The Geometry of What Is Learned: Realization Maps, Fiber Structure, and the Semantic Manifold

Flyxion
Independent Researcher

March 2026

Abstract

The parameters of a neural network are not its meaning. This essay develops the geometric consequences of that observation. The central object is the realization map $\Phi : \Theta \to \mathcal{N}$, sending parameter vectors to the functions they compute. Its fibers are the equivalence classes of implementations of the same function, forming smooth submanifolds of Θ whose dimension is determined by the architectural symmetry group $G_{\mathcal{A}}$. The quotient $\mathcal{N} \cong \Theta/G_{\mathcal{A}}$ is the semantic manifold: the space of distinct realizable functions equipped with a canonical Riemannian metric derived from the data distribution. Standard quantities used to analyze learning—flatness, parameter magnitude, gradient norms—are not invariant under $G_{\mathcal{A}}$ and carry no semantic content. The correct geometry is that of $\mathcal{N}$: learning is gradient flow on $(\mathcal{N}, g)$, generalization depends on the semantic dimension $d = \dim \mathcal{N}$ and the connection curvature κ, and compression is lossless if and only if it preserves the fiber. Entanglement—the failure of the horizontal distribution to be integrable, measured by the connection curvature Ω—simultaneously degrades optimization, generalization, and compression capacity. The semantic manifold carries a canonical statistical structure, a universal categorical characterization, a gauge interpretation of redundancy, and a global topology that constrains optimization. The correct unit of analysis is the equivalence class under Φ: meaning is invariant under implementation.

1. Introduction

The parameters of a neural network are not its meaning. This distinction, though obvious when stated, is almost never taken seriously as a mathematical commitment. The prevailing practice treats the parameter vector $\theta \in \Theta$ as the primary object of study: loss functions are defined over Θ, optimization trajectories are traced through Θ, generalization bounds are stated in terms of properties of Θ, and pruning methods operate by removing coordinates of θ. The implicit assumption is that the geometry of Θ is the geometry that matters. It is not.

A neural network with parameters θ computes a function $f_\theta : \mathcal{X} \to \mathcal{Y}$. Two parameter vectors θ and θ' that compute the same function are, from every behavioral and semantic standpoint, identical. Yet the Euclidean geometry on Θ treats them as distinct points, assigns them different gradient norms, and would have an optimizer treat moving from one to the other as meaningful. None of this is meaningful. It is an artifact of the coordinate system, not a property of the model.

The failure runs deeper than a notational inconvenience. The standard interpretation of flat minima as a signature of generalization is not invariant under reparameterization of Θ: a sharp minimum in one parameterization can be made arbitrarily flat by a smooth coordinate change that leaves f_θ unchanged. The same problem infects pruning, where parameters with small magnitude are removed under the assumption that they contribute little to the realized function—but magnitude is coordinate-dependent and can be redistributed arbitrarily within an equivalence class of identical functions.

The resolution begins with a single observation: there is a map

$$\Phi : \Theta \to \mathcal{N}$$

sending each parameter vector to the function it computes, where $\mathcal{N}$ is the space of realizable functions. This map is the fundamental object of study. Its fibers $\Phi^{-1}(f)$ are the sets of all parameter vectors computing the same function. The quotient $\mathcal{N} \cong \Theta/\sim$ is the space where meaning lives.

This essay develops the consequences of taking Φ seriously as a mathematical object across its full geometric, statistical, categorical, and topological dimensions. The argument proceeds through local differential structure, global topology, statistical grounding, categorical universality, variational derivation of learning, stability theory, gauge interpretation of symmetry, and information geometry. The goal throughout is not maximum generality but the minimal structure sufficient to make the core claims precise.

2. System Model

Overview. This section collects all primitive objects into a single formal tuple and introduces the notation used throughout.

Table 1: Principal notation.

Symbol	Meaning
$\Theta \subseteq \mathbb{R}^n$	Parameter space
$\mathcal{X}, \mathcal{Y}$	Input and output spaces
$\mathcal{F}(\mathcal{X}, \mathcal{Y})$	Function space
$\Phi : \Theta \to \mathcal{N}$	Realization map
$\mathcal{N}$	Semantic manifold
$G_{\mathcal{A}}$	Architectural symmetry group
$\Phi^{-1}(f)$	Implementation fiber over f
$V_\theta = \ker D\Phi(\theta)$	Vertical subspace
$H_\theta = V_\theta^\perp$	Horizontal subspace
g_f	Semantic metric on $\mathcal{N}$
G_θ	Pullback metric on Θ
Ω	Connection curvature
$d = \dim \mathcal{N}$	Semantic dimension
$\kappa = \|\Omega\|$	Entanglement measure
$\mathcal{C}(f)$	Policy complexity
$\mathcal{S}(f)$	Policy entropy
$\mathcal{F}(f)$	Semantic free energy

Definition 2.1 (Realization Map)**.** The *realization map* of architecture $\mathcal{A}$ is $\Phi : \Theta \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$, defined by $\Phi(\theta) = f_\theta$. The *semantic space* $\mathcal{N} := \text{image}(\Phi)$ is the set of all realizable functions.

Definition 2.2 (Implementation Fiber)**.** For $f \in \mathcal{N}$, the *implementation fiber* is $\Phi^{-1}(f) = \{\theta \in \Theta \mid \Phi(\theta) = f\}$. The equivalence relation $\theta \sim \theta'$ if and only if $\Phi(\theta) = \Phi(\theta')$ partitions Θ and yields the quotient $\mathcal{N} \cong \Theta/\sim$.

Definition 2.3 (Architectural Symmetry Group)**.** The *symmetry group* $G_{\mathcal{A}}$ is the group of diffeomorphisms $\sigma : \Theta \to \Theta$ satisfying $\Phi \circ \sigma = \Phi$.

Proposition 2.4 (Smooth Quotient)**.** *When $G_{\mathcal{A}}$ acts freely and properly on the regular locus Θ^{reg}, the quotient $\mathcal{N}^{\text{reg}} = \Theta^{\text{reg}}/G_{\mathcal{A}}$ is a smooth manifold of dimension $n - \dim G_{\mathcal{A}}$, and Φ is a smooth principal $G_{\mathcal{A}}$-bundle.*

Definition 2.5 (Horizontal and Vertical Subspaces)**.** At each regular θ, define the *vertical subspace* $V_\theta := \ker D\Phi(\theta)$ and the *horizontal subspace* $H_\theta := V_\theta^\perp$, so that $T_\theta\Theta = H_\theta \oplus V_\theta$.

Remark 2.6 (Geometric levels)**.** Four levels of structure are active simultaneously. Level (A): parameter space $\Theta \subset \mathbb{R}^n$, finite-dimensional and Euclidean. Level (B): function space $\mathcal{F}(\mathcal{X}, \mathcal{Y})$, infinite-dimensional; we do not work here directly. Level (C): graph of Φ, diffeomorphic to Θ. Level (D): semantic space $\mathcal{N} = \text{image}(\Phi)$, the primary object of study. All claims about manifold structure and learning dynamics refer to level (D) unless stated otherwise.

3. Differential Structure of the Realization Map

Overview. All subsequent structure depends on the local behavior of $D\Phi(\theta)$. This section establishes the rank stratification, introduces the singular set, and derives a local normal form making the fiber geometry explicit.

Definition 3.1 (Rank and Regularity)**.** The *rank* of Φ at θ is $\text{rank}(D\Phi(\theta)) := \dim \text{image}(D\Phi(\theta))$. A point θ is *regular* if this rank equals $d = \dim \mathcal{N}$; otherwise it is *singular*. The *singular set* is $\Sigma := \{\theta \in \Theta \mid \text{rank}(D\Phi(\theta)) < d\}$.

Proposition 3.2 (Generic Regularity)**.** *Under mild smoothness assumptions on $\mathcal{A}$, the singular set Σ has measure zero in Θ. This follows from Sard's theorem applied to the smooth map Φ, which for typical architectures is polynomial or piecewise smooth.*

All subsequent statements are understood to hold on the regular locus $\Theta^{\text{reg}} := \Theta \setminus \Sigma$.

Proposition 3.3 (Local Normal Form)**.** *Let $\theta \in \Theta^{\text{reg}}$. Then there exist local coordinates (u, v) on a neighborhood $U \subset \Theta$ with $u \in \mathbb{R}^d$, $v \in \mathbb{R}^{n-d}$, such that $\Phi(u, v) = u$.*

Sketch. Since θ is regular, $D\Phi(\theta)$ has full rank d. By the constant rank theorem, there exist coordinates in which Φ reduces to projection onto the first d coordinates. □

Three corollaries follow immediately.

Corollary 3.4 (Explicit Fiber, Vertical Space, Semantic Dimension)**.** *In the coordinates of Proposition 3.3: the fiber through (u_0, v_0) is $\{(u_0, v) \mid v \in \mathbb{R}^{n-d}\}$; the vertical space is $V_\theta = \text{span}\{\partial/\partial v_i\}$ and the horizontal space is $H_\theta = \text{span}\{\partial/\partial u_j\}$; and the semantic dimension $d = \text{rank}(D\Phi(\theta))$ is constant on Θ^{reg}.*

Proposition 3.5 (Rank Stratification)**.** *The parameter space decomposes into strata $\Theta = \bigsqcup_{k=0}^{d} \Theta_k$ where $\Theta_k := \{\theta \mid \text{rank}(D\Phi(\theta)) = k\}$, each a submanifold. Points in lower-rank strata correspond to degenerate parameterizations: dead neurons, redundant layers, or collapsed representations. The regular locus $\Theta^{\text{reg}} = \Theta_d$ is open and dense.*

The realization map behaves locally as a projection: $\Theta \approx \mathbb{R}^d \times \mathbb{R}^{n-d} \xrightarrow{\Phi} \mathbb{R}^d$. The first factor carries semantic degrees of freedom; the second carries implementation degrees of freedom. All subsequent constructions are global consequences of this local structure. At singular points, the semantic dimension collapses, the fiber dimension increases, and the geometry of $\mathcal{N}$ becomes non-smooth.

4. Fiber Geometry

Overview. Fibers are the geometric expression of implementation redundancy. Their dimension, curvature, and relationship to the symmetry group determine what optimization can do without changing the model.

At a regular θ, the fiber has dimension $n - \operatorname{rank}(D\Phi(\theta))$, which quantifies the irreducible redundancy of the parameterization.

Proposition 4.1 (Fiber-Orbit Coincidence)**.** *If $G_{\mathcal{A}}$ acts freely and transitively on the fibers of Φ at regular points, then $\Phi^{-1}(\Phi(\theta)) = G_{\mathcal{A}} \cdot \theta$ for all regular θ, and $\Theta \to \mathcal{N}$ is a principal $G_{\mathcal{A}}$-bundle.*

Definition 4.2 (Fiber Curvature)**.** The *fiber curvature* at θ is the second fundamental form $II_\theta : V_\theta \times V_\theta \to H_\theta$ of the fiber $\Phi^{-1}(\Phi(\theta))$ as a submanifold of (Θ, g). It measures the rate at which the fiber's tangent space rotates as one moves along it.

Proposition 4.3 (Pruning as Fiber Projection)**.** *A pruning map $P : \Theta \to \Theta' \subseteq \Theta$ is* lossless at θ *if and only if $\Phi(P(\theta)) = \Phi(\theta)$, i.e., P moves θ within its fiber.*

For ReLU networks, the activation pattern is locally constant within each polyhedral region of Θ, making Φ locally linear and the fiber locally affine. Fiber curvature is concentrated at activation boundaries, where the local linear structure of Φ changes discontinuously.

5. The Failure of Parameter Geometry

Overview. A quantity $Q : \Theta \to \mathbb{R}$ is a *semantic invariant* if $Q(\theta) = Q(\sigma(\theta))$ for all $\sigma \in G_{\mathcal{A}}$, equivalently if Q factors through Φ. Three standard proxies for semantic properties fail this test.

Proposition 5.1 (Flatness is not a semantic invariant)**.** *For any architecture with scaling symmetries, the spectrum of $\nabla^2(\mathcal{L} \circ \Phi)(\theta^*)$ is not invariant under $G_{\mathcal{A}}$. For any $M > 0$ there exists $\sigma \in G_{\mathcal{A}}$ with $\Phi(\sigma(\theta^*)) = \Phi(\theta^*)$ and $\lambda_{\max}(\nabla^2(\mathcal{L} \circ \Phi)(\sigma(\theta^*))) > M$. The flatness measure is therefore a property of the coordinate representation, not of the function.*

Sketch. For a two-layer ReLU network, the transformation $(w, v) \mapsto (\alpha w, v/\alpha)$ is in $G_{\mathcal{A}}$ for any $\alpha > 0$ and leaves f_θ unchanged. The dominant term of the Hessian scales as α^2 in the rescaled neuron direction. Taking $\alpha \to \infty$ drives $\lambda_{\max}$ to infinity while $\Phi(\sigma_\alpha(\theta^*))$ remains fixed. □

Proposition 5.2 (Parameter magnitude is not a semantic invariant)**.** *For any architecture with a scaling symmetry and any $\epsilon > 0$, there exists $\sigma \in G_{\mathcal{A}}$ with $\Phi(\sigma(\theta)) = \Phi(\theta)$ and $\|\sigma(\theta)\|_1 < \epsilon$. Parameter magnitude can be redistributed arbitrarily within a fiber.*

Proposition 5.3 (Gradient descent is not reparameterization-invariant)**.** *Let $\psi : \Theta' \to \Theta$ be a smooth diffeomorphism that is not an isometry. Then the gradient descent trajectory in Θ' and the image under ψ^{-1} of the trajectory in Θ are generically distinct paths, even though they converge to the same semantic minimizers.*

Proposition 5.4 (Non-invariance implies semantic vacuity)**.** *Let $Q : \Theta \to \mathbb{R}$ not be constant on fibers of Φ. Then there exist θ_1, θ_2 with $\Phi(\theta_1) = \Phi(\theta_2)$ and $Q(\theta_1) \neq Q(\theta_2)$. Any conclusion drawn from Q about $f = \Phi(\theta)$ is either coincidental or assumes a canonical fiber representative.*

Parameter space has two legitimate roles: as a computational substrate for optimization, and as the total space of a fiber bundle over $\mathcal{N}$. In both roles it is studied relative to Φ, not as an intrinsic Euclidean space.

6. The Quotient Manifold

Overview. This section establishes the smooth structure, metric, geodesics, curvature, and free energy of $\mathcal{N}$.

Definition 6.1 (Semantic Metric)**.** Let $\mathcal{D}$ be a distribution on $\mathcal{X}$. For $\delta f_1, \delta f_2 \in T_f\mathcal{N}$, define

$$g_f(\delta f_1, \delta f_2) := \mathbb{E}_{x \sim \mathcal{D}}[\delta f_1(x)\, \delta f_2(x)].$$

The *semantic distance* is $d(f_1, f_2) := (\mathbb{E}_{x \sim \mathcal{D}}[(f_1(x) - f_2(x))^2])^{1/2}$.

Proposition 6.2 (Pullback Metric and Fisher Information)**.** *The pullback* $G_\theta(u,v) := g_{\Phi(\theta)}(D\Phi(\theta)u, D\Phi(\theta)v)$ *equals the Fisher information matrix up to a positive constant for probabilistic models [3].*

Proposition 6.3 (Vertical Directions are Metrically Null)**.** *For all* $\theta \in \Theta^{\text{reg}}$ *and all* $v \in V_\theta$, $G_\theta(v,w) = 0$ *for all* w*. The pullback metric descends to a non-degenerate metric on* $H_\theta \cong T_{\Phi(\theta)}\mathcal{N}$*.*

Proposition 6.4 (Curvature Decomposition)**.** *Let* $\tilde{e}_1, \tilde{e}_2$ *be horizontal lifts of orthonormal* $e_1, e_2 \in T_f\mathcal{N}$*. Then*

$$K^{\mathcal{N}}(e_1,e_2) = K^{\Theta}(\tilde{e}_1,\tilde{e}_2) + \|\Omega(\tilde{e}_1,\tilde{e}_2)\|^2 - B(\tilde{e}_1,\tilde{e}_2),$$

where Ω *is the connection curvature and* B *depends on the second fundamental form of the fibers [6]. The term* $\|\Omega\|^2$ *contributes positively: fiber twisting increases the intrinsic curvature of* $\mathcal{N}$*.*

Definition 6.5 (Semantic Free Energy)**.** The *policy complexity* is $\mathcal{C}(f) := \int_{\mathcal{X}} \|\nabla_x f(x)\|^2 \, d\mu(x)$, the *policy entropy* is $\mathcal{S}(f) := -\int p_f(y) \log p_f(y)\, dy$, and the *semantic free energy* is $\mathcal{F}(f) := \mathcal{C}(f) - \tau\mathcal{S}(f)$ for $\tau > 0$. Low free energy indicates a favorable balance between simplicity and informativeness.

7. Pushforward Measures and Statistical Structure

Overview. The semantic metric arises naturally from the pushforward of the data distribution through Φ. This section shows that the Fisher metric is not an additional assumption but a consequence of the realization map, and connects $\mathcal{N}$ to information geometry and optimal transport.

Definition 7.1 (Model Distribution)**.** For $\theta \in \Theta$, define the *model distribution* $\mu_\theta := f_{\theta\#}\mathcal{D}$, the pushforward of $\mathcal{D}$ through f_θ. The *statistical realization map* $\Phi_{\text{stat}} : \Theta \to \mathcal{P}(\mathcal{Y})$, $\Phi_{\text{stat}}(\theta) = \mu_\theta$, factors through Φ: $\mu_\theta = \mu_{\theta'}$ whenever $\Phi(\theta) = \Phi(\theta')$. Thus statistical structure depends only on the semantic class.

Proposition 7.2 (Second-Order Expansion of Divergence)**.** *Let* f_t *be a smooth curve in* $\mathcal{N}$ *with* $f_0 = f$*. Then*

$$D(\mu_{f_t}\|\mu_f) = \tfrac{1}{2}t^2\, g_f(\dot{f},\dot{f}) + o(t^2),$$

where g_f *is the semantic metric of Definition 6.1. The semantic metric is therefore the infinitesimal form of statistical distinguishability.*

Sketch. Expand the divergence to second order. The first variation vanishes at $t = 0$, and the second variation yields a quadratic form that, by direct computation, equals the Fisher information inner product. The identification follows from Definition 6.1. □

Proposition 7.3 (Metric Equivalence)**.** *For probabilistic models with likelihood* $p_\theta(y \mid x)$*, the pullback metric* G_θ *coincides up to a positive constant with the Fisher information matrix* $\mathbb{E}_{(x,y)}[\nabla_\theta \log p_\theta(y \mid x) \nabla_\theta \log p_\theta(y \mid x)^\top]$*.*

Proposition 7.4 (Semantic Divergence is Well-Defined)**.** *The divergence* $D(f_1,f_2) := D(\mu_{f_1}\|\mu_{f_2})$ *depends only on* $[f_1],[f_2] \in \mathcal{N}$ *and is invariant under* $G_{\mathcal{A}}$*. All statistical distinguishability lives on* $\mathcal{N}$*, not on* Θ*.*

The Wasserstein-2 metric $W_2(\mu_{f_1}, \mu_{f_2})$ defines an alternative geometry on $\mathcal{N}$ based on optimal transport [8], measuring global transport cost rather than local distinguishability. Both are induced by Φ and are therefore semantic. The geometry of $\mathcal{N}$ is simultaneously geometric, functional, and statistical.

8. The Quotient as a Universal Construction

Overview. The semantic manifold $\mathcal{N}$ is characterized by a universal property: it is the unique object through which all semantic invariants factor. This places the invariance thesis on categorical footing.

The equivalence relation induced by Φ can be expressed as a fiber product $\Theta \times_{\mathcal{N}} \Theta := \{(\theta_1,\theta_2) \in \Theta \times \Theta \mid \Phi(\theta_1) = \Phi(\theta_2)\}$, with two canonical projections $\pi_1, \pi_2 : \Theta \times_{\mathcal{N}} \Theta \rightrightarrows \Theta$.

Proposition 8.1 (Coequalizer Characterization)**.** *The semantic space* $\mathcal{N}$ *is the coequalizer of the diagram* $\Theta \times_{\mathcal{N}} \Theta \rightrightarrows \Theta \xrightarrow{\Phi} \mathcal{N}$*. For any space* Z *and map* $Q : \Theta \to Z$ *satisfying* $Q \circ \pi_1 = Q \circ \pi_2$*, there exists a unique map* $\tilde{Q} : \mathcal{N} \to Z$ *with* $Q = \tilde{Q} \circ \Phi$*.*

Corollary 8.2 (Factorization of Invariants)**.** *Every semantic invariant* Q *factors uniquely through* $\mathcal{N}$*. The semantic manifold is the minimal representation of all semantically meaningful quantities.*

Proposition 8.3 (Non-Invariance as Obstruction)**.** *Let* $Q : \Theta \to Z$ *not be constant on fibers. Then there does not exist any map* $\tilde{Q} : \mathcal{N} \to Z$ *with* $Q = \tilde{Q} \circ \Phi$*. Non-invariant quantities cannot be interpreted as*

functions of the learned model; they are functions of the implementation.

The following statements from earlier sections become immediate corollaries of the universal property: Proposition 5.4 holds because any non-invariant Q fails to factor through $\mathcal{N}$; the semantic metric g is well-defined because it depends only on $\Phi(\theta)$; and the pushforward measure μ_θ depends only on the equivalence class $[\theta]$. The quotient $\mathcal{N}$ is not a convenient abstraction; it is the unique space in which semantically meaningful quantities can be defined.

9. Learning as Geometric Motion

Overview. This section analyzes what optimization does when viewed through Φ, showing that gradient descent decomposes into horizontal (semantic) and vertical (redundant) components, and that the correct algorithm is natural gradient descent.

Proposition 9.1 (Gradient Descent is Horizontal)**.** *The standard gradient $\nabla_\theta(\mathcal{L}\circ\Phi)(\theta) = D\Phi(\theta)^*\nabla_f\mathcal{L}$ lies entirely in $H_\theta = (\ker D\Phi(\theta))^\perp$. The vertical component vanishes identically.*

Proof. The image of $D\Phi(\theta)^*$ is $(\ker D\Phi(\theta))^\perp = H_\theta$, since $\mathrm{im}(A^*) = (\ker A)^\perp$ for any linear map between inner product spaces. □

Despite this, standard gradient descent is not semantically correct: the horizontal component corresponds to a direction in $T_f\mathcal{N}$ via $D\Phi(\theta)$, but when the horizontal subspace is defined by the Euclidean metric on Θ rather than the semantic metric g, the resulting motion in $\mathcal{N}$ is not the Riemannian gradient of $\mathcal{L}$. This is the core mismatch.

Definition 9.2 (Natural Gradient)**.** The *natural gradient* update is $\dot\theta = -G_\theta^+\nabla_\theta\ell(\theta)$, where $G_\theta^+ = D\Phi(\theta)^+(g_{\Phi(\theta)}^{-1})D\Phi(\theta)^{+*}$ is the pseudoinverse of the pullback metric.

Proposition 9.3 (Natural Gradient is Semantically Correct)**.** *The natural gradient satisfies: $\dot\theta \in H_\theta$; $D\Phi(\theta)\dot\theta = -\mathrm{grad}_g\mathcal{L}(\Phi(\theta))$; and the update is $G_\mathcal{A}$-invariant. Standard gradient descent is not $G_\mathcal{A}$-invariant.*

Proposition 9.4 (Vertical Drift and Connection Curvature)**.** *For a gradient step of size η,*

$$(\theta' - \theta)_V = -\tfrac{\eta^2}{2}\,\Omega(\dot\theta,\dot\theta) + O(\eta^3).$$

High connection curvature generates significant vertical drift per step, causing the optimizer to wander within fibers while the semantic trajectory converges.

Proposition 9.5 (Curvature and Effective Learning Rate)**.** *In a region with $|K^\mathcal{N}| \le K_0$, the effective step size for guaranteed semantic progress satisfies $\eta_{\text{eff}} \le (L + K_0\, d(f, f^*))^{-1}$. High curvature forces smaller steps and degrades the convergence rate.*

10. Projected Dynamics and Variational Principle

Overview. This section derives the natural gradient from a constrained variational principle, removing any arbitrariness from the learning rule and establishing that learning is a projection problem induced by Φ.

The desired semantic evolution is $\dot f = -\mathrm{grad}_g\mathcal{L}(f)$, which induces the constraint on parameter velocities: $D\Phi(\theta)\,\dot\theta = -\mathrm{grad}_g\mathcal{L}(\Phi(\theta))$. Among all velocities satisfying this constraint, the natural choice is the one of minimal norm.

Proposition 10.1 (Constrained Variational Characterization)**.** *The natural gradient flow is the solution to*

$$\min_{\dot\theta\in T_\theta\Theta} \|\dot\theta\|^2 \quad \textit{subject to} \quad D\Phi(\theta)\,\dot\theta = -\mathrm{grad}_g\mathcal{L}(\Phi(\theta)).$$

Proof. Introduce a Lagrange multiplier $\lambda \in T_f^*\mathcal{N}$ and consider $\mathcal{J}(\dot\theta,\lambda) = \|\dot\theta\|^2 + \langle\lambda, D\Phi(\theta)\dot\theta + \mathrm{grad}_g\mathcal{L}\rangle$. Stationarity in $\dot\theta$ gives $2\dot\theta + D\Phi(\theta)^*\lambda = 0$, so $\dot\theta = -\frac{1}{2}D\Phi(\theta)^*\lambda$. Substituting into the constraint gives $D\Phi(\theta)D\Phi(\theta)^*\lambda = 2\,\mathrm{grad}_g\mathcal{L}$, determining λ. Substituting back yields the natural gradient. □

Corollary 10.2 (Natural Gradient as Pseudoinverse Projection)**.** *The solution is $\dot\theta = -D\Phi(\theta)^+\mathrm{grad}_g\mathcal{L}$, where $D\Phi(\theta)^+ := D\Phi(\theta)^*(D\Phi(\theta)D\Phi(\theta)^*)^{-1}$ is the Moore-Penrose pseudoinverse. The natural gradient is the least-squares solution to the semantic descent equation.*

Proposition 10.3 (Projection Interpretation)**.** *The operator $\Pi_\theta := D\Phi(\theta)^*(D\Phi(\theta)D\Phi(\theta)^*)^{-1}D\Phi(\theta)$ is the orthogonal projector onto H_θ. Learning proceeds by projecting the gradient onto H_θ and discarding all vertical components.*

Proposition 10.4 (Uniqueness of the Natural Gradient)**.** *Among all parameter updates satisfying*

$D\Phi(\theta)\dot{\theta} = -\mathrm{grad}_g\mathcal{L}$, *the natural gradient is the unique update minimizing* $\|\dot{\theta}\|$. *It is therefore not a heuristic but a canonical construction.*

Proposition 10.5 (First-Order Consistency)**.** *For step size* η, *the discrete update satisfies* $\Phi(\theta_{k+1}) = \Phi(\theta_k) - \eta\,\mathrm{grad}_g\mathcal{L} + O(\eta^2)$, *inducing the correct first-order semantic motion.*

Standard gradient descent fails because it does not solve the constrained variational problem: it uses the Euclidean metric to measure parameter velocity instead of finding the minimum-norm velocity that achieves the desired semantic change. The geometry of Φ determines both the constraint and its solution.

11. Identifiability and Observability

Overview. The realization map determines not only semantic content but everything that can be observed. The equivalence relation induced by Φ coincides with observational equivalence: no experiment on inputs and outputs can distinguish points within the same fiber.

Definition 11.1 (Observational Equivalence)**.** Parameters $\theta, \theta' \in \Theta$ are *observationally equivalent* if $f_\theta(x) = f_{\theta'}(x)$ for all $x \in \mathcal{X}$.

Proposition 11.2 (Equivalence with Fiber Relation)**.** $\theta \sim_{\mathrm{obs}} \theta'$ *if and only if* $\Phi(\theta) = \Phi(\theta')$.

Fibers are exactly the sets of parameters indistinguishable by any input-output experiment.

Proposition 11.3 (Failure of Identifiability)**.** *For architectures with nontrivial* $G_{\mathcal{A}}$, *statistical identifiability fails:* $\sigma(\theta) \neq \theta$ *but* $\mu_{\sigma(\theta)} = \mu_\theta$ *for all* $\sigma \in G_{\mathcal{A}}$. *The model is identifiable only up to equivalence classes, and this is the strongest identifiability condition that can hold.*

Theorem 11.4 (Indistinguishability under Φ)**.** *Let* O *be any observer with access only to input-output pairs* $(x, f_\theta(x))$ *drawn from* $\mathcal{D}$. *For any* θ, θ' *with* $\Phi(\theta) = \Phi(\theta')$, *the distributions of observations are identical:* $(x, f_\theta(x)) \overset{d}{=} (x, f_{\theta'}(x))$. *No statistical test can distinguish* θ *from* θ', *regardless of dataset size, test procedure, or computational resources.*

Proof. $\Phi(\theta) = \Phi(\theta')$ implies $f_\theta = f_{\theta'}$, so the joint distribution of $(x, f_\theta(x))$ is identical. □

Proposition 11.5 (Observables are Semantic Invariants)**.** *Every function* $Q : \Theta \to \mathbb{R}$ *estimable from input-output data is constant on fibers and therefore factors through* Φ. *Observable and semantic invariant are equivalent conditions.*

The invariance thesis thus has an observational reading: meaning is not what remains after ignoring implementation—it is what remains after exhausting all possible observations. Parameter norms, individual weights, and neuron activations tied to specific parameterizations are unobservable and cannot be inferred from data.

12. Stability and Perturbation Theory

Overview. The geometry induced by Φ determines how sensitive model outputs are to perturbations. This section develops a perturbation theory for Φ, connecting stability to the singular structure of $D\Phi$ and the curvature of $\mathcal{N}$.

For a perturbation $\delta\theta$, the induced change in function space is $\delta f = D\Phi(\theta)\,\delta\theta$. A perturbation $\delta\theta$ has zero semantic effect if and only if $\delta\theta \in V_\theta$; vertical directions are exactly the zero-sensitivity directions.

Proposition 12.1 (Sensitivity Spectrum)**.** *Let* $D\Phi(\theta) = U\Sigma V^\top$ *be the singular value decomposition with singular values* $\sigma_1 \geq \cdots \geq \sigma_d > 0$. *Large singular values correspond to highly sensitive directions; small singular values to weakly sensitive ones; and zero singular values to redundant (vertical) directions. The condition number* $\kappa(\theta) := \sigma_{\max}/\sigma_{\min}$ *controls the anisotropy of sensitivity and the stability of inversion.*

Proposition 12.2 (Second-Order Effects)**.** *Even when* $\delta\theta \in V_\theta$, *higher-order terms may induce nonzero semantic changes:* $D\Phi(\theta)\delta\theta = 0$ *does not imply* $\Phi(\theta + \delta\theta) = \Phi(\theta)$. *The second-order expansion is* $\Phi(\theta + \delta\theta) = \Phi(\theta) + D\Phi(\theta)\delta\theta + \frac{1}{2}D^2\Phi(\theta)(\delta\theta, \delta\theta) + O(\|\delta\theta\|^3)$, *so vertical directions are only infinitesimally null.*

Proposition 12.3 (Inverse Stability)**.** *The minimal-norm solution to* $D\Phi(\theta)\delta\theta = \delta f$ *is* $\delta\theta = D\Phi(\theta)^+\delta f$, *satisfying* $\|\delta\theta\| \leq \sigma_{\min}^{-1}\|\delta f\|$. *Small* $\sigma_{\min}$ *produces large parameter changes for small functional changes.*

Proposition 12.4 (Noise Filtering and Geodesic Deviation)**.** *If parameters are perturbed by noise* $\delta\theta$, *the output variance is* $\mathrm{Var}(\delta f) = D\Phi(\theta)\,\Sigma_\theta\,D\Phi(\theta)^\top$. *Vertical noise is completely filtered:* $\delta\theta \in V_\theta$ *implies*

$\mathrm{Var}(\delta f) = 0$. *Curvature of $\mathcal{N}$ introduces geodesic deviation: $D^2\delta f/dt^2 = -R(\dot\gamma, \delta f)\dot\gamma$, so positive curvature contracts nearby trajectories and negative curvature causes divergence.*

The realization map acts as a filter: it removes redundant directions, attenuates weakly sensitive ones, and amplifies sensitive ones. Stability is therefore a property of how Φ transforms parameter perturbations into functional changes, not a property of parameters per se. Flat regions—where $\|D\Phi(\theta)\|$ is small in relevant directions—correspond to low sensitivity, robustness to noise, and improved generalization.

13. Gauge Symmetry and Redundancy

Overview. The redundancy of parameter space is structured by symmetry. This section shows that the fibers of Φ arise from a gauge symmetry and that learning dynamics have a natural gauge-fixed interpretation.

Proposition 13.1 (Infinitesimal Gauge Generators are Vertical)**.** *For $\xi \in \mathfrak{g} = \mathrm{Lie}(G_{\mathcal{A}})$, define the vector field $X_\xi(\theta) = (d/dt)|_{t=0} \exp(t\xi)\cdot\theta$. Then $X_\xi(\theta) \in V_\theta = \ker D\Phi(\theta)$.*

Proof. Since $\Phi(\exp(t\xi)\cdot\theta) = \Phi(\theta)$ for all t, differentiating at $t = 0$ gives $D\Phi(\theta)X_\xi(\theta) = 0$. □

Vertical directions are therefore generated by infinitesimal symmetries. Gauge equivalence implies semantic equivalence: $\theta \sim_{\text{gauge}} \theta'$ implies $\Phi(\theta) = \Phi(\theta')$, though the converse need not hold if the full fiber is larger than the group orbit.

Definition 13.2 (Connection as Gauge Fixing)**.** A choice of horizontal subspace H_θ complementing V_θ is a *connection* on the bundle $\Theta \to \mathcal{N}$. Choosing H_θ is equivalent to selecting a gauge: it picks a unique representative of each equivalence class for infinitesimal motion. Natural gradient descent is motion in a fixed gauge.

Proposition 13.3 (Curvature as Gauge Obstruction)**.** *The curvature $\Omega(X, Y) = \mathrm{proj}_V([X^H, Y^H])$ is the obstruction to integrability of the horizontal distribution. Nonzero curvature implies that parallel transport is path-dependent and that no global gauge fixing exists. This corresponds to entanglement between semantic and implementation degrees of freedom.*

Proposition 13.4 (Gauge-Projected Dynamics)**.** *The natural gradient flow satisfies $\dot\theta \in H_\theta$ at every point: learning is constrained to remain orthogonal to all gauge directions, eliminating gauge redundancy at the level of dynamics.*

Every gauge-invariant function descends to $\mathcal{N}$, and gauge invariance is sufficient for semantic invariance. The central thesis acquires a gauge-theoretic reading: parameters are coordinates in a gauge space; models are points in the quotient.

14. Generalization and Flat Minima

Overview. Generalization depends on the semantic dimension of the hypothesis class and the connection curvature, not on the parameter count.

Definition 14.1 (Semantic Hypothesis Class)**.** $\mathcal{H}_{B,\kappa} := \{f \in \mathcal{N} \mid \mathcal{F}(f) \le B,\ \|\Omega_f\| \le \kappa\}$, where $\mathcal{F}$ is the semantic free energy of Definition 6.5.

Proposition 14.2 (Semantic Dimension Reduction)**.** $\log N(\varepsilon, \mathcal{H}_{B,\kappa}, d) \le C_1 d \log(A_B/\varepsilon) + C_2\kappa\,\varepsilon^{-1}$. *The covering number depends on the semantic dimension d, not the parameter count n.*

Theorem 14.3 (Semantic Generalization Bound)**.** *With probability $\ge 1-\delta$, uniformly for all $f \in \mathcal{H}_{B,\kappa}$,*

$$\mathrm{GenErr}(f) \le \inf_{\varepsilon>0}\left\{\varepsilon + \sqrt{\frac{C_1 d\log(A_B/\varepsilon) + C_2\kappa\,\varepsilon^{-1} + \log(1/\delta)}{n}}\right\}.$$

When κ is small the bound reduces to $\sqrt{d\log n/n}$ (dimension-dominant regime); when κ is large the bound degrades to $(\kappa/n)^{1/3}$ (curvature-dominant regime), showing that entanglement worsens the generalization rate independently of model size.

Theorem 14.4 (Rademacher Bound on the Semantic Class)**.** *Under the assumptions of the hypothesis class definition,*

$$\mathrm{GenErr}(f) \le \frac{C'L}{\sqrt{n}}\left(\sqrt{d} + \sqrt{\kappa}\,A_B^{1/2}\right) + \sqrt{\frac{\log(1/\delta)}{n}},$$

up to logarithmic factors. The two terms measure semantic complexity and geometric entanglement respectively.

Definition 14.5 (Semantic Flatness)**.** A minimizer $f^* \in \mathcal{N}$ is *semantically flat* if $\nabla^2_g\mathcal{L}(f^*)$ has small eigenvalues in the metric g. This is $G_{\mathcal{A}}$-invariant; the flat minima heuristic in Θ is its coordinate-dependent approximation, correct only when $D\Phi(\theta)$ is approximately an isometry.

15. Projection and Compression

Overview. The realization map induces a compression from parameter space to semantic space. This section formalizes all compression schemes as instances of a single construction and derives the geometric criterion for lossless compression.

Definition 15.1 (Compression Map)**.** A *compression map* $\rho : \Theta \to \Theta'$ is *lossless at* θ^* if $\Phi(\rho(\theta^*)) = \Phi(\theta^*)$. The *compression error* is $\mathcal{E}(\rho, \theta^*) := d(\Phi(\rho(\theta^*)), \Phi(\theta^*))$.

Pruning, distillation, low-rank factorization, and quantization are all instances: they differ in the form of Θ' and the degree to which they are fiber-aware. Distillation directly minimizes $\mathcal{E}$; the others use coordinate-dependent proxies. A compression is *fiber-aware* if $\rho(\theta^*) - \theta^*$ lies approximately in V_{θ^*}; in that case $\mathcal{E}(\rho, \theta^*) \leq \delta \|D\Phi(\theta^*)\| \cdot \|\rho(\theta^*) - \theta^*\|$ for small $\delta > 0$.

Proposition 15.2 (Lossless Compression Existence)**.** *A lossless compression of θ^* into Θ' exists if and only if $\Phi^{-1}(\Phi(\theta^*)) \cap \Theta' \neq \emptyset$.*

Proposition 15.3 (Universal Compression Criterion)**.** *Φ factors through Θ' (universally lossless compression) if and only if the fibers of ρ contain the fibers of Φ:* $\ker(\rho_*) \supseteq V_\theta$ *at every regular point.*

Proposition 15.4 (Information Bottleneck)**.** *The differential $D\Phi(\theta)$ acts as a local information bottleneck: its rank is the effective dimension of information flow, and directions with small singular values are attenuated. Fibers encode lost information; large fibers correspond to high redundancy and improved robustness.*

Proposition 15.5 (Fiber Size, Compression, and Generalization)**.** *The fiber dimension $n-d$ simultaneously upper-bounds the lossless compression gain and lower-bounds the generalization improvement from working in $\mathcal{N}$ rather than Θ. The geometry of the fiber governs both phenomena through a single object.*

16. Entanglement and Curvature

Overview. This section examines the condition in which the fiber is geometrically twisted so that semantic and implementation degrees of freedom cannot be cleanly separated. Entanglement is measured by the connection curvature Ω and the second fundamental form II.

Proposition 16.1 (Entanglement as Non-integrability)**.** *The horizontal distribution H_θ is integrable if and only if $\Omega \equiv 0$. When $\Omega \neq 0$, horizontal vector fields fail to close under the Lie bracket, so Θ does not locally decompose as $\mathcal{N} \times G_{\mathcal{A}}$ and semantic motion necessarily generates vertical displacement.*

Sketch. Frobenius' theorem: a distribution is integrable if and only if it is involutive. The obstruction to involutivity for a principal bundle connection is exactly the curvature Ω. □

Proposition 16.2 (Holonomy and Implementation Drift)**.** *Let γ be a loop in $\mathcal{N}$ based at f. The horizontal lift starting at $\theta \in \Phi^{-1}(f)$ returns to $\tilde{\gamma}(1) = \sigma_\gamma(\theta)$, where $\sigma_\gamma \in G_{\mathcal{A}}$ is the holonomy element. Even when the semantic state returns to its starting point, the implementation may not. The fiber is dynamically active under learning.*

Proposition 16.3 (Curvature Obstructs Factorization)**.** *If $\Omega \neq 0$ in a neighborhood of θ^*, then no coordinate system on Θ exists in which Φ decomposes locally as $\Phi(\theta_{\text{sem}}, \theta_{\text{impl}}) = \Phi(\theta_{\text{sem}})$. Any factored approximation incurs semantic error proportional to $\|\Omega\|$.*

By the O'Neill formula (Proposition 6.4), $\|\Omega\|^2$ contributes positively to $K^{\mathcal{N}}$: nearby semantic directions couple under transport and geodesics diverge from linear approximations more rapidly in entangled regions.

Proposition 16.4 (Entanglement Degrades Optimization, Generalization, and Compression)**.** *When $\|\Omega\|$ is large: horizontal updates induce vertical drift of order $\eta^2 \|\Omega\|$; the discrepancy operator becomes ill-conditioned; gradient directions misalign with semantic gradients; covering numbers and Rademacher complexity increase via the curvature term $\sqrt{\kappa} A_B^{1/2} / \sqrt{n}$; and fiber navigation becomes geometrically costly, preventing compression schemes from finding optimal fiber representatives. Two models with identical parameter counts and similar training loss can differ dramatically in generalization: the difference is curvature, not size.*

Definition 16.5 (Geometric Disentanglement)**.** A representation is *disentangled* in a region if $\|\Omega\| \approx 0$ and $\|II\| \approx 0$. In this regime fibers are flat, horizontal motion is path-independent, compression is fiber-aware and efficient, and optimization aligns with semantics. Normalization schemes, residual connections, and prescribed architectural symmetries func-

tion as curvature-reduction mechanisms in precisely this sense.

Entanglement is not an abstract property of representations. It is a measurable geometric quantity—the curvature of the realization bundle—that determines whether redundancy can be exploited, whether learning aligns with meaning, and whether models can be simplified without loss.

17. Global Topology and Nontrivial Structure of $\mathcal{N}$

Overview. All preceding sections describe local properties of Φ. This section studies the global topology induced by Φ, establishing that $\mathcal{N}$ is a stratified space with nontrivial structure that constrains optimization at a topological level.

The quotient topology on $\mathcal{N}$ is the finest topology making Φ continuous. It may fail to be Hausdorff if fibers intersect nontrivially under limits. The rank stratification of Section 3 induces a stratification $\mathcal{N} = \bigsqcup_{k=0}^{d} \mathcal{N}_k$ where $\mathcal{N}_k := \Phi(\Theta_k)$ and each $\mathcal{N}_k$ is a manifold of dimension k.

Proposition 17.1 (Topological Obstruction to Optimization)**.** *$\mathcal{N}$ may have multiple connected components, corresponding to function classes that cannot be continuously deformed into one another. If f_1 and f_2 lie in different components, no continuous optimization path in Θ can transform one into the other without leaving* image(Φ)*. Optimization is therefore constrained by topology, not only by geometry.*

Examples of distinct components include classifiers with different decision boundary topology, functions with different global symmetry properties, and representations with distinct topological invariants.

Proposition 17.2 (Holonomy and Homotopy)**.** *Parallel transport of θ along a loop γ in $\mathcal{N}$ returns to $\sigma_\gamma(\theta) \in \Phi^{-1}(\Phi(\theta))$: loops in $\mathcal{N}$ correspond to gauge transformations in Θ. Nontrivial homotopy classes in $\mathcal{N}$ indicate topological holes or higher-dimensional features corresponding to global invariants of function space.*

Proposition 17.3 (Morse-Theoretic Constraints)**.** *The loss functional $\mathcal{L} : \mathcal{N} \to \mathbb{R}$ defines a landscape over $\mathcal{N}$. The number and type of critical points are constrained by Morse inequalities applied to the topology of $\mathcal{N}$. The topological complexity of $\mathcal{N}$ therefore imposes lower bounds on the number of critical points of any training objective.*

The semantic manifold $\mathcal{N}$ is not a simple smooth space. It is stratified, possibly non-Hausdorff, and carries nontrivial homotopy. Learning is navigation on a compressed, curved, and topologically structured space of functions, constrained not only by local geometry but by global invariants. This completes the picture: local structure (Sections 3–12) and gauge structure (Section 13) are embedded in a global topology that governs what can be reached at all.

18. Meaning as Invariant

Overview. This section states what all preceding sections are collectively establishing. The claim is philosophical in form but carries precise mathematical content: meaning is invariant under implementation, and the correct unit of analysis is the equivalence class under Φ.

The invariance thesis has teeth because it is violated systematically in current practice. Loss landscapes are analyzed in Θ, not $\mathcal{N}$. Generalization bounds are stated in terms of parameter norms. Pruning is performed by magnitude thresholding. Interpretability research analyzes individual neurons. Each produces the same error: a coordinate-dependent quantity is mistaken for a semantic property.

Definition 18.1 (Semantic Content)**.** The *semantic content* of θ is the point $\Phi(\theta) \in \mathcal{N}$. Two parameters have the same semantic content if and only if they lie in the same fiber.

Theorem 18.2 (Realization Principle)**.** *All quantities invariant under $G_{\mathcal{A}}$ factor uniquely through Φ and are functions on $\mathcal{N}$. Conversely, any non-invariant quantity cannot be interpreted as a property of the learned function: it is a property of the implementation.*

Proof. The first claim is Corollary 8.2. The second is Proposition 5.4 together with the universal property of the coequalizer (Proposition 8.1). □

Working in $\mathcal{N}$ does not require choosing a canonical fiber representative. It requires performing operations whose output is independent of which representative is chosen. Natural gradient descent satisfies this by construction. Distillation satisfies this by minimizing semantic distance directly. The generalization bound satisfies this by depending on d and

κ, both defined on $\mathcal{N}$. The indistinguishability theorem (Theorem 11.4) gives the observational reading: meaning is what remains after exhausting all possible observations, not merely what remains after ignoring implementation.

Everything in the essay compresses into a single chain:

$$\Phi(\theta_1) = \Phi(\theta_2) \;\Rightarrow\; [\theta_1] = [\theta_2] \;\Rightarrow\; \text{same meaning} \;\Rightarrow\; \text{same generalization, compression, and dynamics.} \tag{1}$$

19. Connections and Open Problems

Overview. The framework connects to information geometry, algebraic geometry, the loss landscape literature, equivariant architectures, mechanistic interpretability, optimal transport, and gauge theory.

Open Problem 19.1 (Information-geometric characterization of Ω)**.** Express the connection curvature Ω in terms of the α-connection structure of information geometry [3] and determine whether entanglement conditions correspond to known properties of statistical model classes.

Open Problem 19.2 (Algebraic fiber structure)**.** For architectures with polynomial activations, compute the degree and dimension of the fibers of Φ as algebraic varieties [4] and extend to ReLU networks via semialgebraic geometry.

Open Problem 19.3 (Semantic local minima)**.** Determine whether every local minimum of ℓ in Θ projects to a local minimum of $\mathcal{L}$ in $\mathcal{N}$, whether spurious local minima correspond to degenerate implementations, and whether Morse theory on $\mathcal{N}$ accounts for the topology of sublevel sets of ℓ on Θ [1].

Open Problem 19.4 (Optimal symmetry group design)**.** Given target $\mathcal{N}^*$ and parameter budget n, characterize symmetry groups G achieving $\mathcal{N} \supseteq \mathcal{N}^*$ while maximizing $\dim G$ [7, 9].

Open Problem 19.5 (Invariant feature construction)**.** Develop systematic methods for constructing $G_{\mathcal{A}}$-invariant quantities from network weights and determine when such invariants separate points of $\mathcal{N}$ [2, 5].

Open Problem 19.6 (Wasserstein structure of $\mathcal{N}$)**.** For probabilistic models, determine the relationship between the semantic metric g and the Wasserstein-2 metric on model distributions, and connect the semantic free energy $\mathcal{F}$ to displacement convexity in Wasserstein space [8].

Open Problem 19.7 (Gauge-theoretic optimization)**.** Formulate natural gradient as a gauge-covariant gradient flow and determine whether Yang-Mills equations have a meaningful analogue in the optimization of $\mathcal{L}$ on $\mathcal{N}$.

Open Problem 19.8 (Global topology of $\mathcal{N}$)**.** Compute the fundamental group $\pi_1(\mathcal{N})$ and higher homotopy groups for specific architectures. Determine whether non-trivial loops correspond to observable training phenomena and whether the Morse theory of $\mathcal{L}$ on $\mathcal{N}$ provides a complete account of optimization landscape topology.

Open Problem 19.9 (Effective computation on $\mathcal{N}$)**.** Develop efficient algorithms for optimization, generalization assessment, compression, and interpretation directly on $\mathcal{N}$ without computing the full Jacobian $D\Phi(\theta)$. This is the primary technical obstacle between the theoretical framework and its practical application.

20. Conclusion: The Geometry of What Is Learned

The parameters of a neural network are not its meaning. From this observation, pursued through differential geometry, probability theory, category theory, variational analysis, gauge theory, and global topology, a complete geometric theory has emerged. Every result in the preceding sections is a consequence of taking the realization map $\Phi : \Theta \to \mathcal{N}$ seriously as a mathematical object.

The fibers of Φ are smooth submanifolds of Θ with dimension determined by $G_{\mathcal{A}}$. The local normal form makes fibers explicit as coordinate planes, and Sard's theorem guarantees they are well-behaved generically. The quotient $\mathcal{N} \cong \Theta/G_{\mathcal{A}}$ carries the Fisher information metric, characterized as the infinitesimal form of statistical distinguishability via the pushforward of the data distribution. The universal property of the coequalizer makes the invariance thesis categorically inevitable: semantic invariants are precisely the quantities that factor through Φ, and non-invariants are provably unobservable.

Learning is the unique minimum-norm solution to a constrained variational problem—natural gradient descent—derived without appeal to heuristics. Stability is governed by the singular values of $D\Phi$ and the

curvature of $\mathcal{N}$, with vertical directions filtering noise and curvature causing geodesic deviation. The symmetry group $G_{\mathcal{A}}$ generates the vertical directions as infinitesimal gauge transformations, and the connection defines a gauge-fixed dynamics. Generalization depends on $d = \dim \mathcal{N}$ and the connection curvature κ; compression is lossless iff it preserves fibers; entanglement—measured by Ω—degrades all three simultaneously. The semantic manifold carries a global topology that constrains what can be reached by any continuous optimization.

The framework implies three things for practice. The relevant dimension is $d = n - \dim G_{\mathcal{A}}$, not n. The symmetry group $G_{\mathcal{A}}$ is a design parameter: larger symmetry groups produce larger fibers, smaller d, and better generalization per parameter. Training procedures that reduce curvature—normalization, equivariant architectures, residual connections—improve not just optimization stability but compression capacity and generalization simultaneously, because all three are governed by the same geometric object.

The central open problem is computational: the Jacobian $D\Phi(\theta)$ has dimensions $|\mathcal{N}| \times n$ and is prohibitively large for modern architectures. Efficient approximations remain the primary obstacle between this theoretical framework and practice.

What the framework provides is a correct starting point: a precise identification of the objects that matter and a rigorous account of why they matter. The geometry of what a neural network learns is the geometry of the semantic manifold $\mathcal{N}$, not the geometry of the parameter space Θ. The realization map $\Phi : \Theta \to \mathcal{N}$ is the structure connecting the two.

Parameters are coordinates in a gauge space. Models are points in the quotient. Every question about learning, generalization, compression, and interpretation that is asked in Θ should be asked instead in $\mathcal{N}$. Meaning is invariant under implementation.

References

[1] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[2] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, N. DasSarma, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021.

[3] S. ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[4] J. Kileel, M. Trager, and J. Bruna. On the expressive power of deep polynomial neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[5] C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001.

[6] B. O'Neill. The fundamental equations of a submersion. *Michigan Mathematical Journal*, 13(4): 459–469, 1966.

[7] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. In *arXiv preprint arXiv:1802.08219*, 2018.

[8] C. Villani. *Optimal Transport: Old and New*. Springer, 2009.

[9] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R. Salakhutdinov, and A. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.