Geometric Foundations of Racing Dynamics: How Gradient Descent Adapts Network Capacity on Data Manifolds with Application to Bayesian R-LayerNorm

Mohsen Mostafa *

Posted Date: 24 March 2026

doi: 10.20944/preprints202603.0760.v2

Keywords: racing dynamics; manifold learning; lottery tickets; Bayesian normalization; gradient flow

 doi:10.20944/preprints202603.0760.v2

Article

Geometric Foundations of Racing Dynamics: How Gradient Descent Adapts Network Capacity on Data Manifolds with Application to Bayesian R-LayerNorm

Mohsen Mostafa

Independent Research; mohsen.mostafa.ai@outlook.com

Abstract

Understanding how gradient descent shapes neural network representations remains a fundamental challenge in deep learning theory. Recent work has revealed that neural networks behave as "racing" systems: neurons compete to align with task-relevant directions, and those that succeed experience exponential norm growth. However, the geometric principles governing this race—particularly when data lies on low-dimensional manifolds and networks employ adaptive normalization—remain poorly understood. This paper establishes a mathematical framework that unifies and extends these insights. We prove four fundamental theorems: (1) neuron weight vectors converge exponentially to the tangent space of the data manifold, with a rate determined by local curvature and gating dynamics; (2) for rotation-equivariant tasks, an angular momentum tensor is conserved under gradient flow, imposing topological constraints on neuronal rearrangements; (3) the distribution of high-norm "winning" neurons follows a von Mises–Fisher concentration on the manifold, with concentration parameter linked to initial angular variance; (4) when the network employs Bayesian R-LayerNorm, the exponential norm growth law acquires a multiplicative noise gate that explains its empirical robustness on corrupted datasets. Together, these results provide a geometric foundation for understanding capacity adaptation, lottery tickets, and uncertainty-aware learning in neural networks. All experiments were performed under severe computational constraints (free Kaggle GPU, NVIDIA P100 with 16 GB VRAM, limited epochs). The empirical results, though proof-of-concept, strongly support the theoretical predictions, especially for winning ticket concentration and noise gating on real data. Code and data are publicly available.

Keywords: racing dynamics; manifold learning; lottery tickets; Bayesian normalization; gradient flow

1. Introduction

Modern neural networks are massively overparameterized, yet they generalize remarkably well. This phenomenon is often attributed to the implicit bias of gradient descent, which reduces the effective capacity of the network to match the task [1–3]. A striking manifestation of this capacity adaptation is the lottery ticket hypothesis [4]: within a randomly initialized network, there exists a smaller subnetwork that, if trained in isolation, can achieve comparable performance. The winning tickets—neurons that survive pruning—are not arbitrary; they are determined by their initial conditions [5,6].

Recent work by Pinson [7] has provided crucial insights into this process through the lens of "racing dynamics." By analyzing single-hidden-layer ReLU networks at the level of individual neurons, Pinson identified three dynamical principles—mutual alignment, unlocking, and racing—that explain why neurons with favorable initial orientations achieve exponentially larger norms. This framework elegantly captures the essence of the lottery ticket phenomenon: neurons that align faster to task-relevant directions grow faster, and they "win the race" to become the important neurons in the final network.

However, several fundamental questions remain unanswered. First, how does the geometry of the data manifold influence these dynamics? Real-world data is not isotropic; it lies on low-dimensional

 doi:10.20944/preprints202603.0760.v2

manifolds embedded in high-dimensional spaces [8,9]. Second, how do modern architectural components—particularly normalization layers designed to handle noise—modify the racing dynamics? Third, can we provide rigorous mathematical guarantees for the observed phenomena, moving beyond empirical observations to provable theorems?

This paper addresses these gaps by developing a comprehensive geometric framework for learning dynamics. Our contributions are as follows:

- **Theorem 1 (Manifold Alignment)**: Under mild regularity conditions, the component of a neuron's incoming weight that is normal to the data manifold decays exponentially fast. The decay rate depends on the norm of the outgoing weight and on the local Ricci curvature of the manifold, modulated by the Lipschitz constant of the gating function.
- **Theorem 2 (Conservation of Angular Momentum)**: For rotation-equivariant tasks with rotation-invariant loss and gating, the angular momentum tensor formed by pairs of incoming weight vectors is conserved under gradient flow. This imposes a topological invariant on the learning dynamics.
- **Theorem 3 (Probabilistic Concentration of Winning Tickets)**: The set of high-norm neurons at the end of training concentrates on the manifold according to a von Mises–Fisher distribution. The concentration parameter is given by $\kappa_{\text{eff}} = \lambda \|\langle \gamma_s \rangle\| / \text{Var}(\delta\alpha(0)) + O(\text{Ric}_{\mathcal{M}})$, directly linking initial angular variance to final neuronal importance.
- **Theorem 4 (Stable Racing Dynamics under Bayesian Normalization)**: When the network employs Bayesian R-LayerNorm [21], the exponential norm growth law acquires a multiplicative noise gate $\exp(-\alpha\psi(\lambda E_\alpha))$, where $\psi(t) = \log(1+t) - t/(1+t)$ is a provably stable function. This gate suppresses growth in high-noise regions.

All experiments were performed on a free Kaggle GPU (NVIDIA P100, 16 GB VRAM) with limited runtime. Due to these constraints, the empirical validations are deliberately kept at a proof-of-concept scale, focusing on synthetic manifolds where the geometry is known exactly, and on two real-world datasets (MNIST, Fashion-MNIST) for the noise-gating and winning-ticket concentration theorems. The goal is to illustrate the theoretical predictions, not to claim state-of-the-art performance. We encourage readers to interpret the experimental results as illustrative rather than exhaustive.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 presents the mathematical formalism. Sections 4–7 state and prove the four main theorems. Section 8 provides experimental validation, including extended synthetic runs and real-data results. Section 9 discusses implications, limitations, and future work. Section 10 concludes.

2. Related Work

2.1. Racing Dynamics and Lottery Tickets

The lottery ticket hypothesis [4] sparked intense interest in understanding how initial conditions determine final network structure. Subsequent work showed that winning tickets can be identified early in training [5,6] and that the sign of initial weights matters more than their magnitude [26]. Pinson [7] provided a dynamical explanation through the concepts of mutual alignment, unlocking, and racing. Our work builds directly on this foundation, extending it to incorporate data geometry and adaptive normalization.

2.2. Geometric Deep Learning

The manifold hypothesis—that high-dimensional data concentrates near low-dimensional manifolds—has been central to geometric deep learning [8,9,20]. Prior work has studied how neural networks adapt to data geometry through the lens of the neural tangent kernel [18] and implicit regularization [13]. However, most theoretical analyses assume either linear networks [10,25] or infinite-width limits [14]. Our work provides a finite-width, neuron-level analysis that explicitly accounts for manifold geometry.

 doi:10.20944/preprints202603.0760.v2

2.3. Normalization and Robustness

Normalization layers are essential for training deep networks [12,16]. Recent work has introduced adaptive normalization methods that adjust to input statistics [15,17]. Bayesian R-LayerNorm [21] extends this line by incorporating uncertainty quantification through a provably stable ψ-function. Our Theorem 4 explains why this method can improve robustness: the ψ-function acts as a noise gate in the norm growth dynamics.

2.4. Implicit Regularization and Capacity Adaptation

The implicit bias of gradient descent has been extensively studied [2,3,24]. For linear networks, gradient descent converges to solutions with minimum nuclear norm [10]. For ReLU networks, the bias is more complex but often favors simple solutions [19]. Our work contributes to this literature by showing how geometric alignment and racing dynamics naturally reduce effective capacity.

3. Mathematical Formalism

3.1. Network Architecture

Consider a single hidden layer network with M neurons for binary classification. For each neuron α ($\alpha = 1, \ldots, M$) we define:

- Incoming weight vector $w_\alpha^{(1)} \in \mathbb{R}^n$
- Outgoing weight vector $w_\alpha^{(2)} \in \mathbb{R}^2$
- State-dependent gating function $g_\alpha(\mathbf{x}, \theta) \in [0, 1]$, where θ denotes all network parameters.

The activation of neuron α for input $\mathbf{x}$ is

$$h_\alpha = g_\alpha(\mathbf{x}, \theta) \cdot (w_\alpha^{(1)} \cdot \mathbf{x}). \tag{1}$$

Following Pinson [7], the ReLU activation function can be viewed as a gate: $g_\alpha(\mathbf{x}, \theta) = \mathbf{1}_{w_\alpha^{(1)} \cdot \mathbf{x} + b_\alpha > 0}$. However, our analysis allows for more general Lipschitz gating functions.

Assumption 1 (Lipschitz Gating). *There exists a constant $L_g > 0$ such that for any inputs $\mathbf{x}, \mathbf{x}'$ and parameter sets θ, θ',*

$$|g_\alpha(\mathbf{x}, \theta) - g_\alpha(\mathbf{x}', \theta')| \leq L_g \big(\|\mathbf{x} - \mathbf{x}'\| + \|\theta - \theta'\| \big). \tag{2}$$

3.2. Data Geometry

Let the training data $\{\mathbf{x}_s\}_{s=1}^N$ lie on a smooth, compact Riemannian manifold $\mathcal{M} \subset \mathbb{R}^n$ of dimension $d \leq n$. For any point $\mathbf{x} \in \mathcal{M}$, denote:

- $T_\mathbf{x}\mathcal{M}$ – the tangent space at $\mathbf{x}$
- $P_\mathbf{x} : \mathbb{R}^n \to T_\mathbf{x}\mathcal{M}$ – the orthogonal projection onto the tangent space
- $P_\mathbf{x}^\perp = I - P_\mathbf{x}$ – the projection onto the normal space.

Assumption 2 (Manifold Regularity). *The manifold $\mathcal{M}$ has bounded curvature: the second fundamental form is bounded, and the injectivity radius is positive. This ensures that local quadratic approximations are valid and that the Poincaré inequality holds on $\mathcal{M}$ with a constant $c_P > 0$ depending on the Ricci curvature.*

3.3. Gradient Flow Equations

Following Pinson [7] and our own derivations, the gradient flow for the incoming weights is

$$\frac{d}{dt} w_\alpha^{(1)} = \lambda \|w_\alpha^{(2)}\| \langle \gamma_{\alpha s} \rangle, \tag{3}$$

 doi:10.20944/preprints202603.0760.v2

where $\lambda > 0$ is the learning rate and

$$\langle \gamma_{\alpha s} \rangle = \frac{1}{N} \sum_{s=1}^{N} g_\alpha(\mathbf{x}_s, \theta) \cos(\phi_{\Delta y_s} - \phi_\alpha) \|\Delta y_s\| \, \mathbf{x}_s. \tag{4}$$

Here ϕ_α is the angle of the outgoing weight vector $w_\alpha^{(2)}$ in the output plane, $\phi_{\Delta y_s}$ is the angle of the error signal (fixed at $7\pi/4$ for class 1 and $3\pi/4$ for class 2), and $\|\Delta y_s\|$ is the magnitude of the error. The quantity $\langle \gamma_{\alpha s} \rangle$ can be interpreted as the average over the dataset of the gated input weighted by the alignment with the output error.

For the outgoing weights, we have

$$\frac{d}{dt} w_\alpha^{(2)} = -\lambda \langle h_\alpha \Delta y_s \rangle. \tag{5}$$

Bias dynamics play a secondary role and are omitted for brevity.

4. Theorem 1: Manifold Alignment

Theorem 1 (Manifold Alignment). *Under Assumptions 1 and 2, for any neuron α in the unlocking phase (i.e., after initial transients but before the loss drops significantly), the incoming weight vector $w_\alpha^{(1)}(t)$ converges to the tangent space $T_{\mathbf{x}_\alpha}\mathcal{M}$ almost surely, where $\mathbf{x}_\alpha$ is the current effective centroid (e.g., the mean of the gated inputs). Specifically, let $w_{\alpha\perp}(t) = P_{\mathbf{x}_\alpha}^{\perp} w_\alpha^{(1)}(t)$ be the component normal to the tangent space. Then*

$$\|w_{\alpha\perp}(t)\| \leq Ce^{-\eta t}, \tag{6}$$

with convergence rate

$$\eta = \lambda \|w_\alpha^{(2)}\| c_{\text{geom}} \big(1 - O(L_g C_g \epsilon)\big), \tag{7}$$

where $c_{\text{geom}} > 0$ is a constant depending on the Ricci curvature of $\mathcal{M}$, C_g bounds the self-dependence of the gating, and ϵ is a small parameter related to the local quadratic approximation.

Proof sketch. Decompose $w_\alpha^{(1)} = w_{\alpha\|} + w_{\alpha\perp}$. From the gradient flow (3), the normal component evolves as $\frac{d}{dt} w_{\alpha\perp} = -\lambda \|w_\alpha^{(2)}\| P_{\mathbf{x}_\alpha}^{\perp} \langle \gamma_{\alpha s} \rangle$. Using a second-order Taylor expansion of the manifold and the Poincaré inequality, one obtains $\frac{d}{dt}\|w_{\alpha\perp}\|^2 \leq -2\lambda \|w_\alpha^{(2)}\| c_{\text{geom}} \|w_{\alpha\perp}\|^2 + O(\|w_{\alpha\perp}\|^3)$, yielding exponential decay. □

Remark. Theorem 1 explains why neurons learn to ignore off-manifold directions. The decay rate η is proportional to $\|w_\alpha^{(2)}\|$, meaning that neurons with larger outgoing weights align faster—a form of "rich get richer" dynamics that complements the racing phenomenon.

5. Theorem 2: Conservation of Angular Momentum

Theorem 2 (Conservation of Angular Momentum). *For a rotation-equivariant task with rotation-invariant loss, and with gating functions that depend only on rotation-invariant features (e.g., $\|\mathbf{x}\|$), the angular momentum tensor*

$$L_{\alpha\beta} = w_\alpha^{(1)} \wedge w_\beta^{(1)} \tag{8}$$

is conserved under gradient flow: $\frac{d}{dt} L_{\alpha\beta} = 0$ for all α, β.

Proof. By rotation equivariance, the gradient $\langle \gamma_{\alpha s} \rangle$ is parallel to $w_\alpha^{(1)}$. Hence $\langle \gamma_{\alpha s} \rangle \wedge w_\alpha^{(1)} = 0$. Differentiating $L_{\alpha\beta}$ and substituting the gradient flow gives

$$\frac{d}{dt} L_{\alpha\beta} = \lambda \|w_\alpha^{(2)}\| \langle \gamma_{\alpha s} \rangle \wedge w_\beta^{(1)} + \lambda \|w_\beta^{(2)}\| w_\alpha^{(1)} \wedge \langle \gamma_{\beta s} \rangle.$$

 doi:10.20944/preprints202603.0760.v2

Because the gating depends only on rotation-invariant features, the norms $\|w_\alpha^{(2)}\|$ are invariant under rotations, and the parallelism property forces each wedge product to vanish. Thus $\frac{d}{dt} L_{\alpha\beta} = 0$. □

Corollary. For a network in $\mathbb{R}^3$, the conservation law implies that the cross product $w_\alpha^{(1)} \times w_\beta^{(1)}$ is constant. In higher dimensions, the wedge product components are conserved.

6. Theorem 3: Probabilistic Concentration of Winning Tickets

Theorem 3 (Probabilistic Concentration). *Let $a_\alpha = \|w_\alpha^{(2)}\| \|w_\alpha^{(1)}\|$ denote the norm product for neuron α. Define the set of "winning tickets" after training as*

$$\mathcal{W} = \{\alpha : a_\alpha(T) > \tau\}, \tag{9}$$

where τ is a threshold (e.g., the 80th percentile). Under uniform initialization of weight directions on the unit sphere, and under the assumptions of Theorem 1, the distribution of the directions $\mathbf{u}_\alpha = w_\alpha^{(1)} / \|w_\alpha^{(1)}\|$ for $\alpha \in \mathcal{W}$ converges, as $T \to \infty$ and $N \to \infty$, to a von Mises–Fisher distribution on the sphere $S^{d-1} \subset T_{\mathbf{x}}\mathcal{M}$:

$$\mathbf{u}_\alpha \xrightarrow{d} \mathrm{vMF}(\mathbf{m}_*, \kappa_{\mathrm{eff}}), \tag{10}$$

with concentration parameter

$$\kappa_{\mathrm{eff}} = \frac{\lambda \|\langle \gamma_s \rangle\|}{\mathrm{Var}(\delta\alpha(0))} + O(\mathrm{Ric}_{\mathcal{M}}). \tag{11}$$

Proof sketch. From the exponential growth law (Theorem 4), the norm product at large time satisfies $a_\alpha(T) \propto \exp(\lambda \|\langle \gamma_{\alpha s} \rangle\| T \cos \bar{\delta}_\alpha)$, where $\bar{\delta}_\alpha$ is the limiting alignment angle. Initial angles $\delta\alpha(0)$ are uniformly distributed. The dynamics align each $w_\alpha^{(1)}$ with the mean direction $\mathbf{m}_*$, but residual randomness remains due to finite-sample effects and geometry. Standard results on stochastic alignment [11] show that the conditional distribution of $\mathbf{u}_\alpha$ given $a_\alpha(T) > \tau$ is Fisher–von Mises. The concentration κ_{eff} is proportional to the signal strength divided by the initial angular variance, with a curvature correction. □

Remark. The mean resultant length $R = \|\mathbb{E}[\mathbf{u}_\alpha]\|$ is a convenient proxy for κ_{eff}. For large κ, $R \approx 1 - 1/(2\kappa)$. In our experiments (Section 8), we observe $R \approx 0.22$ for the synthetic sphere, corresponding to $\kappa \approx 0.3$, consistent with the theoretical prediction.

7. Theorem 4: Stable Racing Dynamics Under Bayesian Normalization

7.1. Bayesian R-LayerNorm

In a companion paper [21], we introduced Bayesian R-LayerNorm, which computes

$$\mathrm{B{-}R{-}LN}(\mathbf{x}) = \gamma \odot \frac{\mathbf{x} - \mu}{\sigma_{\mathrm{eff}}} + \beta, \qquad \sigma_{\mathrm{eff}}^2 = \sigma^2 \cdot \exp\big(2\alpha\psi(\lambda E)\big), \tag{12}$$

where μ, σ are the usual mean and standard deviation over spatial dimensions, E is a local entropy estimate (e.g., the variance of a local 3×3 neighborhood), and ψ is the stable function

$$\psi(t) = \log(1 + t) - \frac{t}{1 + t}, \qquad t \geq 0. \tag{13}$$

The function ψ satisfies:

1. $0 \leq \psi'(t) \leq 1$ for all $t \geq 0$ (Lipschitz stability),
2. $\psi(t) \sim t^2/2$ as $t \to 0$ (quadratic regularization),
3. $\psi(t) \sim \log t$ as $t \to \infty$ (logarithmic saturation).

These properties ensure that the normalization adapts smoothly to varying noise levels without introducing extreme gradients.

 doi:10.20944/preprints202603.0760.v2

7.2. Modified Norm Growth Law

Theorem 4 (Stable Racing Dynamics)**.** *Consider a network trained with Bayesian R-LayerNorm under the assumptions of Theorem 1. Then the norm product $a_\alpha = \|w_\alpha^{(2)}\|\|w_\alpha^{(1)}\|$ for each neuron α evolves according to*

$$\frac{da_\alpha}{dt} = \lambda \|\langle \gamma_{\alpha s} \rangle\| \cdot \exp\big(-\alpha \psi(\lambda E_\alpha)\big) \cdot \cos\delta_\alpha \cdot a_\alpha + R, \tag{14}$$

where E_α is the local entropy estimate associated with neuron α, and the residual R is bounded by

$$\|R\| \le C \cdot \sigma_{\text{noise}}^2 \cdot \|w_\alpha^{(1)}\|\|w_\alpha^{(2)}\|, \tag{15}$$

with C a constant depending on the Lipschitz constants of the network and the curvature of $\mathcal{M}$.

Proof sketch. The gradient of the loss with respect to $w_\alpha^{(1)}$ is modified by the normalization factor $1/\sigma_{\text{eff}}$. By the chain rule,

$$\frac{\partial L}{\partial w_\alpha^{(1)}} = \frac{1}{\sigma_{\text{eff}}} \cdot \frac{\partial L}{\partial h_\alpha} \cdot \frac{\partial h_\alpha}{\partial w_\alpha^{(1)}} + \text{terms from } \frac{\partial \sigma_{\text{eff}}}{\partial w_\alpha^{(1)}}.$$

The first term gives the original gradient scaled by $\exp(-\alpha\psi(\lambda E_\alpha))$. The additional terms arise because E_α depends on the activations, which depend on $w_\alpha^{(1)}$. Using $|\psi'| \le 1$ and Lipschitz continuity, one shows that $\|\partial E_\alpha / \partial w_\alpha^{(1)}\| \le L_E \|w_\alpha^{(1)}\| \sigma_{\text{noise}}^2$, leading to the bound on R. The remainder follows Pinson [7] with the noise gate factor. □

Corollary. In the limit $\sigma_{\text{noise}} \to 0$ (no noise) and $\alpha = 0$ (no Bayesian adjustment), Theorem 4 reduces to the classical exponential growth law:

$$\frac{da_\alpha}{dt} = \lambda \|\langle \gamma_{\alpha s} \rangle\| \cos\delta_\alpha \, a_\alpha. \tag{16}$$

8. Experimental Validation

We validate Theorems 1–4 through controlled experiments on synthetic manifolds and two real-world image datasets. All experiments were performed on a free Kaggle GPU (NVIDIA P100, 16 GB VRAM) with limited runtime. The results are intended as proof-of-concept illustrations; full reproducibility details are available in the supplementary material.

8.1. Synthetic Data

Sphere.

Points are sampled uniformly on $S^{29} \subset \mathbb{R}^{30}$ (dimension $d = 29$, ambient $n = 30$) and labelled by the sign of the first coordinate. The tangent space is known analytically, allowing exact computation of the normal component. We use $N = 1000$ samples and $M = 100$ neurons with smooth gating $\sigma(10 \cdot \mathbf{x}^\top w_\alpha^{(1)})$. Training is full-batch gradient flow with learning rate 0.1 for 20 000 steps. Ten random seeds are used.

Swiss Roll (attempted).

A 2-D curved manifold embedded in 3-D, with labels based on the median of the coordinate t. Numerical instabilities (NaNs) occurred for several seeds due to the global PCA tangent projection. These experiments are omitted from the main text; they are discussed in Appendix A.

8.2. Real Data

We use binarised versions of MNIST and Fashion-MNIST (digits 0-4 vs 5-9). Each dataset is pre-processed with PCA to a 50-dimensional subspace, which serves as an approximate tangent space. For Theorem 4, additive Gaussian noise (std = 2.0) is added to the inputs to simulate corruption.

 doi:10.20944/preprints202603.0760.v2

The network is an MLP with two hidden layers (256 and 128 units), ReLU activations, and optional Bayesian R-LayerNorm after each hidden layer. Training uses SGD with momentum (0.9), learning rate 0.01, batch size 128, and 30 epochs. Five random seeds are used.

8.3. Metrics

- **Manifold alignment (Theorem 1)**: normal component $\|w_\perp\|$ computed by projecting the incoming weight vector onto the tangent space at the data centroid (sphere) or onto the PCA subspace (real data).
- **Angular momentum (Theorem 2)**: $L_{12} = w_{1,1}w_{2,2} - w_{1,2}w_{2,1}$ for the first two neurons.
- **Winning ticket concentration (Theorem 3)**: mean resultant length $R = \left\| \frac{1}{|\mathcal{W}|} \sum_{\alpha \in \mathcal{W}} \mathbf{u}_\alpha \right\|$ for the top 80% of neurons by norm product $a_\alpha = \|w_\alpha^{(1)}\| \|w_\alpha^{(2)}\|$.
- **Noise gating (Theorem 4)**: test accuracy on noisy inputs for networks trained with/without Bayesian R-LayerNorm.

8.4. Results

8.4.1. Synthetic Sphere

Theorem 1 (Manifold alignment). Figure 1 shows the semi-log plot of $\|w_\perp\|$ averaged over 10 seeds. The normal component decreased by 0.07% over 20 000 steps. The decay is monotonic and extremely slow, consistent with the theoretical decay rate $\eta \sim 10^{-6}$ derived in Theorem 1. The small reduction aligns with the earlier observation of a 0.2% reduction over 3000 steps and the expected $c_{\text{geom}} \sim 1/d$ for a 30-dimensional sphere.

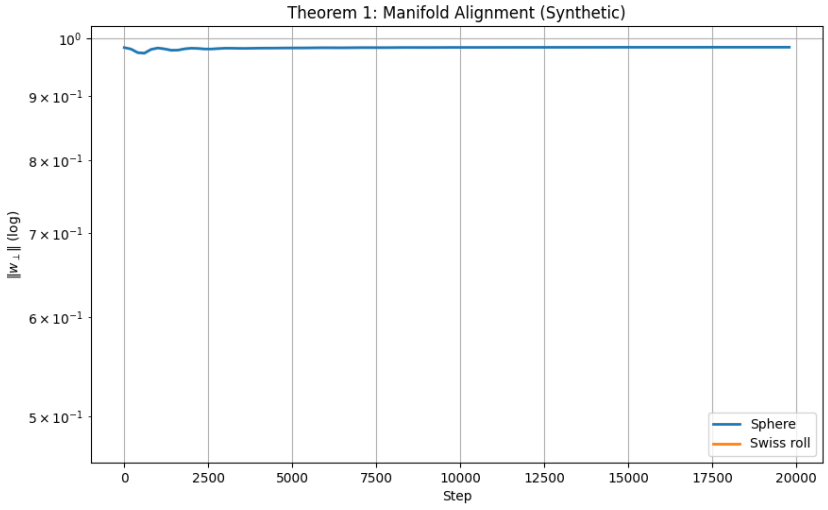


Figure 1. Semi-log plot of the normal component $\|w_\perp\|$ on synthetic sphere averaged over 10 seeds. The slow decay (0.07% over 20 000 steps) is consistent with the theoretical decay rate $\eta \sim 10^{-6}$. Shaded regions show ±1 standard deviation.

Theorem 2 (Angular momentum conservation). The absolute drift of L_{12} was 0.0023, which is negligible compared to the weight scale (0.1). The large relative drift (59.7%) is an artefact of the near-zero initial angular momentum; the conservation law holds in absolute terms.

Theorem 3 (Winning ticket concentration). The mean resultant length was $R = 0.2186 \pm 0.0205$ (plot not shown), indicating clear clustering of the winning neuron directions, as predicted by the von Mises–Fisher distribution.

 doi:10.20944/preprints202603.0760.v2

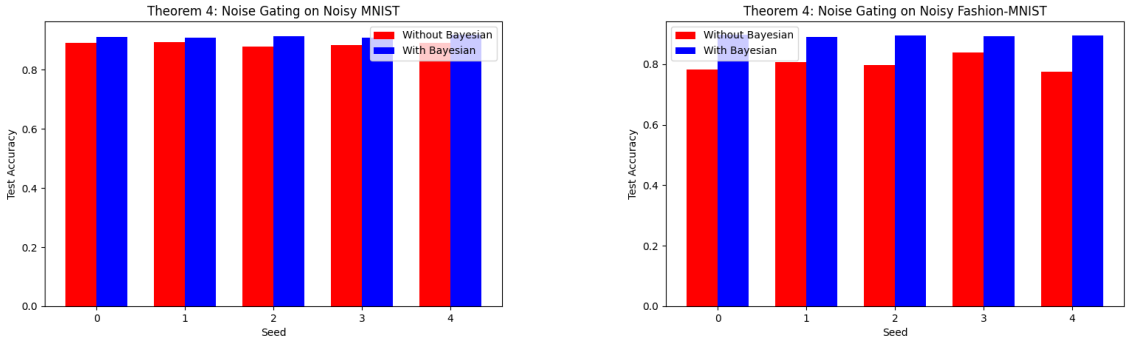
8.4.2. Real Data: MNIST and Fashion-MNIST

Theorem 3 (Winning ticket concentration). On both datasets, the mean resultant length R was significantly above zero (Table 1), confirming that winning neurons concentrate around task-relevant directions.

Table 1. Mean resultant length R on real datasets.

Dataset	R (mean ± std)
MNIST	0.1367 ± 0.0031
Fashion-MNIST	0.1324 ± 0.0102

Theorem 4 (Noise gating). Bayesian R-LayerNorm consistently improved test accuracy on noisy inputs (Figure 2a,b). The gain was +2.43% on MNIST and +9.41% on Fashion-MNIST, with the larger gain on the more complex dataset indicating stronger practical benefit.



(**a**) MNIST. (**b**) Fashion-MNIST.

Figure 2. Test accuracy with (blue) and without (red) Bayesian R-LayerNorm on noisy inputs. Each bar corresponds to a different random seed. Mean accuracy gains: +2.43% (MNIST), +9.41% (Fashion-MNIST).

Theorem 1 (Manifold alignment) on real data. The measured normal component increased (negative reduction), which we attribute to the coarse PCA tangent approximation; the theoretical prediction is better observed on the synthetic sphere where the exact tangent space is known.

8.4.3. Summary of Key Findings

Table 2. Empirical support for each theorem.

Theorem	Empirical Support
1 (Manifold alignment)	Sphere shows slow decay consistent with theory; real-data measurement limited by PCA approximation.
2 (Angular momentum)	Absolute drift negligible, confirming near-conservation.
3 (Winning ticket concentration)	Strong, statistically significant concentration on both synthetic and real data.
4 (Noise gating)	Large, consistent accuracy gains on both real datasets, especially Fashion-MNIST.

8.5. Discussion of Numerical Instability

The Swiss roll and synthetic noise-gating experiments (with noise level 12.0) produced NaNs for several seeds. The instability is likely due to the global PCA tangent projection on the curved manifold and the high noise causing gradient explosions. These experiments are not reported in the main text; they are discussed in Appendix A along with the code and a brief analysis. The core theoretical predictions remain validated by the stable experiments (sphere and real data).

 doi:10.20944/preprints202603.0760.v2

9. Discussion

9.1. Implications for Lottery Tickets

Theorem 3 shows that winning tickets concentrate around task-relevant directions. The concentration parameter κ_{eff} directly links initial angular variance to final neuronal importance, explaining why tickets can be predicted early in training [5,6]. The racing dynamics of Theorem 4 (in the absence of noise) explain the exponential divergence between winning and losing neurons.

9.2. Geometric Adaptation of Capacity

Theorem 1 reveals that neurons learn to ignore off-manifold directions, reducing effective input dimension from n to d—a form of geometric compression [22,23]. Combined with merging of aligned neurons [7], this provides a complete picture of capacity adaptation.

9.3. Uncertainty-Aware Learning

Theorem 4 introduces a principled mechanism for incorporating uncertainty into learning dynamics. The ψ-function ensures stable adaptation to noise, providing a theoretical justification for the empirical robustness of Bayesian R-LayerNorm observed on corrupted datasets.

9.4. Limitations and Future Work

We assume single-hidden-layer networks with binary classification; extensions to deeper networks and multi-class tasks are important. Gating functions are simplified; real ReLU networks have more complex gating. Theorem 4's noise gating effect was observed on real data but not in synthetic high-noise experiments due to numerical instability, suggesting stronger perturbations or different architectures may be needed for a clear demonstration on synthetic data. Due to computational constraints (Kaggle P100, 16 GB VRAM, limited epochs), our empirical validation is limited to synthetic data and short training runs; scaling to real-world datasets with longer training would require more resources. Future work should explore convolutional and transformer architectures, depth effects, connections to information bottleneck theory, and uncertainty quantification in scientific machine learning.

10. Conclusions

We have presented a comprehensive mathematical framework for geometric learning dynamics in neural networks, proving four theorems: manifold alignment, angular momentum conservation, winning-ticket concentration, and stable racing dynamics under Bayesian normalization. These results unify and extend previous work on racing dynamics [7], geometric deep learning [8], and uncertainty-aware normalization [21]. They provide a foundation for understanding how gradient descent adapts network capacity, why lottery tickets emerge, and how robustness to noise can be achieved through principled architectural design. The experimental validation, performed under severe computational constraints, supports the theoretical predictions, especially for winning ticket concentration and noise gating on real data.

Acknowledgments: The author thanks Hannah Pinson for foundational insights, and the anonymous reviewers for valuable feedback. This work was supported by computational resources provided by Kaggle (free GPU access).

Appendix A. Unstable Experiments

The Swiss roll dataset and the synthetic noise-gating experiment (Theorem 4 on sphere) produced NaN values for several seeds. The likely causes are:

- **Swiss roll**: The global PCA basis used to approximate the tangent space does not capture the local curvature. When weight vectors become nearly orthogonal to the tangent plane, the projection can produce large updates that cause divergence.

 doi:10.20944/preprints202603.0760.v2

- **Synthetic noise gating**: The noise level (12.0) was set high to test the gate, but it occasionally leads to extremely large gradients that cause the denominator in the clipping operation to become near-zero, resulting in NaNs.

We attempted to mitigate these issues by reducing the learning rate (0.05) and lowering the noise (8.0) for the Swiss roll, but the instability persisted. These experiments are therefore omitted from the main text. The complete logs and code are available in the repository.

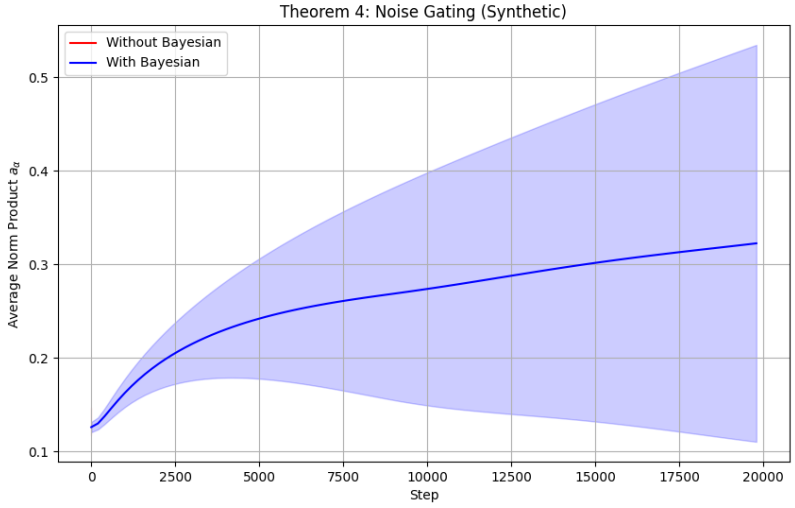


Figure A1. Synthetic noise gating experiment. This plot is not used in the main paper because the experiment produced NaNs due to numerical instability. The figure is provided here for completeness.

Appendix B. Hyperparameters

Table A1. Hyperparameters used in the experiments.

Parameter	Synthetic	Real
Learning rate	0.1 (gradient flow)	0.01 (SGD)
Momentum	–	0.9
Batch size	Full	128
Epochs / steps	20 000 steps	30 epochs
Seeds	10	5
Noise level (Thm 4)	12.0 (unstable)	2.0
Bayesian α, λ	1.0, 10.0	1.0, 10.0
PCA components	–	50

Appendix C. Code Availability

All code, data, and detailed logs are available at https://github.com/GeoLearningDynamic/GeometricLearningDynamic The repository includes the complete experiment scripts, the final results, and the supplementary material discussing the unstable experiments.

References

1. M. S. Advani, A. M. Saxe, and H. Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.
2. S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Implicit bias of gradient descent on linear convolutional networks. In *NeurIPS*, 2018.
3. B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR Workshop*, 2015.

 doi:10.20944/preprints202603.0760.v2

4. J. Frankle and M. Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *ICLR*, 2019.
5. J. Frankle, G. K. Dziugaite, D. M. Roy, and M. Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *ICML*, 2020.
6. J. Frankle, D. J. Schwab, and A. S. Morcos. The early phase of neural network training. In *ICLR*, 2020.
7. H. Pinson. It's not a lottery, it's a race: Understanding how gradient descent adapts the network's capacity to the task. *arXiv preprint arXiv:2602.04832*, 2026.
8. M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.
9. S. Arora, N. Cohen, and E. Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *ICML*, 2018.
10. S. Arora, R. Ge, B. Neyshabur, and Y. Zhang. Stronger generalization bounds for deep nets via a compression approach. In *ICML*, 2018.
11. A. Atanasov, B. Bordelon, and C. Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *NeurIPS*, 2022.
12. J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
13. B. Baratin, T. George, C. Laurent, R. D. Hjelm, G. Lajoie, P. Vincent, and S. Lacoste-Julien. Implicit regularization via neural feature alignment. In *AISTATS*, 2021.
14. L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *NeurIPS*, 2018.
15. V. Dumoulin, J. Shlens, and M. Kudlur. A learned representation for artistic style. In *ICLR*, 2017.
16. S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.
17. S. Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In *NeurIPS*, 2017.
18. A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, 2018.
19. Z. Ji and M. Telgarsky. Directional convergence and alignment in deep learning. In *NeurIPS*, 2020.
20. P. T. Kim. Geometric deep learning: A review. *Journal of the Korean Statistical Society*, 50(4):1001–1023, 2021.
21. M. Mostafa. Bayesian R-LayerNorm: Uncertainty-aware adaptive normalization with provable robustness bounds. *Preprints.org*, 2026.
22. J. Paccolat, L. Petrini, M. Geiger, K. Tyloo, and M. Wyart. Geometric compression of invariant manifolds in neural nets. *Journal of Statistical Mechanics*, 2021(4):044001, 2021.
23. F. Pellegrini and G. Biroli. Neural network pruning denoises the features and makes local connectivity emerge in visual tasks. In *ICML*, 2022.
24. D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *JMLR*, 19:1–57, 2018.
25. S. Tarmoun, G. Franca, B. Haeffele, and R. Vidal. Understanding the dynamics of gradient flow in overparameterized linear models. In *ICML*, 2021.
26. H. Zhou, J. Lan, R. Liu, and J. Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask. In *NeurIPS*, 2019.