ON THE ASYMPTOTIC BEHAVIOUR OF STOCHASTIC PROCESSES, WITH APPLICATIONS TO SUPERMARTINGALE CONVERGENCE, DVORETZKY'S APPROXIMATION THEOREM, AND STOCHASTIC QUASI-FEJÉR MONOTONICITY

MORENIKEJI NERI[a], NICHOLAS PISCHKE[a], THOMAS POWELL[b]

[a] Department of Mathematics, Technische Universität Darmstadt,
Schlossgartenstraße 7, 64289 Darmstadt, Germany,
[b] Department of Computer Science, University of Bath,
Claverton Down, Bath, BA2 7AY, United Kingdom,
E-mails: {neri,pischke}@mathematik.tu-darmstadt.de, trjp20@bath.ac.uk

Abstract. We prove a novel and general result on the asymptotic behavior of stochastic processes which conform to a certain relaxed supermartingale condition. Our result provides quantitative information in the form of an explicit and effective construction of a rate of convergence for this process, both in mean and almost surely, that is moreover highly uniform in the sense that it only depends on very few data of the surrounding objects involved in the iteration. We then apply this result to derive new quantitative versions of well-known concepts and theorems from stochastic approximation, in particular providing effective rates for a variant of the Robbins-Siegmund theorem, Dvoretzky's convergence theorem, as well as the convergence of stochastic quasi-Fejér monotone sequences, the latter of which formulated in a novel and highly general metric context. We utilize the classic and widely studied Robbins-Monro procedure as a template to evaluate our quantitative results and their applicability in greater detail. We conclude by illustrating the breadth of potential further applications with a brief discussion on a variety of other well-known iterative procedures from stochastic approximation, covering a range of different applied scenarios to which our methods can be immediately applied. Throughout, we isolate and discuss special cases of our results which even allow for the construction of fast, and in particular linear, rates.

Keywords: Stochastic processes, rates of convergence, stochastic approximation, proof mining
MSC2020 Classification: 62L20, 90C15, 60G42, 03F10

1. Introduction

Stochastic approximation methods are fundamental to modern science, with applications in optimization, statistics, machine learning, control theory, and many other areas. Accordingly, a rich mathematical theory on stochastic approximation has been developed, some of the most useful results being general convergence theorems that can be used to guarantee good asymptotic behaviour for large classes of approximation methods.

Several of the most important abstract convergence results for stochastic approximation utilise *supermartingale convergence* (as embodied by the famous Robbins-Siegmund theorem [86]), or similar abstract descent properties (as in Dvoretzky's convergence theorem [32]), along with the closely related phenomenon of *stochastic quasi-Fejér monotonicity* [26, 34]. Each of these represents, in a slightly different guise, a formal perspective on the strategy for proving convergence that exploits the fact that for many stochastic algorithms, the distance of the iterates from a target point decreases in some way, where here both the notion of *distance* and *decreasing* are rather general and can take various forms. The power and broad applicability of supermartingale convergence and quasi-Fejér monotonicity has been recently surveyed in

Date: April 17, 2025.

[41], and there exists a large body of literature where these methods are applied in concrete scenarios, either explicitly or (even more commonly) implicitly.

Aside from their general utility, abstract convergence theorems take on additional significance in that they give insight into the relationship between different stochastic algorithms and can therefore guide the development of new methods. Even more interestingly, *strengthenings* of abstract theorems with, for example, quantitative information, propagate down to the many concrete methods whose convergence has been established using those theorems, providing new insights and results for both well-known and newly discovered stochastic algorithms.

This paper is concerned with abstract convergence results of this type for stochastic algorithms, and in particular their quantitative aspects. Concretely, we develop new abstract convergence theorems and strengthen existing ones, before demonstrating the applicability of these results across the field of stochastic optimization.

Our contributions fall into three main parts. We first provide a novel and very general convergence theorem (Theorem 2.8) under which each of the aforementioned strategies (that is supermartingale convergence, Dvoretzky's convergence theorem, and stochastic quasi-Fejér monotonicty) is subsumed. Most importantly, we equip our convergence theorem with concrete *convergence rates* which can be effectively constructed in terms of a modest amount of surrounding quantitative data, which can essentially always be supplied in practice. We then show that these rates are inherited by all three of the aforementioned strategies for analysing stochastic methods, where we formulate, in turn, quantitative variants of the Robbins-Siegmund theorem (Theorem 3.4) and Dvoretzky's convergence theorem (Theorem 4.2), along with a quantitative convergence result for stochastic quasi-Fejér monotone sequences in metric spaces (Theorem 5.8), where the latter is even qualitatively new. Finally, we present an array of case studies where explicit convergence rates are obtained for concrete stochastic algorithms by utilizing these results. In many cases, these rates are provided for the first time, or at a higher level of generality than before.

Our overall aim is to provide a general methodology for obtaining quantitative convergence guarantees for stochastic methods, a methodology that is broadly applicable whenever convergence can be established using classic descent-based methods. Our intention is that this methodology can be applied to analyse specific algorithms as well as to produce new abstract quantitative convergence results for various classes of approximation methods, as illustrated through our study of the well-known Robbins-Monro procedure in Section 6 and the many further examples which are given in Section 7.

Convergence rates for general stochastic methods are of obvious practical importance given that such methods underlie numerous concrete algorithms (as a canonical example, the Robbins-Monro procedure is instantiated as stochastic gradient descent, which is in turn of fundamental importance in modern machine learning). Though the convergence rates attached to our abstract convergence theorems are necessarily general, we in particular highlight how our framework can be readily refined to produce *fast* (e.g. linear) rates in special cases. In this way, we demonstrate that our abstract approach is capable of generating quantitative information that is of clear practical use.

Finally, and as elaborated further in Section 1.3 below, the aforementioned results are obtained using the logic-based proof mining program [50, 51], where this paper represents one of the first applications of proof-theoretic techniques to stochastic methods. We therefore consider this implicit connection between proof theory and stochastic approximation to be a further novel contribution of the paper.

1.1. **Overview of the main results.** We now describe our main results in more detail, with a particular emphasis on our most general abstract convergence result, which is both qualitatively and quantitatively new. Concretely, at our most abstract, we consider general nonnegative real-valued stochastic processes (X_n) that conform to the almost-supermartingale property that

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \leqslant (1 + A_n)X_n + C_n$$

almost surely for all $n \in \mathbb{N}$, formulated relative to a filtration $(\mathcal{F}_n)$ of an underlying probability space. The general idea is that X_n represents the distance between the element x_n of a stochastic algorithm and some target point, with the perturbation $1+A_n$ and error term C_n arising in some way from stochastic noise. Together with natural conditions on the errors and the perturbations, we provide a convergence result for such processes, both in expectation and almost surely, in the presence of an additional approximation assumption

$$\liminf_{n\to\infty} \mathbb{E}[f(X_n)] = 0$$

where f belongs to a certain class of well-behaved functions (containing in particular n-th roots and logarithms) that *slow down* the process X_n, and hence substantially weaken the stronger assumption that $\liminf_{n\to\infty} \mathbb{E}[X_n] = 0$, typically required to establish convergence. Our ability to slow down the process in this way is crucial for the applications that follow.

The proof of our general result is rather elementary, relying only on fundamental properties of martingales and conditional expectations (notably Ville's inequality for nonnegative supermatingales), and the corresponding quantitative convergence result not only guarantees the convergence to zero of $f(X_n)$ in mean and of X_n almost surely, but also provides rates in both cases in the form of explicitly constructed functions $\rho : (0, \infty) \to \mathbb{N}$ with

$$\forall \varepsilon > 0 \; \forall n \geqslant \rho(\varepsilon) \, (\mathbb{E}[f(X_n)] < \varepsilon)$$

and $\rho' : (0, \infty)^2 \to \mathbb{N}$ with

$$\forall \lambda, \varepsilon > 0 \, (\mathbb{P}(\exists n \geqslant \rho'(\lambda, \varepsilon)(X_n \geqslant \varepsilon)) < \lambda) \, ,$$

respectively.

The motivation behind our approximation assumption is the observation from practice that in many concrete scenarios, the above approximation property for the process $f(X_n)$ for a suitable "slowdown" function f naturally comes into existence through a decomposition into a similar property $\liminf_{n\to\infty} \mathbb{E}[V_n] = 0$ for a secondary process (V_n), which in many cases can be naturally and easily obtained, together with a type of regularity assumption relating $\mathbb{E}[V_n]$ to $\mathbb{E}[f(X_n)]$, which in this paper will be concretely represented through a modulus $\tau : (0, \infty) \to (0, \infty)$ satisfying

$$\forall n \in \mathbb{N} \; \forall \varepsilon > 0 \, (\mathbb{E}[V_n] < \tau(\varepsilon) \to \mathbb{E}[f(X_n)] < \varepsilon) \, .$$

As we will show, this modulus τ represents a powerful abstraction of many commonly made regularity or uniqueness assumptions from the literature, a range of concrete examples of which will be discussed in detail.

Going from the abstract to the concrete, we then illustrate that the three distinct (but conceptually overlapping) approaches to convergence as discussed above, that is supermartingale convergence in the form of the Robbins-Siegmund theorem, Dvoretzky's convergence theorem and convergence of stochastic quasi-Fejér monotone sequences, are all (in their respective quantitative variants) consequences of our abstract theorem, where the regularity moduli as detailed above naturally manifest themselves through commonly made assumptions in all these cases. This provides not only a uniform proof strategy for these three central approaches to stochastic

approximation, but furthermore yields rates in each case, phrased in terms of corresponding regularity moduli.

We then demonstrate the applicability of these concrete instantiations of our general result by considering a series of examples. We first consider classic stochastic approximation algorithms based on the Robbins-Monro scheme, where in particular our results manage to reproduce the well-known $O(1/\sqrt{n})$ bounds for strongly monotone operators under various commonly made assumptions in that case, but can also be used to formulate more abstract rates under much weaker assumptions.

We then conclude the paper by indicating further applications of our results in stochastic optimization by surveying

(1) a range of natural optimization and approximation scenarios together with commonly made assumptions on the objective functions that entail the existence of our general regularity moduli discussed above, including (quasi-) contractions, functions with sharp minima as well as uniformly accretive set-valued operators;
(2) a variety of stochastic methods commonly studied in the literature which are stochastically quasi-Fejér monotone and hence entail the almost-supermartingale condition discussed above, and where the other surrounding quantitative data can be naturally instantiated, including the Krasnoselskii-Mann iteration in general Hadamard spaces and a variant of the proximal point algorithm in Hilbert spaces (both with stochastic noise) along with random-order metric splitting algorithms.

1.2. **This paper as a foundation for future work.** The main purpose of our abstract quantitative convergence results is to develop a framework that can be applied to provide quantitative insights into stochastic algorithms across stochastic optimization and related fields. For that reason, already in this paper we provide a detailed account of how our results are applicable to both classic and modern methods of a wide variety (as will be discussed in Sections 6 and 7). However, the methods discussed here should merely be viewed as illustrative of our framework's potential, and a great deal of further work is anticipated, exploiting in particular the fact that we cover not one but three distinct convergence techniques that are of central importance in the literature, and that our approach is highly versatile and modular.

Our brief study of the Robbins-Monro scheme via the quantitative Robbins-Siegmund theorem in Section 6 could be readily extended to a more detailed quantitative study of gradient descent methods with more general conditions, as represented, for example, by the classic paper [13], or in the more abstract setting of Hilbert spaces as in [7]. A more recent example along these lines is the unified supermartingale convergence theorem of [63], which is tailored to establishing convergence for a range of stochastic methods, including varieties of stochastic gradient descent.

Indeed, we envisage that a particularly effective use of our results would be to provide abstract, quantitative versions of the various unified convergence theorems already known in the literature, such as the aforementioned result from [63], which are themselves often adaptations of the well-known convergence techniques covered in this paper. In this way, any rates of convergence immediately propagate down to the class of algorithms covered by those unified theorems. Another important example along these lines, this time connected to Dvoretzky's theorem, is the widely used convergence theorem presented in [45], originally conceived to provide a rigorous convergence proof for dynamic programming based learning algorithms such as Q-learning [98], and which continues to be used in this way to prove convergence of new reinforcement learning algorithms.

Several concrete directions for future applications of our quantitative approach to stochastic quasi-Fejér monotonicity beyond those provided in Section 7 also immediately present themselves. For example, our work can be utilized to provide quantitative results for stochastic splitting methods for monotone inclusions as in particular studied in [23, 27, 77, 88, 97] as well as for stochastic variants of the proximal point method as in particular studied in [4, 14], both of which have become staples in the applied literature.

Finally, we emphasise that the basic convergence strategies studied in this paper, though classically applied to algorithms in $\mathbb{R}^d$, continue to be applicable in more exotic settings. Two such examples are already surveyed as applications in this paper, both of which can be found in Section 7: Concretely, we discuss the classical Kransnoselskii-Mann iteration as well as a metric splitting algorithm in the setting of general Hadamard spaces, a class of spaces that beyond infinite dimensional Hilbert spaces also in particular encompasses the Hilbert ball and general Hadamard manifolds as well as more unique examples like the Billera-Holmes-Vogtmann tree which has fundamental applications in computational biology. We therefore anticipate that our results (or extensions thereof) will also be applicable to more such sophisticated geometric settings, such as those of e.g. [2, 17, 18], which are concerned with establishing the convergence of gradient descent-type algorithms on various classes of Riemannian manifolds using similar methods.

1.3. **This paper as part of the proof mining program.** As already mentioned, the results developed in this paper have been established using the logic-based methodology of *proof mining* [50, 51] (going back to G. Kreisel's program of *unwinding of proofs* [57, 58]). In particular, the present paper is part of a recent and ongoing effort by the authors to bring methods from proof mining to bear in probability theory and stochastic optimization for the first time. In particular, it builds on the recent work [73] by Neri and Pischke which provides the logical foundations for such an approach, the works [75, 76] by Neri and Powell which provided an initial quantitative approach to martingales as well as to the Robbins-Siegmund theorem from the perspective of proof mining, and the work [82] which was likewise concerned with the (fast) asymptotic behavior stochastic approximation procedures and underlying general principles for stochastic process. Other applications of proof mining and related logical perspectives to probability theory and the study of random processes can be found in [71, 72, 74], which are methodologically similar to the present paper, as well as [3, 5, 6]. Furthermore, the present results, in particular the parts focused on stochastic quasi-Fejér monotonicity and the corresponding derivation of rates through quantitative notions of regularity and uniqueness, are influenced and inspired by preceding work on abstract convergence results for quasi-Fejér monotone sequences as developed in [53, 54, 80] (see also [78]), similarly produced using methods from proof mining. As is common in proof mining, all the results and proofs given in this work are formulated in a way which avoids any reference to mathematical logic.

1.4. **Structure of the paper.** The paper is now organized as follows: In Section 2, we present our general theorem on rates of convergence for stochastic processes. In Section 3 – 5, we then apply these results to the Robbins-Siegmund theorem, Dvoretzky's convergence theorem, and to stochastic quasi-Fejér monotone sequences, respectively. These results are then gauged and compared in Section 6 against the backdrop of the well-known Robbins-Monro process and further, more novel, applications are sketched in Section 7.

2. A general theorem on rates of convergence for stochastic processes

We begin the paper with a very general convergence theorem for stochastic processes that satisfy an almost-supermartingale condition. Crucially, this theorem is not just concerned

with plain convergence, but in particular also provides a rate of convergence for the associated process, both in expectation and almost surely, as described in the introduction above. This convergence theorem will then be instantiated in the subsequent sections to yield various applications in stochastic approximation, all of which inherit the convergence rates formulated here. We also take the opportunity to provide some useful results concerning regularity moduli, which will be essential for the various applications of our main theorem.

For both this section and the remainder of the paper, we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and we understand all measure-theoretic and probabilistic notions, such as random variables, (conditional) expectations $\mathbb{E}$ and almost-sureness (which we abbreviate by a.s.), to be defined relative to it. All random variables, unless specified otherwise, are assume to be real-valued and (in-)equalities between random variables, if not specified otherwise, are understood to hold almost surely.

We now discuss the main ingredients required for our main theorem. We begin by recalling some fundamental results, the first of which is a well-known concentration inequality for nonnegative supermartingales due to Ville [96] (see also [67] for a detailed overview of concentration inequalities for martingales):

Lemma 2.1. *Let (U_n) be a nonnegative supermartingale. Then for any $a > 0$ we have*

$$\mathbb{P}\left(\sup_{n\in\mathbb{N}} U_n \geqslant a\right) \leqslant \frac{\mathbb{E}[U_0]}{a}.$$

Further, we will rely on Jensen's inequality for (conditional) expectations (the following formulation of which is taken from [48], see Theorems 7.9 and 8.20 therein):

Lemma 2.2. *Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$ and let X be a nonnegative integrable random variable. Let $\varphi : [0,\infty) \to [0,\infty)$ be a measurable function. If φ is convex, then*

$$\varphi(\mathbb{E}[X \mid \mathcal{G}]) \leqslant \mathbb{E}[\varphi(X) \mid \mathcal{G}] \text{ a.s. and } \varphi(\mathbb{E}[X]) \leqslant \mathbb{E}[\varphi(X)],$$

where if φ is concave, then

$$\varphi(\mathbb{E}[X \mid \mathcal{G}]) \geqslant \mathbb{E}[\varphi(X) \mid \mathcal{G}] \text{ a.s. and } \varphi(\mathbb{E}[X]) \geqslant \mathbb{E}[\varphi(X)].$$

Our convergence result and the construction of associated rates of convergence rests on a key descent condition of lim inf-type for the associated process, forcing it to have expectation below ε for each $\varepsilon > 0$ infinitely often. Crucially, this descent condition however is "slowed down" as already discussed in the introduction, i.e. in the sense that only the expectation of the process under a suitable given function f is expected to decrease asymptotically in this sense. This generality will be crucial for some of our applications that follow.

We now introduce the relevant class of admissible functions f. For that, we begin with the following general notion:

Definition 2.3. A function $f : [0,\infty) \to [0,\infty)$ is called ψ-supermultiplicative for a function $\psi : [0,1] \to [0,1]$ if

$$f(xa) \geqslant f(x)\psi(a)$$

for all $x \in [0,\infty)$ and $a \in [0,1]$ and ψ satisfies $\psi(x) > 0$ for $x > 0$.

Definition 2.4. A function $f : [0,\infty) \to [0,\infty)$ is called continuous at $0 = f(0)$ with a modulus $\kappa : (0,\infty) \to (0,\infty)$ if

$$x < \kappa(\varepsilon) \text{ implies } f(x) < \varepsilon$$

for all $x \in [0,\infty)$ and $\varepsilon > 0$.

The general class of functions f that we in the following want to allow is then the following:

Definition 2.5. A function $f : [0,\infty) \to [0,\infty)$ is called s.i.c.c. (with moduli ψ and κ) if

(1) f is ψ-supermultiplicative for a function $\psi : [0,1] \to [0,1]$,
(2) f is strictly increasing,
(3) f is concave,
(4) f is continuous, and $\kappa : (0,\infty) \to (0,\infty)$ is a modulus of continuity at $0 = f(0)$.

Before we discuss the main theorem, we shortly give some illuminating examples for typical s.i.c.c. functions and discuss some closure properties of that class.

Example 2.6. (1) The function x^q for $q \in (0,1]$ is s.i.c.c. with moduli a^q and $\sqrt[q]{\varepsilon}$. In particular, $\sqrt{x}$ and x are s.i.c.c. To see this, note that x^q is clearly increasing and concave. Further, since

$$(xa)^q = x^q a^q$$

for any $x, a \geqslant 0$, we get that x^q is a^q-supermultiplicative. Lastly, x^q is clearly continuous and since

$$x < \sqrt[q]{\varepsilon} \text{ implies } x^q < \varepsilon,$$

it immediately follows that $\sqrt[q]{\varepsilon}$ is a modulus of continuity for x^q at $0 = 0^q$.

(2) The function $\log_c(1+x)$ for $c > 1$ is s.i.c.c. with moduli a and $c^\varepsilon - 1$. To see this, note that $\log_c(1+x)$ is increasing and concave. Further, $\log_c(1+x)$ is clearly continuous and since

$$x < c^\varepsilon - 1 \text{ implies } \log_c(1+x) < \varepsilon,$$

the function $c^\varepsilon - 1$ is a modulus of continuity for $\log_c(1+x)$ at $0 = \log_c(1)$. Lastly, $\log_c(1+x)$ is also a-supermultiplicative since we have

$$(1+x)^a \leqslant 1 + xa$$

for all $x \geqslant 0$ and all $a \in [0,1]$, so that

$$a \log_c(1+x) \leqslant \log_c(1+xa)$$

for all such x, a.

The next proposition, which lists some closure properties of the class of s.i.c.c. functions, follows immediately from some simple calculations. We hence omit the proof.

Proposition 2.7. *Let f be s.i.c.c. with moduli ψ and κ and let g be s.i.c.c. with moduli ψ' and κ'. Then:*

(1) $\alpha f + \beta g$, given $\alpha, \beta > 0$, is s.i.c.c. with moduli $\min\{\psi, \psi'\}$ and $\min\{\kappa(\varepsilon/2\alpha), \kappa'(\varepsilon/2\beta)\}$.
(2) $f \circ g$ is s.i.c.c. with moduli $\psi \circ \psi'$ and $\kappa' \circ \kappa$.
(3) $\min\{f, g\}$ is s.i.c.c. with moduli $\min\{\psi, \psi'\}$ and $\max\{\kappa, \kappa'\}$.

We now state and prove the main theorem of this section:

Theorem 2.8. *Let $(\mathcal{F}_n)$ be a filtration of $\mathcal{F}$ and let (X_n), (A_n) and (C_n) be sequences of nonnegative, integrable real-valued random variables adapted to $(\mathcal{F}_n)$. Suppose that for all $n \in \mathbb{N}$:*

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \leqslant (1 + A_n)X_n + C_n \text{ a.s.}$$

Also, suppose that there exist $K \geqslant 1$ and $\chi : (0,\infty) \to \mathbb{N}$ satisfying

$$\prod_{i=0}^{\infty}(1 + A_i) < K \text{ a.s. and } \forall \varepsilon > 0 \left(\sum_{i=\chi(\varepsilon)}^{\infty} \mathbb{E}[C_i] < \varepsilon \right).$$

Further, let $f : [0,\infty) \to [0,\infty)$ *be s.i.c.c. with moduli* ψ *and* κ*. Finally, suppose that* φ *is a* $\liminf$*-modulus for* $(f(X_n))$ *in expectation in the sense that*

$$\forall \varepsilon > 0 \; \forall N \in \mathbb{N} \; \exists n \in [N; \varphi(\varepsilon, N)] \, (\mathbb{E}[f(X_n)] < \varepsilon) \, .$$

Then $\mathbb{E}[f(X_n)] \to 0$ *with rate*

$$\rho(\varepsilon) := \varphi\left(\frac{\varepsilon\psi(K^{-1})}{2}, \chi\left(\kappa\left(\frac{\varepsilon\psi(K^{-1})}{2}\right)\right)\right)$$

and $X_n \to 0$ *a.s. with rate*

$$\rho'(\lambda, \varepsilon) := \rho(\lambda f(\varepsilon)).$$

Proof. For any $n \in \mathbb{N}$, define

$$U_n := \frac{X_n}{B_{n-1}} + \mathbb{E}\left[\sum_{i=n}^{\infty} \frac{C_i}{B_i} \mid \mathcal{F}_n\right] \text{ where } B_j := \prod_{i=0}^{j}(1 + A_i),$$

with the convention that $B_{-1} := 1$. The stochastic process (U_n) is a nonnegative supermartingale, since for any $n \in \mathbb{N}$ we have

$$\begin{aligned}
\mathbb{E}[U_{n+1} \mid \mathcal{F}_n] &= \mathbb{E}\left[\frac{X_{n+1}}{B_n} \mid \mathcal{F}_n\right] + \mathbb{E}\left[\mathbb{E}\left[\sum_{i=n+1}^{\infty} \frac{C_i}{B_i} \mid \mathcal{F}_{n+1}\right] \mid \mathcal{F}_n\right] \\
&= \frac{\mathbb{E}[X_{n+1} \mid \mathcal{F}_n]}{B_n} + \mathbb{E}\left[\sum_{i=n+1}^{\infty} \frac{C_i}{B_i} \mid \mathcal{F}_n\right] \\
&\leqslant \frac{X_n}{B_{n-1}} + \frac{C_n}{B_n} + \mathbb{E}\left[\sum_{i=n+1}^{\infty} \frac{C_i}{B_i} \mid \mathcal{F}_n\right] \\
&= U_n,
\end{aligned}$$

using that $B_j \geqslant 1$ for every j. Thereby $(f(U_n))$ is also a nonnegative supermartingale since for an arbitrary $n \in \mathbb{N}$, we have

$$\mathbb{E}[f(U_{n+1}) \mid \mathcal{F}_n] \leqslant f(\mathbb{E}[U_{n+1} \mid \mathcal{F}_n]) \leqslant f(U_n)$$

using Jensen's inequality (and that f is continuous as well as concave) as well as that f is monotone. Further, for any $n \in \mathbb{N}$, we get

$$f(U_n) \leqslant f\left(\frac{X_n}{B_{n-1}}\right) + f\left(\mathbb{E}\left[\sum_{i=n}^{\infty} \frac{C_i}{B_i} \mid \mathcal{F}_n\right]\right) \leqslant f(X_n) + f\left(\mathbb{E}\left[\sum_{i=n}^{\infty} C_i \mid \mathcal{F}_n\right]\right)$$

using that f is subadditive (since f is concave) and that f is monotone together with $B_j \geqslant 1$ for every j. In particular, using Jensen's inequality (and so the concavity and continuity of f) again, we have

$$(+) \qquad \mathbb{E}[f(U_n)] \leqslant \mathbb{E}[f(X_n)] + f\left(\mathbb{E}\left[\sum_{i=n}^{\infty} C_i\right]\right).$$

Now, let $\varepsilon > 0$ and choose

$$n(\varepsilon) \in \left[\chi\left(\kappa\left(\frac{\varepsilon\psi(K^{-1})}{2}\right)\right); \varphi\left(\frac{\varepsilon\psi(K^{-1})}{2}, \chi\left(\kappa\left(\frac{\varepsilon\psi(K^{-1})}{2}\right)\right)\right)\right]$$

such that $\mathbb{E}[f(X_{n(\varepsilon)})] < \varepsilon\psi(K^{-1})/2$. Let $m \geqslant n(\varepsilon)$ be arbitrary. Then

$$\begin{aligned}\mathbb{E}[f(U_m)] &\leqslant \mathbb{E}[f(U_{n(\varepsilon)})] \\ &\leqslant \mathbb{E}[f(X_{n(\varepsilon)})] + f\left(\mathbb{E}\left[\sum_{i=n(\varepsilon)}^{\infty} C_i\right]\right) \\ &\leqslant \varepsilon\psi(K^{-1})\end{aligned}$$

using that $(f(U_n))$ is a supermartingale as well as $(+)$ and that f is monotone and continuous at 0 together with the defining property of χ, which implies

$$\mathbb{E}\left[\sum_{i=n(\varepsilon)}^{\infty} C_i\right] < \kappa\left(\frac{\varepsilon\psi(K^{-1})}{2}\right).$$

Finally, we get that

$$\psi(K^{-1})f(X_m) \leqslant f\left(\frac{X_m}{K}\right) < f\left(\frac{X_m}{B_{m-1}}\right) \leqslant f(U_m)$$

as $B_{m-1} < K$, using also the monotonicity and ψ-supermultiplicativity of f, and after taking expectations we get

$$\psi(K^{-1})\mathbb{E}[f(X_m)] \leqslant \mathbb{E}[f(U_m)] < \varepsilon\psi(K^{-1})$$

and so $\mathbb{E}[f(X_m)] < \varepsilon$. As m was arbitrary, this yields that ρ is a rate of convergence for $\mathbb{E}[f(X_n)] \to 0$. For the rate that $X_n \to 0$ a.s., note that

$$\begin{aligned}\mathbb{P}(\exists m \geqslant n(\varepsilon)(X_m \geqslant a)) &\leqslant \mathbb{P}(\exists m \geqslant n(\varepsilon)(U_m \geqslant a/K)) \\ &\leqslant \mathbb{P}(\exists m \geqslant n(\varepsilon)(f(U_m) \geqslant f(a/K))) \\ &\leqslant \mathbb{P}(\exists m \geqslant n(\varepsilon)(f(U_m) \geqslant f(a)\psi(K^{-1}))) \\ &\leqslant \frac{\mathbb{E}[f(U_{n(\varepsilon)})]}{f(a)\psi(K^{-1})}\end{aligned}$$

where the first inequality follows from the fact that

$$\frac{X_m}{K} < \frac{X_m}{B_{m-1}} \leqslant U_m$$

for all $m \in \mathbb{N}$ and the last inequality follows from Ville's inequality. This immediately implies that $X_n \to 0$ a.s. with rate ρ'. □

Disregarding the quantitative information in the above result, we in particular get the following qualitative theorem on the convergence of certain stochastic processes:

Corollary 2.9. Suppose that (X_n), (A_n) and (C_n) satisfy the supermartingale-type property of Theorem 2.8 and assume that $\prod_{i=0}^{\infty}(1 + A_i) < \infty$ a.s., that $\sum_{i=0}^{\infty}\mathbb{E}[C_i] < \infty$, and that $\liminf_{n\to\infty}\mathbb{E}[f(X_n)] = 0$ for some s.i.c.c. function f. Then $\mathbb{E}[f(X_n)] \to 0$ and $X_n \to 0$ a.s.

Remark 2.10. The conclusion of $\mathbb{E}[f(X_n)] \to 0$ in the above Theorem 2.8 is in general the best one can hope for when allowing for general s.i.c.c. functions f other than the identity. Concretely, already in the case of $f = \sqrt{\cdot}$, it can naturally not be guaranteed that we further also have $\mathbb{E}[X_n] \to 0$. For that, consider the following example: Let (Y_n) be a nonnegative i.i.d. stochastic process with $\mathbb{E}[Y_n] = 1$ but $\mathbb{E}[\sqrt{Y_n}] = \eta \in (0, 1)$ for all $n \in \mathbb{N}$ (arising e.g. naturally through a Bernoulli process with values 0 and 2, each with probability $\frac{1}{2}$). Defining $X_n := \prod_{k=0}^{n} Y_k$ as well as choosing $\mathcal{F}_n := \sigma(Y_0, \ldots, Y_n)$, i.e. $\mathcal{F}_n$ is the σ-algebra generated by

$Y_0, \dots, Y_n$, we immediately obtain that (X_n) is a martingale w.r.t. the filtration $(\mathcal{F}_n)$. As (Y_n) is independent, we have

$$\mathbb{E}[\sqrt{X_n}] = \prod_{k=0}^{n} \mathbb{E}[\sqrt{Y_k}] = \prod_{k=0}^{n} \eta \to 0.$$

In particular, the conditions of Theorem 2.8 are trivially satisfied for $A_n := C_n := 0$. However, for similar reasons we have $\mathbb{E}[X_n] = \prod_{k=0}^{n} \mathbb{E}[Y_k] = 1$ for any $n \in \mathbb{N}$, i.e. $(\mathbb{E}[X_n])$ does not converge to 0.

In many cases where we subsequently apply the above Theorem 2.8, the crucial lim inf-property for $(f(X_n))$ in expectation and its corresponding modulus arise through a composition of two other properties, quantitatively witnessed by corresponding moduli in an analogous way. Concretely, in many applications a lim inf-modulus φ for $(f(X_n))$ in expectation comes into existence through a lim inf-modulus φ' for an auxiliary sequence (V_n) in expectation, i.e.

$$\forall \varepsilon > 0 \; \forall N \in \mathbb{N} \; \exists n \in [N; \varphi'(\varepsilon, N)] \, (\mathbb{E}[V_n] < \varepsilon) \, ,$$

together with a type of regularity modulus $\tau : (0, \infty) \to (0, \infty)$ connecting (V_n) and $f(X_n)$ in expectation, by which we here concretely mean that

$$\forall n \in \mathbb{N} \; \forall \varepsilon > 0 \, (\mathbb{E}[V_n] < \tau(\varepsilon) \to \mathbb{E}[f(X_n)] < \varepsilon) \, .$$

In the presence of these two moduli, we immediately get that

$$\varphi(\varepsilon) := \varphi'(\tau(\varepsilon), N)$$

is a lim inf-modulus for $(f(X_n))$ in expectation.

We later discuss various concrete situations where such moduli naturally arise. For now, we set out some abstract conditions under which a regularity modulus τ as above exists and can be determined.

Proposition 2.11. *Let I be a non-empty index set and let $(X_i)_{i \in I}$, $(V_i)_{i \in I}$ be two families of nonnegative real-valued random variables with*

$$V_i \geqslant \tau(X_i) \text{ a.s.}$$

for all $i \in I$, where $\tau : [0, \infty) \to [0, \infty)$ is a convex and strictly increasing function. Assume that each X_i is integrable. Then τ satisfies

$$\forall i \in I \; \forall \varepsilon > 0 \, (\mathbb{E}[V_i] < \tau(\varepsilon) \to \mathbb{E}[X_i] < \varepsilon) \, .$$

Proof. Take $\varepsilon > 0$ and $i \in I$ with $\mathbb{E}[V_i] < \tau(\varepsilon)$. Thus $\mathbb{E}[\tau(X_i)] < \tau(\varepsilon)$ by the above inequality. Using Jensen's inequality, as τ is convex, we have

$$\tau(\mathbb{E}[X_i]) \leqslant \mathbb{E}[\tau(X_i)] < \tau(\varepsilon).$$

As τ is increasing, we have $\mathbb{E}[X_i] < \varepsilon$. □

Remark 2.12. The above condition that $V_i \geqslant \tau(X_i)$ a.s. for all $i \in I$ can, for a strictly increasing τ, be equivalently recognized as requiring that τ satisfies

$$\forall \varepsilon > 0 \, (V_i < \tau(\varepsilon) \to X_i < \varepsilon) \text{ a.s.}$$

for all $i \in I$ and is hence equivalent to a pointwise version of the regularity property in expectation. Clearly $V_i \geqslant \tau(X_i)$ a.s. implies the above property as τ is strictly increasing and for the converse, note that $X_i \geqslant X_i$ a.s. and so $V_i \geqslant \tau(X_i)$ a.s. follows by the above property.

The previous proposition can be generalized to avoid the assumption of convexity of the modulus τ and allow for a more general pointwise premise, at the expense of a uniform integrability assumption.

Proposition 2.13. *Let I be a non-empty index set and let $(X_i)_{i\in I}$, $(V_i)_{i\in I}$ be two families of nonnegative real-valued random variables with*

$$\forall \varepsilon, l > 0 \, (X_i \in [\varepsilon, l] \to V_i \geqslant \pi(\varepsilon, l)) \;\; a.s.$$

for all $i \in I$ and where $\pi : (0,\infty)^2 \to (0,\infty)$. Let $K \geqslant \mathbb{E}[X_i]$ for all $i \in I$ and let $(X_i)_{i\in I}$ be uniformly integrable with modulus $\mu : (0,\infty) \to (0,\infty)$, i.e.

$$\forall A \in \mathcal{F} \; \forall \varepsilon > 0 \; \forall i \in I \, (\mathbb{P}(A) \leqslant \mu(\varepsilon) \to \mathbb{E}[X_i \mathbf{1}_A] \leqslant \varepsilon) \, .$$

Then we have

$$\forall i \in I \; \forall \varepsilon > 0 \, (\mathbb{E}[V_i] < \tau(\varepsilon) \to \mathbb{E}[X_i] < \varepsilon)$$

for τ defined by

$$\tau(\varepsilon) := \pi\left(\frac{\varepsilon}{4}, \frac{K}{\mu(\varepsilon/4)}\right) \frac{\mu(\varepsilon/4)}{2}.$$

Proof. Let $i \in I$ be given and assume that $\mathbb{E}[X_i] > 0$. For any $L, \delta > 0$ we have

$$\begin{aligned}
\mathbb{E}[X_i] &= \mathbb{E}[X_i \mathbf{1}_{X_i \geqslant \delta}] + \mathbb{E}[X_i \mathbf{1}_{X_i < \delta}] \\
&\leqslant \mathbb{E}[X_i \mathbf{1}_{X_i \geqslant \delta}] + \delta \\
&\leqslant \mathbb{E}[X_i \mathbf{1}_{\delta \leqslant X_i \leqslant L}] + \mathbb{E}[X_i \mathbf{1}_{X_i \geqslant L}] + \delta \\
&\leqslant L\mathbb{P}(\delta \leqslant X_i \leqslant L) + \mathbb{E}[X_i \mathbf{1}_{X_i \geqslant L}] + \delta.
\end{aligned}$$

Setting $L_i^\delta := \mathbb{E}[X_i]/\mu(\delta)$ and using Markov's inequality, we have $\mathbb{P}(X_i \geqslant L_i^\delta) \leqslant \mu(\delta)$. Therefore we get $\mathbb{E}[X_i \mathbf{1}_{X_i \geqslant L_i^\delta}] < \delta$ and so

$$\mathbb{E}[X_i] \leqslant L_i^\delta \mathbb{P}(\delta \leqslant X_i \leqslant L_i^\delta) + 2\delta$$

for any $\delta > 0$. In particular, we have shown that

$$\mathbb{P}\left(\delta \leqslant X_i \leqslant \frac{K}{\mu(\delta)}\right) \geqslant \mathbb{P}\left(\delta \leqslant X_i \leqslant L_i^\delta\right) \geqslant \frac{\mathbb{E}[X_i] - 2\delta}{L_i^\delta} = \mu(\delta)\left(1 - \frac{2\delta}{\mathbb{E}[X_i]}\right).$$

Now, let $\varepsilon > 0$ be given and assume that $\mathbb{E}[X_i] \geqslant \varepsilon$. Setting $\delta := \varepsilon/4$ in the above, we obtain

$$\mathbb{P}\left(\frac{\varepsilon}{4} \leqslant X_i \leqslant \frac{K}{\mu(\varepsilon/4)}\right) \geqslant \mu(\varepsilon/4)\left(1 - \frac{\varepsilon}{2\mathbb{E}[X_i]}\right) \geqslant \frac{\mu(\varepsilon/4)}{2}.$$

This in particular implies that

$$\begin{aligned}
\mathbb{E}[V_i] &\geqslant \mathbb{E}[V_i \mathbf{1}_{\varepsilon/4 \leqslant X_i \leqslant K/\mu(\varepsilon/4)}] \\
&\geqslant \pi\left(\frac{\varepsilon}{4}, \frac{K}{\mu(\varepsilon/4)}\right) \mathbb{P}\left(\frac{\varepsilon}{4} \leqslant X_i \leqslant \frac{K}{\mu(\varepsilon/4)}\right) \\
&\geqslant \pi\left(\frac{\varepsilon}{4}, \frac{K}{\mu(\varepsilon/4)}\right) \frac{\mu(\varepsilon/4)}{2}
\end{aligned}$$

using the properties of π. □

Remark 2.14. If $(X_i)_{i\in I}$ is even almost surely bounded in the above Proposition 2.13, say with $K \geqslant X_i$ a.s. for all $i \in I$, then the uniform integrability assumption can be omitted and we can simply defined τ via

$$\tau(\varepsilon) := \pi(\varepsilon/2, K)\frac{\varepsilon}{2K}.$$

To see this, note that by the reverse Markov inequality we have

$$\mathbb{P}\left(\frac{\varepsilon}{2} \leqslant X_i \leqslant K\right) = \mathbb{P}\left(\frac{\varepsilon}{2} \leqslant X_i\right) \geqslant \frac{\mathbb{E}[X_i] - \varepsilon/2}{K - \varepsilon/2} \geqslant \frac{\varepsilon}{2K}$$

for $\mathbb{E}[X_i] \geqslant \varepsilon$ and therefore

$$\mathbb{E}[V_i] = \mathbb{E}[V_i \mathbf{1}_{\varepsilon/2 \leqslant X_i} \leqslant K] \geqslant \pi(\varepsilon/2, K)\mathbb{P}\left(\frac{\varepsilon}{2} \leqslant X_i \leqslant K\right) \geqslant \pi(\varepsilon/2, K)\frac{\varepsilon}{2K}.$$

3. Applications to the Robbins-Siegmund theorem

The first application of our general result that we discuss is concerned with the seminal theorem of Robbins-Siegmund on supermartingale convergence:

Theorem 3.1 ([86])**.** *Let (X_n), (A_n), (B_n) and (C_n) be sequences of nonnegative integrable real-valued random variables adapted to a filtration $(\mathcal{F}_n)$. Assume that $\sum_{i=0}^{\infty} A_i < \infty$ and $\sum_{i=0}^{\infty} C_i < \infty$ a.s. and*

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \leqslant (1 + A_n)X_n - B_n + C_n$$

a.s. for all $n \in \mathbb{N}$. Then (X_n) converges a.s. and $\sum_{i=0}^{\infty} B_i < \infty$ a.s.

As anticipated by the range of applications presented in the original work [86] by Robbins and Siegmund, this result has become a fundamental tool for establishing the convergence of stochastic approximation algorithms, and we refer to the recent survey paper [41] for a broad overview of the many applications up to 2022.

Our starting point here is the recent work [75] by the first and third author, where a quantitative variant of the "full" Robbins-Siegmund theorem is given, deriving so-called rates of uniform learnability (which are in turn closely related to fluctuation or oscillation bounds) for the convergence in the conclusion under very general conditions. These rates of uniform learnability represent a quantitative formulation of convergence which is computationally weaker than direct rates of convergence, and computable rates of convergence for the Robbins-Siegmund theorem in the sense of the present paper are in general not possible[1].

However, in many applied cases where the Robbins-Siegmund theorem is utilized, one is actually in the situation that $B_n := u_n V_n$ for (u_n) a sequence of nonnegative control parameters with divergent sum, where moreover the circumstances of the problem at hand allow one to derive a modulus of regularity connecting the process V_n with the process X_n, in the style of the preceding Section 2. As the particular form of the relaxed supermartingale condition in the Robbins-Siegmund theorem additionally guarantees that $\liminf_{n\to\infty} \mathbb{E}[V_n] = 0$ a.s. with a corresponding modulus, we can apply Theorem 2.8 in these situations to derive the convergence of X_n in mean and almost surely, together with explicit rates of convergence. This is in particular the case in the context of the first four applications in the original paper of Robbins and Siegmund [86], including their well-known proof and generalization of Dvoretzky's theorem, along with the Robbins-Monro procedure, both of which will be discussed in more detail below.

To obtain our quantitative version of the Robbins-Siegmund theorem in the above circumstances, we first require two preliminary quantitative results. The first is a quantitative version of slight strengthening of a lemma of Qihou [83], which itself is a non-stochastic analogue of the Robbins-Siegmund theorem.

[1]In more detail, this limitation arises as the Robbins-Siegmund theorem in particular allows one to derive the convergence of nonnegative supermartingales, and so in particular of monotone sequences of real numbers, in which context well-known results on the computability theory of real analysis due to Specker [93] rule out general computable rates of convergence. See [75] for a more detailed discussion of this point.

Lemma 3.2. *Let (x_n), (α_n), (β_n) and (γ_n) be sequences of nonnegative reals with*

$$x_{n+1} \leqslant (1+\alpha_n)x_n - \beta_n + \gamma_n$$

for all $n \in \mathbb{N}$. If $\prod_{i=0}^{\infty}(1+\alpha_i) < \infty$ and $\sum_{i=0}^{\infty}\gamma_i < \infty$, then (x_n) converges and $\sum_{i=0}^{\infty}\beta_i < \infty$.

Further, if $K, L, M > 0$ satisfy $x_0 < K$, $\prod_{i=0}^{\infty}(1+\alpha_i) < L$ and $\sum_{i=0}^{\infty}\gamma_i < M$, then $\sum_{i=0}^{\infty}\beta_i < L(K+M)$.

The first part of the above result appears e.g. as Lemma 5.31 in [8] and the bound on the series $\sum_{i=0}^{\infty}\beta_i$ appears as Theorem 3.2 in [75] (together with a quantitative result on the convergence of the sequence (x_n) which, in the absence of the terms (β_n), was already discussed in [52]).

Second, we also require the following result on the asymptotic behavior of the summands of certain series.

Lemma 3.3. *Suppose that (u_n), (v_n) are sequences of nonnegative reals with*

$$\sum_{n=0}^{\infty} u_n v_n < L \quad \text{and} \quad \forall b > 0\ \forall k \in \mathbb{N}\left(\sum_{n=k}^{\theta(k,b)} u_n \geqslant b\right)$$

for $L > 0$ and $\theta : \mathbb{N} \times (0,\infty) \to \mathbb{N}$. Then $\liminf_{n\to\infty} v_n = 0$ with

$$\forall \varepsilon > 0\ \forall N \in \mathbb{N}\ \exists n \in [N; \theta(N, L/\varepsilon)](v_n < \varepsilon).$$

Proof. Fix $\varepsilon > 0$ and $N \in \mathbb{N}$. If $v_n \geqslant \varepsilon$ for all $n \in [N; \theta(N, L/\varepsilon)]$, we would have

$$L \leqslant \varepsilon \sum_{n=N}^{\theta(N,L/\varepsilon)} u_n \leqslant \sum_{n=N}^{\theta(N,L/\varepsilon)} u_n v_n \leqslant \sum_{n=0}^{\infty} u_n v_n < L$$

which is a contradiction. □

We now give a version of the Robbins-Siegmund theorem [86], slightly adapted to use-cases in stochastic approximation theory such as Dvoretzky's theorem which will be discussed in the coming section. An alternative quantitative variant of the Robbins-Siegmund theorem in its fully general form was recently given by the first and third author in [75]: Here, direct rates of convergence are not generally possible, and a result connected to fluctuation bounds is given instead, as mentioned before, and we refer to [75, Section 4.2] for a discussion of this type of result as well as the recent related literature on quantitative supermartingale convergence.

Theorem 3.4. *Let (X_n), (V_n) and (C_n) be sequences of nonnegative integrable real-valued random variables adapted to $(\mathcal{F}_n)$, and (a_n), (u_n) be sequences of nonnegative reals. Suppose that*

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \leqslant (1+a_n)X_n - u_n V_n + C_n \text{ a.s.}$$

for all $n \in \mathbb{N}$. Also, suppose that there exist $K, L, M > 0$ satisfying

$$\mathbb{E}[X_0] < L,\ \prod_{i=0}^{\infty}(1+a_i) < K, \quad \text{and} \quad \sum_{i=0}^{\infty}\mathbb{E}[C_i] < M$$

along with $\chi : (0,\infty) \to \mathbb{N}$ and $\theta : \mathbb{N} \times (0,\infty) \to \mathbb{N}$ such that

$$\forall b > 0\,\forall k \in \mathbb{N}\left(\sum_{i=k}^{\theta(k,b)} u_i \geqslant b\right) \quad \text{and} \quad \forall \varepsilon > 0\left(\sum_{i=\chi(\varepsilon)}^{\infty}\mathbb{E}[C_i] < \varepsilon\right).$$

Finally, suppose that $\tau : (0,\infty) \to (0,\infty)$ satisfies

$$\forall \varepsilon > 0\,\forall n \in \mathbb{N}\,(\mathbb{E}[V_n] < \tau(\varepsilon) \to \mathbb{E}[f(X_n)] < \varepsilon)$$

for some s.i.c.c. $f : [0,\infty) \to [0,\infty)$ *with moduli* ψ *and* κ*. Then* $\mathbb{E}[f(X_n)] \to 0$ *with rate*

$$\rho(\varepsilon) := \theta\left(\chi\left(\kappa\left(\varepsilon'\right)\right), \frac{K(L+M)}{\tau(\varepsilon')}\right) \quad \textit{for} \quad \varepsilon' := \frac{\varepsilon\psi(K^{-1})}{2}$$

and $X_n \to 0$ *a.s. with rate*

$$\rho'(\lambda,\varepsilon) := \rho(\lambda f(\varepsilon)).$$

Proof. Integrating both sides of the supermartingale property we obtain

$$\mathbb{E}[X_{n+1}] \leqslant (1+a_n)\mathbb{E}[X_n] - u_n\mathbb{E}[V_n] + \mathbb{E}[C_n].$$

Lemma 3.2 yields

$$\sum_{n=0}^{\infty} u_n\mathbb{E}[V_n] \leqslant K(L+M).$$

Therefore by Lemma 3.3 we have $\liminf_{n\to\infty} \mathbb{E}[V_n] = 0$ modulus $\varphi'(\varepsilon, N) := \theta(N, K(L+M)/\varepsilon)$ and thus a lim inf-modulus for $(f(X_n))$ is given by $\varphi(\varepsilon, N) := \varphi'(\tau(\varepsilon), N)$. Applying Theorem 2.8 with $A_n := a_n$ gives the result. □

If the regularity modulus is linear in the above special case of the Robbins-Siegmund theorem, then we can also obtain linear rates of convergence under suitable assumptions on the other parameters. For that, we make use of a quantitative result about real numbers that is a slight generalisation of a standard argument (see e.g. [70]).

Lemma 3.5. *Suppose that* (x_n) *is a sequence of nonnegative reals such that for* $c > 1$*,* $d \geqslant 0$ *and* $r \in \mathbb{N}\backslash\{0\}$*, we have*

$$x_{n+1} \leqslant \left(1 - \frac{c}{n+r}\right) x_n + \frac{d}{(n+r)^2}$$

for all $n \in \mathbb{N}$*. Then for all* $n \in \mathbb{N}$*:*

$$x_n \leqslant \frac{u}{n+r} \quad \textit{for} \quad u \geqslant \max\left\{\frac{d}{c-1}, rx_0\right\}.$$

Proof. We show the claim by induction. The case $n = 0$ holds by definition and for the induction step we note that since $d \leqslant u(c-1)$, we have

$$d \leqslant u\left(c - \frac{n+r}{n+r+1}\right)$$

for all n, r, and therefore

$$\begin{aligned}
x_{n+1} &\leqslant \left(1 - \frac{c}{n+r}\right) x_n + \frac{d}{(n+r)^2} \\
&\leqslant \left(1 - \frac{c}{n+r}\right) \frac{u}{n+r} + \frac{d}{(n+r)^2} \\
&\leqslant u\left(\left(1 - \frac{c}{n+r}\right) \frac{1}{n+r} + \left(c - \frac{n+r}{n+r+1}\right) \frac{1}{(n+r)^2}\right) \\
&= u\left(\frac{1}{n+r} - \frac{1}{(n+r+1)(n+r)}\right) \\
&= \frac{u}{n+r+1}\left(\frac{n+r+1}{n+r} - \frac{1}{n+r}\right) \\
&= \frac{u}{n+r+1}
\end{aligned}$$

which completes the induction. □

This gives the following result on fast rates for the Robbins-Siegmund theorem in the context of suitably fast parameters and a linear regularity modulus τ, i.e. with $\tau(\varepsilon) = t\varepsilon$ for some $t > 0$.

Theorem 3.6. *Let (X_n), (V_n) and (C_n) be sequences of nonnegative integrable real-valued random variables adapted to $(\mathcal{F}_n)$, and (a_n), (u_n) be sequences of nonnegative reals. Suppose that*

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \leqslant (1 + a_n)X_n - u_nV_n + C_n \text{ a.s.}$$

for all $n \in \mathbb{N}$. Also, let $K \geqslant 1$ and $L > 0$ be such that $\prod_{i=0}^{\infty}(1 + a_i) < K$ and $L \geqslant \mathbb{E}[X_0]$ and assume that

$$\mathbb{E}[C_n] \leqslant \frac{d}{(n+r)^2} \quad \text{and} \quad a_n + \frac{c}{n+r} \leqslant tu_n$$

for all $n \in \mathbb{N}$ where $c > 1$, $d \geqslant 0$ and $r \in \mathbb{N}\backslash\{0\}$. Finally, suppose we have $\mathbb{E}[V_n] \geqslant t\mathbb{E}[X_n]$ for all $n \in \mathbb{N}$ with $t > 0$. Then

$$\mathbb{E}[X_n] \leqslant \frac{u}{n+r} \quad \text{for} \quad u \geqslant \max\left\{\frac{d}{c-1}, rL\right\}$$

as well as

$$\mathbb{P}\left(\exists m \geqslant n(X_m \geqslant \varepsilon)\right) \leqslant \frac{1}{\varepsilon} \cdot \frac{K(u+2d)}{n+r}.$$

Proof. Integrating the inequality, we get

$$\begin{aligned}\mathbb{E}[X_{n+1}] &\leqslant (1 + a_n)\mathbb{E}[X_n] - u_n\mathbb{E}[V_n] + \mathbb{E}[C_n] \\ &\leqslant (1 + a_n - tu_n)\mathbb{E}[X_n] + \mathbb{E}[C_n] \\ &\leqslant \left(1 - \frac{c}{n+r}\right)\mathbb{E}[X_n] + \frac{d}{(n+r)^2}.\end{aligned}$$

Applying Lemma 3.5 yields the rate for $\mathbb{E}[X_n]$. For the second claim, we proceed similar to the proof of Theorem 2.8. Concretely, note first that

$$\mathbb{E}[C_n] = \frac{d}{(n+r)^2} \leqslant \frac{2d}{(n+r)(n+r+1)} = 2d\left(\frac{1}{n+r} - \frac{1}{n+r+1}\right)$$

so that

$$\sum_{i=n}^{\infty}\mathbb{E}[C_i] \leqslant 2d\sum_{i=n}^{\infty}\left(\frac{1}{n+r} - \frac{1}{n+r+1}\right) = \frac{2d}{n+r}$$

Now defining

$$U_n := \frac{X_n}{b_{n-1}} + \mathbb{E}\left[\sum_{i=n}^{\infty}\frac{C_i}{b_i} \mid \mathcal{F}_n\right], \text{ where } b_j := \prod_{i=0}^{j}(1 + a_i)$$

with $b_{-1} := 1$, analogously to Theorem 2.8, we have that (U_n) is a supermartingale. Since

$$\mathbb{E}[U_n] = \frac{\mathbb{E}[X_n]}{b_{n-1}} + \sum_{i=n}^{\infty}\frac{\mathbb{E}[C_i]}{b_i} \leqslant \mathbb{E}[X_n] + \sum_{i=n}^{\infty}\mathbb{E}[C_i] \leqslant \frac{u+2d}{n+r},$$

using Ville's inequality, this yields

$$\begin{aligned}\mathbb{P}\left(\exists m \geqslant n(X_m \geqslant \varepsilon)\right) &\leqslant \mathbb{P}\left(\exists m \geqslant n(U_m \geqslant \varepsilon/K)\right) \\ &\leqslant \frac{K}{\varepsilon} \cdot \mathbb{E}[U_n] \\ &\leqslant \frac{1}{\varepsilon} \cdot \frac{K(u+2d)}{n+r}.\end{aligned}$$

□

4. Applications to Dvoretzky's theorem

A notable application of the Robbins-Siegmund theorem, as already highlighted in the original paper [86], is a generalization of the seminal theorem of Dvoretzky [32] in stochastic approximation to Hilbert spaces[2]. Concretely, the following theorem is established in [86] via an application of the Robbins-Siegmund theorem:

Theorem 4.1 ([86], based on [32])**.** *Let X be a separable Hilbert space and let, for any $n \geqslant 1$, $T_n : X^n \to X$ be a Borel-measurable function such that there exists a point $\theta \in X$ along with sequences of nonnegative real numbers $(a_n), (b_n), (c_n)$ such that*

$$\|T_n(z_0, \dots, z_{n-1}) - \theta\| \leqslant \max\{a_{n-1}, (1 + b_{n-1}) \|z_{n-1} - \theta\| - c_{n-1}\}$$

for all $z_0, \dots, z_{n-1} \in X^n$. Fix X-valued random variables x_0 and (y_n) such that $\mathbb{E}[y_n \mid \mathcal{F}_n] = 0$ for each $n \in \mathbb{N}$, where $\mathcal{F}_n := \sigma(x_0, y_0, \dots, y_{n-1})$, i.e. the σ-algebra generated by $x_0, y_0, \dots, y_{n-1}$. In x_0 and (y_n), define the iteration

$$x_{n+1} := T_{n+1}(x_0, \dots, x_n) + y_n.$$

Suppose that $a_n \to 0$, $\sum_{n=0}^{\infty} b_n < \infty$, $\sum_{n=0}^{\infty} c_n = \infty$ and $\sum_{n=0}^{\infty} \mathbb{E}[\|y_n\|^2] < \infty$. Then $x_n \to \theta$ a.s.

A number of stochastic approximation algorithms take the form of the general iterative procedure

$$x_{n+1} = T_{n+1}(x_0, \dots, x_n) + y_n$$

of the above theorem, most notably the Kiefer-Wolfowitz scheme [47] and (many variants of) the Robbins-Monro procedure [85]. Even further, Dvoretzky's result presented general conditions on the parameters of the previously introduced iterative scheme that unified many stochastic approximation results in the literature.

In this section, we provide a quantitative version of Dvoretzky's theorem by combining an analysis of the proof of this result given in [86] with our previously established quantitative version of the Robbins-Siegmund theorem. Before we do this, we briefly discuss the assumptions in the above result. Concretely, for one, note that as in the proof presented in [86], beyond the assumption that $\sum_{n=0}^{\infty} c_n = \infty$, the sequence (c_n) can be assumed, without loss of generality, to additionally satisfy $\sum_{n=0}^{\infty} c_n^2 < \infty$ (and actually even $c_n \leqslant 1$ for all $n \in \mathbb{N}$, although we will not really rely on this). In that way, the quantitative results presented below will feature this additional assumption together with a quantitative rendering thereof. For another, note that it is an immediate consequence of the assumptions of the above result that $\mathbb{E}[\|x_n\|^2] < \infty$ for all $n \in \mathbb{N}$. While, in the following quantitative result, we could also calculate a bound on these means in terms of the other parameters recursively in $n \in \mathbb{N}$, we will for simplicity assume bounds $L_n > 0$ on these means as primitive inputs.

Our quantitative version of Dvoretzky's theorem in stochastic approximation now takes the form of the following Theorem 4.2, and the proof in particular heavily relies on the fact that our general result, as well as the resulting quantitative version of the Robbins-Siegmund theorem, allow for processes slowed down by a s.i.c.c. function, which here will be instantiated by the root, as it is only the presence of that function which in the end allows us to produce a corresponding modulus of regularity.

Theorem 4.2. *Let X be a separable Hilbert space and let, for any $n \geqslant 1$, $T_n : X^n \to X$ be a Borel-measurable function such that there exists a point $\theta \in X$ along with sequences of nonnegative real numbers $(a_n), (b_n), (c_n)$ such that*

$$\|T_n(z_0, \dots, z_{n-1}) - \theta\| \leqslant \max\{a_{n-1}, (1 + b_{n-1}) \|z_{n-1} - \theta\| - c_{n-1}\}$$

[2]Dvoretzky later offered an alternative proof of this generalization in [33], see also [95].

for all $z_0, \dots, z_{n-1} \in X^n$. Fix X-valued random variables x_0 and (y_n) such that $\mathbb{E}[y_n \mid \mathcal{F}_n] = 0$ for each $n \in \mathbb{N}$, where $\mathcal{F}_n := \sigma(x_0, y_0, \dots, y_{n-1})$, i.e. $\mathcal{F}_n$ is the σ-algebra generated by $x_0, y_0, \dots, y_{n-1}$. In x_0 and (y_n), define the iteration

$$x_{n+1} := T_{n+1}(x_0, \dots, x_n) + y_n.$$

Suppose that $a_n \to 0$, $\sum_{n=0}^{\infty} b_n < \infty$, $\sum_{n=0}^{\infty} c_n^2 < \infty$ and $\sum_{n=0}^{\infty} \mathbb{E}[\|y_n\|^2] < \infty$ with upper bounds $A, B, C, M > 0$ and rates of convergence $\varphi, \beta, \gamma, \mu : (0, \infty) \to \mathbb{N}$, i.e.

$$\forall n \geqslant \varphi(\varepsilon)\,(a_n < \varepsilon) \;\text{ as well as }\; \sum_{n=\beta(\varepsilon)}^{\infty} b_n,\ \sum_{n=\gamma(\varepsilon)}^{\infty} c_n^2,\ \sum_{n=\mu(\varepsilon)}^{\infty} \mathbb{E}[\|y_n\|^2] < \varepsilon$$

for any $\varepsilon > 0$. Further, assume $\sum_{n=0}^{\infty} c_n = \infty$ with a rate of divergence $\theta : \mathbb{N} \times (0, \infty) \to \mathbb{N}$, i.e. $\sum_{n=k}^{\theta(k,b)} c_n \geqslant b$ for any $k \in \mathbb{N}$ and $b > 0$. Lastly, assume that $\mathbb{E}[\|x_n\|^2] < \infty$ for all $n \in \mathbb{N}$ with upper bounds $L_n > 0$. Then $x_n \to \theta$ a.s. with a rate

$$\rho(\lambda, \varepsilon) := \theta\left(\max\left\{\chi_{\varepsilon/2}\left(\frac{\lambda^2 \varepsilon}{4K_{\varepsilon/2}}\right), \varphi(\varepsilon/2)\right\}, \frac{2K_{\varepsilon/2}\sqrt{K_{\varepsilon/2}}(L_{\varphi(\varepsilon/2)} + M_{\varepsilon/2})}{\lambda\varepsilon}\right)$$

where, for any $\delta > 0$, we define

$$K_\delta := (1 + B^2)e^{\delta B},$$
$$M_\delta := (1 + \delta B)C + \delta(1 + \delta B)B + M,$$
$$\chi_\delta(\varepsilon) := \max\left\{\mu\left(\frac{\varepsilon}{3}\right), \gamma\left(\frac{\varepsilon}{3(1 + \delta B)}\right), \beta\left(\frac{\varepsilon}{3\delta(1 + \delta B)}\right)\right\}.$$

Proof. We reduce the result to the Robbins-Siegmund theorem, following (a slight modification of) the well-known argument given in [86]. For this, we first fix a $\delta > 0$ and, writing $(x)^+$ for $\max\{x, 0\}$, we set

$$W_{n,\delta} := \left((\|x_n - \theta\| - \delta)^+\right)^2$$

as well as the X-valued random variables $T_n := T_n(x_0, \dots, x_{n-1}) - \theta$ for $n \geqslant 1$ and

$$u_n := T_{n+1}\mathbf{1}_{\|T_{n+1}\| \leqslant \delta} + \delta\frac{T_{n+1}}{\|T_{n+1}\|}\mathbf{1}_{\|T_{n+1}\| > \delta}$$

for $n \in \mathbb{N}$. By definition, T_n is $\mathcal{F}_{n-1}$-measurable and so u_n is $\mathcal{F}_n$-measurable. Further, we immediately have $\|u_n\| \leqslant \delta$ pointwise everywhere which implies

$$W_{n+1,\delta} = \left((\|x_{n+1} - \theta\| - \delta)^+\right)^2 \leqslant \left((\|x_{n+1} - \theta - u_n\| + \|u_n\| - \delta)^+\right)^2 \leqslant \|x_{n+1} - \theta - u_n\|^2$$

for all $n \in \mathbb{N}$. Further, note that by definition we have

$$T_{n+1} - u_n = \left(1 - \frac{\delta}{\|T_{n+1}\|}\right) T_{n+1}\mathbf{1}_{\|T_{n+1}\| > \delta}$$

so that $\|T_{n+1} - u_n\| = (\|T_{n+1}\| - \delta)^+$. Combined with the assumption that $\mathbb{E}[y_n \mid \mathcal{F}_n] = 0$, we can then derive that

$$\begin{aligned}
\mathbb{E}[W_{n+1,\delta} \mid \mathcal{F}_n] &\leqslant \mathbb{E}[\|x_{n+1} - \theta - u_n\|^2 \mid \mathcal{F}_n] \\
&= \mathbb{E}[\|T_{n+1} + y_n - u_n\|^2 \mid \mathcal{F}_n] \\
&= \mathbb{E}[\|T_{n+1} - u_n\|^2 \mid \mathcal{F}_n] + \mathbb{E}[\|y_n\|^2 \mid \mathcal{F}_n] + 2\mathbb{E}[\langle y_n, T_{n+1} - u_n\rangle \mid \mathcal{F}_n] \\
&= \|T_{n+1} - u_n\|^2 + \mathbb{E}[\|y_n\|^2 \mid \mathcal{F}_n] \\
&= \left((\|T_{n+1}\| - \delta)^+\right)^2 + \mathbb{E}[\|y_n\|^2 \mid \mathcal{F}_n]
\end{aligned}$$

for all $n \in \mathbb{N}$. Setting $N := \varphi(\delta)$ yields that $a_{N+n} \leqslant \delta$ for all $n \in \mathbb{N}$, so that our main assumption on the mappings T_n yield

$$\begin{aligned}(\|T_{N+n+1}\| - \delta)^+ &\leqslant \max\{a_{N+n} - \delta, (1 + b_{N+n})\,\|x_{N+n} - \theta\| - c_{N+n} - \delta, 0\}\\ &\leqslant \max\{(1 + b_{N+n})\,\|x_{N+n} - \theta\| - c_{N+n} - \delta, 0\}\\ &= ((1 + b_{N+n})(\|x_{N+n} - \theta\| - \delta) - c_{N+n} + \delta b_{N+n})^+ .\end{aligned}$$

Now, either we have $\|T_{N+n+1}\| > \delta$ so that $(\|T_{N+n+1}\| - \delta)^+ = \|T_{N+n+1}\| - \delta > 0$ and therefore

$$\begin{aligned}0 < (\|T_{N+n+1}\| - \delta)^+ &\leqslant ((1 + b_{N+n})(\|x_{N+n} - \theta\| - \delta) - c_{N+n} + \delta b_{N+n})^+\\ &= (1 + b_{N+n})(\|x_{N+n} - \theta\| - \delta) - c_{N+n} + \delta b_{N+n}\\ &\leqslant (1 + b_{N+n})(\|x_{N+n} - \theta\| - \delta)^+ - c_{N+n} + \delta b_{N+n}\end{aligned}$$

which yields

$$\left((\|T_{N+n+1}\| - \delta)^+\right)^2 \leqslant \left((1 + b_{N+n})(\|x_{N+n} - \theta\| - \delta)^+ - c_{N+n} + b_{N+n}\delta\right)^2$$

in this case. However, if $\|T_{N+n+1}\| \leqslant \delta$, then it holds that $(\|T_{N+n+1}\| - \delta)^+ = 0$ and so the above inequality is true unconditionally. Finally, from this we get that

$$\begin{aligned}\left((\|T_{N+n+1}\| - \delta)^+\right)^2 \leqslant{}& (1 + \delta b_{N+n})\,(1 + b_{N+n})^2 W_{N+n,\delta} - 2(1 + b_{N+n}) c_{N+n}\sqrt{W_{N+n,\delta}}\\ &+ (1 + \delta b_{N+n})\,c_{N+n}^2 + \delta b_{N+n}\,(1 + \delta b_{N+n})\,.\end{aligned}$$

which we derive utilizing that $(x + y)^2 \leqslant (1 + y)x^2 + y(1 + y)$ for any $x \in \mathbb{R}$ and $y \geqslant 0$, instantiated with

$$x := (1 + b_{N+n})(\|x_{N+n}\| - \delta)^+ - c_{N+n} = (1 + b_{N+n})\sqrt{W_{N+n,\delta}} - c_{N+n}$$

as well as $y := \delta b_{N+n}$, and noting that $1 + \delta b_{N+n} \geqslant 0$.

All in all, we have derived that

$$\mathbb{E}[W_{N+n+1,\delta} \mid \mathcal{F}_{N+n}] \leqslant (1 + \alpha_{N+n,\delta})W_{N+n,\delta} - u_{N+n}\sqrt{W_{N+n,\delta}} + C_{N+n,\delta}$$

holds for all $n \in \mathbb{N}$, where $u_n := 2(1 + b_n)c_n$ as well as

$$\begin{aligned}\alpha_{n,\delta} &:= (1 + \delta b_n)\,(1 + b_n)^2 - 1,\\ C_{n,\delta} &:= (1 + \delta b_n)\,c_n^2 + \delta b_n\,(1 + \delta b_n) + \mathbb{E}[\|y_n\|^2 \mid \mathcal{F}_n].\end{aligned}$$

It is elementary to verify that, by construction, we have $\sum_{n=0}^{\infty} \mathbb{E}[C_{N+n,\delta}] < \infty$ with a bound M_δ and rate of convergence $\max\{\chi_\delta(\varepsilon) - N, 0\}$. Further, we have $\prod_{n=0}^{\infty}(1 + \alpha_{N+n,\delta}) < \infty$ with a bound K_δ and it can be immediately verified that $\theta(N + n, b) - N$ is a rate of divergence for (u_{N+n}).

Hence, we can apply Theorem 3.4 with $X_n := W_{N+n,\delta}$, $V_n := \sqrt{W_{N+n,\delta}}$ and $\mathcal{F}_n := \mathcal{F}_{N+n}$ (as well as $f(x) := \sqrt{x}$ so that we set $\psi(a) := \sqrt{a}$ and $\kappa(\varepsilon) := \varepsilon^2$ following Example 2.6, (1)) to derive that $W_{N+n,\delta} \to 0$ a.s. with a certain rate $\Delta^\delta(\lambda, \varepsilon)$ arising from Theorem 3.4. At last, let $\varepsilon, \lambda > 0$ be given. Then for $\delta = \varepsilon/2$, we obtain thereby that

$$\begin{aligned}&\mathbb{P}\left(\exists n \geqslant \Delta^{\varepsilon/2}(\lambda, \varepsilon^2/4) + N\left(\|x_n - \theta\| \geqslant \varepsilon\right)\right)\\ &\quad= \mathbb{P}\left(\exists n \geqslant \Delta^{\varepsilon/2}(\lambda, \varepsilon^2/4)\left(\left((\|x_{N+n} - \theta\| - \varepsilon/2)^+\right)^2 \geqslant \varepsilon^2/4\right)\right)\\ &\quad= \mathbb{P}\left(\exists n \geqslant \Delta^{\varepsilon/2}(\lambda, \varepsilon^2/4)\left(W_{N+n,\varepsilon/2} \geqslant \varepsilon^2/4\right)\right) < \lambda\end{aligned}$$

and so $\|x_n\| \to 0$ a.s. with a rate given by

$$\rho(\lambda, \varepsilon) = \Delta^{\varepsilon/2}(\lambda, \varepsilon^2/4) + N = \Delta^{\varepsilon/2}(\lambda, \varepsilon^2/4) + \varphi(\varepsilon/2).$$

The rate presented in the above theorem follows from simplifying the corresponding expressions arising from Theorem 3.4. □

It should be observed that there exist alternative proofs of Dvoretzky's theorem, notably that of Derman and Sacks [29], which are more closely tailored to the specific assumptions of the theorem and do not make direct use of supermartingale convergence. A careful analysis of such proofs might result in a different rate of convergence and thus an alternative quantitative version of Dvoretzky's theorem than the one give here.[3]

5. Stochastic quasi-Fejér monotonicity in the presence of uniqueness

We now apply our general results on rates of convergence to one further abstract scenario, by providing a general result on rates of convergence, under a certain stochastic uniqueness assumption, for stochastic quasi-Fejér monotone sequences, a class of (in this paper potentially metric-valued) stochastic processes that encompasses a wide array of common methods employed in these areas, such as the Robbins-Monro method, stochastic variants of the proximal point method, Krasnoselskii-Mann iterations and many others.

The notion of (quasi-)Fejér monotonicity[4] belongs to one of the most central notions for the study of iterative methods in nonlinear analysis and optimization, and we refer to the well-known expositions in [8, 24, 25] for further information on this concept in these contexts. In rather general contexts, in particular encompassing a metric setting, Fejér monotonicity (and generalizations thereof) was studied in [53, 54, 55, 80], where these works are particularly relevant as they are both of a quantitative nature and logically motivated by the perspective of proof mining similar as the present paper.

The first stochastic variants of quasi-Fejér monotonicity seem to have been considered in the pioneering works [34, 35, 37], set in Euclidean spaces, which subsequently where refined in [26, 28] to general separable Hilbert spaces. The latter work [28] in particular is concerned with fast rates for resulting processes under a quasi-contractivity assumption on the involved operators, which as will be discussed later is a particular example of a situation where one naturally obtains the existence of a strongly unique solution in expectation, so that these results in particular fall under the breadth of the present paper.

Throughout this section, let (X, d) be a fixed metric space. At our most abstract, we will be concerned with finding a zero $F(z) = 0$ of a general measurable function $F : X \to [0, \infty]$, assuming that the set of zeros $\mathrm{zer}F$ is nonempty and actually a singleton, i.e. $\mathrm{zer}F = \{z\}$, and that, moreover, this property is quantitatively witnessed in the following uniform way:

Definition 5.1. Let D be a collection of X-valued random variables. We say that the zero z of F is strongly unique in expectation (over D) if there exists a function $\tau : (0, \infty) \to (0, \infty)$ such that

$$\forall x \in D \; \forall \varepsilon > 0 \, (\mathbb{E}[F(x)] < \tau(\varepsilon) \to \mathbb{E}[d(x, z)] < \varepsilon) \, .$$

Such a τ is then called a modulus of strong uniqueness in expectation for F (over D).

More illustratively, F has a strongly unique zero in expectation if any random variable which x is a good-enough approximate zero of F in expectation already is close to z in expectation. This notion of a strongly unique zero is closely related, in the presence of a uniqueness assumption, to the abstract general regularity notion studied in [54].

[3] This observation is due to R. Arthan and P. Oliva (private communication), who are working on an analysis Derman-Sacks proof.

[4] The notion seems to have first been named as such in [68], after the use of a similar concept made by L. Fejér in [39].

This property is a strengthening of the plain uniqueness of the zero of F and this strength re-manifests itself in the general type of convergence result we are able to prove for certain stochastic iterations approximating such a zero. However, as will be illustrated in the following part of this section where we discuss a rather general situation as well as an array of examples where such a modulus of strong uniqueness in expectation can be naturally given, the condition, while being strong, is still rather comprehensive and allows for the uniform treatment of a variety of regularity conditions commonly assumed in the literature on stochastic optimization.

Now, as mentioned before, to solve this problem we study the special class of stochastic iterations (x_n) (i.e. (x_n) is a sequence of X-value random variables) which are stochastically quasi-Fejér monotone in the following sense:

Definition 5.2. Let $(\mathcal{F}_n)$ be a filtration and let (x_n) be a sequence of X-valued $\mathcal{F}_n$-measurable random variables. Then (x_n) is called stochastically quasi-Fejér monotone w.r.t. $S \subseteq X$ and $(\mathcal{F}_n)$ if

$$\mathbb{E}[d(x_{n+1}, z) \mid \mathcal{F}_n] \leqslant (1 + \zeta_n)d(x_n, z) + \xi_n \text{ a.s.}$$

for all $z \in S$ and all $n \in \mathbb{N}$, where $\sum_{n=0}^{\infty} \xi_n, \sum_{n=0}^{\infty} \zeta_n < \infty$ a.s.

In analogy to the non-stochastic quasi-Fejér monotonicity, also here the only further property needed to induce convergence of a stochastically quasi-Fejér monotone sequence relative to the zero of a function F which is strongly unique in expectation is that the iteration has approximate zeros for the function F infinitely often, which we require to be in expectation in this stochastic context in the following sense:

Definition 5.3. Let (x_n) be a sequence of X-valued random variables. We say that (x_n) has the lim inf-property in expectation relative to F if $\liminf_{n\to\infty} \mathbb{E}[F(x_n)] = 0$. A function $\varphi : (0, \infty) \times \mathbb{N} \to (0, \infty)$ witnessing this property quantitatively in the sense that

$$\forall \varepsilon > 0 \; \forall N \in \mathbb{N} \; \exists n \in [N; \varphi(\varepsilon, N)] \, (\mathbb{E}[F(x_n)] < \varepsilon)$$

is called a lim inf-bound in expectation for (x_n) relative to F.

As it will turn out, most properties of the metric are actually very inessential for the proofs of our main arguments and so we in the following will actually follow the approach of the recent work [80] and consider this notion to be relativized to a more general measurable mapping $\phi : X \times X \to [0, \infty)$ which measures the distances in the quasi-Fejér monotonicity property instead, and which allows us to capture even wider classes of iterations uniformly and abstractly with this approach. To connect back to the metric, the strongest assumption placed on that distance function in this section will be the property of uniform consistency similarly introduced in [80]:[5]

Definition 5.4. A mapping $\phi : X \times X \to [0, \infty)$ is called uniformly consistent if there exists a function $\kappa : (0, \infty) \to (0, \infty)$ such that

$$\forall \varepsilon > 0 \; \forall x, y \in X \, (\phi(x, y) < \kappa(\varepsilon) \to d(x, y) < \varepsilon) \, .$$

Such a κ is then called a modulus of uniform consistency for ϕ.

We refer to [80] (and to recent applications [79, 81] featuring this or similar notions) for further illustrations on the breadth of this perspective.

In the context of this modification of the distance function, both the strong uniqueness in expectation as well as the stochastic quasi-Fejér monotonicity have to be relativized to yield a

[5]It should be noted that the notion of uniform consistency introduced in [80] also includes a converse to the property highlighted here, which will however not be needed for the results of this paper and is hence omitted from the definition.

sensible convergence result. For the strong uniqueness in expectation, this yields the following relativization:

Definition 5.5. Let D be a collection of X-valued random variables. Let $\phi : X \times X \to [0, \infty)$ be a measurable mapping. We say that the zero z of F is strongly ϕ-unique in expectation (over D) if there exists a function $\tau : (0, \infty) \to (0, \infty)$ such that

$$\forall x \in D \; \forall \varepsilon > 0 \, (\mathbb{E}[F(x)] < \tau(\varepsilon) \to \mathbb{E}[\phi(x, z)] < \varepsilon) \, .$$

Such a τ is then called a modulus of strong ϕ-uniqueness in expectation for F (over D).

Remark 5.6. The above notion can be recognized as an instantiation of the notion of a modulus of regularity from Section 2 by regarding $(F(x))$ and $(\phi(x, z))$ as families of nonnegative real-valued random variables over the index set D. Thereby, akin to the previous Remark 2.12 (and akin to the circumstances of the general notion of a modulus of regularity introduced and studied in [54], see Remark 3.2 therein), this kind of regularity notion induces a growth condition on the premise in terms of the conclusion in the sense that for a strictly increasing τ, requiring that τ is a modulus of strong ϕ-uniqueness in expectation for F over D is equivalent to requiring

$$\mathbb{E}[F(x)] \geqslant \tau(\mathbb{E}[\phi(x, z)]) \text{ for all } x \in D.$$

For the stochastic quasi-Fejér monotonicity, this yields the following relativized version:

Definition 5.7. Let $\phi : X \times X \to [0, \infty)$ be a measurable mapping. Let $(\mathcal{F}_n)$ be a filtration and let (x_n) be a sequence of X-valued $\mathcal{F}_n$-measurable random variables. Then (x_n) is called stochastically ϕ-quasi-Fejér monotone w.r.t. $S \subseteq X$ and $(\mathcal{F}_n)$ if

$$\mathbb{E}[\phi(x_{n+1}, z) \mid \mathcal{F}_n] \leqslant (1 + \zeta_n)\phi(x_n, z) + \xi_n \text{ a.s.}$$

for all $z \in S$ and all $n \in \mathbb{N}$, where $(\xi_n), (\zeta_n)$ are suitable sequences of nonnegative, integrable real-valued random variables.

Under these main assumptions, together with some minor assumptions on surrounding (quantitative) data, we get the following result on the convergence of such sequences as well as their speed from our main Theorem 2.8.

Theorem 5.8. *Let (X, d) be a metric space and let $\phi : X \times X \to [0, \infty)$ be a measurable map. Let $F : X \to [0, \infty]$ be measurable with $\mathrm{zer}F = \{z\}$, where the zero z is strongly ϕ-unique in expectation over a collection D of X-valued random variables with a modulus $\tau : (0, \infty) \to (0, \infty)$, i.e.*

$$\forall x \in D \; \forall \varepsilon > 0 \, (\mathbb{E}[F(x)] < \tau(\varepsilon) \to \mathbb{E}[\phi(x, z)] < \varepsilon) \, .$$

Let $(\mathcal{F}_n)$ be a filtration and let $(x_n) \subseteq D$ be a sequence of X-valued $\mathcal{F}_n$-measurable random variables such that $\phi(x_n, z)$ is integrable for all $n \in \mathbb{N}$. Suppose further that (x_n) is stochastically ϕ-quasi-Fejér monotone w.r.t. $\mathrm{zer}F$ and $(\mathcal{F}_n)$, i.e.

$$\mathbb{E}[\phi(x_{n+1}, z) \mid \mathcal{F}_n] \leqslant (1 + \zeta_n)\phi(x_n, z) + \xi_n \text{ a.s.}$$

for all $n \in \mathbb{N}$, where $(\xi_n), (\zeta_n)$ are sequences of integrable real-valued random variables such that there exist $K > 0$ and $\chi : (0, \infty) \to \mathbb{N}$ with

$$\prod_{n=0}^{\infty}(1 + \zeta_n) < K \text{ a.s.} \quad \text{and} \quad \forall \varepsilon > 0 \left(\sum_{n=\chi(\varepsilon)}^{\infty} \mathbb{E}[\xi_n] < \varepsilon \right).$$

Suppose furthermore that (x_n) *has the* $\liminf$*-property in expectation relative to* F *with a bound* $\varphi : (0,\infty) \times \mathbb{N} \to (0,\infty)$*, i.e.*

$$\forall \varepsilon > 0 \; \forall N \in \mathbb{N} \; \exists n \in [N; \varphi(\varepsilon, N)] \, (\mathbb{E}[F(x_n)] < \varepsilon) \, .$$

Then $\mathbb{E}[\phi(x_n, z)] \to 0$ *with rate*

$$\rho(\varepsilon) := \varphi\left(\tau\left(\frac{\varepsilon}{2K}\right), \chi\left(\frac{\varepsilon}{2K}\right)\right)$$

and $\phi(x_n, z) \to 0$ *a.s. with rate* $\rho'(\lambda, \varepsilon) := \rho(\lambda\varepsilon)$*. If furthermore* ϕ *is uniformly consistent with modulus* $\kappa : (0,\infty) \to (0,\infty)$*, i.e.*

$$\forall \varepsilon > 0 \; \forall x, y \in X \, (\phi(x,y) < \kappa(\varepsilon) \to d(x,y) < \varepsilon) \, ,$$

then $d(x_n, z) \to 0$ *a.s. with rate* $\rho'(\lambda, \kappa(\varepsilon))$*.*

Proof. Define $X_n := \phi(x_n, z)$ for any $n \in \mathbb{N}$. As (x_n) is adapted to $(\mathcal{F}_n)$ and ϕ is measurable, also (X_n) is adapted to $(\mathcal{F}_n)$. As in the discussion after Theorem 2.8, by the assumptions on φ and τ, we have that $\varphi(\tau(\varepsilon), N)$ is a $\liminf$-modulus for $\mathbb{E}[X_n] = \mathbb{E}[\phi(x_n, z)]$. For $A_n := \zeta_n$ and $C_n := \xi_n$ as well as $f := \mathrm{id}$, Theorem 2.8 then yields the respective rates for $\mathbb{E}[\phi(x_n, z)] \to 0$ and $\phi(x_n, z) \to 0$ a.s. The latter in particular means

$$\forall \varepsilon, \lambda > 0 \, (\mathbb{P}(\exists n \geqslant \rho'(\lambda, \varepsilon)(\phi(x_n, z) \geqslant \varepsilon)) < \lambda) \, .$$

If ϕ is then also uniformly consistent, we further have for any $\varepsilon, \lambda > 0$ that

$$\mathbb{P}(\exists n \geqslant \rho'(\lambda, \kappa(\varepsilon))(d(x_n, z) \geqslant \varepsilon)) \leqslant \mathbb{P}(\exists n \geqslant \rho'(\lambda, \kappa(\varepsilon))(\phi(x_n, z) \geqslant \kappa(\varepsilon))) < \lambda$$

so that $\rho'(\lambda, \kappa(\varepsilon))$ is a rate for $d(x_n, z) \to 0$. □

In analogy to Corollary 2.9, we also here want to highlight the following qualitative theorem on the convergence of stochastically quasi-Fejér monotone sequences under a strong uniqueness assumption, which we obtain from the above result by disregarding the quantitative information:

Corollary 5.9. Given a metric space (X, d) and measurable functions $\phi : X \times X \to [0,\infty)$ and $F : X \to [0,\infty]$, suppose that F has a unique zero z which is strongly ϕ-unique in expectation over a collection D of X-valued random variables. Suppose that $(x_n) \subseteq D$ is a sequence of X-valued random variables which is stochastically ϕ-quasi-Fejér monotone w.r.t. $\mathrm{zer}F$ (and a suitable filtration) and that it has the $\liminf$-property in expectation relative to F. Then $\mathbb{E}[\phi(x_n, z)] \to 0$ and $\phi(x_n, z) \to 0$ a.s. If ϕ is uniformly consistent, then $d(x_n, z) \to 0$ a.s.

Similar to Theorem 3.6, we also want to highlight a result on fast rates of convergence for certain stochastic quasi-Fejér monotone sequences in the context of linear moduli of strong uniqueness.

Theorem 5.10. *Let* (X, d) *be a metric space and let* $\phi : X \times X \to [0,\infty)$ *be a measurable map. Let* $F : X \to [0,\infty]$ *be measurable with* $\mathrm{zer}F = \{z\}$*, where the zero* z *is strongly* ϕ*-unique in expectation over a collection* D *of* X*-valued random variables where further* $\mathbb{E}[F(x)] \geqslant t\mathbb{E}[\phi(x, z)]$ *for all* $x \in D$ *with* $t > 0$*. Let* $(\mathcal{F}_n)$ *be a filtration and let* $(x_n) \subseteq D$ *be a sequence of* X*-valued* $\mathcal{F}_n$*-measurable random variables such that* $\phi(x_n, z)$ *is integrable for all* $n \in \mathbb{N}$*. Suppose further that* (x_n) *satisfies*

$$\mathbb{E}[\phi(x_{n+1}, z) \mid \mathcal{F}_n] \leqslant (1 + \zeta_n)\phi(x_n, z) - \eta_n F(x_n) + \xi_n \;\; a.s.$$

for all $n \in \mathbb{N}$*, where* (ζ_n)*,* (η_n) *are sequences of nonnegative reals and* (ξ_n) *is a sequence of integrable real-valued random variables such that*

$$\mathbb{E}[\xi_n] \leqslant \frac{d}{(n+r)^2} \quad \textit{and} \quad \zeta_n + \frac{c}{n+r} \leqslant t\eta_n$$

for all $n \in \mathbb{N}$ where $c > 1$, $d \geqslant 0$ and $r \in \mathbb{N}\backslash\{0\}$. Finally suppose that $K \geqslant 1$ and $L > 0$ are such that $\prod_{i=0}^{\infty}(1+\zeta_i) < K$ and $L \geqslant \mathbb{E}[\phi(x_0, z)]$. Then

$$\mathbb{E}[\phi(x_n, z)] \leqslant \frac{u}{n+r} \quad \textit{for} \quad u \geqslant \max\left\{\frac{d}{c-1}, rL\right\}$$

as well as

$$\mathbb{P}\left(\exists m \geqslant n(\phi(x_m, z) \geqslant \varepsilon)\right) \leqslant \frac{1}{\varepsilon} \cdot \frac{K(u+2d)}{n+r}.$$

As the above is an immediate corollary of Theorem 3.6, we omit the proof.

We end this section with some general consideration on situations where the unique zero z of a function $F : X \to [0, \infty]$ is even strongly unique in expectation, akin to the previous Propositions 2.11 and 2.13. Note for this again as in Remark 5.6 that the notion of strong uniqueness in expectation can be recognized as an instantiation of the notion of a modulus of regularity from Section 2 by regarding $(F(x))$ and $(\phi(x, z))$ as families of nonnegative real-valued random variables over the index set D. As the proofs are then rather immediate, we omit them here.

Proposition 5.11. *Let D be a collection of X-valued random variables. Let $\phi : X \times X \to [0, \infty)$ be a measurable mapping and let $F : X \to [0, \infty]$ with $z \in \mathrm{zer}F$ be such that $\phi(x, z)$ is integrable and*

$$F(x) \geqslant \tau(\phi(x, z)) \text{ a.s.}$$

for all $x \in D$, where $\tau : [0, \infty) \to [0, \infty)$ is a convex and strictly increasing function. Then z is strongly ϕ-unique in expectation over D and τ is a modulus.

Proposition 5.12. *Let D be a collection of X-valued random variables. Let $\phi : X \times X \to [0, \infty)$ be a measurable mapping and let $F : X \to [0, \infty]$ with $z \in \mathrm{zer}F$ be such that*

$$F(x) \geqslant \pi(\phi(x, z)) \text{ a.s.}$$

for all $x \in D$ and where $\pi : (0, \infty) \to (0, \infty)$. Let $K \geqslant \mathbb{E}[\phi(x, z)]$ for all $x \in D$ and let $(\phi(x, z))_{x \in D}$ be uniformly integrable with modulus $\mu : (0, \infty) \to (0, \infty)$. Then z is strongly ϕ-unique in expectation over D with a modulus τ defined by

$$\tau(\varepsilon) := \pi\left(\frac{\varepsilon}{4}, \frac{K}{\mu(\varepsilon/4)}\right)\frac{\mu(\varepsilon/4)}{2}.$$

In both cases, note that the above requirement that $F(x) \geqslant \tau(\phi(x, z))$ a.s. for all $x \in D$ is in particular true if $F(x) \geqslant \tau(\phi(x, z))$ for all $x \in X$.

6. The Robbins-Monro procedure

Having presented three abstract quantitative convergence frameworks, we now focus on applications. These fall into two categories. In this section, we deliberately focus on one of the best known stochastic approximation algorithms, namely the Robbins-Monro scheme, in order to illustrate our work with a familiar example, and to thereby demonstrate how algorithms based on Robbins-Monro fit into our general approach. In the section that follows this, we then list a wide range of different iterations, many of which arise from recent work, in order to emphasise the scope and potential of our approach in producing new quantitative information for stochastic algorithms.

The Robbins-Monro scheme [85] represents one of the first stochastic approximation algorithms, originally designed as a recursive procedure for finding the unique root z of a function

$M : \mathbb{R} \to \mathbb{R}$ in situations where we do not have direct access to the function M but can only take noise-corrupted samples. The basic scheme is defined by

$$x_{n+1} := x_n - a_n(M(x_n) + \varepsilon_n)$$

where (ε_n) represents the random errors made by (approximately) calculating $M(x_n)$ and (a_n) is a positive sequence of reals satisfying the (now classic) conditions

$$\sum_{n=0}^{\infty} a_n = \infty \quad \text{and} \quad \sum_{n=0}^{\infty} a_n^2 < \infty.$$

A similar scheme for finding the maximum of a function was considered by Kiefer and Wolfowitz [47], and these early results were strengthened by Blum in [15] who established almost sure convergence of the scheme under weakened assumptions. A detailed survey of the Robbins-Monro scheme and well-known algorithms connected to it, along with the various different approaches to establishing convergence, can be found in all standard references on stochastic approximation (see in particular [31, 59, 60]).

Our aim in what follows is to give a version of the Robbins-Monro scheme that matches our level of abstraction, and in line with our overall approach, comes equipped with a quantitative convergence theorem with the corresponding rate formulated in terms of abstract moduli. Concretely, we consider a generalized variant of the Robbins-Monro scheme taking values in a separable Hilbert space H (the expansion of Robbins-Monro algorithms to general Hilbert spaces was first considered in [84, 89] and then developed later in e.g. [100]) which, starting at some arbitrary initial point $x_0 \in H$, takes the form of a sequence (x_n) of H-valued random variables recursively defined by

$$x_{n+1} := x_n - a_n y_n \tag{RM}$$

where (y_n) another sequence of H-valued random variables and (a_n) a sequence of nonnegative reals. The classic Robbins-Monro scheme as discussed above is then just an instance of (RM) for $H = \mathbb{R}$ and $y_n = M(x_n) + \varepsilon_n$.

There are numerous strategies for establishing the convergence of (RM), and indeed, each of our three abstract convergence results can be applied in this context. Supermartingale convergence in the form of the Robbins-Siegmund theorem is possibly the most well known technique in this regard (as detailed in e.g. [60]), while Dvoretzky's theorem [32] was conceived specifically as a generalisation of the Robbins-Monro and Kiefer-Wolfowitz procedures. Approaches to convergence of (RM) focusing on stochastic quasi-Fejér monotonicity as a central principle essentially coincide with the introduction of stochastic quasi-Fejér monotonicity as a concept (see e.g. [34]), where stochastic gradient descent was one of the first algorithms studied from this perspective. Other well known approaches not covered in this paper include ODE methods, an overview of which can be found in [59].

With Theorem 6.1, we now present an abstract convergence result for (RM) based on our quantitative Robbins-Siegmund theorem (recall Theorem 3.4), although the core of our argument could also be used, mutatis mutandis, to present a similar such result through the use of the other abstract approaches to the convergence of stochastic approximation schemes, like e.g. stochastic quasi-Fejér monotonicity. The specific conditions we impose on (x_n) and (y_n) are essentially "reverse-engineered" to ensure that the premises of Theorem 3.4 are satisfied in this case, but nevertheless work very well as abstractions of various well-known conditions for establishing convergence of Robbins-Monro schemes, and we give several examples of this

in what follows. For details on how standard concepts like conditional expectations extend to Hilbert spaces, the reader is directed to [61].

Theorem 6.1. *Suppose that (x_n), (y_n) and (a_n) satisfy* (RM). *Let $z \in H$ and suppose that $L > 0$ satisfies $\|x_0 - z\|^2 < L$. Define $\mathcal{F}_n := \sigma(x_0, y_0, \dots, x_{n-1}, y_{n-1})$, i.e. $\mathcal{F}_n$ is the σ-algebra generated by $x_0, y_0, \dots, x_{n-1}, y_{n-1}$, and suppose that the following conditions are satisfied:*

(i) there exist $c, d > 0$ and a sequence of nonnegative random variables (d_n) satisfying $\sup_{n\in\mathbb{N}} \mathbb{E}[d_n] \leqslant d$ such that

$$\mathbb{E}[\|y_n\|^2 \mid \mathcal{F}_n] \leqslant c\,\|x_n - z\|^2 + d_n \text{ a.s.}$$

for all $n \in \mathbb{N}$;

(ii) $\langle x_n - z, \mathbb{E}[y_n \mid \mathcal{F}_n]\rangle \geqslant 0$ for all $n \in \mathbb{N}$;

(iii) there exists a function $\tau : (0,\infty) \to (0,\infty)$ such that

$$\mathbb{E}[\langle x_n - z, \mathbb{E}[y_n \mid \mathcal{F}_n]\rangle] < \tau(\varepsilon) \to \mathbb{E}[f(\|x_n - z\|^2)] < \varepsilon$$

for all $\varepsilon > 0$ and $n \in \mathbb{N}$, where f is a s.i.c.c. function with moduli ψ and κ;

(iv) $\sum_{n=0}^{\infty} a_n = \infty$ with rate of divergence θ and $\sum_{n=0}^{\infty} a_n^2 < M$ with rate of convergence χ, i.e.

$$\sum_{n=k}^{\theta(k,b)} a_n \geqslant b \text{ and } \sum_{n=\chi(\varepsilon)}^{\infty} a_n^2 < \varepsilon$$

for any $\varepsilon, b > 0$ and any $k \in \mathbb{N}$.

Then $\mathbb{E}[f(\|x_n - z\|^2)] \to 0$ with rate

$$\rho(\varepsilon) := \theta\left(\chi\left(\frac{\kappa\,(K_1\varepsilon)}{d}\right), \frac{K_2}{\tau(K_1\varepsilon)}\right)$$

for $K_1 := \frac{1}{2}\psi\left(e^{-cM}\right)$ and $K_2 := e^{cM}(L + dM)$ as well as $x_n \to z$ a.s. with rate

$$\rho'(\lambda, \varepsilon) := \rho(\lambda f(\varepsilon^2)).$$

Proof. We first observe that x_n is $\mathcal{F}_n$ measurable for all $n \in \mathbb{N}$, and thus

$$\begin{aligned}\mathbb{E}[\|x_{n+1} - z\|^2 \mid \mathcal{F}_n] &= \mathbb{E}[\|x_n - a_n y_n - z\|^2 \mid \mathcal{F}_n]\\ &= \mathbb{E}[\|x_n - z\|^2 \mid \mathcal{F}_n] - 2a_n\mathbb{E}[\langle x_n - z, y_n\rangle \mid \mathcal{F}_n] + a_n^2\mathbb{E}[\|y_n\|^2 \mid \mathcal{F}_n]\\ &= \|x_n - z\|^2 - 2a_n\langle x_n - z, \mathbb{E}[y_n \mid \mathcal{F}_n]\rangle + a_n^2\mathbb{E}[\|y_n\|^2 \mid \mathcal{F}_n]\end{aligned}$$

almost surely for any $n \in \mathbb{N}$. Using property (i) we therefore have

$$\mathbb{E}[\|x_{n+1} - z\|^2 \mid \mathcal{F}_n] \leqslant (1 + ca_n^2)\,\|x_n - z\|^2 - 2a_n\langle x_n - z, \mathbb{E}[y_n \mid \mathcal{F}_n]\rangle + a_n^2 d_n \text{ a.s.}$$

for any $n \in \mathbb{N}$. We now apply Theorem 3.4 for $X_n := \|x_n - z\|^2$ and $V_n := \langle x_n - z, \mathbb{E}[y_n \mid \mathcal{F}_n]\rangle$, by which the given rates follow. For that, we simply note that nonnegativity of V_n follows from (ii) and we have $\prod_{i=0}^{\infty}(1 + ca_n^2) < e^{cM}$ as well as

$$\sum_{n=k}^{\theta(k,b/2)} 2a_n \geqslant 2\sum_{n=k}^{\theta(k,b/2)} a_n \geqslant b$$

and

$$\sum_{n=\chi(\varepsilon/d)}^{\infty} \mathbb{E}[a_n^2 z_n] \leqslant d\sum_{n=\chi(\varepsilon/d)}^{\infty} a_n^2 < \varepsilon.$$

□

Setting $f = \sqrt{\cdot}$ the gives the following corollary:

Corollary 6.2. Assuming that condition (iii) of Theorem 6.1 is simplified to

$$\mathbb{E}[\langle x_n - z, \mathbb{E}[y_n \mid \mathcal{F}_n]\rangle] < \tau(\varepsilon) \to \mathbb{E}[\|x_n - z\|] < \varepsilon,$$

then the conclusions of Theorem 6.1 simplify to $\mathbb{E}[\|x_n - z\|] \to 0$ with rate

$$\rho(\varepsilon) := \theta\left(\frac{(K_1\varepsilon)^2}{d}, \frac{K_2}{\tau(K_1\varepsilon)}\right)$$

for $K_1 := \frac{1}{2}e^{-cM/2}$ and $K_2 := e^{cM}(L + dM)$ as well as $x_n \to z$ a.s. with rate $\rho'(\lambda, \varepsilon) := \rho(\lambda\varepsilon)$.

We now give a series of examples where we demonstrate how a number of standard instances of (RM) follow from Theorem 6.1, and more importantly, inherit rates of convergence formulated in terms of general moduli.

6.1. **Computing zeroes of set-valued monotone operators.** The canonical use of (RM) is to compute solutions to equations of the form $M(z) = 0$, where the function $M : H \to H$ typically satisfies some kind of strong monotonicity property, and y_n represents an estimator for $M(x_n)$. One can readily extend this to set-valued operators $M : H \to 2^H$ where we are then in turn, in this more general context, interested in solving inclusions of the form $0 \in M(z)$. In that case, y_n is then an estimator for an element of $M(x_n)$, which we represent abstractly via the condition

$$\mathbb{E}[y_n \mid \mathcal{F}_n] \in M(x_n).$$

Suppose now that $z \in \mathrm{zer}M$, i.e. $0 \in M(z)$. A rather general condition on the operator M which ensures that the zero z is unique, and which hence ensures condition (iii) above, is the τ-uniform monotonicity of M at the zero z, i.e. M satisfies

$$\forall (x, u) \in M\left(\langle x - z, u\rangle \geqslant \tau(\|x - z\|)\right),$$

with a modulus $\tau : [0, \infty) \to [0, \infty)$ which is convex, strictly increasing and satisfies $\tau(0) = 0$. This assumption is a particular special case of the notion of uniform monotonicity (see e.g. [8]), restricted to a particular zero at hand and we refer to Section 7.1 below where this notion, and in particular its generalization to Banach spaces and so-called accretive operators, is discussed in further detail. In any way, by Proposition 2.11, the conditions (ii) and (iii) of Theorem 6.1 immediately follow from this assumption, with the latter even in the simplified form of Corollary 6.2. Thus, assuming that (i) and (iv) also hold, both $\mathbb{E}[\|x_n - z\|] \to 0$ and $x_n \to z$ a.s. follow from Corollary 6.2 with corresponding rates as indicated therein.

For the special but still widely studied case that M is β-strongly monotone (see again e.g. [8]) in the sense that

$$\langle x - y, u - v\rangle \geqslant \beta \|x - y\|^2$$

whenever $(x, u), (y, v) \in M$, a simple adaptation of the proof of Theorem 6.1 puts us in the scope of fast rates as in Theorem 3.6: As in the proof of Theorem 6.1, we have

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \leqslant (1 + ca_n^2)X_n - 2a_nV_n + a_n^2 d_n \text{ a.s.}$$

for all $n \in \mathbb{N}$, where $X_n := \|x_n - z\|^2$ and $V_n := \langle x_n - z, \mathbb{E}[y_n \mid \mathcal{F}_n]\rangle$, but now, using the β-strong monotonicity of M, we can derive the stronger relation that $V_n \geqslant \beta X_n$ hold almost surely, and so in particular that $\mathbb{E}[V_n] \geqslant \beta\mathbb{E}[X_n]$ for all $n \in \mathbb{N}$. Defining

$$a_n := \frac{1}{\beta(n + r)} \quad \text{for} \quad r \geqslant \frac{2c}{\beta^2},$$

it is easy to show that

$$\mathbb{E}\left[a_n^2 d_n\right] \leqslant \frac{d}{\beta^2(n+r)^2} \quad \text{and} \quad ca_n^2 + \frac{3}{2(n+r)} \leqslant 2a_n\beta$$

and so Theorem 3.6 applies, yielding in particular

$$\mathbb{E}\left[\|x_n - z\|^2\right] \leqslant \frac{u}{n+r}$$

for all $n \in \mathbb{N}$ where $u \geqslant \max\{2d/\beta, rL\}$.

In this way, we derive the asymptotic estimate $\mathbb{E}[\|x_n - z\|] = O(1/\sqrt{n})$ for Robbins-Monro schemes on strongly monotone operators, further noting rather in passing that it is also applicable in general seperable Hilbert spaces, and so in particular under the conditions of e.g. Theorem 1 of [89].

6.2. **Stochastic subgradient methods.** The discussion of the previous section instantiates naturally to stochastic subgradient algorithms, first introduced in [36] (see also [92]). Starting as before on the more abstract level, suppose that $f : H \to \mathbb{R}$ is a τ-uniformly convex function, i.e. f satisfies

$$f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda)f(y) - \lambda(1-\lambda)\tau(\|x-y\|)$$

for all $x, y \in H$ and $\lambda \in [0,1]$. In that case, it is well-known (see e.g. [8]) that the subderivative $\partial f : H \to 2^H$, defined by

$$\partial f(x) := \{u \in H \mid f(y) - f(x) \geqslant \langle u, y - x\rangle \text{ for all } y \in H\},$$

is 2τ-uniformly monotone, and so in particular 2τ-uniformly monotone at a zero $z \in \operatorname{zer}\partial f = \operatorname{argmin} f$. The corresponding stochastic subgradient method

$$x_{n+1} := x_n - a_n y_n \quad \text{for} \quad \mathbb{E}[y_n \mid \mathcal{F}_n] \in \partial f(x_n) \tag{SG}$$

hence reduces to RM for $M = \partial f$ and so converges in mean and almost surely to the minimizer z of f by the discussion of the preceeding Section 6.1 (where condition (i) of Theorem 6.1 is in particular trivially satisfied under standard bounding assumptions such as $\mathbb{E}[\|y_n\|^2 \mid \mathcal{F}_n] \leqslant d$ for all $n \in \mathbb{N}$, cf. [92, Theorem 2.19]). In this way, standard convergence results for stochastic subgradient algorithms (in a very general setting) emerge naturally from our framework, and come equipped with rates.

In the special case that f is β-strongly convex, i.e. f satisfies

$$f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda)f(y) - \frac{\lambda(1-\lambda)}{2}\beta\|x-y\|^2$$

for all $x, y \in H$ and $\lambda \in [0,1]$, it is again well-known (see e.g. [8]) that ∂f is then β-strongly monotone, and so, similarly following the discussion of Section 6.1, we can derive that particular asymptotic estimate $\mathbb{E}[\|x_n - z\|] = O(1/\sqrt{n})$ for the subgradient algorithm (SG) under similar assumptions on the scalars in that case, where z is a minimizer of f as before. We thus also here reobtain standard convergence rates for stochastic subgradient methods (as discussed in e.g. [70]), noting that (SG) reduces to the standard stochastic gradient descent algorithm if f is differentiable.

6.3. **Robbins-Monro under Blum's conditions.** While the previous conditions such as uniform and strong monotonicity are certainly prevalent in the literature, various other generalist approaches to the convergence of the Robbins-Monro scheme commonly feature more involved assumptions on the mapping M which are of a certain local, pointwise, nature. For an example where this is the case, we give a brief discussion of Blum's conditions [15] which are perhaps the most prominent set of assumptions of that nature. Concretely, in the context of the simple scheme

$$x_{n+1} := x_n - a_n(M(x_n) + \varepsilon_n)$$

for an operator $M : \mathbb{R} \to \mathbb{R}$, the key regularity condition in that context commonly takes (following the presentation of [60]) the form of assuming

$$\inf_{\varepsilon \leqslant |x-z| \leqslant \varepsilon^{-1}} \{M(x)(x-z)\} > 0 \text{ a.s.}$$

for all $0 < \varepsilon < 1$. Now, working in a general seperable Hilbert space with a single-valued operator $M : H \to H$ and letting f be some s.i.c.c. function as in Theorem 6.1, pointwise conditions of this kind can be lifted to our general setting in a quantitative way by supposing that we have a modulus $\pi : (0,\infty)^2 \to (0,\infty)$ satisfying

$$f(\|x-z\|^2) \in [\varepsilon, l] \to \langle M(x), x-z\rangle \geqslant \pi(\varepsilon, l) \text{ a.s.}$$

for all $x \in H$ and $0 < \varepsilon < l$. Now letting (x_n) be defined as in (RM) for $y_n := M(x_n) + \varepsilon_n$ with $\mathbb{E}[\varepsilon_n \mid \mathcal{F}_n] = 0$, a modulus τ satisfying condition (iii) of Theorem 6.1 can be constructed using Proposition 2.13: Specifically, under the additional assumptions that $\mathbb{E}[f(\|x_n - z\|^2)] \leqslant K$ for some $K > 0$, along with the existence of a modulus of uniform integrability μ for the sequence $(f(\|x_n - z\|^2))$, i.e. μ satisfies

$$\forall A \in \mathcal{F}\ \forall \varepsilon > 0\ \forall n \in \mathbb{N}\ \big(\mathbb{P}(A) \leqslant \mu(\varepsilon) \to \mathbb{E}[f(\|x_n - z\|^2)\mathbf{1}_A] \leqslant \varepsilon\big),$$

it then follows from Proposition 2.13 that

$$\mathbb{E}[\langle x_n - z, M(x_n)\rangle] < \tau(\varepsilon) \to \mathbb{E}[f(\|x_n - z\|^2)] < \varepsilon$$

for all $\varepsilon > 0$ and $n \in \mathbb{N}$, with τ as defined in Proposition 2.13, which is exactly condition (iii) of Theorem 6.1 in this case (note also Remark 2.14 by which the modulus of uniform integrability is no longer required if $f(\|x_n - z\|^2) \leqslant K$ a.s. for all $n \in \mathbb{N}$).

In the special case of $H = \mathbb{R}$ and $f(x) = \sqrt{x}$, such a modulus π can be easily defined in terms of a function $\delta : (0,1) \to (0,\infty)$ witnessing the regularity condition of Blum via

$$\inf_{\varepsilon \leqslant |x-z| \leqslant \varepsilon^{-1}} \{M(x)(x-z)\} \geqslant \delta(\varepsilon) \text{ a.s.}$$

for all $0 < \varepsilon < 1$. Further, the standard conditions on the regression function M as stated in e.g. [60] can be easily shown to imply the remaining conditions of Theorem 6.1 in this case, where (ii) follows from the monotonicity of M and (i) from the standard bounding conditions

$$|M(x)| \leqslant c_1|x-z| + c_2 \quad \text{and} \quad \mathbb{E}[\varepsilon_n^2 \mid \mathcal{F}_n] \leqslant c_3$$

for all $x \in \mathbb{R}$ and $n \in \mathbb{N}$ with respective constants $c_1, c_2, c_3 > 0$, since in that latter case we in particular have

$$\begin{aligned}\mathbb{E}[y_n^2 \mid \mathcal{F}_n] &= M(x_n)^2 + 2M(x_n)\mathbb{E}[\varepsilon_n \mid \mathcal{F}_n] + \mathbb{E}[\varepsilon_n^2 \mid \mathcal{F}_n] \\ &\leqslant (c_1|x-z| + c_2)^2 + c_3 \\ &\leqslant c|x-z|^2 + d\end{aligned}$$

for suitable $c, d > 0$. Therefore our main abstract result can be used to provide a quantitative convergence theorem for the Robbins-Monro scheme under classic conditions.

7. Further applications to stochastic optimization

In this last section, we survey a range of further applications of our results to stochastic optimization and approximation. All of these examples are discussed and presented as instantiations of our abstract considerations of the convergence of stochastically quasi-Fejér monotone sequences under strong uniqueness assumptions in expectation, so that we also illustrate the applicability of these results, although in these context we could have similarly applied our quantitative variant of the Robbins-Siegmund theorem as we did in our presentation of the Robbins-Monro scheme. In that vein, we separately both present a range of examples of concrete optimization scenarios where moduli of strong uniqueness in expectation can be attained, as well as a range of concrete methods which are stochastically quasi-Fejér monotone and which further satisfy a corresponding approximation property.

7.1. **Examples of moduli of strong uniqueness in expectation.** We begin by surveying various concrete scenarios and assumptions which entail the strong uniqueness of the desired solution in expectation and in particular allow for the derivation of a corresponding modulus witnessing that property. In that context, our list of examples here is largely motivated by the range of examples of the so-called moduli of regularity studied in the context of quantitative results on ordinary Fejér monotone sequences as seminally pioneered and discussed in [54]. The cases surveyed in that paper provide further examples, not discussed here, where moduli of strong uniqueness in expectation naturally occur through classical, deterministic assumptions although the notion is certainly not limited to them and we refer to e.g. [4, 46, 69, 91] for further, intrinsically stochastic, assumptions which entail this special uniqueness property.

Fixed-point problems and quasi-contractions. Let (X, d) be a metric space and $T : X \to X$ with $\mathrm{Fix}T \neq \varnothing$. We are then interested in solving the equation $Tx = x$, i.e. we want to find a fixed point $x \in \mathrm{Fix}T$ of T. For that, assume additionally that T is a quasi-contraction, i.e. for some $z \in \mathrm{Fix}T$ it holds that

$$d(Tx, z) \leqslant rd(x, z)$$

for all $x \in X$, where $r \in [0, 1)$. This not only implies $\mathrm{Fix}T = \{z\}$ but further yields

$$d(x, z) \leqslant d(x, Tx) + d(Tx, z) \leqslant d(x, Tx) + rd(x, z).$$

So, defining $F(x) := d(x, Tx)$, we have $\mathrm{zer}F = \mathrm{Fix}T = \{z\}$ (so that the problem of finding zeros of F correspondingly represents the problem at hand) and the above implies

$$(1 - r)d(x, z) \leqslant F(x)$$

for all $x \in X$. Thus the assumption of Proposition 5.11 is satisfied with $\tau(\varepsilon) := (1 - r)\varepsilon$ which is thereby a modulus for z being a strongly unique zero in expectation for $F(x) := d(x, Tx)$.

Beyond its ubiquity in deterministic optimization, such (quasi-)contractivity assumptions for example feature in an essential way in the seminal work [28] on rates of convergence for stochastic block-coordinate methods or, more broadly, in the context of dynamic programming algorithms for reinforcement learning (see e.g. [45, 94]), just to name a few.

Minimization problems and functions with sharp minima. Let (X, d) be a metric space and $f : X \to (-\infty, +\infty]$ with $\mathrm{argmin}f \neq \varnothing$ be a function. We are then interested in solving the problem $\min_{x \in X} f(x)$, i.e. we want to find a minimizer $x \in \mathrm{argmin}f$ of f. For that, assume additionally that f has a τ-global sharp minimum, i.e. for some $z \in \mathrm{argmin}f$ it holds that

$$f(x) \geqslant \min f + \tau(d(x, z))$$

for all $x \in X$ where τ is convex, strictly increasing and satisfies $\tau(0) = 0$, a slightly stronger condition than the well-known notion of weak sharp minima pioneered in [22], see also the works

[21, 40, 62]. Again, this implies that $\operatorname{argmin} f = \{z\}$. Further, defining $F(x) := f(x) - \min f$, we have that $\operatorname{zer} F = \operatorname{argmin} f = \{z\}$ (so that the problem of finding zeros of F correspondingly represents the problem at hand) and

$$\tau(d(x,z)) \leqslant F(x)$$

for all $x \in X$. Thus the assumption of Proposition 5.11 is satisfied with τ, which is thereby a modulus for z being a strongly unique zero in expectation for $F(x) := f(x) - \min f$.

Note that the inverse τ^{-1} of τ satisfies

$$\tau^{-1}(f(x) - \min f) \geqslant d(x,z)$$

for all $x \in X$. Inequalities of that type are commonly called error bounds, as introduced in the seminal work [16], and we thereby in particular get that if f satisfies an error bound of the form

$$\tau(f(x) - \min f) \geqslant d(x,z)$$

for a concave and strictly increasing function with $\tau(0) = 0$, then τ^{-1} is a modulus for z being a strongly unique zero in expectation for $F(x) := f(x) - \min f$ as well.

Sharp minima (viz. error bounds) feature as crucial assumptions at various places in the (stochastic) approximation literature, especially also in the context of stochastically quasi-Fejér monotone iterations, and we refer to e.g. [38, 42, 44, 64, 90, 99, 101] for a few pointers to such circumstances beyond the works mentioned above.

A particularly prominent situation where a function f has sharp minima is in the case of uniformly quasiconvex functions (generalizing the discussion from Section 6.2 on minimizers of uniformly convex functions on Hilbert spaces), i.e. $f : X \to (-\infty, +\infty]$ with $\operatorname{argmin} f \neq \varnothing$ on a normed space $(X, \|\cdot\|)$ where

$$f(\lambda x + (1-\lambda)y) \leqslant \max\{f(x), f(y)\} - \lambda(1-\lambda)\tau(\|x-y\|)$$

for all $x, y \in X$ and $\lambda \in [0,1]$ for a convex and strictly increasing τ with $\tau(0) = 0$. Then for $z \in \operatorname{argmin} f$ and $x \in X$, we have

$$\min f \leqslant f\left(\frac{x+z}{2}\right) \leqslant \max\{f(x), f(z)\} - \frac{1}{4}\tau(\|x-z\|) \leqslant f(x) - \frac{1}{4}\tau(\|x-z\|)$$

and so

$$\frac{1}{4}\tau(\|x-z\|) \leqslant F(x) := f(x) - \min f$$

which implies that $\frac{1}{4}\tau$ is a respective modulus for z being a strongly unique zero in expectation for F. One specific such situation contains functions which are strongly quasiconvex, i.e. where

$$f(\lambda x + (1-\lambda)y) \leqslant \max\{f(x), f(y)\} - \lambda(1-\lambda)\frac{\mu}{2}\|x-y\|^2$$

in which case we simply have $\tau(\varepsilon) := \frac{\mu}{2}\varepsilon^2$ in the above. These considerations also generalize to nonlinear hyperbolic spaces.

Set-valued inclusion problems and uniformly accretive operators. Let X be a Banach space and $A : X \to 2^X$ be m-accretive with $\operatorname{zer} A \neq \varnothing$. We are then interested in solving the inclusion $0 \in Ax$, i.e. we want to find a zero $x \in \operatorname{zer} A$ of A. For that, assume additionally that A is τ-uniformly accretive at a zero $z \in \operatorname{zer} A$ (as already discussed in the context of Hilbert spaces in Section 6.1), i.e. it holds that

$$\langle x - z, u \rangle_+ \geqslant \tau(\|x-z\|)$$

for all $(x,u) \in A$, where τ is convex, strictly increasing and satisfies $\tau(0) = 0$ and

$$\langle x, y\rangle_+ := \max\{j(y) \mid j \in J(x)\}$$

where $J : X \to 2^{X^*}$ is the normalized duality map of X defined by

$$J(x) := \{j \in X^* \mid j(x) = \|x\|^2 = \|j\|^2\}.$$

This property implies $\mathrm{zer}A = \{z\}$ and we in particular have

$$\tau(\|x - z\|) \leqslant F(x) := \inf_{u \in Ax} \langle x - z, u\rangle_+$$

for all $x \in X$. Thus the assumption of Proposition 5.11 is satisfied with τ, which thereby is a modulus for z being a strongly unique zero in expectation for $F(x) := \inf_{u\in Ax}\langle x - z, u\rangle_+$ (so that the problem of finding zeros of F represents the problem of finding zeros of A in a suitable way). The above is in particular true if A is τ-uniformly accretive in the sense that

$$\langle x - y, u - v\rangle_+ \geqslant \tau(\|x - y\|)$$

for all $(x,u),(y,v) \in A$.

Further, the above also applies where A is a τ-strongly accretive operator, i.e. where

$$\langle x - y, u - v\rangle_+ \geqslant \tau(\|x - y\|)\,\|x - y\|$$

for all $(x,u),(y,v) \in A$, which in particular includes the case of β-strongly accretive operator, i.e. operators A where

$$\langle x - y, u - v\rangle_+ \geqslant \beta\,\|x - y\|^2$$

for all $(x,u),(y,v) \in A$ for some $\beta > 0$. In the latter two cases, we can also move to a function F which is defined independently of z as we in particular have

$$\tau(\|x - z\|) \leqslant \frac{\langle x - z, u\rangle_+}{\|x - z\|} \leqslant \|u\|$$

for all $(x,u) \in A$ with $\|x - z\| > 0$. Defining $F(x) := \mathrm{dist}(0, Ax)$ (the problem of finding zeros of which similarly representing the problem of finding zeros of A), we get

$$\tau(\|x - z\|) \leqslant \inf_{u\in Ax} \|u\| = \mathrm{dist}(0, Ax) = F(x)$$

for all $x \in \mathrm{dom}A$ with $\|x - z\| > 0$. We further get $\tau(\|x - z\|) \leqslant F(x)$ for all $x \in X$ as the inequality is trivial for $x \notin \mathrm{dom}A$ or for $x \in \mathrm{dom}A$ with $\|x - z\| = 0$. Thus the assumption of Proposition 5.11 is satisfied with τ, which thereby is a modulus for z being a strongly unique zero in expectation for $F(x) := \mathrm{dist}(0, Ax)$.

Uniform accretivity (viz. monotonicity) assumptions are, similarly to uniform convexity assumptions, rather prominent in the (stochastic) approximation literature, and besides the Robbins-Monro procedure as presented in Section 6 and the works on stochastic splitting methods [23, 27, 77, 88, 97] mentioned in the introduction, we in particular further refer to works such as [14, 44] as representative, more genuinely stochastic, examples.

7.2. **Examples of stochastically quasi-Fejér monotone iterations.** We now discuss a variety of iterations commonly studied in stochastic approximation and optimization which are stochastically quasi-Fejér monotone in the above sense and where we can naturally obtain a corresponding lim inf-modulus.

A Krasnoselskii-Mann iteration with stochastic noise. Let (X, d) be a metric space and let $T : X \to X$ be a given nonexpansive mapping, i.e. T satisfies

$$d(Tx, Ty) \leqslant d(x, y)$$

for all $x, y \in X$. We are interested in stochastically approximating fixed points of T, i.e. finding elements of the set $\mathrm{Fix}T$ which we in the following assume to be a singleton, i.e. $\mathrm{Fix}T = \{z\}$. Rephrasing this in the above formalism, we have $\mathrm{Fix}T = \mathrm{zer}F$ for $F(x) := d(Tx, x)$.

Naturally, most methods for solving the respective problem require a certain amount of geometry on the space. In that vein, we here assume that the space (X, d) is a CAT(0)-space.[6] The prominent CAT(0)-spaces, introduced by Aleksandrov [1] and named as such by Gromov [43], are commonly defined as geodesic spaces with a nonpositive curvature as captured by the Bruhat-Tits CN-inequality [20]. Complete CAT(0)-spaces are also referred to as Hadamard spaces. Examples of such spaces in particular include Hilbert spaces, $\mathbb{R}$-trees as well as complete simply connected Riemannian manifolds with nonpositive sectional curvature, among many more (as also discussed in the introduction already), and hence provide a strong common metric generalization to study these various objects. We refer in particular to the seminal monographs [11, 19] for further details on these spaces and here only discuss the bare necessities required for the paper: Crucially, if the reader prefers, they can simply imagine everything that follows taking place in a Hilbert space, in which case the various inequalities that follow are completely standard.

At first, we deem it helpful to follow a slightly different approach to these spaces. Namely, more axiomatically, CAT(0)-spaces can also be equivalently characterized as a special subclass of so-called hyperbolic spaces in the sense of Kohlenbach [49] which are metric spaces (X, d) endowed with a further mapping $W : X \times X \times [0, 1] \to X$ that, via $W(x, y, \lambda)$, represents an abstract access to a convex combination in X essentially replacing the geodesics (see [49] for further details). Accordingly, in the following we utilize the intuitive notation $(1 - \lambda)x \oplus \lambda y$ for $x, y \in X$ and $\lambda \in [0, 1]$ to denote a convex combination of x and y formally represented by $W(x, y, \lambda)$. This convex combination satisfies various intuitive properties with respect to the metric, notably

$$d((1 - \lambda)x \oplus \lambda y, z) \leqslant (1 - \lambda)d(x, z) + \lambda d(y, z).$$

In the context of hyperbolic spaces, CAT(0)-spaces arise equivalently as those hyperbolic spaces which satisfy the so-called CN^--inequality (which is in that context actually equivalent to the CN-inequality), given by

$$d^2\left(\frac{1}{2}x \oplus \frac{1}{2}y, z\right) \leqslant \frac{1}{2}d^2(x, z) + \frac{1}{2}d^2(y, z) - \frac{1}{4}d^2(x, y)$$

for all $x, y, z \in X$ and $\lambda \in [0, 1]$. Alternatively, CAT(0)-spaces can be characterized by the so-called quasi-linearization function

$$\langle \overrightarrow{xy}, \overrightarrow{uv} \rangle := \frac{1}{2}\left(d^2(x, v) + d^2(y, u) - d^2(x, u) - d^2(y, v)\right)$$

for $x, y, u, v \in X$ from the work of Berg and Nikolaev [12], wherein it is shown that CAT(0)-spaces are precisely those metric spaces which satisfy the following nonlinear variant of the Cauchy-Schwarz inequality:

$$\langle \overrightarrow{xy}, \overrightarrow{uv} \rangle \leqslant d(x, y)d(u, v). \tag{$*$)$_1$}$$

[6]While this class of spaces does not exhaust the generality in which the method studied in this part, the so-called Krasnoselskii-Mann iteration, can be studied, we here nevertheless focus on this prominent class of spaces as it still illustrates a wide generalization beyond linear spaces and still allows for a smooth technical development of the method.

The only other property that we will here require of the quasi-linearization function is that it satisfies the following equality

$$(*)_2 \qquad d^2(x,z) = d^2(x,y) + d^2(y,z) + 2\langle \overrightarrow{xy}, \overrightarrow{yz} \rangle,$$

which immediately follows from the above definition. Lastly, we will require in the following that the CN^--inequality extends beyond the midpoint (see e.g. Lemma 2.5 in [30]), i.e. that

$$(*)_3 \qquad d^2((1-\lambda)x \oplus \lambda y, z) \leqslant (1-\lambda)d^2(x,z) + \lambda d^2(y,z) - \lambda(1-\lambda)d^2(x,y).$$

holds for all $x, y, z \in X$ and $\lambda \in [0,1]$.

In the context of CAT(0)-spaces, we then can employ the following variant of the well-known Krasnoselskii-Mann iteration [56, 65] with stochastic errors to approximate the fixed points of T: given an initial X-valued random variable x_0, we define recursively

$$\begin{cases} x_{n+1} := (1-\lambda_n)x_n \oplus \lambda_n y_n \\ \text{with } y_n \text{ such that } d(y_n, Tx_n) \leqslant \varepsilon_n \text{ a.s.} \end{cases}$$

where (ε_n) is a sequence of nonnegative real valued random variables. Both in a linear and nonlinear context, this methods belongs to one of the most well-studied procedures for approximating fixed points and is consequently widely studied. We mainly refer to the works [8, 24, 26] for further reading, with a particular emphasis on [26] for the presentation, phrased over linear spaces, of a variant with stochastic noise similar to the above. Indeed, writing $y_n = Tx_n + e_n$ over a linear space X for an X-valued noise sequence (e_n) as in [26] reduces the method studied therein to the above iteration (with $\varepsilon_n = \|e_n\|$).

For an analysis of the above method, we note first that we have

$$\begin{aligned} d(x_{n+1}, z) &\leqslant (1-\lambda_n)d(x_n, z) + \lambda_n d(y_n, z) \\ &\leqslant (1-\lambda_n)d(x_n, z) + \lambda_n d(Tx_n, z) + \lambda_n d(y_n, Tx_n) \\ &\leqslant d(x_n, z) + \lambda_n \varepsilon_n \end{aligned}$$

by the properties of CAT(0)-spaces and using that T is nonexpansive, so that for $\mathcal{F}_n := \sigma(x_0, \ldots, x_n)$, i.e. the σ-algebra generated by $x_0, \ldots, x_n$, as before, we have

$$\mathbb{E}[d(x_{n+1}, z) \mid \mathcal{F}_n] \leqslant d(x_n, z) + \lambda_n \sqrt{\mathbb{E}[\varepsilon_n^2 \mid \mathcal{F}_n]}$$

almost surely for all $n \in \mathbb{N}$, which yields that the sequence is stochastically quasi-Fejér monotone w.r.t. the filtration $(\mathcal{F}_n)$. We in the following assume the existence of a $K > 0$ such that $\mathbb{E}[d^2(x_n, z)] \leqslant K$.

Using the properties $(*)_1$ – $(*)_3$ of CAT(0)-spaces discussed above, along with standard properties of the quasi-linearization function such as $\langle \overrightarrow{xy}, \overrightarrow{uv} \rangle = -\langle \overrightarrow{xy}, \overrightarrow{vu} \rangle$, $\langle \overrightarrow{xy}, \overrightarrow{uv} \rangle = \langle \overrightarrow{uv}, \overrightarrow{xy} \rangle$,

$\langle \overrightarrow{xy}, \overrightarrow{xy} \rangle = d^2(x,y)$ and $\langle \overrightarrow{xy}, \overrightarrow{uv} \rangle + \langle \overrightarrow{xy}, \overrightarrow{vw} \rangle = \langle \overrightarrow{xy}, \overrightarrow{uw} \rangle$, we get

$$\begin{aligned}
d^2(x_{n+1}, z) &\leqslant (1-\lambda_n) d^2(x_n, z) + \lambda_n d^2(y_n, z) - \lambda_n(1-\lambda_n) d^2(x_n, y_n) \\
&\leqslant d^2(x_n, z) - \lambda_n(1-\lambda_n) d^2(x_n, Tx_n) + \lambda_n d^2(y_n, Tx_n) + 2\lambda_n \left\langle \overrightarrow{y_n Tx_n}, \overrightarrow{Tx_n z} \right\rangle \\
&\quad - \lambda_n(1-\lambda_n) d^2(y_n, Tx_n) - 2\lambda_n(1-\lambda_n) \left\langle \overrightarrow{y_n Tx_n}, \overrightarrow{Tx_n x_n} \right\rangle \\
&\leqslant d^2(x_n, z) - \lambda_n(1-\lambda_n) d^2(x_n, Tx_n) + \lambda_n^2 d^2(y_n, Tx_n) + 2\lambda_n \left\langle \overrightarrow{y_n Tx_n}, \overrightarrow{Tx_n z} \right\rangle \\
&\quad - 2\lambda_n(1-\lambda_n) \left\langle \overrightarrow{y_n Tx_n}, \overrightarrow{Tx_n z} \right\rangle - 2\lambda_n(1-\lambda_n) \left\langle \overrightarrow{y_n Tx_n}, \overrightarrow{z x_n} \right\rangle \\
&\leqslant d^2(x_n, z) - \lambda_n(1-\lambda_n) d^2(x_n, Tx_n) + \lambda_n^2 d^2(y_n, Tx_n) + 2\lambda_n d(y_n, Tx_n) d(Tx_n, z) \\
&\quad + 2\lambda_n d(y_n, Tx_n) d(x_n, z) \\
&\leqslant d^2(x_n, z) - \lambda_n(1-\lambda_n) d^2(x_n, Tx_n) + \lambda_n \varepsilon_n^2 + 4\lambda_n \varepsilon_n d(x_n, z)
\end{aligned}$$

almost surely for any $n \in \mathbb{N}$. In particular, integrating the inequality and applying Hölder's inequality yields

$$\begin{aligned}
\mathbb{E}[d^2(x_{n+1}, z)] &\leqslant \mathbb{E}[d^2(x_n, z)] - \lambda_n(1-\lambda_n)\mathbb{E}[d^2(Tx_n, x_n)] + 4\lambda_n \mathbb{E}[\varepsilon_n d(x_n, z)] + \lambda_n \mathbb{E}[\varepsilon_n^2] \\
&\leqslant \mathbb{E}[d^2(x_n, z)] - \lambda_n(1-\lambda_n)\mathbb{E}[d^2(Tx_n, x_n)] + 4\sqrt{\mathbb{E}[d^2(x_n, z)]}\lambda_n \sqrt{\mathbb{E}[\varepsilon_n^2]} + \lambda_n \mathbb{E}[\varepsilon_n^2] \\
&\leqslant \mathbb{E}[d^2(x_n, z)] - \lambda_n(1-\lambda_n)\mathbb{E}[d^2(Tx_n, x_n)] + 4\sqrt{K}\lambda_n \sqrt{\mathbb{E}[\varepsilon_n^2]} + \lambda_n \mathbb{E}[\varepsilon_n^2].
\end{aligned}$$

Applying Lemma 3.2, we get $\sum_{n=0}^{\infty} \lambda_n(1-\lambda_n)\mathbb{E}[d^2(Tx_n, x_n)] < L := K + M(4\sqrt{K}+1)$ for $M > 0$ such that $\sum_{n=0}^{\infty} \lambda_n \sqrt{\mathbb{E}[\varepsilon_n^2]} < M$ and assuming $\mathbb{E}[\varepsilon_n^2] \leqslant 1$ for all $n \in \mathbb{N}$ (similar to [26]). Lemma 3.3 together with Jensen's inequality then yields

$$\forall \varepsilon > 0 \; \forall N \in \mathbb{N} \; \exists n \in [N; \theta(N, L/\varepsilon^2)] (\mathbb{E}[F(x_n)] < \varepsilon)$$

for a rate of divergence $\theta : \mathbb{N} \times (0, \infty) \to \mathbb{N}$ for $\sum_{n=0}^{\infty} \lambda_n(1-\lambda_n) = \infty$, i.e.

$$\forall b > 0 \; \forall k \in \mathbb{N} \left(\sum_{n=k}^{\theta(k,b)} \lambda_n(1-\lambda_n) \geqslant b \right).$$

Theorem 5.8 can then be used in combination with any modulus of strong uniqueness in expectation for $\mathrm{Fix}T = \mathrm{zer}F = \{z\}$ with $F(x) = d(Tx, x)$ (together with a rate of convergence for the summability of the errors in expectation) to produce a rate of convergence for the above iteration.

A proximal point algorithm with stochastic noise. Let $(X, \langle \cdot, \cdot \rangle)$ be a Hilbert space and let $A : X \to 2^X$ be a given maximally monotone operator, i.e. A satisfies

$$\langle x - y, u - v \rangle \geqslant 0$$

for all $(x,u), (y,v) \in A$ and no proper extension of A satisfies this property. We are interested in stochastically approximating zeros of A, i.e. finding elements of the set $\mathrm{zer}A$. Similar to before, we assume that $\mathrm{zer}A = \{z\}$ and further write $\mathrm{zer}A = \mathrm{zer}F$ for $F(x) := \mathrm{dist}(0, Ax)$. The most prominent method for approximating zeros of such operators is the proximal point method [66, 87] with utilizes the so-called resolvent,

$$J_{\gamma A} := (\mathrm{Id} + \gamma A)^{-1}$$

which is single-valued as A is monotone, to define a Picard-like iteration. Here, we consider the following variant incorporating stochastic errors to model inaccuracies in the evaluation of the resolvent: given an initial X-valued random variable x_0, we define recursively

$$x_{n+1} := J_{\gamma_n A} x_n + \varepsilon_n,$$

where (ε_n) is a sequence of X-valued random variables and $(\gamma_n) \subseteq (0, \infty)$ is a sequence of parameters with $\inf\{\gamma_n \mid n \in \mathbb{N}\} \geqslant \underline{\gamma} > 0$. The analysis of this method is now similar to that of the Krasnoselskii-Mann iteration. For one, utilizing that the resolvent is a nonexpansive and total mapping (as A is maximally monotone) with $\mathrm{Fix}(J_{\gamma A}) = \mathrm{zer} A$, analogously to before we get

$$\mathbb{E}[\|x_{n+1} - z\| \mid \mathcal{F}_n] \leqslant \|x_n - z\| + \sqrt{\mathbb{E}[\|\varepsilon_n\|^2 \mid \mathcal{F}_n]}$$

almost surely for all $n \in \mathbb{N}$, where $\mathcal{F}_n := \sigma(x_0, \ldots, x_n)$, i.e. $\mathcal{F}_n$ is the σ-algebra generated by $x_0, \ldots, x_n$, which yields that the sequence is stochastically quasi-Fejér monotone w.r.t. the filtration $(\mathcal{F}_n)$. As before, in the following we assume the existence of a $K > 0$ such that $\mathbb{E}[\|x_n - z\|^2] \leqslant K$.

Next, we recall the fact that the resolvent is not only nonexpansive, but even firmly nonexpansive in the sense that

$$\|J_{\gamma A} x - J_{\gamma A} y\|^2 + \|(x - J_{\gamma A} x) - (y - J_{\gamma A} y)\|^2 \leqslant \|x - y\|^2$$

for any $\gamma > 0$ and $x, y \in X$ (following the definitions of [8]). Similar to before, we then get

$$\begin{aligned} \|x_{n+1} - z\|^2 &\leqslant (\|J_{\gamma_n A} x_n - z\| + \|\varepsilon_n\|)^2 \\ &= \|J_{\gamma_n A} x_n - z\|^2 + 2\,\|J_{\gamma_n A} x_n - z\|\,\|\varepsilon_n\| + \|\varepsilon_n\|^2 \\ &\leqslant \|x_n - z\|^2 - \|x_n - J_{\gamma_n A} x_n\|^2 + 2\,\|x_n - z\|\,\|\varepsilon_n\| + \|\varepsilon_n\|^2 \end{aligned}$$

almost surely for any $n \in \mathbb{N}$, which by Hölder's inequality yields

$$\mathbb{E}[\|x_{n+1} - z\|^2] \leqslant \mathbb{E}[\|x_n - z\|^2] - \mathbb{E}[\|x_n - J_{\gamma_n A} x_n\|^2] + 2\sqrt{K}\sqrt{\mathbb{E}[\|\varepsilon_n\|^2]} + \mathbb{E}[\|\varepsilon_n\|^2]$$

similarly to before. Applying Lemma 3.2, we get

$$\sum_{n=0}^{\infty} \mathbb{E}[\|x_n - J_{\gamma_n A} x_n\|^2] < L := K + M(2\sqrt{K} + 1)$$

for $M > 0$ such that $\sum_{n=0}^{\infty} \sqrt{\mathbb{E}[\|\varepsilon_n\|^2]} < M$ and assuming $\mathbb{E}[\|\varepsilon_n\|^2] \leqslant 1$ for all $n \in \mathbb{N}$, also analogously to the previous section. Lemma 3.3 together with Jensen's inequality then yields

$$\forall \varepsilon > 0\ \forall N \in \mathbb{N}\ \exists n \in [N; N + L/\varepsilon^2](\mathbb{E}[\|x_n - J_{\gamma_n A} x_n\|] < \varepsilon).$$

In particular, using the resolvent inequality

$$\|J_{\lambda\gamma A} x - x\| \leqslant (2 - \lambda)\,\|J_{\gamma A} x - x\|$$

for $\gamma, \lambda > 0$ with $\lambda \leqslant 1$ (see e.g. Proposition 23.31 in [8]), we have

$$\left\|J_{\underline{\gamma} A} x - x\right\| = \left\|J_{\frac{\underline{\gamma}}{\gamma_n}\gamma_n A} x - x\right\| \leqslant \left(2 - \frac{\underline{\gamma}}{\gamma_n}\right)\|J_{\gamma_n A} x - x\| \leqslant 2\,\|J_{\gamma_n A} x - x\|$$

so that the above actually implies

$$\forall \varepsilon > 0\ \forall N \in \mathbb{N}\ \exists n \in [N; N + 4L/\varepsilon^2]\left(\mathbb{E}\left[\left\|x_n - J_{\underline{\gamma} A} x_n\right\|\right] < \varepsilon\right).$$

Theorem 5.8 can then be used in combination with any modulus τ of strong uniqueness in expectation for $\mathrm{zer} A = \mathrm{zer} F = \{z\}$ with $F(x) = \mathrm{dist}(0, Ax)$ (together with a rate of convergence

for the absolute summability of the errors in expectation) to produce a rate of convergence for the above iteration. To see this, note that we have to use the following slightly modified construction: Instead of F, we consider the map $F'(x) = \|x - J_{\underline{\gamma}A}x\|$. The above then shows that

$$\forall \varepsilon > 0\ \forall N \in \mathbb{N}\ \exists n \in [N; N + 4L/\varepsilon^2](\mathbb{E}[F'(x_n)] < \varepsilon).$$

Further, the modulus τ can be converted to a modulus of strong uniqueness in expectation for $\mathrm{zer}A = \mathrm{zer}F' = \{z\}$ with the above F'. To see that, note that $\underline{\gamma}^{-1}(x - J_{\underline{\gamma}A}x) \in A(J_{\underline{\gamma}A}x)$ and so

$$\mathrm{dist}(0, A(J_{\underline{\gamma}A}x)) \leqslant \underline{\gamma}^{-1} \left\| x - J_{\underline{\gamma}A}x \right\|$$

for any $x \in X$. Therefore, if $\mathbb{E}[F'(x)] < \min\{\tau(\varepsilon/2) \cdot \underline{\gamma}, \varepsilon/2\}$, we have $\mathbb{E}[F(J_{\underline{\gamma}A}x)] < \tau(\varepsilon/2)$. As τ is a modulus relative to F, we have $\mathbb{E}[\|J_{\underline{\gamma}A}x - z\|] < \varepsilon/2$ and hence

$$\mathbb{E}[\|x - z\|] \leqslant \mathbb{E}\left[\left\|J_{\underline{\gamma}A}x - z\right\|\right] + \mathbb{E}\left[\left\|J_{\underline{\gamma}A}x - x\right\|\right] < \varepsilon.$$

Hence $\min\{\tau(\varepsilon/2) \cdot \underline{\gamma}, \varepsilon/2\}$ is a modulus of strong uniqueness in expectation for the alternative F' and Theorem 5.8 can be used with that F' instead.

A stochastic metric splitting algorithm. Let (X, d) be a CAT(0)-space (recall the previous discussion) and $f : X \to (-\infty, +\infty]$ be a function of the form

$$f = \sum_{n=1}^{N} f_n$$

where each $f_n : X \to (-\infty, +\infty]$ is convex and lower-semicontinuous. We want to find minimizers of said function but instead of proceeding in the usual style of the proximal point algorithm, where we refer to the excellent exposition in [9], and utilizing (appropriate metric analogs of) the resolvent of f directly, which might be costly, we follow the splitting approach of [10] and study the algorithm

$$x_{k+1} := J_{\lambda_k}^{r_k}(x_k)$$

for an initial point $x_0 \in X$, parameters (λ_k) with $\sum_{k=0}^{\infty} \lambda_k = \infty$ and $\sum_{k=0}^{\infty} \lambda_k^2 < \infty$, and (r_k) a sequence of pairwise independent random variables taking values in $\{1, \ldots, N\}$ according to the uniform distribution which facilitates a random order selection of the individual resolvents

$$J_\lambda^n(x) := \mathrm{argmin}_{y \in X} \left(f_n(y) + \frac{1}{2\lambda} d^2(x, y) \right).$$

This kind of stochastic splitting-type proximal point algorithm is, although phrased in terms of CAT(0)-spaces and for a finite collection of convex functions, structurally similar to variants of the proximal point algorithm which include stochastic perturbation of the involved set-valued operator or convex function as e.g. studied in [4, 14] which, as mentioned in the introduction, will be studied in more detail through the lens of this paper in subsequent work.

As shown in Lemma 3.6 in [10], if $\mathrm{argmin}(f) \neq \varnothing$ and we have

$$f_n(x_k) - f_n(x_{k+1}^n) \leqslant L d(x_k, x_{k+1}^n) \text{ a.s.}$$

for some $L > 0$, where x_{k+1}^n is the result of the iteration with x_k if $r_k = n$ (which in particular is validated if the functions f_n are locally Lipschitz, see Remark 3.8 in [10]), then it holds that

$$\mathbb{E}[d^2(x_{k+1}, z) \mid \mathcal{F}_k] \leqslant d^2(x_k, z) - \frac{2\lambda_k}{N}(f(x_k) - \min f) + 4\lambda_k^2 L^2 \text{ a.s.}$$

for all $k \in \mathbb{N}$ and $z \in \operatorname{argmin} f$, where $\mathcal{F}_k = \sigma(x_0, \ldots, x_k)$, i.e. $\mathcal{F}_n$ is the σ-algebra generated by $x_0, \ldots, x_k$. In particular, the sequence is stochastically quasi-Fejér monotone w.r.t. $(\mathcal{F}_k)$. Again, integrating this inequality yields

$$\mathbb{E}[d^2(x_{k+1}, z)] \leqslant \mathbb{E}[d^2(x_k, z)] - \frac{2\lambda_k}{N}(\mathbb{E}[f(x_k) - \min f]) + 4\lambda_k^2 L^2$$

for all $k \in \mathbb{N}$, which by Lemma 3.2 yields $\sum_{k=0}^{\infty} \frac{2\lambda_k}{N}(\mathbb{E}[f(x_k) - \min f]) < L := K + 4ML^2$ for $K \geqslant \mathbb{E}[d^2(x_k, z)]$ and $M \geqslant \sum_{k=0}^{\infty} \lambda_k^2$. Lemma 3.3 then yields

$$\forall \varepsilon > 0 \; \forall N \in \mathbb{N} \; \exists n \in [N; \theta(N, L/\varepsilon)](\mathbb{E}[F(x_n)] < \varepsilon)$$

for $F(x) := f(x) - \min f$ and for a rate of divergence $\theta : \mathbb{N} \times (0, \infty) \to \mathbb{N}$ for $\sum_{n=0}^{\infty} \frac{2\lambda_k}{N} = \infty$, i.e.

$$\forall b > 0 \; \forall k \in \mathbb{N} \left(\sum_{n=k}^{\theta(k,b)} \frac{2\lambda_k}{N} \geqslant b \right).$$

As before, Theorem 5.8 can now be used in combination with any modulus of strong uniqueness in expectation for $\operatorname{argmin} f = \operatorname{zer} F = \{z\}$ with $F(x) = f(x) - \min f$ (together with a rate of convergence for the summability of the errors in expectation) to produce a rate of convergence for the above iteration. Such a modulus of strong uniqueness can in particular be given in the special case where the above method is used to compute Fréchet means in Hadamard spaces X, that is finding the (unique, as will be discussed below) point

$$\operatorname{argmin}_{x \in X} \sum_{n=1}^{N} w_n d^2(x, a_n),$$

given finitely many weights $w_1, \ldots, w_N$ with $\sum_{n=1}^{N} w_n = 1$ and points $a_1, \ldots, a_N \in X$. As discussed in [10], this problem plays a central role in quite a few different and highly relevant circumstances such as computational biology, more concretely phylogenetics (via a particular Hadamard space known as the Billera-Holmes-Vogtmann tree), diffusion tensor imaging or consensus algorithms. We refer to the discussions in [10] for more details on these applied aspects of Fréchet means.

In any way, as shown in Theorem 2.4 in [10], there exists a unique point $z \in X$ minimizing the above expression which moreover satisfies

$$d^2(x, z) + \sum_{n=1}^{N} w_n d^2(z, a_n) \leqslant \sum_{n=1}^{N} w_n d^2(x, a_n)$$

for any $x \in X$, i.e. we have $d^2(x, z) \leqslant F(x) := f(x) - \min f$ for

$$f(x) := \sum_{n=1}^{N} w_n d^2(x, a_n).$$

Hence, using Jensen's inequality, we have $(\mathbb{E}[d(x, z)])^2 \leqslant \mathbb{E}[d^2(x, z)] \leqslant \mathbb{E}[F(x)]$ for any $x \in X$ so that $\tau(\varepsilon) := \varepsilon^2$ is a modulus of strong uniqueness for the solution z.

Acknowledgments: The authors want to thank Miroslav Bačák and Paulo Oliva for helpful discussions on the topic of this paper. The first author was partially supported by the EPSRC Centre for Doctoral Training in Digital Entertainment EP/L016540/1, and the third author was partially supported by the EPSRC grant EP/W035847/1.

References

[1] A.D. Aleksandrov. A theorem on triangles in a metric space and some of its applications. *Trudy Matematicheskogo Instituta imeni V.A. Steklova*, 38:5–23, 1951.

[2] M. Arnaudon, C. Dombry, A. Phan, and L. Yang. Stochastic algorithms for computing means of probability measures. *Stochastic Processes and their Applications*, 122(4):1437–1455, 2012.

[3] R. Arthan and P. Oliva. On the Borel-Cantelli Lemma, the Erdős-Rényi Theorem, and the Kochen-Stone Theorem. *Journal of Logic and Analysis*, 13(6):1–23, 2021.

[4] H. Asi and J.C. Duchi. Stochastic (Approximate) Proximal Point Methods: Convergence, Optimality, and Adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

[5] J. Avigad, E.T. Dean, and J. Rute. A metastable dominated convergence theorem. *Journal of Logic and Analysis*, 4(3):1–19, 2012.

[6] J. Avigad, P. Gerhardy, and H. Towsner. Local stability of ergodic averages. *Transactions of the American Mathematical Society*, 362(1):261–288, 2010.

[7] K. Barty, J.-S. Roy, and C. Strugarek. Hilbert-valued perturbed subgradient algorithms. *Mathematics of Operations Research*, 32(3):551–562, 2007.

[8] H.H. Bauschke and P.L. Combettes. *Convex Analysis and Montone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer Cham, 2nd edition, 2017.

[9] M. Bačák. The proximal point algorithm in metric spaces. *Israel Journal of Mathematics*, 194:689–701, 2013.

[10] M. Bačák. Computing medians and means in Hadamard spaces. *SIAM Journal on Optimization*, 24(3):1542–1566, 2014.

[11] M. Bačák. *Convex Analysis and Optimization in Hadamard Spaces*, volume 22 of *Series in Nonlinear Analysis and Applications*. Berlin, München, Boston: De Gruyter, 2014.

[12] I.D. Berg and I.G. Nikolaev. Quasilinearization and curvature of Aleksandrov spaces. *Geometriae Dedicata*, 133:195–218, 2008.

[13] D.P. Bertsekas and J.N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

[14] P. Bianchi. Ergodic Convergence of a Stochastic Proximal Point Algorithm. *SIAM Journal on Optimization*, 26(4):2235–2260, 2016.

[15] J.R. Blum. Approximation methods which converge with probability one. *Annals of Mathematical Statistics*, 25:382–386, 1954.

[16] J. Bolte, T.P. Nguyen, J. Peypouquet, and B.W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming, Series A*, 165:471–507, 2017.

[17] S. Bonnabel. Stochastic Gradient Descent on Riemannian Manifolds. *IEEE Transaction on Automatic Control*, 58(9):2217–2229, 2013.

[18] S. Bonnabel. Averaging Stochastic Gradient Descent on Riemannian Manifolds. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 650–687, 2018.

[19] M.R. Bridson and A. Haefliger. *Metric Spaces of Non-Positive Curvature*, volume 319 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin, Heidelberg, 1999.

[20] F. Bruhat and J. Tits. Groupes réductifs sur un corps local. I. Données radicielles valuées. *Publications Mathématiques de l'Institut des Hautes Études Scientifiques*, 41:5–251, 1972.

[21] J.V. Burke and S. Deng. Weak sharp minima revisited. I. Basic theory. *Control and Cybernetics*, 31:439–469, 2002.

[22] J.V. Burke and M.C. Ferris. Weak sharp minima in mathematical programming. *SIAM Journal on Control and Optimization*, 31:1340–1359, 1993.

[23] V. Cevher, B.C. Vũ, and A. Yurtsever. Stochastic Forward Douglas-Rachford Splitting Method for Monotone Inclusions. In P. Giselsson and A. Rantzer, editors, *Large-Scale and Distributed Optimization*, volume 2227 of *Lecture Notes in Mathematics*, pages 149–179. Springer, Cham, 2018.

[24] P.L. Combettes. Quasi-Fejérian analysis of some optimization algorithms. In D. Butnariu, Y. Censor, and S. Reich, editors, *Inherently Parallel Algorithms for Feasibility and Optimization*, pages 115–152. Elsevier, New York, 2001.

[25] P.L. Combettes. Fejér monotonicity in convex optimization. In C.A. Floudas and P.M. Pardalos, editors, *Encyclopedia of Optimization*, pages 1016–1024. Springer, New York, 2nd edition, 2009.

[26] P.L. Combettes and J.-C. Pesquet. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, 25(2):1221–1248, 2015.
[27] P.L. Combettes and J.-C. Pesquet. Stochastic approximations and perturbations in forward-backward splitting for monotone operators. *Pure and Applied Functional Analysis*, 1(1):13–37, 2016.
[28] P.L. Combettes and J.-C. Pesquet. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping II: mean-square and linear convergence. *Mathematical Programming , Series B*, 174:433–451, 2019.
[29] C. Derman and J. Sacks. On Dvoretsky's stochastic approximation theorem. *Annals of Mathematical Statistics*, 30(2):601–606, 1959.
[30] S. Dhompongsa and B. Panyanak. On Δ-convergence theorems in CAT(0) spaces. *Computers & Mathematics with Applications*, 56(10):2572–2579, 2008.
[31] M. Duflo. *Random Iterative Models*, volume 34 of *Stochastic Modelling and Applied Probability*. Springer Berlin Heidelberg, 1997.
[32] A. Dvoretzky. On stochastic approximation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 39–56. University of California Press, 1956.
[33] A. Dvoretzky. Stochastic approximation revisited. *Advances in Applied Mathematics*, 7(2):220–227, 1986.
[34] Y.M. Ermol'ev. On the method of generalized stochastic gradients and quasi-Fejér sequences. *Cybernetics*, 5:208–220, 1969.
[35] Y.M. Ermol'ev. On convergence of random quasi-Fejér sequences. *Cybernetics*, 7:655–656, 1971.
[36] Y.M. Ermol'ev and N.Z. Shor. A random search method for two-stage problems of stochastic programming and its generalization. *Kibernetika*, 1:90–92, 1968.
[37] Y.M. Ermol'ev and A.D. Tuniev. Random fejér and quasi-fejér sequences. *Theory of Optimal Solutions – Akademiya Nauk Ukrainskoĭ, SSR Kiev*, 2:76–83, 1968. in Russian; English translation in Amer. Math. Soc. Select. Translat. Math. Statist. Probab., 13 (1973), pp. 143–148.
[38] M.J. Fabian, R. Henrion, A.Y. Kruger, and J.V. Outrata. Error bounds: necessary and sufficient conditions. *Set-Valued and Variational Analysis*, 18(2):121–149, 2010.
[39] L. Fejér. Über die Lage der Nullstellen von Polynomen, die aus Minimumforderungen gewisser Art entspringen. *Mathematische Annalen*, 85:41–48, 1922.
[40] M.C. Ferris. Finite termination of the proximal point algorithm. *Mathematical Programming, Series A*, 50:359–366, 1991.
[41] B. Franci and S. Grammatico. Convergence of sequences: A survey. *Annual Reviews in Control*, 53:161–186, 2022.
[42] B. Grimmer. Convergence rates for deterministic and stochastic subgradient methods without Lipschitz continuity. *SIAM Journal on Optimization*, 29(2):1350–1365, 2019.
[43] M. Gromov. Hyperbolic groups. In S.M. Gersten, editor, *Essays in group theory*, volume 8 of *Mathematical Sciences Research Institute Publications*, pages 75–263. Springer, New York, 1987.
[44] A.N. Iusem, A. Jofré, and P. Thompson. Incremental constraint projection methods for monotone stochastic variational inequalities. *Mathematics of Operations Research*, 44(1):236–263, 2019.
[45] T. Jaakkola, M.I. Jordan, and S.P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6(6):1185–1201, 1994.
[46] H. Jiang and U.V. Shanbhag. On the solution of stochastic optimization and variational problems in imperfect information regimes. *SIAM Journal on Optimization*, 26(4):2394–2429, 2016.
[47] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistic*, 23:462–466, 1952.
[48] A. Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer Cham, 3rd edition, 2020.
[49] U. Kohlenbach. Some logical metatheorems with applications in functional analysis. *Transactions of the American Mathematical Society*, 357(1):89–128, 2005.
[50] U. Kohlenbach. *Applied Proof Theory: Proof Interpretations and their Use in Mathematics*. Springer Monographs in Mathematics. Springer-Verlag Berlin Heidelberg, 2008.
[51] U. Kohlenbach. Proof-theoretic Methods in Nonlinear Analysis. In B. Sirakov, P. Ney de Souza, and M. Viana, editors, *Proceedings ICM 2018*, volume 2, pages 61–82. World Scientific, 2019.
[52] U. Kohlenbach and B. Lambov. Bounds on iterations of asymptotically quasi-nonexpansive mappings. In *Proceedings of the International Conference on Fixed Point Theory and Applications*, pages 143–172. Yokohama Publishers, 2004.
[53] U. Kohlenbach, L. Leuştean, and A. Nicolae. Quantitative results on Fejér monotone sequences. *Communications in Contemporary Mathematics*, 20(2), 2018. 1750015, 42pp.

[54] U. Kohlenbach, G. López-Acedo, and A. Nicolae. Moduli of regularity and rates of convergence for Fejér monotone sequences. *Israel Journal of Mathematics*, 232:261–297, 2019.
[55] U. Kohlenbach and P. Pinto. Fejér monotone sequences revisited. *Journal of Convex Analysis*, 2025. to appear.
[56] M.A. Krasnosel'skii. Two remarks on the method of successive approximations. *Uspekhi Matematicheskikh Nauk*, 10(1(63)):123–127, 1955.
[57] G. Kreisel. On the Interpretation of Non-Finitist Proofs–Part I. *The Journal of Symbolic Logic*, 16(4):241–267, 1951.
[58] G. Kreisel. On the Interpretation of Non-Finitist Proofs–Part II. Interpretation of Number Theory. Applications. *The Journal of Symbolic Logic*, 17(1):43–58, 1952.
[59] H.J. Kushner and G.G. Yin. *Stochastic approximation and recursive algorithms and applications*. Stochastic Modelling and Applied Probability. Springer, 2 edition, 2003.
[60] T.L. Lai. Stochastic approximation: invited paper. *Annals of Statistics*, 31(2):391–406, 2003.
[61] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.
[62] C. Li, B.S. Mordukhovich, J.H. Wang, and J.C. Yao. Weak sharp minima on Riemannian manifolds. *SIAM Journal on Optimization*, 21:1523–1560, 2011.
[63] X. Li and A. Milzarek. A unified convergence theorem for stochastic optimization methods. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 33017–33119. Curran Associates Inc., USA, 2022.
[64] M. Liu, X. Zhang, L. Zhang, R. Jin, and T. Yang. Fast rates of ERM and stochastic approximation: adaptive to error bound conditions. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 4683–4694. Curran Associates Inc., USA, 2018.
[65] W.R. Mann. Mean value methods in iteration. *Proceedings of the American Mathematical Society*, 4:506–510, 1953.
[66] B. Martinet. Régularisation dinéquations variationnelles par approximations successives. *Revue française d'informatique et de recherche opérationnelle*, 4:154–158, 1970.
[67] M. Métivier. *Semimartingales*, volume 2 of *De Gruyter Studies in Mathematics*. De Gruyter, 1982.
[68] T. Motzkin and I. Schoenberg. The Relaxation Method for Linear Inequalities. *Canadian Journal of Mathematics*, 6:393–404, 1954.
[69] I. Necoara and A. Nedić. Minibatch stochastic subgradient-based projection algorithms for feasibility problems with convex inequalities. *Computational Optimization and Applications*, 80(1):121–152, 2021.
[70] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.
[71] M. Neri. A finitary Kronecker's lemma and large deviations in the Strong Law of Large numbers on Banach spaces. *Annals of Pure and Applied Logic*, 176(6), 2025. 103569, 31pp.
[72] M. Neri. Quantitative Strong Laws of Large Numbers. *Electronic Journal of Probability*, 30, 2025. 20, 22pp.
[73] M. Neri and N. Pischke. Proof mining and probability theory. Preprint, available at `https://arxiv.org/abs/2403.00659`, 2024.
[74] M. Neri, N. Pischke, and T. Powell. Generalized learnability of stochastic principles. In *Proceedings of Computability in Europe (CiE 2025)*, Lecture Notes in Computer Science. Springer, 2025. To appear, preprint available at `https://nicholaspischke.github.io/`.
[75] M. Neri and T. Powell. A quantitative Robbins-Siegmund theorem. Preprint, available at `https://arxiv.org/abs/2410.15986`, 2024.
[76] M. Neri and T. Powell. On quantitative convergence for stochastic processes: Crossings, fluctuations and martingales. *Transactions of the American Mathematical Society*, 2025. To appear, preprint available at `https://arxiv.org/abs/2406.19979`.
[77] V.D. Nguyen, B.C. Vũ, and D. Papadimitriou. A stochastic primal-dual splitting algorithm with variance reduction for composite optimization problems. *Applicable Analysis*, 2024. to appear, 25pp.
[78] N. Pischke. Quantitative Results on Algorithms for Zeros of Differences of Monotone Operators in Hilbert Space. *Journal of Convex Analysis*, 30(1):295–315, 2023.
[79] N. Pischke. Duality, Fréchet differentiability and Bregman distances in hyperbolic spaces. *Israel Journal of Mathematics*, 2025. To appear, 40pp.
[80] N. Pischke. Generalized Fejér monotone sequences and their finitary content. *Optimization*, 2025. To appear, 68pp., doi:10.1080/02331934.2024.2390114.

[81] N. Pischke and U. Kohlenbach. Effective rates for iterations involving Bregman strongly nonexpansive operators. *Set-Valued and Variational Analysis*, 32(4), 2024. 33, 58pp.
[82] N. Pischke and T. Powell. Asymptotic regularity of a generalised stochastic Halpern scheme with applications. Preprint, available at `https://arxiv.org/abs/2411.04845`, 2024.
[83] L. Qihou. Iteration sequences for asymptotically quasi-nonexpansive mappings with error member. *Journal of Mathematical Analysis and Applications*, 259:18–24, 2001.
[84] P. Révész. Robbins-Monro procedure in a Hilbert space and its application in the theory of learning processes. I. *Studia Scientiarum Mathematicarum Hungarica*, 8:391–398, 1973.
[85] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.
[86] H. Robbins and D. Siegmund. A convergence theorem for non-negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.
[87] R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal of Control and Optimization*, 14:877–898, 1976.
[88] L. Rosasco, S. Villa, and B.C. Vũ. Stochastic forward–backward splitting for monotone inclusions. *Journal of Optimization Theory and Applications*, 169:388–406, 2016.
[89] G.I. Salov. On a stochastic approximation theorem in a Hilbert space and its applications. *Theory of Probability and its Applications*, 24(2):413–419, 1980.
[90] A. Shapiro. Quantitative stability in stochastic programming. *Mathematical Programming*, 67:99–108, 1994.
[91] A. Shapiro and T. Homem de Mello. On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs. *SIAM Journal on Optimization*, 11(1):70–86, 2000.
[92] N.Z. Shor. *Minimization Methods for Non-Differentiable Functions*. Springer Series in Computational Mathematics. Springer, 1985.
[93] E. Specker. Nicht konstruktiv beweisbare Sätze der Analysis. *Journal of Symbolic Logic*, 14:145–158, 1949.
[94] R.S. Sutton and A.G. Barto. *Reinforcement Leaning: An Introduction*. MIT Press, 2nd edition, 2018.
[95] J. H. Venter. On Dvoretzky Stochastic Approximation Theorems. *The Annals of Mathematical Statistics*, 37(6):1534–1544, 1966.
[96] J. Ville. *Étude Critique de la Notion de Collectif*. PhD thesis, École Polytechnique, 1939.
[97] B.C. Vũ. Almost sure convergence of the forward–backward–forward splitting algorithm. *Optimization Letters*, 10:781–803, 2016.
[98] C. Watkins. *Learning from delayed rewards*. PhD thesis, University of Cambridge, 1989.
[99] Y. Xu, Q. Lin, and T. Yang. Accelerate stochastic subgradient method by leveraging local growth condition. *Analysis and Applications*, 17(5):773–818, 2019.
[100] G. Yin and Y.M. Zhu. On H-valued Robbins-Monro processes. *Journal of Multivariate Analysis*, 34(1):116–140, 1990.
[101] H. Zhang, Z. Zheng, and J. Lavaei. Stochastic $L^\natural$-convex function minimization. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pages 13004–13018. Curran Associates Inc., USA, 2021.