Hyperbolic Busemann Neural Networks

Supplementary Material

Appendix Contents

Table 9. Summary of notation.

Notation	Description
$\mathcal{M}$	Riemannian manifold
$T_x\mathcal{M}$	Tangent space at x
$\mathrm{Exp}_x(v)$	Exponential map at x
$\mathrm{Log}_x(y)$	Logarithmic map at x
$\mathrm{PT}_{x\to y}(v)$	Parallel transport of v from x to y
$g_x(u,v)$	Riemannian metric at x
$\mathrm{d}(x,y)$	Geodesic distance
$\gamma(t)$	Unit speed geodesic ray
$B^\gamma(x)$	Busemann function associated with the geodesic ray γ
$(\mathcal{X},\mathrm{d})$	Metric space and its distance function
$\partial\mathcal{X}$	Boundary at infinity of $\mathcal{X}$
$[x,y]$	Geodesic segment joining x and y
$\Delta(x,y,z)$	Geodesic triangle in $\mathcal{X}$
(M_K^n,d_K)	Model space of constant curvature K with distance d_K
$\mathrm{CAT}(K)$	$\mathrm{CAT}(K)$ space
$HB_\tau^\gamma,\ H_\tau^\gamma$	Horoball and horosphere of γ at level τ
$\mathbb{R}^n$	Euclidean space of dimension n
$\langle\cdot,\cdot\rangle,\ \|\cdot\|$	Euclidean inner product and norm
$\mathbf{0}$	Zero vector in $\mathbb{R}^n$
$\mathcal{H}_K^n$	Hyperbolic space, either $\mathbb{P}_K^n$ or $\mathbb{L}_K^n$
$K<0$	Constant sectional curvature
$\oplus_\mathcal{H}$ and $\odot_\mathcal{H}$	Gyroaddition and scalar gyromultiplication on $\mathcal{H}_K^n$
e	Origin in $\mathcal{H}_K^n$
$\mathbb{P}_K^n$	Poincaré ball model in $\mathbb{R}^n$ with curvature K
λ_x^K	Conformal factor $\frac{2}{1+K\|x\|^2}$
$\oplus_\mathrm{M},\ \odot_\mathrm{M}$	Möbius gyroaddition and scalar gyromultiplication on $\mathbb{P}_K^n$
$\mathbb{L}_K^n$	Lorentz model in $\mathbb{R}^{n+1}$ with curvature K
$\langle x,y\rangle_\mathcal{L},\ \|x\|_\mathcal{L}$	Lorentzian inner product and norm
$x=\left[x_t,x_s^\top\right]^\top$	Temporal and spatial decomposition in the Lorentz ambient space
$\overline{\mathbf{0}}$	Origin of $\mathbb{L}_K^n$
$\oplus_\mathbb{L},\ \odot_\mathbb{L}$	Gyroaddition and scalar gyromultiplication on $\mathbb{L}_K^n$
$\mathbb{S}^{n-1}$	Unit sphere in $\mathbb{R}^n$
$B^v(x)$	Busemann function associated with a unit direction $v\in\mathbb{S}^{n-1}$
$H_{v,\alpha,b}$	Horosphere $\{x\in\mathcal{H}_K^n \mid -\alpha B^v(x)+b=0\}$
$\bar{\mathrm{d}}(y,H)$	Signed point-to-hyperplane distance
$\mathrm{softmax}$	Softmax operator
$u_k(x)$	Logit for class k
$e_k\in\mathbb{R}^m$	k-th standard basis vector
ϕ	Activation function $\mathbb{R}\to\mathbb{R}$

A. Notation

Tab. 9 summarizes the notation used throughout the paper.

B. Preliminaries

B.1. Riemannian geometry

We briefly review Riemannian geometry. For in-depth discussions, please refer to [19].

Tangent space. Given a smooth manifold $\mathcal{M}$, the *tangent space* $T_x\mathcal{M}$ at $x\in\mathcal{M}$ is a Euclidean space, consisting of

velocities of smooth curves through x, namely $v \in T_x\mathcal{M}$ if there exists a smooth γ with $\gamma(0) = x$ and $\dot{\gamma}(0) = v$.

Riemannian manifold. A *Riemannian manifold* is a smooth manifold $\mathcal{M}$ endowed with a *Riemannian metric* g, that is, for each $x \in \mathcal{M}$ an inner product $g_x(\cdot, \cdot)$ or $\langle \cdot, \cdot \rangle_x$ on the tangent space $T_x\mathcal{M}$ varying smoothly with x. The metric induces the length of a smooth curve $\gamma : [0, 1] \to \mathcal{M}$ as $L(\gamma) = \int_0^1 \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))}dt$, and the *geodesic distance* $\mathrm{d}(x, y)$ is the infimum of lengths over all smooth curves joining x and y.

Geodesic. Straight lines generalize to constant-speed curves that are locally length minimizing between points $x, y \in \mathcal{M}$, known as *geodesics*:

$$\gamma^* = \arg\min_{\gamma} L(\gamma) \quad \text{subject to } \gamma(0) = x, \gamma(1) = y, \|\dot{\gamma}(t)\|_{\gamma(t)} = c > 0. \tag{26}$$

We focus on unit-speed geodesics, *i.e.*, $c = 1$. In hyperbolic space, between any two points there exists a unique geodesic segment that is globally length minimizing.

Exponential and Logarithmic Maps. For $x \in \mathcal{M}$ and $v \in T_x\mathcal{M}$, let $\gamma_{x,v}$ denote the unique geodesic with $\gamma_{x,v}(0) = x$ and $\dot{\gamma}_{x,v}(0) = v$. The exponential map $\mathrm{Exp}_x : T_x\mathcal{M} \supset \mathcal{V} \to \mathcal{M}$ is defined by $\mathrm{Exp}_x(v) = \gamma_{x,v}(1)$, where $\mathcal{V}$ is an open neighborhood of the origin in $T_x\mathcal{M}$. Its local inverse, defined for y in a neighborhood $\mathcal{U} \subset \mathcal{M}$ of x, is the logarithmic map $\mathrm{Log}_x : \mathcal{U} \to T_x\mathcal{M}$, satisfying $\mathrm{Exp}_x \circ \mathrm{Log}_x = \mathbb{1}_{\mathcal{U}}$. In hyperbolic space, these maps are globally well-defined, that is, Exp_x is defined on all of $T_x\mathcal{M}$ and $\mathrm{Log}_x(y)$ exists for every $y \in \mathcal{M}$.

Parallel transport. *Parallel transport* moves tangent vectors along a curve while preserving the norm. Given a geodesic γ from x to y, the parallel transport of a tangent vector $v \in T_x\mathcal{M}$ is the unique vector $\mathrm{PT}_{x\to y}(v) \in T_y\mathcal{M}$ obtained by transporting v along γ so that its covariant derivative along γ vanishes. Parallel transport defines a linear isometry between $T_x\mathcal{M}$ and $T_y\mathcal{M}$.

B.2. Metric geometry

We present a concise overview of metric geometry, which generalizes Riemannian concepts to metric spaces. The theory develops geodesics and curvature without smooth structure, providing the tools used in our analysis of point-to-horosphere distances. We follow [6, Ch. I.1, I.2, II.1, II.2 and II.8] for definitions and results.

B.2.1. Geodesic metric spaces

We begin by recalling some basic notions in metric spaces.

Definition B.1 (Metric space)**.** A *metric space* is a pair $(\mathcal{X}, \mathrm{d})$ where $\mathcal{X}$ is a set and the distance function $\mathrm{d} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfies, for all $x, y, z \in \mathcal{X}$,

$$\text{Positivity: } \mathrm{d}(x, y) \geq 0, \quad \mathrm{d}(x, y) = 0 \iff x = y, \tag{27}$$

$$\text{Symmetry: } \mathrm{d}(x, y) = \mathrm{d}(y, x), \tag{28}$$

$$\text{Triangle inequality: } \mathrm{d}(x, z) \leq \mathrm{d}(x, y) + \mathrm{d}(y, z) \tag{29}$$

Geodesics, rays, and lines generalize unit-speed minimizing geodesics to metric spaces.

Definition B.2 (Geodesic, geodesic ray and line)**.** Let $(\mathcal{X}, \mathrm{d})$ be a metric space and let $I = [0, l] \subseteq \mathbb{R}$ be a closed interval. A *geodesic* joining x to y is a map $\gamma : I \to \mathcal{X}$ with $\gamma(0) = x$, $\gamma(l) = y$ such that

$$\mathrm{d}(\gamma(t), \gamma(t')) = |t - t'| \quad \text{for all } t, t' \in I \cap [0, l]. \tag{30}$$

A *geodesic ray* is a map $\gamma : [0, \infty) \to \mathcal{X}$ such that $\mathrm{d}(\gamma(t), \gamma(t')) = |t - t'|$ for all $t, t' \geq 0$. A *geodesic line* is a map $\gamma : \mathbb{R} \to \mathcal{X}$ such that $\mathrm{d}(\gamma(t), \gamma(t')) = |t - t'|$ for all $t, t' \in \mathbb{R}$.

Geodesic metric spaces extend the Riemannian premise that any two points can be joined by a geodesic to metric spaces.

Definition B.3 (Geodesic metric space)**.** The metric space $(\mathcal{X}, \mathrm{d})$ is a *geodesic metric space* (or, more briefly, a *geodesic space*) if every two points are joined by a geodesic. We say that $(\mathcal{X}, \mathrm{d})$ is *uniquely geodesic* if there is exactly one geodesic joining x to y for all $x, y \in \mathcal{X}$.

Convexity is defined through geodesics, generalizing linear convexity in Euclidean space and geodesic convexity on manifolds.

Definition B.4 (Convex subset)**.** Let $(\mathcal{X}, \mathrm{d})$ be a metric space. A subset $C \subseteq \mathcal{X}$ is *convex* if every pair $x, y \in C$ can be joined by a geodesic in $\mathcal{X}$ and the image of every such geodesic is contained in C.

B.2.2. CAT(0) spaces

We next review concepts that extend nonpositive curvature from manifolds to metric spaces. As a starting point, we recall the model spaces, that is, manifolds of constant curvature, including hyperbolic, Euclidean, and spherical geometries. These serve as reference spaces for metric spaces.

Definition B.5 (Model space M_K^n)**.** For $K \in \mathbb{R}$, the *model space* (M_K^n, d_K) is given by

$$(M_K^n, \mathrm{d}_K) = \begin{cases} (\mathbb{L}^n,\ 1/\sqrt{-K}\,\mathrm{d}), & K > 0, \\ (\mathbb{R}^n,\ \mathrm{d}), & K = 0, \\ (\mathbb{S}^n, 1/\sqrt{K}\,\mathrm{d}), & K < 0, \end{cases} \tag{31}$$

where $\mathbb{L}^n$ and $\mathbb{S}^n$ are the n-dimensional unit Lorentz and sphere manifolds, respectively, and d is the geodesic distance in the corresponding manifold. The diameter of M_k^2 is denoted D_k, which is equal to $\pi/\sqrt{K}$ if $K > 0$, and ∞ otherwise.

Definition B.6 (Comparison triangle)**.** Let $\triangle(x, y, z)$ be a geodesic triangle in $\mathcal{X}$ with side lengths $a = \mathrm{d}(y, z)$, $b = \mathrm{d}(x, z)$, and $c = \mathrm{d}(x, y)$. A *comparison triangle* for $\triangle(x, y, z)$ is a triangle $\triangle(\bar{x}, \bar{y}, \bar{z})$ in M_K^2 with $\mathrm{d}_K(\bar{y}, \bar{z}) = a$, $\mathrm{d}_K(\bar{x}, \bar{z}) = b$, and $\mathrm{d}_K(\bar{x}, \bar{y}) = c$. When $a + b + c < 2D_K$, the comparison triangle exists.

CAT(K) spaces encode curvature through triangle comparison with the model plane M_K^2. Intuitively, a $\mathrm{CAT}(K)$ space is a metric space where triangles are "thinner" than the corresponding comparison triangles in the model space M_K^2.

Definition B.7 ($\mathrm{CAT}(K)$ space)**.** Let $\mathcal{X}$ be a metric space and let K be a real number. Let Δ be a geodesic triangle in $\mathcal{X}$ with perimeter less than $2D_K$. Let $\bar{\Delta} \subset M_K^2$ be a comparison triangle for Δ. Then, Δ is said to satisfy the $\mathrm{CAT}(K)$ *inequality* if for all $x, y \in \Delta$ and all comparison points $\bar{x}, \bar{y} \in \bar{\Delta}$,

$$d(x, y) \leq d_K(\bar{x}, \bar{y}). \tag{32}$$

Then, the $\mathrm{CAT}(K)$ space is defined as follows.

- If $K \leq 0$, then $\mathcal{X}$ is called a $\mathrm{CAT}(K)$ space (more briefly, "$\mathcal{X}$ is $\mathrm{CAT}(K)$") if $\mathcal{X}$ is a geodesic space all of whose geodesic triangles satisfy the $\mathrm{CAT}(K)$ inequality.
- If $K > 0$, then $\mathcal{X}$ is called a $\mathrm{CAT}(K)$ space if $\mathcal{X}$ is D_K-geodesic and all geodesic triangles in $\mathcal{X}$ of perimeter less than $2D_K$ satisfy the $\mathrm{CAT}(K)$ inequality.

Here, D_K-geodesic means that for every pair of points $x, y \in \mathcal{X}$ with $d(x, y) < D_K$ there is a geodesic joining x to y.

Definition B.8 (Hadamard space)**.** A *Hadamard space* is a complete $\mathrm{CAT}(0)$ space.

Proposition B.9 (Orthogonal projection)**.** *Let $(\mathcal{X}, \mathrm{d})$ be a $\mathrm{CAT}(0)$ space and let $C \subseteq \mathcal{X}$ be a convex subset that is complete in the induced metric. For every $x \in \mathcal{X}$, there exists a unique point $\pi_C(x) \in C$ such that*

$$\mathrm{d}\left(x, \pi_C(x)\right) = \inf_{y \in C} \mathrm{d}\left(x, y\right) = \mathrm{d}\left(x, C\right). \tag{33}$$

If x' belongs to the geodesic segment $[x, \pi_C(x)]$, then

$$\pi_C(x') = \pi_C(x). \tag{34}$$

The map $\pi_C : \mathcal{X} \to C$ is a called an orthogonal projection*, or simply a* projection*.*

With geodesics, we can extend the asymptote, Busemann function and horosphere in Sec. 2 to metric spaces.

Definition B.10 (Asymptote)**.** Let $(\mathcal{X}, \mathrm{d})$ be a metric space. Two geodesic rays $\gamma, \eta : [0, \infty) \to \mathcal{X}$ are *asymptotic* if

$$\sup_{t \geq 0} \mathrm{d}\left(\gamma(t), \eta(t)\right) < \infty. \tag{35}$$

The set $\partial\mathcal{X}$ of *boundary* points of $\mathcal{X}$, which we shall also call the *points at infinity* or *ideal points*, is the set of equivalence classes of geodesic rays: two geodesic rays being equivalent if and only if they are asymptotic.

Definition B.11 (Busemann function, horoball, horosphere)**.** Let $(\mathcal{X}, \mathrm{d})$ be a metric space and let $\gamma : [0, \infty) \to \mathcal{X}$ be a geodesic ray. The *Busemann function* associated with γ is

$$B^\gamma(x) = \lim_{t\to\infty} \left(\mathrm{d}\left(x, \gamma(t)\right) - t\right), \quad x \in \mathcal{X}. \tag{36}$$

Its sublevel sets $HB^\gamma_\tau = \{x \in \mathcal{X} : B^\gamma(x) \leq \tau\}$ are *horoballs* and the level sets $H^\gamma_\tau = \{x \in \mathcal{X} : B^\gamma(x) = \tau\}$ are *horospheres*.

As shown by Bridson and Haefliger [6, Lem. II. 8.18], the limits in the Busemann function exist. Besides, Busemann functions in Hadamard spaces are invariant to the choice of asymptotic geodesic ray Bridson and Haefliger [6, Cor. II. 8.20].

Corollary B.12. *If $\mathcal{X}$ is a Hadamard space, then the Busemann functions associated to asymptotic rays in $\mathcal{X}$ are equal up to addition of a constant.*

B.3. Gyrovector space

Classical vector spaces can be characterized as a commutative group together with a compatible scalar multiplication. By analogy, a gyrovector space is built from a gyrocommutative gyrogroup endowed with a compatible scalar gyromultiplication [55].

Definition B.13 (Gyrogroup [55])**.** Given a nonempty set G with a binary operation $\oplus : G \times G \to G$, the pair $(G, \oplus)$ is a *gyrogroup* if, for all $x, y, z \in G$, the following axioms hold:
(G1) There is at least one element $e \in G$ called a left identity (or neutral element) such that $e \oplus x = x$.
(G2) There is an element $\ominus x \in G$ called a left inverse of x such that $\ominus x \oplus x = e$.
(G3) There is an automorphism $\mathrm{gyr}[x, y] : G \to G$ for each $x, y \in G$ such that

$$x \oplus (y \oplus z) = (x \oplus y) \oplus \mathrm{gyr}[x, y]z \quad \text{(Left gyroassociative law)}.$$

The map $\mathrm{gyr}[x, y]$ is the *gyration* of G generated by x and y.
(G4) Left reduction law: $\mathrm{gyr}[x, y] = \mathrm{gyr}[x \oplus y, y]$.

Definition B.14 (Gyrocommutative gyrogroup [55])**.** A gyrogroup $(G, \oplus)$ is *gyrocommutative* if

$$x \oplus y = \mathrm{gyr}[x, y]\,(y \oplus x) \quad \text{(Gyrocommutative law)}.$$

Definition B.15 (Gyrovector space [15])**.** A gyrocommutative gyrogroup $(G, \oplus)$ equipped with a scalar gyromultiplication $\odot : \mathbb{R} \times G \to G$ is a *gyrovector space* if, for $s, t \in \mathbb{R}$ and $x, y, z \in G$, the following axioms hold:
(V1) Identity scalar multiplication: $1 \odot x = x$.
(V2) Scalar distributive law: $(s + t) \odot x = s \odot x \oplus t \odot x$.
(V3) Scalar associative law: $(st) \odot x = s \odot (t \odot x)$.
(V4) Gyroautomorphism homogeneity: $\mathrm{gyr}[x, y]\,(t \odot z) = t \odot \mathrm{gyr}[x, y]z$.
(V5) Identity gyroautomorphism: $\mathrm{gyr}\left[s \odot x, t \odot x\right] = \mathbb{1}$, where $\mathbb{1}$ denotes the identity map.

Intuition. Gyrogroups generalize groups: they are nonassociative, yet obey a controlled form of associativity governed by gyrations. Since gyrations in a group are the identity, every group is a gyrogroup. In the same spirit, gyrovector spaces extend vector spaces and provide an algebraic toolkit that has proved effective for modeling hyperbolic geometry [55].

B.4. Hyperbolic geometry

Given $x, y \in \mathcal{H}^n_K$ and tangent vectors $v, w \in T_x\mathcal{H}^n_K$, Tab. 10 summarizes the Riemannian operators.

Gyrovector operators. The gyro-structure over the hyperbolic space can be defined by its Riemannian operators [15, 23]. Given $x, y, z \in \mathcal{H}^n_K$ and $t \in \mathbb{R}$, the gyroaddition and gyromultiplication are defined as

$$x \oplus_{\mathcal{H}} y = \mathrm{Exp}_x\left(\mathrm{PT}_{e\to x}\left(\mathrm{Log}_e\, y\right)\right), \tag{37}$$

$$t \odot_{\mathcal{H}} x = \mathrm{Exp}_e\left(t\,\mathrm{Log}_e\, x\right), \tag{38}$$

$$\mathrm{gyr}[x, y]z = \ominus_{\mathcal{H}}\left(x \oplus_{\mathcal{H}} y\right) \oplus_{\mathcal{H}} \left(x \oplus_{\mathcal{H}} \left(y \oplus_{\mathcal{H}} z\right)\right), \tag{39}$$

where e denotes the origin in $\mathcal{H}^n_K$.

Table 10. Riemannian operators on Poincaré ball and Lorentz ($K < 0$).

Operator	Poincaré ball $\mathbb{P}_K^n$	Lorentz $\mathbb{L}_K^n$
Definition	$\mathbb{P}_K^n = \{x \in \mathbb{R}^n \mid \|x\|^2 < -1/K\}$	$\mathbb{L}_K^n = \{x \in \mathbb{R}^{n+1} \mid \langle x, x\rangle_{\mathcal{L}} = 1/K, x_t > 0\}$
$g_x(w, v)$	$\left(\lambda_x^K\right)^2 \langle w, v\rangle, \quad \lambda_x^K = \frac{2}{1 + K\|x\|^2}$	$\langle w, v\rangle_{\mathcal{L}} = \langle v_s, w_s\rangle - v_t w_t$
$\mathrm{d}(x, y)$	$\frac{2}{\sqrt{\lvert K\rvert}} \tanh^{-1}\left(\sqrt{\lvert K\rvert}\, \lVert -x \oplus_{\mathrm{M}} y\rVert\right)$	$\frac{1}{\sqrt{\lvert K\rvert}} \cosh^{-1}\left(K \langle x, y\rangle_{\mathcal{L}}\right)$
$\mathrm{Log}_x\, y$	$\frac{2}{\sqrt{\lvert K\rvert}\lambda_x^K} \tanh^{-1}\left(\sqrt{\lvert K\rvert}\, \lVert -x \oplus_{\mathrm{M}} y\rVert\right) \frac{-x \oplus_{\mathrm{M}} y}{\lVert -x \oplus_{\mathrm{M}} y\rVert}$	$\frac{\cosh^{-1}(\beta)}{\sqrt{\beta^2 - 1}}(y - \beta x), \quad \beta = K \langle x, y\rangle_{\mathcal{L}}$
$\mathrm{Exp}_x\, v$	$x \oplus_{\mathrm{M}} \left(\tanh\left(\sqrt{\lvert K\rvert}\frac{\lambda_x^K \lVert v\rVert}{2}\right) \frac{v}{\sqrt{\lvert K\rvert}\lVert v\rVert}\right)$	$\cosh(\alpha)x + \frac{\sinh(\alpha)}{\alpha} v, \quad \alpha = \sqrt{\lvert K\rvert}\lVert v\rVert_{\mathcal{L}}$
$\mathrm{PT}_{x \to y}(v)$	$\frac{\lambda_x^K}{\lambda_y^K} \mathrm{gyr}[y, -x]v$	$v - \frac{K \langle y, v\rangle_{\mathcal{L}}}{1 + K \langle x, y\rangle_{\mathcal{L}}}(x + y)$

On the Poincaré ball $\mathbb{P}_K^n$, such gyro-structure is known as the Möbius gyrovector space [55, Ch. 6.14]:

$$x \oplus_{\mathrm{M}} y = \frac{\left(1 - 2K\langle x, y\rangle - K\|y\|^2\right) x + \left(1 + K\|x\|^2\right) y}{1 - 2K\langle x, y\rangle + K^2\|x\|^2\|y\|^2},$$
$$t \odot_{\mathrm{M}} x = \frac{\tanh\left(t \tanh^{-1}(\sqrt{\lvert K\rvert}\|x\|)\right)}{\sqrt{\lvert K\rvert}} \frac{x}{\|x\|}$$

where $\ominus_{\mathrm{M}} x = -1 \odot_{\mathrm{M}} x = -x$ is the gyroinverse and $\mathbf{0}$ is the gyro identity: $\mathbf{0} \oplus_{\mathrm{M}} x = x, \forall x \in \mathbb{P}_K^n$. As shown by Chen et al. [15, Props. 24-25], the Lorentz gyroaddition and gyromultiplication also admit closed-form expressions:

$$x \oplus_{\mathbb{L}} y = \begin{cases} x, & y = \overline{\mathbf{0}}, \\ y, & x = \overline{\mathbf{0}}, \\ \begin{bmatrix} \frac{1}{\sqrt{\lvert K\rvert}} \frac{D - KN}{D + KN} \\ \frac{2(A_s x_s + A_y y_s)}{D + KN} \end{bmatrix}, & \text{Otherwise.} \end{cases} \tag{40}$$

$$t \odot_{\mathbb{L}} x = \begin{cases} \overline{\mathbf{0}}, & t = 0 \vee x = \overline{\mathbf{0}} \\ \frac{1}{\sqrt{\lvert K\rvert}} \begin{bmatrix} \cosh(t\theta) \\ \frac{\sinh(t\theta)}{\|x_s\|} x_s \end{bmatrix}, & \text{Otherwise,} \end{cases} \tag{41}$$

Here, $\theta = \cosh^{-1}(\sqrt{\lvert K\rvert}x_t)$, $A_s = ab^2 - 2Kbs_{xy} - Kan_y$ and $A_y = b(a^2 + Kn_x)$ with the following notation:

$$\begin{aligned} &a = 1 + \sqrt{\lvert K\rvert}x_t, b = 1 + \sqrt{\lvert K\rvert}y_t, \\ &n_x = \|x_s\|^2, n_y = \|y_s\|^2, s_{xy} = \langle x_s, y_s\rangle, \\ &D = a^2b^2 - 2Kabs_{xy} + K^2 n_x n_y, \\ &N = a^2 n_y + 2abs_{xy} + b^2 n_x. \end{aligned} \tag{42}$$

In particular, the gyro identity is $\overline{\mathbf{0}}$ and the gyroinverse is $\ominus_{\mathbb{L}} x = -1 \odot_{\mathbb{L}} x = [x_t, -x_s^\top]^\top$.

C. Comparison with existing hyperbolic MLR

Hyperbolic MLRs follow a point-to-hyperplane formulation, of which Chen et al. [12, Eqs. 4-6] provides the Riemannian prototype. Therefore, the key difference lies in hyperplanes and point-to-hyperplane distances across methods. In addition to Tab. 2, Tabs. 11 and 12 further make this comparison. We draw the following three conclusions.

Table 11. Comparison of hyperplanes. Compact params indicates whether the parameterization requires an additional manifold-valued point.

Method	Hyperplane	Formulation	Applied manifolds	Compact params
Euclidean MLR	Euclidean	$\{x \in \mathbb{R}^n \mid \langle a, x\rangle + b = 0\}$ $a \in \mathbb{R}^n,\ b \in \mathbb{R}$	$\mathbb{R}^n$	✓
Poincaré MLR [23]	Geodesic [23, Def. 3.1]	$\left\{x \in \mathbb{P}^n_K \mid \left\langle \mathrm{Log}_p(x), a\right\rangle_p = 0\right\}$ $p \in \mathbb{P}^n_K,\ a \in T_p\mathbb{P}^n_K$	$\mathbb{P}^n_K$	✗
Pseudo-Busemann MLR [45]	Busemann & gyro [45, Def. 4.1]	$\{x \in \mathbb{P}^n_K \mid B^v(-p \oplus_{\mathrm{M}} x) = 0\}$ $v \in \mathbb{S}^{n-1},\ p \in \mathbb{P}^n_K$	$\mathbb{P}^n_K$	✗
Lorentz MLR [3]	Ambient Minkowski [3, Eq. (7)]	$\{x \in \mathbb{L}^n_K \mid \langle w, x\rangle_{\mathcal{L}} = 0\}$ $p \in \mathbb{L}^n_K,\ w \in T_p\mathbb{L}^n_K$	$\mathbb{L}^n_K$	✗
BMLR	Horosphere	$\{x \in \mathcal{H}^n_K \mid -\alpha B^v(x) + b = 0\}$ $\alpha > 0,\ v \in \mathbb{S}^{n-1},\ b \in \mathbb{R}$	$\mathbb{P}^n_K, \mathbb{L}^n_K$	✓

Table 12. Comparison of point-to-hyperplane distances. **Real** means the point-to-hyperplane distance is the real distance, obtained by $\inf_{y\in H} \mathrm{d}(x, y)$ with H as a hyperplane and d as the geodesic distance. Instead, **Pseudo** means the point-to-hyperplane distance is a surrogate, which only equals to the real distance under the Euclidean geometry.

Method	Point-to-hyperplane distance	Applied manifolds	Dist
Euclidean MLR	$\frac{\lvert\langle a, x\rangle + b\rvert}{\lVert a\rVert}$	$\mathbb{R}^n$	**Real**
Poincaré MLR [23]	$\frac{1}{\sqrt{-K}} \sinh^{-1}\left(\frac{2\sqrt{-K}\,\lvert\langle -p \oplus_{\mathrm{M}} x, a\rangle\rvert}{\left(1 + K\lVert -p \oplus_{\mathrm{M}} x\rVert^2\right)\lVert a\rVert}\right)$ [23, Thm. 5]	$\mathbb{P}^n_K$	**Real**
Pseudo-Busemann MLR [45]	$\mathrm{d}(x, p)\frac{B^v(-p \oplus_{\mathrm{M}} x)}{\lVert -p \oplus_{\mathrm{M}} x\rVert}$ [45, Cor. 4.3]	$\mathbb{P}^n_K$	**Pseudo**
Lorentz MLR [3]	$\frac{1}{\sqrt{-K}}\left\lvert\sinh^{-1}\left(\sqrt{-K}\frac{\langle v, x\rangle_{\mathcal{L}}}{\lVert v\rVert_{\mathcal{L}}}\right)\right\rvert$ [3, Eq. (44)]	$\mathbb{L}^n_K$	**Real**
BMLR	$\frac{\lvert -\alpha B^v(x) + b\rvert}{\alpha}$	$\mathbb{P}^n_K, \mathbb{L}^n_K$	**Real**

1. **Hyperplanes.** Our BMLR uses Busemann-based horospheres that simultaneously satisfy three desiderata: (i) compact parameterization without a per-class manifold-valued point, whereas other hyperbolic ones[2] are over-parameterized; (ii) a natural generalization of Euclidean hyperplanes via horospheres, while the Lorentz MLR relies on the ambient Minkowski space, failing to fully respect the intrinsic geometry; and (iii) applicability across hyperbolic models, whereas the Lorentz MLR is tailored to the Lorentz model.
2. **Point-to-hyperplane distances.** Although pseudo-Busemann MLR also exploits Busemann functions, it relies on a pseudo point-to-hyperplane distance that only coincides with the real distance in Euclidean geometry. In contrast, our BMLR calculates the real point-to-horosphere distance, ensuring geometric fidelity across hyperbolic models.
3. **Batch efficiency.** Recalling Tab. 2, the Poincaré MLR [23] computes logits using $\langle -p_k \oplus_{\mathrm{M}} x, a_k\rangle$ and $\lVert -p_k \oplus_{\mathrm{M}} x\rVert^2$. For a batch $X \in \mathbb{R}^{\mathrm{bs}\times n}$ and C classes, evaluating $-p_k \oplus_{\mathrm{M}} X$ for every k yields an intermediate tensor of shape $[\mathrm{bs}, C, n]$,

[2]We note that Bdeir et al. [3], Shimizu et al. [51] mitigate this issue through re-parameterization: $a = \mathrm{PT}_{e\to p}(z)$ and $p = \exp_e\left(b\frac{z}{\lVert z\rVert}\right)$, where $e \in \mathcal{H}^n_K$ denotes the origin, $z \in \mathbb{R}^n$, and $b \in \mathbb{R}$. Nevertheless, the underlying definitions are over-parameterized.

and materializing this tensor can cause GPU out of memory (OOM) when n or C is large. The same limitation holds for the pseudo-Busemann MLR. Consequently, their official implementations compute per class in a for-loop, which is batch inefficient. By contrast, BMLR uses logits $-\alpha_k B^{v_k}(x) + b_k$, whose Busemann term reduces to class-wise inner products $\langle v_k, x\rangle$ (or $\langle v_k, x_s\rangle$). With $X \in \mathbb{R}^{\mathrm{bs}\times n}$ and $V = [v_1, \dots, v_C] \in \mathbb{R}^{n\times C}$, such inner products can be efficiently implemented as a single matrix multiplication XV without any $[\mathrm{bs}, C, n]$ intermediate, yielding high throughput and low memory usage.

D. Busemann fully connected layers and point-to-horosphere distances

An apparently natural attempt to define a hyperbolic FC layer is to replace the LHS of Eq. (22) by the *signed point-to-horosphere* distance. The Euclidean hyperplane passing through the origin and orthogonal to $e_k \in \mathbb{R}^m$ is $H_{e_k,0} = \{y \in \mathbb{R}^m \mid \langle e_k, y\rangle = 0\}$. The corresponding hyperbolic horosphere, following Eq. (18), is $H_{e_k,1,0} = \{y \in \mathcal{H}^n_K \mid -B^{e_k}(y) = 0\}$. By Eq. (20), the signed point-to-horosphere distance to $H_{e_k,1,0}$ equals $-B^{e_k}(y)$. Accordingly, we can define an alternative FC mapping $\mathcal{F} : \mathcal{H}^n_K \ni x \mapsto y \in \mathcal{H}^m_K$ via

$$B^{e_k}(y) = \alpha_k B^{v_k}(x) - b_k, \quad k = 1, \dots, m, \tag{43}$$

where $u_k(x) = -\alpha_k B^{v_k}(x) + b_k$, and $\alpha_k > 0$, $v_k \in \mathbb{S}^{n-1}$, $b_k \in \mathbb{R}$ are learnable. Although this only differs from Eq. (22) on the LHS, the following discussion shows that such a definition is infeasible in general and fails to deliver a valid hyperbolic FC layer.

D.1. Poincaré model

Using Eq. (3) with $v = e_k$, Eq. (43) becomes

$$\frac{1}{\sqrt{-K}} \log\left(\frac{\left\|e_k - \sqrt{-K}y\right\|^2}{1 + K\|y\|^2}\right) = -u_k(x). \tag{44}$$

Define $t_k = \exp\left(-\sqrt{-K}u_k(x)\right) > 0$. Exponentiating Eq. (44) gives

$$t_k\left(1 + K\|y\|^2\right) = 1 - 2\sqrt{-K}y_k - K\|y\|^2, \quad k = 1, \dots, m. \tag{45}$$

Writing $R = \|y\|^2$, Eq. (45) yields an affine expression for each coordinate

$$y_k = c_k + d_k R, \quad c_k = \frac{1 - t_k}{2\sqrt{-K}}, \quad d_k = \frac{\sqrt{-K}}{2}(1 + t_k). \tag{46}$$

Imposing $R = \sum_{k=1}^m y_k^2 = \sum_{k=1}^m (c_k + d_k R)^2$ gives a quadratic in R:

$$A_2 R^2 + (A_1 - 1) R + A_0 = 0, \tag{47}$$

where, denoting $T = \sum_{k=1}^m t_k$ and $q = \sum_{k=1}^m t_k^2$,

$$A_2 = \frac{-K}{4}(m + 2T + q), \quad A_1 = \frac{m - q}{2}, \quad A_0 = \frac{m - 2T + q}{4(-K)} \geq 0. \tag{48}$$

Its discriminant is

$$\begin{aligned}
\Delta_{\mathrm{P}} &= (A_1 - 1)^2 - 4A_2A_0 \\
&= \left(\frac{m - q}{2} - 1\right)^2 - 4\left(\frac{-K}{4}(m + 2T + q)\right)\left(\frac{m - 2T + q}{4(-K)}\right) \\
&= \left(\frac{m - 2 - q}{2}\right)^2 - \frac{(m + 2T + q)(m - 2T + q)}{4} \\
&= \frac{1}{4}\left[(m - 2 - q)^2 - \left((m + q)^2 - (2T)^2\right)\right] \\
&= \frac{1}{4}\left[((m - 2 - q) - (m + q))((m - 2 - q) + (m + q)) + 4T^2\right] \\
&= \frac{1}{4}\left[(-2 - 2q)(2m - 2) + 4T^2\right] \\
&= T^2 - (m - 1)(1 + q).
\end{aligned} \tag{49}$$

A real solution R exists only if $\Delta_{\mathrm{P}} \geq 0$. In addition, feasibility requires $0 \leq R < -1/K$ so that $y \in \mathbb{P}_K^m$. Such conditions can fail for generic $\{u_k(x)\}$, in which case no $y \in \mathbb{P}_K^m$ satisfies Eq. (43).

D.2. Lorentz model

Using Eq. (4) with $v = e_k$, Eq. (43) becomes

$$\frac{1}{\sqrt{-K}} \log\left(\sqrt{-K}\,(y_t - (y_s)_k)\right) = -u_k(x). \tag{50}$$

Define $t_k = \exp\left(-\sqrt{-K}u_k(x)\right) > 0$ and, denoting $T = \sum_{k=1}^m t_k$ and $q = \sum_{k=1}^m t_k^2$, we obtain from Eq. (50),

$$(y_s)_k = y_t - \frac{t_k}{\sqrt{-K}}, \quad k = 1, \ldots, m, \quad \Longrightarrow \quad y_s = y_t \mathbf{1} - \frac{1}{\sqrt{-K}} t. \tag{51}$$

Enforcing the hyperboloid constraint $\|y_s\|^2 - y_t^2 = 1/K$ yields a quadratic in y_t:

$$(m-1)Ky_t^2 + 2\sqrt{-K}Ty_t - (1+q) = 0. \tag{52}$$

The discriminant is

$$\begin{aligned} \Delta_{\mathrm{L}} &= \left(2\sqrt{-K}T\right)^2 - 4(m-1)K\,(-(1+q)) \\ &= 4(-K)T^2 + 4(m-1)K\,(1+q) \\ &= 4(-K)\left[T^2 - (m-1)\,(1+q)\right]. \end{aligned} \tag{53}$$

Hence, a real y_t exists only if $T^2 - (m-1)\,(1+q) \geq 0$. In addition, feasibility requires $y_t > 0$. Such conditions can fail for generic $\{u_k(x)\}$, where no $y \in \mathbb{L}_K^m$ satisfies Eq. (43).

D.3. Summary

Equating Busemann coordinates as in Eq. (43) requires nontrivial inequalities on the responses $\{u_k(x)\}$. These constraints are not guaranteed during learning, so the system can become infeasible and the output y undefined. This motivates our choice in Eq. (22) to use signed point-to-hyperplane distances on the LHS, which admit closed-form solutions that are feasible for all inputs and parameters in both the Poincaré and Lorentz models.

E. Experimental details and additional results

E.1. Image classification

E.1.1. Datasets

The CIFAR-10 [36] and CIFAR-100 [36] datasets each contain 60,000 32×32 color images from 10 and 100 classes, respectively. We use the standard PyTorch splits: 50,000 training images and 10,000 test images. Tiny-ImageNet [37] is a subset of ImageNet with 100,000 images from 200 classes, resized to 64×64. We use the official validation split for evaluation. The large-scale ImageNet-1k [17] dataset contains 1.28M training images, 50K validation images, and 100K test images distributed across 1k classes.

E.1.2. Implementation details

For CIFAR-10/100 and Tiny-ImageNet, we follow Bdeir et al. [3, App. C.1]. For ImageNet-1k, we follow Guo et al. [28, Sec. 4]. Tab. 13 summarizes the dataset-specific hyperparameters. For the hyperbolic MLR, before mapping into the hyperbolic space, we clip the feature vector by

$$\mathrm{CLIP}\,(x; r) = \min\left\{1, \frac{r}{\|x\|}\right\} x \tag{54}$$

where $r > 0$ is a hyperparameter. The clipped Euclidean embedding is projected via the exponential map to the target hyperbolic space: $\mathrm{Exp}_e\left(\mathrm{CLIP}(x; r)\right)$. For the Lorentz model, the clipping parameter is $r = 1$ on CIFAR-10/100 and $r = 4$ on Tiny-ImageNet and ImageNet-1k. On the Poincaré ball, $r = 1$ on all four datasets.

All methods are implemented in PyTorch and trained with cross-entropy loss. The results of MLR, PMLR, and LMLR on CIFAR-10/100 and Tiny-ImageNet are copied from Bdeir et al. [3, Tab. 1], while the ones of PBMLR-P on CIFAR-10/100 are copied from Nguyen et al. [45, Tab. 2]. The remaining results are obtained by our careful implementation.

Table 13. Summary of hyperparameters used in the image classification task.

Hyperparameter	CIFAR-10/100 Tiny-ImageNet	ImageNet-1k
Epochs	200	100
Batch size	128	256
Initial learning rate	0.1	0.1
LR schedule	$60, 120, 160; \gamma = 0.2$	$30, 60, 90; \gamma = 0.1$
Weight decay	$5e^{-4}$	$1e^{-4}$
Optimizer	SGD	SGD
Precision	32-bit	32-bit
#GPUs	1 (RTX A6000)	2 (RTX A100)
Curvature K	-1	-1

Table 14. Summary statistics for TEB.

Task	Species	Datasets	Num. classes	Max length	Train / Dev / Test
Retrotransposons	Plant	LTR Copia	2	500	7666 / 682 / 568
		LINEs		1000	22502 / 2030 / 1782
		SINEs		500	21152 / 1836 / 1784
DNA transposons	Plant	CMC-EnSpm	2	200	19912 / 1872 / 1808
		hAT-Ac		1000	17322 / 1822 / 1428
Pseudogenes	Human	processed	2	1000	17956 / 1046 / 1740
		unprocessed			12938 / 766 / 884

Table 15. Summary statistics for the adopted GUE datasets.

Task	Species	Dataset	Num. classes	Length	Train / Dev / Test
Core promoter detection	Human	tata	2	70	4904 / 613 / 613
		notata			42452 / 5307 / 5307
		all			47356 / 5920 / 5920
Promoter detection	Human	tata	2	300	4904 / 613 / 613
		notata			42452 / 5307 / 5307
		all			47356 / 5920 / 5920
Covid variant classification	Virus	Covid	9	1000	77669 / 7000 / 7000
Species classification	Fungi	Fungi	25	5000	8000 / 1000 / 1000
	Virus	Virus	20	10000	4000 / 500 / 500

E.2. Genome sequence learning

E.2.1. Datasets

TEB. The Transposable Elements Benchmark (TEB) [33] comprises seven binary classification datasets that investigate transposable elements (TEs) across plant and human genomes. The seven datasets are LTR Copia, LINEs, and SINEs for plant retrotransposons; CMC-EnSpm and hAT-Ac for plant DNA transposons; and processed and unprocessed pseudogenes for human pseudogenes. TEB targets a less explored area of genome organization in genomics deep learning and provides a novel resource for benchmarking models. For each dataset, positive examples are sequences spanning annotated elements of interest, and negatives are randomly sampled, non-overlapping genomic segments outside these regions. We adopt chromosome-level training, validation, and test splits, using chromosomes 8 and 9 for validation and test in plant genomes, and chromosomes 20 to 22 and 17 to 19 for validation and test in human genomes, respectively. Summary statistics are provided in Tab. 14.

GUE. The Genome Understanding Evaluation (GUE) benchmark [62] is a recently published tool that contains seven

Table 16. Hyperparameters for genome sequence learning.

Batch size	100
Epochs	100
Optimizer	Adam
β_1, β_2	0.9, 0.999
Initial learning rate	$1e^{-4}$
LR schedule	$60, 85; \gamma = 0.1$
Weight decay	0.1
#GPUs	1 (RTX A6000)
Initial curvature K	-1

Table 17. Summary statistics for the node classification datasets.

Dataset	#Nodes	#Edges	#Classes	#Features
Disease	1044	1043	2	1000
Airport	3188	18631	4	4
PubMed	19717	44338	3	500
Cora	2708	5429	7	1433

biologically significant genome analysis tasks that span 28 datasets. GUE prioritizes genomic datasets that are challenging enough to discern differences between models. The datasets contain sequences ranging from 70 to 1000 base pairs in length and originating from yeast, mouse, human, and virus genomes. In our experiments, we select Core promoter detection and Promoter detection, and the multi-class tasks Covid variant classification and species classification, totaling nine datasets from GUE. Summary statistics are provided in Tab. 15.

E.2.2. Implementation details

We mainly follow the official implementations of Khan et al. [33] for data processing, model architecture, and training. We adopt a simple CNN with three convolutional blocks followed by dense, ReLU activated layers to extract features [33, Fig. 4]. Before classification, the features are clipped and mapped to the target hyperbolic space as in Eq. (54), then passed to either a prior hyperbolic MLR or our BMLR head. The clipping factor defaults to $r = 1$, with the following exceptions on GUE: on the Lorentz model, $r = 2.0$ for Covid variant classification and $r = 5.0$ for species classification; on the Poincaré ball, $r = 2.0$ for species classification. Following Khan et al. [33], we treat the curvature K as a learnable parameter initialized as -1. The remaining hyperparameters are listed in Tab. 16. All models share these hyperparameters, except LMLR on Covid variant classification, for which we set the weight decay to $1e^{-3}$ to ensure convergence.

All methods are implemented in PyTorch and trained with cross-entropy loss. Results are obtained from our reimplementation.

E.3. Node classification

E.3.1. Datasets

Disease [1]. It represents a disease propagation tree, simulating the SIR disease transmission model, with each node representing either an infection or a non-infection state.

Airport [61]. It is a transductive dataset where nodes represent airports and edges represent the airline routes as from OpenFlights.org.

PubMed [44]. This is a standard benchmark describing citation networks where nodes represent scientific papers in the area of medicine, edges are citations between them, and node labels are academic (sub)areas.

Cora [50]. It is a citation network where nodes represent scientific papers in the area of machine learning, edges are citations between them, and node labels are academic (sub)areas.

Tab. 17 summarizes the statistics of the datasets.

E.3.2. Implementation details

We follow the official implementations of HGCN [8] and PBMLR [45] and conduct experiments on the Poincaré ball and the Lorentz model, respectively. We adhere to their experimental settings. The only changes are weight decay and dropout.

Table 18. Hyperparameters for node classification on Disease, Airport, PubMed, and Cora.

Space	Weight decay	Dropout
$\mathbb{P}_K^n$	$1e-4, 1e-5, 1e-3, 1e-3$	$0.3, 0, 0, 0.2$
$\mathbb{L}_K^n$	$1e-4, 5e-5, 1e-3, 1e-3$	$0, 0, 0, 0.3$

Table 19. Efficiency comparison: fit time (s/epoch) and parameter count. Slowest results and largest parameter counts are in **red**.

Space	Method	Disease		Airport		PubMed		Cora	
		Fit Time	#Params	Fit Time	#Params	Fit Time	#Params	Fit Time	#Params
$\mathbb{P}_K^n$	Möbius	0.0200	464	0.0535	480	0.1120	8288	0.0229	23216
	Poincaré FC	0.0198	528	0.0536	544	0.1176	8352	0.0248	23280
	BFC-P	0.0201	528	0.0512	544	0.1123	8352	0.0231	23280
$\mathbb{L}_K^n$	LTFC	**0.0343**	464	**0.0818**	480	**0.1633**	8288	**0.0370**	23216
	Lorentz FC	0.0232	**563**	0.0715	**580**	0.1537	**8876**	0.0261	**24737**
	BFC-L	0.0244	528	0.0713	544	0.1525	8352	0.0280	23280

We train with cross-entropy loss and the Adam optimizer [35] for 5000 epochs with a learning rate of $1e^{-2}$, curvature as -1, embedding dimension 16, and three GCN layers. We tune weight decay and dropout and report the values in Tab. 18.

All methods are implemented in PyTorch and run on a single RTX A6000 GPU. The results of HGCN-PMLR and HGCN-PBMLR-P on the Poincaré ball are taken from Nguyen et al. [45, Tab. 11]. Results for the remaining baselines are obtained from our reimplementation following the original settings.

E.4. Link prediction

The datasets are identical to those used for node classification. We provide implementation details and additional ablation studies below.

E.4.1. Implementation details

We follow the official implementations of HNN [23], HNN++ [51], and HyboNet [11], and adopt the experimental protocol of Chami et al. [8] for link prediction. The encoder consists of two fully connected layers: the first maps the input features to 16, and the second maps 16 to 16. Each FC layer is instantiated as either our BFC or an existing hyperbolic FC layer. After each FC, we apply the activation $\exp_e\left(\mathrm{ReLU}(\log_e(x))\right)$, where e denotes the model origin. Following Chami et al. [8], this activation is disabled on Cora. As in the Möbius layer, we apply a gyro bias after each FC, that is, $x \oplus_{\mathcal{H}} b$. We train with Adam [35] at a learning rate of $1e^{-2}$ and tune weight decay and FC dropout. For BFC, we set $\phi = \tanh$ on Airport and Cora, which yields better performance, while we use the identity map on the other two datasets.

E.4.2. Ablations on training time and parameter count

Tab. 19 summarizes fit time per epoch and parameter counts. Our BFC layers achieve training time and model size comparable to existing layers. In particular, LTFC is the slowest due to costly logarithmic and exponential maps, and LFC uses the largest number of parameters among Lorentz variants.

E.4.3. Ablations on the activation in BFC layers

Table 20. Ablation on activation ϕ in BFC layers.

Space	ϕ	Airport	Cora
$\mathbb{P}_K^n$	None	94.57 ± 0.29	87.55 ± 0.34
	tanh	**94.88 ± 0.39**	**91.94 ± 0.32**
$\mathbb{L}_K^n$	None	94.91 ± 0.24	87.45 ± 1.39
	tanh	**95.37 ± 0.17**	**92.28 ± 0.12**

We find that $\phi = \tanh$ in our BFC layers (see Sec. 4.2) is beneficial on Airport and Cora. This is validated by Tab. 20, which shows that setting $\phi = \tanh$ consistently improves AUC over the identity activation on the Poincaré and Lorentz models.

F. Proofs

F.1. Proof of Thm. 3.1

Proof. Denoting $\kappa^2 = -K$ with $\kappa > 0$, we rewrite the Busemann functions in Eqs. (3) and (4) as

$$\text{(Poincaré)} \quad B^v(x) = \frac{1}{\kappa} \log \left(\frac{\|v - \kappa x\|^2}{1 - \kappa^2 \|x\|^2} \right), \tag{55}$$

$$\text{(Lorentz)} \quad B^v(x) = \frac{1}{\kappa} \log \left(\kappa x_t - \kappa \langle x_s, v \rangle \right). \tag{56}$$

For the Poincaré case,

$$\frac{\|v - \kappa x\|^2}{1 - \kappa^2 \|x\|^2} = \frac{1 - 2\kappa \langle v, x \rangle + \kappa^2 \|x\|^2}{1 - \kappa^2 \|x\|^2} = 1 + \frac{-2\kappa \langle v, x \rangle + 2\kappa^2 \|x\|^2}{1 - \kappa^2 \|x\|^2}. \tag{57}$$

Using $\log(1 + u) = u + O(u^2)$ as $u \to 0$ and $\left(1 - \kappa^2 \|x\|^2\right)^{-1} = 1 + O(\kappa^2)$, we obtain

$$\begin{aligned} B^v(x) &= \frac{1}{\kappa} \left[\left(-2\kappa \langle v, x \rangle + 2\kappa^2 \|x\|^2 \right) \left(1 + O(\kappa^2) \right) + O(\kappa^2) \right] \\ &= -2 \langle v, x \rangle + 2\kappa \|x\|^2 + O(\kappa), \\ &= -2 \langle v, x \rangle + O(\kappa). \end{aligned} \tag{58}$$

Therefore, $B^v(x) \xrightarrow{K \to 0^-} -2 \langle v, x \rangle$.

For the Lorentz case, the hyperboloid constraint $-x_t^2 + \|x_s\|^2 = -\kappa^{-2}$ and $x_t > 0$ yield

$$\kappa x_t = \sqrt{1 + \kappa^2 \|x_s\|^2} = 1 + \frac{1}{2}\kappa^2 \|x_s\|^2 + O(\kappa^4). \tag{59}$$

Set $z = -\kappa \langle x_s, v \rangle + \frac{1}{2}\kappa^2 \|x_s\|^2 + O(\kappa^4)$. Then

$$\log(\kappa x_t - \kappa \langle x_s, v \rangle) = \log(1 + z) = z + O(z^2) = -\kappa \langle x_s, v \rangle + O(\kappa^2), \tag{60}$$

which implies $B^v(x) = -\langle x_s, v \rangle + O(\kappa)$ and therefore $B^v(x) \xrightarrow{K \to 0^-} -\langle x_s, v \rangle$.

Substituting the above two limits into $u_k(x) = -\alpha_k B^{v_k}(x) + b_k$ gives the stated Euclidean limits of the logits. □

F.2. Proof of Thm. 3.3

We first review two useful characterizations of Busemann functions in Hadamard spaces.

Definition F.1. Let $\mathcal{B}$ be the set of functions $h : \mathcal{X} \to \mathbb{R}$ on $(\mathcal{X}, \mathrm{d})$ satisfying:

(i) h is convex;
(ii) 1-Lipschitz: $|h(x) - h(y)| \le \mathrm{d}(x, y)$ for all $x, y \in \mathcal{X}$;
(iii) for any $x_0 \in \mathcal{X}$ and $r > 0$, the function h attains its minimum on the sphere $S_r(x_0)$ at a unique point y with $h(y) = h(x_0) - r$.

Proposition F.2. *For a function $h : \mathcal{X} \to \mathbb{R}$, the following conditions are equivalent:*

(1) h is a Busemann;
(2) $h \in \mathcal{B}$;
(3) h is convex, and for every $t \in \mathbb{R}$, the set $h^{-1}(-\infty, t]$ is nonempty; moreover, for each $x \in \mathcal{X}$, the curve $c_x : [0, \infty) \to \mathcal{X}$ defined by $t \mapsto \pi_{h^{-1}(-\infty, h(x)-t]}(x)$ is a geodesic ray.

Now, we are ready to prove Thm. 3.3.

Proof. As $\tau_2 = \tau_1$ is trivial, we only consider $\tau_2 \neq \tau_1$. We assume $\tau_2 > \tau_1$, and discuss the other direction at last.

Step 1: symmetry. By definition,

$$\mathrm{d}\left(H^\gamma_{\tau_1}, H^\gamma_{\tau_2}\right) = \inf_{x\in H^\gamma_{\tau_1}, y\in H^\gamma_{\tau_2}} \mathrm{d}(x,y) = \inf_{y\in H^\gamma_{\tau_2}, x\in H^\gamma_{\tau_1}} \mathrm{d}(y,x) = \mathrm{d}\left(H^\gamma_{\tau_2}, H^\gamma_{\tau_1}\right). \tag{61}$$

Step 2: lower bound. For any $x \in H^\gamma_{\tau_2}$ and $y \in H^\gamma_{\tau_1}$, the 1-Lipschitz property of B^γ gives

$$|B^\gamma(x) - B^\gamma(y)| \leq \mathrm{d}(x,y). \tag{62}$$

With $B^\gamma(x) = \tau_2$ and $B^\gamma(y) = \tau_1$ this yields

$$\tau_2 - \tau_1 \leq \mathrm{d}(x,y) \qquad \forall y \in H^\gamma_{\tau_1}. \tag{63}$$

Taking infimum in Eq. (63) over $y \in H^\gamma_{\tau_1}$ gives

$$\tau_2 - \tau_1 \leq \mathrm{d}\left(x, H^\gamma_{\tau_1}\right). \tag{64}$$

Step 3: upper bound. For any $x \in H^\gamma_{\tau_2}$, by property (3) in Prop. F.2, the projection map

$$c_x(t) = \pi_{\{B^\gamma \leq B^\gamma(x)-t\}}(x) = \pi_{HB^\gamma_{\tau_2-t}}(x), \quad t \in [0,\infty), \tag{65}$$

is a unit-speed geodesic ray: $\mathrm{d}(x, c_x(t)) = t$.

Let $t = \tau_2 - \tau_1 > 0$ and $z = c_x(t) \in HB^\gamma_{\tau_1}$. If $B^\gamma(z) < \tau_1$, then z lies in the interior of the horoball $HB^\gamma_{\tau_1}$. We can move slightly from z toward x along the geodesic segment $\overline{xz}$ to obtain a point z_ε with $\mathrm{d}(x, z_\varepsilon) < \mathrm{d}(x,z)$, contradicting the minimality of the projection. Hence, the projected point $z = c_x(t)$ indeed lies on $H^\gamma_{\tau_1}$: $B^\gamma(z) = \tau_1$. Then, we have the following:

$$\mathrm{d}(x, H^\gamma_{\tau_1}) \leq \mathrm{d}(x, HB^\gamma_{\tau_1}) = \mathrm{d}(x,z) = \mathrm{d}(x, c_x(t)) = t = \tau_2 - \tau_1. \tag{66}$$

Step 4: sandwich closure. Combining Eqs. (64) and (66) gives

$$\mathrm{d}(x, H^\gamma_{\tau_1}) = \tau_2 - \tau_1. \tag{67}$$

Combining Eq. (66) and Eq. (64),

$$\mathrm{d}\left(x, H^\gamma_{\tau_1}\right) = \tau_2 - \tau_1, \qquad \text{for every } x \in H^\gamma_{\tau_2}. \tag{68}$$

The right-hand side does not depend on x, hence

$$\mathrm{d}\left(H^\gamma_{\tau_2}, H^\gamma_{\tau_1}\right) = \tau_2 - \tau_1. \tag{69}$$

Step 5: opposite direction. If $\tau_1 > \tau_2$, we can swap the roles of τ_1, τ_2 above due to the symmetry of the distance between horospheres, which brings

$$\mathrm{d}\left(H^\gamma_{\tau_2}, H^\gamma_{\tau_1}\right) = |\tau_2 - \tau_1|. \tag{70}$$

□

F.3. Proof of Thm. 4.1

Proof. The hyperplane and point-to-hyperplane distance are presented in Tabs. 11 and 12. The origin of the Poincaré ball model is $e = \mathbf{0} \in \mathbb{P}^m_K$. The specific ones w.r.t. the origin [51, Def. 1 and Eq. (56)] are

$$H_{e_k,\mathbf{0}} = \{y \in \mathbb{P}^m_K \mid \langle e_k, y\rangle = 0\}, \tag{71}$$

$$\bar{\mathrm{d}}\left(y, H_{e_k,\mathbf{0}}\right) = \frac{1}{\sqrt{-K}} \sinh^{-1}\left(\frac{2\sqrt{-K}y_k}{1 + K\left\|y\right\|^2}\right). \tag{72}$$

Equating $\bar{\mathrm{d}}\left(y, H_{e_k,\mathbf{0}}\right)$ with $u_k(x)$ from Eq. (22) gives

$$\sinh^{-1}\left(\frac{2\sqrt{-K}y_k}{1+K\left\|y\right\|^2}\right) = \sqrt{-K}u_k(x),\ \forall 1 \le k \le m. \tag{73}$$

Note that Eq. (73) takes the same form as Shimizu et al. [51, Eq. (56)], except that their responses are different. The below proof are inspired by their derivation.

Applying $\sinh(\cdot)$ on both sides of Eq. (73) yields

$$\frac{2\sqrt{-K}y_k}{1+K\left\|y\right\|^2} = \sinh\left(\sqrt{-K}u_k(x)\right). \tag{74}$$

Define $\omega_k := \frac{1}{\sqrt{-K}}\sinh\left(\sqrt{-K}u_k(x)\right)$ and $\omega = [\omega_k]_{k=1}^m$. Then, we have

$$2y_k = \left(1+K\left\|y\right\|^2\right)\omega_k, \quad \forall k, \tag{75}$$

which is equivalent to the vector identity

$$2y = \left(1+K\left\|y\right\|^2\right)\omega. \tag{76}$$

Hence, y is collinear with ω. Write $y = \lambda\omega$ with $\lambda \ge 0$. Substituting into Eq. (76) and taking norms gives a quadratic in λ:

$$K\left\|\omega\right\|^2\lambda^2 - 2\lambda + 1 = 0. \tag{77}$$

Solving and selecting the branch that satisfies $y \to 0$ as $\omega \to 0$ yields

$$\lambda = \frac{1-\sqrt{1-K\left\|\omega\right\|^2}}{K\left\|\omega\right\|^2} = \frac{1}{1+\sqrt{1-K\left\|\omega\right\|^2}}. \tag{78}$$

Therefore,

$$y = \frac{\omega}{1+\sqrt{1-K\left\|\omega\right\|^2}}, \quad \omega_k = \frac{\sinh\left(\sqrt{-K}u_k(x)\right)}{\sqrt{-K}}, \tag{79}$$

which proves the claim. One can check that $y \in \mathbb{P}_K^m$. □

F.4. Proof of Thm. 4.2

Proof. Recalling Tab. 11, a Lorentz hyperplane is

$$H_{w,p} = \{x \in \mathbb{L}_K^m \mid \langle w, x\rangle_{\mathcal{L}} = 0\},\ \text{with}\ p \in \mathbb{L}_K^m,\ w \in T_p\mathbb{L}_K^m. \tag{80}$$

The canonical origin is $\overline{\mathbf{0}} \in \mathbb{L}_K^m$. The tangent space at the origin is

$$T_{\overline{\mathbf{0}}}\mathbb{L}_K^m = \{[0, v^\top]^\top, v \in \mathbb{R}^m\}, \tag{81}$$

where each tangent vector has a zero time component. Therefore, the coordinate hyperplane through the origin and orthogonal to the k-th axis is

$$\begin{aligned} H_{\bar{e}_k,e} &= \{y \in \mathbb{L}_K^m \mid \langle \bar{e}_k, y\rangle_{\mathcal{L}} = 0\} \\ &= \{y = (y_t, y_s) \in \mathbb{L}_K^m \mid (y_s)_k = 0\}, \end{aligned} \tag{82}$$

where $\bar{e}_k = [0, e_k^\top]^\top \in T_{\overline{\mathbf{0}}}\mathbb{L}_K^m$. From Tab. 12, the associated signed point-to-hyperplane distance is

$$\begin{aligned} \bar{\mathrm{d}}\left(y, H_{\bar{e}_k,e}\right) &= \operatorname{sign}\left(\langle \bar{e}_k, y\rangle_{\mathcal{L}}\right)\mathrm{d}\left(y, H_{\bar{e}_k,e}\right) \\ &= \frac{1}{\sqrt{-K}}\sinh^{-1}\left(\sqrt{-K}(y_s)_k\right). \end{aligned} \tag{83}$$

Equating $\bar{\mathrm{d}}\left(y, H_{\bar{e}_k, e}\right)$ with $u_k(x)$ from Eq. (22) gives

$$\sinh^{-1}\left(\sqrt{-K}(y_s)_k\right) = \sqrt{-K}u_k(x), \quad 1 \le k \le m. \tag{84}$$

Applying $\sinh(\cdot)$ to both sides of Eq. (84) yields

$$(y_s)_k = \frac{1}{\sqrt{-K}} \sinh\left(\sqrt{-K}u_k(x)\right), \quad 1 \le k \le m. \tag{85}$$

Stacking the coordinates gives

$$y_s = \frac{1}{\sqrt{-K}} \sinh\left(\sqrt{-K}u(x)\right), \quad u(x) = \left(u_1(x), \ldots, u_m(x)\right)^\top. \tag{86}$$

Since $y \in \mathbb{L}_K^m$, the hyperboloid constraint $\langle y, y\rangle_{\mathcal{L}} = 1/K$ implies $-y_t^2 + \|y_s\|^2 = 1/K$. Taking the positive time component yields

$$y_t = \sqrt{\frac{1}{-K} + \|y_s\|^2}. \tag{87}$$

Combining the expressions for y_t and y_s proves the claim. □

F.5. Proof of Thm. 4.3

Proof. Set $K = -\kappa^2$ with $\kappa > 0$.

Poincaré case. Recall that

$$y = \frac{\omega}{1 + \sqrt{1 + \kappa^2 \|\omega\|^2}}, \quad \omega_k = \frac{\sinh\left(\kappa u_k(x)\right)}{\kappa}. \tag{88}$$

For any bounded scalar z,

$$\sinh(\kappa z) = \kappa z + \frac{\kappa^3 z^3}{3!} + O\left(\kappa^5\right) \Rightarrow \frac{\sinh(\kappa z)}{\kappa} = z + \frac{\kappa^2 z^3}{6} + O\left(\kappa^4\right). \tag{89}$$

Applying this to $z = u_k(x)$ yields

$$\omega_k = u_k(x) + \frac{\kappa^2}{6}u_k(x)^3 + O\left(\kappa^4\right) = u_k(x) + O\left(\kappa^2\right). \tag{90}$$

Thus, $\omega = u(x) + O\left(\kappa^2\right)$ and $\|\omega\|^2 = \|u(x)\|^2 + O\left(\kappa^2\right)$.

For the denominator,

$$\sqrt{1 + \kappa^2 \|\omega\|^2} = 1 + \frac{1}{2}\kappa^2 \|\omega\|^2 + O\left(\kappa^4\right), \tag{91}$$

which gives

$$1 + \sqrt{1 + \kappa^2 \|\omega\|^2} = 2 + \frac{1}{2}\kappa^2 \|\omega\|^2 + O\left(\kappa^4\right). \tag{92}$$

Taking the reciprocal produces

$$\frac{1}{1 + \sqrt{1 + \kappa^2 \|\omega\|^2}} = \frac{1}{2} + O\left(\kappa^2\right). \tag{93}$$

Multiplying with $\omega = u(x) + O\left(\kappa^2\right)$ gives

$$y = \frac{1}{2}u(x) + O\left(\kappa^2\right). \tag{94}$$

By Thm. 3.1, $u_k(x) \to 2\alpha_k \langle v_k, x\rangle + b_k$, hence

$$y_k \to \alpha_k \langle v_k, x\rangle + \frac{1}{2}b_k. \tag{95}$$

Lorentz case. Recall that

$$y_s = \frac{1}{\kappa}\sinh\left(\kappa u(x)\right), \quad y_t = \sqrt{\frac{1}{\kappa^2} + \|y_s\|^2}. \tag{96}$$

Using the same expansion as above,

$$y_{s,k} = \frac{1}{\kappa}\left(\kappa u_k(x) + \frac{\kappa^3}{3!}u_k(x)^3 + O\left(\kappa^5\right)\right) = u_k(x) + O\left(\kappa^2\right). \tag{97}$$

Therefore, $y_s = u(x) + O\left(\kappa^2\right)$ and $\|y_s\|^2 = \|u(x)\|^2 + O\left(\kappa^2\right)$.

Factor out κ^{-1} and expand the square root:

$$\begin{aligned} y_t &= \frac{1}{\kappa}\sqrt{1 + \kappa^2\|y_s\|^2} \\ &= \frac{1}{\kappa}\left(1 + \frac{1}{2}\kappa^2\|y_s\|^2 + O\left(\kappa^4\right)\right) \\ &= \frac{1}{\kappa} + \frac{\kappa}{2}\|y_s\|^2 + O\left(\kappa^3\right) \\ &= \frac{1}{\kappa} + O\left(\kappa\right) \to \infty. \end{aligned} \tag{98}$$

By Thm. 3.1, $u_k(x) \to \alpha_k \langle v_k, x_s\rangle + b_k$. Using the spatial expansion yields

$$(y_s)_k \to \alpha_k \langle v_k, x_s\rangle + b_k. \tag{99}$$

This completes the proof. □