FORWARD-BACKWARD SPLITTING IN BILATERALLY BOUNDED ALEXANDROV SPACES

Heikki von Koch[*]   Tuomo Valkonen[†]

Abstract   With the goal of solving optimisation problems on non-Riemannian manifolds, such as geometrical surfaces with sharp edges, we develop and prove the convergence of a forward-backward method in Alexandrov spaces with curvature bounded both from above and from below. This bilateral boundedness is crucial for the availability of both the gradient and proximal steps, instead of just one or the other. We numerically demonstrate the behaviour of the proposed method on simple geometrical surfaces in $\mathbb{R}^3$.

1 INTRODUCTION

Our goal is to develop proximal gradient (i.e., forward-backward) methods for the problem

$$(1.1) \qquad \min_{x \in X} \ F(x) + G(x)$$

in Alexandrov spaces, in particular, on manifolds whose natural atlas may not possess a Riemannian metric, such as two-dimensional surfaces with kinks, embedded in $\mathbb{R}^3$. Here both F and G are convex with F smooth, in a suitable sense that we will discuss in Section 3. Alexandrov spaces, as will introduce in Section 2, roughly speaking, are manifolds with curvature bounded from either above or below.

No such method yet exists in the literature. In [22], gradient methods are defined in Alexandrov spaces with lower curvature bounds, and proximal point methods are defined in spaces with upper curvature bounds. With appropriate assumptions on the objective function and step sizes, both methods converge to a minimiser. It stands to reason that by combining the methods, we could create a proximal gradient method in *bilaterally bounded* Alexandrov spaces, i.e., spaces with both upper and lower curvature bounds.

Alexandrov spaces with just upper curvature bounds include Hadamard spaces, such as the manifold of positive semidefinite matrices. Several recent works discuss optimisation in Hadamard spaces [3, 4, 6, 8, 9, 16, 29]. Hadamard spaces, which have nonpositive curvature, are easy from the point of view of optimisation, as geodesics are unique and an inequality reminiscent of the Pythagoras identity in Hilbert spaces, holds in the correct direction for standard convergence proofs to go through with little modifications [27].

Informally speaking, the curvature bounds of an Alexandrov space guarantee that its distance function has a certain amount of concavity or convexity when compared to its counterparts in the model spaces. For example, the distance function of a nonnegatively curved Alexandrov space is not less concave than that of the Euclidean plane, whereas for nonpositively curved Alexandrov spaces it is

[*]Department of Mathematics and Statistics, University of Helsinki, Finland, heikki.vonkoch@helsinki.fi, ORCID: 0009-0006-6906-9702

[†]Research Center in Mathematical Modeling and Optimization (MODEMAT), Quito, Ecuador *and* Department of Mathematics and Statistics, University of Helsinki, Finland, tuomo.valkonen@iki.fi, ORCID: 0000-0001-6683-3572

not less convex. More geometrically, the triangles in a nonnegatively/nonpositively curved Alexandrov space are not thinner/thicker than their Euclidean counterparts. Even though at first glance these purely metric spaces have very little to do with the well-behaved model spaces they are compared to, it turns out that much can be said about their local and global structure. For instance, perhaps the most important shared property of all Alexandrov spaces is that of the well-defined notion of an angle. It allows us to develop differential calculus in the form of differentials and gradients which further allow us to talk about optimality conditions and gradient flow.

Earlier works on optimisation in general Alexandrov spaces [19, 22], which we review in Section 4, assume the effective domain of the map G to be small or, how it is written in those works, they calculate the proximal map not for just any G, but for $G + \delta_A$, where A is assumed to have a small diameter. This has the effect of forcing the existence of geodesics between all points of A, allowing for simple "convex optimisation" style proofs even in non-Hadamard spaces. We do not make that restriction. Instead our proofs are more reminiscent of convergence proofs for nonconvex composite optimisation [13–15], based on a priori and a posteriori estimates. The a priori estimate proves that, if our current iterate is close enough to a solution (i.e., deep enough in the interior of A), then the next iterate does not escape from A. The a posteriori estimate then proves that we do not escape at all, and even advance towards the solution. The effect is that our results only require initialising close to solution of (1.1), instead restricting the problem to a set of bounded diameter. The restriction is a common assumption in optimisation on manifolds and is necessitated by the uniqueness of geodesics between points in the set where our iterates lie; the maximum diameter of this set depends on the upper curvature bound.

The bilateral bounds of an Alexandrov space forces the space to be a topological manifold, which along with completeness, implies the completeness of the tangent spaces of its points. This allows us to use the exponential map instead of the rather technical gradient exponential map in the definition of the gradient descent map. This has the benefit of being easier to compute in simple examples. One could also potentially use the gradient exponential in the gradient descent map, but as the gradient exponential and the exponential maps agree on whenever the latter is defined, these result in the same map, at least, in geodesically convex sets.

After the aforementioned introductory material of Sections 2 to 4, our main contributions start in Section 5 with a study of contractivity properties of a proximal gradient map in bilaterally bounded Alexandrov spaces. We then use this work in Section 6 to prove the convergence of a proximal gradient method based on the iterative application of the eponymous map. We finish in Section 7 by working out and numerically illustrating the proximal gradient map for a Lasso-type problem on the surfaces of a cube and a capped cylinder in $\mathbb{R}^3$.

2 ALEXANDROV SPACES

We start by introducing definitions and notation used in this paper. For a comprehensive treatment of metric geometry and basics of spaces with curvature bounded from above and below, see for example [11]. For a fundamental study of general Alexandrov spaces, see [1].

Let (X, d) be a metric space. A continuous curve $\gamma : [0, 1] \to X$ is called a unit-speed geodesic (or simply, geodesic) if for all $s, t \in [0, 1]$ we have $d(\gamma(s), \gamma(t)) = |s - t|$, i.e., it is globally distance-preserving. Metric spaces (X, d) that admit a unit-speed geodesic between any two points x and y are called geodesic, and such unit-speed geodesics from x to y will be denoted by $\gamma(t) = x \#_t y$ for $t \in [0, 1]$ (note that the curve may not be unique). A subset $A \subset X$ of a geodesic space X is called (geodesically) convex if for every two points $x, y \in A$, any unit-speed geodesic $x \#_t y$ lies in A.

Given $\kappa \in \mathbb{R}$, the model κ-plane, denoted by $\mathbb{M}^2(\kappa)$, is a complete simply-connected 2-dimensional Riemannian manifold of constant curvature κ. This limits the metric space $(\mathbb{M}^2(\kappa), d_{\mathbb{M}^2(\kappa)})$ to be one of three possibilities (up to isometry): the hyperbolic space of constant curvature, if $\kappa < 0$; the Euclidean plane, if $\kappa = 0$; or the sphere of constant curvature, if $\kappa > 0$. Set $D_\kappa := \operatorname{diam} \mathbb{M}^2(\kappa)$, where the diameter

of a metric space (X, d) is defined by $\operatorname{diam} X := \sup_{x,y \in X} d(x,y)$. We have $D_\kappa = \infty$, if $\kappa \leq 0$ and $D_\kappa = \pi/\sqrt{\kappa}$, if $\kappa > 0$.

To define geodesic metric spaces with bounded sectional curvature, we need to compare the metric triangles defined in X to those defined in a model κ-plane $\mathbb{M}^2(\kappa)$. For any three points $x, y, z \in X$ satisfying $d(x,y) + d(y,z) + d(z,x) < 2D_\kappa$, there are unique points (up to isometry, [10, 2.14 Lemma]) $\tilde{x}, \tilde{y}, \tilde{z} \in \mathbb{M}^2(\kappa)$ such that

$$d_{\mathbb{M}^2(\kappa)}(\tilde{x}, \tilde{y}) = d(x,y), \quad d_{\mathbb{M}^2(\kappa)}(\tilde{y}, \tilde{z}) = d(y,z), \quad d_{\mathbb{M}^2(\kappa)}(\tilde{z}, \tilde{x}) = d(z,x).$$

The triangle $\triangle \tilde{x} \tilde{y} \tilde{z}$ is called the comparison triangle of $\triangle xyz$ in $\mathbb{M}^2(\kappa)$. A unit-speed geodesic $\tilde{\gamma} : [0,1] \to \mathbb{M}^2(\kappa)$ from $\tilde{x}$ to $\tilde{y}$ will be denoted by $\tilde{x} \#_t \tilde{y}$.

Definition 2.1 (Alexandrov spaces). Let (X, d) be a geodesic metric space, $\kappa \in \mathbb{R}$, and $x, y, z \in X$ any three points satisfying $d(x,y) + d(y,z) + d(z,x) < 2D_\kappa$. The space (X, d) is called an Alexandrov space with curvature bounded above by κ, if

$$(2.1) \qquad\qquad d(y\#_t z, x) \leq d_{\mathbb{M}^2(\kappa)}(\tilde{y}\#_t\tilde{z}, \tilde{x})$$

for any unit-speed geodesic $y\#_t z$.

Likewise, (X, d) is called an Alexandrov space with curvature bounded below by κ, if

$$(2.2) \qquad\qquad d(y\#_t z, x) \geq d_{\mathbb{M}^2(\kappa)}(\tilde{y}\#_t\tilde{z}, \tilde{x})$$

for any unit-speed geodesic $y\#_t z$.

We call both spaces bounded from above and spaces bounded from below by the generic name Alexandrov space. We will usually distinguish the bounds from each other by denoting the upper bound by $\overline{\kappa}$ and the lower bound by $\underline{\kappa}$. Some typical examples are given here.

Example 2.2. 1. Complete Riemannian manifolds with sectional curvature at most $\overline{\kappa}$ and their Gromov-Hausdorff limits are Alexandrov spaces with curvature bounded from above by $\overline{\kappa}$. Others include locally simply connected subsets of $\mathbb{R}^2$, complements of closed convex bodies in $\mathbb{R}^n$, isometric gluings of spaces bounded from above along convex sets, Hadamard manifolds and Hilbert spaces (such as $\mathbb{R}^n$; bounded from above by 0), and metric trees (bounded from above by any $\overline{\kappa} \in \mathbb{R}$). Away from the vertices, polyhedral surfaces in $\mathbb{R}^n$ are bounded from above with $\overline{\kappa} = 0$.

2. Complete Riemannian manifolds with sectional curvature at least $\underline{\kappa}$ and their Gromov-Hausdorff limits are Alexandrov spaces with curvature bounded from below by $\underline{\kappa}$. Other examples include Hilbert spaces (bounded from below by 0), L^2-Wasserstein spaces over nonnegatively curved spaces, and convex hypersurfaces of Riemannian manifolds (such as graphs of convex functions of the form $F : \mathbb{R}^n \to \mathbb{R}$ and boundaries of closed convex bodies in $\mathbb{R}^n$). Importantly, convex polyhedral surfaces in $\mathbb{R}^n$ are bounded from below by 0 and the tangent spaces of non-vertex points are isometric to $\mathbb{R}^n$. Equipped with distortion metrics, spaces of metric measure spaces are also Alexandrov spaces with curvature bounded from below [26]. Such spaces have relevance to, e.g., geometric shape interpolation [5].

3. Spaces with bilaterally bounded sectional curvature in the sense of Alexandrov include Hilbert spaces (bounded from above and below by 0) and, for example, a finitely long flat cylinder with hemispheres isometrically glued to its ends (bounded from below by 0 and above by the (constant) curvature of the sphere). In fact, spaces with bilateral curvature bounds are Gromov-Hausdorff limit spaces of smooth Riemannian manifolds with uniformly bounded bilateral sectional curvature.

To study the local geometry of Alexandrov spaces, we need to replace the concept of tangent spaces from the theory of smooth manifolds with the notion of a space of directions. Fix a point $x \in X$ and two unit-speed geodesics starting from it: $\gamma(0) = x = \eta(0)$. The angle between γ and η is defined as

$$\angle_x(\gamma, \eta) := \lim_{t,s \searrow 0} \angle \tilde{\gamma}(t)\tilde{x}\tilde{\eta}(s),$$

where $\angle \tilde{\gamma}(t)\tilde{x}\tilde{\eta}(s) \in \mathbb{M}^2(\kappa)$ is the well-defined angle at $\tilde{x}$ of a comparison triangle $\triangle \tilde{\gamma}(t)\tilde{x}\tilde{\eta}(s)$ in $\mathbb{M}^2(\kappa)$. The angle always exists and is independent of the choice of $\kappa \in \mathbb{R}$. Setting the geodesics γ and η to be equivalent if $\angle_x(\gamma, \eta) = 0$ defines an equivalence class and a pseudometric space $(\Sigma'_x X, \angle_x)$, where $\Sigma'_x X$ is the set of all nontrivial unit-speed geodesics (strictly speaking, equivalence classes of unit-speed geodesics) starting from x. The completion of $\Sigma'_x X$ is called the space of directions at x and is denoted by $\Sigma_x X$.

Definition 2.3 (The tangent cone). The tangent cone $(T_x X, d_{\angle_x})$ at $x \in X$ is defined as the Euclidean cone over $(\Sigma_x X, \angle_x)$, i.e., $T_x X := (\Sigma_x X \times [0, \infty))/\sim$, where $(\gamma, t) \sim (\eta, s)$ if and only if $t = s = 0$. The metric on $T_x X$ is defined as

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

for $(\gamma, t), (\eta, s) \in T_x X$.

Similarly, the Euclidean cone over the pseudometric space $(\Sigma'_x X, \angle_x)$ is denoted by $(T'_x X, d_{\angle_x})$. The definition of the tangent cone metric allows us to define an "inner product" of two elements $(\gamma, t), (\eta, s) \in T_x X$ as

$$(2.3) \qquad \langle (\gamma, t), (\eta, s) \rangle := ts \cos \angle_x(\gamma, \eta) = \frac{t^2 + s^2 - d_{\angle_x}^2((\gamma, t), (\eta, s))}{2},$$

where $\langle (\gamma, t), (\eta, s) \rangle := 0$, if $t = 0$ or $s = 0$.

The elements of the tangent cone $T_x X$ will be called tangent vectors (despite the fact that it is a cone and not necessarily a vector space). A vector $(\gamma, t) \in T'_x X$ with $t = d(x, y)$, corresponding to a geodesic $\gamma(t) = x \#_t y$, is called logarithm (of the geodesic γ) and is denoted by $\log_x y$ (not necessarily uniquely defined). In general, we have the following definition for velocity of curves.

Definition 2.4 ([1, Definition 6.9]). Let X be a metric space, $a > 0$, and $\alpha : [0, a) \to X$ be a function, not necessarily continuous, such that $\alpha(0) = x$. We say that $v \in T_x X$ is the right derivative of α at 0, briefly $\alpha^+(0) = v$, if for some (and therefore any) sequence of vectors $v_n \in T'_x X$ such that $v_n \to v$ as $n \to \infty$, and corresponding geodesics γ_n, we have

$$\limsup_{t \searrow 0} \frac{d(\alpha(t), \gamma_n(t))}{t} \to 0 \quad \text{as} \quad n \to \infty.$$

Especially, if we choose α to be a geodesic $\gamma(t) = x \#_t y$ and $v_n = v$ for all n, then $\gamma^+(0) = v = \log_x y \in T'_x X$.

Remark 2.5. For a Riemannian manifold (X, d), the space of directions $(\Sigma_x X, \angle_x)$ and the tangent cone $(T_x X, d_{\angle_x})$ coincide with the unit tangent sphere and the tangent space at x, respectively [12].

The definition of the angle allows for the following fundamental rule for differentiating the length of a variable curve; see [1, 8.42, 9.36].

Theorem 2.6 (First variation formula). *Let (X, d) be an Alexandrov space with curvature bounded from above or below by κ, and assume that X is locally compact when the curvature is bounded from below. For any point $z \in X$ and geodesic $\gamma(t) = x \#_t y$ with $x, y \in X$, and $0 < d(x, z) < \pi/\sqrt{\kappa}$ if $\kappa > 0$, we have*

$$\lim_{t \searrow 0} \frac{d_z(\gamma(t)) - d_z(x)}{t} = -\cos \theta_{\min},$$

where $d_z(a) := d(z, a)$ for any $a \in X$ and $\theta_{\min}$ is the minimum of angles between γ and all geodesics from x to z.

Remark 2.7. The first variation inequality "$\leq$" is very general as it holds for metric spaces with defined angles [1, 6.7] (no need for curvature bounds or compactness). This turns out to be sufficient for the trigonometric bound for the squared distance function Lemma 3.7, whereas the other trigonometric bound Lemma 3.9 requires the full equality.

It turns out that bilaterally bounded Alexandrov spaces are Riemannian in the sense that the metric tensor is of lower smoothness than is usually required.

Theorem 2.8 ([7, Theorem 5]). *Let (X, d) be a complete locally compact Alexandrov space with curvature bounded above by $\overline{\kappa}$ and below by $\underline{\kappa}$ ($\underline{\kappa} \leq \overline{\kappa}$). Then X is finite dimensional and admits a natural atlas of distance coordinates of class C^1. Furthermore, the distance on X is defined by a metric of class C^0 (continuous).*

In fact, from the construction of the distance coordinates, a little more is true; the atlas is $C^{1,1}$-smooth making the components of the metric tensor $C^{0,1}$-smooth. No more regularity can be expected, however, as can be seen from the classic example of gluing a hemisphere to the top of a flat cylinder: the curvature is not continuous and hence the atlas cannot be C^2-smooth.

Remark 2.9. The notion of a boundary of a finite dimensional Alexandrov space (denoted by ∂X) with curvature bounded from below is required for the results regarding the gradient descent map. The boundary is defined inductively with respect to dimension and space of directions, and can be found in, for example, [12, 7.19]. However, since we will be assuming bilateral bounds for our results, the space will have the structure of a differentiable manifold (of low regularity) and as such the boundary is defined in the standard manner with upper half-spaces.

We also need the following concept of sequences satisfying a monotonicity property with an error term to analyse the convergence of optimisation algorithms in generic metric spaces.

Definition 2.10. A sequence $\{x^k\}_{k\in\mathbb{N}}$ in a complete metric space (X, d) is called quasi-Fejér with respect to a set $U \subset X$ if, for every $\bar{x} \in U$, there exist a sequence $\{\varepsilon_k\}_{k\in\mathbb{N}} \subset \mathbb{R}$ such that $\varepsilon_k \geq 0$, $\sum_{k=0}^{\infty} \varepsilon_k < \infty$, and $d^2(x^{k+1}, \bar{x}) \leq d^2(x^k, \bar{x}) + \varepsilon_k$, for all $k = 0, 1, \ldots$.

If, for every point $\bar{x} \in U$, we can choose the zero sequence $\varepsilon_k = 0$ for all $k = 0, 1, \ldots$, then the sequence $\{x^k\}_{k\in\mathbb{N}}$ is simply called Fejér (with respect to U). The fundamental property of quasi-Fejér sequences is the following extension of Opial's lemma to metric spaces.

Theorem 2.11. *Let $\{x^k\}_{k\in\mathbb{N}}$ be a sequence in a complete locally compact metric space (X, d). If $\{x^k\}_{k\in\mathbb{N}}$ is quasi-Fejér with respect to a nonempty set $U \subset X$, then $\{x^k\}_{k\in\mathbb{N}}$ is bounded. If furthermore, an accumulation point x^* of $\{x^k\}_{k\in\mathbb{N}}$ belongs to U, then $\lim_{k\to\infty} x^k = x^*$.*

Proof. Let $\{x^k\}_{k\in\mathbb{N}}$ be a quasi-Fejér sequence with respect to a nonempty set $U \subset X$ and $\bar{x} \in U$. By definition, we have $d^2(x^1, \bar{x}) \leq d^2(x^0, \bar{x}) + \varepsilon_0$ so that $d^2(x^2, \bar{x}) \leq d^2(x^1, \bar{x}) + \varepsilon_1 \leq d^2(x^0, \bar{x}) + \varepsilon_0 + \varepsilon_1$. Iterating this further we have that $d^2(x^k, \bar{x}) \leq d^2(x^0, \bar{x}) + \sum_{i=0}^{k-1} \varepsilon_i \leq d^2(x^0, \bar{x}) + \sum_{i=0}^{\infty} \varepsilon_i < \infty$ for all $k = 0, 1, \ldots$ by the definition of the sequence $\{\varepsilon_k\}_{k\in\mathbb{N}}$; that is, $\{x^k\}_{k\in\mathbb{N}}$ is bounded.

Now let $x^* \in U$ be an accumulation point of $\{x^k\}_{k\in\mathbb{N}}$. Since the sequence is bounded, by local compactness, it has a subsequence $\{x^{k_i}\}_{i\in\mathbb{N}}$ such that $\lim_{i\to\infty} x_{k_i} = x^*$. Let $\delta > 0$. By definition, there exists an N_1 such that $\sum_{i=N_1}^{\infty} \varepsilon_i < \delta/2$ and an $N_2 \geq N_1$ such that $d^2(x^{k_j}, x^*) < \delta/2$ for any $j \geq N_2$. From these we get that

$$d^2(x^k, x^*) \leq d^2(x^{k_j}, x^*) + \sum_{i=j}^{k-1} \varepsilon_i \leq d^2(x^{k_j}, x^*) + \sum_{i=j}^{\infty} \varepsilon_i < \frac{\delta}{2} + \frac{\delta}{2} = \delta,$$

for any $k \geq j$. Since $\delta > 0$ was arbitrary, $\lim_{k\to\infty} x^k = x^*$. $\qquad\square$

3 CONVEX ANALYSIS IN ALEXANDROV SPACES

Let (X, d) be an Alexandrov space with curvature bounded from above or below by $\kappa \in \mathbb{R}$. We define extended real-valued functions as $F : X \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ (the codomain excludes the value $-\infty$). As it is natural to consider extended real-valued functions in convex analysis, it is also useful to specify

the set in which the function attains finite values. These points constitute the effective domain of a function F, denoted by $\mathrm{dom}\, F$, i.e.,

$$\mathrm{dom}\, F := \{x \in X \mid F(x) < \infty\}.$$

If $\mathrm{dom}\, F \neq \emptyset$, the function F is called proper.

Extended real-valued λ-convex functions are defined in the usual manner.

Definition 3.1 (λ-convexity). A function $F : X \to \overline{\mathbb{R}}$ is called λ-convex for $\lambda \in \mathbb{R}$ if

$$F(\gamma(t)) \leq (1-t)F(x) + tF(y) - \frac{\lambda}{2}t(1-t)d^2(x, y)$$

holds for any $x, y \in X$, $t \in [0, 1]$ and any geodesic $\gamma : [0, 1] \to X$ with $\gamma(t) = x \#_t y$.

Note that geodesics in Alexandrov spaces are always assumed to be minimal (distance minimising) and unit-speed (parametrised by arc length). Functions are called convex and strongly convex when $\lambda = 0$ and $\lambda > 0$, respectively. It should also be noted that functions that are strongly convex are also strictly convex (the converse is not true however). Perhaps the most important λ-convex functions are the squared distance functions. We define the closed metric ball of radius $r > 0$ around a point $x \in X$ as: $\overline{B}(x, r) := \{y \in X \mid d(x, y) \leq r\}$.

Proposition 3.2 ([20, Proposition 3.1]). *Let* (X, d) *be an Alexandrov space with curvature bounded from above by* $\overline{\kappa} > 0$. *Then, for any* $a \in X$, *the function* $d_a^2 := d^2(a, x)$, *for* $x \in X$, *is* λ-convex *on the geodesically convex metric ball* $\overline{B}(a, r)$ *with* $2r = (\pi/2 - \varepsilon)/\sqrt{\overline{\kappa}}$ *and* $\lambda = (\pi - 2\varepsilon)\tan\varepsilon$ *for arbitrary* $\varepsilon \in (0, \pi/2)$.

When the space is an Alexandrov space with curvature bounded from above by 0, the squared distance function is 2-convex [1, Corollary 9.26] (in fact, these are equivalent). Moreover, since a complete Alexandrov space bounded from above by $\overline{\kappa}$, is an Alexandrov space bounded from above by $\overline{K}$, for $\overline{\kappa} < \overline{K}$ [1, Corollary 9.18], we have a strong convexity factor for the squared distance function for all $\overline{\kappa} \in \mathbb{R}$ (λ is given by Proposition 3.2 when $\overline{\kappa} > 0$, and $\lambda = 2$, when $\overline{\kappa} \leq 0$).

Proposition 3.3 ([21, Lemma 3.3]). *Let* (X, d) *be an Alexandrov space with curvature bounded from below by* $\underline{\kappa} < 0$. *Then, for any* $a \in X$, *the function* $-d_a^2$ *is* λ-convex *on the metric ball* $\overline{B}(a, r)$ *with* $\lambda = -2(1 - \underline{\kappa}(2r)^2)$ *for all* $r > 0$.

When the space is a complete Alexandrov space with curvature bounded from below by 0, the negative squared distance function is -2-convex [1, Corollary 8.24] (in fact, these are equivalent). Moreover, since a complete Alexandrov space bounded from below by $\underline{\kappa}$, is an Alexandrov space bounded from below by $\underline{K}$, for $\underline{K} < \underline{\kappa}$ [1, Corollary 8.33], we have a strong concavity factor for the squared distance function for all $\underline{\kappa} \in \mathbb{R}$ (λ is given by Proposition 3.3 when $\underline{\kappa} < 0$, and $\lambda = -2$, when $\underline{\kappa} \geq 0$).

Due to the existence of angles in Alexandrov spaces, functions have well-defined directional derivatives.

Definition 3.4 (Directional derivative and differential). [22, Definition 3.2] Let X be an Alexandrov space and $F : X \to \overline{\mathbb{R}}$. The directional derivative of F at $x \in \mathrm{dom}\, F$ in the direction $h \in T_x X$ is defined as

$$D_x F(h) := \liminf_{(\gamma, s) \to h} \left\{ \lim_{t \searrow 0} \frac{F(\gamma(st)) - F(x)}{t} \right\},$$

where $(\gamma, s) \in \Sigma'_x X \times [0, \infty) \subset T_x X$. The function $D_x F : T_x X \to \mathbb{R}$ is called the differential of F at x.

This limit can be shown to exist for proper, λ-convex, and lower semicontinuous functions [21, 4.1], while its uniqueness is established in [1, Proposition 6.16] for functions with additional Lipschitz continuity. For the former functions, it immediately follows from the definition of λ-convexity that

$$(3.1) \qquad\qquad D_x F(\gamma^+(0)) \leq F(y) - F(x) - \frac{\lambda}{2}d^2(x, y),$$

for a geodesic $\gamma(t) = x \#_t y$ and any $y \in X$. This immediately yields a Fermat principle for convex and strongly convex functions.

Corollary 3.5 (Fermat principle). *Let (X, d) be an Alexandrov space and $F : X \to \overline{\mathbb{R}}$ proper, lower semicontinuous, and λ-convex with $\lambda \geq 0$. Furthermore, let $x^* \in \operatorname{dom} F$ and γ be a geodesic with $\gamma(0) = x^*$. The following are equivalent:*

1. *$D_{x^*} F(\gamma^+(0)) \geq 0$, for all $\gamma^+(0) \in T_{x^*} X$;*

2. *$F(x^*) = \min_{x \in X} F(x)$.*

Proof. Let $D_{x^*} F(\gamma^+(0)) \geq 0$, for any geodesic $\gamma(t) = x^* \#_t y$. Directly from (3.1) and $\lambda \geq 0$ we get

$$0 \leq D_{x^*} F(\gamma^+(0)) \leq F(y) - F(x^*) - \frac{\lambda}{2} d^2(x^*, y) \leq F(y) - F(x^*) \quad \text{for all } y \in X,$$

i.e., $F(x^*) \leq F(y)$ for all $y \in X$.

Now let $F(x^*) = \min_{x \in X} F(x)$. This implies $F(\gamma(t)) \geq F(x^*)$, for any $y \in X$ with $\gamma(t) = x^* \#_t y$ and $t \in [0, 1]$; that is

$$\frac{F(\gamma(t)) - F(x^*)}{t} \geq 0,$$

for small $t > 0$. Taking $t \searrow 0$ yields the claim. $\qquad\square$

We also need the following definition of the descending absolute gradient of F.

Definition 3.6 (Absolute gradients). [22, Definition 3.1] The descending absolute gradient of $F : X \to \overline{\mathbb{R}}$ at $x \in \operatorname{dom} F$ is defined by

$$|\nabla_- F|(x) := \max \left\{ 0, \limsup_{y \to x} \frac{F(x) - F(y)}{d(x, y)} \right\}.$$

It holds that $|\nabla_- F|(x) \in [0, \infty]$ and $|\nabla_- F|(x) \leq L$ if F is L-Lipschitz. A point x is called critical if $|\nabla_- F|(x) = 0$, which holds if $F(x) = \inf_{y \in X} F(y)$.

It can be shown in Alexandrov spaces bounded from below, that for a lower semicontinuous, λ-convex function F with $0 < |\nabla_- F|(x) < \infty$, there exists a unique direction $\alpha \in \Sigma_x X$ such that $D_x F((\alpha, 1)) = -|\nabla_- F|(x)$ and

$$(3.2) \qquad\qquad D_x F((\beta, 1)) \geq -|\nabla_- F|(x) \langle (\alpha, 1), (\beta, 1) \rangle$$

for all $\beta \in \Sigma_x X$ [21, Lemma 4.3]. As such, the tangent vector

$$(3.3) \qquad\qquad \nabla(-F)(x) := (\alpha, |\nabla_- F|(x)) \in T_x X$$

can be thought as a subgradient vector of $-F$ at x. This will be used in the definition of the gradient descent map in the lower curvature bound case. Normally, we would use the descent directions of F but since the tangent space $T_x X$ is not (in general) a vector space, the negative of the gradient may not exist. Instead, we will use a subgradient of negative F instead.

3.1 TRIGONOMETRIC BOUNDS

Using Propositions 3.2 and 3.3, we can now derive some very useful trigonometric bounds related to the squared distance functions; effectively generalisations of the law of cosine. By Proposition 3.3, in a complete Alexandrov space with curvature bounded from below by $\underline{\kappa} < 0$ (note that for $\underline{\kappa} \geq 0$, we

can choose $\lambda = -2$ as is mentioned after Proposition 3.3), the negative distance squared function is λ-convex with $\lambda < 0$:

$$-d^2(z, \gamma(t)) \leq -(1-t)d^2(z, x) - td^2(z, y) - \frac{\lambda}{2}t(1-t)d^2(x, y),$$

where $\gamma(t) = x\#_t y$ is a geodesic and $z \in X$ with $d(z, x) > 0$. This gives us

$$-\frac{d^2(z, \gamma(t)) - d^2(z, x)}{t} \leq d^2(z, x) - d^2(z, y) - \frac{\lambda}{2}(1-t)d^2(x, y),$$

which by taking $t \searrow 0$ yields

$$-D_x(d^2(z, x))(\gamma^+(0)) \leq d^2(z, x) - d^2(z, y) - \frac{\lambda}{2}d^2(x, y).$$

The differential exists (as d_z^2 is also proper and continuous) and the first variation formula Theorem 2.6 gives $D_x(d^2(z, x))(\gamma^+(0)) \leq -2\langle \gamma^+(0), \eta^+(0)\rangle$, where $\eta : [0, 1] \to X$ is the geodesic from x to z that realises the minimum angle between γ and all geodesics from x to z. The definition of the inner product (2.3) further gives $-2\langle \gamma^+(0), \eta^+(0)\rangle = -2d(x, y)d(x, z)\cos \angle_x(\gamma, \eta)$ so by rearranging we finally get

$$d^2(z, y) \leq d^2(z, x) - \frac{\lambda}{2}d^2(x, y) - 2d(x, y)d(x, z)\cos \angle_x(\gamma, \eta),$$

which is, in fact, nothing but a modified cosine rule in disguise (since $\lambda < 0$). The factor $\lambda/2$ can be given a more explicit expression as has been done in [30].

Lemma 3.7 ([30, Lemma 6]). *If a, b, c are the sides (i.e., side lengths) of a geodesic triangle in an Alexandrov space with curvature bounded below by $\underline{\kappa} \in \mathbb{R}$, and θ is the angle between sides b and c, then*

$$a^2 \leq \frac{\sqrt{|\underline{\kappa}|}c}{\tanh(\sqrt{|\underline{\kappa}|}c)}b^2 + c^2 - 2bc\cos(\theta).$$

That is, the factor $-\lambda/2$ can be chosen to be $\sqrt{|\underline{\kappa}|}d(x, z)/\tanh(\sqrt{|\underline{\kappa}|}d(x, z))$. Keeping the curvature $\underline{\kappa} < 0$ fixed, we see that $\sqrt{-\underline{\kappa}}d(x, z)/\tanh(\sqrt{-\underline{\kappa}}d(x, z))$ is an increasing function of the side length $c = d(x, z)$. In particular, if $x, z \in A \subset X$ with $\operatorname{diam} A < \infty$, we have that $c \leq \operatorname{diam} A$ and $\sqrt{-\underline{\kappa}}d(x, z)/\tanh(\sqrt{-\underline{\kappa}}d(x, z)) \leq \lambda_l/2$ for

$$(3.4) \qquad \frac{\lambda_l}{2} := \frac{\sqrt{-\underline{\kappa}}\operatorname{diam} A}{\tanh(\sqrt{-\underline{\kappa}}\operatorname{diam} A)}.$$

Remark 3.8. Although the lemma is stated for all $\underline{\kappa} \in \mathbb{R}$, we will only use it for $\underline{\kappa} < 0$ and choose $\lambda_l = 2$, when $\underline{\kappa} \geq 0$.

Repeating the same arguments with Proposition 3.2, in the case of a complete Alexandrov spaces with curvature bounded from above by $\overline{\kappa} > 0$, we get

$$(3.5) \qquad d^2(z, y) \geq d^2(z, x) + \frac{\lambda}{2}d^2(x, y) - 2d(x, y)d(x, z)\cos \angle_x(\gamma, \eta),$$

where $\lambda > 0$. The factor $\lambda/2$ can be chosen to be $\sqrt{\overline{\kappa}}d(x, z)/\tan(\sqrt{\overline{\kappa}}d(x, z))$, for $\overline{\kappa} > 0$. (As mentioned after Proposition 3.2, for $\overline{\kappa} \leq 0$, we can choose $\lambda = 2$.)

Lemma 3.9. *Let (X, d) be an Alexandrov space with curvature bounded from above by $\overline{\kappa} > 0$ and a, b, c the sides (i.e., side lengths) of a geodesic triangle with vertices and sides in a geodesically convex set $A \subset X$ with $\operatorname{diam} A < \pi/(2\sqrt{\overline{\kappa}})$. For the angle θ between sides b and c we have*

$$a^2 \geq \frac{\sqrt{\overline{\kappa}}c}{\tan(\sqrt{\overline{\kappa}}c)}b^2 + c^2 - 2bc\cos(\theta).$$

For curvature bounded above by $\overline{\kappa} \leq 0$, we can choose $(\sqrt{\overline{\kappa}}c)/\tan(\sqrt{\overline{\kappa}}c) = 1$.

Proof. From the previous discussion, we only need to show that the factor $\lambda/2$ can be chosen to be $\sqrt{\overline{\kappa}}c/\tan(\sqrt{\overline{\kappa}}c)$, for $\overline{\kappa} > 0$ and some side length c, with $c < \pi/(2\sqrt{\overline{\kappa}})$. By Proposition 3.2, we have $\varepsilon = \pi/2 - 2r\sqrt{\overline{\kappa}}$ and

$$\frac{\lambda}{2} = \frac{(\pi - 2\varepsilon)\tan\varepsilon}{2} = 2r\sqrt{\overline{\kappa}}\tan\left(\frac{\pi}{2} - 2r\sqrt{\overline{\kappa}}\right) = \frac{2r\sqrt{\overline{\kappa}}}{\tan(2r\sqrt{\overline{\kappa}})},$$

for $\overline{\kappa} > 0$ and an arbitrary $\varepsilon \in (0, \pi/2)$. Since for any $\varepsilon \in (0, \pi/2)$ we have $2r < \pi/(2\sqrt{\overline{\kappa}})$, the claim follows. $\qquad\square$

Again, keeping the curvature $\overline{\kappa}$ fixed, we see that $\sqrt{\overline{\kappa}}c/\tan(\sqrt{\overline{\kappa}}c)$ is a decreasing function of the side length $c = d(x, z)$. That is, if $x, z \in A \subset X$ with $\operatorname{diam} A < \pi/(2\sqrt{\overline{\kappa}})$, we have that $c \le \operatorname{diam} A$ and $\sqrt{\overline{\kappa}}c/\tan(\sqrt{\overline{\kappa}}c) \ge \lambda_u/2$ for

$$(3.6) \qquad\qquad \frac{\lambda_u}{2} := \frac{\sqrt{\overline{\kappa}}\operatorname{diam} A}{\tan(\sqrt{\overline{\kappa}}\operatorname{diam} A)}.$$

These bounds allow us to write the key lemmas related to the proximal, gradient, and proximal gradient methods in an explicit manner.

Remark 3.10. The assumption on the diameter of the subset is an important standard assumption in convex analysis in positively curved spaces as it forces the set to be convex [1, Corollary 9.27]. In the case of Riemannian manifolds (of sufficient smoothness), this also means that the exponential map as well as its inverse is uniquely defined in the set. Practically, convergence may be observed in a larger set.

4 PROXIMAL AND GRADIENT DESCENT MAPS

To analyse the proximal gradient map, we need to analyse the proximal and gradient descent maps seperately first. It turns out that the curvature bound determines which method is more appropriate for the analysis of the gradient flow; the proximal map behaves reasonabley in spaces with upper curvature bounds, while the gradient descent map is suitable in the lower bounded case. This would also imply that a bilateral bounded Alexandrov space is appropriate for the analysis of the proximal gradient map.

4.1 THE PROXIMAL MAP

Let (X, d) be a complete Alexandrov space with curvature bounded from above by $\overline{\kappa} \in \mathbb{R}$ and $G : X \to \overline{\mathbb{R}}$ a proper, convex, and lower semicontinuous function. Fix the step size $\tau > 0$ and a closed, (geodesically) convex set $A \subset X$ ($\operatorname{diam} A < \pi/(2\sqrt{\overline{\kappa}})$, if $\overline{\kappa} > 0$) containing a nonempty sublevel set of G.

Definition 4.1 (Constrained proximal map [22, Definition 4.1]). For each $x \in X$, we define the constrained proximal map

$$\operatorname{prox}_{\tau G}^A : X \to X, \qquad \operatorname{prox}_{\tau G}^A(x) := \arg\min_{y \in A}\left\{G(y) + \frac{1}{2\tau}d^2(x, y)\right\}.$$

Remark 4.2. It is important to note the nonstandard restriction to the set A in the definition of the proximal map. We will in Section 5 instead work with standard unrestricted proximal maps. To remedy the lack of this restriction in our convergence proofs, we will instead perform locality analysis and restrict the step sizes to not stray too far from the reference points. We will do this for the forward-backward method but one could just as well do it for the standard proximal map (with just an upper curvature bound and without the gradient step) and get similar results.

It can be shown that the proximal map is uniquely determined if G is λ-convex for $\lambda \geq 0$. This follows from the strong convexity of the squared distance function (λ-convex for $\lambda > 0$) and hence from the strong convexity of the whole expression. The following fundamental lemma is key in the analysis of the proximal point method.

Lemma 4.3 ([22, Lemma 4.6(I)])**.** *Let (X, d) be a complete Alexandrov space with curvature bounded from above by $\overline{\kappa} > 0$ and $\operatorname{diam} A < \pi/(2\sqrt{\overline{\kappa}})$. Then we have*

$$d^2(y, \operatorname{prox}^A_{\tau G}(x)) \leq d^2(y, x) + 2\tau(G(y) - G(\operatorname{prox}^A_{\tau G}(x))) \quad \text{for all} \quad x, y \in A.$$

We can consider the algorithm generated by the proximal map by taking an arbitrary starting point $x^0 \in A$ and iterating

$$(4.1) \qquad\qquad x^{k+1} = \operatorname{prox}^A_{\tau_k G}(x^k), \quad k \geq 0,$$

for an appropriate positive sequence $\{\tau_k\}_{k\in\mathbb{N}}$ of step sizes. Using Lemma 4.3, the convergence of the function values $G(x^k)$ to the infimum is shown in [22, Theorem 5.1] for a positive step size sequence $\{\tau_k\}_{k\in\mathbb{N}}$ with $\sum_{k=0}^{\infty} \tau_k = \infty$. Let us also show that the sequence of points with the same step sizes converge to a minimiser of G when X is assumed locally compact.

Theorem 4.4. *Let (X, d) be a complete locally compact Alexandrov space with curvature bounded from above by $\overline{\kappa} \in \mathbb{R}$. Let G be lower semicontinuous and $\{x^k\}_{k\in\mathbb{N}}$ be the sequence generated by (4.1) with $\sum_{k=0}^{\infty} \tau_k = \infty$. If the minimiser set U^* of G is nonempty, then $\lim_{k\to\infty} x^k = x^*$ and $x^* \in U^*$.*

Proof. Since U^* is nonempty, we have $G(y) \leq G(x^k)$ for any $y \in U^*$ and $k \geq 0$. By Lemma 4.3 we have that $d^2(y, x^{k+1}) \leq d^2(y, x^k)$ and hence $d(y, x^{k+1}) \leq d(y, x^k)$ for all $k = 0, 1, \ldots$, i.e., $\{x^k\}_{k\in\mathbb{N}}$ is a Fejér sequence with respect to the nonempty set $U^* \subset X$. By Theorem 2.11, Fejér sequences are bounded and since X is locally compact and complete, the sequence $\{x^k\}_{k\in\mathbb{N}}$ contains a convergent subsequence $\{x^{k_i}\}_{i\in\mathbb{N}}$ such that $\lim_{i\to\infty} x^{k_i} = x^*$. By lower semicontinuity and [22, Theorem 5.1] we have $G(x^*) \leq \liminf_{i\to\infty} G(x^{k_i}) = \inf_{x\in X} G(x)$, which implies that $x^* \in U$. Theorem 2.11 now gives $\lim_{k\to\infty} x^k = x^*$. $\qquad\square$

Without local compactness but assuming strong convexity from G, the algorithm can also be shown to converge to the unique minimiser of G [22, Theorem 5.7]. For completeness, let us also state an iteration-complexity bound for the method.

Proposition 4.5. *Let $\{x^k\}_{k\in\mathbb{N}}$ be the sequence generated by (4.1) with $\tau_k \geq \tau > 0$, for $k = 0, 1, \ldots$. Then for every $N \in \mathbb{N}$, and x^* a point where G achieves a minimum, we have*

$$G(x^{N+1}) - G(x^*) \leq \frac{d^2(x^*, x^0)}{2\tau(N+1)}.$$

Proof. Choose $y = x^*$ in Lemma 4.3 and use $\tau_k \geq \tau > 0$ to get

$$G(x^{k+1}) - G(x^*) \leq \frac{1}{2\tau_k}(d^2(x^*, x^k) - d^2(x^*, x^{k+1})) \leq \frac{1}{2\tau}(d^2(x^*, x^k) - d^2(x^*, x^{k+1})).$$

Summing over from 0 to N and using the fact that the sequence $\{G(x^k)\}_{k\in\mathbb{N}}$ is monotone nonincreasing (choose $y = x$ in Lemma 4.3 to get $G(x^N) \leq G(x^k)$ for all $k = 0, 1\ldots, N$) gives us the result

$$(G(x^{N+1}) - G(x^*))(N+1) = \sum_{k=0}^{N}(G(x^{N+1}) - G(x^*)) \leq \sum_{k=0}^{N}(G(x^{k+1}) - G(x^*)) \leq \frac{d^2(x^*, x^0)}{2\tau}. \quad \square$$

4.2 GRADIENT DESCENT MAP

For a complete finite dimensional Alexandrov space (X, d) with $\partial X = \emptyset$ and curvature bounded from below by $\underline{\kappa} \in \mathbb{R}$, we can define the gradient descent map of a proper, convex, and lower semicontinuous function $F : X \to \overline{\mathbb{R}}$. Fix the step size $\tau > 0$ and a closed, (geodesically) convex set $A \subset X$ containing a nonempty sublevel set of F.

Definition 4.6 (Gradient descent map [22, Definition 4.5]). For each $x \in X$ with $|\nabla_- F|(x) < \infty$, we define

$$\text{g-desc}_{\tau F} : X \to X, \qquad \text{g-desc}_{\tau F}(x) := \text{g-exp}(\tau \nabla(-F)(x)),$$

where g-exp is the gradient exponential map and $\nabla(-F)$ is as given in (3.3).

Remark 4.7. The rigorous definition of the gradient exponential map is very technical and we will not give it here as we will will not use it for our results. Informally, the gradient exponential map $\text{g-exp} : T_p X \to X$ maps a given $v \in T_p X$ to a gradient curve of the distance function from p, i.e., $\text{g-exp}(tv) = \gamma(t)$, where $\gamma : [0, \infty) \to X$ is a radial curve for $d_p(\cdot)$. These radial curves have a parametrisation which coincides with the arc length parametrisation as long as γ is a geodesic, meaning that the gradient exponential agrees with the exponential map at all points where the latter is defined (the exponential map will be defined in Section 5). The notable difference between the maps is that the gradient exponential map is defined for all times $t \geq 0$, whereas the exponential map (if defined at all) is defined only in the cut-locus (which can be arbitrary small). See [1, Chapter 16 J] for the definition and some of its properties.

As in the upper curvature bound case, to prove convergence results we need the corresponding key lemma for the gradient descent map.

Lemma 4.8 ([22, Lemma 4.6(II)]). *Let (X, d) be a finite dimensional complete Alexandrov space with curvature bounded from below by $\underline{\kappa} < 0$, $\partial X = \emptyset$, and $\operatorname{diam} A < \infty$. Then we have*

$$d^2(y, \text{g-desc}_{\tau F}(x)) \leq d^2(y, x) + 2\tau(F(y) - F(x)) + \frac{\lambda_l}{2}(\tau |\nabla_- F|(x))^2$$

for all $x, y \in A$ satisfying $\text{g-desc}_{\tau F}(x) \in A$, where $\lambda_l = \lambda_l(\underline{\kappa}, \operatorname{diam} A)$ is the constant (3.4).

As before, we consider the algorithm generated by the gradient descent map $\text{g-desc}_{\tau F}$ by choosing an arbitrary starting point $x^0 \in A$ and iterating

$$(4.2) \qquad\qquad x^{k+1} = \text{g-desc}_{\tau_k F}(x^k), \quad k \geq 0,$$

for an appropriate positive sequence $\{\tau_k\}_{k \in \mathbb{N}}$. This algorithm is shown to converge to a minimiser of a Lipschitz continuous F [22, Theorem 5.5] and to its unique minimiser when it is also strongly convex [22, Theorem 5.7]. These are done with the same step size assumptions as in the upper curvature bound case ($\tau_k > 0$, $\sum_{k=0}^{\infty} \tau_k = \infty$, and $\sum_{k=0}^{\infty} \tau_k^2 < \infty$). For these step sizes, we have the following iteration-complexity bound for the gradient descent algorithm.

Proposition 4.9. *Let F be Lipschitz continuous with constant L_F and $\{x^k\}_{k \in \mathbb{N}}$ be the sequence generated by (4.2). Let $\alpha_k := \tau_k |\nabla_- F|(x^k)$, for $k = 0, 1, \dots$. Then for every $N \in \mathbb{N}_{\geq 0}$, the following holds*

$$\min\{F(x^k) - F(x^*) \mid k = 0, 1, \dots, N\} \leq L_F \frac{d^2(x^0, x^*) + \frac{\lambda_l}{2} \sum_{k=0}^{N} \alpha_k^2}{2 \sum_{k=0}^{N} \alpha_k},$$

where x^* is a point in which F achieves a minimum.

Proof. Choose $y = x^*$ and insert $\tau_k = \alpha_k / |\nabla_- F|(x^k)$ in Lemma 4.8 to get

$$2 \frac{\alpha_k}{|\nabla_- F|(x^k)} (F(x^k) - F(x^*)) \leq d^2(x^k, x^*) - d^2(x^{k+1}, x^*) + \frac{\lambda_l}{2} \alpha_k^2$$

for all $k = 0, 1, \ldots, N$. From the Lipschitz continuity of F we have that $|\nabla_- F|(x^k) \leq L_F$, so summing over from 0 to N we get

$$\frac{2}{L_F} \sum_{k=0}^{N} \alpha_k (F(x^k) - F(x^*)) \leq d^2(x^0, x^*) - d^2(x^{N+1}, x^*) + \frac{\lambda_l}{2} \sum_{k=0}^{N} \alpha_k^2.$$

Simply noticing now that $\sum_{k=0}^{N} \alpha_k (F(x^k) - F(x^*)) \geq \min\{F(x^k) - F(x^*) \mid k = 0, 1, \ldots, N\} \sum_{k=0}^{N} \alpha_k$ gives us

$$\min\{F(x^k) - F(x^*) \mid k = 0, 1, \ldots, N\} \sum_{k=0}^{N} \alpha_k \leq L_F \frac{d^2(x^0, x^*) + \frac{\lambda_l}{2} \sum_{k=0}^{N} \alpha_k^2}{2}$$

which when divided by the sum $\sum_{k=0}^{N} \alpha_k$ gives us the result. $\qquad\square$

5 PROXIMAL GRADIENT MAP

Now we are ready to define and analyse the proximal gradient map. Let (X, d) be a complete locally compact Alexandrov space with $\partial X = \emptyset$ and curvature bounded from above and below by $\overline{\kappa}, \underline{\kappa} \in \mathbb{R}$, respectively ($\underline{\kappa} \leq \overline{\kappa}$). Fix the step size $\tau > 0$ and let $G, F : X \to \overline{\mathbb{R}}$ be proper, convex, and lower semicontinuous functions. The proximal gradient map will use a different version of the gradient map relying on the exponential map instead of the gradient exponential map.

Definition 5.1 (Exponential map). The exponential map is a mapping from a subset of $T_x X$ into X defined by $\exp_x(tv) = \gamma(t)$, where $\gamma : [0, \delta] \to X$ is a geodesic, with $\gamma(0) = x$ and $\gamma^+(0) = v$.

Geometrically, the exponential map takes a geodesic direction v at a starting point x and travels to that direction for distance t arriving at the point $\gamma(t)$. Clearly this is well-defined for at least sufficiently short vectors (as long as the curve γ stays minimal). In fact, as X is a bilaterally bounded Alexandrov space, it is a topological manifold and hence has infinitely extendible (to both sides) local geodesics by [1, Exercise 9.9]. This, along with local compactness, implies that the space of directions $\Sigma'_x X$ (and hence $T'_x X$ as well) is complete for any $x \in X$ [1, Exercise 9.10]. Thus, if we allow our geodesics to be only locally minimising, we see that the exponential map is defined for any vector in $T_x X$. This will be used in the a priori locality Lemma 6.4 but will subsequently be strengthened to geodesics that lie in a geodesically convex set. The exponential map also enjoys Lipschitz continuity.

Lemma 5.2 ([18, 2.3]). *The exponential map is Lipschitz continuous with a constant $L_{\exp} := 1 + L(\overline{\kappa}, r)$, where $L(\overline{\kappa}, r) \geq 0$ and r is a radius of the origin where the exponential map is defined, with $L(\overline{\kappa}, r) \to 0$ as $r \to 0$ (and naturally as $\overline{\kappa} \to 0$).*

We can now reformulate the gradient descent map.

Definition 5.3 (Gradient descent map, bilateral bounds). For each $x \in X$ with $|\nabla_- F|(x) < \infty$, we define

$$\mathrm{desc}_{\tau F} : X \to X, \qquad \mathrm{desc}_{\tau F}(x) := \exp_x(\tau \nabla(-F)(x)),$$

where exp is the exponential map and $\nabla(-F)$ is as given in (3.3).

Combining our newly defined gradient descent map with the standard unrestricted proximal map gives us the proximal gradient map.

Definition 5.4 (Proximal gradient map). For each $x \in X$ with $|\nabla_- F|(x) < \infty$, we define

$$J : X \to X, \qquad J(x) := \mathrm{prox}_{\tau G}(\mathrm{desc}_{\tau F}(x)) = \underset{y \in X}{\arg\min} \left\{ G(y) + \frac{1}{2\tau} d^2(\exp_x(\tau \nabla(-F)(x)), y) \right\}.$$

To analyse the proximal gradient map, we need a result akin to Lemmas 4.3 and 4.8. To do so, we need an optimality condition, refine Lemma 4.3 to include the strong convexity factor of the squared distance function, modify Lemma 4.8 to include the squared distance function (replacing the squared vector term), and use a smoothness condition for F to move from $F(x)$ to $F(J(x))$.

We can provide a necessary condition characterisation of the point $J(x)$ similarly to the basic proximal map in [22, (4.2)].

Lemma 5.5. *Let (X, d) be a locally compact Alexandrov space with an (upper or lower) curvature bound κ and $G : X \to \overline{\mathbb{R}}$ proper, lower semicontinuous, and λ-convex with $\lambda \geq 0$. Then for all unit-speed geodesics $\gamma, \eta : [0,1] \to X$ with $\gamma(0) = \mathrm{prox}_{\tau G}(x)$ and $\eta(t) = \mathrm{prox}_{\tau G}(x) \#_t x$ such that $0 < d(\mathrm{prox}_{\tau G}(x), x) < \pi/\sqrt{\kappa}$, we have*

$$D_{\mathrm{prox}_{\tau G}(x)} G(\gamma^+(0)) - \frac{1}{\tau} \langle \gamma^+(0), \eta^+(0) \rangle \geq 0.$$

Proof. By Definition 5.4 and the Fermat principle of Corollary 3.5, $\mathrm{prox}_{\tau G}(x)$ solves the minimisation problem for the proximal map if and only if $D_{\mathrm{prox}_{\tau G}(x)}(G + (2\tau)^{-1} d_x^2)(\gamma^+(0)) \geq 0$ for all $\gamma^+(0) \in T_{\mathrm{prox}_{\tau G}(x)} X$. The first variation formula of Theorem 2.6 then shows that $D_{\mathrm{prox}_{\tau G}(x)}(2\tau)^{-1} d_x^2(\gamma^+(0)) = -\tau^{-1} \langle \gamma^+(0), \eta^+(0) \rangle$ for all unit-speed geodesics $\eta : [0,1] \to X$ with $\eta(t) = \mathrm{prox}_{\tau G}(x) \#_t x$. □

To refine Lemma 4.3, we use Lemma 3.9, the optimality condition of the proximal map, and the convexity of G to get the appropriate estimate. The following few lemmas are a posteriori estimates that will require first showing that $\mathrm{prox}_{\tau G}(x), \mathrm{desc}_{\tau F}(x) \in A$.

Lemma 5.6. *Let (X, d) be a complete Alexandrov space with curvature bounded from above by $\overline{\kappa}$ and $G : X \to \overline{\mathbb{R}}$ proper, lower semicontinuous, and convex. Let $A \subset X$ be such that $\mathrm{diam}\, A < \pi/(2\sqrt{\overline{\kappa}})$. Then we have*

$$d^2(y, \mathrm{prox}_{\tau G}(x)) \leq d^2(y, x) + 2\tau(G(y) - G(\mathrm{prox}_{\tau G}(x))) - \frac{\lambda_u}{2} d^2(x, \mathrm{prox}_{\tau G}(x)),$$

for all $x, y \in A$ satisfying $\mathrm{prox}_{\tau G}(x) \in A$, where $\lambda_u = \lambda_u(\overline{\kappa}, \mathrm{diam}\, A)$ is given in (3.6).

Proof. Choose $a = d(y, x)$, $b = d(x, \mathrm{prox}_{\tau G}(x))$, and $c = d(y, \mathrm{prox}_{\tau G}(x))$ in Lemma 3.9 to get

$$
\begin{aligned}
(5.1) \quad d^2(y, x) &\geq \frac{\sqrt{\overline{\kappa}}c}{\tan(\sqrt{\overline{\kappa}}c)} d^2(x, \mathrm{prox}_{\tau G}(x)) + d^2(y, \mathrm{prox}_{\tau G}(x)) \\
&\quad - 2d(x, \mathrm{prox}_{\tau G}(x)) d(y, \mathrm{prox}_{\tau G}(x)) \cos(\theta) \\
&\geq \frac{\lambda_u}{2} d^2(x, \mathrm{prox}_{\tau G}(x)) + d^2(y, \mathrm{prox}_{\tau G}(x)) - 2d(x, \mathrm{prox}_{\tau G}(x)) d(y, \mathrm{prox}_{\tau G}(x)) \cos(\theta),
\end{aligned}
$$

where $\theta = \angle_{\mathrm{prox}_{\tau G}(x)}(\gamma, \eta)$ is the angle between the sides b and c, and $\gamma(t) = \mathrm{prox}_{\tau G}(x) \#_t y$ and $\eta(t) = \mathrm{prox}_{\tau G}(x) \#_t x$. The definition of the "inner product", followed by Lemma 5.5 and convexity (3.1) of G along the geodesic γ, establish

$$
\begin{aligned}
(5.2) \quad -2d(x, \mathrm{prox}_{\tau G}(x)) d(y, \mathrm{prox}_{\tau G}(x)) \cos(\theta) &= -2\langle \gamma^+(0), \eta^+(0) \rangle \\
&\geq -2\tau D_{\mathrm{prox}_{\tau G}(x)} G(\gamma^+(0)) \\
&\geq 2\tau(G(\mathrm{prox}_{\tau G}(x)) - G(y)).
\end{aligned}
$$

Combining (5.1) and (5.2) establishes the claim. □

To modify Lemma 4.8, we use Lemma 3.7, the optimality condition of the gradient map (3.2), and convexity of F to get the appropriate estimate.

Lemma 5.7. *Let (X, d) be a complete Alexandrov space with curvature bounded from above by $\overline{\kappa}$ and below by $\underline{\kappa}$, and $F : X \to \overline{\mathbb{R}}$ proper, lower semicontinuous, and convex. Let $A \subset X$ be such that $\operatorname{diam} A < \pi/(2\sqrt{\overline{\kappa}})$ and $0 < |\nabla_- F|(x) < \infty$, for all $x \in A$. Then we have*

$$d^2(y, \operatorname{desc}_{\tau F}(x)) \leq d^2(y, x) + 2\tau(F(y) - F(x)) + \frac{\lambda_l}{2}d^2(x, \operatorname{desc}_{\tau F}(x)),$$

for all $x, y \in A$ satisfying $\operatorname{desc}_{\tau F}(x) \in A$, where $\lambda_l = \lambda_l(\underline{\kappa}, \operatorname{diam} A)$ is the constant (3.4).

Proof. Since A is assumed geodesically convex and $x, y, \operatorname{desc}_{\tau F}(x) \in A$, we have by Lemma 3.7

$$d^2(y, \operatorname{desc}_{\tau F}(x)) \leq d^2(y, x) + \frac{\lambda_l}{2}d^2(x, \operatorname{desc}_{\tau F}(x)) - 2d(x, y)d(x, \operatorname{desc}_{\tau F}(x)) \cos(\theta).$$

It follows from $\operatorname{desc}_{\tau F}(x) = \exp_x(\tau \nabla(-F)(x))$ and the definition of the "inner product" for geodesics $\eta(t) = x \#_t \exp_x(\tau \nabla(-F)(x))$ and $\gamma(t) = x \#_t y$ that

$$-2d(x, y)d(x, \operatorname{desc}_{\tau F}(x)) \cos(\theta) = -2\langle \eta^+(0), \gamma^+(0)\rangle = -2\langle \tau \nabla(-F)(x), \gamma^+(0)\rangle.$$

Furthermore, by convexity along γ and the characterisation (3.2) we have

$$F(y) - F(x) \geq D_x F(\gamma^+(0)) \geq -\langle \nabla(-F)(x), \gamma^+(0)\rangle.$$

The claim follows by combining the estimates. $\qquad\square$

Finally, we need to introduce the concept of uniform smoothness of a function, which is a natural reformulation of the same concept in Banach spaces and a necessary assumption in analysing the proximal gradient map.

Definition 5.8 (Uniform smoothness). A function $F : X \to \overline{\mathbb{R}}$ is called uniformly smooth with factor $L > 0$, if for all $x, y \in X$ and geodesics $\gamma(t) = x \#_t y$ with $t \in [0, 1]$, we have

$$F(\gamma(t)) + \frac{L}{2}t(1 - t)d^2(x, y) \geq (1 - t)F(x) + tF(y).$$

Example 5.9. Let X be an Alexandrov space with curvature bounded from above. The beginning of Section 3.1 shows that the squared distance function d_z^2, for a fixed $z \in X$, is $-\lambda$-smooth, where λ is the factor from Proposition 3.3.

Using uniform smoothness to derive a suitable descent lemma, we need to also assume differentiability from F. This assumption is justified as our space X is bilaterally bounded and thus contains the structure of a differentiable manifold by Theorem 2.8. For a differentiable F with a gradient $\nabla F(x)$ at x, we have the correspondence

$$\nabla(-F)(x) = -\nabla F(x) \quad \text{and} \quad D_x F(\gamma^+(0)) = \langle \nabla F(x), \gamma^+(0)\rangle,$$

for all geodesics γ with $\gamma(0) = x$ [2, Chapter 1.3]. Note that for critical points x^* of F we have $\nabla F(x^*) = 0$.

Lemma 5.10 (Descent lemma). *Let (X, d) be a complete locally compact Alexandrov space and F be uniformly smooth with factor $L > 0$. We have (for any instance of the possibly non-unique $\log_x y$ when (X, d) is not upper-bounded)*

$$(5.3) \qquad\qquad D_x F(\log_x y) + \frac{L}{2}d^2(x, y) \geq F(y) - F(x).$$

If, (X, d) is, moreover, bilaterally bounded, and F differentiable, then, for $\gamma(t) = x \#_t y$, we have

$$-\langle \gamma^+(0), \nabla(-F)(x)\rangle + \frac{L}{2}d^2(x, y) + F(x) \geq F(y).$$

Proof. By the uniform smoothness we have

$$\frac{F(\gamma(t)) - F(x)}{t} + \frac{L}{2}(1-t)d^2(x, y) \geq F(y) - F(x),$$

which letting $t \searrow 0$ yields

$$D_x F(\gamma^+(0)) + \frac{L}{2}d^2(x, y) \geq F(y) - F(x).$$

That is, (5.3) holds. Lastly, the bilateral bounds and the differentiability of F give $D_x F(\gamma^+(0)) = \langle \gamma^+(0), \nabla F(x) \rangle = -\langle \gamma^+(0), \nabla(-F)(x) \rangle$. □

5.1 THE KEY ESTIMATE

We now have the necessary tools to derive our key inequality for the proximal gradient map. The plan is to combine Lemmas 5.6 and 5.7 at specific points, using the uniform smoothness of F to move from the estimates on F from being based at x to $J(x)$ (to be able combine with the estimate of G at $J(x)$ in Lemma 5.6).

By Lemma 5.7 we have

$$(5.4) \qquad d^2(y, w) \leq d^2(y, x) + 2\tau(F(y) - F(x)) + \frac{\lambda_l}{2}d^2(x, w),$$

where $w = \mathrm{desc}_{\tau F}(x) = \exp_x(\tau \nabla(-F)(x))$. Assuming that F is uniformly smooth with factor $L > 0$, we have by Lemma 5.10 that

$$(5.5) \qquad -2\tau F(x) \leq -2\tau F(J(x)) + L\tau d^2(x, J(x)) - 2\langle \gamma^+(0), \tau \nabla(-F)(x) \rangle$$

for a geodesic $\gamma(t) = x \#_t J(x)$. By Lemma 3.9 we have

$$(5.6) \qquad -2\langle \gamma^+(0), \tau \nabla(-F)(x) \rangle = -2d(x, J(x))d(x, w)\cos(\theta)$$

$$\leq d^2(w, J(x)) - \frac{\lambda_u}{2}d^2(x, J(x)) - d^2(x, w),$$

where θ is the angle between geodesics from x to $J(x)$ and w. Combining expressions (5.4)–(5.6) gives us

$$(5.7) \qquad d^2(y, w) \leq d^2(y, x) + 2\tau(F(y) - F(J(x))) + (L\tau - \frac{\lambda_u}{2})d^2(x, J(x))$$

$$+ \left(\frac{\lambda_l}{2} - 1\right)d^2(x, w) + d^2(w, J(x)).$$

Finally, let $x = w$ in Lemma 5.6 to obtain

$$(5.8) \qquad d^2(y, J(x)) \leq d^2(y, w) + 2\tau(G(y) - G(J(x))) - \frac{\lambda_u}{2}d^2(w, J(x)),$$

and sum estimates (5.7) and (5.8) to obtain

$$d^2(y, J(x)) \leq d^2(y, x) + 2\tau(H(y) - H(J(x))) + (L\tau - \frac{\lambda_u}{2})d^2(x, J(x))$$

$$+ \left(\frac{\lambda_l}{2} - 1\right)d^2(x, w) + \left(1 - \frac{\lambda_u}{2}\right)d^2(w, J(x)),$$

where $H(x) = G(x) + F(x)$. We have proved the following key result. It is an a posteriori result, which still requires showing that $w, J(x) \in A$.

Lemma 5.11 (Key lemma, convex G and F). *Let (X, d) be a complete locally compact Alexandrov space with $\partial X = \emptyset$ and curvature bounded from above by $\overline{\kappa} \in \mathbb{R}$ and below by $\underline{\kappa} \in \mathbb{R}$ ($\underline{\kappa} \le \overline{\kappa}$). Fix the step size $\tau > 0$ and a closed, geodesically convex set $A \subset X$ (diam $A < \pi/(2\sqrt{\overline{\kappa}})$, if $\overline{\kappa} > 0$) containing a nonempty sublevel set of $G + F$, where $G, F : X \to \overline{\mathbb{R}}$ are proper, convex, and lower semicontinuous functions. Assume further that F is differentiable and uniformly smooth with factor $L > 0$. Then we have*

$$(5.9) \qquad d^2(y, J(x)) \le d^2(y, x) + 2\tau(H(y) - H(J(x))) + (L\tau - \frac{\lambda_u}{2})d^2(x, J(x))$$

$$+ \left(\frac{\lambda_l}{2} - 1\right)d^2(x, w) + \left(1 - \frac{\lambda_u}{2}\right)d^2(w, J(x))$$

for all $x, y \in A$ satisfying $w, J(x) \in A$, where $H(x) = G(x) + F(x)$, $w = \exp_x(\tau\nabla(-F)(x))$, along with

$$\frac{\lambda_u}{2} := \frac{\sqrt{\overline{\kappa}} \operatorname{diam} A}{\tan(\sqrt{\overline{\kappa}} \operatorname{diam} A)} \quad and \quad \frac{\lambda_l}{2} = \frac{\sqrt{-\underline{\kappa}} \operatorname{diam} A}{\tanh(\sqrt{-\underline{\kappa}} \operatorname{diam} A)}.$$

Much like the corresponding Lemmas 4.3 and 4.8 for the proximal and gradient descent maps, respectively, this estimate will be used to prove convergence results related to the iteration scheme defined by the proximal gradient map.

6 FORWARD–BACKWARD METHOD

We can now consider the algorithm generated by the proximal gradient map J of Definition 5.4. Fixing an arbitrary starting point $x^0 \in X$, the forward-backward method is defined as the iterative process

$$(6.1) \qquad\qquad x^{k+1} = J(x^k), \quad k \ge 0$$

for an appropriately chosen positive sequence $\{\tau_k\}_{k \in \mathbb{N}}$. We generally denote the intermediate step by $w^k = \operatorname{desc}_{\tau F}(x^k) = \exp_{x^k}(\tau_k \nabla(-F)(x^k))$, so that $x^{k+1} = \operatorname{prox}_{\tau_k G}(w^k)$.

Before showing that method (6.1) converges to a minimum of the function H, let us do some preliminary analysis for cases of specific curvature bounds. Denote complete locally compact Alexandrov spaces bounded from above by $\overline{\kappa}$ by CAT($\overline{\kappa}$), and by those bounded from below by $\underline{\kappa}$ by CBB($\underline{\kappa}$) ($\underline{\kappa} \le \overline{\kappa}$). Let us also assume that we can use Lemma 5.11 for all $k \ge 0$.

- CAT(0) + CBB(0) (flat case: e.g. $\mathbb{R}^n$ or a Hilbert space): now $\lambda_u/2 = 1 = \lambda_l/2$, giving us the standard estimate

$$(6.2) \qquad d^2(y, x^{k+1}) \le d^2(y, x^k) + 2\tau_k(H(y) - H(x^{k+1})) + (L\tau_k - 1)d^2(x^k, x^{k+1})$$

  for all y. Assuming $0 < L\tau_{\min} \le L\tau_k < 1$ for all $k \ge 0$ establishes the Fejér property of the sequence and leads to the usual convergence results.

- CAT(0) + CBB($\underline{\kappa}$) (e.g. a Hadamard space with an additional lower curvature bound): $\lambda_u/2 = 1$, giving us the estimate

$$(6.3) \qquad d^2(y, x^{k+1}) \le d^2(y, x^k) + 2\tau_k(H(y) - H(x^{k+1}))$$

$$+ (L\tau_k - 1)d^2(x^k, x^{k+1}) + \left(\frac{\lambda_l}{2} - 1\right)d^2(x^k, w^k).$$

  Since $\lambda_l/2 - 1 \ge 0$, we need to estimate the term $d^2(x^k, w^k)$, for which, by Lemma 5.2, we get from Lipschitz continuity of the exponential map (see the dicussion after Definition 5.1) that $d^2(x^k, w^k) \le (L_{\exp}\tau_k|\nabla_-F|(x^k))^2$. By further assuming that F is, for example, Lipschitz continuous with factor L_F, we have $|\nabla_-F|(x^k) \le L_F$ and $(L_{\exp}\tau_k|\nabla_-F|(x^k))^2 \le (L_{\exp}\tau_k L_F)^2$. This shows that along with the previous assumption of $L\tau_k < 1$, we also need to assume $\sum_{k=0}^{\infty} \tau_k^2 < \infty$ to obtain convergence results.

- CAT($\overline{\kappa}$) + CBB(0): now $\lambda_l/2 = 1$, giving us the estimate

$$
\begin{aligned}
(6.4) \qquad d^2(y, x^{k+1}) &\leq d^2(y, x^k) + 2\tau_k(H(y) - H(x^{k+1})) \\
&\quad + (L\tau_k - \frac{\lambda_u}{2})d^2(x^k, x^{k+1}) + \left(1 - \frac{\lambda_u}{2}\right) d^2(w^k, x^{k+1}).
\end{aligned}
$$

Since $1 - \lambda_u/2 \geq 0$, we need to estimate the term $d^2(w^k, x^{k+1})$ as well as assume that $L\tau_k < \lambda_u/2$. One way of dealing with $d^2(w^k, x^{k+1})$ is to use the definition of the proximal map to get $G(x^{k+1}) + (2\tau_k)^{-1}d^2(w^k, x^{k+1}) \leq G(w^k)$. Assume further that G is Lipschitz continuous with factor L_G so that $d^2(w^k, x^{k+1}) \leq 2\tau_k[G(w^k) - G(x^{k+1})] \leq 2\tau_k L_G d(w^k, x^{k+1})$ giving $d^2(w^k, x^{k+1}) \leq 4\tau_k^2 L_G^2$. This again suggests that we need to require $\sum_{k=0}^{\infty} \tau_k^2 < \infty$. However, the Lipschitz assumption on G is rather unpleasant as it excludes, for example, the indicator function.

- CAT($\overline{\kappa}$) + CBB($\underline{\kappa}$): here we cannot simplify at all and simply have the original estimate (5.9). For convergence results, we should need all of the previous assumptions: $L\tau_k < \lambda_u/2$, $\sum_{k=0}^{\infty} \tau_k^2 < \infty$, and a bound for $d^2(w^k, x^{k+1})$.

It appears to be clear from these points, that in the presence of nonzero curvature, the rather strict assumption of square summability of the step sizes is essential in obtaining convergence results. This ultimately arises from the strong convexity and strong concavity factors of the squared distance function in spaces with upper and lower curvature bounds, respectively.

6.1 CONVERGENCE ANALYSIS

Before analysing convergence, we gather the fundamental assumptions regarding the forward-backward algorithm here.

Assumption 6.1. The space (X, d) is a complete locally compact bilaterally bounded Alexandrov space with $\partial X = \emptyset$ (from below by $\underline{\kappa} \in \mathbb{R}$ and above by $\overline{\kappa} \in \mathbb{R}$, $\underline{\kappa} \leq \overline{\kappa}$).

Assumption 6.2. The functions $G, F : X \to \overline{\mathbb{R}}$ are proper, convex, and lower semicontinuous and F is differentiable and uniformly smooth with factor $L > 0$ (to be precise, these only need to apply in a closed geodesically convex set containing a nonempty sublevel set of $H = G + F$).

We first start by showing that the forward-backward method (6.1) creates a monotone sequence of function values of H which in turn shows the convergence of the iterates. This is still an a posteriori result, which requires first showing that $x^{k+1}, \text{desc}_{\tau F}(x^k) \in A$.

Proposition 6.3 (A posteriori monotonicity). *Along with Assumption 6.1, let G be lower semicontinuous and F differentiable, and uniformly smooth with factor $L > 0$. Fix a closed, geodesically convex set $A \subset X$ containing a nonempty sublevel set of $G + F$ and, if $\overline{\kappa} > 0$, such that $\text{diam}\, A < \pi/(2\sqrt{\overline{\kappa}})$. Let $\{\tau_k\}_{k \in \mathbb{N}}$ be a positive sequence with $L\tau_k < \lambda_u/2$ for all $k \geq 0$ If the points x^k, x^{k+1}, and $w^k = \text{desc}_{\tau F}(x^k)$ generated by the forward-backward method (6.1) belong to A, then*

$$
(6.5) \qquad H(x^{k+1}) \leq H(x^k) + (\frac{L}{2} - \frac{\lambda_u}{4\tau_k})d^2(x^k, x^{k+1}).
$$

In particular, the sequence $\{H(x^k)\}_{k \in \mathbb{N}}$ is monotonically nonincreasing. Furthermore, if $x^0 \in \text{dom}\, H$ and $\inf H > -\infty$, then $d(x^k, x^{k+1}) \to 0$ as $k \to \infty$.

Proof. Definition 5.4 and the iteration scheme (6.1) gives

$$
(6.6) \qquad G(x^{k+1}) + \frac{1}{2\tau_k}d^2(w^k, x^{k+1}) \leq G(x^k) + \frac{1}{2\tau_k}d^2(w^k, x^k),
$$

where $w^k = \mathrm{desc}_{\tau_k F}(x^k) = \exp_{x^k}(\tau_k \nabla(-F)(x^k))$. By (6.6) and Lemma 5.10 we have

$$(6.7) \quad H(x^{k+1}) = G(x^{k+1}) + F(x^{k+1})$$

$$\leq H(x^k) + \frac{L}{2} d^2(x^k, x^{k+1}) + \frac{1}{2\tau_k}\left(d^2(w^k, x^k) - d^2(w^k, x^{k+1})\right) - \langle \gamma^+(0), \nabla(-F)(x^k)\rangle,$$

for a geodesic $\gamma(t) = x^k \#_t x^{k+1}$. Now use Lemma 3.9, (3.5) (i.e., our replacement of the Pythagoras' identity in Hilbert spaces) and (2.3) to estimate

$$(6.8) \qquad d^2(w^k, x^k) - d^2(w^k, x^{k+1}) \leq -\frac{\lambda_u}{2} d^2(x^k, x^{k+1}) + 2 d(x^k, x^{k+1}) d(x^k, w^k) \cos(\theta)$$

$$= -\frac{\lambda_u}{2} d^2(x^k, x^{k+1}) + 2\langle \gamma^+(0), \tau_k \nabla(-F)(x^k)\rangle.$$

Putting (6.7) and (6.8) together gives (6.5). The latter implies that the sequence $\{H(x^k)\}_{k \in \mathbb{N}}$ is monotonically nonincreasing when $L\tau_k < \lambda_u/2$ for all $k \geq 0$.

Now let ε satisfy $\lambda_u/(4\tau_k) - L/2 \geq \varepsilon > 0$ so that (6.5) yields

$$\varepsilon d^2(x^k, x^{k+1}) \leq \left(\frac{\lambda_u}{4\tau_k} - \frac{L}{2}\right) d^2(x^k, x^{k+1}) \leq H(x^k) - H(x^{k+1}).$$

Let $x^0 \in \mathrm{dom}\, H$ and sum over from 0 to N to get

$$\varepsilon \sum_{k=0}^{N} d^2(x^k, x^{k+1}) \leq \sum_{k=0}^{N} (H(x^k) - H(x^{k+1})) = H(x^0) - H(x^{N+1}) \leq H(x^0) - \inf H < \infty.$$

It follows that $d(x^k, x^{k+1}) \to 0$ since $\varepsilon > 0$. $\qquad\square$

As Lemma 5.11 and Proposition 6.3 need the iterates to stay in the same neighborhood as the reference points, we need to perform some locality analysis to determine when this is true, i.e., how close do we need to start from the reference point and how small do the step sizes need to be for the statement to hold. The next a priori lemma proves that $x^{k+1}, w^k = \mathrm{desc}_{\tau F}(x^k) \in A$ as required by the former a posteriori results. Note also that here we allow the geodesics from x^k to w^k to be just locally minimizing. This discrepancy will be resolved in the convergence proofs with an appropriate diameter bound for A.

Lemma 6.4 (A priori locality). *Along with Assumption 6.1, let $G, F : X \to \overline{\mathbb{R}}$, where G is proper and F is differentiable and Lipschitz continuous with factor L_F. Let $\{x^k\}_{k \in \mathbb{N}}$ be generated by (6.1) and x^* be a minimiser of $G + F$. Let $x^k \in \bar{B}(x^*, r)$ for some $r > 0$. If $R > 2r$ and*

$$(6.9) \qquad\qquad 0 < \tau_k \leq \frac{R - 2r}{2 L_F (1 + L_{\exp})},$$

then $w^k, x^{k+1} \in \bar{B}(x^*, R)$.

Proof. The definition of the proximal map gives

$$G(x^{k+1}) + \frac{1}{2\tau_k} d^2(x^{k+1}, w^k) \leq G(y) + \frac{1}{2\tau_k} d^2(y, w^k) \quad \text{for all} \quad y \in X,$$

where $x^{k+1} = \mathrm{prox}_{\tau_k G}(w^k)$ and $w^k = \exp_{x^k}(\tau_k \nabla(-F)(x^k))$. Since $G(y) + F(y) \leq G(x^{k+1}) + F(x^{k+1})$ for a minimising point $y = x^*$, then immediately

$$d^2(x^{k+1}, w^k) \leq 2\tau_k(F(x^{k+1}) - F(x^*)) + d^2(x^*, w^k) \leq 2\tau_k L_F d(x^{k+1}, x^*) + d^2(x^*, w^k)$$

by the Lipschitz continuity of F. By this and the triangle inequality

$$d(x^{k+1}, x^*) \leq d(x^{k+1}, w^k) + d(x^*, w^k) \leq \sqrt{2\tau_k L_F d(x^{k+1}, x^*) + d^2(x^*, w^k)} + d(x^*, w^k).$$

Thus, rearranged and squared

$$d^2(x^{k+1}, x^*) + d^2(x^*, w^k) - 2d(x^{k+1}, x^*)d(x^*, w^k) = (d(x^{k+1}, x^*) - d(x^*, w^k))^2$$
$$\leq 2\tau_k L_F d(x^{k+1}, x^*) + d^2(x^*, w^k).$$

That is,

$$d^2(x^{k+1}, x^*) - 2d(x^{k+1}, x^*)d(x^*, w^k) \leq 2\tau_k L_F d(x^{k+1}, x^*),$$

which divided by $d(x^{k+1}, x^*)$ when this quantity is non-zero, gives

$$d(x^{k+1}, x^*) \leq 2\tau_k L_F + 2d(x^*, w^k).$$

Obviously, this inequality also holds when $d(x^{k+1}, x^*) = 0$. By the triangle inequality, Lemma 5.2, and the Lipschitz continuity of F and of the exponential map (see the dicussion after Definition 5.1), we have

$$(6.10) \qquad d(x^*, w^k) \leq d(x^*, x^k) + d(x^k, \exp_{x^k}(\tau_k \nabla(-F)(x^k)))$$
$$\leq d(x^*, x^k) + L_{\exp} \tau_k |\nabla_- F|(x^k)$$
$$\leq d(x^*, x^k) + L_F L_{\exp} \tau_k.$$

Thus,

$$(6.11) \qquad d(x^{k+1}, x^*) \leq 2d(x^*, x^k) + 2L_F(1 + L_{\exp})\tau_k.$$

Now, using (6.9) and $x^k \in \bar{B}(x^*, r)$ in (6.10) and (6.11), respectively, we obtain

$$d(x^*, w^k) \leq r + L_F L_{\exp} \frac{R - 2r}{2L_F(1 + L_{\exp})} \leq r + L_{\exp} \frac{R - 2r}{2L_{\exp}} = \frac{R}{2} < R,$$

as well as

$$d(x^{k+1}, x^*) \leq 2r + 2L_F(1 + L_{\exp}) \frac{R - 2r}{2L_F(1 + L_{\exp})} = R.$$

In other words, $x^{k+1}, w^k \in \bar{B}(x^*, R)$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

Remark 6.5 (Combined step with lower bound only). For any instance of the possibly non-unique (when (X, d) is not upper-bounded) $\log_{x^k} x$, let us define x^{k+1} as a minimiser of the *total surrogate*

$$\varphi_k(x) := G(x) + F(x^k) + D_{x^k} F(\log_{x^k} x) + \frac{1}{2\tau_k} d^2(x^k, x).$$

In a Hilbert space, this step is equivalent to (6.1). In a general Alexandrov space, we may not have equivalency. Nevertheless, we have by definition that $\varphi_k(x^{k+1}) \leq \varphi_k(x^k)$, which combined with the descent inequality (5.3) for an uniformly smooth F yields

$$[F + G](x^{k+1}) + \frac{\tau_k^{-1} - L}{2} d^2(x^k, x^{k+1}) \leq [F + G](x^k).$$

Thus, by standard technique, if $F + G > -\infty$ and $\sup_k \tau_k L < 1$, we obtain $\sum_{k=0}^{\infty} d^2(x^k, x^{k+1}) < \infty$, hence $d^2(x^k, x^{k+1}) \to 0$. In a Hilbert space, this property readily implies convergence of subdifferentials to zero; see, e.g, [15, Theorem 4.19]. In a general Alexandrov space, such a characterisation appears more challenging to obtain, however, assuming for a subsequence that both x^{k_ℓ} and $x^{k_\ell+1}$ are, e.g., on the same polyhedron of a polyhedral complex, we can use the Hilbert space characterisation to obtain subdifferential convergence for the subsequence. Through the equivalency to (6.1) in this case, this holds even for our proposed method with split steps.

When G and F are assumed to be Lipschitz continuous, the convergence of the forward-backward method to a minimiser of H is an application of Lemma 6.4 and the following deterministic variant of the Siegmund–Robbins lemma.

Lemma 6.6 ([24]). *Let $\{a_k\}_{k\in\mathbb{N}}$, $\{b_k\}_{k\in\mathbb{N}}$ and $\{c_k\}_{k\in\mathbb{N}}$ be sequences of nonnegative real numbers such that $a_{k+1} \leq a_k - b_k + c_k$ for each $k \geq 0$ and $\sum_{k=0}^{\infty} c_k < \infty$. Then the sequence $\{a_k\}_{k\in\mathbb{N}}$ converges and $\sum_{k=0}^{\infty} b_k < \infty$.*

Theorem 6.7 (Convergence to a minimiser, Lipschitz continuous G and F). *Along with Assumptions 6.1 and 6.2, let G and F be Lipschitz continuous with factors L_G and L_F, respectively. Let $x^0 \in \bar{B}(x^*, r)$ for $r > 0$ and $R > 2r$ such that $R < \pi/(4\sqrt{\kappa})$. Let $\{\tau_k\}_{k\in\mathbb{N}}$ be a positive sequence of step sizes satisfying $\sum_{k=0}^{\infty} \tau_k = \infty$, $\sum_{k=0}^{\infty} \tau_k^2 < \infty$, and*

$$(6.12) \qquad \tau_k \leq \min\left\{ \frac{\lambda_u}{2L}, \frac{R - 2r}{2L_F(1 + L_{\exp})} \right\}$$

for all $k \geq 0$. If $\inf_{x\in X} H(x)$ is attained at some point, then the sequence $\{x^k\}_{k\in\mathbb{N}}$ created by the iterative process (6.1) converges to some minimiser of H in $\bar{B}(x^*, R)$ as $k \to \infty$.

Proof. Set $A := \bar{B}(x^*, R)$ so that $\operatorname{diam} A \leq 2R < \pi/(2\sqrt{\kappa})$. Let $c := \sum_{k=0}^{\infty} c_k$, as well as $R := \sqrt{r^2 + c}$ for

$$c_k = \left(\left(\frac{\lambda_l}{2} - 1 \right) L_{\exp}^2 L_F^2 + \left(1 - \frac{\lambda_u}{2} \right) 4L_G^2 \right) \tau_k^2.$$

We will, along the course of the proof, establish by induction that $x^k \in \bar{B}(x^*, R)$ for all $k \in \mathbb{N}$. By assumption, we have $x^0 \in \bar{B}(x^*, r) \subset A$, taking care of the inductive basis. So, for the inductive step, suppose $\{x^j\}_{j=0}^k \subset \bar{B}(x^*, r)$. By Lemma 6.4, we have $x^{k+1} \in A$, so we may apply Lemma 5.11. The Lipschitz continuity of G and the definition of the proximal map establish

$$\frac{1}{2\tau_k} d^2(x^{k+1}, w^k) \leq G(w^k) - G(x^{k+1}) \leq L_G d(x^{k+1}, w^k),$$

hence $d^2(w^k, x^{k+1}) \leq 4\tau_k^2 L_G^2$. Likewise, the Lipschitz continuity of F and the exponential map yield $d^2(x^k, w^k) \leq L_{\exp}^2 \tau_k^2 L_F^2$ (see the examples from the beginning of Section 6). Using these estimates in the claim (5.9) of Lemma 5.11 establishes

$$(6.13) \qquad d^2(x^*, x^{k+1}) \leq d^2(x^*, x^k) - 2\tau_k(H(x^{k+1}) - H(x^*)) + c_k$$

for all $k \in \mathbb{N}$. Through our inductive argument, we thus have $d^2(x^*, x^{j+1}) \leq d^2(x^*, x^j) + c_j$ for all $j = 0, \dots, k$. Summing, this establishes $x^{k+1} \in \bar{B}(x^*, R)$, completing the inductive step.

Thus $x^k \in \bar{B}(x^*, R)$ and (6.13) for all $k \in \mathbb{N}$. Set $a_k := d^2(x^*, x^k)$, $b_k := 2\tau_k(H(x^{k+1}) - H(x^*))$. The sequence $\{b_k\}_{k\in\mathbb{N}}$ is positive due to $H(x^*) \leq H(x^{k+1})$ and $\tau_k > 0$, whereas the positivity of $\{c_k\}_{k\in\mathbb{N}}$ follows from $\lambda_l/2 \geq 1$ and $\lambda_u/2 \leq 1$. By (6.13), we have $a_{k+1} \leq a_k - b_k + c_k$. Applying Lemma 6.6 now implies the convergence of the sequence $\{d^2(x^*, x^k)\}_{k\in\mathbb{N}}$, boundedness of $\{x^k\}_{k\in\mathbb{N}}$, as well as

$$\sum_{k=0}^{\infty} 2\tau_k(H(x^{k+1}) - H(x^*)) < \infty.$$

This, with the assumption $\sum_{k=0}^{\infty} \tau_k = \infty$, implies that there exists a subsequence $\{x^{k_i}\}_{i\in\mathbb{N}}$ of $\{x^k\}_{k\in\mathbb{N}}$ such that $\lim_{i\to\infty} H(x^{k_i}) = H(x^*)$. Since $\{x^{k_i}\}_{i\in\mathbb{N}}$ is bounded, it has by local compactness a subsequence converging to a point $\hat{x} \in \bar{B}(x^*, R)$. By the lower semicontinuity of H it must be a minimiser thus implying that the sequence $\{d^2(\hat{x}, x^k)\}_{k\in\mathbb{N}}$ converges to 0, since a subsequence of $\{x^k\}_{k\in\mathbb{N}}$ converges to $\hat{x}$. Hence we have that $\lim_{k\to\infty} x^k = \hat{x}$. □

With these assumptions, this method enjoys the following iteration-complexity bound.

Corollary 6.8 (Iteration-complexity, Lipschitz continuous G and F). *Under the assumptions of Theorem 6.7, for all $N \in \mathbb{N}$, we have*

$$H(x^N) - H(x^*) \leq \frac{d^2(x^*, x^0) + \left(\left(\frac{\lambda_l}{2} - 1\right) L_{\exp}^2 L_F^2 + \left(1 - \frac{\lambda_u}{2}\right) 4L_G^2\right) \sum_{k=0}^{N-1} \tau_k^2}{2 \sum_{k=0}^{N-1} \tau_k},$$

where x^* is a point in which H achieves a minimum.

Proof. Let $y = x^*$ be a minimiser of H and sum over (6.13) from 0 to $N - 1$ to get

$$\sum_{k=0}^{N-1} 2\tau_k (H(x^{k+1}) - H(x^*)) \leq d^2(x^*, x^0) + \left(\left(\frac{\lambda_l}{2} - 1\right) L_{\exp}^2 L_F^2 + \left(1 - \frac{\lambda_u}{2}\right) 4L_G^2\right) \sum_{k=0}^{N-1} \tau_k^2.$$

Since $H(x^k)$ is monotone nonincreasing by Proposition 6.3, $H(x^{N-1}) \leq H(x^k)$ for all $k = 0, 1 \ldots, N-1$ gives us the result

$$2(H(x^N) - H(x^*)) \sum_{k=0}^{N-1} \tau_k \leq \sum_{k=0}^{N-1} 2\tau_k (H(x^{k+1}) - H(x^*))$$

$$\leq d^2(x^*, x^0) + \left(\left(\frac{\lambda_l}{2} - 1\right) L_{\exp}^2 L_F^2 + \left(1 - \frac{\lambda_u}{2}\right) 4L_G^2\right) \sum_{k=0}^{N-1} \tau_k^2. \qquad \square$$

Remark 6.9. As noted before, the Lipschitz condition on G is not necessary when the curvature is bounded from above by 0 (since $1 - \lambda_u/2 = 0$).

With an additional restriction on the diameter of the neighborhood of the minimising point and some penalty on the step sizes (due to said restriction), we can also prove convergence to a minimiser without the rather restrictive Lipschitz assumption on G.

Theorem 6.10 (Convergence to a minimiser, Lipschitz continuous F). *Along with Assumptions 6.1 and 6.2, let F be Lipschitz continuous with factor L_F. Let $x^0 \in \bar{B}(x^*, r)$ for $r > 0$ and $R > 2r$ such that $R < \pi/(4\sqrt{\kappa})$ and $\lambda_u/2 = 2R\sqrt{\kappa}/\tan(2R\sqrt{\kappa}) > 1/2$ hold (see Remark 6.11). Let $\{\tau_k\}_{k \in \mathbb{N}}$ be a positive sequence of step sizes satisfying $\sum_{k=0}^\infty \tau_k = \infty$, $\sum_{k=0}^\infty \tau_k^2 < \infty$, and*

$$(6.14) \qquad \tau_k < \min\left\{\frac{\lambda_u - 1}{L}, \frac{R - 2r}{2L_F(1 + L_{\exp})}\right\}$$

for all $k \geq 0$. If $\inf_{x \in X} H(x)$ is attained at some point, then the sequence $\{x^k\}_{k \in \mathbb{N}}$ created by the iterative process (6.1) converges to some minimiser of H in $\bar{B}(x^*, R)$ as $k \to \infty$.

Proof. The overall outline of the proof is analogous to Theorem 6.7, however the constants c_k, estimating the final terms of the claim of Lemma 5.11, will be different. In particular, we will estimate $d^2(w^k, x^{k+1})$ differently, to avoid imposing the Lipschitz requirement on G.

Since the inequalities $\lambda_u/2 > 1/2$ and $L\tau_k < \lambda_u - 1$ are strict, there exists an $\varepsilon > 0$ such that $\lambda_u/2 > (\varepsilon + 1)/(\varepsilon + 2)$ and $L\tau_k < \lambda_u(\varepsilon + 2)/2 - 1 - \varepsilon$. By the triangle and Young's inequalities, we have

$$d^2(w^k, x^{k+1}) \leq d^2(w^k, x^k) + d^2(x^k, x^{k+1}) + 2d(x^k, x^{k+1})d(w^k, x^k)$$

$$\leq (1 + \varepsilon^{-1})d^2(w^k, x^k) + (1 + \varepsilon)d^2(x^k, x^{k+1}).$$

This and Lemma 5.11 with $y = x^*$ and $L\tau_k < \lambda_u(\varepsilon + 2)/2 - 1 - \varepsilon$ give, for $x^k, w^k, x^{k+1} \in \bar{B}(x^*, R)$,

$$d^2(x^*, x^{k+1}) \leq d^2(x^*, x^k) + 2\tau_k(H(x^*) - H(x^{k+1})) + (L\tau_k - \frac{\lambda_u}{2})d^2(x^k, x^{k+1}) + \left(\frac{\lambda_l}{2} - 1\right)d^2(x^k, w^k)$$

$$+ \left(1 - \frac{\lambda_u}{2}\right)d^2(w^k, x^{k+1})$$

$$\leq d^2(x^*, x^k) + 2\tau_k(H(x^*) - H(x^{k+1})) + (L\tau_k - \frac{\lambda_u}{2})d^2(x^k, x^{k+1}) + \left(\frac{\lambda_l}{2} - 1\right)d^2(x^k, w^k)$$

$$+ (1 - \frac{\lambda_u}{2})(1 + \varepsilon^{-1})d^2(w^k, x^k) + (1 - \frac{\lambda_u}{2})(1 + \varepsilon)d^2(x^k, x^{k+1})$$

$$\leq d^2(x^*, x^k) - 2\tau_k(H(x^{k+1}) - H(x^*)) + (\frac{\lambda_l}{2} + \varepsilon^{-1} - \frac{\lambda_u(1 + \varepsilon)}{2\varepsilon})d^2(x^k, w^k).$$

Using the Lipschitz continuity of the exponential map and F grants us the estimate $d^2(x^k, w^k) \leq L_{\exp}^2 \tau_k^2 L_F^2$, further giving

(6.15) $$d^2(x^*, x^{k+1}) \leq d^2(x^*, x^k) - 2\tau_k(H(x^{k+1}) - H(x^*)) + c_k,$$

i.e., (6.13), now with

$$c_k = \left(\left(\frac{\lambda_l}{2} + \varepsilon^{-1} - \frac{\lambda_u(1 + \varepsilon)}{2\varepsilon}\right)L_{\exp}^2 L_F^2\right)\tau_k^2.$$

Note that $c_k \geq 0$ for all $k \geq 0$, since $\lambda_l/2 \geq 1$, $\lambda_u/2 \leq 1$, and thus $\lambda_l/2 + \varepsilon^{-1} - \lambda_u(1 + \varepsilon)/(2\varepsilon) \geq 1 + \varepsilon^{-1} - (1 + \varepsilon)/\varepsilon = 0$.

The rest of the proof is exactly as that of Theorem 6.7. □

Remark 6.11. The assumption $\lambda_u/2 > 1/2$ is necessitated by the requirements: $L\tau_k + 1 + -\lambda_u < 0$ and $L\tau_k > 0$. This can be met, as the factor $\lambda_u/2 = 2R\sqrt{\kappa}/\tan(2R\sqrt{\kappa})$ can be made to be as close to 1 as we wish by making R smaller (and hence diam $\bar{B}(x^*, R)$ smaller). Preferably, the factor would be as small as possible to have the diameter be as large as possible.

For these step size restrictions, we have the following iteration-complexity bound for the forward-backward method.

Corollary 6.12 (Iteration-complexity, Lipschitz continuous F). *Let F be Lipschitz continuous and set $\alpha_k = \tau_k|\nabla_- F|(x^k)L_{\exp}$, for $k = 0, 1, \ldots$ where $\{x^k\}_{k \in \mathbb{N}}$ is the sequence generated by (6.1). Then for every $N \in \mathbb{N}_{\geq 0}$ and some choice of $\varepsilon > 0$, the following holds*

$$H(x^N) - H(x^*) \leq L_F \frac{d^2(x^0, x^*) + (\frac{\lambda_l}{2} + \varepsilon^{-1} - \frac{\lambda_u(1+\varepsilon)}{2\varepsilon})\sum_{k=0}^{N-1}\alpha_k^2}{2\sum_{k=0}^{N-1}\alpha_k},$$

where x^* is a point in which H achieves a minimum.

Proof. Same as Proposition 4.9 with added monotonicity of $H(x^k)$ as in Corollary 6.8. □

Remark 6.13. As noted before, in the presence of a nonzero curvature bound, the square summability of the step sizes seems to always be necessary in obtaining convergence results; square summability ensures that the step sizes are small enough while the nonsummability of the general sequence keeps them from being too small (guaranteeing convergence).

For a more general case of the key inequality, we may consider the instance when G and F can also be strongly convex.

Lemma 6.14 (Key lemma, strongly convex G and F). *Along with Assumptions 6.1 and 6.2, let G and F be λ-convex with factors $\lambda_G, \lambda_F \geq 0$, respectively. Fix the step size $\tau > 0$ and a closed, geodesically convex set $A \subset X$ (diam $A < \pi/(2\sqrt{\overline{\kappa}})$, if $\overline{\kappa} > 0$) containing a nonempty sublevel set of $G + F$. The key inequality (5.9) now reads*

$$(6.16) \qquad (1 + \lambda_G \tau)d^2(y, J(x)) \leq (1 - \lambda_F \tau)d^2(y, x) + 2\tau(H(y) - H(J(x))) + (L\tau - \frac{\lambda_u}{2})d^2(x, J(x))$$

$$+ \left(\frac{\lambda_l}{2} - 1\right) d^2(x, w) + \left(1 - \frac{\lambda_u}{2}\right) d^2(w, J(x))$$

for all $x, y \in A$ satisfying $w, J(x) \in A$, where $H(x) = G(x) + F(x)$, $w = \exp_{x^k}(\tau \nabla(-F)(x))$, $\lambda_u/2 = \sqrt{\overline{\kappa}} \operatorname{diam} A/\tan(\sqrt{\overline{\kappa}} \operatorname{diam} A)$, and $\lambda_l/2 = \sqrt{-\underline{\kappa}} \operatorname{diam} A/\tanh(\sqrt{-\underline{\kappa}} \operatorname{diam} A)$.

Furthermore, when x^* is a minimiser of H, we have

$$(6.17) \qquad (1 + \lambda_G \tau + \lambda_H \tau)d^2(x^*, J(x)) \leq (1 - \lambda_F \tau)d^2(x^*, x) + (L\tau - \frac{\lambda_u}{2})d^2(x, J(x))$$

$$+ \left(\frac{\lambda_l}{2} - 1\right) d^2(x, w) + \left(1 - \frac{\lambda_u}{2}\right) d^2(w, J(x)),$$

where λ_H is the convexity factor of H ($\lambda_H = \lambda_G + \lambda_F$).

Proof. For a λ_G-convex G, we have by (3.1)

$$D_{\operatorname{prox}_{\tau G}(x)} G(\gamma^+(0)) \leq G(y) - G(\operatorname{prox}_{\tau G}(x)) - (\lambda_G/2)d^2(y, \operatorname{prox}_{\tau G}(x)),$$

so Lemma 5.6 now improves to

$$d^2(y, \operatorname{prox}_{\tau G}(x)) \leq d^2(y, x) + 2\tau(G(y) - G(\operatorname{prox}_{\tau G}(x))) - \frac{\lambda_u}{2}d^2(x, \operatorname{prox}_{\tau G}(x)) - \lambda_G \tau d^2(y, \operatorname{prox}_{\tau G}(x)).$$

Analogously Lemma 5.7 improves for λ_F-convex F to

$$d^2(y, \operatorname{desc}_{\tau F}(x)) \leq d^2(y, x) + 2\tau(F(y) - F(x)) + \frac{\lambda_l}{2}d^2(x, \operatorname{desc}_{\tau F}(x)) - \lambda_F \tau d^2(y, x)$$

By following the proof of Lemma 5.11 with these estimates now gives us (6.16). For (6.17), let λ_H be the convexity factor of H and x^* its minimiser. For a geodesic γ between $J(x)$ and x^*, we have $H(\gamma(t)) \leq (1-t)H(J(x)) + tH(x^*) - (\lambda_H/2)t(1-t)d^2(J(x), x^*)$ by λ-convexity and $H(x^*) \leq H(\gamma(t))$ by minimality, for all $t \in [0, 1]$ and any $J(x) \in A$. Dividing by $1 - t$ and letting $t \to 1$ gives us

$$\frac{\lambda_H}{2}d^2(J(x), x^*) \leq H(J(x)) - H(x^*).$$

The claim now follows by choosing $y = x^*$ in (6.16). □

As strongly convex functions in complete geodesic spaces have unique minima [1, Lemma 14.4], one can again show using Lemma 6.14, that the forward-backward method converges to the unique minimiser of H (one might even allow $\lambda_F < 0$ as long as $\lambda_H > 0$). As an example, we again take G and F to be Lipschitz continuous with G being convex and F strongly convex.

Theorem 6.15 (Convergence to a minimiser, Lipschitz continuous G and F, and strongly convex F). *Along with the assumptions of Theorem 6.7, let F be strongly convex with factor $\lambda_F > 0$ and $\tau_k \lambda_F < 1$ for all $k \geq 0$. If $\inf_{x \in X} H(x)$ is attained at some point, then the sequence $\{x^k\}_{k \in \mathbb{N}}$ created by the iterative process (6.1) converges to the unique minimiser of H in $\bar{B}(x^*, R)$ as $k \to \infty$.*

Furthermore, we have

(6.18)
$$d^2(x^{k+1}, x^*) \leq \prod_{i=0}^{k} \frac{1 - \tau_i \lambda_F}{1 + \tau_i \lambda_F} d^2(x^0, x^*) + \sum_{i=0}^{k} \left(\left((\frac{\lambda_l}{2} - 1) \frac{L_{\exp}^2 L_F^2 \tau_i^2}{1 + \tau_i \lambda_F} + \left(1 - \frac{\lambda_u}{2}\right) \frac{4 L_G^2 \tau_i^2}{1 + \tau_i \lambda_F} \right) \prod_{j=i+1}^{k} \frac{1 - \tau_j \lambda_F}{1 + \tau_j \lambda_F} \right),$$

with the convention that $\prod_{j=m}^{n}$ is the empty product when $m > n$, i.e., equal to 1.

The proof follows the same locality and convergence arguments as in the proof of Theorem 6.7, while uniqueness of the minimiser is due to the strong convexity of H. The expression (6.18) follows by iterating (6.17) with the Lipschitz estimates from Theorem 6.7.

7 EXAMPLES

We test our algorithm on a Lasso-like problem on two surfaces: the surface of a cube and that of a right circular cylinder capped from both ends with a flat disc of the same radius. For a regularisation parameter $\lambda > 0$, a choice of "origin" $z \in X$, and data points $\{y_i\}_{i=1}^{n} \subset X$, the objective function on both surfaces will be

(7.1)
$$H(x) = F(x) + G(x) = \sum_{i=1}^{n} \frac{1}{2} d^2(x, y_i) + \lambda d(x, z),$$

To apply forward-backward splitting method to minimise this function, we need to calculate the gradient of F and the proximal operator of G. The latter will be a type of soft thresholding operator. Practically, to compute these maps, we need to form the exponential and logarithmic maps of the surfaces. The software implementation of our experiments (in Rust) is available on Zenodo [28].

7.1 AUXILIARY LEMMA

For $y, z \in B(x, D_{\bar{\kappa}})$, where $D_{\bar{\kappa}}$ is the diameter of the model space (see Section 2), we have by [1, Chapter 13, sections D and E], the gradients

(7.2)
$$\nabla F(x) = -\sum_{i=1}^{n} \log_x y_i \quad \text{and} \quad \nabla G(x) = -\lambda \frac{\log_x z}{d(x,z)}, \quad \text{if } x \neq z.$$

The gradient step can now be written as

$$\text{desc}_{\tau F}(x) = \exp_x(\tau \nabla(-F)(x)) = \exp_x(\tau \sum_{i=1}^{n} \log_x y_i),$$

whereas the proximal step needs a little bit more work and has the following closed-form solution.

Lemma 7.1. *Let $G(x) = \lambda d(x, z)$, with $x, z \in A \subset X$, $\text{diam } A < \pi/(2\sqrt{\kappa})$, and $\lambda > 0$. Then*

(7.3)
$$\text{prox}_{\tau G}(x) = y^* = \begin{cases} z, & \text{if } d(x, z) \leq \lambda \tau; \\ \exp_x(\frac{\lambda \tau}{d(x,z)} \log_x z), & \text{if } d(x, z) > \lambda \tau. \end{cases}$$

For consistency, we could also write $z = \exp_x(\log_x z)$.

Proof. Consider first $d(x, z) \leq \lambda \tau$. Let $y \in X$. We have by the triangle inequality

$$(d(x, z) - d(y, z))^2 \leq d^2(x, y) \leq d^2(x, y) + d^2(y, z)$$
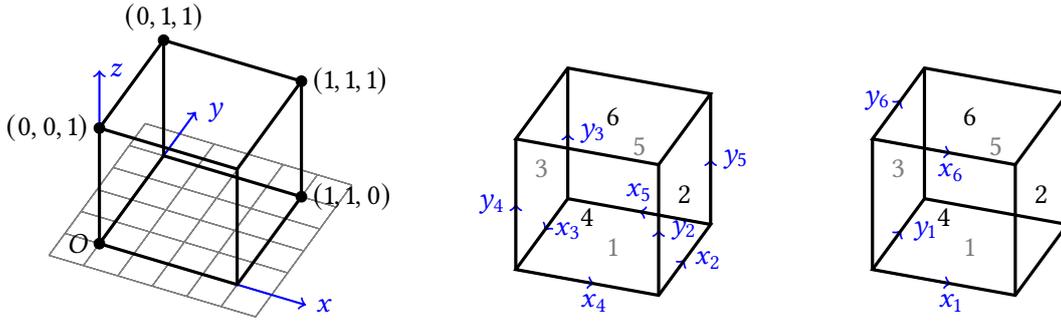
Figure 1: A unit cube with labeled faces and their local coordinate systems.

so that

$$d^2(x, z) \leq 2d(x, z)d(y, z) + d^2(x, y).$$

Using the assumption $d(x, z) \leq \lambda\tau$, and adding $2\lambda\tau d(z, z) = 0$ on both sides leads to

$$2\lambda\tau d(z, z) + d^2(x, z) \leq 2\lambda\tau d(y, z) + d^2(x, y) \quad \text{for all} \quad y \in X.$$

This is to say, $\text{prox}_{\tau G}(x) = z$.

Now, suppose $d(x, z) > \lambda\tau$. By (7.2), for any $y \neq z$, we have

$$\nabla(G(y) + (2\tau)^{-1}d^2(x, y))(y) = -\lambda(d(y, z))^{-1}\log_y z - \tau^{-1}\log_y x,$$

which for $y = y^* \neq z$ to be a minimiser, must equal to 0, that is,

(7.4) $$\log_{y^*} x = -\lambda\tau(d(y^*, z))^{-1}\log_{y^*} z.$$

By the assumption $x, z \in A$, we can write the geodesic $\gamma(t) = x\#_t z$, for $t \in [0, 1]$, as $\gamma(t) = \exp_x(t\log_x z)$. Since the lengths of tangent vectors are always positive and the geodesic γ is unique, the equivalence of the left- and right-hand sides of (7.4) implies that y^* lies on the geodesic γ, i.e. $y^* = \exp_x(t^*\log_x z)$, for some $0 < t^* < 1$. More explicitly, we have by (7.4)

$$d(y^*, x) = d(\exp_x(t^*\log_x z), x) = t^*|\log_x z| \implies t^*|\log_x z| = \lambda\tau \implies t^* = \lambda\tau/|\log_x z|,$$

which, by assumption $d(x, z) > \lambda\tau > 0$, satisfies $0 < t^* < 1$. Although the formula $y^* = \exp_x(t^*\log_x z)$ would still hold with $t^* = 0$ in the case $y^* = z$, we have just shown that this case never happens. □

7.2 CUBE

Let our surface be a unit cube with labeled faces and the origin as one of the vertices. In addition, let each face F have their own fixed oriented local two-dimensional coordinate system denoted by $(x_F, y_F)_F \in [0, 1]^2$, where $F \in \{1, 2, 3, 4, 5, 6\}$, as in Figure 1. We shall use the triple (x, y, z) for the standard coordinates in three-dimensional Euclidean space. As the surface of a convex body, the curvature of the cube is bounded from below by $\underline{\kappa} = 0$ [11, Theorem 10.2.6]. By the Reshetnyak gluing theorem [11, Theorem 9.1.21], *away from the corners*, but including the remainder of the edges, the curvature is also bounded from above by $\overline{\kappa} = 0$. Since geodesics are not unique at the corners, the curvature cannot be bounded from above near the corners [11, Theorem 9.1.17]. Practically, we will never reach the corners[1], so our convergence theory is applicable. Alternatively, we could smoothed

---

[1] Since geodesics do not pass through corners, the proximal map ensures that x^k will never be in a corner. The gradient descent map may attempt to pass through a corner, but due to the finite number of corner points, this is easily avoided by a numerically insignificant perturbation.

$\mathbb{R}^3$	$(x, y, 0)$	$(1, x, y)$	$(0, 1-x, y)$	$(x, 0, y)$	$(1-x, 1, y)$	$(x, y, 1)$
Face #	1	2	3	4	5	6
1	$(x, y)_1$	$(y+1, x)_1$	$(-y, 1-x)_1$	$(x, -y)_1$	$(1-x, y+1)_1$	
2	$(y, x-1)_2$	$(x, y)_2$		$(x-1, y)_2$	$(x+1, y)_2$	$(y, 2-x)_2$
3	$(1-y, -x)_3$		$(x, y)_3$	$(x+1, y)_3$	$(x-1, y)_3$	$(1-y, x+1)_3$
4	$(x, -y)_4$	$(x+1, y)_4$	$(x-1, y)_4$	$(x, y)_4$		$(x, y+1)_4$
5	$(1-x, y-1)_5$	$(x-1, y)_5$	$(x+1, y)_5$		$(x, y)_5$	$(1-x, 2-y)_5$
6		$(2-y, x)_6$	$(y-1, 1-x)_6$	$(x, y-1)_6$	$(1-x, 2-y)_6$	$(x, y)_6$

Table 1: Coordinate transformations between the faces, with $x, y \in [0, 1]$, as well as their standard three-dimensional coordinates. Identity transformations are denoted in grey. Note that the empty places correspond to non-adjacent faces.

the corners (but, importantly, not the edges), to have a the entire manifold bounded from both above and below.

Geodesics on the 3D cube form straight lines that do not go through vertices [25, Lemma 4.1], when the cube is unfolded into a two-dimensional surface. To determine the geodesics, we need transformations between local coordinate systems, which can then be composed together to make longer chains between faces. A coordinate transformation from one face to another is made by rotating (with respect to the common edge) the target face onto the same plane as the starting face (keeping this face fixed) such that there is no overlap. Table 1 contains all coordinate transformations between adjacent faces as well as their standard three-dimensional coordinates in each face. Distances between points can be calculated with the help of Table 1 by transforming the coordinate of the endpoint to that of the starting point's system and using the Pythagorean theorem for all of the possible paths between these points. The possible transformations also facilitates the formation of the exponential and logarithm maps. An illustration of geodesics between two points on faces 1 and 2 in different unfoldings can be seen in Figure 2. We illustrate the numerical behaviour of our forward-backward algorithm on the cube in Figure 4a.

7.3 CYLINDER

For our second example, we consider the capped circular cylinder

$$C_{r,h} := \{(t\cos\varphi, t\sin\varphi, z) \mid \varphi \in [0, 2\pi), \ 0 \le t \le r, \ -h \le z \le h, \ (t-r)(h-|z|) = 0\} \subset \mathbb{R}^3$$

of height $2h$ and radius r. Again, as the surface of a convex body, the curvature is bounded from below by $\underline{\kappa} = 0$ [11, Theorem 10.2.6]. *Away from the edges*, i.e., independently on the cap and the side, the curvature is also bounded from above $\overline{\kappa} = 0$. This can be easily seen from the definition. Since geodesics are not unique near the edge between the cylindrical surface and the cap, the curvature cannot be bounded from above at the edge [11, Theorem 9.1.17]. Thus, our convergence theory only applies if, after a finite number of steps, the geodesics joining the iterates no longer cross the edges of the caps, all the iterates staying on either the side or one of the caps. The computationally more difficult "total surrogate" approach of Remark 6.5 would be applicable without such a restriction. Moreover, the same considerations as for the cube in Section 7.2, apply to a polyhedral approximation of the cylinder.

We will equivalently work with the cylindrical coordinates (z, φ, t). Geodesics between two points
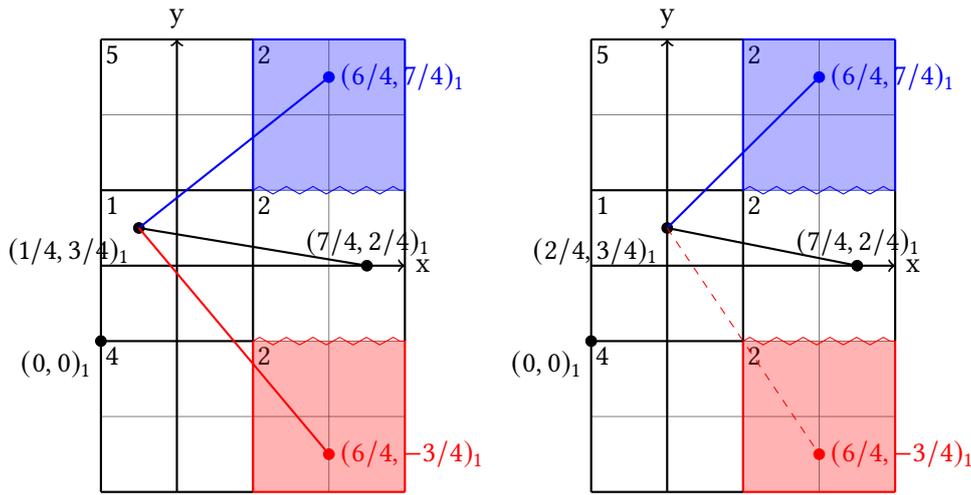
Figure 2: Two copies of an unfolding of the cube with the labels of the faces marked on their upper left corners. Included are three colour-coded paths from face 1 to face 2, with different starting points but same endpoints, in the coordinate system of face 1. The paths on the left are all geodesics, whereas the dashed red path on the right is not, as it goes through the vertex $(1, 0)_1$. Note that the two zigzag edges between faces 2 are not real but merely artefacts of the drawing.

$p_1 = (z_1, \varphi_1, r)$ and $p_2 = (z_2, \varphi_2, r)$ on the side are either helices of the form

$$\gamma(\varphi) = \left(r \cos \varphi, r \sin \varphi, \frac{z_1 - z_2}{\varphi_1 - \varphi_2} \varphi + \frac{\varphi_1 z_2 - \varphi_2 z_1}{\varphi_1 - \varphi_2} \right)$$

or, in the degenerate cases, straight lines or arcs of circles (this follows from Clairaut's Theorem [23, Proposition 9.3.2]), whereas geodesics on the disk-shaped caps are simply straight lines. Geodesics on the whole surface are thus concatenations of these types such that the geodesic forms a straight line in an unfolding of the cylinder [17, Chapter III, Theorem 3]; the side forms a rectangle and the disk-shaped caps are tangent to this on opposite sides (see Figure 3). The two difficult types of geodesics on the whole surface are: a path going over the whole side or a whole cap. A simple example shows that the second case is indeed possible and that instead of a helix on the cylindrical surface, we sometimes take a shortcut through a cap.

1. A geodesic starting from p_1 at the top cap and ending at p_2 at the bottom cap. Points x_1 and x_2, where the geodesic crosses the top and bottom caps, respectively, need to be chosen such that the curve going through the points p_1, x_1, x_2 and p_2 is straight in the unfolded picture.

2. A geodesic between points p_1 and p_2, both on side of the cylinder. If the geodesic crosses the boundary of a cap at x_1 and x_2, then these points need to be chosen such that the curve going through the points p_1, x_1, x_2, and p_2 is straight in the unfolded picture.

Both of these problems are solved from the optimisation problem

$$d(p_1, p_2) = \min_{(x_1, x_2)} \left\{ d(p_1, x_1) + d(x_1, x_2) + d(x_2, p_2) \right\}.$$

With a bit of trigonometry[2], this can be solved with several applications of the law of cosines and minimising with respect to the angles where the geodesic crosses the edges. The resulting solutions

---

[2]A detailed derivation is included with our implementation [28]

(a) Flattened 2D view   (b) 3D view
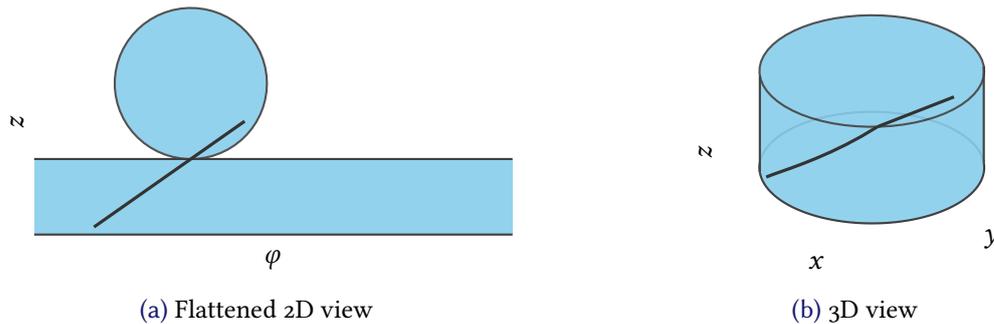
Figure 3: Cylinder tangents and geodesics. In the flattened 2D view of (a), geodesics are straight lines; for any two points on the geodesic, the tangent vector given by the logarithmic map is parallel to this line. In the 3D view of (b), the same geodesic is a helical path on the side, and again a straight line on the top.

are used to form the exponential and logarithm maps on the closed cylinder. However, these maps can only be solved numerically, as the solutions to the geodesic optimisation problems form transcendental equations. We illustrate the numerical behaviour of our forward-backward algorithm on the cylinder in Figure 4b. These simple numerical experiments validate our theory.

REFERENCES

[1] S. Alexander, V. Kapovitch, and A. Petrunin, Alexandrov geometry: foundations, 2023.

[2] L. Ambrosio, N. Gigli, and G. Savare, *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*, Lectures in Mathematics. ETH Zürich, Birkhäuser Basel, 2006.

[3] M. Bačák, *Convex Analysis and Optimization in Hadamard Spaces*, Nonlinear Analysis and Applications, De Gruyter, 2014.

[4] S. Banert, Backward–backward splitting in Hadamard spaces, *Journal of Mathematical Analysis and Applications* 414 (2014), 656–665, doi:https://doi.org/10.1016/j.jmaa.2014.01.054, https://www.sciencedirect.com/science/article/pii/S0022247X14000626.

[5] F. Beier and R. Beinert, Tangential Fixpoint Iterations for Gromov-Wasserstein Barycenters, 2024, arXiv:2403.08612.

[6] G. C. Bento, O. P. Ferreira, and J. G. Melo, Iteration-Complexity of Gradient, Subgradient and Proximal Point Methods on Riemannian Manifolds, *Journal of Optimization Theory and Applications* 173 (2017), 548––562, doi:10.1007/s10957-017-1093-4.

[7] V. N. Berestovskii, Introduction of a Riemann structure into certain metric spaces, *Siberian Mathematical Journal* 16 (1975), 499–507, doi:10.1007/bf00967122.

[8] R. Bergmann, R. Herzog, M. Silva Louzeiro, D. Tenbrinck, and J. Vidal-Núñez, Fenchel Duality Theory and a Primal-Dual Algorithm on Riemannian Manifolds, *Foundations of Computational Mathematics* 21 (2021), 1465–1504, doi:10.1007/s10208-020-09486-5.

[9] R. Bergmann, J. Persch, and G. Steidl, A Parallel Douglas–Rachford Algorithm for Minimizing ROF-like Functionals on Images with Values in Symmetric Hadamard Manifolds, *SIAM Journal on Imaging Sciences* 9 (2016), 901–937, doi:10.1137/15m1052858.

(a) Cube   (b) Cylinder
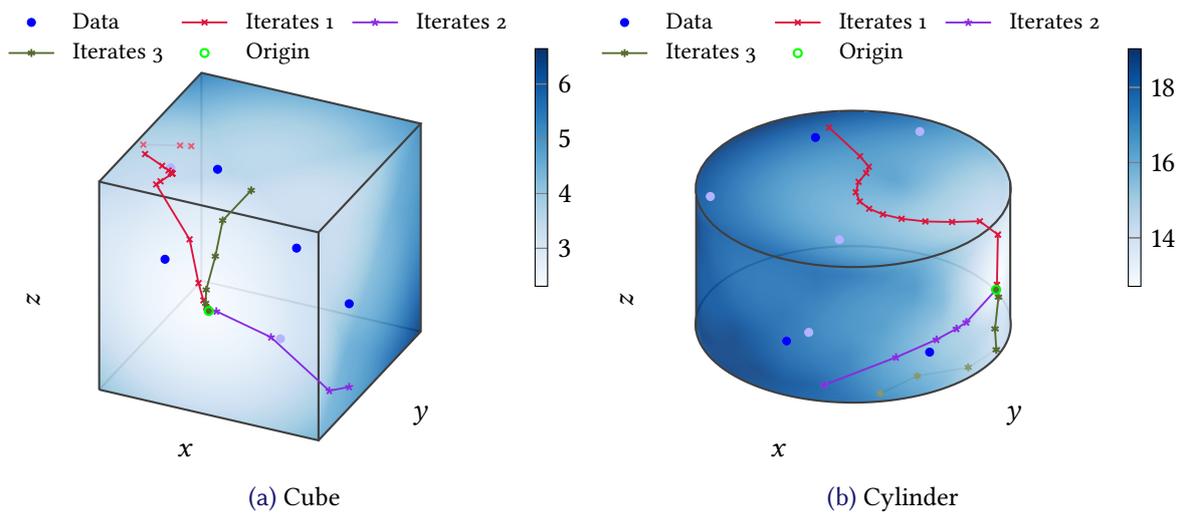
Figure 4: Numerical illustrations. The surface shading indicates function value. Blue dots indicate data points for squared distance terms on exposed faces, while dimmer dots indicate data points on hidden faces. The green circle indicates the "origin" data point for the unsquared distance term. Algorithm iterates and connecting lines are shown for three different starting points.

[10] M. Bridson and A. Häfliger, *Metric Spaces of Non-Positive Curvature*, Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg, 2011.

[11] D. Burago, I. Burago, and S. Ivanov, *A Course in Metric Geometry*, Crm Proceedings & Lecture Notes, American Mathematical Society, 2001.

[12] Y. Burago, M. Gromov, and G. Perel'man, A.D. Alexandrov spaces with curvature bounded below, *Russian Mathematical Surveys* 47 (1992), 1, doi:10.1070/rm1992v047n02abeh000877.

[13] C. Clason, S. Mazurenko, and T. Valkonen, Acceleration and global convergence of a first-order primal-dual me thod for nonconvex problems, *SIAM Journal on Optimization* 29 (2019), 933–963, doi:10.1137/18m1170194, arXiv:1802.03347.

[14] N. Dizon, J. Jauhiainen, and T. Valkonen, Online optimisation for dynamic electrical impedance tomography, 2024, arXiv:2412.12944. submitted.

[15] N. Dizon and T. Valkonen, Differential estimates for fast first-order multilevel nonconvex optimisation, 2024, arXiv:2412.01481. submitted.

[16] O. P. Ferreira and P. R. Oliveira, Proximal Point Algorithm On Riemannian Manifolds, *Optimization* 51 (2002), 257–270, doi:10.1080/02331930290019413.

[17] S. Kutateladze, *A.D. Alexandrov: Selected Works Part II: Intrinsic Geometry of Convex Surfaces*, Classics of Soviet Mathematics, CRC Press, 2005.

[18] K. Kuwae, Y. Machigashira, and T. Shioya, Sobolev spaces, Laplacian, and heat kernel on Alexandrov spaces, *Mathematische Zeitschrift* 238 (2001), 269–316, doi:10.1007/s002090100252, https://doi.org/10.1007/s002090100252.

[19] F. Lauster and D. R. Luke, Convergence of proximal splitting algorithms in CAT(κ) spaces and beyond, *Fixed Point Theory and Algorithms for Sciences and Engineering* 2021 (2021), 13, doi:10.1186/s13663-021-00698-0, https://doi.org/10.1186/s13663-021-00698-0.

[20] S. I. Ohta, Convexities of metric spaces, *Geometriae Dedicata* 125 (2007), 225–250, doi:10.1007/s10711-007-9159-3.

[21] S. I. Ohta, Gradient Flows on Wasserstein Spaces over Compact Alexandrov Spaces, *American Journal of Mathematics* 131 (2009), 475–516.

[22] S. I. Ohta and M. Pálfia, Discrete-time gradient flows and law of large numbers in Alexandrov spaces, *Calculus of Variations* (2014), doi:10.1007/s00526-015-0837-y.

[23] A. Pressley, *Elementary Differential Geometry*, Springer Undergraduate Mathematics Series, Springer London, 2010.

[24] H. Robbins and D. Siegmund, A convergence theorem for non negative almost supermartingales and some applications, *Optimizing Methods in Statistics* (1971), 233–257, doi:10.1016/b978-0-12-604550-5.50015-8.

[25] M. Sharir and A. Schorr, On Shortest Paths in Polyhedral Spaces, *SIAM Journal on Computing* 15 (1986), 193–215, doi:10.1137/0215014, https://doi.org/10.1137/0215014.

[26] K. Sturm, *The Space of Spaces: Curvature Bounds and Gradient Flows on the Space of Metric Measure Spaces*, Memoirs of the American Mathematical Society, American Mathematical Society, 2023.

[27] T. Valkonen, First-order primal-dual methods for nonsmooth nonconvex optimisation, in *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging*, K. Chen, C. B. Schönlieb, X. C. Tai, and L. Younes (eds.), Springer, Cham, 2021, doi:10.1007/978-3-030-03009-4_93-1.

[28] T. Valkonen and H. von Koch, Non-Riemannian optimisation, 2025, doi:10.5281/zenodo.15102168. software on Zenodo.

[29] A. Weinmann, L. Demaret, and M. Storath, Total Variation Regularization for Manifold-Valued Data, *SIAM Journal on Imaging Sciences* 7 (2014), 222–2257, doi:10.1137/130951075.

[30] H. Zhang and S. Sra, First-order Methods for Geodesically Convex Optimization, in *29th Annual Conference on Learning Theory*, V. Feldman, A. Rakhlin, and O. Shamir (eds.), volume 49 of Proceedings of Machine Learning Research, PMLR, Columbia University, New York, New York, USA, 2016, 1617–1638.