MEAN-SQUARE AND LINEAR CONVERGENCE OF A STOCHASTIC PROXIMAL POINT ALGORITHM IN METRIC SPACES OF NONPOSITIVE CURVATURE

NICHOLAS PISCHKE

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

Abstract. We define a stochastic variant of the proximal point algorithm in the general setting of nonlinear (separable) Hadamard spaces for approximating zeros of the mean of a stochastically perturbed monotone vector field and prove its convergence under a suitable strong monotonicity assumption, together with a probabilistic independence assumption and a separability assumption on the tangent spaces. As a particular case, our results transfer previous work by P. Bianchi on that method in Hilbert spaces for the first time to Hadamard manifolds. Moreover, our convergence proof is fully effective and allows for the construction of explicit rates of convergence for the iteration towards the (unique) solution both in mean and almost surely. These rates are moreover highly uniform, being independent of most data surrounding the iteration, space or distribution. In that generality, these rates are novel already in the context of Hilbert spaces. Linear nonasymptotic guarantees under additional second-moment conditions on the Yosida approximates and special cases of stochastic convex minimization are discussed.

Keywords: Proximal point algorithm; stochastic approximation; rates of convergence; Hadamard spaces; proof mining
MSC2020 Classification: 47J25, 90C15, 62L20, 03F10

1. Introduction

1.1. **Background and motivation.** One of the fundamental problems in stochastic approximation is solving

$$\min_{x\in X}\int f(\xi,x)\,d\mu(\xi),$$

for a function $f : E \times X \to (-\infty, +\infty]$ on a probability space $(E, \mathcal{E}, \mu)$ and some other target space X. Indeed, this problem is widely studied for various classes of spaces X and functions f, with particular focus being placed on approximation methods and their complexity, and we refer to [14, 15, 44] and the references therein for various discussions along those lines. If X is a Hilbert space and f is a normal convex integrand (see [56]), some of the most prominent methods employed in that context are variants of the well-known stochastic proximal point method, that is the iteration

$$x_{n+1} = \operatorname{prox}^f_{\lambda_n}(\xi_{n+1}, x_n) \tag{+}$$

of random variables over an auxiliary probability space $(\Omega, \mathcal{F}, \mathbb{P})$, assuming a starting point $x_0 \in X$, a sequence of parameters $(\lambda_n) \subseteq (0, \infty)$ with certain growth conditions, a sequence (ξ_{n+1}) of random variables $\xi_{n+1} : \Omega \to E$ which are independent and identically distributed (i.i.d.) with (common) distribution μ, and writing $\operatorname{prox}^f_\lambda(\xi, x) := \operatorname{argmin}_{y\in X}\left\{f(\xi, y) + \frac{1}{2\lambda}\|x - y\|^2\right\}$ for the proximal map of f. This iteration and variants thereof, and in particular their complexities, are

Date: October 12, 2025.

widely studied under various assumptions on the convexity of f (we refer to [12, 14, 15, 44, 58] among many others), prominently in particular when f is strongly convex where then also fast rates of convergence can be achieved under additional moment conditions.

In a non-probabilistic setting, as is well-known (see e.g. [8]), proximal maps of convex functions are just a special instantiation of the general notion of a resolvent of a monotone operator. On this general level of monotone operators, Bianchi [17] (see also [16]) studied a corresponding variant of the proximal point algorithm phrased with resolvents for general monotone operators which are now stochastically perturbed, similar as to f above.

Concretely, let $(E, \mathcal{E}, \mu)$ and $(\Omega, \mathcal{F}, \mathbb{P})$ be probability spaces as before, let X be a separable Hilbert space and $A : E \times X \to 2^X$ be a set-valued map. Under a suitable measurability assumption on A and assuming the maximal monotonicity of $A(\xi, \cdot)$ (the precise assumptions will be discussed later), Bianchi [17] studied the iteration

$$(*) \qquad x_{n+1} = J_{\lambda_n}(\xi_{n+1}, x_n)$$

for an i.i.d. sequence (ξ_{n+1}) of random variables $\Omega \to E$ with distribution μ and a suitable sequence of parameters (λ_n) as before, where now $J_\lambda(\xi, x) := (\mathrm{Id} + \lambda A(\xi, \cdot))^{-1}(x)$.

This process indeed generalizes the method (+) discussed above by setting $A = \partial f$, with ∂f being the (stochastic) subgradient of f. Moreover, as highlighted in [17], this method bears resemblance in form to the seminal Robbins-Monro method [55] for finding roots of integral functions $\int V(s, x)\, d\mu(s)$, that is $x_{n+1} = x_n - \lambda_n V(\xi_{n+1}, x_n)$, with similar constants as above, as $(*)$ can also be equivalently written as $x_{n+1} = x_n - \lambda_n A_{\lambda_n}(\xi_{n+1}, x_n)$ where $A_\lambda(\xi, x) := (x - J_\lambda(\xi, x))/\lambda$ is the so-called Yosida approximation of the operator A.

While the iteration (+) approximates a minimizer of the mean of the function in question, the iteration generated by $(*)$ approximates a zero of the mean operator

$$\underline{A}(x) := \int A(s, x)\, d\mu(s),$$

where the integral refers to the Aumann integral [7]. Indeed, as shown in [17], the weighted averages of the sequence (x_n), that is $\overline{x}_n := \sum_{k=0}^n \lambda_k x_k / \sum_{k=0}^n \lambda_k$, converge weakly to a zero of $\underline{A}$. Approaching this convergence result faces considerable difficulties, resulting among others in an additional uniform integrability assumption, and the proof given by Bianchi in [17] is very sophisticated.

As illustrated in [17], these additional uniform integrability assumptions and some of the difficulties of the proof can be circumvented under the assumption of strong monotonicity of A, and the convergence can then be improved to the strong convergence of (x_n) towards the (in that context unique) zero of $\underline{A}$. However, even in that context the proof given in [17] is nontrivial, relying on various well-known results from stochastic approximation, like the Robbins-Siegmund theorem on supermartingale convergence, that are not immediately recognized to be effective. As put forward by Bianchi in [17], while the work [17] is set in a highly general context, "the price to pay with our approach is the absence of convergence rate certificates". Indeed, while rates have been given for various special cases, they have (essentially) always focused on the case $A = \partial f$ for some suitable convex function f and the general case remains, to our knowledge, quantitatively untreated already for strongly monotone operators A over Hilbert spaces.

1.2. **The contributions of the present paper and related work.** In the present paper, we augment the strong convergence result given in [17] under a strong monotonicity assumption with explicit rates of convergence both in expectation and almost surely. However, we move

considerably beyond simply quantitatively outfitting the results of Bianchi by lifting the algorithm to the general nonlinear setting of Hadamard spaces, that is complete geodesic metric spaces of nonpositive curvature.

These metric spaces of nonpositive curvature were originally introduced by Aleksandrov and are commonly called CAT(0) spaces, after the work of Gromov. Examples range from Hilbert spaces, $\mathbb{R}$-trees and Hadamard manifolds (i.e. complete simply connected Riemannian manifolds of nonpositive sectional curvature), to intricate examples like the Billera-Holmes-Vogtmann tree space prominently used in phylogenetics [18]. As illustrated by this plethora of spaces, extending tools and results from convex analysis to such metric contexts is particularly well motivated through applied considerations, not the least of which being the extensive developments of machine learning (where optimization over manifolds and other nonlinear spaces plays a key role, as discussed e.g. in [64]). We refer to the seminal monograph [19] for a comprehensive overview of CAT(0) and Hadamard spaces and further refer to [11] for a shorter treatment focused on aspects of convex analysis and optimization and to [2] for a recent treatment of geodesic metric spaces.

Even defining the results of Bianchi in this general context is rather subtle, requiring a synthesis of a range of different notions and results. Concretely, at first we rely on the theory of monotone vector fields in these general geodesic contexts, as introduced by Chaipunya, Kohsaka and Kumamin [22],[1] simultaneously generalizing monotone operators on Hilbert spaces and monotone vector fields on Hadamard manifolds (see Section 2.2). These in turn further require various considerations on the geometry of geodesic metric spaces, including in particular the notion of tangent spaces introduced in this general context by Nikolaev [50], generalizing the respective central notion from Riemannian manifolds (see Section 2.1). Beyond that, we naturally require a theory of integration in the context of Hadamard spaces which was largely developed in the seminal work of Sturm [60, 61] (see Section 2.3). In particular, we require an extension of Sturm's integral to set-valued mappings in this metric context, that is an Aumann-Sturm type integral.

All of these considerations come together to define stochastically perturbed monotone vector fields on separable Hadamard spaces (see Section 3) and with that a metric analogue of the stochastic proximal point algorithm of Bianchi (see Section 4), for which we prove a strong convergence result (see Theorem 4.7) under a strong monotonicity assumption together with an additional probabilistic independence assumption as well as a separability assumption on the metric tangent spaces, both discussed in full detail later on. This auxiliary independence assumption is in particular naturally satisfied in the context where the tangent spaces have flat curvature. Our result hence immediately covers both Bianchi's original setting of (separable) Hilbert spaces as well as Hadamard manifolds, in which case our results seem to be in particular qualitatively novel as, to our knowledge, such a stochastic variant of the general proximal point algorithm in the style of Bianchi was not considered in any kind of nonlinear context before.

In that way, our results also extend the previous seminal work of Li, López and Martín-Márquez [38] on the proximal point method for monotone vector fields in Hadamard manifolds for the first time to the stochastic context, at least in the special case of strong monotonicity. In the special case of the subgradient of a strongly convex function, which will be discussed

[1] As mentioned in [22], this notion seems to be distinct from the notion of a monotone operator on a CAT(0) space as introduced by Khatibzadeh and Ranjbar [31], relying on a previous notion of dual space for a CAT(0) space by Kakavandi and Amini [29].

throughout, our method in particular reduces to a stochastic proximal point algorithm in separable Hadamard spaces previous studied by Bačák [12] (extending [10] as well as [9]).[2]

Also in this generalized setting, our convergence proof is in fact fully effective and allows for the construction of explicit rates of convergence for the iteration towards the (unique) solution both in expectation and almost surely, which are highly uniform, being independent of most data surrounding the iteration, space or distribution. Even in the context of strong monotonicity assumptions, rates of convergence for the stochastic proximal point method are largely restricted to the setting of a convex function (where the assumption translates to a strong convexity assumption), such as in the well-known works [4, 52] as well as [25]. The only work known to us that treats general random monotone operators quantitatively is the recent preprint [59] where fast rates of convergence are derived under a strong monotonicity assumption in Hilbert spaces. This work however focuses on the method $x_{n+1} = J_\lambda(\xi_{n+1}, x_n)$, i.e. where the λ_n are kept constant, which is distinct from the one studied by Bianchi in [17] as the results there require $\lambda_n \to 0$ (by virtue of assuming $\sum_{n\in\mathbb{N}} \lambda_n^2 < \infty$). In that way, in the context of Bianchi's method from [17], the rates presented in this paper seem to us to be novel already in the context of Hilbert spaces in their generality.

In the end, we briefly discuss applications to the case of the minimization of the expectation of a strongly convex function (see Corollary 4.10). Also, we discuss additional uniform boundedness conditions on the second moments of the Yosida approximates, akin to some of the assumptions considered in [4, 25] in the special case of a convex function and restricted to a linear setting, which allow us to derive fast nonasymptotic guarantees (see Theorem 4.11).

The methods we employ here to derive the rates of convergence follows a general approach introduced recently by Neri, Powell and the author [46] towards constructing rates of convergence for very general classes stochastic approximation methods,[3] including ones that pertain to metric generalizations of stochastic quasi-Fejér monotonicity (as studied in Hilbert spaces in the seminal works of Combettes and Pesquet [23, 24]) of which the present algorithm is, crucially, a particular instance. Indeed, our paper in that way also serves as a case study to illustrate how the abstract approach from [46] can be used in a very concrete situation to give a perspicuous quantitative analysis of a rather involved algorithm and also how the general metric setting of [46] can be practically of use. We hope that the concrete applications presented in the present paper help to develop future applications of [46], such as potentially to the recent works of Karimi, Hsieh, Mertikopoulos and Krause [27, 30], where variants of the Robbins-Monro method over Riemannian manifolds were studied. In particular, the stochastic considerations on metric spaces of the present paper might be of help in lifting these results to a broader metric context.

Also for the stochastic proximal point method studied here, various questions remain which we hope can be answered by future research instigated by the present work, such as whether the weak convergence results from [17] or the fast rates derived for the distinct method $x_{n+1} = J_\lambda(\xi_{n+1}, x_n)$ in the previously mentioned preprint [59] lift to the metric setting.

[2]Indeed, our approach is quite different to that of [12]. In the context of a strongly convex function, our results dispense of the local compactness assumption and Lipschitz-like conditions from [12] while imposing separability and probabilistic independence assumptions on the tangent space. Comparing, and perhaps unifying, these works would prove for interesting future work.

[3]The results from [46], and likewise the present results, have been obtained using the logic-based methodology of *proof mining* [33, 34]. More precisely, they are part of a recent advance to apply these logical methods in probability theory and stochastic optimization for the first time [45, 46, 47, 48, 53]. As common in proof mining however, this paper avoids any reference to mathematical logic.

2. Preliminaries

In this section, we give the necessary preliminaries to objects involved in the present paper, such as geodesic metric spaces and their tangent bundles, monotone vector fields on these spaces, and in particular the theory of integration on geodesic spaces of nonpositive curvature. The range of different notions involved, the care needed to bring them together later, and the fact that such a combination has not been considered before, results in rather complex preliminaries which we however have tried to keep as minimal as possible.

2.1. **Geodesics,** CAT(0) **spaces and tangent spaces.** We begin with the background on geodesic metric spaces. Let (X, d) be a metric space. A geodesic is an isometry $\gamma : [0, l] \to X$. We call the image $\gamma([0, l])$ a geodesic segment and say that it joins $x = \gamma(0)$ and $y = \gamma(l)$, as well as that γ issues from x. Note that necessarily $l = d(x, y)$. X is called (uniquely) geodesic is every two points are joined by a (unique) geodesic. If such geodesics are unique, we denote the (unique) geodesic connecting two points $x, y \in X$ by $\gamma_{x,y}$.

For the purpose of this paper, a geodesic metric space (X, d) is now called a CAT(0) space (also called a space of nonpositive curvature in the sense of Aleksandrov) if it satisfies

$$d^2(\gamma(tl), x) \leqslant (1-t)d^2(\gamma(0), x) + td^2(\gamma(l), x) - t(1-t)d^2(\gamma(0), \gamma(l))$$

for all $x \in X$ and all geodesics $\gamma : [0, l] \to X$ (that is, an extension of the so-called Bruhat-Tits CN-inequality [20] to geodesics, see e.g. Proposition 2.3 in [61]). Any CAT(0) space is uniquely geodesic. A complete CAT(0) space is called a Hadamard space. As mentioned in the introduction, we refer to [2, 11, 19] for comprehensive overviews of geodesic metric spaces, CAT(0) spaces and Hadamard spaces, including alternative definitions.

Another characterization of CAT(0) spaces that will be useful in this paper was given by Berg and Nikolaev [13] using their so-called quasi-inner product (also called quasi-linearization function), that is the map defined by

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

for all $x, y, u, v \in X$, where we wrote $\overrightarrow{xy}, \overrightarrow{uv}$ as a shorthand for pairs $(x, y), (u, v) \in X^2$. As shown in [13], in any metric space (X, d) this function is the unique function $X^2 \times X^2 \to \mathbb{R}$ such that for all $x, y, u, v, w \in X$: (1) $\langle \overrightarrow{xy}, \overrightarrow{xy} \rangle = d^2(x, y)$; (2) $\langle \overrightarrow{xy}, \overrightarrow{uv} \rangle = \langle \overrightarrow{uv}, \overrightarrow{xy} \rangle$; (3) $\langle \overrightarrow{xy}, \overrightarrow{uv} \rangle = -\langle \overrightarrow{yx}, \overrightarrow{uv} \rangle$; (4) $\langle \overrightarrow{xy}, \overrightarrow{uv} \rangle + \langle \overrightarrow{xy}, \overrightarrow{vw} \rangle = \langle \overrightarrow{xy}, \overrightarrow{uw} \rangle$. It then follows from the results in [13] that a geodesic metric space (X, d) is a CAT(0) space if, and only if,

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

for all $x, y, u, v \in X$, i.e. where a metric version of the Cauchy-Schwarz inequality holds.

The most important notion for our paper regarding CAT(0) spaces is that of their tangent spaces as developed in the work of Nikolaev [50], as they can be used to provide analogs of fundamental notions from duality theory of linear spaces and manifolds. We essentially follow the exposition and (mostly) the notation of [37] and generally refer to [2, 19] for further exposition and proofs. Throughout, let (X, d) be a CAT(0) space. For nonconstant geodesics γ and η issuing from a point $x \in X$, their Aleksandrov angle $\angle_x(\gamma, \eta)$ is defined by

$$\angle_x(\gamma, \eta) := \lim_{s,t \to 0^+} \bar{\angle}_x(\gamma(s), \eta(t)),$$

with $\bar{\angle}_x(y, z)$, generally, referring to the comparison angle, defined via the comparison triangle $\bar{\Delta}(\bar{x}, \bar{y}, \bar{z})$ of the geodesic triangle $\Delta(x, y, z) \subseteq X$, as usual. The Aleksandrov angle $\angle_x$ now defines a pseudometric on the set of all nonconstant geodesics issuing from x. We write $\Sigma'_x X$ for the set of all equivalence classes of such geodesics under the equivalence relation defined

by $\angle_x(\gamma,\eta)=0$ and we still write $\angle_x$ for the Aleksandrov angle extended to these equivalence classes in the obvious way. The completion $(\Sigma_xX,\angle_x)$ of the space $(\Sigma'_xX,\angle_x)$ is called the metric space of directions from x and we denote the elements of it still by letters also used for geodesics, that is γ, η, etc. The tangent space T_xX of X at x is then the Euclidean cone over Σ_xX, that is $T_xX:=(\Sigma_xX\times[0,\infty))/\sim$ where $(\gamma,t)\sim(\eta,s)$ if, and only if, $t=s=0$ or $t=s>0$ and $\gamma=\eta$. For brevity, we write $t\gamma$ for the equivalence class $[(\gamma,t)]_\sim$ and given $u=t\gamma$ and $\lambda\geqslant 0$, we write $\lambda u:=(\lambda t)\gamma$. On T_xX, we define a metric

$$d_x(t\gamma,s\eta):=\sqrt{t^2+s^2-2ts\cos\angle_x(\gamma,\eta)}.$$

The space $TX=\bigcup_{x\in X}T_xX$ is called the tangent bundle of X. It in particular follows from the results of Nikolaev [50] that T_xX is a complete CAT(0) space, that is a Hadamard space. If X is a Hilbert space, then T_xX reduces to X and if X is a Hadamard manifold, then T_xX reduces to the usual Riemannian tangent space of X at x.

We now require some further notation and structure on T_xX. We write $0_x:=0\gamma$ and we introduce the notations $\|t\gamma\|_x:=d_x(0_x,t\gamma)=t$ and

$$g_x(t\gamma,s\eta):=\frac{1}{2}\left(\|t\gamma\|_x^2+\|s\eta\|_x^2-d_x^2(t\gamma,s\eta)\right)=ts\cos\angle_x(\gamma,\eta).$$

Note that $g_x(t\gamma,s\eta)=\left\langle\overrightarrow{0_xt\gamma},\overrightarrow{0_xs\eta}\right\rangle_x$, where we wrote $\langle\vec{\cdot},\vec{\cdot}\rangle_x$ for the quasi-inner product on T_xX, and so $g_x(t\gamma,s\eta)\leqslant\|t\gamma\|_x\|s\eta\|_x$ by (CS), as well as $g_x(t\gamma,t\gamma)=\|t\gamma\|_x^2$, $g_x(t\gamma,s\eta)=g_x(s\eta,t\gamma)$ and $g_x(t\gamma,s\eta)=tg_x(\gamma,s\eta)$.

Following e.g. [51] (see also [35, 26, 40]), we define the function $\log_x:X\to T_xX$ by $\log_x a:=d(x,a)\gamma_{x,a}$ for $a\neq x$ as well as $\log_x x:=0_x$, which provides an extension of the well-known inverse exponential map, crucial to the study of Riemannian manifolds and their curvature, to this metric setting.[4] Crucially, note that $\log_x$ is nonexpansive (see e.g. [26], eq. (2.4)), that is

$$d_x(\log_x a,\log_x b)\leqslant d(a,b).$$

The most important property of the pseudo-inner product g_x on T_xX is the following:

Lemma 2.1 (essentially Proposition 2.16 in [22])**.** *For any $x,a,b\in X$:*

$$g_x(t\log_x a,s\log_x b)\geqslant\frac{ts}{2}(d^2(x,a)+d^2(x,b)-d^2(a,b)).$$

Also, we will in the following use that g_x is Lipschitz continuous in both arguments, which we show in the following lemma:

Lemma 2.2. *For $x\in X$ and $u,v,w\in T_xX$: $|g_x(u,v)-g_x(u,w)|\leqslant\|u\|_x\,d_x(v,w)$.*

Proof. Note that

$$\begin{aligned}g_x(u,v)-g_x(u,w)&=\frac{1}{2}\|v\|_x^2-\frac{1}{2}\|w\|_x^2-\frac{1}{2}d_x^2(u,v)+\frac{1}{2}d_x^2(u,w)\\&=\frac{1}{2}\left(d_x^2(0_x,v)+d_x^2(u,w)-d_x^2(0_x,w)-d_x^2(u,v)\right)=\left\langle\overrightarrow{0_xu},\overrightarrow{vw}\right\rangle_x.\end{aligned}$$

Using that T_xX is a CAT(0) space, (CS) applied to $\left\langle\overrightarrow{0_xu},\overrightarrow{vw}\right\rangle_x$ yields

$$g_x(u,v)-g_x(u,w)=\left\langle\overrightarrow{0_xu},\overrightarrow{vw}\right\rangle_x\leqslant d_x(0_x,u)d_x(v,w)=\|u\|_x\,d_x(v,w).$$

Analogously, we obtain $g_x(u,w)-g_x(u,v)\leqslant\|u\|_x\,d_x(v,w)$. This yields the claim. □

[4] Indeed, under suitable assumptions on the extendibility of geodesics, one can consider the function $\exp_x t\gamma:=\gamma(t)$ for $\gamma\in T_xX$, of which $\log_x$ is an inverse of and which provides a metric analog of the exponential map of Riemannian manifolds. We will however not rely on this map in the rest of this paper.

2.2. **Monotone vector fields, maximality and resolvents.** We now discuss monotone vector fields in this metric context as introduced in [22],[5] extending monotone operators on Hilbert spaces and monotone vector fields on Hadamard manifolds as introduced in [43, 49] (see also [38, 39, 63]). Precisely, a monotone vector field on a CAT(0) space X is a mapping $A : X \to 2^{TX}$ such that $A(x) \subseteq T_xX$ and

$$g_x(u, \log_x y) \leqslant -g_y(v, \log_y x)$$

for all $(x, u), (y, v) \in A$. While not introduced in [22], in analogy to [38] we call the mapping strongly monotone (with parameter $\alpha > 0$) if

$$g_x(u, \log_x y) \leqslant -g_y(v, \log_y x) - \alpha d^2(x, y)$$

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

We denote the set of zeros of A by $\operatorname{zer} A := \{x \in X \mid 0_x \in A(x)\}$. If A is strongly monotone, then it immediately follows that its zero is unique if it exists.

Example 2.3. The canonical example for a monotone vector field is the subdifferential of a proper, convex and lower-semicontinuous function $f : X \to (-\infty, +\infty]$, with convexity expressed by requiring that

$$f(\gamma(tl)) \leqslant (1-t)f(\gamma(0)) + tf(\gamma(l))$$

for any geodesic $\gamma : [0, l] \to X$ and any $t \in [0, 1]$. This object was defined in the general setting of Hadamard spaces and their tangent bundles in [22] via

$$\partial f(x) := \{u \in T_xX \mid f(y) \geqslant f(x) + g_x(u, \log_x y) \text{ for all } y \in X\}.$$

In the usual settings of Hilbert spaces or Hadamard manifolds, this object naturally reduces to the subdifferential studied there. Further, in the recent work of Lewis, López-Acedo and Nicolae [37], an alternative characterization of this object is given, over locally compact spaces with the geodesic extension property, via normal cones and further substantial structure theory is provided (all the while performing the considerably more subtle task of providing these results on metric spaces with general upper bounded curvature). Indeed, that the subdifferential studied in [37] coincides with the one studied in [22] over suitable Hadamard spaces as considered in [37] follows from Proposition 4.4 therein. The monotonicity of ∂f follows immediately from the definition, see also Proposition 3.7 in [22].

Further, this object also provides a suitable example for a strongly monotone vector field in the case where f is a strongly convex function with constant $\alpha > 0$, i.e. where

$$f(\gamma(tl)) \leqslant (1-t)f(\gamma(0)) + tf(\gamma(l)) - t(1-t)\frac{\alpha}{2}d^2(\gamma(0), \gamma(l))$$

for any geodesic $\gamma : [0, l] \to X$ and any $t \in [0, 1]$, as the following Proposition 2.4 shows. In particular, this result generalizes a similar result given in [63] in the setting of monotone vector fields over Hadamard manifolds (which also requires a rather involved proof compared to the corresponding result in Hilbert spaces, originally due to Rockafellar [57], see also Example 22.4 in [8]).

Proposition 2.4. *If f is a strongly convex function with constant $\alpha > 0$, then ∂f is strongly monotone with constant α in this case*

[5]The work [22] relies on a slightly different approach towards the tangent spaces of a CAT(0) space, which however has no impact on the present paper. All results cited from [22] hold true in our setup as well.

Proof. Assuming w.l.o.g. that $x \neq y$, and writing $l = d(x,y)$, note that for $(x,u),(y,v) \in \partial f$, we have

$$(1-t)f(x) + tf(y) - t(1-t)\frac{\alpha}{2}d^2(x,y) \geqslant f(\gamma_{x,y}(tl)) \geqslant f(x) + g_x(u, \log_x \gamma_{x,y}(tl))$$

and so

$$f(y) - f(x) \geqslant \frac{g_x(u, \log_x \gamma_{x,y}(tl))}{t} + (1-t)\frac{\alpha}{2}d^2(x,y)$$

for any $t \in (0,1]$. Similarly, we get

$$f(x) - f(y) \geqslant \frac{g_y(v, \log_y \gamma_{y,x}(tl))}{t} + (1-t)\frac{\alpha}{2}d^2(x,y)$$

for all $t \in (0,1]$ so that combined, using $g_x(u, \log_x \gamma_{x,y}(tl))/t = g_x(u, \log_x y)$ (and similarly for y and v), we have

$$-g_y(v, \log_y x) - (1-t)\alpha d^2(x,y) \geqslant g_x(u, \log_x y).$$

Sending $t \to 0$ yields $-g_y(v, \log_y x) - \alpha d^2(x,y) \geqslant g_x(u, \log_x y)$ by which ∂f is strongly monotone with constant $\alpha > 0$. To see that $g_x(u, \log_x \gamma_{x,y}(tl))/t = g_x(u, \log_x y)$ (and similarly for y and v), note that we have $\log_x \gamma_{x,y}(tl) = d(x, \gamma_{x,y}(tl))\gamma_{x,\gamma_{x,y}(tl)} = td(x,y)\gamma_{x,\gamma_{x,y}(tl)}$ so that using Lemma 2.2, it holds that

$$\begin{aligned} |\tfrac{1}{t} g_x(u, \log_x \gamma_{x,y}(tl)) - g_x(u, \log_x y)| &= |g_x(u, \gamma_{x,\gamma_{x,y}(tl)}) - g_x(u, \gamma_{x,y})| \\ &\leqslant \|u\|_x \, d_x(\gamma_{x,\gamma_{x,y}(tl)}, \gamma_{x,y}) \\ &= \|u\|_x \sqrt{2 - 2\cos \angle_x(\gamma_{x,\gamma_{x,y}(tl)}, \gamma_{x,y})}. \end{aligned}$$

Now, by definition we have

$$\begin{aligned} \angle_x(\gamma_{x,\gamma_{x,y}(tl)}, \gamma_{x,y}) &= \lim_{s,s' \to 0^+} \bar{\angle}_x(\gamma_{x,\gamma_{x,y}(tl)}(s), \gamma_{x,y}(s')) \\ &= \lim_{s,s' \to 0^+} \bar{\angle}_x(\gamma_{x,y}(s), \gamma_{x,y}(s')) \\ &= \angle_x(\gamma_{x,y}, \gamma_{x,y}) = 0, \end{aligned}$$

using that $\gamma_{x,\gamma_{x,y}(tl)}(s) = \gamma_{x,y}(stl)$ for suitably small s. Hence, we have

$$|\tfrac{1}{t} g_x(u, \log_x \gamma_{x,y}(t)) - g_x(u, \log_x y)| \leqslant \|u\|_x \sqrt{2 - 2\cos \angle_x(\gamma_{x,\gamma_{x,y}(tl)}, \gamma_{x,y})} = 0.$$

□

Example 2.5. As follows from (the proof of) Corollary 4.5 in [37], in the case of the subdifferential ∂f of a proper, convex and lower-semicontinuous function $f : X \to (-\infty, +\infty]$ one has that

$$\operatorname{argmin} f = \{x \in X \mid 0_x \in \partial f(x)\} = \operatorname{zer} \partial f.$$

Following [22] (which in turn generalizes [39]), we define the resolvent J_λ via

$$J_\lambda x := \{z \in X \mid \tfrac{1}{\lambda} \log_z x \in A(z)\}.$$

As shown in (the proof of) Proposition 3.4 in [22], it follows from Lemma 2.1 that if A is monotone, then for any $x \in X$ and $\lambda > 0$, a $z \in X$ such that $\frac{1}{\lambda} \log_z x \in A(z)$ is necessarily unique. In that case, we identify J_λ with the corresponding (potentially partial) function from X to X.

A monotone vector field A is called maximal if its graph $\operatorname{gra} A := \{(x,u) \in X \times TX \mid u \in Ax\}$ cannot be extended properly while preserving monotonicity. By the well-known theorem of Minty [42], the maximality of a monotone operator over a Hilbert space is equivalent to the

totality of the resolvent. This extends to the setting of Hadamard manifolds as shown in [38] under the condition that the domain $\mathrm{dom}A := \{x \in X \mid A(x) \neq \varnothing\}$ is the whole space (see Remark 4.4 in [38]).

As mentioned in [22], it is unknown whether this equivalence between maximality and totality extends to this general setting of CAT(0) spaces. One direction however remains valid, as shown in Proposition 3.5 in [22]: if the resolvents are all total, that is for any $\lambda > 0$ and $x \in X$ there exists a $z \in X$ with $\frac{1}{\lambda}\log_z x \in A(z)$, then A is maximal. Following [22], we say that A satisfies the surjectivity condition if all resolvents are total.

Example 2.6. As shown in Proposition 3.8 in [22], in the case of the subdifferential ∂f of a proper, convex and lower-semicontinuous function $f : X \to (-\infty, +\infty]$, its resolvent is given by

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

that is the proximal map of f (or also called the Moreau-Yosida resolvent), which in this context of Hadamard spaces was first defined by Jost [28]. In particular, as each prox_λ^f is total, ∂f satisfies the surjectivity condition (and so is also maximal).

In terms of essential properties of the resolvent required in this paper, we have for any $\lambda > 0$ that J_λ is nonexpansive, that is

$$d(J_\lambda x, J_\lambda y) \leqslant d(x,y)$$

for any $x, y \in \mathrm{dom}(J_\lambda)$, and that $\mathrm{Fix}(J_\lambda) = \mathrm{zer}A$. The resolvent is in fact even firmly nonexpansive in the sense of Ariza-Ruiz, Leuştean and López-Acedo [3], but we will not rely on this here. For all these properties, we refer to Proposition 4.3 in [22].

At last, we will rely on the so-called Yosida approximate of the operator A. We define this object here via

$$A_\lambda x := \tfrac{1}{\lambda}\log_{J_\lambda x} x.$$

This seems to be distinct from the variant of the Yosida approximate introduced in [22], which relies on the so-called negative geodesics constructed using the geodesic extension property. However, the above definition precisely serves our purpose here as we have the following two crucial properties:

Lemma 2.7. *For any $\lambda > 0$ and any $x \in \mathrm{dom}(J_\lambda)$:*

(1) $A_\lambda x \in A(J_\lambda x)$,
(2) $\|A_\lambda x\|_{J_\lambda x} = \left\|\frac{1}{\lambda}\log_{J_\lambda x} x\right\|_{J_\lambda x} = \frac{1}{\lambda}d(x, J_\lambda x)$.

Proof. The first item is immediate by definition of J_λ as if $x \in \mathrm{dom}(J_\lambda)$, then $A_\lambda x = \frac{1}{\lambda}\log_{J_\lambda x} x \in A(J_\lambda x)$ since J_λ is single-valued. The second item is immediate from the definition of $\|\cdot\|$ and log, by which we have $\left\|\frac{1}{\lambda}\log_{J_\lambda x} x\right\|_{J_\lambda x} = \frac{1}{\lambda}d(x, J_\lambda x)$. □

On the contrary, the definition given in [22] does not seem to satisfy the above essential inclusion given in item (1) (while it does satisfy the norm property given in item (2)). Indeed, the present definition of the Yosida approximate seems to be a dual variant of that given in [22] which lives in the tangent space of the resolvent instead of that of the point. However, one benefit of the present object is that it does not rely on the geodesic extension property.

In any way, this mapping will however only serve a technical and in a way auxiliary purpose here and is not our main object of study so that no compatibility issues between the present work and [22] arise.

2.3. **Measurability and integration on** CAT(0) **spaces.** We now introduce the necessary background from the theory of random variables with values in Hadamard spaces. This theory goes back to the seminal work of Sturm and covers a range of advanced areas such as martingales and Markov processes. We here only rely on Sturm's relatively early works [60, 61] for the development of the L^1- and L^2-theory of these random variables. We also refer to the exposition of these matters given by Bačák [11]. For that, let $(T, \mathcal{T}, \tau)$ be a probability space and (X, d) be a separable Hadamard space. An (X-valued) random variable is a map $x : T \to X$ which is $\mathcal{T}/\mathcal{B}(X)$-measurable, where $\mathcal{B}(X)$ is the Borel σ-algebra of X. Write $\mathcal{P}(X)$ for the set of all probability measures on X and for $p \in [1, \infty)$, write $\mathcal{P}^p(X)$ for the set of all measures $P \in \mathcal{P}(X)$ such that $\int d^p(w, z)\, dP(w) < \infty$ for some/any $z \in X$. As usual, the push-forward measure τ_x, given by $\tau_x(A) := \tau(x^{-1}(A))$ for $A \in \mathcal{B}(X)$, is called the distribution of x. Naturally, we have $\tau_x \in \mathcal{P}(X)$.

Fundamentally (compare e.g. Theorem 2.3.1 in [11]), we have the following result in Hadamard spaces: For any $P \in \mathcal{P}^1(X)$ and some/any $y \in X$, there is a unique minimizer

$$b(P) := \operatorname{argmin}_{z \in X} \int \left(d^2(z, w) - d^2(w, y) \right) dP(w),$$

which is independent of y, called the barycenter of P. If $P \in \mathcal{P}^2(X)$, we further have

$$b(P) = \operatorname{argmin}_{z \in X} \int d^2(w, z)\, dP(w).$$

Given $p \in [1, \infty)$, the space $\mathcal{L}^p(T, X, \tau)$ is the space of all $\mathcal{T}/\mathcal{B}(X)$-measurable maps $x : T \to X$ with $d_p(x, z) < \infty$ for some/any $z \in X$, where $d_p^p(x, y) := \int d^p(x, y)\, d\tau$ for $\mathcal{T}/\mathcal{B}(X)$-measurable maps $x, y : T \to X$. The space $L^p(T, X, \tau)$ arises from this space by considering equivalence classes under the equivalence relation defined by $d_p(x, y) = 0$. Note that $x \in L^p(T, X, \tau)$ if, and only if, $\tau_x \in \mathcal{P}^p(X)$.

The expectation of a random variable $x \in L^1(T, X, \tau)$ is now defined via

$$\mathbb{E}[x] := \int x\, d\tau := b(\tau_x)$$

where $b(\tau_x)$ is the barycenter of τ_x as before.

We can similarly define the conditional expectation of a random variable $x : T \to X$ relative to a sub-σ-algebra $\mathcal{T}_0 \subseteq \mathcal{T}$. Concretely, as shown in [60], for any $x \in L^2(T, X, \tau)$ there is a unique equivalence class of $\mathcal{T}_0/\mathcal{B}(X)$-measurable random variables $z \in L^2(T, X, \tau)$ which minimizes $d_2(z, x)$. We denote this equivalence class by $\mathbb{E}[x \,|\, \mathcal{T}_0]$. Note that for $\mathcal{T}_0 = \{\varnothing, T\}$, this notion reduces to the previously defined expectation. This L^2-theory of conditional expectations then continuously extends to the L^1-case (see Corollary 2.4 in [60]).

Indeed, we require only relatively little theory of the above objects beyond their definitions. The first result is the following transformation theorem (which crucially employs separability):

Lemma 2.8 (p. 371 in [61])**.** *Let $(T', \mathcal{T}')$ be another measure space. For any $\mathcal{T}/\mathcal{T}'$-measurable $\zeta : T \to T'$ and any $\mathcal{T}'/\mathcal{B}(X)$-measurable $x : T' \to X$ such that $x \circ \zeta$ and x are integrable, it holds that*

$$\int x \circ \zeta\, d\tau = \int x\, d\tau_\zeta.$$

The second result that we mention is a version of Jensen's inequality, formulated even for conditional expectations:

Lemma 2.9 (Proposition 3.4 in [60])**.** *Let $\mathcal{T}_0 \subseteq \mathcal{T}$ be a sub-σ-algebra. Further, let $x \in L^1(T, X, \tau)$ and let $\varphi : X \to \mathbb{R}$ be lower-semicontinuous and convex with $(\varphi \circ x)_+ \in L^1(T, \tau)$. Then*

$$\mathbb{E}[\varphi \circ x \mid \mathcal{T}_0] \geqslant \varphi\left(\mathbb{E}[x \mid \mathcal{T}_0]\right).$$

The last result is the following independence property for conditional expectations, which lifts a well-known result from real-valued conditional expectations (see e.g. Theorem 8.14 in [32]) to the case of Hadamard spaces:

Lemma 2.10. *Let $\mathcal{T}_0 \subseteq \mathcal{T}$ be a sub-σ-algebra and let $x \in L^2(T, X, \tau)$ be independent of $\mathcal{T}_0$, that is its generated σ-algebra $\sigma(x)$ is independent of $\mathcal{T}_0$ in the usual sense (see e.g.* [32]*). Then*

$$\mathbb{E}[x \mid \mathcal{T}_0] = \int x \, d\tau.$$

Proof. Note that $\int x \, d\tau$ as a constant map is clearly in $L^2(T, X, \tau)$ and $\mathcal{T}_0/\mathcal{B}(X)$-measurable. To see $\mathbb{E}[x \mid \mathcal{T}_0] = \int x \, d\tau$ and as $x \in L^2(T, X, \tau)$, it thus remains to see that $\int x \, d\tau$ minimizes $d_2(z, x)$ amongst all $\mathcal{T}_0/\mathcal{B}(X)$-measurable $z \in L^2(T, X, \tau)$. For that, simply note that

$$\int d^2(z, x) \, d\tau = \int\int d^2(z(t'), x(t)) \, d\tau(t) \, d\tau(t') \geqslant \int d^2\left(\int x \, d\tau, x(t)\right) d\tau(t),$$

where the equality follows as z is $\mathcal{T}_0/\mathcal{B}(X)$-measurable and hence independent of x, and the inequality follows from the fact that $\int x \, d\tau = \operatorname{argmin}_{w \in X} \int d^2(x, w) \, d\tau$ as $x \in L^2(T, X, \tau)$. □

As we will later be concerned with these probabilistic notions for random variables taking values not only in a Hadamard space X but also in its tangent spaces $T_x X$, we comment on some measurability aspects and some crucial assumptions related to this already in the following remark:

Remark 2.11. As seen above, we crucially rely on the separability and completeness of the underlying geodesic space to develop probability theory over it. As a consequence, we will later throughout require that each tangent space $T_x X$ is separable, given an underlying separable Hadamard space X. As discussed in [26] (see also [1]), there is rather little structural theory on when the space $T_x X$ inherits these properties from the space X. While this assumption is clearly true for separable Hilbert space and Hadamard manifolds, the only other result in that vein we are aware of is given in [40] (see Corollary 5.7 and 5.8; see also Proposition 2.3.23 in [35]), where it is shown that if X is a separable, locally compact and geodesically complete $\mathrm{CAT}(\kappa)$ space, then $T_x X$ is a separable, locally compact and geodesically complete Hadamard space. However, we in general do not assume local compactness here.

In any case, if $T_x X$ is separable, then d_x is jointly measurable and so $\|\cdot\|_x$, g_x and $\langle \overrightarrow{xy}, \overrightarrow{uv} \rangle_x$ are (jointly) measurable as well. Further, note that $\log_x$ is measurable as it is nonexpansive and hence uniformly continuous.[6]

As usual, we say that a sequence (x_n) of (X-valued) random variables is independent if the generated σ-algebras $\sigma(x_n)$ are independent and (x_n) is called identically distributed if all distributions τ_{x_n} coincide. We abbreviate the property that a sequence is independent and identically distributed by i.i.d., also as usual.

[6]Note that as $T_x X$ is separable, $\log_x$ is also measurable in the context of lower bounded curvature as discussed in Remark 2.3 in [26], see also [51].

In a separable Hilbert space $(X, \langle\cdot,\cdot\rangle)$, it is well-known that independent (X-valued) random variables $u, v : T \to X$ satisfy

$$\int \langle u, v\rangle \, d\tau = \left\langle \int u \, d\tau, \int v \, d\tau \right\rangle,$$

where the integral in that case reduces to the Bochner integral on X. We will later require a fractional part of this property for random variables taking values in the tangent space T_xX of a Hadamard space X, relative to the mapping g_x considered above. While this therefore naturally holds for Hilbert spaces and Hadamard manifolds, it is not clear if this property holds in this full generality. Indeed, the only partial answer we can give here is that equality always holds when T_xX has flat curvature: If T_xX has also nonnegative curvature, in addition to the nonpositive curvature it natively has, then following similar arguments as in the proof of Theorem 2.4, (4) in [26], it is rather immediate to show that $g_x(u, v)$ is both convex and concave, in both arguments. Thus, for independent integrable random variables $u, v : T \to T_xX$, we have[7]

$$\int g_x(u, v) \, d\tau = \int\int g_x(u(t), v(t')) \, d\tau(t) \, d\tau(t')$$

by independence and thus further

$$\int\int g_x(u(t), v(t')) \, d\tau(t) \, d\tau(t') \leqslant \int g_x\left(\int u \, d\tau, v(t')\right) \, d\tau(t') \leqslant g_x\left(\int u \, d\tau, \int v \, d\tau\right)$$

by applying Jensen's inequality (recall Lemma 2.9) twice, using concavity. Using convexity, we obtain the above inequalities with opposite sign and so combined we get that

$$\int g_x(u, v) \, d\tau = g_x\left(\int u \, d\tau, \int v \, d\tau\right)$$

holds in that case. Further, by virtue of using Lemma 2.9, the above clearly remains true for conditional expectations, that is we have

$$\mathbb{E}[g_x(u, v) \mid \mathcal{T}_0] = g_x\left(\mathbb{E}[u \mid \mathcal{T}_0], \mathbb{E}[v \mid \mathcal{T}_0]\right)$$

for a sub-σ-algebra $\mathcal{T}_0 \subseteq \mathcal{T}$, in the case where T_xX has flat curvature.

As this issue is thereby rather delicate in this metric context, we will be highly explicit of any assumption relating to this property throughout. However, to emphasize this here again, the above equality and the results of this paper for that matter are in particular true in separable Hilbert spaces and Hadamard manifolds. We do not know whether the present result extends to prominent spaces beyond these cases, such as e.g. the Billera-Holmes-Vogtmann tree space [18], but recent characterizations of tangents spaces thereof given in [35] might be of help to establish the above independence property there.

We now transfer the above notion of an integral to set-valued operators in an analogous way as the seminal work of Aumann [7] did for random variables taking values in (separable) Hilbert spaces, replacing the use of the Bochner integral therein with the integral of Sturm.

Concretely, given a set-valued operator $F : T \to 2^X$, a function $\phi : T \to X$ is a measurable selection of F if it is $\mathcal{T}/\mathcal{B}(X)$-measurable and $\phi(s) \in F(s)$ for all $s \in T$. The set of all measurable selections of F is denoted by $\mathcal{S}(F)$ and we write $\mathcal{S}^p(F) := \mathcal{S}(F) \cap L^p(T, X, \tau)$ where $L^p(T, X, \tau)$

[7]Note that as u, v are integrable and independent, also $g_x(u, v)$ is integrable as we have $\int |g_x(u, v)| \, d\tau \leqslant \int \|u\|_x \|v\|_x \, d\tau = \left(\int \|u\|_x \, d\tau\right)\left(\int \|v\|_x \, d\tau\right) < \infty$, where the equality follows by independence.

is the L^p-space defined as above via Sturm's integral. The Aumann-Sturm integral of F is then defined as

$$\int F\,d\mu = \left\{ \int \phi\,d\mu \mid \phi \in \mathcal{S}^1(F) \right\}.$$

3. Random monotone vector fields on nonlinear spaces

We now bring the previous preliminaries together to define stochastically perturbed monotone vector fields on CAT(0) spaces. For that, let $(T, \mathcal{T}, \tau)$ be a probability space and X be a separable Hadamard space such that each T_xX is separable (recall the previous Remark 2.11).

Consider a set-valued operator $A : T \times X \to 2^{TX}$ with $A(s,x) \subseteq T_xX$ for all $s \in T$ and $x \in X$ such that $A(s,\cdot)$ is monotone for any $s \in T$.

For such an operator, we define the resolvent similar to before via

$$J_\lambda(s,x) := \{z \in X \mid \tfrac{1}{\lambda}\log_z x \in A(s,z)\}$$

for $\lambda > 0$, $s \in T$ and $x \in X$. As before, if $A(s,\cdot)$ is monotone, then such a z is necessarily unique if it exists and we denote it by $J_\lambda(s,x)$, similar to before.

The Yosida approximate is then lifted to this setting via

$$A_\lambda(s,x) := \tfrac{1}{\lambda}\log_{J_\lambda(s,x)} x,$$

so that Lemma 2.7 yields $A_\lambda(s,x) \in A(s, J_\lambda(s,x))$ as well as $\|A_\lambda(s,x)\|_{J_\lambda(s,x)} = \frac{1}{\lambda} d(x, J_\lambda(s,x))$ for all $\lambda > 0$, $s \in T$ and $x \in \mathrm{dom}(J_\lambda(s,\cdot))$.

There are various possible measurability properties for such operators which can be imposed. In the context of a Hilbert space with an operator $A : T \times X \to 2^X$, the most direct is perhaps to assume that A satisfies

$$(\%) \qquad \{(s,x) \in T \times X \mid A(s,x) \cap U \neq \varnothing\} \in \mathcal{T} \otimes \mathcal{B}(X)$$

for any open set $U \subseteq X$, a property that is often called (Effros) measurability. We refer to [6, 21] for further discussions on measurable set-valued operators. In Hilbert spaces and when A is maximally monotone, as also outlined in [17], it follows by Lemma 2.1 from [5] that this assumption implies (and is in fact equivalent to) the property that $J_\lambda(\cdot,x)$ is $\mathcal{T}/\mathcal{B}(X)$-measurable for any $x \in X$ and some (or any) $\lambda > 0$.

However, this equivalence does not seem to readily transfer to this hyperbolic setting. It is at first not completely clear how to adequately transfer the above measurability assumption to the nonlinear context, and if phrased as satisfying (%) for any open set $U \subseteq TX$, then the question for a suitable topology on TX remains. Further, for any immediate such choice, neither direction of the above equivalence seems to hold.

As it will be critical for us to guarantee the measurability of the resolvent, we will hence focus on this latter property. So we arrive at the following official definition:

Definition 3.1. An operator $A : T \times X \to 2^{TX}$ with $A(s,x) \subseteq T_xX$ is called a random monotone vector field if $A(s,\cdot)$ is monotone for any $s \in T$ and $J_\lambda(\cdot,x)$ is $\mathcal{T}/\mathcal{B}(X)$-measurable for any $x \in X$ and some (or any) $\lambda > 0$.

Note that this property fully suffices for our purposes here. In particular, we will make no further measurability assumptions on such set-valued operators A in the following.

As will be discussed in Example 3.2 below, the above condition can be immediately verified for the fundamental example of the subdifferential of a convex function.

In that context of a random monotone vector field A, now assume that $A(s,\cdot)$ also satisfies the surjectivity condition, i.e. for any $\lambda > 0$ and $x \in X$, there exists a $z \in X$ with $\frac{1}{\lambda}\log_z x \in A(s,z)$. Then, note that $J_\lambda(s,\cdot)$ is single-valued, total and nonexpansive as discussed before

and therefore, $J_\lambda(s,\cdot)$ is uniformly continuous for any $s \in E$. So, J_λ is a Carathéodory map whenever A is a random monotone vector field with the surjectivity condition. In particular, we then have that J_λ is $\mathcal{T}\otimes\mathcal{B}(X)/\mathcal{B}(X)$-measurable in that case (see e.g. Lemma 8.2.6 in [6]).

For such a random monotone vector field A, we define its mean $\underline{A}$ via

$$\underline{A}(x) := \int A(s,x)\, d\tau(s)$$

where the integral refers to the Aumann-Sturm integral as defined before, now on T_xX. Clearly, $\underline{A}$ is monotone if $A(s,\cdot)$ is monotone for every $s \in E$. Further, we introduce the notation $S_A(x) := \mathcal{S}(A(\cdot,x))$ and $S_A^p(x) := \mathcal{S}^p(A(\cdot,x))$. We write $\operatorname{zer}(\underline{A}) := \{x \in X \mid 0_x \in \underline{A}(x)\}$ for the set of zeros of $\underline{A}$ and for $p \geqslant 1$, we write

$$\mathcal{Z}_A(p) := \left\{ x \in X \mid \exists \phi \in S_A^p(x) \left(\int \phi \, d\mu = 0 \right) \right\}.$$

It should be noted that $\mathcal{Z}_A(p) \subseteq \mathcal{Z}_A(1) = \operatorname{zer}(\underline{A})$.

Later on, we will in particular assume that $A(s,\cdot)$ is strongly monotone with a modulus $\alpha(s)$, where $\alpha : E \to \mathbb{R}_+$ forms a measurable and integrable function such that $\int \alpha \, d\tau > 0$. Similar to [17], this implies that $\underline{A}$ is strongly monotone with modulus $\underline{\alpha} = \int \alpha \, d\mu > 0$.

Example 3.2. In analogy to [56], let $f : T \times X \to (-\infty, +\infty]$ be a normal convex integrand, i.e. $f(s,\cdot)$ is proper, lower-semicontinuous and convex for all $s \in T$ and f is $T\otimes\mathcal{B}(X)$-measurable. Then for its associated subdifferential

$$\partial f(s,x) := \{u \in T_xX \mid f(s,y) \geqslant f(s,x) + g_x(u, \log_x y) \text{ for all } y \in X\}$$

as in Example 2.3, its resolvents are given, following Example 2.6, by the proximal maps

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

It is thereby easy to see that ∂f satisfies the previous discussed measurability condition on the resolvents and hence is a random monotone vector field. Define

$$F(x) := \int f(s,x)\, d\tau(s)$$

and assume that F is proper. It further follows immediately that F is convex and lower-semicontinuous. If we now in analogy to [17] assume that

$$\underline{\partial f}(x) = \int \partial f(s,x)\, d\tau(s) = \partial \int f(s,x)\, d\tau(s) = \partial F(x),$$

then we in particular have

$$\operatorname{zer}\underline{\partial f} = \operatorname{zer}\partial F = \operatorname{argmin} F$$

by Example 2.5. Lastly, note that if it is further assumed that $f(s,\cdot)$ is strongly convex with constant $\alpha(s) > 0$ such that α is integrable with $\int \alpha \, d\tau > 0$, then $\partial f(s,\cdot)$ is strongly monotone with constant $\alpha(s)$ by Proposition 2.4, and so $\underline{\partial f}$ is strongly monotone with constant $\underline{\alpha} := \int \alpha \, d\tau$ and F is strongly convex with constant $\underline{\alpha}$.

4. A stochastic proximal point algorithm

Extending the previous work of Bianchi [17], we now consider a stochastic variant of the proximal point algorithm. Let $(E, \mathcal{E}, \mu)$ be a probability space and let X be a separable Hadamard space where each T_xX is also separable. Further, let $A : E \times X \to 2^{TX}$ be a random monotone vector field such that $A(s, \cdot)$ satisfies the surjectivity condition for any $s \in E$.

The stochastic proximal point method is now given by the iteration

$$x_{n+1} := J_{\lambda_n}(\xi_{n+1}, x_n) \tag{SPPA}$$

for a given starting value $x_0 \in X$, a sequence of parameters $(\lambda_n) \subseteq (0, \infty)$ and a sequence (ξ_{n+1}) of random variables $\xi_n : \Omega \to E$ for an ambient probability space $(\Omega, \mathcal{F}, \mathbb{P})$ over which the iteration takes place. Note that each x_n is thereby an (X-valued) random variable as J_{λ_n} is a Carathéodory map.

In terms of the parameters, we make the assumptions that

$$(\lambda_n) \in \ell^2_+ \backslash \ell^1_+ \text{ and that } (\xi_{n+1}) \text{ is i.i.d. with distribution } \mu, \tag{A0}$$

where we write ℓ^p_+ for the space of nonnegative p-summable sequences. To distinguish the two notions of integration we get from the two probability spaces, we use $\int$ to denote integrals over $(E, \mathcal{E}, \mu)$ and $\mathbb{E}$ to denote integrals over $(\Omega, \mathcal{F}, \mathbb{P})$. In further terms of notation, we in the following write $\mathcal{F}_n := \sigma(\xi_1, \dots, \xi_n)$ as well as $\mathbb{E}_n[\cdot]$ as a shorthand for the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}_n]$. Also, all (in)equalities are understood to hold almost surely (over the suitable probability space), if not stated otherwise.

Now, in the remainder of this section, we will establish a quantitative convergence result for the above stochastic proximal point method in the context of a strong monotonicity assumption. Concretely, we in the following assume that

$$A(s, \cdot) \text{ is strongly monotone with modulus } \alpha(s) > 0 \text{ such that } \int \alpha \, d\mu > 0. \tag{A1}$$

As discussed before, this property in fact entails that $\underline{A}$, the mean of the fields $A(s, \cdot)$, is strongly monotone with modulus $\underline{\alpha} = \int \alpha \, d\mu > 0$. We assume w.l.o.g. that $\alpha(s) \leqslant 1$ for any $s \in E$.

Motivated by the assumptions of Theorem 4 in [17], we assume that there exists a (hence unique) zero x^* of $\underline{A}$ which satisfies

$$x^* \in \mathcal{Z}_A(2). \tag{A2}$$

Further, we fix a $\phi^* \in S^2_A(x^*)$ with $\int \phi^* \, d\mu = 0_{x^*}$. Lastly, assume that $T_{x^*}X$ satisfies a partial independence property given by

$$\mathbb{E}_n[g_{x^*}(\phi^*(\xi_{n+1}), \log_{x^*} x_n)] = 0 \tag{A3}$$

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

We want to shortly discuss this assumption (A3) further. Note first that

$$g_{x^*}(\mathbb{E}_n[\phi^*(\xi_{n+1})], \mathbb{E}_n[\log_{x^*} x_n]) = 0.$$

Indeed, this follows from the assumption that $\int \phi^* \, d\mu = 0$ as follows: By Lemma 2.10, the independence of ξ_{n+1} from $\mathcal{F}_n$ yields that $\mathbb{E}_n[\phi^*(\xi_{n+1})] = \mathbb{E}[\phi^*(\xi_{n+1})]$ (where one should note that $\phi^*(\xi_{n+1}) \in L^2(\Omega, X, \mathbb{P})$ as $\phi^* \in S^2_A(x^*)$). Further, by Lemma 2.8 we have $\mathbb{E}[\phi^*(\xi_{n+1})] = \int \phi^* \, d\mu = 0_{x^*}$ and so $g_{x^*}(\mathbb{E}_n[\phi^*(\xi)], \mathbb{E}_n[\log_{x^*} x_n]) = 0$.

Thereby, the assumption (A3) is indeed a fragment of the principle

$$\mathbb{E}[g_{x^*}(u, v) \mid \mathsf{F}] = g_{x^*}(\mathbb{E}[u \mid \mathsf{F}], \mathbb{E}[v \mid \mathsf{F}])$$

for independent random variables $u, v : \Omega \to T_{x^*}X$ and $\mathsf{F} \subseteq \mathcal{F}$ a sub-σ-algebra, where above one concretely has $u = \phi^*(\xi_{n+1})$, $v = \log_{x^*} x_n$ and $\mathsf{F} = \mathcal{F}_n$.

In particular, combined with the discussion from Section 2.3, we get that (A3) holds in particular whenever $T_{x^*}X$ has flat curvature.

Under the assumptions (A0) – (A3), we now provide a strong convergence result for the stochastic proximal point algorithm presented in (SPPA). Moreover, this convergence result comes equipped with a highly uniform rate of convergence of the iteration both in mean and almost surely which seems to be, in this general context, even novel in Hilbert spaces (as already discussed in the introduction). The qualitative convergence result generalizes Theorem 4 in [17] and in that way at least seems to be novel in all classes of Hadamard spaces transcending Hilbert spaces satisfying our standing assumptions, in particular including Hadamard manifolds.

Our key analytical ingredients will be two central almost sure inequalities, modeled after inequalities established in [17].

The first is the inequality given in Lemma 4.1, establishing a type of stochastic quasi-Fejér monotonicity for the iteration in question. This inequality is modeled after the proof of Proposition 1 in [17] (see in particular p. 2244 therein).[8]

Lemma 4.1. *Let $\beta \in (0, \frac{1}{2}]$. For any $n \in \mathbb{N}$, we have*

$$\mathbb{E}_n[d^2(x_{n+1}, x^*)] \leqslant d^2(x_n, x^*) - \lambda_n^2(1-2\beta)\mathbb{E}_n[\|A_{\lambda_n}(\xi_{n+1}, x_n)\|_{x_{n+1}}^2] + \lambda_n^2 \frac{\int \|\phi^*\|_{x^*}^2 \, d\mu}{2\beta}$$

Proof. If not stated otherwise, all equalities and inequalities are understood to hold almost surely (if applicable). Using Lemma 2.1, we get

$$d^2(x_{n+1}, x^*) + d^2(x_{n+1}, x_n) - 2g_{x_{n+1}}(\log_{x_{n+1}} x_n, \log_{x_{n+1}} x^*) \leqslant d^2(x_n, x^*).$$

Now, note that

$$\begin{aligned} g_{x_{n+1}}(\log_{x_{n+1}} x_n, \log_{x_{n+1}} x^*) &= \lambda_n g_{x_{n+1}}(\tfrac{1}{\lambda_n}\log_{x_{n+1}} x_n, \log_{x_{n+1}} x^*) \\ &= \lambda_n g_{x_{n+1}}(A_{\lambda_n}(\xi_{n+1}, x_n), \log_{x_{n+1}} x^*) \\ &\leqslant -\lambda_n g_{x^*}(\phi^*(\xi_{n+1}), \log_{x^*} x_{n+1}), \end{aligned}$$

where the last inequality used the monotonicity of A, and that

$$A_{\lambda_n}(\xi_{n+1}, x_n) \in A(\xi_{n+1}, J_{\lambda_n}(\xi_{n+1}, x_n)) = A(\xi_{n+1}, x_{n+1})$$

as well as $\phi^*(\xi_{n+1}) \in A(\xi_{n+1}, x^*)$. Now, note that

$$\begin{aligned} &-\lambda_n g_{x^*}(\phi^*(\xi_{n+1}), \log_{x^*} x_{n+1}) \\ &\quad = -\lambda_n g_{x^*}(\phi^*(\xi_{n+1}), \log_{x^*} x_n) + \lambda_n(g_{x^*}(\phi^*(\xi_{n+1}), \log_{x^*} x_n) \\ &\qquad - g_{x^*}(\phi^*(\xi_{n+1}), \log_{x^*} x_{n+1})) \\ &\quad \leqslant -\lambda_n g_{x^*}(\phi^*(\xi_{n+1}), \log_{x^*} x_n) + \lambda_n \|\phi^*(\xi_{n+1})\|_{x^*} \, d_{x^*}(\log_{x^*} x_n, \log_{x^*} x_{n+1}) \\ &\quad \leqslant -\lambda_n g_{x^*}(\phi^*(\xi_{n+1}), \log_{x^*} x_n) + \lambda_n \|\phi^*(\xi_{n+1})\|_{x^*} \, d(x_n, x_{n+1}) \end{aligned}$$

using Lemma 2.2 and the fact that $\log_{x^*}$ is nonexpansive in a Hadamard space. Using (analogously to the proof of Lemma 2 in [17]) that

$$\lambda_n \|\phi^*(\xi_{n+1})\|_{x^*} \, d(x_n, x_{n+1}) \leqslant \frac{\lambda_n^2}{4\beta} \|\phi^*(\xi_{n+1})\|_{x^*}^2 + \beta d^2(x_n, x_{n+1}),$$

[8]The expression $\|A_{\lambda_n}(\xi_{n+1}, x_n)\|_{x_{n+1}}^2$ featuring in Lemma 4.1 could prove problematic from a measurability point of view, as it involves a random variable in the index of the tangent space norm, i.e. as the point of issue of a tangent space. However, note that $\lambda_n^2 \|A_{\lambda_n}(\xi_{n+1}, x_n)\|_{x_{n+1}}^2 = d^2(x_n, x_{n+1})$ as also crucially used in the proof, so that this expression is immediately measurable by the measurability of the resolvent and the joint measurability of the metric, using separability of X.

we get that

$$\begin{aligned} &g_{x_{n+1}}(\log_{x_{n+1}} x_n, \log_{x_{n+1}} x^*) \\ &\quad\leqslant -\lambda_n g_{x^*}(\phi^*(\xi_{n+1}), \log_{x^*} x_n) + \frac{\lambda_n^2}{4\beta} \left\|\phi^*(\xi_{n+1})\right\|_{x^*}^2 + \beta d^2(x_n, x_{n+1}) \end{aligned}$$

and so, using that $\lambda_n^2 \left\|A_{\lambda_n}(\xi_{n+1}, x_n)\right\|_{x_{n+1}}^2 = d^2(x_n, x_{n+1})$, we get

$$\begin{aligned} d^2(x_{n+1}, x^*) \leqslant d^2(x_n, x^*) &- \lambda_n^2(1-2\beta) \left\|A_{\lambda_n}(\xi_{n+1}, x_n)\right\|_{x_{n+1}}^2 \\ &- 2\lambda_n g_{x^*}(\phi^*(\xi_{n+1}), \log_{x^*} x_n) + \frac{\lambda_n^2}{2\beta} \left\|\phi^*(\xi_{n+1})\right\|_{x^*}^2 . \end{aligned}$$

We now apply the conditional expectation $\mathbb{E}_n$. Using the usual transformation rule, we get that $\mathbb{E}_n[\|\phi^*(\xi_{n+1})\|_{x^*}^2] = \int \|\phi^*\|_{x^*}^2 \, d\mu$. We get $\mathbb{E}_n[g_{x^*}(\phi^*(\xi_{n+1}), \log_{x^*} x_n)] = 0$ using assumption (A3). This yields

$$\mathbb{E}_n[d^2(x_{n+1}, x^*)] \leqslant d^2(x_n, x^*) - \lambda_n^2(1-2\beta)\mathbb{E}_n[\|A_{\lambda_n}(\xi_{n+1}, x_n)\|_{x_{n+1}}^2] + \lambda_n^2 \frac{\int \|\phi^*\|_{x^*}^2 \, d\mu}{2\beta}$$

which was the claim. □

Importantly, the above inequality yields the square integrability of the sequence (x_n):

Corollary 4.2. $\mathbb{E}[d^2(x_n, x^*)] \leqslant \mathbb{E}[d^2(x_0, x^*)] + \int \|\phi^*\|_{x^*}^2 \, d\mu \sum_{n=0}^{\infty} \lambda_n^2 < \infty$ for any $n \in \mathbb{N}$.

The next inequality again establishes a type of stochastic quasi-Fejér monotonicity for the iteration in question, however now with a different selection of error terms in the recurrence inequality based on the strong monotonicity of the field which now plays a crucial role, compared to the former inequality where it was not used. While at first sight perhaps redundant, it is exactly the interplay between this and the former inequality that will allow us to establish rates of convergence of the iteration in the end. This inequality is modeled after the proof of Theorem 4 in [17] (see in particular p. 2253 therein).

Lemma 4.3. *For any $n \in \mathbb{N}$, we have*

$$\mathbb{E}_n[d^2(x_{n+1}, x^*)] \leqslant (1 + 2\lambda_n^2) d^2(x_n, x^*) - 2\lambda_n \underline{\alpha} d^2(x_n, x^*) + \lambda_n^2 V_n$$

for $V_n = 2\mathbb{E}_n[\|A_{\lambda_n}(\xi_{n+1}, x_n)\|_{x_{n+1}}^2] + \int \|\phi^*\|_{x^*}^2 \, d\mu$.

Proof. We proceed similarly to the proof of Lemma 4.1 and get

$$d^2(x_{n+1}, x^*) + d^2(x_{n+1}, x_n) - 2g_{x_{n+1}}(\log_{x_{n+1}} x_n, \log_{x_{n+1}} x^*) \leqslant d^2(x_n, x^*)$$

as before using Lemma 2.1. Applying strong monotonicity in place of monotonicity then yields

$$\begin{aligned} &g_{x_{n+1}}(\log_{x_{n+1}} x_n, \log_{x_{n+1}} x^*) \\ &\quad\leqslant -\lambda_n g_{x^*}(\phi^*(\xi_{n+1}), \log_{x^*} x_{n+1}) - \lambda_n \alpha(\xi_{n+1}) d^2(x_{n+1}, x^*). \end{aligned}$$

As before in the proof of Lemma 4.1 (now with $\beta = \frac{1}{2}$), we get

$$\begin{aligned} &g_{x_{n+1}}(\log_{x_{n+1}} x_n, \log_{x_{n+1}} x^*) \\ &\quad\leqslant -\lambda_n g_{x^*}(\phi^*(\xi_{n+1}), \log_{x^*} x_n) + \frac{\lambda_n^2}{2} \left\|\phi^*(\xi_{n+1})\right\|_{x^*}^2 + \frac{1}{2} d^2(x_n, x_{n+1}) \end{aligned}$$

and so

$$\begin{aligned} d^2(x_{n+1}, x^*) \leqslant d^2(x_n, x^*) &- 2\lambda_n \alpha(\xi_{n+1}) d^2(x_{n+1}, x^*) \\ &- 2\lambda_n g_{x^*}(\phi^*(\xi_{n+1}), \log_{x^*} x_n) + \lambda_n^2 \left\|\phi^*(\xi_{n+1})\right\|_{x^*}^2 . \end{aligned}$$

Now, using Lemma 2.1 again, we get

$$\begin{aligned} d^2(x_{n+1},x^*) &\geqslant d^2(x_n,x_{n+1}) + d^2(x_n,x^*) - 2g_{x_n}(\log_{x_n} x_{n+1}, \log_{x_n} x^*) \\ &\geqslant d^2(x_n,x^*) - 2\lambda_n g_{x_n}(\tfrac{1}{\lambda_n}\log_{x_n} x_{n+1},) \\ &\geqslant d^2(x_n,x^*) - \lambda_n \left\|\tfrac{1}{\lambda_n}\log_{x_n} x_{n+1}\right\|_{x_n}^2 - \lambda_n \left\|\log_{x_n} x^*\right\|_{x_n}^2 \end{aligned}$$

where the third inequality follows using the definition of g_{x_n}. Combined, this yields

$$\begin{aligned} d^2(x_{n+1},x^*) \leqslant d^2(x_n,x^*) - 2\lambda_n\alpha(\xi_{n+1})d^2(x_n,x^*) + 2\lambda_n^2 \left\|\tfrac{1}{\lambda_n}\log_{x_n} x_{n+1}\right\|_{x_n}^2 \\ +2\lambda_n^2 \left\|\log_{x_n} x^*\right\|_{x_n}^2 - 2\lambda_n g_{x^*}(\phi^*(\xi_{n+1}),\log_{x^*} x_n) + \lambda_n^2 \left\|\phi^*(\xi_{n+1})\right\|_{x^*}^2 . \end{aligned}$$

where we have in particular used that $\alpha(s) \leqslant 1$. Note now that $\left\|\log_{x_n} x^*\right\|_{x_n} = d(x_n,x^*)$ and that $\|\frac{1}{\lambda_n}\log_{x_n} x_{n+1}\|_{x_n} = \frac{1}{\lambda_n} d(x_n,x_{n+1}) = \|A_{\lambda_n}(\xi_{n+1},x_n)\|_{x_{n+1}}$, so that we have

$$\begin{aligned} d^2(x_{n+1},x^*) \leqslant (1+2\lambda_n^2)d^2(x_n,x^*) - \lambda_n\alpha(\xi_{n+1})d^2(x_n,x^*) \\ +2\lambda_n^2 \left\|A_{\lambda_n}(\xi_{n+1},x_n)\right\|_{x_{n+1}}^2 - 2\lambda_n g_{x^*}(\phi^*(\xi_{n+1}),\log_{x^*} x_n) + \lambda_n^2 \left\|\phi^*(\xi_{n+1})\right\|_{x^*}^2 . \end{aligned}$$

We now again apply the conditional expectation $\mathbb{E}_n$. Using the usual transformation rule, we get $\mathbb{E}_n[\|\phi^*(\xi_{n+1})\|_{x^*}^2] = \int \|\phi^*\|_{x^*}^2 \, d\mu$ and $\mathbb{E}_n[g_{x^*}(\phi^*(\xi_{n+1}),\log_{x^*} x_n)] = 0$ follows from the assumption (A3), both as before. Also, using the independence of ξ_{n+1} and x_n as well as the usual transformation rule yields $\mathbb{E}_n[\alpha(\xi_{n+1})d^2(x_n,x)] = \underline{\alpha} d^2(x_n,x^*)$. Combined, we have

$$\mathbb{E}_n[d^2(x_{n+1},x^*)] \leqslant (1+2\lambda_n^2)d^2(x_n,x^*) - 2\lambda_n\underline{\alpha}d^2(x_n,x^*) + \lambda_n^2 V_n$$

which was the claim. □

Next, we endow our qualitative assumptions on the parameter sequence (λ_n) with moduli witnessing their quantitative content. To be precise, we in the following assume that we have functions $\chi : (0,\infty) \to \mathbb{N}$ and $\theta : \mathbb{N} \times (0,\infty) \to \mathbb{N}$ such that

(1) $\sum_{n=\chi(\varepsilon)}^{\infty} \lambda_n^2 < \varepsilon$ for all $\varepsilon > 0$,
(2) $\sum_{n=k}^{\theta(k,b)} \lambda_n \geqslant b$ for all $b > 0$ and $k \in \mathbb{N}$.

We also assume a bound $\Lambda > \sum_{n=0}^{\infty} \lambda_n^2$ and that we are given a $c > 0$ with $c > \int \|\phi^*\|_{x^*}^2 \, d\mu$. Lastly, we assume that $b > 0$ satisfies $b > \mathbb{E}[d^2(x_0,x^*)]$.

The second-to-last key quantitative result that we quote from the literature is the following quantitative version of a lemma of Qihou [54] (see also Lemma 5.31 in [8]):

Lemma 4.4 (Theorem 3.2 in [47])**.** *Let (x_n), (α_n), (β_n) and (γ_n) be sequences of nonnegative reals with*

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

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

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

Finally, we only require the following rather immediate result (which can for example be found in [46]).

Lemma 4.5. *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 $\sum_{n=k}^{\theta(k,b)} u_n \geqslant b$ 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 $v_n \geqslant \varepsilon$ for all $n \in [N; \theta(N, L/\varepsilon)]$. Then $L \leqslant \varepsilon \sum_{n=N}^{\theta(N,L/\varepsilon)} u_n \leqslant \sum_{n=N}^{\theta(N,L/\varepsilon)} u_n v_n \leqslant \sum_{n=0}^{\infty} u_n v_n < L$, which is a contradiction. □

We can now employ this to derive a so-called lim inf-rate in expectation for the sequence $d^2(x_n, x^*)$:

Lemma 4.6. *It holds that* $\liminf_{n\to\infty} \mathbb{E}[d^2(x_n, x^*)] = 0$ *with*

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

where $C := 4\,(b + \Lambda 2c) + \Lambda c$ *and* $D := e^{2\Lambda}(b + C)/2\underline{\alpha}$.

Proof. By Lemma 4.1 with $\beta = \frac{1}{4}$, we have

$$\mathbb{E}[d^2(x_{n+1}, x^*)] \leqslant \mathbb{E}[d^2(x_n, x^*)] - \frac{\lambda_n^2}{2}\mathbb{E}[\|A_{\lambda_n}(\xi_{n+1}, x_n)\|^2_{x_{n+1}}] + \lambda_n^2 2c.$$

Applying Lemma 4.4 yields $\sum_{n=0}^{\infty} \frac{\lambda_n^2}{2}\mathbb{E}[\|A_{\lambda_n}(\xi_{n+1}, x_n)\|^2_{x_{n+1}}] < 2\,(b + \Lambda 2c)$ and so

$$\sum_{n=0}^{\infty} \lambda_n^2 \left(\mathbb{E}[\|A_{\lambda_n}(\xi_{n+1}, x_n)\|^2_{x_{n+1}}] + c\right) < 4\,(b + \Lambda 2c) + \Lambda c =: C.$$

By Lemma 4.3, we have

$$\begin{aligned}&\mathbb{E}[d^2(x_{n+1}, x^*)]\\ &\quad\leqslant (1 + 2\lambda_n^2)\mathbb{E}[d^2(x_n, x^*)] - 2\lambda_n\underline{\alpha}\mathbb{E}[d^2(x_n, x^*)] + \lambda_n^2 \left(\mathbb{E}[\|A_{\lambda_n}(\xi_{n+1}, x_n)\|^2_{x_{n+1}}] + c\right)\end{aligned}$$

and so Lemma 4.4 implies that $\sum_{n=0}^{\infty} \lambda_n \mathbb{E}[d^2(x_n, x^*)] < \frac{e^{2\Lambda}(b+C)}{2\underline{\alpha}} =: D$. Finally, Lemma 4.5 implies $\liminf_{n\to\infty} \mathbb{E}[d^2(x_n, x^*)] = 0$ together with the respective rate as claimed. □

We can now already present our main theorem:

Theorem 4.7. *Let* $(E, \mathcal{E}, \mu)$ *and* $(\Omega, \mathcal{F}, \mathbb{P})$ *be probability spaces. Let* X *be a separable Hadamard space and assume that each* $T_x X$ *is also separable. Let* $A : E \times X \to 2^{TX}$ *with* $A(s, x) \subseteq T_x X$ *be a random monotone vector field and assume that* $A(s, \cdot)$ *satisfies the surjectivity condition for any* $s \in E$. *Let* (x_n) *be the iteration given by* (SPPA). *Assume* (A0) – (A3). *Then it holds that*

$$\mathbb{E}[d^2(x_n, x^*)] \to 0 \text{ and } d^2(x_n, x^*) \to 0 \text{ a.s.}$$

Moreover, the following rates of convergence apply: Let $\chi : (0, \infty) \to \mathbb{N}$ *and* $\theta : \mathbb{N} \times (0, \infty) \to \mathbb{N}$ *be such that*

$$\forall \varepsilon > 0 \left(\sum_{n=\chi(\varepsilon)}^{\infty} \lambda_n^2 < \varepsilon\right) \text{ and } \forall b > 0 \; \forall k \in \mathbb{N} \left(\sum_{n=k}^{\theta(k,b)} \lambda_n \geqslant b\right).$$

Let $\Lambda > \sum_{n=0}^{\infty} \lambda_n^2$. *Also, let* $\phi^* \in S_A^2(x^*)$ *such that* $\int \phi^* \, d\mu = 0_{x^*}$ *and* $c > 0$ *with* $c > \int \|\phi^*\|^2_{x^*} \, d\mu$. *Lastly, let* $b > 0$ *be such that* $b > \mathbb{E}[d^2(x_0, x^*)]$. *Then*

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

with rate $\rho(\varepsilon) := \theta(\chi(\varepsilon/2c), 2D/\varepsilon)$ *and*

$$\forall \lambda, \varepsilon > 0 \left(\mathbb{P}\left(\exists n \geqslant \rho'(\lambda, \varepsilon) \left(d^2(x_n, x^*) \geqslant \varepsilon\right)\right) < \lambda\right)$$

with rate $\rho'(\lambda, \varepsilon) := \rho(\lambda\varepsilon)$. *Here:* $C := 4\,(b + \Lambda 2c) + \Lambda c$ *and* $D := e^{2\Lambda}(b + C)/2\underline{\alpha}$.

Proof. It obviously suffices to establish the quantitative results. For any $n \in \mathbb{N}$, define $X_n := d^2(x_n, x^*) + c\sum_{m=n}^{\infty} \lambda_m^2$. As (x_n) is adapted to $(\mathcal{F}_n)$, also (X_n) is adapted to $(\mathcal{F}_n)$. As we have

$$\mathbb{E}_n[d^2(x_{n+1}, x^*)] \leqslant d^2(x_n, x^*) + \lambda_n^2 c \text{ a.s.}$$

by Lemma 4.1, the stochastic process (X_n) is a nonnegative supermartingale. Indeed, note that

$$\mathbb{E}_n[X_{n+1}] = \mathbb{E}_n\left[d^2(x_{n+1}, x^*)\right] + c\sum_{m=n+1}^{\infty} \lambda_m^2 \leqslant d^2(x_n, x^*) + c\sum_{m=n}^{\infty} \lambda_m^2 = X_n.$$

Now, let $\varepsilon > 0$ be arbitrary. Using Lemma 4.6, we choose an

$$n \in [\chi(\varepsilon/2c); \theta(\chi(\varepsilon/2c), 2D/\varepsilon)]$$

such that $\mathbb{E}[d^2(x_n, x^*)] < \varepsilon/2$. Let $m \geqslant n$ be arbitrary. Then

$$\mathbb{E}[d^2(x_m, x^*)] \leqslant \mathbb{E}[X_m] \leqslant \mathbb{E}[X_n] = \mathbb{E}[d^2(x_n, x^*)] + c\sum_{m=n}^{\infty} \lambda_m^2 < \varepsilon$$

using that (X_m) is a supermartingale and the properties of χ. As m was arbitrary, this yields $\mathbb{E}[\|x_n - x^*\|^2] \to 0$ and that ρ is a rate of convergence for that limit. For $d^2(x_n, x^*) \to 0$ a.s., note that

$$\mathbb{P}(\exists m \geqslant n(d^2(x_m, x^*) \geqslant a)) \leqslant \mathbb{P}(\exists m \geqslant n(X_m \geqslant a)) \leqslant \frac{\mathbb{E}[X_n]}{a}$$

where the second inequality follows from Ville's inequality [62] (see also [41]). This immediately implies that $d^2(x_n, x^*) \to 0$ a.s. with rate ρ'. □

While the proof is presented here in a self-contained style, it follows the outline and is effectively an instance of the proof of Theorem 2.8 in [46].

Remark 4.8. Note that the above Theorem 4.7 in particular holds whenever all tangent spaces are flat, where the crucial independence assumption (A3) is then validated as discussed before, so that the assumptions (A0) – (A2) suffice to establish the result. This in particular includes separable Hilbert spaces and separable Hadamard manifolds. To our knowledge, in the generality presented here, the quantitative aspects of the above Theorem 4.7 are already novel in the Hilbert space context while even the qualitative aspects of the above results seem to be novel in the context of Hadamard manifolds.

Remark 4.9. If ρ is invertible and decreasing, Theorem 4.7 immediately implies the nonasymptotic guarantee $\mathbb{E}[d^2(x_n, x^*)] \leqslant \rho^{-1}(n)$ for all $n \in \mathbb{N}$. We can then also derive a similar estimate for $\mathbb{P}(\exists m \geqslant n(d^2(x_m, x^*) \geqslant \varepsilon))$. However, the complexity of the rates that arise from Theorem 4.7 is without further assumptions rather dire, namely exponential:[9] For the canonical choice of $\lambda_n = \frac{1}{n+1}$, a quick calculation shows that we get

$$\mathbb{E}[d^2(x_n, x^*)] \leqslant \frac{4\max\{C, D\}}{\ln(n+2)} \text{ for all } n \in \mathbb{N}$$

in that case, with a similar bound on $\mathbb{P}(\exists m \geqslant n(d^2(x_m, x^*) \geqslant \varepsilon))$. This is because we make no other assumptions on A besides the strong monotonicity assumption and the minor underlying measurability assumptions. A brief discussion on fast rates under additional assumptions is given below.

[9] As such, the rates are similar for deterministic variants of the proximal point algorithm in metric settings without further assumptions than strong monotonicity, as e.g. obtained in [36] for the special case of strongly (even uniformly) convex functions.

As a particular corollary, we get the following result on minimizing strongly convex integrands as discussed in Example 3.2:

Corollary 4.10. Let $(E, \mathcal{E}, \mu)$ and $(\Omega, \mathcal{F}, \mathbb{P})$ be probability spaces. Let X be a separable Hadamard space and assume that each $T_x X$ is also separable. Let $f : T \times X \to (-\infty, +\infty]$ be a normal convex integrand such that $f(s, \cdot)$ is strongly convex with constant $\alpha(s) > 0$ such that α is integrable with $\int \alpha \, d\mu > 0$. Assume that $F(x) := \int f(s, x) \, d\nu(s)$ is proper and that $\int \partial f(s, x) \, d\tau(s) = \partial \int f(s, x) \, d\tau(s)$. Let (x_n) be the iteration given by $x_{n+1} := \operatorname{prox}^f_{\lambda_n}(\xi_{n+1}, x_n)$, and assume (A0), (A2) and (A3).

Then $\mathbb{E}[d^2(x_n, x^*)] \to 0$ and $d^2(x_n, x^*) \to 0$ a.s. Moreover, rates of convergence can be computed for these limits in similarity to Theorem 4.7.

Note that assumption (A1) is immediately satisfied in the context of Corollary 4.10 by virtue of Proposition 2.4. Again, the above is immediately true in separable Hilbert spaces and Hadamard manifolds, in which case the independence assumption (A3) disappears.

As a last result, we briefly discuss an additional assumption on A that allows us to derive fast rates of convergence. Concretely, assume the following uniform boundedness property for the second moments of the Yosida approximates along the iteration: There is a $\sigma > 0$ such that

$$\mathbb{E}[\|A_{\lambda_n}(\xi_{n+1}, x_n)\|^2_{x_{n+1}}] \leqslant \sigma \tag{A4}$$

for all $n \in \mathbb{N}$. This is a special case of a general uniform boundedness assumption of second moments of *arbitrary* selections from the random field A. In that way, assumption (A4) is akin to the uniform boundedness assumptions for second moments of subgradient selections as considered over linear spaces, e.g., in [4, 25]. In that context, we can derive the following fast nonasymptotic guarantees:

Theorem 4.11. *In the context of Theorem 4.7, assume* (A4). *Then for* $\lambda_n := 1/\underline{\alpha}(n+2)$, *it holds that*

$$\mathbb{E}[d^2(x_n, x^*)] \leqslant \frac{u}{n+2} \text{ and } \mathbb{P}\left(\exists m \geqslant n \left(d^2(x_m, x^*) \geqslant \varepsilon\right)\right) \leqslant \frac{1}{\varepsilon} \cdot \frac{e^{2\Lambda}(u + 4\sigma + 2c)}{n+2}$$

for all $n \in \mathbb{N}$, *where* $u = \max\{4\sigma + 2c, \lceil 4/\underline{\alpha}^2 \rceil (b + c\Lambda)\}$ *and* $b \geqslant \mathbb{E}[d^2(x_0, x^*)]$.

Proof. Note that $\mathbb{E}[d^2(x_n, x^*)] < b + c\Lambda$ by Corollary 4.2. Hence, the claim is clear for $n \leqslant n_0 := \lceil 4/\underline{\alpha}^2 \rceil - 2$. Using (A4), it follows from Lemma 4.3 that

$$\mathbb{E}[d^2(x_{n+1}, x^*)] \leqslant (1 + 2\lambda_n^2 - 2\lambda_n \underline{\alpha}) d^2(x_n, x^*) + \lambda_n^2 (2\sigma + c).$$

It is straightforward, albeit a bit tedious, to verify that $1 + 2\lambda_n^2 - 2\lambda_n \underline{\alpha} \leqslant 1 - 1.5/(n+2)$ for $n \geqslant n_0$, so that we obtain

$$\mathbb{E}[d^2(x_{n+1}, x^*)] \leqslant \left(1 - \frac{1.5}{n+2}\right) d^2(x_n, x^*) + \lambda_n^2 (2\sigma + c).$$

for all such n. We then get $\mathbb{E}[d^2(x_n, x^*)] \leqslant u/(n+2)$ for the u above by induction on n (see also Lemma 3.5 in [46]). Akin to the proof of Theorem 4.7, we can then also derive

$$\mathbb{P}(\exists m \geqslant n (d^2(x_m, x^*) \geqslant a)) \leqslant \frac{1}{\varepsilon} \cdot \frac{e^{2\Lambda}(u + 4\sigma + 2c)}{n+2}$$

by first moving to the supermartingale $X_n = d^2(x_n, x^*) + c \sum_{m=n}^{\infty} \lambda_m^2$ and then applying Ville's inequality. We omit the details here. □

As with the proof of Theorem 4.7, while we have presented the proof in a style tailored to (SPPA), it follows the outline and is effectively an instance of the proof of Theorem 3.6 in [46].

Acknowledgments: I want to thank Ulrich Kohlenbach, Morenikeji Neri and in particular Thomas Powell for many helpful comments on a previous draft of this paper.

References

[1] A. Ahidar-Coutrix, T. Le Gouic, and Q. Paris. Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics. *Probability Theory and Related Fields*, 177(1):323–368, 2020.
[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] D. Ariza-Ruiz, L. Leuştean, and G. López-Acedo. Firmly nonexpansive mappings in classes of geodesic spaces. *Transactions of the American Mathematical Society*, 366(8):4299–4322, 2014.
[4] H. Asi and J.C. Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.
[5] H. Attouch. Familles d'opérateurs maximaux monotones et mesurabilité. *Annali di Matematica Pura ed Applicata*, 120:35–111, 1979.
[6] J.-P. Aubin and H. Frankowska. *Set-Valued Analysis*. Springer, New York, 2009.
[7] R.J. Aumann. Integrals of set-valued functions. *Journal of Mathematical Analysis and Applications*, 12(1):1–12, 1965.
[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] M. Bačák. The proximal point algorithm in metric spaces. *Israel Journal of Mathematics*, 194(2):689–701, 2013.
[10] M. Bačák. Computing medians and means in Hadamard spaces. *SIAM Journal of Optimization*, 24(3):1542–1566, 2014.
[11] 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.
[12] M. Bačák. A variational approach to stochastic minimization of convex functionals. *Pure and Applied Functional Analysis*, 3(2):287–295, 2018.
[13] I.D. Berg and I.G. Nikolaev. Quasilinearization and curvature of Aleksandrov spaces. *Geometriae Dedicata*, 133:195–218, 2008.
[14] D.P. Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming. Series B*, 129:163–195, 2011.
[15] 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.
[16] P. Bianchi. A stochastic proximal point algorithm: convergence and application to convex optimization. In *Proceedings of the 2015 IEEE 6th International Workshop on Computa- tional Advances in Multi-Sensor Adaptive Processing*, pages 1–4. IEEE Press, Piscataway, NJ, 2015.
[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] M.R. Bridson and A. Haefliger. *Metric Spaces of Non-Positive Curvature*, volume 319 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin, Heidelberg, 1999.
[20] F. Bruhat and J. Tits. Groupes réductifs sur un corps local. I. Données radicielles valuées. *Publications Mathématiques de l'Institut des Hautes Études Scientifiques*, 41:5–251, 1972.
[21] C. Castaing and M. Valadier. *Convex Analysis and Measurable Multifunctions*. Lecture Notes in Mathematics. Springer Berlin, Heidelberg, 1977.
[22] 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.
[23] 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.

[24] 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.
[25] M. Eisenmann, T. Stillfjord, and M. Williamson. Sub-linear convergence of a stochastic proximal iteration method in Hilbert space. *Computational Optimization and Applications*, 83:181–210, 2022.
[26] T. Le Gouic, Q. Paris, P. Rigollet, and A.J. Stromme. Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space. *Journal of the European Mathematical Society*, 25(6):2229–2250, 2022.
[27] Y.P. Hsieh, M.R. Karimi, A. Krause, and P. Mertikopoulos. Riemannian stochastic optimization methods avoid strict saddle points. *Advances in Neural Information Processing Systems*, 36:29580–29601, 2023.
[28] J. Jost. Convex functionals and generalized harmonic maps into spaces of nonpositive curvature. *Commentarii Mathematici Helvetici*, 70:659–673, 1995.
[29] B.A. Kakavandi and M. Amini. Duality and subdifferential for convex functions on complete CAT(0) metric spaces. *Nonlinear Analysis: Theory, Methods & Applications*, 73(10):3450–3455, 2010.
[30] M.R. Karimi, Y.P. Hsieh, P. Mertikopoulos, and A. Krause. The dynamics of Riemannian Robbins-Monro algorithms, 2022. 33pp., math.OC, arXiv:2206.06795v3. Abstract in Proceedings of Thirty Fifth Conference on Learning Theory. PMLR 178:3503-3503.
[31] H. Khatibzadeh and S. Ranjbar. Monotone operators and the proximal point algorithm in complete CAT(0) metric spaces. *Journal of the Australian Mathematical Society*, 103(1):70–90, 2017.
[32] A. Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer Cham, 3rd edition, 2020.
[33] U. Kohlenbach. *Applied Proof Theory: Proof Interpretations and their Use in Mathematics*. Springer Monographs in Mathematics. Springer-Verlag Berlin Heidelberg, 2008.
[34] U. Kohlenbach. Proof-theoretic Methods in Nonlinear Analysis. In B. Sirakov, P. Ney de Souza, and M. Viana, editors, *Proceedings ICM 2018*, volume 2, pages 61–82. World Scientific, Singapore, 2019.
[35] L. Lammers. *Exploring Stickiness in* CAT(κ) *Spaces*. PhD thesis, Georg-August-Universität Göttingen, 2024.
[36] 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.
[37] A.S. Lewis, G. López-Acedo, and A. Nicolae. Basic convex analysis in metric spaces with bounded curvature. *SIAM Journal on Optimization*, 34(1):366–388, 2024.
[38] 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.
[39] C. Li, G. López, V. Martín-Márquez, and J.H. Wang. Resolvents of set-valued monotone vector fields in Hadamard manifolds. *Set-Valued and Variational Analysis*, 19(3):361–383, 2011.
[40] A. Lytchak and K. Nagano. Geodesically complete spaces with an upper curvature bound. *Geometric and Functional Analysis*, 29:295–342, 2019.
[41] M. Métivier. *Semimartingales*, volume 2 of *De Gruyter Studies in Mathematics*. Walter de Gruyter GmbH, Berlin/Boston, 1982.
[42] G.J. Minty. Monotone (nonlinear) operators in Hilbert spaces. *Duke Mathematical Journal*, 29:341–346, 1962.
[43] S.Z. Németh. Monotone vector fields. *Publicationes Mathematicae Debrecen*, 54:437–449, 1999.
[44] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal of Optimization*, 19:1574–1609, 2009.
[45] M. Neri and N. Pischke. Proof mining and probability theory, 2024. Preprint, `https://arxiv.org/abs/2403.00659`.
[46] M. Neri, N. Pischke, and T. Powell. On the asymptotic behaviour of stochastic processes, with applications to supermartingale convergence, Dvoretzky's approximation theorem, and stochastic quasi-Fejér monotonicity, 2025. Preprint, `https://arxiv.org/abs/2504.12922`.
[47] M. Neri and T. Powell. A quantitative Robbins-Siegmund theorem. *Annals of Applied Probability*, 2024. To appear, `https://arxiv.org/abs/2410.15986`.
[48] M. Neri and T. Powell. On quantitative convergence for stochastic processes: Crossings, fluctuations and martingales. *Transactions of the American Mathematical Society, Series B*, 12:974–1019, 2025.
[49] J.X. Da Cruz Neto, O.P. Ferreira, and L.R. Lucambio Pérez. Monotone point-to-set vector fields. *Balkan Journal of Geometry and Its Applications*, 5:69–79, 2000.
[50] I.G. Nikolaev. The tangent cone of an Aleksandrov space of curvature $\leqslant K$. *Manuscripta Mathematica*, 86(1):137–147, 1995.

[51] S. Ohta. Barycenters in Alexandrov spaces of curvature bounded below. *Advances in Geometry*, 12(4):571–587, 2012.
[52] A. Patrascu and I. Necoara. Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *Journal of Machine Learning Research*, 18(198):1–42, 2018.
[53] N. Pischke and T. Powell. Asymptotic regularity of a generalised stochastic Halpern scheme , 2025. Preprint, `https://arxiv.org/abs/2411.04845`.
[54] L. Qihou. Iteration sequences for asymptotically quasi-nonexpansive mappings with error member. *Journal of Mathematical Analysis and Applications*, 259:18–24, 2001.
[55] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:400–407, 1951.
[56] 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.
[57] R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal of Control and Optimization*, 14:877–898, 1976.
[58] 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`.
[59] A. Sadiev, L. Condat, and P. Richtárik. Stochastic Proximal Point Methods for Monotone Inclusions under Expected Similarity, 2024. Preprint, `https://arxiv.org/abs/2405.14255`.
[60] K.-T. Sturm. Nonlinear martingale theory for processes with values in metric spaces of nonpositive curvature. *The Annals of Probability*, 30(3):1195–1222, 2002.
[61] K.-T. Sturm. Probability measures on metric spaces of nonpositive curvature. In P. Auscher, T. Coulhon, and A. Grigoryan, editors, *Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces*, volume 338 of *Contemporary Mathematics*, pages 357–390. American Mathematical Society, Providence, RI, 2003.
[62] J. Ville. *Étude Critique de la Notion de Collectif*. PhD thesis, École Polytechnique, 1939.
[63] J.H. Wang, G. López, V. Martín-Márquez, and C. Li. Monotone and accretive vector fields on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 146(3):691–708, 2010.
[64] 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)*, volume 49 of *Proceedings of Machine Learning Research*, pages 1617–1638. PMLR, 2016.