}$-measurable, let $\Phi\colon\mathbb{X}\times\mathbb{Y}\to\mathbb{R}$ be $(\mathcal{X}\otimes\mathcal{Y})/\mathcal{B}(\mathbb{R})$-measurable, assume that $\mathbb{E}\big[|\Phi(X,Y)|\big]<\infty$, let $c\in\mathbb{R}$, let $\phi\colon\mathbb{Y}\to\mathbb{R}$ be a function, assume for all $y\in\mathbb{Y}$ with $\mathbb{E}\big[|\Phi(X,y)|\big]<\infty$ that $\phi(y)=\mathbb{E}\big[\Phi(X,y)\big]$, and assume for all $y\in\mathbb{Y}$ with $\mathbb{E}\big[|\Phi(X,y)|\big]=\infty$ that $\phi(y)=c$. Then*

(i) it holds that $\big\{y\in\mathbb{Y}\colon\mathbb{E}\big[|\Phi(X,y)|\big]<\infty\big\}\in\mathcal{Y}$,

(ii) it holds that $\mathbb{P}\big(Y\in\big\{y\in\mathbb{Y}\colon\mathbb{E}\big[|\Phi(X,y)|\big]<\infty\big\}\big)=1$,

(iii) it holds that the function ϕ is $\mathcal{Y}/\mathcal{B}(\mathbb{R})$-measurable,

(iv) it holds that $\mathbb{E}\big[|\phi(Y)|\big]<\infty$, and

(v) it holds for all $A\in\mathcal{G}$ that

$$\mathbb{E}\big[\Phi(X,Y)\mathbb{1}_A\big]=\mathbb{E}\big[\phi(Y)\mathbb{1}_A\big]. \tag{55}$$

Proof of Corollary 2.9. Throughout this proof let $\Phi_k\colon\mathbb{X}\times\mathbb{Y}\to[0,\infty)$, $k\in\{1,2\}$, be the functions which satisfy for all $k\in\{1,2\}$, $x\in\mathbb{X}$, $y\in\mathbb{Y}$ that

$$\Phi_k(x,y)=\max\big\{(-1)^{k+1}\Phi(x,y),0\big\}, \tag{56}$$

let $B\subseteq\mathbb{Y}$ be the set given by

$$B=\big\{y\in\mathbb{Y}\colon\mathbb{E}\big[|\Phi(X,y)|\big]<\infty\big\}, \tag{57}$$

let $\mu : \mathcal{Y} \to [0,1]$ be the measure which satisfies for all $E \in \mathcal{Y}$ that

$$\mu(E) = \mathbb{P}(Y^{-1}(E)) = \mathbb{P}(Y \in E), \tag{58}$$

let $\Psi_k \colon \mathbb{X} \times \mathbb{Y} \to [0,\infty)$, $k \in \{1,2\}$, be the functions which satisfy for all $k \in \{1,2\}$, $x \in \mathbb{X}$, $y \in \mathbb{Y}$ that

$$\Psi_k(x,y) = \begin{cases} \Phi_k(x,y) & : y \in B \\ 0 & : y \in \mathbb{Y} \setminus B, \end{cases} \tag{59}$$

and let $\psi_k \colon \mathbb{Y} \to [0,\infty)$, $k \in \{1,2\}$, be the functions which satisfy for all $k \in \{1,2\}$, $y \in \mathbb{Y}$ that

$$\psi_k(y) = \mathbb{E}[\Psi_k(X,y)]. \tag{60}$$

Observe that the hypothesis that the function $\Phi \colon \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ is $(\mathcal{X} \otimes \mathcal{Y})/\mathcal{B}(\mathbb{R})$-measurable and the hypothesis that the function $X \colon \Omega \to \mathbb{X}$ is $\mathcal{F}/\mathcal{X}$-measurable assure that the function

$$\Omega \times \mathbb{Y} \ni (\omega, y) \mapsto |\Phi(X(\omega), y)| \in [0,\infty) \tag{61}$$

is $(\mathcal{F} \otimes \mathcal{Y})/\mathcal{B}([0,\infty))$-measurable. Fubini's theorem (cf., e.g., Klenke [50, (14.6) in Theorem 14.16]) hence proves that

$$B = \big\{ y \in \mathbb{Y} \colon \mathbb{E}\big[|\Phi(X,y)|\big] < \infty \big\} \in \mathcal{Y}. \tag{62}$$

This establishes item (i). In addition, observe that for all $y \in \mathbb{Y}$ it holds that

$$\phi(y) = \begin{cases} \mathbb{E}\big[\Phi(X,y)\big] & : y \in B \\ c & : y \in \mathbb{Y} \backslash B. \end{cases} \tag{63}$$

Next observe that the hypothesis that the function $\Phi : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ is $(\mathcal{X} \otimes \mathcal{Y})/\mathcal{B}(\mathbb{R})$-measurable and the fact that $B \in \mathcal{Y}$ ensure that the functions $\Psi_k \colon \mathbb{X} \times \mathbb{Y} \to [0,\infty)$, $k \in \{1,2\}$, are $(\mathcal{X} \otimes \mathcal{Y})/\mathcal{B}([0,\infty))$-measurable. Moreover, note that (57) implies that for all $k \in \{1,2\}$, $y \in B$ it holds that

$$\mathbb{E}[\Psi_k(X,y)] = \mathbb{E}[|\Psi_k(X,y)|] = \mathbb{E}[\Phi_k(X,y)] \leq \mathbb{E}[|\Phi(X,y)|] < \infty. \tag{64}$$

Hence, we obtain that for all $y \in B$ it holds that

$$\begin{aligned} \psi_1(y) - \psi_2(y) + c\mathbb{1}_{\mathbb{Y} \backslash B}(y) &= \psi_1(y) - \psi_2(y) \\ &= \mathbb{E}[\Psi_1(X,y)] - \mathbb{E}[\Psi_2(X,y)] \\ &= \mathbb{E}[\Psi_1(X,y) - \Psi_2(X,y)] \\ &= \mathbb{E}[\Phi_1(X,y) - \Phi_2(X,y)] \\ &= \mathbb{E}[\Phi(X,y)] = \phi(y). \end{aligned} \tag{65}$$

Furthermore, observe that (59), (60), and (63) ensure that for all $y \in \mathbb{Y} \setminus B$ it holds that

$$\psi_1(y) - \psi_2(y) + c\mathbb{1}_{\mathbb{Y}\setminus B}(y) = c\mathbb{1}_{\mathbb{Y}\setminus B}(y) = c = \phi(y). \tag{66}$$

Moreover, note that Fubini's theorem (cf., e.g., Klenke [50, (14.6) in Theorem 14.16]), the fact that the function $X\colon \Omega \to \mathbb{X}$ is $\mathcal{F}/\mathcal{X}$-measurable, and the fact that the functions $\Psi_k\colon \mathbb{X}\times\mathbb{Y} \to [0,\infty)$, $k \in \{1,2\}$, are $(\mathcal{X}\otimes\mathcal{Y})/\mathcal{B}([0,\infty))$-measurable demonstrate that the functions $\psi_k\colon \mathbb{Y} \to [0,\infty)$, $k \in \{1,2\}$, are $\mathcal{Y}/\mathcal{B}([0,\infty))$-measurable. Combining this and the fact that $B \in \mathcal{Y}$ with (65) and (66) demonstrates that the function ϕ is $\mathcal{Y}/\mathcal{B}(\mathbb{R})$-measurable. This establishes item (iii). Next observe that Lemma 2.8, (63), and the hypothesis that $\mathbb{E}[|\Phi(X,Y)|] < \infty$ ensure that

$$\mathbb{E}\big[|\phi(Y)|\big] \leq |c| + \mathbb{E}\Big[\big(\mathbb{E}\big[|\Phi(X,y)|\big]\big)\big|_{y=Y}\Big] = |c| + \mathbb{E}\big[|\Phi(X,Y)|\big] < \infty. \tag{67}$$

This establishes item (iv). Moreover, note that the hypothesis that $\mathbb{E}[|\Phi(X,Y)|] < \infty$ and Lemma 2.8 assure that

$$\int_{\mathbb{Y}} \mathbb{E}[|\Phi(X,y)|]\,\mu(\mathrm{d}y) = \mathbb{E}\Big[\big(\mathbb{E}[|\Phi(X,y)|]\big)\big|_{y=Y}\Big] = \mathbb{E}[|\Phi(X,Y)|] < \infty. \tag{68}$$

Combining this with (57) shows that

$$\mu(B) = \mu(\{y \in \mathbb{Y} : \mathbb{E}[|\Phi(X,y)|] < \infty\}) = 1. \tag{69}$$

Hence, we obtain that

$$\mathbb{P}(Y \in B) = 1. \tag{70}$$

This establishes item (ii). It thus remains to prove item (v). For this observe that (56), (59), (70), and the fact that $\mathbb{E}[\Psi_1(X,Y) + \Psi_2(X,Y)] \leq \mathbb{E}[\Phi_1(X,Y) + \Phi_2(X,Y)] = \mathbb{E}[|\Phi(X,Y)|] < \infty$ ensure that for all $A \in \mathcal{G}$ it holds that

$$\begin{aligned}\mathbb{E}\big[\Phi(X,Y)\mathbb{1}_A\big] &= \mathbb{E}\big[(\Phi_1(X,Y) - \Phi_2(X,Y))\mathbb{1}_A\big]\\ &= \mathbb{E}\big[\Phi_1(X,Y)\mathbb{1}_A\big] - \mathbb{E}\big[\Phi_2(X,Y)\mathbb{1}_A\big]\\ &= \mathbb{E}\big[\Phi_1(X,Y)\mathbb{1}_B(Y)\mathbb{1}_A\big] - \mathbb{E}\big[\Phi_2(X,Y)\mathbb{1}_B(Y)\mathbb{1}_A\big]\\ &= \mathbb{E}\big[\Psi_1(X,Y)\mathbb{1}_A\big] - \mathbb{E}\big[\Psi_2(X,Y)\mathbb{1}_A\big].\end{aligned} \tag{71}$$

Combining the fact that $\mathbb{E}[\Psi_1(X,Y) + \Psi_2(X,Y)] \leq \mathbb{E}[\Phi_1(X,Y) + \Phi_2(X,Y)] = \mathbb{E}[|\Phi(X,Y)|] < \infty$ and Lemma 2.8 with (64), (66), and (70) demonstrate that for all

$A \in \mathcal{G}$ it holds that

$$\begin{aligned}\mathbb{E}\big[\Phi(X,Y)\mathbb{1}_A\big] &= \mathbb{E}\Big[\big(\mathbb{E}\big[\Psi_1(X,y)\big]\big)\big|_{y=Y}\mathbb{1}_A\Big] - \mathbb{E}\Big[\big(\mathbb{E}\big[\Psi_2(X,y)\big]\big)\big|_{y=Y}\mathbb{1}_A\Big] \\ &= \mathbb{E}\big[\psi_1(Y)\mathbb{1}_A\big] - \mathbb{E}\big[\psi_2(Y)\mathbb{1}_A\big] \\ &= \mathbb{E}\big[\psi_1(Y)\mathbb{1}_A\big] - \mathbb{E}\big[\psi_2(Y)\mathbb{1}_A\big] + c\,\mathbb{P}(\{Y \in \mathbb{Y} \setminus B\} \cap A) \\ &= \mathbb{E}\big[(\psi_1(Y) - \psi_2(Y))\mathbb{1}_A\big] + c\,\mathbb{E}\big[\mathbb{1}_{\mathbb{Y}\setminus B}(Y)\mathbb{1}_A\big] \\ &= \mathbb{E}\big[(\psi_1(Y) - \psi_2(Y) + c\mathbb{1}_{\mathbb{Y}\setminus B}(Y))\mathbb{1}_A\big] \\ &= \mathbb{E}\big[\phi(Y)\mathbb{1}_A\big].\end{aligned} \tag{72}$$

This establishes item (v). The proof of Corollary 2.9 is thus completed. □

2.4 On convergence properties of a specific class of sequences

In this subsection we present in Lemma 2.10 an elementary auxiliary result on the convergence of a specific class of sequences. Lemma 2.10 is used in the proof of Lemma 4.1 in Subsection 4.1 below.

Lemma 2.10. *Let $\beta, \delta \in (0,\infty)$ with $\beta < \delta + 1$. Then*

$$\limsup_{n\to\infty}\left[\frac{\big|n^{-\delta} - (n-1)^{-\delta}\big|}{n^{-\beta}}\right] = 0. \tag{73}$$

Proof of Lemma 2.10. First, note that the fundamental theorem of calculus ensures that for all $n \in \{2, 3, \ldots\}$ it holds that

$$\begin{aligned} 0 \geq \frac{n^{-\delta} - (n-1)^{-\delta}}{n^{-\beta}} &= n^{\beta}\left(\frac{1}{n^{\delta}} - \frac{1}{(n-1)^{\delta}}\right) = n^{\beta}\big(\,\big[x^{-\delta}\big]_{x=n-1}^{x=n}\big) \\ &= n^{\beta}(-\delta)\left[\int_{n-1}^{n}\frac{1}{x^{\delta+1}}\,\mathrm{d}x\right] \geq -\frac{\delta n^{\beta}}{(n-1)^{\delta+1}}.\end{aligned} \tag{74}$$

The assumption that $\beta < \delta + 1$ therefore implies that

$$\begin{aligned} 0 \leq \limsup_{n\to\infty}\left[\frac{\big|n^{-\delta} - (n-1)^{-\delta}\big|}{n^{-\beta}}\right] &\leq \limsup_{n\to\infty}\left[\frac{\delta n^{\beta}}{(n-1)^{\delta+1}}\right] \\ = \limsup_{n\to\infty}\left[\frac{\delta(n+1)^{\beta}}{n^{\delta+1}}\right] &= \limsup_{n\to\infty}\left[\frac{\delta(1+{}^{1}\!/\!{}_{n})^{\beta}}{n^{\delta+1-\beta}}\right] = 0.\end{aligned} \tag{75}$$

This completes the proof of Lemma 2.10. □

2.5 On stability properties of the Euler scheme for ordinary differential equations

In this subsection we study in the elementary observations in Lemmas 2.12–2.15 and Proposition 2.16 below necessary and sufficient conditions which ensure that the Euler scheme admits a suitable Lyapunov-stability-type property (cf. Lemma 2.11 below). Similar results can be found, e.g., in Dereich & Müller-Gronbach [31, Remark 2.1] and the references mentioned therein. Lemma 2.12 is employed in our strong error analysis in Proposition 3.4 in Subsection 3.3 and Proposition 3.6 in Subsection 3.4 below.

Lemma 2.11 (Lyapunov-stability for the Euler scheme)**.** *Let $d \in \mathbb{N}$, $\vartheta \in \mathbb{R}^d$, $c, \varrho \in (0, \infty)$, let $\|\cdot\| \colon \mathbb{R}^d \to [0, \infty)$ be a norm, let $g \colon \mathbb{R}^d \to \mathbb{R}^d$ and $V \colon \mathbb{R}^d \to \mathbb{R}$ be functions which satisfy of all $\theta \in \mathbb{R}^d$ that*

$$V(\theta) = \|\theta - \vartheta\|^2, \tag{76}$$

and let $(\Theta_n^{r,\theta})_{n \in \mathbb{N}_0} \colon \mathbb{N}_0 \to \mathbb{R}^d$, $r \in [0, \infty)$, $\theta \in \mathbb{R}^d$, be the functions which satisfy for all $r \in [0, \infty)$, $\theta \in \mathbb{R}^d$, $n \in \mathbb{N}$ that

$$\Theta_0^{r,\theta} = \theta \qquad \textit{and} \qquad \Theta_n^{r,\theta} = \Theta_{n-1}^{r,\theta} + rg(\Theta_{n-1}^{r,\theta}). \tag{77}$$

Then the following three statements are equivalent:

(i) It holds for all $r \in [0, \varrho]$, $\theta \in \mathbb{R}^d$, $n \in \mathbb{N}$ that

$$V(\Theta_n^{r,\theta}) \leq (1 - cr)V(\Theta_{n-1}^{r,\theta}) \leq e^{-cr}\, V(\Theta_{n-1}^{r,\theta}). \tag{78}$$

(ii) It holds for all $r \in [0, \varrho]$, $\theta \in \mathbb{R}^d$ that

$$V(\Theta_1^{r,\theta}) \leq (1 - cr)V(\theta) \leq e^{-cr}\, V(\theta). \tag{79}$$

(iii) It holds for all $r \in [0, \varrho]$, $\theta \in \mathbb{R}^d$ that

$$\|\theta + rg(\theta) - \vartheta\|^2 \leq (1 - cr)\|\theta - \vartheta\|^2. \tag{80}$$

The proof of Lemma 2.11 is obvious. The next result, Lemma 2.12, provides a condition (see (81) in Lemma 2.12 below) which is sufficient to ensure that the stability property in item (iii) in Lemma 2.11 holds (see item (v) in Lemma 2.12 below).

Lemma 2.12. *Let $d \in \mathbb{N}$, $\vartheta \in \mathbb{R}^d$, $c_1, c_2 \in (0, \infty)$, let $\langle \cdot, \cdot \rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a scalar product, let $\|\cdot\| \colon \mathbb{R}^d \to [0, \infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that $\|\theta\| = \sqrt{\langle \theta, \theta \rangle}$, and let $g \colon \mathbb{R}^d \to \mathbb{R}^d$ be a function which satisfies for all $\theta \in \mathbb{R}^d$ that*

$$\langle \theta - \vartheta, g(\theta) \rangle \leq -\max\bigl\{c_1 \|\theta - \vartheta\|^2, c_2 \|g(\theta)\|^2\bigr\}\,. \tag{81}$$

Then

(i) it holds that

$$\{\theta \in \mathbb{R}^d \colon g(\theta) = 0\} = \{\vartheta\}, \tag{82}$$

(ii) it holds that $c_1 c_2 \leq 1$,

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

$$c_1 \|\theta - \vartheta\| \leq \|g(\theta)\| \leq \tfrac{1}{c_2} \|\theta - \vartheta\|, \tag{83}$$

(iv) it holds for all $\theta \in \mathbb{R}^d$, $r \in [0, 2c_2]$ that

$$\|\theta + r g(\theta) - \vartheta\|^2 \leq \bigl(1 - c_1 r (2 - \tfrac{r}{c_2})\bigr) \|\theta - \vartheta\|^2, \tag{84}$$

and

(v) it holds for all $\theta \in \mathbb{R}^d$, $r \in [0, c_2]$ that

$$\|\theta + r g(\theta) - \vartheta\|^2 \leq (1 - c_1 r)\, \|\theta - \vartheta\|^2. \tag{85}$$

Proof of Lemma 2.12. First, note that (81) (with $\theta = \vartheta$ in the notation of (81)) implies that

$$0 \leq -\max\{0, c_2 \|g(\vartheta)\|^2\} \leq -c_2 \|g(\vartheta)\|^2. \tag{86}$$

Hence, we obtain that $0 \geq \|g(\vartheta)\|^2$. This assures that $g(\vartheta) = 0$. Next observe that (81) and the Cauchy-Schwarz inequality ensure that for all $\theta \in \mathbb{R}^d$ it holds that

$$c_2 \|g(\theta)\|^2 \leq \max\left\{c_1 \|\theta - \vartheta\|^2, c_2 \|g(\vartheta)\|^2\right\} \leq -\langle \theta - \vartheta, g(\theta) \rangle \leq \|\theta - \vartheta\| \|g(\theta)\|. \tag{87}$$

Therefore, we obtain that for all $\theta \in \mathbb{R}^d$ it holds that

$$c_1 \|\theta - \vartheta\|^2 \leq -\langle \theta - \vartheta, g(\theta) \rangle \leq \|\theta - \vartheta\| \|g(\theta)\|. \tag{88}$$

Combining this with (87) and the fact that $g(\vartheta) = 0$ proves that for all $\theta \in \mathbb{R}^d$ it holds that

$$c_1 \|\theta - \vartheta\| \leq \|g(\theta)\| \leq \tfrac{1}{c_2} \|\theta - \vartheta\|. \tag{89}$$

Therefore, we obtain that for all $\theta \in \mathbb{R}^d$ it holds that $c_1c_2\|\theta - \vartheta\| \leq \|\theta - \vartheta\|$. This demonstrates that $c_1c_2 \leq 1$. Combining (89) and the fact that $g(\vartheta) = 0$ hence establishes items (i)–(iii). Next observe that for all $r \in [0, 2c_2]$ it holds that $r(2 - \frac{r}{c_2}) \geq 0$. This and (87) imply that for all $\theta \in \mathbb{R}^d$, $r \in [0, 2c_2]$ it holds that

$$\begin{aligned}
\|\theta + rg(\theta) - \vartheta\|^2 &= \|\theta - \vartheta\|^2 + 2r\langle\theta - \vartheta, g(\theta)\rangle + r^2\|g(\theta)\|^2 \\
&\leq \|\theta - \vartheta\|^2 + 2r\langle\theta - \vartheta, g(\theta)\rangle - \tfrac{r^2}{c_2}\langle\theta - \vartheta, g(\theta)\rangle \\
&= \|\theta - \vartheta\|^2 + r\left(2 - \frac{r}{c_2}\right)\langle\theta - \vartheta, g(\theta)\rangle \\
&\leq \|\theta - \vartheta\|^2 - r\left(2 - \frac{r}{c_2}\right)c_1\|\theta - \vartheta\|^2 \\
&= \left(1 - c_1r\left(2 - \frac{r}{c_2}\right)\right)\|\theta - \vartheta\|^2.
\end{aligned} \tag{90}$$

This proves item (iv). Moreover, note that item (iv) and the fact that for all $r \in [0, c_2]$ it holds that $2 - \frac{r}{c_2} \geq 1$ establish item (v). The proof of Lemma 2.12 is thus completed. □

Lemma 2.13 (On the monotonicity of a property)**.** *Let $d \in \mathbb{N}$, $\vartheta \in \mathbb{R}^d$, $c, \varrho \in (0, \infty)$, let $\langle\cdot,\cdot\rangle\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a scalar product, let $\|\cdot\|\colon \mathbb{R}^d \to [0, \infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that $\|\theta\| = \sqrt{\langle\theta, \theta\rangle}$, and let $g\colon \mathbb{R}^d \to \mathbb{R}^d$ be a function which satisfies for all $\theta \in \mathbb{R}^d$ that*

$$\|\theta + \varrho g(\theta) - \vartheta\|^2 \leq (1 - c\varrho)\|\theta - \vartheta\|^2. \tag{91}$$

Then it holds for all $\theta \in \mathbb{R}^d$, $r \in [0, \varrho]$ that

$$\|\theta + rg(\theta) - \vartheta\|^2 \leq (1 - cr)\|\theta - \vartheta\|^2. \tag{92}$$

Proof of Lemma 2.13. First, observe that (91) implies that for all $\theta \in \mathbb{R}^d$ it holds that

$$\begin{aligned}
&\|\theta - \vartheta\|^2 + 2\varrho\langle\theta - \vartheta, g(\theta)\rangle + \varrho^2\|g(\theta)\|^2 \\
&= \|\theta - \vartheta\|^2 + 2\langle\theta - \vartheta, \varrho g(\theta)\rangle + \|\varrho g(\theta)\|^2 \\
&= \|(\theta - \vartheta) + \varrho g(\theta)\|^2 \\
&= \|\theta + \varrho g(\theta) - \vartheta\|^2 \\
&\leq (1 - c\varrho)\|\theta - \vartheta\|^2 \\
&= \|\theta - \vartheta\|^2 - c\varrho\|\theta - \vartheta\|^2.
\end{aligned} \tag{93}$$

Therefore, we obtain that for all $\theta \in \mathbb{R}^d$ it holds that

$$2\langle \theta - \vartheta, g(\theta) \rangle + \varrho \|g(\theta)\|^2 \leq -c\|\theta - \vartheta\|^2. \tag{94}$$

This ensures that for all $\theta \in \mathbb{R}^d$, $r \in [0, \varrho]$ it holds that

$$\begin{aligned} \|\theta + rg(\theta) - \vartheta\|^2 &= \|(\theta - \vartheta) + rg(\theta)\|^2 \\ &= \|\theta - \vartheta\|^2 + 2\langle \theta - \vartheta, rg(\theta) \rangle + \|rg(\theta)\|^2 \\ &= \|\theta - \vartheta\|^2 + 2r\langle \theta - \vartheta, g(\theta) \rangle + r^2\|g(\theta)\|^2 \\ &= \|\theta - \vartheta\|^2 + r\left(2\langle \theta - \vartheta, g(\theta) \rangle + r\|g(\theta)\|^2\right) \\ &\leq \|\theta - \vartheta\|^2 + r\left(2\langle \theta - \vartheta, g(\theta) \rangle + \varrho\|g(\theta)\|^2\right) \\ &\leq \|\theta - \vartheta\|^2 + r(-c\|\theta - \vartheta\|^2) \\ &= (1 - cr)\|\theta - \vartheta\|^2. \end{aligned} \tag{95}$$

The proof of Lemma 2.13 is thus completed. □

Lemma 2.14. *Let $d \in \mathbb{N}$, $\vartheta \in \mathbb{R}^d$, $c, \varrho \in (0, \infty)$, let $\langle \cdot, \cdot \rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a scalar product, let $\|\cdot\| \colon \mathbb{R}^d \to [0, \infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that $\|\theta\| = \sqrt{\langle \theta, \theta \rangle}$, and let $g \colon \mathbb{R}^d \to \mathbb{R}^d$ be a function which satisfies for all $\theta \in \mathbb{R}^d$, $r \in [0, \varrho]$ that*

$$\|\theta + rg(\theta) - \vartheta\|^2 \leq (1 - cr)\|\theta - \vartheta\|^2. \tag{96}$$

Then it holds that $g(\vartheta) = 0$ and

$$\inf_{r \in (0,\infty)} \left(\sup_{\theta \in \mathbb{R}^d \setminus \{\vartheta\}} \left[\frac{\|\theta + rg(\theta) - \vartheta\|^2}{\|\theta - \vartheta\|^2} \right] \right) \leq 1 - c\varrho < 1. \tag{97}$$

Proof of Lemma 2.14. Observe that (96) (with $\theta = \vartheta$, $r = \varrho$ in the notation of (96)) implies that $\|\varrho g(\vartheta)\| \leq 0$. The hypothesis that $\varrho \in (0, \infty)$ hence demonstrates that $g(\vartheta) = 0$. Moreover, note that (96) ensures that

$$\begin{aligned} \inf_{r \in (0,\infty)} \left(\sup_{\theta \in \mathbb{R}^d \setminus \{\vartheta\}} \left[\frac{\|\theta + rg(\theta) - \vartheta\|^2}{\|\theta - \vartheta\|^2} \right] \right) &\leq \sup_{\theta \in \mathbb{R}^d \setminus \{\vartheta\}} \left[\frac{\|\theta + \varrho g(\theta) - \vartheta\|^2}{\|\theta - \vartheta\|^2} \right] \\ &\leq 1 - c\varrho \\ &< 1. \end{aligned} \tag{98}$$

The proof of Lemma 2.14 is thus completed. □

Lemma 2.15. *Let $d \in \mathbb{N}$, $\vartheta \in \mathbb{R}^d$, $C, r \in (0,\infty)$, let $\langle \cdot,\cdot \rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a scalar product, let $\|\cdot\| \colon \mathbb{R}^d \to [0,\infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that $\|\theta\| = \sqrt{\langle \theta, \theta \rangle}$, and let $g \colon \mathbb{R}^d \to \mathbb{R}^d$ be a function which satisfies that $g(\vartheta) = 0$ and*

$$\sup_{\theta \in \mathbb{R}^d \setminus \{\vartheta\}} \left[\frac{2\langle \theta - \vartheta, g(\theta) \rangle + r\|g(\theta)\|^2}{\|\theta - \vartheta\|^2} \right] \leq -C. \tag{99}$$

Then it holds for all $\theta \in \mathbb{R}^d$ that

$$\langle \theta - \vartheta, g(\theta) \rangle \leq -\left[\tfrac{\min\{C,r\}}{2}\right] \max\bigl\{\|\theta - \vartheta\|^2, \|g(\theta)\|^2\bigr\}. \tag{100}$$

Proof of Lemma 2.15. First, note that (99) implies that for all $\theta \in \mathbb{R}^d \setminus \{\vartheta\}$ it holds that

$$2\langle \theta - \vartheta, g(\theta) \rangle + r\|g(\theta)\|^2 \leq -C\|\theta - \vartheta\|^2. \tag{101}$$

Therefore, we obtain that for all $\theta \in \mathbb{R}^d \setminus \{\vartheta\}$ it holds that

$$\begin{aligned} \langle \theta - \vartheta, g(\theta) \rangle &\leq -\frac{r}{2}\|g(\theta)\|^2 - \frac{C}{2}\|\theta - \vartheta\|^2 \\ &\leq -\left[\tfrac{\min\{C,r\}}{2}\right]\|g(\theta)\|^2 - \left[\tfrac{\min\{C,r\}}{2}\right]\|\theta - \vartheta\|^2 \\ &= -\left[\tfrac{\min\{C,r\}}{2}\right]\bigl[\|g(\theta)\|^2 + \|\theta - \vartheta\|^2\bigr] \\ &\leq -\left[\tfrac{\min\{C,r\}}{2}\right] \max\bigl\{\|\theta - \vartheta\|^2, \|g(\theta)\|^2\bigr\}. \end{aligned} \tag{102}$$

The assumption that $g(\vartheta) = 0$ hence shows that for all $\theta \in \mathbb{R}^d$ it holds that

$$\langle \theta - \vartheta, g(\theta) \rangle \leq -\left[\tfrac{\min\{C,r\}}{2}\right] \max\bigl\{\|\theta - \vartheta\|^2, \|g(\theta)\|^2\bigr\}. \tag{103}$$

This completes the proof of Lemma 2.15. □

Proposition 2.16 (Equivalence of properties)**.** *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 $\|\cdot\| \colon \mathbb{R}^d \to [0,\infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that $\|\theta\| = \sqrt{\langle \theta, \theta \rangle}$, and let $g \colon \mathbb{R}^d \to \mathbb{R}^d$ be a function. Then the following five statements are equivalent:*

(i) There exists $c \in (0,\infty)$ such that for all $\theta \in \mathbb{R}^d$ it holds that

$$\langle \theta - \vartheta, g(\theta) \rangle \leq -c \max\bigl\{\|\theta - \vartheta\|^2, \|g(\theta)\|^2\bigr\}. \tag{104}$$

(ii) There exist $c, \varrho \in (0,\infty)$ such that for all $\theta \in \mathbb{R}^d$ it holds that

$$\|\theta + \varrho g(\theta) - \vartheta\|^2 \leq (1 - c\varrho)\|\theta - \vartheta\|^2. \tag{105}$$

(iii) There exist $c, \varrho \in (0, \infty)$ such that for all $\theta \in \mathbb{R}^d$, $r \in [0, \varrho]$ it holds that

$$\|\theta + rg(\theta) - \vartheta\|^2 \leq (1 - cr)\|\theta - \vartheta\|^2. \tag{106}$$

(iv) It holds that $g(\vartheta) = 0$ and

$$\inf_{r \in (0,\infty)} \left(\sup_{\theta \in \mathbb{R}^d \setminus \{\vartheta\}} \left[\frac{\|\theta + rg(\theta) - \vartheta\|^2}{\|\theta - \vartheta\|^2} \right] \right) < 1. \tag{107}$$

(v) It holds that $g(\vartheta) = 0$ and

$$\inf_{r \in (0,\infty)} \left(\sup_{\theta \in \mathbb{R}^d \setminus \{\vartheta\}} \left[\frac{2\langle \theta - \vartheta, g(\theta) \rangle + r\|g(\theta)\|^2}{\|\theta - \vartheta\|^2} \right] \right) < 0. \tag{108}$$

Proof of Proposition 2.16. First, note that item (v) in Lemma 2.12 ensures that ((i) $\Rightarrow$ (ii)). Next observe that Lemma 2.13 implies that ((ii) $\Rightarrow$ (iii)). Moreover, note that Lemma 2.14 demonstrates that ((iii) $\Rightarrow$ (iv)). In addition, observe that the fact that for all $r \in (0, \infty)$ and all functions $h \colon \mathbb{R}^d \to \mathbb{R}^d$ it holds that

$$\begin{aligned}
&\sup_{\theta \in \mathbb{R}^d \setminus \{\vartheta\}} \left[\frac{\|\theta + rh(\theta) - \vartheta\|^2}{\|\theta - \vartheta\|^2} \right] \\
&= \sup_{\theta \in \mathbb{R}^d \setminus \{\vartheta\}} \left[\frac{\|(\theta - \vartheta) + rh(\theta)\|^2}{\|\theta - \vartheta\|^2} \right] \\
&= \sup_{\theta \in \mathbb{R}^d \setminus \{\vartheta\}} \left[\frac{\|\theta - \vartheta\|^2 + 2\langle \theta - \vartheta, rh(\theta) \rangle + \|rh(\theta)\|^2}{\|\theta - \vartheta\|^2} \right] \\
&= \sup_{\theta \in \mathbb{R}^d \setminus \{\vartheta\}} \left[1 + \frac{2r\langle \theta - \vartheta, h(\theta) \rangle + r^2\|h(\theta)\|^2}{\|\theta - \vartheta\|^2} \right] \\
&= 1 + \sup_{\theta \in \mathbb{R}^d \setminus \{\vartheta\}} \left[\frac{2r\langle \theta - \vartheta, h(\theta) \rangle + r^2\|h(\theta)\|^2}{\|\theta - \vartheta\|^2} \right] \\
&= 1 + r \left(\sup_{\theta \in \mathbb{R}^d \setminus \{\vartheta\}} \left[\frac{2\langle \theta - \vartheta, h(\theta) \rangle + r\|h(\theta)\|^2}{\|\theta - \vartheta\|^2} \right] \right)
\end{aligned} \tag{109}$$

implies that ((iv) $\Leftrightarrow$ (v)). Furthermore, note that Lemma 2.15 implies that ((v) $\Rightarrow$ (i)). The proof of Proposition 2.16 is thus completed. □

2.6 A Gronwall-type inequality

In this subsection we establish in Lemma 2.17 a certain Gronwall-type inequality. Lemma 2.17 is used in our strong error analysis in Proposition 3.2 in Subsection 3.2 below.

Lemma 2.17. *Let $N \in \mathbb{N}_0$, $k, \kappa, c, C \in (0,\infty)$, $(e_n)_{n\in\mathbb{N}_0} \subseteq [0,\infty)$, $(\gamma_n)_{n\in\mathbb{N}_0} \subseteq (0,\infty)$ satisfy for all $n \in \mathbb{N} \cap (N,\infty)$ that*

$$e_n \leq (1 - c\gamma_n)e_{n-1} + \kappa(\gamma_n)^{k+1}, \qquad \sup_{l\in\mathbb{N}\cap(N,\infty)} \gamma_l \leq {}^1\!/\!{}_c, \tag{110}$$

$$\textit{and} \qquad C = \inf_{l\in\mathbb{N}\cap(N,\infty)} \left[\tfrac{(\gamma_l)^k - (\gamma_{l-1})^k}{(\gamma_l)^{k+1}} + \tfrac{c(\gamma_{l-1})^k}{(\gamma_l)^k}\right]. \tag{111}$$

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

$$e_n \leq \left[\max\!\left\{\frac{e_0}{(\gamma_0)^k}, \frac{e_1}{(\gamma_1)^k}, \ldots, \frac{e_N}{(\gamma_N)^k}, \frac{\kappa}{C}\right\}\right] (\gamma_n)^k. \tag{112}$$

Proof of Lemma 2.17. Throughout this proof let $\lambda \in (0,\infty)$ satisfy

$$\lambda = \max\!\left\{\frac{e_N}{(\gamma_N)^k}, \frac{\kappa}{C}\right\}. \tag{113}$$

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

$$e_n \leq \lambda(\gamma_n)^k. \tag{114}$$

We now prove (114) by induction on $n \in \{N, N+1, \ldots\}$. For the base case $n = N$ observe that

$$e_N = \left[\frac{e_N}{(\gamma_N)^k}\right] (\gamma_N)^k \leq \lambda(\gamma_N)^k. \tag{115}$$

This proves (114) in the base case $n = N$. For the induction step $\{N, N+1, \ldots\} \ni (n-1) \to n \in \mathbb{N} \cap (N,\infty)$ note that (110), (111), and (113) demonstrate that for all

$n \in \mathbb{N} \cap (N, \infty)$ with $e_{n-1} \leq \lambda(\gamma_{n-1})^k$ it holds that

$$\begin{aligned} e_n &\leq (1 - c\gamma_n)e_{n-1} + \kappa(\gamma_n)^{k+1} \\ &\leq (1 - c\gamma_n)\lambda(\gamma_{n-1})^k + \kappa(\gamma_n)^{k+1} \\ &= \lambda(\gamma_{n-1})^k - \lambda c\gamma_n(\gamma_{n-1})^k + \kappa(\gamma_n)^{k+1} \\ &= \lambda(\gamma_n)^k - (\gamma_n)^{k+1}\left(\frac{\lambda c\gamma_n(\gamma_{n-1})^k}{(\gamma_n)^{k+1}} + \frac{\lambda(\gamma_n)^k}{(\gamma_n)^{k+1}} - \frac{\lambda(\gamma_{n-1})^k}{(\gamma_n)^{k+1}} - \kappa\right) \\ &= \lambda(\gamma_n)^k - (\gamma_n)^{k+1}\left(\lambda\left[\frac{c(\gamma_{n-1})^k}{(\gamma_n)^k} + \frac{(\gamma_n)^k}{(\gamma_n)^{k+1}} - \frac{(\gamma_{n-1})^k}{(\gamma_n)^{k+1}}\right] - \kappa\right) \\ &= \lambda(\gamma_n)^k - (\gamma_n)^{k+1}\left(\lambda\left[\frac{(\gamma_n)^k - (\gamma_{n-1})^k}{(\gamma_n)^{k+1}} + \frac{c(\gamma_{n-1})^k}{(\gamma_n)^k}\right] - \kappa\right). \end{aligned} \tag{116}$$

Hence, we obtain that for all $n \in \mathbb{N} \cap (N, \infty)$ with $e_{n-1} \leq \lambda(\gamma_{n-1})^k$ it holds that

$$\begin{aligned} e_n &\leq \lambda(\gamma_n)^k - (\gamma_n)^{k+1}(\lambda C - \kappa) \\ &\leq \lambda(\gamma_n)^k - (\gamma_n)^{k+1}\left(\left[\frac{\kappa}{C}\right]C - \kappa\right) \\ &= \lambda(\gamma_n)^k - (\gamma_n)^{k+1}(\kappa - \kappa) = \lambda(\gamma_n)^k. \end{aligned} \tag{117}$$

Induction thus proves (114). Next note that (114) ensures that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{aligned} e_n &\leq \left[\max\left\{\frac{e_0}{(\gamma_0)^k}, \frac{e_1}{(\gamma_1)^k}, \dots, \frac{e_{N-1}}{(\gamma_{N-1})^k}, \lambda\right\}\right](\gamma_n)^k \\ &= \left[\max\left\{\frac{e_0}{(\gamma_0)^k}, \frac{e_1}{(\gamma_1)^k}, \dots, \frac{e_N}{(\gamma_N)^k}, \frac{\kappa}{C}\right\}\right](\gamma_n)^k. \end{aligned} \tag{118}$$

The proof of Lemma 2.17 is thus completed. □

Corollary 2.18. *Let $k, \kappa, c \in (0, \infty)$, $(e_n)_{n\in\mathbb{N}_0} \subseteq [0, \infty)$, $(\gamma_n)_{n\in\mathbb{N}_0} \subseteq (0, \infty)$ satisfy for all $n \in \mathbb{N}$ that*

$$e_n \leq (1 - c\gamma_n)e_{n-1} + \kappa(\gamma_n)^{k+1} \qquad \text{and} \tag{119}$$

$$\limsup_{l\to\infty} \gamma_l = 0 < \liminf_{l\to\infty}\left[\frac{(\gamma_l)^k - (\gamma_{l-1})^k}{(\gamma_l)^{k+1}} + \frac{c(\gamma_{l-1})^k}{(\gamma_l)^k}\right]. \tag{120}$$

Then there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}_0$ it holds that

$$e_n \leq C(\gamma_n)^k. \tag{121}$$

Proof of Corollary 2.18. Observe that (120) ensures that there exists $N \in \mathbb{N}_0$ such that

$$\sup_{l \in \mathbb{N} \cap (N,\infty)} \gamma_l \leq {}^1\!/\!_c \qquad \text{and} \qquad \inf_{l \in \mathbb{N} \cap (N,\infty)} \left[\frac{(\gamma_l)^k - (\gamma_{l-1})^k}{(\gamma_l)^{k+1}} + \frac{c(\gamma_{l-1})^k}{(\gamma_l)^k} \right] > 0. \tag{122}$$

Lemma 2.17 therefore assures that for all $n \in \mathbb{N}_0$ it holds that

$$e_n \leq \left[\max\left\{ \frac{e_0}{(\gamma_0)^k}, \frac{e_1}{(\gamma_1)^k}, \ldots, \frac{e_N}{(\gamma_N)^k}, \frac{\kappa}{\inf_{l \in \mathbb{N} \cap (N,\infty)} \left[\frac{(\gamma_l)^k - (\gamma_{l-1})^k}{(\gamma_l)^{k+1}} + \frac{c(\gamma_{l-1})^k}{(\gamma_l)^k} \right]} \right\} \right] (\gamma_n)^k. \tag{123}$$

This completes the proof of Corollary 2.18. □

3 Error analysis for stochastic approximation algorithms (SAAs)

In this section we establish in Theorem 3.7 in Subsection 3.4 below for every $p \in (0,\infty)$ strong L^p-convergence rates for stochastic approximation algorithms.

3.1 Main setting for the strong error analysis

Throughout this section the following setting is frequently used.

Setting 3.1. *Let $d \in \mathbb{N}$, $(\gamma_n)_{n \in \mathbb{N}_0} \subseteq (0,\infty)$, let $g\colon \mathbb{R}^d \to \mathbb{R}^d$ be $\mathcal{B}(\mathbb{R}^d)/\mathcal{B}(\mathbb{R}^d)$-measurable, let $(\Omega, \mathcal{F}, \mathbb{P}, (\mathbb{F}_n)_{n \in \mathbb{N}_0})$ be a filtered probability space, let $D\colon \mathbb{N} \times \Omega \to \mathbb{R}^d$ be an $(\mathbb{F}_n)_{n \in \mathbb{N}}/\mathcal{B}(\mathbb{R}^d)$-adapted stochastic process, let $\Theta\colon \mathbb{N}_0 \times \Omega \to \mathbb{R}^d$ be a function, assume that Θ_0 is $\mathbb{F}_0/\mathcal{B}(\mathbb{R}^d)$-measurable, and assume for all $n \in \mathbb{N}$ that*

$$\Theta_n = \Theta_{n-1} + \gamma_n (g(\Theta_{n-1}) + D_n). \tag{124}$$

Note that in Setting 3.1 the hypothesis that the function g is $\mathcal{B}(\mathbb{R}^d)/\mathcal{B}(\mathbb{R}^d)$-measurable, the hypothesis that the function Θ_0 is $\mathbb{F}_0/\mathcal{B}(\mathbb{R}^d)$-measurable, the hypothesis that D is an $(\mathbb{F}_n)_{n \in \mathbb{N}}/\mathcal{B}(\mathbb{R}^d)$-adapted stochastic process, and (124) imply that Θ is an $(\mathbb{F}_n)_{n \in \mathbb{N}_0}/\mathcal{B}(\mathbb{R}^d)$-adapted stochastic process.

3.2 Lyapunov based convergence for SAAs

Proposition 3.2 (Lyapunov based convergence for stochastic approximation)**.** *Assume Setting 3.1 and let $N \in \mathbb{N}_0$, $k, \kappa, c, C \in (0,\infty)$, $V \in C^1(\mathbb{R}^d, [0,\infty))$ satisfy for* *all $m \in \mathbb{N}_0$, $n \in \mathbb{N} \cap (N, \infty)$, $t \in [0,1]$, $\theta \in \mathbb{R}^d$ that*

$$\mathbb{E}\big[V(\Theta_m) + |V'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))(D_n)|\big] < \infty, \tag{125}$$

$$\textstyle\int_0^1 \mathbb{E}\big[|V'(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + sD_n))(D_n)|\big]\,\mathrm{d}s < \infty, \tag{126}$$

$$\mathbb{E}\big[V'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))(D_n)\big] = 0, \qquad V(\theta + \gamma_n g(\theta)) \leq (1 - c\gamma_n)V(\theta), \tag{127}$$

$$\begin{split}&\mathbb{E}\big[|V'(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + tD_n))(D_n) - V'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))(D_n)|\big] \\ &\leq \kappa\big((\gamma_n)^k + \gamma_n \mathbb{E}\big[V(\Theta_{n-1})\big]\big),\end{split} \tag{128}$$

$$\sup_{l \in \mathbb{N} \cap (N,\infty)} \gamma_l \leq \min\{\tfrac{c}{2\kappa}, \tfrac{2}{c}\}, \qquad \textit{and} \qquad C = \inf_{l \in \mathbb{N} \cap (N,\infty)} \Big[\tfrac{(\gamma_l)^k - (\gamma_{l-1})^k}{(\gamma_l)^{k+1}} + \tfrac{c(\gamma_{l-1})^k}{2(\gamma_l)^k}\Big]. \tag{129}$$

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

$$\mathbb{E}\big[V(\Theta_n)\big] \leq \left[\max\!\left(\left\{\frac{\kappa}{C}\right\} \cup \left\{\frac{\mathbb{E}\big[V(\Theta_l)\big]}{(\gamma_l)^k} : l \in \{0, 1, \ldots, N\}\right\}\right)\right](\gamma_n)^k < \infty. \tag{130}$$

Proof of Proposition 3.2. Throughout this proof let $(e_n)_{n \in \mathbb{N}_0} \subseteq [0, \infty]$ satisfy for all $n \in \mathbb{N}_0$ that

$$e_n = \mathbb{E}\big[V(\Theta_n)\big]. \tag{131}$$

Note that (125) ensures that for all $n \in \mathbb{N}_0$ it holds that $e_n < \infty$. Moreover, observe that (125) and (127) assure that for all $n \in \mathbb{N} \cap (N, \infty)$ it holds that

$$\mathbb{E}\big[V(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))\big] \leq \mathbb{E}\big[(1 - c\gamma_n)V(\Theta_{n-1})\big] \leq \mathbb{E}\big[V(\Theta_{n-1})\big] < \infty. \tag{132}$$

This, (124), and (131) imply that for all $n \in \mathbb{N} \cap (N, \infty)$ it holds that

$$\begin{split}e_n &= \mathbb{E}\big[V(\Theta_n)\big] \\ &= \mathbb{E}\big[V(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + D_n))\big] \\ &= \mathbb{E}\big[V(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + D_n)) - V(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))\big] \\ &\quad + \mathbb{E}\big[V(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))\big].\end{split} \tag{133}$$

Combining this with (132) assures that for all $n \in \mathbb{N} \cap (N, \infty)$ it holds that

$$\begin{split}\mathbb{E}\big[V(\Theta_n)\big] &\leq \mathbb{E}\big[V(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + D_n)) - V(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))\big] \\ &\quad + \mathbb{E}\big[(1 - c\gamma_n)V(\Theta_{n-1})\big] \\ &= \mathbb{E}\big[V(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}) + \gamma_n D_n) - V(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))\big] \\ &\quad + \mathbb{E}\big[(1 - c\gamma_n)V(\Theta_{n-1})\big].\end{split} \tag{134}$$

Next note that the assumption that $V \in C^1(\mathbb{R}^d, [0,\infty))$, the chain rule, and the fundamental theorem of calculus ensure that for all $x, y \in \mathbb{R}^d$ it holds that $(\mathbb{R} \ni t \mapsto V(x+ty) \in [0,\infty)) \in C^1(\mathbb{R}, [0,\infty))$ and

$$V(x+y) - V(x) = \big[V(x+ty)\big]_{t=0}^{t=1} = \int_0^1 V'(x+sy)(y)\,\mathrm{d}s. \tag{135}$$

Combining (126), (127), and (134) with Fubini's theorem hence shows that for all $n \in \mathbb{N} \cap (N,\infty)$ it holds that

$$\begin{aligned}
&\mathbb{E}\big[V(\Theta_n)\big] \\
&\leq \mathbb{E}\Bigg[\int\limits_0^1 V'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}) + s\gamma_n D_n)(\gamma_n D_n)\,\mathrm{d}s\Bigg] + (1 - c\gamma_n)\,\mathbb{E}\big[V(\Theta_{n-1})\big] \\
&= \int\limits_0^1 \mathbb{E}\big[V'(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + sD_n))(\gamma_n D_n)\big]\,\mathrm{d}s + (1 - c\gamma_n)\,\mathbb{E}\big[V(\Theta_{n-1})\big] \\
&= \gamma_n \int_0^1 \mathbb{E}\big[V'(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + sD_n))(D_n)\big]\,\mathrm{d}s \\
&\quad - \gamma_n \int_0^1 \mathbb{E}\big[V'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))(D_n)\big]\,\mathrm{d}s + (1 - c\gamma_n)\,\mathbb{E}\big[V(\Theta_{n-1})\big] \\
&\leq \gamma_n \sup_{s\in[0,1]} \mathbb{E}\big[|V'(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + sD_n))(D_n) - V'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))(D_n)|\big] \\
&\quad + (1 - c\gamma_n)\,\mathbb{E}\big[V(\Theta_{n-1})\big].
\end{aligned} \tag{136}$$

The fact that $\sup_{l\in\mathbb{N}\cap(N,\infty)} \gamma_l \leq \frac{c}{2\kappa}$ and (128) therefore ensure that for all $n \in \mathbb{N} \cap (N,\infty)$ it holds that

$$\begin{aligned}
\mathbb{E}\big[V(\Theta_n)\big] &\leq (1 - c\gamma_n)\,\mathbb{E}\big[V(\Theta_{n-1})\big] + \gamma_n\kappa\big((\gamma_n)^k + \gamma_n\mathbb{E}\big[V(\Theta_{n-1})\big]\big) \\
&= \big(1 - c\gamma_n + \kappa(\gamma_n)^2\big)\,\mathbb{E}\big[V(\Theta_{n-1})\big] + \kappa(\gamma_n)^{k+1} \\
&\leq \Big(1 - c\gamma_n + \frac{\gamma_n\kappa c}{2\kappa}\Big)\,\mathbb{E}\big[V(\Theta_{n-1})\big] + \kappa(\gamma_n)^{k+1} \\
&= \Big(1 - \frac{\gamma_n c}{2}\Big)\,\mathbb{E}\big[V(\Theta_{n-1})\big] + \kappa(\gamma_n)^{k+1}.
\end{aligned} \tag{137}$$

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

$$\begin{aligned}
e_n = \mathbb{E}\big[V(\Theta_n)\big] &\leq \Big(1 - \frac{\gamma_n c}{2}\Big)\,\mathbb{E}\big[V(\Theta_{n-1})\big] + \kappa(\gamma_n)^{k+1} \\
&= \Big(1 - \frac{\gamma_n c}{2}\Big)e_{n-1} + \kappa(\gamma_n)^{k+1}.
\end{aligned} \tag{138}$$

Combining this with (129) and Lemma 2.17 (with $N = N$, $k = k$, $\kappa = \kappa$, $c = {}^{c}\!/\!_{2}$, $e_n = e_n$, $\gamma_n = \gamma_n$ for $n \in \mathbb{N}_0$ in the notation of Lemma 2.17) demonstrates that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{aligned} \mathbb{E}\big[V(\Theta_n)\big] &= e_n \\ &\leq \left[\max\left\{\frac{e_0}{(\gamma_0)^k}, \frac{e_1}{(\gamma_1)^k}, \dots, \frac{e_N}{(\gamma_N)^k}, \frac{\kappa}{C}\right\}\right](\gamma_n)^k \\ &= \left[\max\left(\left\{\frac{\kappa}{C}\right\} \cup \left\{\frac{\mathbb{E}\big[V(\Theta_l)\big]}{(\gamma_l)^k} : l \in \{0, 1, \dots, N\}\right\}\right)\right](\gamma_n)^k. \end{aligned} \tag{139}$$

The proof of Proposition 3.2 is thus completed. □

Corollary 3.3. *Assume Setting 3.1 and let $N \in \mathbb{N}_0$, $k, \kappa, c \in (0, \infty)$, $\varrho \in (0, {}^{1}\!/\!_{c}]$, $V \in C^1(\mathbb{R}^d, [0, \infty))$ satisfy for all $m \in \mathbb{N}_0$, $n \in \mathbb{N} \cap (N, \infty)$, $r \in [0, \varrho]$, $t \in [0, 1]$, $\theta \in \mathbb{R}^d$ that*

$$\mathbb{E}\big[V(\Theta_m) + |V'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))(D_n)|\big] < \infty, \tag{140}$$

$$\textstyle\int_0^1 \mathbb{E}\big[|V'(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + sD_n))(D_n)|\big]\,\mathrm{d}s < \infty, \tag{141}$$

$$\mathbb{E}\big[V'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))(D_n)\big] = 0, \qquad V(\theta + rg(\theta)) \leq (1 - cr)V(\theta), \tag{142}$$

$$\begin{aligned} &\mathbb{E}\big[|V'(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + tD_n))(D_n) - V'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))(D_n)|\big] \\ &\leq \kappa\big((\gamma_n)^k + \gamma_n \mathbb{E}\big[V(\Theta_{n-1})\big]\big), \end{aligned} \tag{143}$$

$$\text{and} \qquad \limsup_{l \to \infty} \gamma_l = 0 < \liminf_{l \to \infty} \left[\tfrac{(\gamma_l)^k - (\gamma_{l-1})^k}{(\gamma_l)^{k+1}} + \tfrac{c(\gamma_{l-1})^k}{2(\gamma_l)^k}\right]. \tag{144}$$

Then there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}_0$ it holds that

$$\mathbb{E}\big[V(\Theta_n)\big] \leq C(\gamma_n)^k. \tag{145}$$

Proof of Corollary 3.3. First, note that (144) ensures that there exists $M \in \{N, N+1, \dots\}$ such that $\sup_{l \in \mathbb{N} \cap (M, \infty)} \gamma_l \leq \min\{{}^{c}\!/\!_{2\kappa}, \varrho\}$ and

$$\inf_{l \in \mathbb{N} \cap (M, \infty)} \left[\frac{(\gamma_l)^k - (\gamma_{l-1})^k}{(\gamma_l)^{k+1}} + \frac{c(\gamma_{l-1})^k}{2(\gamma_l)^k}\right] > 0. \tag{146}$$

Next observe that (142) and the fact that $\forall\, n \in \mathbb{N} \cap (M, \infty)\colon \gamma_n \leq \varrho$ demonstrate that for all $n \in \mathbb{N} \cap (M, \infty)$, $\theta \in \mathbb{R}^d$ it holds that

$$V(\theta + \gamma_n g(\theta)) \leq (1 - c\gamma_n)V(\theta). \tag{147}$$

The fact that $\sup_{l\in\mathbb{N}\cap(M,\infty)}\gamma_l \leq \min\{\frac{c}{2\kappa},\varrho\} \leq \min\{\frac{c}{2\kappa},\frac{2}{c}\}$, (140)–(143), (146), and Proposition 3.2 (with $N = M$ in the notation of Proposition 3.2) hence assure that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{split}&\mathbb{E}\big[V(\Theta_n)\big]\\ &\leq \max\!\left(\left\{\frac{\kappa}{\inf\limits_{l\in\mathbb{N}\cap(M,\infty)}\Big[\frac{(\gamma_l)^k-(\gamma_{l-1})^k}{(\gamma_l)^{k+1}}+\frac{c(\gamma_{l-1})^k}{2(\gamma_l)^k}\Big]}\right\}\cup\left\{\tfrac{\mathbb{E}[V(\Theta_l)]}{(\gamma_l)^k} : l\in\{0,1,\dots,M\}\right\}\right)(\gamma_n)^k.\end{split} \tag{148}$$

Combining this with (140) establishes (145). Corollary 3.3 is thus completed. □

3.3 Strong L^2-convergence rate for SAAs

Proposition 3.4 (Mean square error of stochastic approximation)**.** *Assume Setting 3.1, let $\langle\cdot,\cdot\rangle\colon\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ be a scalar product, let $\|\cdot\|\colon\mathbb{R}^d\to[0,\infty)$ be the function which satisfies for all $\theta\in\mathbb{R}^d$ that $\|\theta\| = \sqrt{\langle\theta,\theta\rangle}$, let $N\in\mathbb{N}_0$, $c,\kappa,C\in(0,\infty)$, $\vartheta\in\mathbb{R}^d$, assume for all $n\in\mathbb{N}\cap(N,\infty)$, $A\in\mathbb{F}_{n-1}$ with $\mathbb{E}[\|D_n\|]<\infty$ that $\mathbb{E}[D_n\mathbb{1}_A]=0$, and assume for all $n\in\mathbb{N}$, $\theta\in\mathbb{R}^d$ that*

$$\mathbb{E}\big[\|\Theta_0\|^2\big]<\infty, \qquad \mathbb{E}\big[\|D_n\|^2\big]\leq\kappa\big(1+\mathbb{E}\big[\|\Theta_{n-1}-\vartheta\|^2\big]\big), \tag{149}$$

$$\sup_{l\in\mathbb{N}\cap(N,\infty)}\gamma_l\leq\min\big\{\tfrac{c}{4\kappa},c\big\}, \qquad C=\inf_{l\in\mathbb{N}\cap(N,\infty)}\left[\tfrac{\gamma_l-\gamma_{l-1}}{(\gamma_l)^2}+\tfrac{c\gamma_{l-1}}{2\gamma_l}\right], \tag{150}$$

$$\text{and} \qquad \langle\theta-\vartheta,g(\theta)\rangle\leq -c\max\big\{\|\theta-\vartheta\|^2,\|g(\theta)\|^2\big\}. \tag{151}$$

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

$$\mathbb{E}\big[\|\Theta_n-\vartheta\|^2\big]\leq\gamma_n\max\!\Big(\big\{\tfrac{2\kappa}{C}\big\}\cup\Big\{\tfrac{\mathbb{E}[\|\Theta_l-\vartheta\|^2]}{\gamma_l}:l\in\{0,1,\dots,N\}\Big\}\Big)<\infty. \tag{152}$$

Proof of Proposition 3.4. Throughout this proof let $V\colon\mathbb{R}^d\to[0,\infty)$ be the function which satisfies for all $\theta\in\mathbb{R}^d$ that

$$V(\theta)=\|\theta-\vartheta\|^2. \tag{153}$$

Observe that (151) and Lemma 2.12 imply that for all $\theta\in\mathbb{R}^d$, $r\in[0,c]$ it holds that

$$c\leq 1\leq {}^{1}\!/_{c}, \qquad \|g(\theta)\|\leq\tfrac{1}{c}\|\theta-\vartheta\|, \qquad \text{and} \tag{154}$$

$$V(\theta+rg(\theta))=\|\theta+rg(\theta)-\vartheta\|^2\leq(1-cr)\|\theta-\vartheta\|^2=(1-cr)V(\theta). \tag{155}$$

This and (150) ensure that for all $n \in \mathbb{N} \cap (N, \infty)$, $\theta \in \mathbb{R}^d$ it holds that

$$V(\theta + \gamma_n g(\theta)) \leq (1 - c\gamma_n)V(\theta). \tag{156}$$

In addition, note that (150) and (154) show that

$$\sup_{l \in \mathbb{N} \cap (N,\infty)} \gamma_l \leq \min\bigl\{\tfrac{c}{4\kappa}, c\bigr\} \leq \min\bigl\{\tfrac{c}{4\kappa}, 1\bigr\} \leq \min\bigl\{\tfrac{c}{4\kappa}, \tfrac{1}{c}\bigr\} \leq \min\bigl\{\tfrac{c}{4\kappa}, \tfrac{2}{c}\bigr\}. \tag{157}$$

Next we claim that for all $n \in \mathbb{N}$ it holds that

$$\mathbb{E}\bigl[V(\Theta_{n-1})\bigr] = \mathbb{E}\bigl[\|\Theta_{n-1} - \vartheta\|^2\bigr] < \infty \qquad \text{and} \qquad \mathbb{E}\bigl[\|D_n\|^2\bigr] < \infty. \tag{158}$$

We now prove (158) by induction on $n \in \mathbb{N}$. For the base case $n = 1$ note that (149) implies that

$$\mathbb{E}\bigl[\|\Theta_0 - \vartheta\|^2\bigr] < \infty \qquad \text{and} \qquad \mathbb{E}\bigl[\|D_1\|^2\bigr] \leq \kappa\bigl(1 + \mathbb{E}\bigl[\|\Theta_0 - \vartheta\|^2\bigr]\bigr) < \infty. \tag{159}$$

This establishes (158) in the base case $n = 1$. For the induction step $\mathbb{N} \ni n \to n + 1 \in \{2, 3, \ldots\}$ observe that (124) and (154) ensure that for all $n \in \mathbb{N}$ with $\mathbb{E}\bigl[V(\Theta_{n-1}) + \|D_n\|^2\bigr] < \infty$ it holds that

$$\begin{aligned} \mathbb{E}\bigl[\|\Theta_n - \vartheta\|^2\bigr] &= \mathbb{E}\bigl[\|\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + D_n) - \vartheta\|^2\bigr] \\ &\leq \mathbb{E}\Bigl[\bigl(\|\Theta_{n-1} - \vartheta\| + \gamma_n\|g(\Theta_{n-1})\| + \gamma_n\|D_n\|\bigr)^2\Bigr] \\ &\leq \mathbb{E}\Bigl[\bigl((1 + \tfrac{\gamma_n}{c})\|\Theta_{n-1} - \vartheta\| + \gamma_n\|D_n\|\bigr)^2\Bigr] < \infty. \end{aligned} \tag{160}$$

This and (149) imply that for all $n \in \mathbb{N}$ with $\mathbb{E}\bigl[V(\Theta_{n-1}) + \|D_n\|^2\bigr] < \infty$ it holds that

$$\mathbb{E}\bigl[\|D_{n+1}\|^2\bigr] \leq \kappa\bigl(1 + \mathbb{E}\bigl[\|\Theta_n - \vartheta\|^2\bigr]\bigr) = \kappa\bigl(1 + \mathbb{E}[V(\Theta_n)]\bigr) < \infty. \tag{161}$$

Induction thus proves (158). Next note that Lemma 2.4 implies that for all $\theta, v \in \mathbb{R}^d$ it holds that

$$V \in C^1(\mathbb{R}^d, [0, \infty)) \qquad \text{and} \qquad V'(\theta)(v) = 2\langle\theta - \vartheta, v\rangle. \tag{162}$$

Furthermore, observe that (154) and (158) prove that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \mathbb{E}\bigl[\|\Theta_{n-1} + \gamma_n g(\Theta_{n-1}) - \vartheta\|^2\bigr] &\leq \mathbb{E}\bigl[(\|\Theta_{n-1} - \vartheta\| + \gamma_n\|g(\Theta_{n-1})\|)^2\bigr] \\ &\leq \mathbb{E}\bigl[(\|\Theta_{n-1} - \vartheta\| + \tfrac{\gamma_n}{c}\|\Theta_{n-1} - \vartheta\|)^2\bigr] \\ &= \mathbb{E}\bigl[([1 + \tfrac{\gamma_n}{c}]\,\|\Theta_{n-1} - \vartheta\|)^2\bigr] \\ &= [1 + \tfrac{\gamma_n}{c}]^2\,\mathbb{E}\bigl[\|\Theta_{n-1} - \vartheta\|^2\bigr] \\ &= [1 + \tfrac{\gamma_n}{c}]^2\,\mathbb{E}\bigl[V(\Theta_{n-1})\bigr] < \infty. \end{aligned} \tag{163}$$

The Cauchy-Schwarz inequality, (158), and (162) therefore ensure that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}&\mathbb{E}\big[|V'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))(D_n)|\big] = \mathbb{E}\big[2|\langle \Theta_{n-1} + \gamma_n g(\Theta_{n-1}) - \vartheta, D_n\rangle|\big]\\ &\leq 2\,\mathbb{E}\big[\|\Theta_{n-1} + \gamma_n g(\Theta_{n-1}) - \vartheta\|\|D_n\|\big]\\ &\leq 2\left(\mathbb{E}\big[\|\Theta_{n-1} + \gamma_n g(\Theta_{n-1}) - \vartheta\|^2\big]\right)^{1/2}\left(\mathbb{E}\big[\|D_n\|^2\big]\right)^{1/2} < \infty.\end{aligned} \tag{164}$$

Combining (158) and (162) hence demonstrates that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}&\int_0^1 \mathbb{E}\big[|V'(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + sD_n))(D_n)|\big]\,\mathrm{d}s\\ &= \int_0^1 \mathbb{E}\big[2|\langle \Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + sD_n) - \vartheta, D_n\rangle|\big]\,\mathrm{d}s\\ &= \int_0^1 \mathbb{E}\big[2|\langle \Theta_{n-1} + \gamma_n g(\Theta_{n-1}) - \vartheta, D_n\rangle + s\gamma_n\|D_n\|^2|\big]\,\mathrm{d}s\\ &\leq \mathbb{E}\big[2|\langle \Theta_{n-1} + \gamma_n g(\Theta_{n-1}) - \vartheta, D_n\rangle|\big] + \gamma_n\,\mathbb{E}\big[\|D_n\|^2\big]\\ &\leq 2\big(\mathbb{E}\big[\|\Theta_{n-1} + \gamma_n g(\Theta_{n-1}) - \vartheta\|^2\big]\big)^{1/2}\left(\mathbb{E}\big[\|D_n\|^2\big]\right)^{1/2} + \gamma_n\,\mathbb{E}\big[\|D_n\|^2\big] < \infty.\end{aligned} \tag{165}$$

Moreover, observe that the fact that for all $n \in \mathbb{N}$ it holds that the function Θ_{n-1} is $\mathbb{F}_{n-1}/\mathcal{B}(\mathbb{R}^d)$-measurable, (158), the fact that for all $n \in \mathbb{N}\cap(N,\infty)$, $A \in \mathbb{F}_{n-1}$ it holds that $\mathbb{E}[D_n\mathbb{1}_A] = 0$, (162), (164), and Lemma 2.5 prove that for all $n \in \mathbb{N}\cap(N,\infty)$ it holds that

$$\mathbb{E}\big[V'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))(D_n)\big] = 2\,\mathbb{E}\big[\langle \Theta_{n-1} + \gamma_n g(\Theta_{n-1}) - \vartheta, D_n\rangle\big] = 0. \tag{166}$$

Furthermore, note that (149) and (162) imply that for all $n \in \mathbb{N}$, $t \in [0,1]$ it holds that

$$\begin{aligned}&\mathbb{E}\big[|V'(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + tD_n))(D_n) - V'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))(D_n)|\big]\\ &= \mathbb{E}\big[2|\langle \Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + tD_n) - (\Theta_{n-1} + \gamma_n g(\Theta_{n-1})), D_n\rangle|\big]\\ &= \mathbb{E}\big[2t\gamma_n\|D_n\|^2\big] = 2t\gamma_n\,\mathbb{E}\big[\|D_n\|^2\big] \leq 2\gamma_n\,\mathbb{E}\big[\|D_n\|^2\big]\\ &\leq 2\gamma_n\kappa\big(1 + \mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^2\big]\big) = 2\kappa\big(\gamma_n + \gamma_n\,\mathbb{E}\big[V(\Theta_{n-1})\big]\big).\end{aligned} \tag{167}$$

Combining the fact that $\sup_{l\in\mathbb{N}\cap(N,\infty)}\gamma_l \leq \min\{\frac{c}{4\kappa}, \frac{2}{c}\}$, (150), (156), (158), (164), (165), and (166) with Proposition 3.2 (with $N = N$, $k = 1$, $\kappa = 2\kappa$, $c = c$, and $V = V$ in the notation of Proposition 3.2) therefore demonstrates that for all $n \in \mathbb{N}_0$

it holds that

$$\begin{aligned}
&\mathbb{E}\big[\|\Theta_n - \vartheta\|^2\big] \\
&= \mathbb{E}\big[V(\Theta_n)\big] \\
&\le \gamma_n \max\!\left(\left\{\frac{2\kappa}{C}\right\} \cup \left\{\frac{\mathbb{E}\big[V(\Theta_l)\big]}{\gamma_l} : l \in \{0, 1, \ldots, N\}\right\}\right) \\
&= \gamma_n \max\!\left(\left\{\frac{2\kappa}{C}\right\} \cup \left\{\frac{\mathbb{E}\big[\|\Theta_l - \vartheta\|^2\big]}{\gamma_l} : l \in \{0, 1, \ldots, N\}\right\}\right) < \infty.
\end{aligned} \tag{168}$$

The proof of Proposition 3.4 is thus completed. □

Corollary 3.5. *Assume Setting 3.1, let $\langle \cdot, \cdot \rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a scalar product, let $\|\cdot\| \colon \mathbb{R}^d \to [0, \infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that $\|\theta\| = \sqrt{\langle \theta, \theta \rangle}$, let $c, \kappa \in (0, \infty)$, $\vartheta \in \mathbb{R}^d$, assume for all $n \in \mathbb{N}$, $A \in \mathbb{F}_{n-1}$ with $\mathbb{E}[\|D_n\|] < \infty$ that $\mathbb{E}[D_n \mathbb{1}_A] = 0$, and assume for all $n \in \mathbb{N}$, $\theta \in \mathbb{R}^d$ that*

$$\mathbb{E}\big[\|\Theta_0\|^2\big] < \infty, \qquad \mathbb{E}\big[\|D_n\|^2\big] \le \kappa\big(1 + \mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^2\big]\big), \tag{169}$$

$$\limsup_{l \to \infty} \gamma_l = 0 < \liminf_{l \to \infty} \left[\tfrac{\gamma_l - \gamma_{l-1}}{(\gamma_l)^2} + \tfrac{c\gamma_{l-1}}{2\gamma_l}\right], \tag{170}$$

$$\text{and} \qquad \langle \theta - \vartheta, g(\theta) \rangle \le -c \max\big\{\|\theta - \vartheta\|^2, \|g(\theta)\|^2\big\}. \tag{171}$$

Then there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}_0$ it holds that

$$\mathbb{E}\big[\|\Theta_n - \vartheta\|^2\big] \le C\gamma_n. \tag{172}$$

Proof of Corollary 3.5. Observe that (170) ensures that there exists $N \in \mathbb{N}_0$ such that

$$\sup_{l \in \mathbb{N} \cap (N, \infty)} \gamma_l \le \min\!\left\{\frac{c}{4\kappa}, c\right\} \qquad \text{and} \qquad \inf_{l \in \mathbb{N} \cap (N, \infty)} \left[\frac{\gamma_l - \gamma_{l-1}}{(\gamma_l)^2} + \frac{c\gamma_{l-1}}{2\gamma_l}\right] > 0. \tag{173}$$

Proposition 3.4 therefore establishes that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{aligned}
&\mathbb{E}\big[\|\Theta_n - \vartheta\|^2\big] \\
&\le \gamma_n \max\!\left(\left\{\frac{2\kappa}{\inf_{l \in \mathbb{N} \cap (N, \infty)} \left[\frac{\gamma_l - \gamma_{l-1}}{(\gamma_l)^2} + \frac{c\gamma_{l-1}}{2\gamma_l}\right]}\right\} \cup \left\{\tfrac{\mathbb{E}[\|\Theta_l - \vartheta\|^2]}{\gamma_l} : l \in \{0, 1, \ldots, N\}\right\}\right) < \infty.
\end{aligned} \tag{174}$$

This completes the proof of Corollary 3.5. □

3.4 Strong L^p-convergence rate for SAAs

Proposition 3.6 (L^p-convergence rate for stochastic approximation)**.** *Assume Setting 3.1, let $\langle \cdot, \cdot \rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a scalar product, let $\|\cdot\| \colon \mathbb{R}^d \to [0, \infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that $\|\theta\| = \sqrt{\langle \theta, \theta \rangle}$, let $p \in \{2, 4, 6, \ldots\}$, $c, \kappa \in (0, \infty)$, $\vartheta \in \mathbb{R}^d$, assume for all $n \in \mathbb{N}$, $A \in \mathbb{F}_{n-1}$ with $\mathbb{E}[\|D_n\|] < \infty$ that $\mathbb{E}[D_n \mathbb{1}_A] = 0$, and assume for all $n \in \mathbb{N}$, $A \in \mathbb{F}_{n-1}$, $\theta \in \mathbb{R}^d$ that*

$$\mathbb{E}\big[\|\Theta_0\|^p\big] < \infty, \qquad \mathbb{E}\big[\|D_n\|^p \mathbb{1}_A\big] \leq \kappa \, \mathbb{E}\big[(1 + \|\Theta_{n-1} - \vartheta\|^p) \mathbb{1}_A\big], \tag{175}$$

$$\limsup_{l \to \infty} \gamma_l = 0 < \min_{k \in \{1, 2, \ldots, p/2\}} \Big(\liminf_{l \to \infty} \Big[\tfrac{(\gamma_l)^k - (\gamma_{l-1})^k}{(\gamma_l)^{k+1}} + \tfrac{c(\gamma_{l-1})^k}{2(\gamma_l)^k} \Big] \Big), \tag{176}$$

$$\textit{and} \qquad \langle \theta - \vartheta, g(\theta) \rangle \leq -c \max\big\{\|\theta - \vartheta\|^2, \|g(\theta)\|^2\big\}. \tag{177}$$

Then there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}_0$ it holds that

$$\big(\mathbb{E}\big[\|\Theta_n - \vartheta\|^p\big]\big)^{1/p} \leq C(\gamma_n)^{1/2}. \tag{178}$$

Proof of Proposition 3.6. Throughout this proof assume w.l.o.g. that $\kappa \geq 1$, let $N \in \mathbb{N}$ satisfy $\sup_{n \in \mathbb{N} \cap (N, \infty)} \gamma_n \leq c$, and let $V_q \colon \mathbb{R}^d \to [0, \infty)$, $q \in \mathbb{N}$, be the functions which satisfy for all $q \in \mathbb{N}$, $\theta \in \mathbb{R}^d$ that

$$V_q(\theta) = \|\theta - \vartheta\|^q. \tag{179}$$

Note that Lemma 2.12 implies that for all $q \in \{2, 4, 6, \ldots\} \cap [2, p]$, $\theta \in \mathbb{R}^d$, $r \in [0, c]$ it holds that

$$c \leq 1 \leq 1/c, \qquad \|g(\theta)\| \leq \tfrac{1}{c}\|\theta - \vartheta\|, \qquad \text{and} \tag{180}$$

$$V_q(\theta + rg(\theta)) = \|\theta + rg(\theta) - \vartheta\|^q \leq (1 - cr)^{q/2} \|\theta - \vartheta\|^q \leq (1 - cr) V_q(\theta). \tag{181}$$

In the next step we claim that for all $n \in \mathbb{N}$ it holds that

$$\mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^p\big] < \infty \qquad \text{and} \qquad \mathbb{E}\big[\|D_n\|^p\big] < \infty. \tag{182}$$

We now prove (182) by induction on $n \in \mathbb{N}$. For the base case $n = 1$ note that (175) implies that

$$\mathbb{E}\big[\|\Theta_0 - \vartheta\|^p\big] < \infty \qquad \text{and} \qquad \mathbb{E}\big[\|D_1\|^p\big] \leq \kappa\big(1 + \mathbb{E}\big[\|\Theta_0 - \vartheta\|^p\big]\big) < \infty. \tag{183}$$

This establishes (182) in the base case $n = 1$. For the induction step $\mathbb{N} \ni n \to n + 1 \in \{2, 3, \ldots\}$ observe that (124) and (180) ensure that for all $n \in \mathbb{N}$ with $\mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^p + \|D_n\|^p\big] < \infty$ it holds that

$$\begin{aligned} \mathbb{E}\big[\|\Theta_n - \vartheta\|^p\big] &= \mathbb{E}\big[\|\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + D_n) - \vartheta\|^p\big] \\ &\leq \mathbb{E}\big[\big(\|\Theta_{n-1} - \vartheta\| + \gamma_n\|g(\Theta_{n-1})\| + \gamma_n\|D_n\|\big)^p\big] \\ &\leq \mathbb{E}\big[\big((1 + \tfrac{\gamma_n}{c})\|\Theta_{n-1} - \vartheta\| + \gamma_n\|D_n\|\big)^p\big] < \infty. \end{aligned} \tag{184}$$

This and (175) imply that for all $n \in \mathbb{N}$ with $\mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^p + \|D_n\|^p\big] < \infty$ it holds that

$$\mathbb{E}\big[\|D_{n+1}\|^p\big] \le \kappa\big(1 + \mathbb{E}\big[\|\Theta_n - \vartheta\|^p\big]\big) < \infty. \tag{185}$$

Induction thus proves (182). Next note that the conditional Jensen inequality (see, e.g., Klenke [50, Theorem 8.20]) and the fact that for all $q \in \{1, 2, \ldots\} \cap [0, p]$ it holds that the function $\mathbb{R} \ni z \mapsto |z|^{p/q} \in [0, \infty)$ is convex ensure that for all $q \in \{1, 2, \ldots\} \cap [0, p]$, $n \in \mathbb{N}$ it holds $\mathbb{P}$-a.s. that

$$\big|\mathbb{E}\big[\|D_n\|^q \,|\, \mathbb{F}_{n-1}\big]\big|^{p/q} \le \mathbb{E}\big[\|D_n\|^p \,|\, \mathbb{F}_{n-1}\big]. \tag{186}$$

Hence, we obtain that for all $q \in \{1, 2, \ldots\} \cap [0, p]$, $n \in \mathbb{N}$ it holds $\mathbb{P}$-a.s. that

$$\mathbb{E}\big[\|D_n\|^q \,|\, \mathbb{F}_{n-1}\big] \le \big|\mathbb{E}\big[\|D_n\|^p \,|\, \mathbb{F}_{n-1}\big]\big|^{q/p}. \tag{187}$$

Moreover, observe that (175) and (182) demonstrate that for all $n \in \mathbb{N}$ it holds $\mathbb{P}$-a.s. that

$$\mathbb{E}\big[\|D_n\|^p \,|\, \mathbb{F}_{n-1}\big] \le \kappa\big(1 + \|\Theta_{n-1} - \vartheta\|^p\big). \tag{188}$$

Combining this with (187) proves that for all $q \in \{1, 2, \ldots\} \cap [0, p]$, $n \in \mathbb{N}$ it holds $\mathbb{P}$-a.s. that

$$\mathbb{E}\big[\|D_n\|^q \,|\, \mathbb{F}_{n-1}\big] \le \big|\mathbb{E}\big[\|D_n\|^p \,|\, \mathbb{F}_{n-1}\big]\big|^{q/p} \le \big[\kappa\big(1 + \|\Theta_{n-1} - \vartheta\|^p\big)\big]^{q/p}. \tag{189}$$

The fact that $\forall\, x, y \in [0, \infty)$, $r \in (0, 1]\colon (x + y)^r \le x^r + y^r$ and the assumption that $\kappa \ge 1$ hence assure that for all $q \in \{1, 2, \ldots\} \cap [0, p]$, $n \in \mathbb{N}$ it holds $\mathbb{P}$-a.s. that

$$\begin{aligned}
\mathbb{E}\big[\|D_n\|^q \,|\, \mathbb{F}_{n-1}\big] &\le \big[\kappa\big(1 + \|\Theta_{n-1} - \vartheta\|^p\big)\big]^{q/p} \\
&\le \kappa^{q/p}\big(1 + \|\Theta_{n-1} - \vartheta\|^q\big) \\
&\le \kappa\big(1 + \|\Theta_{n-1} - \vartheta\|^q\big).
\end{aligned} \tag{190}$$

The tower property for conditional expectations therefore shows that for all $q \in \{1, 2, \ldots\} \cap [0, p]$, $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
\mathbb{E}[\|D_n\|^q] &= \mathbb{E}\big[\mathbb{E}[\|D_n\|^q \,|\, \mathbb{F}_{n-1}]\big] \\
&\le \mathbb{E}\big[\kappa\big(1 + \|\Theta_{n-1} - \vartheta\|^q\big)\big] \\
&= \kappa\big(1 + \mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^q\big]\big).
\end{aligned} \tag{191}$$

Next we claim that for all $q \in \{2, 4, 6, \ldots\} \cap [0, p]$ there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}_0$ it holds that

$$\mathbb{E}\big[\|\Theta_n - \vartheta\|^q\big] \le C(\gamma_n)^{q/2}. \tag{192}$$

We now prove (192) by induction on $q \in \{2, 4, 6, \ldots\} \cap [0, p]$. Observe that Corollary 3.5 and (191) establish (192) in the base case $q = 2$. For the induction step $\{2, 4, 6, \ldots\} \cap [0, p-2] \ni (q-2) \to q \in \{4, 6, 8, \ldots\} \cap [0, p]$ let $q \in \{4, 6, 8, \ldots\} \cap [0, p]$, $C \in (0, \infty)$ satisfy for all $n \in \mathbb{N}_0$ that

$$\mathbb{E}\big[\|\Theta_n - \vartheta\|^{q-2}\big] \leq C(\gamma_n)^{(q-2)/2}. \tag{193}$$

Note that (182) and Jensen's inequality ensure that for all $n \in \mathbb{N}_0$ it holds that

$$\mathbb{E}\big[V_q(\Theta_n)\big] = \mathbb{E}\big[\|\Theta_n - \vartheta\|^q\big] < \infty. \tag{194}$$

Next observe that (180) implies that for all $\theta \in \mathbb{R}^d$, $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \|\theta + \gamma_n g(\theta) - \vartheta\|^{q-1} &\leq (\|\theta - \vartheta\| + \|\gamma_n g(\theta)\|)^{q-1} \\ &\leq \big(\|\theta - \vartheta\| + |\tfrac{\gamma_n}{c}|\|\theta - \vartheta\|\big)^{q-1} \\ &= \big(1 + \tfrac{\gamma_n}{c}\big)^{q-1} \|\theta - \vartheta\|^{q-1}. \end{aligned} \tag{195}$$

Combining this and (182) with Jensen's inequality ensures that for all $n \in \mathbb{N}$ it holds that

$$\mathbb{E}\big[\|\Theta_{n-1} + \gamma_n g(\Theta_{n-1}) - \vartheta\|^{q-1}\big] \leq \big(1 + \tfrac{\gamma_n}{c}\big)^{q-1} \mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^{q-1}\big] < \infty. \tag{196}$$

Moreover, note that (195), the tower property for conditional expectations, and the fact that for all $n \in \mathbb{N}$ it holds that Θ_{n-1} is $\mathbb{F}_{n-1}/\mathcal{B}(\mathbb{R}^d)$-measurable assure that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} &\mathbb{E}\big[\|\Theta_{n-1} + \gamma_n g(\Theta_{n-1}) - \vartheta\|^{q-1}\|D_n\|\big] \\ &\leq \big(1 + \tfrac{\gamma_n}{c}\big)^{q-1} \mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^{q-1}\|D_n\|\big] \\ &= \big(1 + \tfrac{\gamma_n}{c}\big)^{q-1} \mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^{q-1}\mathbb{E}[\|D_n\| \,|\, \mathbb{F}_{n-1}]\big]. \end{aligned} \tag{197}$$

Combining this with (182), (190), and Jensen's inequality proves that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} &\mathbb{E}\big[\|\Theta_{n-1} + \gamma_n g(\Theta_{n-1}) - \vartheta\|^{q-1}\|D_n\|\big] \\ &\leq \kappa \big(1 + \tfrac{\gamma_n}{c}\big)^{q-1} \mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^{q-1} (1 + \|\Theta_{n-1} - \vartheta\|)\big] \\ &= \kappa \big(1 + \tfrac{\gamma_n}{c}\big)^{q-1} \big(\mathbb{E}[\|\Theta_{n-1} - \vartheta\|^{q-1}] + \mathbb{E}\,[\|\Theta_{n-1} - \vartheta\|^q]\big) < \infty. \end{aligned} \tag{198}$$

Furthermore, observe that Lemma 2.4 implies that for all $\theta, v \in \mathbb{R}^d$ it holds that

$$V_q \in C^1(\mathbb{R}^d, [0, \infty)) \qquad \text{and} \qquad V_q'(\theta)(v) = q\|\theta - \vartheta\|^{q-2}\langle \theta - \vartheta, v\rangle. \tag{199}$$

This, (198), and the Cauchy-Schwarz inequality demonstrate that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} &\mathbb{E}\big[|V_q'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))(D_n)|\big] \\ &= q\,\mathbb{E}\big[\|\Theta_{n-1} + \gamma_n g(\Theta_{n-1}) - \vartheta\|^{q-2}|\langle \Theta_{n-1} + \gamma_n g(\Theta_{n-1}) - \vartheta, D_n\rangle|\big] \\ &\leq q\,\mathbb{E}\big[\|\Theta_{n-1} + \gamma_n g(\Theta_{n-1}) - \vartheta\|^{q-1}\|D_n\|\big] < \infty. \end{aligned} \tag{200}$$

Furthermore, note that (199), the Cauchy-Schwarz inequality, and Lemma 2.1 (with $p = q-1$ in the notation of Lemma 2.1) imply that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} &\int_0^1 \mathbb{E}\big[|V_q'(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + sD_n))(D_n)|\big]\,\mathrm{d}s \\ &= \int_0^1 \mathbb{E}\big[q\|\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + sD_n) - \vartheta\|^{q-2} \\ &\quad \cdot |\langle \Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + sD_n) - \vartheta, D_n\rangle|\big]\,\mathrm{d}s \\ &\leq q\int_0^1 \mathbb{E}\big[\|\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + sD_n) - \vartheta\|^{q-1}\|D_n\|\big]\,\mathrm{d}s \\ &\leq q2^{q-1}\int_0^1 \mathbb{E}\big[\big(\|\Theta_{n-1} + \gamma_n g(\Theta_{n-1}) - \vartheta\|^{q-1} + s^{q-1}|\gamma_n|^{q-1}\|D_n\|^{q-1}\big)\|D_n\|\big]\,\mathrm{d}s. \end{aligned} \tag{201}$$

This and (195) ensure that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} &\int_0^1 \mathbb{E}\big[|V_q'(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + sD_n))(D_n)|\big]\,\mathrm{d}s \\ &\leq q2^{q-1}\int_0^1 \mathbb{E}\Big[\big(\big[1 + \tfrac{\gamma_n}{c}\big]^{q-1}\|\Theta_{n-1} - \vartheta\|^{q-1} + s^{q-1}|\gamma_n|^{q-1}\|D_n\|^{q-1}\big)\|D_n\|\Big]\,\mathrm{d}s \\ &\leq q2^{q-1}\Big(\big[1 + \tfrac{\gamma_n}{c}\big]^{q-1}\mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^{q-1}\|D_n\|\big] + |\gamma_n|^{q-1}\mathbb{E}\big[\|D_n\|^q\big]\Big). \end{aligned} \tag{202}$$

The tower property for conditional expectations, (182), (190), and Jensen's inequality hence demonstrate that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
&\int_0^1 \mathbb{E}\big[|V_q'(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + sD_n))(D_n)|\big]\,\mathrm{d}s \\
&\leq q2^{q-1}\big[1 + \tfrac{\gamma_n}{c}\big]^{q-1}\mathbb{E}\Big[\|\Theta_{n-1} - \vartheta\|^{q-1}\mathbb{E}\big[\|D_n\|\,|\,\mathbb{F}_{n-1}\big]\Big] + q2^{q-1}|\gamma_n|^{q-1}\mathbb{E}\big[\|D_n\|^q\big] \\
&\leq \kappa q2^{q-1}\big[1 + \tfrac{\gamma_n}{c}\big]^{q-1}\mathbb{E}\Big[\|\Theta_{n-1} - \vartheta\|^{q-1}\big(1 + \|\Theta_{n-1} - \vartheta\|\big)\Big] + q2^{q-1}|\gamma_n|^{q-1}\mathbb{E}\big[\|D_n\|^q\big] \\
&\leq q2^{q-1}\big[1 + \tfrac{\gamma_n}{c}\big]^{q-1}\max\{\kappa, (\gamma_n)^{q-1}\}\,\mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^{q-1} + \|\Theta_{n-1} - \vartheta\|^q + \|D_n\|^q\big] \\
&< \infty.
\end{aligned} \tag{203}$$

Next observe that Jensen's inequality, the hypothesis that for all $n \in \mathbb{N}$, $A \in \mathbb{F}_{n-1}$ with $\mathbb{E}[\|D_n\|] < \infty$ it holds that $\mathbb{E}[D_n \mathbb{1}_A] = 0$, and (182) ensure that for all $n \in \mathbb{N}$, $A \in \mathbb{F}_{n-1}$ it holds that

$$\mathbb{E}[\|D_n\|] < \infty \qquad \text{and} \qquad \mathbb{E}[D_n \mathbb{1}_A] = 0. \tag{204}$$

This, the fact that for all $n \in \mathbb{N}$ it holds that the function Θ_{n-1} is $\mathbb{F}_{n-1}/\mathcal{B}(\mathbb{R}^d)$-measurable, (182), (196), (198), (199), (200), and Lemma 2.5 assure that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
&\mathbb{E}[V_q'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))(D_n)] \\
&= q\,\mathbb{E}\big[\big\langle \|\Theta_{n-1} + \gamma_n g(\Theta_{n-1}) - \vartheta\|^{q-2}(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}) - \vartheta), D_n\big\rangle\big] = 0.
\end{aligned} \tag{205}$$

In addition, note that (199) ensures that for all $n \in \mathbb{N} \cap (N, \infty)$, $t \in [0, 1]$ it holds that

$$
\begin{aligned}
&\mathbb{E}\Big[\big|V_q'(\Theta_{n-1}+\gamma_n(g(\Theta_{n-1})+tD_n))(D_n)-V_q'(\Theta_{n-1}+\gamma_n g(\Theta_{n-1}))(D_n)\big|\Big]\\
&=\mathbb{E}\Big[q\,\big|\|\Theta_{n-1}+\gamma_n(g(\Theta_{n-1})+tD_n)-\vartheta\|^{q-2}\langle\Theta_{n-1}+\gamma_n(g(\Theta_{n-1})+tD_n)-\vartheta,D_n\rangle\\
&\quad-\|\Theta_{n-1}+\gamma_n g(\Theta_{n-1})-\vartheta\|^{q-2}\langle\Theta_{n-1}+\gamma_n g(\Theta_{n-1})-\vartheta,D_n\rangle\big|\Big]\\
&\le q\,\mathbb{E}\Big[\big|\|\Theta_{n-1}+\gamma_n(g(\Theta_{n-1})+tD_n)-\vartheta\|^{q-2}-\|\Theta_{n-1}+\gamma_n g(\Theta_{n-1})-\vartheta\|^{q-2}\big|\\
&\quad\cdot|\langle\Theta_{n-1}+\gamma_n g(\Theta_{n-1})-\vartheta,D_n\rangle|\Big]\\
&\quad+q\,\mathbb{E}\Big[\|\Theta_{n-1}+\gamma_n(g(\Theta_{n-1})+tD_n)-\vartheta\|^{q-2}|\langle\gamma_n tD_n,D_n\rangle|\Big]\\
&=q\,\mathbb{E}\Big[\big|\|\Theta_{n-1}+\gamma_n(g(\Theta_{n-1})+tD_n)-\vartheta\|^{q-2}-\|\Theta_{n-1}+\gamma_n g(\Theta_{n-1})-\vartheta\|^{q-2}\big|\\
&\quad\cdot|\langle\Theta_{n-1}+\gamma_n g(\Theta_{n-1})-\vartheta,D_n\rangle|\Big]\\
&\quad+q\gamma_n t\,\mathbb{E}\Big[\|\Theta_{n-1}+\gamma_n(g(\Theta_{n-1})+tD_n)-\vartheta\|^{q-2}\|D_n\|^2\Big].
\end{aligned}
\tag{206}
$$

Lemma 2.1 (with $p=q-2$ in the notation of Lemma 2.1), Lemma 2.3 (with $p=q-2$ in the notation of Lemma 2.3), and the Cauchy-Schwarz inequality hence demonstrate that for all $n\in\mathbb{N}\cap(N,\infty)$, $t\in[0,1]$ it holds that

$$
\begin{aligned}
&\mathbb{E}\Big[\big|V_q'(\Theta_{n-1}+\gamma_n(g(\Theta_{n-1})+tD_n))(D_n)-V_q'(\Theta_{n-1}+\gamma_n g(\Theta_{n-1}))(D_n)\big|\Big]\\
&\le q\,\mathbb{E}\Big[2^{q-2}\|\gamma_n tD_n\|\big(\|\Theta_{n-1}+\gamma_n g(\Theta_{n-1})-\vartheta\|^{q-3}+\|\gamma_n tD_n\|^{q-3}\big)\\
&\quad\cdot\|\Theta_{n-1}+\gamma_n g(\Theta_{n-1})-\vartheta\|\|D_n\|\Big]\\
&\quad+q\gamma_n t\,\mathbb{E}\Big[2^{q-2}\big(\|\Theta_{n-1}+\gamma_n g(\Theta_{n-1})-\vartheta\|^{q-2}+\|\gamma_n tD_n\|^{q-2}\big)\|D_n\|^2\Big]\\
&=q2^{q-2}\,\mathbb{E}\Big[\gamma_n t\|D_n\|^2\big(\|\Theta_{n-1}+\gamma_n g(\Theta_{n-1})-\vartheta\|^{q-3}+(\gamma_n t)^{q-3}\|D_n\|^{q-3}\big)\\
&\quad\cdot\|\Theta_{n-1}+\gamma_n g(\Theta_{n-1})-\vartheta\|\\
&\quad+\gamma_n t\big(\|\Theta_{n-1}+\gamma_n g(\Theta_{n-1})-\vartheta\|^{q-2}+(\gamma_n t)^{q-2}\|D_n\|^{q-2}\big)\|D_n\|^2\Big].
\end{aligned}
\tag{207}
$$

In addition, observe that (180) and (181) ensure that for all $r\in[0,c]$, $\theta\in\mathbb{R}^d$ it holds that $\|\theta+rg(\theta)-\vartheta\|\le\|\theta-\vartheta\|$. This, the fact that $\sup_{n\in\mathbb{N}\cap(N,\infty)}\gamma_n\le c$, and (207) assure that for all $n \in \mathbb{N} \cap (N, \infty)$, $t \in [0,1]$ it holds that

$$\begin{aligned}
&\mathbb{E}\Big[\big|V_q'(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + tD_n))(D_n) - V_q'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))(D_n)\big|\Big] \\
&\quad \leq q2^{q-2}\gamma_n\,\mathbb{E}\Big[t\|D_n\|^2\big(\|\Theta_{n-1} - \vartheta\|^{q-3} + (\gamma_n t)^{q-3}\|D_n\|^{q-3}\big)\|\Theta_{n-1} - \vartheta\| \\
&\qquad + t\big(\|\Theta_{n-1} - \vartheta\|^{q-2} + (\gamma_n t)^{q-2}\|D_n\|^{q-2}\big)\|D_n\|^2\Big] \\
&\quad = q2^{q-2}\gamma_n t\,\mathbb{E}\Big[\|D_n\|^2\|\Theta_{n-1} - \vartheta\|^{q-2} + (\gamma_n t)^{q-3}\|D_n\|^{q-1}\|\Theta_{n-1} - \vartheta\| \\
&\qquad + \|D_n\|^2\|\Theta_{n-1} - \vartheta\|^{q-2} + (\gamma_n t)^{q-2}\|D_n\|^q\Big] \\
&\quad \leq q2^{q-1}\gamma_n\,\mathbb{E}\Big[\|D_n\|^2\|\Theta_{n-1} - \vartheta\|^{q-2} + (\gamma_n)^{q-3}\|D_n\|^{q-1}\|\Theta_{n-1} - \vartheta\| + (\gamma_n)^{q-2}\|D_n\|^q\Big].
\end{aligned} \tag{208}$$

The tower property for conditional expectations and (190) hence imply that for all $n \in \mathbb{N} \cap (N, \infty)$, $t \in [0,1]$ it holds that

$$\begin{aligned}
&\mathbb{E}\Big[\big|V_q'(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + tD_n))(D_n) - V_q'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))(D_n)\big|\Big] \\
&\quad \leq q2^{q-1}\gamma_n\,\mathbb{E}\Big[\mathbb{E}\big[\|D_n\|^2 \,|\, \mathbb{F}_{n-1}\big]\|\Theta_{n-1} - \vartheta\|^{q-2} \\
&\qquad + (\gamma_n)^{q-3}\,\mathbb{E}\big[\|D_n\|^{q-1} \,|\, \mathbb{F}_{n-1}\big]\|\Theta_{n-1} - \vartheta\| + (\gamma_n)^{q-2}\,\mathbb{E}\big[\|D_n\|^q \,|\, \mathbb{F}_{n-1}\big]\Big] \\
&\quad \leq q2^{q-1}\gamma_n\,\mathbb{E}\Big[\kappa\big(1 + \|\Theta_{n-1} - \vartheta\|^2\big)\|\Theta_{n-1} - \vartheta\|^{q-2} \\
&\qquad + (\gamma_n)^{q-3}\kappa\big(1 + \|\Theta_{n-1} - \vartheta\|^{q-1}\big)\|\Theta_{n-1} - \vartheta\| + (\gamma_n)^{q-2}\kappa\big(1 + \|\Theta_{n-1} - \vartheta\|^q\big)\Big] \\
&\quad = q2^{q-1}\kappa\gamma_n\Big(\mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^{q-2}\big] + (\gamma_n)^{q-3}\,\mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|\big] + (\gamma_n)^{q-2} \\
&\qquad + \big(1 + (\gamma_n)^{q-3} + (\gamma_n)^{q-2}\big)\,\mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^q\big]\Big).
\end{aligned} \tag{209}$$

The fact that $\forall\, x \in [0,\infty)\colon x \leq 1 + x^q$ and the induction hypothesis (see (193)) therefore ensure that for all $n \in \mathbb{N} \cap (N, \infty)$, $t \in [0,1]$ it holds that

$$\begin{aligned}
&\mathbb{E}\Big[\big|V_q'(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + tD_n))(D_n) - V_q'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))(D_n)\big|\Big]\\
&\le q2^{q-1}\kappa\gamma_n\Big(C(\gamma_n)^{(q-2)/2} + (\gamma_n)^{q-3}\,\mathbb{E}\big[1 + \|\Theta_{n-1} - \vartheta\|^q\big] + (\gamma_n)^{q-2}\\
&\quad + \big(1 + (\gamma_n)^{q-3} + (\gamma_n)^{q-2}\big)\,\mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^q\big]\Big)\\
&= q2^{q-1}\kappa\gamma_n\Big(C(\gamma_n)^{(q-2)/2} + (\gamma_n)^{q-3} + (\gamma_n)^{q-2}\\
&\quad + \big(1 + 2(\gamma_n)^{q-3} + (\gamma_n)^{q-2}\big)\,\mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^q\big]\Big).
\end{aligned}\tag{210}$$

The fact that $\sup_{n\in\mathbb{N}\cap(N,\infty)} \gamma_n \le c \le 1$ and the fact that ${(q-2)}/{2} \le q - 3$ hence demonstrate that for all $n \in \mathbb{N} \cap (N, \infty)$, $t \in [0,1]$ it holds that

$$\begin{aligned}
&\mathbb{E}\Big[\big|V_q'(\Theta_{n-1} + \gamma_n(g(\Theta_{n-1}) + tD_n))(D_n) - V_q'(\Theta_{n-1} + \gamma_n g(\Theta_{n-1}))(D_n)\big|\Big]\\
&\le q2^{q-1}\kappa\gamma_n\big((C+2)(\gamma_n)^{(q-2)/2} + 4\,\mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^q\big]\big)\\
&= q2^{q-1}\kappa\big((C+2)(\gamma_n)^{q/2} + 4\gamma_n\,\mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^q\big]\big)\\
&\le q2^{q-1}\kappa\max\{C+2, 4\}\big((\gamma_n)^{q/2} + \mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^q\big]\big)\\
&\le q2^{q+1}\kappa\max\{C, 1\}\big((\gamma_n)^{q/2} + \mathbb{E}\big[\|\Theta_{n-1} - \vartheta\|^q\big]\big)\\
&= q2^{q+1}\kappa\max\{C, 1\}\big((\gamma_n)^{q/2} + \mathbb{E}\big[V_q(\Theta_{n-1})\big]\big).
\end{aligned}\tag{211}$$

Combining this, (176), (180), (181), (194), (199), (200), (203), and (205) with Corollary 3.3 (with $N = N$, $k = {q}/{2}$, $\kappa = q2^{q+1}\kappa\max\{1, C\}$, $c = c$, $\varrho = c$, $V = V_q$ in the notation of Corollary 3.3) yields that there exists $\mathfrak{C} \in (0, \infty)$ such that for all $n \in \mathbb{N}_0$ it holds that

$$\mathbb{E}\big[\|\Theta_n - \vartheta\|^q\big] = \mathbb{E}\big[V_q(\Theta_n)\big] \le \mathfrak{C}(\gamma_n)^{q/2}. \tag{212}$$

Induction thus proves (192). Next note that (192) demonstrates that for all $q \in \{2, 4, 6, \ldots\} \cap [0, p]$ there exists $C \in (0, \infty)$ such that for all $n \in \mathbb{N}_0$ it holds that

$$\big(\mathbb{E}\big[\|\Theta_n - \vartheta\|^q\big]\big)^{1/q} \le \big(C(\gamma_n)^{q/2}\big)^{1/q} = C^{1/q}\,(\gamma_n)^{1/2}. \tag{213}$$

This completes the proof of Proposition 3.6. □

Theorem 3.7. *Let $d \in \mathbb{N}$, $p \in \{2, 4, 6, \ldots\}$, $\kappa, c \in (0, \infty)$, $(\gamma_n)_{n\in\mathbb{N}} \subseteq (0, \infty)$, $\vartheta \in \mathbb{R}^d$, let $\langle\cdot,\cdot\rangle\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a scalar product, let $\|\cdot\|\colon \mathbb{R}^d \to [0, \infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that $\|\theta\| = \sqrt{\langle\theta, \theta\rangle}$, let $g\colon \mathbb{R}^d \to \mathbb{R}^d$ be $\mathcal{B}(\mathbb{R}^d)/\mathcal{B}(\mathbb{R}^d)$-measurable, let $(\Omega, \mathcal{F}, \mathbb{P}, (\mathbb{F}_n)_{n\in\mathbb{N}_0})$ be a filtered probability space, let $D\colon \mathbb{N}\times\Omega \to \mathbb{R}^d$ be* *an $(\mathbb{F}_n)_{n\in\mathbb{N}}/\mathcal{B}(\mathbb{R}^d)$-adapted stochastic process which satisfies for all $n \in \mathbb{N}$, $A \in \mathbb{F}_{n-1}$ with $\mathbb{E}[\|D_n\|] < \infty$ that $\mathbb{E}[D_n \mathbb{1}_A] = 0$, let $\Theta\colon \mathbb{N}_0 \times \Omega \to \mathbb{R}^d$ be a function, assume that Θ_0 is $\mathbb{F}_0/\mathcal{B}(\mathbb{R}^d)$-measurable, and assume for all $n \in \mathbb{N}$, $A \in \mathbb{F}_{n-1}$, $\theta \in \mathbb{R}^d$ that*

$$\langle \theta - \vartheta, g(\theta) \rangle \leq -c \max\bigl\{ \|\theta - \vartheta\|^2, \|g(\theta)\|^2 \bigr\}, \tag{214}$$

$$\limsup_{l\to\infty} \gamma_l = 0 < \min_{k\in\{1,2,\ldots,p/2\}} \Bigl(\liminf_{l\to\infty} \Bigl[\tfrac{(\gamma_l)^k - (\gamma_{l-1})^k}{(\gamma_l)^{k+1}} + \tfrac{c(\gamma_{l-1})^k}{2(\gamma_l)^k} \Bigr] \Bigr), \tag{215}$$

$$\Theta_n = \Theta_{n-1} + \gamma_n (g(\Theta_{n-1}) + D_n), \qquad \mathbb{E}\bigl[\|\Theta_0\|^p\bigr] < \infty, \tag{216}$$

$$\textit{and} \qquad \mathbb{E}\bigl[\|D_n\|^p \mathbb{1}_A\bigr] \leq \kappa\, \mathbb{E}\bigl[(1 + \|\Theta_{n-1}\|^p)\mathbb{1}_A\bigr]. \tag{217}$$

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

$$\{\theta \in \mathbb{R}^d \colon g(\theta) = 0\} = \{\vartheta\} \qquad \textit{and} \qquad \bigl(\mathbb{E}\bigl[\|\Theta_n - \vartheta\|^p\bigr]\bigr)^{1/p} \leq C(\gamma_n)^{1/2}. \tag{218}$$

Proof of Theorem 3.7. Observe that Lemma 2.1 and (217) ensure that for all $n \in \mathbb{N}$, $A \in \mathbb{F}_{n-1}$ it holds that

$$\begin{aligned}
\mathbb{E}\bigl[\|D_n\|^p \mathbb{1}_A\bigr] &\leq \kappa\, \mathbb{E}\bigl[(1 + \|\Theta_{n-1}\|^p)\mathbb{1}_A\bigr] \\
&= \kappa\, \mathbb{E}\bigl[(1 + \|\Theta_{n-1} - \vartheta + \vartheta\|^p)\mathbb{1}_A\bigr] \\
&\leq \kappa\, \mathbb{E}\bigl[(1 + 2^p\|\Theta_{n-1} - \vartheta\| + 2^p\|\vartheta\|^p)\mathbb{1}_A\bigr] \\
&\leq \kappa\, \mathbb{E}\bigl[(\max\{1 + 2^p\|\vartheta\|^p, 2^p\} + 2^p\|\Theta_{n-1} - \vartheta\|^p)\mathbb{1}_A\bigr] \\
&\leq \kappa\, \max\{1 + 2^p\|\vartheta\|^p, 2^p\}\, \mathbb{E}\bigl[(1 + \|\Theta_{n-1} - \vartheta\|^p)\mathbb{1}_A\bigr].
\end{aligned} \tag{219}$$

Combining item (i) in Lemma 2.12 and Proposition 3.6 hence establishes (218). The proof of Theorem 3.7 is thus completed. □

Remark 3.8 (A comment on assumption (217))**.** *Let $d \in \mathbb{N}$, $p \in \{2, 4, 6, \ldots\}$, $\kappa \in (0, \infty)$, let $D\colon \mathbb{N} \times \Omega \to \mathbb{R}^d$ be an $(\mathbb{F}_n)_{n\in\mathbb{N}}/\mathcal{B}(\mathbb{R}^d)$-adapted stochastic process, and let $\Theta\colon \mathbb{N}_0 \times \Omega \to \mathbb{R}^d$ be an $(\mathbb{F}_n)_{n\in\mathbb{N}_0}/\mathcal{B}(\mathbb{R}^d)$-adapted stochastic process. Then the following two statements are equivalent:*

(i) For all $n \in \mathbb{N}$, $A \in \mathbb{F}_{n-1}$ it holds that

$$\mathbb{E}\bigl[\|D_n\|^p \mathbb{1}_A\bigr] \leq \kappa\, \mathbb{E}\bigl[(1 + \|\Theta_{n-1}\|^p)\mathbb{1}_A\bigr]. \tag{220}$$

(ii) For all $n \in \mathbb{N}$ it holds $\mathbb{P}$-a.s. that

$$\mathbb{E}\bigl[\|D_n\|^p \,|\, \mathbb{F}_{n-1}\bigr] \leq \kappa\bigl(1 + \|\Theta_{n-1}\|^p\bigr). \tag{221}$$

4 Applications

In this section we present several consequences of Theorem 3.7 above. In particular, we prove in this section for every arbitrarily small $\varepsilon \in (0,\infty)$ and every arbitrarily large $p \in (0,\infty)$ that SGD optimization algorithms converge in the strong L^p-sense with order $1/2 - \varepsilon$ to the global minimum of the objective function of a suitable stochastic optimization problem.

4.1 Strong L^p-convergence rate for a specific type of SAAs

In order to apply Theorem 3.7 to a SGD optimization algorithm we need to verify that the sequence $\gamma_n \in (0,\infty)$, $n \in \mathbb{N}$, of learning rates of the considered SGD optimization algorithm satisfies the hypothesis in (215) in Theorem 3.7. For this we employ the next result, Lemma 4.1 below, which, in particular, provides explicit examples of sequences that satisfy the hypothesis in (215) in Theorem 3.7.

Lemma 4.1 (Example of suitable learning rates)**.** *Let $k, \alpha, c \in (0,\infty)$, $\nu \in \mathbb{R} \setminus \{1\}$, $(\gamma_n)_{n\in\mathbb{N}} \subseteq (0,\infty)$ satisfy for all $n \in \mathbb{N}$ that $\gamma_n = \alpha n^{-\nu}$. Then the following two statements are equivalent:*

(i) It holds that $\nu \in (0,1)$.

(ii) It holds that

$$\limsup_{n\to\infty} \gamma_n = 0 < \liminf_{n\to\infty} \left[\tfrac{(\gamma_n)^k - (\gamma_{n-1})^k}{(\gamma_n)^{k+1}} + \tfrac{c(\gamma_{n-1})^k}{(\gamma_n)^k} \right]. \tag{222}$$

Proof of Lemma 4.1. Throughout this proof let $(\Gamma_{n,r})_{n\in\mathbb{N}, r\in\mathbb{R}} \subseteq (0,\infty)$ satisfy for all $n \in \mathbb{N}$, $r \in \mathbb{R}$ that $\Gamma_{n,r} = \alpha n^{-r}$. Note that for all $r \in (0,\infty)$ it holds that

$$\limsup_{n\to\infty} \Gamma_{n,r} = \limsup_{n\to\infty} \left[\frac{\alpha}{n^r} \right] = 0. \tag{223}$$

Moreover, observe that Lemma 2.10 (with $\beta = (k+1)r$, $\delta = kr$ for $r \in (0,1)$ in the notation of Lemma 2.10) and the fact that for all $r \in (0,1)$ it holds that $(k+1)r = kr + r < kr + 1$ prove that for all $r \in (0,1)$ it holds that

$$\begin{aligned} \liminf_{n\to\infty} \left[\frac{(\Gamma_{n,r})^k - (\Gamma_{n-1,r})^k}{(\Gamma_{n,r})^{k+1}} \right] &= \liminf_{n\to\infty} \left[\frac{\left(\frac{\alpha}{n^r}\right)^k - \left(\frac{\alpha}{(n-1)^r}\right)^k}{\left(\frac{\alpha}{n^r}\right)^{k+1}} \right] \\ &= \frac{1}{\alpha} \liminf_{n\to\infty} \left[\frac{n^{-kr} - (n-1)^{-kr}}{n^{-(k+1)r}} \right] \\ &= 0. \end{aligned} \tag{224}$$

In addition, note that for all $r \in (0, \infty)$ it holds that

$$\begin{aligned}\liminf_{n\to\infty}\left[\frac{(\Gamma_{n-1,r})^k}{(\Gamma_{n,r})^k}\right] &= \liminf_{n\to\infty}\left[\frac{\left(\frac{\alpha}{(n-1)^r}\right)^k}{\left(\frac{\alpha}{n^r}\right)^k}\right] \\ &= \liminf_{n\to\infty}\left[\frac{n^{kr}}{(n-1)^{kr}}\right] \\ &= \liminf_{n\to\infty}\left[\frac{(n+1)^{kr}}{n^{kr}}\right] \\ &= \liminf_{n\to\infty}\left[(1+{}^{1}/{}_{n})^{kr}\right] = 1.\end{aligned} \tag{225}$$

Therefore, we obtain that for all $r \in (0, 1)$ it holds that

$$\begin{aligned}&\liminf_{n\to\infty}\left[\frac{(\Gamma_{n,r})^k - (\Gamma_{n-1,r})^k}{(\Gamma_{n,r})^{k+1}} + \frac{c(\Gamma_{n-1,r})^k}{(\Gamma_{n,r})^k}\right] \\ &\geq \liminf_{n\to\infty}\left[\frac{(\Gamma_{n,r})^k - (\Gamma_{n-1,r})^k}{(\Gamma_{n,r})^{k+1}}\right] + \liminf_{n\to\infty}\left[\frac{c(\Gamma_{n-1,r})^k}{(\Gamma_{n,r})^k}\right] = c > 0.\end{aligned} \tag{226}$$

Next observe that for all $r \in (-\infty, 0)$ it holds that

$$\limsup_{n\to\infty}\Gamma_{n,r} = \limsup_{n\to\infty}\left[\alpha n^{|r|}\right] = \infty. \tag{227}$$

Moreover, note that for all $r \in (1, \infty)$, $n \in \{2, 3, \dots\}$ it holds that

$$\begin{aligned}\frac{(\Gamma_{n,r})^k - (\Gamma_{n-1,r})^k}{(\Gamma_{n,r})^{k+1}} &= \frac{1}{\alpha}\left[\frac{n^{-kr} - (n-1)^{-kr}}{n^{-(k+1)r}}\right] \\ &= \frac{n^{(k+1)r}}{\alpha}\left[\frac{1}{n^{kr}} - \frac{1}{(n-1)^{kr}}\right] \\ &= -\frac{(kr)n^{(k+1)r}}{\alpha}\int_{n-1}^{n}\frac{1}{x^{kr+1}}\,\mathrm{d}x \\ &\leq -\frac{(kr)n^{(k+1)r}}{\alpha n^{kr+1}} \\ &= -\frac{(kr)n^{r-1}}{\alpha}.\end{aligned} \tag{228}$$

This assures that for all $r \in (1,\infty)$ it holds that

$$\begin{aligned}
&\liminf_{n\to\infty}\left[\frac{(\Gamma_{n,r})^k-(\Gamma_{n-1,r})^k}{(\Gamma_{n,r})^{k+1}}+\frac{c(\Gamma_{n-1,r})^k}{(\Gamma_{n,r})^k}\right]\\
&\leq \liminf_{n\to\infty}\left[-\frac{krn^{r-1}}{\alpha}+\frac{c\big(\frac{\alpha}{(n-1)^r}\big)^k}{\big(\frac{\alpha}{n^r}\big)^k}\right]\\
&= \liminf_{n\to\infty}\left[-\frac{krn^{r-1}}{\alpha}+\frac{cn^{rk}}{(n-1)^{rk}}\right]\\
&= \liminf_{n\to\infty}\left[-\frac{kr(n+1)^{r-1}}{\alpha}+\frac{c(n+1)^{rk}}{n^{rk}}\right]\\
&= \liminf_{n\to\infty}\left[-\frac{kr(n+1)^{r-1}}{\alpha}+c\left[1+{}^1\!/\!{}_n\right]^{rk}\right]\\
&= c+\liminf_{n\to\infty}\left[-\frac{krn^{r-1}}{\alpha}\right]=-\infty.
\end{aligned}\tag{229}$$

Combining this, (223), (226), and (227) completes the proof of Lemma 4.1. □

Lemma 4.2. *Let $(\Omega,\mathcal{F},\mu)$ be a sigma-finite measure space, let $(\mathbb{X},\mathcal{X})$ and $(\mathbb{Y},\mathcal{Y})$ be measurable spaces, let $Y:\Omega\to\mathbb{Y}$ be $\mathcal{F}/\mathcal{Y}$ measurable, let $d\in\mathbb{N}$, let $G:\mathbb{X}\times\mathbb{Y}\to\mathbb{R}^d$ be $(\mathcal{X}\otimes\mathcal{Y})/\mathcal{B}(\mathbb{R}^d)$-measurable, let $\|\cdot\|:\mathbb{R}^d\to[0,\infty)$ be a norm, and assume for all $x\in\mathbb{X}$ that $\int_\Omega\|G(x,Y(\omega))\|\,\mu(\mathrm{d}\omega)<\infty$. Then it holds that the function*

$$\mathbb{X}\ni x\mapsto\int_\Omega G(x,Y(\omega))\,\mu(\mathrm{d}\omega)\in\mathbb{R}^d \tag{230}$$

is $\mathcal{X}/\mathcal{B}(\mathbb{R}^d)$-measurable.

Proof. Throughout this proof assume w.l.o.g. that $\mathbb{X}\neq\emptyset$, let $G_i:\mathbb{X}\times\mathbb{Y}\to\mathbb{R}$, $i\in\{1,2,\dots,d\}$, be the functions which satisfy for all $x\in\mathbb{X}$, $y\in\mathbb{Y}$ that

$$G(x,y)=(G_1(x,y),G_2(x,y),\dots,G_d(x,y)), \tag{231}$$

let $c\in(0,\infty)$ satisfy

$$c=\sup_{\theta=(\theta_1,\dots,\theta_d)\in\mathbb{R}^d\setminus\{0\}}\left(\frac{\big(\sum_{i=1}^d|\theta_i|\big)}{\|\theta\|}\right), \tag{232}$$

let $v\in\mathbb{X}$, and let $\nu:\mathcal{X}\to[0,\infty]$ be the measure which satisfies for all $A\in\mathcal{X}$ that

$$\nu(A)=\begin{cases}1 & : v\in A\\ 0 & : v\in\mathbb{X}\setminus A.\end{cases} \tag{233}$$

Observe that (231), (232), and the hypothesis that for all $x \in \mathbb{X}$ it holds that $\int_\Omega \|G(x, Y(\omega))\| \,\mu(\mathrm{d}\omega) < \infty$ ensure that for all $i \in \{1, 2, \ldots, d\}$, $x \in \mathbb{X}$ it holds that

$$\begin{aligned}
&\int_\Omega \max\bigl\{G_i(x, Y(\omega)), 0\bigr\} \,\mu(\mathrm{d}\omega) + \int_\Omega \max\bigl\{-G_i(x, Y(\omega)), 0\bigr\} \,\mu(\mathrm{d}\omega) \\
&= \int_\Omega \bigl|G_i(x, Y(\omega))\bigr| \,\mu(\mathrm{d}\omega) \\
&\leq \int_\Omega \Biggl[\sum_{j=1}^{d} \bigl|G_j(x, Y(\omega))\bigr|\Biggr] \,\mu(\mathrm{d}\omega) \\
&\leq c \int_\Omega \|G(x, Y(\omega))\| \,\mu(\mathrm{d}\omega) < \infty.
\end{aligned} \tag{234}$$

Hence, we obtain that for all $i \in \{1, 2, \ldots, d\}$, $x \in \mathbb{X}$ it holds that

$$\begin{aligned}
&\int_\Omega G_i(x, Y(\omega)) \,\mu(\mathrm{d}\omega) \\
&= \int_\Omega \Bigl[\max\bigl\{G_i(x, Y(\omega)), 0\bigr\} + \min\bigl\{G_i(x, Y(\omega)), 0\bigr\}\Bigr] \,\mu(\mathrm{d}\omega) \\
&= \int_\Omega \Bigl[\max\bigl\{G_i(x, Y(\omega)), 0\bigr\} - \max\bigl\{-G_i(x, Y(\omega)), 0\bigr\}\Bigr] \,\mu(\mathrm{d}\omega) \\
&= \int_\Omega \max\bigl\{G_i(x, Y(\omega)), 0\bigr\} \,\mu(\mathrm{d}\omega) - \int_\Omega \max\bigl\{-G_i(x, Y(\omega)), 0\bigr\} \,\mu(\mathrm{d}\omega).
\end{aligned} \tag{235}$$

Next note that Fubini's theorem and the fact that the measure $(\nu \otimes \mu)\colon (\mathcal{X} \otimes \mathcal{F}) \to [0, \infty]$ is sigma-finite prove that for all $i \in \{1, 2, \ldots, d\}$ it holds that the functions

$$\mathbb{X} \ni x \mapsto \int_\Omega \max\bigl\{G_i(x, Y(\omega)), 0\bigr\} \,\mu(\mathrm{d}\omega) \in [0, \infty] \tag{236}$$

and

$$\mathbb{X} \ni x \mapsto \int_\Omega \max\bigl\{-G_i(x, Y(\omega)), 0\bigr\} \,\mu(\mathrm{d}\omega) \in [0, \infty] \tag{237}$$

are $\mathcal{X}/\mathcal{B}([0, \infty])$-measurable. Combining this with (234) demonstrates that for all $i \in \{1, 2, \ldots, d\}$ it holds that the functions

$$\mathbb{X} \ni x \mapsto \int_\Omega \max\bigl\{G_i(x, Y(\omega)), 0\bigr\} \,\mu(\mathrm{d}\omega) \in [0, \infty) \tag{238}$$

and

$$\mathbb{X} \ni x \mapsto \int_\Omega \max\bigl\{-G_i(x, Y(\omega)), 0\bigr\} \,\mu(\mathrm{d}\omega) \in [0, \infty) \tag{239}$$

are $\mathcal{X}/\mathcal{B}([0,\infty))$-measurable. This and (235) ensure that for all $i \in \{1,2,\dots,d\}$ it holds that the function $G_i \colon \mathbb{X} \to \mathbb{R}$ is $\mathcal{X}/\mathcal{B}(\mathbb{R})$-measurable. The proof of Lemma 4.2 is thus completed. □

Proposition 4.3. *Let $d \in \mathbb{N}$, $p \in \{2,4,6,\dots\}$, $\alpha, \kappa, c \in (0,\infty)$, $\nu \in (0,1)$, $\xi, \vartheta \in \mathbb{R}^d$, let $\langle \cdot,\cdot \rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a scalar product, let $\|\cdot\| \colon \mathbb{R}^d \to [0,\infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that $\|\theta\| = \sqrt{\langle \theta, \theta \rangle}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(S, \mathcal{S})$ be a measurable space, let $X_n \colon \Omega \to S$, $n \in \mathbb{N}$, be i.i.d. random variables, let $G \colon \mathbb{R}^d \times S \to \mathbb{R}^d$ be $(\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{S})/\mathcal{B}(\mathbb{R}^d)$-measurable, let $g \colon \mathbb{R}^d \to \mathbb{R}^d$ be a function, assume for all $\theta \in \mathbb{R}^d$ that*

$$\mathbb{E}\big[\|G(\theta, X_1) - g(\theta)\|^p\big] \leq \kappa\big(1 + \|\theta\|^p\big), \qquad g(\theta) = \mathbb{E}\big[G(\theta, X_1)\big], \tag{240}$$

$$\text{and} \qquad \langle \theta - \vartheta, g(\theta) \rangle \leq -c \max\big\{\|\theta - \vartheta\|^2, \|g(\theta)\|^2\big\}, \tag{241}$$

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

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} + \tfrac{\alpha}{n^\nu} G(\Theta_{n-1}, X_n). \tag{242}$$

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

$$\{\theta \in \mathbb{R}^d \colon g(\theta) = 0\} = \{\vartheta\} \qquad \text{and} \qquad \big(\mathbb{E}\big[\|\Theta_n - \vartheta\|^p\big]\big)^{1/p} \leq C n^{-\nu/2}. \tag{243}$$

Proof of Proposition 4.3. Throughout this proof let $(\gamma_n)_{n \in \mathbb{N}} \subseteq (0,\infty)$ satisfy for all $n \in \mathbb{N}$ that $\gamma_n = \alpha n^{-\nu}$, let $D \colon \mathbb{N} \times \Omega \to \mathbb{R}^d$ be the function which satisfies for all $n \in \mathbb{N}$ that

$$D_n = G(\Theta_{n-1}, X_n) - g(\Theta_{n-1}), \tag{244}$$

let $\mathfrak{G} \colon \mathbb{R}^d \to [0,\infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that

$$\mathfrak{G}(\theta) = \mathbb{E}\big[\|G(\theta, X_1) - g(\theta)\|^p\big], \tag{245}$$

and let $\mathbb{F}_n \subseteq \mathcal{F}$, $n \in \mathbb{N}_0$, be the sigma-algebras which satisfy for all $n \in \mathbb{N}$ that

$$\mathbb{F}_0 = \{\{\}, \Omega\} \qquad \text{and} \qquad \mathbb{F}_n = \sigma_\Omega(X_1, X_2, \dots, X_n). \tag{246}$$

Observe that (240), the hypothesis that the function $G \colon \mathbb{R}^d \times S \to \mathbb{R}^d$ is $(\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{S})/\mathcal{B}(\mathbb{R}^d)$-measurable, and Lemma 4.2 (with $\Omega = \Omega$, $\mathcal{F} = \mathcal{F}$, $\mu = \mathbb{P}$, $\mathbb{X} = \mathbb{R}^d$, $\mathcal{X} = \mathcal{B}(\mathbb{R}^d)$, $\mathbb{Y} = S$, $\mathcal{Y} = \mathcal{S}$, $Y = X_1$, $d = d$, $G = G$, $\|\cdot\| = \|\cdot\|$in the notation of Lemma 4.2) prove that the function $g \colon \mathbb{R}^d \to \mathbb{R}^d$ is $\mathcal{B}(\mathbb{R}^d)/\mathcal{B}(\mathbb{R}^d)$-measurable. Next note that the hypothesis that the function $G \colon \mathbb{R}^d \times S \to \mathbb{R}^d$ is $(\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{S})/\mathcal{B}(\mathbb{R}^d)$-measurable, the hypothesis that $\Theta_0 = \xi$, and (242) imply that Θ is an $(\mathbb{F}_n)_{n \in \mathbb{N}_0}/\mathcal{B}(\mathbb{R}^d)$-adapted stochastic process. Combining (244) and the fact that the function $g\colon \mathbb{R}^d \to \mathbb{R}^d$ is $\mathcal{B}(\mathbb{R}^d)/\mathcal{B}(\mathbb{R}^d)$-measurable with the hypothesis that the function $G\colon \mathbb{R}^d \times S \to \mathbb{R}^d$ is $(\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{S})/\mathcal{B}(\mathbb{R}^d)$-measurable hence demonstrates that D is an $(\mathbb{F}_n)_{n\in\mathbb{N}}/\mathcal{B}(\mathbb{R}^d)$-adapted stochastic process. Furthermore, note that (242) proves that

$$\mathbb{E}\big[\|\Theta_0\|^p\big] = \|\xi\|^p < \infty. \tag{247}$$

In addition, observe that (244) and (242) ensure that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}\Theta_n &= \Theta_{n-1} + \tfrac{\alpha}{n^\nu} G(\Theta_{n-1}, X_n)\\ &= \Theta_{n-1} + \tfrac{\alpha}{n^\nu}\big(g(\Theta_{n-1}) + [G(\Theta_{n-1}, X_n) - g(\Theta_{n-1})]\big)\\ &= \Theta_{n-1} + \tfrac{\alpha}{n^\nu}\big(g(\Theta_{n-1}) + D_n\big)\\ &= \Theta_{n-1} + \gamma_n\big(g(\Theta_{n-1}) + D_n\big).\end{aligned} \tag{248}$$

Next observe that (240), item (ii) in Lemma 2.8, the fact that for all $n \in \mathbb{N}$ it holds that the function Θ_{n-1} is $\mathbb{F}_{n-1}/\mathcal{B}(\mathbb{R}^d)$-measurable, and the fact that for all $n \in \mathbb{N}$ it holds that X_n is independent of $\mathbb{F}_{n-1}$ ensure that for all $n \in \mathbb{N}$, $A \in \mathbb{F}_{n-1}$ it holds that

$$\begin{aligned}\mathbb{E}\big[\|D_n\|^p \mathbb{1}_A\big] &= \mathbb{E}\big[\|G(\Theta_{n-1}, X_n) - g(\Theta_{n-1})\|^p \mathbb{1}_A\big]\\ &= \mathbb{E}\Big[\mathbb{E}\big[\|G(\Theta_{n-1}, X_n) - g(\Theta_{n-1})\|^p \mathbb{1}_A \,|\, \mathbb{F}_{n-1}\big]\Big]\\ &= \mathbb{E}\Big[\mathbb{E}\big[\|G(\Theta_{n-1}, X_n) - g(\Theta_{n-1})\|^p \,|\, \mathbb{F}_{n-1}\big]\mathbb{1}_A\Big]\\ &= \mathbb{E}\big[\mathfrak{G}(\Theta_{n-1})\mathbb{1}_A\big]\\ &= \kappa\,\mathbb{E}\big[(1 + \|\Theta_{n-1}\|^p)\mathbb{1}_A\big].\end{aligned} \tag{249}$$

Moreover, note that Corollary 2.9, the fact that for all $n \in \mathbb{N}$ it holds that the function Θ_{n-1} is $\mathbb{F}_{n-1}/\mathcal{B}(\mathbb{R}^d)$-measurable, the fact that for all $\theta \in \mathbb{R}^d$ it holds that $\mathbb{E}\big[\|G(\theta, X_1) - g(\theta)\|\big] < \infty$, and the fact that for all $n \in \mathbb{N}$ it holds that X_n is independent of $\mathbb{F}_{n-1}$ prove that for all $n \in \mathbb{N}$, $A \in \mathbb{F}_{n-1}$ with $\mathbb{E}[\|D_n\|] < \infty$ it holds that

$$\begin{aligned}\mathbb{E}\big[D_n \mathbb{1}_A\big] &= \mathbb{E}\Big[\big(G(\Theta_{n-1}, X_n) - g(\Theta_{n-1})\big)\mathbb{1}_A\Big]\\ &= \mathbb{E}\Big[\mathbb{E}\big[\big(G(\Theta_{n-1}, X_n) - g(\Theta_{n-1})\big)\mathbb{1}_A \,|\, \mathbb{F}_{n-1}\big]\Big]\\ &= \mathbb{E}\Big[\big(\mathbb{E}\big[G(\Theta_{n-1}, X_n) \,|\, \mathbb{F}_{n-1}\big] - g(\Theta_{n-1})\big)\mathbb{1}_A\Big]\\ &= \mathbb{E}\Big[\big(g(\Theta_{n-1}) - g(\Theta_{n-1})\big)\mathbb{1}_A\Big] = 0.\end{aligned} \tag{250}$$

Furthermore, observe that Lemma 4.1 ensures that for all $k \in (0,\infty)$ it holds that

$$\limsup_{n\to\infty} \gamma_n = 0 < \liminf_{n\to\infty} \left[\frac{(\gamma_n)^k - (\gamma_{n-1})^k}{(\gamma_n)^{k+1}} + \frac{c(\gamma_{n-1})^k}{2(\gamma_n)^k} \right]. \tag{251}$$

This implies that

$$\limsup_{l\to\infty} \gamma_l = 0 < \min_{k\in\{1,2,\dots,p/2\}} \left(\liminf_{l\to\infty} \left[\frac{(\gamma_l)^k - (\gamma_{l-1})^k}{(\gamma_l)^{k+1}} + \frac{c(\gamma_{l-1})^k}{2(\gamma_l)^k} \right] \right). \tag{252}$$

Combining the fact that D is an $(\mathbb{F}_n)_{n\in\mathbb{N}}/\mathcal{B}(\mathbb{R}^d)$-adapted stochastic process, the fact that the function Θ_0 is $\mathbb{F}_0/\mathcal{B}(\mathbb{R}^d)$-measurable, (241), and (247)–(250) with Theorem 3.7 hence demonstrates that there exists $C \in (0,\infty)$ such that for all $n \in \mathbb{N}$ it holds that

$$\{\theta \in \mathbb{R}^d \colon g(\theta) = 0\} = \{\vartheta\} \tag{253}$$

and

$$\big(\mathbb{E}\big[\|\Theta_n - \vartheta\|^p\big]\big)^{1/p} \le C(\gamma_n)^{1/2} = [C\sqrt{\alpha}]n^{-\nu/2}. \tag{254}$$

This establishes (243). The proof of Proposition 4.3 is thus completed. □

4.2 Strong L^p-convergence rate for stochastic gradient descent

Lemma 4.4. *Let $d \in \mathbb{N}$, let $(S,\mathcal{S})$ be a measurable space, let $F = (F(\theta,x))_{\theta\in\mathbb{R}^d, x\in S} \colon \mathbb{R}^d \times S \to \mathbb{R}$ be $(\mathcal{B}(\mathbb{R}^d)\otimes\mathcal{S})/\mathcal{B}(\mathbb{R})$-measurable, and assume for all $x \in S$ that*

$$(\mathbb{R}^d \ni \theta \mapsto F(\theta,x) \in \mathbb{R}) \in C^1(\mathbb{R}^d,\mathbb{R}). \tag{255}$$

Then it holds that the function

$$\mathbb{R}^d \times S \ni (\theta,x) \mapsto (\nabla_\theta F)(\theta,x) \in \mathbb{R}^d \tag{256}$$

is $(\mathcal{B}(\mathbb{R}^d)\otimes\mathcal{S})/\mathcal{B}(\mathbb{R}^d)$-measurable.

Proof of Lemma 4.4. Throughout this proof let $G = (G_1,\dots,G_d) \colon \mathbb{R}^d \times S \to \mathbb{R}^d$ be the function which satisfies for all $\theta \in \mathbb{R}^d$, $x \in S$ that

$$G(\theta,x) = (\nabla_\theta F)(\theta,x). \tag{257}$$

The hypothesis that the function $F \colon \mathbb{R}^d \times S \to \mathbb{R}$ is $(\mathcal{B}(\mathbb{R}^d)\otimes\mathcal{S})/\mathcal{B}(\mathbb{R})$-measurable implies that for all $i \in \{1,\dots,d\}$, $h \in \mathbb{R}\backslash\{0\}$ it holds that the function

$$\mathbb{R}^d \times S \ni (\theta,x) = ((\theta_1,\dots,\theta_d),x) \mapsto \left(\tfrac{F((\theta_1,\dots,\theta_{i-1},\theta_i+h,\theta_{i+1},\dots,\theta_d),x)-F(\theta,x)}{h}\right) \in \mathbb{R} \tag{258}$$

is $(\mathcal{B}(\mathbb{R}^d)\otimes\mathcal{S})/\mathcal{B}(\mathbb{R})$-measurable. The fact that for all $i \in \{1,\dots,d\}$, $\theta = (\theta_1,\dots,\theta_d) \in \mathbb{R}^d$, $x \in S$ it holds that

$$G_i(\theta, x) = \lim_{n\to\infty} \left(\tfrac{F((\theta_1,\dots,\theta_{i-1},\theta_i+2^{-n},\theta_{i+1},\dots,\theta_d),x)-F(\theta,x)}{2^{-n}} \right) \tag{259}$$

hence ensures that for all $i \in \{1,\dots,d\}$ it holds that the function $G_i \colon \mathbb{R}^d \times S \to \mathbb{R}$ is $(\mathcal{B}(\mathbb{R}^d)\otimes\mathcal{S})/\mathcal{B}(\mathbb{R})$-measurable. This and (257) complete the proof of Lemma 4.4. □

Corollary 4.5. *Let $d \in \mathbb{N}$, $p \in \{2,4,6,\dots\}$, $\alpha,\kappa,c \in (0,\infty)$, $\nu \in (0,1)$, $\xi,\vartheta \in \mathbb{R}^d$, let $\langle\cdot,\cdot\rangle \colon \mathbb{R}^d\times\mathbb{R}^d \to \mathbb{R}$ be a scalar product, let $\|\cdot\| \colon \mathbb{R}^d \to [0,\infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that $\|\theta\| = \sqrt{\langle\theta,\theta\rangle}$, let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, let $(S,\mathcal{S})$ be a measurable space, let $X_n \colon \Omega \to S$, $n \in \mathbb{N}$, be i.i.d. random variables, let $F = (F(\theta,x))_{\theta\in\mathbb{R}^d, x\in S} \colon \mathbb{R}^d\times S \to \mathbb{R}$ be $(\mathcal{B}(\mathbb{R}^d)\otimes\mathcal{S})/\mathcal{B}(\mathbb{R})$-measurable, let $g \colon \mathbb{R}^d \to \mathbb{R}^d$ be a function, assume for all $x \in S$ that $(\mathbb{R}^d \ni \theta \mapsto F(\theta,x) \in \mathbb{R}) \in C^1(\mathbb{R}^d,\mathbb{R})$, assume for all $\theta \in \mathbb{R}^d$ that*

$$\mathbb{E}\big[\|(\nabla_\theta F)(\theta,X_1) - g(\theta)\|^p\big] \le \kappa\big(1+\|\theta\|^p\big), \tag{260}$$

$$\langle\theta-\vartheta, g(\theta)\rangle \le -c\max\big\{\|\theta-\vartheta\|^2, \|g(\theta)\|^2\big\}, \tag{261}$$

and $g(\theta) = \mathbb{E}\big[(\nabla_\theta F)(\theta,X_1)\big]$, 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

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} + \tfrac{\alpha}{n^\nu}(\nabla_\theta F)(\Theta_{n-1},X_n). \tag{262}$$

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

$$\{\theta\in\mathbb{R}^d \colon g(\theta)=0\} = \{\vartheta\} \qquad \text{and} \qquad \big(\mathbb{E}\big[\|\Theta_n-\vartheta\|^p\big]\big)^{1/p} \le Cn^{-\nu/2}. \tag{263}$$

Proof of Corollary 4.5. Combining Lemma 4.4 and Proposition 4.3 (with $G(\theta,x) = \nabla_\theta F(\theta,x)$, $g(\theta) = g(\theta)$ for $\theta \in \mathbb{R}^d$, $x \in S$ in the notation of Proposition 4.3) establishes (263). The proof of Corollary 4.5 is thus completed. □

Corollary 4.6. *Let $d \in \mathbb{N}$, $p \in \{2,4,6,\dots\}$, $\alpha,\kappa,c \in (0,\infty)$, $\nu \in (0,1)$, $\xi,\vartheta \in \mathbb{R}^d$, let $\langle\cdot,\cdot\rangle \colon \mathbb{R}^d\times\mathbb{R}^d \to \mathbb{R}$ be a scalar product, let $\|\cdot\| \colon \mathbb{R}^d \to [0,\infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that $\|\theta\| = \sqrt{\langle\theta,\theta\rangle}$, let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, let $(S,\mathcal{S})$ be a measurable space, let $X_n \colon \Omega \to S$, $n \in \mathbb{N}$, be i.i.d. random variables, and let $F = (F(\theta,x))_{\theta\in\mathbb{R}^d, x\in S} \colon \mathbb{R}^d\times S \to \mathbb{R}$ be $(\mathcal{B}(\mathbb{R}^d)\otimes\mathcal{S})/\mathcal{B}(\mathbb{R})$-measurable, assume for all $x \in S$ that $(\mathbb{R}^d \ni \theta \mapsto F(\theta,x) \in \mathbb{R}) \in C^1(\mathbb{R}^d,\mathbb{R})$, assume for all $\theta \in \mathbb{R}^d$ that*

$$\mathbb{E}\big[\|(\nabla_\theta F)(\theta,X_1)\|\big] < \infty, \tag{264}$$

$$\langle \theta - \vartheta, \mathbb{E}[(\nabla_\theta F)(\theta, X_1)] \rangle \leq -c \max\bigl\{ \|\theta - \vartheta\|^2, \|\mathbb{E}[(\nabla_\theta F)(\theta, X_1)]\|^2 \bigr\}, \tag{265}$$

$$\mathbb{E}\bigl[\|(\nabla_\theta F)(\theta, X_1) - \mathbb{E}[(\nabla_\theta F)(\theta, X_1)]\|^p \bigr] \leq \kappa\bigl(1 + \|\theta\|^p\bigr), \tag{266}$$

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*

$$\Theta_0 = \xi \qquad \textit{and} \qquad \Theta_n = \Theta_{n-1} + \tfrac{\alpha}{n^\nu} (\nabla_\theta F)(\Theta_{n-1}, X_n). \tag{267}$$

Then

(i) it holds that $\bigl\{ \theta \in \mathbb{R}^d \colon \mathbb{E}[(\nabla_\theta F)(\theta, X_1)] = 0 \bigr\} = \{\vartheta\}$ *and*

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

$$\bigl(\mathbb{E}\bigl[\|\Theta_n - \vartheta\|^p \bigr] \bigr)^{1/p} \leq C n^{-\nu/2}. \tag{268}$$

Proof of Corollary 4.6. Corollary 4.5 (with $g(\theta) = \mathbb{E}[(\nabla_\theta F)(\theta, X_1)]$ for $\theta \in \mathbb{R}^d$ in the notation of Corollary 4.5) establishes (268). The proof of Corollary 4.6 is thus completed. □

Lemma 4.7. *Let* $d \in \mathbb{N}$, $p \in [1, \infty)$, *let* $\langle \cdot, \cdot \rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *be a scalar product, let* $\|\cdot\| \colon \mathbb{R}^d \to [0, \infty)$ *be the function which satisfies for all* $\theta \in \mathbb{R}^d$ *that* $\|\theta\| = \sqrt{\langle \theta, \theta \rangle}$, *let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space, let* $\mathbb{I}$ *be a non-empty set, let* $X_i \colon \Omega \to \mathbb{R}^d$, $i \in \mathbb{I}$, *be random variables, assume for all* $i \in \mathbb{I}$ *that* $\mathbb{E}[\|X_i\|] < \infty$, *and assume that* $\sup_{i \in \mathbb{I}} \mathbb{E}[\|X_i - \mathbb{E}[X_i]\|^p] < \infty$ *and* $\mathbb{P}(\sup_{i \in \mathbb{I}} \|X_i\| < \infty) > 0$. *Then it holds that*

$$\sup\nolimits_{i \in \mathbb{I}} \mathbb{E}\bigl[\|X_i\|^p \bigr] < \infty. \tag{269}$$

Proof of Lemma 4.7. Throughout this proof let $Y_i \colon \Omega \to [0, \infty)$, $i \in \mathbb{I}$, be the random variables which satisfy for all $i \in \mathbb{I}$ that

$$Y_i = \|X_i - \mathbb{E}[X_i]\|^p \tag{270}$$

and let $j = (j_k)_{k \in \mathbb{N}} \colon \mathbb{N} \to \mathbb{I}$ be a function which satisfies for all $k \in \mathbb{N}$ that

$$\mathbb{E}\bigl[\|X_{j_k}\|^p \bigr] \leq \mathbb{E}\bigl[\|X_{j_{k+1}}\|^p \bigr] \qquad \text{and} \qquad \limsup_{l \to \infty} \mathbb{E}\bigl[\|X_{j_l}\|^p \bigr] = \sup_{i \in \mathbb{I}} \mathbb{E}\bigl[\|X_i\|^p \bigr]. \tag{271}$$

Observe that the hypothesis that $\sup_{i \in \mathbb{I}} \mathbb{E}[\|X_i - \mathbb{E}[X_i]\|^p] < \infty$ and Fatou's lemma imply that for all functions $(n_k)_{k \in \mathbb{N}} \colon \mathbb{N} \to \mathbb{I}$ it holds that

$$\begin{aligned} \mathbb{E}\bigl[\liminf\nolimits_{k \to \infty} Y_{n_k} \bigr] &\leq \liminf_{k \to \infty} \mathbb{E}\bigl[Y_{n_k} \bigr] \\ &= \liminf_{k \to \infty} \mathbb{E}\bigl[\|X_{n_k} - \mathbb{E}[X_{n_k}]\|^p \bigr] \\ &\leq \sup\nolimits_{i \in \mathbb{I}} \mathbb{E}[\|X_i - \mathbb{E}[X_i]\|^p] < \infty. \end{aligned} \tag{272}$$

Moreover, note that the triangle inequality assures that for all functions $(n_k)_{k\in\mathbb{N}}\colon \mathbb{N}\to\mathbb{I}$ it holds that

$$\begin{aligned}\mathbb{E}\big[\liminf_{k\to\infty} Y_{n_k}\big] &= \mathbb{E}\big[\liminf_{k\to\infty} \|\mathbb{E}[X_{n_k}] - X_{n_k}\|^p\big]\\ &\geq \mathbb{E}\Big[\liminf_{k\to\infty} \big|\|\mathbb{E}[X_{n_k}]\| - \|X_{n_k}\|\big|^p\Big].\end{aligned}\tag{273}$$

The fact that $\forall\, x\in[0,\infty)\colon x^p \geq \max\{x,0\}-1$ therefore implies that for all functions $(n_k)_{k\in\mathbb{N}}\colon \mathbb{N}\to\mathbb{I}$ it holds that

$$\begin{aligned}&\mathbb{E}\big[\liminf_{k\to\infty} Y_{n_k}\big]\\ &\geq \mathbb{E}\Big[\liminf_{k\to\infty} \max\big\{\big|\|\mathbb{E}[X_{n_k}]\| - \|X_{n_k}\|\big|, 0\big\}\Big] - 1\\ &\geq \mathbb{E}\Big[\liminf_{k\to\infty} \max\big\{\|\mathbb{E}[X_{n_k}]\| - \|X_{n_k}\|, 0\big\}\Big] - 1\\ &\geq \mathbb{E}\Big[\liminf_{k\to\infty} \max\big\{\|\mathbb{E}[X_{n_k}]\| - \sup_{i\in\mathbb{I}}\|X_i\|, 0\big\}\Big] - 1.\end{aligned}\tag{274}$$

Combining this with (272) proves that for all functions $(n_k)_{k\in\mathbb{N}}\colon \mathbb{N}\to\mathbb{I}$ it holds that

$$\mathbb{E}\Big[\liminf_{k\to\infty} \max\big\{\|\mathbb{E}[X_{n_k}]\| - \sup_{i\in\mathbb{I}}\|X_i\|, 0\big\}\Big] < \infty.\tag{275}$$

Hence, we obtain that for all functions $(n_k)_{k\in\mathbb{N}}\colon \mathbb{N}\to\mathbb{I}$ it holds that

$$\mathbb{P}\Big(\liminf_{k\to\infty} \max\big\{\|\mathbb{E}[X_{n_k}]\| - \sup_{i\in\mathbb{I}}\|X_i\|, 0\big\} < \infty\Big) = 1.\tag{276}$$

Therefore, we obtain that for all functions $(n_k)_{k\in\mathbb{N}}\colon \mathbb{N}\to\mathbb{I}$ it holds that

$$\mathbb{P}\Big(\liminf_{k\to\infty} \big[\|\mathbb{E}[X_{n_k}]\| - \sup_{i\in\mathbb{I}}\|X_i\|\big] < \infty\Big) = 1.\tag{277}$$

Next note that for all $A, B\in\mathcal{F}$ with $\mathbb{P}(A)=1$ and $\mathbb{P}(B)>0$ it holds that

$$\mathbb{P}(\Omega\setminus(A\cap B)) = \mathbb{P}([\Omega\setminus A]\cup[\Omega\setminus B]) \leq \mathbb{P}(\Omega\setminus A) + \mathbb{P}(\Omega\setminus B) = \mathbb{P}(\Omega\setminus B) < 1.\tag{278}$$

Hence, we obtain that for all $A, B\in\mathcal{F}$ with $\mathbb{P}(A)=1$ and $\mathbb{P}(B)>0$ it holds that $\mathbb{P}(A\cap B)>0$. Combining this and the hypothesis that $\mathbb{P}(\sup_{i\in\mathbb{I}}\|X_i\|<\infty)>0$ with (277) proves that for all functions $(n_k)_{k\in\mathbb{N}}\colon \mathbb{N}\to\mathbb{I}$ it holds that

$$\mathbb{P}\Big(\big\{\liminf_{k\to\infty} \big[\|\mathbb{E}[X_{n_k}]\| - \sup_{i\in\mathbb{I}}\|X_i\|\big] < \infty\big\} \cap \big\{\sup_{i\in\mathbb{I}}\|X_i\| < \infty\big\}\Big) > 0.\tag{279}$$

Therefore, we obtain that for all functions $(n_k)_{k\in\mathbb{N}}\colon \mathbb{N}\to\mathbb{I}$ it holds that

$$\begin{aligned}&\big\{\liminf_{k\to\infty}\big[\|\mathbb{E}[X_{n_k}]\|-\sup_{i\in\mathbb{I}}\|X_i\|\big]<\infty\big\}\cap\big\{\sup_{i\in\mathbb{I}}\|X_i\|<\infty\big\}\\&=\Big\{\omega\in\Omega\colon\big(\liminf_{k\to\infty}\big[\|\mathbb{E}[X_{n_k}]\|-\sup_{i\in\mathbb{I}}\|X_i(\omega)\|\big]<\infty,\ \sup_{i\in\mathbb{I}}\|X_i(\omega)\|<\infty\big)\Big\}\\&\neq\emptyset.\end{aligned}\tag{280}$$

Moreover, note that for all functions $(n_k)_{k\in\mathbb{N}}\colon \mathbb{N}\to\mathbb{I}$ and all $\omega\in(\{\liminf_{k\to\infty}[\|\mathbb{E}[X_{n_k}]\|-\sup_{i\in\mathbb{I}}\|X_i\|]<\infty\}\cap\{\sup_{i\in\mathbb{I}}\|X_i\|<\infty\})$ it holds that

$$\liminf_{k\to\infty}\|\mathbb{E}[X_{n_k}]\|<\infty.\tag{281}$$

Combining this with (280) proves that for all functions $(n_k)_{k\in\mathbb{N}}\colon \mathbb{N}\to\mathbb{I}$ it holds that

$$\liminf_{k\to\infty}\|\mathbb{E}[X_{n_k}]\|<\infty.\tag{282}$$

Moreover, observe that Lemma 2.1 ensures that for all functions $(n_k)_{k\in\mathbb{N}}\colon \mathbb{N}\to\mathbb{I}$ it holds that

$$\begin{aligned}\liminf_{k\to\infty}\mathbb{E}\big[\|X_{n_k}\|^p\big]&=\liminf_{k\to\infty}\mathbb{E}\big[\|X_{n_k}-\mathbb{E}[X_{n_k}]+\mathbb{E}[X_{n_k}]\|^p\big]\\&\le 2^p\liminf_{k\to\infty}\Big(\mathbb{E}\big[\|X_{n_k}-\mathbb{E}[X_{n_k}]\|^p\big]+\|\mathbb{E}[X_{n_k}]\|^p\Big)\\&\le 2^p\big(\sup_{i\in\mathbb{I}}\mathbb{E}[\|X_i-\mathbb{E}[X_i]\|^p]\big)+2^p\big[\liminf_{k\to\infty}\|\mathbb{E}[X_{n_k}]\|\big]^p.\end{aligned}\tag{283}$$

The hypothesis that $\sup_{i\in\mathbb{I}}\mathbb{E}[\|X_i-\mathbb{E}[X_i]\|^p]<\infty$ and (282) hence imply that for all functions $(n_k)_{k\in\mathbb{N}}\colon \mathbb{N}\to\mathbb{I}$ it holds that

$$\liminf_{k\to\infty}\mathbb{E}\big[\|X_{n_k}\|^p\big]<\infty.\tag{284}$$

This and (271) prove that

$$\sup_{i\in\mathbb{I}}\mathbb{E}\big[\|X_i\|^p\big]=\limsup_{k\to\infty}\mathbb{E}\big[\|X_{j_k}\|^p\big]=\liminf_{k\to\infty}\mathbb{E}\big[\|X_{j_k}\|^p\big]<\infty.\tag{285}$$

The proof of Lemma 4.7 is thus completed. □

Lemma 4.8. *Let $d\in\mathbb{N}$, $p\in(1,\infty)$, $\kappa\in(0,\infty)$, $\vartheta\in\mathbb{R}^d$, let $\langle\cdot,\cdot\rangle\colon\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ be a scalar product, let $\|\cdot\|\colon\mathbb{R}^d\to[0,\infty)$ be the function which satisfies for all $\theta\in\mathbb{R}^d$ that $\|\theta\|=\sqrt{\langle\theta,\theta\rangle}$, let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, let $(S,\mathcal{S})$ be a measurable* *space, let $X\colon \Omega \to S$ be a random variable, let $F = (F(\theta, x))_{\theta\in\mathbb{R}^d, x\in S}\colon \mathbb{R}^d \times S \to \mathbb{R}$ be $(\mathcal{B}(\mathbb{R}^d)\otimes\mathcal{S})/\mathcal{B}(\mathbb{R})$-measurable, let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a function, assume for all $x \in S$ that $(\mathbb{R}^d \ni \theta \mapsto F(\theta, x) \in \mathbb{R}) \in C^1(\mathbb{R}^d, \mathbb{R})$, and assume for all $\theta \in \mathbb{R}^d$ that*

$$\mathbb{E}\big[|F(\theta, X)| + \|(\nabla_\theta F)(\theta, X)\|\big] < \infty, \tag{286}$$

$$\mathbb{E}\big[\|(\nabla_\theta F)(\theta, X) - \mathbb{E}[(\nabla_\theta F)(\theta, X)]\|^p\big] \le \kappa\big(1 + \|\theta\|^p\big), \tag{287}$$

and $f(\theta) = \mathbb{E}[F(\theta, X)]$. Then

(i) it holds that $f \in C^1(\mathbb{R}^d, \mathbb{R})$ and

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

$$(\nabla f)(\theta) = \mathbb{E}\big[(\nabla_\theta F)(\theta, X)\big]. \tag{288}$$

Proof of Lemma 4.8. Throughout this proof let $c \in (0, \infty)$ satisfy

$$c = \sup_{\theta=(\theta_1,\dots,\theta_d)\in\mathbb{R}^d\backslash\{0\}} \left(\frac{\big(\sum_{j=1}^d |\theta_j|\big)}{\|\theta\|}\right), \tag{289}$$

let $q = p - 1 \in (0, \infty)$, and let $e_1 = (1, 0, \dots, 0)$, $e_2 = (0, 1, 0, \dots, 0)$, $\dots$, $e_d = (0, \dots, 0, 1) \in \mathbb{R}^d$. Observe that (286), (287), and Lemma 2.1 imply that for all $\theta \in \mathbb{R}^d$, $v \in [-q, q]^d$ it holds that

$$\begin{aligned} &\mathbb{E}\big[\|(\nabla_\theta F)(\theta + v, X)\|^p\big] \\ &\le 2^p\Big(\mathbb{E}\big[\|(\nabla_\theta F)(\theta + v, X) - \mathbb{E}[(\nabla_\theta F)(\theta + v, X)]\|^p\big] + \|\mathbb{E}[(\nabla_\theta F)(\theta + v, X)]\|^p\Big) \\ &\le 2^p\Big(\kappa\big(1 + \|\theta + v\|^p\big) + |\mathbb{E}[\|(\nabla_\theta F)(\theta + v, X)\|]|^p\Big) < \infty. \end{aligned} \tag{290}$$

Moreover, note that (287) assures that for all $\theta \in \mathbb{R}^d$ it holds that

$$\begin{aligned} &\sup_{v\in[-q,q]^d} \Big(\mathbb{E}\big[\|(\nabla_\theta F)(\theta + v, X) - \mathbb{E}[(\nabla_\theta F)(\theta + v, X)]\|^p\big]\Big) \\ &\le \sup_{v\in[-q,q]^d} \Big(\kappa\big(1 + \|\theta + v\|^p\big)\Big) < \infty. \end{aligned} \tag{291}$$

Furthermore, observe that the hypothesis that for all $x \in S$ it holds that $(\mathbb{R}^d \ni \theta \mapsto F(\theta, x) \in \mathbb{R}) \in C^1(\mathbb{R}^d, \mathbb{R})$ ensures that for all $\theta \in \mathbb{R}^d$ it holds that

$$\mathbb{P}\bigg(\sup_{v\in[-q,q]^d} \|(\nabla_\theta F)(\theta + v, X)\| < \infty\bigg) = 1. \tag{292}$$

Lemma 4.7 (with $\mathbb{I} = [-q,q]^d$, $X_i = (\nabla_\theta F)(\theta + i, X)$ for $i \in [-q,q]^d$, $\theta \in \mathbb{R}^d$ in the notation of Lemma 4.7), (286), and (291) therefore imply that for all $\theta \in \mathbb{R}^d$ it holds that

$$\sup_{v \in [-q,q]^d} \mathbb{E}\big[\|(\nabla_\theta F)(\theta + v, X)\|^p\big] < \infty. \tag{293}$$

This and (289) demonstrate that for all $i \in \{1, 2, \ldots, d\}$, $\theta \in \mathbb{R}^d$ it holds that

$$\begin{aligned}\sup_{v \in [-q,q]^d} \mathbb{E}\big[|(\tfrac{\partial}{\partial \theta_i} F)(\theta + v, X)|^p\big] &\leq \sup_{v \in [-q,q]^d} \mathbb{E}\Big[\big(\textstyle\sum_{j=1}^d |(\tfrac{\partial}{\partial \theta_j} F)(\theta + v, X)|\big)^p\Big] \\ &\leq c^p \Bigg[\sup_{v \in [-q,q]^d} \mathbb{E}\big[\|(\nabla_\theta F)(\theta + v, X)\|^p\big]\Bigg] < \infty.\end{aligned} \tag{294}$$

Next observe that the hypothesis that for all $x \in S$ it holds that $(\mathbb{R}^d \ni \theta \mapsto F(\theta, x) \in \mathbb{R}) \in C^1(\mathbb{R}^d, \mathbb{R})$ and the fundamental theorem of calculus ensure that for all $i \in \{1, 2, \ldots, d\}$, $\theta \in \mathbb{R}^d$, $h \in \mathbb{R}$ it holds that

$$\begin{aligned} f(\theta + he_i) - f(\theta) &= \mathbb{E}[F(\theta + he_i, X) - F(\theta, X)] \\ &= \mathbb{E}\Big[\big[F(\theta + ue_i, X)\big]_{u=0}^{u=h}\Big] \\ &= \mathbb{E}\bigg[\int_0^h (\tfrac{\partial}{\partial \theta} F)(\theta + ue_i, X) e_i \, \mathrm{d}u\bigg] \\ &= \mathbb{E}\bigg[\int_0^h (\tfrac{\partial}{\partial \theta_i} F)(\theta + ue_i, X) \, \mathrm{d}u\bigg]. \end{aligned} \tag{295}$$

Moreover, note that Fubini's theorem (see, e.g., Klenke [50, Theorem 14.16]), (294), the hypothesis that $p > 1$, and Jensen's inequality assure that for all $i \in \{1, 2, \ldots, d\}$, $\theta \in \mathbb{R}^d$, $h \in [-q, q]$ it holds that

$$\begin{aligned} \mathbb{E}\bigg[\int_{\min\{h,0\}}^{\max\{h,0\}} |(\tfrac{\partial}{\partial \theta_i} F)(\theta + ue_i, X)| \, \mathrm{d}u\bigg] &= \int_{\min\{h,0\}}^{\max\{h,0\}} \mathbb{E}\big[|(\tfrac{\partial}{\partial \theta_i} F)(\theta + ue_i, X)|\big] \, \mathrm{d}u \\ &\leq |h| \Bigg[\sup_{v \in [-q,q]^d} \mathbb{E}\big[|(\tfrac{\partial}{\partial \theta_i} F)(\theta + v, X)|\big]\Bigg] \\ &\leq |h| \Bigg[\sup_{v \in [-q,q]^d} \big(\mathbb{E}\big[|(\tfrac{\partial}{\partial \theta_i} F)(\theta + v, X)|^p\big]\big)^{1/p}\Bigg] \\ &< \infty. \end{aligned} \tag{296}$$

This, (295), and again Fubini's theorem (see, e.g., Klenke [50, Theorem 14.16]) imply that for all $i \in \{1, 2, \dots, d\}$, $\theta \in \mathbb{R}^d$, $h \in [-q, q]$ it holds that

$$\begin{aligned} f(\theta + he_i) - f(\theta) &= \mathbb{E}\bigg[\int_0^h (\tfrac{\partial}{\partial \theta_i} F)(\theta + ue_i, X) \, \mathrm{d}u \bigg] \\ &= \int_0^h \mathbb{E}\big[(\tfrac{\partial}{\partial \theta_i} F)(\theta + ue_i, X) \big] \, \mathrm{d}u. \end{aligned} \tag{297}$$

In addition, note that (294), the hypothesis that $p > 1$, and the de la Vallée Poussin theorem (cf., e.g., Klenke [50, Corollary 6.21]) ensure that for all $i \in \{1, 2, \dots, d\}$, $\theta \in \mathbb{R}^d$ it holds that the family of random variables

$$\big(\Omega \ni \omega \mapsto (\tfrac{\partial}{\partial \theta_i} F)(\theta + v, X(\omega)) \in \mathbb{R} \big), \qquad v \in [-q, q]^d, \tag{298}$$

is uniformly integrable. The fact that for all $i \in \{1, 2, \dots, d\}$, $x \in S$ it holds that the function $\mathbb{R}^d \ni \theta \mapsto (\frac{\partial}{\partial \theta_i} F)(\theta, x) \in \mathbb{R}$ is continuous and the Vitali convergence theorem (cf., e.g., Klenke [50, Theorem 6.25]) hence imply that for all $i \in \{1, 2, \dots, d\}$, $\theta \in \mathbb{R}^d$ and all functions $v = (v_n)_{n \in \mathbb{N}} \colon \mathbb{N} \to \mathbb{R}$ with $\limsup_{n \to \infty} |v_n| = 0$ it holds that

$$\limsup_{n \to \infty} \mathbb{E}\Big[\big| (\tfrac{\partial}{\partial \theta_i} F)(\theta + v_n e_i, X) - (\tfrac{\partial}{\partial \theta_i} F)(\theta, X) \big| \Big] = 0. \tag{299}$$

Hence, we obtain that for all $i \in \{1, 2, \dots, d\}$, $\theta \in \mathbb{R}^d$, $\varepsilon \in (0, \infty)$ there exists $\delta \in (0, \infty)$ such that

$$\sup_{u \in [-\delta, \delta]} \mathbb{E}\Big[\big| (\tfrac{\partial}{\partial \theta_i} F)(\theta + ue_i, X) - (\tfrac{\partial}{\partial \theta_i} F)(\theta, X) \big| \Big] \le \varepsilon. \tag{300}$$

Therefore, we obtain that for all $i \in \{1, 2, \dots, d\}$, $\theta \in \mathbb{R}^d$ and all functions $h = (h_n)_{n \in \mathbb{N}} \colon \mathbb{N} \to \mathbb{R}$ with $\limsup_{n \to \infty} |h_n| = 0$ it holds that

$$\limsup_{n \to \infty} \sup_{u \in [-|h_n|, |h_n|]} \mathbb{E}\Big[\big| (\tfrac{\partial}{\partial \theta_i} F)(\theta + ue_i, X) - (\tfrac{\partial}{\partial \theta_i} F)(\theta, X) \big| \Big] = 0. \tag{301}$$

This and (297) demonstrate that for all $i \in \{1, 2, \dots, d\}$, $\theta \in \mathbb{R}^d$ and all functions

$h = (h_n)_{n\in\mathbb{N}} \colon \mathbb{N} \to \mathbb{R} \setminus \{0\}$ with $\limsup_{n\to\infty} |h_n| = 0$ it holds that

$$\begin{aligned}
&\limsup_{n\to\infty} \left| \tfrac{f(\theta+h_n e_i)-f(\theta)}{h_n} - \mathbb{E}\big[(\tfrac{\partial}{\partial\theta_i}F)(\theta,X)\big] \right| \\
&= \limsup_{n\to\infty} \left[\frac{1}{|h_n|} \left| f(\theta+h_n e_i) - f(\theta) - h_n\, \mathbb{E}\big[(\tfrac{\partial}{\partial\theta_i}F)(\theta,X)\big] \right| \right] \\
&= \limsup_{n\to\infty} \left[\frac{1}{|h_n|} \left| f(\theta+h_n e_i) - f(\theta) - \int_0^{h_n} \mathbb{E}\big[(\tfrac{\partial}{\partial\theta_i}F)(\theta,X)\big]\, \mathrm{d}u \right| \right] \\
&= \limsup_{n\to\infty} \left[\frac{1}{|h_n|} \left| \int_0^{h_n} \mathbb{E}\big[(\tfrac{\partial}{\partial\theta_i}F)(\theta+ue_i,X)\big]\, \mathrm{d}u - \int_0^{h_n} \mathbb{E}\big[(\tfrac{\partial}{\partial\theta_i}F)(\theta,X)\big]\, \mathrm{d}u \right| \right] \\
&= \limsup_{n\to\infty} \left[\frac{1}{|h_n|} \left| \int_0^{h_n} \mathbb{E}\big[(\tfrac{\partial}{\partial\theta_i}F)(\theta+ue_i,X) - (\tfrac{\partial}{\partial\theta_i}F)(\theta,X)\big]\, \mathrm{d}u \right| \right] \\
&\le \limsup_{n\to\infty} \left[\frac{1}{|h_n|} \int_{\min\{h_n,0\}}^{\max\{h_n,0\}} \mathbb{E}\Big[\big|(\tfrac{\partial}{\partial\theta_i}F)(\theta+ue_i,X) - (\tfrac{\partial}{\partial\theta_i}F)(\theta,X)\big|\Big]\, \mathrm{d}u \right] \\
&\le \limsup_{n\to\infty} \left[\sup_{u\in[-|h_n|,|h_n|]} \mathbb{E}\Big[\big|(\tfrac{\partial}{\partial\theta_i}F)(\theta+ue_i,X) - (\tfrac{\partial}{\partial\theta_i}F)(\theta,X)\big|\Big] \right] = 0.
\end{aligned} \tag{302}$$

Next observe that (298), the fact that for all $i \in \{1,2,\dots,d\}$, $x \in S$ it holds that the function $\mathbb{R}^d \ni \theta \mapsto (\frac{\partial}{\partial\theta_i}F)(\theta,x) \in \mathbb{R}$ is continuous, and the Vitali convergence theorem (cf., e.g, Klenke [50, Theorem 6.25]) assure that for all $i \in \{1,2,\dots,d\}$, $\theta \in \mathbb{R}^d$ and all sequences $v = (v_n)_{n\in\mathbb{N}} \colon \mathbb{N} \to \mathbb{R}^d$ with $\limsup_{n\to\infty} \|v_n\| = 0$ it holds that

$$\begin{aligned}
&\limsup_{n\to\infty} \big|\mathbb{E}\big[(\tfrac{\partial}{\partial\theta_i}F)(\theta+v_n,X)\big] - \mathbb{E}\big[(\tfrac{\partial}{\partial\theta_i}F)(\theta,X)\big]\big| \\
&\le \limsup_{n\to\infty} \mathbb{E}\Big[\big|(\tfrac{\partial}{\partial\theta_i}F)(\theta+v_n,X) - (\tfrac{\partial}{\partial\theta_i}F)(\theta,X)\big|\Big] = 0.
\end{aligned} \tag{303}$$

Combining this and (302) establishes items (i) and (ii). The proof of Lemma 4.8 is thus completed. □

Corollary 4.9. *Let $d \in \mathbb{N}$, $p \in \{2,4,6,\dots\}$, $\alpha,\kappa,c \in (0,\infty)$, $\nu \in (0,1)$, $\xi,\vartheta \in \mathbb{R}^d$, let $\langle\cdot,\cdot\rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the d-dimensional Euclidean scalar product, let $\|\cdot\| \colon \mathbb{R}^d \to [0,\infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that $\|\theta\| = \sqrt{\langle\theta,\theta\rangle}$, let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, let $(S,\mathcal{S})$ be a measurable space, let $X_n \colon \Omega \to S$, $n \in \mathbb{N}$, be i.i.d. random variables, let $F = (F(\theta,x))_{\theta\in\mathbb{R}^d,x\in S} \colon \mathbb{R}^d \times S \to \mathbb{R}$ be $(\mathcal{B}(\mathbb{R}^d)\otimes\mathcal{S})/\mathcal{B}(\mathbb{R})$-measurable, assume for all $x \in S$ that $(\mathbb{R}^d \ni \theta \mapsto F(\theta,x) \in \mathbb{R}) \in C^1(\mathbb{R}^d,\mathbb{R})$, assume* *for all $\theta \in \mathbb{R}^d$ that*

$$\mathbb{E}\big[|F(\theta, X_1)| + \|(\nabla_\theta F)(\theta, X_1)\|\big] < \infty, \tag{304}$$

$$\langle \theta - \vartheta, \mathbb{E}[(\nabla_\theta F)(\theta, X_1)] \rangle \geq c \max\big\{\|\theta - \vartheta\|^2, \|\mathbb{E}[(\nabla_\theta F)(\theta, X_1)]\|^2\big\}, \tag{305}$$

$$\mathbb{E}\big[\|(\nabla_\theta F)(\theta, X_1) - \mathbb{E}[(\nabla_\theta F)(\theta, X_1)]\|^p\big] \leq \kappa\big(1 + \|\theta\|^p\big), \tag{306}$$

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

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \tfrac{\alpha}{n^\nu}(\nabla_\theta F)(\Theta_{n-1}, X_n). \tag{307}$$

Then

(i) it holds that $\big\{\theta \in \mathbb{R}^d \colon \big(\mathbb{E}[F(\theta, X_1)] = \inf_{v \in \mathbb{R}^d} \mathbb{E}[F(v, X_1)]\big)\big\} = \{\vartheta\}$ and

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

$$\big(\mathbb{E}\big[\|\Theta_n - \vartheta\|^p\big]\big)^{1/p} \leq C n^{-\nu/2}. \tag{308}$$

Proof of Corollary 4.9. Throughout this proof let $f\colon \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that

$$f(\theta) = \mathbb{E}[F(\theta, X_1)]\,. \tag{309}$$

Lemma 4.8 assures that for all $\theta \in \mathbb{R}^d$ it holds that

$$f \in C^1(\mathbb{R}^d, \mathbb{R}) \qquad \text{and} \qquad (\nabla f)(\theta) = \mathbb{E}\big[(\nabla_\theta F)(\theta, X_1)\big]. \tag{310}$$

This and (305) imply that for all $\theta \in \mathbb{R}^d$ it holds that

$$\langle \theta - \vartheta, (\nabla f)(\theta) \rangle \geq c \max\big\{\|\theta - \vartheta\|^2, \|(\nabla f)(\theta)\|^2\big\}. \tag{311}$$

Hence, we obtain that for all $\theta \in \mathbb{R}^d$ it holds that

$$\langle \theta - \vartheta, (\nabla f)(\theta) \rangle \geq c\|\theta - \vartheta\|^2. \tag{312}$$

This proves that for all $v \in \mathbb{R}^d$ it holds that

$$\langle v, (\nabla f)(\vartheta + v) \rangle \geq c\|v\|^2. \tag{313}$$

The fundamental theorem of calculus therefore ensures that for all $\theta \in \mathbb{R}^d$ it holds that

$$\begin{aligned} f(\theta) &= f(\vartheta) + \big[f(\vartheta + t(\theta - \vartheta))\big]_{t=0}^{t=1} \\ &= f(\vartheta) + \int_0^1 f'(\vartheta + s(\theta - \vartheta))(\theta - \vartheta)\,\mathrm{d}s \\ &= f(\vartheta) + \int_0^1 \big\langle (\nabla f)\big(\vartheta + s(\theta - \vartheta)\big), \theta - \vartheta \big\rangle\,\mathrm{d}s \\ &= f(\vartheta) + \int_0^1 \frac{1}{s}\,\big\langle (\nabla f)\big(\vartheta + s(\theta - \vartheta)\big), s(\theta - \vartheta) \big\rangle\,\mathrm{d}s \\ &\geq f(\vartheta) + \int_0^1 \frac{c}{s}\,\|s(\theta - \vartheta)\|^2\,\mathrm{d}s \\ &= f(\vartheta) + \frac{c}{2}\,\|\theta - \vartheta\|^2. \end{aligned} \tag{314}$$

The hypothesis that $c \in (0, \infty)$ hence demonstrates that for all $\theta \in \mathbb{R}^d \setminus \{\vartheta\}$ it holds that

$$f(\theta) \geq f(\vartheta) + \frac{c}{2}\|\theta - \vartheta\|^2 > f(\vartheta). \tag{315}$$

This establishes item (i). Moreover, observe that Corollary 4.6 (with $F = -F$ in the notation of Corollary 4.6) establishes item (ii). The proof of Corollary 4.9 is thus completed. □

4.3 Stochastic approximation for linear regression

Lemma 4.10. *Let $d \in \mathbb{N}$, let $\langle \cdot, \cdot \rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the d-dimensional Euclidean scalar product, let $\|\cdot\| \colon \mathbb{R}^d \to [0, \infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that $\|\theta\| = \sqrt{\langle \theta, \theta \rangle}$, let $A \in \mathbb{R}^{d \times d}$, assume that A is invertible and symmetric, and assume for all $\theta \in \mathbb{R}^d$ that $\langle \theta, A\theta \rangle \geq 0$. Then there exists $c \in (0, \infty)$ such that for all $\theta \in \mathbb{R}^d$ it holds that*

$$\langle \theta, A\theta \rangle \geq c \max\{\|\theta\|^2, \|A\theta\|^2\}. \tag{316}$$

Proof of Lemma 4.10. Throughout this proof for every $v \in \mathbb{R}^d$ let $v^* \in \mathbb{R}^{1 \times d}$ be the transpose of v, for every $M \in \mathbb{R}^{d \times d}$ let $M^* \in \mathbb{R}^{d \times d}$ be the transpose of M, let $e_1 = (1, 0, \dots, 0)$, $e_2 = (0, 1, 0, \dots, 0)$, $\dots$, $e_d = (0, \dots, 0, 1) \in \mathbb{R}^d$, let $E \in \mathbb{R}^{d \times d}$ be the $(d \times d)$-identity matrix, let $T \in \mathbb{R}^{d \times d}$ and $D = (\delta_{i,j})_{(i,j) \in \{1, \dots, d\}^2} \in \mathbb{R}^{d \times d}$ be $(d \times d)$-matrices such that D is a diagonal matrix and such that

$$TT^* = E \qquad \text{and} \qquad A = TDT^*, \tag{317}$$

let $c_0 = \min_{i \in \{1,\dots,d\}} \delta_{i,i} \in \mathbb{R}$, and let $c_1 = \max_{i \in \{1,\dots,d\}} \delta_{i,i} \in \mathbb{R}$. Note that (317) implies that for all $\theta \in \mathbb{R}^d$ it holds that

$$\|T^*\theta\|^2 = \langle T^*\theta, T^*\theta \rangle = (T^*\theta)^*(T^*\theta) = \theta^* T T^* \theta = \theta^* E \theta = \langle \theta, \theta \rangle = \|\theta\|^2. \tag{318}$$

Furthermore, observe that (317) demonstrates that T is invertible with $T^{-1} = T^*$. Hence, we obtain that for all $\theta \in \mathbb{R}^d$ it holds that

$$\|T\theta\|^2 = \langle T\theta, T\theta \rangle = (T\theta)^*(T\theta) = \theta^* T^* T \theta = \theta^* T^{-1} T \theta = \theta^* E \theta = \langle \theta, \theta \rangle = \|\theta\|^2. \tag{319}$$

Combining this with (317) and the hypothesis that for all $\theta \in \mathbb{R}^d$ it holds that $\langle \theta, A\theta \rangle \geq 0$ ensures that for all $i \in \{1, \dots, d\}$ it holds that

$$\begin{split} \delta_{i,i} = \delta_{i,i} \langle e_i, e_i \rangle = \langle \delta_{i,i} e_i, e_i \rangle = \langle D e_i, e_i \rangle = \langle T D e_i, T e_i \rangle \\ = \langle T D T^{-1} T e_i, T e_i \rangle = \langle A(T e_i), T e_i \rangle \geq 0. \end{split} \tag{320}$$

Furthermore, observe that (317), the fact that A is invertible, and the fact that T is invertible prove that for all $i \in \{1, \dots, d\}$ it holds that

$$\delta_{i,i} e_i = D e_i = T^{-1} A T e_i \neq 0. \tag{321}$$

This assures that for all $i \in \{1, \dots, d\}$ it holds that

$$\delta_{i,i} \neq 0. \tag{322}$$

Combining this with (320) shows that for all $i \in \{1, \dots, d\}$ it holds that

$$\delta_{i,i} > 0. \tag{323}$$

Hence, we obtain that

$$0 < c_0 = \min_{i \in \{1,2,\dots,d\}} \delta_{i,i} \leq \max_{i \in \{1,2,\dots,d\}} \delta_{i,i} = c_1 < \infty. \tag{324}$$

Next note that for all $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$ it holds that

$$\langle \theta, D\theta \rangle = \sum_{i=1}^{d} \big(\theta_i (\delta_{i,i} \theta_i) \big) = \sum_{i=1}^{d} \big(\delta_{i,i} (\theta_i)^2 \big) \geq c_0 \left[\sum_{i=1}^{d} (\theta_i)^2 \right] = c_0 \langle \theta, \theta \rangle = c_0 \|\theta\|^2. \tag{325}$$

Moreover, observe that (324) ensures that for all $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$ it holds that

$$\begin{split} \|D\theta\|^2 = \langle D\theta, D\theta \rangle = \sum_{i=1}^{d} \big((\delta_{i,i}\theta_i)(\delta_{i,i}\theta_i) \big) = \sum_{i=1}^{d} \big((\delta_{i,i})^2 (\theta_i)^2 \big) \\ \leq (c_1)^2 \left[\sum_{i=1}^{d} (\theta_i)^2 \right] = (c_1)^2 \langle \theta, \theta \rangle = (c_1)^2 \|\theta\|^2. \end{split} \tag{326}$$

Furthermore, note that (317) implies that for all $\theta \in \mathbb{R}^d$ it holds that

$$\langle \theta, A\theta \rangle = \langle \theta, TDT^*\theta \rangle = \theta^* TDT^*\theta = (T^*\theta)^* D(T^*\theta) = \langle T^*\theta, D(T^*\theta) \rangle. \tag{327}$$

This, (318), and (325) ensure that for all $\theta \in \mathbb{R}^d$ it holds that

$$\langle \theta, A\theta \rangle \geq c_0 \|T^*\theta\|^2 = c_0 \|\theta\|^2. \tag{328}$$

Furthermore, observe that for all $\theta \in \mathbb{R}^d$ it holds that

$$\|A\theta\| \leq \left[\sup_{v \in \mathbb{R}^d \setminus \{0\}} \frac{\|Av\|}{\|v\|} \right] \|\theta\|. \tag{329}$$

Hence, we obtain that for all $\theta \in \mathbb{R}^d$ it holds that

$$\left[\inf_{v \in \mathbb{R}^d \setminus \{0\}} \frac{\|A^{-1}v\|}{\|v\|} \right] \|A\theta\| = \left[\inf_{v \in \mathbb{R}^d \setminus \{0\}} \frac{\|v\|}{\|Av\|} \right] \|A\theta\| = \frac{\|A\theta\|}{\left[\sup_{v \in \mathbb{R}^d \setminus \{0\}} \frac{\|Av\|}{\|v\|} \right]} \leq \|\theta\|. \tag{330}$$

This shows that for all $\theta \in \mathbb{R}^d$ it holds that

$$\|\theta\|^2 \geq \left[\inf_{v \in \mathbb{R}^d \setminus \{0\}} \frac{\|A^{-1}v\|}{\|v\|} \right]^2 \|A\theta\|^2 \geq \left[\min\left\{ 1, \left[\inf_{v \in \mathbb{R}^d \setminus \{0\}} \tfrac{\|A^{-1}v\|}{\|v\|} \right]^2 \right\} \right] \|A\theta\|^2. \tag{331}$$

Hence, we obtain that for all $\theta \in \mathbb{R}^d$ it holds that

$$\|\theta\|^2 \geq \left[\min\left\{ 1, \inf_{v \in \mathbb{R}^d \setminus \{0\}} \tfrac{\|A^{-1}v\|^2}{\|v\|^2} \right\} \right] \max\{ \|\theta\|^2, \|A\theta\|^2 \}. \tag{332}$$

Combining this with (328) demonstrates that for all $\theta \in \mathbb{R}^d$ it holds that

$$\langle \theta, A\theta \rangle \geq \left[c_0 \min\left\{ 1, \inf_{v \in \mathbb{R}^d \setminus \{0\}} \tfrac{\|A^{-1}v\|^2}{\|v\|^2} \right\} \right] \max\{ \|\theta\|^2, \|A\theta\|^2 \}. \tag{333}$$

The proof of Lemma 4.10 is thus completed. □

Corollary 4.11 (Stochastic approximation for linear regression)**.** *Let $d \in \mathbb{N}$, $p \in \{2, 4, 6, \ldots\}$, $\alpha \in (0, \infty)$, $\nu \in (0, 1)$, $\xi \in \mathbb{R}^d$, let $\langle \cdot, \cdot \rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the d-dimensional Euclidean scalar product, let $\|\cdot\| \colon \mathbb{R}^d \to [0, \infty)$ be the function which satisfies for all $\theta \in \mathbb{R}^d$ that $\|\theta\| = \sqrt{\langle \theta, \theta \rangle}$, 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, let $h \colon \mathbb{R}^d \to \mathbb{R}$ be $\mathcal{B}(\mathbb{R}^d)/\mathcal{B}(\mathbb{R})$-measurable, assume that $\mathbb{E}\big[\|h(X_1)X_1\|^p + |h(X_1)|^2 + \|X_1\|^{2p} \big] < \infty$, for every $v \in \mathbb{R}^d$*

let $v^* \in \mathbb{R}^{1\times d}$ be the transpose of v, assume that $\mathbb{E}[X_1(X_1)^*] \in \mathbb{R}^{d\times d}$ is invertible, let $\vartheta \in \mathbb{R}^d$ satisfy

$$\vartheta = \big(\mathbb{E}\big[X_1(X_1)^*\big]\big)^{-1}\,\mathbb{E}\big[h(X_1)X_1\big], \tag{334}$$

let $F = (F(\theta,x))_{\theta\in\mathbb{R}^d, x\in\mathbb{R}^d} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the function which satisfies for all $\theta, x \in \mathbb{R}^d$ that $F(\theta,x) = [\langle \theta, x\rangle - h(x)]^2$, 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

$$\Theta_0 = \xi \qquad \text{and} \qquad \Theta_n = \Theta_{n-1} - \tfrac{\alpha}{n^\nu}(\nabla_\theta F)(\Theta_{n-1}, X_n). \tag{335}$$

Then

(i) it holds for all $\theta \in \mathbb{R}^d$ that $\mathbb{E}\big[|F(\theta, X_1)|\big] < \infty$,

(ii) it holds that $\big\{\theta \in \mathbb{R}^d \colon \big(\mathbb{E}[F(\theta,X_1)] = \inf_{v\in\mathbb{R}^d}\mathbb{E}[F(v,X_1)]\big)\big\} = \{\vartheta\}$, and

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

$$\big(\mathbb{E}\big[\|\Theta_n - \vartheta\|^p\big]\big)^{1/p} \leq C n^{-\nu/2}. \tag{336}$$

Proof of Corollary 4.11. Observe that the chain rule and the hypothesis that for all $\theta, x \in \mathbb{R}^d$ it holds that $F(\theta,x) = [\langle\theta,x\rangle - h(x)]^2$ ensure that for all $\theta, x, v \in \mathbb{R}^d$ it holds that

$$(\tfrac{\partial}{\partial\theta}F)(\theta,x)(v) = 2[\langle\theta,x\rangle - h(x)]\langle v, x\rangle. \tag{337}$$

Hence, we obtain that for all $\theta, x \in \mathbb{R}^d$ it holds that

$$(\nabla_\theta F)(\theta,x) = 2[\langle\theta,x\rangle - h(x)]x = 2x\langle x,\theta\rangle - 2h(x)x = 2xx^*\theta - 2h(x)x. \tag{338}$$

This, the hypothesis that $\mathbb{E}\big[\|h(X_1)X_1\|^p + |h(X_1)|^2 + \|X_1\|^{2p}\big] < \infty$, the Cauchy-Schwarz inequality, and Jensen's inequality assure that for all $\theta \in \mathbb{R}^d$ it holds that

$$\begin{aligned}
\mathbb{E}\big[\|(\nabla_\theta F)(\theta,X_1)\|\big] &= 2\,\mathbb{E}\big[\|X_1(X_1)^*\theta - h(X_1)X_1\|\big] \\
&\leq 2\,\mathbb{E}\big[\|X_1(X_1)^*\theta\|\big] + 2\,\mathbb{E}\big[\|h(X_1)X_1\|\big] \\
&= 2\,\mathbb{E}\big[\|X_1\langle X_1,\theta\rangle\|\big] + 2\,\mathbb{E}\big[\|h(X_1)X_1\|\big] \\
&= 2\,\mathbb{E}\big[|\langle X_1,\theta\rangle|\|X_1\|\big] + 2\,\mathbb{E}\big[\|h(X_1)X_1\|\big] \\
&\leq 2\,\mathbb{E}\big[\|X_1\|\|\theta\|\|X_1\|\big] + 2\,\mathbb{E}\big[\|h(X_1)X_1\|\big] \\
&= 2\,\|\theta\|\,\mathbb{E}\big[\|X_1\|^2\big] + 2\,\mathbb{E}\big[\|h(X_1)X_1\|\big] < \infty.
\end{aligned} \tag{339}$$

Next observe that the hypothesis that for all $\theta, x \in \mathbb{R}^d$ it holds that $F(\theta, x) = [\langle \theta, x \rangle - h(x)]^2$ implies that for all $\theta, x \in \mathbb{R}^d$ it holds that

$$\begin{aligned} F(\theta, x) &= [\langle \theta, x \rangle - h(x)][\langle \theta, x \rangle - h(x)] \\ &= |\langle \theta, x \rangle|^2 - 2\langle \theta, x \rangle h(x) + |h(x)|^2 \\ &= \langle \theta, x \rangle \langle x, \theta \rangle - 2\langle \theta, h(x)x \rangle + |h(x)|^2 \\ &= \theta^* x x^* \theta - 2\theta^* h(x) x + |h(x)|^2. \end{aligned} \tag{340}$$

This, the triangle inequality, the Cauchy-Schwarz inequality, Jensen's inequality, and the hypothesis that $\mathbb{E}\big[\|h(X_1)X_1\|^p + |h(X_1)|^2 + \|X_1\|^{2p}\big] < \infty$ ensure that for all $\theta \in \mathbb{R}^d$ it holds that

$$\begin{aligned} &\mathbb{E}\big[|F(\theta, X_1)|\big] \\ &= \mathbb{E}\Big[\big|[\langle \theta, X_1 \rangle]^2 - 2\langle \theta, h(X_1)X_1 \rangle + [h(X_1)]^2\big|\Big] \\ &\leq \mathbb{E}\big[|\langle \theta, X_1 \rangle|^2 + 2|\langle \theta, h(X_1)X_1 \rangle| + |h(X_1)|^2\big] \\ &\leq \mathbb{E}\big[\|\theta\|^2 \|X_1\|^2 + 2\|\theta\| \|h(X_1)X_1\| + |h(X_1)|^2\big] \\ &\leq \big(1 + 2\|\theta\| + \|\theta\|^2\big)\Big(\mathbb{E}\big[\|X_1\|^2\big] + \mathbb{E}\big[\|h(X_1)X_1\|\big] + \mathbb{E}\big[|h(X_1)|^2\big]\Big) < \infty. \end{aligned} \tag{341}$$

Next note that (334), (338), and (339) imply that for all $\theta \in \mathbb{R}^d$ it holds that

$$\begin{aligned} \mathbb{E}\big[(\nabla_\theta F)(\theta, X_1)\big] &= 2\,\mathbb{E}\big[X_1(X_1)^*\big]\theta - 2\,\mathbb{E}\big[h(X_1)X_1\big] \\ &= 2\,\mathbb{E}\big[X_1(X_1)^*\big]\theta - 2\,\mathbb{E}\big[X_1(X_1^*)\big]\big(\mathbb{E}\big[X_1(X_1^*)\big]\big)^{-1}\mathbb{E}\big[h(X_1)X_1\big] \\ &= 2\,\mathbb{E}\big[X_1(X_1)^*\big]\theta - 2\,\mathbb{E}\big[X_1(X_1)^*\big]\vartheta \\ &= 2\,\mathbb{E}\big[X_1(X_1)^*\big](\theta - \vartheta). \end{aligned} \tag{342}$$

In addition, observe that for all $\theta \in \mathbb{R}^d$ it holds that

$$\big\langle \theta, 2\,\mathbb{E}\big[X_1(X_1)^*\big]\theta \big\rangle = 2\,\mathbb{E}\big[\theta^* X_1(X_1)^*\theta\big] = 2\,\mathbb{E}\big[|\theta^* X_1|^2\big] \geq 0. \tag{343}$$

The fact that $2\,\mathbb{E}[X_1(X_1)^*]$ is symmetric and invertible, (342), and Lemma 4.10 hence ensure that there exists $c \in (0, \infty)$ such that for all $\theta \in \mathbb{R}^d$ it holds that

$$\begin{aligned} \big\langle \theta - \vartheta, \mathbb{E}[(\nabla_\theta F)(\theta, X_1)] \big\rangle &= \big\langle \theta - \vartheta, (2\,\mathbb{E}[X_1(X_1)^*])(\theta - \vartheta) \big\rangle \\ &\geq c \max\big\{\|\theta - \vartheta\|^2, \|(2\,\mathbb{E}[X_1(X_1)^*])(\theta - \vartheta)\|^2\big\} \\ &= c \max\big\{\|\theta - \vartheta\|^2, \|\mathbb{E}[(\nabla_\theta F)(\theta, X_1)]\|^2\big\}. \end{aligned} \tag{344}$$

This and item (iii) in Lemma 2.12 (with $g(\theta) = -\mathbb{E}[(\nabla_\theta F)(\theta, X_1)]$ for $\theta \in \mathbb{R}^d$ in the notation of Lemma 2.12) ensure that there exists $\mathfrak{C} \in (0, \infty)$ such that for all $\theta \in \mathbb{R}^d$ it holds that

$$\|\mathbb{E}[(\nabla_\theta F)(\theta, X_1)]\| \leq \mathfrak{C}\|\theta - \vartheta\| \leq \mathfrak{C}\|\vartheta\| + \mathfrak{C}\|\theta\|. \tag{345}$$

Lemma 2.1 and (338) hence imply that for all $\theta \in \mathbb{R}^d$ it holds that

$$\begin{aligned}
&\mathbb{E}\big[\|(\nabla_\theta F)(\theta, X_1) - \mathbb{E}[(\nabla_\theta F)(\theta, X_1)]\|^p\big] \\
&\leq \mathbb{E}\big[2^p\big(\|(\nabla_\theta F)(\theta, X_1)\|^p + \|\mathbb{E}[(\nabla_\theta F)(\theta, X_1)]\|^p\big)\big] \\
&\leq 2^p\big(\mathbb{E}\big[\|2X_1(X_1)^*\theta - 2h(X_1)X_1\|^p\big] + (\mathfrak{C}\|\vartheta\| + \mathfrak{C}\|\theta\|)^p\big) \\
&\leq 2^p\big(2^p\,\mathbb{E}\big[\|X_1(X_1)^*\theta - h(X_1)X_1\|^p\big] + 2^p(\mathfrak{C}^p\|\vartheta\|^p + \mathfrak{C}^p\|\theta\|^p)\big) \\
&= 2^{2p}\big(\mathbb{E}\big[\|X_1(X_1)^*\theta - h(X_1)X_1\|^p\big] + \mathfrak{C}^p\|\vartheta\|^p + \mathfrak{C}^p\|\theta\|^p\big).
\end{aligned} \tag{346}$$

The triangle inequality, Lemma 2.1, and the Cauchy-Schwarz inequality therefore demonstrate that for all $\theta \in \mathbb{R}^d$ it holds that

$$\begin{aligned}
&\mathbb{E}\big[\|(\nabla_\theta F)(\theta, X_1) - \mathbb{E}[(\nabla_\theta F)(\theta, X_1)]\|^p\big] \\
&\leq 2^{2p}\Big(\mathbb{E}\big[2^p\big(\|X_1(X_1)^*\theta\|^p + \|h(X_1)X_1\|^p\big)\big] + \mathfrak{C}^p\|\vartheta\|^p + \mathfrak{C}^p\|\theta\|^p\Big) \\
&\leq 2^{2p}\Big(2^p\,\mathbb{E}\big[\|\theta\|^p\|X_1\|^{2p} + \|h(X_1)X_1\|^p\big] + \mathfrak{C}^p\|\vartheta\|^p + \mathfrak{C}^p\|\theta\|^p\Big) \\
&\leq 2^{3p}\Big(\mathbb{E}\big[\|X_1\|^{2p}\big]\|\theta\|^p + \mathbb{E}\big[\|h(X_1)X_1\|^p\big] + \mathfrak{C}^p\|\vartheta\|^p + \mathfrak{C}^p\|\theta\|^p\Big) \\
&\leq 2^{3p}\Big(\mathbb{E}\big[\|X_1\|^{2p}\big] + \mathbb{E}\big[\|h(X_1)X_1\|^p\big] + \mathfrak{C}^p\|\vartheta\|^p + \mathfrak{C}^p\Big)(1 + \|\theta\|^p) \\
&\leq \bigg[2^{3p+2}\max\Big\{\mathbb{E}\big[\|X_1\|^{2p}\big], \mathbb{E}\big[\|h(X_1)X_1\|^p\big], \mathfrak{C}^p\|\vartheta\|^p, \mathfrak{C}^p\Big\}\bigg](1 + \|\theta\|^p).
\end{aligned} \tag{347}$$

Combining this, the fact that for all $x \in \mathbb{R}^d$ it holds that $(\mathbb{R}^d \ni \theta \mapsto F(\theta, x) \in \mathbb{R}) \in C^1(\mathbb{R}^d, \mathbb{R})$, (335), (339), (341), and (344) with Corollary 4.9 (with $\kappa = 2^{3p+2}\max\big\{\mathbb{E}[\|X_1\|^{2p}], \mathbb{E}[\|h(X_1)X_1\|^p], \mathfrak{C}^p\|\vartheta\|^p, \mathfrak{C}^p\big\}$, $c = c$ in the notation of Corollary 4.9) establishes items (i), (ii), and (iii). The proof of Corollary 4.11 is thus completed. □

References

[1] Amari, S.-I. Natural gradient works efficiently in learning. *Neural computation 10*, 2 (1998), 251–276.

[2] Amari, S.-i., Park, H., and Fukumizu, K. Adaptive method of realizing natural gra- dient learning for multilayer perceptrons. *Neural Computation 12*, 6 (2000), 1399–1409.

[3] Bach, F. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research 15* (2014), 595–627.

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

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

[6] Benaï m, M. *Dynamics of stochastic approximation algorithms*, vol. 1709 of *Lecture Notes in Math.* Springer, Berlin, 1999.

[7] Benveniste, A., Métivier, M., and Priouret, P. *Adaptive algorithms and stochastic approximations*, vol. 22 of *Applications of Mathematics (New York).* Springer-Verlag, Berlin, 1990. Translated from the French by Stephen S. Wilson.

[8] Bercu, B., and Fort, J.-C. Generic stochastic gradient methods. *Wiley Encyclopedia of Operations Research and Management Science* (2013), 1–8.

[9] Bhatnagar, S., Prasad, H. L., and Prashanth, L. A. *Stochastic recursive algorithms for optimization*, vol. 434 of *Lecture Notes in Control and Information Sciences.* Springer, London, 2013. Simultaneous perturbation methods.

[10] Bordes, A., Bottou, L., and Gallinari, P. SGD-QN: Careful quasi-newton stochastic gradient descent. *Journal of Machine Learning Research 10* (2009), 1737–1754.

[11] Borkar, V. S. *Stochastic approximation.* Cambridge University Press, Cambridge; Hindustan Book Agency, New Delhi, 2008. A dynamical systems viewpoint.

[12] Bottou, L. On-line learning in neural networks. Cambridge University Press, New York, NY, USA, 1998, ch. On-line Learning and Stochastic Approximations, pp. 9–42.

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

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

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

[16] 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).

[17] Bottou, L., and LeCun, Y. On-line learning for very large datasets. *Apllied Stochastic Models in Business and Industry 21* (2005), 137–151.

[18] Broadie, M. N., Cicek, D. M., and Zeevi, A. General bounds and finite-time improvement for stochastic approximation algorithms. *Technical report, Columbia University* (2009).

[19] Brosse, N., Durmus, A., Moulines, E., and Sabanis, S. The tamed unadjusted langevin algorithm. *preprint, arXiv:1710.05559* (2017).

[20] Chapelle, O., and Erhan, D. Improved preconditioner for hessian free optimization. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning* (2011).

[21] 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).

[22] Chen, H.-F. *Stochastic approximation and its applications*, vol. 64 of *Nonconvex Optimization and its Applications.* Kluwer Academic Publishers, Dordrecht, 2002.

[23] Chung, K. L. On a stochastic approximation method. *Ann. Math. Statistics 25* (1954), 463–483.

[24] Da Prato, G., and Zabczyk, J. *Stochastic equations in infinite dimensions*, vol. 44 of *Encyclopedia of Mathematics and its Applications.* Cambridge University Press, Cambridge, 1992.

[25] Darken, C., Chang, J., and Moody, J. Learning rate schedules for faster stochastic gradient search. *Neural Networks for Signal Processing II Proceedings of the 1992 IEEE Workshop* (1992), 1–11.

[26] Dauphin, Y., de Vries, H., and Bengio, Y. Equilibrated adaptive learning rates for non- convex optimization. *Advances in Neural Information Processing Systems* (2015), 1504–1512.

[27] Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *preprint, arXiv:1406.2572* (2014).

[28] Dean, J., Corrado, G. S., Monga, R., Chen, K., Devin, M., Le, Q. V., Mao, M. Z., Ranzato, M. A., Senior, A., Tucker, P., Yang, K., and Ng., A. Y. Large scale distributed deep networks. *Advances in Neural Information Processing Systems (NIPS)* (2012), 1–11.

[29] Défossez, A., and Bach, F. Adabatch: Efficient gradient aggregation rules for sequential and parallel stochastic gradient methods. *preprint, arXiv:1711.01761* (2017).

[30] Deng, L., Li, J., Huang, J.-T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., and Williams, J. Recent advances in deep learning for speech research at microsoft. *ICASSP 2013* (2013).

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

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

[33] Dippon, J. Accelerated randomized stochastic optimization. *Ann. Statist. 31*, 4 (2003), 1260–1281.

[34] Dippon, J., and Renz, J. Weighted means in stochastic approximation of minima. *SIAM J. Control Optim. 35*, 5 (1997), 1811–1827.

[35] Dozat, T. Incorporating nesterov momentum into adam. *ICLR Workshop* (2016), 2013–2016.

[36] Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research 12* (2011), 2121–2159.

[37] Duflo, M. *Algorithmes stochastiques*, vol. 23 of *Mathématiques & Applications (Berlin)*. Springer-Verlag, Berlin, 1996.

[38] Duflo, M. *Random iterative models*, vol. 34 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1997. Translated from the 1990 French original by Stephen S. Wilson and revised by the author.

[39] Fabian, V. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics 39*, 4 (1968), 1327–1332.

[40] Gaposhkin, V. F., and Krasulina, T. P. On the law of the iterated logarithm in stochastic approximation processes. *Theory Prob. Appl. 19*, 4 (1974), 844–850.

[41] Gerencsér, L. Convergence rate of moments in stochastic approximation with simultaneous perturbation gradient approximation and resetting. *IEEE Trans. on Automatic Control* (1999), 894–905.

[42] Graves, A. Generating sequences with recurrent neural networks. *preprint, arXiv:1308.0850* (2013).

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

[44] 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.

[45] Hinton, G. E., and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science 313*, 5786 (2006), 504–507.

[46] Inoue, M., Park, H., and Okada, M. On-line learning theory of soft committee machines with correlated hidden units steepest gradient descent and natural gradient descent. *Journal of the Physical Society of Japan 72*, 4 (2003), 805–810.

[47] Jentzen, A., and Pusnik, P. Exponential moments for numerical approximations of stochastic partial differential equations. *arXiv:1609.07031* (2016), 44 pages.

[48] Kiefer, J., and Wolfowitz, J. Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist. 23*, 3 (09 1952), 462–466.

[49] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980* (2014), 15 pages.

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

[51] Komlós, J., and Révész, P. On the rate of convergence of the Robbins-Monro method. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 25* (1972/73), 39–47.

[52] Konda, V. R., and Tsitsiklis, J. N. Convergence rate of linear two-time-scale stochastic approximation. *Ann. Appl. Probab. 14*, 2 (2004), 796–819.

[53] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* (2012), 1097–1105.

[54] Kushner, H. J., and Clark, D. S. *Stochastic approximation methods for constrained and unconstrained systems*, vol. 26 of *Applied Mathematical Sciences*. Springer-Verlag, New York-Berlin, 1978.

[55] Kushner, H. J., and Huang, H. Rates of convergence for stochastic approximation type algorithms. *SIAM J. Control Optim. 17*, 5 (1979), 607–617.

[56] Kushner, H. J., and Yang, J. Stochastic approximation with averaging of the iterates: optimal asymptotic rate of convergence for general processes. *SIAM J. Control Optim. 31*, 4 (1993), 1045–1062.

[57] Kushner, H. J., and Yin, G. G. *Stochastic approximation and recursive algorithms and applications*, second ed., vol. 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 2003. Stochastic Modelling and Applied Probability.

[58] Lai, T. L. Stochastic approximation. *Ann. Statist. 31*, 2 (2003), 391–406. Dedicated to the memory of Herbert E. Robbins.

[59] Lai, T. L., and Robbins, H. Limit theorems for weighted sums and stochastic approximation processes. *Proc. Nat. Acad. Sci. U.S.A. 75* (1978).

[60] Langford, J., Li, L., and Zhang, T. Sparse online learning via truncated gradient. *Journal of Machine Learning Research 10* (2009), 777–801.

[61] Le Breton, A., and Novikov, A. Averaging for estimating covariances in stochastic approximation. *Math. Methods Statist. 3*, 3 (1994), 244–266.

[62] Le Breton, A., and Novikov, A. Some results about averaging in stochastic approximation. *Metrika 42*, 3–4 (1995), 153–171.

[63] Le Roux, N., and Fitzgibbon, A. W. A fast natural newton method. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (2010), 623–630.

[64] Le Roux, N., Manzagol, P.-A., and Bengio, Y. Topmoumoute online natural gradient algorithm. *Advances in Neural Information Processing Systems (NIPS) 20* (2008), 849–856.

[65] Le Roux, N., Schmidt, M., and Bach, F. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. *preprint, hal-00674995v3* (2012).

[66] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324.

[67] LeCun, Y., Bottou, L., Orr, G., and Muller, K. Efficient backprop. *In Orr, G. and K., Muller (eds.), Neural Networks: Tricks of the trade. Springer* (1998), 9–50.

[68] L'Ecuyer, P., and Yin, G. Budget-dependent convergence rate of stochastic approximation. *SIAM J. Optim. 8*, 1 (1998), 217–247.

[69] Ljung, L., Pflug, G., and Walk, H. *Stochastic approximation and optimization of random systems*, vol. 17 of *DMV Seminar*. Birkhäuser Verlag, Basel, 1992.

[70] Martens, J, Sutskever, I., and Swersky, K. Estimating the hessian by back-propagating curvature. *preprint, arXiv:1206.6464* (2012).

[71] Martens, J. Deep learning via hessian-free optimization. *ICML* (2010), 735–742.

[72] Mcmahan, H. B., and Streeter, M. Delay-tolerant algorithms for asynchronous distributed online learning. *Advances in Neural Information Processing Systems (Proceedings of NIPS)* (2014), 1–9.

[73] Mizutani, E., and Dreyfus, S. An analysis on negative curvature induced by singularity in multi-layer neural-network learning. *Advances in Neural Information Processing Systems* (2010), 1669–1677.

[74] Mokkadem, A., and Pelletier, M. A generalization of the averaging procedure: the use of two-time-scale algorithms. *SIAM J. Control Optim. 49*, 4 (2011), 1523–1543.

[75] Neelakantan, A., Vilnis, L., Le, Q. V., Sutskever, I., Kaiser, L., Kurach, K., and Martens, J. Adding gradient noise improves learning for very deep networks. *preprint, arXiv:1511.06807* (2015).

[76] Nemirovski, A., Juditski, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Control and Optimization 19*, 4 (2009), 1574–1609.

[77] Nesterov, Y. A method for unconstrained convex minimization problem with the rate of convergence o(1/k2). *Doklady ANSSSR (translated as Soviet.Math.Docl.) 269* (1983), 543–547.

[78] Nevel'son, M. B. On properties of moments of stochastic approximation procedures. *Theory of Probability and its Applications 30*, 2 (1986), 407–413.

[79] Niu, F., Recht, B., Christopher, R., and Wright, S. J. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *preprint, arXiv:1106.5730* (2011).

[80] Pascanu, R., and Bengio, Y. Revisiting natural gradient for deep networks. *International Conference on Learning Representations* (2014).

[81] Pelletier, M. On the almost sure asymptotic behaviour of stochastic algorithms. *Stochastic Process. Appl. 78*, 2 (1998), 217–244.

[82] Pelletier, M. Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing. *Ann. Appl. Probab. 8*, 1 (1998), 10–44.

[83] Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing 1* (2014), 1532–1543.

[84] Pillaud-Vivien, L., Rudi, A., and Bach, F. Exponential convergence of testing error for stochastic gradient methods. *preprint, hal-01662278* (2017).

[85] Polyak, B. T. A new method of stochastic approximation type. *Avtomat. i Telemekh. 51*, 7 (1998), 937–1008.

[86] Polyak, B. T., and Juditsky, A. B. Acceleration of stochastic approximation by averaging. *Automation and Remote Control 30*, 4 (1992), 838–855.

[87] Polyak, B. T., and Tsypkin, Y. Z. Optimal pseudogradient adaptation algorithms. *Avtomat. i Telemekh. 8* (1980), 74–84.

[88] Qian, N. On the momentum term in gradient descent learning algorithms. *Neural networks : the official journal of the International Neural Network Society 12*, 1 (1999), 145–151.

[89] Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. *preprint, arXiv:1109.5647* (2012).

[90] Rattray, M., Saad, D., and Amari, S. I. Natural gradient descent for on-line learning. *Physical Review Letters 81*, 24 (1998), 5461–5464.

[91] Robbins, H., and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics 22*, 3 (1951), 400–407.

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

[93] Ruppert, D. Almost sure approximations to the robbins-monro and kiefer-wolfowitz processes with dependent noise. *Ann. Probab. 20* (1982).

[94] Ruppert, D. Efficient estimations from a slowly convergent robbins-monro process. *Technical Report 781, Cornell University Operations Research and Industrial Engineering* (1988).

[95] Ruppert, D. Stochastic approximation. *Handbook of sequential analysis, volume 118 of Statist. Textbooks Monogr., Dekker, New York* (1991), 503–529.

[96] Schaul, T., Zhang, S., and LeCun, Y. Generating sequences with recurrent neural networks. *preprint, arXiv:1206.1106* (2012).

[97] Schmetterer, L. Stochastic approximation. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (Berkeley, Calif., 1961), University of California Press, pp. 587–609.

[98] Schraudolph, N. N. Local gain adaptation in stochastic gradient descent. *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470) 2* (1999), 569–574.

[99] Schraudolph, N. N. Fast curvature matrix-vector products for second-order gradient descent. *Neural computation 14*, 7 (2002), 1723–1738.

[100] Schraudolph, N. N., Yu, J., and Günter, S. A stochastic quasi-newton method for online convex optimization. *Proceedings of the 9th International Conference on Artificial Intelligence and Statistics (AISTAT)* (2007), 433–440.

[101] Shalev-Shwartz, S., Shinger, Y., and Nathan Srebro, A. C. Pegasos: primal estimated sub-gradient solver for svm. *Mathematical Programming 127*, 1 (2011), 3–30.

[102] Shalev-Shwartz, S., and Tewari, A. Stochastic methods for l1 regularized loss minimization. *Proceedings of the 26st International Conference on Machine Learning (ICML)* (2009).

[103] Sohl-Dickstein, J., Poole, B., and Ganguli, S. Fast large-scale optimization by unifying stochastic gradient and quasi-newton methods. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (2014), 604–612.

[104] Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* (2013), 1139–1147.

[105] Sutton, R. S. Two problems with backpropagation and other steepest-descent learning procedures for networks. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society, Hillsdale, NJ: Erlbaum* (1986).

[106] 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.

[107] Vinyals, O., and Povey, D. Krylov subspace descent for deep learning. *AISTATS* (2012).

[108] Xu, W. Towards optimal one pass large scale learning with averaged stochastic gradient descent. *preprint, arXiv:1301.3584* (2011).

[109] Zeiler, M. D. Adadelta: An adaptive learning rate method. *preprint, arXiv:1212.5701* (2012).

[110] Zhang, S., Choromanska, A., and LeCun, Y. Deep learning with elastic averaging SGD. *Neural Information Processing Systems Conference (NIPS 2015)* (2015), 1–24.

[111] Zhang, T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. *Proceedings of the 21st International Conference on Machine Learning (ICML)* (2004).