Representation Theory of Parametrized Neural Models on Spaces

Statistical families, equivariant layers, and homotopy-coherent models

November 2026

Abstract. A parametrized network over a space is a family of evaluation maps. We represent its realizations and reparametrizations by a slice over an internal mapping object, prove base change and descent, and extend the construction to stochastic outputs for a strong probability monad. On smooth charts this gives deterministic manifold networks and conditional statistical models, together with pullback Fisher geometry. For compact symmetries we classify bounded layers linear in a moving parameter by multiplicity-space intertwiners. The classification exhibits frequency shifts that cannot occur for a fixed equivariant layer. For parameters with local symmetries we retain stabilizers and transition data in a quotient stack. A neural decoder on a twisted real line illustrates why global probability laws can lack a global smooth weight. Mapping spectra supply a stable version of equivariance with higher intertwiners.

Keywords. Parametrized networks, representation theory, equivariance, Fisher information, descent, quotient stacks.

1 From statistical families to representations

A supervised model commonly starts with a parameter manifold P, a feature space X, and a map $f(p, x)$ valued in labels, predictions, or probability distributions. A convolutional network has additional structure. Features may live in bundles on an observation space, and a symmetry can transport both features and weights. Two local weights can also present the same model via a change of frame. These cases require tracking evaluation, transport, and equivalences of presentations.

Representation Theory in Higher Categories develops ordinary and coherent group actions, induced feature spaces, and equivariant convolutions [2, Chs. 17, 20, 27–28]. We start with its idea of encoding actions by functors and relate evaluation, symmetry, and descent. Evaluation records what a parameter computes. Representation theory controls how evaluations transform. Descent determines when local evaluations and parameter presentations assemble across a base. Theorem 2.2 gives the general comparison. Theorem 5.1 computes the moving-parameter part in a compact linear setting.

Parametrized composition and supervised learning have categorical formulations through learners, lenses, and reverse differentiation [6, 5]. Markov categories organize channels and sufficient statistics [7], while equivariant convolution on homogeneous spaces is described by induced representations [4]. Here we apply representation constraints to the evaluation families themselves. This permits the same typing of a manifold layer, a statistical decoder, and a locally presented system of layers.

2 Relative parametrized models

Let $\mathcal{E}$ be an ∞-topos and B an object of $\mathcal{E}$. Possible choices are the ∞-category $\mathcal{S}$ of spaces and the ∞-topos of sheaves of spaces on a smooth site. Work first in $\mathcal{E}_{/B}$. For X, Y over B, write $[X, Y]_B$ for their exponential in that slice. Its evaluation map is $[X, Y]_B \times_B X \to Y$ over B. A choice of site specifies which families of sections are admitted. Smooth manifolds will be embedded by their smooth functors of points.

Definition 2.1 (A parametrized layer)**.** A layer from X to Y over B consists of $P \to B$ and an evaluation $f : P \times_B X \to Y$. A map to (Q, g) is a morphism $u : P \to Q$ over B with a specified homotopy $f \simeq g(u \times_B \mathrm{id}_X)$ over B. Higher morphisms are the compatibilities in the mapping space of the resulting slice.

Write $r_f : P \to [X, Y]_B$ for the transpose of f. Given a second layer $(Q, g) : Y \to Z$, the series composite has parameter $P \times_B Q$ and evaluation $(p, q, x) \mapsto g(q, f(p, x))$. Here and below parametric formulas describe generalized points in the slice. A shared parameter P can be imposed by reparametrization along the diagonal $P \to P \times_B P$. More general tying is expressed by a map $P \to \prod_{i,B} P_i$ into the product of layer parameters.

Theorem 2.2 (Relative representation and descent)**.** *The ∞-category of layers from X to Y is $\mathcal{E}_{/[X,Y]_B}$. On these categories, series composition is induced by internal composition of exponentials and respects reparametrizations. The construction commutes with pullback in B. For an effective epimorphism $U \to B$ with Čech nerve $U_\bullet$, the global layer category is equivalent to*

$$\lim_{[n]\in\mathbf{\Delta}} \mathcal{E}_{/[X_n,Y_n]_{U_n}}, \qquad X_n = X \times_B U_n, \quad Y_n = Y \times_B U_n. \tag{1}$$

$$\begin{array}{ccc} P \times_B X & \xrightarrow{f} & Y \\ {\scriptstyle u \times_B \mathrm{id}_X}\downarrow & & \| \\ Q \times_B X & \xrightarrow[g]{} & Y \end{array}$$

Figure 1: A reparametrization includes a commuting comparison of evaluations. In an ∞-topos the comparison is a specified homotopy.

Proof. Currying yields the equivalence of mapping spaces

$$\mathrm{Map}_B(P \times_B X, Y) \simeq \mathrm{Map}_B(P, [X, Y]_B).$$

It is natural in P. The two definitions of a morphism of layers therefore give the same homotopy fiber of a map of mapping spaces, including its higher simplices. The relative double slice $(\mathcal{E}_{/B})_{/[X,Y]_B}$ is $\mathcal{E}_{/[X,Y]_B}$.

Evaluation in series transposes to a morphism

$$[X, Y]_B \times_B [Y, Z]_B \longrightarrow [X, Z]_B.$$

It sends the classifying maps of the two layers to the classifying map of their composite. The identity is classified by the transpose of id_X with parameter B. The associativity and unit compatibilities follow by uncurrying the iterated compositions. Both ways to combine a string yield evaluation in the same order on the product of its parameters. All product comparisons come from their limit universal properties.

Let $a : C \to B$. If $T \to C$, the left adjoint $a_!$ to a^* is composition of structure maps. Using the pullback identification $a_!(T \times_C a^*X) \simeq (a_!T) \times_B X$ gives

$$\begin{aligned} \mathrm{Map}_C(T, a^*[X, Y]_B) &\simeq \mathrm{Map}_B(a_!T, [X, Y]_B) \\ &\simeq \mathrm{Map}_B(a_!T \times_B X, Y) \\ &\simeq \mathrm{Map}_C(T \times_C a^*X, a^*Y). \end{aligned}$$

By Yoneda these equivalences identify $a^*[X, Y]_B$ with $[a^*X, a^*Y]_C$, preserving evaluation. Pullback also preserves the products that define series composition.

Finally, the effective epimorphism is the colimit of its Čech nerve. Colimits in an ∞-topos satisfy descent for slices [11, §6.1.3]. Hence $\mathcal{E}_{/B}$ is the limit of the categories $\mathcal{E}_{/U_n}$. Taking the slice at the compatible exponential commutes with this limit, since a slice is the category of arrows with fixed target. Base change identifies the local targets with the ones in (1). □

A point of $[X, Y]_B$ over b is one evaluation on the fibers. The classifying map retains a distinction between a space of these evaluations and a space of weights mapping into it. An architecture selects a classifying map, for instance by composing affine maps, an equivariant convolution, and a nonlinearity. The ambient exponential also includes maps that the selected architecture cannot realize.

2.1 Transport and actions

In spaces, straightening identifies $\mathcal{S}_{/B}$ with $\mathrm{Fun}(B, \mathcal{S})$, where the base B is regarded as an ∞-groupoid. Transport acts by equivalences and the slice exponential is computed fiberwise [11, §3.2]. For a path $\gamma : b \to c$ with transports t_X, t_Y, exponentials transport by $h \mapsto t_Y h t_X^{-1}$. A classifying map is consequently a coherent family of maps $P_b \to [X_b, Y_b]$ commuting with transport. For an ordinary action of an abstract group G on P, X, Y, the resulting comparison reads

$$f(gp, gx) = gf(p, x). \tag{2}$$

It is a condition on the entire family. At a fixed p, it forces equivariance under the stabilizer G_p. In a smooth site one can instead use the classifying stack of a Lie group. Sections of that stack are torsors with smooth local trivializations, and (2) then holds for smooth test families. The homotopy type of a smooth base retains path transport and may identify smoothly distinct data.

2.2 Realization on a quotient

Suppose a group object G acts on P and a realization $r : P \to M$ is coherently invariant with trivial action on M. Here M can itself be a space of functions or probability laws. The action groupoid has simplicial levels $P \times G^n$. Its colimit is $P//G$, and the colimit universal property gives

$$\mathrm{Map}(P//G, M) \simeq \mathrm{Tot}\, \mathrm{Map}(P \times G^{\bullet}, M).$$

Thus r factors through $P//G$ exactly when equipped with descent, whose first comparison identifies $r(p)$ and $r(gp)$. Equivalences between local parameter presentations are part of a family in the quotient. Over $\mathcal{S}$, the fiber of $P//G \to BG$ over the base point is the parameter space P, with transport the original action. For smooth P and Lie G, the same construction in sheaves gives the smooth quotient stack. The example in Section 6 distinguishes the quotient presentations from their common observable laws.

2.3 Stochastic evaluation

In a cartesian closed category, a strong monad T has a natural strength $\mathrm{st}_{A,Y} : A \times TY \to T(A \times Y)$ compatible with products, unit η, and multiplication μ. When T is a probability monad, the layer $P \times X \to TY$ assigns a probability-valued prediction. Currying represents it by $P \to [X, TY]$. For $f : P \times X \to TY$ and $g : Q \times Y \to TZ$, their composite with parameter $P \times Q$ is

$$\begin{aligned} P \times Q \times X &\longrightarrow Q \times TY \\ &\xrightarrow{\mathrm{st}} T(Q \times Y) \xrightarrow{Tg} T^2 Z \xrightarrow{\mu_Z} TZ, \end{aligned}$$

with first map $(p, q, x) \mapsto (q, f(p, x))$. For a formal check, fix a context A and form the category whose arrows

$X \to Y$ are maps $A \times X \to Y$. It is the co-Kleisli category of the comonad $A \times (-)$ with diagonal and projection. Strength is a mixed distributive law $A \times T(-) \to T(A \times (-))$ for this comonad and the monad T. Its compatibility with the diagonal and projection is product coherence and naturality in A. Monad unit and multiplication compatibility are the strength axioms. Thus T acts as a monad on the category with context A. Kleisli composition for that monad extends a map $h : A \times Y \to TZ$ using strength followed by Th and μ_Z. Taking $A = P \times Q$ and restricting the layers along its two projections gives the displayed formula. Context maps preserve the construction by naturality of strength. A compatible system of monads on slices gives the same construction over B when their strengths and monad operations commute with base change.

For instance, the finitely supported distribution monad on sets gives finite channels by restricting the feature and output sets to be finite. If p, q are fixed, the composite has law

$$k_{p,q}(z \mid x) = \sum_y g_q(z \mid y) f_p(y \mid x).$$

Dirac laws embed deterministic evaluations. Repeated sampling can share an unknown parameter by pulling a product family back along its diagonal. For a coin of bias $1/5$ or $4/5$ with equal prior probability, shared-parameter sampling gives probability $17/50$ of two successes. Independently drawing the parameter for each observation gives $1/4$. The single-observation law therefore needs to be supplemented by a parameter sharing convention.

For smooth conditional models, the probability-valued evaluator can be defined on a represented family of probability laws. The next section uses finite outcome laws with positive probabilities, where the open simplex is a manifold and the Fisher tensor can be computed in ordinary coordinates.

3 Statistical and manifold models

A parametrized smooth map $F : P \times X \to M$ is a manifold-valued prediction, with a family of maps $F_p : X \to M$. A deterministic network may also retain intermediate feature bundles by factoring F through smooth maps of total spaces. With an ordinary parameter manifold and no prescribed equivalences of parameters, these are ordinary smooth families. For stochastic predictions on a finite set Y, a smooth readout is a map

$$q : P \times X \longrightarrow \Delta^\circ(Y), \qquad (p, x) \longmapsto (q_i(p, x))_{i \in Y},$$

where $\Delta^\circ(Y)$ is the simplex of strictly positive probability vectors. Evaluation at p gives the conditional law of the outcome given x. Taking $X = *$ recovers the usual smooth finite statistical model. A deterministic prediction into a standard Borel space M has the associated kernel $x \mapsto \delta_{F(p,x)}$. Dirac measures on a finite set with at least two elements lie on the boundary of the simplex. Strict positivity permits differentiation of the finite log likelihood.

A stochastic decoder can be composed with a succeeding stochastic layer by summing over the intermediate outcome. If $K_{ji}(p, x)$ is the conditional probability of output j given outcome i, the new readout is $r_j = \sum_i K_{ji} q_i$. Associativity is the associativity of finite summation, and deterministic relabelings are the special case of Dirac kernels [7]. Smooth charts of P and X present these models locally: if the predictions agree on the overlaps after coordinate change, the coordinate functions q_i glue to a unique smooth readout. For a manifold target M, the same assertion uses transition maps of M on overlapping target charts.

3.1 The Fisher form and stochastic postprocessing

At a point q of the positive simplex, write u, v for tangent vectors with $\sum_i u_i = \sum_i v_i = 0$. The Fisher metric is

$$g_q(u, v) = \sum_{i \in Y} \frac{u_i v_i}{q_i}.$$

At a fixed input x, the score in direction $v \in T_p P$ is $s_v(i) = d_p \log q_i(p, x)(v)$. Its mean is zero by differentiation of $\sum_i q_i = 1$. Thus the covariance of the scores in directions v, w is

$$I_{p,x}(v, w) = \sum_{i \in Y} \frac{d_p q_i(v)\, d_p q_i(w)}{q_i(p, x)}. \tag{3}$$

Equivalently, $I_x = q_x^* g$, where $q_x(p) = q(p, x)$. For a parameter-independent Borel probability measure μ on X, the joint experiment observes both the input and the label with law $d\mu(x)\, q_i(p, x)$. Its Fisher form is $I_p^\mu = \int_X I_{p,x}\, d\mu(x)$, assuming the scores are square integrable and the experiment is differentiable under the integral. These conditions hold for a smooth positive readout when X is compact, locally in the parameter.

For a smooth reparametrization $r : P' \to P$, the conditional Fisher form of $q \circ (r \times \mathrm{id}_X)$ is $r^* I_x$ by the chain rule. This holds for I^μ as well. The identity gives the change-of-coordinates rule $I' = Dr^\top I\, Dr$ on charts. On input charts the same joint form is computed by transporting μ with the coordinates. Thus the tensor descends with the model and the chosen design measure. Its pointwise null space at x is $\ker d_p q_x$, because all $q_i > 0$. The joint form vanishes on v precisely when $d_p q_i(v)$ vanishes μ-almost everywhere for every i.

Proposition 3.1 (Finite Fisher contraction)**.** *A stochastic postprocessor K, fixed with respect to p, weakly decreases both the conditional Fisher form and its design average.*

Proof. Omit output values j that are unreachable under K. For $r_j = \sum_i K_{ji} q_i > 0$, the score of r_j in direction v is

$$d_p \log r_j(v) = \sum_i \frac{K_{ji} q_i}{r_j}\, s_v(i).$$

The weights are the conditional probabilities of i given j. Jensen's inequality bounds the square by the weighted sum of squares. Multiplying by r_j and summing over j yields $I^r_{p,x}(v,v) \le I^q_{p,x}(v,v)$. Integrating yields the claim for I^μ. At a fixed x, equality in direction v holds exactly when s_v is constant on the intermediate outcomes that contribute to each reachable output. □

The parameter-independence hypothesis keeps the post-processor from introducing its own score. Formula (3) is the finite instance of the parametrized-measure Fisher tensor [3].

Example 3.2 (Gaussian observations of a manifold family)**.** Let $m : P \times X \to \mathbb{R}^k$ be a smooth readout and let the conditional outcome have distribution $N(m(p,x), \Sigma(x))$, where $\Sigma(x)$ is positive definite and independent of p. For an observed outcome z, differentiation of the Gaussian log density at fixed x gives the score

$$d_p m(v)^\top \Sigma(x)^{-1}\big(z - m(p,x)\big).$$

The expectation of the product of two scores uses the covariance $\Sigma(x)$ and gives

$$I_{p,x}(v,w) = d_p m(v)^\top \Sigma(x)^{-1} d_p m(w).$$

Thus an immersed mean manifold at fixed x has a Riemannian metric by pullback of the Euclidean metric weighted by $\Sigma(x)^{-1}$. If the mean map has a kernel in its differential, the pullback has the same null directions. An invertible affine coordinate change of the output with linear part A replaces Σ by $A\Sigma A^\top$ and $d_p m$ by $A\, d_p m$, leaving the form invariant. For a general smooth output coordinate change, one transports the entire conditional law. The transformed density contains a Jacobian factor independent of p, so its score is the transport of the original score even when the law exits the Gaussian family.

3.2 What the Fisher form forgets

Consider a two-feature decoder on S^1, with angle t understood modulo 2π and parameters $p = (a,b,c) \in \mathbb{R}^3$. Its logit is $L_p(t) = a\cos t + bc \sin t$ and its Bernoulli success probability is $s_p(t) = (1 + e^{-L_p(t)})^{-1}$. For a tangent vector $v = (v_a, v_b, v_c)$, the score at label $y \in \{0,1\}$ is $(y - s_p(t))\, dL_p(t)(v)$. With any full-support design probability measure μ on the circle, the joint Fisher form is

$$I^\mu_p(v,v) = \int_{S^1} s_p(t)(1 - s_p(t))\, (\delta_v(t))^2\, d\mu(t),$$
$$\delta_v(t) = v_a \cos t + (c v_b + b v_c) \sin t.$$

Positivity of the Bernoulli variance and continuity in t imply that the integral vanishes exactly when δ_v vanishes identically. The independent harmonics give the two equations $v_a = 0$ and $c v_b + b v_c = 0$. The rank is two away from $b = c = 0$ and one along that locus. In particular the Fisher tensor is the pullback of a positive definite form on the two-dimensional logistic family with coordinates (a,d), along the map $(a,b,c) \mapsto (a, bc)$. The strict monotonicity of the logistic function and the independent harmonics also show that this map labels exactly the distinct output laws. This is an elementary instance of singular parametrization [13].

The multiplicative group $\mathbb{R}^\times$ acts by $\lambda \cdot (a,b,c) = (a, \lambda b, \lambda^{-1} c)$, preserving the readout. The action groupoid has stabilizer $\mathbb{R}^\times$ at $b = c = 0$ and trivial stabilizer elsewhere. Its three distinct orbits over a given prediction with $d = 0$ correspond to the origin and the two punctured coordinate axes in the (b,c)-plane. Thus the quotient groupoid of parameter symmetries contains information beyond the set of distinct output laws. Even the infinitesimal null space can include directions that change the output at higher order: along $(a, \varepsilon, \varepsilon)$ through $(a,0,0)$, $d = \varepsilon^2$ changes the prediction to second order. Consequently equality of output laws, Fisher-null directions, and parameter equivalence answer three different questions.

3.3 A manifold decoder as a statistical model

Let Z be a latent manifold with fixed probability ν and $d : P \times Z \to M$ a smooth decoder. In the absence of observation noise the law at p is $(d_p)_* \nu$. A noise channel $N : M \to Y$ then gives the observed law

$$\mathbb{P}_p(A) = \int_Z N(d_p(z), A)\, d\nu(z)$$

for measurable $A \subseteq Y$. This is ordinary composition of Markov kernels on these measurable spaces. The manifold evaluation d can be represented by the smooth slice construction of Section 2. The pushforward additionally uses the chosen law ν and noise N. If $a : Z \to Z$ preserves ν, then replacing d_p by $d_p \circ a$ leaves $\mathbb{P}_p$ unchanged, by the change-of-variables identity for pushforwards. A law that observes z and the output together can distinguish the two. Thus the observation protocol determines which parameter symmetries are invisible.

For an explicit case take $Z = S^1$ with Haar probability and the ellipse $e(t) = (a\cos t, b \sin t)$ in $\mathbb{R}^2$ for $a, b > 0$ with $a \neq b$. A hidden phase θ in $t \mapsto e(t + \theta)$ is entirely unidentifiable after integrating out t. Adding centered isotropic Gaussian noise with covariance $\sigma^2 I$ preserves this invariance. The integrated Gaussian mixture density is independent of θ, so its Fisher information in the phase direction is zero. In contrast, an output orientation in

$t \mapsto R_\theta e(t)$ rotates the signal covariance:

$$\mathrm{Cov}_\theta(Y) = \tfrac{1}{2} R_\theta \begin{pmatrix} a^2 & 0 \\ 0 & b^2 \end{pmatrix} R_\theta^\top + \sigma^2 I.$$

Because $a^2 \neq b^2$, equality of these covariances forces coincidence of the principal axes and hence equality of θ modulo π. Conversely a rotation by π is the latent shift $t \mapsto t+\pi$ on this ellipse, so orientations differing by π give the same law. A symmetry of latent coordinates and a rotation of the observed manifold therefore have distinct statistical effects even within the same two-dimensional harmonic decoder.

4 Homogeneous feature bundles

Let G be a compact Lie group, $H, K \subseteq G$ closed subgroups, and P a topological G-space of parameters. Write $X = G/H$, $Y = G/K$, and let ρ and σ be finite-dimensional unitary representations of H and K on V and W. Use $E = G \times_H V$ and $F = G \times_K W$ as input and output feature bundles. Our convention is $[bh, v] = [b, \rho(h)v]$. An input field is therefore encoded by $f : G \to V$ with $f(bh) = \rho(h)^{-1} f(b)$, and a group element g acts by left translation $(gf)(b) = f(g^{-1}b)$. Haar probability induces the invariant measures on X, Y. The fixed-parameter kernel model is the familiar homogeneous-space convolution [4, 2]; the parameter action below allows the kernel itself to move with the symmetry.

Proposition 4.1 (Kernels with moving parameters)**.** *Let $L_p(y, x) : E_x \to F_y$ be jointly continuous in the parameter and the two base variables, with diagonal covariance*

$$L_{gp}(gy, gx) = g L_p(y, x) g^{-1}.$$

Such families correspond bijectively to continuous functions $B : P \times G \to \mathrm{Hom}(V, W)$ satisfying

$$B(kp, kzh) = \sigma(k) B(p, z) \rho(h). \tag{4}$$

Here $k \in K, h \in H$. The corresponding family of bounded linear layers is

$$(T_p f)(a) = \int_G B(a^{-1}p, a^{-1}b) f(b)\, db. \tag{5}$$

It satisfies $T_{gp}(gf) = g(T_p f)$.

Proof. Evaluate L_p at the output basepoint $1K$ and write $L_p(1K, zH)[z, v] = [1, B(p, z)v]$. Changing input representative z to zh introduces $\rho(h)$. Acting by $k \in K$ fixes the output basepoint and acts there by $\sigma(k)$. Together these give (4). Conversely, define

$$L_p(aK, bH)[b, v] = [a, B(a^{-1}p, a^{-1}b)v].$$

If a, b are replaced by ak, bh, then the arguments of B are replaced by $k^{-1}a^{-1}p, k^{-1}a^{-1}bh$, and v by $\rho(h)^{-1}v$. The two associated-bundle relations and (4) cancel these changes. Continuity follows in local sections of $G \to X, Y$, and diagonal covariance follows since $(ga)^{-1}gp = a^{-1}p$. The two formulas are inverse.

Integrating L_p over X gives (5). The integrand is independent of the input representative, and the output has the required right K-transformation. For fixed p the kernel is continuous on $Y \times X$, hence square-integrable in the Hilbert–Schmidt norm on the finite-dimensional fibers. Cauchy–Schwarz gives $\|T_p\| \leq \|L_p\|_{L^2(Y \times X)}$. Joint continuity and compactness of $Y \times X$ give uniform convergence of $L_{p'}$ to L_p as $p' \to p$, so $p \mapsto T_p$ is operator-norm continuous. Changing variables $b \mapsto gb$ proves the stated operator covariance. If two continuous kernels give the same operators, testing in local frames against smooth sections supported in coordinate charts identifies their distributional pairings. The resulting equality almost everywhere extends everywhere because the measures have full support. □

When G fixes p, each T_p is an intertwiner. Equation (4) then constrains the permissible value at z to intertwiners for $K \cap zHz^{-1}$, where the input representation is transported by $k \mapsto z^{-1}kz$. This follows by setting $h = z^{-1}k^{-1}z$ for a k in the intersection. When a parameter moves, the analogous constraint uses $K_p \cap zHz^{-1}$, where K_p fixes p. Thus a parameter with a larger stabilizer can impose more equations on the kernel at an orbit representative.

Linear layers. For a fixed parameter, Peter–Weyl and compact Frobenius reciprocity decompose $L^2(E)$ as the Hilbert direct sum of $U_\lambda \otimes M_\lambda^E$, where U_λ runs through irreducibles of G and $M_\lambda^E = \mathrm{Hom}_H(U_\lambda|_H, V)$. For F use $M_\lambda^F = \mathrm{Hom}_K(U_\lambda|_K, W)$. Give the multiplicity spaces the inner products that make the isotypic decompositions unitary. Schur's lemma identifies bounded intertwiners with the blocks $1_{U_\lambda} \otimes A_\lambda$, where $A_\lambda : M_\lambda^E \to M_\lambda^F$ and $\sup_\lambda \|A_\lambda\| < \infty$. Finite spectral support gives smooth kernels by finite sums of matrix coefficients. A general bounded intertwiner has a distributional kernel by applying the Schwartz kernel theorem to $\Gamma^\infty(E) \to L^2(F) \to \Gamma^{-\infty}(F)$ [9, Thm. 1.16]. If $E = F$ and X is positive-dimensional with E of positive rank, the identity has a diagonal delta kernel. Continuous kernels on the compact product give Hilbert–Schmidt operators, whereas the identity on this infinite-dimensional space fails that condition.

Example 4.2 (A learned axis)**.** Take $P = X = Y = S^2$ with the rotation action and scalar channels. For $p \in P$ define $q_p(x) = p \cdot x$ and choose the kernel $L_p(y, x) = 3q_p(y)q_p(x)$. Rotational invariance of Haar probability on the sphere gives $\int_{S^2} x_i x_j\, dx = \delta_{ij}/3$, so $3\|q_p\|_2^2 = 1$. The layer T_p is the orthogonal projection onto $\mathbb{C}q_p$, a line of degree-one spherical harmonics. Every g carries q_p to

q_{gp}, hence $T_{gp} = gT_pg^{-1}$. A fixed-p layer commutes with rotations preserving the axis $\mathbb{R}p$. In particular $T_{-p} = T_p$, and its symmetry group in G is strictly larger than the stabilizer of p. The parameter evaluation identifies opposite poles and realizes the family of rank-one projections parameterized by $\mathbb{RP}^2$. This is a model symmetry that arises from the map that realizes the parameters.

Spin channels. For $S^2 = \mathrm{SO}(3)/\mathrm{SO}(2)$, let L_m correspond to $\rho_m(e^{it}) = e^{imt}$. The ℓ-th spherical type U_ℓ restricts to the stabilizer with weights $-\ell, \dots, \ell$, each of multiplicity one. Thus $L^2(L_m)$ contains one copy of U_ℓ for $\ell \geq |m|$. The section representations at weights 1 and -1 are therefore unitarily equivalent. The oriented complex tangent line is L_1, and L_m is its m-th tensor power, using duals for $m < 0$. The tangent Euler number 2 gives

$$c_1(L_m) = 2m\, u, \qquad \langle u, [S^2] \rangle = 1$$

[8, Ch. 3]. The two opposite channels therefore carry inequivalent complex bundles. Fiberwise nonlinear layers detect the characters directly. A continuous fiber map induced by $\phi : \mathbb{C} \to \mathbb{C}$ from L_2 to L_3 would obey $\phi(e^{2it}v) = e^{3it}\phi(v)$. Taking $t = \pi$ forces $\phi = 0$. A nonzero global intertwiner from L_2 to L_3 is obtained by retaining the common $\ell = 3$ block. The equivariance constraints on the two operations arise at different levels of the representation.

Manifolds with local frames. The kernel description also makes sense without a transitive global action. On a smooth manifold M, let $Q \to M$ be a principal H-bundle and use ρ, σ as its two feature representations. Local frames of Q identify a kernel $E_x \to F_y$ with a matrix $A(y, x)$. Replacing the frames independently by right translates by $h_x, h_y \in H$ transforms the matrix by

$$A'(y, x) = \sigma(h_y)^{-1} A(y, x) \rho(h_x).$$

Indeed, new input coordinates must first be converted to the old ones by $\rho(h_x)$, and old output coordinates to new ones by $\sigma(h_y)^{-1}$. This relation is exactly the gluing condition for a section of the bundle with fibers $\mathrm{Hom}(E_x, F_y)$ on $M \times M$. Integration against a chosen smooth positive density defines a global layer whenever the kernel satisfies the needed integrability conditions. Setting $x = y$ yields the corresponding frame change for a fiber map. A fiberwise nonlinearity can also be glued if it obeys the same change of coordinates in the sense of intertwining the two actions.

Quotients and descent. For a smooth action on a manifold P, the quotient stacks give

$$[\,(P \times G/H)/G\,] \;\simeq\; [P/H]. \tag{6}$$

An object over a test manifold consists of a right G-torsor with a G-equivariant map to $P \times G/H$, with inverse action on the target in the right-torsor convention. The inverse image of $1H$ under the second component is an H-reduction of the torsor. The first component restricts to an equivariant map from that reduction to P. Conversely, extending the H-torsor along $H \hookrightarrow G$ produces the original object and both maps. The constructions commute with pullback and identify torsor isomorphisms [2, Chs. 20, 28]. For $P = *$ this gives the smooth classifying stack BH. A representation $H \to \mathrm{U}(V)$ then gives a morphism of smooth classifying stacks and a descended feature bundle. More generally a homotopy-coherent coefficient system on an action groupoid retains the diagrams of identifications over all composable arrows. Formulas (4) and (5) illustrate the ordinary continuous case with a chosen invariant measure. The descent data determine how to transport fibers, while the measure enters the integration operation separately.

5 Harmonic decomposition with moving weights

A fixed-parameter equivariant layer can mix multiplicities of an irreducible representation but preserves its isotypic component. A moving parameter can carry the difference between the input and output representation types. This section classifies that effect for layers linear in the parameter and in the features. We work with complex Hilbert spaces. The representation W of the compact group G is finite dimensional and unitary, and H_X, H_Y have strongly continuous unitary actions. A linear filter family is a complex-linear map $A : W \to \mathcal{B}(H_X, H_Y)$, with the equivariance condition

$$A(gw) = \rho_Y(g) A(w) \rho_X(g)^{-1}. \tag{7}$$

Boundedness of the map into the operator norm is automatic for finite-dimensional W.

Let $\widehat{G}$ index the irreducible unitary representations V_π. Write M_π^X for the Hilbert multiplicity space of V_π in H_X, and likewise M_π^Y. Compact harmonic analysis gives the orthogonal decompositions

$$H_X \simeq \widehat{\bigoplus_{\pi \in \widehat{G}}} V_\pi \otimes M_\pi^X, \qquad H_Y \simeq \widehat{\bigoplus_{\tau \in \widehat{G}}} V_\tau \otimes M_\tau^Y.$$

For intertwiners $S, T : V \to H$ with common irreducible source V, we use the scalar inner product defined by $S^*T = \langle S, T \rangle \, \mathrm{id}_V$.

Theorem 5.1 (Classification of linear filter families)**.** *For each $\tau \in \widehat{G}$, set*

$$C_{\tau,\pi} = \mathrm{Hom}_G(V_\tau, W \otimes V_\pi),$$
$$N_\tau = \widehat{\bigoplus_{\pi \in \widehat{G}}} C_{\tau,\pi} \otimes M_\pi^X.$$

The families satisfying (7) are in bijection with families of bounded maps $B_\tau : N_\tau \to M_\tau^Y$ satisfying $\sup_\tau \|B_\tau\| < \infty$. The norm of the linearized operator $W \otimes H_X \to H_Y$ equals that supremum. In particular the transition from π to τ can contribute only if V_τ is a summand of $W \otimes V_\pi$.

Proof. Send $w \otimes x$ to $A(w)x$. For an orthonormal basis $(w_j)_{j=1}^d$ of W, Cauchy–Schwarz gives

$$\left\| \sum_j A(w_j)x_j \right\| \le \left(\sum_j \|A(w_j)\|^2 \right)^{1/2} \left(\sum_j \|x_j\|^2 \right)^{1/2}.$$

Thus linearization extends to the Hilbert tensor product. Conversely, a bounded operator $L : W \otimes H_X \to H_Y$ gives $A(w)x = L(w \otimes x)$. Applying (7) to a decomposable tensor is precisely the intertwining condition for L with the diagonal action on its source.

The compact isotypic decomposition follows from complete reducibility of continuous unitary representations [2, Chs. 2–4]. Tensoring with W and decomposing each finite-dimensional $W \otimes V_\pi$ gives

$$W \otimes H_X \simeq \widehat{\bigoplus_{\tau \in \widehat{G}}} V_\tau \otimes N_\tau.$$

Evaluation $V_\tau \otimes C_{\tau,\pi} \to W \otimes V_\pi$ identifies the image with the τ-isotypic part: orthogonality of the summands is Schur's lemma and the chosen Hilbert structure on $C_{\tau,\pi}$. Schur's lemma applied using multiplicity coordinates then forces L to have block $\mathrm{id}_{V_\tau} \otimes B_\tau$ and zero blocks between distinct types. The norm of each diagonal block is $\|B_\tau\|$. Conversely the Hilbert sum of such blocks is bounded with norm their supremum, and commutes with G on each block. □

With W trivial of dimension one this gives the usual classification of a bounded intertwiner. A trivial parameter representation of larger dimension adds multiplicity but preserves the same restriction on which types are allowed. A nontrivial W changes that restriction. In a network with two successive linear parameter families, the evaluation is bilinear in their weights. It linearizes in $W_2 \otimes W_1$, with its diagonal action. A path of intermediate irreducible types is allowed when each step obeys the tensor multiplicity condition. Its actual linear map is the composition dictated by the original layer maps. Not every block of a free $W_2 \otimes W_1$ model factors through a prescribed architecture.

Example 5.2 (A learned Fourier shift)**.** Let $G = S^1$ act on $L^2(S^1)$ by translation $(\rho(\theta)x)(t) = x(t-\theta)$. Let $W = \mathbb{C}$ transform by the character $e^{im\theta}$ for a nonzero integer m. Then

$$(A(w)x)(t) = w\, e^{-imt} x(t)$$

has norm $|w|$ and satisfies (7). For the Fourier vector $e_n(t) = e^{int}$, it maps e_n to we_{n-m}. Indeed, the representation character on $w \otimes e_n$ is $e^{-i(n-m)\theta}$, which is the character on e_{n-m}. This is the only possible output frequency for input e_n in any such family with scalar weight m. More generally $A(w)e_n = wa_n e_{n-m}$ classifies these families by bounded sequences $(a_n)_{n\in\mathbb{Z}}$. If w is frozen at a nonzero value without transforming, the displayed operator fails to commute with rotations that act nontrivially on that weight.

Pointwise nonlinearities must preserve equivariance at the feature representations on which they operate. For a complex line of a circle weight, the radial map $z \mapsto z\,\psi(|z|^2)$ for continuous ψ is an equivariant activation. A scalar real activation applied separately to real and imaginary parts need not commute with rotations. For instance, coordinatewise rectification fails already at rotation by π. Nonlinearities can mix Fourier modes even though the *whole* model respects (2). The theorem describes the linear layers and their allowed type transitions, rather than a Fourier diagonalization of the nonlinear network.

6 A sign-symmetric decoder on the circle

A tied-weight scalar neural decoder gives a direct comparison of observable predictions, parameter coordinates, and descent data. Let $G = \{1, -1\}$ act on a real weight by $w \mapsto -w$. For input $x \in \mathbb{R}$, define the hidden feature and Bernoulli decoder by

$$h_w(x) = \tanh(wx), \qquad \ell_w(x) = wh_w(x), \quad (8)$$
$$K_w(1 \mid x) = \frac{1}{1 + \exp(-\ell_w(x))},$$
$$K_w(0 \mid x) = 1 - K_w(1 \mid x).$$

Under a sign change the hidden feature changes sign and the tied readout changes sign, so the logit and the two probabilities are invariant. The statistical model is a Markov kernel with a finite output space. Its local parameters encode a choice of frame for the one-dimensional hidden feature.

Proposition 6.1 (Observational equivalence of scalar weights)**.** *Two weights u, w give the same kernel for every real input x if and only if $u = \pm w$. The gauge stabilizer of a weight is G at $w = 0$ and trivial elsewhere.*

Proof. Equality of probabilities gives equality of logits by invertibility of the logistic function. Differentiation in the input gives $\ell_w'(0) = w^2$, so equality implies $u^2 = w^2$. The sign invariance of (8) gives the converse. A sign fixes w precisely when its product with w is w. □

6.1 Descent of weights and observables

For a smooth base B, a family in the quotient stack $[\mathbb{R}/G]$ consists of a principal G-bundle $P \to B$ and a smooth equivariant function $w : P \to \mathbb{R}$. Morphisms over B are

principal-bundle isomorphisms compatible with w. This is the torsor description of a quotient stack, using open covers and local smooth trivializations in this setting [12, cf. the algebraic version in Tag 04UV]. Equivalently, choose local frames over an open cover U_i and write the coordinates as $w_i : U_i \to \mathbb{R}$. The locally constant signs ε_{ij} satisfy

$$\varepsilon_{ij}\varepsilon_{jk} = \varepsilon_{ik}, \qquad w_i = \varepsilon_{ij} w_j.$$

Changing frames acts by a family of local signs $\alpha_i : U_i \to G$, replacing these data by $\alpha_i \varepsilon_{ij} \alpha_j^{-1}$ and $\alpha_i w_i$. The same signs glue the hidden features h_{w_i}, and the signs cancel in $w_i h_{w_i}$. Hence the observable kernel is a single smooth family $K : B \times \mathbb{R} \to \mathrm{Prob}(\{0,1\})$. The invariant function $r(b) = w(p)^2$, for $p \in P_b$, can be recovered from K by differentiating the logit in x at 0. Both constructions commute with pullback along smooth base maps. They give maps from the groupoid of families to the set of smooth nonnegative functions and then to the set of observational laws. For a fixed base, the second map is injective on the functions obtained from families by Proposition 6.1. The first map can identify nonisomorphic families, as the zero example below shows.

6.2 Monodromy in the zero fiber

Over $S^1 = \mathbb{R}/(2\pi\mathbb{Z})$, take the connected double cover P_- with total space $\mathbb{R}/(4\pi\mathbb{Z})$ and deck transformation $\theta \mapsto \theta + 2\pi$. The associated sign line is the Möbius bundle [8, §1.1]. Transport once around the circle switches sheets. For comparison, the trivial double cover $P_+ = S^1 \times G$ preserves each sheet. The connectedness of P_- distinguishes it from P_+. Both admit the zero equivariant weight, and both resulting decoders output probability $1/2$ at every base point and input.

Proposition 6.2 (Families over the zero observable)**.** *The groupoid of smooth families on S^1 with constant observable $K(1 \mid x) = 1/2$ is equivalent to $BG \amalg BG$.*

Proof. The recovery formula for r forces the equivariant function w to be zero, so the whole groupoid consists of G-torsors and their isomorphisms. Cut the circle at one point and trivialize a torsor on the intervening interval. The two ends are glued by one element of G acting by multiplication, and changing the trivialization conjugates that element. Here G is abelian, so its two elements determine exactly the trivial and connected double covers. A torsor automorphism is multiplication by a locally constant sign on the base. The circle is connected, so each object has automorphism group G. □

At a single base point the corresponding fiber is BG. The extra isomorphism class over the circle comes from sign monodromy. The statistic r identifies orbits at each point, while the stack retains both their automorphisms and their compatibility around a loop.

6.3 A family with local weights

On P_-, the smooth function $w([\theta]) = \cos(\theta/2)$ changes sign under the deck transformation, so it is an equivariant weight. The descended family has

$$r([\theta]) = \cos^2(\theta/2) = \tfrac{1}{2}(1 + \cos\theta).$$

Its logits and probabilities are smooth by the local gluing above. This family cannot be represented by a single smooth weight on the circle in the same architecture. Indeed, any such weight $v : S^1 \to \mathbb{R}$ must satisfy $v^2 = r$ by Proposition 6.1. The sole zero of r is at $[\pi]$, where

$$r([\pi + t]) = \tfrac{1}{4}t^2 + O(t^4).$$

Thus smoothness forces $v'([\pi])^2 = 1/4$. A smooth weight with this derivative changes sign at its sole zero. The complement of that zero in the circle is connected, and a continuous nonvanishing real function on it has constant sign, a contradiction. The two chartwise signed weights do define the model in the quotient stack, with the nontrivial transition sign recording the absence of a global frame. The absolute value of the local weight yields a continuous scalar lift, which shows the role of the smooth category in the obstruction.

Information geometry is sensitive to the parameter chart at the fixed point. The Fisher information of one Bernoulli observation at fixed x is $K_w(1 \mid x)K_w(0 \mid x)(\partial_w \ell_w(x))^2$; at $w = 0$ it vanishes because $\partial_w \ell_w(x) = 0$. For fixed x, the expansion

$$\ell_w(x) = w^2 x - \tfrac{1}{3} w^4 x^3 + O(w^6)$$

gives $I_w(x) = x^2 w^2 + O(w^4)$. In the nonnegative coordinate $r = w^2$, the same logit starts as $rx - \frac{1}{3} r^2 x^3$. Its derivative from the right gives Fisher information $I_r(x) = x^2/4$ at $r = 0$, which is positive for $x \neq 0$. Away from zero the parameters satisfy the pullback identity $I_w(x) = 4w^2 I_r(x)$, with the right-hand information evaluated at $r = w^2$. This extends to zero by the displayed expansions. The sign automorphism at zero, the monodromy of P_-, and the Fisher degeneracy measure three different features of the same family. In particular, the finite gauge group has discrete orbits, while the map from smooth weights to observable laws loses its entire first derivative at the origin.

7 Equivariant comparison and stable features

At the level of an ∞-category, an action of a discrete group G on features takes the form of a functor from the groupoid BG. A layer carrying symmetry specifies a natural transformation from its source action to its target action. If its parameter and feature objects can be combined by a tensor product, the source is the corresponding diagonal action.

For a concrete stable target, take $D(k)$ for a commutative ring k. A linear layer from X to Y with linear parameter W is now a map $W \otimes_k^L X \to Y$. A coherent G-action on these objects equips the ordinary mapping spectrum $\mathrm{map}_{D(k)}(W \otimes_k^L X, Y)$ with conjugation transport. In a strict chain model, an element g sends φ to $\rho_Y(g)\varphi(\rho_W(g^{-1}) \otimes \rho_X(g^{-1}))$. The mapping spectrum of equivariant layers is

$$\begin{aligned}\mathrm{map}_{\mathrm{Fun}(BG,D(k))}(W \otimes_k^L X, Y)\\ \simeq \mathrm{map}_{D(k)}(W \otimes_k^L X, Y)^{hG}.\end{aligned} \tag{9}$$

The superscript hG denotes the limit of a BG-action. The formula also holds without a parameter, by taking the trivial tensor unit $W = k$.

To see (9), the homotopy end for natural transformations resolves a transformation along strings of group arrows. In cosimplicial degree n the term is a product over G^n of copies of the underlying mapping spectrum. Transporting the source of each component along its associated group arrows converts the two outer cofaces into conjugation at one end and deletion at the other. The inner cofaces compose successive arrows. This is the cosimplicial replacement for the homotopy limit on the right-hand side. Taking the totalization imposes all the coherences of an equivariant layer [2, Chs. 17–18].

The formula compares intertwiners between representations in a stable setting. It can model layers between complexes of features, for example when grading records features and relations between them. It introduces no probability law by itself. A stochastic decoder still requires its own evaluation and law.

Example 7.1 (A split underlying comparison with an equivariant obstruction)**.** Let $G = C_2$ and $k = \mathbb{F}_2$. Give k the trivial action, and let $R = k[C_2] \simeq k[u]/(u^2)$ with $u = g - 1$. The exact sequence of R-modules

$$0 \longrightarrow k \xrightarrow{1 \mapsto u} R \xrightarrow{\varepsilon} k \longrightarrow 0$$

splits over k. An R-linear splitting would choose a lift of 1 in R killed by u. Every such lift is $1 + au$, and $u(1 + au) = u \neq 0$. Hence the sequence gives a nonzero class of equivariant maps $k \to k[1]$, becoming zero when the action is forgotten. The free resolution obtained by repeating multiplication by u shows $\mathrm{Ext}_R^j(k, k) \simeq k$ for every $j \geq 0$. This gives the positive cohomological degrees in the homotopy fixed point spectrum of the endomorphism spectrum of k.

The regular representation in the example and the trivial module are objects of the same stable coefficient category. To identify the action category with derived $k[G]$-modules, evaluation at the unique object of BG is conservative and preserves colimits. Its left adjoint is the free action, whose underlying monad is $k[G] \otimes_k^L (-)$ with multiplication from G. The same monad comes from extension and restriction of scalars for $k \to k[G]$. Monadicity gives the identification [10, §4.7.3]. Over a ring where $|G|$ is a unit, the averaging idempotent makes the trivial module projective and the higher Ext groups in this example vanish. This clarifies the relation with classical averaging. With real or complex Hilbert features and continuous compact actions, averaging and decomposition follow instead from Haar integration under the analytic hypotheses of Section 5.

8 Losses, derivatives, and symmetry

The classifying map of a layer can be used for optimization when the parameter and feature objects carry differential structure. Work in smooth manifolds, or use smooth charts of bundles over a common base. Let $h = f(p, x)$, $y = g(q, h)$, and let a smooth loss have output covector $\lambda_y = d_y\ell$. The partial reverse derivatives of the two-layer realization are

$$\begin{aligned}\lambda_q &= (D_q g)^* \lambda_y, \quad \lambda_h = (D_h g)^* \lambda_y,\\ \lambda_p &= (D_p f)^* \lambda_h, \quad \lambda_x = (D_x f)^* \lambda_h.\end{aligned}$$

The dual maps in this display are cotangent pullbacks, with no metric identification. The chain rule proves that they glue across smooth changes of charts, using the usual pullback of a loss along the realization. Under a tying map $u : P \to \prod_i P_i$ the resulting parameter covector is $(Du)^*(\lambda_{p_1}, \ldots, \lambda_{p_n})$. In particular, tying two copies by a diagonal adds their covectors at the shared point. The same map specifies sharing in the forward composite of Section 2. This is the cotangent form of the compatibility exploited by categorical formulations of backpropagation [6, 5].

A scalar expected loss can be formed from the conditional laws of Section 3, using a finite dataset or a distribution for which integration and differentiation are justified. Gradient vectors additionally require a positive metric or another update rule. For example, take a G-invariant statistical realization $r : P \to \mathcal{M}$, where $\mathcal{M}$ is a smooth space of laws with Fisher metric $F_{\mathcal{M}}$. The pullback $F_P = r^* F_{\mathcal{M}}$ is positive semidefinite. It annihilates vectors tangent to G-orbits and, more generally, $\ker Dr$, which can be larger than the orbit tangent space. The degeneracies in Sections 3 and 6 arise even when the relevant symmetry group at the parameter is discrete or has tangent orbit of smaller dimension. The inverse-Fisher natural gradient is defined where the chosen statistical coordinates have nondegenerate information [1].

Proposition 8.1 (An equivariant gradient field)**.** *Suppose a compact Lie group G acts smoothly on a finite-dimensional, paracompact parameter manifold P and the smooth scalar loss L is invariant. If F_P is a smooth invariant positive semidefinite tensor, there exists a G-invariant Riemannian metric j such that for every $\lambda > 0$*

the gradient of L with respect to $g_\lambda = F_P + \lambda j$ is equivariant. Its local flow is equivariant. If the action is free, this field is horizontal for $P \to P/G$ in g_λ and projects to the gradient of the descended loss for the induced quotient metric.

Proof. Choose any Riemannian metric and average its pullbacks using normalized Haar measure. Integration of smooth functions over a compact Lie group gives smoothness. Positive definiteness follows by averaging positive values on a nonzero tangent vector. Invariance follows by the right invariance of compact Haar measure. This yields j, and the sum g_λ remains invariant and positive definite.

For each $a \in G$, differentiate $L \circ a = L$ and apply metric invariance. Nondegeneracy of g_λ then gives $Da \operatorname{grad} L(p) = \operatorname{grad} L(ap)$. Uniqueness for local ordinary differential equations supplies the flow compatibility. In the free case the compact action is proper, so $P \to P/G$ is a smooth principal bundle. The loss differential vanishes on vertical vectors. Hence $\operatorname{grad} L$ is horizontal, and the invariant horizontal metric induces a quotient metric by the differential of the projection. Pairing the projected field with arbitrary quotient tangent vectors by horizontal lift gives the gradient identity on P/G. □

For a nonfree action the equivariant evolution carries presentations to presentations and induces local maps of the quotient stack on the invariant domains of the flow. For the scalar deck-symmetric model, a metric can make optimization well defined at the origin, but the first derivative of a smooth loss of its law still vanishes there. Regularization changes the identification of covectors with vectors. It does not enlarge the derivative of the realization.

9 Conclusion

Over an ordinary smooth base the construction recovers smooth networks and statistical families, with cotangent composition and Fisher geometry when present. Over a homogeneous feature space it recovers induced bundles and intertwining layers. Allowing the weights to move contributes additional harmonic types, determined exactly by the tensor multiplicities in Theorem 5.1. These mechanisms are compatible because composition and families are represented in the same slices.

The higher model distinguishes observable equality from equivalence of parameter presentations. Descent records the transition arrows and their coherence, while homotopy fixed points describe the comparison spaces for stable features. The circle decoder makes that distinction visible without changing the input or output law. For each specialization the relevant structures enter explicitly: a smooth site for smooth realizations, a probability interpretation for sampling, and a compact unitary action for the Hilbert decomposition.

References

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

[2] Anonymous. *Representation Theory in Higher Categories: From Compact Harmonic Analysis to Coherent Group Actions*. Higher Categorical Semantics, 2026. Chapters 2–4, 17–18, 20, and 27–28. Published edition.

[3] N. Ay, J. Jost, H. V. Lê, and L. Schwachhöfer. Information geometry and sufficient statistics. *Probability Theory and Related Fields* 162, 327–364, 2015. DOI.

[4] T. S. Cohen, M. Geiger, and M. Weiler. A general theory of equivariant CNNs on homogeneous spaces. *Advances in Neural Information Processing Systems* 32, 2019. Proceedings.

[5] G. S. H. Cruttwell, B. Gavranović, N. Ghani, P. Wilson, and F. Zanasi. Categorical foundations of gradient-based learning. In *Programming Languages and Systems*, ESOP 2022. DOI.

[6] B. Fong, D. I. Spivak, and R. Tuyéras. Backprop as functor: A compositional perspective on supervised learning. In *Proceedings of LICS*, 2019. Author manuscript.

[7] T. Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics* 370, 107239, 2020. DOI.

[8] A. Hatcher. *Vector Bundles and K-Theory*. Version 2.2, 2017. Author's notes.

[9] C. Kottke. *Linear Analysis on Manifolds*. Notes for Math 7376, 2016. Lecture notes.

[10] J. Lurie. *Higher Algebra*. Manuscript, 2017. Author's manuscript.

[11] J. Lurie. *Higher Topos Theory*. Annals of Mathematics Studies 170. Princeton University Press, 2009. Author's copy.

[12] The Stacks Project Authors. *The Stacks Project*, section on quotients by group actions. Tag 04UV.

[13] S. Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009. DOI.