

# ON BUSEMANN SUBGRADIENT METHODS FOR STOCHASTIC MINIMIZATION IN HADAMARD 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 study the recently introduced Busemann subgradient method due to Goodwin, Lewis, Nicolae and López-Acedo, extending it to minimize the mean of a stochastic function over general Hadamard spaces. We prove a strong convergence theorem under a local compactness assumption and further prove weak ergodic convergence of the method over Hadamard spaces satisfying condition  $(\bar{Q}_4)$ , a slight extension of the  $(Q_4)$  condition of Kirk and Payanak, which in particular includes Hilbert spaces,  $\mathbb{R}$ -trees and spaces of constant curvature. The proof is based on a general (weak) convergence theorem for stochastic processes in Hadamard spaces which confine to a stochastic variant of quasi-Fejér monotonicity, together with a nonlinear variant of Pettis' theorem, which are of independent interest. Lastly, we provide a strong convergence result under a strong convexity assumption, and in that case in particular derive explicit rates of convergence.

**Keywords:** Subgradient method, Busemann function, stochastic approximation, weak convergence, Hadamard spaces, proof mining

**MSC2020 Classification:** 47J25, 90C15, 90C25, 62L20, 03F10

## 1. INTRODUCTION

**1.1. Background and motivation.** One of the most general and productive formulations of (stochastic) approximation is the problem of minimizing a convex integral function, that is solving the problem

$$(P) \quad \min_{x \in C} \int f(e, x) d\mu(e),$$

for a given normal convex integrand  $f : E \times C \rightarrow \mathbb{R}$  (see [56]) on a complete probability space  $(E, \mathcal{E}, \mu)$  and some target set  $C \subseteq X$  in a suitable space  $X$ , say a Hilbert or Banach space. There are various prevalent modern tools for approaching this problem, among them being stochastic variants of the proximal point algorithm and of projected subgradient methods, and we refer to [14, 15, 45], among many others, for various such discussions.

Indeed, most of these methods already are concerned with the particularly important and motivating special case of minimizing a finite sum of convex functions  $f_i : C \rightarrow \mathbb{R}$ , i.e.

$$(P)_f \quad \min_{x \in C} \sum_{i=1}^m f_i(x).$$

These problems gain further relevance if considered outside of linear contexts such as Hilbert or Banach spaces. Concretely, latest since the extensive developments of machine learning in recent years, where optimization over nonlinear spaces such as manifolds plays a key role (we refer e.g. to the discussion in [59]), extensions of these tools from (stochastic) convex analysis to nonlinear contexts are of high practical relevance. Further, this relevance naturally transcends

the realm of spaces with differentiable structure such as manifolds, as illustrated by e.g. the Billera-Holmes-Vogtmann tree space [17] prominently used in phylogenetics.

The present paper is concerned with a method recently introduced by Goodwin, Lewis, López-Acedo and Nicolae [32] to solve the above minimization problem  $(P)_f$  in the context of the general class of geodesic metric spaces with nonpositive curvature, as introduced in the work of Aleksandrov [1]. These spaces, often called CAT(0) spaces after the work of Gromov [33] and Hadamard spaces if complete, uniformly cover examples such as Hilbert spaces,  $\mathbb{R}$ -trees and Hadamard manifolds (i.e. complete simply connected Riemannian manifolds of nonpositive sectional curvature) and the Billera-Holmes-Vogtmann tree space mentioned before, as well as further involved examples. As such, they have continuously been a focus of attention for extending tools from convex analysis on linear spaces to nonlinear contexts, most notably recently in other related work by Goodwin, Lewis, López-Acedo and Nicolae [31, 41, 42]. We refer to [2, 18] for a comprehensive overview of geodesic and CAT(0) spaces and further refer to [9] for a shorter treatment focused on aspects of convex analysis and optimization.

Concretely, the work [32] provides a projected subgradient method for the problem  $(P)_f$ . However, lack of linear structure makes constructions such as subgradients, which naturally rely on duality theory, complicated in general geodesic metric spaces such as Hadamard spaces. In that way, the approach of [32] diverges from usual subgradient methods by making particularly novel use of the boundary cone  $CX^\infty$  of a Hadamard space  $X$  and associated Busemann functions, and other advanced geometric tools from Hadamard spaces, to define a novel type of subgradient for the associated functions.

Indeed, as shown in [32], this resulting notion of a Busemann subgradient, which in Euclidean space coincides with the usual notion of subgradients, supports a broad theory and in particular allows for the derivation and effective analysis of the following stochastic projected subgradient method:<sup>1</sup> Given a starting point  $x_0 \in X$ , an i.i.d. sequence  $(i_n)$  of random selections distributed uniformly over  $\{1, \dots, m\}$  and a sequence of step-sizes  $(t_n)$  of positive reals with  $\sum_{n \in \mathbb{N}} t_n = +\infty$  and  $\sum_{n \in \mathbb{N}} t_n^2 < +\infty$ , one defines the iteration

$$(SB_0) \quad x_{n+1} := P_C(r_{x_n, \xi_n}(s_n t_n))$$

where  $[\xi_n, s_n] = \text{Busemann}_{f_{i_n}}(x_n) \in CX^\infty$  represents a Busemann subgradient of  $f_{i_n}$  at  $x_n$ , chosen using an oracle function  $\text{Busemann}$ , and  $r_{x_n, \xi_n}(s_n t_n)$  represents the point reached by following the geodesic ray starting from  $x_n$  with direction  $\xi_n$  and speed  $s_n$ , and with step size  $t_n$ .<sup>2</sup>

For this stochastic method, they in particular establish the following result:

**Theorem 1.1** (Theorem 6.2 in [32]). *Let  $X$  be a Hadamard space with the geodesic extension property and at least two points and let  $C \subseteq X$  be nonempty, closed and convex. Let  $F(x) := \sum_{i=1}^m f_i(x)$  be given, where each  $f_i : C \rightarrow \mathbb{R}$  is Busemann subdifferentiable, and such that  $\text{argmin} F \neq \emptyset$ . Assume further that there is some  $L \geq 0$  such that every Busemann subgradient  $[\xi, s]$  of any  $f_i$  at any point in  $C$  satisfies  $s \leq L$ .*

*Let  $(x_n)$  be the sequence generated by  $(SB_0)$  with  $(t_n)$  and  $(i_n)$ , where  $(t_n) \subseteq (0, +\infty)$  is such that<sup>3</sup>  $\sum_{n \in \mathbb{N}} t_n = +\infty$  and  $\sum_{n \in \mathbb{N}} t_n^2 < +\infty$ , and  $(i_n)$  is a sequence of i.i.d. random variables*

<sup>1</sup>The work [32] also considers a non-stochastic incremental variant of this Busemann subgradient method, but we will not be concerned with such deterministic methods in the present paper.

<sup>2</sup>The above is just intended as a sketch. In particular, all these currently undefined notions will be detailed in Section 5 later on.

<sup>3</sup>In fact, they already provide this result under the assumption that  $\sum_{k=0}^n t_k^2 / \sum_{k=0}^n t_k \rightarrow 0$ .

distributed uniformly over  $\{1, \dots, m\}$ . Then

$$\mathbb{E} \left[ \min_{i=0, \dots, n} F(x_i) \right] \rightarrow \min F.$$

Further, they also obtain the following quantitative result: For the special case that  $C$  has diameter bounded by  $D > 0$ , Theorem 6.2 in [32] further establishes the non-asymptotic guarantee

$$\mathbb{E} \left[ \min_{i=0, \dots, n} F(x_i) \right] \leq \frac{2(1 + \log(3))mLD}{\sqrt{n+2}}$$

for all  $n \geq 2$ , using the parameter sequence  $t_n := \frac{D}{mL\sqrt{n+1}}$ .

While the above result presents a (quantitative) approximation result for the method  $(\mathbf{SB}_0)$ , the possibilities of Busemann subgradients and methods such as  $(\mathbf{SB}_0)$  are only beginning to be explored (see e.g. also the recent [25]), and many questions remain, such as convergence behavior of the whole sequence, rates under regularity conditions, and extensions for solving general stochastic minimization problems. We address precisely those questions in the present paper.

**1.2. Main results and related work.** In this paper, we adapt the previous Busemann subgradient method introduced by Goodwin, Lewis, López-Acedo and Nicolae in [32] to solve general stochastic minimization problems as in (P), and moreover provide a more detailed study on its asymptotic behavior.

Concretely, let  $(\Omega, \mathcal{F}, \mathbb{P})$  and  $(E, \mathcal{E}, \mu)$  be probability spaces, with  $(E, \mathcal{E}, \mu)$  complete, and let  $X$  be a separable Hadamard space with the geodesic extension property and at least two points.<sup>4</sup> Further, fix a closed convex nonempty subset  $C \subseteq X$  together with a functional  $f : E \times C \rightarrow \mathbb{R}$ .

We now want to solve (P), that is we want to minimize  $F(x) := \int f(e, x) d\mu(e)$  over  $C$ , assuming that  $F$  is proper and that such a minimum actually exists, i.e. that  $\operatorname{argmin} F \neq \emptyset$ .

For that, we assume that  $f$  has the following properties: At first, we assume that  $f$  actually has Busemann subgradients and further is suitably measurable, i.e.

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

Further, we impose a Lipschitz condition on the Busemann subgradients, i.e.

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

These two assumptions on  $f$  are derived from [32] (see Assumption A therein, and recall Theorem 1.1), suitably extended to the present, more general context. However, both are also natural if seen in the broader context of (stochastic) subgradient methods, where measurability and a Lipschitz condition on the subgradients are among the most basic common assumptions (see e.g. [45]). As mentioned before, further details on all involved objects will be given in Section 5 later on.

Motivated by [32], we consider the following stochastic Busemann subgradient method: Define

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

---

<sup>4</sup>These assumption pertain to the existence of boundary points and the well-definedness of various notions, as discussed in detail later.

given a starting point  $x_0 \in C$  and sequences  $(t_n)$  of positive reals as well as  $(\zeta_{n+1})$  of random variables  $\Omega \rightarrow E$ , where  $[\xi_n, s_n] = \text{Busemann}_f(\zeta_{n+1}, x_n)$  represents a Busemann subgradient of  $f(\zeta_{n+1}, \cdot)$  at  $x_n$ , chosen using an oracle function  $\text{Busemann}_f$  for  $f$ . Crucially, we assume that we utilize an oracle which preserves measurability, namely

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

While such an oracle also already appears in [32] as discussed before, measurability thereof or of the associated iteration is not discussed, which however will occupy us here. Indeed, the question whether such a map always exists seems to be rather subtle, but it can be guaranteed in proper spaces, as will be discussed later.

Lastly, for the parameters we assume that

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

While not immediately obvious at first, we will later show that the assumptions (A1) – (A3) in fact suffice to guarantee that  $x_n$  is measurable for any  $n \in \mathbb{N}$ .

It should be noted that, as mentioned before, (P) is a generalisation of the minimization problem for finite sums of functions  $(P)_f$  as studied in [32], which is reobtained by considering finite measure spaces. Further, the above method (SB) subsumes the method  $(SB_0)$  in the same way, so that the present results also pertain to the method studied in [32].

Towards an asymptotic analysis of the method (SB), we will show the following results. At first, we show that the sequence strongly converges in the presence of a local compactness assumption.

**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 locally compact Hadamard space with the geodesic extension property and at least two points and a closed convex nonempty subset  $C \subseteq X$ . Let  $f : E \times C \rightarrow \mathbb{R}$  be a function with properties (A1) – (A3) as above. Define  $F(x) := \int f(e, x) d\mu(e)$  and assume  $\text{argmin} F \neq \emptyset$ . Let  $(x_n)$  be the iteration given by (SB) with  $(t_n)$  and  $(\zeta_{n+1})$ , and assume (Par).*

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

Lifting the local compactness assumption, the best we in general can expect is a weak convergence result. However, lifting such a result to a stochastic context like the above is highly nontrivial. Indeed, for related methods like the proximal point algorithm, it is a fundamental open problem whether its stochastic analog converges weakly even over separable Hilbert spaces (see [11]) and this also seems to be true for subgradient methods such as the above.

We instead show an ergodic convergence result, i.e. a convergence result for the ergodic average sequence  $(\bar{x}_n)$  defined recursively by

$$\bar{x}_0 := x_0 \text{ and } \bar{x}_{n+1} := \frac{\bar{t}_n}{\bar{t}_{n+1}} \bar{x}_n \oplus \frac{t_{n+1}}{\bar{t}_{n+1}} x_{n+1}$$

where  $\bar{t}_n := \sum_{k=0}^n t_k$ , and writing  $(1 - \lambda)x \oplus \lambda y$  for the point  $\gamma(\lambda d(x, y))$  on the unique geodesic  $\gamma : [0, d(x, y)] \rightarrow X$  joining  $x$  and  $y$ .

This result however comes at the expense of a narrower class of spaces, being derived for separable Hadamard spaces satisfying condition  $(\bar{Q}_4)$  as introduced in [35], a slight extension of the  $(Q_4)$  condition of Kirk and Panyanak [37], which in particular includes Hilbert spaces,  $\mathbb{R}$ -trees and spaces of constant curvature, but is known to exclude certain gluings of Hadamard spaces (see [30]).

**Theorem 1.3.** *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 satisfying  $(\overline{Q}_4)$  with the geodesic extension property and at least two points and a closed convex nonempty subset  $C \subseteq X$ . Let  $f : E \times C \rightarrow \mathbb{R}$  be a function with properties (A1) – (A3) as above. Define  $F(x) := \int f(e, x) d\mu(e)$  and assume  $\operatorname{argmin} F \neq \emptyset$ . Let  $(x_n)$  be the iteration given by (SB) with  $(t_n)$  and  $(\zeta_{n+1})$ , and assume (Par).*

*Then  $(\bar{x}_n)$ , defined with  $(t_n)$ , a.s. weakly converges to an  $\operatorname{argmin} F$ -valued random variable.*

The question whether the method (SB) satisfies a weak ergodic convergence result in all separable Hadamard spaces remains an open question.

Our arguments for both theorems above are based on a general (weak) convergence theorem for stochastic processes in Hadamard spaces which confine to a stochastic variant of quasi-Fejér monotonicity (see Proposition 4.3 later on), which is also of independent interest. Fejér monotonicity is a fundamental notion in modern (convex) analysis and optimization, with many if not most iterative procedures confining to a variant thereof (see e.g. [21, 22] and also [6]). In particular, Fejér monotonicity has played a crucial role in the weak convergence proof for the deterministic proximal point method in Hadamard spaces given in [7], which relied on a previous (deterministic) weak convergence result for such sequences given in [12]. Stochastic variants of (quasi-)Fejér monotonicity go back at least to the work of Ermol'ev [27, 28, 29], but the notion was subsequently further refined in the seminal work of Combettes and Pesquet [23] (see also [24]), where a general and abstract (weak) convergence theorem for such sequences is established over separable Hilbert spaces on which we have modeled our result on. Indeed, the approach to our central weak convergence result given in Proposition 4.3 is a rather immediate synthesis of the stochastic work [23] set in Hilbert spaces and the deterministic work of Bačák, Searston and Sims [12] set in a metric context. The extensions for weak ergodic convergence over Hilbert spaces are due to Passty [50] and related arguments over Hadamard spaces, involving condition  $(\overline{Q}_4)$ , have recently been given in the work of Khatibzadeh and Moosavi [36], on which we base our ergodic convergence result (see Proposition 4.5 later on). In the course of these results, we further rely on a nonlinear variant of (the consequence of) Pettis' theorem [51], stating that a weak limit of a sequence of measurable functions taking values in a separable Hadamard space is again measurable (see Proposition 3.1 later on). To our knowledge, this result is also new to the literature of Hadamard spaces.

Further considerations on stochastic quasi-Fejér monotonicity in a metric context can e.g. be found in the recent work [47], which while phrased in the context of strong stochastic regularity conditions (which among other things induce strong convergence) nevertheless further illustrates the range of different methods that immediately fall under this paradigm. Indeed, even though we only consider the above method (SB) in the present paper, we think that the general convergence result formulated here will be of use for the convergence analysis of further methods from stochastic convex optimization over nonlinear spaces.

One such method which we in particular want to highlight derives from the work of Bianchi [16], in particular its recent extension to a metric setting of nonpositive curvature given in [52]. Concretely, the work [16] studies a corresponding stochastic proximal point algorithm for stochastically perturbed monotone operators over separable Hilbert spaces. Only almost sure weak ergodic convergence is known in Hilbert spaces (see [16]), which can be improved to almost sure strong convergence of the original iteration in the context of a strong monotonicity assumption. This latter result was extended to a metric setting in [52], and it remains an interesting open problem if also the almost sure weak ergodic convergence proven in [16] can be lifted to the metric setting, where we hope that the present considerations will be helpful, at least over spaces satisfying condition  $(\overline{Q}_4)$ . While such a method might offer additional difficulties due to the use of general monotone vector fields, a more intermediate goal along

the same vein that we want to highlight is the stochastic proximal point method as discussed in Hadamard spaces by Bačák [10], generalising related work on the minimization of a finite sum of functions as discussed in the well-known work [8] (in a similar way as the present paper generalises [32]). Indeed, as mentioned before, it remains a fundamental open question whether already that method almost surely weakly converges even in separable Hilbert spaces (recall [11]), but a related result on almost sure weak ergodic convergence, at least over spaces satisfying condition  $(\overline{Q}_4)$ , could hopefully be derived via the present results (in a similar way as  $(\overline{Q}_4)$  has already been used in [36] to provide such a result for the deterministic proximal splitting method considered in [8]).

Lastly, we will prove the following result on rates under a strong convexity assumption:

**Theorem 1.4.** *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 with the geodesic extension property and at least two points and a closed convex nonempty subset  $C \subseteq X$ . Let  $f : E \times C \rightarrow \mathbb{R}$  be a function with properties (A1) – (A3) as above and assume additionally that  $f(e, \cdot)$  is strongly convex with parameter  $\alpha(e) > 0$ , i.e.*

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

for any geodesic  $\gamma : [0, 1] \rightarrow C$  and any  $t \in [0, 1]$ , where additionally  $\underline{\alpha} := \int \alpha d\mu > 0$ . Define  $F(x) := \int f(e, x) d\mu(e)$  and assume  $\operatorname{argmin} F \neq \emptyset$ . Let  $(x_n)$  be the iteration given by (SB) with  $(t_n)$  and  $(\zeta_{n+1})$ , and assume (Par).

Then  $(x_n)$  a.s. and in mean strongly converges to the unique minimizer  $x^*$  of  $F$ . Moreover, the following rates of convergence apply: Let  $\chi : (0, \infty) \rightarrow \mathbb{N}$  and  $\theta : \mathbb{N} \times (0, \infty) \rightarrow \mathbb{N}$  be such that

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

Let  $T > \sum_{n=0}^{\infty} t_n^2$ . Lastly, let  $b > 0$  be such that  $b > d^2(x_0, x^*)$ . Then

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

with rate  $\rho(\varepsilon) := \theta(\chi(\varepsilon/2L^2), 8(b + L^2T)/\varepsilon\underline{\alpha})$  and

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

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

The proof essentially relies on a similar approach as the results given by Neri, Powell and the author in [47] for the quantitative asymptotic behavior for general stochastic processes subscribing to a monotonicity condition.<sup>5</sup>

Indeed, the above result can in fact be substantially extended to other regularity conditions on the mean function  $\underline{f}(x) := \int f(e, x) d\mu(e)$  beyond strong convexity, including generalized weak sharp minima, as will be discussed in more detail in forthcoming work by the author and Thomas Powell on the general theory of such regularity conditions and their relation to quantitative convergence analyses for stochastic processes, extending [47].

<sup>5</sup>The results from [47], and likewise the present results regarding rates of convergence, have been obtained using the logic-based methodology of *proof mining* [39, 40], and are part of a series of recent applications of these methods to probability theory and stochastic optimization [46, 47, 48, 49, 53]. As common in proof mining however, this paper avoids any reference to mathematical logic.

## 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 in this paper we refer to [2, 9, 18] for a comprehensive overview of geodesic metric spaces and their properties, in particular to [9] for aspects of (stochastic) optimization. Further, beyond the results indicated, we refer to e.g. [38] for a standard textbook on probability theory.

Let  $(X, d)$  be a metric space. A geodesic is an isometry  $\gamma : [0, l] \rightarrow X$ . We say that it joins  $x = \gamma(0)$  and  $y = \gamma(l)$  (where necessarily  $l = d(x, y)$ ).  $X$  is called (uniquely) geodesic if every two points are joined by a (unique) geodesic. We call the image of a geodesic such as the above a geodesic segment, and in the uniquely geodesic case denote it by  $[x, y]$ . A geodesic ray is an isometry  $r : [0, \infty) \rightarrow X$ , and we say that  $r$  issues from  $r(0)$ . A space  $X$  has the geodesic extension property if for all  $x \neq y \in X$ , there is a ray  $r : [0, \infty) \rightarrow X$  issuing from  $x$  such that  $r(t) = y$  for some  $t > 0$ .

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  $\gamma : [0, l] \rightarrow X$  (that is, an extension of the so-called Bruhat-Tits CN-inequality [19] to geodesics). Any CAT(0) space is uniquely geodesic and a complete CAT(0) space is called a Hadamard space. Further, a CAT(0) space has nonpositive curvature in the sense of Busemann (see e.g. [9, 18]) and hence the metric is jointly convex in the sense that  $d(\gamma(tl), \eta(tm))$  is convex on  $t \in [0, 1]$  for all geodesics  $\gamma : [0, l] \rightarrow X$  and  $\eta : [0, m] \rightarrow X$  (see e.g. Proposition 1.1.5 in [9]).

Weak convergence in CAT(0) spaces goes back to the work of Jost [34] and is often called  $\Delta$ -convergence following the work of Kirk and Panyanak [37] (we refer in particular to the discussion in [7] on that matter). We define weak convergence here as follows (see e.g. [9]):

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

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

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 [26]). Further, each bounded sequence has a weak cluster point (see e.g. Proposition 3.1.2 in [9]).

We say that a bounded sequence  $(x_n)$  weakly converges to  $x \in X$ , written  $x_n \rightarrow^w x$ , 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  $x_{n_k} \rightarrow^w x$ .

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  $(x_n)$ , the latter defined as usual using the metric.

A set  $C \subseteq X$  is called convex if the geodesics between any two points in  $C$  are contained in  $C$ , and we call a function  $g : C \rightarrow \mathbb{R}$  convex if

$$g(\gamma(tl)) \leq (1-t)g(\gamma(0)) + tg(\gamma(l))$$

for any geodesic  $\gamma : [0, l] \rightarrow C$  and any  $t \in [0, 1]$ . Further, we call  $g$  lower-semicontinuous (lsc) if

$$\liminf_{n \rightarrow \infty} g(x_n) \geq g(x)$$

whenever  $(x_n) \subseteq C$  is a sequence such that  $x_n \rightarrow x$ . Crucially, every convex lsc function on a Hadamard space is also weakly lower-semicontinuous (weakly lsc, see [7]), i.e. it satisfies the above inequality even if  $(x_n) \subseteq C$  is a sequence such that  $x_n \rightarrow^w x$ . A particular example of a convex function on a Hadamard spaces is the Busemann function  $b_r : X \rightarrow \mathbb{R}$  associated to a ray  $r$ , defined by

$$b_r(x) := \lim_{t \rightarrow \infty} (d(x, r(t)) - t),$$

which is nonexpansive (1-Lipschitz), convex and satisfies  $b_r(r(0)) = 0$  (see Example 2.2.10 in [9]).

Throughout this paper, if not stated otherwise, we let  $(X, d)$  be a separable Hadamard space and  $(\Omega, \mathcal{F}, \mathbb{P})$  as well as  $(E, \mathcal{E}, \mu)$  be two probability spaces, 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  $x : \Omega \rightarrow X$  which is measurable relative to  $\mathcal{F}$  and the Borel  $\sigma$ -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. Sometimes we are working on general measurable spaces  $(T, \mathcal{T})$  or even measure spaces  $(T, \mathcal{T}, \tau)$ . In such cases, we are always quite explicit on regarding measurability notions, and in particular use the expression “almost everywhere” (a.e.) instead.

### 3. A NONLINEAR VARIANT OF PETTIS’ THEOREM

In this section, we prove the following measurability result for weak limits in Hadamard spaces:

**Proposition 3.1.** *Let  $(T, \mathcal{T}, \tau)$  be a finite measure space and let  $X$  be a separable Hadamard space. Let  $(x_n)$  be a sequence of  $\mathcal{T}/\mathcal{B}(X)$ -measurable functions and let  $x : T \rightarrow X$  be a function such that the  $(x_n)$  are bounded and  $x_n \rightarrow^w x$  almost everywhere. Then there is a  $\mathcal{T}/\mathcal{B}(X)$ -measurable  $y$  which is equal to  $x$  almost everywhere.*

*If  $T$  is complete, then  $x$  itself is measurable. Further,  $T$  can be  $\sigma$ -finite in that case.*

In separable Hilbert spaces, the above result is a consequence (see Corollary 1.13 in [51]) of the seminal theorem of Pettis on weakly measurable functions (see Theorem 1.1 in [51]). Indeed, compared to [51], we are here confined to finite measure spaces.

Our proof is rather simple, relying only on a few results from measurable selection theory. We generally refer to [3, 4, 20] for further background on that area. Given a complete separable metric space  $X$  and a measurable space  $(T, \mathcal{T})$ , call a set-valued map  $\varphi : T \rightarrow 2^X$  graph measurable if

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

Over complete  $\sigma$ -finite measure spaces, and if  $\varphi$  has nonempty closed images, this is equivalent to the (weak) measurability of  $\varphi : T \rightarrow 2^X$ , that is that

$$\varphi^{-1}(C) := \{t \in T \mid \varphi(t) \cap C \neq \emptyset\} \in \mathcal{T}$$

for all open sets  $C \subseteq X$  (see e.g. Theorem 8.1.4 in [4]).

The key result we need is the following on the graph measurability of minimizing maps.

**Lemma 3.2.** *Let  $(T, \mathcal{T})$  be a measurable space and let  $X$  be a complete separable metric space. Further, let  $g : T \times X \rightarrow \mathbb{R}$  be a Carathéodory function. Then the functions*

$$m(t) := \inf_{x \in X} g(t, x) \text{ and } M(t) := \operatorname{argmin}_{x \in X} g(t, x)$$

*are measurable and graph measurable, respectively.*



This result will then be combined with the following selection theorem of Aumann [5].

**Theorem 3.3** ([5], see also Corollary 18.27 in [3]). *Let  $(T, \mathcal{T}, \tau)$  be a finite measure space and let  $X$  be a complete separable metric space. Let  $\varphi : T \rightarrow 2^X$  be graph measurable with nonempty values. Then there is a measurable function  $x : T \rightarrow X$  such that  $x(t) \in \varphi(t)$  almost everywhere.*

Lemma 3.2 commonly appears in the literature under the assumption that the measurable space is complete and  $\sigma$ -finite for some measure, with the (in that context equivalent) conclusion of measurability of  $M$  (see e.g. Theorem 8.2.11 in [4]). However, we want to dispense of this completeness assumption here, as it would later (unnecessarily) require us to assume completeness of  $(\Omega, \mathcal{F}, \mathbb{P})$ . In that way, we rederive that result here which requires some care on measurability and on the assumptions.

To prove Lemma 3.2, we first need the following result on Carathéodory functions:

**Lemma 3.4** (Corollary 18.8 in [3]). *Let  $g : T \times X \rightarrow (-\infty, +\infty]$  be Carathéodory function. Then*

$$\varphi(t) := \{x \in X \mid g(t, x) = 0\}$$

*is graph measurable.*

We can now provide the proof of Lemma 3.2:

*Proof of Lemma 3.2.* Using that  $X$  is separable, fix a countable dense set  $(z_n)$ . Then

$$m(t) = \inf_{x \in X} g(t, x) = \inf_{n \in \mathbb{N}} g(t, z_n)$$

by continuity of  $g(t, \cdot)$ . Thus,  $m$  is measurable as every  $g(\cdot, z_n)$  is measurable. As  $g$  is a Carathéodory function, so is  $g'(t, x) = g(t, x) - m(t)$ . Using Lemma 3.4, we then get that

$$M(t) = \{y \in X \mid g(t, y) = m(t)\} = \{y \in X \mid g'(t, y) = 0\}$$

is graph measurable. □

The key observation is then that  $r(x_n(t), y)$  defines a Carathéodory function, leading to the following proof of our nonlinear variant of Pettis's theorem:

*Proof of Proposition 3.1.* Fix a set  $T_0$  of measure zero such that  $(x_n(t))$  is bounded and  $x_n(t) \rightarrow^w x(t)$  for all  $t \in T_0^c$ . First note that  $(t, y) \mapsto r(x_n(t), y)$  for  $(t, y) \in T_0^c \times X$  is a Carathéodory function. Indeed, recall that by definition  $r(x_n(t), y) = \limsup_{n \rightarrow \infty} d^2(x_n(t), y)$ . First fix  $t \in T_0^c$ . As each  $d^2(x_n(t), \cdot)$  is Lipschitz continuous with the same constant, it follows that  $r(x_n(t), \cdot)$  is locally Lipschitz and hence continuous (see e.g. Example 2.2.8 in [9]). Now fix  $y \in X$ . As each  $d^2(x_n(\cdot), y)$  is measurable, we immediately get that  $r(x_n(\cdot), y)$  is measurable.

Lemma 3.2 then yields that

$$\varphi'(t) := \operatorname{argmin}_{y \in X} r(x_n(t), y),$$

defined for  $t \in T_0^c$ , is graph measurable. As  $(x_n(t))$  is bounded, we have by the existence of asymptotic centers that  $\varphi'(t)$  is nonempty for all  $t \in T_0^c$  (note that they are also closed by continuity of  $r(x_n(t), \cdot)$ ). Now define  $\varphi(t) := \varphi'(t)$  for  $t \in T_0^c$  and  $\varphi(t) := X$  otherwise. Then

$$\operatorname{gra}(\varphi) = \operatorname{gra}(\varphi') \cup (T_0 \times X) \in \mathcal{T} \otimes \mathcal{B}(X)$$

so that  $\varphi$  is still graph measurable, with  $\varphi(t)$  nonempty (and closed) for all  $t \in T$ . Using Theorem 3.3, there exists a measurable function  $y$  such that  $y(t) \in \varphi(t)$  almost everywhere, say on  $T_1^c$  with  $T_1$  of measure zero. For  $t \in T_0^c \cap T_1^c$ , by the uniqueness of asymptotic centers, we get that  $x(t) = y(t)$ . Thus  $x = y$  almost everywhere

If  $(T, \mathcal{T}, \tau)$  is complete and  $\sigma$ -finite, then one can apply the Kuratowski–Ryll–Nardzewski selection theorem (see e.g. Theorem 8.1.3 in [4]) in place of Theorem 3.3, which yields a measurable function  $y$  with  $x = y$  almost everywhere as before. As the space is now complete, we get that  $x$  is measurable as well.  $\square$

#### 4. STOCHASTIC QUASI-FEJÉR MONOTONICITY AND WEAK (ERGODIC) CONVERGENCE

As outlined in the introduction, the main technical convergence result we present is a general proposition on the weak and strong convergence of stochastic quasi-Fejér monotone sequences in metric spaces.

For that, we first introduce some convenient notation: Given a filtration  $\mathbf{F} = (\mathbf{F}_n)$  of  $\mathcal{F}$ , that is a sequence of sub- $\sigma$ -algebras of  $\mathcal{F}$  such that  $\mathbf{F}_n \subseteq \mathbf{F}_m$  for  $n \leq m$ , we write  $\ell_+(\mathbf{F})$  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  $\mathbf{F}_n$ -measurable for all  $n \in \mathbb{N}$ . Further, we write  $\ell_+^1(\mathbf{F})$  for the set of all  $(e_n) \in \ell_+(\mathbf{F})$  such that  $\sum_{n \in \mathbb{N}} e_n < +\infty$  a.s.

Now, fixing a solution set  $Z \subseteq X$  and a filtration  $\mathbf{F} = (\mathbf{F}_n)$ , stochastic quasi-Fejér monotonicity for a sequence  $(x_n)$  adapted to  $\mathbf{F}$  in our context takes the form of requiring

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

for all  $z \in Z$  and  $n \in \mathbb{N}$ , where  $\phi$  is a suitable perturbation function and  $(\chi_n(z)), (\eta_n(z)) \in \ell_+^1(\mathbf{F})$  as well as  $(\theta_n(z)) \in \ell_+(\mathbf{F})$  are error sequences which might depend on the point  $z$  in question.

As such, our notion is a direct lift of the rather general stochastic quasi-Fejér monotonicity considered in separable Hilbert spaces in the seminal work of Combettes and Pesquet [23] (see equation (2.5) therein and recall the further references given in the introduction), to the metric setting.

We can then obtain the following result on weak and strong convergence based on weak and strong cluster points, which is itself an extension of a corresponding result in separable Hilbert spaces given by Combettes and Pesquet [23] (see Proposition 2.3 therein). As mentioned in the introduction, as such our result is in particular an immediate synthesis of Proposition 2.3 from [23] (especially concerning items (1) – (3) of the following Proposition 4.3, which are highly derivative of it and essentially are a direct lift to the metric setting) and Proposition 3.3 from [12] on weak convergence of Fejér monotone sequences in Hadamard spaces. Indeed, the key aspect of the convergence of Fejér monotone sequences in Hadamard spaces is the following result essentially derived (in the proof of) Proposition 3.3 in [12], and it immediately allows us to derive items (4) and (5) of the following Proposition 4.3. For self-containedness, we rederive it here, with a slightly different argument.

**Lemma 4.1** (essentially Proposition 3.3 in [12]). *Let  $X$  be a Hadamard space and let  $Z \subseteq X$  be a non-empty closed subset of  $X$  and  $(x_n) \subseteq X$  be a given sequence such that  $d(x_n, z)$  converges for all  $z \in Z$ .*

- (1) *If  $\mathfrak{W}(x_n) \subseteq Z$ , then  $(x_n)$  weakly converges a.s. to some point in  $Z$ .*
- (2) *If  $\mathfrak{S}(x_n) \cap Z \neq \emptyset$ , then  $(x_n)$  strongly converges to some point in  $Z$ .*

*Proof.* Note that  $(x_n)$  is bounded and therefore has a weak cluster point. For the first item, it hence suffices to show that  $(x_n)$  has a unique weak cluster point. To that effect, let  $(x_{n_k})$  and  $(x_{m_k})$  be subsequences of  $(x_n)$  with asymptotic centers  $c_1$  and  $c_2$ , respectively. As  $\mathfrak{W}(x_n) \subseteq Z$ , we have  $c_1, c_2 \in Z$ . Therefore  $d(x_n, c_1)$  and  $d(x_n, c_2)$  and so also  $d^2(x_n, c_1)$  and  $d^2(x_n, c_2)$

converge. Now, assume w.l.o.g. that  $r(x_{n_k}) \leq r(x_{m_k})$ . We then have

$$\begin{aligned} r(x_{m_k}) &\geq r(x_{n_k}) = r(x_{n_k}, c_1) \\ &= \limsup_{k \rightarrow \infty} d^2(x_{n_k}, c_1) \\ &= \lim_{n \rightarrow \infty} d^2(x_n, c_1) \\ &= \limsup_{k \rightarrow \infty} d^2(x_{m_k}, c_1) \\ &= r(x_{m_k}, c_1) \geq r(x_{m_k}), \end{aligned}$$

where the third (and fourth) line follows from the fact that  $d^2(x_n, c_1)$  (and so every subsequence of it) converges. Thus in particular  $r(x_{m_k}) = r(x_{m_k}, c_1)$  and so  $c_1$  is also an asymptotic center of  $(x_{m_k})$ , next to  $c_2$ . As asymptotic centers are unique, we get  $c_1 = c_2$  which completes the argument.

For the strong convergence result, assume  $\mathfrak{S}(x_n) \cap Z \neq \emptyset$ . Concretely, let  $x \in Z$  be a strong accumulation point of  $(x_n)$ , i.e.

$$\liminf_{n \rightarrow \infty} d(x_n, x) = 0.$$

As  $d(x_n, x)$  converges, we get  $\lim_{n \rightarrow \infty} d(x_n, x) = 0$ , which completes the proof.  $\square$

Like in [23], and any other result on stochastic quasi-Fejér monotonicity for that matter, the key statistical ingredient of our convergence result is the seminal Robbins-Siegmund theorem on almost-supermartingale convergence:

**Lemma 4.2** (Theorem 1 in [55]). *Let  $\mathbf{F} = (\mathbf{F}_n)$  be a filtration, and let  $(\alpha_n), (\theta_n) \in \ell_+(\mathbf{F})$  as well as  $(\eta_n), (\chi_n) \in \ell_+^1(\mathbf{F})$  be such that*

$$\mathbb{E}[\alpha_{n+1} \mid \mathbf{F}_n] \leq (1 + \chi_n)\alpha_n - \theta_n + \eta_n \text{ a.s.}$$

*for any  $n \in \mathbb{N}$ . Then  $(\alpha_n)$  a.s. converges to a nonnegative real-valued random variable and  $(\theta_n) \in \ell_+^1(\mathbf{F})$ .*

**Proposition 4.3** (extending Proposition 2.3 in [23] and Proposition 3.3 in [12]). *Let  $X$  be a separable Hadamard space and let  $Z \subseteq X$  be a nonempty closed subset of  $X$  and let  $\phi : [0, +\infty) \rightarrow [0, +\infty)$  be strictly increasing such that  $\lim_{t \rightarrow +\infty} \phi(t) = +\infty$ . Let  $\mathbf{F} = (\mathbf{F}_n)$  be a filtration and let  $(x_n)$  be a sequence of  $X$ -valued random variables adapted to  $\mathbf{F}$  such that it is stochastically quasi-Fejér monotone w.r.t.  $Z$ , that is for any  $z \in Z$  there are  $(\chi_n(z)), (\eta_n(z)) \in \ell_+^1(\mathbf{F})$  and  $(\theta_n(z)) \in \ell_+(\mathbf{F})$  such that for all  $n \in \mathbb{N}$ :*

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

*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  $\tilde{\Omega}$  with  $\mathbb{P}(\tilde{\Omega}) = 1$  such that for all  $\omega \in \tilde{\Omega}$  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.

*Proof.* Ad (1): For any  $z \in Z$ , the above inequality  $(*)$  and Lemma 4.2 immediately yield  $\sum_{n \in \mathbb{N}} \theta_n(z) < +\infty$  a.s.

Ad (2): Let  $z \in Z$  be arbitrary. The above inequality  $(*)$  and Lemma 4.2 immediately yield that  $\phi(d(x_n, z))$  converges a.s. to some nonnegative real-valued random variable  $\alpha_z$ . As

$\lim_{t \rightarrow +\infty} \phi(t) = +\infty$ , we get that the sequence  $d(x_n, z)$  is bounded a.s., which is hence true also for  $(x_n)$ .

Ad (3): Let us first note that given any  $z \in Z$ ,  $d(x_n, z)$  converges a.s., namely at every point where the sequence  $d(x_n, z)$  is bounded. For suppose not, then there is a point  $\omega \in \Omega$  such that the sequence  $d(x_n(\omega), z)$  is bounded and where there are subsequences  $d(x_{n_k}(\omega), z)$  and  $d(x_{m_k}(\omega), z)$  with respective distinct limits  $\tau_1(\omega)$  and  $\tau_2(\omega)$ . W.l.o.g. suppose  $\tau_2(\omega) > \tau_1(\omega)$  and take  $\varepsilon(\omega) \in (0, (\tau_2(\omega) - \tau_1(\omega))/2)$ . For large enough  $n$ , we will have

$$d(x_{n_k}(\omega), z) \leq \tau_1(\omega) + \varepsilon(\omega) < \tau_2(\omega) - \varepsilon(\omega) \leq d(x_{m_k}(\omega), z).$$

As  $\phi$  is strictly increasing, we immediately get

$$\phi(d(x_{n_k}(\omega), z)) \leq \phi(\tau_1(\omega) + \varepsilon(\omega)) < \phi(\tau_2(\omega) - \varepsilon(\omega)) \leq \phi(d(x_{m_k}(\omega), z)).$$

Taking the limit as  $n \rightarrow \infty$  gives

$$\alpha_z(\omega) \leq \phi(\tau_1(\omega) + \varepsilon(\omega)) < \phi(\tau_2(\omega) - \varepsilon(\omega)) \leq \alpha_z(\omega)$$

for the random variable  $\alpha_z$  from item (2), which is a contradiction.

Now, we move on to actually showing item (3). As  $X$  is separable, also  $Z$  is separable. Fix a countable dense set  $Z_0$  in  $Z$ . By the previous, for every  $z \in Z_0$  there exists a set  $\Omega_z$  of measure one such that  $d(x_n(\omega), z)$  converges for any  $\omega \in \Omega_z$ . Define  $\tilde{\Omega} := \bigcap_{z \in Z_0} \Omega_z$ . Clearly  $\tilde{\Omega}$  has measure one. Let now  $z \in Z$  and  $\omega \in \tilde{\Omega}$  be arbitrary. As  $Z_0$  is dense in  $Z$ , there exists a sequence  $(z_k)$  in  $Z_0$  such that  $z_k \rightarrow z$ . As  $\omega \in \tilde{\Omega} \subseteq \Omega_{z_k}$ , we in particular have that  $d(x_n(\omega), z_k)$  converges for any  $k \in \mathbb{N}$  and for any  $\omega \in \tilde{\Omega}$ , say with limit  $\tau_k(\omega)$ . For any  $k \in \mathbb{N}$  and any such  $\omega \in \tilde{\Omega}$ , we thus have

$$-d(z_k, z) \leq d(x_n(\omega), z) - d(x_n(\omega), z_k) \leq d(z_k, z)$$

using triangle inequality, so that we get

$$\begin{aligned} -d(z_k, z) &\leq \liminf_{n \rightarrow \infty} d(x_n(\omega), z) - \lim_{n \rightarrow \infty} d(x_n(\omega), z_k) \\ &= \liminf_{n \rightarrow \infty} d(x_n(\omega), z) - \tau_k(\omega) \\ &\leq \limsup_{n \rightarrow \infty} d(x_n(\omega), z) - \tau_k(\omega) \\ &= \limsup_{n \rightarrow \infty} d(x_n(\omega), z) - \lim_{n \rightarrow \infty} d(x_n(\omega), z_k) \\ &\leq d(z_k, z). \end{aligned}$$

Taking the limit as  $k \rightarrow \infty$  yields that  $d(x_n(\omega), z)$  converges with

$$\lim_{n \rightarrow \infty} d(x_n(\omega), z) = \lim_{k \rightarrow \infty} \tau_k(\omega).$$

Ad (4): Let  $\tilde{\Omega}$  be the set with measure one from item (3) and suppose that  $\mathfrak{W}(x_n) \subseteq Z$  a.s., say  $\mathfrak{W}(x_n(\omega)) \subseteq Z$  for all  $\omega \in \tilde{\Omega}$  for  $\tilde{\Omega}$  with measure one. Lemma 4.1 now yields that for any  $\omega \in \tilde{\Omega} \cap \tilde{\Omega}$ , the sequence  $(x_n(\omega))$  converges weakly to an element  $x(\omega)$  of  $Z$ . This then yields the claim as follows:  $\tilde{\Omega} \cap \tilde{\Omega}$  still has measure one. By Proposition 3.1, we get that there is a measurable  $y$  with  $x = y$  a.s. Thus  $y$  is a.s.  $Z$ -valued and so w.l.o.g. we can assume that  $y$  is  $Z$ -valued. Clearly,  $(x_n)$  converges weakly to  $y$  almost surely.

Ad (5): In a similar way as item (4) above, using in particular item (3), Lemma 4.1 now yields that  $x_n(\omega)$  strongly converges to some point  $x(\omega) \in Z$  on a set of measure one. The measurability of  $x$  follows immediately as  $X$  is separable.  $\square$

Under suitable restrictions on the space, we can now also provide a stochastic weak ergodic convergence result. We establish this result not for all separable Hadamard spaces, but only for those that satisfy the condition  $(\overline{Q}_4)$ , i.e. for points  $x, y, p, q \in X$  and any point  $m \in [x, y]$ :

$$d(p, x) \leq d(x, q) \text{ and } d(p, y) \leq d(y, q) \text{ implies } d(p, m) \leq d(m, q).$$

Equivalently, and more crucially for the present paper, the condition expresses that

$$(\overline{Q}_4) \quad F(p, q) := \{x \in X \mid d(p, x) \leq d(x, q)\} \text{ is convex}$$

for all  $p, q \in X$ . This condition was introduced by Kakavandi in [35] for studying topologies capturing weak convergence (which was recently, and by quite a different method, substantially extended by Lytchak and Petrunin [43]) and is a slight strengthening of the condition  $(Q_4)$  introduced by Kirk and Panyanak in [37] for establishing convergence of midpoint sequences in geodesic segments (which was subsequently weakened in [30]). Examples of spaces that satisfy  $(\overline{Q}_4)$  include Hilbert spaces,  $\mathbb{R}$ -trees, Hadamard spaces with constant curvature and their closed convex subsets (see [35] and see also [30] for related results for the condition  $(Q_4)$ ). However, not all Hadamard spaces even satisfy the condition  $(Q_4)$  as shown in [30]. In particular, any CAT(0) gluing space containing two spaces with constant but different curvatures does not.

In the recent paper [36], this condition was used to establish the weak ergodic convergence of a nonlinear variant, first studied by Bačák [8], of the proximal splitting method originally considered by Passty [50].

While presented as tailored to that particular iteration, the approach given in [36] essentially establishes the following general convergence principle, which can be seen a nonlinear variant of Passty's ergodic convergence condition given as Lemma 1 in [50]. However, note that compared to [50], the present result requires the asymptotic regularity condition  $d(x_n, x_{n+1}) \rightarrow 0$  on the sequence  $(x_n)$  in addition to the usual condition that all weak cluster points of the weighted averages are solutions. Even though this requirement is slightly cumbersome, especially in a stochastic context, both properties can often be established simultaneously, as will also be the case in this paper.

As the result only appears implicitly in [36], we reprove it here for the readers convenience:

**Lemma 4.4** (essentially the proof of Theorem 4.2 in [36]). *Let  $X$  be a Hadamard space satisfying  $(\overline{Q}_4)$  and let  $Z \subseteq X$  be a non-empty closed convex subset of  $X$ . Let  $(x_n) \subseteq X$  be a given sequence such that  $d(x_n, z)$  converges for all  $z \in Z$ . Further, let  $(t_n) \subseteq (0, +\infty)$  be such that  $\sum_{n \in \mathbb{N}} t_n = +\infty$  and let  $(\bar{x}_n)$  be the associated sequence of weighted averages. If  $\mathfrak{W}(\bar{x}_n) \subseteq Z$  and  $d(x_n, x_{n+1}) \rightarrow 0$ , then  $(\bar{x}_n)$  weakly converges a.s. to some point in  $Z$ .*

*Proof.* The sequence  $(x_n)$  and hence also  $(\bar{x}_n)$  is bounded and therefore has a weak cluster point. It thus suffices to show that  $(\bar{x}_n)$  has no further weak cluster point. Suppose that  $c_1, c_2$  are two weak cluster points and let  $(\bar{x}_{n_k})$  and  $(\bar{x}_{m_k})$  be subsequences of  $(\bar{x}_n)$  with asymptotic centers  $c_1$  and  $c_2$ , respectively. As  $\mathfrak{W}(\bar{x}_n) \subseteq Z$ , we have  $c_1, c_2 \in Z$ . Therefore  $d(x_n, c_1)$  and  $d(x_n, c_2)$  and so also  $d^2(x_n, c_1)$  and  $d^2(x_n, c_2)$  converge. Assume w.l.o.g. that  $\lim_{n \rightarrow \infty} d^2(x_n, c_1) \leq \lim_{n \rightarrow \infty} d^2(x_n, c_2)$ . We now distinguish two cases, namely  $\lim_{n \rightarrow \infty} d^2(x_n, c_1) < \lim_{n \rightarrow \infty} d^2(x_n, c_2)$  and  $\lim_{n \rightarrow \infty} d^2(x_n, c_1) = \lim_{n \rightarrow \infty} d^2(x_n, c_2)$ .

We begin with the former. There hence exists an  $n_0 \geq 1$  with  $d^2(x_n, c_1) < d^2(x_n, c_2)$  and so in particular  $x_n \in F(c_1, c_2)$  for all  $n \geq n_0$ . Given  $i \in \mathbb{N}$ , denote the weighted average of  $(x_{n+i})$  with  $(t_n)$  by  $\bar{x}_m^i$ . Convexity of  $F(c_1, c_2)$ , that is  $(\overline{Q}_4)$ , yields that  $\bar{x}_m^{n_0} \in F(c_1, c_2)$ . We now show that  $d(\bar{x}_m, \bar{x}_m^{n_0}) \rightarrow 0$ . By triangle inequality, we have  $d(\bar{x}_m, \bar{x}_m^{n_0}) \leq \sum_{i=0}^{n_0-1} d(\bar{x}_m^i, \bar{x}_m^{i+1})$

it is enough to show that  $d(\bar{x}_m^i, \bar{x}_m^{i+1}) \rightarrow 0$  for all  $i \in \mathbb{N}$ . For that, we show

$$d(\bar{x}_m^i, \bar{x}_m^{i+1}) \leq \sum_{k=0}^m \frac{t_k}{\bar{t}_m} d(x_{i+k}, x_{i+k+1}),$$

which in turn follows by induction on  $m$ . Concretely, we have  $d(\bar{x}_0^i, \bar{x}_0^{i+1}) = d(x_i, x_{i+1})$  and, using joint convexity of the metric in Hadamard spaces, further

$$\begin{aligned} d(\bar{x}_{m+1}^i, \bar{x}_{m+1}^{i+1}) &= d\left(\frac{\bar{t}_m}{\bar{t}_{m+1}} \bar{x}_m^i \oplus \frac{t_{m+1}}{\bar{t}_{m+1}} x_{i+m+1}, \frac{\bar{t}_m}{\bar{t}_{m+1}} \bar{x}_m^{i+1} \oplus \frac{t_{m+1}}{\bar{t}_{m+1}} x_{i+m+2}\right) \\ &\leq \frac{\bar{t}_m}{\bar{t}_{m+1}} d(\bar{x}_m^i, \bar{x}_m^{i+1}) + \frac{t_{m+1}}{\bar{t}_{m+1}} d(x_{i+m+1}, x_{i+m+2}) \\ &\leq \frac{\bar{t}_m}{\bar{t}_{m+1}} \sum_{k=0}^m \frac{t_k}{\bar{t}_m} d(x_{i+k}, x_{i+k+1}) + \frac{t_{m+1}}{\bar{t}_{m+1}} d(x_{i+m+1}, x_{i+m+2}) \\ &= \sum_{k=0}^{m+1} \frac{t_k}{\bar{t}_{m+1}} d(x_{i+k}, x_{i+k+1}), \end{aligned}$$

with  $\oplus$  defined as before. As we have  $d(x_n, x_{n+1}) \rightarrow 0$  and  $\sum_{n \in \mathbb{N}} t_n = +\infty$ , this yields that  $d(\bar{x}_m^i, \bar{x}_m^{i+1}) \rightarrow 0$  by the convergence of weighted averages (e.g. the Silverman-Toeplitz theorem). As  $\bar{x}_{m_k} \rightarrow^w c_2$ , we therefore get  $\bar{x}_{m_k}^{n_0} \rightarrow^w c_2$ . As  $F(c_1, c_2)$  is convex and closed, it is weakly closed (see e.g. Lemma 3.2.1 in [9]) and we hence have  $c_2 \in F(c_1, c_2)$ . So in particular  $d(c_2, c_1) \leq d(c_2, c_2) = 0$ , that is  $c_1 = c_2$ .

Suppose now that  $\lim_{n \rightarrow \infty} d^2(x_n, c_1) = \lim_{n \rightarrow \infty} d^2(x_n, c_2) = L$  and suppose for a contradiction that  $c_1 \neq c_2$ . Let  $\gamma$  be the unique geodesic joining  $c_1$  and  $c_2$ , and write  $\gamma'(\lambda) := \gamma(\lambda d(c_1, c_2))$ . As  $Z$  is convex, we get  $\gamma'(\lambda) \in Z$  for all  $\lambda \in [0, 1]$ . By assumption, we thus have that  $d^2(x_n, \gamma'(\lambda))$  converges. The strong convexity of  $d^2$  yields that

$$d^2(x_n, \gamma'(\lambda)) \leq (1 - \lambda)d^2(x_n, c_1) + \lambda d^2(x_n, c_2) - \lambda(1 - \lambda)d^2(c_1, c_2).$$

Taking the limit as  $n \rightarrow \infty$  thus yields

$$\lim_{n \rightarrow \infty} d^2(x_n, \gamma'(\lambda)) \leq L - \lambda(1 - \lambda)d^2(c_1, c_2) < L$$

for all  $0 < \lambda < 1$ , as  $d^2(c_1, c_2) > 0$ . Therefore, one has  $\lim_{n \rightarrow \infty} d^2(x_n, \gamma'(\lambda)) < \lim_{n \rightarrow \infty} d^2(x_n, c_2)$  and hence can show as in the first part that  $c_2 \in F(\gamma'(\lambda), c_2)$ . This gives

$$d(c_2, \gamma'(\lambda)) \leq d(c_2, c_2) = 0$$

for all  $0 < \lambda < 1$ , so that sending  $\lambda$  to 0 yields  $c_1 = c_2$ , which is a contradiction.  $\square$

A direct lift to the stochastic is then the following proposition:

**Proposition 4.5.** *Under the assumptions of Proposition 4.3, suppose additionally that  $X$  satisfies  $(\bar{Q}_4)$ , that  $Z$  is convex and that  $\mathfrak{W}(\bar{x}_n) \subseteq Z$  as well as  $d(x_n, x_{n+1}) \rightarrow 0$  a.s. for a sequence of weighted averages  $(\bar{x}_n)$  defined via some  $(t_n) \subseteq (0, +\infty)$  with  $\sum_{n \in \mathbb{N}} t_n = +\infty$ .*

*Then  $(\bar{x}_n)$  weakly converges a.s. to a  $Z$ -valued random variable.*

*Proof.* The previous Proposition 4.3, item (3), yields that there exists a set  $\tilde{\Omega}$  with measure one such that for all  $\omega \in \tilde{\Omega}$  and  $z \in Z$ , the sequence given by  $d(x_n(\omega), z)$  converges. By assumption, we have  $\mathfrak{W}(\bar{x}_n(\omega)) \subseteq Z$  as well as  $d(x_n(\omega), x_{n+1}(\omega)) \rightarrow 0$  for all  $\omega \in \tilde{\Omega}$ , and  $\hat{\Omega}$  with measure one. In particular, we therefore have  $\mathfrak{W}(\bar{x}_n(\omega)) \subseteq Z$ ,  $d(x_n(\omega), x_{n+1}(\omega)) \rightarrow 0$  and  $d(x_n(\omega), z)$  converges for all  $z \in Z$ , for all  $\omega \in \hat{\Omega} \cap \tilde{\Omega}$ , and  $\hat{\Omega} \cap \tilde{\Omega}$  still has measure one.

Lemma 4.4 applied for each  $\omega \in \widehat{\Omega} \cap \widetilde{\Omega}$  individually yields that  $(\bar{x}_n(\omega))$  weakly converges to some  $x(\omega) \in Z$ . As each  $\bar{x}_n$  is measurable, Proposition 3.1 yields that there is a measurable  $y$  with  $x = y$  a.s. Thus  $y$  is a.s.  $Z$ -valued and so w.l.o.g. we can assume that  $y$  is  $Z$ -valued. Clearly,  $(\bar{x}_n)$  converges weakly to  $y$  almost surely.  $\square$

## 5. INITIAL CONSIDERATIONS ON THE METHOD (SB)

After these rather general considerations on the convergence of stochastic processes in Hadamard spaces, we now turn to the method (SB). The present section is concerned with preliminary observations on that method which we derive before turning to our convergence results, in particular the measurability of the iteration (SB) under the assumptions (A1) – (A3) and the existence of measurable maps satisfying (A3). In particular, from now on we assume, if not stated otherwise, that  $X$  is a separable Hadamard space with the geodesic extension property and at least two points.

**5.1. Boundary cones and Busemann subgradients.** We first recall the basic definitions surrounding the method (SB). We begin with the notions of boundary  $X^\infty$  and boundary cone  $CX^\infty$ , where we refer to Chapter II.8 in [18] for further discussions (see also [32]).

Recall that the boundary of  $X$  at infinity  $X^\infty$  is the set of all equivalence classes of rays in  $X$  under the equivalence relation of being asymptotic, where two rays  $r, r'$  are called asymptotic if

$$d(r(t), r'(t)) \leq K \text{ for all } t \geq 0$$

for some constant  $K \geq 0$ . As  $X$  has the geodesic extension property and contains at least two points,  $X^\infty$  is nonempty. We say that a ray  $r$  has direction  $\xi \in X^\infty$  if it belongs to the respective equivalence class. In particular, given any  $x \in X$  and  $\xi \in X^\infty$ , there exists a unique ray issuing at  $x$  with direction  $\xi$ . Following [32], we consider  $X^\infty$  to be endowed with the cone topology (see Definition II.8.6 in [18]). This makes the space first-countable, and moreover the topology is completely specified by its convergent sequences, with the key result in that vein discussed later in Lemma 5.2. Indeed, we refer to [32] for discussions on the particular benefits that the cone topology offers for optimisation purposes, compared to the perhaps more common topology induced by the angular metric.

The boundary cone  $CX^\infty$  is now the usual Euclidean cone over  $X^\infty$ , i.e.  $CX^\infty$  is the quotient of  $X^\infty \times [0, \infty)$  under the equivalence relation

$$(\xi, s) \sim (\xi', s') \text{ if, and only if, } s = s' = 0 \text{ or } (\xi, s) = (\xi', s').$$

Topologically,  $X^\infty \times [0, \infty)$  is endowed with the product of the cone topology on  $X^\infty$  and the usual metric topology on  $[0, \infty)$ , and  $CX^\infty$  is endowed with the resulting quotient topology, also referred to as the cone topology. We denote an equivalence class of  $(\xi, s) \in X^\infty \times [0, \infty)$  in  $CX^\infty$  by  $[\xi, s]$ , and write  $[0]$  for the equivalence class of  $(\xi, 0)$  for some/any  $\xi \in X^\infty$ . Again, convergence in  $CX^\infty$  has a particularly useful characterization which we discuss later in Lemma 5.3. Examples of  $X^\infty$  and  $CX^\infty$  for various spaces  $X$  are discussed in [32].

We follow [32] and define the “pairing” function  $\langle \cdot, \cdot \rangle : X \times CX^\infty \rightarrow \mathbb{R}$  defined by

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

where we wrote  $b_\xi$  for the Busemann function corresponding to the (unique) ray  $r_{\bar{x}, \xi}$  with direction  $\xi$  and some arbitrary but fixed origin  $\bar{x}$ . Crucially, this pairing is continuous and  $\langle \cdot, [\xi, s] \rangle$  is convex,  $s$ -Lipschitz and positively homogeneous in  $s$ , meaning that  $\langle \cdot, [\xi, \alpha s] \rangle = \alpha \langle \cdot, [\xi, s] \rangle$  for all  $[\xi, s] \in CX^\infty$  and  $\alpha \geq 0$  (see Proposition 2.3 in [32]).

We now turn to Busemann subgradients. Given a function  $g : C \rightarrow \mathbb{R}$ , a Busemann subgradient of  $g$  at  $x \in C$  is a point  $[\xi, s] \in CX^\infty$  such that

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

We call  $g$  Busemann subdifferentiable if  $g$  has a Busemann subgradient at every  $x \in C$ . As discussed in [32],  $x$  minimizes  $g$  if, and only if,  $[0]$  is a Busemann subgradient of  $g$  at  $x$ . The work [32] provides quite a broad theory for such subgradients, including a chain rule, notions of conjugacy and geometry of level sets. We also refer to [32] for various examples of Busemann subdifferentiable functions and further discussion on relation to subgradients in the sense of [41], defined using tangent cones of Alexandrov spaces.

As for properties of Busemann subdifferentiable maps, we here mention the following: First, a Busemann subdifferentiable  $g$  is lsc (see p. 11 in [32]) and if  $C$  is convex, then  $g$  is also convex (see Proposition 3.3 in [32]). Further, if  $g$  is Busemann subdifferentiable such that each point  $x \in C$  admits a subgradient  $[\xi, s]$  with  $s \leq L$ , then  $g$  is  $L$ -Lipschitz, with the converse being true if  $C$  is open (see Proposition 4.3 in [32]).

**5.2. Measurability of (SB).** We now show the measurability of the main iteration under the assumptions (A1) – (A3). For that, we recall the following folklore result on Carathéodory functions.

**Lemma 5.1** (folklore, see e.g. Lemma 8.2.6 in [4]). *Let  $X, Y$  be complete separable metric spaces and let  $(T, \mathcal{T})$  be a measurable space. If  $g : T \times X \rightarrow Y$  is a Carathéodory function, then it is  $\mathcal{T} \otimes \mathcal{B}(X)/\mathcal{B}(Y)$  measurable.*

As a function  $f : E \times C \rightarrow \mathbb{R}$  satisfying (A1) and (A2) is measurable in its left argument by (A1) and continuous in its right by (A2) (recall the above discussion by which  $f(e, \cdot)$  is in particular  $L$ -Lipschitz), it is thereby a Carathéodory function and so is  $\mathcal{E} \otimes \mathcal{B}(C)$ -measurable.

Further, we require the following two characterizations of convergence in  $X^\infty$  as well as  $CX^\infty$ , both established in [32] (see also the discussion therein for previous works mentioning this and related results).

**Lemma 5.2** (Proposition 2.2 in [32]). *Fix  $x \in X$ . For any  $(\xi_n) \subseteq X^\infty$  and  $\xi \in X^\infty$ , we have  $\xi_n \rightarrow \xi$  if, and only if,*

$$r_{x, \xi_n}(\delta) \rightarrow r_{x, \xi}(\delta)$$

for all  $\delta > 0$ .

**Lemma 5.3** (Lemma 2.4 in [32]). *For any  $([\xi_n, s_n]) \subseteq CX^\infty$  and  $[\xi, s] \in CX^\infty$ , if  $[\xi_n, s_n] \rightarrow [\xi, s]$  in  $CX^\infty$ , then*

$$\begin{cases} s_n \rightarrow s & \text{if } s = 0, \\ s_n \rightarrow s \text{ and } \xi_n \rightarrow \xi & \text{if } s > 0. \end{cases}$$

Immediately, we can now see that rays are continuous in both direction and origin.

**Lemma 5.4.** *The ray  $r_{x, \xi}(s)$ , seen as a function  $X \times CX^\infty \rightarrow X$ , is continuous both in  $x \in X$  and  $[\xi, s] \in CX^\infty$ .*

*Proof.* As geodesic rays depend continuously on its origin,  $r_{x, \xi}(s)$  is continuous in  $x$ . Now let  $[\xi_n, s_n] \rightarrow [\xi, s]$  in  $CX^\infty$ . We have

$$\begin{aligned} d(r_{x, \xi_n}(s_n), r_{x, \xi}(s)) &\leq d(r_{x, \xi_n}(s_n), r_{x, \xi_n}(s)) + d(r_{x, \xi_n}(s), r_{x, \xi}(s)) \\ &= |s_n - s| + d(r_{x, \xi_n}(s), r_{x, \xi}(s)) \end{aligned}$$



using the fact that the rays are geodesics. Lemma 5.3 now implies that either  $s_n \rightarrow s$  if  $s = 0$ , or  $s_n \rightarrow s$  and  $\xi_n \rightarrow \xi$  if  $s > 0$ . In the former case, i.e. if  $s = 0$ , we have  $d(r_{x,\xi_n}(s), r_{x,\xi}(s)) = d(x, x) = 0$  so that  $d(r_{x,\xi_n}(s_n), r_{x,\xi}(s)) \leq |s_n - s| \rightarrow 0$ . In the second case, i.e. if  $s > 0$ , we get that  $\xi_n \rightarrow \xi$  implies that  $r_{x,\xi_n}(\delta) \rightarrow r_{x,\xi}(\delta)$  for all  $\delta > 0$  using Lemma 5.2. In particular  $d(r_{x,\xi_n}(s), r_{x,\xi}(s)) \rightarrow 0$  as  $s > 0$ . As  $|s_n - s| \rightarrow 0$  as before, we get  $d(r_{x,\xi_n}(s_n), r_{x,\xi}(s)) \rightarrow 0$  all the same.  $\square$

We now establish measurability of the sequence (SB).

**Lemma 5.5.** *Let  $X$  be a separable Hadamard space with the geodesic extension property and at least two points, and let  $C \subseteq X$  be closed convex nonempty. Let  $f : E \times C \rightarrow \mathbb{R}$  be a given functional and assume (A1) – (A3).*

*Then the sequence  $(x_n)$  defined by (SB) with  $(t_n)$  and  $(\zeta_{n+1})$ , satisfying (Par), is measurable.*

*Proof.* By Lemma 5.4, we have that  $r_{x,\xi}(s)$  is continuous in  $x \in X$  and  $[\xi, s] \in CX^\infty$ . In particular, it is  $\mathcal{B}(CX^\infty)/\mathcal{B}(X)$  measurable in  $[\xi, s]$ . Thereby,  $r_{x,\xi}(s)$  is a Carathéodory function and so is  $\mathcal{B}(X) \otimes \mathcal{B}(CX^\infty)/\mathcal{B}(X)$  measurable by Lemma 5.1 (taking  $(CX^\infty, \mathcal{B}(CX^\infty))$  as  $(T, \mathcal{T})$ ). As  $P_C$  is continuous, it is measurable. Clearly  $x_0 \in X$  is measurable and so, combined with assumption (A3), we get by induction that

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

is measurable.  $\square$

**5.3. Measurable Busemann oracles.** It remains a question whether (A1) and (A2) suffice to guarantee the existence of an oracle satisfying (A3). We here give a partial answer to this question for proper spaces  $X$ , relying on a few further parts of measurable selection theory as well as on the geometry of  $CX^\infty$ .

First, note that  $CX^\infty$  becomes Polish when  $X$  is locally compact. More concretely, if  $X$  is a locally compact separable Hadamard space, then  $CX^\infty$  is a locally compact Polish topological space, i.e. it is metrizable so that the resulting metric space is locally compact, complete and separable (see p. 7 of [13]). Further, if  $X$  is proper, then  $X^\infty$  is actually compact (see Definition II.8.6 in [18]).

Now, define the Busemann subgradient via

$$\partial_B f : E \times C \rightarrow 2^{CX^\infty}, (e, x) \mapsto \{[\xi, s] \in CX^\infty \mid x \in \operatorname{argmin}_{y \in C} (f(e, y) - \langle y, [\xi, s] \rangle)\}.$$

The existence of an oracle Busemann satisfying assumption (A3) amounts to showing that for any  $x : \Omega \rightarrow C$  and  $\zeta : \Omega \rightarrow E$  measurable, the map  $\partial_B f(\zeta, x) : \Omega \rightarrow 2^{CX^\infty}$  has a measurable selection, which amounts to  $\partial_B f(\zeta, x)$  being measurable.

Note that assumption (A2) yields that

$$\partial_B f(e, x) \subseteq \{[\xi, s] \in CX^\infty \mid s \leq L\} = (X^\infty \times [0, L]) / \sim$$

for all  $e \in E$  and  $x \in C$ . If  $X$  is proper, then  $X^\infty$  is compact and therefore also  $X^\infty \times [0, L]$  and  $(X^\infty \times [0, L]) / \sim$  are compact.

We now first note that  $\partial_B f(e, \cdot)$  has a closed graph. For that, we rely on a characterization using Fenchel conjugacy. Following [32] (see p. 16 therein), we define the Fenchel conjugate  $f^\bullet : E \times CX^\infty \rightarrow (-\infty, +\infty]$  of  $f$  by

$$f^\bullet(e, [\xi, s]) := \sup_{y \in C} (\langle y, [\xi, s] \rangle - f(e, y)).$$

Note that  $f^\bullet(e, \cdot)$  is lsc as  $\langle y, [\xi, s] \rangle$  is continuous in  $[\xi, s]$  and that  $f^\bullet(\cdot, [\xi, s])$  is measurable as  $f(\cdot, x)$  is measurable for all  $x$ .

Crucially, we have the following result on Fenchel duality:

**Lemma 5.6** (eq. (15) in [32]). *For any  $e \in E$ ,  $x \in C$  and  $[\xi, s] \in CX^\infty$ :*

$$f^\bullet(e, [\xi, s]) + f(e, x) \geq \langle x, [\xi, s] \rangle,$$

*with equality if, and only if,  $[\xi, s] \in \partial_B f(e, x)$ .*

We can now use this result to prove that  $\partial_B f(e, \cdot)$  has a closed graph (following essentially the usual argument for subdifferentials, see e.g. Proposition 16.36 in [6]):

**Lemma 5.7.** *Fix  $e \in E$ . Let  $(x_n) \subseteq C$  and  $([\xi_n, s_n]) \subseteq CX^\infty$  be given with  $[\xi_n, s_n] \in \partial_B f(e, x_n)$  as well as  $x_n \rightarrow x$  and  $[\xi_n, s_n] \rightarrow [\xi, s]$ . Then  $[\xi, s] \in \partial_B f(e, x)$ .*

*Proof.* Recall that both  $f$  and  $f^\bullet$  are lsc in their right argument. By Lemma 5.6, we get  $f^\bullet(e, [\xi_n, s_n]) + f(e, x_n) = \langle x_n, [\xi_n, s_n] \rangle$  from  $[\xi_n, s_n] \in \partial_B f(e, x_n)$ . Hence, using Lemma 5.6 again, we obtain

$$\begin{aligned} \langle x, [\xi, s] \rangle &\leq f^\bullet(e, [\xi, s]) + f(e, x) \\ &\leq \liminf_{n \rightarrow \infty} f^\bullet(e, [\xi_n, s_n]) + \liminf_{n \rightarrow \infty} f(e, x_n) \\ &\leq \liminf_{n \rightarrow \infty} (f^\bullet(e, [\xi_n, s_n]) + f(e, x_n)) \\ &= \lim_{n \rightarrow \infty} \langle x_n, [\xi_n, s_n] \rangle = \langle x, [\xi, s] \rangle, \end{aligned}$$

using in particular also again that  $\langle \cdot, \cdot \rangle$  is continuous. We hence have  $f^\bullet(e, [\xi, s]) + f(e, x) = \langle x, [\xi, s] \rangle$  and Lemma 5.6 yields  $[\xi, s] \in \partial_B f(e, x)$ .  $\square$

We can now derive that the graph of  $\partial_B f$  is measurable, and so using completeness of the space  $(E, \mathcal{E}, \mu)$  that  $e \mapsto \text{gra}(\partial_B f(e, \cdot))$  is a measurable set-valued map.

**Lemma 5.8.** *The function  $e \mapsto \text{gra}(\partial_B f(e, \cdot))$  is (weakly) measurable.*

*Proof.* Define

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

Then for  $g(e, x, [\xi, s]) = (e, [\xi, s], x)$ , which is measurable, we have

$$\text{gra}(\partial_B f) = g^{-1}(\text{gra}(\varphi)).$$

Note that  $f(e, y) - \langle y, [\xi, s] \rangle$  is jointly measurable in  $(e, [\xi, s])$  as well as continuous in  $y$ . Hence  $\text{gra}(\varphi)$  is measurable by Lemma 3.2 and so  $\text{gra}(\partial_B f)$  is measurable. The result now follows using completeness of  $(E, \mathcal{E}, \mu)$  (recall Theorem 8.1.4 in [4]).  $\square$

The last result we need is the following composition lemma which in works such as [4] (see Theorem 8.2.8 therein, which provides even more general results) is usually derived under a completeness assumption. We however here need to dispense of that, and for that follow the approach of Rockafellar [57] (whose proof goes through for our special case, albeit being originally phrased for Euclidean spaces):

**Lemma 5.9** (essentially Theorem 1N in [57]). *Let  $X, Y$  be complete separable metric spaces, with  $Y$  proper, and  $(T, \mathcal{T})$  be a measurable space. Let  $A : T \times X \rightarrow 2^Y$  be such that  $t \mapsto \text{gra}(A(t, \cdot))$  is (weakly) measurable with closed values. Then for all measurable  $x : T \rightarrow X$ , the map*

$$\varphi(t) := A(t, x(t))$$

*is (weakly) measurable.*

*Proof.* Let  $C \subseteq Y$  be open. Using that  $Y$  is proper, we express  $C$  as the union over a sequence of compact sets  $(C_k)$ . Define the map  $\varphi_k(t) := \{x(t)\} \times C_k$ , and note that  $\varphi_k$  has compact images for every  $k$ . By Lemma 18.4 in [3], each  $\varphi_k$  is (weakly) measurable. As  $C$  is open and  $\varphi(t)$  is closed, we have

$$\begin{aligned} \varphi^{-1}(C) &= \{t \in T \mid C \cap A(t, x(t)) \neq \emptyset\} \\ &= \bigcup_{k \in \mathbb{N}} \{t \in T \mid (x, y) \in \text{gra}(A(t, \cdot)) \text{ and } x = x(t), y \in C_k\} \\ &= \bigcup_{k \in \mathbb{N}} \{t \in T \mid \text{gra}(A(t, \cdot)) \cap \varphi_k(t) \neq \emptyset\}. \end{aligned}$$

By Lemma 18.4 in [3], we have that  $\text{gra}(A(t, \cdot)) \cap \varphi_k(t)$  is (weakly) measurable (where it is crucial that  $\varphi_k$  has compact images). In particular, the set

$$\{t \in T \mid \text{gra}(A(t, \cdot)) \cap \varphi_k(t) \neq \emptyset\} = (\text{gra}(A(t, \cdot)) \cap \varphi_k(t))^{-1}(X \times Y)$$

and hence  $\varphi^{-1}(C)$  is measurable. As  $C$  was arbitrary, we get that  $\varphi$  is (weakly) measurable.  $\square$

We recall the Kuratowski–Ryll–Nardzewski selection theorem:

**Lemma 5.10** (see e.g. Theorem 8.1.3 in [4]). *Let  $X$  be a complete separable metric space and let  $(T, \mathcal{T})$  be a measurable space. If  $\varphi : T \rightarrow 2^X$  is a set-valued measurable map such that  $\varphi(t)$  is non-empty and closed for any  $t \in T$ , then there exists a measurable function  $x : T \rightarrow X$  such that  $x(t) \in \varphi(t)$  for all  $t \in T$ .*

We now obtain the following result on the existence of Busemann subgradient oracles:

**Proposition 5.11.** *Let  $X$  be a separable Hadamard space with the geodesic extension property and at least two points, and assume that  $X$  is proper. Let  $C \subseteq X$  be closed convex nonempty. Further, let  $f : E \times C \rightarrow \mathbb{R}$  be a given functional, assuming (A1) and (A2).*

*Then for any  $x : \Omega \rightarrow C$  and  $\zeta : \Omega \rightarrow E$  measurable, the map  $\partial_B f(\zeta, x) : \Omega \rightarrow 2^{CX^\infty}$  is measurable. In particular, it has a measurable selection.*

*Proof.* Lemma 5.8 yields that  $e \mapsto \text{gra}(\partial_B f(e, \cdot))$  is (weakly) measurable and Lemma 5.7 yields that it has closed values, so that measurability of  $\partial_B f(\zeta, x)$  follows using Lemma 5.9 (recall that  $\partial_B f$  maps to subsets of a compact space). As  $\partial_B f(\zeta, x)$  is closed, Lemma 5.10 yields the existence of a measurable selection.  $\square$

## 6. PROOFS OF THE MAIN RESULTS

We now turn to our convergence results for (SB). For that, we now briefly recall the setup: Let  $(E, \mathcal{E}, \mu)$  and  $(\Omega, \mathcal{F}, \mathbb{P})$  be probability spaces, with  $(E, \mathcal{E}, \mu)$  complete, exactly as before, and let  $X$  be a separable Hadamard space with the geodesic extension property and at least two points. Further, fix a closed convex nonempty subset  $C \subseteq X$  and let  $f : E \times C \rightarrow \mathbb{R}$  be a function with properties (A1) – (A3) as above. Let  $(x_n)$  be the iteration given by (SB), and assume (Par).

The convergence proof now proceeds by showing that the iteration in question is stochastically quasi-Fejér monotone. For that, we introduced some notation. Define the filtration

$$\mathbb{F}_n := \sigma(\zeta_1, \dots, \zeta_n, x_0, \dots, x_n)$$

and write  $\mathbb{E}_n$  for the conditional expectation  $\mathbb{E}[\cdot \mid \mathbb{F}_n]$ .

The key geometric ingredient is the following property of the Busemann subgradients:

**Lemma 6.1** (see e.g. Lemma 6.1 in [32]). *Let  $g : C \rightarrow \mathbb{R}$  be a given function for a non-empty closed and convex set  $C \subseteq X$  and let  $[\xi, s]$  be a Busemann subgradient of  $f$  at  $x \in C$ . Given  $t > 0$ , define*

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

*Then for any  $y \in C$ :*

$$d^2(x^+, y) \leq d^2(x, y) - 2t(f(x) - f(y)) + s^2t^2.$$

This immediately allows us to derive the following stochastic variant of quasi-Fejér monotonicity.

**Lemma 6.2.** *For any  $n \in \mathbb{N}$  and any  $y \in C$ :*

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

*Proof.* Fix  $y \in C$  and  $n \in \mathbb{N}$ . Using Lemma 6.1, we get

$$d^2(x_{n+1}, y) \leq d^2(x_n, y) - 2t_n(f(\zeta_{n+1}, x_n) - f(\zeta_{n+1}, y)) + s_n^2t_n^2.$$

By assumption (A2), we get  $s_n \leq L$  and applying conditional expectations yields

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

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

$$\mathbb{E}_n[f(\zeta_{n+1}, x_n)](\omega) = \int f(\zeta_{n+1}(\omega'), x_n(\omega)) d\mathbb{P}(\omega') = F(x_n(\omega))$$

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

We can now prove Theorems 1.2 and Theorem 1.3:

*Proof of Theorem 1.2.* By Lemma 6.2, we have

$$\mathbb{E}_n[d^2(x_{n+1}, z)] \leq d^2(x_n, z) - 2t_n(F(x_n) - \min F) + L^2t_n^2$$

for any  $z \in \operatorname{argmin} F$ , for which we fix one in the following (using  $\operatorname{argmin} F \neq \emptyset$ ). Therefore, using (Par), we find that (\*) in Proposition 4.3 is satisfied. Items (1) and (2) of that result now yield that  $(x_n)$  is bounded and  $\sum_{n=0}^{\infty} t_n(F(x_n) - \min F) < +\infty$  a.s., say jointly on a set  $\widehat{\Omega}$  of measure one. Fix one such  $\omega \in \widehat{\Omega}$ . Using (Par) again, this in particular implies

$$\liminf_{n \rightarrow \infty} F(x_n(\omega)) = \min F.$$

Let  $F(x_{n_k}(\omega)) \rightarrow \min F$ . Using local compactness, choose a convergent subsequence  $x_{n_{k_j}}(\omega) \rightarrow x(\omega)$ . In particular, using that  $F$  is lsc by Fatou's lemma, we have

$$F(x(\omega)) \leq \liminf_{j \rightarrow \infty} F(x_{n_{k_j}}(\omega)) = \lim_{j \rightarrow \infty} F(x_{n_{k_j}}(\omega)) = \min F$$

so that  $x(\omega) \in \operatorname{argmin} F$  for any  $\omega \in \widehat{\Omega}$ . In particular, we thus have that  $\mathfrak{S}(x_n) \cap \operatorname{argmin} F \neq \emptyset$  a.s., so that item (5) of Proposition 4.3 yields that  $(x_n)$  strongly converges a.s. to an  $\operatorname{argmin} F$ -valued random variable.  $\square$

Further, we also immediately get the following proof for the corresponding result on weak ergodic convergence:

*Proof of Theorem 1.3.* As in the proof of Theorem 1.2, we derive

$$\mathbb{E}_n[d^2(x_{n+1}, z)] \leq d^2(x_n, z) - 2t_n(F(x_n) - \min F) + L^2 t_n^2$$

for any  $z \in \operatorname{argmin} F$  from Lemma 6.2, which using (Par) further yields

$$\sum_{n=0}^{\infty} t_n(F(x_n(\omega)) - \min F) < +\infty$$

for any  $\omega \in \widehat{\Omega}$ , with some  $\widehat{\Omega}$  of measure one. Using the convexity of  $F$ , we now have that

$$F(\bar{x}_n(\omega)) - \min F \leq \frac{\sum_{k=0}^n t_k(F(x_k(\omega)) - \min F)}{\bar{t}_n} \rightarrow 0$$

as  $\bar{t}_n \rightarrow +\infty$  by (Par) and  $\sum_{n=0}^{\infty} t_n(F(x_n(\omega)) - \min F) < +\infty$ , for any  $\omega \in \widehat{\Omega}$ . Therefore, if  $\bar{x}_{n_k}(\omega) \rightarrow^w x(\omega)$  we get

$$F(x(\omega)) \leq \liminf_{k \rightarrow \infty} F(\bar{x}_{n_k}(\omega)) = \lim_{n \rightarrow \infty} F(\bar{x}_n(\omega)) = \min F$$

so that  $x(\omega) \in \operatorname{argmin} F$  for any  $\omega \in \widehat{\Omega}$ , using that  $F$  is lsc by Fatou's lemma and hence also weakly lsc (recall Section 2). We have thus shown  $\mathfrak{W}(\bar{x}_n) \subseteq \operatorname{argmin} F$  a.s. Further, note that  $\operatorname{argmin} F$  is convex as  $F$  is convex. Lastly, note that Lemma 6.1 in particular implies

$$d^2(x_{n+1}, y) \leq d^2(x_n, y) - 2t_n(f(\zeta_{n+1}, x_n) - f(\zeta_{n+1}, y)) + s_n^2 t_n^2$$

for all  $y \in C$ , so that setting  $y = x_n$  yields  $d^2(x_{n+1}, x_n) \leq L^2 t_n^2$  and hence

$$\sum_{n=0}^{\infty} d^2(x_{n+1}, x_n) < +\infty \text{ a.s.}$$

and so  $d(x_{n+1}, x_n) \rightarrow 0$  a.s. The assumptions of Proposition 4.5 are thereby met, which yields that  $(\bar{x}_n)$  weakly converges a.s. to an  $\operatorname{argmin} F$ -valued random variable.  $\square$

At last, we turn to the quantitative result given in Theorem 1.4. For that, we require the following quantitative version of a lemma of Qihou [54] (see also Lemma 5.31 in [6]):

**Lemma 6.3** (Theorem 3.2 in [48]). *Let  $(x_n)$ ,  $(\alpha_n)$ ,  $(\beta_n)$  and  $(\gamma_n)$  be sequences of nonnegative reals with*

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

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

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

The next result on approximation properties for summable sequences is folklore:

**Lemma 6.4.** *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  $\theta : \mathbb{N} \times (0, \infty) \rightarrow \mathbb{N}$  such that  $\sum_{n=k}^{\theta(k,b)} u_n \geq b$  for all  $b > 0$  and  $k \in \mathbb{N}$ . Then  $\liminf_{n \rightarrow \infty} v_n = 0$  with*

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

*Proof.* For arbitrary  $\varepsilon > 0$  and  $N \in \mathbb{N}$ , suppose for a contradiction that  $v_n \geq \varepsilon$  for all  $n \in [N; \theta(N, L/\varepsilon)]$ . Then  $L \leq \varepsilon \sum_{n=N}^{\theta(N, L/\varepsilon)} u_n \leq \sum_{n=N}^{\theta(N, L/\varepsilon)} u_n v_n \leq \sum_{n=0}^{\infty} u_n v_n < L$ , which is a contradiction.  $\square$

With these in place, we now derive a first asymptotic approximation result, extending Theorem 1.1:

**Lemma 6.5.** *Let  $\theta : \mathbb{N} \times (0, \infty) \rightarrow \mathbb{N}$  be such that  $\sum_{n=k}^{\theta(k,b)} t_n \geq b$  for all  $b > 0$  and  $k \in \mathbb{N}$ . Further, let  $T > \sum_{n=0}^{\infty} t_n^2$ . Lastly, let  $b > 0$  be such that  $b > d^2(x_0, x^*)$  for some minimizer  $x^*$  of  $F$ . Then  $\liminf_{n \rightarrow \infty} \mathbb{E}[F(x_n)] = \min F$  with*

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

*Proof.* We have

$$\mathbb{E}_n[d^2(x_{n+1}, x^*)] \leq d^2(x_n, x^*) - 2t_n(F(x_n) - \min F) + L^2 t_n^2$$

by Lemma 6.2. Therefore, Lemma 6.3 yields  $\sum_{n=0}^{\infty} t_n \mathbb{E}[F(x_n) - \min F] < b + L^2 T$  and so Lemma 6.4 yields the result.  $\square$

We can now prove Theorem 1.4:

*Proof of Theorem 1.4.* For any  $n \in \mathbb{N}$ , define  $X_n := d^2(x_n, x^*) + L^2 \sum_{m=n}^{\infty} t_m^2$ . As  $(x_n)$  is adapted to  $(\mathcal{F}_n)$ , also  $(X_n)$  is adapted to  $(\mathcal{F}_n)$ . As we have

$$\mathbb{E}_n[d^2(x_{n+1}, x^*)] \leq d^2(x_n, x^*) - 2t_n(F(x_n) - \min F) + L^2 t_n^2 \leq d^2(x_n, x^*) + L^2 t_n^2$$

by Lemma 6.2, the stochastic process  $(X_n)$  is a nonnegative supermartingale. Indeed, note that

$$\mathbb{E}_n[X_{n+1}] = \mathbb{E}_n[d^2(x_{n+1}, x^*)] + L^2 \sum_{m=n+1}^{\infty} t_m^2 \leq d^2(x_n, x^*) + L^2 \sum_{m=n}^{\infty} t_m^2 = X_n.$$

Using the fact that  $f(e, \cdot)$  is strongly convex, we get that  $F$  is strongly convex with parameter  $\underline{\alpha}$ , i.e.

$$F(\gamma(tl)) \leq (1-t)F(\gamma(0)) + tF(\gamma(1)) - t(1-t)\frac{\underline{\alpha}}{2}d^2(\gamma(0), \gamma(1))$$

for any geodesic  $\gamma : [0, l] \rightarrow X$ . For  $\gamma$  being the unique geodesic joining  $x_n$  and  $x^*$ , and  $l = d(x_n, x^*)$ , we get

$$\min F \leq F\left(\gamma\left(\frac{l}{2}\right)\right) \leq \frac{1}{2}F(x_n) + \frac{1}{2}\min F - \frac{\underline{\alpha}}{8}d^2(x_n, x^*)$$

so that

$$\frac{\underline{\alpha}}{4}d^2(x_n, x^*) \leq F(x_n) - \min F.$$

Now, let  $\varepsilon > 0$  be arbitrary and using Lemma 6.5, choose an

$$n \in [\chi(\varepsilon/2L^2); \theta(\chi(\varepsilon/2L^2), 8(b + L^2 T)/\varepsilon\underline{\alpha})]$$

such that  $\mathbb{E}[F(x_n)] - \min F < \varepsilon\underline{\alpha}/8$ . Then  $\mathbb{E}[d^2(x_n, x^*)] < \varepsilon/2$ . Let  $m \geq n$  be arbitrary. Then

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

using that  $(X_m)$  is a supermartingale and the properties of  $\chi$ . As  $m$  was arbitrary, this yields  $\mathbb{E}[d^2(x_n, x^*)] \rightarrow 0$  and that  $\rho$  is a rate of convergence for that limit.

For  $d^2(x_n, x^*) \rightarrow 0$  a.s., note that

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

where the second inequality follows from Ville's inequality [58] (see also [44]). This immediately implies that  $d^2(x_n, x^*) \rightarrow 0$  a.s. with rate  $\rho'$ .  $\square$

**Acknowledgments.** I want to thank Morenikeji Neri and Thomas Powell for comments on a previous draft of this paper.

## REFERENCES

- [1] A.D. Aleksandrov. A theorem on triangles in a metric space and some of its applications. *Trudy Matematicheskogo Instituta imeni V.A. Steklova*, 38:5–23, 1951.
- [2] S. Alexander, V. Kapovitch, and A. Petrunin. *Alexandrov Geometry: Foundations*, volume 236 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2024.
- [3] C.D. Aliprantis and K.C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Berlin, Heidelberg, 2006.
- [4] J.-P. Aubin and H. Frankowska. *Set-Valued Analysis*. Birkhäuser Boston, MA, 2009.
- [5] R.J. Aumann. Measurable utility and the measurable choice problem. In G.T. Guilbaud, editor, *La Décision*, pages 15—26. Editions du Centre National de la Recherche Scientifique, Paris, 1969.
- [6] 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.
- [7] M. Bačák. The proximal point algorithm in metric spaces. *Israel Journal of Mathematics*, 194(2):689–701, 2013.
- [8] M. Bačák. Computing medians and means in Hadamard spaces. *SIAM Journal of Optimization*, 24(3):1542–1566, 2014.
- [9] 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.
- [10] M. Bačák. A variational approach to stochastic minimization of convex functionals. *Pure and Applied Functional Analysis*, 3(2):287–295, 2018.
- [11] M. Bačák. Old and new challenges in Hadamard spaces. *Japanese Journal of Mathematics*, 18(2):115–168, 2023.
- [12] M. Bačák, I. Searston, and B. Sims. Alternating projections in CAT(0) spaces. *Journal of Mathematical Analysis and Applications*, 385:599–607, 2012.
- [13] J. Bertrand and B. Kloeckner. A geometric study of Wasserstein spaces: Hadamard spaces. *Journal of Topology and Analysis*, 4(4):515–542, 2012.
- [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. Ergodic convergence of a stochastic proximal point algorithm. *SIAM Journal on Optimization*, 26(4):2235–2260, 2016.
- [17] 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.
- [18] M.R. Bridson and A. Haefliger. *Metric Spaces of Non-Positive Curvature*, volume 319 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin, Heidelberg, 1999.
- [19] 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.
- [20] C. Castaing and M. Valadier. *Convex Analysis and Measurable Multifunctions*, volume 580 of *Lecture Notes in Mathematics*. Springer Berlin, Heidelberg, 1977.
- [21] P.L. Combettes. Quasi-Fejérian analysis of some optimization algorithms. In D. Butnariu, Y. Censor, and S. Reich, editors, *Inherently Parallel Algorithms for Feasibility and Optimization*, pages 115–152. Elsevier, New York, 2001.
- [22] P.L. Combettes. Fejér monotonicity in convex optimization. In C.A. Floudas and P.M. Pardalos, editors, *Encyclopedia of Optimization*, pages 1016–1024. Kluwer, Boston, 2009.
- [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] C. Criscitiello and J. Kim. Horospherically Convex Optimization on Hadamard Manifolds Part I: Analysis and Algorithms, 2025. Preprint, <https://arxiv.org/abs/2505.16970>.
- [26] S. Dhompongsa, W.A. Kirk, and B. Sims. Fixed points of uniformly Lipschitzian mappings. *Nonlinear Analysis. Theory, Methods & Applications*, 65:762–772, 2006.

- [27] Y.M. Ermol'ev. On the method of generalized stochastic gradients and quasi-Fejér sequences. *Cybernetics*, 5:208–220, 1969.
- [28] Y.M. Ermol'ev. On convergence of random quasi-Fejér sequences. *Cybernetics*, 7:655–656, 1971.
- [29] Y.M. Ermol'ev and A.D. Tuniev. Random fejer and quasi-fejer sequences. *Theory of Optimal Solutions – Akademiya Nauk Ukrainskoi, SSR Kiev*, 2:76–83, 1968. in Russian; English translation in Amer. Math. Soc. Select. Translat. Math. Statist. Probab., 13 (1973), pp. 143–148.
- [30] R. Espínola and A. Fernández-León. CAT(k)-spaces, weak convergence and fixed points. *Journal of Mathematical Analysis and Applications*, 353(1):410–427, 2009.
- [31] A. Goodwin, A.S. Lewis, G. López-Acedo, and A. Nicolae. Convex optimization on CAT(0) cubical complexes. *Advances in Applied Mathematics*, 165, 2025. 102849, 20pp.
- [32] A. Goodwin, A.S. Lewis, G. López-Acedo, and A. Nicolae. Stochastic and incremental subgradient methods for convex optimization on Hadamard spaces, 2025. Preprint, <https://arxiv.org/abs/2412.06730>.
- [33] 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.
- [34] J. Jost. Equilibrium maps between metric spaces. *Calculus of Variations and Partial Differential Equations*, 2:173–204, 1994.
- [35] B. Kakavandi. Weak topologies in complete CAT(0) metric spaces. *Proceedings of the American Mathematical Society*, 141(3):1029–1039, 2013.
- [36] H. Khatibzadeh and M. Moosavi. Two proximal splitting methods in Hadamard spaces. *Applicable Analysis*, 102(2):635–650, 2023.
- [37] W.A. Kirk and B. Panyanak. A concept of convergence in geodesic spaces. *Nonlinear Analysis: Theory, Methods & Applications*, 68(12):3689–3696, 2008.
- [38] A. Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer Cham, 3rd edition, 2020.
- [39] U. Kohlenbach. *Applied Proof Theory: Proof Interpretations and their Use in Mathematics*. Springer Monographs in Mathematics. Springer-Verlag Berlin Heidelberg, 2008.
- [40] 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.
- [41] 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.
- [42] A.S. Lewis, G. López-Acedo, and A. Nicolae. Horoballs and the subgradient method, 2024. Preprint, <https://arxiv.org/abs/2403.15749>.
- [43] A. Lytchak and A. Petrunin. Weak topology on CAT(0) spaces. *Israel Journal of Mathematics*, 255(2):763–781, 2023.
- [44] M. Métivier. *Semimartingales*, volume 2 of *De Gruyter Studies in Mathematics*. Walter de Gruyter GmbH, Berlin/Boston, 1982.
- [45] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal of Optimization*, 19:1574–1609, 2009.
- [46] M. Neri and N. Pischke. Proof mining and probability theory. *Forum of Mathematics, Sigma*, 13, 2025. e187, 47pp.
- [47] 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>.
- [48] M. Neri and T. Powell. A quantitative Robbins-Siegmund theorem. *Annals of Applied Probability*, 2024. To appear, <https://arxiv.org/abs/2410.15986>.
- [49] 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.
- [50] G.B. Passty. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 72(2):383–390, 1979.
- [51] B.J. Pettis. On integration in vector spaces. *Transactions of the American Mathematical Society*, 44:277–304, 1938.
- [52] N. Pischke. Mean-square and linear convergence of a stochastic proximal point algorithm in metric spaces of nonpositive curvature, 2025. Preprint, <https://arxiv.org/abs/2510.10697>.
- [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 D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In J.S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257. Academic Press, New York, 1971.
- [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. Integral functionals, normal integrands and measurable selections. In J.P. Gossez, E.J. Lami Dozo, J. Mawhin, and L. Waelbroeck, editors, *Nonlinear Operators and the Calculus of Variations*, volume 543 of *Lecture Notes in Mathematics*, pages 157–207. Springer, Berlin, Heidelberg, 1976.
- [58] J. Ville. *Étude Critique de la Notion de Collectif*. PhD thesis, École Polytechnique, 1939.
- [59] 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.