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

Abstract

We study the implicit bias of Riemannian gradient flow for hyperbolic multiclass classification with fixed class prototypes in hyperbolic space $\mathbb{H}^n$. Our framework accommodates general *permutation invariant relative margin (PERM)* losses [1], a class that includes cross entropy and other standard multiclass losses. Our analysis is based on a decomposition: at large radius, the distance to each prototype splits into a radial term and a direction dependent term described by the Busemann function. This yields two main results. First, we prove a radial dichotomy: the sign of a drift coefficient μ determines whether the radius is pushed toward the ideal boundary or back toward the interior; if the positive drift persists, then $r(t) = \frac{1}{2}\log t + O(1)$, while persistent negative drift returns the trajectory to the large radius threshold in finite time. Second, we show that large-radius directions can approach critical points of the Busemann risk on $\partial\mathbb{H}^n$, even during radial return. These results provide a rigorous large-radius perspective on two phenomena we refer to as *boundary saturation* and *near-boundary clustering* in hyperbolic representation learning.

1 Introduction

Over the past decade, deep learning has achieved remarkable success across diverse domains, from computer vision and natural language processing to scientific discovery [2–4]. However, a central mystery underlying this success has captivated the research community: modern neural networks are typically overparameterized, and even in the absence of explicit regularization terms in the training objective, gradient based optimization nevertheless tends to find solutions that generalize well [5, 6]. This phenomenon is known as **implicit regularization** (or implicit bias) [7]. This perspective has led to a more precise understanding of several models. For example, in Euclidean settings, gradient descent on separable data with cross-entropy loss converges directionally to max margin solutions [8], while matrix factorized models implicitly favor low rank structures [9, 10].

In parallel, **hyperbolic representation learning** has become an important tool for modeling hierarchical, tree like, and long tailed data. Due to its negative curvature and exponential volume growth, hyperbolic space can represent latent hierarchical structures with substantially lower distortion than Euclidean space [11–13]. This geometric advantage has motivated a broad family of hyperbolic models, ranging from Poincaré embeddings to hyperbolic neural networks, graph neural networks, and Transformers [14–17], with applications across vision, graph learning, recommendation, reinforcement learning, and beyond [18–22]. However, despite rapid progress in model design and empirical applications, the large radius training dynamics of hyperbolic models remain poorly understood.

Understanding implicit bias in hyperbolic space is not a direct extension of Euclidean theory. Unlike Euclidean models, whose analyses of implicit bias often rely on norm direction decompositions and max margin characterizations [8, 23, 24], hyperbolic models are governed by Riemannian gradient dynamics on a noncompact curved manifold [25, 26]. This geometry permits representations to reach large radii while their angular directions become governed by the ideal boundary, even when

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

radial motion is ultimately bounded. Moreover, hyperbolic architectures commonly rely on geodesic distances, exponential and logarithmic maps, Möbius operations, and other nonlinear geometric constructions [14, 27], making their losses fundamentally different from convex surrogates composed with linear predictors. As a result, standard Euclidean tools do not directly apply.

Beyond the theoretical challenge, this gap leaves concrete empirical phenomena unexplained. For example, hyperbolic classifiers are often associated with two behaviors in the large-radius regime: ***boundary saturation***, where embeddings reach large hyperbolic radii near the Poincaré ball boundary and numerical instabilities may arise [28, 29]; and ***near-boundary clustering***, where class discrimination at large radii depends less on radial positions and more on boundary directions [30, 31].

Motivated by these observations, we study the large radius dynamics of hyperbolic representations under Riemannian gradient flow for multiclass classification. Our central questions are: *when do hyperbolic representations undergo outward excursions, and what governs their angular directions during radial return?* Our analysis starts from a simple geometric observation: along a large radius ray in $\mathbb{H}^n$, the distance to any fixed prototype decomposes into a common radial term and a direction dependent Busemann term. The common radial term cancels in relative margin based multiclass losses, or *PERM losses* [1], leaving the leading order classification behavior governed by Busemann functions on the ideal boundary $\partial\mathbb{H}^n$. This decomposition yields a clean asymptotic picture. Persistent positive drift conditionally yields the radial law $r(t) = \frac{1}{2}\log t + O(1)$, while angular alignment can reverse outward motion and induce a slow return toward the interior. In the large radius regime, the angular dynamics are governed approximately by a *Busemann loss* on the ideal boundary. Thus, at large radius, classification is governed primarily by the boundary direction rather than the full finite radius hyperbolic position.

Contributions. To the best of our knowledge, we provide the first rigorous asymptotic analysis of Riemannian gradient flow for hyperbolic multiclass classification under general PERM losses [1], including cross-entropy, multiclass exponential, and Gamma-Phi losses. Our main contributions are:

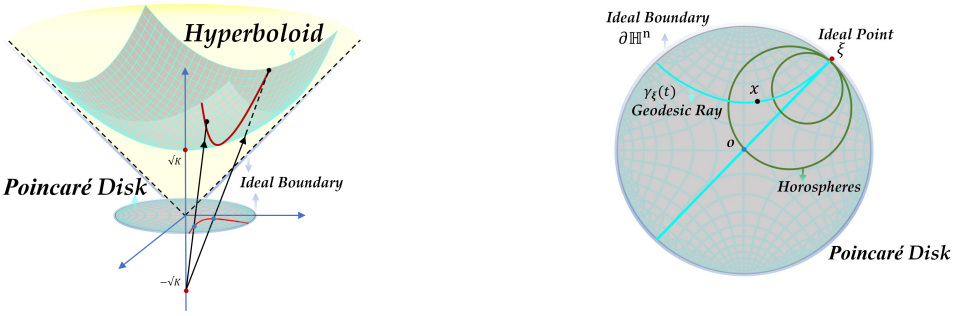
- We develop a Busemann asymptotic framework for hyperbolic multiclass classification with PERM losses, showing that, at large radius, each sample loss is governed to leading order by a Busemann loss on the ideal boundary.
- We establish a radial dichotomy governed by the drift coefficient μ: persistent positive drift conditionally yields $r(t) = \frac{1}{2}\log t + O(1)$, while angular alignment can reverse outward motion and induce slow radial return, suggesting a possible mechanism for boundary saturation.
- We show escaping trajectories with ideal prototypes converge to global minimizers of Busemann risk, while finite prototype trajectories approach local minima arbitrarily closely during three-phase return, suggesting a possible mechanism for near-boundary clustering.
- We characterize the asymptotic classifier and its decision boundaries on the ideal boundary.

2 Related Work

Due to the limited space in the main text, we provide a brief overview of related literature here and defer a more detailed discussion to Appendix [A.1]. For **implicit bias**: Gradient based optimization has been shown to select max margin solutions in separable linear classification [8], with extensions to nonseparable data [32], homogeneous networks [23], deep directional convergence [24], optimization geometry dependent biases [33], and low rank biases in matrix factorization [9]. For **multiclass classification**, the PERM framework extends implicit bias theory beyond binary losses [34], while recent work also studies algorithm dependent margin biases in homogeneous networks [35]. These Euclidean results characterize margin or complexity biases, but not radial drift or directional alignment in curved noncompact spaces. For **hyperbolic representation learning**: Hyperbolic learning has developed a broad range of models, including Poincaré embeddings [12], hyperbolic neural networks [14], hyperbolic neural network variants [20], hyperbolic graph neural networks [15], and hyperbolic Transformers [16]; see [17] for a survey. Empirical studies have also reported boundary related phenomena, including numerical instability near the Poincaré boundary [28], the effectiveness of clipping [29], common radius concentration [30], and metric learning induced class separation [31]. These works highlight hyperbolic geometry, but focus mainly on model design and empirical phenomena rather than the large radius dynamics governing boundary approach and return.

3 Hyperbolic Space

In this section, we provide a brief review of the essential concepts and a theorem about hyperbolic geometry that underlie our proposed method.



(a) Hyperboloid model and Poincaré Disk

(b) Geodesic rays, ideal boundary, and horospheres

Figure 1: Geometric illustration of hyperbolic representations and their boundary asymptotics.

The *Minkowski inner product*, defined for vectors $\mathbf{x} = (x_0, x_1, \ldots, x_n) = (x_0, \bar{\mathbf{x}}), \mathbf{y} = (y_0, y_1, \ldots, y_n) = (y_0, \bar{\mathbf{y}}) \in \mathbb{R} \times \mathbb{R}^n$, is given by: $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{M}} = -x_0 y_0 + \bar{\mathbf{x}}^\top \bar{\mathbf{y}} = \mathbf{x}^\top \boldsymbol{\eta} \mathbf{y}$ where the *Minkowski metric tensor* of hyperboloid model is $\boldsymbol{\eta} = \mathrm{diag}(-1, 1, \ldots, 1)$. The n-dimensional *hyperbolic space* (hyperboloid model) is defined as:

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

where $K > 0$ is the *Minkowski constant*, which determines the constant negative *curvature* $\kappa = -1/K$. The *hyperbolic distance* of $\mathbf{x},\ \mathbf{y} \in \mathbb{H}^{n,K}$ is given by

$$d_{\mathbb{H}}^K(\mathbf{x}, \mathbf{y}) = \sqrt{K}\, \mathrm{arcosh}\left(-\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{M}} / K\right). \tag{2}$$

The *north pole* is defined as: $\mathbf{o} = (\sqrt{K}, 0, \ldots, 0)$. Every $\mathbf{x} = (x_0, \bar{\mathbf{x}}) \in \mathbb{H}^{n,K} \setminus \{\mathbf{o}\}$ admits a unique polar decomposition of the form

$$\mathbf{x}(r) = \left(\sqrt{K} \cosh \frac{r}{\sqrt{K}}, \sqrt{K} \sinh \frac{r}{\sqrt{K}} \hat{\boldsymbol{\xi}}\right), \tag{3}$$

where radius $r = d_{\mathbb{H}}^K(\mathbf{o}, \mathbf{x})$ and direction $\hat{\boldsymbol{\xi}} = \bar{\mathbf{x}}/\|\bar{\mathbf{x}}\| \in \mathbb{S}^{n-1}$. The *ideal boundary* is defined as the set $\partial\mathbb{H}^{n,K} = \{\boldsymbol{\xi} = (1, \hat{\boldsymbol{\xi}}) : \hat{\boldsymbol{\xi}} \in \mathbb{S}^{n-1}\}$, where each element $\boldsymbol{\xi}$ represents an equivalence class of geodesic rays. **As $r \to \infty$, the ray $\mathbf{x}(r)$ in (3) converges to the *ideal point* $\boldsymbol{\xi} = (1, \hat{\boldsymbol{\xi}})$ along the direction $\hat{\boldsymbol{\xi}}$.** For an ideal point $\boldsymbol{\xi} = (1, \hat{\boldsymbol{\xi}}) \in \partial\mathbb{H}^{n,K}$ and $\mathbf{w} \in \mathbb{H}^{n,K}$, the *Busemann function* is defined as:

$$b_{\boldsymbol{\xi}}^K(\mathbf{w}) := \lim_{t \to \infty} \left[d_{\mathbb{H}}^K(\mathbf{w}, \gamma_{\boldsymbol{\xi}}(t)) - t\right] = \sqrt{K}\, \log\!\left(\frac{-\langle \mathbf{w}, \boldsymbol{\xi} \rangle_{\mathcal{M}}}{\sqrt{K}}\right), \tag{4}$$

where $\gamma_{\boldsymbol{\xi}}(t) = \left(\sqrt{K} \cosh \frac{t}{\sqrt{K}},\ \sqrt{K} \sinh \frac{t}{\sqrt{K}}\, \hat{\boldsymbol{\xi}}\right)$ is the unit-speed geodesic ray toward $\boldsymbol{\xi}$.

Intuitively, $b_{\boldsymbol{\xi}}^K(\mathbf{w})$ measures the asymptotic position of $\mathbf{w}$ along the direction toward the ideal point $\boldsymbol{\xi}$. The level sets $\{\mathbf{w} : b_{\boldsymbol{\xi}}^K(\mathbf{w}) = c\}$ are *horospheres* centred at $\boldsymbol{\xi}$.

Clarification. In fact, the hyperboloid model and the Poincaré ball (disk) model are equivalent representations of the same hyperbolic space [36], while the gyrovector formalism provides an algebraic language for operations on the hyperbolic space [37]. Since the quantities studied in this paper are intrinsic geometric objects, their asymptotic behavior is independent of the chosen representation. We therefore carry out the analysis in the hyperboloid model, where these quantities admit the simpler closed form expressions. And we present visualizations in the Poincaré ball model for better geometric intuition. **Also, without loss of generality, we set Minkowski constant $K = 1$ throughout the main development to simplify the presentation, and thus omit K from the notation.** The general case follows by a straightforward rescaling argument, see Appendix [A.2].

4 Hyperbolic Multi-classification Loss

4.1 Classification Setup

We consider a C-class classification problem in $\mathbb{H}^n$ with $C \geq 2$. For each class $c \in [C] := \{1, \ldots, C\}$, let $\mathbf{w}_c \in \mathbb{H}^n$ denote its *prototype*, a fixed representative point of class c in the embedding space. We are given N training samples $\{(\mathbf{z}_i, y_i)\}_{i=1}^N$, where $\mathbf{z}_i$ is the input and $y_i \in [C]$ is its class label. An encoder f_θ maps each input $\mathbf{z}_i$ to a point $\mathbf{x}_i = f_\theta(\mathbf{z}_i) \in \mathbb{H}^n$, which we write in polar form as $\mathbf{x}_i = (\cosh r_i,\ \sinh r_i\, \hat{\boldsymbol{\xi}}_i)$, and we omit parameters θ throughout the paper for brevity.

We define class score functions $v_c(\mathbf{x}),\ c \in [C]$, producing a score vector $\mathbf{v}(\mathbf{x}) = (v_1(\mathbf{x}), \ldots, v_C(\mathbf{x}))^\top : \mathbb{H}^n \to \mathbb{R}^C$. Let $\boldsymbol{\phi}(\mathbf{v}) = (\phi_1(\mathbf{v}), \ldots, \phi_C(\mathbf{v}))^\top : \mathbb{R}^C \to \mathbb{R}^C$ be a vector-valued multiclass loss function, where $\phi_c(\mathbf{v})$ denotes the c-th component of $\boldsymbol{\phi}(\mathbf{v})$, corresponding to the loss when the true class is c. The loss for sample $(\mathbf{x}_i, y_i)$ takes the *class-score form*:

$$\ell(\mathbf{x}_i, y_i) = \phi_{y_i}\big(\mathbf{v}(\mathbf{x}_i)\big) = \phi_{y_i}\big(v_1(\mathbf{x}_i), \ldots, v_C(\mathbf{x}_i)\big). \tag{5}$$

The total loss is $L(\mathbf{X}) = \frac{1}{N}\sum_{i=1}^N \ell(\mathbf{x}_i, y_i)$, where $\mathbf{X} := (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ collects all training embeddings. The choice of class score function $v_c(\mathbf{x})$ determines how geometric information is used for classification. We adopt the distance to prototype score, a common and geometrically direct choice based on proximity in the intrinsic hyperbolic metric. Other hyperbolic scoring rules, such as hyperplane based scores, hypersphere based scores, and Lorentzian inner products, are also closely related to the underlying hyperbolic distance geometry. Thus, throughout this work, we set $v_c(\mathbf{x}) = -d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}_c) = -\operatorname{arcosh}\big(-\langle \mathbf{x}, \mathbf{w}_c\rangle_{\mathcal{M}}\big)$. We define $\mathbf{W} := [\mathbf{w}_1, \ldots, \mathbf{w}_C]$ to collect the prototypes column wise. The resulting *hyperbolic multiclass classification objective* is:

$$L_{\mathbf{W}}(\mathbf{X}) = \frac{1}{N}\sum_{i=1}^N \phi_{y_i}\Big(-d_{\mathbb{H}}(\mathbf{x}_i, \mathbf{w}_1),\ \ldots,\ -d_{\mathbb{H}}(\mathbf{x}_i, \mathbf{w}_C)\Big). \tag{6}$$

4.2 PERM Losses Frame

We employ the PERM framework of Wang and Scott [1]. We first recall some central concepts. Let $\mathfrak{S}_C$ denote the symmetric group on $[C]$, which is the group of all $C!$ permutations of $\{1, \ldots, C\}$. For $\sigma \in \mathfrak{S}_C$, the associated permutation matrix $S_\sigma \in \{0,1\}^{C\times C}$ has entries $(S_\sigma)_{ij} = \mathbf{1}[j = \sigma(i)]$, so that $(S_\sigma \mathbf{v})_i = v_{\sigma(i)}$. The matrix $D = [I_{C-1} \mid -\mathbf{1}_{C-1}] \in \mathbb{R}^{(C-1)\times C}$ extracts differences with respect to the last component: $(D\mathbf{v})_k = v_k - v_C,\ k = 1, \ldots, C{-}1$. For each class $y \in [C]$, the matrix label code $\Upsilon_y \in \{0, \pm 1\}^{(C-1)\times(C-1)}$ is the unique invertible matrix satisfying

$$\Upsilon_y D\mathbf{v} = \big(v_y - v_{j_1},\ \ldots,\ v_y - v_{j_{C-1}}\big)^\top, \quad \{j_1, \ldots, j_{C-1}\} = [C] \setminus \{y\}. \tag{7}$$

That is, Υ_y re-expresses the base differences $\{v_k - v_C\}$ as the class y margins $\{v_y - v_j\}_{j\neq y}$.

Definition 4.1 (PERM loss)**.** A multiclass loss $\boldsymbol{\phi}(\mathbf{v}) = (\phi_1(\mathbf{v}), \ldots, \phi_C(\mathbf{v}))^\top : \mathbb{R}^C \to \mathbb{R}^C$ is a *PERM loss* if it is: (1) *Permutation Equivariant*: $\boldsymbol{\phi}(S_\sigma \mathbf{v}) = S_\sigma\, \boldsymbol{\phi}(\mathbf{v})$ for all $\sigma \in \mathfrak{S}_C$; and (2) *Relative Margin based*: $\phi_y(\mathbf{v})$ depends on $\mathbf{v}$ only through the margins $\{v_y - v_j\}_{j\neq y}$.

By Theorem 2.5 in [1], a PERM loss is uniquely characterised by a symmetric function $\psi : \mathbb{R}^{C-1} \to \mathbb{R}$, called its *template function*, via

$$\phi_y(\mathbf{v}) = \psi(\Upsilon_y D\mathbf{v}). \tag{8}$$

Because ψ is symmetric, the particular ordering of the margins in (7) does not affect ϕ_y.

Definition 4.2 (Totally regular PERM loss)**.** Let $\psi : \mathbb{R}^{C-1} \to \mathbb{R}$ be a PERM template. For each $n \in \{2, \ldots, C-1\}$, define its n-ary truncation $\psi^{(n)} : \mathbb{R}^n \to \mathbb{R}$ by restricting ψ to any n coordinates. The loss is *totally regular* if, for all such n, $\psi^{(n)}$ is nonnegative, twice differentiable, strictly convex, semi-coercive, and satisfies $\nabla\psi^{(n)} \prec 0$ everywhere.

Then we introduce our first assumption:

(A1) ***The multiclass loss function $\boldsymbol{\phi}$ is a totally regular PERM loss with template function ψ.***

(A1) are satisfied by cross-entropy ($\psi(\mathbf{u}) = \log\big(1 + \sum_{j=1}^{C-1} e^{-u_j}\big)$), multiclass exponential loss, and all Gamma-Phi losses with exponentially decaying ϕ (Proposition 4.6 in [1]).

5 Asymptotic Expansion and Busemann Risk

5.1 The Radial Busemann Expansion

The cornerstone of our analysis is the following large radius decomposition of hyperbolic distance into a radial term and a direction dependent Busemann term.

Lemma 5.1 (Radial Busemann expansion)**.** *For* $\mathbf{x} = (\cosh r, \sinh r\,\hat{\boldsymbol{\xi}}) \in \mathbb{H}^n \setminus \{\mathbf{o}\}$ *and* $\mathbf{w} = (w_0, w_1, \ldots, w_n) = (w_0, \bar{\boldsymbol{w}}) \in \mathbb{H}^n$*, let* $\boldsymbol{\xi} = (1, \hat{\boldsymbol{\xi}}) \in \partial\mathbb{H}^n$*. There exists a constant* $R_0(\mathbf{w}) > 0$ *such that for all* $r \geq R_0$*,*

$$d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}) = r + b_{\boldsymbol{\xi}}(\mathbf{w}) + O(e^{-2r}), \tag{9}$$

where $b_{\boldsymbol{\xi}}(\mathbf{w}) = \log(w_0 - \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}})$ *is the Busemann function.*

Proof sketch. Using the hyperboloid distance formula and the expansions of $\cosh r$ and $\sinh r$, we obtain $-\langle \mathbf{x}, \mathbf{w} \rangle_{\mathcal{M}} = \frac{e^r}{2}(w_0 - \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}})\big(1 + O(e^{-2r})\big)$. Since $w_0 - \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}} > 0$, applying $\operatorname{arcosh}(u) = \log(2u) + O(u^{-2})$ gives $d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}) = r + \log(w_0 - \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}}) + O(e^{-2r})$, which is exactly the claimed expansion. The full and detailed proof is in Appendix [B.1]. □

This decomposition allows us to expand the nonconvex loss ℓ and separate the radial and angular mechanisms behind the later convergence analysis.

5.2 Loss Decomposition

Using Lemma 5.1, we derive an asymptotic expansion of the loss for each sample in (5), consisting of a leading Busemann term and an exponentially small correction. Recall from Assumption (A1) that $\boldsymbol{\phi}$ is a totally regular PERM loss with template ψ. Hence $\ell_i(\mathbf{x}_i, y_i) = \psi(\Upsilon_{y_i} D\mathbf{v}(\mathbf{x}_i))$, where $\mathbf{v}(\mathbf{x}_i) = (v_1(\mathbf{x}_i), \ldots, v_C(\mathbf{x}_i)) = (-d_{\mathbb{H}}(\mathbf{x}_i, \mathbf{w}_1), \ldots, -d_{\mathbb{H}}(\mathbf{x}_i, \mathbf{w}_C))$.

Definition 5.2 (Busemann loss, risk and correction coefficient)**.** For a direction $\boldsymbol{\xi} = (1, \hat{\boldsymbol{\xi}}) \in \partial\mathbb{H}^n$ with label $y \in [C]$, and prototype $\{\mathbf{w}_c\}_{c=1}^C$, define:

$$\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) \;:=\; \psi\big(\Upsilon_y D\,\mathbf{b}(\boldsymbol{\xi})\big), \qquad \text{where } \mathbf{b}(\boldsymbol{\xi}) \;:=\; (-b_{\boldsymbol{\xi}}(\mathbf{w}_1), \ldots, -b_{\boldsymbol{\xi}}(\mathbf{w}_C))^\top, \tag{10}$$

$$\boldsymbol{\alpha}(\boldsymbol{\xi}) \;:=\; \big(\alpha_{\boldsymbol{\xi}}(\mathbf{w}_1), \ldots, \alpha_{\boldsymbol{\xi}}(\mathbf{w}_C)\big)^\top, \qquad \text{where } \alpha_{\boldsymbol{\xi}}(\mathbf{w}_c) \;:=\; \frac{\|\bar{\boldsymbol{w}}_c^{\perp}\|^2}{\langle \mathbf{w}_c, \boldsymbol{\xi} \rangle_{\mathcal{M}}^2}, \tag{11}$$

here $\bar{\boldsymbol{w}}_c^{\perp} := \bar{\boldsymbol{w}}_c - (\hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}}_c)\,\hat{\boldsymbol{\xi}}$ is the component of $\bar{\boldsymbol{w}}_c$ orthogonal to the boundary direction $\hat{\boldsymbol{\xi}}$.

We call $\mathcal{L}^{\mathrm{BS}} : \partial\mathbb{H}^n \times [C] \to \mathbb{R}$ the *Busemann loss* function, and $\boldsymbol{\alpha} : \partial\mathbb{H}^n \to \mathbb{R}^C$ the *correction coefficient*. Moreover, the *Busemann risk* is

$$\mathcal{R}^{\mathrm{BS}}_{\mathbf{W}}(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N) \;:=\; \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}_i, y_i) \tag{12}$$

Proposition 5.3 (Loss expansion)**.** *Under Assumption (A1), there exists a constant* $R_0(\mathbf{W}) > 0$ *such that for each* $i \in [N]$ *and all* $r_i \geq R_0$*,*

$$\ell_i(\mathbf{x}_i, y_i) \;=\; \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}_i, y_i) \;+\; \mu(\boldsymbol{\xi}_i, y_i)\, e^{-2r_i} \;+\; O(e^{-4r_i}), \tag{13}$$

where

$$\mu(\boldsymbol{\xi}, y) \;:=\; -\big\langle \nabla\psi\big(\Upsilon_y D\,\mathbf{b}(\boldsymbol{\xi})\big),\; \Upsilon_y D\,\boldsymbol{\alpha}(\boldsymbol{\xi}) \big\rangle, \tag{14}$$

and we call $\mu : \partial\mathbb{H}^n \times [C] \to \mathbb{R}$ *the radial drift coefficient.*

Proof sketch. By extending Lemma 5.1, $v_c(\mathbf{x}_i) = -d_{\mathbb{H}}(\mathbf{x}_i, \mathbf{w}_c) = -r_i - b_{\boldsymbol{\xi}_i}(\mathbf{w}_c) - \alpha_{\boldsymbol{\xi}_i}(\mathbf{w}_c)\, e^{-2r_i} + O(e^{-4r_i})$. Since ℓ_i is a PERM loss by Assumption (A1), it depends on $\mathbf{v}$ only through the relative margin vector $\Upsilon_{y_i} D\mathbf{v}$. The common radial term $-r_i$ is annihilated by D, so:

$$\Upsilon_{y_i} D\,\mathbf{v}(\mathbf{x}_i) \;=\; \Upsilon_{y_i} D\,\mathbf{b}(\boldsymbol{\xi}_i) \;-\; \Upsilon_{y_i} D\,\boldsymbol{\alpha}(\boldsymbol{\xi}_i)\, e^{-2r_i} \;+\; O(e^{-4r_i}). \tag{15}$$

Then applying Taylor expansion of ψ around $\Upsilon_{y_i} D\mathbf{b}(\boldsymbol{\xi}_i)$ to the relative margin expansion in (15), and using the definition of $\mu(\boldsymbol{\xi}_i, y_i)$ in (14), yields the loss expansion (13).

The full and detailed proof is in Appendix [B.2]. □

The common radial term cancels in the relative margin vector $\Upsilon_{y_i} D\,\mathbf{v}$ by the relative margin structure in Assumption (A1). Therefore, the leading order loss depends only on the boundary direction $\boldsymbol{\xi}_i$, not on the radius r_i. This yields the radial angular decoupling in leading order, which drives our asymptotic analysis. Summing over i, the empirical loss admits $L_{\mathbf{W}}(\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N}\left[\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}_i, y_i) + \mu(\boldsymbol{\xi}_i, y_i)e^{-2r_i} + O(e^{-4r_i})\right]$. In particular, as $\min_i r_i \to \infty$,

$$L_{\mathbf{W}}(\mathbf{X}) = \mathcal{R}^{\mathrm{BS}}_{\mathbf{W}}(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N) + o(1). \tag{16}$$

This decoupling motivates our dynamical analysis below, where we study how the radial and angular components of the embeddings evolve under gradient flow. Throughout the analysis, we fix the class prototypes $\{\mathbf{w}_c\}_{c=1}^{C}$ and study the gradient flow dynamics of the embeddings $\{\mathbf{x}_i\}_{i=1}^{N}$ directly on $\mathbb{H}^n$. The setting of fixed prototypes isolates the geometric mechanism behind boundary divergence in distance based hyperbolic classifiers, is natural in view of the symmetry of hyperbolic distance, and is consistent with methods using fixed prototypes in prototypical learning [38–40]. It also parallels classical implicit bias analyses of linear classification, while shifting the object of study from classifier weights to the radial and angular dynamics of hyperbolic embeddings.

6 Radial and Angular Behavior

We study the Riemannian gradient flow on $(\mathbb{H}^n)^N$:

$$\dot{\mathbf{x}}_i(t) \;=\; -\operatorname{grad}_{\mathbb{H}^n}\, \ell_i(\mathbf{x}_i(t), y_i), \qquad i = 1, \ldots, N, \tag{17}$$

with initial conditions $\mathbf{x}_i(0) \in \mathbb{H}^n$, where for any smooth function f on a Riemannian manifold (M, g) with metric matrix G, the Riemannian gradient is defined as $\operatorname{grad}_M f = G^{-1}\nabla f$, with ∇f denoting the vector of partial derivatives in local coordinates. This is the continuous time analogue of Riemannian gradient descent $\mathbf{x} \leftarrow \exp_{\mathbf{x}}(-\eta\, \operatorname{grad}_M f)$. Recall that each point on the hyperboloid is $\mathbf{x}_i = (\cosh r_i,\, \sinh r_i\, \hat{\boldsymbol{\xi}}_i)$ with $r_i > 0$ and $\hat{\boldsymbol{\xi}}_i \in \mathbb{S}^{n-1}$. In these polar coordinates $(r_i, \hat{\boldsymbol{\xi}}_i)$, the induced metric takes the block diagonal form $ds^2 = dr^2 + \sinh^2(r)\, d\omega^2$, where $d\omega^2$ is the round metric on $\mathbb{S}^{n-1}$. Since the metric is block diagonal between r and $\hat{\boldsymbol{\xi}}$, the gradient flow decouples into a radial and an angular equation:

$$\dot{r}_i \;=\; -\frac{\partial \ell_i}{\partial r_i}, \qquad \dot{\hat{\boldsymbol{\xi}}}_i \;=\; -\frac{1}{\sinh^2(r_i)}\, \operatorname{grad}_{\mathbb{S}^{n-1}}\, \ell_i\big|_{\hat{\boldsymbol{\xi}}_i}, \tag{18}$$

where $\operatorname{grad}_{\mathbb{S}^{n-1}} \ell_i \;=\; \nabla_{\hat{\boldsymbol{\xi}}}\ell_i - (\nabla_{\hat{\boldsymbol{\xi}}}\ell_i \cdot \hat{\boldsymbol{\xi}}_i)\,\hat{\boldsymbol{\xi}}_i$ is the Euclidean gradient with respect to $\hat{\boldsymbol{\xi}}_i \in \mathbb{R}^n$ projected onto the tangent plane of the sphere at $\hat{\boldsymbol{\xi}}_i$. We first analyze a single trajectory that fix $i \in [N]$ and suppress the subscript i. The extension to all samples follows by applying the same argument independently to each i.

6.1 Radial Dichotomy

Lemma 6.1 (Radial gradient)**.** *Under Assumptions (A1), there exists a constant $R_0(\mathbf{W}) > 0$ such that for all $r \geq R_0$,*

$$\frac{\partial \ell}{\partial r} \;=\; -2\mu(\boldsymbol{\xi}, y)\, e^{-2r} \;+\; O(e^{-4r}), \tag{19}$$

where μ is the radial drift coefficient defined in (14)*.*

Proof. By Lemma 5.1, the expansion in Proposition 5.3 holds with a C^1 remainder ρ satisfying $\rho = O(e^{-4r}), \partial_r\rho = O(e^{-4r})$ uniformly. Differentiating $\ell = \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) + \mu(\boldsymbol{\xi}, y)e^{-2r} + \rho$ in r, and noting that $\mathcal{L}^{\mathrm{BS}}$ and μ are r-independent, yields $\partial_r\ell = -2\mu(\boldsymbol{\xi}, y)e^{-2r} + O(e^{-4r})$. □

Theorem 6.2 (Radial dichotomy)**.** *Under Assumption (A1), for any drift lower bound $\underline{\mu} > 0$, there exists $R_0 = R_0(\mathbf{W}, \underline{\mu}) > 0$ such that:*

(a) ***Boundary divergence.*** *If there exist $T_2 > T_1 \geq 0$ such that $r(T_1) \geq R_0$ and $\mu(\boldsymbol{\xi}(t), y) \geq \underline{\mu}$ on the time interval $I = [T_1, T_2]$, then $r(t) \geq \frac{1}{2}\log\big(e^{2r(T_1)} + 2\underline{\mu}(t - T_1)\big)$ for all $t \in I$. Moreover, if $r(T_1) \geq R_0$ and $\mu(\boldsymbol{\xi}(t), y) \geq \underline{\mu}$ for all $t \geq T_1$, then*

$$r(t) = \tfrac{1}{2}\log t + O(1), \qquad t \to \infty. \tag{20}$$

(b) ***Boundary repulsion.*** *If there exist $T_2 > T_1 \geq 0$ such that $\mu(\boldsymbol{\xi}(t), y) \leq -\underline{\mu}$ and $r(t) \geq R_0$ for all $t \in I = [T_1, T_2]$, then $r(t) \leq \frac{1}{2}\log\big(e^{2r(T_1)} - 2\underline{\mu}(t - T_1)\big)$ for all $t \in I$. Moreover, if $r(T_1) \geq R_0$ and $\mu(\boldsymbol{\xi}(t), y) \leq -\underline{\mu}$ whenever $r(t) \geq R_0$, then the first hitting time $T^* := \inf\{t \geq T_1 : r(t) \leq R_0\}$ satisfies*

$$T^* \leq T_1 + \frac{e^{2r(T_1)} - e^{2R_0}}{2\underline{\mu}} < \infty. \tag{21}$$

Proof sketch. Firstly, by Proposition 5.3, differentiating the loss expansion with respect to r in geodesic polar coordinates gives $\dot{r} = 2\mu(t)\, e^{-2r} + O(e^{-4r})$. For $r \geq R_0$ sufficiently large, the remainder is absorbed, yielding $\underline{\mu}\, e^{-2r} \leq \dot{r} \leq 3\bar{\mu}\, e^{-2r}$ when $\mu \geq \underline{\mu} > 0$. (i) Part (a): From the radial ODE, for $r \geq R_0$ large enough, $\dot{r} \geq \underline{\mu}\, e^{-2r}$. Multiplying both sides by e^{2r} and recognizing $\frac{d}{dt}[e^{2r}] = 2e^{2r}\dot{r}$, we obtain $\frac{d}{dt}[e^{2r}] \geq 2\underline{\mu}$. Integrating from T_1 to t gives $e^{2r(t)} \geq 2\underline{\mu}(t - T_1) + e^{2r(T_1)}$, hence $r(t) \geq \frac{1}{2}\log t + O(1)$ if the condition holds for all $t \geq T_1$. The matching upper bound follows from $\dot{r} \leq 3\bar{\mu}\, e^{-2r}$ by the same argument.

(ii) Part (b): When $\mu \leq -\underline{\mu} < 0$ and $r \geq R_0$, the same technique gives $\frac{d}{dt}[e^{2r}] \leq -2\underline{\mu}$. Integrating from T_1 to t yields $e^{2r(t)} \leq e^{2r(T_1)} - 2\underline{\mu}(t - T_1)$, which reaches e^{2R_0} no later than $T^* = T_1 + (e^{2r(T_1)} - e^{2R_0})/(2\underline{\mu})$.

The full and detailed proof is in Appendix [C.1]. □

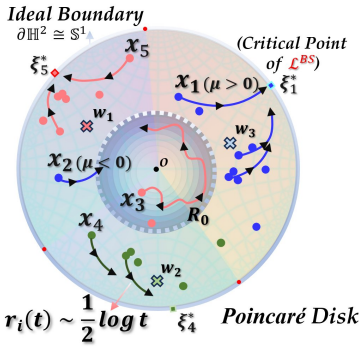


Figure 2: Radial dichotomy and boundary convergence.

Figure 2 illustrates the radial dichotomy and angular convergence on the Poincaré disk. The dashed circle denotes the fixed threshold radius R_0, outside which the large radius expansion is valid. The crosses w_1, w_2, w_3 are fixed prototypes, and the colored points are embeddings. For x_1, positive drift $\mu(\boldsymbol{\xi}_1, y_1) > 0$ drives the trajectory toward the ideal boundary, while for x_2, negative drift $\mu(\boldsymbol{\xi}_2, y_2) < 0$ pushes the trajectory back toward the interior, as characterized by Theorem 6.2. The point x_3 starts inside R_0, where the large radius expansion does not control the dynamics and the motion is instead governed by the full nonconvex loss ℓ.

6.2 Angular Convergence

We first consider the large-radius dynamics with fixed finite prototypes. Although trajectories remain bounded, they may undergo outward radial excursions before returning toward the interior. During this process, angular alignment can reverse the radial drift, while the directions approach local minima of $\mathcal{L}^{\mathrm{BS}}$ to arbitrary accuracy at sufficiently large initial radii. To formalize this behavior, we consider the following configuration:

(A2) ***Noncollinear prototypes: The fixed prototypes $\mathbf{W} = \{w_c\}_{c=1}^{C} \subset \mathbb{H}^n$, $C \geq 3$, are pairwise distinct, and there exists no hyperbolic geodesic containing all w_c, $c \in [C]$.***

For sufficiently large initial radii, trajectories with initially positive drift can undergo long time outward motion, allowing angular alignment to reverse the sign of μ and induce radial return. During this slow return, the angular deviation from a Busemann local minimum decreases exponentially with the return radius R. We formalize this behavior through the following aligned region.

Definition 6.3 (Sufficiently aligned region)**.** Let $\boldsymbol{\xi}^* \in \partial\mathbb{H}^n$ be a nondegenerate local minimum of $\mathcal{L}^{\mathrm{BS}}(\cdot, y)$. Define

$$\mathcal{A}(\boldsymbol{\xi}^*) := \{\boldsymbol{\xi} \in \partial\mathbb{H}^n : d_{\mathbb{S}^{n-1}}(\boldsymbol{\xi}, \boldsymbol{\xi}^*) \leq \theta_*\}, \tag{22}$$

where $d_{\mathbb{S}^{n-1}}(\boldsymbol{\xi}, \boldsymbol{\xi}^*) = \arccos(\hat{\boldsymbol{\xi}}^\top \hat{\boldsymbol{\xi}}^*)$ is the spherical geodesic distance, and $\theta_* > 0$ is sufficiently small that $\mu(\boldsymbol{\xi}, y) < 0$ and the Busemann Hessian is positive definite throughout $\mathcal{A}(\boldsymbol{\xi}^*)$.

The angular width of this region depends on the prototype geometry. For any number of distinct prototype directions with common radius ρ, the maximal alignment angle near the Busemann minimum approaching the true class direction scales as $\Theta(e^{-2\rho})$ under cross-entropy.

Theorem 6.4 (Three-phase Busemann dynamics). *Under Assumptions (A1) and (A2), with ψ real analytic, suppose $\boldsymbol{\xi}(T_0)$ lies in the Busemann gradient-flow basin of a nondegenerate local minimum $\boldsymbol{\xi}^*$, with $\mu(\boldsymbol{\xi}(T_0), y) > 0$. For a fixed initial angular direction, for any $R \geq R_0$, if $r(T_0) \geq 2R$, then there exist finite times $T_0 < T_1 < T_2 < T_3$ satisfying:*

(a) Large-radius excursion. *The trajectory initially moves outward, $\dot{r}(T_0) > 0$, while $\|\dot{\hat{\boldsymbol{\xi}}}(t)\| = O(e^{-2r(t)})$ in the large-radius regime.*

(b) Angular alignment and drift reversal. *At T_1, $\mu(\boldsymbol{\xi}(T_1), y) = 0$, and $\mu(\boldsymbol{\xi}(t), y) < 0$ for $t \in (T_1, T_3]$. At T_2, the trajectory first enters the sufficiently aligned region $\mathcal{A}(\boldsymbol{\xi}^*)$.*

(c) Slow radial return. *For $t \in [T_2, T_3]$, $\dot{r}(t) < 0$ and $|\dot{r}(t)| = O(e^{-2r(t)})$, with $r(T_3) = R$. For any $\varepsilon > 0$, the returning trajectory reaches*

$$d_{\mathbb{S}^{n-1}}(\boldsymbol{\xi}(T_3), \boldsymbol{\xi}^*) \leq \varepsilon \tag{23}$$

in finite time by choosing R sufficiently large.

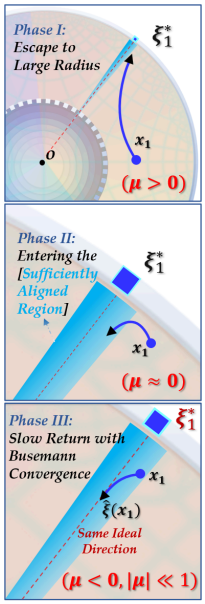


Figure 3: Three-phase Busemann dynamics.

Proof sketch. For sufficiently large r, the radial and angular dynamics satisfy $\dot{r} = 2\mu(\boldsymbol{\xi}, y)e^{-2r} + O(e^{-4r})$ and $\dot{\hat{\boldsymbol{\xi}}} = -(\sinh^{-2} r)(\operatorname{grad}_{\mathbb{S}^{n-1}} \mathcal{L}^{\mathrm{BS}} + O(e^{-2r}))$. At a Busemann critical point, $\mu(\boldsymbol{\xi}^*, y) = -\sum_{c \neq y} g_c \| \operatorname{grad}_{\mathbb{S}^{n-1}} b_{\boldsymbol{\xi}^*}(w_c) - \operatorname{grad}_{\mathbb{S}^{n-1}} b_{\boldsymbol{\xi}^*}(w_y)\|^2 < 0$, where $g_c > 0$, by noncollinearity. Under the time change $ds = dt/\sinh^2 r$, the angular dynamics is an $O(e^{-2r})$ perturbation of Busemann gradient flow. Since the initial direction lies in the basin of $\boldsymbol{\xi}^*$, finite-time approximation ensures entry into $\mathcal{A}(\boldsymbol{\xi}^*)$ while $r \geq 2R - O(1)$, forcing a preceding drift reversal. The positive definite Busemann Hessian then gives $\frac{d}{ds} d_{\mathbb{S}^{n-1}}(\boldsymbol{\xi}, \boldsymbol{\xi}^*) \leq -\lambda d_{\mathbb{S}^{n-1}}(\boldsymbol{\xi}, \boldsymbol{\xi}^*) + Ce^{-2r}$, ensuring continued angular alignment and negative radial drift. Since $|dr/ds|$ is bounded, the slow return to R provides at least $O(R)$ rescaled time for angular contraction. Consequently, $d_{\mathbb{S}^{n-1}}(\boldsymbol{\xi}(T_3), \boldsymbol{\xi}^*) \leq C' e^{-\kappa R}$, which is at most ε for sufficiently large R.

The full and detailed proof is in Appendix [C.2]. □

Theorem 6.4 establishes that trajectories with finite prototypes can undergo radial return while approaching Busemann local minima to arbitrary accuracy. We now consider ideal boundary prototypes, for which permanent radial escape becomes possible. In this setting, Busemann scores replace hyperbolic distance scores, while preserving the PERM loss structure.

(A3) ***Ideal boundary prototypes: The fixed class prototypes $w_c^\infty = (1, \hat{w}_c) \in \partial\mathbb{H}^n$ are pairwise distinct.***

In the Busemann risk, we use naturally continuous extension $b_\xi(w_c^\infty) = \log(-\langle w_c^\infty, \xi \rangle_M)$ for $\xi \neq w_c^\infty$.

Theorem 6.5 (Angular convergence to Busemann minimizers in the escaping regime)**.** *Under Assumptions (A1) and (A3), the Riemannian gradient flow satisfies*

$$r(t) \longrightarrow \infty, \qquad \boldsymbol{\xi}(t) \longrightarrow \boldsymbol{\xi}^*, \quad \textit{where} \quad \boldsymbol{\xi}^* \in \arg\min_{\boldsymbol{\xi} \in \partial\mathbb{H}^n} \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y). \tag{24}$$

*If Assumption (A3) holds for all $i \in [N]$, then $(\boldsymbol{\xi}_1^*, \ldots, \boldsymbol{\xi}_N^*)$ is a **global minimizer** of the Busemann risk $\mathcal{R}^{\mathrm{BS}}_{\mathbf{W}^\infty} = \frac{1}{N}\sum_{i=1}^N \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}_i, y_i)$ on $(\partial\mathbb{H}^n)^N$.*

Proof sketch. Rotate coordinates so $w_y^\infty = (1, 1, 0, \ldots, 0)$ and write $x = (x_0, x_1, u)$, $s = x_0 - x_1 > 0$, and $a_c = \hat{w}_{c,\perp}/(1 - \hat{w}_{c,1})$. The relative margins are $m_c = \log[(1 + \|u - sa_c\|^2)/(s^2(1 + \|a_c\|^2))]$. With $q_c = -\partial_c \psi(m) > 0$, the Riemannian flow yields $\dot{s} = -2s \sum_{c \neq y} q_c/(1 + \|u - sa_c\|^2) < 0$; du/ds is a weighted average of the fixed a_c. The margins remain bounded below, ensuring global existence, while bounded du/ds implies $u(t)$ converges. If $s(t)$ approached a positive limit, the trajectory would remain compact, contradicting a uniformly negative $\dot{s}$. Thus $s(t) \to 0$. Since $x_0 + x_1 = (1 + \|u\|^2)/s$, we obtain $r(t) \to \infty$ and $\boldsymbol{\xi}(t) \to w_y^\infty$. All relative Busemann margins then diverge to $+\infty$. Nonnegativity and coordinatewise monotonicity of ψ yield a global Busemann minimum under continuous extension. Samplewise separability establishes the joint risk result.

The full and detailed proof is in Appendix [C.3]. □

6.3 Asymptotic Decision Boundary

Theorem 6.6 (Asymptotic classifier)**.** *Under Assumption (A1), let* $\{\mathbf{w}_c\}_{c=1}^C$ *be fixed prototypes, either all in* $\mathbb{H}^n$ *or all in* $\partial\mathbb{H}^n$*, with distance or Busemann scores, respectively. For a test point* $\mathbf{x} = (\cosh r, \sinh r\,\hat{\boldsymbol{\xi}})$ *with* $\boldsymbol{\xi} = (1, \hat{\boldsymbol{\xi}}) \in \partial\mathbb{H}^n$*, define* $\hat{y}(\mathbf{x}) := \arg\min_{y\in[C]} \psi(\Upsilon_y D\,\mathbf{v}(\mathbf{x}))$*. If* $\arg\min_y \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y)$ *is unique, then for all sufficiently large* r*,* $\hat{y}(\mathbf{x}) = \arg\min_{y\in[C]} \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) = \arg\min_{c\in[C]} b_{\boldsymbol{\xi}}(\mathbf{w}_c)$*. For ideal prototypes, this equality holds for all* $r > 0$*.*

Moreover, the decision boundary between classes c *and* c' *on* $\partial\mathbb{H}^n \cong \mathbb{S}^{n-1}$ *is*

$$\mathcal{B}_{c,c'} = \left\{\hat{\boldsymbol{\xi}} \in \mathbb{S}^{n-1} : \hat{\boldsymbol{\xi}}^\top(\bar{\boldsymbol{w}}_c - \bar{\boldsymbol{w}}_{c'}) = w_{c,0} - w_{c',0},\quad b_{\boldsymbol{\xi}}(\mathbf{w}_c) \le b_{\boldsymbol{\xi}}(\mathbf{w}_k),\quad \forall k \in [C]\right\}, \tag{25}$$

which is a subset of the intersection of $\mathbb{S}^{n-1}$ *with an affine hyperplane in* $\mathbb{R}^n$*. For ideal prototypes,* $w_{c,0} = 1$*, so the corresponding spherical sections are great subspheres.*

Proof sketch. By the radial Busemann expansion (9), the common term $-r$ cancels in the relative margin vector. Hence, uniformly over y, $\psi(\Upsilon_y D\mathbf{v}(\mathbf{x})) = \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) + O(e^{-2r})$. If the minimizer of $\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, \cdot)$ is unique, the $O(e^{-2r})$ error cannot change the argmin for sufficiently large r. Therefore $\hat{y}(\mathbf{x}) = \arg\min_y \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y)$. It remains to identify this argmin. Writing $\beta_c = b_{\boldsymbol{\xi}}(\mathbf{w}_c)$, the Busemann loss depends on the relative margins $(\beta_c - \beta_y)_{c\neq y}$. Since a totally regular PERM template is permutation invariant and strictly decreasing in each coordinate, the minimum is achieved exactly by the smallest Busemann value: $\arg\min_y \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) = \arg\min_c b_{\boldsymbol{\xi}}(\mathbf{w}_c)$. Finally, $b_{\boldsymbol{\xi}}(\mathbf{w}_c) = b_{\boldsymbol{\xi}}(\mathbf{w}_{c'})$ is equivalent to $\hat{\boldsymbol{\xi}}^\top(\bar{\boldsymbol{w}}_c - \bar{\boldsymbol{w}}_{c'}) = w_{c,0} - w_{c',0}$. The additional minimum condition ensures that the pairwise bisector belongs to the multiclass decision boundary rather than separating two nonwinning classes. The full and detailed proof is in Appendix [C.4]. □

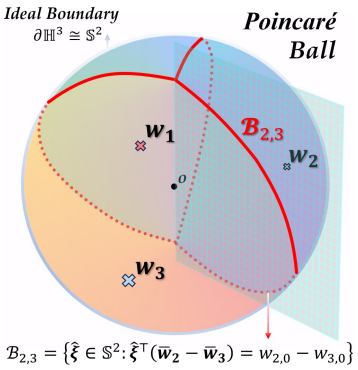


Figure 4: Asymptotic decision boundary on the ideal boundary.

Figure 4 illustrates the pairwise boundary $\mathcal{B}_{2,3}$ between prototypes w_2 and w_3, which is the spherical section obtained by intersecting $\partial\mathbb{H}^3 \simeq \mathbb{S}^2$ with the affine hyperplane $\hat{\xi}^\top(\bar{w}_2 - \bar{w}_3) = w_{2,0} - w_{3,0}$. And the Busemann loss $\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}_i, y_i)$ evaluates the classification quality of the boundary direction $\boldsymbol{\xi}_i$. This can be interpreted through the following Busemann margin:

Definition 6.7 (Busemann margin)**.** Let $\boldsymbol{\xi} = (1, \hat{\boldsymbol{\xi}}) \in \partial\mathbb{H}^n$. The *Busemann margin* of sample i at direction $\boldsymbol{\xi}$ is

$$m_i(\boldsymbol{\xi}) := \min_{c\neq y_i} \left[b_{\boldsymbol{\xi}}(\mathbf{w}_c) - b_{\boldsymbol{\xi}}(\mathbf{w}_{y_i})\right]. \tag{26}$$

Sample i is correctly classified at the ideal direction $\boldsymbol{\xi}$ if and only if $m_i(\boldsymbol{\xi}) > 0$.

Corollary 6.8 (Correct classification under cross-entropy)**.** *Under cross-entropy loss, for any class label* y_i *and any boundary direction* $\boldsymbol{\xi} \in \partial\mathbb{H}^n$*, if*

$$\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y_i) < \log 2, \tag{27}$$

then the Busemann margin is strictly positive, $m_i(\boldsymbol{\xi}) > 0$*, and the asymptotic Busemann classifier correctly predicts class* y_i*.*

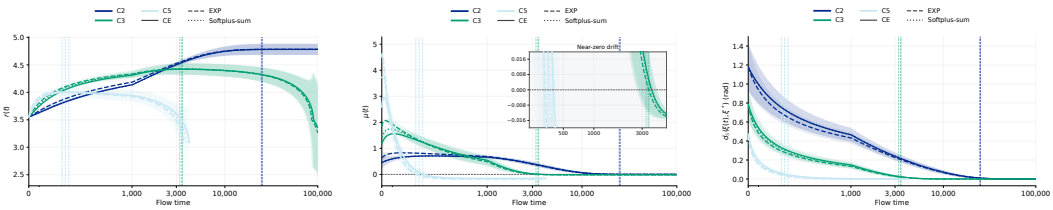
Proof. By the cross-entropy formulation of the Busemann loss,

$\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y_i) = \log\left(1 + \sum_{c\neq y_i} e^{b_{\boldsymbol{\xi}}(w_{y_i}) - b_{\boldsymbol{\xi}}(w_c)}\right) \ge \log(1 + e^{-m_i(\boldsymbol{\xi})})$. Rearranging yields the stated margin bound. Since $\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y_i) < \log 2$, the lower bound is strictly positive, implying $b_{\boldsymbol{\xi}}(w_{y_i}) < b_{\boldsymbol{\xi}}(w_c)$ for all $c \neq y_i$. Thus the Busemann classifier assigns the correct label. □

Together, these results characterize the asymptotic decision boundaries and classification margins through Busemann geometry. We next examine the predicted radial and angular behavior through numerical experiments.

7 Experiments

We evaluate hyperboloid gradient dynamics using $C = 3$ prototypes on $\mathbb{H}^2$ at radius $\rho = 1$ with $N = 15$ embeddings, and ablate $C \in \{2, 3, 5\}$, PERM templates $\psi \in \{\psi_{\mathrm{CE}}, \psi_{\mathrm{EXP}}, \psi_{\mathrm{SS}}\}$. We integrate both the full coupled gradient flow and the separate radial dynamics.



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

Figure 5: Three-phase Busemann dynamics under coupled Riemannian gradient flow.

Figure 5 confirms the three phases predicted by Theorem 6.4 across configurations with $C \geq 3$. The trajectories initially move outward, undergo angular alignment with radial drift reversal, and subsequently return slowly toward the interior. Across all configurations, the directions approach Busemann local minima, supporting the predicted Busemann implicit bias across prototype geometries and PERM losses, with weaker drift reversal in degenerate cases. Shaded regions indicate standard deviations across seeds. Vertical dashed lines indicate entry into sufficiently aligned regions $\mathcal{A}(\boldsymbol{\xi}^*)$.

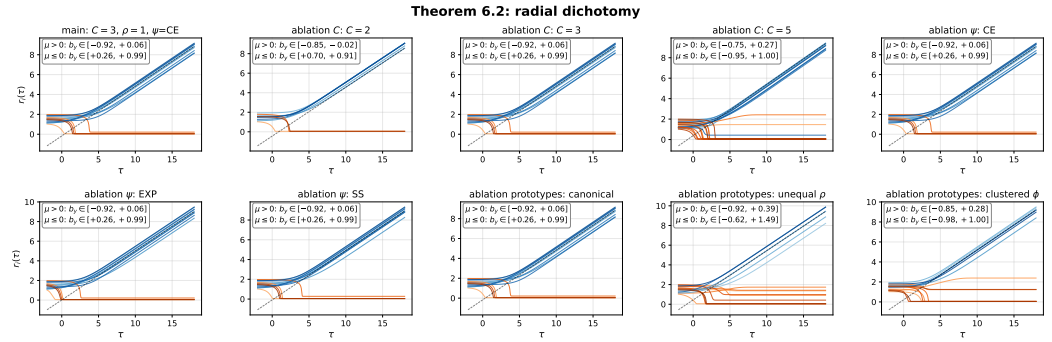


Figure 6: Numerical verification of radial dichotomy under fixed-direction gradient flow.

Figure 6 confirms the predicted $r_i(t) = \frac{1}{2} \log t + O(1)$ scaling under persistent positive drift, while negative drift induces inward motion. These results support the radial dichotomy and three-phase behavior across the tested settings. Full experimental details, neural network experiments, and additional diagnostics are provided in Appendix [D].

8 Conclusions

In this work, we analyzed the implicit bias of Riemannian gradient flow for hyperbolic multiclass classification with fixed prototypes. Through a Radial Busemann decomposition, we showed that large-radius dynamics are governed by an interaction between radial drift and boundary direction optimization. This yields a radial dichotomy and a three-phase regime of outward motion, angular alignment, and slow return, during which the directions approach local minima of the Busemann loss. We also established radial escape and convergence to global Busemann risk minimizers. We further characterized the asymptotic classifier and its spherical decision boundaries.

Controlled experiments support the predicted radial and angular dynamics, while coupled integration illustrates drift reversal and three-phase behavior. Neural network experiments further validate the Busemann asymptotic approximation. Together, these results provide a unified perspective on boundary saturation and near-boundary clustering through their shared Busemann directional bias. More broadly, our analysis identifies the ideal boundary as the natural geometric object governing optimization bias in hyperbolic representation learning and suggests a boundary-based perspective for studying more general hyperbolic learning dynamics.

Acknowledgments and Disclosure of Funding

Funding. This work was partially supported by JST Moonshot R&D, Grant Number JPMJPS2011, the New Energy and Industrial Technology Development Organization (NEDO) under Project JPNP26007, JSPS KAKENHI, Grant Number JP25K21268 and JST SPRING, Grant Number JPMJSP2108.

Competing interests. The authors declare no competing interests.

References

[1] Yutong Wang and Clayton Scott. Unified binary and multiclass margin-based classification. *Journal of Machine Learning Research*, 25(143):1–51, 2024. URL http://jmlr.org/papers/v25/23-1599.html.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017. ISSN 0001-0782. doi: 10.1145/3065386. URL https://doi.org/10.1145/3065386.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:52967399.

[4] John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021. URL https://api.semanticscholar.org/CorpusID:235959867.

[5] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *ArXiv*, abs/1611.03530, 2016. URL https://api.semanticscholar.org/CorpusID:6212000.

[6] Mikhail Belkin, Daniel J. Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116:15849 – 15854, 2018. URL https://api.semanticscholar.org/CorpusID:198496504.

[7] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *CoRR*, abs/1412.6614, 2014. URL https://api.semanticscholar.org/CorpusID:6021932.

[8] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19 (70):1–57, 2018. URL http://jmlr.org/papers/v19/18-188.html.

[9] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. *Implicit regularization in deep matrix factorization*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[10] Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6152–6160, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[11] Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In Marc van Kreveld and Bettina Speckmann, editors, *Graph Drawing*, pages 355–366, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-25878-7.

[12] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *ArXiv*, abs/1705.08039, 2017. URL https://api.semanticscholar.org/CorpusID:25418227.

[13] Frederic Sala, Chris De Sa, Albert Gu, and Christopher Re. Representation tradeoffs for hyperbolic embeddings. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4460–4469. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/sala18a.html.

[14] Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 5350–5360, Red Hook, NY, USA, 2018. Curran Associates Inc.

[15] Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:202771753.

[16] Menglin Yang, Harshit Verma, Delvin Ce Zhang, Jiahong Liu, Irwin King, and Rex Ying. Hypformer: Exploring efficient transformer fully in hyperbolic space. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024. URL https://api.semanticscholar.org/CorpusID:271961739.

[17] Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. Hyperbolic deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10023–10044, 2022. doi: 10.1109/TPAMI.2021.3136921.

[18] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6417–6427, 2020. doi: 10.1109/CVPR42600.2020.00645.

[19] Edoardo Cetin, Benjamin Paul Chamberlain, Michael M. Bronstein, and Jonathan J Hunt. Hyperbolic deep reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TfBHFLgv77.

[20] Ryohei Shimizu, Yusuke Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. *ArXiv*, abs/2006.08210, 2020. URL https://api.semanticscholar.org/CorpusID:260519400.

[21] Xin Yang, Xingrun Li, Heng Chang, Yang Jinze, Xihong Yang, Shengyu Tao, Maiko Shigeno, Ningkang Chang, Junfeng Wang, Dawei Yin, and Erxue Min. Hgformer: Hyperbolic graph transformer for collaborative filtering. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 70813–70832. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/yang25o.html.

[22] Pascal Mettes, Mina Ghadimi Atigh, Martin Keller-Ressel, Jeffrey Gu, and Serena Yeung. Hyperbolic deep learning in computer vision: A survey. *International Journal of Computer Vision*, 132:3484 – 3508, 2023. URL https://api.semanticscholar.org/CorpusID:258615705.

[23] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *ArXiv*, abs/1906.05890, 2019. URL https://api.semanticscholar.org/CorpusID:189898036.

[24] Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *ArXiv*, abs/2006.06657, 2020. URL https://api.semanticscholar.org/CorpusID:219573767.

[25] Gary Bécigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. *ArXiv*, abs/1810.00760, 2018. URL https://api.semanticscholar.org/CorpusID:52898806.

[26] Silvère Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013. doi: 10.1109/TAC.2013.2254619.

[27] Maximilian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. *ArXiv*, abs/1806.03417, 2018. URL https://api.semanticscholar.org/CorpusID:47016889.

[28] Gal Mishne, Zhengchao Wan, Yusu Wang, and Sheng Yang. The numerical stability of hyperbolic representation learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24925–24949. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/mishne23a.html.

[29] Yunhui Guo, Xudong Wang, Yubei Chen, and Stella X. Yu. Clipped hyperbolic classifiers are super-hyperbolic classifiers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10, 2022. doi: 10.1109/CVPR52688.2022.00010.

[30] Gabriel Moreira, Manuel Marques, João Paulo Costeira, and Alexander Hauptmann. Hyperbolic vs euclidean embeddings in few-shot learning: Two sides of the same coin. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2071–2079, 2024. doi: 10.1109/WACV57701.2024.00208.

[31] Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7399–7409, 2022. doi: 10.1109/CVPR52688.2022.00726.

[32] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1772–1798. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/ji19a.html.

[33] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1832–1841. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/gunasekar18a.html.

[34] Hrithik Ravi, Clayton Scott, Daniel Soudry, and Yutong Wang. The implicit bias of gradient descent on separable multiclass data. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 81324–81359. Curran Associates, Inc., 2024. doi: 10.52202/079017-2585. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/942d5e890e4e0ab0e7a5f8cbcef5b9ff-Paper-Conference.pdf.

[35] Nikolaos Tsilivis, Gal Vardi, and Julia Kempe. Flavors of margin: Implicit bias of steepest descent in homogeneous neural networks. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 10609–10635, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/1c9ee8c9e1a0c7de49e15720c4ce6f78-Paper-Conference.pdf.

[36] James W. Anderson. Hyperbolic geometry. 1999. URL https://api.semanticscholar.org/CorpusID:221600251.

[37] Abraham Albert Ungar. *Gyrovector Spaces*, pages 55–110. Springer International Publishing, Cham, 2009. ISBN 978-3-031-02396-5. doi: 10.1007/978-3-031-02396-5_3. URL https://doi.org/10.1007/978-3-031-02396-5_3.

[38] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:309759.

[39] Pascal Mettes, Elise van der Pol, and Cees G. M. Snoek. Hyperspherical prototype networks. *ArXiv*, abs/1901.10514, 2019. URL https://api.semanticscholar.org/CorpusID:59413788.

[40] Paul Berg, Léo Buecher, Björn Michele, Minh-Tan Pham, Laetitia Chapel, and Nicolas Courty. Multi-prototype hyperbolic learning guided by class hierarchy: Multi-prototype hyperbolic learning guided by class hierarchy. *Int. J. Comput. Vision*, 133(11):7969–7984, August 2025. ISSN 0920-5691. doi: 10.1007/s11263-025-02571-8. URL https://doi.org/10.1007/s11263-025-02571-8.

[41] Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3420–3428. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/nacson19b.html.

[42] Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *ArXiv*, abs/1803.07300, 2018. URL https://api.semanticscholar.org/CorpusID:4803831.

[43] Lénaïc Chizat and Francis R. Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Annual Conference Computational Learning Theory*, 2020. URL https://api.semanticscholar.org/CorpusID:211076120.

[44] Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3635–3673. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/woodworth20a.html.

[45] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3779–3788. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/nickel18a.html.

[46] Mina Ghadimi Atigh, Martin Keller-Ressel, and Pascal Mettes. Hyperbolic busemann learning with ideal prototypes. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:235658959.

[47] Xiran Fan, Chun-Hao Yang, and Baba C. Vemuri. Horospherical decision boundaries for large margin classification in hyperbolic space. *Advances in neural information processing systems*, 36:11194–11204, 2023. URL https://api.semanticscholar.org/CorpusID:259138291.

[48] Ziheng Chen, Bernhard Schölkopf, and Niculae Sebe. Hyperbolic busemann neural networks. *ArXiv*, abs/2602.18858, 2026. URL https://api.semanticscholar.org/CorpusID:285973639.

[49] Xuan Son Nguyen, Shuo Yang, and Aymeric Histace. Neural networks on symmetric spaces of noncompact type. *ArXiv*, abs/2601.01097, 2026. URL https://api.semanticscholar.org/CorpusID:278499232.

A Related Works, Discussion and Additional Results

A.1 Related Works

Implicit bias. Implicit bias studies how optimization algorithms select particular solutions among many empirical risk minimizers, even without explicit regularization. In separable linear classification, gradient descent with exponential or cross entropy losses converges in direction to the maximum margin classifier [8]. This line of work has been extended to refined convergence rates and margin dynamics [41, 42], nonseparable classification [32], homogeneous neural networks [23], directional convergence in deep models [24], general homogeneous or overparameterized models [43, 44], and optimization geometry dependent biases such as steepest descent and mirror descent [33, 35]. Beyond classification, implicit regularization has also been studied in matrix factorization, deep matrix factorization, and deep linear networks, where gradient based methods favor low rank or low complexity solutions [9, 10]. These works mainly characterize Euclidean biases toward maximum margin, low norm, low rank, or other complexity controlled structures.

Multiclass implicit bias and PERM losses. For multiclass classification, recent work extends binary implicit bias theory to permutation invariant relative margin losses through the PERM framework [34, 1]. This framework covers cross entropy, multiclass exponential losses, and broader relative margin losses. We use PERM losses as a general multiclass loss class, but our goal is different from Euclidean multiclass margin theory: instead of characterizing linear maximum margin behavior, we study radial escape and boundary direction selection in hyperbolic space.

Hyperbolic representation learning and Busemann based methods. Hyperbolic representation learning has been widely used for hierarchical, tree like, and long tailed data. Representative works include Poincaré embeddings [12], Lorentz model embeddings [45], hyperbolic neural networks [14], hyperbolic neural network variants [20], hyperbolic graph neural networks [15], and hyperbolic Transformers [16]; see [17] for a survey. Applications include computer vision, few shot learning, metric learning, graph learning, recommendation, and reinforcement learning [31, 30, 21, 19]. Several empirical studies also report boundary related phenomena, including numerical instability near the Poincaré boundary [28], the effectiveness of clipping [29], common radius concentration [30], and metric learning induced class separation [31]. Busemann functions have also been used in hyperbolic machine learning. Hyperbolic Busemann Learning uses ideal prototypes and a Busemann loss [46], and horospherical decision boundary methods use level sets of Busemann functions as classifiers [47]. Recent works further study Busemann based neural network layers [48], and related constructions on symmetric spaces [49]. These works demonstrate the importance of hyperbolic and Busemann geometry, but mainly focus on model design, empirical behavior, or geometric constructions. Our work is complementary: rather than proposing a new Busemann model, we show that Busemann geometry emerges as the large radius limit of distance based hyperbolic multiclass classification. Under PERM losses, the common radial term cancels in the relative margins, yielding a Busemann risk on the ideal boundary and leading to radial dichotomy, angular convergence, and an asymptotic boundary classifier.

A.2 Discussion and Limitation

A.2.1 Discussion.

Our results provide an asymptotic description of the boundary regime in hyperbolic representation learning. This regime is different from the local Euclidean behavior near the origin: near the ideal boundary, hyperbolic distances admit an exponential expansion governed by Busemann functions. Our analysis exploits this expansion to separate the loss into a common radial term and a term that depends on direction through the Busemann function. This separation is the geometric mechanism behind the radial dichotomy, the Busemann risk on the ideal boundary, and the resulting asymptotic classifier.

This perspective helps interpret several empirical phenomena in hyperbolic learning. First, the radial dichotomy gives a mechanism for boundary saturation: when the drift coefficient is positive, embeddings are pushed toward the boundary at rate $r(t) = \frac{1}{2}\log t + O(1)$, consistent with observations that embeddings concentrate near the Poincaré boundary and that numerical instabilities arise in this region [28]. The coefficient $\frac{1}{2}$ is universal in our setting, in the sense that it does not depend on the

data geometry or margin. This contrasts with Euclidean separable linear classification, where the logarithmic growth coefficient is determined by the data margin γ^* [8]. Second, the angular convergence theorem suggests a mechanism for near boundary clustering: once embeddings have escaped sufficiently, their discriminative information is encoded mainly in boundary directions governed by the Busemann risk on $\partial\mathbb{H}^n$, consistent with observations of concentration at a common radius and directional class separation [30, 31].

Our theory also provides an interpretation of radius clipping [29]. Since the leading classification behavior near the boundary depends on Busemann values and boundary directions, extremely large radii are not necessary once embeddings have entered the asymptotic regime. Clipping can therefore be viewed as preventing excessive radial growth while preserving the boundary direction structure that determines the classifier in the late stage of training.

The discussion above was stated for the normalized case $K = 1$. For completeness, we now record the corresponding curvature scaling, which justifies the $K = 1$ convention used in the main text.

Curvature Dependence. For general Minkowski constant $K > 0$ (equivalently, curvature $\kappa = -1/K$), the model $\mathbb{H}^{n,K}$ and all preceding quantities rescale by $\sqrt{K}$. Substituting $r \mapsto r/\sqrt{K}$, $d_{\mathbb{H}}^K = \sqrt{K}\, d_{\mathbb{H}}^{K=1}$, and $b_{\hat{\boldsymbol{\xi}}}^K = \sqrt{K}\, b_{\hat{\boldsymbol{\xi}}}^{K=1}$ throughout:

Corollary A.1 (Curvature scaling)**.** *For any $K > 0$, the rescaled properties are:*

(a) (Radial rate) Under the persistent positive drift condition of Theorem 6.2, $r_i(t) = \frac{\sqrt{K}}{2} \log t + O(1)$ as $t \to \infty$.

(b) (Angular limit) Under the assumptions of Theorem 6.5 for ideal prototypes, $\hat{\boldsymbol{\xi}}_i(t) \to \hat{\boldsymbol{\xi}}_i^{*,K} \in \arg\min_{\hat{\boldsymbol{\xi}} \in \mathbb{S}^{n-1}} \mathcal{L}^{\mathrm{BS},K}(\boldsymbol{\xi}, y_i)$, where $\mathcal{L}^{\mathrm{BS},K}(\boldsymbol{\xi}, y_i) := \psi(\Upsilon_{y_i} D\mathbf{b}^K(\boldsymbol{\xi}))$.

(c) (Decision boundary) The asymptotic decision boundaries on $\partial\mathbb{H}^{n,K} \cong \mathbb{S}^{n-1}$ are identical to the $K = 1$ case.

(d) (Horospherical margin) $m_i^K(\boldsymbol{\xi}) = \sqrt{K}\, m_i^{K=1}(\boldsymbol{\xi})$.

A.2.2 Limitations.

Our analysis focuses on Riemannian gradient flow for embeddings with fixed class prototypes. This setting follows a common strategy in implicit bias theory, where part of the learning problem is fixed to isolate the optimization mechanism of interest. For example, separable linear classification fixes the data and studies classifier weight dynamics [8], while matrix factorization fixes observed entries and analyzes the implicit preference of learned factors [9, 10]. In our case, fixing the prototypes isolates the geometric mechanism governing radial drift and angular alignment. We further investigate shared encoder parameterization numerically. Extending the theoretical analysis to joint optimization of embeddings, prototypes, and encoder parameters introduces additional interactions through the encoder Jacobian and remains an important direction for future work.

Our experiments combine controlled numerical integration with neural network investigations. Coupled gradient flow experiments illustrate the predicted three-phase dynamics across prototype configurations and PERM losses. Shared encoder experiments reveal exponential proximity between hyperbolic and Busemann stationary points, while ResNet-18 experiments on CIFAR-10 provide complementary evidence through parameter gradient comparisons. Together, these experiments extend the numerical evidence beyond independent embedding dynamics, while broader architectures and jointly learned prototypes remain natural directions for further investigation.

B Proof of Section 5

B.1 Proof of Lemma 5.1

Lemma 5.1 (Radial-Busemann expansion) *For* $\mathbf{x} = (\cosh r, \sinh r\,\hat{\boldsymbol{\xi}}) \in \mathbb{H}^n \setminus \{\mathbf{o}\}$ *and* $\mathbf{w} = (w_0, \bar{\boldsymbol{w}}) \in \mathbb{H} \times \mathbb{H}^{n-1}$*, let* $\boldsymbol{\xi} = (1, \hat{\boldsymbol{\xi}}) \in \partial\mathbb{H}^n$*. There exists a constant* $R_0(\mathbf{w}) > 0$ *such that for all* $r \geq R_0$*,*

$$d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}) = r + b_{\boldsymbol{\xi}}(\mathbf{w}) + O(e^{-2r}),$$

where $b_{\boldsymbol{\xi}}(\mathbf{w}) = \log(w_0 - \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}})$ *is the Busemann function.*

Proof. Recall that the hyperboloid model is defined as $\mathbb{H}^n = \{\mathbf{x} = (x_0, x_1, \ldots, x_n) = (x_0, \bar{\mathbf{x}}) \in \mathbb{R}^{n+1} : \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{M}} = -1,\ x_0 > 0\}$, where the Minkowski inner product is $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{M}} = -x_0 y_0 + \sum_{i=1}^n x_i y_i = -x_0 y_0 + \bar{\mathbf{x}}^\top \bar{\mathbf{y}}$. The geodesic distance between $\mathbf{x}, \mathbf{w} \in \mathbb{H}^n$ is

$$d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}) = \operatorname{arcosh}\big(-\langle \mathbf{x}, \mathbf{w} \rangle_{\mathcal{M}}\big). \tag{28}$$

We are given $\mathbf{x} = (\cosh r, \sinh r\,\hat{\boldsymbol{\xi}})$ with $r > 0$ and $\hat{\boldsymbol{\xi}} \in \mathbb{S}^{n-1}$, and $\mathbf{w} = (w_0, w_1, \ldots, w_n) = (w_0, \bar{\boldsymbol{w}})$. Substituting into the Minkowski inner product:

$$\begin{aligned} -\langle \mathbf{x}, \mathbf{w} \rangle_{\mathcal{M}} &= -\big(-\cosh r \cdot w_0 + \sinh r\,\hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}}\big) \\ &= w_0 \cosh r - (\hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}}) \sinh r. \end{aligned} \tag{29}$$

Now we use the exponential forms of hyperbolic functions:

$$\cosh r = \frac{e^r + e^{-r}}{2}, \qquad \sinh r = \frac{e^r - e^{-r}}{2}.$$

Substituting into (29):

$$\begin{aligned} -\langle \mathbf{x}, \mathbf{w} \rangle_{\mathcal{M}} &= w_0 \cdot \frac{e^r + e^{-r}}{2} - (\hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}}) \cdot \frac{e^r - e^{-r}}{2} \\ &= \frac{e^r}{2}\big(w_0 - \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}}\big) + \frac{e^{-r}}{2}\big(w_0 + \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}}\big). \end{aligned} \tag{30}$$

Since $\mathbf{w} \in \mathbb{H}^n$, we have $\langle \mathbf{w}, \mathbf{w} \rangle_{\mathcal{M}} = -w_0^2 + \|\bar{\boldsymbol{w}}\|^2 = -1$, so $w_0 = \sqrt{1 + \|\bar{\boldsymbol{w}}\|^2}$. In particular $w_0 \geq 1$ and $w_0 > \|\bar{\boldsymbol{w}}\|$. By the Cauchy–Schwarz inequality in $\mathbb{R}^n$:

$$\|\hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}}\| \leq \|\hat{\boldsymbol{\xi}}\|\,\|\bar{\boldsymbol{w}}\| = \|\bar{\boldsymbol{w}}\| < w_0,$$

since $\hat{\boldsymbol{\xi}} \in \mathbb{S}^{n-1}$ has $\|\hat{\boldsymbol{\xi}} = 1$. Therefore $w_0 - \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}} > 0$. We define

$$A \;:=\; w_0 - \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}} \;>\; 0. \tag{31}$$

Similarly, we define $B := w_0 + \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}} > 0$. Then substitute into (30):

$$-\langle \mathbf{x}, \mathbf{w} \rangle_{\mathcal{M}} = \frac{e^r}{2} A + \frac{e^{-r}}{2} B = \frac{e^r A}{2}\left(1 + \frac{B}{A}\, e^{-2r}\right). \tag{32}$$

Since the distance function is $d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}) = \operatorname{arcosh}\big(-\langle \mathbf{x}, \mathbf{w} \rangle_{\mathcal{M}}\big)$, we also need to expand the arcosh function. By definition, $\operatorname{arcosh}(u) = \log(u + \sqrt{u^2 - 1})$. We have

$$\begin{aligned} u + \sqrt{u^2 - 1} &= u + u\sqrt{1 - 1/u^2} \\ &= u\left(1 + \sqrt{1 - 1/u^2}\right) \\ &= 2u \cdot \frac{1 + \sqrt{1 - 1/u^2}}{2}. \end{aligned} \tag{33}$$

Therefore

$$\operatorname{arcosh}(u) = \log(2u) + \log\left(\frac{1 + \sqrt{1 - 1/u^2}}{2}\right). \tag{34}$$

Let $u \geq 2$ and set $\delta = 1/u^2$, then $\delta \in (0, 1/4]$. Using the Taylor bound $\sqrt{1-\delta} \geq 1 - \delta$ for $\delta \in [0, 1]$:

$$\frac{1 + \sqrt{1-\delta}}{2} \geq \frac{1 + 1 - \delta}{2} = 1 - \frac{\delta}{2}. \tag{35}$$

Also $\sqrt{1-\delta} \leq 1$, so:

$$\frac{1 + \sqrt{1-\delta}}{2} \leq 1. \tag{36}$$

Hence

$$1 - \frac{\delta}{2} \leq \frac{1 + \sqrt{1-\delta}}{2} \leq 1. \tag{37}$$

Since $\log$ is monotone and $|\log(1-s)| \leq 2s$ for $s \in [0, 1/2]$:

$$|\log\left(\frac{1 + \sqrt{1-\delta}}{2}\right)| \leq |\log(1 - \delta/2)| \leq 2 \cdot \frac{\delta}{2} = \delta = \frac{1}{u^2}. \tag{38}$$

So we get that for $u \geq 2$, we have:

$$\operatorname{arcosh}(u) = \log(2u) + \varepsilon(u), \qquad \text{where } |\varepsilon(u)| \leq \frac{1}{u^2}. \tag{39}$$

Then we need $u = -\langle \mathbf{x}, \mathbf{w} \rangle_{\mathcal{M}} \geq 2$ for applying 39 . From (32), $u = \frac{e^r A}{2}\left(1 + \frac{B}{A}\, e^{-2r}\right)$, since $A > 0$ and $B > 0$, $1 + \frac{B}{A}\, e^{-2r} \geq 1$ for all $r > 0$, so

$$u \;\geq\; \frac{e^r A}{2}. \tag{40}$$

Since $\hat{\boldsymbol{\xi}} \in \mathbb{S}^{n-1}$ is a unit vector, the Cauchy–Schwarz inequality gives

$$A \;=\; w_0 - \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}} \;\geq\; w_0 - \|\bar{\boldsymbol{w}}\|. \tag{41}$$

Using the hyperboloid constraint $w_0^2 - \|\bar{\boldsymbol{w}}\|^2 = 1$, we factor

$$w_0 - \|\bar{\boldsymbol{w}}\| \;=\; \frac{w_0^2 - \|\bar{\boldsymbol{w}}\|^2}{w_0 + \|\bar{\boldsymbol{w}}\|} \;=\; \frac{1}{w_0 + \|\bar{\boldsymbol{w}}\|} \;>\; 0. \tag{42}$$

Hence $A \geq \frac{1}{w_0 + \|\bar{\boldsymbol{w}}\|}$ uniformly over all directions $\hat{\boldsymbol{\xi}} \in \mathbb{S}^{n-1}$. Then the condition $u \geq \frac{e^r A}{2} \geq 2$ is guaranteed whenever

$$r \;\geq\; R_0(\mathbf{w}) \;:=\; \log\big(4(w_0 + \|\bar{\boldsymbol{w}}\|)\big), \tag{43}$$

which depends only on the prototype $\mathbf{w}$, not on the direction $\hat{\boldsymbol{\xi}}$.

Finally, for r large enough, all $r \geq R_0(\mathbf{w})$:

$$\begin{aligned} d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}) &= \operatorname{arcosh}(u) \\ &= \log(2u) + O(u^{-2}). \end{aligned} \tag{44}$$

We compute $\log(2u)$:

$$\begin{aligned} \log(2u) &= \log\left(2 \cdot \frac{e^r A}{2}\left(1 + \frac{B}{A} e^{-2r}\right)\right) \\ &= \log\left(e^r A\left(1 + \frac{B}{A} e^{-2r}\right)\right) \\ &= \log(e^r) + \log(A) + \log\left(1 + \frac{B}{A} e^{-2r}\right) \\ &= r + \log A + \log\left(1 + \frac{B}{A} e^{-2r}\right). \end{aligned} \tag{45}$$

The last logarithm is bounded by e^{-2r} using $\log(1+s) \leq s$ for $s > 0$:

$$0 \leq \log\left(1 + \frac{B}{A}e^{-2r}\right) \leq \frac{B}{A}\, e^{-2r} \leq C_1(\mathbf{w})e^{-2r}, \tag{46}$$

where $C_1(\mathbf{w}) := (w_0 + \|\bar{\boldsymbol{w}}\|)^2$ since $\frac{B}{A} = \frac{w_0+\hat{\boldsymbol{\xi}}\bar{\boldsymbol{w}}}{w_0-\hat{\boldsymbol{\xi}}\bar{\boldsymbol{w}}} \leq (w_0 + \|\bar{\boldsymbol{w}}\|)^2$. Also, the $O(u^{-2})$ term in (44) is also bounded by e^{-2r}, since $u \geq \frac{e^r A}{2}$ for $r \geq R_0$, we have $u^{-2} \leq \frac{4}{A^2}\, e^{-2r} \leq C_2(\mathbf{w})e^{-2r}$, where $C_2(\mathbf{w}) := 4(w_0 + \|\bar{\boldsymbol{w}}\|)^2$. Combining (44), (45), and (46):

$$\begin{aligned} d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}) &= r + \log A + \underbrace{\log\left(1 + \frac{B}{A}e^{-2r}\right)}_{O(e^{-2r})} + \underbrace{O(u^{-2})}_{O(e^{-2r})} \\ &= r + \log A + O(e^{-2r}). \end{aligned}$$

Recalling $A = w_0 - \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}}$ from (31), we have $\log A = \log(w_0 - \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}})$. This is precisely the Busemann function $b_{\boldsymbol{\xi}}(\mathbf{w})$ centered at the ideal boundary point $\boldsymbol{\xi} = (1, \hat{\boldsymbol{\xi}}) \in \partial\mathbb{H}^n$:

$$b_{\boldsymbol{\xi}}(\mathbf{w}) = \log(w_0 - \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}}).$$

So, we have shown that for $r \geq R_0(\mathbf{w})$,

$$d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}) = r + b_{\boldsymbol{\xi}}(\mathbf{w}) + O(e^{-2r}),$$

which completed the proof. □

B.2 Proof of Proposition 5.3

Proposition 5.3 (Loss expansion) *Under Assumption (A1), there exists a constant $R_0(\mathbf{W}) > 0$ such that for each $i \in [N]$ and all $r_i \geq R_0$,*

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

where

$$\mu(\boldsymbol{\xi}, y) \;:=\; -\big\langle \nabla\psi\big(\Upsilon_y D\,\mathbf{b}(\boldsymbol{\xi})\big),\; \Upsilon_y D\,\boldsymbol{\alpha}(\boldsymbol{\xi})\big\rangle,$$

and we call $\mu : \partial\mathbb{H}^n \times [C] \to \mathbb{R}$ the radial drift coefficient.

Proof. Recall from the proof of Lemma 5.1 (Appendix B.1) that for a sample point $\mathbf{x}_i = (\cosh r_i,\, \sinh r_i\, \hat{\boldsymbol{\xi}}_i)$ and a prototype $\mathbf{w}_c = (w_{c,0},\, \bar{\boldsymbol{w}}_c)$, we defined

$$A_c := w_{c,0} - \hat{\boldsymbol{\xi}}_i \cdot \bar{\boldsymbol{w}}_c > 0, \qquad B_c := w_{c,0} + \hat{\boldsymbol{\xi}}_i \cdot \bar{\boldsymbol{w}}_c > 0,$$

and showed that

$$u_c \;:=\; -\langle \mathbf{x}_i, \mathbf{w}_c\rangle_{\mathcal{M}} \;=\; \frac{e^{r_i} A_c}{2}\big(1 + s_c\big), \qquad s_c := \frac{B_c}{A_c}\, e^{-2r_i}. \tag{47}$$

Lemma 5.1 only kept the leading term, here we need the exact e^{-2r_i} coefficient, so we expand more carefully. Here by using the Cauchy inequality and the hyperboloid constraint $w_{c,0}^2 - \|\bar{\boldsymbol{w}}_c\|^2 = 1$, we established in Lemma 5.1 that

$$A_c \;\geq\; w_{c,0} - \|\bar{\boldsymbol{w}}_c\| \;=\; \frac{1}{w_{c,0} + \|\bar{\boldsymbol{w}}_c\|} \;>\; 0. \tag{48}$$

Similarly, $B_c \leq w_{c,0} + \|\bar{\boldsymbol{w}}_c\|$, so the ratio B_c/A_c is bounded above by $(w_{c,0} + \|\bar{\boldsymbol{w}}_c\|)^2$. Define the global threshold

$$R_0(\mathbf{W}) \;:=\; \max_{c\in[C]} \log\big(4(w_{c,0} + \|\bar{\boldsymbol{w}}_c\|)\big). \tag{49}$$

For any $r_i \geq R_0(\mathbf{W})$ and every $c \in [C]$, two conditions hold simultaneously:

$$u_c \geq \frac{e^{r_i} A_c}{2} \geq 2, \tag{50}$$

ensuring that the arcosh expansion (equation (39) in Lemma 5.1) is valid, and the perturbation s_c satisfies

$$0 < s_c \leq (w_{c,0} + \|\bar{\boldsymbol{w}}_c\|)^2 \cdot \frac{1}{16(w_{c,0} + \|\bar{\boldsymbol{w}}_c\|)^2} = \frac{1}{16}, \tag{51}$$

so all Taylor remainders for $s_c \to 0$ are uniformly $O(e^{-4r_i})$. We now refine the two pieces of $d_{\mathbb{H}}(\mathbf{x}_i, \mathbf{w}_c) = \log(2u_c) + \varepsilon(u_c)$. From (47):

$$\begin{aligned} \log(2u_c) &= \log\big(e^{r_i}\, A_c\,(1 + s_c)\big) \\ &= \log(e^{r_i}) + \log(A_c) + \log(1 + s_c). \end{aligned} \tag{52}$$

Since $0 < s_c \leq 1/16$, we use the Taylor expansion $\log(1+s) = s - \frac{s^2}{2} + \cdots$ with the bound $|\log(1+s) - s| \leq s^2$ for $s \in [0, 1/2]$. This gives

$$\log(1 + s_c) = s_c + O(s_c^2) = \frac{B_c}{A_c}\, e^{-2r_i} + O(e^{-4r_i}). \tag{53}$$

Substituting (53) into (52):

$$\log(2u_c) = r_i + \log A_c + \frac{B_c}{A_c}\, e^{-2r_i} + O(e^{-4r_i}). \tag{54}$$

Recall from the Appendix [B.1] that $\varepsilon(u) = \log\left(\frac{1+\sqrt{1-1/u^2}}{2}\right)$. We first compute $1/u_c^2$. From (47), $u_c = \frac{e^{r_i} A_c}{2}(1 + s_c)$, so

$$\frac{1}{u_c^2} = \frac{4}{e^{2r_i}\, A_c^2\,(1 + s_c)^2}. \tag{55}$$

Since $0 < s_c \le 1/16$, we expand $(1+s_c)^{-2} = 1 - 2s_c + O(s_c^2) = 1 + O(s_c)$, giving

$$\delta_c \;:=\; \frac{1}{u_c^2} \;=\; \frac{4}{A_c^2}\, e^{-2r_i} + O(e^{-4r_i}). \tag{56}$$

Now we expand ε in powers of δ_c. First, $\sqrt{1-\delta} = 1 - \frac{\delta}{2} + O(\delta^2)$ for $\delta \to 0$:

$$\sqrt{1-\delta_c} = 1 - \frac{2}{A_c^2}\, e^{-2r_i} + O(e^{-4r_i}). \tag{57}$$

Hence

$$\frac{1+\sqrt{1-\delta_c}}{2} = 1 - \frac{1}{A_c^2}\, e^{-2r_i} + O(e^{-4r_i}). \tag{58}$$

Next, $\log(1-s) = -s + O(s^2)$ for $s \to 0$, with $s = \frac{1}{A_c^2}\, e^{-2r_i}$, we get:

$$\varepsilon(u_c) = -\frac{1}{A_c^2}\, e^{-2r_i} + O(e^{-4r_i}). \tag{59}$$

Adding (54) and (59) we can get:

$$d_{\mathbb{H}}(\mathbf{x}_i, \mathbf{w}_c) = r_i + \underbrace{\log A_c}_{=b_{\boldsymbol{\xi}_i}(\mathbf{w}_c)} + \underbrace{\left(\frac{B_c}{A_c} - \frac{1}{A_c^2}\right)}_{:=\alpha_{\boldsymbol{\xi}_i}(\mathbf{w}_c)} e^{-2r_i} + O(e^{-4r_i}). \tag{60}$$

We combine the two fractions over a common denominator:

$$\alpha_{\boldsymbol{\xi}_i}(\mathbf{w}_c) = \frac{B_c}{A_c} - \frac{1}{A_c^2} = \frac{A_c\, B_c - 1}{A_c^2} = \frac{\|\bar{\boldsymbol{w}}_c\|^2 - (\hat{\boldsymbol{\xi}}_i \cdot \bar{\boldsymbol{w}}_c)^2}{(w_{c,0} - \hat{\boldsymbol{\xi}}_i \cdot \bar{\boldsymbol{w}}_c)^2} \;\ge\; 0. \tag{61}$$

Denoting by $\bar{\boldsymbol{w}}_c^{\perp} := \bar{\boldsymbol{w}}_c - (\hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}}_c)\, \hat{\boldsymbol{\xi}}$ the component of $\bar{\boldsymbol{w}}_c$ orthogonal to the boundary direction $\hat{\boldsymbol{\xi}}$, and recalling that $A_c = -\langle \mathbf{w}_c, \boldsymbol{\xi} \rangle_{\mathcal{M}}$ with $\boldsymbol{\xi} = (1, \hat{\boldsymbol{\xi}})$, we obtain the compact form

$$\alpha_{\boldsymbol{\xi}}(\mathbf{w}_c) := \frac{\|\bar{\boldsymbol{w}}_c^{\perp}\|^2}{\langle \mathbf{w}_c, \boldsymbol{\xi} \rangle_{\mathcal{M}}^2}, \tag{62}$$

with equality if and only if $\bar{\boldsymbol{w}}_c \parallel \hat{\boldsymbol{\xi}}$, which means that the prototype points is in the same direction as the boundary point or $\mathbf{w}_c = \mathbf{o}$. We collect these coefficients into a vector:

$$\boldsymbol{\alpha}(\boldsymbol{\xi}_i) := \big(\alpha_{\boldsymbol{\xi}_i}(\mathbf{w}_1),\; \ldots,\; \alpha_{\boldsymbol{\xi}_i}(\mathbf{w}_C)\big)^{\top} \in \mathbb{R}^C. \tag{63}$$

By the PERM loss definition in (8), the per-sample loss is

$$\ell_i(\mathbf{x}_i, y_i) = \psi\big(\Upsilon_{y_i}\, D\, \mathbf{v}_i\big), \tag{64}$$

where the score vector $\mathbf{v}_i = (v_{i,1}, \ldots, v_{i,C})^{\top} \in \mathbb{R}^C$ has entries $v_{i,c} = -d_{\mathbb{H}}(\mathbf{x}_i, \mathbf{w}_c)$. Substituting the second-order expansion (60) for each component:

$$\begin{aligned} v_{i,c} &= -d_{\mathbb{H}}(\mathbf{x}_i, \mathbf{w}_c) \\ &= -r_i - b_{\boldsymbol{\xi}_i}(\mathbf{w}_c) - \alpha_{\boldsymbol{\xi}_i}(\mathbf{w}_c)\, e^{-2r_i} + O(e^{-4r_i}). \end{aligned} \tag{65}$$

Recall from Definition 5.2 that $\mathbf{b}(\boldsymbol{\xi}_i)_c = -b_{\boldsymbol{\xi}_i}(\mathbf{w}_c)$. Therefore $-b_{\boldsymbol{\xi}_i}(\mathbf{w}_c) = \mathbf{b}(\boldsymbol{\xi}_i)_c$, and we can rewrite (65) in vector form:

$$\mathbf{v}_i = -r_i\, \mathbf{1} + \mathbf{b}(\boldsymbol{\xi}_i) - \boldsymbol{\alpha}(\boldsymbol{\xi}_i)\, e^{-2r_i} + O(e^{-4r_i}), \tag{66}$$

where $\mathbf{1} = (1, \ldots, 1)^{\top} \in \mathbb{R}^C$. Now we apply $\Upsilon_{y_i}\, D$ to both sides of (66). The key algebraic property of the relative-margin matrix D is

$$D\, \mathbf{1} = \mathbf{0}. \tag{67}$$

This holds because D extracts pairwise differences $v_y - v_c$: adding any constant to all logits leaves all differences unchanged. Therefore:

$$\begin{aligned} \Upsilon_{y_i}\, D\, \mathbf{v}_i &= -r_i\, \underbrace{\Upsilon_{y_i}\, D\, \mathbf{1}}_{=\mathbf{0}} + \Upsilon_{y_i}\, D\, \mathbf{b}(\boldsymbol{\xi}_i) \;-\; \Upsilon_{y_i}\, D\, \boldsymbol{\alpha}(\boldsymbol{\xi}_i)\, e^{-2r_i} \;+\; O(e^{-4r_i}) \\ &= \underbrace{\Upsilon_{y_i}\, D\, \mathbf{b}(\boldsymbol{\xi}_i)}_{=:\, \mathbf{z}_0 \,\in\, \mathbb{R}^{C-1}} \;-\; \underbrace{\Upsilon_{y_i}\, D\, \boldsymbol{\alpha}(\boldsymbol{\xi}_i)}_{=:\, \mathbf{h} \,\in\, \mathbb{R}^{C-1}}\, e^{-2r_i} \;+\; O(e^{-4r_i}). \end{aligned} \tag{68}$$

By Assumption A1, the template $\psi : \mathbb{R}^{C-1} \to \mathbb{R}$ is twice differentiable. We take the Taylor expansion of ψ around the point $\mathbf{z}_0 \in \mathbb{R}^{C-1}$ and step $\boldsymbol{\delta} \in \mathbb{R}^{C-1}$:

$$\psi(\mathbf{z}_0 + \boldsymbol{\delta}) = \psi(\mathbf{z}_0) + \langle \nabla\psi(\mathbf{z}_0), \boldsymbol{\delta} \rangle + O\big(\|\boldsymbol{\delta}\|^2\big). \tag{69}$$

From (68), the argument of ψ is $\mathbf{z}_0 + \boldsymbol{\delta}$, where the step is

$$\boldsymbol{\delta} := -\mathbf{h}\, e^{-2r_i} + O(e^{-4r_i}). \tag{70}$$

Its norm satisfies $\|\boldsymbol{\delta}\| = O(e^{-2r_i})$, and hence $\|\boldsymbol{\delta}\|^2 = O(e^{-4r_i})$.

Substitute (70) into (69):

$$\begin{aligned}\psi(\mathbf{z}_0 + \boldsymbol{\delta}) &= \psi(\mathbf{z}_0) + \big\langle \nabla\psi(\mathbf{z}_0),\ -\mathbf{h}\, e^{-2r_i} + O(e^{-4r_i}) \big\rangle + O(e^{-4r_i}) \\ &= \psi(\mathbf{z}_0) - \langle \nabla\psi(\mathbf{z}_0), \mathbf{h} \rangle\, e^{-2r_i} + O(e^{-4r_i}),\end{aligned} \tag{71}$$

where in the second line we used the fact that $\langle \nabla\psi(\mathbf{z}_0), O(e^{-4r_i}) \rangle = O(e^{-4r_i})$ since $\nabla\psi(\mathbf{z}_0)$ is a fixed vector, independent of r_i. Adding (64) and (71) we get:

$$\ell_i(\mathbf{x}_i, y_i) = \psi(\mathbf{z}_0) - \langle \nabla\psi(\mathbf{z}_0), \mathbf{h} \rangle\, e^{-2r_i} + O(e^{-4r_i}). \tag{72}$$

And we substitute the definition $\mathbf{z}_0 = \Upsilon_{y_i}\, D\, \mathbf{b}(\boldsymbol{\xi}_i)$:

$$\psi(\mathbf{z}_0) = \psi\big(\Upsilon_{y_i}\, D\, \mathbf{b}(\boldsymbol{\xi}_i)\big) = \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}_i, y_i), \tag{73}$$

which by (Definition 5.2), it is the Busemann loss. Then we substitute $\mathbf{h} = \Upsilon_{y_i}\, D\, \boldsymbol{\alpha}(\boldsymbol{\xi}_i)$ into the coefficient of e^{-2r_i}:

$$-\langle \nabla\psi(\mathbf{z}_0), \mathbf{h} \rangle = -\big\langle \nabla\psi\big(\Upsilon_{y_i}\, D\, \mathbf{b}(\boldsymbol{\xi}_i)\big),\ \Upsilon_{y_i}\, D\, \boldsymbol{\alpha}(\boldsymbol{\xi}_i) \big\rangle =: \mu(\boldsymbol{\xi}_i, y_i). \tag{74}$$

This is the *radial drift coefficient* defined in (14).

Then we obtain:

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

valid for all $r_i \geq R_0(\mathbf{W}) = \max_{c\in[C]} \log\big(4(w_{c,0} + \|\bar{\boldsymbol{w}}_c\|)\big)$. This completed the proof. □

C Proof of Section 6

C.1 Proof of Theorem 6.2

Theorem 6.2 (Radial dichotomy) *Under Assumption (A1), for any drift lower bound $\underline{\mu} > 0$, there exists $R_0 = R_0(\mathbf{W}, \underline{\mu}) > 0$ such that:*

(a) ***Boundary divergence.*** *If there exist $T_2 > T_1 \geq 0$ such that $r(T_1) \geq R_0$ and $\mu(\boldsymbol{\xi}(t), y) \geq \underline{\mu}$ on the time interval $I = [T_1, T_2]$, then $r(t) \geq \frac{1}{2}\log\big(e^{2r(T_1)} + 2\underline{\mu}(t - T_1)\big)$ for all $t \in I$. Moreover, if $r(T_1) \geq R_0$ and $\mu(\boldsymbol{\xi}(t), y) \geq \underline{\mu}$ for all $t \geq T_1$, then*

$$r(t) = \tfrac{1}{2}\log t + O(1), \qquad t \to \infty.$$

(b) ***Boundary repulsion.*** *If there exist $T_2 > T_1 \geq 0$ such that $\mu(\boldsymbol{\xi}(t), y) \leq -\underline{\mu}$ and $r(t) \geq R_0$ for all $t \in I = [T_1, T_2]$, then $r(t) \leq \frac{1}{2}\log\big(e^{2r(T_1)} - 2\underline{\mu}(t - T_1)\big)$ for all $t \in I$. Moreover, if $r(T_1) \geq R_0$ and $\mu(\boldsymbol{\xi}(t), y) \leq -\underline{\mu}$ whenever $r(t) \geq R_0$, then the first hitting time $T^* := \inf\{t \geq T_1 : r(t) \leq R_0\}$ satisfies*

$$T^* \leq T_1 + \frac{e^{2r(T_1)} - e^{2R_0}}{2\underline{\mu}} < \infty.$$

Proof. Firstly recall that any point $\mathbf{x} \in \mathbb{H}^n \setminus \{\mathbf{o}\}$ can be written in geodesic polar coordinates as $\mathbf{x} = (\cosh r,\, \sinh r\, \hat{\boldsymbol{\xi}})$ with $r > 0$ and $\hat{\boldsymbol{\xi}} \in \mathbb{S}^{n-1}$. In these coordinates, the Riemannian metric of $\mathbb{H}^n$ takes the form

$$ds^2 = dr^2 + \sinh^2(r)\, d\hat{\boldsymbol{\xi}}^2, \tag{75}$$

where $d\hat{\boldsymbol{\xi}}^2$ is the round metric on $\mathbb{S}^{n-1}$. For any smooth function $f : \mathbb{H}^n \to \mathbb{R}$, the Riemannian gradient in polar coordinates is

$$\operatorname{grad}_{\mathbb{H}^n} f = \frac{\partial f}{\partial r}\mathbf{e}_r + \frac{1}{\sinh^2(r)}\, \operatorname{grad}_{\mathbb{S}^{n-1}} f, \tag{76}$$

where $\mathbf{e}_r$ denotes the unit radial vector field. Since the prototypes $\{\mathbf{w}_c\}$ are fixed, the loss ℓ_i depends only on $\mathbf{x}_i$. The Riemannian gradient flow is

$$\dot{\mathbf{x}}_i(t) = -\operatorname{grad}_{\mathbb{H}^n}\, \ell_i(\mathbf{x}_i(t), y_i). \tag{77}$$

Projecting onto the radial direction using (76):

$$\dot{r}(t) = -\frac{\partial \ell_i}{\partial r}\bigg|_{\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\xi}}(t)}. \tag{78}$$

From Proposition 5.3, for $r \geq R_0(\mathbf{W})$:

$$\ell_i = \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y_i) + \mu(\boldsymbol{\xi}, y_i)\, e^{-2r} + O(e^{-4r}). \tag{79}$$

When we differentiate with respect to r while holding $\hat{\boldsymbol{\xi}}$ fixed. Since the first term $\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y_i)$ depends only on the direction $\boldsymbol{\xi} = (1, \hat{\boldsymbol{\xi}})$, not on r, so $\frac{\partial}{\partial r}\mathcal{L}^{\mathrm{BS}} = 0$. The second term $\mu = \mu(\boldsymbol{\xi}, y_i)$ also does not depend on r: $\frac{\partial}{\partial r}\big[\mu\, e^{-2r}\big] = \mu \cdot (-2)\, e^{-2r} = -2\mu\, e^{-2r}$. The remainder: $\frac{\partial}{\partial r}O(e^{-4r}) = O(e^{-4r})$. Substituting into (78):

$$\dot{r}(t) = 2\mu(t)\, e^{-2r(t)} + O(e^{-4r(t)}). \tag{80}$$

The sign of $\mu(t)$ determines whether the sample is pushed toward the boundary ($\mu > 0$) or pulled back toward the interior ($\mu < 0$).

Proof of part (a): boundary divergence.

Firstly, the $O(e^{-4r})$ in (80) can be bounded that there exists a constant $M = M(\mathbf{W}) > 0$, depending only on the prototypes, such that for all $r \geq R_0$ and all $\hat{\boldsymbol{\xi}} \in \mathbb{S}^{n-1}$:

$$|\dot{r}(t) - 2\mu(t)\, e^{-2r(t)}| \leq M\, e^{-4r(t)}. \tag{81}$$

The constant M can be extracted from the explicit second-order expansion in Proposition 5.3.

Assume $r(T_1) \geq R_0$ and $\mu(t) \geq \underline{\mu} > 0$ on some interval $I = [T_1, T_2]$. From (80) and (81):

$$\dot{r}(t) \geq 2\underline{\mu}\, e^{-2r(t)} - M\, e^{-4r(t)} = e^{-2r(t)}\big(2\underline{\mu} - M\, e^{-2r(t)}\big). \tag{82}$$

We can choose $R_0 \geq \frac{1}{2}\log(M/\underline{\mu})$ so that $M\, e^{-2R_0} \leq \underline{\mu}$. Then for all $r(t) \geq R_0$:

$$\dot{r}(t) \geq e^{-2r(t)}\big(2\underline{\mu} - \underline{\mu}\big) = \underline{\mu}\, e^{-2r(t)} > 0. \tag{83}$$

So $r(t)$ is strictly increasing whenever $r(t) \geq R_0$. Since $r(T_1) \geq R_0$, we have $r(t) \geq R_0$ for all $t \in I$, and $\dot{r}(t) > 0$ for all $t \in I$, which means once the trajectory is above R_0, it never comes back down on this interval.

Then we try to prove $r(t) \to \infty$ under the global condition. Suppose $\mu(t) \geq \underline{\mu} > 0$ for all $t \geq T_1$. Suppose for contradiction that $r(t) \to r_B < \infty$ as $t \to \infty$. Since $r(t)$ is increasing and $r(t) \to r_B$, we have $r(t) \leq r_B$ for all $t \geq T_1$, and hence $e^{-2r(t)} \geq e^{-2r_B}$. Substituting into (83) and integrating from T_1 to T:

$$r(T) - r(T_1) = \int_{T_1}^{T} \dot{r}(t)\, dt \geq \int_{T_1}^{T} \underline{\mu}\, e^{-2r(t)}\, dt \geq \underline{\mu}\, e^{-2r_B} \cdot (T - T_1). \tag{84}$$

Sending $T \to \infty$, the right-hand side diverges, so $r(T) \to \infty$, contradicting $r(T) \to r_B < \infty$.

Therefore $r(t) \to \infty$.

Then we prove the rate of divergence. Firstly, we prove the lower bound that $r(t) \geq \frac{1}{2}\log t + O(1)$.

We multiply both sides of (83) by $e^{2r(t)} > 0$:

$$e^{2r(t)}\, \dot{r}(t) \;\geq\; \underline{\mu}\, e^{-2r(t)}\, e^{2r(t)} \;=\; \underline{\mu}.$$

The left-hand side is a total derivative. By the chain rule, $\frac{d}{dt}\big[e^{2r(t)}\big] = 2\, e^{2r(t)}\, \dot{r}(t)$, so the inequality becomes

$$\frac{d}{dt}\big[e^{2r(t)}\big] \;\geq\; 2\underline{\mu}. \tag{85}$$

This says that $e^{2r(t)}$ grows at least as fast as a linear function. Integrating both sides from T_1 to T:

$$e^{2r(T)} - e^{2r(T_1)} \;=\; \int_{T_1}^{T} \frac{d}{dt}\big[e^{2r(t)}\big]\, dt \;\geq\; \int_{T_1}^{T} 2\underline{\mu}\, dt \;=\; 2\underline{\mu}\,(T - T_1). \tag{86}$$

Rearranging:

$$e^{2r(T)} \;\geq\; 2\underline{\mu}\,(T - T_1) + e^{2r(T_1)}. \tag{87}$$

Taking logarithms and dividing by 2:

$$r(T) \;\geq\; \frac{1}{2}\log\big(2\underline{\mu}\,(T - T_1) + e^{2r(T_1)}\big). \tag{88}$$

For $T \geq T_1 + 1$, since $2\underline{\mu}\,(T - T_1) + e^{2r(T_1)} \geq 2\underline{\mu}\,(T - T_1)$:

$$r(T) \;\geq\; \frac{1}{2}\log(2\underline{\mu}\,(T - T_1)) \;=\; \frac{1}{2}\log T + O(1). \tag{89}$$

Then we prove the upper bound that $r(t) \leq \frac{1}{2}\log t + O(1)$.

Since $\mu(t) \geq \underline{\mu} > 0$ and $\hat{\boldsymbol{\xi}}(t)$ stays in the compact set $\mathbb{S}^{n-1}$, there exists a constant $\bar{\mu} := \sup_{\hat{\boldsymbol{\xi}} \in \mathbb{S}^{n-1}} \mu(\boldsymbol{\xi}, y_i) < \infty$ (by continuity of μ on the compact set $\mathbb{S}^{n-1}$). From (80) and (81):

$$\dot{r}(t) \leq 2\bar{\mu}\, e^{-2r(t)} + M\, e^{-4r(t)}. \tag{90}$$

For $r(t) \geq R_0 \geq \frac{1}{2}\log(M/\bar{\mu})$, we have $M\, e^{-4r} \leq M\, e^{-2R_0}\, e^{-2r} \leq \bar{\mu}\, e^{-2r}$, therefore:

$$\dot{r}(t) \leq 3\bar{\mu}\, e^{-2r(t)}. \tag{91}$$

From (91), $\dot{r}(t) \leq 3\bar{\mu}\, e^{-2r(t)}$. Similarly, multiplying both sides by $e^{2r(t)}$:

$$e^{2r(t)}\, \dot{r}(t) \;\leq\; 3\bar{\mu}, \tag{92}$$

hence for $T \geq T_1 + 1$:

$$r(T) \;\leq\; \frac{1}{2}\log\!\big(6\bar{\mu}\,(T - T_1) + e^{2r(T_1)}\big) \;\leq\; \frac{1}{2}\log T + O(1). \tag{93}$$

Finally, from (89) and (93), there exist constants c_1, c_2 (depending on $\underline{\mu}$, $\bar{\mu}$, $r(T_1)$, and T_1) such that for all $t \geq T_1 + 1$:

$$\frac{1}{2}\log t + c_1 \;\leq\; r(t) \;\leq\; \frac{1}{2}\log t + c_2. \tag{94}$$

This gives $r(t) = \frac{1}{2}\log t + O(1)$, proving part (a).

Proof of part (b): boundary repulsion.

Assume $r(T_1) \geq R_0$ and $\mu(t) \leq -\underline{\mu} < 0$ whenever $r(t) \geq R_0$.

From (80) and (81), whenever $r(t) \geq R_0$:

$$\dot{r}(t) \leq -2\underline{\mu}\, e^{-2r(t)} + M\, e^{-4r(t)} = e^{-2r(t)}\big(-2\underline{\mu} + M\, e^{-2r(t)}\big). \tag{95}$$

Also, we can choose $R_0 \geq \frac{1}{2}\log(M/\underline{\mu})$ so that $M\, e^{-2R_0} \leq \underline{\mu}$. Then for all $r(t) \geq R_0$:

$$\dot{r}(t) \leq -\underline{\mu}\, e^{-2r(t)} < 0. \tag{96}$$

So $r(t)$ is *strictly decreasing* whenever $r(t) \geq R_0$. Similarly, multiplying both sides by $e^{2r(t)}$:

$$2\dot{r}(t)\, e^{2r(t)} \leq 2\big(-\underline{\mu}\, e^{-2r(t)}\big)\, e^{2r(t)} = -2\underline{\mu}. \tag{97}$$

Integrating from T_1 to t:

$$e^{2r(t)} \leq e^{2r(T_1)} - 2\underline{\mu}\,(t - T_1). \tag{98}$$

The condition $e^{2r(t)} \leq e^{2R_0}$ (i.e., $r(t) \leq R_0$) is achieved when

$$e^{2r(T_1)} - 2\underline{\mu}\,(t - T_1) \leq e^{2R_0}, \qquad t \geq T^* := T_1 + \frac{e^{2r(T_1)} - e^{2R_0}}{2\underline{\mu}}. \tag{99}$$

Since $r(T_1) \geq R_0$, we have $T^* \geq T_1$, and $T^* < \infty$. So $r(t)$ drops below R_0 no later than time T^*.

This completed the proof of part (b), and with it, we completed the proof of the whole theorem. $\square$

C.2 Proof of Theorem 6.4

Theorem 6.4 *Under Assumptions (A1) and (A2), with ψ real analytic, suppose $\boldsymbol{\xi}(T_0)$ lies in the Busemann gradient-flow basin of a nondegenerate local minimum $\boldsymbol{\xi}^*$, with $\mu(\boldsymbol{\xi}(T_0), y) > 0$. For a fixed initial angular direction, there exist $R_0 > 0$ and constants $C, \lambda, \kappa > 0$ such that, for any $R \geq R_0$, if $r(T_0) \geq 2R$, then there exist finite times $T_0 < T_1 < T_2 < T_3$ satisfying:*

(a) Large-radius excursion. *The trajectory initially moves outward, $\dot{r}(T_0) > 0$, while $\|\dot{\hat{\boldsymbol{\xi}}}(t)\| = O(e^{-2r(t)})$ in the large-radius regime.*

(b) Angular alignment and drift reversal. *At T_1, $\mu(\boldsymbol{\xi}(T_1), y) = 0$, and $\mu(\boldsymbol{\xi}(t), y) < 0$ for $t \in (T_1, T_3]$. At T_2, the trajectory first enters the sufficiently aligned region $\mathcal{A}(\boldsymbol{\xi}^*)$.*

(c) Slow radial return. *For $t \in [T_2, T_3]$, $\dot{r}(t) < 0$ and $|\dot{r}(t)| = O(e^{-2r(t)})$, with $r(T_3) = R$. For any $\varepsilon > 0$, the returning trajectory reaches*

$$d_{\mathbb{S}^{n-1}}(\boldsymbol{\xi}(T_3), \boldsymbol{\xi}^*) \leq \varepsilon$$

in finite time by choosing R sufficiently large.

Proof. Firstly recall that, in geodesic polar coordinates $x = (\cosh r, \sinh r\, \hat{\boldsymbol{\xi}})$, the Riemannian gradient flow satisfies

$$\dot{r} = -\partial_r \ell, \qquad \dot{\hat{\boldsymbol{\xi}}} = -\frac{1}{\sinh^2 r} \operatorname{grad}_{\mathbb{S}^{n-1}} \ell. \tag{100}$$

From Proposition 5.3, $\ell(x, y) = \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) + \mu(\boldsymbol{\xi}, y) e^{-2r} + O(e^{-4r})$. Differentiating gives

$$\dot{r} = 2\mu(\boldsymbol{\xi}, y) e^{-2r} + O(e^{-4r}), \qquad \dot{\hat{\boldsymbol{\xi}}} = -\frac{1}{\sinh^2 r} \left(\operatorname{grad}_{\mathbb{S}^{n-1}} \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) + O(e^{-2r}) \right). \tag{101}$$

We first establish the sign of the radial drift at a Busemann critical direction. Write $g_c = -\partial_c \psi > 0$ at the Busemann margins and $b_c = b_{\boldsymbol{\xi}}(w_c)$. By the definitions of μ and $\mathcal{L}^{\mathrm{BS}}$,

$$\mu = \sum_{c \neq y} g_c \left(\| \operatorname{grad}_{\mathbb{S}^{n-1}} b_y \|^2 - \| \operatorname{grad}_{\mathbb{S}^{n-1}} b_c \|^2 \right), \ \operatorname{grad}_{\mathbb{S}^{n-1}} \mathcal{L}^{\mathrm{BS}} = \sum_{c \neq y} g_c (\operatorname{grad}_{\mathbb{S}^{n-1}} b_y - \operatorname{grad}_{\mathbb{S}^{n-1}} b_c).$$

Consequently, at any Busemann critical direction $\boldsymbol{\xi}^*$,

$$\mu(\boldsymbol{\xi}^*, y) = -\sum_{c \neq y} g_c \left\| \operatorname{grad}_{\mathbb{S}^{n-1}} b_{\boldsymbol{\xi}^*}(w_c) - \operatorname{grad}_{\mathbb{S}^{n-1}} b_{\boldsymbol{\xi}^*}(w_y) \right\|^2 < 0. \tag{102}$$

Indeed, equality would require identical Busemann gradients for all prototypes, placing them on a common hyperbolic geodesic asymptotic to $\boldsymbol{\xi}^*$, contrary to Assumption (A2). Since $\boldsymbol{\xi}^*$ is a nondegenerate local minimum, we may choose $\mathcal{A}(\boldsymbol{\xi}^*)$ sufficiently small that $\mu(\boldsymbol{\xi}, y) \leq -m < 0$ and $\operatorname{Hess}_{\mathbb{S}^{n-1}} \mathcal{L}^{\mathrm{BS}} \succeq \lambda I$ throughout this region, for some $m, \lambda > 0$.

We next compare the angular dynamics with Busemann gradient flow. Introduce the rescaled time $s(t) = \int_{T_0}^{t} \sinh^{-2} r(\tau)\, d\tau$. Equation (101) becomes

$$\frac{dr}{ds} = \frac{1}{2}\mu(\boldsymbol{\xi}, y) + O(e^{-2r}), \qquad \frac{d\hat{\boldsymbol{\xi}}}{ds} = -\operatorname{grad}_{\mathbb{S}^{n-1}} \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) + O(e^{-2r}). \tag{103}$$

Let $\boldsymbol{\zeta}(s)$ solve the unperturbed Busemann gradient flow starting from $\boldsymbol{\xi}(T_0)$. By the basin-of-attraction hypothesis, $\boldsymbol{\zeta}(s) \to \boldsymbol{\xi}^*$. Hence there exists a finite $S > 0$ such that $d_{\mathbb{S}^{n-1}}(\boldsymbol{\zeta}(S), \boldsymbol{\xi}^*) < \theta_*/2$.

Since μ is uniformly bounded, the rescaled radial velocity is bounded by some $M > 0$ in the large-radius regime. Thus $r(s) \geq r(T_0) - Ms$. For sufficiently large R, the trajectory remains above R on $[0, S]$, and the angular perturbation is uniformly $O(e^{-2R})$. By smoothness and a Gronwall estimate,

$$\sup_{0 \leq s \leq S} d_{\mathbb{S}^{n-1}}(\boldsymbol{\xi}(s), \boldsymbol{\zeta}(s)) \leq C_S e^{-2R}. \tag{104}$$

Choosing R sufficiently large, the full angular trajectory therefore enters $\mathcal{A}(\boldsymbol{\xi}^*)$ within finite rescaled time, while remaining at large radius.

We now prove the three parts.

Proof of part (a): large-radius excursion. At T_0, $\mu(\boldsymbol{\xi}(T_0), y) > 0$. From (101), for sufficiently large $r(T_0)$, the positive leading term dominates the remainder, giving $\dot{r}(T_0) > 0$. Thus the trajectory initially moves outward.

Moreover, smoothness of $\mathcal{L}^{\mathrm{BS}}$ and compactness of $\mathbb{S}^{n-1}$ imply that its spherical gradient is uniformly bounded. Hence

$$\|\dot{\hat{\boldsymbol{\xi}}}(t)\| = O(\sinh^{-2} r(t)) = O(e^{-2r(t)})$$

throughout the large-radius regime. Intermediate drift reversals are allowed. This proves part (a).

Proof of part (b): angular alignment and drift reversal. Let T_2 be the first entrance time into $\mathcal{A}(\boldsymbol{\xi}^*)$. The finite-time approximation above shows that $T_2 < \infty$ and, since $s(T_2) \le S$,

$$r(T_2) \ge 2R - MS > R$$

for sufficiently large R.

By (102) and the definition of $\mathcal{A}(\boldsymbol{\xi}^*)$, $\mu(\boldsymbol{\xi}(T_2), y) \le -m < 0$. Together with $\mu(\boldsymbol{\xi}(T_0), y) > 0$, continuity gives at least one zero crossing on (T_0, T_2). Let T_1 be the last such crossing. Then $T_0 < T_1 < T_2$ and $\mu(\boldsymbol{\xi}(t), y) < 0$ for $t \in (T_1, T_2]$.

Furthermore, positive definiteness of the local Busemann Hessian and (103) yield

$$\frac{d}{ds} d_{\mathbb{S}^{n-1}}(\boldsymbol{\xi}, \boldsymbol{\xi}^*) \le -\lambda d_{\mathbb{S}^{n-1}}(\boldsymbol{\xi}, \boldsymbol{\xi}^*) + Ce^{-2r}. \tag{105}$$

At zero distance, the inequality is understood in the upper Dini derivative sense. Choosing R so that $Ce^{-2R} < \lambda\theta_*$ makes $\mathcal{A}(\boldsymbol{\xi}^*)$ forward invariant while $r \ge R$. Therefore $\mu < 0$ persists throughout the subsequent large-radius return. This proves part (b).

Proof of part (c): slow radial return. Within $\mathcal{A}(\boldsymbol{\xi}^*)$, $\mu(\boldsymbol{\xi}, y) \le -m < 0$. From (103), for sufficiently large R,

$$-M \le \frac{dr}{ds} \le -\frac{m}{4} < 0$$

whenever $r \ge R$. Thus the radius decreases strictly and reaches R in finite rescaled time. Let T_3 denote the first hitting time. Since r remains bounded on $[T_2, T_3]$, the corresponding original time is also finite. Moreover, (101) gives $|\dot{r}(t)| = O(e^{-2r(t)})$.

It remains to establish the angular convergence. Since r decreases on $[T_2, T_3]$, $e^{-2r(t)} \le e^{-2R}$. Integrating (105) therefore gives

$$d_{\mathbb{S}^{n-1}}(\boldsymbol{\xi}(t), \boldsymbol{\xi}^*) \le \theta_* e^{-\lambda(s(t)-s(T_2))} + \frac{C}{\lambda} e^{-2R}. \tag{106}$$

On the other hand, the lower bound on the radial velocity and $r(T_2) \ge 2R - MS$ imply

$$s(T_3) - s(T_2) \ge \frac{r(T_2) - R}{M} \ge \frac{R - MS}{M}.$$

Substituting into (106) yields

$$d_{\mathbb{S}^{n-1}}(\boldsymbol{\xi}(T_3), \boldsymbol{\xi}^*) \le \theta_* e^{\lambda S} e^{-\lambda R/M} + \frac{C}{\lambda} e^{-2R} \le C' e^{-\kappa R}, \tag{107}$$

where $\kappa = \min\{\lambda/M, 2\} > 0$ and $C' > 0$ is independent of R.

Consequently, for any $\varepsilon > 0$, choosing R sufficiently large gives

$$d_{\mathbb{S}^{n-1}}(\boldsymbol{\xi}(T_3), \boldsymbol{\xi}^*) \le \varepsilon.$$

Thus the returning trajectory approaches the Busemann local minimum to arbitrary accuracy in finite time, despite the decreasing radius. This completes the proof. □

C.3 Proof of Theorem 6.5

Theorem 6.5 (Angular convergence to Busemann minimizers in the escaping regime) *Under Assumptions (A1) and (A3), the Riemannian gradient flow satisfies*

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

*If Assumption (A3) holds for all $i \in [N]$, then $(\boldsymbol{\xi}_1^*, \dots, \boldsymbol{\xi}_N^*)$ is a **global minimizer** of the Busemann risk $\mathcal{R}^{\mathrm{BS}}_{\mathbf{W}^\infty} = \frac{1}{N}\sum_{i=1}^N \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}_i, y_i)$ on $(\partial \mathbb{H}^n)^N$.*

Proof. We first consider a single sample with label y. Under Assumption (A3), the class scores are $v_c(\mathbf{x}) = -b_{w_c^\infty}(\mathbf{x})$, where the ideal prototypes are pairwise distinct. For each ideal prototype $w_c^\infty = (1, \hat{w}_c)$, consider $w_c(R) = (\cosh R, \sinh R\, \hat{w}_c) \in \mathbb{H}^n$. By the radial Busemann expansion,

$$d_{\mathbb{H}^n}(x, w_c(R)) = R + b_{w_c^\infty}(x) + O(e^{-2R}).$$

Since all prototypes have the same radius R, the common term $-R$ cancels in the relative margins of any PERM loss. Consequently, the limiting class scores are

$$v_c(x) = -b_{w_c^\infty}(x).$$

Recall that $\mathbf{x} = (x_0, \bar{\mathbf{x}}) \in \mathbb{H}^n$. By a spatial rotation, we may assume that the true-class ideal prototype is $w_y^\infty = (1, 1, 0, \dots, 0)$. Write $\bar{\mathbf{x}} = (x_1, \mathbf{u})$ and $s = x_0 - x_1 > 0$. The hyperboloid constraint $\langle \mathbf{x}, \mathbf{x} \rangle_M = -1$ then gives $x_0 + x_1 = (1 + \|\mathbf{u}\|^2)/s$.

For each $c \neq y$, write $w_c^\infty = (1, \hat{w}_{c,1}, \hat{\mathbf{w}}_{c,\perp})$ and define $a_c = \hat{\mathbf{w}}_{c,\perp}/(1 - \hat{w}_{c,1})$. Since the ideal prototypes are distinct, $1 - \hat{w}_{c,1} > 0$. Using the Minkowski inner product, we obtain $-\langle \mathbf{x}, w_y^\infty \rangle_M = s$ and

$$m_c(\mathbf{x}) = b_{w_c^\infty}(\mathbf{x}) - b_{w_y^\infty}(\mathbf{x}) = \log \frac{1 + \|\mathbf{u} - s a_c\|^2}{s^2(1 + \|a_c\|^2)}, \qquad c \neq y. \tag{108}$$

By Assumption (A1), $\ell(\mathbf{x}, y) = \psi(m)$. Let $q_c = -\partial_c \psi(m) > 0$. For any ideal prototype w_c^∞, the Riemannian gradient of its Busemann function is $\operatorname{grad}_{\mathbb{H}^n} b_{w_c^\infty}(\mathbf{x}) = \mathbf{x} - w_c^\infty/(-\langle \mathbf{x}, w_c^\infty \rangle_M)$. Consequently, the full Riemannian gradient flow satisfies

$$\begin{aligned} \dot{s} &= -2s \sum_{c \neq y} \frac{q_c}{1 + \|\mathbf{u} - s a_c\|^2} < 0, \\ \dot{\mathbf{u}} &= -2s \sum_{c \neq y} \frac{q_c a_c}{1 + \|\mathbf{u} - s a_c\|^2}. \end{aligned} \tag{109}$$

Thus $s(t)$ is strictly decreasing.

We first verify global existence. Since $0 < s(t) \leq s(0)$, the margins in (108) are uniformly bounded below. By convexity, nonnegativity, and coordinatewise monotonicity of ψ, we have $q_c \leq \psi(m - e_c)$, where e_c is the corresponding coordinate vector. The lower bounds on the margins therefore imply that all q_c remain bounded. Moreover, since each Busemann function has unit Riemannian gradient norm, $\|\operatorname{grad}_{\mathbb{H}^n} m_c\| \leq 2$. Hence $\|\operatorname{grad}_{\mathbb{H}^n} \ell\|$ is uniformly bounded. By completeness of $\mathbb{H}^n$, the gradient flow exists for all $t \geq 0$.

We next establish radial divergence. Dividing the two equations in (109) yields

$$\frac{d\mathbf{u}}{ds} = \frac{\displaystyle\sum_{c \neq y} \frac{q_c a_c}{1 + \|\mathbf{u} - s a_c\|^2}}{\displaystyle\sum_{c \neq y} \frac{q_c}{1 + \|\mathbf{u} - s a_c\|^2}}. \tag{110}$$

The right-hand side is a weighted average of the fixed vectors a_c, and is therefore uniformly bounded. Since $s(t)$ is positive and decreasing, equation (110) implies that $\mathbf{u}(t)$ remains bounded and converges to a finite limit.

Suppose, for contradiction, that $s(t) \to s_\infty > 0$. Then s and $\mathbf{u}$ remain in a compact region, and so does $\mathbf{x}$ by the hyperboloid constraint. Since $q_c > 0$ and the vector field is continuous, the first

equation in (109) has a strictly negative upper bound on this compact region. This contradicts the convergence of $s(t)$ to a positive limit. Therefore $s(t) \to 0$.

By the hyperboloid constraint, $x_0 + x_1 = (1 + \|\mathbf{u}\|^2)/s \to \infty$, while $x_0 - x_1 = s \to 0$. Thus $x_0 \to \infty$, $x_1/x_0 \to 1$, and $\mathbf{u}/x_0 \to 0$. Recalling $x_0 = \cosh r$ and $\bar{\mathbf{x}} = \sinh r\, \hat{\boldsymbol{\xi}}$, we conclude that

$$r(t) \longrightarrow \infty, \qquad \boldsymbol{\xi}(t) \longrightarrow \boldsymbol{\xi}^* \in \partial\mathbb{H}^n, \quad t \to \infty. \tag{111}$$

It remains to establish the minimizing property of $\boldsymbol{\xi}^*$. Recall that the Busemann loss retains the original PERM form $\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) = \psi(\Upsilon_y D\mathbf{b}(\boldsymbol{\xi}))$. For directions distinct from the ideal prototypes, the relative margins are $\log((1 - \hat{\boldsymbol{\xi}}^\top \hat{w}_c)/(1 - \hat{\boldsymbol{\xi}}^\top \hat{w}_y))$, $c \neq y$.

As $\boldsymbol{\xi}$ approaches the true-class ideal direction, all relative margins diverge to $+\infty$. Since ψ is nonnegative and strictly decreasing in each coordinate, $L_* = \lim_{a\to\infty} \psi(a\mathbf{1})$ exists and equals the infimum of ψ. The Busemann loss therefore extends continuously to $\boldsymbol{\xi}^*$ with value L_*.

For any direction with finite relative margins, strict coordinatewise monotonicity gives $\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) > L_*$. At any other ideal prototype direction, at least one relative margin tends to $-\infty$, while the remaining margins stay bounded above. Monotonicity then also excludes the value L_*. Consequently,

$$\boldsymbol{\xi}^* \in \arg\min_{\boldsymbol{\xi}\in\partial\mathbb{H}^n} \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y). \tag{112}$$

Finally, restoring the sample index i, the same argument applies independently to every trajectory. Since the Busemann risk is separable across samples, $\mathcal{R}^{\mathrm{BS}}_{\mathbf{W}^\infty} = \frac{1}{N}\sum_{i=1}^N \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}_i, y_i)$, the joint limit $(\boldsymbol{\xi}_1^*, \ldots, \boldsymbol{\xi}_N^*)$ is a global minimizer on $(\partial\mathbb{H}^n)^N$. This completes the proof. □

C.4 Proof of Theorem 6.6

Theorem 6.6 *Under Assumption (A1), let $\{\mathbf{w}_c\}_{c=1}^C$ be fixed prototypes, either all in $\mathbb{H}^n$ or all in $\partial\mathbb{H}^n$, with distance or Busemann scores, respectively. For a test point $\mathbf{x} = (\cosh r, \sinh r\,\hat{\boldsymbol{\xi}})$ with $\boldsymbol{\xi} = (1, \hat{\boldsymbol{\xi}}) \in \partial\mathbb{H}^n$, define $\hat{y}(\mathbf{x}) := \arg\min_{y\in[C]} \psi(\Upsilon_y D\,\mathbf{v}(\mathbf{x}))$. If $\arg\min_y \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y)$ is unique, then for all sufficiently large r, $\hat{y}(\mathbf{x}) = \arg\min_{y\in[C]} \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) = \arg\min_{c\in[C]} b_{\boldsymbol{\xi}}(\mathbf{w}_c)$. For ideal prototypes, this equality holds for all $r > 0$.*

Moreover, the decision boundary between classes c and c' on $\partial\mathbb{H}^n \cong \mathbb{S}^{n-1}$ is

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

which is a subset of the intersection of $\mathbb{S}^{n-1}$ with an affine hyperplane in $\mathbb{R}^n$. For ideal prototypes, $w_{c,0} = 1$, so the corresponding spherical sections are great subspheres.

Proof. We first consider the case of finite prototypes. Throughout this part, fix a boundary direction $\boldsymbol{\xi} = (1, \hat{\boldsymbol{\xi}}) \in \partial\mathbb{H}^n,\ \hat{\boldsymbol{\xi}} \in \mathbb{S}^{n-1}$, with $\mathbf{x} = (\cosh r, \sinh r\,\hat{\boldsymbol{\xi}})$. We write

$$\beta_c := b_{\boldsymbol{\xi}}(\mathbf{w}_c), \qquad c \in [C], \tag{113}$$

for the Busemann value of the c-th prototype.

For each candidate label $y \in [C]$, the PERM loss at the test point is

$$\psi(\Upsilon_y D\mathbf{v}(\mathbf{x}))\,, \qquad v_c(\mathbf{x}) = -d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}_c). \tag{114}$$

By the radial Busemann expansion (9), for each prototype $\mathbf{w}_c$,

$$d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}_c) = r + b_{\boldsymbol{\xi}}(\mathbf{w}_c) + O(e^{-2r}) = r + \beta_c + O(e^{-2r}). \tag{115}$$

By the definition of the Busemann loss,

$$\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) = \psi\left(\Upsilon_y D \begin{pmatrix} -\beta_1 \\ \vdots \\ -\beta_C \end{pmatrix}\right). \tag{116}$$

Since ψ is smooth and the number of labels is finite, we have

$$\psi(\Upsilon_y D\mathbf{v}(\mathbf{x})) = \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) + O(e^{-2r}), \tag{117}$$

uniformly over $y \in [C]$. Thus, there exist constants $C_0 > 0$ and $R_0 > 0$ such that, for all $r \ge R_0$ and all $y \in [C]$,

$$\big|\psi(\Upsilon_y D\mathbf{v}(\mathbf{x})) - \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y)\big| \le C_0 e^{-2r}. \tag{118}$$

Let

$$y^* = \arg\min_{y\in[C]} \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y). \tag{119}$$

By assumption, this minimizer is unique. Therefore, the gap between the best label and the second best label is strictly positive:

$$\Delta := \min_{y\neq y^*} \big[\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) - \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y^*)\big] > 0. \tag{120}$$

Choose $R_1 \ge R_0$ so large that

$$2C_0 e^{-2r} < \Delta \qquad \text{for all } r \ge R_1. \tag{121}$$

Then, for any $y \neq y^*$ and any $r \ge R_1$,

$$\begin{aligned} &\psi(\Upsilon_y D\mathbf{v}(\mathbf{x})) - \psi(\Upsilon_{y^*} D\mathbf{v}(\mathbf{x})) \\ &\quad \ge \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) - \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y^*) - 2C_0 e^{-2r} \\ &\quad \ge \Delta - 2C_0 e^{-2r} > 0. \end{aligned} \tag{122}$$

Thus, for all sufficiently large r,

$$\arg\min_{y\in[C]} \psi(\Upsilon_y D\mathbf{v}(\mathbf{x})) = \arg\min_{y\in[C]} \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y). \tag{123}$$

We now prove that

$$\arg\min_{y\in[C]} \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) = \arg\min_{c\in[C]} b_{\boldsymbol{\xi}}(\mathbf{w}_c). \tag{124}$$

Recall that $\beta_c = b_{\boldsymbol{\xi}}(\mathbf{w}_c)$. For a candidate label y, the vector of Busemann margins has the form

$$\mathbf{m}^y = \big(\beta_c - \beta_y\big)_{c\neq y} \in \mathbb{R}^{C-1}, \tag{125}$$

up to a permutation of coordinates. Hence

$$\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) = \psi(\mathbf{m}^y). \tag{126}$$

Let

$$y^* \in \arg\min_{c\in[C]} \beta_c. \tag{127}$$

We compare the loss of y^* with the loss of any other label $y \neq y^*$.

If $\beta_y > \beta_{y^*}$, then the margin vector for y^* is strictly larger than the margin vector for y, after a possible permutation of coordinates. Indeed, for every $c \neq y, y^*$,

$$\beta_c - \beta_{y^*} > \beta_c - \beta_y, \tag{128}$$

because $\beta_y > \beta_{y^*}$. For the remaining coordinate pair, we have

$$\beta_y - \beta_{y^*} > \beta_{y^*} - \beta_y. \tag{129}$$

Thus $\mathbf{m}^{y^*}$ is strictly larger than $\mathbf{m}^y$ componentwise, up to permutation.

Since ψ is a PERM template, it is invariant under permutations of its coordinates. Since the loss is totally regular, $\nabla\psi \prec 0$, so ψ is strictly decreasing in each coordinate. Therefore,

$$\psi(\mathbf{m}^{y^*}) < \psi(\mathbf{m}^y), \tag{130}$$

or equivalently,

$$\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y^*) < \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y). \tag{131}$$

Hence any label y with $\beta_y > \beta_{y^*}$ cannot minimize the Busemann loss.

If instead $\beta_y = \beta_{y^*}$, then the two margin vectors $\mathbf{m}^y$ and $\mathbf{m}^{y^*}$ differ only by a permutation of coordinates. By permutation invariance of ψ,

$$\psi(\mathbf{m}^y) = \psi(\mathbf{m}^{y^*}). \tag{132}$$

Therefore the minimizers of the Busemann loss are exactly the labels whose Busemann value β_y is minimal. This proves (124). In particular, uniqueness of $\arg\min_y \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y)$ is equivalent to uniqueness of $\arg\min_c b_{\boldsymbol{\xi}}(\mathbf{w}_c)$.

Combining (123) and (124), we obtain, for all sufficiently large r,

$$\hat{y}(\mathbf{x}) = \arg\min_{y\in[C]} \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) = \arg\min_{c\in[C]} b_{\boldsymbol{\xi}}(\mathbf{w}_c). \tag{133}$$

Let

$$c^* = \arg\min_{c\in[C]} b_{\boldsymbol{\xi}}(\mathbf{w}_c). \tag{134}$$

By uniqueness, the Busemann gap

$$\Gamma := \min_{c\neq c^*} [b_{\boldsymbol{\xi}}(\mathbf{w}_c) - b_{\boldsymbol{\xi}}(\mathbf{w}_{c^*})] \tag{135}$$

is strictly positive. Again by the radial Busemann expansion, $d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}_c) = r + b_{\boldsymbol{\xi}}(\mathbf{w}_c) + O(e^{-2r})$ for every $c \in [C]$. Since there are only finitely many prototypes, the error can be bounded uniformly over c. Hence there exist constants $C_1 > 0$ and $R_2 > 0$ such that, for all $r \geq R_2$,

$$|d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}_c) - r - b_{\boldsymbol{\xi}}(\mathbf{w}_c)| \leq C_1 e^{-2r} \qquad \text{for all } c \in [C]. \tag{136}$$

Choose $R_3 \geq R_2$ such that

$$2C_1 e^{-2r} < \Gamma \qquad \text{for all } r \geq R_3. \tag{137}$$

Then, for any $c \neq c^*$ and any $r \geq R_3$,

$$\begin{aligned} d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}_c) - d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}_{c^*}) &\geq b_{\boldsymbol{\xi}}(\mathbf{w}_c) - b_{\boldsymbol{\xi}}(\mathbf{w}_{c^*}) - 2C_1 e^{-2r} \\ &\geq \Gamma - 2C_1 e^{-2r} > 0. \end{aligned} \tag{138}$$

Thus

$$\arg\min_{c\in[C]} d_{\mathbb{H}}(\mathbf{x}, \mathbf{w}_c) = \arg\min_{c\in[C]} b_{\boldsymbol{\xi}}(\mathbf{w}_c) \tag{139}$$

for all sufficiently large r. Therefore the asymptotic PERM prediction also coincides with the nearest prototype prediction.

We next consider ideal prototypes $\mathbf{w}_c = w_c^\infty \in \partial\mathbb{H}^n$. Their class scores satisfy

$$v_c(\mathbf{x}) = -b_{w_c^\infty}(\mathbf{x}) = -\log\Big(\cosh r - \sinh r\, \hat{\boldsymbol{\xi}}^\top \bar{\boldsymbol{w}}_c\Big). \tag{140}$$

For every $r > 0$, the ordering of these scores is determined exactly by $\hat{\boldsymbol{\xi}}^\top \bar{\boldsymbol{w}}_c$. The same PERM comparison argument above therefore gives $\hat{y}(\mathbf{x}) = \arg\max_{c\in[C]} \hat{\boldsymbol{\xi}}^\top \bar{\boldsymbol{w}}_c$. Moreover, the extended Busemann values satisfy $b_{\boldsymbol{\xi}}(w_c^\infty) = \log(1 - \hat{\boldsymbol{\xi}}^\top \bar{\boldsymbol{w}}_c)$ away from the ideal prototype directions. Their ordering is identical, and hence

$$\hat{y}(\mathbf{x}) = \arg\min_{y\in[C]} \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) = \arg\min_{c\in[C]} b_{\boldsymbol{\xi}}(w_c^\infty), \qquad r > 0. \tag{141}$$

The equality at ideal prototype directions follows from the continuous extension of the Busemann loss. Thus, unlike the finite case, the classification rule is exact for every positive radius.

Finally, we characterize the decision boundary. For either prototype configuration, a boundary direction separating classes c and c' must satisfy

$$b_{\boldsymbol{\xi}}(\mathbf{w}_c) = b_{\boldsymbol{\xi}}(\mathbf{w}_{c'}), \tag{142}$$

with the common value minimal among all classes. Using the explicit formula

$$b_{\boldsymbol{\xi}}(\mathbf{w}_c) = \log\big(w_{c,0} - \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}}_c\big), \tag{143}$$

the equality becomes

$$\log\big(w_{c,0} - \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}}_c\big) = \log\big(w_{c',0} - \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}}_{c'}\big). \tag{144}$$

The quantities inside the logarithms are positive for finite prototypes and for ideal prototypes away from their own directions. Since distinct ideal prototypes cannot be tied at either endpoint, the logarithms are well defined at any pairwise boundary. Their injectivity gives

$$w_{c,0} - \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}}_c = w_{c',0} - \hat{\boldsymbol{\xi}} \cdot \bar{\boldsymbol{w}}_{c'}. \tag{145}$$

Rearranging gives

$$\hat{\boldsymbol{\xi}} \cdot (\bar{\boldsymbol{w}}_c - \bar{\boldsymbol{w}}_{c'}) = w_{c,0} - w_{c',0}. \tag{146}$$

The additional minimum condition requires $b_{\boldsymbol{\xi}}(\mathbf{w}_c) \leq b_{\boldsymbol{\xi}}(\mathbf{w}_k)$ for every $k \in [C]$. Therefore,

$$\mathcal{B}_{c,c'} = \big\{\hat{\boldsymbol{\xi}} \in \mathbb{S}^{n-1} : \hat{\boldsymbol{\xi}} \cdot (\bar{\boldsymbol{w}}_c - \bar{\boldsymbol{w}}_{c'}) = w_{c,0} - w_{c',0},\ \ b_{\boldsymbol{\xi}}(\mathbf{w}_c) \leq b_{\boldsymbol{\xi}}(\mathbf{w}_k), \quad \forall k \in [C]\big\}. \tag{147}$$

This is a subset of the intersection of $\mathbb{S}^{n-1}$ with the affine hyperplane

$$\big\{z \in \mathbb{R}^n : z \cdot (\bar{\boldsymbol{w}}_c - \bar{\boldsymbol{w}}_{c'}) = w_{c,0} - w_{c',0}\big\}. \tag{148}$$

For ideal prototypes, $w_{c,0} = w_{c',0} = 1$, so the affine hyperplane passes through the origin and its spherical section is a great subsphere.

Then we completed the proof. □

D Experiments

In this appendix, we present the numerical setup, integration procedures, ablations, and additional diagnostics for the radial and angular dynamics. We first integrate the radial and limiting Busemann angular equations separately to examine their asymptotic behavior. We then study the full coupled Riemannian gradient flow, focusing on the evolution of the drift coefficient μ and the three-phase dynamics characterized by Theorem 6.4. Finally, we report classification results from neural network training.

D.1 Setup

Throughout the numerical experiments, we use the hyperboloid model

$$\mathbb{H}^n = \left\{ \mathbf{x} \in \mathbb{R}^{n+1} : \langle \mathbf{x}, \mathbf{x} \rangle_M = -1,\ x_0 > 0 \right\}, \tag{149}$$

with $n = 2$ and constant sectional curvature -1. We represent each point in polar coordinates $\mathbf{x} = (\cosh r, \sinh r\, \hat{\boldsymbol{\xi}})$, where $r \geq 0$ and $\hat{\boldsymbol{\xi}} \in \mathbb{S}^{n-1}$, and visualize the trajectories on the Poincaré disk through $\tanh(r/2)\hat{\boldsymbol{\xi}}$.

For the canonical prototype configuration, we use $C = 3$ fixed prototypes on the geodesic sphere of radius $\rho = 1.0$, equally spaced at angles $\phi_c = 2\pi c/C$ for $c \in \{0, 1, 2\}$. The prototypes remain fixed throughout the numerical integration, while the sample embeddings evolve according to the corresponding gradient dynamics.

Dataset. The dataset consists of $N = 15$ samples, with five samples per class assigned by $\mathrm{label}_i = i \bmod C$. For each sample, the initial radius is drawn from $r_0 \sim \mathrm{Uniform}[1.0, 2.0]$, and the initial angle from $\phi_0 \sim \mathrm{Uniform}[0, 2\pi)$, with $\hat{\boldsymbol{\xi}}_0 = (\cos\phi_0, \sin\phi_0)$. We use rejection sampling with angular exclusion intervals of ± 0.10 radians around the prototype directions. All random initializations use `np.random.default_rng(seed=11)`.

The default loss is the PERM cross-entropy $\psi_{\mathrm{CE}}(u) = \log(1 + \sum_j e^{-u_j})$, where the relative margins are $u_k = v_y - v_{j_k}$ with $v_c = -d_{\mathbb{H}^n}(\mathbf{x}, \mathbf{w}_c)$ and $j_k \in [C] \setminus \{y\}$.

D.2 Separate Radial and Angular Flows

We first examine the radial and angular components of the large-radius dynamics through separate numerical integration. The radial trajectories characterize the dependence of radial motion on the drift coefficient μ, while the angular trajectories characterize optimization of the Busemann loss on $\partial\mathbb{H}^n$. The resulting behavior is visualized in Fig. 6.

Radial flow. We fix the angular direction $\hat{\boldsymbol{\xi}}_i(t) = \hat{\boldsymbol{\xi}}_i(0)$ and integrate $\dot{r} = -\partial_r \ell$. Introducing the logarithmic time coordinate $\tau = \log t$ allows us to examine the asymptotic radial scaling $r(t) = \frac{1}{2}\log t + O(1)$ predicted by Theorem 6.2.

We sample $\tau \in [-2, 18]$ at 300 points using `scipy.integrate.solve_ivp` with LSODA, relative tolerance 10^{-7}, and absolute tolerance 10^{-9}. Integration terminates when $r \geq 14$ or $r \leq 0.05$. We partition the trajectories according to the sign of the initial drift $\mu(\boldsymbol{\xi}_i(0), y_i)$ and estimate their late-time slopes using linear regression over the final 30% of the integration interval.

Busemann angular flow. We integrate the limiting angular dynamics on $\mathbb{S}^{n-1}$:

$$\frac{d\hat{\boldsymbol{\xi}}}{ds} = -\operatorname{grad}_{\mathbb{S}^{n-1}} \mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y). \tag{150}$$

The Busemann loss is evaluated as $\mathcal{L}^{\mathrm{BS}}(\boldsymbol{\xi}, y) = \psi((b_{\boldsymbol{\xi}}(\mathbf{w}_c) - b_{\boldsymbol{\xi}}(\mathbf{w}_y))_{c \neq y})$, and its spherical gradient is computed analytically.

We sample $s \in [0, 40]$ at 120 linearly spaced points using `scipy.integrate.solve_ivp` with RK45, relative tolerance 10^{-8}, absolute tolerance 10^{-10}, and maximum step size 0.5. The angular directions are normalized during integration. We evaluate convergence through the limiting directions and the spherical gradient norm $\| \operatorname{grad}_{\mathbb{S}^{n-1}} \mathcal{L}^{\mathrm{BS}} \|$.

Numerical implementation. We evaluate the Minkowski inner product using the identity

$$-\langle \mathbf{x}, \mathbf{w} \rangle_M = \frac{A}{2} e^r + \frac{B}{2} e^{-r}, \tag{151}$$

where $A = w_0 - \hat{\boldsymbol{\xi}}^\top \bar{\boldsymbol{w}}$ and $B = w_0 + \hat{\boldsymbol{\xi}}^\top \bar{\boldsymbol{w}}$. Hyperbolic distances are evaluated through $\operatorname{arccosh}(\max\{-\langle \mathbf{x}, \mathbf{w} \rangle_M, 1\})$, with analytic derivatives in the radial and angular coordinates. For the drift coefficient, we use

$$\alpha_{\boldsymbol{\xi}}(\mathbf{w}_c) = \frac{\|\bar{\boldsymbol{w}}_c\|^2 - (\hat{\boldsymbol{\xi}}^\top \bar{\boldsymbol{w}}_c)^2}{(w_{c,0} - \hat{\boldsymbol{\xi}}^\top \bar{\boldsymbol{w}}_c)^2} = \frac{\|\bar{\boldsymbol{w}}_c^\perp\|^2}{\langle \mathbf{w}_c, \boldsymbol{\xi} \rangle_M^2}. \tag{152}$$

We evaluate the PERM losses and their derivatives using the same numerical expressions across prototype configurations.

D.3 Coupled Gradient Flow and Three-Phase Dynamics

We next integrate the full Riemannian gradient flow, allowing both the radius and angular direction to evolve:

$$\dot{r} = -\frac{\partial \ell}{\partial r}, \qquad \dot{\hat{\boldsymbol{\xi}}} = -\frac{1}{\sinh^2 r} \operatorname{grad}_{\mathbb{S}^{n-1}} \ell. \tag{153}$$

This experiment examines the interaction between radial drift and angular alignment under the dynamics of Theorem 6.4.

Initialization and evaluation. We consider prototype counts $C \in \{2, 3, 5\}$ with equally spaced prototypes at radius $\rho = 1$, and compare cross-entropy, multiclass exponential, and sum-of-softplus PERM losses. For the nondegenerate configurations, initial directions are selected within the attraction basin of a Busemann local minimum $\boldsymbol{\xi}^*$, with $\mu(\boldsymbol{\xi}(0), y) > 0$. We initialize trajectories at large radii and integrate the coupled radial and angular equations simultaneously.

Along each trajectory, we record the radius $r(t)$, drift coefficient $\mu(\boldsymbol{\xi}(t), y)$, and spherical distance $d_{\mathbb{S}^{n-1}}(\boldsymbol{\xi}(t), \boldsymbol{\xi}^*)$. We also evaluate the radial velocity $\dot{r}(t)$ and its large-radius approximation $2\mu(\boldsymbol{\xi}(t), y) e^{-2r(t)}$. Figure 5 presents the trajectories across prototype counts and PERM losses, with shaded regions indicating standard deviations across seeds.

Sufficiently aligned regions. For nondegenerate configurations, we examine the sufficiently aligned region defined in Definition 6.3:

$$\mathcal{A}(\boldsymbol{\xi}^*) = \{\boldsymbol{\xi} \in \partial \mathbb{H}^n : d_{\mathbb{S}^{n-1}}(\boldsymbol{\xi}, \boldsymbol{\xi}^*) \le \theta_*\},$$

where $\theta_* > 0$ is chosen such that $\mu(\boldsymbol{\xi}, y) < 0$ and the Busemann Hessian is positive definite throughout the region. We record the first entry time

$$T_2 = \inf\{t : \boldsymbol{\xi}(t) \in \mathcal{A}(\boldsymbol{\xi}^*)\},$$

indicated by vertical dashed lines in Figure 5. The inset in the drift panel magnifies the near-zero behavior of μ for $C = 3$ and $C = 5$.

Three-phase evolution. The coupled trajectories exhibit three successive phases. Initially, positive drift produces outward radial motion while the directions evolve toward Busemann local minima. As angular alignment proceeds, the drift decreases and crosses zero, reversing the leading radial motion. The trajectories subsequently enter the sufficiently aligned regions, where negative drift induces slow radial return.

Figure 5 illustrates these phases across the nondegenerate configurations, while the degenerate case exhibits weaker drift reversal. The angular distances decrease toward zero across the tested configurations, providing numerical evidence of the Busemann directional bias.

Radial and angular coupling. The large-radius radial dynamics satisfy

$$\dot{r} = 2\mu(\boldsymbol{\xi}, y) e^{-2r} + O(e^{-4r}).$$

The numerical trajectories illustrate how angular alignment changes the sign of the leading radial term, connecting outward excursions with subsequent slow return. Together, these observations support the coupled radial and angular behavior characterized by Theorem 6.4.

D.4 Ablations

We examine the radial and Busemann angular dynamics across three factors: the number of prototypes, the PERM loss template, and the geometric configuration of the prototypes. The results are presented column-wise in Fig. 6.

(i) Number of prototypes C. We consider $C \in \{2, 3, 5\}$, with five samples per class, giving $N \in \{10, 15, 25\}$. The prototype radius is fixed at $\rho = 1.0$, and the loss is cross-entropy. All configurations use seed 11. This experiment examines the radial dichotomy and the structure of angular attractors as the number of classes increases.

(ii) PERM loss template ψ. We compare three PERM loss templates:

1. Cross-entropy: $\psi_{\mathrm{CE}}(u) = \log(1 + \sum_j e^{-u_j})$.
2. Multiclass exponential: $\psi_{\mathrm{EXP}}(u) = \sum_j e^{-u_j}$.
3. Sum-of-softplus: $\psi_{\mathrm{SS}}(u) = \sum_j \log(1 + e^{-u_j})$.

We use numerically stable implementations of these functions, including `logsumexp` for cross-entropy. Their gradients are checked against `scipy.optimize.approx_fprime`, with relative error below 10^{-6}. All three losses are evaluated on the same canonical configuration ($C = 3$, $\rho = 1.0$, seed 11).

(iii) Prototype configuration. We consider three geometric layouts:

1. Canonical: equal radii $\rho_c = 1.0$, with $\phi_c = 2\pi c/3$.
2. Unequal radii: $\rho_c \in \{0.7, 1.0, 1.5\}$, with equally spaced angles.
3. Clustered angles: equal radii $\rho_c = 1.0$, with $\phi_c \in \{0, \pi/3, 2\pi/3\}$.

We use cross-entropy and seed 11 throughout these configurations.

Evaluation. For each panel of Fig. 6, we fit a linear regression to $r_i(\tau)$ over the final 30% of the logarithmic time interval. We report the late-time slope with bootstrap confidence intervals. The predicted positive-drift slope is $1/2$.

We further group the radial trajectories by $\mu(\boldsymbol{\xi}_i(0), y_i) > 0$ and $\mu(\boldsymbol{\xi}_i(0), y_i) \leq 0$, corresponding to the two radial behaviors. Angular convergence is evaluated through the spherical gradient norm and the limiting Busemann directions.

Across the tested configurations, the positive-drift radial trajectories exhibit the predicted logarithmic scaling, while the negative-drift trajectories move inward. The angular trajectories converge to Busemann critical points, with $\| \operatorname{grad}_{\mathbb{S}^{n-1}} \mathcal{L}^{\mathrm{BS}} \| \leq 10^{-6}$ across the tested settings. For the canonical configuration, the limiting gradient norms reach approximately 10^{-16}.

D.5 Additional Diagnostics

We further examine the drift coefficient at Busemann angular attractors, visualize the trajectories on the Poincaré disk, and characterize the angular attraction basins. These diagnostics complement the radial and angular results in Fig. 6 and the coupled three-phase analysis.

Drift coefficient at angular attractors. For each panel, we cluster the limiting angular directions into bins of width 0.05 radians and evaluate the drift coefficient at each cluster centroid:

$$\mu(\boldsymbol{\xi}^*, y) = -\left\langle \nabla\psi(\Upsilon_y D\mathbf{b}(\boldsymbol{\xi}^*)), \Upsilon_y D\boldsymbol{\alpha}(\boldsymbol{\xi}^*) \right\rangle .$$

In nine of the ten panels, all observed angular attractors have negative drift. For the canonical configuration ($C = 3$, $\rho = 1$, cross-entropy), we obtain six attractors with $\mu \in [-0.965, -0.059]$. For $C = 5$, the nine observed attractors satisfy $\mu \in [-1.033, -0.303]$.

In the $C = 2$ configuration, the degenerate angular structure produces $\mu \approx 10^{-15}$ at the corresponding antipodal directions. The distribution of drift coefficients illustrates the negative-drift structure of nondegenerate Busemann attractors and its connection to the radial return mechanism.

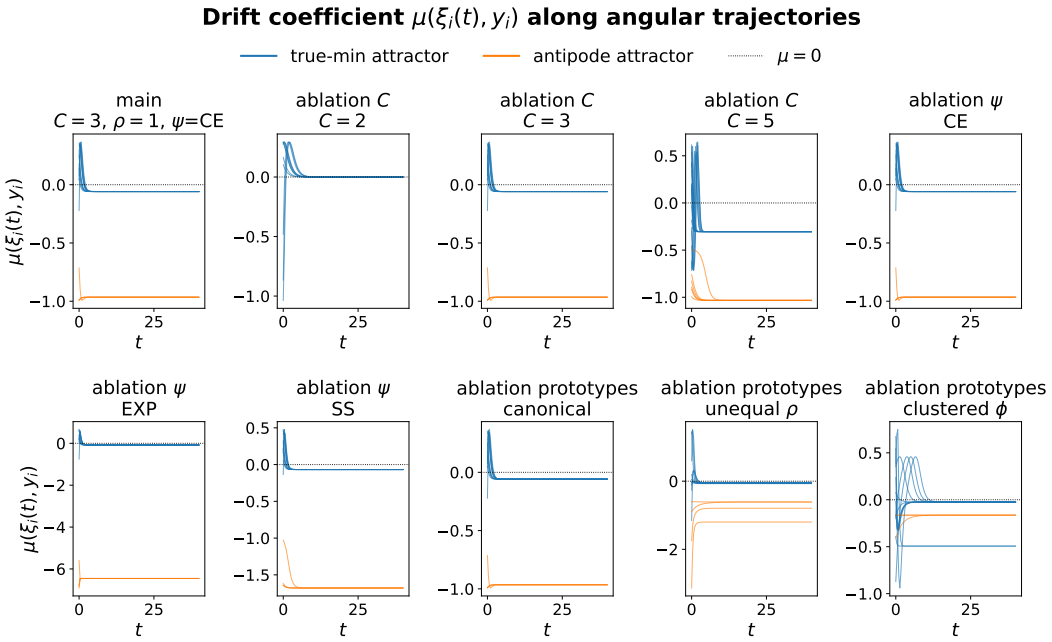


Figure 7: Evolution of the drift coefficient $\mu(\boldsymbol{\xi}(t), y)$ along the limiting Busemann angular trajectories. Blue: class-aligned attractors; orange: antipodal attractors; dashed line: $\mu = 0$. The observed nondegenerate attractors have negative drift, illustrating the angular alignment mechanism underlying radial return.

Poincaré disk visualizations. Fig. 8 presents the same ten configurations as Fig. 6, arranged in a 4×5 grid. The upper two rows show the radial trajectories, while the lower two rows show the Busemann angular trajectories.

We use $\mathbf{X}$ to denote prototype locations, $\circ$ for initial directions, and $\star$ for the limiting angular directions. For the angular trajectories, we apply a radial visualization coordinate in $[0.55, 0.97]$ to display their evolution on $\mathbb{S}^1$ within the Poincaré disk.

Initialization basin map. We consider the canonical configuration with $C = 3$, $\rho = 1.0$, cross-entropy, and target class $y = 0$. We sample 200 equally spaced initial directions $\theta_0 \in [-\pi, \pi)$ and integrate the Busemann angular flow up to $s = 40$.

Each limiting direction is classified as class-aligned when $b_{\boldsymbol{\xi}^*}(\mathbf{w}_y) < \min_{c \neq y} b_{\boldsymbol{\xi}^*}(\mathbf{w}_c)$, and as antipodal otherwise. We obtain 145 class-aligned trajectories and 55 antipodal trajectories, corresponding to 72.5% and 27.5% of the initial directions. The basin boundary occurs at approximately $|\theta_0| = 132.5°$, giving the antipodal basin a width of approximately $95°$.

These results characterize how the initial boundary direction determines the limiting Busemann attractor and, through μ, influences the subsequent radial dynamics.

Software, hardware, and reproducibility. The numerical integration experiments use Python 3.9, NumPy 2.0, SciPy 1.13, and matplotlib 3.9 on a CPU. The original separated-flow figures, tables, and CSV outputs are generated in approximately 30 seconds. The implementation uses fixed random seeds and analytic gradients for the hyperbolic dynamics. The experimental code is anonymously released for review.

D.6 Neural Network Training

We further investigate Busemann asymptotic behavior under shared neural network parameterization. We first compare stationary solutions of the exact hyperbolic loss and Busemann risk using a shared angular encoder, then examine their gradients on ResNet-18 checkpoints trained on CIFAR-10.

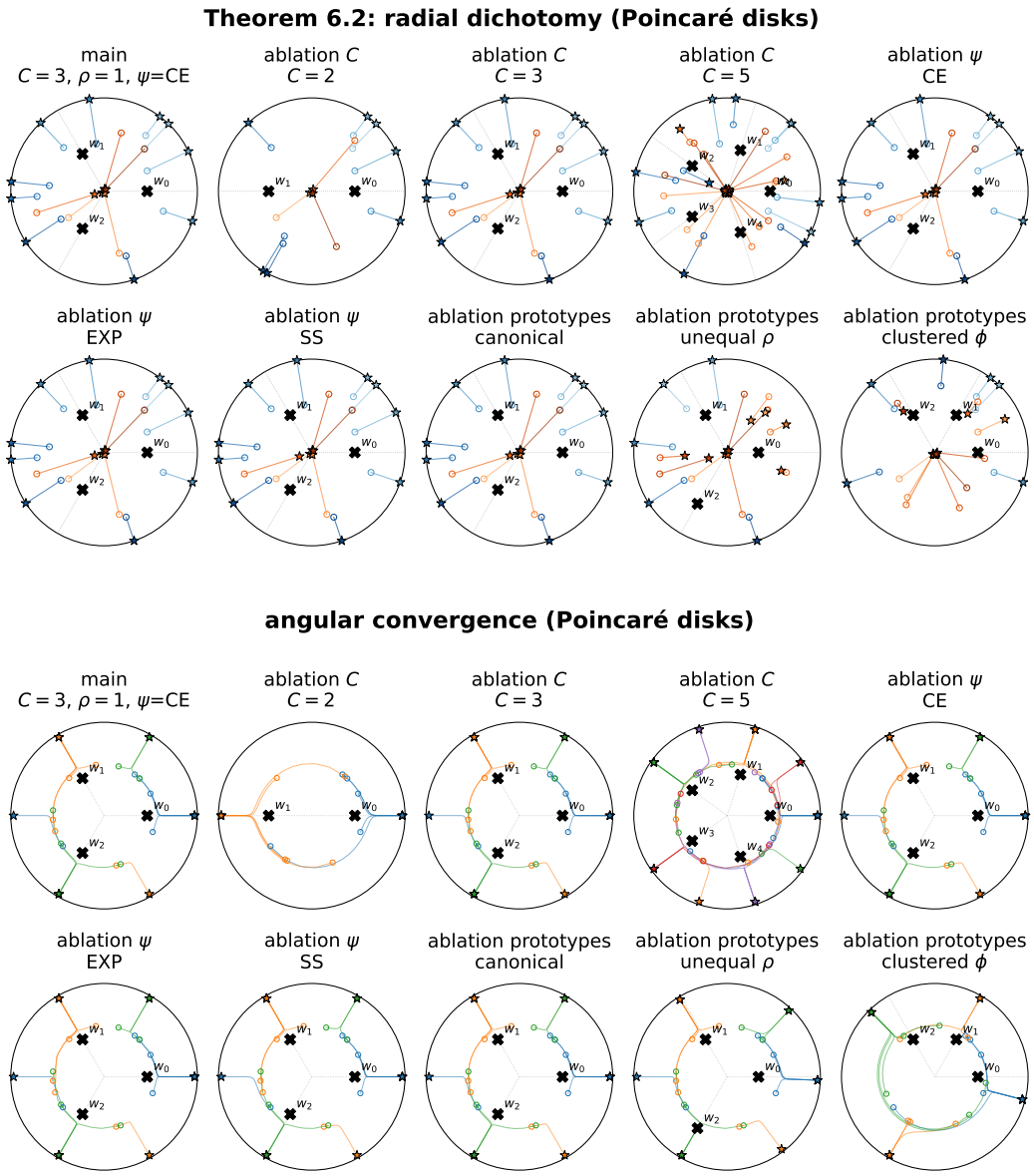


Figure 8: Poincaré disk visualization of the radial and Busemann angular trajectories across ten experimental configurations. Top two rows: fixed-direction radial dynamics. Bottom two rows: limiting angular dynamics, displayed with radial separation for visualization. Prototype locations, initial directions, and limiting directions are marked by $\mathbf{X}$, $\circ$, and $\star$, respectively.

Shared-encoder stationarity. We consider a three-parameter shared linear angular encoder with 900 samples, cross-entropy loss, and fixed prototypes. We use canonical and irregular prototype configurations, five independent seeds, and seven radii $R \in \{3, 3.5, 4, 4.5, 5, 5.5, 6\}$, yielding 70 stationary-point checkpoints.

From each exact hyperbolic stationary point θ_R^*, we recover a nearby Busemann stationary point θ_{BS}^* by solving $\nabla_\theta \mathcal{R}_{\mathbf{W}}^{\mathrm{BS}}(\boldsymbol{\Xi}_\theta) = 0$ using full-batch float64 L-BFGS and step-restricted Newton iterations. All 70 recovered points have gradient residuals below 1.11×10^{-13}.

Prototypes	Checkpoints	Slope of $\log D_R$	95% CI
Canonical	All	-1.989442	$[-2.017487, -1.961398]$
Irregular	All	-2.000482	$[-2.001438, -1.999526]$
Canonical	Converged	-2.001996	$[-2.007563, -1.996428]$
Irregular	Converged	-2.000017	$[-2.000328, -1.999706]$

Table 1: Angular proximity between stationary representations of the exact hyperbolic loss and Busemann risk. The converged subset satisfies $\|\nabla_\theta \mathcal{L}_{\mathbf{W}}\| < 10^{-8}$. The observed slopes closely follow e^{-2R} scaling.

Seed	R_θ	Mean radius	$\|\nabla_\theta \mathcal{L}_{\mathbf{W}}\|$	$\|\nabla_\theta \mathcal{R}^{\mathrm{BS}}_{\mathbf{W}}\|$
0	3.0677	3.1514	0.096812	0.095768
1	3.0591	3.1493	0.435698	0.429801
2	3.0258	3.0369	0.461785	0.457161

Table 2: Exact hyperbolic and Busemann parameter gradient norms at CIFAR-10 ResNet-18 checkpoints, evaluated on a fixed 1024-sample training probe.

We measure the angular proximity between their representations:

$$D_R = \left[\frac{1}{N} \sum_{i=1}^{N} d_{\mathbb{S}^{n-1}} \left(\hat{\boldsymbol{\xi}}_i(\theta_R^*), \hat{\boldsymbol{\xi}}_i(\theta_{\mathrm{BS}}^*) \right)^2 \right]^{1/2}. \tag{154}$$

Motivated by the large-radius expansion, we examine the scaling $D_R = O(e^{-2R})$. We fit the slope of $\log D_R$ against R for each seed and compute 95% confidence intervals from the five seed-level slopes.

Table 1 shows slopes close to -2 in both configurations. The corresponding slopes of $\log \|\nabla_\theta \mathcal{R}^{\mathrm{BS}}_{\mathbf{W}}\|$ are -1.990407 and -2.000416 for all checkpoints, and -2.001859 and -2.000007 for converged subsets.

We also examine the local first-order stationary-point displacement. The recovered Busemann Hessians have smallest absolute eigenvalues between 0.424545 and 0.558863, with maximum condition number 2.13155. Median relative errors of the first-order parameter and angular displacement approximations are 0.3374% and 0.3375%, respectively. The recovered Busemann critical parameters agree across radii within 1.72×10^{-13}. These observations provide additional numerical evidence for $O(e^{-2R})$ stationary-point proximity under shared parameterization.

ResNet-18 training. We additionally evaluate saved ResNet-18 checkpoints trained on CIFAR-10 with freely varying hyperbolic embedding radii. We compute exact hyperbolic and Busemann risk parameter gradients on a fixed unaugmented 1024-sample training probe using float64 arithmetic.

Table 2 shows closely matched gradient norms across all three seeds. Together with the shared-encoder results, these observations provide complementary numerical evidence for Busemann asymptotic behavior under neural network parameterization.

E Societal Impact

This work is primarily theoretical and studies the implicit bias of Riemannian gradient flow in hyperbolic representation learning. It does not collect user data, involve human subjects, release pretrained models, or deploy a system. The potential positive impact is indirect: a better understanding of boundary saturation, near boundary clustering, and asymptotic decision boundaries may improve the stability, interpretability, and reliability of hyperbolic models for structured or hierarchical data. Potential negative impacts are also indirect. More effective representation learning methods could be used in sensitive domains such as recommendation, ranking, social network analysis, or decision support, where insufficient validation, fairness evaluation, or privacy safeguards may lead to harmful outcomes. Since our experiments are controlled synthetic simulations and our contribution is theoretical, we do not identify direct risks beyond the broader responsible use concerns associated with representation learning.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, we claim it clearly.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer:[Yes]

   Justification: We discuss our limitations in Appendix [A.2].

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide all the detailed proofs in Appendix [B.1],[B.2],[C.1],[C.3], and [C.4].

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the details of experiments in Appendix [D].

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide them in the supplementary materials.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all the details of experiments in Appendix [D].

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report inter-seed standard deviations for coupled gradient flow trajectories and 95% confidence intervals for neural network stationary point scaling estimates. These statistics quantify the variability of our numerical verification of theoretical predictions, rather than statistical superiority over baselines.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide all the details of experiments in Appendix [D].

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes. Our research is primarily a theoretical analysis of optimization dynamics in hyperbolic classification, supported by controlled synthetic experiments.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss our impacts in Appendix [E]

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The experiments use only controlled synthetic data for numerical verification of theoretical results, and no sensitive, private, or user collected data are involved.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cites prior work and theoretical frameworks used in the analysis. The numerical experiments use synthetic data generated by the authors and do not use external datasets, pretrained models, or third party assets beyond standard open source scientific computing libraries.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new datasets, pretrained models, or other standalone assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments, human subjects, participant instructions, screenshots for user studies, or participant compensation. The experiments are controlled synthetic numerical simulations.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, user studies, or participant data. Therefore, there are no study participant risks to disclose and no IRB or equivalent human subjects approval is required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer:[N/A]

    Justification: The research does not use LLMs as part of the core methodology. Any LLM usage was limited to writing, editing, and experimental assistance.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.