WEAK CONVERGENCE OF THE STOCHASTIC PROXIMAL POINT METHOD IN METRIC SPACES

NICHOLAS PISCHKE

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

Abstract. We prove the almost sure weak convergence of a stochastic proximal point method for minimizing a convex integral function in the general nonlinear context of complete geodesic metric spaces of nonpositive curvature (so-called Hadamard spaces), solving a problem of M. Bačák. This method, formulated in the context of a mild growth condition on the function which generalizes Lipschitz continuity, was previously only considered in the context of strong metric regularity conditions or in the context of locally compact spaces. The proof is a combination of a weak almost sure convergence theorem for stochastic processes in Hadamard spaces which confine to a stochastic variant of quasi-Fejér monotonicity, due to previous work of the author, together with a new argument for proving the almost sure convergence of the mean function values of the process towards the minimal value.

Keywords: Proximal point algorithm; stochastic approximation; weak convergence; Hadamard spaces
MSC2020 Classification: 47J25, 90C15, 90C25, 62L20

1. Introduction

1.1. **Background and motivation.** The problem of minimizing a convex integral function, that is solving the problem

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

for a given normal convex integrand $f : E \times X \to (-\infty, +\infty]$ (see [37]) on a complete probability space $(E, \mathcal{E}, \mu)$ and some target Hilbert space X, is one of the most important general formulations of stochastic approximation. Motivated by the seminal proximal point algorithm for approximating minimizers of "ordinary" convex functions on such spaces, which goes back to the work of Rockafellar [38], Martinet [28] as well as Brézis and Lions [15], one of the prevalent modern tools for approaching this problem is the stochastic proximal point method

$$x_{n+1} = \operatorname{argmin}_{y\in X} \left\{ f(\xi_{n+1}, y) + \frac{1}{2\lambda_n}\|x_n - y\|^2 \right\},$$

formulated over an auxiliary probability space $(\Omega, \mathcal{F}, \mathbb{P})$, given a starting point $x_0 \in X$, a sequence of parameters $(\lambda_n) \subseteq (0, \infty)$ with certain growth conditions and 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. This iteration and its variants are widely studied and we refer to [4, 12, 13, 30, 39], among many others, for various such discussions (in particular regarding their complexities).

Extensions of these tools from (stochastic) convex analysis to nonlinear settings are of high practical relevance, particularly because of the extensive developments of machine learning in recent years, where optimization over nonlinear spaces such as manifolds plays a key role (see

Date: May 19, 2026.

e.g. [41]). However, the need for suitable methods in nonlinear contexts is not only driven through optimization on manifolds or other domains with differentiable structure, but also in particular by spaces with much sparcer structure naturally occurring in applications, such as e.g. the Billera-Holmes-Vogtmann tree space [14] prominently used in phylogenetics (see also e.g. [7]).

In this paper, we study the stochastic proximal point method in the context of the general class of geodesic metric spaces with nonpositive curvature, as introduced in the work of Alexandrov [2]. Also known as CAT(0) spaces, following the work of Gromov [22], these spaces uniformly cover spaces such as Hilbert spaces, $\mathbb{R}$-trees and Hadamard manifolds (i.e. complete simply connected Riemannian manifolds of nonpositive sectional curvature) as well as the previously mentioned Billera-Holmes-Vogtmann tree space or the Hilbert ball, and further involved examples. Authoritative references for geodesic and CAT(0) spaces are in particular the works [3, 16] as well as [8], with the latter providing a shorter treatment focused in particular on convex analysis and optimization.

The deterministic proximal point method was lifted to the setting of Hadamard spaces (that is *complete* geodesic metric spaces with nonpositive curvature) by Bačák [6], relying on metric variants of the proximal mappings. In particular, the work [6] established weak convergence in Hadamard spaces of this deterministic proximal point method. Naturally, by Güler's seminal work [23], this is the most one can hope for already in Hilbert spaces.

The stochastic proximal point method, as sketched over Hilbert spaces above, was extended to the setting of (separable) Hadamard spaces in the work [9], building on preceding work [7] on a splitting proximal point method with random order for finite sums of convex functions over similar spaces.

1.2. **Main results and related work.** We now fix the formal framework. In similarity to the above, let $(\Omega, \mathcal{F}, \mathbb{P})$ and $(E, \mathcal{E}, \mu)$ be probability spaces, with $(E, \mathcal{E}, \mu)$ complete, and let X now be a separable Hadamard space. In analogy to [37] (see also [18]), 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[1] for all $e \in E$ and f is $\mathcal{E} \otimes \mathcal{B}(X)$-measurable. 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. [25, 29]). Further, $\operatorname{prox}_\lambda^f(e, \cdot)$ is nonexpansive for any $e \in E$ and $\lambda > 0$ (see e.g. Lemma 4 in [25]), 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 $\mathcal{E} \otimes \mathcal{B}(X)$-measurable (see e.g. Lemma 8.2.6 in [5]).

The stochastic proximal point method is then given by the iteration

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

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

$$(\xi_{n+1}) \text{ is i.i.d. with distribution } \mu \text{ and } \sum_{n \in \mathbb{N}} \lambda_n = +\infty, \sum_{n \in \mathbb{N}} \lambda_n^2 < +\infty. \tag{A1}$$

The work [9] 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:

[1]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 in X.

Assume there exists a positive function $L \in L^2(E,\mu)$ and a point $p \in X$ such that

$$f(e,x) - f(e,y) \leq L(e)(1+d(x,p))d(x,y) \tag{A2}$$

for all $x, y \in X$ and almost all $e \in E$. We refer to [9] for further discussions of this condition as well as on the benefits of the proximal point method at large compared to methods such as gradient descent (see also e.g. [12]), especially in contexts such as Hadamard spaces, where usual stochastic gradient methods cannot be used without additional differential structure that one would normally have access to on e.g. manifolds. Beyond that, we refer to the excellent exposition in [6, 7, 9] for further discussions on these methods and relations other works.

The only previous *general* convergence result on the stochastic proximal point method in Hadamard spaces formulated above, that is *without* any further regularity assumptions such as strong convexity or weak sharp minima,[2] is a strong convergence result in the context of a local compactness assumption, given in [9] (see also [10]).

Theorem 1.1 (Theorem 3.1 in [9])**.** *Let $(E,\mathcal{E},\mu)$ and $(\Omega,\mathcal{F},\mathbb{P})$ be probability spaces, with $(E,\mathcal{E},\mu)$ complete, and let X be a locally compact Hadamard space. Let $f : E\times X \to (-\infty,+\infty]$ be a normal convex integrand such that $F(x) := \int f(e,x)\,d\mu(e)$ is proper and* $\operatorname{argmin} F \neq \emptyset$*. Let (x_n) be the iteration given by* (SPPA)*, and assume* (A1) *as well as* (A2)*.*

Then (x_n) a.s. strongly converges to an $\operatorname{argmin} F$*-valued random variable.*

In this paper, we prove the weak convergence of the stochastic proximal point method as formulated in [9] in the context of general (separable) Hadamard spaces.

Theorem 1.2. *Let $(E,\mathcal{E},\mu)$ and $(\Omega,\mathcal{F},\mathbb{P})$ be probability spaces, with $(E,\mathcal{E},\mu)$ complete, and let X be a separable Hadamard space. Let $f : E \times X \to (-\infty,+\infty]$ be a normal convex integrand such that $F(x) := \int f(e,x)\,d\mu(e)$ is proper and* $\operatorname{argmin} F \neq \emptyset$*. Let (x_n) be the iteration given by* (SPPA)*, and assume* (A1) *as well as* (A2)*.*

Then (x_n) a.s. weakly converges to an $\operatorname{argmin} F$*-valued random variable.*

This in particular solves Problem 6.6 of Bačák [10]. Further, as the stochastic proximal point method generalizes the splitting proximal point method with random order given in [7] (see Theorem 3.7 therein), we also get the following:

Theorem 1.3. *Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space and let X be a separable Hadamard space. Let $F(x) := \sum_{k=1}^{N} f_k(x)$, where each $f_k : X \to (-\infty,+\infty]$ is proper convex lsc, and where we assume that* $\operatorname{argmin} F \neq \emptyset$*. Let (x_n) be the iteration given by*

$$x_{n+1} := \operatorname{prox}_{\lambda_n}^{f_{r_{n+1}}}(x_n)$$

where (r_{n+1}) is a sequence of independent random variables which attain values in $\{1,\dots,N\}$ according to the uniform distribution. Assume (A1) *and* (A2)*.*[3]

Then (x_n) a.s. weakly converges to an $\operatorname{argmin} F$*-valued random variable.*

[2] We refer e.g. to the recent work [36] which provides quantitative results for the proximal point method considered herein under stronger regularity assumptions, such as weak sharp minima (and beyond), relying on rather general considerations on stochastic quasi-Fejér monotonicity in a metric context (see also the related [31]).

[3] By (A2), we here mean the condition that $f_k(x) - f_k(y) \leq L(k)(1+d(x,p))d(x,y)$ for all $k = 1,\dots,N$ and $x, y \in X$, for some $L(k) > 0$, see also the proof of Theorem 1.3 given later on.

This in particular solves Problem 6.4 of Bačák [10].[4] It should be noted that according to [10], both results are new even in Hilbert spaces which seems to be true, to our knowledge.

Our argument is based on two key ingredients: The first is a general (weak) convergence theorem for stochastic processes in Hadamard spaces which confine to a stochastic variant of quasi-Fejér monotonicity (see Lemma 3.1 later on), derived in recent work of the author [35]. This result extends similar such general stochastic convergence theorems obtained by Combettes and Pesquet [19] over Hilbert spaces to the setting of Hadamard spaces, and simultaneously extends related deterministic convergence theorems over Hadamard spaces obtained by Bačák, Searston and Sims [11] to the stochastic setting. This result in particular relies on a nonlinear variant of Pettis' theorem [34], similarly derived in [35]. The second key ingredient is a new convergence result for stochastic recursive inequalities, used for proving the almost sure convergence of the mean function values of the process towards the minimal value (see Lemma 3.4 later on). This result can be considered as a stochastic extension of a similar deterministic result due to Alber, Iusem and Solodov [1] (discussed in more detail later), and in particular combines some of the arguments found there with Kolmogorov's two-series theorem (see Lemma 3.3 later on).

Even though we only consider the stochastic proximal point method over Hadamard spaces as given in [9] as well as its special case from [7] in the present paper, we think that the general approach taken here will be of use for the convergence analysis of further methods from stochastic convex optimization already over Hilbert spaces, but in particular also over nonlinear spaces. One particular example we want to mention here is the Busemann subgradient method introduced recently by Goodwin, Lewis, López-Acedo and Nicolae [21], and its extension for stochastic minimization as considered in [35].

2. Preliminaries

We now discuss the few preliminary definitions, results and notations that we require throughout. As mentioned in the introduction, beyond the results indicated here, we refer to [3, 8, 16] for a comprehensive overview of geodesic metric spaces and their properties, in particular to [8] for aspects of (stochastic) optimization. Further, we refer to e.g. [27] for a standard textbook on probability theory.

Let (X, d) be a metric space. A geodesic is an isometry $\gamma : [0, l] \to X$ (where necessarily $l = d(x, y)$). We say that it joins $x = \gamma(0)$ and $y = \gamma(l)$. X is called (uniquely) geodesic is 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

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

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 [17] to geodesics). Any CAT(0) space is uniquely geodesic and a complete CAT(0) space is called a Hadamard space.

Weak convergence in CAT(0) spaces goes back to the work of Jost [24] and is often called Δ-convergence following the work of Kirk and Panyanak [26] (we refer in particular to the discussion in [6] on that matter). We define weak convergence here as follows (see e.g. [8]):

[4]The choice to allow general $x, y \in X$ in (A2) is purely for simplicity. Throughout the paper, it would suffice to assume this condition only along the sequence (x_n), akin to [7]. In that context, it is then immediate to see that the assumption (A2) in Theorem 1.3 generalizes the growth condition used in Theorem 3.7 in [7]. We however stay with this formulation of (A2) here, for simplicity.

Given a bounded sequence $(x_n) \subseteq X$ and a point $x \in X$, their asymptotic radius is given by

$$r(x_n, x) := \limsup_{n\to\infty} d^2(x_n, x)$$

and the general asymptotic radius of the sequence (x_n) is given by

$$r(x_n) := \inf_{x\in X} r(x_n, x).$$

A point $x \in X$ is called an asymptotic center of (x_n) if $r(x_n, x) = r(x_n)$. In Hadamard spaces, asymptotic centers exist and are unique (see e.g. Proposition 7 in [20]).

We say that (x_n) weakly converges to $x \in X$, written $x_n \to^w x$, if x is the asymptotic center of each subsequence of (x_n). A point $x \in X$ is a weak cluster point of (x_n) if there is a subsequence (x_{n_k}) of (x_n) with $x_{n_k} \to^w x$.

We write $\mathfrak{W}(x_n)$ for the set of all weak cluster points of (x_n) and $\mathfrak{S}(x_n)$ for the set of all strong cluster points of (x_n), defined as usual using the metric.

Throughout this paper, we now fix a separable Hadamard space (X, d) and two probability spaces, $(\Omega, \mathcal{F}, \mathbb{P})$ and $(E, \mathcal{E}, \mu)$, with $(E, \mathcal{E}, \mu)$ complete. All probabilistic notions such as measurability, random variables, almost sureness (a.s.), expectation, etc., are understood relative to the space $(\Omega, \mathcal{F}, \mathbb{P})$, if not stated otherwise. In particular, an X-valued random variable is a map $x : \Omega \to X$ which is measurable relative to $\mathcal{F}$ and the Borel σ-algebra $\mathcal{B}(X)$ of that space. We denote (conditional) expectations over $(\Omega, \mathcal{F}, \mathbb{P})$ by $\mathbb{E}$. All properties as well as (in-)equalities between random variables are understood to hold only almost surely, if not stated otherwise.

3. Main results

3.1. **Key lemmas.** As outlined in the introduction, the first main technical ingredient we need is a general result on the weak convergence of stochastic quasi-Fejér monotone sequences in metric spaces.

For that, and for the rest of the paper, we use the following notation (similar to [35]): Let $\mathsf{F} = (\mathsf{F}_n)$ be a given filtration of $\mathcal{F}$, that is a sequence of sub-σ-algebras of $\mathcal{F}$ such that $\mathsf{F}_n \subseteq \mathsf{F}_m$ for $n \leq m$. We write $\ell_+(\mathsf{F})$ for the set of sequences of non-negative real-valued random variables (e_n) that are adapted to the filtration, i.e. where e_n is F_n-measurable for all $n \in \mathbb{N}$. Further, we write $\ell^1_+(\mathsf{F})$ for the set of all $(e_n) \in \ell_+(\mathsf{F})$ such that $\sum_{n\in\mathbb{N}} e_n < +\infty$ a.s.

Lemma 3.1 (Proposition 4.3 in [35])**.** *Let X be a separable Hadamard space and let $Z \subseteq X$ be a nonempty closed subset of X and let $\phi : [0, +\infty) \to [0, +\infty)$ be strictly increasing such that $\lim_{t\to+\infty} \phi(t) = +\infty$. Let $\mathsf{F} = (\mathsf{F}_n)$ be a filtration and let (x_n) be a sequence of X-valued random variables adapted to F such that it is stochastically quasi-Fejér monotone w.r.t. Z, that is for any $z \in Z$ there are $(\chi_n(z)), (\eta_n(z)) \in \ell^1_+(\mathsf{F})$ and $(\theta_n(z)) \in \ell_+(\mathsf{F})$ such that for all $n \in \mathbb{N}$:*

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

Then we have the following assertions:

1. *$\sum_{n\in\mathbb{N}} \theta_n(z) < +\infty$ a.s. for all $z \in Z$.*
2. *(x_n) is bounded a.s.*
3. *There exists a set $\widetilde{\Omega}$ with $\mathbb{P}(\widetilde{\Omega}) = 1$ such that for all $\omega \in \widetilde{\Omega}$ and $z \in Z$, the sequence given by $d(x_n(\omega), z)$ converges.*
4. *If $\mathfrak{W}(x_n) \subseteq Z$ a.s., then (x_n) weakly converges a.s. to a Z-valued random variable.*
5. *If $\mathfrak{S}(x_n) \cap Z \neq \emptyset$ a.s., then (x_n) strongly converges a.s. to a Z-valued random variable.*

In fact, we later rely on a slightly strengthened variant of item (2) in Lemma 3.1 above, showing that $\sup_{n\in\mathbb{N}} d(x_n, z) < +\infty$ a.s. for some (any) $z \in Z$. Indeed, this follows immediately from item (3) above, since $d(x_n, z)$ converges a.s. for any $z \in Z$.

It is easy to see (see e.g. Section 4.1 in [32]) that for a general nonnegative stochastic process (X_n), being almost surely uniformly bounded in the sense of $\sup_{n\in\mathbb{N}} X_n < +\infty$ a.s. is equivalent to the statement that for any $\lambda > 0$, there exists an $N > 0$ such that $\mathbb{P}(\sup_{n\in\mathbb{N}} X_n > N) \leq \lambda$. Following [32], we call a function $\psi : (0,1) \to (0,\infty)$ that provides such an N in terms of $\lambda > 0$ a modulus of uniform boundedness. The following lemma just collects the fact that such a modulus exists for $X_n := d(x_n, z)$ in the setting of Lemma 3.1 for later use.

Lemma 3.2. *In the setting of Lemma 3.1, fix $z \in Z$. Then $\sup_{n\in\mathbb{N}} d(x_n, z) < +\infty$ a.s. In particular, there exists a function $\psi : (0,1) \to (0,\infty)$ such that*

$$\mathbb{P}\left(\sup_{n\in\mathbb{N}} d(x_n, z) > \psi(\lambda)\right) \leq \lambda$$

for any $\lambda \in (0,1)$.

In fact, such a modulus can be explicitly constructed from associated quantitative data witnessing the key assumptions of Lemma 3.1, and details for such constructions can be found in an abstract setting in [33].

The second main ingredient of the present work is a general result on the almost sure convergence of stochastic processes satisfying a Lipschitz-type condition controlled by a summable i.i.d. process.

For that, we rely on Kolmogorov's two-series theorem:

Lemma 3.3 (Theorem 12.2 in [40])**.** *Let (X_n) be an independent sequence of square-integrable random variables with means $\mathbb{E}[X_n] = 0$ and variances $\mathsf{Var}(X_n) = \sigma_n^2$ such that $\sum_{n\in\mathbb{N}} \sigma_n^2$ converges. Then $\sum_{n\in\mathbb{N}} X_n$ converges a.s.*

Our second main ingredient now takes the following form:

Lemma 3.4. *Let $(\lambda_n) \subseteq (0,\infty)$ with $\sum_{n\in\mathbb{N}} \lambda_n = +\infty$ and $\sum_{n\in\mathbb{N}} \lambda_n^2 < +\infty$. Assume (α_n) is a sequence of nonnegative square-integrable random variables which is i.i.d. with mean μ. Further, assume that γ_n satisfies $\sup_{n\in\mathbb{N}} \gamma_n < +\infty$ a.s. Let (β_n) be a sequence of nonnegative random variables such that $\sum_{n\in\mathbb{N}} \lambda_n \beta_n < +\infty$ a.s. and*

$$(+) \qquad \beta_{n+1} - \beta_n \leq \theta \lambda_n \gamma_n \alpha_n$$

a.s. for all $n \in \mathbb{N}$, and some constant $\theta > 0$. Then $\beta_n \to 0$ a.s.

This result can be considered to be a stochastic variant of a result due to Alber, Iusem and Solodov (see Proposition 2 in [1]), by which for sequences of nonnegative real numbers $(\alpha_n), (\beta_n)$ which satisfy $\sum_{n\in\mathbb{N}} \alpha_n = +\infty$ and $\sum_{n\in\mathbb{N}} \alpha_n \beta_n < +\infty$ as well as

$$\beta_{n+1} - \beta_n \leq \theta \alpha_n \text{ for all } n \in \mathbb{N}$$

for some constant $\theta > 0$, it holds that $\beta_n \to 0$ for $n \to \infty$. While our proof crucially uses many ideas from the associated proof of this result given in [1], we were not able to fully reduce our stochastic Lemma 3.4 to this deterministic result. We now turn to the proof itself.

Proof of Lemma 3.4. Consider the sequence of partial sums given by $E_n := \sum_{k=0}^{n} \lambda_k(\alpha_k - \mu)$. Then $\mathbb{E}[\lambda_n(\alpha_n - \mu)] = 0$ and the variance of $\lambda_n(\alpha_n - \mu)$ is given by $\mathbb{E}[\lambda_n^2(\alpha_n - \mu)^2] = \lambda_n^2 \mathbb{E}[(\alpha_n - \mu)^2]$. As each α_n is square-integrable and i.i.d., it holds that $\mathbb{E}[(\alpha_n - \mu)^2] < +\infty$ and the value is independent of $n \in \mathbb{N}$. Therefore $\sum_{n\in\mathbb{N}} \lambda_n^2 \mathbb{E}[(\alpha_n - \mu)^2] < +\infty$ since $\sum_{n\in\mathbb{N}} \lambda_n^2 < +\infty$. It follows by Kolmogorov's two-series theorem that E_n converges a.s.

In particular, for any $\delta > 0$, there exists an N_δ such that for all $n \geq m \geq N_\delta$, it holds that

$$\left| \sum_{k=m}^{n} \lambda_k (\alpha_k - \mu) \right| \leq \delta \text{ a.s.}$$

Hence, in particular the following holds a.s., say on a set Ω_0 of measure one: For any $\delta > 0$, there exists an N_δ such that for all $n \geq m \geq N_\delta$:

$$(\dagger) \qquad \sum_{k=m}^{n} \lambda_k \alpha_k \leq \mu \sum_{k=m}^{n} \lambda_k + \delta.$$

Suppose now that β_n does not converge to 0 a.s. Since we have $\sum_{n\in\mathbb{N}} \lambda_n \beta_n < +\infty$ a.s., as well as $\sum_{n\in\mathbb{N}} \lambda_n = +\infty$, we have $\liminf_{n\to\infty} \beta_n = 0$ a.s., say on a set of measure one Ω_1. By that however, we hence do not have $\limsup_{n\to\infty} \beta_n = 0$ a.s. Correspondingly, this fails on a set of positive measure Ω_2, say with $\mathbb{P}(\Omega_2) \geq p_0$ where $p_0 \in (0,1)$. Further, suppose that $(+)$ holds on a set of measure one Ω_3. Lastly, since $\sup_{n\in\mathbb{N}} \gamma_n < +\infty$ a.s., let $\psi : (0,1) \to (0,\infty)$ be a function such that $\mathbb{P}\left(\sup_{n\in\mathbb{N}} \gamma_n > \psi(\lambda)\right) \leq \lambda$. In particular, we hence have $\mathbb{P}\left(\sup_{n\in\mathbb{N}} \gamma_n \leq \psi(p_0/2)\right) \geq 1 - p_0/2$, and we denote the inner set by Ω_4.

Fix an $\omega \in \Omega_0 \cap \Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$. As $\omega \in \Omega_2$, there is a number ε such that $\limsup_{n\to\infty} \beta_n(\omega) > \varepsilon$, so choose a strictly increasing sequence (n_i) of indices such that $\beta_{n_i}(\omega) \geq \varepsilon$ for all $i \in \mathbb{N}$. As $\omega \in \Omega_1$, we have $\liminf_{n\to\infty} \beta_n(\omega) = 0$, there is a strictly increasing sequence (m_i) such that $\beta_{m_i} < \varepsilon/2$. Moreover, we can choose m_i such that $m_i < n_i$ and $\beta_n(\omega) \geq \varepsilon/2$ for all n such that $m_i < n \leq n_i$. Using $\omega \in \Omega_3$ and so $(+)$ as well as $\omega \in \Omega_4$, we then have

$$\begin{aligned} \frac{\varepsilon}{2} &\leq \beta_{n_i}(\omega) - \beta_{m_i}(\omega) = \sum_{k=m_i}^{n_i-1} \beta_{k+1}(\omega) - \beta_k(\omega) \\ &\leq \theta \sum_{k=m_i}^{n_i-1} \lambda_k \gamma_k(\omega) \alpha_k(\omega) \leq \theta \psi(p_0/2) \sum_{k=m_i}^{n_i-1} \lambda_k \alpha_k(\omega). \end{aligned}$$

For $i \in \mathbb{N}$ suitably large such that $n_i > m_i \geq N_\delta$ for $\delta = \varepsilon/(4\theta\psi(p_0/2))$, the estimate $(\dagger)$ yields

$$\frac{\varepsilon}{2} \leq \theta\psi(p_0/2)\mu \sum_{k=m_i}^{n_i-1} \lambda_k + \frac{\varepsilon}{4}.$$

Without loss of generality, we can now assume that $\mu > 0$, as if $\mu = 0$, then $\alpha_n = 0$ for all $n \in \mathbb{N}$ so that (β_n) would be constant by $(+)$, which combined with $\liminf_{n\to\infty} \beta_n = 0$ would contradict $\limsup_{n\to\infty} \beta_n > 0$. Thus, the above yields

$$\sum_{k=m_i}^{n_i-1} \lambda_k \geq \frac{\varepsilon}{4\theta\psi(p_0/2)\mu}.$$

and so we have

$$\begin{aligned} \sum_{k=m_i}^{n_i-1} \lambda_k \beta_k(\omega) &\geq \sum_{k=m_i+1}^{n_i-1} \lambda_k \beta_k(\omega) \geq \frac{\varepsilon}{2} \sum_{k=m_i+1}^{n_i-1} \lambda_k \\ &= \frac{\varepsilon}{2} \sum_{k=m_i}^{n_i-1} \lambda_k - \frac{\varepsilon}{2}\lambda_{m_i} \geq \frac{\varepsilon^2}{8\theta\psi(p_0/2)\mu} - \frac{\varepsilon}{2}\lambda_{m_i}. \end{aligned}$$

If $i \in \mathbb{N}$ is large enough, then $\lambda_{m_i} < \varepsilon/8\theta\psi(p_0/2)\mu$ since $\lambda_n \to 0$. Hence, for large enough $i \in \mathbb{N}$, we get

$$\sum_{k=m_i}^{n_i-1} \lambda_k \beta_k(\omega) \geq \frac{\varepsilon^2}{16\theta\psi(p_0/2)\mu},$$

so that summing over such large enough $i \in \mathbb{N}$ where the intervals spanning from m_i+1 to n_i-1 are disjoint, we get that $\sum_{n\in\mathbb{N}} \lambda_n\beta_n(\omega) = +\infty$, hence we have shown that $\sum_{n\in\mathbb{N}} \lambda_n\beta_n = +\infty$ on $\Omega' := \Omega_0 \cap \Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$. The set Ω' has positive measure as by the Fréchet inequalities, we have

$$\mathbb{P}(\Omega') \geq \sum_{i=0}^{4} \mathbb{P}(\Omega_i) - 4 = \mathbb{P}(\Omega_2) + \mathbb{P}(\Omega_4) - 1 \geq p_0 - p_0/2 = p_0/2 > 0.$$

This contradicts that $\sum_{n\in\mathbb{N}} \lambda_n\beta_n < +\infty$ holds a.s. □

3.2. **Derivation of the main result.** We now turn to the weak convergence result for the method (SPPA). The key result on the iteration given by (SPPA) is then a quasi-Fejér-type inequality given in Lemma 3.6. In fact, this inequality is already derived in passing in [9], but since the result only appears there in the context of a broad local compactness assumption we prove it here again for the benefit of the reader. Also, we give a slightly different argument than that given in [9].

For that, we first require the following property of the proximal map, which is however immediate from its definition:

Lemma 3.5 (see e.g. Lemma 2.2.23 in [8])**.** *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 key quasi-Fejér-type inequality then takes the form of the following Lemma 3.6. For that, we in the following set $\mathsf{F}_n := \sigma(\xi_1, \ldots, \xi_n)$ 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 3.6 (essentially Bačák [9])**.** *For any $y \in X$, there exists a constant $C_{y,p} > 0$ such that for all $n \in \mathbb{N}$:*

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

Proof. Given $n \in \mathbb{N}$ and $y \in X$, Lemma 3.5 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

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

where the third equality follows by independence of ξ_{n+1} and x_n, as well as the fact that ξ_{n+1} has distribution μ. Now, note that Lemma 3.5 together with (A2) yield

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

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

$$\begin{aligned} f(\xi_{n+1}, x_n) - f(\xi_{n+1}, x_{n+1}) &\leq L(\xi_{n+1})(1 + d(x_n, p))d(x_n, x_{n+1}) \\ &\leq 2\lambda_n L^2(\xi_{n+1})(1 + d(x_n, p))^2 \\ &\leq 4\lambda_n L^2(\xi_{n+1})(1 + d^2(x_n, p)). \end{aligned}$$

In particular, there is a constant $C_{y,p} > 0$ such that

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

so that we have

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

again using the independence of ξ_{n+1} to F_n and x_n. Combined, we get

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

as claimed. □

As in the context of Theorem 1.2, define

$$F(x) := \int f(e, x)\, d\mu(e)$$

and assume that F is proper. By Fatou's lemma, since f is a normal convex integrand, it follows that F is also lsc. Further F retains the convexity of f. Thereby, F is even weakly lower semicontinuous (weakly lsc), i.e.

$$\liminf_{n\to\infty} F(x_n) \geq F(x)$$

whenever $(x_n) \subseteq X$ and $x \in X$ such that $x_n \to^w x$, which immediately follows from the following result of Bačák:

Lemma 3.7 (Lemma 3.1 in [6])**.** *Let X be a Hadamard space. If $f : X \to (-\infty, +\infty]$ is lsc and convex, then it is weakly lsc.*

In the following, we now assume that F has a minimizer. We denote the set of minimizers by $\operatorname{argmin} F$ and the minimal value by $\min F$.

Combining the previous approach for weak convergence of the deterministic proximal point method in Hadamard spaces given in the seminal paper [6] with the Lemmas 3.1 and 3.4, we can then give the following proof of Theorem 1.2:

Proof of Theorem 1.2. Using Lemma 3.6 for some $z \in \operatorname{argmin} F$, we get

$$(\circ) \qquad \mathbb{E}_n[d^2(x_{n+1}, z)] \leq (1 + 2C_{z,p}\lambda_n^2 \underline{L})d^2(x_n, z) - 2\lambda_n(F(x_n) - \min F) + 2C_{z,p}\lambda_n^2 \underline{L}.$$

By the assumptions on the parameters (A1), we get $\sum_{n\in\mathbb{N}} \lambda_n^2 < +\infty$ so that the assumptions of Lemma 3.1, in particular $(*)$, are met. Item (1) of that lemma implies

$$\sum_{n\in\mathbb{N}} \lambda_n[F(x_n) - \min F] < +\infty \text{ a.s.}$$

By assumption (A2), we have

$$F(x) - F(y) \leq \int L(e)\, d\mu(e)(1 + d(x, p))d(x, y),$$

so that for $\beta_n := F(x_n) - \min F$, we have

$$\beta_{n+1} - \beta_n = F(x_{n+1}) - F(x_n) \leq \int L(e)\, d\mu(e)(1 + d(x_{n+1}, p))d(x_{n+1}, x_n).$$

Further recall from the proof of Lemma 3.6 that $d(x_{n+1}, x_n) \leq 2\lambda_n L(\xi_{n+1})(1+d(x_n,p))$, so that we get

$$\beta_{n+1} - \beta_n \leq \int L(e)\, d\mu(e) 2\lambda_n (1 + d(x_{n+1},p))(1+d(x_n,p))L(\xi_{n+1})$$

for all $n \in \mathbb{N}$. Note that by Lemma 3.2 together with inequality $(\circ)$ above, we get that $\sup_{n\in\mathbb{N}} d(x_n, z) < +\infty$ a.s. Hence, for $\gamma_n := (1 + d(x_{n+1},p))(1+d(x_n,p))$ we in particular have

$$\sup_{n\in\mathbb{N}} \gamma_n \leq \left(\sup_{n\in\mathbb{N}}(1+d(x_n,p))\right)^2 \leq \left(\sup_{n\in\mathbb{N}}(1+d(x_n,z)+d(z,p))\right)^2 < +\infty \text{ a.s.}$$

So, also using (A1), the assumptions of Lemma 3.4 are met with β_n and γ_n as above, as well as $\theta := 2\int L(e)\, d\mu(e)$ and $\alpha_n := L(\xi_{n+1})$, so that we obtain $\beta_n \to 0$ a.s., that is

$$F(x_n) \to \min F \text{ a.s.},$$

say on a set $\widehat{\Omega}$ with measure one.

We now show that on that set $\widehat{\Omega}$, we also have $\mathfrak{W}(x_n) \subseteq \operatorname{argmin} F$. For that, let $\omega \in \widehat{\Omega}$ and let $x(\omega)$ be a weak cluster point of $(x_n(\omega))$, i.e. $x_{n_k}(\omega) \to^w x(\omega)$ for some subsequence $(x_{n_k}(\omega))$. Since it follows from Lemma 3.7 that F is weakly lsc, as discussed before, and since $\omega \in \widehat{\Omega}$, we have

$$F(x(\omega)) \leq \liminf_{k\to\infty} F(x_{n_k}(\omega)) = \lim_{n\to\infty} F(x_n(\omega)) = \min F$$

and so $x(\omega) \in \operatorname{argmin} F$. As we thus have $\mathfrak{W}(x_n) \subseteq \operatorname{argmin} F$ a.s., item (4) of Lemma 3.1 implies that (x_n) a.s. weakly converges to an $\operatorname{argmin} F$-valued random variable. □

As a corollary, we obtain from Theorem 1.2 the weak convergence of the splitting proximal point method with random order given in [7], which in particular solves Problem 6.4 of Bačák [10].

Proof of Theorem 1.3. Consider the space $(E, \mathcal{E}, \mu)$, where $E = \{1, \dots, N\}$, $\mathcal{E} = 2^E$ and μ is the uniform distribution. On that space, define $f : E \times X \to (-\infty, +\infty]$ by $f(k, x) := f_k(x)$. The result then immediately follows from Theorem 1.2 under the corresponding assumptions (A1) and (A2), where the latter translates to the condition that $f_k(x) - f_k(y) \leq L(k)(1 + d(x,p))d(x,y)$ for all $k = 1, \dots, N$ and $x, y \in X$, for some $L(k) > 0$ (recall footnote 3). □

Acknowledgements: The author wants to thank Miroslav Bačák and Thomas Powell for helpful conversations on the topic of this paper. Further, he particularly wants to thank Morenikeji Neri not only for many insightful discussions on the topic of this paper, but also for proof reading various parts of this paper.

The author was originally stuck on the proof of Lemma 3.4, having essentially rederived the key parts of Proposition 2 from the work of Alber, Iusem and Solodov [1] (which was unknown to him at the time, but was later pointed out by Morenikeji Neri, for which he is grateful), but unsure how to control the involved series in a stochastic context and unsure about the precise formulation in general. A first version of Lemma 3.4 and its argument was streamlined and finalised only after a series of interactions with OpenAI's GPT-5.5 embedded into Microsoft Copilot, but then later modified and extended without it, especially after having learned of [1]. The author takes full accountability for the work. In particular, no other parts of this paper were worked on or impacted by the use of LLMs, or were written using them in any way.

References

[1] Ya.I. Alber, A.N. Iusem, and M.V. Solodov. On the projected subgradient method for nonsmooth convex optimization in a Hilbert space. *Mathematical Programming*, 81:23–35, 1998.

[2] 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.

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

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

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

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

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

[8] 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.

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

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

[11] M. Bačák, I. Searston, and B. Sims. Alternating projections in CAT(0) spaces. *Journal of Mathematical Analysis and Applications*, 385:599–607, 2012.

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

[13] 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.

[14] 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.

[15] H. Brézis and P.L. Lions. Produits infinis de resolvantes. *Israel Journal of Mathematics*, 29(4):329–345, 1978.

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

[17] 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.

[18] C. Castaing and M. Valadier. *Convex Analysis and Measurable Multifunctions*, volume 580 of *Lecture Notes in Mathematics*. Springer Berlin, Heidelberg, 1977.

[19] 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.

[20] S. Dhompongsa, W.A. Kirk, and B. Sims. Fixed points of uniformly Lipschitzian mappings. *Nonlinear Analysis. Theory, Methods & Applications*, 65:762–772, 2006.

[21] 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.

[22] 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.

[23] O. Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM Journal on Control and Optimization*, 29:403—-419, 1991.

[24] J. Jost. Equilibrium maps between metric spaces. *Calculus of Variations and Partial Differential Equations*, 2:173–204, 1994.

[25] J. Jost. Convex functionals and generalized harmonic maps into spaces of nonpositive curvature. *Commentarii Mathematici Helvetici*, 70:659–673, 1995.

[26] W.A. Kirk and B. Panyanak. A concept of convergence in geodesic spaces. *Nonlinear Analysis. Theory, Methods & Applications*, 68:3689–3696, 2008.

[27] A. Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer Cham, 3rd edition, 2020.

[28] B. Martinet. Régularisation dinéquations variationnelles par approximations successives. *Revue française d'informatique et de recherche opérationnelle*, 4:154–159, 1970.
[29] U. Mayer. Gradient flows on nonpositively curved metric spaces and harmonic maps. *Communications in Analysis and Geometry*, 6:199–253, 1998.
[30] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal of Optimization*, 19:1574–1609, 2009.
[31] M. Neri, N. Pischke, and T. Powell. An abstract effective convergence theorem for stochastic processes, with applications to stochastic approximation, 2026. Preprint, `https://arxiv.org/abs/2504.12922`.
[32] 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.
[33] M. Neri and T. Powell. A quantitative Robbins-Siegmund theorem. *The Annals of Applied Probability*, 36(1):636–651, 2026.
[34] B.J. Pettis. On integration in vector spaces. *Transactions of the American Mathematical Society*, 44:277–304, 1938.
[35] N. Pischke. On Busemann subgradient methods for stochastic minimization in Hadamard spaces, 2026. Preprint, `https://arxiv.org/abs/2602.08127`.
[36] N. Pischke and T. Powell. Convergence guarantees for stochastic algorithms solving non-unique problems in metric spaces, 2026. Preprint, `https://arxiv.org/abs/2605.06129`.
[37] 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.
[38] R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal of Control and Optimization*, 14:877–898, 1976.
[39] E.K. Ryu and S. Boyd. Stochastic Proximal Iteration: A Non-Asymptotic Improvement upon Stochastic Gradient Descent. working draft, accessed 2026, `https://ernestryu.com/papers/spi.pdf`.
[40] D. Williams. *Probability with martingales*. Cambridge University Press, 1991.
[41] 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.