CONVERGENCE GUARANTEES FOR STOCHASTIC ALGORITHMS SOLVING NON-UNIQUE PROBLEMS IN METRIC SPACES

NICHOLAS PISCHKE AND THOMAS POWELL

Department of Computer Science, University of Bath,
Claverton Down, Bath, BA2 7AY, United Kingdom,
E-mails: {nnp39,trjp20}@bath.ac.uk

Abstract. We prove a general quantitative theorem on the asymptotic behavior of stochastic quasi-Fejér monotone sequences in a broad metric context. Concretely, our result explicitly constructs a rate of convergence for such process, both in mean and almost surely, under an abstract stochastic regularity assumption, derived from previous work of Kohlenbach, López-Acedo and Nicolae [Isr. J. Math. 232(1), pp. 261-297, 2019] on such notions in a deterministic context. Our notion of regularity extends and unifies many common conditions from the literature, such as generalized contractivity for self maps, weak sharp minima and error bounds for real-valued functions, uniform monotonicity and global metric subregularity for set-valued operators, related Polyak-Łojasiewicz or Kurdyka-Łojasiewicz conditions, as well as expected sharp growth as e.g. studied by Asi and Duchi [SIAM J. Optim. 29(3), pp. 2257-2290, 2019]. The rate is moreover highly uniform, depending only on very few data of the surrounding objects. We also discuss special cases which allow for the construction of fast rates in the form of linear non-asymptotic guarantees. We conclude by presenting three concrete methods from stochastic approximation where our results yield new rates of convergence, including the classical example of the stochastic proximal point method, a randomized variant of the Krasnoselskii-Mann scheme for solving stochastic fixed-point equations, and a Busemann subgradient method recently introduced by Goodwin, Lewis, López-Acedo and Nicolae [Math. Program., to appear], all of which make use of our metric generality by being formulated over complete geodesic metric spaces of nonpositive curvature.

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

1. Introduction

1.1. **The setup.** We are concerned with the study of stochastic algorithms (x_n) that approximate almost-sure solutions to the abstract problem

$$\text{find a zero } z \in \operatorname{zer} F := \{z \in X \mid F(z) = 0\} \text{ for a function } F : X \to [0, \infty],$$

over arbitrary separable and complete metric spaces (X, d). A wide range of deterministic and stochastic problems can be naturally brought into this form, including almost-sure fixed point problems or mean minimization problems (see Section 3.3 later on).

As with the problem formulation, we consider a broad class of algorithms: Instead of confining ourselves to a specific iteration schema, we consider arbitrary stochastic processes (x_n), adapted to a filtration (F_n) over a probability space $(\Omega, \mathsf{F}, \mathbb{P})$, that are stochastically quasi-Fejér monotone, i.e.

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

for all $n \in \mathbb{N}$ and all $z \in \operatorname{zer} F$ and where $(\zeta_n), (\xi_n)$ are summable nonnegative random variables adapted to (F_n). This represents a relaxed supermartingale condition in line with similar abstract approaches in the literature.

Date: May 6, 2026.

The driving question of our paper is concerned with quantitative aspects of convergence:

Fundamental question: When can effective rates towards solutions, in mean and almost surely, be explicitly described, for general classes of problems F and algorithms (x_n)?

As is well-known, this problem is in particular nontrivial since *effective* rates cannot always be found, a phenomenon that in optimization theory is sometimes referred to as *arbitrary slow convergence*. This can be formally explained using tools from mathematical logic and computability theory, where so-called Specker sequences [88] can be used to show that already in very simple deterministic cases, computable rates of convergence are generally unachievable for general quasi-Fejér monotone methods (x_n) and problems F (see in particular [72]).

1.2. **The theoretical contributions of the paper.** Inspired by the work of Kohlenbach, López-Acedo and Nicolae [53],[1] we consider a broad class of problems that can be characterised, in an abstract way, by considering functions $F : X \to [0, \infty]$ that satisfy a generalized regularity assumption. Indeed, we extend the generalized notion of regularity introduced in [53] here to a stochastic context, taking the form of a modulus $\tau : (0, \infty) \to (0, \infty)$ satisfying

$$\forall \varepsilon > 0 \, \forall x \in D \, (\mathbb{E}[F(x)] < \tau(\varepsilon) \to \mathbb{E}[\mathrm{dist}_{\mathrm{zer}F}(x)] < \varepsilon)$$

for a suitable collection of X-valued random variables D. As we will show, such regularity conditions are intimately related to associated growth conditions in mean

$$\forall x \in D \, (\mathbb{E}[F(x)] \geq \tau(\mathbb{E}[\mathrm{dist}_{\mathrm{zer}F}(x)]))$$

for the mapping F.

As we will discuss in detail (see Section 3.3), these conditions uniformly cover various notions such as: generalized contractions and retractions; generalized weak sharp minima in the sense of [53] (extending [28], see also [27, 43, 58]) and related to that corresponding notions of error bounds (see e.g. [22, 39]) as well as polynomial growth conditions (see e.g. [87]) and expected sharp growth [5]; generalized metric subregularity (see e.g. [49, 59], extending [56, 38]) and related Polyak-Łojasiewicz, or more generally Kurdyka-Łojasiewicz conditions (see e.g. [21, 22, 89]); uniform and strong accretivity and monotonicity for operators and vector fields; uniform and strong convexity for functions. We note that many of these conditions, along with our abstract regularity notion in general, do not necessarily entail that zer$F$ has a unique solution.

It is known that already for the deterministic Picard iteration, such a (deterministic) modulus of regularity can be explicitly constructed from an associated (uniform) rate of convergence for the process towards a solution (see Proposition 4.4 in [53]), illustrating that the presence of this generalized regularity condition is fundamentally connected to the existence of explicit (and suitably uniform) rates of convergence. This in particular justifies the wide generality of the approach not only in the deterministic but also in the stochastic setting.

Our main results are general quantitative convergence theorems for stochastically quasi-Fejér monotone iterations relative to solution sets of problems that satisfy such a stochastic notion of regularity, requiring in addition only a mild approximation property (see Theorems 4.8 and

[1]The present paper is part of an ongoing effort, initiated in the last two years, to bring methods from proof mining [51, 52] systematically to bear on probability theory and stochastic optimization. The works that this paper builds on, notably [53] and [69] together with [75] (as well as the related [44, 70]) are similarly part of this proof-theoretic approach to computational results in mathematics. Here, for example, this logical approach in particular influenced the way in which we formulated the various (stochastic) moduli. The rest of the paper avoids any technical reference to mathematical logic.

4.11). These results provide explicit constructions for rates of convergence, in mean and almost surely, which are moreover highly uniform, depending only on very few data of the surrounding objects, and in particular being independent of the distribution of the space. Furthermore, we provide a similarly general result on fast (that is in this case linear) nonasymptotic guarantees in the context of linear regularity and suitable assumptions on the surrounding parameters (see Theorem 4.13). Finally, all of these results are formulated in the even broader context of general distance functions $\phi : X \times X \to [0, \infty)$, following previous work of the first author in a deterministic context [75] (and of the authors and Neri [69] in certain stochastic settings, as discussed below), which allow us to uniformly cover perturbations of the metric and distance functions beyond that, such as Bregman distances.

Some initial results on effective convergence for stochastic quasi-Fejér monotone sequences are contained in the authors' earlier work [69], confined to the more restrictive class of problems with unique solutions. The present paper extends those results substantially: Along with a much more general and extensive treatment of stochastic regularity, we are confronted with multiple demanding technical difficulties that are entirely absent in the case of unique solutions. For example, to deal with non-unique problems, our main results combine a quantitative martingale argument motivated by [69] with a new measurable selection argument along the filtration (F_n) of the process (x_n). By nature of the filtration, these results from measurable selection theory cannot assume the completeness of the measure space and hence require additional care beyond that of canonical results. In that context, we are forced to consider a strengthened variant of stochastic quasi-Fejér monotonicity, introduced abstractly for the first time here, which allows one to maintain the supermartingale condition

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

even for F_n-measurable and suitably integrable zer$F$-valued random variables z. This notion is often satisfied outright for many methods (in particular for the applications discussed in this paper), but we additionally examine this stronger property in relation to the usual formulation, providing a general lifting result for a class of distances including metric powers (abstracting recent work of Combettes and Madariaga [35]).

1.3. **The applied contributions of the paper.** We provide three example applications of our theoretical work, yielding new results for both classical and recently introduced methods. To illustrate the metric generality of our framework, all applications are set in the context of Hadamard spaces, that is complete geodesic metric spaces with nonpositive curvature in the sense of Alexandrov (see e.g. [25]), covering in particular Hilbert spaces, $\mathbb{R}$-trees, as well as Hadamard manifolds (i.e. complete simply connected Riemannian manifolds of nonpositive sectional curvature). This generality is particularly beneficial since stochastic optimization over nonlinear spaces such as manifolds plays a key role in modern machine learning (see e.g. [90]). However, Hadamard spaces are not limited to such settings and beyond that also cover practically relevant spaces without immediate differentiable structure, such as the Billera-Holmes-Vogtmann tree space [18] prominently used in phylogenetics, and its recent variation for networks [67]. In all cases, our results not only provide rates at a level of generality not considered previously, but also yield the strong convergence of these methods without local compactness assumptions, which for our abstract notion of regularity is qualitatively novel, even disregarding the quantitative aspects.

The first method we discuss is a stochastic variant of the proximal point method for minimizing the mean of a randomized convex function satisfying a relaxed Lipschitz condition as studied e.g. in [13, 73] (see also [11]). Here we extend the strong convergence results provided in [73] to more general regularity assumptions on the function, and quantitatively outfit them with

explicit rates. The second is a randomized variant of the central Krasnoselskii-Mann iteration for solving stochastic common fixed point problems, similar in nature to recent work by Combettes and Madariaga [35] and recently studied over proper Hadamard spaces in [70]. Finally, we consider the recently introduced projected subgradient method of Goodwin, Lewis, López-Acedo and Nicolae [45] and its extension studied in [77] for stochastic minimization, which utilizes the so-called Busemann subgradients introduced in [45], a new type of subgradient that is particularly suitable for nonlinear geometric contexts.

1.4. **Future work.** We envisage this paper as the basis for future work, not only as a tool for providing effective convergence guarantees for various other methods in stochastic optimization and approximation, but also for instigating further theoretical developments. On the applied side, future work should in particular be concerned with stochastic iterations that make use of our generalized distances. For example, a unified approach to stochastic optimization via Bregman distances, extending stochastic quasi-Fejér monotonicity in the particular form treated in [35], has been recently considered in [91], and we anticipate that several of the examples given there can be suitably adapted to our framework. Also, the present work could provide guiding principles for developing effective convergence guarantees for stochastic algorithms with super relaxations in the style of [35] also over metric contexts such as Hadamard spaces, making use of geodesic rays. On the theoretical as well as applied side, future work could include extensions of the present results to continuous-time processes, combining the approach of this paper (as well as [69, 70]) with the recent work [44] of the first author and Freund on similarly broad quantitative considerations of continuous-time dynamical systems in a deterministic setting. These results would presumably rely on a strategy that, in the style of the present paper, combines measurable selection theory with continuous-time martingale theory, a general quantitative approach to the latter being interesting in its own right. Potential applications of these combined results could then in particular include works such as [19, 20, 60, 61, 64, 65] on stochastic differential equations and inclusions and related dynamical systems.

Preliminaries and notation. Throughout, we fix a probability space $(\Omega, \mathsf{F}, \mathbb{P})$ together with a filtration (F_n). We denote the (conditional) expectation over that space by $\mathbb{E}[\cdot]$ and we denote characteristic functions of measurable sets $A \in \mathsf{F}$ by $\mathbf{1}_A$. Further, if not stated otherwise, (X, d) will denote a separable and complete metric space. If seen as a measure space, we assume that X is endowed with its Borel σ-algebra $\mathsf{B}(X)$. We denote closed balls relative to the metric by

$$\overline{B}_r(a) := \{x \in X \mid d(x, a) \leq r\},$$

given $r > 0$ and $a \in X$. We often call a sequence of X-valued random variables an X-valued stochastic process. If a stochastic process (x_n) is such that x_n is F_n-measurable for any $n \in \mathbb{N}$, we call it adapted to (F_n). We write $\ell_+(\mathsf{F}_n)$ for the set of all sequences (ξ_n) of nonnegative random variables adapted to (F_n) and $\ell^1_+(\mathsf{F}_n)$ for all such sequences that further satisfy $\sum_{n=0}^{\infty} \xi_n < \infty$ a.s. Above, and throughout the paper, we generally write ∞ for $+\infty$, unless we are explicitly differentiating $+\infty$ from $-\infty$.

2. Carathéodory distance functions and stochastic approximation properties

At our most abstract, we will be concerned with general distance functions $\phi : X \times X \to [0, \infty)$, potentially distinct from the metric. For such a distance and a non-empty set $S \subseteq X$, we write

$$\operatorname{dist}^{\phi}_S(x) := \inf_{s \in S} \phi(s, x)$$

and we omit the superscript ϕ only if $\phi = d$. Naturally, not all such distance functions will be permissible in stochastic contexts, and we have to place some assumptions on the measurability of ϕ. The central such assumption will be that ϕ is a Carathéodory function:

Assumption 2.1 (Carathéodory distance)**.** Assume ϕ is a Carathéodory distance, in the sense that ϕ is continuous in its left argument and measurable in its right argument.

Throughout, whenever we use ϕ to refer to a general distance function, we always implicitly assume the above even if not stated explicitly. As we ultimately care for metric convergence, we will be concerned with converting from a generic distance ϕ back to the underlying metric d. If points equal in the sense of the distance ϕ are also equal in the sense of the metric, we call such a distance consistent. The following uniform quantitative formulation of consistency defines our main assumption in that direction:

Definition 2.2 (Uniformly consistent distance)**.** A distance ϕ is called uniformly consistent with modulus $\theta : (0, \infty) \to (0, \infty)$ if

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

Such a modulus essentially induces a growth condition on the distance ϕ in terms of the metric, as we highlight in the following remark:

Remark 2.3. Let $\theta : [0, \infty) \to [0, \infty)$ be such that $\theta(0) = 0$ and $\theta(\varepsilon) > 0$ for $\varepsilon > 0$. If θ is a modulus of uniform consistency for ϕ, then $\phi(x, y) \geq \theta(d(x, y))$ for all $x, y \in X$.

Some key examples of uniformly consistent distances are collected in the following example:

Example 2.4. The following distance functions are Carathéodory distances:

(1) Perturbed metric distances, that is $\phi = G \circ d$ where d is the metric of the space and $G : [0, \infty) \to [0, \infty)$ is continuous. In particular, if G is inverse continuous at 0 with a modulus $g : (0, \infty) \to (0, \infty)$, that is

$$\forall a \geq 0 \; \forall \varepsilon > 0 \, (G(a) < g(\varepsilon) \to a < \varepsilon) \, ,$$

then ϕ is uniformly consistent with modulus $\theta(\varepsilon) := g(\varepsilon)$. A common example of a perturbation function is $G = (\cdot)^2$, which is immediately inverse continuous at 0 with modulus $g(\varepsilon) := \varepsilon^2$.

(2) Perturbed Bregman distances, that is $\phi = G \circ D_f$ for G continuous as above and

$$D_f(x, y) := f(x) - f(y) - \langle x - y, \nabla f(y) \rangle$$

over a reflexive Banach space $(X, \|\cdot\|)$, where $f : X \to \mathbb{R}$ is lsc, convex and Fréchet differentiable on X. Suppose further that G is inverse continuous at 0 with modulus g as above, and that f is sequentially consistent[2] with a modulus of sequential consistency $\rho : (0, \infty)^2 \to (0, \infty)$ in the sense of [78], i.e.

$$\forall \varepsilon, b > 0 \; \forall x, y \in \overline{B}_b(0) \, (D_f(x, y) < \rho(\varepsilon, b) \to \|x - y\| < \varepsilon) \, .$$

Then ϕ is uniformly consistent on every ball $\overline{B}_b(0) \subseteq X$ with modulus $\theta(\varepsilon) := g(\rho(\varepsilon, b))$. Indeed, if $x, y \in \overline{B}_b(0)$ are such that $G(D_f(x, y)) < g(\rho(\varepsilon, b))$, then $D_f(x, y) < \rho(\varepsilon, b)$ as before and so $\|x - y\| < \varepsilon$. We refer to [78] for related discussions.

If a distance is consistent, then we can in particular also convert the associated set-distance functions:

[2]Crucially, if X contains at least two points, a function f as above is sequentially consistent iff it is totally convex on bounded sets iff it is uniformly convex on bounded sets (see Theorem 2.10 in [30] and see also [29]).

Lemma 2.5. *Let ϕ be uniformly consistent with modulus θ. Further, let $S \subseteq X$ be non-empty. Then for any $\varepsilon > 0$ and any $x \in X$:*

$$\mathrm{dist}_S^\phi(x) < \theta(\varepsilon) \to \mathrm{dist}_S(x) < \varepsilon.$$

Proof. Suppose $\mathrm{dist}_S^\phi(x) < \theta(\varepsilon)$. Thus, there exists an $s \in S$ with $\phi(s,x) < \theta(\varepsilon)$. By the assumption on θ, we get $d(s,x) < \varepsilon$ and so in particular $\mathrm{dist}_S(x) < \varepsilon$. □

We now turn to measurability properties of Carathéodory distances. Firstly, this assumption entails the following crucial measurability properties on ϕ and dist^ϕ:

Lemma 2.6. *If ϕ is a Carathéodory distance, then*

(1) ϕ is measurable w.r.t. $\mathsf{B}(X) \otimes \mathsf{B}(X)$,
(2) $\mathrm{dist}_S^\phi(x)$ is measurable for any non-empty $S \in \mathsf{B}(X)$.

Proof. Joint measurability of ϕ, that is item (1), follows from e.g. Lemma 8.2.6 in [6]. The measurability of $\mathrm{dist}_S^\phi(x)$ follows from e.g. the first part of Lemma 8.2.11 in [6], noting that for that first part it suffices to just work over a measurable space. □

As a key step in our convergence proofs and construction of associated rates, we will later crucially rely on the property that given a set $S \subseteq X$, we can measurably select points $s \in S$ so that $\phi(s,x)$ approximates $\mathrm{dist}_S^\phi(x)$, and that with arbitrary degree of precision. A result in that vein appears, for $\phi = d$, in the work of Römisch [84] and our results effectively extend his. However, [84] operates under the assumption that the underlying σ-algebra is complete. Indeed, it is exactly this completeness of the underlying probability space, which occurs as a common assumption whenever results from measurable selection theory are used, that is problematic in our context as we will later require this property for all elements F_n of an associated filtration, which will generally not be complete. Nevertheless, the above property can be guaranteed without any completeness assumptions of the underlying probability space for the present Carathéodory distances ϕ, as we will now show.

For this, we rely on a few results from measurable selection theory which we now collect. Given a measurable space (T, T) and a complete separable metric space X, a set-valued map $\varphi : T \to 2^X$ is called graph measurable if

$$\mathrm{gra}(\varphi) := \{(t,x) \in T \times X \mid x \in \varphi(t)\} \in \mathsf{T} \otimes \mathsf{B}(X).$$

Over complete σ-finite measure spaces, and if φ has non-empty closed images, this is equivalent to the (weak) measurability of φ as commonly considered in measurable selection theory but generally, the above presents a weaker notion. We refer to [3, 6, 31] for further background on that area.

We begin with the well-known measurable selection theorem of Aumann [7], which forms a crucial ingredient for dispensing of completeness.

Lemma 2.7 (Aumann [7], see also Corollary 18.27 in [3])**.** *Let (T, T, μ) be a finite measure space and let X be a complete separable metric space. Let $\varphi : T \to 2^X$ be a graph measurable with non-empty values. Then there is a measurable function $x : T \to X$ such that $x(t) \in \varphi(t)$ almost everywhere.*

Beyond this selection theorem, the other main result we need is a variant of the inverse image theorem in measurable selection theory (see e.g. Theorem 8.2.9 in [6]). This result is commonly stated under the assumption that the measure space is complete, an assumption which we, as discussed before, crucially want to avoid. We in the following give a variant which dispenses of that assumption while weakening the conclusion to the graph measurability of the

respective function, which will however suffice in the context of Aumann's selection theorem. That argument essentially follows the usual proof of the inverse image theorem, and so harbors no surprises. Nevertheless, we rederive it here for the convenience of the reader.

Lemma 2.8. *Let (T, T) be a measurable space and let X, Y be complete separable metric spaces. Let $\varphi : T \to 2^X$ and $\psi : T \to 2^Y$ be graph measurable and let $c : T \times X \to Y$ be a Carathéodory function. Then χ defined by*

$$\chi(t) := \{x \in \varphi(t) \mid c(t, x) \in \psi(t)\}$$

is graph measurable.

Proof. Define $d(t, x) := (t, c(t, x))$ and note that

$$\mathrm{gra}(\chi) = \mathrm{gra}(\varphi) \cap d^{-1}(\mathrm{gra}(\psi)).$$

As c is a Carathéodory function, it is $\mathsf{T} \otimes \mathsf{B}(X)/\mathsf{B}(Y)$-measurable. In particular, we thus have

$$d^{-1}(A \times B) = c^{-1}(B) \cap (A \times X) \in \mathsf{T} \otimes \mathsf{B}(X)$$

for any $A \in \mathsf{T}$ and $B \in \mathsf{B}(Y)$. In particular, we thus have $d^{-1}(C) \in \mathsf{T} \otimes \mathsf{B}(X)$ for any $C \in \mathsf{T} \otimes \mathsf{B}(Y)$ and so d is measurable. In particular, as $\mathrm{gra}(\varphi)$ and $\mathrm{gra}(\psi)$ are measurable, we thus have that $\mathrm{gra}(\chi)$ is measurable as well. □

The following result now is an extension of the canonical result that closed balls in complete separable metric spaces are measurable if their origins and radii are (see e.g. Corollary 8.2.13 in [6]). For that, given $r > 0$ and $x \in X$, we define

$$\overline{B}^{\phi}_r(x) := \{y \in X \mid \phi(y, x) \leq r\}$$

as the closed ball around x with radius r, relative to ϕ.

Lemma 2.9. *Let (T, T) be a measurable space and let $f : T \to X$, $\rho : T \to [0, \infty)$ be measurable and let ϕ be a Carathéodory distance. Then $t \mapsto \overline{B}^{\phi}_{\rho(t)}(f(t))$ is graph measurable.*

Proof. Define $c(t, x) := \phi(x, f(t))$. As ϕ is a Carathéodory distance as described above, we get that also c is a Carathéodory function. The result now follows from the previous Lemma 2.8 by setting $\varphi \equiv X$ and $\psi(t) := [0, \rho(t)]$. □

Lemma 2.10. *Let $\mathcal{F}$ be a sub-σ-algebra of F. Further, let ϕ be a Carathéodory distance and let $S \in \mathsf{B}(X)$ be non-empty. Then dist^{ϕ}_S has $\mathcal{F}$-measurable approximations in the sense that for all $\mathcal{F}$-measurable X-valued random variables x and any $\varepsilon > 0$, there exists some X-valued $\mathcal{F}$-measurable random variable s such that*

$$s \in S \text{ and } \phi(s, x) \leq \mathrm{dist}^{\phi}_S(x) + \varepsilon \text{ a.s.}$$

Proof. Let x be an X-valued $\mathcal{F}$-measurable random variable and let $\varepsilon > 0$ be given. Define

$$A(\omega) := \{s \in S \mid \phi(s, x(\omega)) \leq \mathrm{dist}^{\phi}_S(x(\omega)) + \varepsilon\} = S \cap \overline{B}^{\phi}_{r(\omega)}(x(\omega))$$

where $r(\omega) := \mathrm{dist}^{\phi}_S(x(\omega)) + \varepsilon$. As dist_{ϕ} is measurable, $r(\omega)$ is an $\mathcal{F}$-measurable random variable. By Lemma 2.9, we get that $\overline{B}^{\phi}_{r(\omega)}(x(\omega))$ is graph measurable w.r.t. $\mathcal{F}$. As we have

$$\mathrm{gra}A = \mathrm{gra}\left(\overline{B}^{\phi}_r(x)\right) \cap (\Omega \times S)$$

and since S is Borel, we also get that A is graph measurable w.r.t. $\mathcal{F}$. Aumann's measurable selection theorem, that is Lemma 2.7, now yields the existence of an X-valued $\mathcal{F}$-measurable random variable s such that $s \in S$ and $\phi(s, x) \leq \mathrm{dist}^{\phi}_S(x) + \varepsilon$ a.s. □

We will actually only rely on the above property being true in mean:

Corollary 2.11. Let $\mathcal{F}$ be a sub-σ-algebra of F. Further, let ϕ be a Carathéodory distance and let $S \in \mathsf{B}(X)$ be non-empty. Then dist_S^ϕ has $\mathcal{F}$-measurable approximations in mean, in the sense that for all $\mathcal{F}$-measurable X-valued random variables x and any $\varepsilon > 0$, there exists some X-valued $\mathcal{F}$-measurable random variable s such that

$$s \in S \text{ a.s. and } \mathbb{E}[\phi(s,x)] \leq \mathbb{E}[\mathrm{dist}_S^\phi(x)] + \varepsilon.$$

3. Stochastic regularity for abstract problems

3.1. **Regularity in mean.** As motivated in the introduction already, we consider an arbitrary function $F : X \to [0,\infty]$ and the associated problem of finding an element of

$$\mathrm{zer}F := \{z \in X \mid F(z) = 0\}$$

for a general and abstract problem formulation. Naturally, also not all such problem formulations will be permissible in stochastic contexts. We make the following assumption:

Assumption 3.1 (Stochastic problem)**.** Assume F is measurable, and that $\mathrm{zer}F$ is a closed non-empty set.

As we will see, a range of stochastic problems can be formulated in this simple manner, satisfying the above assumption (see in particular Section 3.3 later). In particular, while the function F and its associated zero problem are themselves deterministic, they also naturally cover stochastic zero problems as illustrated abstractly in the following example:

Example 3.2. Let (T, T, μ) be a σ-finite measure space and $h : T \times X \to [0,\infty]$ be a Carathéodory function. The associated stochastic problem

$$\mathrm{zer}h := \{z \in X \mid h(t,z) = 0 \text{ almost everywhere}\}$$

can be recognized as an instance of a problem $\mathrm{zer}F$ as above by setting $F(z) := \int h(t,z)\, d\mu(t)$. In particular, F is measurable by the Fubini-Tonelli theorem as h is $\mathsf{T} \otimes \mathsf{B}(X)$-measurable. Further, note that $\mathrm{zer}F$ is closed: If $(z_n) \subseteq \mathrm{zer}F$ with $\lim_{n\to\infty} z_n = z$, then $h(t,z_n) = 0$ almost everywhere for all $n \in \mathbb{N}$, say on T_n^c with T_n of measure 0. Define $T' := \bigcup_{n\in\mathbb{N}} T_n$. Then T' still has measure 0 and for $t \in (T')^c$, we have $h(t,z_n) = 0$ for all $n \in \mathbb{N}$. As h is continuous in its right argument, we have $h(t,z) = 0$. Therefore $h(t,z) = 0$ almost everywhere, so that $z \in \mathrm{zer}F$.

We are primarily interested in quantitative stochastic regularity conditions on such problems which allow for the construction of explicit rates of convergence of stochastic processes that satisfy a standard almost-supermartingale condition. An important "regularity" assumption that is often imposed in this regard, even though often left implicit, is that the respective problem has a unique solution $\mathrm{zer}F = \{z\}$ quantified by an explicit *modulus of uniqueness* in the following stochastic sense:

Definition 3.3 (Stochastic uniqueness in mean)**.** Let ϕ be a Carathéodory distance and let D be a collection of X-valued random variables. Assume $\mathrm{zer}F = \{z\}$. A modulus of ϕ-uniqueness for F in mean w.r.t. D is a function $\tau : (0,\infty) \to (0,\infty)$ with

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

The more general regularity notion we introduce below arises as a natural extension of this quantitative notion of uniqueness in mean to *non-unique* problems, obtained by replacing the distance $\phi(z,x)$ to the (unique) solution z with the distance $\mathrm{dist}_{\mathrm{zer}F}^\phi(x)$ to the solution set $\mathrm{zer}F$.

Definition 3.4 (Stochastic regularity in mean)**.** Let ϕ be a Carathéodory distance and let D be a collection of X-valued random variables. A modulus of ϕ-regularity for F in mean w.r.t. D is a function $\tau : (0,\infty) \to (0,\infty)$ with

$$\forall \varepsilon > 0 \; \forall x \in D \left(\mathbb{E}[F(x)] < \tau(\varepsilon) \to \mathbb{E}[\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x)] < \varepsilon \right).$$

If D is the class of all X-valued random variables, we simply call such a function τ a modulus of ϕ-regularity for F in mean. We note that this notion coincides with the previous modulus of uniqueness if $\mathrm{zer}F = \{z\}$. Crucially, such a function induces a growth condition on F in mean, as we collect in the following remark.

Remark 3.5. Let $\tau : [0,\infty) \to [0,\infty)$ be such that $\tau(0) = 0$ and $\tau(\varepsilon) > 0$ for $\varepsilon > 0$. If τ is a modulus of ϕ-regularity for F in mean w.r.t. D, then

$$\mathbb{E}[F(x)] \geq \tau(\mathbb{E}[\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x)])$$

for all $x \in D$. Further, if τ is additionally nondecreasing, these two properties are equivalent. To see this equivalence, simply note that for any $\varepsilon > 0$, if $\mathbb{E}[F(x)] < \tau(\varepsilon)$, then it holds that $\tau(\mathbb{E}[\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x)]) < \tau(\varepsilon)$ which yields that $\mathbb{E}[\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x)] < \varepsilon$, as τ is nondecreasing.

Instantiations of such general stochastic moduli of regularity appear in various situations already throughout the literature, as we will survey later in this section. It is the goal of this paper to develop a uniform quantitative theory of these regularity assumptions and in particular to illustrate how they can be used to systematically derive rates of convergence for stochastic algorithms. As discussed in the introduction, related (but slightly different) results for the notion of stochastic uniqueness in mean were obtained by the present authors and Neri in [69] (with moduli of uniqueness in mean called *moduli of strong uniqueness in expectation* therein), derived from more abstract quantitative results for certain stochastic processes, and these also motivate part of our approach here. However, as also highlighted before, the uniqueness assumption heavily simplifies the problem, so that the present paper in particular relies on substantiative additional theory that complements the work [69].

3.2. **Variants of stochastic regularity.** We now discuss abstract ways in which regularity in mean can arise, and how corresponding moduli can be derived and defined in these cases. These abstract results are then used later on to capture well-known regularity notions from the literature as instances of our abstract notion. Our first result in this vein shows how stochastic regularity in mean can be obtained from a pointwise property, for suitable τ.

Definition 3.6 (Pointwise regularity)**.** Let ϕ be a Carathéodory distance. A modulus of ϕ-regularity for F is a function $\tau : (0,\infty) \to (0,\infty)$ with

$$\forall \varepsilon > 0 \; \forall x \in X \left(F(x) < \tau(\varepsilon) \to \mathrm{dist}^{\phi}_{\mathrm{zer}F}(x) < \varepsilon \right).$$

For $\phi = d$, this property coincides with (a special case[3] of) the deterministic notion of a modulus of regularity as defined in [53]. The case of more general distance functions first appeared in [75]. Similarly to before, we can recognize such a regularity property as a different formulation of a growth condition for F. This was already highlighted in [53] (see Remark 3.2 therein), after which the following remark (and in fact already Remark 3.5) are modelled.

[3]In [53], the authors further consider a "local" variant of this notion, with the property required only on a ball around a fixed solution $z \in \mathrm{zer}F$ (see Definition 3.1 therein). In this paper, we omit this additional locality condition, as it has implications on the boundedness of random variables that seem to limit the stochastic theory. In any way, most regularity notions commonly encountered in the literature are even of this "global" form, as we will later survey in Section 3.3, so that this creates no severe limitations.

Remark 3.7. Let $\tau : [0,\infty) \to [0,\infty)$ be such that $\tau(0) = 0$ and $\tau(\varepsilon) > 0$ for $\varepsilon > 0$. If τ is a modulus of ϕ-regularity for F, then

$$F(x) \geq \tau(\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x))$$

for all $x \in X$. Again, if τ is additionally nondecreasing, these two properties are equivalent, which can be shown as in Remark 3.5.

While in [53] (and [75]) this modulus is studied only in the context of nonstochastic problems, we can now show that in the case that τ is nondecreasing and convex and $\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x)$ integrable for all $x \in D$, any such deterministic modulus is also a modulus of ϕ-regularity for F in mean w.r.t. D. In particular, it thus follows that essentially all deterministic regularity notions as studied in [53], which are rather numerous as we will later discuss, immediately entail a stochastic regularity notion with the same modulus under mild conditions that are in most cases satisfied.

Lemma 3.8. *Let $\tau : [0,\infty) \to [0,\infty)$ be convex and nondecreasing with $\tau(0) = 0$ and $\tau(\varepsilon) > 0$ for all $\varepsilon > 0$, and let D be a collection of random variables such that $\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x)$ is integrable for all $x \in D$. Then, if τ is a modulus of ϕ-regularity for F, it also a modulus of ϕ-regularity for F in mean w.r.t. D.*

Proof. Let τ be a modulus of ϕ-regularity for F. Using Remark 3.7, we in particular have $\tau(\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x)) \leq F(x)$ pointwise everywhere for all $x \in D$. Using that τ is convex, Jensen's inequality yields

$$\tau(\mathbb{E}[\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x)]) \leq \mathbb{E}[\tau(\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x))] \leq \mathbb{E}[F(x)]$$

for all $x \in D$. As in Remark 3.5, since τ is nondecreasing, this yields that τ is a modulus of ϕ-regularity for F in mean w.r.t. D. □

As this relation between pointwise and stochastic regularity has a crucial impact on the range of the stochastic theory laid out in this paper, a natural question is when such *convex* moduli of regularity can be obtained. Interestingly, if τ satisfies a certain mild growth condition, then we can always guarantee this:

Remark 3.9. Suppose that $\tau : [0,\infty) \to [0,\infty)$ is strictly increasing with $\tau(0) = 0$. If τ satisfies $\liminf_{x\to\infty} \tau(x)/x > 0$, then its convex envelope

$$\check{\tau}(x) := \sup\{f(x) \mid f \leq \tau \text{ is convex}\}$$

is also strictly increasing. While this result seems folklore, we are not aware of a reference and so provide a proof below. However, first note that in such a case, by virtue of Remark 3.7, the convex envelope $\check{\tau}$ is also a modulus of ϕ-regularity for F, provided τ was one, as we in particular have

$$F(x) \geq \tau(\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x)) \geq \check{\tau}(\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x))$$

for all $x \in X$. Now, to see the above result, note that since $\liminf_{x\to\infty} \tau(x)/x > 0$, we have $\tau(x)/x > c > 0$ for all $x \geq x_0 > 0$ for some c and x_0. Take $x_1 \in (0, x_0)$ and define $m := \min\{c, \tau(x_1)/(x_0 - x_1)\}$. The function $f(x) := m(x - x_1)$ is clearly convex and further satisfies $f \leq \tau$. In particular, we therefore have $0 < f(x) \leq \check{\tau}(x)$ for all $x \in (x_1, \infty)$ and as x_1 can be chosen arbitrarily close to 0, we have $\check{\tau}(x) > 0$ for all $x \in (0,\infty)$. This implies that $\check{\tau}$ is strictly increasing as for $0 < x < y$, using convexity of $\check{\tau}$ and that $\check{\tau}(0) = 0$, we have

$$\check{\tau}(x) = \check{\tau}(\tfrac{x}{y} y) \leq \tfrac{x}{y}\check{\tau}(y) < \check{\tau}(y),$$

using that $\check{\tau}(y) > 0$.

We return to the setting of the previous Example 3.2 and now consider problems of the form $F(z) := \int h(t,z)\,d\mu(t)$ where $h : T \times X \to [0,\infty]$ is a Carathéodory function over a σ-finite measure space (T, T, μ). In that context, we can generally guarantee the existence of a stochastic modulus of regularity already under a relatively broad probabilistic condition, stating that h has a pointwise modulus of regularity with positive probability. This property is readily checked in concrete scenarios, as we discuss briefly in Section 3.3, and in particular bears a resemblance to probabilistic regularity notions recently investigated by Asi and Duchi [5] (see Sections 4.1 and 4.2 therein) or Combettes and Madariaga [35] (see eq. (5.7) therein).

Lemma 3.10. *Let (T, T, μ) be a σ-finite measure space and let $h : T \times X \to [0,\infty]$ be a Carathéodory function, where we set $F(z) := \int h(t,z)\,d\mu(t)$. Suppose that $\tau, \sigma : [0,\infty) \to [0,\infty)$ are nondecreasing functions with $\tau(0), \sigma(0) = 0$ and $\tau(\varepsilon), \sigma(\varepsilon) > 0$ for $\varepsilon > 0$. If*

$$\mu\left(\left\{t \in T \mid h(t,x) \geq \tau(\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x))\right\}\right) \geq \sigma(\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x))$$

for all $x \in X$, then $(\sigma \cdot \tau)(\varepsilon) := \sigma(\varepsilon)\tau(\varepsilon)$ is a modulus of ϕ-regularity for F. In particular, whenever τ and σ are convex and $\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x)$ is integrable for all $x \in D$, then $\sigma \cdot \tau$ is a modulus of ϕ-regularity for F in mean w.r.t. D.

Proof. Given $x \in X$, write S_x for the set $\{t \in T \mid h(t,x) \geq \tau(\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x))\}$. Then we have

$$F(x) \geq \int_{S_x} h(t,x)\,d\mu(t) \geq \int_{S_x} \tau(\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x))\,d\mu(t) \geq \mu(S_x)\cdot\tau(\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x)) \geq (\sigma\cdot\tau)(\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x))$$

which yields the first part using Remark 3.7, as $\sigma \cdot \tau$ must also be nondecreasing. The second part then follows from Lemma 3.8 using the standard fact that the product of two convex, nondecreasing, nonnegative functions is convex. □

Remark 3.11. Let (T, T, μ) be a probability space. In the special case of σ defined as $\sigma(0) := 0$ and $\sigma(\varepsilon) := p$ for $\varepsilon > 0$, given a $p \in (0,1]$, Lemma 3.10 represents the following stochastic variant of Lemma 3.8: If, for any $x \in X$, we have $h(t,x) \geq \tau(\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x))$ with probability p w.r.t. μ, where τ is a suitable (in particular convex) function, then $p \cdot \tau$ is a modulus of ϕ-regularity for F in mean w.r.t. D.

3.3. **Examples from practice.** There are numerous concrete instantiations of the abstract notions of regularity presented above. As already highlighted, by virtue of Lemma 3.8, essentially all of the examples studied in [53] immediately lift to the stochastic setting. Among others, these encompass:

- *Fixed point problems*, formalized via $F(x) := d(Tx, x)$ for some measurable mapping $T : X \to X$ such that $\mathrm{Fix}T \neq \emptyset$ is closed. Here, explicit pointwise moduli of regularity can be constructed when, e.g., T is a quasi-contraction, a continuous orbital contraction, a retraction onto a subset of X or the composition of reflected resolvents for convex semi-algebraic sets in the sense of [23]. Explicit constructions of such pointwise moduli are given in Example 3.6 in [53] (with quasi-contractivity being a simple modification of the construction given therein). Moreover, these moduli of regularity are all linear, and hence in particular convex, and so lift to regularity in mean by Lemma 3.8.
- *Minimization problems*, formalized via $F(x) := f(x) - \min f$ for some measurable function $f : X \to (-\infty, +\infty]$ such that $\mathrm{argmin} f \neq \emptyset$ is closed. Here, explicit moduli of regularity can be constructed when, e.g., f has a τ-global weak sharp minimum for some strictly increasing $\tau : [0,\infty) \to [0,\infty)$ with $\tau(0) = 0$, i.e.

$$f(x) - \min f \geq \tau(\mathrm{dist}_{\mathrm{argmin} f}(x)) \text{ for all } x \in X.$$

Concretely, τ is then immediately a pointwise modulus of regularity. This notion was introduced in [53] (see Example 3.7 therein), extending the well-known notion of weak sharp minima [28] (see also [27, 43, 58]). Inverses of this property, that is increasing functions $\tilde{\tau}$ such that

$$\tilde{\tau}(f(x) - \min f) \geq \mathrm{dist}_{\mathrm{argmin} f}(x) \text{ for all } x \in X,$$

are also known as error bounds [22, 39]. In particular, weak sharp minima as defined in [28] arise from the above by considering $\tau(\varepsilon) = k\varepsilon$ for $k > 0$, so that this immediately induces regularity in mean by Lemma 3.8. However, by that lemma, the above pointwise property of course induces a regularity property in mean also for more general convex moduli τ, which in particular further encompasses polynomial growth conditions (see e.g. [87] among many others), that is

$$c(f(x) - \min f) \geq (\mathrm{dist}_{\mathrm{argmin} f}(x))^{\theta} \text{ for all } x \in X,$$

for some $c > 0$ and $\theta \geq 1$. The quadratic growth condition $\theta = 2$ in particular is closely related (and under suitable condition equivalent) to the well-known Polyak-Łojasiewicz, or more generally Kurdyka-Łojasiewicz conditions, where we refer e.g. to [21, 22, 89]. A particular situation where a function has weak sharp minima, and which moreover guarantees uniqueness of the solution, is e.g. when f is uniformly (and hence in particular strongly) quasiconvex.

- *Set-valued inclusion problems*, formalized via $F(x) := \mathrm{dist}(O_Y, A(x))$ for $A : X \to 2^Y$ where X, Y are two given metric spaces and $O_Y \in Y$ is a designated point such that $A^{-1}(O_Y) := \{x \in X \mid O_Y \in A(x)\} \neq \emptyset$ is closed. Here, explicit moduli of regularity can be constructed when e.g. A is τ-global metrically subregular for some strictly increasing $\tau : [0, \infty) \to [0, \infty)$ with $\tau(0) = 0$, i.e.

$$\mathrm{dist}_{A(x)}(O_Y) \geq \tau(\mathrm{dist}_{A^{-1}(O_Y)}(x)) \text{ for all } x \in X.$$

Concretely, τ is then immediately a pointwise modulus of regularity. A local variant of this generalized notion of metric subregularity was studied recently in [59], extending the well-known notion of (local) metric subregularity [56, 38] (a global variant of which was studied, in the context of stochastic iterations, in [49]). In particular, this usual notion of metric subregularity arises from the above by considering $\tau(\varepsilon) = \varepsilon/k$ for $k > 0$, which immediately implies regularity in mean by Lemma 3.8. However, as before this is not limited to that case but holds more generally whenever τ is convex, in particular encompassing polynomial growth conditions similar to the above, that is

$$c\mathrm{dist}_{A(x)}(O_Y) \geq (\mathrm{dist}_{A^{-1}(O_Y)}(x))^{\theta} \text{ for all } x \in X,$$

for some $c > 0$ and $\theta \geq 1$, and we refer again to [21, 22, 89] for discussions of such conditions for subgradients and other set-valued operators as well as their relation to Polyak-Łojasiewicz or Kurdyka-Łojasiewicz conditions. Another case where such a regularity conditions for set-valued inclusions arise naturally, and moreover guarantee uniqueness of the solution, is when, over Banach spaces, the operator is τ-uniformly accretive for a strictly increasing (convex) τ as above, covering in particular strongly accretive operators (see Example 3.9 in [53]). This moreover applies to other notions of monotonicity for set-valued operators, in particular uniformly and strongly monotone vector fields over Hadamard manifolds (see e.g. [57]) or Hadamard spaces (see [32]).

These examples already encompass many of the standard regularity assumptions from deterministic and stochastic optimization. In the setting of stochastic optimization, some further problem formulations and associated regularity notions, which are of a genuinely probabilistic

nature, feature prominently, and we now illustrate how some of these examples fit into our general notion:

- Consider the generic (sometimes called online) stochastic minimization problem

$$\text{find some minimizer of } \underline{f}(x) := \int f(e,x)\, d\mu(e)$$

for some suitable[4] function $f : E \times X \to (-\infty, +\infty]$, over some suitable probability space (E, E, μ). Assume that a minimizer $z \in \operatorname{argmin} \underline{f}$ exists. Instead of requiring a direct growth condition on $\underline{f}$, such as quadratic growth considered e.g. in [87], or the more general conditions surveyed above, which in particular leads to associated expected growth conditions (recall Lemma 3.8), one can assume a more pointwise and probabilistic condition. Assume for this that the above z is actually a minimizer almost everywhere, that is

$$f(e,z) = \inf_{x \in X} f(e,x) \text{ for } \mu\text{-almost every } e.$$

Problems satisfying this condition are called *interpolation problems* in [4] (or *easy problems* in [5], where it is however required that the above condition holds for all solutions z; see also e.g. [86] for similar such conditions). We can then consider the following probabilistic growth condition

$$\mu\left(\left\{e \in E \mid f(e,x) - f(e,z) \geq \tau(\operatorname{dist}_{\operatorname{argmin}\underline{f}}(x))\right\}\right) \geq p \text{ for all } x \in X,$$

for some $p \in (0,1]$ and a convex strictly increasing function $\tau : [0,\infty) \to [0,\infty)$ with $\tau(0) = 0$, i.e. that f satisfies a growth condition induced by τ with non-zero probability. By Lemma 3.10 (setting $h(e,x) := f(e,x) - f(e,z)$), we thereby in particular get that $p \cdot \tau$ is a modulus of regularity for $F(x) := \underline{f}(x) - \min \underline{f}$ in mean.

This in particular generalizes examples surveyed in [5] (see in particular Sections 4.1 and 4.2 therein), where the special cases of $\tau(\varepsilon) := \lambda\varepsilon$ or $\tau(\varepsilon) := \lambda\varepsilon^2$ are considered, which are reasonably easy to check in some situations as discussed in [5], in particular including (see Section 4.3 in [5]) overdetermined linear systems, data interpolation problems, or convex feasibility problems for suitable sets (such as halfspaces [5], or more generally convex semi-algebraic sets [23], as discussed also above).

- Consider the generic stochastic common fixed point problem

$$\text{find some point } z \text{ such that } T_k z = z \ \mathbb{P}\text{-a.s.}$$

for some suitable[5] family of mappings $(T_k)_{k \in K}$ over some measurable space (K, K) and a K-valued random variable $k : \Omega \to K$ over a probability space $(\Omega, \mathsf{F}, \mathbb{P})$. Assuming that a solution $z \in \operatorname{Fix} T := \{z \in X \mid T_k z = z \ \mathbb{P}\text{-a.s.}\}$ exists, we can consider the probabilistic condition

$$\mathbb{P}\left(\left\{\omega \in \Omega \mid d^2(T_{k(\omega)}x, x) \geq \tau(\operatorname{dist}^2_{\operatorname{Fix}T}(x))\right\}\right) \geq p \text{ for all } x \in X,$$

for some $p \in (0,1]$ and a convex strictly increasing function $\tau : [0,\infty) \to [0,\infty)$ with $\tau(0) = 0$, similar to before. Again, by Lemma 3.10 (setting $h(\omega, x) := d^2(T_k x, x)$), we thereby in particular get that $p \cdot \tau$ is a modulus of regularity for $F(x) := \mathbb{E}[d^2(T_k x, x)]$ in mean. For a linear function $\tau(\varepsilon) := \varepsilon/v$, given some $v \in [1,\infty)$, the above is a

[4] A very common assumption on f in particular is that f is a normal convex integrand, that is f is $\mathsf{E} \otimes \mathsf{B}(X)$-measurable and $f(e,\cdot)$ is proper, lower-semicontinuous and convex, which we discuss in our applications later.

[5] A common assumption might be that each T_k is continuous (e.g. nonexpansive), and that the function $(k,x) \mapsto T_k x$ is $\mathsf{K} \otimes \mathsf{B}(X)/\mathsf{B}(X)$-measurable, as we discuss in our applications later.

probabilistic variant of the common assumption of pointwise *linear regularity* of (T_k) in the sense of

$$\mathrm{dist}^2_{\mathrm{Fix}T}(x) \leq v\mathbb{E}[d^2(T_k x, x)] \text{ for all } x \in X,$$

as e.g. recently considered in the work of Combettes and Madariaga (cf. condition (5.10) in [35], and also Remark 5.6 of that paper which discusses related literature in which variants of this regularity notion appear), which by Lemma 3.8 of course also immediately induces a resulting regularity modulus for F in mean. A simple example in which linear regularity is achieved is for finite families $T_1, \ldots, T_N$ of nonexpansive mappings where $k : \Omega \to \{1, \ldots, N\}$ is some random variable with $0 < p_i := \mathbb{P}(E_i)$ for $E_i := \{k = i\}$ for all $i = 1, \ldots, N$. Then $\mathrm{Fix}T = \bigcap_{i=1}^N \mathrm{Fix}T_i$ and linear regularity follows, for example, from the piecewise property

$$\forall x \in X \left(\mathrm{dist}_{\mathrm{zer}F}(x) \leq ad(T_i x, x) \text{ for some } i \in \{1, \ldots, N\}\right),$$

which by Lemma 3.10 yields linear regularity with $v := a^2/\mu$ for $\mu := \min_{i\in\{1,\ldots,N\}} p_i$.

4. Rates for monotone stochastic processes under regularity conditions

We now utilise moduli of regularity in mean in order to derive explicit rates of convergence for stochastic methods that solve problems of the form $\mathrm{zer}F$. Here, we will focus on methods that are suitably monotone.

4.1. Stochastic quasi-Fejér monotonicity.

Our central notion is motivated by the property of stochastic quasi-Fejér monotonicity:

Definition 4.1 (Stochastic quasi-Fejér monotonicity)**.** Let (F_n) be a filtration and let (x_n) be an X-valued stochastic process adapted to (F_n). Then (x_n) is called stochastically ϕ-quasi-Fejér monotone w.r.t. $S \subseteq X$ and (F_n) if

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

for all $n \in \mathbb{N}$ and all $z \in S$, where $(\zeta_n), (\xi_n) \in \ell^1_+(\mathsf{F}_n)$.

Already in the deterministic context, quasi-Fejér monotonicity is one of the most fundamental concepts in the modern study of numerical algorithms (see e.g. [33, 34]), and this notion retains its relevance in stochastic contexts, where it is typically also referred to simply as a "supermartingale property", and indeed the convergence theory of supermartingales fundamentally underlies this notion (in particular the Robbins-Siegmund theorem [81]). In Euclidean contexts, this stochastic notion appears already in the pioneering works of Ermol'ev [40, 41, 42] together with a wide theory. In the general context of Hilbert spaces, central convergence results (almost surely and in mean) are then presented and streamlined in the seminal work of Combettes and Pesquet [36, 37], which were extended to the metric context of nonlinear Hadamard spaces in the recent work [77] (see also [69, 70] for different metric considerations).

As it appears above, this general notion of stochastic quasi-Fejér monotonicity was already investigated in the preceding work [69]. While this notion proved suitable for constructing rates of convergence under uniqueness conditions, it turns out that the generality gained by the abstract regularity notions considered here, which in particular allow for non-unique problems, requires us to consider a strengthened version of the above quasi-Fejér monotonicity property, where we allow z to be an S-valued F_n-measurable random variable:

Definition 4.2 (Strong stochastic quasi-Fejér monotonicity)**.** Let (F_n) be a filtration and let (x_n) be an X-valued stochastic process adapted to (F_n). Then (x_n) is called strongly stochastically ϕ-quasi-Fejér monotone w.r.t. $S \subseteq X$ and (F_n) if

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

for all $n \in \mathbb{N}$ and all S-valued F_n-measurable random variables z with $z \in S$ a.s. such that $\mathbb{E}[\phi(z, x_n)] < \infty$, where $(\zeta_n), (\xi_n) \in \ell^1_+(\mathsf{F}_n)$.

As we will see in Section 5 below, many standard methods satisfy our strong stochastic quasi-Fejér monotonicity property outright, virtue of the fact that they naturally satisfy an even stronger pointwise inequality. However, we can further justify the reach of our strong stochastic quasi-Fejér property by showing that it immediately arises from the standard property under a relaxed variant of the triangle inequality for ϕ, which is satisfied in many cases. In this context, we in particular generalise an argument set out in [35] to establish a similar fact, tailored to a specific iteration at hand, in Hilbert spaces and for $\phi = \|\cdot\|^2$ (contained in their Proposition 2.4 and Theorem 3.2 (iii)).

Definition 4.3 (Weak quasi-triangle inequality)**.** A distance ϕ is satisfies the weak quasi-triangle inequality if there is a concave and nondecreasing function $H : [0, \infty) \to [0, \infty)$ such that

$$\phi(x, y) \leq H(\phi(x, o) + \phi(y, o))$$

for any $x, y, o \in X$.

Related notions are e.g. studied in [46] in the context of generalized distance functions. We here restrict ourselves to the canonical examples of p-th orders of the metric, similar to Example 3.1 in [46], and certain examples of Bregman distances.

Example 4.4. The following distance functions satisfy the weak quasi-triangle inequality:

(1) In the case that $\phi = d^q$ for some $q \geq 1$, it follows by the (discrete) Jensen's inequality that

$$d^q(x, y) \leq 2^{q-1}(d^q(x, o) + d^q(y, o))$$

for all $x, y, o \in X$. Hence, $\phi = d^q$ satisfies the weak quasi-triangle inequality with function $H(a) := 2^{q-1}a$.

(2) In the case that $\phi = D_f$ over a reflexive Banach space $(X, \|\cdot\|)$, where $f : X \to \mathbb{R}$ is lsc, convex and Fréchet differentiable on X, assume further that there are nondecreasing functions $\theta, \Theta : (0, \infty) \to (0, \infty)$ such that $D_f(x, y) \geq \theta(b) \|x - y\|^2$ and $\|\nabla f(x) - \nabla f(y)\| \leq \Theta(b) \|x - y\|$ for any $b > 0$ and $x, y \in \overline{B}_b(0)$. Such functions are (essentially) considered over Euclidean spaces in particular in [9], where it is shown that they exist whenever f is *very strictly convex*. Related discussions for the infinite dimensional case can be found in [74]. In particular, by Lemma 2.5 in [74], the latter condition on the gradient in particular implies that $D_f(x, y) \leq \Theta(b) \|x - y\|^2$ for any $b > 0$ and $x, y \in \overline{B}_b(0)$. As such, for $b > 0$ and $x, y, o \in \overline{B}_b(0)$, we ultimately have

$$\begin{aligned} D_f(x, y) &\leq \Theta(b) \|x - y\|^2 \\ &\leq \frac{2\Theta(b)}{\theta(b)} \left(\theta(b) \|x - o\|^2 + \theta(b) \|y - o\|^2\right) \\ &\leq \frac{2\Theta(b)}{\theta(b)}(D_f(x, o) + D_f(y, o)), \end{aligned}$$

using in particular also item (1) above. Hence, in such a case, $\phi = D_f$ satisfies the weak quasi-triangle inequality on every ball $\overline{B}_b(0) \subseteq X$ with function $H(a) := (2\Theta(b)/\theta(b))a$.

We now give our central result on the relation between stochastic quasi-Fejér monotone and strongly stochastic quasi-Fejér monotone processes.

Proposition 4.5. *Let X be a separable and complete metric space and let $S \subseteq X$ be non-empty and closed. Further, let (x_n) be an X-valued stochastic process adapted to (F_n) which is stochastically ϕ-quasi-Fejér monotone w.r.t. S and (F_n). Suppose that $\mathbb{E}[\phi(o, x_n)] < \infty$ for all $n \in \mathbb{N}$, where $o \in S$ is fixed. If ϕ is such that $\phi(x, x) = 0$ for all $x \in X$, and ϕ satisfies the weak quasi-triangle inequality with a concave and nondecreasing function H, then (x_n) is strongly stochastically ϕ-quasi-Fejér monotone w.r.t. S and (F_n).*

Proof. We start by showing that the strong stochastic quasi-Fejér property holds for S-valued F_n-simple functions z, that is random variables z such that there are $z_0, \dots, z_k \in S$ and disjoint $A_0, \dots, A_k \in \mathsf{F}_n$ with $\bigcup_{i=0}^k A_i = \Omega$ and $z(\omega) = z_i$ if, and only if, $\omega \in A_i$. To see this, we observe for such z that

$$\begin{aligned}
\mathbb{E}[\phi(z, x_{n+1}) \mid \mathsf{F}_n] &= \mathbb{E}\left[\sum_{i=0}^k \phi(z_i, x_{n+1})\mathbf{1}_{A_i} \mid \mathsf{F}_n\right] \\
&= \sum_{i=0}^k \mathbb{E}[\phi(z_i, x_{n+1}) \mid \mathsf{F}_n]\mathbf{1}_{A_i} \\
&\leq (1+\zeta_n)\sum_{i=0}^k \phi(z_i, x_n)\mathbf{1}_{A_i} + \xi_n \\
&= (1+\zeta_n)\phi(z, x_n) + \xi_n,
\end{aligned}$$

where for the second equality, we used that $A_i \in \mathsf{F}_n$. Next, we show that for any S-valued F_n-measurable z such that $\mathbb{E}[\phi(z, x_n)] < \infty$, there exists a sequence (z_k) of S-valued F_n-simple functions that converge to z a.s. and satisfy $\sup_{k\in\mathbb{N}} \phi(z_k, o) \leq \phi(z, o) + 1$ a.s. To this end, write $\psi(x) := \phi(x, o)$ and let (p_k) be a countable dense subset of S where p_0 is chosen to satisfy $\psi(p_0) \leq \inf_{y\in S} \psi(y) + 1$, and define for $k \in \mathbb{N}$ and $y \in S$ the set $I_{k,y} \subset \mathbb{N}$ by

$$I_{k,y} := \{i \in \{0, \dots, k\} \mid \psi(p_i) \leq \psi(y) + 1\},$$

noting that $0 \in I_{k,y}$ for all k and y. Now, fixing such a S-valued F_n-measurable z with $\mathbb{E}[\phi(z, x_n)] < \infty$, for each $k \in \mathbb{N}$, define $(A_i^{k,z})$ for $i \in \{0, \dots, k\}$ by

$$\begin{aligned}
A_0^{k,z} &:= \left\{\omega \in \Omega \mid d(z(\omega), p_0) = \min_{j\in I_{k,z(\omega)}} d(z(\omega), p_j)\right\}, \\
A_i^{k,z} &:= \left\{\omega \in \Omega \mid i \in I_{k,z(\omega)} \text{ and } \min_{j\in I_{i-1,z(\omega)}} d(z(\omega), p_j) > d(z(\omega), p_i) = \min_{j\in I_{k,z(\omega)}} d(z(\omega), p_j)\right\},
\end{aligned}$$

noting that some of these sets might be empty. It is easy to check that $A_i^{k,z} \in \mathsf{F}_n$ when z is F_n-measurable, using also that ψ is continuous, and moreover we clearly have $\bigcup_{i=0}^k A_i^{k,z} = \Omega$ where this is a union of disjoint sets. Therefore for $k \in \mathbb{N}$, defining $z_k(\omega) := p_i$ if, and only if, $\omega \in A_i^{k,z}$, we have that z_k is an S-valued F_n-simple function. We see by definition that for any $\omega \in \Omega$ we have $\psi(z_k(\omega)) \leq \psi(z(\omega)) + 1$, and moreover

$$z_k(\omega) = p_{i_k} \text{ for the least } i_k \in I_{k,z(\omega)} \text{ such that } d(z(\omega), p_{i_k}) = \min_{j\in I_{k,z(\omega)}} d(z(\omega), p_j),$$

and since (p_k) is dense in S and ψ is continuous, we have $z_k(\omega) \to z(\omega)$. We now finish the proof by combining the first two steps. Fixing an S-valued F_n-measurable z with $\mathbb{E}[\phi(z, x_n)] < \infty$,

and an approximating sequence (z_k) as above, it follows by the first part that

$$(*) \qquad \mathbb{E}[\phi(z_k, x_{n+1}) \mid \mathsf{F}_n] \leq (1+\zeta_n)\phi(z_k, x_n) + \xi_n \quad \text{a.s.}$$

for any $n, k \in \mathbb{N}$. Using the weak quasi-triangle inequality for ϕ, we get

$$\phi(z, o) \leq H(\phi(z, x_n) + \phi(o, x_n)),$$

which, using concavity of H, yields

$$\mathbb{E}[\phi(z, o)] \leq H(\mathbb{E}[\phi(z, x_n)] + \mathbb{E}[\phi(o, x_n)]) < \infty.$$

Using the weak quasi-triangle inequality again, together with $\phi(x, x) = 0$ for all $x \in X$, we get

$$\phi(x_{n+1}, o) \leq H(\phi(x_{n+1}, x_{n+1}) + \phi(o, x_{n+1})) = H(\phi(o, x_{n+1}))$$

and using the quasi-triangle inequality a third time, as well as that H is nondecreasing, we have

$$\begin{aligned} \phi(z_k, x_{n+1}) &\leq H(\phi(z_k, o) + \phi(x_{n+1}, o)) \\ &\leq H(\phi(z, o) + \phi(x_{n+1}, o) + 1) \\ &\leq H(\phi(z, o) + H(\phi(o, x_{n+1})) + 1) =: Y \quad \text{a.s.} \end{aligned}$$

So, using the concavity of H twice, as well as that H is nondecreasing, we get

$$\mathbb{E}[Y] \leq H(\mathbb{E}[\phi(z, o)] + H(\mathbb{E}[\phi(o, x_{n+1})]) + 1) < \infty$$

from the above, together with our integrability assumptions. Since $z_k \to z$ and thus $\phi(z_k, x_{n+1}) \to \phi(z, x_{n+1})$ a.s. by left continuity of ϕ, by the conditional dominated convergence theorem we have

$$\mathbb{E}[\phi(z_k, x_{n+1}) \mid \mathsf{F}_n] \to \mathbb{E}[\phi(z, x_{n+1}) \mid \mathsf{F}_n] \quad \text{a.s.}$$

and so taking limits in $(*)$, using also that $\phi(z_k, x_n) \to \phi(z, x_n)$ a.s., the result is obtained. □

Remark 4.6. The above Proposition 4.5 features the assumption that $\mathbb{E}[\phi(o, x_n)] < \infty$ for all $n \in \mathbb{N}$ and some fixed $o \in S$. This is immediately guaranteed in essentially all cases via the stochastic ϕ-quasi-Fejér monotonicity w.r.t. S, provided that the initial value x_0 of the process satisfies $\mathbb{E}[\phi(o, x_0)] < \infty$ itself.

4.2. **Approximation properties and effective rates of convergence.** In the context of a stochastic regularity condition as considered in the previous section, we can now guarantee the convergence of strongly stochastic quasi-Fejér monotone processes under a very mild asymptotic approximation property, which takes the following (quantitative) form:

Definition 4.7 (lim inf-property in mean)**.** Let $F : X \to [0, \infty]$ be measurable. An X-valued stochastic process (x_n) has the lim inf-property in mean w.r.t. 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 mean for (x_n) w.r.t. F.

Such a lim inf-bound in mean can then be combined with a modulus of regularity to give a general construction for a rate of convergence, both in mean and almost surely, for the respective stochastic process. This construction, which will occupy us for the most part of the rest of this section, further depends on some minor data, quantitatively witnessing the characteristic properties of the associated correction terms in the quasi-Fejér monotonicity condition.

Concretely, recall that (strong) stochastic ϕ-quasi-Fejér monotonicity features two stochastic correction terms broadening the supermartingale condition, a multiplicative term $(1 + \zeta_n)$ and

an additive term ξ_n, with $(\zeta_n), (\xi_n) \in \ell^1_+(\mathsf{F}_n)$. In the following, we will slightly upgrade and simultaneously quantitatively resolve these integrability properties as follows. For $(\xi_n) \in \ell^1_+(\mathsf{F}_n)$, we will further assume that $\sum_{n=0}^{\infty} \mathbb{E}[\xi_n] < \infty$, quantitatively witnessed by a corresponding rate of convergence $\chi : (0,\infty) \to \mathbb{N}$, i.e.

$$\forall \varepsilon > 0 \left(\sum_{n=\chi(\varepsilon)}^{\infty} \mathbb{E}[\xi_n] < \varepsilon \right).$$

For $(\zeta_n) \in \ell^1_+(\mathsf{F}_n)$, which can be equivalently expressed by $\prod_{n=0}^{\infty}(1+\zeta_n) < \infty$ a.s., we will further assume the existence of a uniform almost-sure bound $K > 0$, i.e.

$$\prod_{n=0}^{\infty}(1+\zeta_n) < K \text{ a.s.}$$

While both requirements are actually qualitative strengthenings of the properties $(\zeta_n), (\xi_n) \in \ell^1_+(\mathsf{F}_n)$, they allows for a much smoother development of the associated quantitative results, and at the same time are practically speaking only very mild restrictions, as many algorithms actually confine to a quasi-Fejér monotonicity property where (ζ_n) is a sequence of reals, and in many cases even is constantly 0, and where (ξ_n) is summable in mean. With these minor quantitative moduli in place, we now are in the position to give our main abstract quantitative convergence theorem, formulated for consistent distances:

Theorem 4.8. *Let $F : X \to [0,\infty]$ be measurable, and such that $\mathrm{zer}F$ is a closed non-empty set. Further, let $\phi : X \times X \to [0,\infty)$ be a Carathéodory distance and assume that ϕ is consistent with a modulus $\theta : [0,\infty) \to [0,\infty)$ which is nondecreasing and convex with $\theta(0) = 0$ and $\theta(\varepsilon) > 0$ for $\varepsilon > 0$. Let (F_n) be a filtration and let (x_n) be an X-valued stochastic process adapted to (F_n) such that:*

(1) (x_n) is strongly stochastically ϕ-quasi-Fejér monotone w.r.t. $\mathrm{zer}F$ and (F_n) and error sequences $(\zeta_n), (\xi_n) \in \ell^1_+(\mathsf{F}_n)$, where $K > 0$ is a uniform almost-sure bound for $\prod_{n=0}^{\infty}(1+\zeta_n) < \infty$ and $\chi : (0,\infty) \to \mathbb{N}$ is a rate of convergence for $\sum_{n=0}^{\infty} \mathbb{E}[\xi_n] < \infty$.

(2) (x_n) has the $\liminf$-property in mean w.r.t. F with a $\liminf$-bound $\varphi : (0,\infty) \times \mathbb{N} \to (0,\infty)$.

Lastly, let $\tau : (0,\infty) \to (0,\infty)$ be a modulus of ϕ-regularity for F in mean w.r.t. D, where D is a collection of X-valued random variables with $(x_n) \subseteq D$. Then there is a $\mathrm{zer}F$-valued random variable x such that $d(x_n, x) \to 0$ in mean and a.s., with rates

$$\forall \varepsilon > 0 \; \forall n \geq \rho(\theta(\varepsilon/2)) \left(\mathbb{E}[d(x_n, x)] < \varepsilon \right)$$

as well as

$$\forall \lambda, \varepsilon > 0 \left(\mathbb{P}(\exists n \geq \rho(\lambda\theta(\varepsilon/2))(d(x_n, x) \geq \varepsilon)) < \lambda \right)$$

where $\rho(\varepsilon) := \varphi\left(\tau\left(\varepsilon/3K\right), \chi\left(\varepsilon/3K\right)\right)$.

Proof. Fix $\varepsilon > 0$ and let $\delta := \theta(\varepsilon/2)K^{-1}$ and $N := \chi(\delta/3)$. By the liminf property, we have

$$\mathbb{E}[F(x_n)] < \tau\left(\delta/3\right)$$

for some $n \in [N; \varphi(\tau(\delta/3), N)] = [N, \rho(\theta(\varepsilon/2))]$. Using that $(x_n) \subseteq D$ and that τ is a modulus of regularity, we get

$$\mathbb{E}[\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x_n)] < \delta/3.$$

Using Corollary 2.11, let z be an X-valued F_n-measurable random variable such that $z \in \mathrm{zer}F$ a.s. and

$$\mathbb{E}[\phi(z, x_n)] \leq \mathbb{E}[\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x_n)] + \delta/3 < 2\delta/3.$$

Now consider the stochastic process $(U_k)_{k\geq n}$ defined by

$$U_k := \frac{\phi(z,x_k)}{y_{k-1}} + \mathbb{E}\left[\sum_{i=k}^{\infty} \frac{\xi_i}{y_i} \mid \mathsf{F}_k\right] \text{ where } y_j := \prod_{i=0}^{j}(1+\zeta_i).$$

Since x_k and z are F_k-measurable (the latter since z is already F_n-measurable), by Lemma 2.6, (1) we have that $\phi(z,x_k)$ and thus U_k is F_k-measurable for all $k \geq n$. Using that (x_n) is strongly stochastically ϕ-quasi-Fejér monotone w.r.t. zer$F$ and (F_n) and that z is F_k-measurable for all $k \geq n$ and $\mathbb{E}[\phi(z,x_n)] < \infty$, it follows by induction that $\mathbb{E}[\phi(z,x_k)] < \infty$ for all $k \geq n$. Therefore, the strong stochastic ϕ-quasi-Fejér monotonicity now implies that $(U_k)_{k\geq n}$ is a supermartingale w.r.t. $(\mathsf{F}_k)_{k\geq n}$. Concretely, using basic properties of conditional expectations:

$$\begin{aligned}
\mathbb{E}[U_{k+1} \mid \mathsf{F}_k] &= \mathbb{E}\left[\frac{\phi(z,x_{k+1})}{y_k} \mid \mathsf{F}_k\right] + \mathbb{E}\left[\mathbb{E}\left[\sum_{i=k+1}^{\infty} \frac{\xi_i}{y_i} \mid \mathsf{F}_{k+1}\right] \mid \mathsf{F}_k\right] \\
&= \frac{\mathbb{E}\left[\phi(z,x_{k+1}) \mid \mathsf{F}_k\right]}{y_k} + \mathbb{E}\left[\sum_{i=k+1}^{\infty} \frac{\xi_i}{y_i} \mid \mathsf{F}_k\right] \\
&\leq \frac{(1+\zeta_k)\phi(z,x_k)}{y_k} + \frac{\xi_k}{y_k} + \mathbb{E}\left[\sum_{i=k+1}^{\infty} \frac{\xi_i}{y_i} \mid \mathsf{F}_k\right] \\
&= \frac{\phi(z,x_k)}{y_{k-1}} + \mathbb{E}\left[\sum_{i=k}^{\infty} \frac{\xi_i}{y_i} \mid \mathsf{F}_k\right] = U_k.
\end{aligned}$$

Further, for any $k \geq n$, we have

$$U_k = \frac{\phi(z,x_k)}{y_{k-1}} + \mathbb{E}\left[\sum_{i=k}^{\infty} \frac{\xi_i}{y_i} \mid \mathsf{F}_k\right] \leq \phi(z,x_k) + \mathbb{E}\left[\sum_{i=k}^{\infty} \xi_i \mid \mathsf{F}_k\right]$$

using $y_j \geq 1$ for all $j \in \mathbb{N}$. Therefore, taking expectations, we have for any $k \geq n$ that

$$\mathbb{E}[U_k] \leq \mathbb{E}[U_n] \leq \mathbb{E}[\phi(z,x_n)] + \sum_{i=n}^{\infty} \mathbb{E}[\xi_i] < 2\delta/3 + \delta/3 = \delta,$$

where we use that $n \geq N = \chi(\delta/3)$. As

$$\frac{\phi(z,x_k)}{K} \leq \frac{\phi(z,x_k)}{y_{k-1}} \leq U_k,$$

it follows that $\mathbb{E}[\phi(z,x_k)K^{-1}] < \delta = \theta(\varepsilon/2)K^{-1}$ and we thus also have $\mathbb{E}[\phi(z,x_k)] < \theta(\varepsilon/2)$ for all $k \geq n$. As in Remark 2.3, we have $\phi(x,y) \geq \theta(d(x,y))$ for all $x, y \in X$, so that since θ is convex, we get

$$\theta(\mathbb{E}[d(z,x_k)]) \leq \mathbb{E}[\theta(d(z,x_k))] \leq \mathbb{E}[\phi(z,x_k)] < \theta(\varepsilon/2)$$

by Jensen's inequality. As θ is nondecreasing, we get $\mathbb{E}[d(z,x_k)] < \varepsilon/2$ for all $k \geq n$, and so we have that (x_n) is Cauchy in mean, with rate $\rho(\theta(\varepsilon/2))$. Suppose now $\delta := \lambda\theta(\varepsilon/2)K^{-1}$ instead. Then as above we get some $n \leq \rho(\lambda\theta(\varepsilon/2))$ with $\mathbb{E}[U_k] < \delta$ for all $k \geq n$, so that by Ville's inequality we have

$$\begin{aligned}
\mathbb{P}(\exists k \geq \rho(\lambda\theta(\varepsilon/2))(\phi(z,x_k) \geq a)) &= \mathbb{P}(\exists k \geq \rho(\lambda\theta(\varepsilon/2))(\phi(z,x_k)/K \geq a/K)) \\
&\leq \mathbb{P}(\exists k \geq \rho(\lambda\theta(\varepsilon/2))(U_k \geq a/K)) \\
&\leq \frac{\mathbb{E}[U_{\rho(\lambda\theta(\varepsilon/2))}]}{aK^{-1}} < \frac{\delta}{aK^{-1}}
\end{aligned}$$

for any $a > 0$. Setting $a := \theta(\varepsilon/2)$, we observe that

$$\begin{aligned}\mathbb{P}(\exists k, l \geq \rho(\lambda\theta(\varepsilon/2))(d(x_k, x_l) \geq \varepsilon) &\leq \mathbb{P}(\exists k, l \geq \rho(\lambda\theta(\varepsilon/2))(d(z, x_k) + d(z, x_l) \geq \varepsilon)) \\ &\leq \mathbb{P}(\exists k \geq \rho(\lambda\theta(\varepsilon/2))(d(z, x_k) \geq \varepsilon/2)) \\ &\leq \mathbb{P}(\exists k \geq \rho(\lambda\theta(\varepsilon/2))(\phi(z, x_k) \geq \theta(\varepsilon/2))) \\ &< \frac{\delta}{\theta(\varepsilon/2)K^{-1}} = \lambda.\end{aligned}$$

This shows that (x_n) is Cauchy a.s. with rate $\rho(\lambda\theta(\varepsilon/2))$, and by inspecting the above calculations we also see that $\mathrm{dist}_{\mathrm{zer}F}(x_n) \to 0$ a.s., with the same rate. In particular, we thus have that (x_n) almost surely converges to some measurable x, and it follows immediately that this is with the same rate. Lastly, we also have

$$\mathrm{dist}_{\mathrm{zer}F}(x) \leq \mathrm{dist}_{\mathrm{zer}F}(x_n) + d(x_n, x)$$

so that $\mathrm{dist}_{\mathrm{zer}F}(x) = 0$ a.s. As $\mathrm{zer}F$ is closed, we have $x \in \mathrm{zer}F$ a.s. It follows immediately that (x_n) also converges to this limit in mean with rate $\rho(\theta(\varepsilon/2))$. □

Remark 4.9. While formulated for Caratheodory distances ϕ above, it follows directly from the proof that Theorem 4.8 holds whenever

(1) ϕ is measurable w.r.t. $\mathsf{B}(X) \otimes \mathsf{B}(X)$,
(2) $\mathrm{dist}^{\phi}_{\mathrm{zer}F}$ is measurable for any non-empty closed $S \subseteq X$,
(3) for any $n \in \mathbb{N}$, $\mathrm{dist}^{\phi}_{\mathrm{zer}F}$ has F_n-measurable approximations in mean w.r.t. D, that is for all F_n-measurable $x \in D$ and any $\varepsilon > 0$, there exists some X-valued F_n-measurable random variable z such that $z \in \mathrm{zer}F$ a.s. and $\mathbb{E}[\phi(z, x)] \leq \mathbb{E}[\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x)] + \varepsilon$.

Remark 4.10. It follows immediately by inspection of the proof that the convexity assumption for the consistency modulus θ featuring in Theorem 4.8 is not necessary to establish the almost-sure convergence together with the corresponding rate. In particular, towards establishing convergence in mean, this assumption may hence be bypassed by, instead, assuming the uniform integrability of the sequence. This can in particular be made quantitative by assuming corresponding so-called moduli of uniform integrability for the sequence as studied in [79] (see also [70]), but we do not discuss this here any further.

If we only care for the distance to the solution set, the assumption of strong stochastic quasi-Fejér monotonicity can be weakened to the "normal" variant:

Theorem 4.11. *Let $F : X \to [0, \infty]$ be measurable, and such that $\mathrm{zer}F$ is a closed non-empty set. Further, let $\phi : X \times X \to [0, \infty)$ be a Carathéodory distance. Let (F_n) be a filtration and let (x_n) be an X-valued stochastic process adapted to (F_n) such that:*

(1) (x_n) is stochastically ϕ-quasi-Fejér monotone w.r.t. $\mathrm{zer}F$ and (F_n) and error sequences $(\zeta_n), (\xi_n) \in \ell^1_+(\mathsf{F}_n)$, where $K > 0$ is a uniform almost-sure bound for $\prod_{n=0}^{\infty}(1 + \zeta_n) < \infty$ and $\chi : (0, \infty) \to \mathbb{N}$ is a rate of convergence for $\sum_{n=0}^{\infty} \mathbb{E}[\xi_n] < \infty$.
(2) (x_n) has the $\liminf$-property in mean w.r.t. F with a $\liminf$-bound $\varphi : (0, \infty) \times \mathbb{N} \to (0, \infty)$.

Lastly, let $\tau : (0, \infty) \to (0, \infty)$ be a modulus of ϕ-regularity for F in mean w.r.t. D, where D is a collection of X-valued random variables with $(x_n) \subseteq D$. Then $\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x_n) \to 0$ in mean and a.s., with rates

$$\forall \varepsilon > 0 \; \forall n \geq \rho(\varepsilon) \left(\mathbb{E}[\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x_n)] < \varepsilon\right)$$

as well as

$$\forall \lambda, \varepsilon > 0 \left(\mathbb{P}(\exists n \geq \rho(\lambda\varepsilon)(\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x_n) \geq \varepsilon)) < \lambda \right)$$

where ρ is as in Theorem 4.8. If ϕ is uniformly consistent with a modulus $\theta : [0,\infty) \to [0,\infty)$ which is nondecreasing and convex with $\theta(0) = 0$ and $\theta(\varepsilon) > 0$ for $\varepsilon > 0$, then we further have $\mathrm{dist}_{\mathrm{zer}F}(x_n) \to 0$ in mean and a.s., with rates $\rho(\theta(\varepsilon))$ and $\rho(\lambda\theta(\varepsilon))$ respectively.

Proof. Given any $z \in \mathrm{zer}F$, we have $\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x_{n+1}) \leq \phi(z, x_{n+1})$ and hence we get

$$\mathbb{E}[\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x_{n+1}) \mid \mathsf{F}_n] \leq \mathbb{E}[\phi(z, x_{n+1}) \mid \mathsf{F}_n] \leq (1+\zeta_n)\phi(z, x_n) + \xi_n$$

for any such $z \in \mathrm{zer}F$. Taking the infimum over z, we get

$$\mathbb{E}[\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x_{n+1}) \mid \mathsf{F}_n] \leq (1+\zeta_n)\mathrm{dist}^{\phi}_{\mathrm{zer}F}(x_n) + \xi_n.$$

The remainder of the proof now follows (a simplification of) the arguments for Theorem 4.8 (now using Lemma 2.6, (2) to establish that the main supermartingale is adapted to (F_n)) and is omitted. □

It is important to stress that our abstract results can be refined to fit more specific conditions on (x_n) and F, and in that way be used to guarantee stronger convergence guarantees. For example, if the associated regularity modulus is linear, and the error sequences are decaying suitably fast, then we can also obtain linear rates of convergence, even in the form of non-asymptotic guarantees. This can be done by utilizing an almost identical strategy as that applied to the case of unique zeros in [69].[6] In that case, we will assume that the process (x_n) is stochastically quasi-Fejér monotone in a *strict sense*, that is it satisfies the stricter inequality

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

for all $n \in \mathbb{N}$ and all $z \in \mathrm{zer}F$, matching more closely the notion of stochastic quasi-Fejér monotonicity studied e.g. over Hilbert spaces in [36]. Beyond this slightly extended property, we further rely on a folklore quantitative result for real recursive inequalities (see e.g. [68] for similar such results):

Lemma 4.12 (see e.g. Lemma 3.5 in [69])**.** *Suppose that (x_n) is a sequence of nonnegative reals such that for $c > 1$, $d \geq 0$ and $r \in \mathbb{N}\backslash\{0\}$, we have*

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

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

$$x_n \leq \frac{u}{n+r} \quad \textit{for} \quad u \geq \max\left\{\frac{d}{c-1}, r x_0\right\}.$$

Our result on fast non-asymptotic guarantees is then readily derived:

Theorem 4.13. *Let $F : X \to [0,\infty]$ be measurable, and such that $\mathrm{zer}F$ is a closed non-empty set. Further, let $\phi : X \times X \to [0,\infty)$ be a Carathéodory distance. Let (F_n) be a filtration and let (x_n) be an X-valued stochastic process adapted to (F_n) which is strictly stochastically ϕ-quasi-Fejér monotone w.r.t. $\mathrm{zer}F$ and (F_n), i.e.*

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

[6] See Theorem 5.7 therein. In fact, the present result arises as a direct application of Theorem 3.6 given in [69]. However, we present the (rather short) argument tailored to the present situation for completeness.

for all $n \in \mathbb{N}$ and all $z \in \operatorname{zer} F$, where $(\zeta_n), (\eta_n)$ are sequences of nonnegative reals and (ξ_n) are nonnegative random variables such that

$$\mathbb{E}[\xi_n] \leq d/(n+r)^2 \text{ and } \zeta_n + c/(n+r) \leq t\eta_n$$

for all $n \in \mathbb{N}$ where $c > 1$, $d \geq 0$ and $r \in \mathbb{N} \setminus \{0\}$. Further suppose that $K \geq 1$ and $L > 0$ are such that $\prod_{i=0}^{\infty}(1+\zeta_i) < K$ and $L \geq \mathbb{E}[\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x_0)]$. Lastly, let D be a collection of X-valued random variables with $(x_n) \subseteq D$ and such that $\mathbb{E}[F(x)] \geq t\mathbb{E}[\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x)]$ for all $x \in D$. Then

$$\mathbb{E}[\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x_n)] \leq \frac{u}{n+r} \text{ for } u \geq \max\left\{\frac{d}{c-1}, rL\right\}$$

as well as

$$\mathbb{P}\left(\exists m \geq n(\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x_n) \geq \varepsilon)\right) \leq \frac{1}{\varepsilon} \cdot \frac{K(u+2d)}{n+r}.$$

Proof. Similarly to Theorem 4.11, we obtain

$$\mathbb{E}[\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x_{n+1}) \mid \mathsf{F}_n] \leq (1+\zeta_n)\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x_n) - \eta_n F(x_n) + \xi_n.$$

Integrating the inequality, we get

$$\begin{aligned}\mathbb{E}[\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x_{n+1})] &\leq (1+\zeta_n)\mathbb{E}[\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x_n)] - \eta_n \mathbb{E}[F(x_n)] + \mathbb{E}[\xi_n] \\ &\leq (1+\zeta_n - t\eta_n)\mathbb{E}[\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x_n)] + \mathbb{E}[\xi_n] \\ &\leq \left(1 - \frac{c}{n+r}\right)\mathbb{E}[\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x_n)] + \frac{d}{(n+r)^2}.\end{aligned}$$

Applying Lemma 4.12 yields the rate for $\mathbb{E}[\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x_n)]$. For the second claim, we proceed similar to the proof of Theorem 4.8. Concretely, note first that

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

so that

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

We define

$$U_n := \frac{\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x_n)}{y_{n-1}} + \mathbb{E}\left[\sum_{i=n}^{\infty}\frac{\xi_i}{y_i} \mid \mathsf{F}_n\right], \text{ where } y_j := \prod_{i=0}^{j}(1+\zeta_i)$$

and, analogously to Theorem 4.8 (now using Lemma 2.6, (2) to establish measurability), we can then derive that (U_n) is a supermartingale. Combining the rate for $\mathbb{E}[\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x_n)]$ with the above bounds yields $\mathbb{E}[U_n] \leq (u+2d)(n+r)$. Using Ville's inequality, this yields

$$\mathbb{P}\left(\exists m \geq n(\operatorname{dist}^{\phi}_{\operatorname{zer} F}(x_m) \geq \varepsilon)\right) \leq \mathbb{P}\left(\exists m \geq n(U_m \geq \varepsilon/K)\right) \leq \frac{K}{\varepsilon} \cdot \mathbb{E}[U_n] \leq \frac{1}{\varepsilon} \cdot \frac{K(u+2d)}{n+r}. \quad \square$$

5. Applications to stochastic algorithms

In this final section we apply our results to concrete stochastic algorithms. In each case, we work in the rather general context of geodesic metric spaces with nonpositive curvature. These spaces were introduced by Alexandrov [1] and are often called CAT(0) spaces after Gromov [47]. We refer to [2, 25] for a comprehensive overview of geodesic and CAT(0) spaces and further refer to [12] for a shorter treatment focused on aspects of convex analysis and optimization.

We introduce background from this class of spaces only as needed in each application. Beyond this, we only need the following few notions. In a metric space (X, d), geodesics are isometries

$\gamma : [0, l] \to X$, said to join $\gamma(0)$ and $\gamma(l)$. The space is called (uniquely) geodesic if every two points are joined by a (unique) geodesic. A geodesic metric space (X, d) is called a CAT(0) space (also called a space of nonpositive curvature in the sense of Alexandrov) if it satisfies

$$d^2(\gamma(tl), x) \leq (1-t)d^2(\gamma(0), x) + td^2(\gamma(l), x) - t(1-t)d^2(\gamma(0), \gamma(l)) \tag{CN}$$

for all $x \in X$, $t \in [0, 1]$ and all geodesics $\gamma : [0, l] \to X$, (an extension of) the so-called Bruhat-Tits CN-inequality [26]. Any CAT(0) space is uniquely geodesic, and a complete CAT(0) space is called a Hadamard space. In Hadamard spaces, we generally write $(1-\lambda)x \oplus \lambda y$ for the point $\gamma(\lambda d(x, y))$ on the unique geodesic $\gamma : [0, d(x, y)] \to X$ joining x and y.

5.1. **Stochastic proximal point methods.** The first method we study will be the classic stochastic proximal point method. In a deterministic context, where the method originates with the work of Rockafellar [83], Martinet [63] as well as Brézis and Lions [24], the proximal point method was extended to Hadamard spaces by Bačák [10], establishing weak convergence (which, by Güler's seminal work [48], is the most one can hope for already in Hilbert spaces).

The stochastic proximal point method, widely studied over Euclidean and Hilbert spaces (we refer to [5, 15, 16, 17, 68, 85], among many others, for various such discussions), was lifted to the setting of (separable) Hadamard spaces in the work [13], building on preceding work [11] on a splitting proximal point method with random order for finite sums of convex functions over similar spaces (see also [76] for a recent related variant for general perturbed strongly monotone vector fields over Hadamard spaces). As it is in particular over such spaces, without additional differential structure, where proximal point methods gain relevance compared to gradient descent (we refer to [13] as well as [15] for further discussions), we focus on these extensions.

Let $(\Omega, \mathsf{F}, \mathbb{P})$ and (E, E, μ) be probability spaces, with (E, E, μ) complete, and let X now be a separable Hadamard space. In analogy to [82] (see also [31]), let $f : E \times X \to (-\infty, +\infty]$ be a normal convex integrand, i.e. $f(e, \cdot)$ is proper, lower-semicontinuous (lsc) and convex[7] for all $e \in E$ and f is $\mathsf{E} \otimes \mathsf{B}(X)$-measurable. Defining $\underline{f}(x) := \int f(e, x)\, d\mu(e)$ and assuming that $\underline{f}$ is proper and that $\operatorname{argmin}\underline{f} \neq \emptyset$, our problem is to

$$\text{find some element of } \operatorname{argmin}\underline{f}.$$

We capture this problem in our general setup via $F(x) := \underline{f}(x) - \min \underline{f}$ (recall Section 3.3). Note that $\underline{f}(x)$ and hence F are measurable by Fubini's theorem and that $\underline{f}(x)$ is lsc by Fatou's lemma, so that $\operatorname{argmin}\underline{f} = \operatorname{zer} F$ is closed.

To now introduce the method, define the proximal map of f via

$$\operatorname{prox}_\lambda^f(e, x) := \operatorname{argmin}_{y \in X} \left\{ f(e, y) + \frac{1}{2\lambda} d^2(x, y) \right\},$$

which is well-defined for all $e \in E$, $x \in X$ and $\lambda > 0$ (see e.g. [50] or [66]). Further, $\operatorname{prox}_\lambda^f(e, \cdot)$ is nonexpansive for any $e \in E$ and $\lambda > 0$ (see e.g. Lemma 4 in [50]), and also $\operatorname{prox}_\lambda^f(\cdot, x)$ is measurable for any $x \in X$ and $\lambda > 0$. Hence, $\operatorname{prox}_\lambda^f$ is a Carathéodory function and so in particular $\mathsf{E} \otimes \mathsf{B}(X)$-measurable.

The stochastic proximal point method is then given by the iteration

$$x_{n+1} := \operatorname{prox}_{\lambda_n}^f(\xi_{n+1}, x_n), \tag{SPPA}$$

[7]Given a Hadamard space X, recall that a function $f : X \to (-\infty, +\infty]$ is called lsc if $\liminf_{n\to\infty} f(x_n) \geq f(x)$ whenever $x_n \to x$ in X, and convex if $f \circ \gamma$ is convex for any geodesic γ in X.

given a starting point $x_0 \in X$ and sequences (λ_n) of positive reals as well as (ξ_{n+1}) of random variables $\Omega \to E$, for which we assume that

(SPPA-A1) (ξ_{n+1}) are i.i.d. with distribution μ and $\sum_{n\in\mathbb{N}} \lambda_n = \infty, \sum_{n\in\mathbb{N}} \lambda_n^2 < \infty$.

The work [13] in particular relies on a certain weak growth condition on the integrand introduced therein, which is a generalization of many of the common growth conditions from the literature, in particular of Lipschitz continuity of the functional. For simplicity however, we here assume the following (slightly stronger) Lipschitz-type assumption (used throughout the literature on this method in linear spaces, see also e.g. [73] for spaces with bounded curvature): Assume there exists a positive function $L \in L^2(E, \mu)$ such that

(SPPA-A2) $$f(e, x) - f(e, y) \leq L(e)d(x, y)$$

for all $x, y \in X$ and almost all $e \in E$.[8]

Now, as mentioned above, without any regularity assumptions the deterministic proximal point method in general only converges weakly. For the stochastic proximal point method the situation is even more dire, as without regularity convergence can in general only be guaranteed on locally compact spaces and a.s. weak convergence remains, even on infinite dimensional Hilbert spaces, an open problem (see [14]).

In both cases, there are no effective convergence guarantees in general. These however can be obtained via additional regularity assumptions. The most common assumption used in the literature is that of strong or at least uniform convexity of the function. Next to a very large number of works in linear spaces, such rates for the deterministic case over Hadamard spaces are also discussed in [10] (see also [55]). The stochastic case, in particular over Hadamard spaces, is considerably less populated and the main strong convergence result under a regularity assumption in that vein appears, to our knowledge, in [73]. However, no explicit rates are given in that work.

We here present the following general result on the effective convergence of (SPPA) under a stochastic regularity assumption:[9]

Theorem 5.1. *Let (E, E, μ) and $(\Omega, \mathsf{F}, \mathbb{P})$ be probability spaces, with (E, E, μ) complete, and let X be a separable Hadamard space. Let $f : E \times X \to (-\infty, +\infty]$ be a normal convex integrand such that $\underline{f}(x) := \int f(e, x)\, d\mu(e)$ is proper and $\operatorname{argmin}\underline{f} \neq \emptyset$. Write $F(x) := \underline{f}(x) - \min \underline{f}$. Let (x_n) be the iteration given by (SPPA), and assume (SPPA-A1) as well as (SPPA-A2). Lastly, let $\tau : (0, \infty) \to (0, \infty)$ be a modulus of regularity for F in mean w.r.t. D, i.e.*

$$\forall \varepsilon > 0\ \forall x \in D \left(\mathbb{E}[\underline{f}(x) - \min \underline{f}] < \tau(\varepsilon) \to \mathbb{E}[\operatorname{dist}^2_{\operatorname{argmin}\underline{f}}(x)] < \varepsilon \right),$$

where D is a collection of X-valued random variables with $(x_n) \subseteq D$. Then (x_n) a.s. strongly converges to an $\operatorname{argmin}\underline{f}$-valued random variable x. Moreover, the following rates of convergence apply: Let $z \in \operatorname{argmin}\underline{f}$ and let $b > d(x_0, z)$. Assume that $T > \sum_{n=0}^{\infty} \lambda_n^2$ and let $\theta : \mathbb{N} \times (0, \infty) \to \mathbb{N}$ as well as $\chi : (0, \infty) \to \mathbb{N}$ be such that $\sum_{n=\chi(\varepsilon)}^{\infty} \lambda_n^2 < \varepsilon$ for all $\varepsilon > 0$ and $\sum_{n=k}^{\theta(k,b)} \lambda_n \geq b$ for all $b > 0$ and $k \in \mathbb{N}$. Then $\mathbb{E}[d(x_n, x)] \to 0$ with rate $\rho(\varepsilon^2/4)$ and $d(x_n, x) \to 0$ a.s. with rate

[8]Note that, in similarity to [13], the above assumption misses absolute values which makes it still weaker than full Lipschitz-continuity assumptions known from the usual literature.

[9]See also [53] for a result with similar generality for the deterministic proximal point method in Hilbert spaces.

$\rho(\lambda\varepsilon^2/4)$, *where*

$$\rho(\varepsilon) := \theta\left(\chi\left(\frac{\varepsilon}{24\underline{L}}\right), \frac{b+4L^2T}{\tau(\varepsilon/6)}\right).$$

This result in particular covers the setup of [73] for strongly convex functions, at least in Hadamard spaces, and as such also that of [11], in particular for finding Fréchet means. Concretely, assume that $f(e,\cdot)$ is strongly convex with parameter $\alpha(e) > 0$, i.e.

$$f(e,(1-t)x \oplus ty) \leq (1-t)f(e,x) + tf(e,y) - t(1-t)\frac{\alpha(e)}{2}d^2(x,y)$$

for any $x, y \in X$ and any $t \in [0,1]$, where additionally $\underline{\alpha} := \int \alpha \, d\mu > 0$. Then $\underline{f}$ is strongly convex with parameter $\underline{\alpha}$ and so we obtain a modulus of regularity by setting $\tau(\varepsilon) := \frac{\underline{\alpha}}{8}\varepsilon^2$. However, the regularity assumption is not restricted to such assumptions, and covers in particular notions such as weak sharp minima or error bounds (recall Section 3.3). In those contexts, already the a.s. strong convergence of the iteration (without any quantitative information) outside of locally compact Hadamard spaces seems to be novel to the literature. Further, under linear regularity assumptions and suitable conditions on the parameters, our result on fast rates (recall Theorem 4.13) can be used to obtain linear non-asymptotic guarantees.

To prove this result, we have to establish the strong stochastic quasi-Fejér monotonicity of the sequence and derive a corresponding lim inf-bound. Both of these rest on the following fundamental property of the proximal map:

Lemma 5.2 (see e.g. Lemma 2.2.23 in [12])**.** *For any $\lambda > 0$, $x, y \in X$ and $e \in E$:*

$$f(e, \operatorname{prox}^f_\lambda(e,x)) - f(e,y) \leq \frac{1}{2\lambda}d^2(x,y) - \frac{1}{2\lambda}d^2(\operatorname{prox}^f_\lambda(e,x), y).$$

The strong stochastic quasi-Fejér monotonicity then follows rather immediately. For that, we in the following set $\mathsf{F}_n := \sigma(\xi_1, \dots, \xi_n)$ and we abbreviate $\mathbb{E}[\cdot \mid \mathsf{F}_n]$ by $\mathbb{E}_n$. Further, we write $\underline{L} := \int L^2 \, d\mu < \infty$.

Lemma 5.3 (extending [13])**.** *Let $n \in \mathbb{N}$. Then for any X-valued F_n-measurable random variable y:*

$$\mathbb{E}_n[d^2(x_{n+1}, y)] \leq d^2(x_n, y) - 2\lambda_n(\underline{f}(x_n) - \underline{f}(y)) + 4\lambda_n^2\underline{L} \text{ a.s.}$$

In particular, if y is additionally such that $y \in \operatorname{argmin}\underline{f}$ a.s., then

$$\mathbb{E}_n[d^2(x_{n+1}, y)] \leq d^2(x_n, y) - 2\lambda_n(\underline{f}(x_n) - \min \underline{f}) + 4\lambda_n^2\underline{L} \text{ a.s.}$$

Proof. Given $n \in \mathbb{N}$ and $y \in X$, Lemma 5.2 implies

$$d^2(x_{n+1}, y) \leq d^2(x_n, y) - 2\lambda_n[f(\xi_{n+1}, x_{n+1}) - f(\xi_{n+1}, y)]$$

and so we immediately have

$$\begin{aligned}
\mathbb{E}_n[d^2(x_{n+1}, y)] &\leq d^2(x_n, y) - 2\lambda_n\mathbb{E}_n[f(\xi_{n+1}, x_{n+1}) - f(\xi_{n+1}, y)] \\
&= d^2(x_n, y) - 2\lambda_n\mathbb{E}_n[f(\xi_{n+1}, x_n) - f(\xi_{n+1}, y)] \\
&\quad + 2\lambda_n\mathbb{E}_n[f(\xi_{n+1}, x_n) - f(\xi_{n+1}, x_{n+1})] \\
&= d^2(x_n, y) - 2\lambda_n[\underline{f}(x_n) - \underline{f}(y)] \\
&\quad + 2\lambda_n\mathbb{E}_n[f(\xi_{n+1}, x_n) - f(\xi_{n+1}, x_{n+1})]
\end{aligned}$$

where the third equality follows by independence of ξ_{n+1} to x_n and y (using that y is F_n-measurable), as well as the fact that ξ_{n+1} has distribution μ. Now, note that Lemma 5.2

together with (SPPA-A2) yields

$$\begin{aligned} d^2(x_{n+1}, x_n) &\leq 2\lambda_n[f(\xi_{n+1}, x_n) - f(\xi_{n+1}, x_{n+1})] \\ &\leq 2\lambda_n L(\xi_{n+1}) d(x_n, x_{n+1}) \end{aligned}$$

so that we have $d(x_{n+1}, x_n) \leq 2\lambda_n L(\xi_{n+1})$. Using (SPPA-A2), we further have

$$f(\xi_{n+1}, x_n) - f(\xi_{n+1}, x_{n+1}) \leq L(\xi_{n+1}) d(x_n, x_{n+1}) \leq 2\lambda_n L^2(\xi_{n+1}).$$

In particular, we have $2\lambda_n \mathbb{E}_n[f(\xi_{n+1}, x_n) - f(\xi_{n+1}, x_{n+1})] \leq 4\lambda_n^2 \underline{L}$, again using the independence of ξ_{n+1} to F_n. Combined, we get

$$\mathbb{E}_n[d^2(x_{n+1}, y)] \leq d^2(x_n, y) - 2\lambda_n(\underline{f}(x_n) - \underline{f}(y)) + 4\lambda_n^2 \underline{L}$$

as claimed. □

We can now use that result to derive a lim inf-bound. For that require two preliminary results which we shall use later on again. The first is a quantitative version of a lemma of Qihou [80] (see also Lemma 5.31 in [8]):

Lemma 5.4 (Theorem 3.2 in [71])**.** *Let (x_n), (α_n), (β_n) and (γ_n) be sequences of nonnegative reals with*

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

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 next result is folklore, but we include the very brief proof for completeness:

Lemma 5.5. *Suppose that (u_n), (v_n) are sequences of nonnegative reals with $L > 0$ such that $\sum_{n=0}^{\infty} u_n v_n < L$ and $\theta : \mathbb{N} \times (0, \infty) \to \mathbb{N}$ such that $\sum_{n=k}^{\theta(k,b)} u_n \geq b$ for all $b > 0$ and $k \in \mathbb{N}$. 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. For arbitrary $\varepsilon > 0$ and $N \in \mathbb{N}$, suppose for a contradiction that $v_n \geq \varepsilon$ for all $n \in [N; \theta(N, L/\varepsilon)]$. Then $L \leq \varepsilon \sum_{n=N}^{\theta(N,L/\varepsilon)} u_n \leq \sum_{n=N}^{\theta(N,L/\varepsilon)} u_n v_n \leq \sum_{n=0}^{\infty} u_n v_n < L$, which is a contradiction. □

Lemma 5.6. *Let $z \in \operatorname{argmin} \underline{f}$ and let $b > d(x_0, z)$. Let $\theta : \mathbb{N} \times (0, \infty) \to \mathbb{N}$ be such that $\sum_{n=k}^{\theta(k,b)} \lambda_n \geq b$ for all $b > 0$ and $k \in \mathbb{N}$, and let $T > \sum_{n=0}^{\infty} \lambda_n^2$. Then we have*

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

where $\varphi(\varepsilon, N) := \theta(N, (b + 4L^2T)/\varepsilon)$.

Proof. Taking expectations in Lemma 5.3, applied to z, and applying Lemma 5.4 yields the bound $\sum_{n=0}^{\infty} \lambda_n \mathbb{E}[\underline{f}(x_n) - \min \underline{f}] < b + 4L^2T$ and so Lemma 5.5 yields the claim. □

Proof of Theorem 5.1. Note that $\chi(\varepsilon/4\underline{L})$ is a rate of convergence for $\sum_{n\in\mathbb{N}} 4\lambda_n^2 \underline{L} < \infty$ as we have $\sum_{n=\chi(\varepsilon/4\underline{L})}^{\infty} \lambda_n^2 < \frac{\varepsilon}{4\underline{L}}$ and so $\sum_{n=\chi(\varepsilon/4\underline{L})}^{\infty} 4\lambda_n^2 \underline{L} < \varepsilon$ for all $\varepsilon > 0$. The result now immediately follows from Theorem 4.8, using in particular Lemmas 5.3 and 5.6, and using that d^2 is uniformly consistent with modulus $\theta(\varepsilon) := \varepsilon^2$ (recall Example 2.4). □

5.2. **Krasnoselkii-Mann schemes for common fixed point problems.** The second method we study is a randomized Krasnoselskii-Mann scheme for solving stochastic common fixed point problems. Going back to [54, 62] in a deterministic context, the Krasnoselskii-Mann scheme is one of the most central fixed point iterations in modern optimization.

Various stochastic versions of that scheme, tailored to different problem settings, have been proposed, notably (relatively straightforward) modifications by stochastic noise (see e.g. [36, 37]). We here focus on a variant of the Krasnoselskii-Mann scheme featuring a randomized selection from a class of operators as e.g. recently investigated (in a broad context) over Hilbert spaces by Combettes and Madariaga [35]. The variant we study here has been previously introduced over Hadamard spaces in work of the authors together with Neri [70], where in particular almost-sure convergence of that method over proper Hadamard spaces is established (see Theorem 5.12 therein), though this is based on a compactness argument rather than a regularity assumption, leading to a very different proof strategy which does not allow us to produce such effective convergence rates as in the present paper.

Let $(\Omega, \mathsf{F}, \mathbb{P})$ be an arbitrary probability space, (K, K) some other measurable space, and X be a separable Hadamard space. Let $(T_k)_{k\in K}$ be a family of mappings on X, and furthermore assume that

$$\text{(SKM-A1)} \qquad \begin{cases} \text{each } T_k \text{ is nonexpansive and the mapping } K \times X \to X \\ \text{defined by } (k,x) \mapsto T_k x \text{ is } \mathsf{K} \otimes \mathsf{B}(X)/\mathsf{B}(X)\text{-measurable.} \end{cases}$$

Let $k : \Omega \to K$ be a K-valued random variable. Our problem is to

$$\text{find some element of } \mathrm{Fix}T := \{x \in X \mid T_k x = x \ \mathbb{P}\text{-a.s.}\},$$

assuming that $\mathrm{Fix}T \neq \emptyset$. We capture this problem in our general setup via $F(x) := \mathbb{E}[d^2(T_k x, x)]$ (recall Section 3.3). Indeed, note that the map $(\omega, x) \mapsto d^2(T_{k(\omega)}x, x)$ is a Carathéodory function since d^2 is continuous in both arguments and $T_{k(\omega)}$ is nonexpansive. Hence, it is in particular $\mathsf{F} \otimes \mathsf{B}(X)$-measurable, and so F is measurable, again by Fubini's theorem. Further, note that $\mathrm{Fix}T = \mathrm{zer}F$ is closed (see e.g. Lemma 5.2 in [70]).

To solve the above problem, we now consider a randomised Krasnoselkii-Mann scheme

$$\text{(SKM)} \qquad x_{n+1} := (1-\lambda_n)x_n \oplus \lambda_n T_{k_n} x_n,$$

given some starting point $x_0 \in X$, sequences (λ_n) of random variables $\Omega \to (0,1]$ and (k_n) of random variables $\Omega \to X$, such that

$$\text{(SKM-A2)} \qquad \begin{cases} (k_n) \text{ are i.i.d. and distributed as } k \text{ and } (\lambda_n) \text{ are} \\ \text{independent of } (k_n) \text{ with } \sum_{n\in\mathbb{N}} \mathbb{E}[\lambda_n(1-\lambda_n)] = \infty. \end{cases}$$

Analogous to the previous section, we now present a general result on the effective convergence of (SKM) under a stochastic regularity assumption.

Theorem 5.7. *Let $(\Omega, \mathsf{F}, \mathbb{P})$ be a probability space, (K, K) a measurable space, and X a separable Hadamard space. Let $(T_k)_{k\in K}$ be a family of mappings satisfying* (SKM-A1)*, and let $k : \Omega \to K$ be a K-valued random variable. Write $F(x) := \mathbb{E}[d^2(T_k x, x)]$. Let (x_n) be the iteration given by* (SKM) *and assume additionally* (SKM-A2)*. Lastly, let $\tau : (0,\infty) \to (0,\infty)$ be a modulus of regularity for F in mean w.r.t. D, i.e.*

$$\forall \varepsilon > 0 \ \forall x \in D \left(\mathbb{E}_{\omega\sim\mathbb{P}}[\mathbb{E}_{\omega'\sim\mathbb{P}}[d^2(T_{k(\omega')}x(\omega), x(\omega))]] < \tau(\varepsilon) \to \mathbb{E}[\mathrm{dist}^2_{\mathrm{Fix}T}(x)] < \varepsilon \right),$$

where D is a collection of X-valued random variables with $(x_n) \subseteq D$.

Then (x_n) a.s. strongly converges to a $\mathrm{Fix}T$-valued random variable x. Moreover, the following rates of convergence apply: Let $z \in \mathrm{Fix}T$ with $b > d(x_0, z)$, and assume that $\theta : \mathbb{N}\times(0,\infty) \to$

$\mathbb{N}$ *is such that* $\sum_{n=k}^{\theta(k,b)} \mathbb{E}[\lambda_n(1-\lambda_n)] \geq b$ *for all* $b > 0$ *and* $k \in \mathbb{N}$. *Then* $\mathbb{E}[d(x_n, x)] \to 0$ *with rate* $\rho(\varepsilon^2/4)$ *and* $d(x_n, x) \to 0$ *a.s. with rate* $\rho(\lambda\varepsilon^2/4)$ *where*

$$\rho(\varepsilon) := \theta\left(0, \frac{b}{\tau(\varepsilon/6)}\right).$$

When (T_k) is linearly regular (recall Section 3.3), and the step sizes satisfy a growth condition like $\frac{v}{n+r} \leq \mathbb{E}[\lambda_n(1-\lambda_n)]$, our result on fast rates (recall Theorem 4.13) in particular apply, which yield linear non-asymptotic guarantees of the form $\mathbb{E}[\mathrm{dist}^2_{\mathrm{zer}F}(x_n)] \leq u/(n+r)$ for a suitable constant $u > 0$, and similarly in the almost sure case.

The result is proven using a similar strategy to that of the previous subsection, establishing first the strong stochastic quasi-Fejér monotonicity, and then a suitable lim inf-bound. These both echo standard calculations associated with Krasnoselkii-Mann schemes, generalised to Hadamard spaces and the stochastic mappings (T_k). Define the filtration $\mathsf{F}_n := \sigma(x_0, \ldots, x_n, k_0, \ldots, k_{n-1}, \lambda_0, \ldots, \lambda_{n-1})$ and write $\mathbb{E}_n$ for the conditional expectation $\mathbb{E}[\cdot \mid \mathsf{F}_n]$.

Lemma 5.8 (extending [70])**.** *Let* $n \in \mathbb{N}$. *Then for any* X*-valued* F_n*-measurable random variable* y*:*

$$\begin{aligned}\mathbb{E}_n[d^2(x_{n+1}, y)] \leq & d^2(x_n, y) - \mathbb{E}[\lambda_n(1-\lambda_n)]\mathbb{E}_n[d^2(T_{k_n}x_n, x_n)] \\ & + \mathbb{E}_n[d^2(T_{k_n}y, y)] + d(x_n, y)\mathbb{E}_n[d(T_{k_n}y, y)] \ \text{a.s.}\end{aligned}$$

In particular, if y *is additionally such that* $y \in \mathrm{Fix}T$ *a.s., we have*

$$\mathbb{E}_n[d^2(x_{n+1}, y)] \leq d^2(x_n, y) - \mathbb{E}[\lambda_n(1-\lambda_n)]\mathbb{E}_n[d^2(T_{k_n}x_n, x_n)] \ \text{a.s.}$$

Proof. Given $n \in \mathbb{N}$ and $y \in X$, by the triangle inequality and the fact that T_k is nonexpansive, we have

$$d^2(T_{k_n}x_n, y) \leq d^2(x_n, y) + d^2(T_{k_n}y, y) + d(x_n, y)d(T_{k_n}y, y)$$

Using (CN) (see also Lemma 5.7 in [70]), we get

$$d^2(x_{n+1}, y) \leq (1-\lambda_n)d^2(x_n, y) + \lambda_n d^2(T_{k_n}x_n, y) - \lambda_n(1-\lambda_n)d^2(T_{k_n}x_n, x_n).$$

Combined, we obtain

$$\begin{aligned}d^2(x_{n+1}, y) &\leq d^2(x_n, y) + \lambda_n d^2(T_{k_n}y, y) + \lambda_n d(x_n, y)d(T_{k_n}y, y) - \lambda_n(1-\lambda_n)d^2(T_{k_n}x_n, x_n) \\ &\leq d^2(x_n, y) + d^2(T_{k_n}y, y) + d(x_n, y)d(T_{k_n}y, y) - \lambda_n(1-\lambda_n)d^2(T_{k_n}x_n, x_n).\end{aligned}$$

If y is now a F_n-measurable random variable, using basic properties of the conditional expectation, and noting that $\lambda_n(1-\lambda_n)$ is independent to both $d^2(T_{k_n}x_n, x_n)$ and F_n, we have

$$\begin{aligned}\mathbb{E}_n[d^2(x_{n+1}, y)] \leq & d^2(x_n, y) - \mathbb{E}[\lambda_n(1-\lambda_n)]\mathbb{E}_n[d^2(T_{k_n}x_n, x_n)] \\ & + \mathbb{E}_n[d^2(T_{k_n}y, y)] + d(x_n, y)\mathbb{E}_n[d(T_{k_n}y, y)]\end{aligned}$$

which was the first claim. If now y also satisfies $y \in \mathrm{Fix}T$ a.s., since k_n is independent of F_n with the same distribution as k, we have

$$\mathbb{E}_n[d^2(T_{k_n}y, y)](\omega) = \mathbb{E}_{\omega' \sim \mathbb{P}}[d^2(T_{k(\omega')}y(\omega), y(\omega))] = 0$$

on a set of measure one. Jensen's inequality yields $\mathbb{E}_n[d(T_{k_n}y, y)] = 0$ a.s., from which the second claim follows. □

Lemma 5.9. *Let* $z \in \mathrm{Fix}T$ *and let* $b > d(x_0, z)$. *Let* $\theta : \mathbb{N} \times (0, \infty) \to \mathbb{N}$ *be such that* $\sum_{n=k}^{\theta(k,b)} \mathbb{E}[\lambda_n(1-\lambda_n)] \geq b$ *for all* $b > 0$ *and* $k \in \mathbb{N}$. *Then we have*

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

where $\varphi(\varepsilon, N) := \theta(N, b/\varepsilon)$.

Proof. Taking expectations in Lemma 5.8, applied to z, and applying Lemma 5.4 yields the bound $\sum_{n=0}^{\infty} \mathbb{E}[\lambda_n(1-\lambda_n)]\mathbb{E}[d^2(T_{k_n}x_n, x_n)] < b$. The result then follows from Lemma 5.5 together with the observation that

$$\mathbb{E}[d^2(T_{k_n}x_n, x_n)] = \mathbb{E}_{\omega\sim\mathbb{P}}[\mathbb{E}_{\omega'\sim\mathbb{P}}[d^2(T_{k(\omega')}x_n(\omega), x_n(\omega))]] = \mathbb{E}[F(x_n)]$$

by Fubini's theorem and the fact that k_n is independent of x_n and is distributed as k. □

Proof of Theorem 5.7. The result is now immediate from Theorem 4.8, using Lemmas 5.8 and 5.9, and again that d^2 is uniformly consistent with modulus $\theta(\varepsilon) := \varepsilon^2$ (recall Example 2.4). □

5.3. **Stochastic Busemann subgradient methods.** With our last method we return to the problem of minimizing the mean of a stochastic convex function, where we now focus on a projected subgradient method recently introduced by Goodwin, Lewis, López-Acedo and Nicolae [45] which employs the novel notion of Busemann subgradients, and in fact focus on an extension of that method recently studied in [77] for general stochastic minimization. This novel type of subgradient coincides with the usual notion of subgradients over Euclidean spaces but, by exploiting the boundary cone CX^∞ of a Hadamard space X, associated Busemann functions, and other advanced geometric tools from Hadamard spaces, supports a theory that is particularly suitable for nonlinear geometric contexts.

To introduce this method, we require a bit more background on Hadamard spaces X: A geodesic ray in X is an isometry $r : [0,\infty) \to X$, said to be issuing from $r(0)$. X has the geodesic extension property if for all $x \neq y \in X$, there is a ray $r : [0,\infty) \to X$ issuing from x such that $r(t) = y$ for some $t > 0$.

The so-called boundary of X at infinity X^∞ is the set of all equivalence classes of rays in X under the equivalence relation of being asymptotic (see Definition 8.1 in [25]). The boundary cone CX^∞ is now the usual Euclidean cone over X^∞, i.e. CX^∞ is the quotient of $X^\infty \times [0,\infty)$ under the equivalence relation $(\xi, s) \sim (\xi', s')$ if, and only if, $s = s' = 0$ or $(\xi, s) = (\xi', s')$. We denote an equivalence class of $(\xi, s) \in X^\infty \times [0,\infty)$ in CX^∞ by $[\xi, s]$, and write $[0]$ for the equivalence class of $(\xi, 0)$ for some/any $\xi \in X^\infty$. We refer to Chapter II.8 in [25] for further information on the boundary X^∞ and the boundary cone CX^∞.[10]

Following [45], we define the "pairing" function $\langle\cdot,\cdot\rangle : X \times CX^\infty \to \mathbb{R}$ by

$$\langle x, [\xi, s]\rangle := \begin{cases} s b_\xi(x) & \text{if } s > 0, \\ 0 & \text{if } s = 0, \end{cases}$$

where we wrote

$$b_\xi(x) := \lim_{t\to\infty} (d(x, r_{\overline{x},\xi}(t)) - t)$$

for the Busemann function $X \to \mathbb{R}$ (see e.g. Example 2.2.10 in [12]) corresponding to the (unique) ray $r_{\overline{x},\xi}$ with direction ξ and some arbitrary but fixed origin $\overline{x}$. Given a set $C \subseteq X$ and a function $g : C \to \mathbb{R}$, a Busemann subgradient of g at $x \in C$ is then a point $[\xi, s] \in CX^\infty$ such that

$$x = \operatorname{argmin}_{y\in C} \{g(y) - \langle y, [\xi, s]\rangle\}.$$

g is called Busemann subdifferentiable if g has a Busemann subgradient at every $x \in C$.

We are now in the position to introduce the related projected Busemann subgradient method. For that, let $(\Omega, \mathsf{F}, \mathbb{P})$ and (E, E, μ) be probability spaces, with (E, E, μ) complete, and let X

[10] Topologically, the spaces X^∞ and CX^∞ actually require a more subtle treatment via the so-called cone topology (see Definition II.8.6 in [25] and the discussion in [45]), but this will not explicitly feature in the following arguments and we hence omit it.

be a separable Hadamard space with the geodesic extension property and at least two points.[11] Further, fix a closed convex non-empty subset $C \subseteq X$ and let $f : E \times C \to \mathbb{R}$ be a given functional. As before, setting $\underline{f}(x) := \int f(e,x)\, d\mu(e)$, we want to

$$\text{find some element of } \operatorname{argmin} \underline{f},$$

where we again assume that $\underline{f}$ is proper and that $\operatorname{argmin}\underline{f} \neq \emptyset$, with $\operatorname{argmin}\underline{f}$ now understood to be a subset of C. We again set $F(x) := \underline{f}(x) - \min \underline{f}$.

In terms of assumption on f, we have to make three distinct types of assumptions, all relating to the Busemann subgradient structure of f. Concretely, in terms of basic regularity of the problem, we assume

(SB-A1) $\begin{cases} f(e,\cdot) \text{ is Busemann subdifferentiable for any } e \in E \\ \text{and } f(\cdot,x) \text{ is measurable for all } x \in X, \end{cases}$

and further, we make the Lipschitz-type assumption that

(SB-A2) $\begin{cases} \text{there exists a constant } L > 0 \text{ such that for any } e \in E \text{ and any} \\ \text{Busemann subgradient } [\xi, s] \in CX^\infty \text{ of } f(e,\cdot) \text{ at } x \in C, \text{ we have } s \leq L. \end{cases}$

Lastly, we want to be able to draw subgradients in measurable way, and to that end assume that there exists an oracle function $\mathsf{Busemann}_f$, where $[\xi, s] = \mathsf{Busemann}_f(e,x)$ represents a Busemann subgradient of $f(e,\cdot)$ at x, such that

(SB-A3) $\begin{cases} \text{whenever } x : \Omega \to C \text{ and } \zeta : \Omega \to E \text{ are measurable functions,} \\ \text{then } [\xi, s] = \mathsf{Busemann}_f(\zeta, x) \text{ is measurable as a function } \Omega \to CX^\infty. \end{cases}$

(SB-A1) and (SB-A2) in particular imply that $f(e,\cdot)$ is convex and L-Lipschitz on C (see [45]). Hence, they guarantee that F is measurable and that $\operatorname{argmin}\underline{f} = \operatorname{zer} F$ is closed, similar to before. (SB-A3) does not explicitly feature in [45], but is crucial to guarantee the measurability of the associated subgradient method (see the discussion in [77]). Due to the different topologies on X^∞ and CX^∞, it is not quite trivial when this assumption can be satisfied, but it can at least be guaranteed in proper Hadamard spaces (see Proposition 5.11 in [77]). We will however not make any local compactness assumptions here, and hence focus on the abstract measurability assumption (SB-A3).

Extending [45] as in [77], we now consider the iteration

(SB) $$x_{n+1} := P_C(r_{x_n,\xi_n}(s_n t_n))$$

where $[\xi_n, s_n] = \mathsf{Busemann}_f(\zeta_{n+1}, x_n)$, given a starting point $x_0 \in X$ and sequences (t_n) of positive reals as well as (ζ_{n+1}) of random variables $\Omega \to E$, for which we assume that

(SB-A4) $$(\zeta_{n+1}) \text{ are i.i.d. with distribution } \mu \text{ and } \sum_{n\in\mathbb{N}} t_n = \infty, \sum_{n\in\mathbb{N}} t_n^2 < \infty.$$

The conditions (SB-A1) – (SB-A4) together imply that x_n is measurable for any $n \in \mathbb{N}$ (see Lemma 5.5 in [77]).

We here present the following general result on the effective convergence of (SB) under a stochastic regularity assumption:

Theorem 5.10. *Let (E, E, μ) and $(\Omega, \mathsf{F}, \mathbb{P})$ be probability spaces, with (E, E, μ) complete, let X be a separable Hadamard space with the geodesic extension property and at least two points, and fix a closed convex non-empty subset $C \subseteq X$. Let $f : E \times C \to \mathbb{R}$ be a function with properties*

[11] See [45] for a discussion of this assumption, which in particular entails that the boundary X^∞ is non-empty.

(SB-A1) – (SB-A3) *as above and assume additionally that* $\underline{f}(x) := \int f(e,x)\,d\mu(e)$ *is proper and* $\operatorname{argmin}\underline{f} \neq \emptyset$. *Write* $F(x) := \underline{f}(x) - \min \underline{f}$. *Let* (x_n) *be the iteration given by* (SB), *and assume* (SB-A4). *Lastly, let* $\tau : (0,\infty) \to (0,\infty)$ *be a modulus of regularity for* F *in mean w.r.t.* D, *i.e.*

$$\forall \varepsilon > 0\ \forall x \in D \left(\mathbb{E}[\underline{f}(x) - \min \underline{f}] < \tau(\varepsilon) \to \mathbb{E}[\operatorname{dist}^2_{\operatorname{argmin}\underline{f}}(x)] < \varepsilon \right),$$

where D *is a collection of* X*-valued random variables with* $(x_n) \subseteq D$. *Then* (x_n) *a.s. strongly converges to an* $\operatorname{argmin}\underline{f}$*-valued random variable* x. *Moreover, the following rates of convergence apply: Let* $z \in \operatorname{argmin}\underline{f}$ *and let* $b > d(x_0, z)$. *Further, let* $\chi : (0,\infty) \to \mathbb{N}$ *and* $\theta : \mathbb{N} \times (0,\infty) \to \mathbb{N}$ *be such that* $\sum_{n=\chi(\varepsilon)}^{\infty} t_n^2 < \varepsilon$ *for all* $\varepsilon > 0$ *and* $\sum_{n=k}^{\theta(k,b)} t_n \geq b$ *for all* $b > 0$ *and* $k \in \mathbb{N}$, *and let* $T > \sum_{n=0}^{\infty} t_n^2$. *Then* $\mathbb{E}[d(x_n, x)] \to 0$ *with rate* $\rho(\varepsilon^2/4)$ *and* $d(x_n, x) \to 0$ *a.s. with rate* $\rho(\lambda\varepsilon^2/4)$, *where*

$$\rho(\varepsilon) := \theta\left(\chi\left(\frac{\varepsilon}{6L^2} \right), \frac{b + L^2 T}{\tau\left(\varepsilon/6\right)} \right).$$

This theorem in particular includes the quantitative result given in [77] (see Theorem 1.4 therein) under the assumption that $f(e,\cdot)$ is strongly convex with parameter $\alpha(e) > 0$ such that $\underline{\alpha} := \int \alpha\,d\mu > 0$ (where we can instantiate the above with $\tau(\varepsilon) := \frac{\underline{\alpha}}{8}\varepsilon^2$ as before). As before, the regularity assumption is however not restricted to strong convexity (recall Section 3.3) and in such broader contexts, already the a.s. strong convergence of the iteration outside of locally compact Hadamard spaces seems to be novel to the literature. By moving to a finite measure space (E, E, μ), our result in particular also applies to the method studied in [45] (see [77] for a comparison). As before, using linear regularity assumptions and suitable conditions on the parameters we can apply our result on fast rates (recall Theorem 4.13) to obtain linear non-asymptotic guarantees.

The proof proceeds similar to both subsections before, and in particular relies on establishing the strong stochastic quasi-Fejér monotonicity of the iteration together with a lim inf-bound. For that, we introduce some notation, similar to before. Define the filtration $\mathsf{F}_n := \sigma(\zeta_1, \dots, \zeta_n, x_0, \dots, x_n)$ and write $\mathbb{E}_n$ for the conditional expectation $\mathbb{E}[\cdot \mid \mathsf{F}_n]$.

As a last preliminary result, we require the following property of the Busemann subgradients:

Lemma 5.11 (Lemma 6.1 in [45])**.** *Let* $g : C \to \mathbb{R}$ *be a given function for a non-empty closed and convex set* $C \subseteq X$ *and let* $[\xi, s]$ *be a Busemann subgradient of* f *at* $x \in C$. *Given* $t > 0$, *define*

$$x^+ := \begin{cases} P_C(r_{x,\xi}(st)) & \text{if } s > 0, \\ x & \text{if } s = 0. \end{cases}$$

Then for any $y \in C$: $d^2(x^+, y) \leq d^2(x,y) - 2t(f(x) - f(y)) + s^2t^2$.

Lemma 5.12 (extending [45] and [77])**.** *Let* $n \in \mathbb{N}$. *Then for any* C*-valued* F_n*-measurable random variable* y:

$$\mathbb{E}_n[d^2(x_{n+1}, y)] \leq d^2(x_n, y) - 2t_n(\underline{f}(x_n) - \underline{f}(y)) + L^2 t_n^2 \ \text{a.s.}$$

In particular, if y *is additionally such that* $y \in \operatorname{argmin}\underline{f}$ *a.s., then*

$$\mathbb{E}_n[d^2(x_{n+1}, y)] \leq d^2(x_n, y) - 2t_n(\underline{f}(x_n) - \min \underline{f}) + L^2 t_n^2 \ \text{a.s.}$$

Proof. Using Lemma 5.11 together with assumption (SB-A2), by which we have $s_n \leq L$, and applying conditional expectations yields

$$\mathbb{E}_n[d^2(x_{n+1}, y)] \leq d^2(x_n, y) - 2t_n(\mathbb{E}_n[f(\zeta_{n+1}, x_n)] - \mathbb{E}_n[f(\zeta_{n+1}, y)]) + L^2 t_n^2.$$

Now, as ζ_{n+1} is independent of x_n and F_n, we have

$$\mathbb{E}_n[f(\zeta_{n+1}, x_n)](\omega) = \mathbb{E}_{\omega' \sim \mathbb{P}}[f(\zeta_{n+1}(\omega'), x_n(\omega))] = \underline{f}(x_n(\omega))$$

and similarly $\mathbb{E}_n[f(\zeta_{n+1}, y)] = \underline{f}(y)$. This yields the claim. □

Lemma 5.13. *Let $z \in \operatorname{argmin}\underline{f}$ and let $b > d(x_0, z)$. Let $\theta : \mathbb{N} \times (0, \infty) \to \mathbb{N}$ be such that $\sum_{n=k}^{\theta(k,b)} t_n \geq b$ for all $b > 0$ and $k \in \mathbb{N}$, and let $T > \sum_{n=0}^{\infty} t_n^2$. Then we have*

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

where $\varphi(\varepsilon, N) := \theta(N, (b + L^2 T)/\varepsilon)$.

Proof. Taking expectations in Lemma 5.12, applied to z, and applying Lemma 5.4 yields the bound $\sum_{n=0}^{\infty} t_n \mathbb{E}[F(x_n) - \min F] < b + L^2 T$ and so Lemma 5.5 yields the claim. □

Proof of Theorem 5.10. Note that $\chi(\varepsilon/L^2)$ is a rate of convergence for $\sum_{n \in \mathbb{N}} L^2 t_n^2 < \infty$ as we have $\sum_{n=\chi(\varepsilon/L^2)}^{\infty} t_n^2 < \frac{\varepsilon}{L^2}$ and so $\sum_{n=\chi(\varepsilon/L^2)}^{\infty} L^2 \lambda_n^2 < \varepsilon$ for all $\varepsilon > 0$. As before, the result then follows from Theorem 4.8 using Lemmas 5.12 and 5.13 and that d^2 is uniformly consistent with modulus $\theta(\varepsilon) := \varepsilon^2$ (recall Example 2.4). □

Funding. The second 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] S. Alexander, V. Kapovitch, and A. Petrunin. *Alexandrov Geometry: Foundations*, volume 236 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2024.

[3] C.D. Aliprantis and K.C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Berlin, Heidelberg, 2006.

[4] H. Asi, K. Chadha, G. Cheng, and J.C. Duchi. Minibatch stochastic approximate proximal point methods. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020)*, pages 21958–21968. Curran Associates Inc., USA, 2020.

[5] H. Asi and J.C. Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

[6] J.-P. Aubin and H. Frankowska. *Set-Valued Analysis*. Springer, New York, 2009.

[7] R.J. Aumann. Measurable utility and the measurable choice problem. In G.T. Guilbaud, editor, *La Décision*, pages 15–26. Editions du Centre National de la Recherche Scientifique, Paris, 1969.

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

[9] H.H. Bauschke and A.S. Lewis. Dykstras algorithm with Bregman projections: A convergence proof. *Optimization*, 48(4):409–427, 2000.

[10] M. Bačák. The proximal point algorithm in metric spaces. *Israel Journal of Mathematics*, 194(2):689–701, 2013.

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

[12] M. Bačák. *Convex analysis and optimization in Hadamard spaces*, volume 22 of *De Gruyter Series in Nonlinear Analysis and Applications*. Walter de Gruyter GmbH, Berlin/Boston, 2014.

[13] M. Bačák. A variational approach to stochastic minimization of convex functionals. *Pure and Applied Functional Analysis*, 3(2):287–295, 2018.

[14] M. Bačák. Old and new challenges in Hadamard spaces. *Japanese Journal of Mathematics*, 18(2):115–168, 2023.

[15] D.P. Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming. Series B*, 129:163–195, 2011.

[16] D.P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. In S. Sra, S. Nowozin, and S.J. Wright, editors, *Optimization for Machine Learning*, Neural Information Processing Series, pages 85–120. The MIT Press, Cambridge, Massachusetts, 2012.

[17] P. Bianchi. Ergodic convergence of a stochastic proximal point algorithm. *SIAM Journal on Optimization*, 26(4):2235–2260, 2016.
[18] L.J. Billera, S.P. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767, 2001.
[19] R.I. Boţ and C. Schindler. On a stochastic differential equation with correction term governed by a monotone and Lipschitz continuous operator. *Evolution Equations and Control Theory*, 14(3):463–493, 2025.
[20] R.I. Boţ and C. Schindler. Long-Time Analysis of Stochastic Heavy Ball Dynamics for Convex Optimization and Monotone Equations. *Discrete and Continuous Dynamical Systems*, 51:439–474, 2026.
[21] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.
[22] 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*, 165:471–507, 2017.
[23] J.M. Borwein, G. Li, and M.K. Tam. Convergence rate analysis for averaged fixed point iterations in common fixed point problems. *SIAM Journal on Optimization*, 27:1–33, 2017.
[24] H. Brézis and P.L. Lions. Produits infinis de resolvantes. *Israel Journal of Mathematics*, 29(4):329–345, 1978.
[25] M.R. Bridson and A. Haefliger. *Metric Spaces of Non-Positive Curvature*, volume 319 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin, Heidelberg, 1999.
[26] 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.
[27] J.V. Burke and S. Deng. Weak sharp minima revisited. I. Basic theory. *Control and Cybernetics*, 31:439–469, 2002.
[28] J.V. Burke and M.C. Ferris. Weak sharp minima in mathematical programming. *SIAM Journal on Control and Optimization*, 31:1340–1359, 1993.
[29] D. Butnariu and A.N. Iusem. *Totally Convex Functions for Fixed Points Computation and Infinite Dimensional Optimization*, volume 40 of *Applied Optimization*. Springer Dordrecht, 2000.
[30] D. Butnariu and E. Resmerita. Bregman distances, totally convex functions and a method for solving operator equations in Banach spaces. *Abstract and Applied Analysis*, 2006. 84919, 39pp.
[31] C. Castaing and M. Valadier. *Convex Analysis and Measurable Multifunctions*, volume 580 of *Lecture Notes in Mathematics*. Springer Berlin, Heidelberg, 1977.
[32] P. Chaipunya, F. Kohsaka, and P. Kumam. Monotone vector fields and generation of nonexpansive semigroups in complete CAT(0) spaces. *Numerical Functional Analysis and Optimization*, 42(9):989–1018, 2021.
[33] P.L. Combettes. Quasi-Fejérian Analysis of Some Optimization Algorithms. In D. Butnariu, Y. Censor, and S. Reich, editors, *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*, volume 8 of *Studies in Computational Mathematics*, pages 115–152. North-Holland, Amsterdam, 2001.
[34] P.L. Combettes. Fejér monotonicity in convex optimization. In C.A. Floudas and P.M. Pardalos, editors, *Encyclopedia of Optimization*, pages 1016–1024. Springer, Boston, MA, 2009.
[35] P.L. Combettes and J.I. Madariaga. A geometric framework for stochastic iterations. *Mathematics of Computation*, 2026. To appear.
[36] 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.
[37] 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*, 174(1):433–451, 2019.
[38] A.L. Dontchev and R.T. Rockafellar. *Implicit functions and solution mappings. A view from variational analysis*. Springer Monographs in Mathematics. Springer, Dordrecht, 2009.
[39] D. Drusvyatskiy and A.S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.
[40] Y.M. Ermol'ev. On the method of generalized stochastic gradients and quasi-Fejér sequences. *Cybernetics*, 5:208–220, 1969.
[41] Y.M. Ermol'ev. On convergence of random quasi-Fejér sequences. *Cybernetics*, 7:655–656, 1971.
[42] 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.
[43] M.C. Ferris. Finite termination of the proximal point algorithm. *Mathematical Programming, Series A*, 50:359–366, 1991.

[44] A. Freund and N. Pischke. Effective rates for continuous-time quasi-Fejér monotone dynamical systems, 2026. Preprint, available at `https://arxiv.org/abs/2603.23708`.
[45] A. Goodwin, A.S. Lewis, G. López-Acedo, and A. Nicolae. Stochastic and incremental subgradient methods for convex optimization on Hadamard spaces. *Mathematical Programming*, 2026. To appear.
[46] M. Grasmair. Generalized Bregman distances and convergence rates for non-convex regularization methods. *Inverse Problems*, 26(11), 2010. 115014, 16pp.
[47] 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.
[48] O. Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM Journal on Control and Optimization*, 29:403–419, 1991.
[49] N. Hermer, D.R. Luke, and A. Sturm. Random function iterations for consistent stochastic feasibility. *Numerical Functional Analysis and Optimization*, 40(4):386–420, 2019.
[50] J. Jost. Convex functionals and generalized harmonic maps into spaces of nonpositive curvature. *Commentarii Mathematici Helvetici*, 70:659–673, 1995.
[51] U. Kohlenbach. *Applied Proof Theory: Proof Interpretations and their Use in Mathematics*. Springer Monographs in Mathematics. Springer-Verlag Berlin Heidelberg, 2008.
[52] U. Kohlenbach. Proof-theoretic Methods in Nonlinear Analysis. In B. Sirakov, P. Ney de Souza, and M. Viana, editors, *Proceedings of ICM 2018*, volume 2, pages 61–82. World Scientific, Singapure, 2019.
[53] 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.
[54] M.A. Krasnoselskii. Two remarks on the method of successive approximations. *Uspekhi Matematicheskikh Nauk*, 10(1(63)):123–127, 1955.
[55] L. Leuştean and A. Sipoş. Effective strong convergence of the proximal point algorithm in CAT(0) spaces. *Journal of Nonlinear and Variational Analysis*, 2(2):219–228, 2018.
[56] D. Leventhal. Metric subregularity and the proximal point method. *Journal of Mathematical Analysis and Applications*, 360:681–688, 2009.
[57] C. Li, G. López, and V. Martín-Márquez. Monotone vector fields and the proximal point algorithm on Hadamard manifolds. *Journal of the London Mathematical Society*, 79(3):663–683, 2009.
[58] 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.
[59] G. Li, B. Mordukhovich, and J. Zhu. Generalized metric subregularity with applications to high-order regularized Newton methods. *Mathematics of Operations Research*, 2026. To appear.
[60] J. Liu, X.J. Long, X.S. Li, and N.J. Huang. Stochastic dual dynamical systems for linear equality constrained convex optimization problems. *Communications in Nonlinear Science and Numerical Simulation*, 154, 2026. 109538, 15pp.
[61] D.R. Luke, J.C. Schnebel, M. Staudigl, J. Peypouquet, and S. Qu. Asymptotic behaviour of coupled random dynamical systems with multiscale aspects, 2026. Preprint, available at `https://arxiv.org/abs/2601.15411`.
[62] W.R. Mann. Mean value methods in iteration. *Proceedings of the American Mathematical Society*, 4:506–510, 1953.
[63] B. Martinet. Régularisation dinéquations variationnelles par approximations successives. *Revue française d'informatique et de recherche opérationnelle*, 4:154–159, 1970.
[64] R. Maulen-Soto, J. Fadili, and H. Attouch. An stochastic differential equation perspective on stochastic convex optimization. *Mathematics of Operations Research*, 50(4):3190–3221, 2025.
[65] R. Maulen-Soto, J. Fadili, H. Attouch, and P. Ochs. Stochastic inertial dynamics via time scaling and averaging. *Stochastic Systems*, 16(1):61–89, 2026.
[66] U. Mayer. Gradient flows on nonpositively curved metric spaces and harmonic maps. *Communications in Analysis and Geometry*, 6:199–253, 1998.
[67] V. Moulton and A. Spillner. Spaces of ranked tree-child networks. *Journal of Mathematical Biology*, 91(3), 2025. 32, 26pp.
[68] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal of Optimization*, 19(4):1574–1609, 2009.
[69] M. Neri, N. Pischke, and T. Powell. An abstract effective convergence theorem for stochastic processes, with applications to stochastic approximation, 2026. Preprint, available at `https://arxiv.org/abs/2504.12922`.

[70] M. Neri, N. Pischke, and T. Powell. Generalized fluctuation bounds for stochastic algorithms in the presence of compactness, 2026. Preprint, available at `https://arxiv.org/abs/2602.22741`.

[71] M. Neri and T. Powell. A quantitative Robbins-Siegmund theorem. *The Annals of Applied Probability*, 36(1):636–651, 2026.

[72] E. Neumann. Computational problems in metric fixed point theory and their Weihrauch degrees. *Logical Methods in Computer Science*, 11(4:20), 2015. 44pp.

[73] S.I. Ohta and M. Pálfia. Discrete-time gradient flows and law of large numbers in Alexandrov spaces. *Calculus of Variations and Partial Differential Equations*, 54(2):1591–1610, 2015.

[74] P. Pinto and N. Pischke. On Dykstra's algorithm with Bregman projections, 2026. Preprint, available at `https://nicholaspischke.github.io`.

[75] N. Pischke. Generalized Fejér monotone sequences and their finitary content. *Optimization*, 74(14):3771–3838, 2025.

[76] N. Pischke. Mean-square and linear convergence of a stochastic proximal point algorithm in metric spaces of nonpositive curvature, 2025. Preprint, available at `https://arxiv.org/abs/2510.10697`.

[77] N. Pischke. On Busemann subgradient methods for stochastic minimization in Hadamard spaces, 2026. Preprint, available at `https://arxiv.org/abs/2602.08127`.

[78] N. Pischke and U. Kohlenbach. Effective rates for iterations involving Bregman strongly nonexpansive operators. *Set-Valued and Variational Analysis*, 32(4), 2024. 33, 58pp.

[79] N. Pischke and T. Powell. Asymptotic regularity of a generalised stochastic Halpern scheme, 2024. Preprint, available at `https://arxiv.org/abs/2411.04845`.

[80] L. Qihou. Iteration sequences for asymptotically quasi-nonexpansive mappings with error member. *Journal of Mathematical Analysis and Applications*, 259:18–24, 2001.

[81] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In J.S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257. Academic Press, New York, 1971.

[82] R.T. Rockafellar. Convex integral functionals and duality. In E.H. Zarantonello, editor, *Contributions to Nonlinear Functional Analysis*, pages 215–236. Academic Press, New York, 1971.

[83] R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal of Control and Optimization*, 14:877–898, 1976.

[84] W. Römisch. On the Convergence of Measurable Selections and an Application to Approximations in Stochastic Optimization. *Zeitschrift für Analysis und ihre Anwendungen*, 5(3):277–288, 1986.

[85] E.K. Ryu and S. Boyd. Stochastic Proximal Iteration: A Non-Asymptotic Improvement upon Stochastic Gradient Descent. working draft, accessed 2025, `https://ernestryu.com/papers/spi.pdf`.

[86] M. Schmidt and N. Le Roux. Fast Convergence of Stochastic Gradient Descent under a Strong Growth Condition, 2013. Preprint, available at `https://arxiv.org/abs/1308.6370`.

[87] A. Shapiro. Quantitative stability in stochastic programming. *Mathematical Programming*, 67(1):99–108, 1994.

[88] E. Specker. Nicht konstruktiv beweisbare Sätze der Analysis. *The Journal of Symbolic Logic*, 14:145–208, 1949.

[89] H. Zhang. New analysis of linear convergence of gradient-type methods via unifying error bound conditions. *Mathematical Programming*, 180(1):371–416, 2020.

[90] H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *Proceedings of the 29th Annual Conference on Learning Theory (COLT 2016)*, volume 49 of *Proceedings of Machine Learning Research*, pages 1617–1638. PMLR, 2016.

[91] J. R. Zhang, X. Mi, G. Du, Q. Sun, S. Wang, J. Li, and W. Zhou. A universal Banach–Bregman framework for stochastic iterations: Unifying stochastic mirror descent, learning and LLM training, 2025. Preprint, available at `https://arxiv.org/abs/2509.14216`.