Singular Information Geometry of Deep Neural Networks

Hanwen Ge*

January 2026

Abstract

What geometry is induced by a deep network through the task it represents? We anchor geometry on the output side by fixing the Fisher–Rao metric on a chosen statistical output manifold and pulling it back through the task map to parameter space and intermediate representation spaces. The resulting pullback Fisher metrics are positive semidefinite and generically degenerate; their null distributions encode task-invisible degrees of freedom and induce pseudometrics whose metric identification yields an effective geometry, smooth on constant-rank strata. We separate population-induced degeneracy from finite-sample artifacts by comparing population and dataset pullbacks. Finally, on frozen-tail slices we obtain a two-sided pullback identity linking representation-side and parameter-side degeneracy layer by layer, enabling layerwise tomography of task-induced equivalence classes.

Contents

1 Introduction

We introduce a geometric framework for deep neural networks that combines *information geometry* (via the Fisher–Rao metric on an output statistical manifold) with *degenerate pullback geometry* (i.e. positive-semidefinite pullback metrics, null distributions/foliations, and metric-identification quotients). The core point is:

- Fix a canonical base metric on the output side: the Fisher–Rao metric G_{FR} on a chosen finite-dimensional manifold of output distributions S°.
- Pull this metric back through the network to obtain task-aware (generally degenerate) geometries on representation spaces M_ℓ and on parameter space Θ.
- Treat the induced null distributions and metric-identification quotients as primary objects: they encode task-induced equivalence classes (invariances, non-identifiability, redundancy) in a coordinate-free way.

What is new (and what is not). This note does *not* claim novelty for the general pullback/quotient machinery (degenerate pullbacks, constant-rank patches, metric identification by zero distance). What is new here is: (i) an *output-first* anchoring in the canonical Fisher–Rao geometry on a chosen output statistical manifold, (ii) an explicit *population-versus-sample* quotient viewpoint that isolates spurious empirical null directions as a coverage phenomenon, and (iii) a *two-sided* (layerwise, frozen-tail) identity that ties representation-side degeneracy to parameter-side degeneracy and supports a layerwise "tomography" of task-invisible directions.

Contributions.

- **Output-first viewpoint:** fix $(S^\circ, G_{\mathrm{FR}})$ and pull back, rather than imposing ad hoc geometries.
- **Quotient-first diagnostics:** treat null leaves/quotients as task-relevant observables, not nuisances.

- **Two-sided unification:** on frozen-tail slices, $g^{(\Theta)}_{x,\ell} = F^*_{\ell,x} g^{(\ell)}_{\theta^\star_{>\ell}}$, enabling layerwise attribution of degeneracy.
- **Population-first clarity:** sample null directions can be spurious; this becomes explicit geometry rather than folklore.

1.1 Regularity and scope

We state the level of regularity implicitly used throughout, so that later patchwise arguments read as deliberate rather than apologetic.

- **Base manifold.** We work on an open set S° of a finite-dimensional statistical manifold S where the Fisher–Rao metric G_{FR} is smooth and positive definite. Boundary singularities (e.g. probabilities approaching 0 in classification) are *not* treated as part of the present singularity analysis.
- **Piecewise smooth networks.** For ReLU-type networks, maps such as $\theta \mapsto \Psi_x(\theta)$ and $\theta \mapsto F_{\ell,x}(\theta)$ are typically only piecewise smooth. All differentials and constant-rank claims are interpreted on connected components of activation/parameter strata where the relevant maps are smooth.
- **Patchwise constant rank.** Quotient statements are made on open patches where the relevant map has constant rank and fibers are path-connected; globally the resulting effective space is generally stratified.
- **Distances.** Pseudodistances are defined via lengths of piecewise C^1 curves; this is the minimal structure needed for metric identification by zero distance.
- **Population objects.** The population metric $g_p = \mathbb{E}_{x\sim p}[g_x]$ is treated as a positive semidefinite field of bilinear forms. We assume measurability and integrability sufficient to define curve lengths and the induced pseudodistance. We do *not* require interchange of derivatives and expectations except where explicitly stated.
- **Discrete symmetries.** Patchwise metric identification captures continuous null directions on regular patches. Discrete symmetries (e.g. neuron permutations) generally require an additional global quotient step.

1.2 Core notation

We list only the symbols needed for Sections 2–3; a full table appears in Appendix A.

- Θ parameter space (open subset of $\mathbb{R}^P$), S° output statistical manifold, G_{FR} Fisher–Rao metric on S°.
- $\Psi_x : \Theta \to S^\circ$, $\Psi_x(\theta) = p_\theta(\cdot \mid x)$; per-input pullback $g^{(\Theta)}_x := \Psi_x^* G_{\mathrm{FR}}$.
- Input measure $p(x)$; population metric $g^{(\Theta)}_p := \mathbb{E}_{x\sim p}[g^{(\Theta)}_x]$.
- Input sample $\mathcal{D}_X = \{x_n\}_{n=1}^N$; sample metric $g^{(\Theta)}_{\mathcal{D}_X} := \frac{1}{N}\sum_{n=1}^N g^{(\Theta)}_{x_n}$.
- Layer-ℓ representation space M_ℓ; representation map $F_{\ell,x}(\theta) = h_\ell(x;\theta)$.

Reader's guide. Section 2 positions the note against adjacent literatures and states the intended novelty boundary. Section 3 fixes the output manifold (S°) and defines per-input, population, and sample parameter metrics. Section 4 records the generic pullback/quotient construction. Sections 5 and 6 apply this to representations and parameters, and Section 7 states the frozen-tail two-sided identity that links the two geometries.

2 Positioning and related work

Information geometry and natural gradient. Our starting point is classical information geometry, where the Fisher–Rao metric provides a canonical Riemannian structure on statistical manifolds and motivates natural-gradient methods [1, 2]. In contrast to work that treats Fisher primarily as an optimization preconditioner (including structured curvature approximations such as K-FAC [8]), we treat Fisher–Rao on the output distribution space as a fixed reference metric and view the resulting pullbacks as geometric observables. This shifts emphasis from step directions to structure: null spaces, induced pseudodistances, and quotient geometries that encode task-invisible degrees of freedom.

Singular pullback geometry for neural networks. The closest technical antecedent is the singular Riemannian geometry line that studies sequences of maps induced by deep networks, pullback degeneracies, pseudometric structures, and metric-identification (Kolmogorov) quotients yielding smooth manifolds under hypotheses [4]. Our contribution is not the pullback/quotient machinery itself. Rather, we (i) anchor the construction in the output statistical manifold with Fisher–Rao as the canonical base metric, (ii) make the sample-versus-population distinction explicit as a source of spurious null directions, and (iii) introduce a two-sided, layerwise coupling via a frozen-tail identity that ties parameter-side and representation-side degeneracies into a single diagnostic program.

Pulling back Fisher–Rao to internal coordinates. Pulling back Fisher–Rao through a learned map has also been pursued for deep generative models, where the Fisher–Rao geometry of decoder distributions is pulled back to latent space to define meaningful latent geometries beyond Gaussian decoders [3]. We share the same output-distribution-first principle but focus on supervised conditional output distributions $p_\theta(\cdot \mid x)$ and on intermediate representations inside a discriminative network. This leads naturally to sample-stacked and population constructions and to the frozen-tail factorization, which enables layerwise attribution of degeneracy that is not present in the standard latent-geometry setup.

Geometric deep learning as a contrast class. A distinct strand of geometric deep learning builds architectures that are equivariant by design, including gauge-equivariant convolutional networks on manifolds [5]. Our goal is different: we do not propose new equivariant operators or symmetry-enforcing architectures. Instead, we diagnose task-induced equivalence classes and degeneracies that arise in a trained model by studying Fisher–Rao pullback geometry and the associated quotient structure.

Relation to singular learning theory. Finally, our emphasis on degeneracy aligns with singular learning theory, which treats neural networks as singular statistical models and explains why classical regular inference approximations can fail [9, 10]. We are not developing RLCT-based generalization theory here. Instead, we provide a differential-geometric interface to singularity in modern networks:

degeneracy appears concretely as null directions of Fisher–Rao pullbacks, and quotienting turns those directions into an effective geometry.

3 Output statistical manifold and Fisher–Rao base geometry

Let X denote inputs and Z denote labels (or targets). We identify the input space X with the layer-0 representation space M_0. For each fixed input x, we assume the model defines an output distribution $p_\theta(\cdot \mid x)$ belonging to a chosen finite-dimensional statistical manifold S. To avoid boundary singularities (e.g. simplex faces), we work on an open region $S^\circ \subseteq S$ on which the Fisher–Rao metric is smooth and positive definite, and assume $p_\theta(\cdot \mid x) \in S^\circ$ on the region of interest.

Per-input pullback metric on parameters. For each x, define the per-input map

$$\Psi_x : \Theta \to S^\circ, \qquad \Psi_x(\theta) := p_\theta(\cdot \mid x),$$

and its pullback metric on Θ:

$$g_x^{(\Theta)} := \Psi_x^* G_{\mathrm{FR}}.$$

Population task parameter geometry. Given an input distribution $p(x)$ (assumed independent of θ), define the population metric:

$$g_{p,\theta}^{(\Theta)}(u, v) := \mathbb{E}_{x\sim p}[g_{x,\theta}^{(\Theta)}(u, v)], \qquad u, v \in T_\theta\Theta. \tag{1}$$

We assume measurability/integrability sufficient for the expectation to exist for each fixed θ, u, v.

Input-sample approximation. Given an input sample $\mathcal{D}_X = \{x_n\}_{n=1}^N$,

$$g_{\mathcal{D}_X}^{(\Theta)} := \frac{1}{N}\sum_{n=1}^{N} g_{x_n}^{(\Theta)}. \tag{2}$$

Equivalently, define $\Psi_{\mathcal{D}_X} : \Theta \to (S^\circ)^N$ by

$$\Psi_{\mathcal{D}_X}(\theta) := (p_\theta(\cdot \mid x_1), \ldots, p_\theta(\cdot \mid x_N)),$$

and equip $(S^\circ)^N$ with the normalized product metric

$$G_{\mathrm{FR}}^{\otimes N} := \frac{1}{N}\bigoplus_{n=1}^{N} G_{\mathrm{FR}}.$$

Then $g_{\mathcal{D}_X}^{(\Theta)} = \Psi_{\mathcal{D}_X}^* G_{\mathrm{FR}}^{\otimes N}$.

Exact Fisher pullback on an input sample. In coordinates, the matrix representation of $g_{\mathcal{D}_X}^{(\Theta)}$ is

$$\mathbf{F}_{\mathcal{D}_X}(\theta) := \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{z\sim p_\theta(\cdot|x_n)}\Big[\nabla_\theta \log p_\theta(z \mid x_n)\, \nabla_\theta \log p_\theta(z \mid x_n)^\top\Big],$$

i.e. the *exact* Fisher pullback of G_{FR} on the empirical input support.

Remark 3.1 (Exact Fisher pullback vs. empirical Fisher). A common object in deep-learning optimization is the *empirical Fisher* formed from observed labels in a labeled dataset $\mathcal{D} = \{(x_n, z_n)\}_{n=1}^N$:

$$\mathbf{F}_{\mathcal{D}}^{\mathrm{emp}}(\theta) := \frac{1}{N} \sum_{n=1}^{N} \nabla_\theta \log p_\theta(z_n \mid x_n) \, \nabla_\theta \log p_\theta(z_n \mid x_n)^\top.$$

This differs from the exact Fisher pullback $\mathbf{F}_{\mathcal{D}_X}(\theta)$ above, which takes expectation over $z \sim p_\theta(\cdot \mid x_n)$. The two coincide only under restrictive conditions (e.g. if labels are sampled from the model, or in certain well-specified limits). In general, empirical Fisher need not approximate the Fisher information and can exhibit distinct pathologies; see [6] for a detailed analysis.

4 Degenerate pullback geometry

We now formulate the general pullback construction and its quotient interpretation.

Definition 4.1 (Pullback metric and degenerate pullback structure)**.** Let (Y, G) be a smooth Riemannian manifold with metric G (positive definite), and let $\Phi : Z \to Y$ be a smooth map between manifolds. The pullback metric $g := \Phi^* G$ is the smooth field of symmetric bilinear forms on TZ defined by

$$g_z(u, v) := G_{\Phi(z)}\big(d\Phi_z(u),\, d\Phi_z(v)\big), \qquad u, v \in T_z Z.$$

In general g_z is only positive semidefinite, so the induced metric may be degenerate. We will say that (Z, g) carries a degenerate pullback geometry when g is positive semidefinite and may have variable rank across Z.

Remark 4.2 (Coordinate expression: Cauchy–Green / Gauss–Newton form). In local coordinates, write the metric on Y as a positive definite matrix $G(y)$ and let $J_\Phi(z)$ denote the Jacobian of Φ at z. Then the pullback metric has the matrix form

$$g(z) = J_\Phi(z)^\top \, G(\Phi(z)) \, J_\Phi(z).$$

In particular, when G is positive definite, $\operatorname{rank}(g(z)) = \operatorname{rank}(J_\Phi(z))$, and degeneracy is equivalent to rank loss of the differential. This is the same "Cauchy–Green" structure that underlies Gauss–Newton and Fisher-type curvature operators in deep learning.

Remark 4.3 (Piecewise-smooth networks and strata). The statements in this section assume Φ is smooth. For piecewise-smooth architectures (e.g. ReLU networks), interpret all differentials and constant-rank claims on each smooth activation stratum, where the forward map is C^∞. Rank variation across strata boundaries yields a stratified geometric picture. Extending beyond strata (e.g. via Clarke differentials) is a separate technical direction.

Remark 4.4 (Parameter-space strata for ReLU-type models). For ReLU-type models, maps such as $F_{\ell,x}(\theta)$ and $\Psi_x(\theta)$ are also piecewise smooth as functions of θ: activation boundaries move in parameter space. Throughout, differentials in θ (e.g. $dF_{\ell,x,\theta}$, $d\Psi_{x,\theta}$) and constant-rank statements should be interpreted on parameter-space strata where the active set is fixed (so the relevant maps are smooth).

Lemma 4.5 (Null space and differential)**.** *Under the above assumptions,*

$$\ker g_z \;=\; \ker d\Phi_z \;\subset\; T_z Z, \qquad \forall z \in Z.$$

Proof. If $v \in \ker d\Phi_z$ then $d\Phi_z(v) = 0$, so for any $u \in T_z Z$,

$$g_z(v, u) = G_{\Phi(z)}(0, d\Phi_z(u)) = 0.$$

Conversely, if $v \in \ker g_z$ then $g_z(v, v) = 0$, i.e.

$$G_{\Phi(z)}(d\Phi_z(v), d\Phi_z(v)) = 0.$$

Since G is positive definite, this implies $d\Phi_z(v) = 0$. □

Definition 4.6 (Pseudometric induced by a pullback)**.** Let (Z, g) be as above. For a piecewise C^1 curve $\gamma : [0, 1] \to Z$ define

$$L_g(\gamma) := \int_0^1 \sqrt{g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))}\, dt.$$

The induced pseudo-distance is

$$d_g(z_0, z_1) := \inf_\gamma L_g(\gamma),$$

where the infimum is over all piecewise C^1 curves $\gamma : [0, 1] \to Z$ with $\gamma(0) = z_0$ and $\gamma(1) = z_1$.

Definition 4.7 (Restricted pseudo-distance on a patch)**.** Let $U \subset Z$ be open. Define

$$d_{g,U}(z_0, z_1) := \inf_\gamma L_g(\gamma),$$

where the infimum is taken over piecewise C^1 curves $\gamma : [0, 1] \to U$ with $\gamma(0) = z_0$ and $\gamma(1) = z_1$. *Standing convention:* we only apply $d_{g,U}$ on path-connected open patches U so that piecewise C^1 connecting curves exist.

Lemma 4.8 (Length identity under pullback)**.** *Let (Y, G) be Riemannian, let $\Phi : Z \to Y$ be C^1, and let $g = \Phi^* G$. For any piecewise C^1 curve $\gamma : [0, 1] \to Z$,*

$$L_g(\gamma) = L_G(\Phi \circ \gamma).$$

Proof. For almost every t,

$$g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t)) = G_{\Phi(\gamma(t))}\big(d\Phi_{\gamma(t)}\dot\gamma(t),\, d\Phi_{\gamma(t)}\dot\gamma(t)\big) = G_{(\Phi\circ\gamma)(t)}\big((\Phi \circ \gamma)'(t), (\Phi \circ \gamma)'(t)\big),$$

so the integrands coincide and hence the lengths coincide. □

Corollary 4.9 (Distance domination)**.** *Let (Y, G) be Riemannian, let $\Phi : Z \to Y$ be C^1, and let $g = \Phi^* G$. Let d_Y be the Riemannian distance on Y induced by G. Then for any $z_0, z_1 \in Z$,*

$$d_Y\big(\Phi(z_0), \Phi(z_1)\big) \le d_g(z_0, z_1).$$

In particular, $d_g(z_0, z_1) = 0 \Rightarrow \Phi(z_0) = \Phi(z_1)$.

Proof. For any piecewise C^1 curve γ from z_0 to z_1,

$$d_Y(\Phi(z_0), \Phi(z_1)) \le L_G(\Phi \circ \gamma) = L_g(\gamma)$$

by Lemma 4.8. Taking the infimum over such γ gives the claim. □

Definition 4.10 (Metric identification (Kolmogorov) quotient)**.** Let (Z, d_g) be a pseudometric space. Define an equivalence relation $z_0 \sim_g z_1 \iff d_g(z_0, z_1) = 0$. The metric-identification quotient (often called the Kolmogorov quotient in degenerate-metric contexts) is the set

$$Z_g^{\text{eff}} := Z/\sim_g,$$

equipped with the induced metric $\overline{d}_g([z_0], [z_1]) := d_g(z_0, z_1)$.

Remark 4.11 (Terminology: "Kolmogorov quotient" vs. T_0 quotient)*.* In general topology, "Kolmogorov quotient" often refers to the T_0 (Kolmogorov) separation quotient of a topological space. In a pseudometric space, identifying points at zero distance is the standard *metric identification* construction; it coincides with the T_0 quotient of the topology induced by the pseudometric. We keep the term "Kolmogorov quotient" only to align with degenerate-metric and singular-Riemannian treatments in the network literature; see [7] for a metric/topological discussion and [4] for an explicit DNN pullback-metric usage.

Definition 4.12 (Patchwise metric-identification quotient)**.** Let $U \subset Z$ be open and path-connected and let $d_{g,U}$ be the restricted pseudo-distance (Definition 4.7). Define $z_0 \sim_{g,U} z_1$ iff $d_{g,U}(z_0, z_1) = 0$, and define the patchwise effective space

$$U_g^{\text{eff}} := U/\sim_{g,U}$$

with induced metric $\overline{d}_{g,U}([z_0]_U, [z_1]_U) := d_{g,U}(z_0, z_1)$.

Assumption 4.13 (Constant rank and regular fibers)**.** The map $\Phi : Z \to Y$ has locally constant rank on an open subset $U \subset Z$. Equivalently, for each $z \in U$ there is a neighbourhood U_z on which $\operatorname{rank}(d\Phi)$ is constant. In this case, by the constant rank theorem, each fiber $\Phi^{-1}(y) \cap U_z$ is a smoothly embedded submanifold of U_z, and its tangent space at z equals $\ker d\Phi_z$.

Proposition 4.14 (Metric and functional equivalence on a patch)**.** *Let (Y, G) be Riemannian, let $\Phi : Z \to Y$ be smooth, and let $g = \Phi^* G$. Assume Φ has locally constant rank on an open subset $U \subset Z$ and that each fiber $\Phi^{-1}(y) \cap U$ is path-connected. Then for $z_0, z_1 \in U$,*

$$d_{g,U}(z_0, z_1) = 0 \quad \iff \quad \Phi(z_0) = \Phi(z_1).$$

Proof. ($\Rightarrow$) (This direction does not use fiber connectivity.) If $d_{g,U}(z_0, z_1) = 0$, then $d_g(z_0, z_1) \leq d_{g,U}(z_0, z_1) = 0$ since d_g takes the infimum over a larger class of curves. By Corollary 4.9, $\Phi(z_0) = \Phi(z_1)$.

($\Leftarrow$) (Here fiber path-connectedness is used to construct a curve inside the fiber.) If $\Phi(z_0) = \Phi(z_1) = y$, then $z_0, z_1 \in \Phi^{-1}(y) \cap U$, which is path-connected by assumption. Thus there exists a piecewise C^1 curve $\gamma : [0, 1] \to \Phi^{-1}(y) \cap U$ from z_0 to z_1. Since $\Phi \circ \gamma$ is constant, Lemma 4.8 gives $L_g(\gamma) = 0$, hence $d_{g,U}(z_0, z_1) = 0$. □

Remark 4.15 (Global quotient picture)*.* Proposition 4.14 is intentionally local. When the rank of Φ varies or fibers change topology, the induced quotient is generally stratified (rank strata, disconnected fiber components, symmetry loci). Throughout we work on patches where constant rank and fiber connectivity hold, and interpret global structure as a stratified gluing of such patchwise effective spaces.

Concretely, one may view these patchwise quotients as local charts of a stratified quotient: the global identification relation is generated by the zero-distance relations on patches and their overlaps (i.e. by taking the transitive closure across an atlas of constant-rank regions).

5 Representation-space geometry

We apply the construction to representation spaces M_ℓ of a fixed network snapshot. For standard architectures, each M_ℓ is naturally identified with a Euclidean space (or an open subset) $\mathbb{R}^{d_\ell}$. The *stratification* arises from the maps (e.g. ReLU gating), not from the ambient manifold: forward maps are piecewise smooth and their differentials (and ranks) can change across activation boundaries. Accordingly, all smoothness and constant-rank statements below are interpreted on connected components of activation strata where the relevant maps are C^∞; on each such patch we work in the usual Euclidean chart on M_ℓ.

Remark 5.1 (Ambient vs reachable representation sets). For fixed θ, the set of reachable representations under the task measure is

$$\mathcal{R}_{\ell,\theta} := \{h_\ell(x;\theta) : x \in \mathrm{supp}(p)\} \subset M_\ell.$$

In applications one often evaluates (or restricts) geometry along $\mathcal{R}_{\ell,\theta}$, since directions never excited by $p(x)$ are practically irrelevant. For an input sample $\mathcal{D}_X = \{x_n\}_{n=1}^N$, the sampled reachable set is $\mathcal{R}_{\ell,\theta,\mathcal{D}_X} := \{h_\ell(x_n;\theta)\}_{n=1}^N$.

Unless stated otherwise, when we interpret representation-side null directions, effective dimensions, or quotients in applications, we mean along $\mathcal{R}_{\ell,\theta}$ (or its sampled version $\mathcal{R}_{\ell,\theta,\mathcal{D}_X}$), since directions never excited by $p(x)$ are operationally irrelevant.

Layered maps and tail. Write parameters as $\theta = (\theta_1, \dots, \theta_L)$ with θ_j governing layer j, and define $\theta_{\le \ell} := (\theta_1, \dots, \theta_\ell)$ and $\theta_{>\ell} := (\theta_{\ell+1}, \dots, \theta_L)$. On each smooth activation stratum, a deep network defines smooth maps

$$M_0 \xrightarrow{\Lambda_{1,\theta_1}} M_1 \xrightarrow{\Lambda_{2,\theta_2}} \cdots \xrightarrow{\Lambda_{L,\theta_L}} M_L,$$

with $f_\theta = \Lambda_{L,\theta_L} \circ \cdots \circ \Lambda_{1,\theta_1}$. For $0 \le \ell < L$, define the tail map

$$N_{\ell,\theta_{>\ell}} := \Lambda_{L,\theta_L} \circ \cdots \circ \Lambda_{\ell+1,\theta_{\ell+1}} : M_\ell \to M_L.$$

Head map and representation-to-output map. Let $\mathcal{H} : M_L \to S^\circ$ be a smooth head map (e.g. softmax, Gaussian map). For each fixed tail parameter value $\theta_{>\ell}$, define

$$\Phi_{\ell,\theta_{>\ell}} := \mathcal{H} \circ N_{\ell,\theta_{>\ell}} : M_\ell \to S^\circ.$$

This induces a tail-indexed family of representation metrics

$$g^{(\ell)}_{\theta_{>\ell}} := (\Phi_{\ell,\theta_{>\ell}})^* G_{\mathrm{FR}}.$$

When isolating layerwise structure, freeze the tail beyond layer ℓ at a snapshot $\theta^\star$, producing the fixed map

$$\Phi_\ell^{\theta^\star_{>\ell}} := \Phi_{\ell,\theta^\star_{>\ell}}, \qquad g^{(\ell)}_{\theta^\star_{>\ell}} := (\Phi_\ell^{\theta^\star_{>\ell}})^* G_{\mathrm{FR}}.$$

Representation pullback metric and patchwise effective space. On the frozen tail, null directions are representation perturbations that do not change the induced output distribution to first order: by Lemma 4.5, $\ker g^{(\ell)}_{\theta^\star_{>\ell},h_\ell} = \ker d(\Phi_\ell^{\theta^\star_{>\ell}})_{h_\ell}$.

Let $U \subset M_\ell$ be a path-connected constant-rank patch on which fibers of $\Phi_\ell^{\theta^\star_{>\ell}}$ are path-connected. Define $h_\ell \sim_{\ell,U} h'_\ell$ iff $d_{g^{(\ell)}_{\theta^\star_{>\ell}},U}(h_\ell, h'_\ell) = 0$, and define the effective representation space on the patch as

$$U_\ell^{\mathrm{eff}} := U^{\mathrm{eff}}_{g^{(\ell)}_{\theta^\star_{>\ell}}} \;=\; U/\!\sim_{\ell,U}\,.$$

Globally, one obtains a stratified quotient assembled from such patchwise effective spaces (Remark 4.15).

Remark 5.2 (Discrete symmetries and disconnected fibers). Many neural-network symmetries (e.g. neuron permutations) act discretely and can produce disconnected components inside a single functional fiber. Patchwise metric-identification quotients collapse connected fiber components on constant-rank patches; identifying disconnected components corresponding to discrete symmetries is a separate global step, typically modeled as an additional quotient by a discrete group action. Thus zero-distance equivalence captures continuous invariances on regular patches; discrete symmetries require a separate identification step.

6 Parameter-space geometry

The parameter-side geometry is induced by the per-input maps $\Psi_x : \Theta \to S^\circ$ and their averages.

Population geometry in squared-norm form. For $v \in T_\theta\Theta$, since $g_x^{(\Theta)} = \Psi_x^* G_{\mathrm{FR}}$ we have

$$g^{(\Theta)}_{p,\theta}(v,v) = \mathbb{E}_{x\sim p}\Big[G_{\mathrm{FR},\,\Psi_x(\theta)}\big(d\Psi_{x,\theta}(v),\, d\Psi_{x,\theta}(v)\big)\Big] = \mathbb{E}_{x\sim p}\Big[\big\|d\Psi_{x,\theta}(v)\big\|^2_{G_{\mathrm{FR}}}\Big], \tag{3}$$

where $\|\cdot\|_{G_{\mathrm{FR}}}$ denotes the norm induced by G_{FR} on S°.

Lemma 6.1 (Zero squared norm implies kernel for PSD forms)**.** *Let g be a symmetric positive semidefinite bilinear form on a real vector space. If $g(v,v) = 0$, then $g(v,w) = 0$ for all w, i.e. $v \in \ker g$.*

Proof. For any $t \in \mathbb{R}$ and any w, positive semidefiniteness gives

$$0 \le g(v+tw,\, v+tw) = g(v,v) + 2t\, g(v,w) + t^2 g(w,w).$$

If $g(v,v) = 0$, the right-hand side is a quadratic in t that is nonnegative for all t. This forces the linear coefficient to vanish, hence $g(v,w) = 0$ for all w. □

Corollary 6.2 (Population null directions are almost-surely null differentials)**.** *Assume G_{FR} is positive definite on S° and the expectation in* (3) *exists. Then for $v \in T_\theta\Theta$,*

$$v \in \ker g^{(\Theta)}_{p,\theta} \quad\Longleftrightarrow\quad d\Psi_{x,\theta}(v) = 0 \text{ for } p\text{-almost-every } x.$$

Proof. ($\Rightarrow$) If $v \in \ker g^{(\Theta)}_{p,\theta}$ then $g^{(\Theta)}_{p,\theta}(v,v) = 0$, hence by (3)

$$\mathbb{E}_{x\sim p}\Big[\big\|d\Psi_{x,\theta}(v)\big\|^2_{G_{\mathrm{FR}}}\Big] = 0.$$

The integrand is nonnegative, so it must vanish p-almost-everywhere. Since G_{FR} is positive definite on S°, $\|d\Psi_{x,\theta}(v)\|_{G_{\mathrm{FR}}} = 0$ implies $d\Psi_{x,\theta}(v) = 0$ for p-a.e. x.

($\Leftarrow$) If $d\Psi_{x,\theta}(v) = 0$ for p-a.e. x, then $g^{(\Theta)}_{x,\theta}(v,w) = 0$ for all w and p-a.e. x. Taking expectations yields $g^{(\Theta)}_{p,\theta}(v,w) = 0$ for all w, so $v \in \ker g^{(\Theta)}_{p,\theta}$. □

Dataset geometry and spurious null directions. The dataset metric $g^{(\Theta)}_{\mathcal{D}_X}$ approximates $g^{(\Theta)}_p$ when $\mathcal{D}_X$ covers p well. Finite input samples can create spurious null directions: directions invisible on sampled inputs but visible elsewhere. This is a geometric expression of dataset coverage (and a practical warning about over-interpreting empirical null spaces).

6.1 Parameter-space equivalence and quotients (population vs dataset)

Population quotient (intrinsic/ideal). Apply Definition 4.6 with $(Z, g) = (\Theta, g^{(\Theta)}_p)$ to obtain a pseudo-distance $d_{g^{(\Theta)}_p}$ on Θ. Define $\theta \sim_p \theta'$ iff $d_{g^{(\Theta)}_p}(\theta, \theta') = 0$, and write the resulting effective space as

$$\Theta^{\text{eff}}_p := \Theta/\!\sim_p .$$

Remark 6.3 (Regularity of $g^{(\Theta)}_p$ as a metric field). Unlike $g^{(\Theta)}_{\mathcal{D}_X}$, which is a pullback of a smooth product map on each smooth/constant-rank stratum, the population object $g^{(\Theta)}_p$ is defined as an expectation. For the purposes of this note we treat $g^{(\Theta)}_p$ as a given positive semidefinite field of bilinear forms and assume enough regularity along piecewise C^1 curves to make the length functional well-defined (e.g. measurability of $t \mapsto g^{(\Theta)}_{p,\gamma(t)}(\dot\gamma(t), \dot\gamma(t))$ and integrability on $[0, 1]$). This is the minimal assumption needed to speak about the induced pseudo-distance and metric-identification quotient at a DL-theory level of formality.

Remark 6.4 (What does $\sim_p$ mean?). Operationally, $\sim_p$ should be read as *task-level indistinguishability under the input measure* p: along directions in $\ker g^{(\Theta)}_p$, the conditional output distributions do not change to first order for p-almost-every x (Corollary 6.2). Thus Θ^{eff}_p is the space of task-distinguishable parameter classes under $p(x)$. In this note, $\sim_p$ is used as an ideal limit object; all rigorous quotient statements are made patchwise (on smooth/constant-rank strata) for finite input samples or for strata where the relevant maps are smooth.

Dataset quotient (probe). Similarly, apply Definition 4.6 to $g^{(\Theta)}_{\mathcal{D}_X}$ to obtain $d_{g^{(\Theta)}_{\mathcal{D}_X}}$ and define

$$\theta \sim_{\mathcal{D}_X} \theta' \quad \Longleftrightarrow \quad d_{g^{(\Theta)}_{\mathcal{D}_X}}(\theta, \theta') = 0, \qquad \Theta^{\text{eff}}_{\mathcal{D}_X} := \Theta/\!\sim_{\mathcal{D}_X} .$$

This identifies parameters that are indistinguishable *on the sampled inputs*. In particular, $\sim_{\mathcal{D}_X}$ can be strictly coarser than $\sim_p$ because finite samples can hide directions that matter off-sample (spurious null directions).

Remark 6.5 (Discrete symmetries (parameters)). Discrete symmetries such as neuron permutations can produce disconnected components inside a functional fiber in parameter space. Patchwise metric-identification quotients collapse connected components; identifying discrete symmetry-related components typically requires a separate global identification (e.g. quotienting by a discrete group action). Thus zero-distance equivalence captures continuous invariances on regular patches; discrete symmetries require a separate identification step.

7 Two-sided unification

The representation-side and parameter-side geometries are linked canonically whenever the task map factors through representations, and freezing makes the dependence on tails explicit and controllable.

Frozen-tail slice and inclusion. Fix a snapshot $\theta^\star = (\theta^\star_{\le \ell}, \theta^\star_{>\ell})$ and consider the frozen-tail slice

$$\Theta^{(\ell,\theta^\star_{>\ell})} := \{(\theta_{\le \ell}, \theta^\star_{>\ell}) : \theta_{\le \ell} \in \Theta_{\le \ell}\}.$$

Let $\iota_{\ell,\theta^\star_{>\ell}} : \Theta^{(\ell,\theta^\star_{>\ell})} \hookrightarrow \Theta$ denote the inclusion map. Let $P_{\le \ell} := \dim(\Theta^{(\ell,\theta^\star_{>\ell})})$.

Restricted per-input map on the frozen-tail slice. Define

$$\Psi_x^{(\ell,\theta^\star_{>\ell})} := \Psi_x\big|_{\Theta^{(\ell,\theta^\star_{>\ell})}}, \qquad g_{x,\ell}^{(\Theta)} := \big(\Psi_x^{(\ell,\theta^\star_{>\ell})}\big)^* G_{\mathrm{FR}}.$$

Equivalently,

$$g_{x,\ell}^{(\Theta)} = \iota_{\ell,\theta^\star_{>\ell}}^* \, g_x^{(\Theta)}.$$

Per-input factorisation and diagram. Fix an input $x \in M_0$ and define the layer-ℓ representation map on the slice

$$F_{\ell,x} : \Theta^{(\ell,\theta^\star_{>\ell})} \to M_\ell, \qquad F_{\ell,x}(\theta) := h_\ell(x;\theta),$$

and the frozen-tail representation-to-output map $\Phi_\ell^{\theta^\star_{>\ell}} : M_\ell \to S^\circ$. On the slice,

$$\Psi_x^{(\ell,\theta^\star_{>\ell})} = \Phi_\ell^{\theta^\star_{>\ell}} \circ F_{\ell,x}. \tag{4}$$

This is captured by the commutative diagram

$$\begin{array}{ccc} \Theta^{(\ell,\theta^\star_{>\ell})} & \xrightarrow{F_{\ell,x}} & M_\ell \\ \Big\downarrow{\scriptstyle \Psi_x^{(\ell,\theta^\star_{>\ell})}} & & \Big\downarrow{\scriptstyle \Phi_\ell^{\theta^\star_{>\ell}}} \\ S^\circ & = & S^\circ \end{array}$$

Lemma 7.1 (Functoriality of pullback for metrics)**.** *Let $F : A \to B$ and $\Phi : B \to Y$ be smooth maps and let (Y, G) be Riemannian. Then*

$$(\Phi \circ F)^* G = F^*(\Phi^* G).$$

Proof. For any $a \in A$ and $u, v \in T_a A$,

$$((\Phi \circ F)^* G)_a(u,v) = G_{\Phi(F(a))}\big(d(\Phi \circ F)_a u,\ d(\Phi \circ F)_a v\big).$$

By the chain rule, $d(\Phi \circ F)_a = d\Phi_{F(a)} \circ dF_a$, hence

$$= G_{\Phi(F(a))}\big(d\Phi_{F(a)}(dF_a u),\ d\Phi_{F(a)}(dF_a v)\big) = (F^*(\Phi^* G))_a(u,v). \qquad \square$$

Corollary 7.2 (Pointwise two-sided unification on a frozen tail)**.** *Let $g^{(\ell)}_{\theta^\star_{>\ell}} := (\Phi_\ell^{\theta^\star_{>\ell}})^* G_{\mathrm{FR}}$ on M_ℓ and $g_{x,\ell}^{(\Theta)} := (\Psi_x^{(\ell,\theta^\star_{>\ell})})^* G_{\mathrm{FR}}$ on $\Theta^{(\ell,\theta^\star_{>\ell})}$. Then for every input x,*

$$g_{x,\ell}^{(\Theta)} \ = \ F_{\ell,x}^* \, g^{(\ell)}_{\theta^\star_{>\ell}}. \tag{5}$$

Remark 7.3 (Unfrozen tails: tail-indexed representation metrics). Without freezing the tail, the factorization (4) still holds, but one should read the identity as a pullback of the tail-indexed family $g^{(\ell)}_{\theta_{>\ell}} := (\Phi_{\ell,\theta_{>\ell}})^* G_{\mathrm{FR}}$ rather than the pullback of a fixed metric. Freezing makes the pullback identity literal in the usual functorial sense and is useful as a layerwise diagnostic: "what geometry does the head impose given a fixed tail/decoder?"

Population and dataset consequences (restricted to the frozen-tail slice). Define the restricted population metric on the slice by

$$g_{p,\ell}^{(\Theta)} := \iota_{\ell,\theta^\star_{>\ell}}^* \, g_p^{(\Theta)} \quad \text{on } \Theta^{(\ell,\theta^\star_{>\ell})},$$

and for an input sample $\mathcal{D}_X = \{x_n\}_{n=1}^N$ define

$$g_{\mathcal{D}_X,\ell}^{(\Theta)} := \iota_{\ell,\theta^\star_{>\ell}}^* \, g_{\mathcal{D}_X}^{(\Theta)} \quad \text{on } \Theta^{(\ell,\theta^\star_{>\ell})}.$$

Since pullback is linear in the metric and expectation is pointwise, we have

$$g_{p,\ell}^{(\Theta)} = \mathbb{E}_{x\sim p}[g_{x,\ell}^{(\Theta)}], \qquad g_{\mathcal{D}_X,\ell}^{(\Theta)} = \frac{1}{N}\sum_{n=1}^N g_{x_n,\ell}^{(\Theta)}.$$

Averaging (5) yields

$$g_{p,\ell}^{(\Theta)} = \mathbb{E}_{x\sim p}[F_{\ell,x}^* \, g_{\theta^\star_{>\ell}}^{(\ell)}], \qquad g_{\mathcal{D}_X,\ell}^{(\Theta)} = \frac{1}{N}\sum_{n=1}^N F_{\ell,x_n}^* \, g_{\theta^\star_{>\ell}}^{(\ell)}.$$

Worked example. Section 8 computes the preceding objects explicitly for a one-hidden-layer softmax classifier, highlighting how task-induced gauge redundancies manifest as null leaves and metric-identification quotients.

8 Worked example: one-hidden-layer softmax classifier and logit-shift gauge redundancy

This section provides a concrete example in which the pullback Fisher–Rao geometry is explicitly computable and the resulting degeneracies admit a simple quotient interpretation. The example also makes the frozen-tail identity (Section 7) fully explicit.

Model. Let $x \in M_0 \subset \mathbb{R}^d$ be an input and let $Z = \{1, \dots, K\}$ be the label set. Consider a one-hidden-layer network with hidden width m:

$$h(x;\theta_{\le 1}) := \sigma(W_1 x + a_1) \in M_1 := \mathbb{R}^m, \tag{6}$$

$$\eta(h;\theta_{>1}) := W_2 h + b_2 \in \mathbb{R}^K, \tag{7}$$

$$p_\theta(\cdot \mid x) = \mathrm{softmax}(\eta(h(x;\theta_{\le 1});\theta_{>1})) \in S^\circ = \Delta_\circ^{K-1}. \tag{8}$$

Here σ is any (piecewise) smooth nonlinearity (e.g. ReLU, GELU), and $\theta = (\theta_{\le 1}, \theta_{>1})$ with $\theta_{\le 1} = (W_1, a_1)$ and $\theta_{>1} = (W_2, b_2)$. The per-input task map is $\Psi_x(\theta) := p_\theta(\cdot \mid x)$.

Factorisation through the hidden representation. Define the representation map $F_{1,x} : \Theta \to M_1$ by

$$F_{1,x}(\theta) := h(x;\theta_{\le 1}),$$

and the tail-induced representation-to-output map $\Phi_{1,\theta_{>1}} : M_1 \to S^\circ$ by

$$\Phi_{1,\theta_{>1}}(h) := \mathrm{softmax}(W_2 h + b_2).$$

Then for every x,

$$\Psi_x(\theta) = \Phi_{1,\theta_{>1}}(F_{1,x}(\theta)). \tag{9}$$

Accordingly, the tail-indexed representation metric is

$$g^{(1)}_{\theta_{>1}} := (\Phi_{1,\theta_{>1}})^* G_{\mathrm{FR}} \quad \text{on } M_1,$$

and the parameter metric is $g^{(\Theta)}_x := \Psi_x^* G_{\mathrm{FR}}$ on Θ.

Remark 8.1 (Working on S°). We assume $p_\theta(\cdot \mid x) \in S^\circ = \Delta_\circ^{K-1}$ so that G_{FR} is smooth and positive definite on S°. As emphasized earlier, this example isolates *pullback-induced* degeneracy from boundary singularities of the base metric.

8.1 Fisher–Rao in logit coordinates and the canonical gauge direction

A key convenience of the softmax head is that Fisher–Rao admits a closed form in logit coordinates. Let $p = \mathrm{softmax}(\eta) \in \Delta_\circ^{K-1}$. Define the $K \times K$ matrix

$$G_\eta(\eta) \;:=\; \mathrm{Diag}(p) - pp^\top. \tag{10}$$

This is the pullback of the Fisher–Rao metric on $\Delta_\circ^{K-1}$ to the redundant logit chart $\eta \in \mathbb{R}^K$. It is positive semidefinite of rank $K-1$.

Lemma 8.2 (Softmax gauge direction)**.** *For any $\eta \in \mathbb{R}^K$ with $p = \mathrm{softmax}(\eta) \in \Delta_\circ^{K-1}$,*

$$G_\eta(\eta)\,\mathbf{1} = 0, \qquad \mathrm{rank}\big(G_\eta(\eta)\big) = K - 1,$$

where $\mathbf{1} \in \mathbb{R}^K$ is the all-ones vector. Equivalently, $\ker G_\eta(\eta) = \mathrm{span}\{\mathbf{1}\}$.

Proof. The identity $(\mathrm{Diag}(p) - pp^\top)\mathbf{1} = p - p(\mathbf{1}^\top p) = p - p = 0$ is immediate. For $p \in \Delta_\circ^{K-1}$, $\mathrm{Diag}(p) - pp^\top$ is the covariance matrix of a categorical distribution and has rank $K-1$, with nullspace equal to $\mathrm{span}\{\mathbf{1}\}$. □

Remark 8.3 (Interpretation). Lemma 8.2 expresses the classical invariance of softmax under logit shifts: $\mathrm{softmax}(\eta + c\mathbf{1}) = \mathrm{softmax}(\eta)$ for any scalar c. Geometrically, $\eta \mapsto \eta + c\mathbf{1}$ is a gauge direction in the redundant logit chart.

8.2 Representation-side pullback metric and null directions

Fix $\theta_{>1} = (W_2, b_2)$ and consider the map $\Phi_{1,\theta_{>1}} : h \mapsto \mathrm{softmax}(W_2 h + b_2)$. Let $p(h) := \mathrm{softmax}(W_2 h + b_2)$ and write $G_\eta(h) := \mathrm{Diag}(p(h)) - p(h)p(h)^\top$. Then the representation metric admits the explicit form

$$g^{(1)}_{\theta_{>1}}(h) \;\equiv\; W_2^\top\, G_\eta(h)\, W_2 \qquad \text{as a PSD bilinear form on } T_h M_1 \cong \mathbb{R}^m. \tag{11}$$

Lemma 8.4 (Characterisation of representation null directions)**.** *Fix $\theta_{>1} = (W_2, b_2)$ and $h \in M_1$. A direction $\delta h \in T_h M_1 \cong \mathbb{R}^m$ satisfies $\delta h \in \ker g^{(1)}_{\theta_{>1},h}$ if and only if*

$$W_2\,\delta h \in \mathrm{span}\{\mathbf{1}\} \subset \mathbb{R}^K. \tag{12}$$

Proof. By (11), $\delta h \in \ker g^{(1)}_{\theta_{>1},h}$ iff $0 = \delta h^\top W_2^\top G_\eta(h) W_2 \delta h$, i.e. $(W_2\delta h)^\top G_\eta(h)(W_2\delta h) = 0$. Since $G_\eta(h)$ is PSD, this is equivalent to $W_2\delta h \in \ker G_\eta(h)$. By Lemma 8.2, $\ker G_\eta(h) = \mathrm{span}\{\mathbf{1}\}$, giving (12). □

Remark 8.5 (Task-invisible representation variations). Lemma 8.4 identifies the task-invisible representation directions: those perturbations δh that change logits only along the logit-shift gauge direction $\mathbf{1}$. In particular, if W_2 has a nontrivial preimage of $\mathrm{span}\{\mathbf{1}\}$, then $g^{(1)}_{\theta_{>1}}$ is degenerate even when W_2 has full row rank.

8.3 A continuous redundant parametrisation of the tail (exact functional symmetry)

The softmax gauge immediately induces a continuous redundancy in the head parameters.

Definition 8.6 (Logit-shift gauge action on the tail)**.** For any $q \in \mathbb{R}^m$ and $c \in \mathbb{R}$, define a transformation of tail parameters

$$(W_2, b_2) \;\mapsto\; (W_2', b_2') := \big(W_2 + \mathbf{1}q^\top,\;\; b_2 + c\,\mathbf{1}\big). \tag{13}$$

Proposition 8.7 (Gauge redundancy yields zero-distance equivalence)**.** *Let $\theta = (\theta_{\le 1}, \theta_{>1})$ and define $\theta' = (\theta_{\le 1}, \theta'_{>1})$ where $\theta'_{>1}$ is obtained from $\theta_{>1}$ by the gauge action* (13)*. Then for every input x,*

$$\Psi_x(\theta') = \Psi_x(\theta), \tag{14}$$

and consequently $d_{g_x^{(\Theta)}}(\theta, \theta') = 0$ for all x. In particular, for any input distribution $p(x)$ and any input sample $\mathcal{D}_X$,

$$d_{g_p^{(\Theta)}}(\theta, \theta') = 0, \qquad d_{g_{\mathcal{D}_X}^{(\Theta)}}(\theta, \theta') = 0.$$

Proof. Let $h := h(x; \theta_{\le 1})$. Under (13),

$$W_2' h + b_2' = (W_2 + \mathbf{1}q^\top)h + (b_2 + c\mathbf{1}) = (W_2 h + b_2) + (q^\top h + c)\,\mathbf{1}.$$

Since $\mathrm{softmax}(\eta + \alpha\mathbf{1}) = \mathrm{softmax}(\eta)$ for all $\alpha \in \mathbb{R}$, we obtain (14). To obtain the zero-distance statements, fix x and consider the smooth curve

$$\gamma(t) := \big(\theta_{\le 1},\; W_2 + t\,\mathbf{1}q^\top,\; b_2 + t\,c\mathbf{1}\big), \qquad t \in [0, 1].$$

Then $\gamma(0) = \theta$ and $\gamma(1) = \theta'$. By (14), $\Psi_x \circ \gamma$ is constant, hence Lemma 4.8 gives $L_{g_x^{(\Theta)}}(\gamma) = 0$ and therefore $d_{g_x^{(\Theta)}}(\theta, \theta') = 0$. Since $g_p^{(\Theta)}$ and $g_{\mathcal{D}_X}^{(\Theta)}$ are averages of the per-input metrics, the same curve has zero length for these averaged metrics as well, yielding the remaining claims. □

Remark 8.8 (Quotient interpretation on constant-rank patches). On a patch where the stacked map $\Psi_{\mathcal{D}_X}$ (or the per-input map Ψ_x) has constant rank and fibers are path-connected, Proposition 8.7 implies that the patchwise effective space Θ^{eff} collapses gauge orbits $\{(W_2 + \mathbf{1}q^\top,\; b_2 + c\mathbf{1}) : q \in \mathbb{R}^m,\; c \in \mathbb{R}\}$. This is the simplest instance of a Kolmogorov (metric-identification) quotient induced by a task-invisible symmetry.

8.4 Frozen-tail identity in closed form (layerwise tomography)

Fix a snapshot $\theta^\star_{>1} = (W_2^\star, b_2^\star)$ and restrict to the frozen-tail slice $\Theta^{(1,\theta^\star_{>1})} = \{(\theta_{\le 1}, \theta^\star_{>1})\}$. Then for each x the per-input restricted parameter metric satisfies the exact identity

$$g_{x,1}^{(\Theta)} = F_{1,x}^*\, g_{\theta^\star_{>1}}^{(1)}. \tag{15}$$

In this example, (15) becomes an explicit matrix formula:

$$g_{x,1}^{(\Theta)}(\theta_{\le 1}) \;\equiv\; J_h(x; \theta_{\le 1})^\top \Big[(W_2^\star)^\top \big(\mathrm{Diag}(p^\star) - p^\star (p^\star)^\top\big) W_2^\star\Big] J_h(x; \theta_{\le 1}), \tag{16}$$

where $J_h(x; \theta_{\le 1}) = \partial h(x; \theta_{\le 1})/\partial \theta_{\le 1}$ is the Jacobian of the hidden representation map and $p^\star = \mathrm{softmax}(W_2^\star h(x; \theta_{\le 1}) + b_2^\star)$. Equation (16) makes the diagnostic content transparent: the tail fixes a PSD metric on M_1 and the parameter geometry up to layer 1 is exactly its pullback through the representation map.

Remark 8.9 (Discrete symmetries). In addition to the continuous gauge symmetry above, one-hidden-layer networks also admit discrete symmetries (e.g. permutations of hidden units accompanied by corresponding permutations of columns of W_2 and entries of a_1). Such symmetries can yield disconnected components in functional fibers; patchwise metric-identification collapses connected components, while identifying discrete components typically requires an additional global quotient (cf. Remarks on discrete symmetries in Sections 5 and 6).

9 Discussion and outlook

9.1 The induced geometry as an output-anchored pullback–quotient structure

The framework fixes a canonical base geometry on the output side: the Fisher–Rao metric G_{FR} on a chosen output statistical manifold S°. For each input x, the task map $\Psi_x : \Theta \to S^\circ$ induces a pullback tensor $g_x^{(\Theta)} = \Psi_x^* G_{\mathrm{FR}}$ on parameter space, and analogously tail-indexed pullbacks on representation spaces. These pullback Fisher metrics are positive semidefinite and generically degenerate. Degeneracy is not treated as an anomaly: the null distribution $\ker g$ is the intrinsic record of task-invisible directions (equivalently, of directions in the kernel of the relevant differential). Because g is only semidefinite, it induces a pseudometric d_g, and the associated metric-identification quotient collapses zero-distance directions into an *effective* geometry. On constant-rank strata this effective space is smooth, while globally it is naturally stratified. In this sense, the "learned geometry" is the output-distinguishability geometry pulled back through the network and then quotientized by task-induced identifications.

9.2 Population versus dataset geometry

A central distinction is between population pullbacks g_p and dataset pullbacks $g_{\mathcal{D}_X}$. Population geometry reflects the task distribution, whereas dataset geometry probes it only on the empirical input support. As a consequence, $g_{\mathcal{D}_X}$ may exhibit additional null directions arising from limited coverage even when such directions are not null in g_p. Comparing kernels and the induced effective quotients therefore separates intrinsic, task-induced degeneracy from finite-sample artifacts. This viewpoint reframes observed "flat" directions as a question of identification and sampling, rather than solely a numerical question about spectra.

9.3 Layerwise localization via frozen tails

On frozen-tail slices $\Theta^{(\ell,\theta^\star_{>\ell})}$, the two-sided identity

$$g_{x,\ell}^{(\Theta)} = F_{\ell,x}^*\, g_{\theta^\star_{>\ell}}^{(\ell)}$$

links parameter-side and representation-side geometry layer by layer (Section 7). Conceptually, the tail $\theta^\star_{>\ell}$ induces a Fisher–Rao pullback metric on the representation space M_ℓ, and the prefix parameters inherit this metric through the representation map $F_{\ell,x}$. Varying ℓ provides a principled way to localize degeneracy across depth and to attribute task-induced identifications to specific depth ranges. The softmax worked example in Section 8 illustrates this logic in closed form: classical gauge redundancies become explicit null leaves and are resolved by the metric-identification quotient.

9.4 Limitations and scope

Several limitations are deliberate. First, we restrict to an open region S° where Fisher–Rao is smooth and positive definite; boundary singularities of the output manifold (e.g. near-zero class probabilities)

are not treated here. Operationally, this means we restrict to regimes where model outputs remain bounded away from the boundary (e.g. in classification, $p_\theta(k \mid x) \geq \varepsilon > 0$ for all classes k on the region of interest), so that $p_\theta(\cdot \mid x) \in S^\circ$ holds throughout the analysis. Second, for piecewise-smooth networks we interpret differentiability and constant-rank arguments patchwise on strata (Section 1.1). Third, patchwise metric identification captures connected null leaves; discrete symmetries (such as neuron permutations) generally require an additional global quotient step (cf. Theorems 5.2 and 6.5). Finally, the pullback Fisher metrics in this note are defined intrinsically from Fisher–Rao on output distributions and should not be conflated with empirical-Fisher approximations common in optimization; see Theorem 3.1.

9.5 Research directions

Natural next steps include: (i) quantitative stability of effective quotients under finite sampling and under optimization dynamics, (ii) boundary-aware extensions suited to highly confident classifiers, (iii) systematic incorporation of discrete symmetry quotients, and (iv) empirical measurement programs comparing layerwise effective dimensions and null-leaf structure across architectures and tasks.

10 Conclusion

We proposed an output-first definition of the geometry induced by a deep network by anchoring Fisher–Rao on an output statistical manifold and pulling it back to parameters and intermediate representations. Because the resulting tensors are generically degenerate, we emphasized null distributions, induced pseudometrics, and metric-identification quotients as the natural effective objects. Separating population and dataset pullbacks clarifies when observed degeneracies are intrinsic versus sample-induced. On frozen-tail slices, a two-sided identity links representation-side and parameter-side degeneracy layer by layer, enabling a layerwise localization of task-induced equivalence classes.

A Full notation table

Symbol	Meaning
$\Theta \subset \mathbb{R}^P$	parameter space; assume Θ is open (smooth manifold)
$T_\theta \Theta$	tangent space at θ
M_ℓ	representation space at layer ℓ (smooth on each activation stratum)
M_0	input space; we identify $X \equiv M_0$
$d_\ell := \dim(M_\ell)$	ambient representation dimension at layer ℓ
$(M_\ell)^N$	N-fold product representation space
S	chosen finite-dimensional manifold of output distributions
$S^\circ \subseteq S$	open region where G_{FR} is smooth and positive definite
G_{FR}	Fisher–Rao metric on S°
$d_{\mathrm{out}} := \dim(S^\circ)$	output-manifold dimension
$\Psi_x : \Theta \to S^\circ$	per-input map $\theta \mapsto p_\theta(\cdot \mid x)$
$g_x^{(\Theta)}$	per-input parameter pullback metric $\Psi_x^* G_{\mathrm{FR}}$
$p(x)$	input distribution (task measure)
$g_p^{(\Theta)}$	population parameter metric $\mathbb{E}_{x\sim p}[g_x^{(\Theta)}]$
$\mathcal{D}_X = \{x_n\}_{n=1}^N$	*input sample* (geometry object; labels not needed)
$\Psi_{\mathcal{D}_X} : \Theta \to (S^\circ)^N$	dataset map $\theta \mapsto (p_\theta(\cdot\|x_1), \ldots, p_\theta(\cdot\|x_N))$
$G_{\mathrm{FR}}^{\otimes N}$	normalized product metric $\frac{1}{N} \bigoplus_{n=1}^N G_{\mathrm{FR}}$
$g_{\mathcal{D}_X}^{(\Theta)}$	sample metric $\frac{1}{N} \sum_{n=1}^N g_{x_n}^{(\Theta)}$
$F_{\ell,x} : \Theta \to M_\ell$	layer-ℓ representation map $\theta \mapsto h_\ell(x;\theta)$
$\theta_{\le \ell}, \theta_{>\ell}$	layerwise parameter split
$\Phi_{\ell,\theta_{>\ell}} : M_\ell \to S^\circ$	representation-to-output map induced by tail parameters $\theta_{>\ell}$
$g_{\theta_{>\ell}}^{(\ell)}$	tail-indexed representation metric $(\Phi_{\ell,\theta_{>\ell}})^* G_{\mathrm{FR}}$
$\Phi_\ell^{\theta^\star_{>\ell}}$	frozen-tail map $\Phi_{\ell,\theta^\star_{>\ell}}$
$g_{\theta^\star_{>\ell}}^{(\ell)}$	frozen-tail representation metric $(\Phi_\ell^{\theta^\star_{>\ell}})^* G_{\mathrm{FR}}$
$\Theta^{(\ell,\theta^\star_{>\ell})}$	frozen-tail slice
$\iota_{\ell,\theta^\star_{>\ell}}$	inclusion $\Theta^{(\ell,\theta^\star_{>\ell})} \hookrightarrow \Theta$

References

[1] Shun-ichi Amari. *Information Geometry and Its Applications*. Springer, 2016.

[2] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. DOI:10.1162/089976698300017746.

[3] Georgios Arvanitidis, Miguel González-Duque, Alison Pouplin, Dimitrios Kalatzis, and Søren Hauberg. Pulling back information geometry. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, *Proceedings of Machine Learning Research* 151:4872–4894, 2022.

[4] Alessandro Benfenati and Alessio Marta. A singular Riemannian geometry approach to Deep Neural Networks I. Theoretical foundations. *Neural Networks*, 158:331–343, 2023. DOI:10.1016/j.neunet.2022.11.022.

[5] Taco S. Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the Icosahedral CNN. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, *Proceedings of Machine Learning Research* 97:1321–1330, 2019.

[6] Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical Fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 4158–4169, 2019. arXiv:1905.12558.

[7] Teemu Pirttimäki. A survey of Kolmogorov quotients. *arXiv preprint arXiv:1905.01157*, 2019.

[8] James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, *Proceedings of Machine Learning Research* 37:2408–2417, 2015.

[9] Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009.

[10] Susan Wei, Daniel Murfet, Mingming Gong, Hui Li, Jesse Gell-Redman, and Thomas Quella. Deep learning is singular, and that's good. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10473–10486, 2023. DOI:10.1109/TNNLS.2022.3167409.