A subgradient splitting algorithm for optimization on nonpositively curved metric spaces

Ariel Goodwin * Adrian S. Lewis † Genaro López-Acedo ‡
Adriana Nicolae §

December 18, 2024

Abstract

Many of the primal ingredients of convex optimization extend naturally from Euclidean to Hadamard spaces — nonpositively curved metric spaces like Euclidean, Hilbert, and hyperbolic spaces, metric trees, and more general CAT(0) cubical complexes. Linear structure, however, and the duality theory it supports are absent. Nonetheless, we introduce a new type of subgradient for convex functions on Hadamard spaces, based on Busemann functions. This notion supports a splitting subgradient method with guaranteed complexity bounds. In particular, the algorithm solves p-mean problems in general Hadamard spaces: we illustrate by computing medians in BHV tree space.

Key words: convex optimization, subgradient, Hadamard space, splitting, complexity, Busemann function, mean, median, tree space

AMS Subject Classification: 90C48, 65Y20, 49M29

1 Introduction

We consider optimization problems posed over a subset C of a complete metric space (X, d). We assume that X is a *Hadamard space*, meaning that it has nonpositive curvature — the "CAT(0)" property from metric geometry [11]. This optimization framework covers the familiar example of Hilbert space, and all complete simply-connected Riemannian manifolds of nonpositive sectional curvature. Examples of such manifolds include Euclidean and hyperbolic spaces, and spaces of positive-definite symmetric matrices with the affine-invariant metric [10]. However, the Hadamard space framework also subsumes interesting examples that are not

*Center for Applied Mathematics, Cornell University, Ithaca, NY. arielgoodwin.github.io

†ORIE, Cornell University, Ithaca, NY. people.orie.cornell.edu/aslewis

‡Department of Mathematical Analysis – IMUS, University of Seville, 41012 Seville, Spain glopez@us.es

§Department of Mathematics, Babeş-Bolyai University, 400084, Cluj-Napoca, Romania anicolae@math.ubbcluj.ro

manifolds, such as the Billera-Holmes-Vogtmann (BHV) tree space [8] and, more generally, all CAT(0) cubical complexes. A simple example to keep in mind is the *tripod*, which consists of three copies of the halfline $\mathbb{R}_+$ with its usual metric, glued together by identifying the three copies of 0.

Any two points x and y in the Hadamard space X are connected by a unique geodesic $[x, y]$. For algorithmic purposes, we assume that we can readily compute the distance $d(x, y)$ and the geodesic $[x, y]$. We also assume the *extension property*: there exists a geodesic ray originating at x that contains y, and furthermore we can find such a ray computationally. These assumptions hold for each of the examples we have mentioned ([10, 16]).

For tractability, as in the classical Euclidean case, we focus on convex minimization problems

$$\inf_C f \tag{1}$$

where both the feasible region C and the objective $f\colon C \to \mathbb{R}$ are geodesically convex. In general Hadamard spaces, as shown in [3], such optimization problems are solvable in principle by iterating the proximal update

$$x \;\leftarrow\; \operatorname{argmin}\{f(y) + \alpha d(x, y)^2 : y \in C\} \tag{2}$$

for any constant $\alpha > 0$. However, this update is only implementable in a few special cases.

For our current purposes, we refine the expression of the optimization problem (1) to allow for additive structure:

$$f(x) \;=\; \sum_{i=1}^{m} f_i(x) \qquad (x \in C), \tag{3}$$

where the component functions $f_i\colon C \to \mathbb{R}$ are geodesically convex. Rather than iterating the proximal update for f, we can then apply a splitting approach, cyclically applying proximal updates to each component f_i in turn. Under mild assumptions on the problem data, convergence is proved in [3] but with no explicit complexity analysis. A closer inspection of the proofs reveals that appropriately chosen stepsizes lead to complexity guarantees for certain problems of interest: an ε-approximate minimizer of the objective can be found within $O(\varepsilon^{-2})$ iterations. Proximal splitting works, for example, for the *p-mean problem* (for $p \geq 1$) associated with given points $a_i \in C$ and nonnegative weights w_i. In that case each component

$$f_i(x) \;=\; w_i d(x, a_i)^p \qquad (x \in C)$$

has an implementable proximal update, albeit requiring a one-dimensional minimization problem unless the exponent p is 1 or 2. The case when each $w_i = 1$ and $p = 1$ is the classical *Weber problem* [28] in location theory.

Unfortunately, proximal updates are rarely implementable in general. Even in the classical Euclidean case we must often rely instead on subgradient-based methods. We therefore consider the following question:

Do classical subgradient methods extend to general Hadamard spaces?

This question presents an immediate conundrum. Whereas the proximal iteration (2) is primal in nature, simply defined on the underlying Hadamard space X, subgradients in the Euclidean setting are dual objects — linear functionals on X. In the absence of any linear structure, the analogue in Hadamard space is unclear.

To handle this fundamental hurdle when the space X is a manifold, the subgradient methods introduced in [15] and the complexity analysis of [29] rely on local linearization, much like the theory of smooth optimization in [10]. These works view gradients and subgradients at a point $x \in X$ as elements of the tangent space $T_x X$, forcing use of the exponential map $\mathrm{Exp}_x \colon T_x X \to X$ or some approximate version. Moreover, the complexity bound in [29] depends unavoidably on a lower curvature bound for X [12].

This approach is technical, and furthermore fails in general Hadamard spaces X because the machinery of local linearization, duality, and lower curvature bounds is unavailable. We argue here for an entirely different approach, one that is simpler, global, primal, and requires no lower curvature bound. We identify subgradients with constant-speed geodesic rays in X, and we escape the curvature-related worst-case bounds of [12] by restricting the class of objectives f, arriving at an algorithm with complexity analogous to the Euclidean case.

A first step in this direction was the recent development of a *horospherical subgradient algorithm* in [22]. This method applies in a general Hadamard space X, but requires a quasiconvexity property: at any point $x \in C$, the level set

$$L_x \;=\; \{y \in C : f(y) \le f(x)\}$$

must be horospherically convex. In non-Euclidean settings, horospherical quasiconvexity is a significant assumption, but one that often holds in practice. The algorithm relies on an oracle that returns a geodesic ray originating from x and in some sense normal to L_x, mimicking the key property of subgradient directions in the Euclidean case. The algorithm then takes a step from x along a "supporting ray": a geodesic ray opposite to the normal ray. The resulting complexity parallels the Euclidean case.

Horospherical ideas had appeared earlier in optimization, in works such as [6, 14, 17]. Horospheres in Hadamard spaces are limits of spheres: in Euclidean space, they are hyperplanes. Horospheres have the form $\{x \in X : b(x) = 0\}$ for *Busemann functions* b, the natural generalizations of affine functions on Euclidean spaces. Noticing these analogies, a notion of Fenchel conjugation based on Busemann functions was introduced for manifolds in [6], and more generally in [17].

Although the supporting ray oracle in [22] is implementable for some interesting functions, the horospherical subgradient algorithm is difficult to apply in much generality, because unlike Euclidean subgradients, supporting rays have no obvious calculus. Specifically, for structured objectives $f = \sum_i f_i$, we cannot easily combine supporting rays for each component f_i into a supporting ray for f. We therefore turn our attention instead to splitting subgradient methods that employ steps computed only from individual components f_i.

Splitting subgradient methods are appealing for their simplicity. Our approach in Hadamard space is inspired by the complexity analysis for a Euclidean splitting (or "incremental") algorithm in [23]. Quasiconvex versions of that algorithm appear in [18, 27], and were considered on manifolds with curvature bounded below in [1], but these algorithms have the drawback that all the components f_i must share a common minimizer. This restriction rules out many interesting examples, including the p-mean problem, but it seems inherent to methods relying only on supporting ray oracles, which cannot distinguish between different component functions having geometrically similar level sets. This drawback suggests the need for a stronger oracle: supporting rays give directional information but that alone does not suffice.

Our main contribution — the new splitting algorithm we introduce here — relies instead on a *Busemann subgradient* oracle. When called for one of the component functions f_i at a point $x \in C$, the oracle returns a Busemann function b and a "speed" $s \geq 0$ such that

$$f_i(y) - f_i(x) \;\geq\; s\big(b(y) - b(x)\big) \qquad \text{for all } y \in C.$$

Analogous subgradient inequalities have appeared previously in [20, 21]. Indeed, the notion of a Busemann subgradient that we introduce corresponds to subgradients in the sense of the [21]. Starting from x, the algorithm uses the information returned by the oracle, following the geodesic ray associated with b at speed s for a judiciously chosen time interval, before repeating the process. The method generalizes the incremental subgradient algorithm of [23], and enjoys the same complexity.

The structure of this paper is as follows. We begin by reviewing some metric geometry, and consequences of nonpositive curvature. Central to our development is the concept of *direction*, so we review some large-scale geometric properties of Hadamard spaces. This exploration leads to a new property, called *Busemann subdifferentiability*, strong enough to support a useful notion of subgradients in nonlinear space. We compare this new notion with earlier ideas about subgradients in Hadamard space, and relate it to horospherical convexity.

Remarkably, even in the Euclidean case, Busemann subdifferentiability gives a new perspective on subgradients of convex functions, allowing a fundamentally geometric or primal understanding, without explicit reference to the inner product. We show that many natural functions defined in terms of metric data are Busemann subdifferentiable, and we explore the calculus of Busemann subgradients. We demonstrate that Busemann subdifferentiability is not preserved under addition, motivating the use of splitting in Busemann subgradient-based algorithms.

Rather than formal convex analysis, our aim here is the development of implementable algorithms with complexity guarantees for structured convex optimization. To this end, we adapt an incremental subgradient method from the Euclidean setting [23] to minimize sums of Busemann subdifferentiable functions on a Hadamard space. The complexity result matches that of the Euclidean algorithm and the cyclic proximal algorithm appearing in [3], requiring $O(\varepsilon^{-2})$ iterations to guarantee an objective function value within ε of the minimum value. To illustrate the approach computationally, we solve the Weber (1-mean) problem for some small examples, from [4], in BHV tree space [8].

2 Geodesic geometry and convexity

A metric space (X, d) is a *geodesic metric space* if every two points $x, y \in X$ can be joined by a *geodesic*, which is to say a map γ from a closed interval $[a, b]$ into X with $d(\gamma(t), \gamma(t')) = |t - t'|$ for all $t, t' \in [a, b]$. A *ray* is a geodesic with domain $\mathbb{R}_+$, and we say r *issues from* x if $r(0) = x$. A geodesic metric space is said to be *CAT(0)* if the map $t \mapsto \frac{1}{2}d(\gamma(t), a)^2$ is 1-strongly convex for every $a \in X$ and geodesic γ. A *Hadamard space* is a complete CAT(0) space. In a Hadamard space, there is exactly one geodesic joining x to y for every $x, y \in X$. We say that a Hadamard space has the *geodesic extension property* if for every $x \neq y \in X$ there exists a ray $r\colon \mathbb{R}_+ \to X$ with $r(0) = x$ and $r(t) = y$ for some $t > 0$. A subset C of a Hadamard space X is *geodesically convex* if the geodesic between any two points in C is contained in C. A metric space (X, d) is *proper* if the closed ball $B_r(x)$ is compact for every $x \in X, r > 0$.

Henceforth, (X, d) will be a Hadamard space with the geodesic extension property, and to avoid degenerate case-splitting in forthcoming results we will assume X has at least two points. The significance of this technical assumption is that there always exist rays. Given a ray r on X we may associate the corresponding *Busemann function* $b_r\colon X \to \mathbb{R}$ defined by

$$b_r(z) = \lim_{t\to\infty}(d(z, r(t)) - t) \quad (z \in X).$$

Busemann functions are 1-Lipschitz, convex, and satisfy $b_r(r(0)) = 0$. Given a ray r, sets of the form $b_r^{-1}((-\infty, 0])$ are called *horoballs*.

We will restrict our attention to a class of functions that interacts nicely with the geometry of the given Hadamard space in a way that goes beyond geodesic convexity. Fundamental to our development is an appropriate notion of *direction* in a Hadamard space; in the Euclidean setting, the fundamental role played by directions and the associated compactification of $\mathbb{R}^n$ has been emphasized as a pillar of modern variational analysis [26, Chapter 3].

Let us review some standard tools to study the geometry at infinity of a Hadamard space, a more detailed discussion of which can be found in [11, Chapter II.8]. Two rays r, r' are said to be *asymptotic* if there exists a positive constant $K \geq 0$ such that $d(r(t), r'(t)) \leq K$ for all $t \geq 0$. This defines an equivalence relation on rays in X: the set of equivalence classes is denoted X^∞ and called the *boundary of X at infinity*, or simply the *boundary*. Note our assumption that X has at least two points and the geodesic extension property guarantees the boundary is nonempty. The equivalence class of a particular ray r is denoted $r(\infty)$, and we say a ray r has *direction* $\xi \in X^\infty$ if r belongs to the equivalence class of ξ. Given $x \in X$ and $\xi \in X^\infty$, there exists a unique ray r issuing from x such that $r(\infty) = \xi$ [11, II.8.2]. If we fix an arbitrary reference point $\bar{x} \in X$, we can thus identify any $\xi \in X^\infty$ with the unique ray $r_{\bar{x},\xi}$ issuing from $\bar{x}$ with direction ξ, and hence to the unique corresponding Busemann function which we may denote either $b_{r_{\bar{x},\xi}}$ or $b_{\bar{x},\xi}$. To avoid cumbersome notation, we henceforth fix such a reference point $\bar{x} \in X$ and denote the corresponding Busemann function $b_{r_{\bar{x},\xi}}$ by simply b_ξ.

The space X^∞ is naturally endowed with the so-called *cone topology* [11, Definition II.8.6], and notably this space is first-countable [11, Proof of Theorem II.8.13]. In particular, the cone topology is completely specified by the convergent sequences, so let us mention that $\xi_n \to \xi \in X^\infty$ if and only if $b_{\xi_n} \to b_\xi$ uniformly on bounded subsets of X. If X is proper then X^∞ is compact [11, Definition II.8.6]. Our interest in topologizing X^∞ comes from optimization: it will be desirable to have a large class of compact subsets of X^∞ so that certain continuous functions defined on X^∞ will attain their maximum. In the literature, the set X^∞ is often endowed with a stronger metric topology induced by an angular metric (see for example [11, Proposition II.9.7], [17]). This angular metric is convenient for the study of large-scale geometry of the space X but is often too strong to be useful for our purposes. For example, the boundary at infinity of the n-dimensional hyperbolic space $\mathbb{H}^n$ is discrete in this angular metric, whereas it is homeomorphic to $\mathbb{S}^{n-1}$ in the cone topology (see Example 2.2). In particular, the only compact sets are finite when the angular metric is used. Henceforth we will only make use of the cone topology.

Convergence in X^∞ can be understood more geometrically in terms of pointwise convergence of rays, a characterization stated in [13, p.7] without proof. We formulate this precisely in the next proposition and give a proof for the sake of completeness. Note the proof and result are similar to [11, Proposition II.8.19], though the result therein deals with unbounded sequences in X that converge to points in X^∞ whereas we emphasize sequences converging from within X^∞.

Proposition 2.1. Let X be a Hadamard space with boundary X^∞, let $\{\xi_n\}_{n=1}^\infty \subseteq X^\infty$ be a sequence, and let $\xi \in X^\infty$. For each $n \in \mathbb{N}$, denote the ray issuing from $\bar{x} \in X$ with direction ξ_n by r_n, and let r be the ray issuing from $\bar{x}$ with direction ξ. Then $\xi_n \to \xi$ in X^∞ if and only if $r_n(\delta) \to r(\delta)$ for all $\delta > 0$.

Proof. Suppose first that $\xi_n \to \xi$ in X^∞, which is to say $b_{\xi_n} \to b_\xi$ uniformly on bounded subsets of X. Fix $\delta > 0$ and let $\varepsilon > 0$ be arbitrary. By assumption, there exists $N \in \mathbb{N}$ such that

$$|b_{\xi_m}(r_n(\delta)) + \delta| = |b_{\xi_m}(r_n(\delta)) - b_{\xi_n}(r_n(\delta))| < \frac{\varepsilon^2}{2\delta} \text{ for all } n, m \geq N. \tag{4}$$

The inequality in [11, Lemma II.8.21(2)] asserts

$$\frac{d(r_n(\delta), r_m(\delta))^2}{2\delta} \leq d(r_m(\delta + t), r_n(\delta)) - d(r_m(\delta + t), r_m(\delta)) \text{ for all } t \geq 0.$$

Making the substitution $s = \delta + t$, this simplifies to

$$\frac{d(r_n(\delta), r_m(\delta))^2}{2\delta} \leq d(r_m(s), r_n(\delta)) - s + \delta \text{ for all } s \geq \delta.$$

Letting $s \to \infty$ we conclude

$$\frac{d(r_n(\delta), r_m(\delta))^2}{2\delta} \leq b_{\xi_m}(r_n(\delta)) + \delta.$$

Now use (4) to bound the righthand side by $\varepsilon^2/(2\delta)$, which leads to

$$d(r_n(\delta), r_m(\delta)) \leq \varepsilon \text{ for all } n, m \geq N.$$

Since $\varepsilon > 0$ was arbitrary, the sequence $\{r_n(\delta)\}_{n=1}^{\infty}$ is Cauchy and thus converges since X is complete. It remains to prove that its limit is $r(\delta)$. It is easy to see from (4) that $\{r_n(\delta)\}_{n=1}^{\infty}$ is an infimizing sequence for b_ξ that lies in the sphere of radius δ around $\bar{x}$, denoted $S_\delta(\bar{x})$. It follows that the limit $\bar{z} = \lim_{n\to\infty} r_n(\delta)$ is in $S_\delta(\bar{x})$. Since $r(\delta)$ is the unique minimizer of b_ξ on $S_\delta(\bar{x})$ ([11, Proposition II.8.22]), we deduce

$$-\delta = b_\xi(r(\delta)) = \min_{z \in S_\delta(\bar{x})} b_\xi(z) \leq b_\xi(\bar{z}) = \lim_{n\to\infty} b_\xi(r_n(\delta)) = -\delta.$$

Thus equality holds throughout, and $b_\xi(\bar{z}) = -\delta$. The uniqueness of b_ξ's minimizer on $S_\delta(\bar{x})$ forces $\bar{z} = \lim_{n\to\infty} r_n(\delta) = r(\delta)$ as desired.

Conversely, let us suppose that $r_n(\delta) \to r(\delta)$ for all $\delta > 0$ and prove $b_{\xi_n} \to b_\xi$ uniformly on bounded subsets. It suffices to prove $b_{\xi_n} \to b_\xi$ uniformly on each ball $B_\delta(\bar{x})$ for all $\delta > 0$. Fix $\delta > 0$ and let $\varepsilon > 0$ be arbitrary. The inequality in [11, Lemma II.8.21(1)] asserts the existence of $R > 0$ such that

$$0 \leq d(z, r_n(R)) + t - d(r_n(R + t), z) \leq \frac{\varepsilon}{2} \text{ for all } z \in B_\delta(\bar{x}), t \geq 0, n \in \mathbb{N}. \tag{5}$$

For any $z \in B_\delta(\bar{x}), n, m \in \mathbb{N}, t \geq 0$, two applications of (5) along with the triangle inequality gives us

$$|d(z, r_n(R + t)) - d(z, r_m(R + t))| \leq \varepsilon + |d(z, r_n(R)) - d(z, r_m(R))|. \tag{6}$$

Letting $t \to \infty$ in (6), we deduce the following inequality for all $z \in B_\delta(\bar{x})$:

$$|b_{\xi_n}(z) - b_{\xi_m}(z)| \leq \varepsilon + |d(z, r_n(R)) - d(z, r_m(R))| \leq \varepsilon + d(r_n(R), r_m(R)).$$

Taking the supremum over $z \in B_\delta(\bar{x})$, it follows that

$$\limsup_{n,m\to\infty} \sup_{z\in B_\delta(\bar{x})} |b_{\xi_n}(z) - b_{\xi_m}(z)| \leq \varepsilon \text{ for all } \varepsilon > 0.$$

As a consequence, $\lim_{n,m\to\infty} \sup_{z\in B_\delta(\bar{x})} |b_{\xi_n}(z) - b_{\xi_m}(z)| = 0$. That is to say $\{b_{\xi_n}\}_{n=1}^\infty$ satisfies the Cauchy criterion for uniform convergence on $B_\delta(\bar{x})$ as desired. □

Consider the product space $X^\infty \times \mathbb{R}_+$ endowed with the product of the cone topology on X^∞ and the usual topology on $\mathbb{R}_+$. Define an equivalence relation $\sim$ on $X^\infty \times \mathbb{R}_+$ by $(\xi, s) \sim (\xi', s')$ if $s = s' = 0$ or $(\xi, s) = (\xi', s')$. Now define the *boundary cone* CX^∞ as the quotient of $X^\infty \times \mathbb{R}_+$ by $\sim$, endowed with the quotient topology which we will also refer to as the cone topology (context should always prevent any confusion). The equivalence class of $(\xi, s) \in X^\infty \times \mathbb{R}_+$ is denoted $[\xi, s]$. We will sometimes use the notation $[0]$ to denote the equivalence class corresponding to $s = 0$. Note that continuity of a function defined on CX^∞ is equivalent to sequential continuity because of the sequentially characterized topology on $X^\infty \times \mathbb{R}_+$. The cone topology on CX^∞ has been used to study the geometry of Wasserstein space $\mathscr{W}_2(X)$ where X is a Hadamard space [7]. When X is Euclidean $\mathbb{R}^n$, the boundary at infinity is homeomorphic to $\mathbb{S}^{n-1}$ and the boundary cone is homeomorphic to $X = \mathbb{R}^n$. The next example serves to illustrate some of these concepts in a space that is very different from $\mathbb{R}^n$.

Example 2.1. The *tripod* is the Hadamard space X consisting of three copies of the half-line $\mathbb{R}_+$ glued together at the common point 0. A natural choice for the reference point $\bar{x}$ is this common origin. Two rays are asymptotic if and only if they eventually lie in the same copy of $\mathbb{R}_+$, so the boundary X^∞ comprises three equivalence classes: $X^\infty = \{\xi_1, \xi_2, \xi_3\}$. If we denote a point in X by (x, j), where $x \in \mathbb{R}_+, j \in \{1, 2, 3\}$ then we can write the Busemann functions explicitly:

$$b_{\xi_i}(x, j) = \begin{cases} x, & i \neq j \\ -x, & i = j. \end{cases}$$

Proposition 2.1 tells us that a sequence of points in X^∞ converges to a limit $\zeta \in X^\infty$ if and only if the corresponding rays issuing from the origin converge pointwise to the ray defined by ζ. Clearly a sequence of rays in the tripod can only converge pointwise if it is eventually constant, i.e. the only convergent sequences in X^∞ are eventually constant. Thus the boundary X^∞ is discrete since every subset is closed.

Example 2.2. In this example we characterize the cone topology for the boundary of hyperbolic space $X = \mathbb{H}^n$. We use the Poincaré ball model, viewing X as the open unit ball in $\mathbb{R}^n$ with metric $d(p,q) = \operatorname{arccosh}\left(1 + 2\frac{\|p-q\|^2}{(1-\|p\|^2)(1-\|q\|^2)}\right)$. Proposition 2.1 says a sequence $\{\xi_n\}_{n=1}^\infty \subseteq X^\infty$ converges to $\xi \in X^\infty$ if and only if the corresponding rays $\{r_n\}_{n=1}^\infty$ issuing from $\bar{x} = 0 \in \mathbb{H}^n$ converge pointwise to the ray defined by ξ. Such rays are radial lines from the origin to the corresponding boundary point ξ_n —we have once again identified X^∞ with the boundary sphere $\mathbb{S}^{n-1}$. Explicitly, we can write $r_n(t) = \tanh(t/2)\xi_n$. Then

$$\begin{aligned} \xi_n \to \xi \text{ in the cone topology } &\iff r_n(\delta) \to r(\delta) \text{ for all } \delta > 0 \\ &\iff \|\tanh(\delta/2)\xi_n - \tanh(\delta/2)\xi\| \to 0 \text{ for all } \delta > 0 \\ &\iff \|\xi_n - \xi\| \to 0. \end{aligned}$$

We see that convergence in the cone topology on X^∞ is precisely norm convergence for the induced Euclidean norm on $\mathbb{S}^{n-1}$. In particular, X^∞ has a familiar and rich topological structure with many compact subsets.

To shed some light on the cone topology for CX^∞ we give a partial characterization of convergence in terms of convergence in $X^\infty \times \mathbb{R}_+$.

Lemma 2.3. Let $\{[\xi_n, s_n]\}_{n=1}^\infty \subseteq CX^\infty$ be a sequence and let $[\xi, s] \in CX^\infty$. Then

$$[\xi_n, s_n] \to [\xi, s] \begin{cases} \implies s_n \to s, & s = 0 \\ \iff s_n \to s \text{ and } \xi_n \to \xi, & s \neq 0. \end{cases}$$

If X is proper then the converse implication also holds in the case $s = 0$.

Proof. Suppose first that $[\xi_n, s_n] \to [0]$. Then $[\xi_n, s_n]$ is eventually contained in any open neighborhood of $[0]$. Let $q\colon X^\infty \times \mathbb{R}_+ \to CX^\infty$ be the quotient map. For any $\varepsilon > 0$ define the set $W_\varepsilon = q(X^\infty \times [0, \varepsilon))$. It is readily verified that $q^{-1}(W_\varepsilon) = X^\infty \times [0, \varepsilon)$, which is open in $X^\infty \times \mathbb{R}_+$. Thus by the definition of the quotient topology on CX^∞, the set W_ε is open in CX^∞ and contains $[0]$. It follows that $[\xi_n, s_n]$ is eventually contained in W_ε, which then implies $s_n < \varepsilon$ eventually. Since $\varepsilon > 0$ was arbitrary, $s_n \to 0$.

For the converse implication when $s = 0$ we assume X is proper, hence X^∞ is compact. Suppose $s_n \to 0$ and fix any neighborhood U of $[0]$. Then for each $\zeta \in X^\infty$ there exists an open set $W_\zeta \subseteq X^\infty$ and $\varepsilon_\zeta > 0$ such that $(\zeta, 0) \in W_\zeta \times [0, \varepsilon_\zeta) \subseteq q^{-1}(U)$ because $q^{-1}(U)$ is open in $X^\infty \times \mathbb{R}_+$. The sets $\{W_\zeta \times [0, \varepsilon_\zeta)\}_{\zeta \in X^\infty}$ form an open cover of the compact set $X^\infty \times \{0\}$, from which we select a finite subcover $\{W_{\zeta_i} \times [0, \varepsilon_{\zeta_i})\}_{i=1}^N$. Let $V := \bigcup_{i=1}^N W_{\zeta_i} \times [0, \varepsilon_{\zeta_i})$. To summarize, $X^\infty \times \{0\} \subseteq V \subseteq q^{-1}(U)$. Since s_n is eventually smaller than $\min_{i=1,\dots,N} \varepsilon_{\zeta_i}$, we have $(\xi_n, s_n) \in V$ for all n sufficiently large. Then $[\xi_n, s_n] \in q(V) \subseteq q(q^{-1}(U)) = U$ for all n sufficiently large, which says $[\xi_n, s_n]$ converges to $[0]$ since the open set U around $[0]$ was arbitrary.

Now suppose $[\xi_n, s_n] \to [\xi, s]$ with $s > 0$. As before, $[\xi_n, s_n]$ is eventually contained in any open neighborhood of $[\xi, s]$. For any $\varepsilon > 0$ with $s-\varepsilon > 0$, and any open set $U \subseteq X^\infty$ containing ξ, consider the set $W_{U,\varepsilon} = q(U \times (s-\varepsilon, s+\varepsilon))$. As before it is easy to check that $q^{-1}(W_{U,\varepsilon}) = U \times (s-\varepsilon, s+\varepsilon)$ which is open in $X^\infty \times \mathbb{R}_+$, hence $W_{U,\varepsilon}$ is open in CX^∞. It follows that $[\xi_n, s_n]$ is eventually contained in $W_{U,\varepsilon}$, which implies $\xi_n \in U$ eventually and $s_n \in (s-\varepsilon, s+\varepsilon)$ eventually. Since U, ε were arbitrary, we conclude $\xi_n \to \xi$ and $s_n \to s$. Conversely, suppose $\xi_n \to \xi$ and $s_n \to s > 0$. If U is an open neighborhood of $[\xi, s]$ then $(\xi, s) \in q^{-1}(U)$ so there exists an open set $W \subseteq X^\infty$ and $\varepsilon > 0$ such that $(\xi, s) \in W \times (s-\varepsilon, s+\varepsilon) \subseteq q^{-1}(U)$. Then $\xi_n \in W$ eventually, and $s_n \in (s-\varepsilon, s+\varepsilon)$ eventually, implying $[\xi_n, s_n] \in q(W \times (s-\varepsilon, s+\varepsilon)) \subseteq q(q^{-1}(U)) = U$ for all n sufficiently large.

□

Remark 2.4. The converse implication in Lemma 2.3 can fail if X is not proper due to the lack of compactness for X^∞. Indeed, if X is an infinite-dimensional Hilbert space then X^∞ is homeomorphic to the unit sphere $S \subseteq X$ with the norm topology. But passing to the quotient topology introduces undesirable open sets. Choose any continuous function $f : S \to \mathbb{R}_{++}$ with infimal value 0 (the existence of such a function hinges on the noncompactness of S), and let $\{\xi_n\}_{n=1}^\infty$ be an infimizing sequence. The set $V = \{(\xi, s) \in X^\infty \times \mathbb{R}_+ \mid s < f(\xi)\}$ is open in $X^\infty \times \mathbb{R}_+$ because f is continuous. Moreover, $X^\infty \times \{0\} \subseteq V$ and $q^{-1}(q(V)) = V$. Thus $q(V)$ is an open neighborhood of $[0]$ in CX^∞, we have $f(\xi_n) \to 0$, but $[\xi_n, f(\xi_n)] \notin q(V)$ for any $n \in \mathbb{N}$ and so does not converge to $[0]$.

Now define a *pairing* on $X \times CX^\infty$ to be the function

$$\langle \cdot, \cdot \rangle : X \times CX^\infty \to \mathbb{R}, \quad \langle x, [\xi, s] \rangle = \begin{cases} s b_\xi(x), & s > 0 \\ 0, & s = 0. \end{cases}$$

This suggestive notation will be put to use in the next section, after we prove here some important properties of the pairing.

Proposition 2.2. The pairing $\langle \cdot, \cdot \rangle : X \times CX^\infty \to \mathbb{R}$ has the following properties:

1. It is continuous.

2. For all $[\xi, s] \in CX^\infty$, the map $\langle \cdot, [\xi, s] \rangle : X \to \mathbb{R}$ is geodesically convex and s-Lipschitz.

3. For all $[\xi, s] \in CX^\infty$ and $\alpha \geq 0$, the pairing is positively homogeneous in the second argument: $\langle \cdot, [\xi, \alpha s] \rangle = \alpha \langle \cdot, [\xi, s] \rangle$.

Proof. (1) Due to our earlier remarks on the sequential nature of the topology of CX^∞, it suffices to prove that the pairing is sequentially continuous. Suppose that

$x_n \to x$ in X and $[\xi_n, s_n] \to [\xi, s]$ in CX^∞. If $s = 0$, then $s_n \to 0$ by Lemma 2.3, and furthermore:

$$|b_{\xi_n}(x_n)| = |b_{\xi_n}(x_n) - b_{\xi_n}(\bar{x})| \leq d(x_n, \bar{x}) \to d(x, \bar{x}).$$

Thus the sequence $\{b_{\xi_n}(x_n)\}_{n=1}^\infty$ is a bounded sequence of real numbers, whence

$$\langle x_n, [\xi_n, s_n]\rangle = s_n b_{\xi_n}(x_n) \to 0 = \langle x, [0]\rangle$$

since $s_n \to 0$. If $s > 0$ then $s_n \to s$ and $\xi_n \to \xi$ by Lemma 2.3, giving

$$|b_{\xi_n}(x_n) - b_{\xi_n}(x)| \leq d(x_n, x) \to 0.$$

Knowing that $b_{\xi_n}(x) \to b_\xi(x)$ from the definition of convergence in X^∞, we deduce that $\{b_{\xi_n}(x_n)\}_{n=1}^\infty$ and $\{b_{\xi_n}(x)\}_{n=1}^\infty$ share the common limit $b_\xi(x)$. It follows that

$$\langle x_n, [\xi_n, s_n]\rangle = s_n b_{\xi_n}(x_n) \to s b_\xi(x) = \langle x, [\xi, s]\rangle .$$

Thus $\langle \cdot, \cdot \rangle$ is continuous.

(2) Let $[\xi, s] \in CX^\infty$ and consider the map $x \mapsto \langle x, [\xi, s]\rangle$ from X to $\mathbb{R}$. If $s = 0$ then this is the constant function zero which is trivially convex and 0-Lipschitz. Otherwise, this map is $x \mapsto s b_\xi(x)$. Busemann functions are geodesically convex and 1-Lipschitz, and $s > 0$ so this map is also geodesically convex and s-Lipschitz.

(3) By definition.

□

It should be noted that continuity of the pairing is further evidence that the cone topology strikes the right balance for our purposes: it has enough open sets to support the nontrivial family of continuous functions $CX^\infty \ni [\xi, s] \mapsto \langle x, [\xi, s]\rangle$ for $x \in X$, without sacrificing too many compact subsets.

3 Busemann subgradients

The rest of this paper relies fundamentally on the following definition.

Definition 3.1. Consider a subset C of X and a real-valued function $f\colon C \to \mathbb{R}$. A *Busemann subgradient* of f at a point $x \in C$ is an element $[\xi, s] \in CX^\infty$ such that x minimizes $y \mapsto f(y) - \langle y, [\xi, s]\rangle$ over C. The function f is *Busemann subdifferentiable* if it has a Busemann subgradient at every point in C.

A few preliminary observations are in order. Immediate from the definition is the simple observation that $x \in C$ minimizes a function $f\colon C \to \mathbb{R}$ if and only if $[0]$ is a Busemann subgradient of f at x. When the space X is Euclidean $\mathbb{R}^n$, the boundary at infinity can be identified with $\mathbb{S}^{n-1}$ and the Busemann function b_ξ has the form $b_\xi(y) = (\bar{x} - y)^T \xi$ (recall the fixed reference point $\bar{x}$). Then Definition 3.1 says $[\xi, s]$ is a Busemann subgradient at x if and only if

$$f(y) - s\xi^T(\bar{x} - y) \geq f(x) - s\xi^T(\bar{x} - x) \text{ for all } y \in C,$$

where we interpret the expression $s\xi$ here and throughout as identically 0 if $s = 0$. This is equivalent to

$$f(y) \geq f(x) - s\xi^T(y - x) \text{ for all } y \in C.$$

In other words, Busemann subgradients for functions on Euclidean space coincide with the usual notion of subgradient via the identification $[\xi, s] \leftrightarrow -s\xi$. More generally, Definition 3.1 says $[\xi, s]$ is a Busemann subgradient at x if and only if

$$f(y) - sb_\xi(y) \geq f(x) - sb_\xi(x) \text{ for all } y \in C. \tag{7}$$

We adopt the same convention that sb_ξ is identically 0 if $s = 0$. The Busemann functions for asymptotic rays differ only by an additive constant [11, II.8.23(1)] so it is easy to check that

$$b_\xi = b_{\bar{x},\xi} = b_{x,\xi} + b_\xi(x). \tag{8}$$

It follows that (7) is equivalent to

$$f(y) \geq f(x) + sb_{x,\xi}(y) \text{ for all } y \in C. \tag{9}$$

This serves as a version of the subgradient inequality in Hadamard space. An immediate consequence of (9) is that $f : C \to \mathbb{R}$ is lower semicontinuous at any point where it has a Busemann subgradient because Busemann functions themselves are continuous. It seems natural to think of Busemann subgradients $[\xi, s]$ for f at x as rays issuing from x in direction ξ with speed $s \geq 0$. Before we proceed to examples let us show that Busemann subdifferentiable functions defined on geodesically convex sets are geodesically convex.

Proposition 3.1. Suppose $C \subseteq X$ is geodesically convex and $f : C \to \mathbb{R}$ is Busemann subdifferentiable. Then f is geodesically convex on C.

Proof. Take any two points $x, y \in C$ and let $\lambda \in [0, 1]$. Let $\gamma : [0, 1] \to X$ parametrize the geodesic segment $[x, y]$, denote the point $\gamma(\lambda)$ by z_λ, and note $z_\lambda \in C$ by geodesic convexity of C. Use the Busemann subdifferentiability of f to procure a Busemann subgradient $[\xi, s]$ at z_λ. Then two applications of Definition 3.1 yield:

$$f(x) - \langle x, [\xi, s] \rangle \geq f(z_\lambda) - \langle z_\lambda, [\xi, s] \rangle$$

$$f(y) - \langle y, [\xi, s] \rangle \geq f(z_\lambda) - \langle z_\lambda, [\xi, s] \rangle .$$

Multiply the first inequality by λ and the second by $1-\lambda$, and sum the resulting inequalities to obtain

$$(10)\quad \lambda f(x)+(1-\lambda)f(y)\geq f(z_\lambda)+\lambda\left\langle x,[\xi,s]\right\rangle+(1-\lambda)\left\langle y,[\xi,s]\right\rangle-\left\langle z_\lambda,[\xi,s]\right\rangle.$$

Convexity of the pairing in the first argument (Proposition 2.2(2)) implies

$$\lambda\left\langle x,[\xi,s]\right\rangle+(1-\lambda)\left\langle y,[\xi,s]\right\rangle\geq\left\langle z_\lambda,[\xi,s]\right\rangle.$$

Then it follows from (10) that

$$\lambda f(x)+(1-\lambda)f(y)\geq f(z_\lambda),$$

so f is geodesically convex on C since $x,y\in C$ were arbitrary. □

Before going any further let us mention a different but related usage of the Busemann function as it pertains to convex analysis in Hadamard spaces. Given a geodesically convex function $f\colon X\to\mathbb{R}$, the paper [17] defines the *asymptotic Legendre-Fenchel conjugate* $f^*\colon CX^\infty\to(-\infty,+\infty]$ by

$$f^*([\xi,s])=\sup_{x\in X}\left\{-\left\langle x,[\xi,s]\right\rangle-f(x)\right\}.$$

Note that this definition depends implicitly on the choice of reference point $\bar{x}$, but only up to an additive term (by (8)). That is, if we use subscripts to denote the choice of reference point then

$$f^*_{\bar{x}}([\xi,s])=f^*_{\hat{x}}([\xi,s])-\left\langle\hat{x},[\xi,s]\right\rangle\ \text{ for all } [\xi,s]\in CX^\infty.$$

The definition of f^* immediately yields a version of the classical Fenchel-Young inequality:

$$f^*([\xi,s])+f(x)\geq-\left\langle x,[\xi,s]\right\rangle\ \text{ for all } [\xi,s]\in CX^\infty, x\in X.$$

If $f\colon X\to\mathbb{R}$ is Busemann subdifferentiable, any Busemann subgradient $[\xi,s]$ at $x\in X$ gives the following inequality by definition:

$$f(y)-\left\langle y,[\xi,s]\right\rangle\geq f(x)-\left\langle x,[\xi,s]\right\rangle\ \text{ for all } y\in X.$$

Rearranging and taking the supremum over $y\in X$ leads to

$$(11)\qquad \left\langle x,[\xi,s]\right\rangle-f(x)\geq\sup_{y\in X}\left\{\left\langle y,[\xi,s]\right\rangle-f(y)\right\}=: f^\bullet([\xi,s]).$$

This defines a different type of conjugate $f^\bullet\colon CX^\infty\to(-\infty,+\infty]$ satisfying the Fenchel-Young-type inequality

$$(12)\qquad f^\bullet([\xi,s])+f(x)\geq\left\langle x,[\xi,s]\right\rangle\ \text{ for all } [\xi,s]\in CX^\infty, x\in X.$$

Inequality (11) shows that (12) holds with equality whenever v is a Busemann subgradient for f at x. Conversely, it is easy to show that whenever (12) holds with equality, v is a Busemann subgradient for f at x. The upshot is that Busemann subgradients can be characterized by equality in a Fenchel-Young-type inequality, in exactly the same way as usual subgradients. Note the only difference between the definitions of f^* and $f^\bullet$ is the sign of the pairing term.

The next few examples tease out the relationship between Busemann and geodesic convexity. Notably, the converse of Proposition 3.1 is false in general but essentially true in Euclidean space. Thus the emphasis on Busemann subdifferentiability in this paper is indeed a restrictive condition and the stronger properties of Busemann subdifferentiable functions in a general Hadamard space are what enable the development of efficient algorithms in the sequel.

Example 3.1. Consider the tripod X and let $C \subseteq X$ be the union of any two out of the three defining rays (including the origin). Then C is geodesically convex so the function $f(x) = \text{dist}(x, C)$ is geodesically convex, but we claim that f has no Busemann subgradient at any point outside of C.

Suppose x_0 lies in the open ray disjoint from C. Assume f has a Busemann subgradient $[\xi, s]$ at x_0, with $s > 0$ because x_0 does not minimize f. Then we have the inequality

$$f(y) - sb_\xi(y) \geq f(x_0) - sb_\xi(x_0) \text{ for all } y \in X. \tag{13}$$

Since x_0 does not minimize f we must have $s > 0$. At least one of the rays defining C must be distinct from the ray emanating from the origin with direction ξ, so take y to be a point in such a ray r. Then $f(y) = 0$, while $-sb_\xi(y)$ blows up to $-\infty$ as y travels along r away from the origin (see Example 2.1). For y sufficiently far along r this leads to a violation of (13).

The situation is no better even if X is a manifold, as the next example shows.

Example 3.2. We construct an example in $X = \mathbb{H}^2$, the Poincaré disk. The subset $F = \{(x, y) \in \mathbb{H}^2 \mid y = 0\}$ is closed and geodesically convex (it is a geodesic line in $\mathbb{H}^2$), so the function $f(p) = \text{dist}(p, F)$ is geodesically convex ([11, II.2 Corollary 2.5]). However, f is not Busemann subdifferentiable on any neighborhood of F.

To see this, let U be a neighborhood of F, consider any $(x_0, y_0) \in U$ with $y_0 \neq 0$, and suppose there exists a Busemann subgradient $[\xi, s]$ at (x_0, y_0), with $s > 0$ because (x_0, y_0) does not minimize f. We can identify $\xi \in X^\infty$ with a point $\xi = (\xi_1, \xi_2) \in \mathbb{S}^1$, and we will use this to write the Busemann function b_ξ explicitly. By definition $f - \langle \cdot, [\xi, s] \rangle$ is minimized at (x_0, y_0), i.e.

$$\text{dist}((t,0), F) - \langle (t,0), [\xi, s] \rangle \geq \text{dist}((x_0, y_0), F) - \langle (x_0, y_0), [\xi, s] \rangle \text{ for all } t \in (-1, 1).$$

Since $(t, 0) \in F$ for all $t \in (-1, 1)$ this is equivalent to

$$-sb_\xi(t, 0) \geq \text{dist}((x_0, y_0), F) - \langle (x_0, y_0), [\xi, s] \rangle \text{ for all } t \in (-1, 1). \tag{14}$$

Choosing our reference point to be $(\bar{x}, \bar{y}) = (0, 0)$, [11, II.8.24(2)] tells us the Busemann function b_ξ has the following form:

$$b_\xi(x, y) = -\log\left(\frac{1 - x^2 - y^2}{(x - \xi_1)^2 + (y - \xi_2)^2}\right). \tag{15}$$

Substituting into (14) gives

$$s \log\left(\frac{1 - t^2}{(t - \xi_1)^2 + \xi_2^2}\right) \geq \operatorname{dist}((x_0, y_0), F) - \langle (x_0, y_0), [\xi, s]\rangle .$$

If $\xi_2 \neq 0$ let $t \to 1$, and if $\xi_2 = 0$ let $t \to -\xi_1$. In either case the lefthand side approaches $-\infty$ while the righthand side remains constant, a contradiction.

Example 3.3. (Euclidean Space) Let $C \subseteq \mathbb{R}^n$ be convex and suppose $f\colon C \to \mathbb{R}$ is convex in the usual sense. A standard convex analysis argument, which can be found in the appendix, shows that if f is locally Lipschitz on C then f has a subgradient at each $x \in C$. By our remarks following Definition 3.1, this says f is Busemann subdifferentiable on C. Together with Proposition 3.1, we conclude that convexity and Busemann subdifferentiability coincide for locally Lipschitz functions on convex sets in Euclidean space.

The next example gives a fundamental way to construct nontrivial Busemann subdifferentiable functions on a general Hadamard space.

Example 3.4. (Busemann envelopes) Suppose $g\colon CX^\infty \to (-\infty, +\infty]$ is a function such that, for any $x \in X$, the function $[\xi, s] \mapsto \langle x, [\xi, s]\rangle - g([\xi, s])$ attains its maximum on CX^∞. Define a function $f\colon X \to \mathbb{R}$ by

$$f(x) = \max\left\{\langle x, [\xi, s]\rangle - g([\xi, s])\colon [\xi, s] \in CX^\infty\right\}.$$

Then f is Busemann subdifferentiable. Indeed, any $[\xi, s]$ attaining the maximum for a given point $x \in X$ gives

$$f(y) - \langle y, [\xi, s]\rangle \geq -g([\xi, s]) = f(x) - \langle x, [\xi, s]\rangle \text{ for all } y \in X.$$

This says $[\xi, s]$ is a Busemann subgradient at x according to Definition 3.1.

Example 3.5. (Busemann functions) Perhaps the most basic example of Busemann subdifferentiable functions are Busemann functions themselves. Suppose r is a ray on X issuing from $x_0 \in X$, and set $\zeta = r(\infty)$. Let $K \subseteq CX^\infty$ be the singleton $\{[\zeta, 1]\}$ and define $g\colon CX^\infty \to (-\infty, +\infty]$ by

$$g([\xi, s]) = \begin{cases} \langle x_0, [\xi, s]\rangle, & [\xi, s] \in K \\ +\infty, & [\xi, s] \notin K. \end{cases}$$

Then we derive:

$$\begin{aligned} b_r(x) &= b_{x_0,\zeta}(x) \\ &= b_\zeta(x) - b_\zeta(x_0) \\ &= \langle x, [\zeta, 1] \rangle - b_\zeta(x_0) \\ &= \max\left\{ \langle x, [\xi, s] \rangle - g([\xi, s]) \colon [\xi, s] \in CX^\infty \right\}. \end{aligned}$$

The first equality is just a consequence of our notation while the second uses (8). We deduce that b_r is a Busemann envelope as in Example 3.4 and is thus Busemann subdifferentiable, with Busemann subgradient $[r(\infty), 1]$ at each $x \in X$.

Example 3.6. (Continuous-compact-representable functions) Our decision to endow X^∞ with the cone topology stems from the family of examples generated by the following natural construction. Suppose $K \subseteq CX^\infty$ is compact and $g \colon CX^\infty \to (-\infty, +\infty]$ is continuous on K and $+\infty$ outside of K. Then g satisfies the assumptions of Example 3.4 because the pairing is also continuous (Proposition 2.2(1)). In Euclidean space we recover the class of "almost sublinear" functions f such that $\sup_{x \in \mathbb{R}^n} |f(x) - h(x)| < \infty$ for some sublinear function h [9, Proposition 4.5]. This should not be viewed as a restricted class, since in practice we optimize functions over closed and bounded feasible regions and can typically extend the functions to have this "almost sublinear" growth off of the feasible set without changing the underlying optimization problem. If the space X is proper, the boundary X^∞ is compact. Then compact subsets of CX^∞ are abundant, because if $F \subseteq X^\infty$ is closed and $I \subseteq \mathbb{R}_+$ is compact, then $q(F \times I)$ is compact in CX^∞ (here $q \colon X^\infty \times \mathbb{R}_+ \to CX^\infty$ is the quotient map).

Example 3.7. (Euclidean convex functions) If $f \colon \mathbb{R}^n \to \mathbb{R}$ is convex, then Fenchel biconjugation allows us to write $f(x) = \sup\left\{ y^T x - f^*(y) \colon y \in \mathbb{R}^n \right\}$ where $f^* \colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is the convex conjugate of f. The supremum is always attained by choosing any subgradient $y \in \partial f(x)$, which is always nonempty since f is continuous. Furthermore, our earlier remarks show that we can identify $C(\mathbb{R}^n)^\infty$ with $\mathbb{R}^n$ itself via the bijection $\varphi \colon C(\mathbb{R}^n)^\infty \to \mathbb{R}^n$, $\varphi([\xi, s]) = -s\xi$ (viewing ξ as an element of $\mathbb{S}^{n-1}$). If we adopt the reference point $\bar{x} = 0$ then our Busemann functions take the form $b_\xi(z) = -\xi^T z$. Then we derive:

$$\begin{aligned} f(x) &= \max\left\{ y^T x - f^*(y) \colon y \in \mathbb{R}^n \right\} \\ &= \max\left\{ \varphi([\xi, s])^T x - (f^* \circ \varphi)([\xi, s]) \colon [\xi, s] \in C(\mathbb{R}^n)^\infty \right\} \\ &= \max\left\{ -s\xi^T x - (f^* \circ \varphi)([\xi, s]) \colon [\xi, s] \in C(\mathbb{R}^n)^\infty \right\} \\ &= \max\left\{ s b_\xi(x) - (f^* \circ \varphi)([\xi, s]) \colon [\xi, s] \in C(\mathbb{R}^n)^\infty \right\} \\ &= \max\left\{ \langle x, [\xi, s] \rangle - (f^* \circ \varphi)([\xi, s]) \colon [\xi, s] \in C(\mathbb{R}^n)^\infty \right\}. \end{aligned}$$

We see that taking $g = f^* \circ \varphi$ we are in the setting of Example 3.4. Thus we recover the class of real-valued convex functions on $\mathbb{R}^n$. More generally, this same argument applies for continuous convex functions on a Hilbert space.

Example 3.8. (Distance functions) Consider the function $f(x) = d(x, a)$ for any fixed $a \in X$. For each $x \neq a$ let $r_{x,a} \colon \mathbb{R}_+ \to X$ be any ray issuing from x that eventually passes through a (at least one exists by the geodesic extension property). Furthermore, define $\xi_x = r_{x,a}(\infty)$, and let $K = \{[\xi_x, 1] \mid x \in X \setminus \{a\}\}$. Define $g \colon CX^\infty \to (-\infty, +\infty]$ by

$$g([\xi, s]) = \begin{cases} \langle a, [\xi, s] \rangle, & [\xi, s] \in K \\ +\infty, & [\xi, s] \notin K. \end{cases}$$

Let $\tilde{r}(t) = r_{x,a}(d(x, a) + t)$ be a ray issuing from a in direction ξ_x, so $b_{a,\xi_x} = b_{\tilde{r}}$. Then

$$b_{a,\xi_x}(x) = b_{\tilde{r}}(x) = \lim_{t \to \infty} (d(x, r_{x,a}(d(x, a) + t)) - t) = d(x, a). \tag{16}$$

Now write:

$$\begin{aligned} \sup \{\langle x, [\xi, s] \rangle - g([\xi, s]) \colon [\xi, s] \in CX^\infty\} &= \sup \{\langle x, [\xi, s] \rangle - g([\xi, s]) \colon [\xi, s] \in K\} \\ &= \sup \{\langle x, [\xi_z, 1] \rangle - b_{\xi_z}(a) \colon z \in X \setminus \{a\}\} \\ &= \sup \{b_{\xi_z}(x) - b_{\xi_z}(a) \colon z \in X \setminus \{a\}\}. \end{aligned} \tag{17}$$

If $x = a$ then this last expression is identically zero, so choose any $z \in X \setminus \{a\}$ and observe $[\xi_z, 1] \in CX^\infty$ attains the supremum in (17). We have the upper bound $b_{\xi_z}(x) - b_{\xi_z}(a) \leq d(x, a)$ for all $z \in X \setminus \{a\}$ because Busemann functions are 1-Lipschitz. On the other hand, if $x \neq a$ then choosing $z = x$ we use (8) and (16) to conclude

$$b_{\xi_x}(x) - b_{\xi_x}(a) = b_{a,\xi_x}(x) = d(x, a).$$

Thus the supremum in (17) is attained by $[\xi_x, 1] \in CX^\infty$. It follows that for all $x \in X$ the supremum in the first line above is always attained, allowing us to say

$$f(x) = d(x, a) = \max \{\langle x, [\xi, s] \rangle - g([\xi, s]) \colon [\xi, s] \in CX^\infty\}$$

is Busemann subdifferentiable by Example 3.4.

Remark 3.9. If $f_1, \ldots, f_m$ are Busemann subdifferentiable functions on a common set C, then $\max\{f_1, \ldots, f_m\}$ is Busemann subdifferentiable on C: a Busemann subgradient at $x \in C$ is obtained by choosing a Busemann subgradient for any function attaining the maximum.

Example 3.10. (Distance to (horo)balls) Generalizing Example 3.8, we can show that the distance functions to balls and horoballs are Busemann subdifferentiable. Fix any $a \in X, \rho \geq 0$ and consider the function $f(x) = \operatorname{dist}(x, B_\rho(a))$ where $B_\rho(a)$ is the closed ball of radius ρ around a. Then $f(x) = \max\{0, d(x, a) - \rho\}$ and the Busemann subdifferentiability of f follows from Example 3.11 and Remark 3.9.

Now consider a ray r in X and the corresponding horoball $H_r = b_r^{-1}((-\infty, 0])$. We will prove that $g(x) = \operatorname{dist}(x, H_r)$ can be written as $g(x) = \max\{0, b_r(x)\}$, proving g is Busemann subdifferentiable by Example 3.5 and Remark 3.9. Indeed, if $x \in H_r$ then $\operatorname{dist}(x, H_r) = 0 = \max\{0, b_r(x)\}$ so it remains to consider $x \notin H_r$. We use the Lipschitz property of b_r and the definition of H_r to deduce

$$b_r(x) \leq b_r(y) + d(x, y) \leq d(x, y) \text{ for all } y \in H_r.$$

Taking the infimum over $y \in H_r$ implies $b_r(x) \leq \operatorname{dist}(x, H_r)$. On the other hand, for any $t \geq 0$ one has $B_t(r(t)) \subseteq H_r$ because $s \mapsto d(x, r(s)) - s$ is nonincreasing. As a consequence, $\operatorname{dist}(x, H_r) \leq \operatorname{dist}(x, B_t(r(t)))$. But since $x \notin H_r$ we must have $x \notin B_t(r(t))$ for any $t \geq 0$, so $\operatorname{dist}(x, B_t(r(t)) = d(x, r(t)) - t$. To summarize,

$$\operatorname{dist}(x, H_r) \leq d(x, r(t)) - t \text{ for all } t \geq 0.$$

As $t \to \infty$ we get $\operatorname{dist}(x, H_r) \leq b_r(x)$, proving $\operatorname{dist}(x, H_r) = b_r(x)$ for $x \notin H_r$.

We now consider some consequences of Busemann subdifferentiability. We recall the notion of horospherical convexity, a geometric property defined for subsets of Hadamard space in terms of Busemann functions.

Definition 3.2. For a closed set $F \subseteq X$ we say F is *horospherically convex* if for each $x \in \operatorname{bdry} F$ there exists a *supporting ray at* x, which is to say a ray r issuing from x such that

$$F \subseteq \{z \in X \mid b_r(z) \leq 0\}.$$

Then we call the righthand side a *supporting horoball* for F at x, and say the ray r *supports* F at x.

In light of the Busemann functions appearing in both Definitions 3.1 and 3.2, it seems plausible that the level sets of a Busemann subdifferentiable function may be horospherically convex in the same way the level sets of a convex function are convex. Indeed, we show that continuous Busemann subdifferentiable functions typically have horospherically convex level sets.

Proposition 3.2. Suppose $f\colon X \to \mathbb{R}$ is continuous and Busemann subdifferentiable with $\inf_X f < M$. Then $F = \{x \in X \mid f(x) \leq M\}$ is horospherically convex.

Proof. Suppose $\bar{x} \in \operatorname{bdry} F$, so $f(\bar{x}) = M$ by continuity. Since f is Busemann subdifferentiable there exists a Busemann subgradient $[\xi, s]$ at $\bar{x}$, so let r be the ray issuing from $\bar{x}$ with direction ξ. We will show that r supports F at $\bar{x}$. Since $M > \inf_X f$ we have $\bar{x} \notin \operatorname{argmin}_X f$, implying $s > 0$. Then apply (9) to deduce

$$s b_{\bar{x},\xi}(z) + f(\bar{x}) \leq f(z) \text{ for all } z \in X.$$

If $z \in F$ notice $f(\bar{x}) = M \geq f(z)$ so in fact $s b_{\bar{x},\xi}(z) \leq 0$. Dividing by $s > 0$ implies $b_{\bar{x},\xi}(z) \leq 0$. The equality $b_{\bar{x},\xi} = b_r$ is only a matter of notation, so in this way we obtain a supporting horoball at every boundary point of F, rendering F horospherically convex. ☐

A basic result in convex analysis says that real-valued convex functions on Euclidean space are continuous, in which case Proposition 3.2 says essentially that convex functions have convex sublevel sets. Proposition 3.2 also gives us another way to reach the conclusions of Examples 3.1 and 3.2: the functions involved are continuous but their sublevel sets are not horospherically convex, so the functions themselves are not Busemann subdifferentiable.

There is also a relationship between Busemann subgradients and Lipschitz continuity of the function, generalizing the relationship between boundedness of subgradients and Lipschitz continuity for convex functions on $\mathbb{R}^n$ (c.f. [5, Theorem 3.61]).

Proposition 3.3. Let $f\colon C \subseteq X \to \mathbb{R}$ be Busemann subdifferentiable, and suppose $L \geq 0$.

(i) If at each point in C the function f admits a Busemann subgradient $[\xi, s]$ with $s \leq L$, then f is L-Lipschitz on C.

(ii) If f is L-Lipschitz on C and C is open, then every Busemann subgradient $[\xi, s]$ for f at x in C satisfies $s \leq L$.

Proof. (i) For $x, y \in C$ choose the assumed Busemann subgradients $[\xi_x, s_x], [\xi_y, s_y]$ at x, y respectively. Applying Definition 3.1 twice gives:

$$\begin{aligned} f(x) - \langle x, [\xi_x, s_x]\rangle &\leq f(y) - \langle y, [\xi_x, s_x]\rangle\,, \\ f(y) - \langle y, [\xi_y, s_y]\rangle &\leq f(x) - \langle x, [\xi_y, s_y]\rangle\,. \end{aligned} \tag{18}$$

Rearranging the first inequality in (18) implies

$$f(x) - f(y) \leq \langle y, [\xi_x, s_x]\rangle - \langle x, [\xi_x, s_x]\rangle \leq s_x d(x, y) \leq L d(x, y),$$

where the second inequality comes from Proposition 2.2(2). Arguing similarly for the second inequality in (18), we find f is L-Lipschitz on C.

(ii) Suppose f is L-Lipschitz on C and take any Busemann subgradient $[\xi, s]$ at $x \in C$. If $s = 0$ there is nothing to prove, so we assume $s > 0$. Let r be the ray issuing from x with direction ξ. Since X has the geodesic extension property r can be extended to a geodesic line $\tilde{r}\colon \mathbb{R} \to X$ by [11, Lemma II.5.8(2)]. The geodesic line $\tilde{r}$ is continuous with $\tilde{r}(0) = x \in C$, and C is open so there exists $\varepsilon > 0$ such that $\tilde{r}(-\varepsilon) \in C$. By (9) we have

$$s b_{x,\xi}(y) + f(x) \leq f(y) \text{ for all } y \in X.$$

Plugging in $y_\varepsilon = \tilde{r}(-\varepsilon)$ and using the definition of $b_{x,\xi} = b_r$ we find $b_{x,\xi}(y_\varepsilon) = \varepsilon$ from which we derive:

$$s\varepsilon = s b_{x,\xi}(y_\varepsilon) \leq f(y_\varepsilon) - f(x) \leq L d(y_\varepsilon, x) = L\varepsilon.$$

Divide through by $\varepsilon > 0$ to conclude $s \leq L$. □

To broaden the class of Busemann subdifferentiable functions on a general Hadamard space, we prove a basic chain rule for compositions of Busemann subdifferentiable functions with scalar convex functions. Then we can leverage previous examples, opening the door to more interesting problems with additive structure.

Proposition 3.4. (Chain rule) Suppose $f\colon C \to \mathbb{R}$ is Busemann subdifferentiable, and $g\colon \mathbb{R} \to \mathbb{R}$ is a nondecreasing convex function. Then $g \circ f$ is Busemann subdifferentiable, and if f has a Busemann subgradient $[\xi, s]$ at $x \in C$ then for any $\alpha \in \partial g(f(x))$, $[\xi, \alpha s]$ is a Busemann subgradient for $g \circ f$ at x.

Proof. Fix $x \in C$ and let $[\xi, s]$ be a Busemann subgradient for f at x. For any $\alpha \in \partial g(f(x))$ we must show $y \mapsto (g \circ f)(y) - \langle y, [\xi, \alpha s]\rangle$ is minimized over C by x. Note that $\alpha \geq 0$ because g is nondecreasing. Since $[\xi, s]$ is a Busemann subgradient for f at x we have

$$f(y) - \langle y, [\xi, s]\rangle \geq f(x) - \langle x, [\xi, s]\rangle \text{ for all } y \in C. \tag{19}$$

Likewise, since $\alpha \in \partial g(f(x))$ we have

$$g(f(y)) - \alpha f(y) \geq g(f(x)) - \alpha f(x) \text{ for all } y \in C. \tag{20}$$

Finally we estimate

$$\begin{aligned} g(f(x)) - \langle x, [\xi, \alpha s]\rangle &= g(f(x)) - \alpha \langle x, [\xi, s]\rangle \\ &\leq \alpha(f(x) - \langle x, [\xi, s]\rangle - f(y)) + g(f(y)) \\ &\leq g(f(y)) - \alpha \langle y, [\xi, s]\rangle \\ &= g(f(y)) - \langle y, [\xi, \alpha s]\rangle . \end{aligned}$$

The first and last equalities use positive homogeneity of the pairing in the second slot (Proposition 2.2(3)), while the second line and third lines use (20) and (19) respectively. The inequality (19) is preserved because the coefficient α is nonnegative. □

Example 3.11. (Reparametrized distance functions) Fix $a \in X$, $\sigma \geq 0, b \in \mathbb{R}$, and $p \geq 1$. Taking $f(x) = d(x, a)$ and $g(s) = \sigma s^p + b$, we combine the result of Example 3.8 with the chain rule (Proposition 3.4) to immediately yield the Busemann subdifferentiability of $g \circ f$ on X. More explicitly, $[0]$ remains a Busemann subgradient at a and at each $x \in X \setminus \{a\}$ we obtain a Busemann subgradient $[r_{x,a}(\infty), p\sigma d(x, a)^{p-1}]$.

Remark 3.12. Example 3.8 showed that $x \mapsto \operatorname{dist}(x, B)$ is Busemann subdifferentiable when B is a singleton, and Example 3.10 generalized this conclusion to sets B that are balls or horoballs. Given the relationship between Busemann subdifferentiability and horospherical convexity, some of which is discussed in Proposition 3.2, one might ask if the distance function to a horospherically convex set is Busemann subdifferentiable. After all, balls, horoballs, and singletons are particular instances of horospherically convex sets. However, this conclusion is false even in Euclidean space. If C is any set of two distinct points in $\mathbb{R}^n$, then C is horospherically convex but not convex—in particular, horospherical convexity is not stronger than geodesic convexity and the distance to C is not convex. It is not hard to see that horospherical convexity does imply geodesic convexity for closed subsets of X with nonempty interior, but we defer a detailed discussion to a future work.

The next example shows that the behaviour illustrated in Remark 3.12 persists even for distances to sets that are both horospherically and geodesically convex.

Example 3.13. (Non-Busemann subdifferentiability of a distance) A simple but interesting example of a Hadamard space (X, d) can be obtained by gluing five Euclidean quadrants along their edges to form a cycle, which can be realized concretely as the union of the following quadrants in $\mathbb{R}^3$ endowed with the intrinsic metric induced by Euclidean distance:

$$\mathbb{R}_+ \times \mathbb{R}_+ \times \{0\}\,,\ \mathbb{R}_+ \times \mathbb{R}_- \times \{0\}\,,\ \mathbb{R}_- \times \mathbb{R}_+ \times \{0\}\,,\ \{0\} \times \mathbb{R}_- \times \mathbb{R}_+,\ \mathbb{R}_- \times \{0\} \times \mathbb{R}_+.$$

This space arises as a subspace of the tree space $\mathcal{T}_4$ (see Section 6), and is illustrated in [8, Figure 12] as well as Figure 4.1. We will revisit this space when discussing the Busemann subdifferentiability of a sum of Busemann subdifferentiable functions in Example 4.1.

Define $C \subseteq X$ to be the geodesic segment joining the points $(1, 0, 0)$ and $(0, 1, 0)$. It is not hard to see that C is horospherically convex by considering rays parallel to $r(t) = (t/\sqrt{2})(1, 1, 0)$, and C is obviously geodesically convex. But one can show that if $f(x) = \operatorname{dist}(x, C)$ then f is not Busemann subdifferentiable at $x_0 = (1, 1, 0)$. The argument is structurally similar to those of Examples 3.1, 3.2, but requires checking a few cases so we relegate this casework to the appendix.

To conclude this section we explore the relationship between Busemann subgradients and *subgradients* in the sense of the recent work [21]. We briefly review the tangent space construction of these objects; many of the subsequent notions are presented in detail in e.g. [11]. Fix $x \in X$ and denote by $\Theta_x X$ the set of all nonconstant geodesics issuing at x.

Given three points $x, y, z \in X$, a *geodesic triangle* $\Delta = \Delta(x, y, z)$ is the union of three geodesic segments (its sides) joining each pair of points. A *comparison triangle* for Δ is a triangle $\Delta(\bar{x}, \bar{y}, \bar{z})$ in $\mathbb{R}^2$ that has side lengths equal to those of Δ. Take two geodesics $\gamma, \eta \in \Theta_x$, let $\gamma_t = \gamma(t), \eta_s = \eta(s)$ for $t, s \geq 0$ and let $\Delta(\bar{\gamma}_t, \bar{x}, \bar{\eta}_s)$ be a comparison triangle in $\mathbb{R}^2$ for $\Delta(\gamma_t, x, \eta_s)$. Then the angle $\angle\bar{\gamma}_t\bar{x}\bar{\eta}_s$ is a nondecreasing function of both t and s and the *Alexandrov angle* between γ and η is well-defined as the following limit:

$$\angle(\gamma, \eta) := \lim_{t,s\downarrow 0} \angle\bar{\gamma}_t\bar{x}\bar{\eta}_s.$$

The Alexandrov angle induces a metric on the set $\Sigma_x X$ of equivalence classes of geodesics in $\Theta_x X$, where two geodesics $\gamma, \eta \in \Theta_x X$ are considered equivalent if $\angle(\gamma, \eta) = 0$. For a geodesic $\gamma \in \Theta_x X$, we denote by $[\gamma]$ its equivalence class. The *tangent space* $T_x X$ of X at x is the Euclidean cone over the metric space $(\Sigma_x X, \angle)$ (see [11, Chapter I, Definition 5.6]). For $v = ([\gamma], r), w = ([\eta], s) \in T_x X$ the *scalar product* of v and w is given by $\langle\langle v, w\rangle\rangle = rs \cos\angle(\gamma, \eta)$. For a geodesically convex function $f\colon X \to \mathbb{R}$, the *subdifferential* of f at x, denoted $\partial f(x)$, can be characterized (c.f. [21, Proposition 4.4]) as the elements $([\eta], s) \in T_x X$ satisfying the *subgradient inequality*

$$\langle\langle([\eta], s), ([\gamma_y], d(x, y))\rangle\rangle + f(x) \leq f(y) \text{ for all } y \in X, \tag{21}$$

where $\gamma_y\colon [0, d(x, y)] \to X$ denotes the geodesic from x to y. Any element $v \in \partial f(x)$ is called a *subgradient* of f at x. It is shown in [21, Theorem 4.11] that continuous geodesically convex functions have subgradients everywhere, whereas we have seen already in Example 3.2 that such functions may not be Busemann subdifferentiable. In the forthcoming Example 4.1 we will examine this failure more closely for a function on the quadrant space of Example 3.13 by demonstrating a subgradient that does not give rise to a Busemann subgradient. On the other hand, the next theorem shows that Busemann subgradients, when they exist, can be understood as subgradients.

Theorem 3.14. Let $f\colon C \to \mathbb{R}$ be geodesically convex and suppose f has a Busemann subgradient $[\xi, s]$ at $x \in C$. Let r be the ray issuing from x with direction ξ. Extend r to a geodesic line $\tilde{r}\colon \mathbb{R} \to X$ and define a new ray $r_-\colon \mathbb{R}_+ \to X$ by $r_-(t) = \tilde{r}(-t)$. Then $([r_-], s) \in \partial f(x)$.

Proof. By definition of the scalar product on $T_x X$, the subgradient inequality we wish to prove is

$$sd(x, y) \cos\angle(r_-, \gamma_y) + f(x) \leq f(y) \text{ for all } y \in X.$$

Busemann subdifferentiability of f implies, by (9),

$$sb_r(y) + f(x) \leq f(y) \text{ for all } y \in X.$$

Thus it is enough to prove (the case $s = 0$ being trivial):

$$b_r(y) \geq d(x, y) \cos\angle(r_-, \gamma_y) \text{ for all } y \in X. \tag{22}$$

We proceed in two steps, first proving $\cos\angle(r_-, \gamma_y) \leq -\cos\angle(r, \gamma_y)$ $(*)$. Using the triangle inequality for angles we find:

$$\pi = \angle(r, r_-) \leq \angle(r_-, \gamma_y) + \angle(r, \gamma_y).$$

Hence $\pi - \angle(r, \gamma_y) \leq \angle(r_-, \gamma_y)$, and taking the cosine reverses the inequality to prove $(*)$:

$$\cos\angle(r_-, \gamma_y) \leq \cos(\pi - \angle(r, \gamma_y)) = -\cos\angle(r, \gamma_y).$$

Next we prove $b_r(y) \geq -d(x, y)\cos\angle(r, \gamma_y)$ $(**)$. By the law of cosines we have:

$$d(r(t), y)^2 \geq d(x, r(t))^2 + d(x, y)^2 - 2d(x, r(t))d(x, y)\cos\angle(r, \gamma_y).$$

Rearrangement leads to

$$\frac{d(r(t), y)^2 - t^2}{2t} \geq \frac{d(x, y)^2}{2t} - d(x, y)\cos\angle(r, \gamma_y). \tag{23}$$

Since r is unit-speed some algebra shows

$$\lim_{t\to\infty} \frac{d(r(t), y)^2 - t^2}{2t} = \lim_{t\to\infty} (d(r(t), y) - t) = b_r(y). \tag{24}$$

Using (24) and sending $t \to \infty$ in (23) readily yields $(**)$. Combining $(*)$ and $(**)$ gives

$$d(x, y)\cos\angle(r_-, \gamma_y) \leq -d(x, y)\cos\angle(r, \gamma_y) \leq b_r(y).$$

This proves (22). □

4 Incremental minimization of sums

We are nearly ready to discuss the role of Busemann subgradients in concrete optimization algorithms of the form (3). Let us pause to mention how an alternative approach based on horospherical convexity is insufficient for tackling such problems, as well as a challenge associated with both Busemann subgradients and horospherical convexity that helped shape the algorithmic paradigm we are soon to present. The iteration described as the horospherical subgradient method in [22] assumes that the function to be minimized is *horospherically quasiconvex*—that is, it has horospherically convex level sets. As discussed in the introduction, this has its limitations and is difficult to extend to the problem of minimizing a sum of functions for at least two reasons:

(i) Horospherical quasiconvexity is not preserved under summation. Simple counterexamples exist for quasiconvex functions on Euclidean space. This necessitates the treatment of each summand individually, for which point (ii) below becomes a concern.

(ii) Different functions can have the same level sets. For example, an algorithm using only level set information cannot distinguish between $x \mapsto d(x, a)^p$ and $x \mapsto d(x, a)$, and thus cannot hope to distinguish p-means from medians.

Horospherical quasiconvexity entails only a supporting ray to a level set at a given point, while Busemann subdifferentiability yields both a ray and an additional scalar—the speed at which that ray is traversed. It is exactly this extra information that allows Busemann subgradients to resolve point (ii) above.

We turn now to the question of whether Busemann subdifferentiability is preserved by addition, the counterpart to point (i) above. Geodesic convexity, of course, is stable under addition. Calculus for sums of convex functions is at the heart of convex analysis and optimization in Euclidean space. One might hope that Busemann subdifferentiability is preserved under addition in general, perhaps with a useful sum rule to employ in algorithms. It turns out this is false, one of the striking differences between Busemann subdifferentiability and geodesic convexity in non-Euclidean spaces. We outline a simple counterexample below, with the details in the appendix. This finding motivates the use of splitting to minimize a sum of Busemann subdifferentiable functions since Busemann subgradients for the sum do not exist.

Example 4.1. (Non-Busemann subdifferentiability of a sum) Recall the quadrant space X of Example 3.13. Let $a_1 = (0, -1, 0), a_2 = (-1, 0, 0)$ be points in X and define $f\colon X \to \mathbb{R}$ by

$$f(x) = \frac{1}{2}d(x, a_1)^2 + \frac{1}{2}d(x, a_2)^2.$$

By Example 3.11, f is a sum of Busemann subdifferentiable functions. Taking $\bar{x} = (1/4, 1/4, 0)$, we claim that the level set $f_{\bar{x}} = \{z \in X \mid f(z) \leq f(\bar{x})\}$ is not horospherically convex. In the appendix, we show that there is no supporting horoball for $f_{\bar{x}}$ at $\bar{x}$. This example is illustrated in Figure 4.1, where the black point is $\bar{x}$, the blue set corresponds to the horoball at $\bar{x}$ generated by the ray moving towards the *spine* (the half-line $x = 0, y = 0, z \geq 0$), and the red set is the level set $f_{\bar{x}}$. Only the parts of the horoball and level set contained in the plane $z = 0$ are shown. Since f is continuous, it cannot be Busemann subdifferentiable else this level set would be horospherically convex by Proposition 3.2. In particular, f is not Busemann subdifferentiable at $\bar{x}$.

As promised before Theorem 3.14, we can still find an explicit subgradient for f at $\bar{x}$. Consider the geodesic $\eta\colon [0,1] \to X$, $\eta(t) = \bar{x} + t(1,1,0)$ in the quadrant $\mathbb{R}_+ \times \mathbb{R}_+ \times \{0\} \subseteq X$. We show in the appendix that $([\eta], 3/\sqrt{2}) \in T_{\bar{x}}X$ is a subgradient for f at $\bar{x}$ by verifying the subgradient inequality (21). Roughly speaking, the ability to convert Busemann subgradients to subgradients but not the other way around can be attributed to the gap incurred by the following inequality used in the proof of Theorem 3.14:

$$b_\eta(y) \geq -d(\bar{x}, y) \cos \angle(\eta, \gamma_y) \text{ for all } y \in X.$$

The expression on the righthand side, defining the subgradient inequality (21), is attentive to the local geometry of the space X because the angle between geodesics

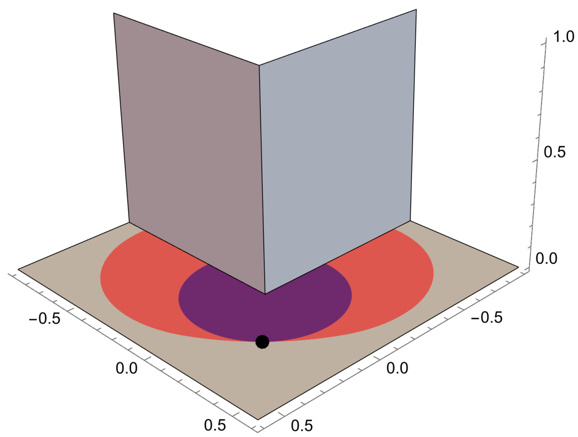

Figure 4.1: Part of a level set that fails to be horospherically convex.

is determined locally at the point of origin. The Busemann function on the lefthand side, defining the Busemann subgradient inequality (9), is determined instead by the large-scale geometry of X rather than the geometry of X near $\bar{x}$. The existence of a nontrivial Busemann subgradient requires the function f to be compatible with the global geometry of the space X.

Let us revisit the original problem of minimizing the objective (3) and collect our assumptions for deriving an incremental subgradient-type method on Hadamard space. We denote the projection onto a nonempty closed geodesically convex set C by P_C (this is well-defined by [11, Proposition II.2.4]). Define an oracle `Busemann` that accepts a Busemann subdifferentiable function $g\colon C \to \mathbb{R}$ and a point $x \in C$, and returns a Busemann subgradient for g at x: $[\xi, s] \leftarrow \texttt{Busemann}(g, x)$. If $[\xi, s]$ is a Busemann subgradient for g at x, we will use the notation r_ξ to denote the ray issuing from x with direction ξ.

Assumption A.

(i) (X, d) is a Hadamard space with the geodesic extension property.

(ii) $C \subseteq X$ is nonempty, closed, and geodesically convex.

(iii) The objective $f\colon C \to \mathbb{R}$ decomposes as $f = \sum_{i=1}^m f_i$, where each $f_i\colon C \to \mathbb{R}, i = 1, \ldots, m$ is Busemann subdifferentiable.

(iv) The optimal set of $\inf_{x \in C} f(x)$ is nonempty, and denoted X^*. The optimal value of the problem is denoted by f_{opt}.

(v) There is a constant $L \geq 0$ such that for all $i = 1, \ldots, m$, every Busemann subgradient $[\xi, s]$ for f_i at every point in C has $s \leq L$.

Under these assumptions, we are now ready to introduce our basic incremental subgradient method:

Algorithm 1 Incremental Busemann Subgradient Method

Require: $x^0 \in C, \{t_k\}_{k=0}^{\infty} \subseteq \mathbb{R}_{++}$
  for $k = 0, 1, 2 \dots$ **do**
    $x^{k,0} \leftarrow x^k$
    for $i = 0, 1, \dots, m-1$ **do**
      $[\xi_{k,i}, s_{k,i}] \leftarrow \texttt{Busemann}(f_i, x^{k,i})$
      $x^{k,i+1} \leftarrow P_C\left(r_{\xi_{k,i}}\left(s_{k,i} t_k\right)\right)$
    end for
    $x^{k+1} \leftarrow x^{k,m}$
  end for

Understanding the complexity of Algorithm 1 is our immediate goal, towards which the next two lemmas take us most of the way.

Lemma 4.2. (Projected Busemann subgradient inequality) Suppose $f \colon X \to \mathbb{R}$ is Busemann subdifferentiable on a nonempty, closed, and geodesically convex set C. Let $x \in C$, $t > 0$ and choose a Busemann subgradient $[\xi, s]$ for f at x. Define the new point

$$x^+ = \begin{cases} P_C\left(r_{x,\xi}\left(st\right)\right), & s > 0 \\ x, & s = 0, \end{cases}$$

where $r_{x,\xi}$ is the ray issuing from x with direction ξ. Then for any $y \in C$,

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

Proof. For brevity we denote $r = r_{x,\xi}$. If $s = 0$ then the desired inequality reduces to $f(x) \leq f(y)$ for all $y \in C$, which holds because x minimizes f over C. Thus we may assume $s > 0$. Since X is Hadamard, $x \mapsto \frac{1}{2} d(x, y)^2$ is 1-strongly convex for any $y \in C$. It follows that for all $\delta \geq st$ we have

$$\begin{aligned} d(x^+, y)^2 &= d\left(P_C(r\left(st\right)), P_C(y)\right)^2 \\ &\leq d\left(r\left(st\right), y\right)^2 \\ &\leq \left(1 - \frac{st}{\delta}\right) d(x, y)^2 + \frac{st}{\delta} d(r(\delta), y)^2 - \left(1 - \frac{st}{\delta}\right) \frac{st}{\delta} d(x, r(\delta))^2 \\ &= \left(1 - \frac{st}{\delta}\right) d(x, y)^2 + \frac{st}{\delta} d(r(\delta), y)^2 + s^2 t^2 - \delta st \\ &= \left(1 - \frac{st}{\delta}\right) d(x, y)^2 + \frac{st}{\delta} (d(r(\delta), y)^2 - \delta^2) + s^2 t^2. \end{aligned}$$

The second line uses nonexpansivity of the projection P_C. Letting $\delta \to \infty$ and using once more (24) we deduce

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

By (9) we have $sb_r(y) \leq -(f(x) - f(y))$. We conclude:

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

□

The following lemma and its proof are straightforwardly adapted from [5, Lemma 8.39], which is itself an adaptation of the original [23, Lemma 2.1].

Lemma 4.3. (Incremental Busemann subgradient inequality) Suppose Assumption A holds, and let $\{x^k\}_{k=0}^{\infty}$ be the sequence of iterates generated by Algorithm 1 with positive stepsizes $\{t_k\}_{k=0}^{\infty}$. Then for any $x^* \in X^*$ and $k \geq 0$,

$$d(x^{k+1}, x^*)^2 \leq d(x^k, x^*)^2 - 2t_k(f(x^k) - f_{\mathrm{opt}}) + t_k^2 m^2 L^2.$$

Proof. Fix $i \in \{0, 1, \ldots, m-1\}$. Lemma 4.2 proves

$$d(x^{k,i+1}, x^*)^2 \leq d(x^{k,i}, x^*)^2 - 2t_k(f_{i+1}(x^{k,i}) - f_{i+1}(x^*)) + s_{k,i}^2 t_k^2.$$

Summing the inequality over $i = 0, 1, \ldots, m-1$ and using the identities $x^{k,0} = x^k, x^{k,m} = x^{k+1}$ we deduce:

$$\begin{aligned} d(x^{k+1}, x^*)^2 &\leq d(x^k, x^*)^2 - 2t_k \sum_{i=0}^{m-1}(f_{i+1}(x^{k,i}) - f_{i+1}(x^*)) + t_k^2 \sum_{i=0}^{m-1} s_{k,i}^2 \\ &\leq d(x^k, x^*)^2 - 2t_k \sum_{i=0}^{m-1}(f_{i+1}(x^{k,i}) - f_{i+1}(x^*)) + t_k^2 m L^2 \\ &= d(x^k, x^*)^2 - 2t_k \left(f(x^k) - f_{\mathrm{opt}} + \sum_{i=0}^{m-1}(f_{i+1}(x^{k,i}) - f_{i+1}(x^k)) \right) + t_k^2 m L^2 \\ &\leq d(x^k, x^*)^2 - 2t_k(f(x^k) - f_{\mathrm{opt}}) + 2t_k L \sum_{i=0}^{m-1} d(x^{k,i}, x^k) + t_k^2 m L^2. \qquad (25) \end{aligned}$$

The second inequality uses Assumption A(v), and the last inequality makes use of L-Lipschitz continuity via Proposition 3.3. We aim to control the size of $d(x^{k,i}, x^k), i = 0, \ldots, m-1$ in the last line above, so we start by estimating

$$d(x^{k,1}, x^k) = d(P_C(r_{\xi_{k,0}}(s_{k,0}t_k)), P_C(x^k)) \leq d(r_{\xi_{k,0}}(s_{k,0}t_k), x^k) = s_{k,0}t_k \leq Lt_k. \qquad (26)$$

This required nonexpansivity of the projection P_C as well as Assumption A(v). Moving on to $x^{k,2}$, we argue similarly

$$d(x^{k,2}, x^k) \leq d(r_{\xi_{k,1}}(s_{k,1}t_k), x^k) \leq d(r_{\xi_{k,1}}(s_{k,1}t_k), x^{k,1}) + d(x^{k,1}, x^k) \leq 2Lt_k.$$

Here we used nonexpansivity of P_C, the triangle inequality, Assumption A(v), and (26). Iterating these estimates for $i = 2, \ldots, m-1$ we conclude

$$d(x^{k,i}, x^k) \leq iLt_k \quad i = 0, 1, \ldots, m-1.$$

Combined with (25) we find:

$$\begin{aligned} d(x^{k+1}, x^*)^2 &\leq d(x^k, x^*)^2 - 2t_k(f(x^k) - f_{\text{opt}}) + 2t_kL\sum_{i=0}^{m-1} d(x^{k,i}, x^k) + t_k^2 mL^2 \\ &\leq d(x^k, x^*)^2 - 2t_k(f(x^k) - f_{\text{opt}}) + 2t_k^2L^2\sum_{i=0}^{m-1} i + t_k^2 mL^2 \\ &= d(x^k, x^*)^2 - 2t_k(f(x^k) - f_{\text{opt}}) + t_k^2L^2m^2. \end{aligned}$$

□

The work above culminates in a standard complexity result under a boundedness assumption on the feasible set C. With Lemma 4.3 in hand, the proof below is a rewrite of [5, Theorem 8.40] with the metric d in place of the Euclidean distance.

Theorem 4.4. (Complexity of incremental Busemann subgradient method) Suppose Assumption A holds, and let $\{x^k\}_{k=0}^{\infty}$ be the sequence of iterates generated by Algorithm 1 with positive stepsizes $\{t_k\}_{k=0}^{\infty}$. Suppose furthermore that the diameter of C is bounded above by $D > 0$. If $t_k = \frac{D}{Lm\sqrt{k+1}}$ then for all $k \geq 2$

$$f_{\text{best}}^k := \min_{i=1,\ldots,k} f(x^i) - f_{\text{opt}} \leq \frac{2(1+\log(3))mLD}{\sqrt{k+2}}.$$

More generally, if $\sum_{k=0}^n t_k^2 / \sum_{k=0}^n t_k \to 0$ as $n \to \infty$ then $f_{\text{best}}^k \to f_{\text{opt}}$ as $k \to \infty$ even if C is unbounded.

Proof. By Lemma 4.3, for any $n \geq 0$

$$d(x^{n+1}, x^*)^2 \leq d(x^n, x^*)^2 - 2t_n(f(x^n) - f_{\text{opt}}) + t_n^2L^2m^2.$$

Summing over $n = \lceil k/2 \rceil, \ldots, k$ we find

$$d(x^{k+1}, x^*)^2 \leq d(x^{\lceil k/2 \rceil}, x^*)^2 - 2\sum_{n=\lceil k/2 \rceil}^{k} t_n(f(x^n) - f_{\text{opt}}) + L^2m^2 \sum_{n=\lceil k/2 \rceil}^{k} t_n^2.$$

Rearranging gives

$$2\sum_{n=\lceil k/2\rceil}^{k} t_n(f(x^n)-f_{\text{opt}}) \leq d(x^{\lceil k/2\rceil},x^*)^2 + L^2m^2\sum_{n=\lceil k/2\rceil}^{k} t_n^2.$$

We readily estimate

$$\min_{i=1,\dots,k} f(x^i)-f_{\text{opt}} \leq \frac{d(x^{\lceil k/2\rceil},x^*)^2 + L^2m^2\sum_{n=\lceil k/2\rceil}^{k} t_n^2}{2\sum_{n=\lceil k/2\rceil}^{k} t_n} \leq \frac{D^2 + L^2m^2\sum_{n=\lceil k/2\rceil}^{k} t_n^2}{2\sum_{n=\lceil k/2\rceil}^{k} t_n}.$$

Plugging in $t_n = D/(Lm\sqrt{n+1})$ we arrive at

$$\min_{i=1,\dots,k} f(x^i)-f_{\text{opt}} \leq \frac{mLD}{2}\frac{\left(1+\sum_{n=\lceil k/2\rceil}^{k}\frac{1}{n+1}\right)}{\sum_{n=\lceil k/2\rceil}^{k}\frac{1}{\sqrt{n+1}}}.$$

Applying [5, Lemma 8.27(b)] gives the bound

$$\frac{\left(1+\sum_{n=\lceil k/2\rceil}^{k}\frac{1}{n+1}\right)}{\sum_{n=\lceil k/2\rceil}^{k}\frac{1}{\sqrt{n+1}}} \leq \frac{4(1+\log(3))}{\sqrt{k+2}}$$

and the final estimate follows. The last statement of the theorem can be proven in exactly the same way as [5, Theorem 8.40(a)]. □

5 Computing medians

On any Hadamard space (X,d) with the geodesic extension property, we specialize to the *median problem*

$$\min\left\{ f(x) := \sum_{i=1}^{m} w_i d(x,a_i) \mid x \in X \right\}, \tag{27}$$

where $\mathcal{A} = \{a_1,\dots,a_m\} \subseteq X$ are given points and $w \in \mathbb{R}^m_+$ is a vector of nonnegative weights summing to one. It is well-known that problem (27) admits at least one minimizer. Each function $f_i := w_i d(x,a_i)$ is 1-Lipschitz so Assumption A(v) is satisfied by Proposition 3.3(ii). Each f_i is Busemann subdifferentiable by Example 3.11, with a Busemann subgradient $[r_{x,a}(\infty), w_i]$ at $x \neq a$ and $[0]$ at $x = a$. To attain the stronger theoretical complexity guarantee in Theorem 4.4 we require a bound on the diameter of the feasible region. The structure of the problem implies that the minimizers cannot be too far from points in $\mathcal{A}$. Without loss of generality we may assume $w_1 > 0$. Then for any minimizer $x^* \in X^*$ and any $x^0 \in X$, we have

$$w_1 d(x^*,a_1) \leq \sum_{i=1}^{m} w_i d(x^*,a_i) \leq \sum_{i=1}^{m} w_i d(x^0,a_i) = f(x^0).$$

It follows that $X^* \subseteq B_{f(x^0)/w_1}(a_1)$. This ball has diameter $D = 2f(x^0)/w_1$, and projecting onto a ball is straightforward. With Assumption A verified, we can use our work above and the stepsize from Theorem 4.4 to specialize Algorithm 1 to the median problem:

Algorithm 2 Incremental Median Algorithm

Require: $x^0 \in X$
  for $k = 0, 1, 2 \ldots$ **do**
    $x^{k,0} \leftarrow x^k$
    for $i = 0, 1, \ldots, m-1$ **do**
      $x^{k,i+1} \leftarrow P_{B_{\frac{f(x^0)}{w_1}}(a_1)}\left(r_{x^{k,i},a_i}\left(\frac{2w_i f(x^0)}{w_1 m\sqrt{k+1}}\right)\right)$
    end for
    $x^{k+1} \leftarrow x^{k,m}$
  end for

Corollary 5.1. (Median complexity) Algorithm 2 satisfies the following complexity bound for all $k \geq 2$:

$$f_{\text{best}}^k \leq \frac{4(1+\log(3))mf\left(x^0\right)}{w_1\sqrt{k+2}}.$$

Proof. Set $D = 2f(x^0)/w_1, L = 1$ in Theorem 4.4. □

In a similar way, one can find a suitable set C (a ball) satisfying Assumption A for the problem of computing p-means and the corresponding Algorithm 1 can be written explicitly in a fashion analagous to Algorithm 2. In closing this section, let us contrast Algorithm 2 with the cyclic proximal point method for computing medians described in [3]. We restate the algorithm below in Algorithm 3:

Algorithm 3 Cyclic Proximal Median Algorithm [3]

Require: $x^0 \in X, \{t_k\}_{k=0}^{\infty} \subseteq \mathbb{R}_{++}$
  for $k = 0, 1, 2 \ldots$ **do**
    $x^{k,0} \leftarrow x^k$
    for $i = 0, 1, \ldots, m-1$ **do**
      $x^{k,i+1} \leftarrow r_{x^{k,i},a_i}\left(\min\left\{d(x^{k,i}, a_i), w_i t_k\right\}\right)$
    end for
    $x^{k+1} \leftarrow x^{k,m}$
  end for

It is shown in [3] that if $\{t_k\}_{k=0}^{\infty}$ is a sequence of positive stepsizes satisfying

$$\sum_{k=0}^{\infty} t_k = \infty, \quad \sum_{k=0}^{\infty} t_k^2 < \infty \tag{28}$$

then the iterates $\{x^k\}_{k=0}^{\infty}$ converge to a median of $\mathcal{A}$. No explicit complexity analysis is given, but it is not hard to see by comparing [3, Inequality (3.7)] with our Lemma 4.3 that the same line of reasoning can be carried out as in Theorem 4.4 to obtain an $O(\varepsilon^{-2})$ complexity bound for p-mean problems. The emphasis on stepsizes satisfying (28) seems to be oriented towards the analysis of a stochastic variant of the incremental proximal algorithm discussed in [3].

Algorithms 1 and 3 are quite similar when C is chosen to be the whole space X, making the projection step in the former redundant. The main difference between Algorithms 1 and 3 is that the latter requires only geodesics instead of rays; this manifests as the thresholding $\min\{d(x^{k,i}, a_i), w_i t_k\}$ in the argument of $r_{x^{k,i},a_i}$. In particular, Algorithm 3 works in Hadamard spaces without the geodesic extension property. In situations where the iterates $\{x^k\}_{k=0}^{\infty}$ remain bounded away from the set $\mathcal{A}$, both algorithms should coincide after sufficiently many iterations because the stepsizes decay to zero. In such cases, the new iterates $x^{k,i+1}$ always lie in the geodesic segment $[x^{k,i}, a_i]$ and the extension to a geodesic ray is redundant; both algorithms are the same. We observe this phenomenon empirically in our computational experiments. This can fail, however, if one of the points in $\mathcal{A}$ happens to be a median.

Example 5.1. Let $X = C = \mathbb{R}$ with $\mathcal{A} = \{0\}$. The unique median of $\mathcal{A}$ is obviously $x^* = 0$, and Algorithm 1 amounts to the classical subgradient method applied to $f(x) = |x|$. Taking $x^0 = 1$ and $t_k = 1/(k+2) + 1/(k+1)$, the sequence of iterates generated by Algorithm 1 is $x^k = (-1)^k/(k+1)$. In particular, x^k alternates between positive and negative values as it overshoots the solution x^* at each iteration, meaning it always requires a ray oracle. Algorithm 3, on the other hand, would generate iterates remaining in the interval $[0, 1]$ and monotonically decreasing to $x^* = 0$ as it only uses a geodesic oracle.

In Example 6.1 we illustrate Algorithm 2 computationally. Under the regime $C = X$ and $t_k = 1/k$, our numerical experiments indicated that this algorithm coincides with Algorithm 3 so we only show the data for the former to avoid obfuscating the plot in Figure 6.2. Based on our discussion above, we expect both algorithms to perform similarly in general. In spite of the stronger theoretical complexity guaranteed by the choice of stepsize in Theorem 4.4, the stepsize $t_k = 1/k$ seems to achieve better performance based on the plots in Figures 6.2 and 6.4.

6 Computational experiments

We consider the BHV tree space $\mathcal{T}_n$ of binary trees on n labelled leaves, introduced in the seminal work [8]. There are $(2n-3)!! = (2n-3)(2n-5)\cdots 3 \cdot 1$ such binary trees, each with $n-2$ internal edges. The space $\mathcal{T}_n$ models all such binary trees by ascribing an $(n-2)$-dimensional orthant $[0, \infty)^{n-2}$ to each tree so that a point in each orthant describes a particular binary tree topology with a prescribed choice of nonnegative internal edge lengths. Since its introduction at the turn of the century, tree space has generated much interest at the intersection of mathematics and computational biology, in particular as a model for comparison and averaging of phylogenetic trees. The space $\mathcal{T}_n$ was shown in [8] to be CAT(0) and is easily seen to be complete, rendering it a Hadamard space. Note that $\mathcal{T}_n$ has the geodesic extension property as a consequence of [11, Proposition II.5.10] since every orthant of dimension at most $n-3$ appears as a boundary face of at least three $(n-2)$-dimensional orthants [8, p. 743] so there are no free faces. The referenced result assumes that the set of isometry classes of the faces of cells in the complex is finite; this is trivially satisfied since $\mathcal{T}_n$ is a finite union of finite-dimensional orthants.

For ease of exposition and visualization we focus on $\mathcal{T}_4$, where the binary trees have 4 labelled leaves. There is no difficulty in extending the computational examples below to $\mathcal{T}_n$, the only input to the code we use is a list of trees in *Newick notation* [2]. The space $\mathcal{T}_4$ is a union of 15 two-dimensional quadrants. In what follows we will demonstrate how Algorithm 2 can be used to estimate the median of a finite set of trees in $\mathcal{T}_n$. To facilitate the examples, we employ the existing software package `SturmMean` [24] which implements a polynomial-time algorithm [25] to compute geodesics and geodesic distances in tree space. Algorithm 2 actually relies on rays which could extend beyond the span of a geodesic between two trees and the choice of extension is not typically unique: consider $\mathcal{T}_3$ which coincides with the tripod of Example 2.1, and imagine extending a geodesic that terminates at the origin. The code from `SturmMean` for computing geodesics does not implement such extensions, but in the examples we consider below all stepsizes are sufficiently small starting from the initial point that the iterates lie in the unextended geodesic segment so a ray oracle is not needed. We also need the value $f(x^0)$ to compute the projection step, so it is often convenient to initialize the algorithm at x^0 being the origin of tree space because it makes $f(x^0)$ particularly easy to compute: the origin is in every orthant so the distance from the origin to each a_i is just the usual Euclidean distance, i.e. the norm of the vector of edge lengths.

Example 6.1. Our first example comes from the documentation of `SturmMean` [24]. This example is convenient because the trees embed simply in $\mathbb{R}^2$ in such a way that we can calculate the true median exactly, allowing us to demonstrate convergence of the best found function value to the optimal value. Figure 6.1 shows the three trees we consider, with equal weights $w_i = 1/3$ assigned to each tree. For x in the top right quadrant we have an explicit representation for f:

$$f(x) = \frac{1}{3}\left(\|x-(1,2)\| + \|x-(2,-3/2)\| + \|x\| + \frac{\sqrt{37}}{2}\right).$$

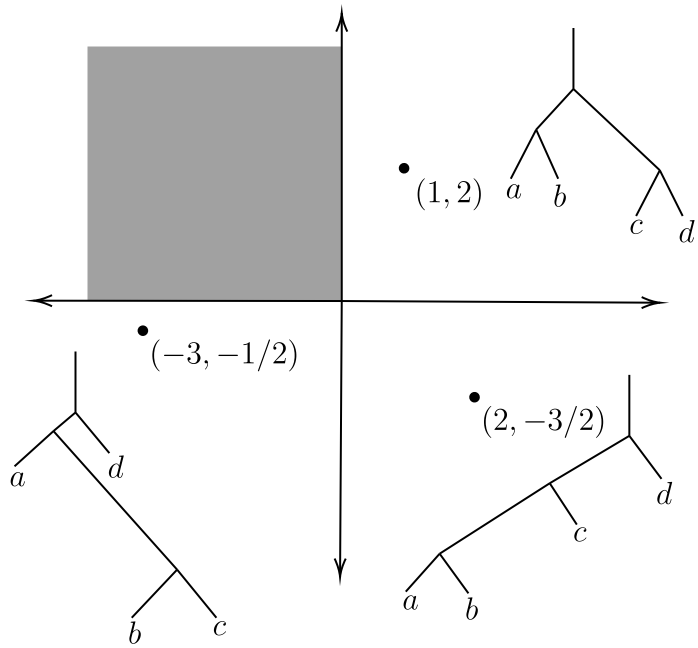

Figure 6.1: Three trees in $\mathcal{T}_4$ with neighboring respective orthants, embedded isometrically in $\mathbb{R}^2$ (recreated from [24]).

From here it is easy to check that

$$x^* = \left(\frac{2657 - 1038\sqrt{3}}{1898}, \frac{3006 - 1369\sqrt{3}}{5694} \right)$$

satisfies the first-order optimality condition $\nabla f(x^*) = 0$. Hence x^* is a local minimizer of f, and the convexity of f on $\mathcal{T}_4$ implies x^* is a global minimizer. Thus we set

$$f_{\text{opt}} = f(x^*) = \frac{1}{3}\sqrt{\frac{1}{2}\left(43 + 11\sqrt{3} + \sqrt{37(49 + 22\sqrt{3})}\right)}.$$

Initialized at the origin, the result of running Algorithm 2 on this problem for 10^5 iterations is shown in Figure 6.2. We also tested the algorithm initialized at a point x^0 far from the median, in the sense that the tree topology of x^0 does not correspond to one of the quadrants containing the given trees and the branch lengths are larger.

Example 6.2. In this example we consider three trees in $\mathcal{T}_4$ whose median lies on the common boundary ray of three quadrants, with each quadrant containing one of the trees. We refer to this common boundary ray as the spine. A figure illustrating the local geometry of this setup can be found in [8, Figure 10]. The

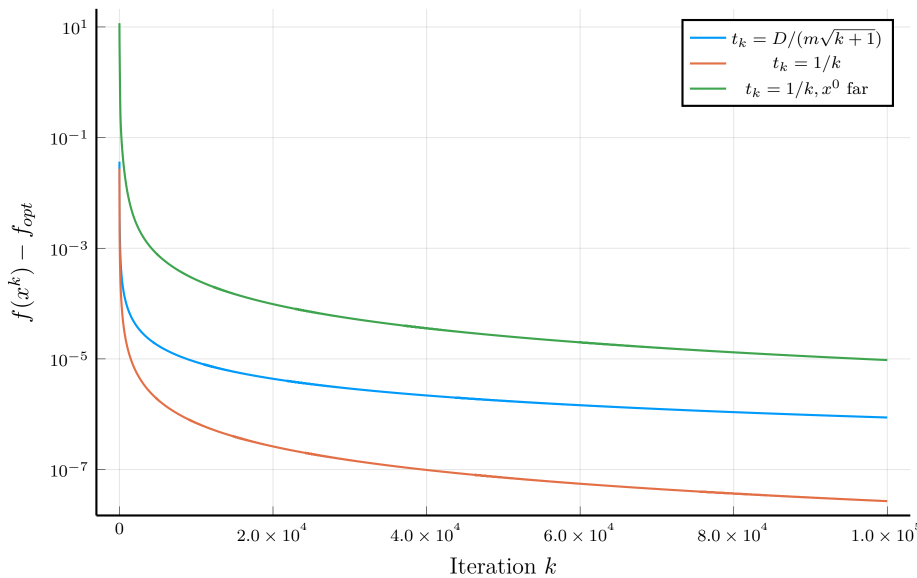

Figure 6.2: Convergence of $f(x^k)$ to f_{opt} in Example 6.1 using two different stepsizes and choices of initial tree.

three neighboring quadrants arise by permuting the labels on the leaves of the left subtree of the leftmost tree in Figure 6.3. The spine corresponds to the second tree in this same figure, obtained by contracting the lower internal edge so that the left subtree leaves become siblings. The label L on the separated leaf remains fixed throughout.

The occurrence of the median on a negligible subset is actually representative of a well-known and studied phenomenon of *stickiness* in certain cubical complexes, whereby the mean of a randomly generated set of points will lie on a lower-dimensional face with positive probability [19]. We observe similarly with the median: letting $p_i, i = 1, 2, 3$ denote the point $(1, 1)$ in some ordering of the three neighboring quadrants, the median clearly lies on the spine, and one can show (e.g. by computing Euclidean directional derivatives) that for $\varepsilon = 1/(1 + 2\sqrt{3})$ the median remains stuck on the spine even after perturbing each of the p_i within the box $p_i + [-\varepsilon, \varepsilon]^2$ in their respective quadrants. In this way we arrive at the following three points whose median lies on the spine:

$$a = \left(\frac{2\sqrt{3}}{1+2\sqrt{3}}, \frac{2\sqrt{3}}{1+2\sqrt{3}} \right), \quad b = (1, 1), \quad c = \left(\frac{2+2\sqrt{3}}{1+2\sqrt{3}}, \frac{2+2\sqrt{3}}{1+2\sqrt{3}} \right).$$

Note that a, b, c correspond directly to trees with the given internal branch lengths and the three tree topologies implicit in Figure 6.3. The height of the median on

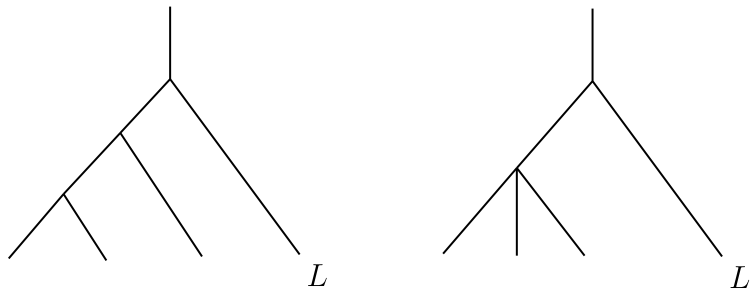

Figure 6.3: The tree topology defining three neighboring quadrants and their common spine.

the spine can be calculated as the optimal solution to the following problem:

$$\min_y \sqrt{(y-a_2)^2 + a_1^2} + \sqrt{(y-b_2)^2 + b_1^2} + \sqrt{(y-c_2)^2 + c_1^2}.$$

Numerically, we find the optimal value to be

$$y^* \approx 0.966816497678259.$$

This gives an estimate for the optimal value:

$$f_{\text{opt}} \approx 1.016799490957051.$$

Initialized at the origin, the result of running Algorithm 2 on this problem for 10^5 iterations is shown in Figure 6.4. We also tested the algorithm initialized at a point x^0 far from the median, in the same sense as Example 6.1.

7 Appendix

7.1 Example 3.3

We will show that if $C \subseteq \mathbb{R}^n$ is convex and $f\colon C \to \mathbb{R}$ is convex and locally Lipschitz on C then f has a (Euclidean) subgradient at each point in C. For any $x \in C$, if f is locally Lipschitz one can find $L, \delta > 0$ such that $|f(y) - f(z)| \leq L\|y - z\|$ for all $y \in C \cap B_\delta(x) =: C'$. Define $g\colon \mathbb{R}^n \to (-\infty, +\infty]$ as follows:

$$g(y) = \begin{cases} f(y), & y \in C' \\ +\infty, & \text{else.} \end{cases}$$

Then set $h(y) = \inf_{z \in \mathbb{R}^n} \{g(z) + L\|y - z\|\}$. We claim that h is finite-valued, convex, and agrees with f on C'. Clearly $h(y) < +\infty$ for any y because C' is nonempty,

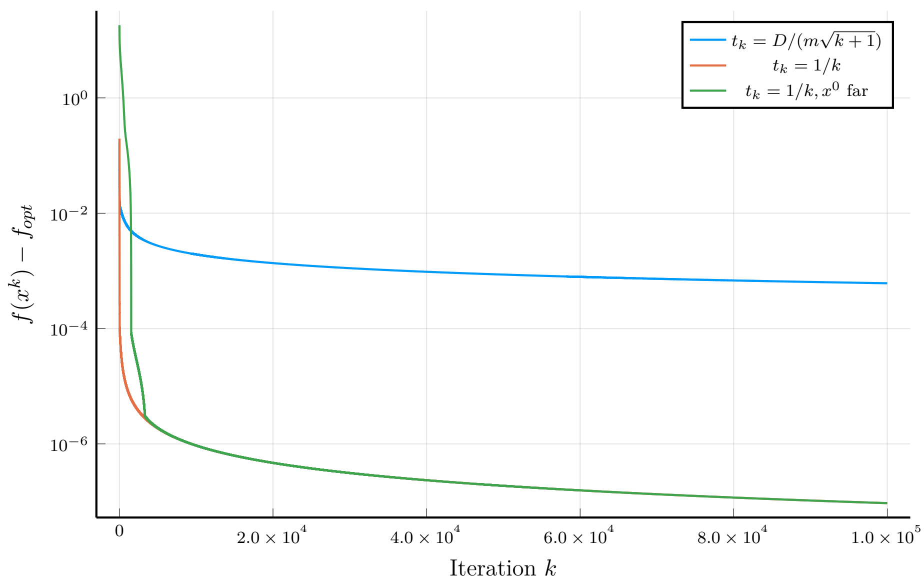

Figure 6.4: Convergence of $f(x^k)$ to f_{opt} in Example 6.2 using two different stepsizes and choices of initial tree.

while $h(y) > -\infty$ because C' is bounded and

$$h(y) = \inf_{z \in C'} \{f(z) + L\|y - z\|\} \geq f(x) + L \inf_{z \in C'} \{\|y - z\| - \|z - x\|\} > -\infty.$$

Thus h is finite-valued, and convexity of h follows by recognizing h as the infimal convolution of the convex function g with the finite-valued convex function $L\|\cdot\|$. Given $y \in C'$ we observe $h(y) \leq f(y)$ because y is feasible for the infimum defining h, while on the other hand the L-Lipschitz property for f on C' implies

$$f(z) + L\|y - z\| \geq f(y) \text{ for all } z \in C'.$$

Taking the infimum over $z \in C'$ implies $h(y) \geq f(y)$, hence $h \equiv f$ on C'.

Now, h is a real-valued convex function on $\mathbb{R}^n$ and thus admits a subgradient v at x. We claim that v is a subgradient of f. The subgradient inequality for h reads

$$h(y) \geq h(x) + v^T(y - x) \text{ for all } y \in \mathbb{R}^n.$$

Restricting to points in C' where $h \equiv f$ this shows $\tilde{f} := f - v^T(\cdot - x)$ is minimized over C' at x. By definition of C' it follows that x is a local minimizer of $\tilde{f}$ on C, from which convexity of $\tilde{f}$ implies x minimizes $\tilde{f}$ on C. This says exactly that v is a subgradient of f at x.

7.2 Example 3.13

Let $\bar{x} = (0,0,0)$ be the reference point with respect to which we define our Busemann functions. Suppose for a contradiction that f has a Busemann subgradient $[\xi, s]$ at $x_0 = (1,1,0)$. As x_0 is not a minimizer of f, we have $s > 0$. Moreover,

$$f(y) - sb_\xi(y) \geq \frac{\sqrt{2}}{2} - sb_\xi(x_0) \text{ for all } y \in X. \tag{29}$$

Case 1: Suppose ξ is the direction of the ray γ issuing from $\bar{x}$ and passing through x_0. Taking $y = (1/2, 1/2, 0)$ we have $f(y) = 0, b_\xi(y) = -\sqrt{2}/2$, and $b_\xi(x_0) = -\sqrt{2}$, in contradiction to (29).

Case 2: Suppose ξ is the direction of a ray r in the plane $\{z = 0\}$ issuing from $\bar{x}$ and obtained by moving γ an angle $\alpha \in (0, 3\pi/4]$ in the trigonometric sense. Taking $y = (3/2, 1/2, 0)$, we have $f(y) = \sqrt{2}/2$ and $d(\bar{x}, y) = \sqrt{5}/\sqrt{2}$. Moreover, for all $t > 0$,

$$d(x_0, r(t))^2 = 2 + t^2 - 2\sqrt{2}t\cos\alpha$$

and

$$d(y, r(t))^2 = \frac{5}{2} + t^2 - 2t(\sqrt{2}\cos\alpha - (1/\sqrt{2})\sin\alpha).$$

Hence, $b_\xi(x_0) = -\sqrt{2}\cos\alpha$ and $b_\xi(y) = -\sqrt{2}\cos\alpha + (1/\sqrt{2})\sin\alpha$, in contradiction to (29).

Case 3: Suppose ξ is the direction of a ray r in the quadrant $\{x \leq 0, y = 0, z \geq 0\}$. Taking $y = (3/2, 1/2, 0)$, we have $f(y) = \sqrt{2}/2$ and $b_\xi(x_0) \leq \sqrt{2}$.

Fix $t > 0$ and let $v = (-s, 0, 0)$, where $s \geq 0$, be such that $d(y, r(t)) = d(y, v) + d(v, r(t))$. Then $s \leq t$ and

$$d(y, r(t)) = d(v, r(t)) + \sqrt{\frac{5}{2} + s^2 + 3s} \geq t - s + \sqrt{\frac{5}{2} + s^2 + 3s} \geq \sqrt{\frac{5}{2} + t^2 + 3t}.$$

We conclude that $b_\xi(y) \geq 3/2$, so $b_\xi(y) > b_\xi(x_0)$, a contradiction.

The rest of the possible cases follow by symmetry.

7.3 Example 4.1

Defining $\tilde{x} = (2/5, 0, 0), \bar{x} = (1/4, 1/4, 0)$, we do some preliminary calculations:

$$\begin{aligned} f(\tilde{x}) &= \frac{1}{2}\left(d(\tilde{x}, a_1)^2 + d(\tilde{x}, a_2)^2\right) = 37/25, \\ f(\bar{x}) &= \frac{1}{2}\left(d(\bar{x}, a_1)^2 + d(\bar{x}, a_2)^2\right) = 13/8. \end{aligned} \tag{30}$$

Any ray issuing from $\bar{x}$ is determined by a choice of unit vector $v = (v_1, v_2, 0)$, or equivalently an angle $\theta \in [0, 2\pi]$ such that $v = (\cos\theta, \sin\theta, 0)$. By symmetry it suffices to consider $\pi/4 \leq \theta \leq 5\pi/4$. We denote a ray r originating from $\bar{x}$ in direction v using the notation r_v.

Case 1: Corresponding to $\theta = 5\pi/4$ consider the direction $v = (-1/\sqrt{2}, -1/\sqrt{2}, 0)$, pointing from $\bar{x}$ to the origin $\mathbf{0} := (0,0,0)$. Since the ray ultimately extends up the spine $\{0\} \times \{0\} \times \mathbb{R}_+$, for $t > 0$ sufficiently large the geodesic between $r_v(t)$ and any $x \in \mathbb{R}_+ \times \mathbb{R}_+ \times \{0\}$ consists of the two segments $[x, \mathbf{0}] \cup [\mathbf{0}, r_v(t)]$. Hence

$$d(x, r_v(t)) = \|x\| + \|(t - \|\bar{x}\|)(0,0,1)\| = \|x\| + t - \|\bar{x}\| \text{ for all } t \geq \|\bar{x}\|.$$

The corresponding Busemann function at such an x is thus

$$b_{r_v}(x) = \|x\| - \|\bar{x}\|.$$

Plugging in $\tilde{x}$ we find $b_{r_v}(\tilde{x}) = 2/5 - \sqrt{2}/4 = (8 - 5\sqrt{2})/20 > 0$. On the other hand, the preliminary calculations (30) show $f(\tilde{x}) < f(\bar{x})$, so $\tilde{x} \in f_{\bar{x}} \setminus \{b_{r_v} \leq 0\}$ i.e. this horoball does not contain the given level set.

Case 2: Now consider any direction v corresponding to $\pi < \theta < 5\pi/4$. After sufficient time, the point $r_v(t)$ will inhabit the upright quadrant $\mathbb{R}_- \times \{0\} \times \mathbb{R}_+$. By imagining the quadrant $\mathbb{R}_- \times \{0\} \times \mathbb{R}_+$ being folded down into $\mathbb{R}_- \times \mathbb{R}_- \times \{0\}$, we see that for $t > 0$ large enough the distance $d(\mathbf{0}, r_v(t))$ is equal to the Euclidean distance $\|\bar{x} + tv\|$ when we identify $\bar{x}, v$ with vectors in $\mathbb{R}^3$. Since the geodesic from $\tilde{x}$ to such an $r_v(t)$ consists again of the two segments $[\tilde{x}, \mathbf{0}] \cup [\mathbf{0}, r_v(t)]$, it follows that

$$d(\tilde{x}, r_v(t)) = \|\tilde{x}\| + \|\bar{x} + tv\|.$$

The corresponding Busemann function evaluated at $\tilde{x}$ is thus

$$b_{r_v}(\tilde{x}) = \|\tilde{x}\| + v^T\bar{x} \geq \|\tilde{x}\| - \|\bar{x}\| > 0.$$

As in the previous case we conclude $\tilde{x} \in f_{\bar{x}} \setminus \{b_{r_v} \leq 0\}$.

Case 3: Finally, consider directions v corresponding to $\pi/4 \leq \theta \leq \pi$. In this case we have $d(\tilde{x}, r_v(t)) = \|\tilde{x} - \bar{x} - tv\|$, from which we deduce $b_{r_v}(\tilde{x}) = v^T(\bar{x} - \tilde{x})$. Using $v = (\cos\theta, \sin\theta, 0)$ we find:

$$b_{r_v}(\tilde{x}) = -\frac{3}{20}\cos\theta + \frac{1}{4}\sin\theta.$$

Using calculus, one can show that the righthand side is a concave function of θ on $[\pi/4, \pi]$ and so attains its minimum at an endpoint. It is easy to check that the values at these endpoints are both positive, so $b_{r_v}(\tilde{x}) > 0$ for all such θ. As before, $\tilde{x} \in f_{\bar{x}} \setminus \{b_{r_v} \leq 0\}$. We conclude that no ray issuing from $\bar{x} \in \operatorname{bdry} f_{\bar{x}}$ supports $f_{\bar{x}}$, so $f_{\bar{x}}$ is not horospherically convex.

Define the geodesic $\eta\colon [0,1] \to X, \eta(t) = \bar{x} + (t/\sqrt{2})(1,1,0)$. We will show that $([\eta], 3/\sqrt{2}) \in T_{\bar{x}}X$ is a subgradient for f at $\bar{x}$. According to (21) and the value computed for $f(\bar{x})$ in (30), it suffices to prove

$$\frac{3}{\sqrt{2}} d(\bar{x}, y) \cos\angle(\eta, \gamma_y) + \frac{13}{8} \leq f(y) \text{ for all } y \in X. \tag{31}$$

Despite the initial requirement that this inequality holds for all $y \in X$, it is shown in [21, Remark 3.3(iv)] that the subdifferential of a geodesically convex function depends only on the function locally. Thus it suffices to verify (31) on a small ball around $\bar{x}$ contained in the quadrant $Q = \mathbb{R}_+ \times \mathbb{R}_+ \times \{0\}$, say $B_{1/5}(\bar{x})$. For $y = (y_1, y_2, 0) \in Q$ we have

$$f(y) = \frac{1}{2}\left((y_1+1)^2 + y_2^2 + y_1^2 + (y_2+1)^2\right). \tag{32}$$

Furthermore, for $y \in B_{1/5}(\bar{x})$ the geodesic segment $[\bar{x}, y]$ is the Euclidean line segment joining these points in Q. It follows that $d(\bar{x}, y) = \|y - \bar{x}\|$ and

$$\cos\angle(\eta, \gamma_y) = \frac{(1,1,0)^T(y-\bar{x})}{\sqrt{2}\|y-\bar{x}\|} = \frac{y_1 + y_2 - 1/2}{\sqrt{2}\|y-\bar{x}\|}. \tag{33}$$

Thus after restricting to $y \in B_{1/5}(\bar{x})$ and substituting (32) and (33) into (31), our desired inequality becomes

$$\frac{3}{2}y_1 + \frac{3}{2}y_2 + \frac{7}{8} \leq \frac{1}{2}\left((y_1+1)^2 + y_2^2 + y_1^2 + (y_2+1)^2\right) \text{ for all } y = (y_1, y_2, 0) \in B_{1/5}(\bar{x}).$$

This now holds by the Euclidean subgradient inequality for the differentiable convex function $h(a,b) = \frac{1}{2}\left((a+1)^2 + b^2 + a^2 + (b+1)^2\right)$ at the point $(1/4, 1/4)$.

References

[1] Q. H. Ansari and M. Uddin. Convergence analysis of incremental quasi-subgradient method on Riemannian manifolds with lower bounded curvature. *Optimization*, 2024.

[2] J. Archie, W.H.E. Day, J. Felsenstein, W. Maddison, C. Meacham, F.J. Rohlf, and D. Swofford. The Newick tree format. *More information: http://evolution.genetics.washington.edu/phylip/newicktree.html*, 1986.

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

[4] M. Bačák. *Convex Analysis and Optimization in Hadamard Spaces*. De Gruyter, Berlin, München, Boston, 2014.

[5] A. Beck. *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2017.

[6] G.d.C. Bento, J.C. Neto, and Í.D.L. Melo. Fenchel conjugate via Busemann function on Hadamard manifolds. *Applied Mathematics and Optimization*, 88(83), 2023.

[7] J. Bertrand and B. Kloeckner. A geometric study of Wasserstein spaces: Hadamard spaces. *Journal of Topology and Analysis*, 04(04):515–542, 2012.

[8] L. J. Billera, S. P. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767, 2001.

[9] J.M. Borwein. Direct theorems in semi-infinite convex programming. *Mathematical Programming*, 21:301–318, 1981.

[10] N. Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

[11] M. R. Bridson and A. Haefliger. *Metric Spaces of Non-Positive Curvature*. Grundlehren der mathematischen Wissenschaften. Springer Berlin, Heidelberg, 1st edition, 1999.

[12] C. Criscitiello and N. Boumal. Curvature and complexity: better lower bounds for geodesically convex optimization. *Proceedings of Machine Learning Research*, 195:1–45, 2023.

[13] B. Duchesne. Groups acting on spaces of non-positive curvature, 2016. arXiv:1603.04573.

[14] X. Fan, C.-H. Yang, and B. C. Vemuri. Horospherical decision boundaries for large margin classification in hyperbolic space. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[15] O.P. Ferreira and P.R. Oliveira. Subgradient algorithm on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 97:93–104, 1998.

[16] K. Hayashi. A polynomial time algorithm to compute geodesics in cat(0) cubical complexes. *Discrete Comput. Geom.*, 65(3):636–654, April 2021.

[17] H. Hirai. Convex analysis on Hadamard spaces and scaling problems. *Foundations of Computational Mathematics*, 24:1979–2016, 2024.

[18] Y. Hu, C. K. W. Yu, and X. Yang. Incremental quasi-subgradient methods for minimizing the sum of quasi-convex functions. *Journal of Global Optimization*, 75(4):1003–1028, 2019.

[19] S. Huckemann, J.C. Mattingly, E. Miller, and J. Nolen. Sticky central limit theorems at isolated hyperbolic planar singularities. *Electronic Journal of Probability*, 20:1–34, 2015.

[20] B. A. Kakavandi and M. Amini. Duality and subdifferential for convex functions on complete CAT(0) metric spaces. *Nonlinear Analysis: Theory, Methods & Applications*, 73(10):3450–3455, 2010.

[21] A. S. Lewis, G. López-Acedo, and A. Nicolae. Basic convex analysis in metric spaces with bounded curvature. *SIAM Journal on Optimization*, 34(1):366–388, 2024.

[22] A. S. Lewis, G. López-Acedo, and A. Nicolae. Horoballs and the subgradient method, 2024. arXiv.2403.15749.

[23] A. Nedic and D.P. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001.

[24] M. Owen, E. Miller, and J.S. Provan. Polyhedral computational geometry for averaging metric phylogenetic trees. *Advances in Applied Mathematics*, 68:51–91, 2015.

[25] M. Owen and S. Provan. A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Transactions on Computational Geometric and Bioinformatics*, 8:2–13, 2011.

[26] R.T. Rockafellar and R.J.B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften. Springer, 1998.

[27] Q. Shi, C. He, and L. Jiang. Normalized incremental subgradient algorithm and its application. *IEEE Transactions on Signal Processing*, 57(10):3759–3774, 2009.

[28] A. Weber. *The Theory of the Location of Industries*. Chicago University Press, 1929.

[29] H. Zhang and S. Sra. First-order methods for geodesically convex optimization. *Journal of Machine Learning Research*, 49:1–22, 2016.