The Implicit Bias of Hyperbolic Representation Learning for Multiclass Data: A Busemann Risk Perspective

Xingrun Li[1], Sho Kuno[1], Yusuke Mukuta[1,2], Xin Yang[1], Tatsuya Harada[1,2]

[1]The University of Tokyo [2]RIKEN

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

TL;DR At large radius, hyperbolic distance splits into the **radius** plus a **Busemann function** of the direction. The radius cancels in multiclass margins, so Riemannian gradient flow is governed by (i) the sign of a **radial drift coefficient** μ and (ii) a **Busemann risk** on the ideal boundary.

Implicit bias. Without explicit regularization, gradient methods still select special solutions, e.g. GD on separable data converges in direction to the max-margin classifier.

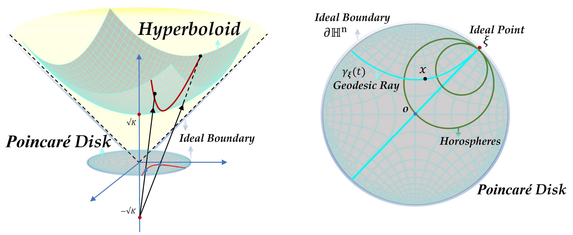
Hyperbolic learning. Negative curvature and exponential volume growth embed hierarchies with low distortion, yet the **large-radius training dynamics** of hyperbolic models remain poorly understood.

Two phenomena are reported in practice:

- **Boundary saturation**: embeddings reach large radii near the Poincaré boundary, where numerical instabilities arise.
- **Near-boundary clustering**: at large radius, classes are separated by boundary directions rather than by radial position.

When do hyperbolic representations move outward, and what governs their directions during radial return?

Hyperbolic Geometry



(a) Hyperboloid and Poincaré disk (b) Geodesic ray, ideal point, horospheres

$$\mathbb{H}^n = \{\mathbf{x} \in \mathbb{R}^{n+1} : \ \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{M}} = -1, \ x_0 > 0\}$$

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{M}} = -x_0 y_0 + \bar{\mathbf{x}}^\top \bar{\mathbf{y}}, \qquad d_{\mathbb{H}}(\mathbf{x}, \mathbf{y}) = \operatorname{arcosh}\big(-\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{M}}\big)$$

$$\mathbf{x}(r) = \big(\cosh r, \ \sinh r\, \hat{\boldsymbol{\xi}}\big), \qquad r = d_{\mathbb{H}}(\mathbf{o}, \mathbf{x}), \quad \hat{\boldsymbol{\xi}} \in \mathbb{S}^{n-1}$$

Each direction defines an **ideal point**, the limit of the geodesic ray:

$$\partial\mathbb{H}^n = \big\{\boldsymbol{\xi} = (1, \hat{\boldsymbol{\xi}}) : \ \hat{\boldsymbol{\xi}} \in \mathbb{S}^{n-1}\big\}, \qquad \mathbf{x}(r) \ \xrightarrow[r\to\infty]{} \ \boldsymbol{\xi}$$

The **Busemann function** measures the asymptotic position of $\mathbf{w}$ along the ray toward ξ; its level sets are horospheres (Fig. b):

$$b_{\boldsymbol{\xi}}(\mathbf{w}) := \lim_{t\to\infty} \big[d_{\mathbb{H}}(\mathbf{w}, \gamma_{\boldsymbol{\xi}}(t)) - t\big] = \log\big(-\langle \mathbf{w}, \boldsymbol{\xi} \rangle_{\mathcal{M}}\big)$$

Setup: Multiclass Classification

C classes with **fixed prototypes** $\mathbf{w}_1, \ldots, \mathbf{w}_C \in \mathbb{H}^n$. An encoder maps each input to $\mathbf{x}_i = (\cosh r_i, \sinh r_i\, \xi_i)$; class scores are $v_c(\mathbf{x}) = -d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}_c)$:

$$L_{\mathbf{W}}(\mathbf{X}) = \frac{1}{N} \sum_{i=1}^{N} \phi_{y_i}\Big(-d_{\mathbb{H}}(\mathbf{x}_i, \mathbf{w}_1), \ \ldots, \ -d_{\mathbb{H}}(\mathbf{x}_i, \mathbf{w}_C)\Big)$$

PERM losses (Wang & Scott, 2024) are **P**ermutation-**E**quivariant and **R**elative-**M**argin based, so a symmetric template ψ generates them:

$$\phi_y(\mathbf{v}) = \psi(\Upsilon_y D\mathbf{v}), \qquad \Upsilon_y D\mathbf{v} = \big(v_y - v_{j_1}, \ \ldots, \ v_y - v_{j_{C-1}}\big)^\top$$

(A1) ϕ is a *totally regular* PERM loss, e.g. cross-entropy, multiclass exponential, Gamma-Phi. $\psi_{\mathrm{CE}}(\mathbf{u}) = \log\Big(1 + \sum_{j=1}^{C-1} e^{-u_j}\Big)$

Radial–Busemann Expansion

Lemma 5.1 (Radial Busemann expansion). For $\mathbf{x} = (\cosh r, \sinh r\, \xi)$ and $r \geq R_0(\mathbf{w})$:

$$d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}) = \underbrace{r}_{\text{radial term}} + \underbrace{b_{\boldsymbol{\xi}}(\mathbf{w})}_{\text{direction-dependent term}} + O(e^{-2r})$$

The common radial term **cancels** in the relative margins $v_y - v_j$, leaving the **Busemann loss** and **risk** on the ideal boundary:

$$\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) := \psi\big(\Upsilon_y D\, \mathbf{b}(\boldsymbol{\xi})\big), \qquad \mathbf{b}(\boldsymbol{\xi}) := \big(-b_{\boldsymbol{\xi}}(\mathbf{w}_1), \ldots, -b_{\boldsymbol{\xi}}(\mathbf{w}_C)\big)^\top$$

$$\mathcal{R}^{\mathrm{BS}}_{\mathbf{W}}(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N) := \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}_i, y_i)$$

Proposition 5.3 (Loss expansion). Under (A1), for $r_i \geq R_0(\mathbf{W})$:

$$\ell_i(\mathbf{x}_i, y_i) = \underbrace{\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}_i, y_i)}_{\text{direction only}} + \mu(\boldsymbol{\xi}_i, y_i)\, e^{-2r_i} + O(e^{-4r_i})$$

$$\mu(\boldsymbol{\xi}, y) := -\big\langle \nabla\psi\big(\Upsilon_y D\, \mathbf{b}(\boldsymbol{\xi})\big), \ \Upsilon_y D\, \boldsymbol{\alpha}(\boldsymbol{\xi}) \big\rangle, \quad \alpha_y(\mathbf{w}_c) := \frac{\|\bar{\mathbf{w}}_c^{\perp}\|^2}{\langle \mathbf{w}_c, \boldsymbol{\xi} \rangle_{\mathcal{M}}^2}$$

μ is the **radial drift coefficient**.

$$L_{\mathbf{W}}(\mathbf{X}) = \mathcal{R}^{\mathrm{BS}}_{\mathbf{W}}(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N) + o(1) \qquad \text{as} \quad \min_i r_i \to \infty$$

At large radius, classification is governed by the **boundary direction**, not the full hyperbolic position.

Decoupled gradient flow. The metric is block-diagonal, $ds^2 = dr^2 + \sinh^2 r\, d\omega^2$, so the Riemannian flow splits into radial and angular parts:

$$\dot{r}_i = -\frac{\partial \ell_i}{\partial r_i} = 2\,\mu(\boldsymbol{\xi}_i, y_i)\, e^{-2r_i} + O(e^{-4r_i})$$

$$\dot{\boldsymbol{\xi}}_i = -\frac{1}{\sinh^2 r_i}\, \operatorname{grad}_{\mathbb{S}^{n-1}} \ell_i\big|_{\boldsymbol{\xi}_i}$$

Result I: Radial Dichotomy

Theorem 6.2. Under (A1), for any drift lower bound $\underline{\mu} > 0$ there is R_0 such that, while $r \geq R_0$:

(a) Boundary divergence

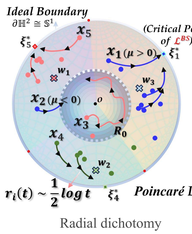
$$\mu(\boldsymbol{\xi}(t), y) \geq \underline{\mu} > 0 \text{ on } [T_1, T_2] \implies r(t) \geq \tfrac{1}{2} \log\big(e^{2r(T_1)} + 2\underline{\mu}\,(t - T_1)\big)$$

persistent positive drift: $r(t) = \frac{1}{2}\log t + O(1), \quad t \to \infty$

(b) Boundary repulsion

$$\mu(\boldsymbol{\xi}(t), y) \leq -\underline{\mu} < 0 \text{ on } [T_1, T_2] \implies r(t) \leq \tfrac{1}{2} \log\big(e^{2r(T_1)} - 2\underline{\mu}\,(t - T_1)\big)$$

first hitting time of R_0 : $T^* \leq T_1 + \dfrac{e^{2r(T_1)} - e^{2R_0}}{2\underline{\mu}} < \infty$



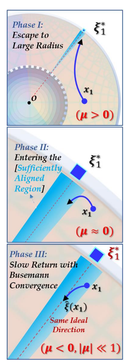
Radial dichotomy

Proof idea. For large r, $\dot{r} \approx 2\mu\, e^{-2r}$, so e^{2r} grows (or shrinks) linearly in t.

- $\mathbf{x}_1$: $\mu > 0$ drives it to the ideal boundary.
- $\mathbf{x}_2$: $\mu < 0$ pushes it back inside R_0.
- $\mathbf{x}_3$: inside R_0 the full non-convex loss governs.

The rate ½ is **universal**, independent of data geometry and margin (in Euclidean separable classification the margin sets it) ⇒ a mechanism for **boundary saturation**.

Result II: Three-Phase Dynamics



Theorem 6.4. Under (A1) and (A2: $C \geq 3$ prototypes, not all on one geodesic), ψ analytic, if $\xi(T_0)$ lies in the basin of a nondegenerate local minimum ξ^* of $L^{\mathrm{BS}}(\cdot, y)$ with $\mu > 0$ and $r(T_0) \geq 2R$, there exist $T_0 < T_1 < T_2 < T_3$:

- **Excursion**: moves outward; angular speed $O(e^{-2r})$.
- **Drift reversal**: $\mu = 0$ at T_1, $\mu < 0$ after; enters the aligned region at T_2.
- **Slow return**: $|\dot{r}| = O(e^{-2r})$ down to $r(T_3) = R$; for R large,

$$d_{\mathbb{S}^{n-1}}\big(\boldsymbol{\xi}(T_3), \boldsymbol{\xi}^*\big) \leq C' e^{-\kappa R} \leq \varepsilon$$

Why μ flips. At a Busemann critical point, noncollinearity gives

$$\mu(\boldsymbol{\xi}^*, y) = -\sum_{c \neq y} g_c \big\| \operatorname{grad}_{\mathbb{S}^{n-1}} b_{\boldsymbol{\xi}^*}(\mathbf{w}_c) - \operatorname{grad}_{\mathbb{S}^{n-1}} b_{\boldsymbol{\xi}^*}(\mathbf{w}_y) \big\|^2 < 0$$

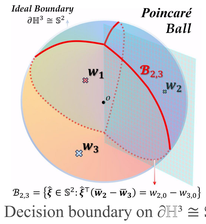
Directions approach **Busemann local minima** *even while the radius returns* ⇒ a mechanism for **near-boundary clustering**.

Result III: Escape & Classifier

Theorem 6.5. With **ideal prototypes** $\mathbf{w}_c^{\infty} \in \partial\mathbb{H}^n$ (A3), under (A1):

$$r(t) \to \infty, \qquad \boldsymbol{\xi}(t) \to \boldsymbol{\xi}^* \in \arg\min_{\boldsymbol{\xi} \in \partial\mathbb{H}^n} \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y)$$

If (A3) holds for all i, $(\xi_1^*, \ldots, \xi_N^*)$ is a **global minimizer** of the Busemann risk.



Decision boundary on $\partial\mathbb{H}^3 \cong \mathbb{S}^2$

Theorem 6.6. For large r, the classifier depends only on the direction:

it picks the prototype with the **smallest Busemann value**.

Decision boundaries are spherical sections (great subspheres for ideal prototypes).

$$\hat{y}(\mathbf{x}) = \arg\min_{y \in [C]} \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) = \arg\min_{c \in [C]} \ b_{\boldsymbol{\xi}}(\mathbf{w}_c)$$

$$\mathcal{B}_{c,c'} = \big\{\hat{\boldsymbol{\xi}} \in \mathbb{S}^{n-1} : \ \hat{\boldsymbol{\xi}}^\top(\bar{\mathbf{w}}_c - \bar{\mathbf{w}}_{c'}) = w_{c,0} - w_{c',0}, \ b_{\boldsymbol{\xi}}(\mathbf{w}_c) \leq b_{\boldsymbol{\xi}}(\mathbf{w}_k) \ \ \forall k \in [C]\big\}$$

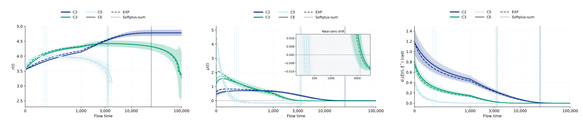
Busemann margin and Corollary 6.8:

$$m_i(\boldsymbol{\xi}) := \min_{c \neq y_i} \big[b_{\boldsymbol{\xi}}(\mathbf{w}_c) - b_{\boldsymbol{\xi}}(\mathbf{w}_{y_i})\big]$$

cross-entropy: $\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y_i) < \log 2 \implies m_i(\boldsymbol{\xi}) > 0$

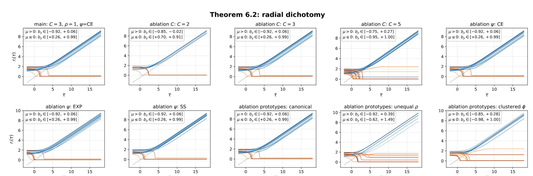
Experiments

Coupled gradient flow on $\mathbb{H}^2$: $C = 3$ prototypes at radius $\rho = 1$, $N = 15$ embeddings; ablations $C \in \{2, 3, 5\}$, $\psi \in \{\mathrm{CE}, \mathrm{EXP}, \mathrm{Softplus\text{-}sum}\}$.



(a) Radial evolution $r(t)$. (b) Drift coefficient $\mu(\boldsymbol{\xi}(t), y)$. (c) Angular $d_{\mathbb{S}^{n-1}}(\boldsymbol{\xi}(t), \boldsymbol{\xi}^*)$.

Outward excursion → drift reversal → slow return, while $\xi(t) \to \xi^*$.



Fixed-direction flow: $\mu > 0$ follows $r \approx \frac{1}{2}\log t$; $\mu \leq 0$ moves inward.

Neural networks. Shared-encoder stationary points of the exact loss vs. the Busemann risk (70 checkpoints): the angular gap $D_R = O(e^{-2R})$.

Prototypes	Checkpoints	Slope of log D_R	95% CI
Canonical	Converged	−2.0020	[−2.0076, −1.9964]
Irregular	Converged	−2.0000	[−2.0003, −1.9997]