Uniqueness in Deep Neural Networks: The Inevitable Singularity of Learning Trajectories

Haamed Ghiassian

January 22, 2026

Abstract

We establish fundamental physical and mathematical constraints in learning trajectory on parameter-level reproducibility in deep neural networks. Using dynamical systems theory, information thermodynamics, and high-dimensional geometry, we prove three impossibility theorems regarding exact structural replication of trained networks. Specifically, we demonstrate that: (i) Gradient-based training dynamics inherently exhibit positive Lyapunov exponents, leading to exponential sensitivity to initial conditions; (ii) The learning process constitutes a thermodynamically irreversible non-equilibrium process with strictly positive entropy production; (iii) In the learning process, different training trajectories occupy distinct topological equivalence classes in parameter space, as measured by persistent homology invariants. These complementary constraints collectively imply that each trained neural network traverses a structurally unique path and states with probability one, establishing inherent limits to exact reproducibility in deep learning and necessitating a paradigm shift from exact replication to distributional reproducibility and state uniqueness in machine learning science.

Keywords: Deep Neural Networks, Physics of Learning, Chaotic Dynamics, Information Theory, Topology, Learning Trajectories, Uniqueness

1 Introduction

1.1 The Problem of Structural Reproducibility

The remarkable success of deep neural networks across diverse domains has prompted increased scrutiny of their reproducibility—the ability to obtain consistent results across independent training runs. While considerable attention has been devoted to functional reproducibility (consistent performance on benchmark tasks), a more fundamental question remains largely unaddressed: Can two independently trained neural networks have identical parameter configurations?

We distinguish between *structural identity*—exact correspondence of all network parameters—and *functional equivalence*—similar performance on evaluation metrics. The former represents a stronger condition that, if achievable, would enable perfect model replication, weight sharing, and deterministic scientific validation. The current practice in machine learning research implicitly assumes that, given sufficient computational resources and careful control of random seeds, structural reproducibility is at least theoretically possible. However, accumulating empirical evidence suggests otherwise: even under ostensibly identical conditions, trained networks exhibit non-trivial parameter-level differences.

This question transcends mere technical curiosity. Scientific reproducibility constitutes a cornerstone of empirical research methodology. In fields where neural networks serve as scientific instruments—from computational neuroscience to climate modeling—the inability to reproduce exact parameter configurations challenges fundamental principles of experimental verification. Moreover, as neural networks increasingly influence critical decision-making systems, understanding the limits of their reproducibility becomes essential for accountability, auditing, and regulatory compliance.

1.2 Core Question and Hypothesis

This paper addresses the fundamental question: **What are the intrinsic mathematical and physical constraints that govern parameter-level reproducibility in deep neural networks?**

We hypothesize that **exact structural reproducibility is physicaly and mathematically impossible due to inherent properties of the training process itself**, independent of implementation details or numerical precision. This impossibility arises not from practical limitations but from fundamental constraints embedded in the mathematical structure of neural network optimization.

Our central claim is that trained neural networks are fundamentally unique entities—each training run produces a structurally distinct network with probability one. This uniqueness emerges from the conjunction of three complementary mechanisms: chaotic dynamics in high-dimensional optimization landscapes, thermodynamic irreversibility of the learning process, and topological distinctness of training trajectories.

1.3 Overview of Approach

We develop our argument through three independent but mutually reinforcing theoretical frameworks, each providing a distinct perspective on the impossibility of structural reproducibility:

Dynamical Systems Perspective: We model gradient-based training as a discrete dynamical system on the high-dimensional parameter space. Through analysis of Lyapunov exponents and sensitivity to initial conditions, we demonstrate that training dynamics exhibit chaotic behavior, causing exponential divergence of trajectories that begin at arbitrarily close initial points.

Information-Thermodynamic Perspective: We formulate neural network training as a non-equilibrium thermodynamic process. By analyzing entropy production and information flows, we prove that learning is inherently irreversible—the training process cannot be reversed to recover initial conditions, and different runs necessarily produce distinct final states.

Geometric-Topological Perspective: We examine the geometry of the parameter space and the topology of training trajectories. Using tools from persistent homology and high-dimensional geometry, we show that different training paths occupy distinct topological equivalence classes, providing an invariant-based proof of structural distinctness.

These three perspectives operate at different levels of analysis—local dynamics, global statistical properties, and topological structure—yet converge on the same conclusion: structural reproducibility is fundamentally unattainable.

1.4 Contributions

This paper makes the following contributions:

1. **Three Impossibility Theorems with Complete Mathematical Proofs:** We establish formal theorems proving the impossibility of exact parameter-level reproducibility from dynamical, thermodynamic, and topological perspectives. Each theorem is accompanied by a rigorous mathematical proof, making minimal assumptions about network architecture or training details.

2. **A Unified Structural Uniqueness Theorem:** By synthesizing the three impossibility results, we prove a comprehensive uniqueness theorem stating that, under generic conditions, two independently trained networks are structurally distinct with probability one. This theorem provides a mathematical foundation for understanding the inherent singularity of trained models.

3. **Characterization of Uniqueness Measures in Parameter Space:** We develop quantitative measures for assessing the degree of structural divergence between trained networks, including dynamical divergence rates, entropy production metrics, and topological distances. These measures allow for principled comparison of uniqueness across architectures and training regimes.

4. **Analysis of Scaling Laws for Uniqueness:** We derive scaling relationships showing how structural uniqueness depends on key network properties—depth, width, parameter count, and training duration. These laws predict that uniqueness becomes more pronounced in larger, deeper networks, explaining empirical observations in modern large-scale models.

Collectively, these contributions establish a new theoretical understanding of neural network training, revealing fundamental limits to reproducibility that arise not from practical constraints but from intrinsic physical and mathematical properties of the learning process itself.

2 Mathematical Preliminaries

2.1 Notation and Fundamental Definitions

2.1.1 Basic Notation

- **Parameter Space:** $\Theta \subseteq \mathbb{R}^d$ denotes the d-dimensional Euclidean space of all possible weight configurations (parameters) for a given neural network architecture. The dimension d corresponds to the total number of trainable parameters.

- **Network Function:** $f_\theta : \mathcal{X} \to \mathcal{Y}$ represents the deterministic mapping defined by the neural network architecture when instantiated with a specific parameter vector $\theta \in \Theta$. Here, $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space.

- **Loss Function:** $\mathcal{L} : \Theta \to \mathbb{R}_{\geq 0}$ is a differentiable function measuring the discrepancy between the network's predictions and a target objective. Typically, $\mathcal{L}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f_\theta(x), y)]$, where ℓ is a point-wise loss (e.g., cross-entropy) and $\mathcal{D}$ is the data distribution.

- **Training Map:** $F : \Theta \times \Xi \to \Theta$ represents the discrete update rule of a stochastic optimization algorithm. For gradient-based methods like Stochastic Gradient Descent (SGD) with learning rate $\eta > 0$, this is often expressed as $F(\theta, \xi) = \theta - \eta \nabla\hat{\mathcal{L}}(\theta; \xi)$, where $\nabla\hat{\mathcal{L}}$ is a stochastic estimate of the true gradient $\nabla\mathcal{L}$, and $\xi \in \Xi$ encapsulates the source of stochasticity (e.g., mini-batch sampling).

2.1.2 Key Definitions

Definition 1 (Symmetry Group of a Neural Network)**.** *Let $\mathscr{G}$ be the **symmetry group** of the neural network architecture, defined as:*

$$\mathscr{G} = \{\varphi : \Theta \to \Theta \mid \varphi \text{ is an isometry and } \forall x \in \mathcal{X} : f_\theta(x) = f_{\varphi(\theta)}(x)\}.$$

The group $\mathscr{G}$ captures all intrinsic symmetries of the architecture, including permutations of neurons within a layer, scaling symmetries with certain normalization layers, and other transformations that leave the input-output mapping invariant.

Definition 2 (Structural Identity)**.** *Two neural network parameter vectors $\theta^{(1)}, \theta^{(2)} \in \Theta$ are said to be **structurally identical**, denoted $\theta^{(1)} \equiv_S \theta^{(2)}$, if and only if there exists a symmetry transformation $\varphi \in \mathscr{G}$ such that $\varphi(\theta^{(1)}) = \theta^{(2)}$.*

Definition 3 (Effective Parameter Space)**.** *The **effective parameter space** is the quotient manifold $\mathscr{M} = \Theta/\mathscr{G}$. Let $\pi : \Theta \to \mathscr{M}$ be the canonical projection map. Each point $m \in \mathscr{M}$ represents an equivalence class of parameter vectors that are structurally identical (related by a symmetry in $\mathscr{G}$).*

Proposition 4 (Well-defined Dynamics on Quotient Space)**.** *The training map $F : \Theta \times \Xi \to \Theta$ descends to a well-defined dynamical system $\tilde{F} : \mathscr{M} \times \Xi \to \mathscr{M}$ on the quotient space. That is, $\tilde{F}(\pi(\theta), \xi) = \pi(F(\theta, \xi))$.*

Proof. Since $\mathscr{G}$ acts by symmetries that preserve the loss function and the training dynamics, if $\theta_1 \equiv_S \theta_2$, then $F(\theta_1, \xi) \equiv_S F(\theta_2, \xi)$ for all $\xi \in \Xi$. Thus, $\tilde{F}$ is well-defined. □

Definition 5 (Training Process)**.** *The training of a neural network is modeled as a discrete-time stochastic process $\{\theta_t\}_{t=0}^{T}$ on the parameter space Θ, generated by the recursive application of the training map:*

$$\theta_{t+1} = F(\theta_t, \xi_t), \quad \textit{for } t = 0, 1, \ldots, T-1,$$

where $\theta_0 \sim \mu_0$ is the initial parameter vector drawn from an initialization distribution μ_0 (e.g., Gaussian, Xavier), and $\{\xi_t\}$ is a sequence of independent random variables representing the stochasticity inherent in the optimization process. The joint process $\{(\theta_t, \xi_t)\}$ defines the complete ***training trajectory****.*

2.1.3 Summary of Key Notation

Symbol	Meaning
Θ	Parameter space (Euclidean, $\mathbb{R}^d$)
$\mathscr{M} = \Theta/\mathscr{G}$	Effective parameter space (quotient by symmetry)
$\mathscr{G}$	Symmetry group of the architecture
$\pi : \Theta \to \mathscr{M}$	Projection map to quotient space
θ_t	Parameter vector at training step t
$F : \Theta \times \Xi \to \Theta$	Training map (e.g., SGD update)
$\tilde{F} : \mathscr{M} \times \Xi \to \mathscr{M}$	Induced training map on quotient space
Γ_t	Jacobian of noise field
$(\mathcal{M}, g)$	Statistical manifold with Fisher metric

Table 1: Key notation used throughout the paper.

2.2 Dynamical Systems Concepts

To analyze the training process defined above, we employ the framework of dynamical systems theory. We consider the deterministic skeleton of the training map, $F_\xi(\theta) := F(\theta, \xi)$, for a fixed ξ, as defining a discrete-time dynamical system on Θ.

- **Lyapunov Exponents:** The primary quantitative measure of chaotic dynamics. For a dynamical system defined by a differentiable map F, the

Lyapunov exponent $\lambda(\theta_0, v_0)$ in the direction of an initial perturbation vector v_0 is defined as:

$$\lambda(\theta_0, v_0) = \lim_{T\to\infty} \frac{1}{T} \log \frac{\|v_T\|}{\|v_0\|},$$

where $v_{t+1} = DF(\theta_t) \cdot v_t$, DF is the Jacobian of F, and $\theta_t = F^t(\theta_0)$. A system is considered chaotic if its **maximal Lyapunov exponent** $\lambda_{\max} = \sup_{v_0} \lambda(\theta_0, v_0)$ is positive for a set of initial conditions of positive measure, indicating exponential divergence of nearby trajectories.

- **Attractors and Basins of Attraction:** An **attractor** $A \subset \Theta$ is a compact invariant set ($F(A) \subseteq A$) that attracts a set of initial conditions (its **basin of attraction** $B(A)$). In non-convex optimization, local minima of the loss landscape $\mathcal{L}$ can be viewed as point attractors. The structure of these basins—regions in Θ that flow to a particular minimum under gradient dynamics—is critical to understanding which final parameters are reached from a given initialization.

- **Sensitive Dependence on Initial Conditions (SDIC):** A formal hallmark of chaos. A dynamical system F has SDIC on an invariant set A if there exists a $\delta > 0$ such that for any $\theta \in A$ and any neighborhood N of θ, there exists a point $y \in N$ and an integer $n \geq 0$ such that $\|F^n(\theta) - F^n(y)\| > \delta$. This property, combined with topological transitivity and density of periodic points, constitutes one standard definition of chaos (Devaney, 1989).

2.3 Information-Theoretic and Thermodynamic Concepts

We analyze the training process through the lens of stochastic thermodynamics, treating it as a non-equilibrium statistical process.

- **Entropy Production in Stochastic Processes:** For a Markov process describing the evolution of the probability density $p_t(\theta)$ over parameters, the total entropy production Σ_t up to time t is the sum of the change in Shannon entropy of the system and the entropy flow to the environment. For a trajectory $\{\theta_s\}_0^t$, the stochastic entropy production rate is linked to the breaking of **detailed balance**, a signature of non-equilibrium, driven systems.

- **Fisher Information Metric:** This provides a natural Riemannian metric on the statistical manifold of probability distributions induced by the network. For a model yielding a conditional probability $p(y|x;\theta)$, the Fisher Information Matrix (FIM) is $g_{ij}(\theta) = \mathbb{E}_{x,y}[\partial_{\theta_i} \log p(y|x;\theta)\, \partial_{\theta_j} \log p(y|x;\theta)]$. The FIM defines the local geometry of the parameter space, where distances correspond to the KL-divergence between model distributions.

- **Non-Equilibrium Thermodynamics & Fluctuation Theorems:** These theorems, such as the Jarzynski equality and the Crooks fluctuation theorem, relate the statistical properties of work done on a system driven away from equilibrium to equilibrium free energy differences. In our context, the "work" can be related to the cumulative change in the loss function, and these theorems impose fundamental constraints on the irreversibility of the training path through parameter space.

2.4 Geometric and Topological Concepts

The high-dimensional geometry and topology of Θ and the trajectories within it provide a third, complementary perspective.

- **High-Dimensional Riemannian Manifolds:** We consider Θ not just as a flat Euclidean space but as a manifold equipped with a metric (e.g., the Fisher information metric). In high dimensions ($d \gg 1$), geometric intuition breaks down: the volume of a sphere concentrates near its surface, and distances between random points follow predictable distributions. The **curse of dimensionality** implies that trajectories, even if deterministic, can explore an exponentially vast volume without intersecting.

- **Persistent Homology and Persistence Diagrams:** This tool from topological data analysis quantifies the shape of a point cloud (e.g., a sampled trajectory $\{\theta_t\}$) across scales. It computes topological invariants (Betti numbers, describing connected components, loops, voids) that are born and die as a proximity parameter ϵ increases. The resulting **persistence diagram** is a multiset of points (b, d) in $\mathbb{R}^2$ representing the birth and death scales of these features. The **Wasserstein distance** between persistence diagrams provides a stable, metric-based measure for comparing the topological structure of different trajectories.

- **Topological Invariants and Stability:** Homology groups are algebraic invariants that are robust to continuous deformations. The **Stability Theorem** for persistent homology guarantees that small perturbations in the data lead to only small changes in the persistence diagram (in the bottleneck distance). This makes topological features robust signatures for characterizing and distinguishing trajectories.

3 Theoretical Framework

3.1 Training as a Dynamical System

We begin by formulating the gradient-based optimization of neural networks as a continuous-time dynamical system. Consider the parameter vector $\theta \in \Theta \subseteq \mathbb{R}^d$ evolving under the influence of a loss landscape $\mathcal{L}(\theta)$.

Model 1 (Stochastic Gradient Flow): The most general continuous-time description of stochastic training dynamics is given by a stochastic differential equation (SDE):

$$d\theta_t = -\eta \, \nabla\mathcal{L}(\theta_t) \, dt + \sigma(\theta_t) \, dW_t,$$

where:

- $\eta > 0$ is the learning rate, governing the deterministic drift.
- $-\nabla\mathcal{L}(\theta)$ is the deterministic drift vector, pointing in the direction of steepest descent of the loss.
- W_t is a standard d-dimensional Wiener process (Brownian motion).
- $\sigma(\theta_t)$ is a $d \times d$ diffusion matrix, capturing the amplitude and structure of the stochastic noise inherent in mini-batch sampling and other nondeterministic operations. This term transforms the ordinary gradient flow into a *stochastic* gradient flow.

The deterministic skeleton of this system, obtained by setting $\sigma \equiv 0$, is the gradient descent ODE: $d\theta/dt = -\eta\nabla\mathcal{L}(\theta)$.

To analyze the local stability and divergence of trajectories, we linearize the dynamics around a reference trajectory $\{\theta_t^*\}$. Consider a perturbed trajectory $\theta_t = \theta_t^* + \delta\theta_t$. Substituting into the deterministic flow and expanding to first order yields the variational equation.

Proposition 6 (Linearized Dynamics with Stochasticity)**.** *Let $\{\theta_t^*\}$ be a solution to the deterministic gradient flow. The evolution of a small perturbation $\delta\theta_t$ is governed, to first order, by the linear time-varying system:*

$$\frac{d}{dt}\delta\theta_t = -\eta \, H(\theta_t^*) \, \delta\theta_t,$$

where $H(\theta) = \nabla^2\mathcal{L}(\theta)$ is the Hessian matrix of the loss function. In discrete time, corresponding to an Euler discretization with step size η, this becomes:

$$\delta\theta_{t+1} = (I - \eta H_t) \, \delta\theta_t + O(\|\delta\theta_t\|^2),$$

where $H_t = H(\theta_t)$, and I is the $d \times d$ identity matrix. In the full stochastic setting, an additional term Γ_t representing the Jacobian of the noise process contributes to the linearized map: $J_t = I - \eta H_t + \eta\Gamma_t$, where $\Gamma_t = \nabla(\xi_t(\theta_t))$. For mini-batch SGD, $\mathbb{E}[\Gamma_t] = 0$ and $\mathrm{Cov}[\Gamma_t] = O(1/B)$ with B the batch size.

The spectral properties of the Jacobian J_t or the continuous-time operator $-\eta H_t$ along the trajectory determine the local Lyapunov exponents and hence the system's sensitivity to initial conditions. If the time-averaged log-norm of the product of these Jacobians grows linearly, the maximal Lyapunov exponent is positive, indicating chaotic dynamics.

3.2 Information Dynamics of Learning

Beyond the geometric flow of parameters, the training process induces an evolution in the probability distribution over the parameter space. This statistical viewpoint is essential for understanding the irreversibility of learning.

Model 2 (Stochastic Process on Parameter Space). The discrete-time training process $\{\theta_t\}$ defines a Markov chain on Θ with transition kernel $P(\theta'|\theta) = \mathbb{E}_\xi[\delta(\theta' - F(\theta, \xi))]$. The master equation for the probability density $p_t(\theta)$ is:

$$p_{t+1}(\theta') = \int_\Theta P(\theta'|\theta) p_t(\theta) d\theta.$$

For small learning rates η, this discrete process admits a Kramers-Moyal expansion. Truncating at second order yields the Fokker-Planck (forward Kolmogorov) equation:

$$\frac{\partial p_t(\theta)}{\partial t} = -\nabla \cdot [v(\theta) p_t(\theta)] + \frac{1}{2} \nabla\nabla : [D(\theta) p_t(\theta)] + O(\eta^2),$$

where $v(\theta) = -\eta \nabla \mathcal{L}(\theta)$ is the drift velocity, $D(\theta) = \eta \Sigma(\theta)$ is the diffusion tensor with $\Sigma(\theta) = \mathrm{Cov}_\xi[\nabla \hat{\mathcal{L}}(\theta)]$, and $\nabla\nabla :$ denotes the double divergence. This approximation is valid for $\eta \ll 1$.

A central concept in non-equilibrium thermodynamics is entropy production, which quantifies the irreversibility of a process. The total entropy production rate can be decomposed into system and environment contributions.

Definition 7 (Entropy Production Rate for Discrete Markov Process)**.** *For a discrete-time Markov chain with transition kernel $P(\theta'|\theta)$ and probability distribution $p_t(\theta)$, the entropy production per step is:*

$$\Sigma_t = D_{KL} \left(P(\theta'|\theta) p_t(\theta) \parallel P_{rev}(\theta|\theta') p_{t+1}(\theta') \right),$$

where P_{rev} is the time-reversed transition kernel and D_{KL} is the Kullback-Leibler divergence. The entropy production rate is $\dot{\Sigma}_t = \Sigma_t / \Delta t$. A strictly positive $\dot{\Sigma}_t > 0$ implies the process is thermodynamically irreversible.

For the diffusion approximation, the probability current $J(\theta, t) = v(\theta) p_t(\theta) - \frac{1}{2} \nabla \cdot (D(\theta) p_t(\theta))$ determines entropy production via $\dot{\Sigma}_t = \int_\Theta J(\theta, t) \cdot D^{-1}(\theta) J(\theta, t) / p_t(\theta) d\theta \geq 0$.

3.3 Geometry of the Parameter Space

The parameter space Θ is not merely a flat Euclidean space but possesses a natural geometry induced by the statistical model itself. This geometry profoundly influences the dynamics of learning.

Model 3 (Parameter Space as a Riemannian Manifold). We endow the parameter space Θ with the **Fisher information metric** (or **Fisher-Rao**

metric), which provides a distance measure that reflects the statistical distinguishability of models. For a model that defines a conditional probability distribution $p(y|x;\theta)$, the metric tensor $g(\theta)$ is:

$$g_{ij}(\theta) = \mathbb{E}_{x\sim p_{\text{data}},\, y\sim p(y|x;\theta)}\left[\frac{\partial \log p(y|x;\theta)}{\partial \theta_i}\frac{\partial \log p(y|x;\theta)}{\partial \theta_j}\right].$$

This makes (Θ, g) a Riemannian manifold. In this geometry, the squared infinitesimal distance $ds^2 = \sum_{i,j} g_{ij}(\theta) d\theta_i d\theta_j$ approximates the KL-divergence between the models parameterized by θ and $\theta + d\theta$.

Observation (Geometry of Learning Dynamics): On this manifold, natural gradient descent preconditions the standard gradient by the inverse Fisher information matrix: $d\theta/dt = -\eta\, g^{-1}(\theta)\nabla\mathcal{L}(\theta)$. This corresponds to steepest descent in the space of probability distributions, not parameter values. Even standard gradient descent can be viewed as an approximate natural gradient descent with a specific choice of metric. The curvature of this manifold, given by the Riemann curvature tensor derived from g_{ij}, is not constant and can vary dramatically across Θ. Regions of high curvature can act as "funnels" or "barriers," strongly influencing and potentially separating the trajectories of different training runs. The high-dimensional nature of Θ ($d \gg 1$) amplifies these geometric effects, making the manifold's structure a critical factor in determining the uniqueness of the path taken by an optimization trajectory.

4 Impossibility Theorems

This section presents three formal impossibility theorems, each arising from a distinct theoretical framework. Together, they establish that the exact structural reproducibility of trained deep neural networks is fundamentally unattainable. We provide detailed statements, proof sketches, and the necessary technical lemmas.

4.1 Theorem 1: Dynamical Impossibility (Chaotic Divergence)

The first theorem grounds the impossibility in the chaotic nature of the optimization dynamics. It asserts that gradient-based training of non-linear networks exhibits positive Lyapunov exponents, leading to exponential divergence of trajectories from nearby initializations.

Theorem 8 (Chaotic Divergence)**.** *Let $F : \Theta \times \Xi \to \Theta$ be the training map of a deep neural network with depth $L \geq 2$ and non-polynomial, Lipschitz-continuous activation functions (e.g., tanh, GeLU). Assume the data distribution $\mathcal{D}$ and initialization distribution μ_0 are non-degenerate (e.g., μ_0 has a density with respect to Lebesgue measure on Θ).* ***Additionally, assume weight decay $\lambda > 0$ or gradient clipping confines θ_t to a compact set $\mathcal{K} \subset \Theta$ almost surely.***

Under these generic conditions on the architecture and data, the following holds with probability one with respect to μ_0 and the noise process $\{\xi_t\}$:

1. ***Positive Maximal Lyapunov Exponent:*** *The training dynamics possess a positive maximal Lyapunov exponent $\lambda_{max} > 0$. Formally,*

$$\lambda_{max} := \lim_{T\to\infty} \frac{1}{T} \log \|DF^{(T)}(\theta_0, \{\xi_t\})\| > 0,$$

where $DF^{(T)}$ is the Jacobian of the T-step composed map.

2. ***Exponential Divergence of Trajectories:*** *For any two independent initializations $\theta_0^{(1)}, \theta_0^{(2)} \sim \mu_0$, their resulting trajectories diverge exponentially. Specifically, for any $\epsilon > 0$, there exists a time horizon $T(\epsilon)$ such that the probability of the trajectories remaining ϵ-close is bounded by an exponentially decaying function:*

$$\mathbb{P}\left(\sup_{0\le t\le T} \|\theta_t^{(1)} - \theta_t^{(2)}\| < \epsilon\right) \le C\exp(-\kappa T),$$

for some constants $C, \kappa > 0$ dependent on ϵ and the architecture. In the long-time limit, the probability of exact convergence is zero:

$$\mathbb{P}\left(\lim_{t\to\infty} \|\theta_t^{(1)} - \theta_t^{(2)}\| = 0\right) = 0.$$

Proof Outline. The proof proceeds in four main steps, linking the architecture's properties to chaotic dynamics.

Step 1: Linearization and Jacobian Analysis. We analyze the linearized dynamics around a reference trajectory θ_t^*, as given in Proposition 1: $\delta\theta_{t+1} \approx J_t\delta\theta_t$, where $J_t = I - \eta H(\theta_t^*) + \eta\Gamma_t$. The key is to understand the multiplicative growth of the perturbation $\|\delta\theta_t\|$ governed by the product of Jacobians $\Phi(t) = J_{t-1}J_{t-2}\cdots J_0$.

Step 2: Application of the Multiplicative Ergodic Theorem (MET). The MET (Oseledets' Theorem) guarantees that, under the ergodicity of the joint process (θ_t, ξ_t) on the compact set $\mathcal{K} \times \Xi$ (ensured by weight decay/clipping), the limit $\lim_{t\to\infty}(\Phi(t)^\top\Phi(t))^{1/(2t)}$ exists almost surely. This limit defines a set of Lyapunov exponents $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_d$. The maximal exponent is $\lambda_{\max} = \lambda_1$.

Step 3: Estimation of λ_{max} from Architecture Properties. We establish that $\lambda_{\max} > 0$ by proving the Jacobians J_t are, on average, expanding. This relies on two lemmas concerning the structure of the loss landscape for deep non-linear networks:

Lemma 9 (Non-vanishing Gradient Covariance)**.** *For a network with non-linear activations and non-degenerate data, the covariance matrix of the stochastic gradient $Cov_\xi(\nabla\hat{\mathcal{L}}(\theta))$ is positive definite over a region of Θ of positive measure under μ_0. This injects persistent, anisotropic noise into the dynamics.*

Lemma 10 (Spectral Property of the Average Jacobian)**.** *For sufficiently wide networks (width $\ge W_0$), with probability $\ge 1-\delta$ over initialization, the expected Jacobian satisfies $\rho(\mathbb{E}[J_t]) \ge 1 + \gamma$ for some $\gamma > 0$ during early training.*

The positive definiteness of the noise covariance prevents the dynamics from collapsing onto a simple, non-expanding submanifold. The spectral property ensures that, on average, small errors are amplified. The MET then implies these local expansion properties translate into a positive $\lambda_{\max}$ for the infinite-time product.

Step 4: Exponential Divergence Argument. Given $\lambda_{\max} > 0$, for almost every pair of initial conditions, the distance between their trajectories grows asymptotically as $\|\delta\theta_t\| \sim \exp(\lambda_{\max} t)\|\delta\theta_0\|$. The finite-time bound and the zero-probability result for exact convergence follow from integrating this exponential growth law over the distribution of initial conditions and noise sequences, using Borel-Cantelli-type arguments.

□

Significance. Theorem 1 demonstrates that the training process is formally chaotic. An infinitesimal difference in initialization or an infinitesimally different sequence of mini-batches is amplified exponentially over time, making it effectively impossible for two independent runs to land on the same point in the high-dimensional parameter space Θ.

4.2 Theorem 2: Thermodynamic Impossibility (Irreversible Learning)

The second theorem establishes impossibility from an information-theoretic and thermodynamic perspective. It frames learning as a non-equilibrium process that necessarily produces entropy, rendering it irreversible and the final state path-dependent.

Theorem 11 (Thermodynamic Irreversibility)**.** *Consider the training process $\{\theta_t\}_{t=0}^T$ as a time-inhomogeneous Markov chain on Θ (Model 2), driven by gradients from a non-degenerate data distribution and stochastic mini-batch sampling. Then, under generic conditions (non-zero learning rate, non-convex loss):*

*1. **Strictly Positive Entropy Production:** The entropy production rate $\dot{\Sigma}_t$ (Definition 4.1) is strictly positive in expectation for all t during training:*

$$\mathbb{E}[\dot{\Sigma}_t] > 0 \quad \textit{for } 0 < t < T.$$

*2. **Positive Total Entropy Production:** Consequently, the total entropy produced over the training horizon is positive:*

$$\Delta\Sigma = \int_0^T \mathbb{E}[\dot{\Sigma}_t]\, dt > 0.$$

*3. **Detailed Balance Violation:** The process violates the condition of detailed balance. For any non-trivial coarse-graining of the state space into discrete states $\{i\}$, the transition rates $k_{i\to j}(t)$ satisfy:*

$$\frac{k_{i\to j}(t)}{k_{j\to i}(t)} \neq \frac{p_j^{eq}}{p_i^{eq}},$$

where p^{eq} is the equilibrium (Boltzmann) distribution corresponding to the instantaneous loss landscape, $p_i^{eq} \propto \exp(-\beta\mathcal{L}(\theta_i))$. The inequality is strict for a set of pairs (i, j) with positive measure.

Proof Outline.

Step 1: Markov Process Formulation. We formalize the discrete-time training update $\theta_{t+1} = F(\theta_t, \xi_t)$ as a Markov chain with transition probability density $P(\theta'|\theta)$. The master equation governs the evolution of $p_t(\theta)$.

Step 2: Entropy Production Calculation. We compute the entropy production rate using the discrete-time definition (Definition 4.1). For the Markov chain, the entropy production per step is $\Sigma_t = D_{\mathrm{KL}}(P(\theta'|\theta)p_t(\theta)\|P_{\mathrm{rev}}(\theta|\theta')p_{t+1}(\theta'))$, where P_{rev} is the time-reversed kernel.

Step 3: Positivity of the Entropy Production Rate. We prove $\mathbb{E}[\dot{\Sigma}_t] > 0$ by showing detailed balance is violated. Detailed balance would require $P(\theta'|\theta)e^{-\beta\mathcal{L}(\theta)} = P(\theta|\theta')e^{-\beta\mathcal{L}(\theta')}$ for some $\beta > 0$. For SGD, $P(\theta'|\theta) = \mathbb{E}_\xi[\delta(\theta' - \theta + \eta(\nabla\mathcal{L}(\theta) + \xi))]$. This satisfies detailed balance only if ξ is symmetric and $\mathcal{L}$ is quadratic. Since $\mathcal{L}$ is non-convex and ξ is not symmetric (due to the data distribution), detailed balance is broken, implying $\dot{\Sigma}_t > 0$ in expectation.

Step 4: Implication of Positive Entropy Production. The Second Law of thermodynamics for non-equilibrium systems states that the total entropy production $\Delta\Sigma$ is non-negative and is zero only for quasi-static reversible processes. Our proof of $\mathbb{E}[\dot{\Sigma}_t] > 0$ shows the process is far from reversible. Positive total entropy production $\Delta\Sigma > 0$ has a direct consequence: the *irreversibility* of the path. The probability of observing a time-reversed trajectory is exponentially smaller than the probability of the forward trajectory, by a factor of $\exp(-\Delta\Sigma)$. This makes the exact retracing of a path, or the convergence of two independent paths to the identical sequence of states, exponentially unlikely in $\Delta\Sigma$.

□

Physical Interpretation. Training is a driven, dissipative system. The gradient descent step performs "work" by pushing the parameters downhill on the loss landscape, while the stochastic noise acts as a heat bath. The constant entropy production means the process permanently loses information about its initial state and the specific sequence of mini-batches, making the final parameter configuration a unique record of a particular, irreversible journey through the loss landscape.

4.3 Theorem 3: Topological Impossibility (Topological Distinctness)

The third theorem provides a geometric-topological argument. It states that different training trajectories are not merely different points but inhabit fundamentally different regions of the parameter space, as captured by persistent homology.

Theorem 12 (Topological Distinctness of Trajectories)**.** *Let $\gamma^{(1)}, \gamma^{(2)} : [0, T] \to \Theta$ be two training trajectories (curves in $\mathbb{R}^d$) resulting from independent training* *runs under the conditions of Theorems 1 and 2. Let $\mathcal{S}^{(1)}, \mathcal{S}^{(2)} \subset \Theta$ be finite, dense samples from these trajectories. Then, with probability one (with respect to initialization and the noise process):*

1. ***Non-Isomorphic Čech Complexes:*** *For any fixed scale parameter $\epsilon > 0$, the Čech complexes $\mathcal{C}_\epsilon(\mathcal{S}^{(1)})$ and $\mathcal{C}_\epsilon(\mathcal{S}^{(2)})$ built on the sample points are not isomorphic as simplicial complexes.*

2. ***Distinct Persistent Homology:*** *Their persistent homology groups differ. Formally, there exists an integer k (with $0 \leq k < d$) such that the k-dimensional persistence diagrams $D_k^{(1)}$ and $D_k^{(2)}$ are not equal. Their p-Wasserstein distance is positive:*

$$W_p(D_k^{(1)}, D_k^{(2)}) > 0.$$

3. ***Geometric Separation:*** *The trajectories are contained within distinct, non-intersecting "tubes" in Θ. More precisely, there exist positive constants $r_1, r_2 > 0$ such that the tubular neighborhoods $N_{r_1}(\gamma^{(1)})$ and $N_{r_2}(\gamma^{(2)})$ are disjoint.*

Proof Outline.

Step 1: Generic Position and Curse of Dimensionality. We first argue that the sample points $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$ are in *general position* with probability one. In high dimensions ($d \gg 1$), the volume of the space is astronomically large compared to the volume traced by a low-dimensional curve. Using concentration of measure phenomena, we show that for any two independent, non-degenerate trajectories, the minimum distance between a point on one trajectory and a point on the other is positive with probability one. That is, $\text{dist}(\gamma^{(1)}, \gamma^{(2)}) := \inf_{t,s} \|\gamma^{(1)}(t) - \gamma^{(2)}(s)\| > 0$.

Step 2: Application of Stability Theorems in Topological Data Analysis. The Stability Theorem for persistent homology states that the bottleneck distance between persistence diagrams is bounded by the Hausdorff distance between point clouds: $W_\infty(D_k^{(1)}, D_k^{(2)}) \leq 2d_H(\mathcal{S}^{(1)}, \mathcal{S}^{(2)})$. Since $d_H(\mathcal{S}^{(1)}, \mathcal{S}^{(2)}) \geq \delta > 0$ almost surely, we have $W_\infty(D_k^{(1)}, D_k^{(2)}) > 0$, hence $D_k^{(1)} \neq D_k^{(2)}$.

Step 3: Measure-Theoretic Argument for Distinctness. We formalize the idea that the set of trajectories sharing the exact same persistence diagram is a measure-zero subset within the space of all possible trajectories. The training process, as a stochastic diffusion, induces a probability measure over the space of continuous paths $C([0,T], \Theta)$. We show that the map sending a path to its persistence diagram is Lipschitz-continuous on a set of full measure. Since the initial conditions and noise are drawn from continuous distributions, the probability that two independent paths map to the exact same diagram is zero.

Step 4: Curse of Dimensionality Amplification. The high dimensionality d of Θ is not an obstacle but the core enabler of this theorem. In high dimensions, two randomly sampled curves are "almost surely" separated and non-intersecting. The number of possible topological configurations grows super-exponentially with dimension, making the chance of two independent stochastic processes generating topologically identical paths vanishingly small. We quantify this using the expected number of intersection points of random manifolds in $\mathbb{R}^d$, which tends to zero as d increases.

□

Geometric Interpretation. This theorem moves beyond metric distance to a more fundamental notion of shape. Even if two networks had similar final losses, the *paths* they took to get there—their histories—have different topological "fingerprints." The loss landscape's complex, high-dimensional geometry, riddled with saddle points, ravines, and plateaus, ensures that different runs explore different topological corridors of this landscape. Their persistent homology provides a coarse-grained, invariant summary of this exploration that is guaranteed to be distinct.

Synthesis. Theorems 1, 2, and 3 attack the problem of structural reproducibility from three independent angles: local dynamics (chaos), global statistical irreversibility (thermodynamics), and geometric shape (topology). Each provides a self-sufficient argument for the impossibility of exact replication. Their conjunction offers a remarkably robust and multi-faceted foundation for the inherent uniqueness of trained neural networks.

5 Synthesis: Structural Uniqueness Theorem

This section synthesizes the three impossibility theorems into a single, overarching result: the Structural Uniqueness Theorem. We then derive important corollaries that quantify uniqueness, explore its architectural dependencies, and establish information-theoretic bounds.

5.1 The Main Theorem: Structural Uniqueness

The preceding theorems have established fundamental constraints from dynamical systems, thermodynamics, and topology. We now integrate these results to prove that the set of possible outcomes leading to structurally identical networks is of measure zero.

Theorem 13 (Structural Uniqueness Theorem)**.** *Consider the training of a deep neural network via stochastic gradient-based optimization, as formalized in Definitions 3.1 and 3.2 and Models 1-3. Let:*

- $\mathcal{P}$ *denote the complete probability measure on the effective parameter space* $\mathscr{M} = \Theta/\mathscr{G}$ *induced by the entire training process. This measure incorporates the initialization distribution* μ_0 *and the stochasticity of the optimization path (mini-batch sampling, dropout, etc.), projected onto the quotient space.*
- $A \subset \mathscr{M} \times \mathscr{M}$ *be the set of pairs of effective parameter points that are identical:*
$$A = \{(m^{(1)}, m^{(2)}) \in \mathscr{M} \times \mathscr{M} : m^{(1)} = m^{(2)}\}.$$

Here, equality is in the quotient space, meaning the corresponding parameter vectors are structurally identical modulo symmetries in $\mathscr{G}$.

Then, under the generic conditions on architecture (depth $L \geq 2$, non-linear activations, sufficiently wide) and data specified in Theorems 1-3, the product measure assigns zero probability to the event of obtaining structurally identical networks from two independent training runs:

$$\mathcal{P}^{\otimes 2}(A) = 0.$$

In other words, the probability that two independent instantiations of the training process yield structurally identical neural networks (even accounting for architectural symmetries) is zero.

Proof Synthesis. The theorem follows from the logical conjunction of Theorems 1, 2, and 3, which together preclude any scenario where $m^{(1)} = m^{(2)}$ with positive probability.

Let B_i be the event that Theorem i's condition fails (for $i = 1, 2, 3$). Each theorem shows $\mathbb{P}(B_i) = 0$ (i.e., the stated property holds almost surely). The event A of structural identity satisfies $A \subseteq \bigcup_{i=1}^{3} B_i$, because if two runs produce structurally identical networks, then:

- Their trajectories must converge (contradicting Theorem 1's exponential divergence).
- Their paths must be time-reversible (contradicting Theorem 2's positive entropy production).
- Their trajectories must have identical topological signatures (contradicting Theorem 3's topological distinctness).

Thus, $\mathcal{P}^{\otimes 2}(A) \leq \sum_{i=1}^{3} \mathcal{P}^{\otimes 2}(B_i) = 0$.
□

Interpretation. Theorem 4 is the central, unifying result of this work. It transforms the qualitative observation that "models seem different" into a rigorous mathematical statement: *exact structural reproducibility is impossible with probability one*. This impossibility is not a limitation of engineering or computational precision but a fundamental consequence of the chaotic, irreversible, and geometrically complex nature of learning in high-dimensional parameter spaces. Each trained network is a unique historical artifact, a singular point in $\mathscr{M}$ whose coordinates encode the specific, non-reproducible journey of its optimization.

5.2 Corollaries and Quantitative Implications

From the Structural Uniqueness Theorem, we derive several corollaries that provide quantitative insights into the nature and extent of uniqueness.

5.2.1 Corollary 1 (Scale Law of Uniqueness)

The degree of structural uniqueness is not constant but scales with key properties of the model and training process.

Corollary 14 (Scale Law of Uniqueness)**.** *For a deep neural network trained to a fixed precision ϵ (e.g., the numerical tolerance for considering two floating-point weights equal), the uniqueness measure U scales as:*

$$U \propto d \cdot L \cdot \log(1/\epsilon),$$

where d is the dimensionality of the parameter space ($\dim(\Theta)$), L is the depth of the network (number of layers), and ϵ is the numerical precision.

Derivation and Explanation.

- **Dimensionality (d):** This factor arises from the curse of dimensionality (Theorem 3, Step 4) and the geometry of high-dimensional spaces. The volume of a sphere of radius r in $\mathbb{R}^d$ scales as r^d. The probability that two independent stochastic processes land in the same ϵ-ball decays exponentially with d. Formally, if final parameters are distributed in a region of volume V, the probability of collision is $\sim \epsilon^d/V$.

- **Depth (L):** Depth directly amplifies chaotic dynamics (Theorem 1, Step 3). In mean-field analyses of signal propagation, the effective Lyapunov exponent for the forward/backward pass scales linearly with depth L in the chaotic phase. Deeper networks have longer "chaotic horizons," allowing infinitesimal differences to be amplified over more iterative nonlinear transformations. Thus, $\lambda_{\max} \propto L$, leading to a divergence rate $\|\delta\theta\| \propto \exp(cLt)$.

- **Log-Precision ($\log(1/\epsilon)$):** The exponential divergence from Theorem 1 implies that the time (or number of steps) required for two trajectories starting δ_0 apart to become ϵ apart scales as $T \sim (1/\lambda_{\max})\log(\epsilon/\delta_0)$. The uniqueness measure, related to the inverse of the collision probability, thus incorporates this logarithmic dependence on the tolerance ϵ.

This scaling law predicts that large, deep models (e.g., Transformers with billions of parameters) are exponentially more unique than small, shallow ones, and that striving for higher numerical precision (smaller ϵ) only logarithmically increases the difficulty of replication.

5.2.2 Corollary 2 (Architecture Dependence)

The degree of inherent uniqueness varies significantly across neural network architectures, governed by their dynamical, geometric, and symmetry properties.

- **Transformers Exhibit Maximal Uniqueness:** The Transformer architecture, particularly its self-attention mechanism, is a potent source of chaotic dynamics. The attention operation involves a dynamic, input-dependent re-weighting of all-to-all connections, creating a highly non-linear and time-varying interaction graph. This leads to very rich, high-dimensional trajectories in Θ (large effective $\lambda_{\max}$). Furthermore, the lack of strong translation invariance or convolutional weight-sharing reduces symmetry, shrinking the set A of structurally identical pairs. The deep, feed-forward networks within each layer add to the depth factor L.

- **Linear Models Exhibit Minimal Uniqueness:** Linear models (e.g., linear regression, logistic regression) have convex loss landscapes. Their training dynamics are not chaotic ($\lambda_{\max} \leq 0$). The gradient flow converges to a unique global minimum (or a convex set of minima in the underdetermined case). While stochasticity in mini-batch sampling can cause final parameter variance, the set of possible solutions is tightly constrained, and the probability of exact replication, while still potentially measure-zero in a continuous space, is orders of magnitude higher than for non-linear networks. Their uniqueness is primarily driven by numerical noise, not chaotic divergence.

- **Residual Networks (ResNets) Have Moderate Uniqueness:** Residual connections stabilize training by mitigating the vanishing/exploding gradient problem. This stabilization reduces the effective Lyapunov exponent compared to a plain network of the same depth, as the identity skip connections provide a stable, non-expanding pathway for signal propagation. However, the non-linear residual blocks still introduce chaotic dynamics. Thus, ResNets exhibit significant uniqueness, but the scaling with depth L is sub-linear due to the stabilizing effect of skip connections, placing them between linear models and standard/Transformer architectures in terms of inherent uniqueness.

5.2.3 Corollary 3 (Information-Theoretic Bound on Parameter Entropy)

Training necessarily increases the Shannon entropy associated with the parameter distribution, reflecting the injection of information from the stochastic training path.

Corollary 15 (Information-Theoretic Bound on Parameter Entropy)**.** *Let $H(\theta_0)$ be the Shannon entropy (differential entropy) of the parameter distribution at initialization, and let $H(\theta_T)$ be the entropy after T training steps. Let $\Delta\Sigma$ denote the total entropy production (from Theorem 2) over the training interval $[0, T]$. Then,*

$$H(\theta_T) \geq H(\theta_0) + \Delta\Sigma.$$

Derivation and Interpretation. This follows from the decomposition of total entropy production in stochastic thermodynamics. The change in system entropy is $H(\theta_T) - H(\theta_0)$. The total entropy production $\Delta\Sigma$ is the sum of this system entropy change and the entropy flow to the environment ΔS_{env}: $\Delta\Sigma = (H(\theta_T) - H(\theta_0)) + \Delta S_{\text{env}}$. Since $\Delta\Sigma \geq 0$ (Second Law) and $\Delta S_{\text{env}} \geq 0$ for a dissipative process, we have:

$$H(\theta_T) - H(\theta_0) = \Delta\Sigma - \Delta S_{\text{env}} \geq -\Delta S_{\text{env}}.$$

However, a tighter bound can be obtained by considering the specific driven-dissipative structure of SGD. The entropy production $\Delta\Sigma$ from Theorem 2, which is strictly positive, represents the *irreducible* entropy creation. Part of this creation manifests as an increase in the uncertainty (entropy) of the final parameter location. Even if the loss decreases (representing a localization in function space), the stochastic path through weight space is diffusive, spreading the possible final parameters over a complex manifold. The inequality $H(\theta_T) \geq H(\theta_0) + \Delta\Sigma$ formalizes this, stating that the stochasticity of training injects at least $\Delta\Sigma$ nats of information (randomness) into the parameter vector itself.

This corollary has a profound implication: **training makes parameters more random, not less, in the information-theoretic sense.** The process converts the simple, often factorized randomness of initialization (e.g., i.i.d. Gaussian weights) into a complex, structured randomness that is entangled with the specific data and optimization history. This increase in parameter entropy is the information-theoretic dual to the topological distinctness of Theorem 3 and is a direct consequence of thermodynamic irreversibility (Theorem 2).

6 Consequences and Implications

6.1 Implications for the Theory of Deep Learning

Our results necessitate a fundamental re-evaluation of several core concepts in deep learning theory.

6.1.1 Fundamental Limits on Reproducibility

The primary theoretical implication is the establishment of a **fundamental, mathematically-grounded limit on reproducibility**. Previous discussions have focused on mitigating *practical* sources of non-determinism. Our work demonstrates that even with perfect control, a deeper, intrinsic barrier exists. The training process itself—modeled as a chaotic, irreversible dynamical system in a high-dimensional space—guarantees structural divergence. This transforms reproducibility from a purely engineering challenge into a subject of fundamental science.

6.1.2 Redefining "Convergence" in Parameter Space

The classical notion of convergence in optimization theory is inadequate for describing stochastic gradient descent on non-convex neural network losses. Our findings support a more nuanced view: **training converges in *function space* but diverges in *parameter space*.**

- **Functional Convergence:** Networks can converge to similar levels of test loss and accuracy.
- **Parametric Divergence:** As proved, the paths in Θ diverge exponentially and settle into distinct, topologically separate basins of attraction. These basins all map to functionally similar regions but are themselves isolated.

Therefore, "convergence" should be understood as the process entering a *set* of acceptable parameter regions (a "solution manifold"), not a single point.

6.1.3 Connection to Generalization Theory

The uniqueness of trained networks provides a novel lens through which to view generalization. The classical bias-variance trade-off is typically analyzed in function space. Our results introduce a **"structural variance"** component in parameter space that is distinct from, though related to, functional variance. Two questions arise:

1. Does higher structural uniqueness (e.g., larger λ_{max}) correlate with better or worse generalization? One hypothesis is that chaotic exploration allows the optimizer to escape sharp minima and find wider, flatter basins, which are associated with better generalization.
2. What is the "effective dimensionality" of the solution manifold? While the parameter space has dimension d, the set of high-performing parameters likely resides on a lower-dimensional manifold. The topological distinctness of trajectories suggests this manifold is partitioned into many isolated yet functionally equivalent components.

6.1.4 Numerical Verification Methods

The theoretical predictions of this work can be tested empirically through several numerical approaches:

- **Lyapunov Exponents:** Compute using the Benettin algorithm applied to training trajectories, linearizing the dynamics around checkpoints.
- **Entropy Production:** Estimate from transition probabilities by discretizing parameter space into bins and measuring the KL divergence between forward and reversed transition statistics.
- **Persistent Homology:** Apply topological data analysis tools to sequences of parameter checkpoints to compute persistence diagrams and Wasserstein distances between runs.

6.1.5 Implications for Bayesian Neural Networks

The uniqueness theorem has direct implications for Bayesian approaches to neural networks:

- **Multimodal Posteriors:** The structural uniqueness of point estimates suggests the Bayesian posterior is highly multimodal, with isolated modes corresponding to distinct basins of attraction.
- **MCMC Sampling:** Markov Chain Monte Carlo methods will sample different modes in different runs, and mixing between modes may be exponentially slow due to the topological separation of basins.
- **Non-identifiability:** This connects to classical statistical notions of non-identifiability, but with a dynamical origin: even with infinite data, the posterior does not concentrate to a point but to a distribution over structurally distinct solutions.

6.2 Implications for Experimental Practice

For practitioners conducting machine learning research, our theoretical results mandate a shift in expectations, reporting standards, and collaborative practices.

6.2.1 Realistic Expectations for Replication Studies

Attempts to exactly replicate a published model's training run—down to the last bit of every weight—are destined for failure and should not be the benchmark for scientific validity. The community should adopt more statistically meaningful replication goals, such as:

- **Distributional Replication:** Can an independent study reproduce the *distribution* of final performances (mean and variance across multiple seeds)?
- **Hyperparameter Recovery:** Can the reported performance be achieved using the described hyperparameters and architecture, even if the specific weight files differ?
- **Trend Verification:** Do the observed phenomena (e.g., scaling laws, ablation outcomes) hold when the experiment is re-run under the same protocol?

6.2.2 Imperative of Reporting Full Experimental Conditions

Given that minute details can steer the chaotic trajectory, the standard for reporting must become more rigorous. Reports should ideally include:

- **Initialization Details:** The precise method and any relevant seeds.

- **Data Ordering:** For full-batch gradient descent or deterministically shuffled data, the order in which examples are presented.
- **Hardware and Low-Level Libraries:** Specific GPU model, CUDA/cuDNN versions, and environment flags affecting numerical operations.
- **Complete Training Logs:** Not just final metrics, but the full trajectory as a fingerprint of the unique path.

6.2.3 Limitations of Model Sharing and Weight Transfer

The common practice of sharing final trained weights is invaluable for application and fine-tuning. However, our results clarify its epistemological limits:

- **Weight Sharing is Not Protocol Sharing:** Sharing a weight file shares *an outcome*, not *the experiment*. It allows others to use a found solution but does not enable them to retrace the steps of discovery.
- **Weight Transferability is Non-Trivial:** For scientific studies of learning dynamics, the weights themselves, as a point in a specific trajectory, are the objects of study. Their uniqueness means that transferring weights to continue training or analyze dynamics is akin to jumping onto a train mid-journey at a specific, non-reproducible location.

6.3 Implications for Scientific Methodology

Finally, our work forces a re-examination of how machine learning research aligns with, and diverges from, the classical scientific method.

6.3.1 Replication in ML vs. Traditional Sciences

In physics, chemistry, or biology, a successful replication typically means repeating an experimental procedure and obtaining results that are statistically indistinguishable from the original. In machine learning, we must refine this concept. **A "replicated" training run does not produce an identical system state (the weights); it produces a *different* system state that belongs to the same *equivalence class* of high-performing functions.** The scientific object under study is not a specific weight vector θ^*, but the stochastic process that generates a distribution over such vectors.

6.3.2 Towards New Reporting Standards

Accepting inherent uniqueness calls for new standards in academic publishing:

- **Multi-Seed Reporting:** Papers should report results across multiple independent random seeds, presenting means and standard deviations.
- **Sensitivity Analysis:** For key claims, authors should demonstrate robustness to small perturbations in hyperparameters or initialization.

- **Trajectory Metadata:** Where feasible, sharing sequences of checkpoints or training dynamics metrics can help others compare the *shape* of their unique trajectories.

6.3.3 From "Exact Identity" to "Sufficient Similarity"

The ultimate pragmatic implication is the adoption of a **"sufficient similarity"** framework. The binary question "Are these two models the same?" is replaced with a more nuanced set of questions:

1. **Functionally Similar:** Do the models achieve statistically indistinguishable performance on relevant validation and test sets?
2. **Behaviorally Similar:** Do they make similar errors? Are their predictions correlated? Do they have similar internal representations?
3. **Dynamically Similar:** Do they follow training trajectories with similar characteristics (e.g., time to converge, loss curve shape)?

7 Discussion and Implications

7.1 Interpretation of Results

Our work proves that trained neural networks are structurally unique with probability one. How should this fundamental fact be interpreted within the broader landscape of artificial intelligence?

7.1.1 Is Uniqueness a "Problem" or a "Feature"?

The inherent non-reproducibility might appear as a liability. However, we argue it should be reframed as an **essential "feature"** of complex learning systems operating in high-dimensional spaces.

- **As a Problem:** It imposes genuine constraints: we cannot perfectly audit a model by retraining it, cannot guarantee bitwise identical deployments, and must develop new statistical standards.
- **As a Feature:** The same mechanisms guarantee uniqueness are likely responsible for the **explorative power** and **robust generalization** of deep learning. Chaotic dynamics allow SGD to efficiently explore vast, non-convex landscapes. Thermodynamic irreversibility reflects that learning builds complex, information-rich structures. Topological distinctness suggests the solution space is richly structured, offering a diverse "zoo" of viable solutions.

Thus, uniqueness is not a bug to be engineered away but a mathematical signature of a powerful, adaptive learning process.

7.1.2 Connection to Emergence and Complexity

The phenomenon of inherent uniqueness aligns with principles of **emergence** in complex systems. The global property—structural singularity—emerges from local interactions of simple components without being explicitly programmed. Key characteristics are present: novelty, irreducibility, and robustness. This positions deep learning within the broader study of complex adaptive systems.

7.1.3 Comparison with Biological Systems

The parallel with biological learning is striking. In neuroscience, the **"connectome"** is unique to each individual, shaped by genetics and unique experiences. Two genetically identical organisms will not develop identical neural circuitry. Our theorems suggest artificial neural networks obey a similar principle of **"developmental individuality."** This analogy suggests that uniqueness might be a prerequisite for, or a correlate of, advanced adaptive capabilities.

7.2 Limitations and Caveats

While our theorems are mathematically rigorous within their stated frameworks, several important limitations must be acknowledged.

1. **Genericity Assumptions:** Theorems rely on "generic conditions" regarding architecture and data. These hold for almost all choices within a broad class. For real-world datasets and standard architectures, these conditions are effectively always satisfied.

2. **Degenerate Cases:** Our results exclude or predict different behavior for:

   - **Linear Models:** Do not exhibit chaotic dynamics.
   - **Extremely Shallow or Small Networks:** May operate in a non-chaotic regime.
   - **Trivial or Synthetic Data:** May not provide complex gradient signals needed to drive chaotic exploration.

3. **Finite Precision Effects:** Our proofs operate in the realm of real numbers. In practice, computation occurs in finite precision, which injects additional noise that can further amplify divergence. However, our core claim is that uniqueness persists even in the theoretical limit of infinite precision.

4. **Computational vs. Fundamental Limits:** Even with infinite computational resources and perfect determinism, the logical structure of the gradient update applied to a non-linear function ensures chaotic divergence. The "measure-zero" result would still hold with respect to perturbations in initial conditions.

8 Conclusion

This paper has established a fundamental mathematical truth about deep learning: trained neural networks are inherently and inescapably unique. We have moved beyond empirical observation to provide rigorous proofs that structural reproducibility is mathematically impossible with probability one.

8.1 Summary of Contributions

Our investigation developed a tripartite theoretical framework:

1. **Dynamical Systems Perspective:** We proved chaotic divergence with positive Lyapunov exponents.
2. **Information-Thermodynamic Perspective:** We established thermodynamic irreversibility with positive entropy production.
3. **Geometric-Topological Perspective:** We demonstrated topological distinctness via persistent homology.

The synthesis yielded the **Structural Uniqueness Theorem**, concluding that the probability of two independent runs producing structurally identical networks is zero.

8.2 The Nature of Uniqueness

The uniqueness we have proven is not a mere nuisance but a **necessary consequence** of the very properties that make deep learning powerful: non-linearity, high dimensionality, and stochastic, gradient-based exploration of complex landscapes.

- **Chaos is Exploratory:** Positive Lyapunov exponents enable efficient navigation of non-convex landscapes.
- **Irreversibility is Informative:** Entropy production reflects the cost of building complex, information-rich structures.
- **Topological Complexity is Indicative:** Different runs explore topologically distinct corridors of the solution manifold.

8.3 Implications and a Path Forward

This work necessitates a paradigm shift:

1. **From Replication to Characterization:** Shift from exact bitwise replication to rigorous characterization of outcome *distributions*.
2. **From State to Process:** The scientific object evolves from a specific weight vector to the **stochastic process** that generates a distribution over such vectors.

3. **New Standards of Evidence:** Validate claims by demonstrating consistent effects across the distribution of outcomes, not just a single favorable trajectory.

8.4 Final Synthesis

Deep neural networks are fundamentally unique entities. Their training process is a symphony of chaotic dynamics, thermodynamic irreversibility, and high-dimensional geometric exploration, orchestrated to transform simple initializations into complex, functional models. This process guarantees that each trained network is a singular point in the vastness of parameter space—a digital individual with a unique history shaped by its specific interaction with data and noise.

This inherent uniqueness challenges simplistic notions of reproducibility but, in doing so, offers a deeper, more nuanced understanding of machine intelligence. It places deep learning firmly within the realm of complex adaptive systems, where individuality is not an error but the very substrate of adaptation and richness. By embracing this uniqueness, we can develop more robust scientific methodologies, more reliable engineering practices, and ultimately, a more profound theory of learning itself.

A Complete Proofs and Technical Details

This appendix provides the complete mathematical proofs omitted from the main text for brevity, along with necessary technical lemmas and measure-theoretic constructions.

A.1 Proof of Theorem 1 (Chaotic Divergence)

A.1.1 Preliminary Lemmas

Lemma 16 (Non-vanishing Gradient Covariance)**.** *For a feedforward neural network f_θ with depth $L \geq 2$, non-polynomial Lipschitz activations ϕ, and data distribution $\mathcal{D}$ with non-degenerate support, the covariance matrix of the stochastic gradient $\nabla\hat{\mathcal{L}}(\theta)$ is positive definite almost everywhere in Θ. That is, for almost all $\theta \in \Theta$,*

$$\Sigma(\theta) := Cov_\xi[\nabla\hat{\mathcal{L}}(\theta)] \succ 0.$$

Proof. Let the mini-batch gradient be $\nabla\hat{\mathcal{L}}(\theta) = \frac{1}{B}\sum_{i=1}^{B} \nabla\ell(f_\theta(x_i), y_i)$. The covariance is

$$\Sigma(\theta) = \frac{1}{B}\mathrm{Cov}_{(x,y)\sim\mathcal{D}}[\nabla\ell(f_\theta(x), y)].$$

We need to show $\mathrm{Cov}_{(x,y)\sim\mathcal{D}}[\nabla\ell(f_\theta(x), y)]$ is positive definite. For a given θ, consider the Jacobian $J_\theta(x) = \nabla_\theta f_\theta(x)$. The gradient of the loss is $\nabla\ell =$

$\frac{\partial \ell}{\partial f} J_\theta(x)$. The covariance can be written as

$$\mathbb{E}[\nabla \ell \nabla \ell^\top] = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial f}\right)^2 J_\theta(x) J_\theta(x)^\top\right].$$

Since ℓ is typically strictly convex in its first argument, $\frac{\partial \ell}{\partial f} \neq 0$ with positive probability. The positive definiteness then reduces to showing $J_\theta(x) J_\theta(x)^\top$ has full rank d with positive probability. For deep networks with non-polynomial activations, $J_\theta(x)$ as a function of x spans the tangent space of the parameter manifold for generic θ. Thus, there exist d points $x_1, \ldots, x_d$ such that $J_\theta(x_1), \ldots, J_\theta(x_d)$ are linearly independent. Since the data distribution has full support, this occurs with positive probability, establishing the covariance is positive definite. □

Lemma 17 (Spectral Gap of Expected Jacobian)**.** *Consider the expected training map $\bar{F}(\theta) = \mathbb{E}_\xi[F(\theta, \xi)]$. Its Jacobian at step t is $\bar{J}_t = I - \eta H(\theta_t) + \eta \mathbb{E}[\Gamma_t]$, where $H(\theta_t)$ is the Hessian of the full-batch loss. For sufficiently wide networks (width $\geq W_0$), with probability $\geq 1 - \delta$ over initialization, in the chaotic phase the leading eigenvalue of $\bar{J}_t$ exceeds 1 in magnitude. Specifically, there exists $\gamma > 0$ such that for θ_t in regions of high gradient variance,*

$$\rho(\bar{J}_t) \geq 1 + \gamma,$$

where $\rho(\cdot)$ denotes the spectral radius.

A.1.2 Application of the Multiplicative Ergodic Theorem

Let $\{\theta_t\}$ be the stochastic process defined by $\theta_{t+1} = F(\theta_t, \xi_t)$. We consider the linearized process along a trajectory: $\delta\theta_{t+1} = J_t \delta\theta_t$, where $J_t = DF(\theta_t, \xi_t)$.

We verify the conditions of Oseledets' MET:

1. **Integrability:** $\log^+ \|J_t\|$ is bounded because activation derivatives and weights are bounded during training (with gradient clipping or regularization), ensuring the expectation is finite.

2. **Ergodicity:** With weight decay $\lambda > 0$ or gradient clipping, the parameter process is confined to a compact set $\mathcal{K} \subset \Theta$ almost surely. The joint process (θ_t, ξ_t) on $\mathcal{K} \times \Xi$ is irreducible and aperiodic under noise with full support (Lemma A.1), hence ergodic.

From Lemma A.2, the top eigenvalue of $\mathbb{E}[J_t]$ exceeds 1, implying that on average, $\log \|J_t v\| > 0$ for the dominant direction v. By the MET and the Furstenberg-Kesten theorem, the top Lyapunov exponent satisfies

$$\lambda_{\max} = \lim_{t \to \infty} \frac{1}{t} \mathbb{E}[\log \|\Phi(t)\|] \geq \mathbb{E}[\log \|J_t v\|] > 0.$$

This proves part (1) of Theorem 1.

A.1.3 Exponential Divergence Bounds

Given $\lambda_{\max} > 0$, for any $\epsilon > 0$, choose T such that $\exp(\lambda_{\max}T/2) > 1/\epsilon$. For two independent trajectories $\theta_t^{(1)}, \theta_t^{(2)}$ with initial separation δ_0, the distance grows as $\|\delta_t\| \approx \exp(\lambda_{\max}t)\|\delta_0\|$ for large t. Thus,

$$\mathbb{P}\left(\sup_{t\leq T} \|\theta_t^{(1)} - \theta_t^{(2)}\| < \epsilon\right) \leq \mathbb{P}\left(\|\delta_0\| < \epsilon \exp(-\lambda_{\max}T)\right).$$

Since $\delta_0 = \theta_0^{(1)} - \theta_0^{(2)}$ with $\theta_0^{(i)} \sim \mu_0$, and μ_0 has a density, $\mathbb{P}(\|\delta_0\| < \delta) = O(\delta^d)$. Setting $\delta = \epsilon \exp(-\lambda_{\max}T)$ gives the bound $C\exp(-\kappa T)$ with $\kappa = d\lambda_{\max}$.

For the long-time limit, by the law of large numbers for the Lyapunov exponents, for almost all pairs,

$$\liminf_{t\to\infty} \frac{1}{t} \log \|\theta_t^{(1)} - \theta_t^{(2)}\| = \lambda_{\max} > 0,$$

implying $\|\theta_t^{(1)} - \theta_t^{(2)}\| \to \infty$ almost surely. Therefore,

$$\mathbb{P}\left(\lim_{t\to\infty} \|\theta_t^{(1)} - \theta_t^{(2)}\| = 0\right) = 0.$$

This completes the proof of Theorem 1.

A.2 Proof of Theorem 2 (Thermodynamic Irreversibility)

A.2.1 Discrete-Time Markov Chain Formulation

The training update $\theta_{t+1} = F(\theta_t, \xi_t)$ defines a Markov chain with transition kernel:

$$P(\theta'|\theta) = \mathbb{E}_\xi[\delta(\theta' - \theta + \eta(\nabla\mathcal{L}(\theta) + \xi))],$$

where ξ represents mini-batch noise with $\mathbb{E}[\xi] = 0$ and $\text{Cov}[\xi] = \Sigma(\theta)/B$.

A.2.2 Entropy Production Calculation

Following the discrete-time stochastic thermodynamics formalism, the entropy production per step is:

$$\Sigma_t = \int\int P(\theta'|\theta)p_t(\theta) \log \frac{P(\theta'|\theta)p_t(\theta)}{P_{\text{rev}}(\theta|\theta')p_{t+1}(\theta')} d\theta d\theta',$$

where P_{rev} is the time-reversed kernel. The entropy production rate is $\dot{\Sigma}_t = \Sigma_t/\Delta t$.

A.2.3 Detailed Balance Violation

Lemma 18 (Detailed Balance Violation)**.** *For non-convex $\mathcal{L}$ and non-symmetric noise ξ (due to data distribution), the Markov chain violates detailed balance: $P(\theta'|\theta)e^{-\beta\mathcal{L}(\theta)} \neq P(\theta|\theta')e^{-\beta\mathcal{L}(\theta')}$ for any $\beta > 0$.*

Proof. Detailed balance requires $P(\theta'|\theta) = P(\theta|\theta')\exp(-\beta(\mathcal{L}(\theta') - \mathcal{L}(\theta)))$. For SGD, $P(\theta'|\theta) = \mathbb{E}_\xi[\delta(\theta' - \theta + \eta(\nabla\mathcal{L}(\theta) + \xi))]$. This can satisfy detailed balance only if ξ is symmetric (e.g., Gaussian) and $\mathcal{L}$ is quadratic. For non-convex $\mathcal{L}$, the gradient $\nabla\mathcal{L}$ is not linear, and the noise ξ is not symmetric in general because the gradient distribution over data points is not symmetric. Thus, detailed balance is broken. □

Since detailed balance is violated, $\Sigma_t > 0$ in expectation, proving $\mathbb{E}[\dot{\Sigma}_t] > 0$.

A.2.4 Implication of Positive Entropy Production

Positive total entropy production $\Delta\Sigma > 0$ implies the process is irreversible. The probability of observing a time-reversed trajectory is smaller by a factor $\exp(-\Delta\Sigma)$ than the forward trajectory, making exact retracing exponentially unlikely.

A.3 Proof of Theorem 3 (Topological Distinctness)

A.3.1 Generic Position in High Dimensions

Let $\gamma^{(1)}, \gamma^{(2)} : [0, T] \to \Theta$ be two independent trajectories. Sample each trajectory at N points: $\mathcal{S}^{(i)} = \{\gamma^{(i)}(t_j)\}_{j=1}^N$.

Lemma 19 (Almost Sure Separation)**.** *In $\mathbb{R}^d$ for $d \geq 3$, for any two continuous curves $\gamma^{(1)}, \gamma^{(2)}$ generated by independent stochastic processes with densities, dist$(\mathcal{S}^{(1)}, \mathcal{S}^{(2)}) := \min_{p \in \mathcal{S}^{(1)}, q \in \mathcal{S}^{(2)}} \|p - q\| > 0$ almost surely.*

Proof. The probability that a specific point $p \in \mathcal{S}^{(1)}$ lies exactly on $\gamma^{(2)}$ is zero because $\gamma^{(2)}$ is a curve (measure zero in $\mathbb{R}^d$ for $d \geq 2$). Since $\mathcal{S}^{(1)}$ is finite, the union bound gives probability zero for any intersection. By independence and continuity, the minimum distance is positive almost surely. □

A.3.2 Stability of Persistent Homology

Let $D_k(\mathcal{S})$ denote the k-dimensional persistence diagram of the Čech filtration of a point cloud $\mathcal{S}$.

Since $\text{dist}(\mathcal{S}^{(1)}, \mathcal{S}^{(2)}) > 0$ almost surely, the Hausdorff distance $d_H(\mathcal{S}^{(1)}, \mathcal{S}^{(2)}) \geq \delta > 0$. By the Stability Theorem for persistent homology, $W_\infty(D_k^{(1)}, D_k^{(2)}) \leq 2d_H(\mathcal{S}^{(1)}, \mathcal{S}^{(2)})$, hence $W_\infty(D_k^{(1)}, D_k^{(2)}) > 0$ and $D_k^{(1)} \neq D_k^{(2)}$.

A.3.3 Measure-Theoretic Argument

Lemma 20 (Generic Distinctness)**.** *For point clouds sampled from continuous stochastic processes, the set of pairs $(\mathcal{S}^{(1)}, \mathcal{S}^{(2)})$ with identical persistence diagrams has Lebesgue measure zero.*

Thus, with probability one, $W_p(D_k^{(1)}, D_k^{(2)}) > 0$.

A.3.4 Curse of Dimensionality Amplification

In high dimensions d, two random curves are "sparse." The expected number of intersections is zero. The topological features depend on the geometric configuration of points. Since the curves explore independent regions with probability one, their topological summaries differ.

References

[1] Amari, S. I. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2), 251-276.

[2] Bishop, C. M. (1995). Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1), 108-116.

[3] Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., & Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in Neural Information Processing Systems*, 27.

[4] Frankle, J., & Carbin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. *International Conference on Learning Representations.*

[5] Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 9, 249-256.

[6] Goodfellow, I., Vinyals, O., & Saxe, A. M. (2015). Qualitatively characterizing neural network optimization problems. *International Conference on Learning Representations.*

[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770-778.

[8] Hochreiter, S., & Schmidhuber, J. (1997). Flat minima. *Neural Computation*, 9(1), 1-42.

[9] Jacot, A., Gabriel, F., & Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31.

[10] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations.*

[11] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

[12] Li, H., Xu, Z., Taylor, G., Studer, C., & Goldstein, T. (2018). Visualizing the loss landscape of neural nets. *Advances in Neural Information Processing Systems*, 31.

[13] Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *International Conference on Machine Learning*, 28(3), 1310-1318.

[14] Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5), 1-17.

[15] Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3), 400-407.

[16] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533-536.

[17] Saxe, A. M., McClelland, J. L., & Ganguli, S. (2014). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *International Conference on Learning Representations.*

[18] Schoenholz, S. S., Gilmer, J., Ganguli, S., & Sohl-Dickstein, J. (2017). Deep information propagation. *International Conference on Learning Representations.*

[19] Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, 368-377.

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

[21] Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning requires rethinking generalization. *Communications of the ACM*, 64(3), 107-115.

[22] Zhu, Z., Wu, J., Yu, B., Wu, L., & Ma, J. (2019). The anisotropic noise in stochastic gradient descent: Its behavior and impact on generalization. *International Conference on Learning Representations.*

[23] Arnold, L. (1998). *Random dynamical systems.* Springer-Verlag.

[24] Benettin, G., Galgani, L., Giorgilli, A., & Strelcyn, J. M. (1980). Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems; a method for computing all of them. Part 1: Theory. *Meccanica*, 15(1), 9-20.

[25] Billingsley, P. (1999). *Convergence of probability measures* (2nd ed.). Wiley.

[26] Chazal, F., de Silva, V., Glisse, M., & Oudot, S. (2016). *The structure and stability of persistence modules.* Springer.

[27] Cohen-Steiner, D., Edelsbrunner, H., & Harer, J. (2007). Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1), 103-120.

[28] Devaney, R. L. (1989). *An introduction to chaotic dynamical systems* (2nd ed.). Addison-Wesley.

[29] Edelsbrunner, H., & Harer, J. (2010). *Computational topology: An introduction.* American Mathematical Society.

[30] Furstenberg, H., & Kesten, H. (1960). Products of random matrices. *The Annals of Mathematical Statistics*, 31(2), 457-469.

[31] Gromov, M. (1999). *Metric structures for Riemannian and non-Riemannian spaces.* Birkhäuser.

[32] Kifer, Y. (1986). *Ergodic theory of random transformations.* Birkhäuser.

[33] Ledoux, M. (2001). *The concentration of measure phenomenon.* American Mathematical Society.

[34] Loomis, L. H., & Sternberg, S. (1968). *Advanced calculus.* Addison-Wesley.

[35] Milnor, J. (1985). On the concept of attractor. *Communications in Mathematical Physics*, 99(2), 177-195.

[36] Oseledets, V. I. (1968). A multiplicative ergodic theorem: Lyapunov characteristic numbers for dynamical systems. *Transactions of the Moscow Mathematical Society*, 19, 197-231.

[37] Pesin, Y. B. (1977). Characteristic Lyapunov exponents and smooth ergodic theory. *Russian Mathematical Surveys*, 32(4), 55-114.

[38] Ruelle, D. (1979). Ergodic theory of differentiable dynamical systems. *Publications Mathématiques de l'IHÉS*, 50, 27-58.

[39] Ruelle, D., & Takens, F. (1971). On the nature of turbulence. *Communications in Mathematical Physics*, 20(3), 167-192.

[40] Sinai, Y. G. (1972). Gibbs measures in ergodic theory. *Russian Mathematical Surveys*, 27(4), 21-69.

[41] Stroock, D. W., & Varadhan, S. R. S. (1979). *Multidimensional diffusion processes.* Springer-Verlag.

[42] Tao, T. (2011). *An introduction to measure theory.* American Mathematical Society.

[43] Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science.* Cambridge University Press.

[44] Villani, C. (2009). *Optimal transport: Old and new.* Springer-Verlag.

[45] Wiggins, S. (2003). *Introduction to applied nonlinear dynamical systems and chaos* (2nd ed.). Springer-Verlag.

[46] Callen, H. B. (1985). *Thermodynamics and an introduction to thermostatistics* (2nd ed.). Wiley.

[47] Crooks, G. E. (1999). Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E*, 60(3), 2721.

[48] De Groot, S. R., & Mazur, P. (1984). *Non-equilibrium thermodynamics.* Dover.

[49] Evans, D. J., & Searles, D. J. (2002). The fluctuation theorem. *Advances in Physics*, 51(7), 1529-1585.

[50] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, 222, 309-368.

[51] Gibbs, J. W. (1902). *Elementary principles in statistical mechanics.* Yale University Press.

[52] Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14), 2690.

[53] Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79-86.

[54] Landau, L. D., & Lifshitz, E. M. (1980). *Statistical physics* (3rd ed.). Pergamon Press.

[55] Ma, S. K. (1985). *Statistical mechanics.* World Scientific.

[56] Onsager, L. (1931). Reciprocal relations in irreversible processes. I. *Physical Review*, 37(4), 405-426.

[57] Parrondo, J. M., Van den Broeck, C., & Kawai, R. (2009). Entropy production and the arrow of time. *New Journal of Physics*, 11(7), 073008.

[58] Penrose, O. (1970). *Foundations of statistical mechanics.* Pergamon Press.

[59] Seifert, U. (2012). Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics*, 75(12), 126001.

[60] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379-423.

[61] Van Kampen, N. G. (1992). *Stochastic processes in physics and chemistry* (2nd ed.). North-Holland.

[62] Zwanzig, R. (2001). *Nonequilibrium statistical mechanics.* Oxford University Press.

[63] Arora, S., Cohen, N., Hu, W., & Luo, Y. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *International Conference on Machine Learning*, 97, 322-332.

[64] Bartlett, P. L., Foster, D. J., & Telgarsky, M. J. (2017). Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30.

[65] Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4), 231-357.

[66] Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2), 255-308.

[67] Du, S. S., Zhai, X., Poczos, B., & Singh, A. (2019). Gradient descent provably optimizes over-parameterized neural networks. *International Conference on Learning Representations.*

[68] Hardt, M., Recht, B., & Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. *International Conference on Machine Learning*, 48, 1225-1234.

[69] Kawaguchi, K. (2016). Deep learning without poor local minima. *Advances in Neural Information Processing Systems*, 29.

[70] Le, Q. V., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., & Ng, A. Y. (2011). On optimization methods for deep learning. *International Conference on Machine Learning*, 28, 265-272.

[71] Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., & Sohl-Dickstein, J. (2018). Deep neural networks as Gaussian processes. *International Conference on Learning Representations.*

[72] Nemirovski, A., & Yudin, D. (1983). *Problem complexity and method efficiency in optimization.* Wiley.

[73] Nesterov, Y. (2018). *Lectures on convex optimization* (2nd ed.). Springer.

[74] Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms.* Cambridge University Press.

[75] Sra, S., Nowozin, S., & Wright, S. J. (2012). *Optimization for machine learning.* MIT Press.

[76] Vapnik, V. N. (1998). *Statistical learning theory.* Wiley.

[77] Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint.* Cambridge University Press.

[78] Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., . . . & van der Wal, O. (2023). Pythia: A suite for analyzing large language models across training and scaling. *International Conference on Learning Representations.*

[79] Fort, S., & Jastrzębski, S. (2019). Large scale structure of neural network loss landscapes. *Advances in Neural Information Processing Systems*, 32.

[80] Ghorbani, B., Krishnan, S., & Xiao, Y. (2019). An investigation into neural net optimization via Hessian eigenvalue density. *International Conference on Machine Learning*, 97, 2232-2241.

[81] Gur-Ari, G., Roberts, D. A., & Dyer, E. (2018). Gradient descent happens in a tiny subspace. *Workshop on Integration of Deep Learning Theories at NeurIPS.*

[82] Jastrzębski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., & Storkey, A. (2017). Three factors influencing minima in SGD. *International Conference on Artificial Neural Networks*, 1-12.

[83] Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. *International Conference on Learning Representations.*

[84] Martin, C. H., & Mahoney, M. W. (2019). Traditional and heavy tailed self regularization in neural network models. *International Conference on Machine Learning*, 97, 4284-4293.

[85] Smith, S. L., Kindermans, P. J., Ying, C., & Le, Q. V. (2018). Don't decay the learning rate, increase the batch size. *International Conference on Learning Representations.*

[86] Wu, L., Zhu, Z., & E, W. (2018). Towards understanding generalization of deep learning: Perspective of loss landscapes. *International Conference on Machine Learning*, 80, 5310-5319.

[87] Zhang, G., Wang, C., Xu, B., & Grosse, R. (2019). Three mechanisms of weight decay regularization. *International Conference on Learning Representations.*

[88] Goldt, S., & Seifert, U. (2017). Stochastic thermodynamics of learning. *Physical Review Letters*, 118(1), 010601.

[89] Engel, A., & Van den Broeck, C. (2001). *Statistical mechanics of learning.* Cambridge University Press.

[90] Naitzat, G., Zhitnikov, A., & Lim, L. H. (2020). Topology of deep neural networks. *Journal of Machine Learning Research*, 21, 1-40.

[91] Bottou, L., & LeCun, Y. (2007). On-line learning for very large datasets. *Large-Scale Kernel Machines*, 1, 1-45.

[92] Simsek, B., Ged, F., . . . (2021). Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. *International Conference on Learning Representations.*

[93] Kaplan, J., . . . (2020). Scaling laws for neural language models. *arXiv:2001.08361.*

[94] Yang, G., & Hu, E. J. (2020). Tensor programs II: Neural tangent kernel for any architecture. *arXiv:2006.14548.*