Topos Actions in Machine Learning and Artificial Intelligence

Prior work and models for local symmetry

November 2026

Abstract. We review uses of topoi, group actions, and descent in machine learning and artificial intelligence. Existing work models neural networks by presheaves and stacks, learns automata with nominal symmetries, and builds architectures from cellular sheaves or groupoid-equivariant kernels. The literature identified here gives little direct treatment of learning with a monoidal Grothendieck topos acting by geometric morphisms. We develop two applications in that setting and its internal counterpart. A central action gives a quotient-type criterion for nuisance-invariant layers and distinguishes quotient features from linear averaging. A nonconstant group sheaf over the circle yields learnable filters whose coefficient bundle has a Möbius summand. Descent determines which parameter choices and update rules give consistent models across contexts. The results are structural and are illustrated by exact learning examples.

Keywords. Topos actions, equivariant learning, descent, sheaf neural networks, nominal automata, nuisance invariance.

1 Which notion of action enters learning?

Equivariance in machine learning usually begins with one group, its representations on data and features, and a trainable map that respects those representations. Three developments motivate a topos description. A local model may depend on a varying context. The observations and the symmetry groups themselves can have nontrivial transition maps. An action can also relate different systems of observations by geometric reindexing. Keeping those distinctions explicit lets one say which facts are architectural and which are consequences of descent.

In the geometric definition of [3], the parameter $\mathcal{T}$ is a monoidal Grothendieck topos. Its action on $\mathcal{X}$ is a geometric morphism $a : \mathcal{T} \times \mathcal{X} \to \mathcal{X}$ with coherent multiplication and unit. Its inverse image a^* pulls an observation on $\mathcal{X}$ back to an observation on the product with the symmetry parameter. A second construction is a weak action of an ordinary group by geometric autoequivalences. It is realized by taking $\mathcal{T}$ to be the sheaf topos of the discrete locale of group elements. Finally, a group object G in a fixed topos $\mathcal{E}$ acts on objects of that topos. The topos $B_{\mathcal{E}}G$ of internal G-objects classifies compatible families with those symmetries. A locally constant sheaf of groups illustrates why its group of global sections may be insufficient to specify a model.

For a finite group in sets this distinction is visible even before learning. The parameter topos Set^G indexed by the underlying discrete set of elements has points indexed by G. The topos of G-sets, denoted $\mathbf{B}G$, classifies G-torsors. Its points form a groupoid with one isomorphism class and automorphism group G. Group laws on the set of elements induce geometric operations on the parameter topos. A Cartesian monoidal law on $\mathbf{B}G$ requires G to be abelian, as proved in [3]. Below we use an actual action of $\mathbf{B}A$ for abelian A and compute what it says about learning a nuisance-invariant readout.

1.1 External symmetry and spaces of layers

Let an ordinary group H act by geometric autoequivalences on a Grothendieck topos $\mathcal{E}$. Write T_h for the inverse image indexed by the inverse element, so that $T_h T_k \simeq T_{hk}$. An equivariant object X has structure $\theta_{h,X} : T_h X \simeq X$. The equivariant topos $\mathcal{E}^H$ is the geometric quotient. For equivariant objects X, Y, the underlying exponential Y^X acquires the conjugation action. The equivalence $T_h(Y^X) \simeq (T_h Y)^{T_h X}$ is followed by precomposition with $\theta_{h,X}^{-1}$ and postcomposition with $\theta_{h,Y}$. The two coherence laws follow by composing these maps on either side of an evaluation. Consequently a layer $P \times X \to Y$ is equivariant precisely when its transpose $P \to Y^X$ is equivariant. The usual equation $F(hp, hx) = hF(p, x)$ is its case in sets. It is an equation about the family. At a fixed parameter the equation applies to its stabilizer.

This elementary result makes the applications testable. An architecture specifies a parameter object, a realization map, and which symmetries it respects. A loss and a learning rule require their own compatibility with those data. Sections 2–4 review relevant work. The later sections develop descent for layer evaluation and compute two specializations where the symmetry changes what can be learned or represented.

2 Topos-theoretic precedents in learning

The closest precedents use three different levels of structure. Some place the states and maps of a learner in a topos of equivariant objects. Others use a classifying topos to organize possible interpretations and changes of context. A literal action in the sense considered here

gives a monoidal Grothendieck topos $\mathcal{T}$ a geometric action $\mathcal{T} \times \mathcal{X} \to \mathcal{X}$ on another topos. One can test this last formulation by looking for the parameter topos, the geometric action map, and the coherent multiplication and unit comparisons. The existing overview of Jia, Peng, Yang, and Chen [20] collects examples of topos-based learning in the broader sense. We examined its references, then checked primary sources for geometric actions, classifying topoi, equivariant sheaves, and isotropy in relation to learning. The sources give a basis for a focused survey at the first two levels. They leave the design of learning procedures using literal monoidal-topos actions or general localic isotropy as a separate question.

2.1 Neural networks in equivariant topoi

A direct precedent relating topoi and neural networks is Belfiore and Bennequin's preprint [5, §2.1–2.2]. A network has an indexing category C that records its architecture. In their constant-symmetry example one replaces set-valued presheaves by presheaves taking values in $\mathbf{B}G$, the topos of left G-sets for a discrete group G. A feature diagram is then

$$F : C^{\mathrm{op}} \longrightarrow \mathbf{B}G.$$

Arrows of C induce restriction maps on feature data and the maps of $\mathbf{B}G$ require equivariance of each transition. Curried evaluation gives

$$\begin{aligned} \mathrm{Fun}(C^{\mathrm{op}}, \mathbf{B}G) &\simeq \mathrm{Fun}(C^{\mathrm{op}} \times BG, \mathbf{Set}) \\ &\simeq C \times \widehat{(BG)}^{\mathrm{op}}. \end{aligned}$$

The last term is the presheaf topos on $C \times (BG)^{\mathrm{op}}$. Here BG is the one-object category with arrow group G. In equivariantization terminology the first expression is the topos of equivariant objects for the trivial external action of G on $\widehat{C}$. This is a concrete use of the group classifying topos as a target for feature data. The geometric theory of arbitrary G-actions has a classifier of its own. Its points describe G-sets, whereas points of $\mathbf{B}G$ describe G-torsors [21, slide 22].

Belfiore and Bennequin also let the categories in the invariance structure vary over the architecture and use a fibration whose fibers carry those categories. Their relative action is given by contravariant functors between fibers commuting with reindexing [5, Eq. (2.1)]. This specifies compatible diagrams of behaviors in the fibers. A geometric action of a parameter topos would further specify a multiplication on the parameter topos and a coherent geometric morphism out of a product. The preprint proposes that invariance explains aspects of extrapolation. The topos description supplies a type of invariance constraint. Statistical generalization still requires an input distribution, a loss, and hypotheses about a training procedure.

A related result gives a topos from the selected interface of a learner. Spivak [28] represents parametrized learners as generalized Moore machines using polynomial interfaces. For a fixed polynomial p, the category of p-coalgebras forms a topos. Its internal language can express propositions about the dynamics, with gradient descent as a worked case. This makes the topos a setting for specifying learner state and behavior. The monoidal structure used to build the interfaces belongs to the category of polynomials. The more recent survey also discusses *The Topos of Transformer Networks*. The canonical record for that preprint is marked withdrawn [30].

2.2 Relative classifying topoi for symbolic AI

Lafforgue [21, slides 20–25] proposes organizing a data language by a geometric theory T. A map from a test topos to its classifying topos classifies a T-model in that test topos. Taking the test topos to vary therefore models families of interpretations and their change of base. His proposed networks of topoi attach geometric object categories and classifiers to different processing stages. A transformation natural in whole families is a candidate for a geometric morphism between classifiers. This provides a way to formulate the transfer of structured representations. The talk gives an architecture proposal rather than an optimization or accuracy result. Caramello's course on relative topoi [7] makes the symbolic application more concrete. It writes invariance constraints for Raven matrices in a geometric language, including an example where an arithmetic rule on positions is expressed by a formula indexed by the graph of that rule. Restricted vocabularies and specified relations make candidate interpretations comparable by proof. This is a formal model of a class of reasoning problems and a distinct application from training an equivariant numerical predictor.

2.3 An algorithmic precedent via nominal sets

Active automata learning gives an implemented example of learning in a symmetry-sensitive setting. The nominal topos is the category of discrete sets with continuous action of the permutation group of countably many names, equipped with the topology of pointwise convergence. Continuity of an orbit map means that an element has a finite support: its stabilizer contains the subgroup fixing some finite set of names pointwise. The set of name pairs, for example, has two orbits, equal and unequal pairs. Such orbit-finite structure permits a finite orbit description even for machines with infinitely many individually named inputs. Iwaniack [19] uses the nominal topos as an instance of topos-enriched automata and derives Myhill–Nerode results there.

Moerman and collaborators [24] adapt Angluin's query algorithm to nominal automata, including preliminary experiments with a Haskell implementation. The teacher

answers queries about acceptance and proposes counterexamples to incorrect hypotheses. Equivariance lets sets of examples and hypotheses be handled by their orbit structure. Barbé [4, Thm. 4.26] generalizes this formulation in a research report. Over an atomic topos with countable coproducts, an alphabet and target state object each having finitely many atom components, and an appropriate teacher, TopL* terminates with a minimal automaton. This conclusion concerns exact learning in the stated query model. An effective implementation still depends on the ability to manipulate the orbit or atom presentations and to supply the teacher. These are meaningful operational assumptions when trying to infer a structured protocol model whose states depend on relationships among names.

3 Sheaf learning and local equivariance

Two established lines of work supply concrete models for applying topos-action ideas. Spectral sheaf learning builds features that satisfy incidence constraints. Equivariant network design builds maps that commute with transports. Their intersection can be described exactly for invertible transports on a graph.

3.1 From sheaf Laplacians to learned incidence maps

Hansen and Ghrist develop the spectral theory of cellular sheaves, including Hodge Laplacians and their relation to sheaf cohomology [16]. Hansen and Gebhart use the sheaf Laplacian in graph neural layers [15]. Bodnar et al. learn the incidence maps from data, study how sheaf geometry changes diffusion and oversmoothing, and compare their models on graphs where adjacent vertices can have different labels [6]. An earlier direction in graph learning uses edge-conditioned filters [25]. Edge attributes there determine message weights. In a sheaf layer, maps from the two endpoints into a common stalk determine which local features agree and how disagreement contributes to the diffusion.

For a finite loopless graph Γ, write V_v for a vertex stalk and V_e for an edge stalk, with maps $r_{ev} : V_v \to V_e$ at each incidence. Give the finite dimensional stalks inner products and orient each edge $e : u \to v$ for purposes of signs. Set

$$(\delta x)_e = r_{ev}x_v - r_{eu}x_u, \qquad L = \delta^*\delta.$$

The Laplacian satisfies $\langle x, Lx\rangle = \|\delta x\|^2$. Thus $\ker L = \ker \delta$ is the vector space of global sections, and e^{-tL} converges to orthogonal projection onto that space as $t \to \infty$. This last statement is for the fixed unnormalized operator on a finite graph. It follows by diagonalizing the positive semidefinite operator and does not describe an arbitrary learned nonlinear diffusion trajectory.

Suppose each incidence map is invertible. Then the edge transport from u to v is $T_e = r_{ev}^{-1}r_{eu}$, and the reverse edge transports by T_e^{-1}. These maps extend to a representation of the fundamental groupoid of the graph, regarded as a one-dimensional complex. Indeed, the incidence category has the vertices and edges as objects and arrows from a vertex to its incident edges. Localizing by those arrows gives the fundamental groupoid of the barycentric subdivision. By the universal property of localization, precisely the diagrams with invertible incidence maps extend. Restricting to the original vertices gives the edge transports T_e. Global sections correspond to parallel families $x_v = T_e x_u$. On each connected component choose a base vertex b. Transport of x_b determines the whole family, and it is independent of the chosen paths precisely when x_b is fixed by every based loop transport. Hence the global sections on a connected graph identify with the holonomy-invariant subspace of V_b. The assertion concerns graph paths. If two-cells are added and their boundary loops are imposed as relations, their transports must satisfy the additional flatness conditions.

A rank-one example is a triangle with two positive unit transports and one negative unit transport. The product around the cycle is -1, so the only real global section is zero. By comparison, using three positive unit transports gives the line of constant sections. The underlying incidence graph is unchanged. This separates the effect of the transport from that of connectivity. When the incidence maps are isometries, each residual has norm $\|x_v - T_e x_u\|$. In that case the operator is a discrete connection Laplacian in the sense relevant to vector diffusion maps [26]. For orthogonal changes of stalk coordinates, let Q_0 and Q_1 be the block diagonal changes at vertices and edges, respectively. The new incidence matrices are $r'_{ev} = Q_e r_{ev} Q_v^{-1}$, giving $\delta' = Q_1 \delta Q_0^{-1}$ and $L' = Q_0 L Q_0^{-1}$. Thus the fixed diffusion is compatible with these changes of frames even when the incidence maps are noninvertible.

The same argument constrains learnable linear parameters. For two representations V, W of the graph path groupoid, a family of local linear maps $A_v : V_v \to W_v$ is compatible with transport exactly when

$$A_v T_e^V = T_e^W A_u, \qquad e : u \to v.$$

Equivalently, A is a parallel section of the bundle of linear maps, with transport $A \mapsto T_e^W A (T_e^V)^{-1}$. On a connected graph its parameter space is therefore the space of intertwiners of the two holonomy representations at one vertex. For sign and trivial line holonomies around the same cycle, a scalar parameter is forced to vanish. For two copies of sign holonomy, any scalar at the chosen vertex extends. Weight tying here is dictated by representation type as well as by the connectivity of the data. The invertibility hypothesis matters: ordinary cellular sheaves also permit noninvertible incidence maps and unequal stalk dimensions, for which the incidence equation

still makes sense and there need not be transport along groupoid arrows.

3.2 From global groups to local bisections

Global group actions determine the shared weights in group-equivariant convolutions [10]. Cohen, Geiger, and Weiler characterize equivariant linear maps between fields on homogeneous spaces by their convolution kernels [9]. On meshes, de Haan et al. use parallel transport to define orientation-sensitive convolutions that respect changes of local frame [12].

For a finite data set carrying a group action, let $K_{x,y} : V_y \to W_x$ be the blocks of a linear layer and let $\rho_V(g; y)$ and $\rho_W(g; x)$ be the feature transports. Equivariance is equivalent to

$$K_{gx,gy} = \rho_W(g; x) K_{x,y} \rho_V(g; y)^{-1}.$$

It follows by comparing coefficients of a feature supported at y. Each diagonal pair orbit has one free block in the intertwiner space of its joint stabilizer, since two transports from a representative to the same pair differ by a stabilizer element. Conversely an intertwiner at each representative propagates to a well-defined equivariant kernel.

Ibort, Jimenez-Vazquez, and Perez-Pardo extend this type of kernel analysis to groupoid representations with a chosen pseudogroup of local bisections, measure, and input and output bundles. Their preprint treats bounded domains, where a geometric transformation can be defined on part of the data space and pair types can depend on the boundary [18]. The choice of which bisections to act by and the integration and support conditions enter the analytical layer construction. In the groupoid descent picture, the representation bundles supply transport data on the features and on the kernel's source and target variables. The kernel then acts on spaces of sections and can mix distinct sites. A morphism of feature representations over the identity base, such as A above, acts fiberwise. This distinction separates useful precedents for learning with groupoid descent from studies of a monoidal Grothendieck topos acting on another topos by geometric morphisms.

4 Compositional training and inference

The categorical literature on training and probability supplies two ingredients for learning in a setting with topos actions. Parametrized maps provide a compositional description of local models, and stochastic kernels provide a description of local inference. To obtain a model over a groupoid of domains or in an action topos, one supplies additional transport for the parameters, inputs, and outputs. Compatibility is a condition on both the local maps and the operations applied to them.

4.1 Reverse differentiation and descent of updates

Fong, Spivak, and Tuyéras make a learner into a compositional morphism with a forward evaluation, a parameter update, and a backward request. Under their choices of error function and fixed learning rate, gradient descent gives a monoidal functor from parametrized functions to learners [13]. Cruttwell et al. separate the optimizer and loss from the architecture using parametrized maps, lenses, and reverse derivative categories [11]. These constructions address how a training rule composes along a network. To use them with an action on a topos, the relevant categories and rules must also support transport along changes of domain. The transport condition is what allows the compositional training rule to respect the chosen action.

For example, in smooth models the cotangent pullback of a loss covector is canonical. Given layers $f : X \to Y$ and $g : Y \to Z$, the cotangent maps satisfy

$$D(g \circ f)^*_x = (Df_x)^* (Dg_{f(x)})^*.$$

The same identity applied to a commutative square of local realizations describes the transformation of the reverse pass. Converting a cotangent vector into a parameter update adds geometric data. In particular, a matrix transpose relative to a Euclidean metric in one coordinate system implements a different conversion when a nonisometric change of coordinates is made and the same Euclidean formula is used in the new coordinates.

An explicit sufficient condition for a global gradient is available in the smooth setting. Suppose a groupoid presents a parameter model by local manifolds P_i with transition diffeomorphisms ϕ_{ij} on the corresponding overlaps. Equip the local manifolds with losses ℓ_i and positive-definite metrics g_i satisfying

$$\ell_i = \ell_j \circ \phi_{ij}, \qquad g_i = \phi_{ij}^* g_j.$$

Then their gradients are compatible on the overlaps:

$$D\phi_{ij}(\operatorname{grad}_{g_i} \ell_i) = (\operatorname{grad}_{g_j} \ell_j) \circ \phi_{ij}.$$

Indeed, pairing the left-hand side with $D\phi_{ij}(v)$ in g_j gives $g_i(\operatorname{grad}_{g_i} \ell_i, v) = d\ell_i(v)$. Differentiating the loss identity gives the same pairing for the right-hand side. Surjectivity of the differential of the transition and nondegeneracy of the metric give equality. Local gradient flows therefore respect the transitions wherever defined, by uniqueness of solutions. A discrete training rule requires its own compatibility. For instance, a step made using the Riemannian exponential inherits the compatibility on the domain where it is defined, because the transitions are local isometries.

Natural gradient supplies a relevant source of metrics [2]. For a smooth family of strictly positive probability

laws $p_\theta(y)$ on a finite outcome set, its Fisher form is

$$I_\theta(v,w) = \sum_y \frac{dp_\theta(y)(v)\, dp_\theta(y)(w)}{p_\theta(y)}.$$

If a transition reparametrizes the family and permutes the outcome labels, the form transforms by metric pullback, by the chain rule and reindexing the sum. When the likelihood map has injective differential this form is positive definite and the descent argument applies. In overparametrized models its degeneracy can require passing to identifiable directions. Invariance of the resulting natural gradient depends on the chosen model or pushforward [29].

In a supervised problem, compatibility of the losses can itself be tested. Let μ_i be a finite distribution on local labelled examples, let F_i be the prediction map, and let λ_i be the example loss. The risk

$$\ell_i(\theta) = \sum_{x,y} \mu_i(x,y)\lambda_i(F_i(\theta,x),y)$$

is preserved if the transition transports the example distribution, intertwines the predictions, and preserves the example loss. The claim follows by substituting the transported examples in the sum. This gives an intrinsic risk for data described in several frames. Distinct datasets require explicitly specifying the risk to be shared across frames.

4.2 Probabilistic models under transport

Markov categories give an abstract semantics for stochastic kernels, conditional independence, and disintegration [14]. Quasi-Borel spaces supply cartesian closed structure with probability distributions, permitting higher-order probabilistic programs such as random functions [17]. Bayesian lenses describe composition of prior-dependent inversions when the corresponding conditionals exist [27]. These frameworks address probabilistic modeling and inference. Applying them to a data-domain groupoid requires specifying what a domain change does to outcomes, states, and probability laws.

In the finite case the extra condition can be checked directly. On domains i, j, take state and observation sets P_i, Y_i and P_j, Y_j, normalized priors q_i, q_j, and likelihood kernels K_i, K_j. A transition consists of bijections $u : P_i \to P_j$ and $v : Y_i \to Y_j$. If $q_j(u\theta) = q_i(\theta)$ and $K_j(vy \mid u\theta) = K_i(y \mid \theta)$, then Bayes' rule commutes with the transition whenever the observed evidence has positive probability:

$$q_j(u\theta \mid vy) = \frac{K_i(y \mid \theta)q_i(\theta)}{\sum_{\theta'} K_i(y \mid \theta')q_i(\theta')} = q_i(\theta \mid y).$$

The denominator follows by reindexing the evidence sum along u. Thus coherent transitions for priors and likelihoods induce coherent transitions for posteriors wherever these are defined. The positivity hypothesis matters when specifying pointwise inference. More general disintegrations may be determined only almost surely and require care about which version is transported.

Compatibility under invertible transports should also be distinguished from consistency under marginalization. Abramsky and Brandenburger's contextuality framework studies when families of distributions on overlapping measurement contexts extend to a global distribution [1]. For an elementary illustration, place an equal mixture of $(0,1)$ and $(1,0)$ on each of the pairs AB, BC, CA of three binary variables. Their singleton marginals agree. Any joint distribution with these marginals would impose all three inequalities with probability one, but three binary variables cannot be pairwise unequal. In multi-view learning this example identifies a concrete modeling distinction: pairwise calibration can hold even when there is no joint law realizing the local models. A useful inference invariant here is existence of a joint law, measured separately from agreement of the observed marginals.

5 Descent of equivariant learning data

An equivariant layer may be specified in local coordinates that change with the observation context. The action topos gives a category in which to compare these specifications. Let $\mathcal{E}$ be a Grothendieck topos, B an object, and G_B an internal group of $\mathcal{D} = \mathcal{E}/B$. For example, G_B can be pulled back from a group on $\mathcal{E}$, or can itself vary over B. Use $\mathcal{C} = B_{\mathcal{D}}G_B$, the category of left G_B-objects in $\mathcal{D}$. It is again a Grothendieck topos [3]. Its products are computed over B with the diagonal action.

A parametrized layer is a morphism $h : P \times X \to Y$ in $\mathcal{C}$, where P is the parameter object. In internal notation the condition is

$$h(gp, gx) = gh(p,x).$$

A fixed parameter accordingly yields an equivariant predictor for the full group when that parameter is fixed by the full action. More generally the condition on the specialized predictor uses the stabilizer of the parameter. Composition of two layers uses the product of their parameter objects, with the same diagonal action. Reusing parameters across layers amounts to inserting an equivariant map into that product.

5.1 Predictions, losses, and updates

Supervision is an object D equipped with input and target maps $\xi : D \to X$ and $\tau : D \to Y$. Let R be a score object from $\mathcal{D}$ with trivial G_B-action, and let $c : Y \times Y \to R$ be an equivariant comparison. The induced loss is

$$\ell : P \times D \to R, \qquad \ell(p,d) = c(h(p,\xi(d)),\tau(d)).$$

Thus $\ell(gp, gd) = \ell(p,d)$. In $\mathcal{D}$ it factors through the joint orbit object $(P \times D)/G_B$, using the coequalizer of the

action and projection. Passing to the product of the two orbit objects can lose values of the loss. For example, in sets take $P = D = \{1, -1\}$ with the sign action by C_2, $X = D$, and $Y = \{1, -1\}$ with trivial action. The layer $h(p, x) = px$, constant target 1, and comparison $c(a, b) = (1-ab)/2$ give two loss values on two joint orbits. The product of the separate orbit objects is a singleton. An objective $P \to R$ obtained by averaging the loss uses a data distribution and an integration operation as additional data. For a finite data set in sets with real scores and probability weights $\mu(d)$, the risk is $\sum_d \mu(d)\ell(p, d)$. If $\mu(gd) = \mu(d)$, simultaneous invariance of the loss and reindexing the sum show that this risk is invariant on P.

An update is separate data $u : P \times D \to P$ in $\mathcal{C}$. Equivariance now says $u(gp, gd) = g\, u(p, d)$. With the same invariant finite sampling law, the law of the next parameter at gp is the pushforward along g of the corresponding law at p. This follows by replacing the random sample d by gd, which has the same law. For a parameter change $r : P' \to P$, the predictor reparametrizes by precomposition. An update u' on P' is compatible if it satisfies

$$r \circ u' = u \circ (r \times \mathrm{id}_D).$$

An isomorphism r transports the update uniquely by this equation. The equation is an additional requirement for a general r, where precomposition alone leaves the result of the old update in P. For example, the update $p \mapsto p + 1$ on $\mathbb{R}$ has no compatible lift to the single parameter value 0.

5.2 A local-to-global criterion

Theorem 5.1 (Descent of layers and updates)**.** *Let $U \to 1$ be an epimorphism in $\mathcal{C}$. Suppose the objects P, X, Y, D are given in $\mathcal{C}/U$ with descent isomorphisms satisfying the cocycle on $U \times U \times U$. Give the restriction of the trivial-action score object R its canonical descent isomorphism. The local layer, input and target maps, comparison, and update descend precisely when each map commutes with the specified descent isomorphisms. The global diagram is unique up to a unique compatible isomorphism. The induced local loss descends through the same comparison maps.*

Proof. In a Grothendieck topos an epimorphism is effective descent for objects. To recall the construction, the transport on $A \to U$ over $U \times U$ identifies a above v with its transported element above w. The unit, inversion, and cocycle define an internal equivalence relation on A. Its effective quotient Q exists. The map $A \to Q \times U$ is monic because each transport within one fiber is the identity. It is epic because locally a representative of each class can be transported to the specified fiber. Hence it is an isomorphism. This is the usual descent calculation using exactness of a topos [23]. A map preserving the transports preserves the relations and descends by the quotient property. Pullback along $U \to 1$ is conservative and faithful, so it detects the required diagrams and their equations. It also preserves products. Consequently the local composite defining the loss is the restriction of the global one. □

A useful cover is the regular G_B-object $G_B \to 1$ in $\mathcal{C}$. The slice $\mathcal{C}/G_B$ is equivalent to $\mathcal{D}$: take the fiber of an equivariant object over the identity. The inverse sends A to $G_B \times A$ over G_B with left translation on the first factor. For an element a of an equivariant object $q : Q \to G_B$, the inverse to this product presentation is $(q(a), q(a)^{-1}a)$. Thus ordinary local data over the regular cover can encode equivariant global data, with the action recovered from the overlap transports. To descend a local parameter choice, treat it as a morphism from the terminal object of $\mathcal{C}/U$. The same compatibility condition then tests whether it gives a global section of P.

5.3 Coordinate compatibility of gradient updates

The update equation also applies to the Riemannian rules of Section 4. If the parameter transition carries the parameter metric and loss, it transports the gradient and the geodesics used for a discrete step [22]. Using a new metric without transporting it can change the update even when the predictions and losses agree.

For example, let $q = 2p$ and $\ell(p) = p^2/2$. The unit metric in p transports to $\frac{1}{4}\, dq^2$, giving gradient component q. Independently choosing the unit metric in q gives component $q/4$ for the same loss $q^2/8$. Thus the descent test for an optimizer concerns both its prediction and its optimization data. For a stateful optimizer the state and its transport can be included in the updated parameter object, with the same comparison condition on the enlarged update.

6 Two uses of quotient and local symmetry

The first use chooses a nuisance action by a monoidal topos and computes the resulting space of feasible layers. The second uses a symmetry that varies as an internal group and computes the admissible coefficients of a linear network.

6.1 A central action and nuisance-invariant learning

Let A be an abelian discrete group, let H be a discrete group, and let $\chi : A \to Z(H)$ be a homomorphism. Write $N = \chi(A)$ and $K = H/N$. The induced action of $\mathbf{B}A$ on $\mathbf{B}H$ has geometric quotient $q : \mathbf{B}H \to \mathbf{B}K$. Its inverse image inflates a K-action along $H \twoheadrightarrow K$. An H-object descends precisely when its N-action is trivial [3]. This

is a literal action between topoi of the kind discussed in Section 1.

Proposition 6.1 (Orbit features from a central quotient)**.** *Let P, W be H-sets with trivial N-action, and V an arbitrary H-set. An H-equivariant map $F : P \times V \to W$ corresponds uniquely to a K-equivariant map $\overline{F} : P \times (V/N) \to W$. For topological actions and continuous layers the same statement holds with the orbit topology on V/N.*

Proof. Equivariance and the hypotheses give $F(p, nv) = nF(p, v) = F(p, v)$. Therefore F is constant on the fibers of $(p, v) \mapsto (p, [v])$ and factors. Normality of N lets H act on the orbits through K. The factor map is consequently equivariant for K, and the converse follows by composing with the quotient. For continuous actions, $V \to V/N$ is an open surjection since the saturation of an open set is the union of its translates. Its product with P is also an open surjection, so the factor map is continuous by the quotient property. □

Suppose now that N is finite and V, W are real linear representations. For W with trivial N-action, an N-invariant *linear* map out of V satisfies $L = L\Pi_N$, where

$$\Pi_N v = \frac{1}{|N|} \sum_{n \in N} nv.$$

The operator is a projector onto V^N, and for an H-representation it commutes with H because N is normal. The same factorization holds for an affine map by adding its constant part. The orbit feature V/N instead retains the information needed for all invariant nonlinear maps. To see the difference, take $H = C_4$ acting by quarter turns on $V = \mathbb{R}^2$ and $N = C_2$ acting by $v \mapsto -v$. This is obtained from $\chi : C_2 \to C_4$ with image the half turn. Here $\Pi_N = 0$. A label on the input set $\{0, \pm e_1, \pm e_2\}$ indicating whether it is nonzero is predicted exactly by the invariant square-activation score $\|v\|^2$. If half of the distribution is at 0 and half is uniform on the other four points, a predictor fed only $\Pi_N v$ is constant. Its minimum mean square error is $1/4$, the variance of the label. The nonlinear invariant predictor has zero error. The quotient determines which information may be discarded, and an implementation still has to choose an orbit encoding.

There is a related operation on predictors rather than input features. If μ is the joint distribution of input and label, assume μ is invariant under $(v, y) \mapsto (nv, y)$. For a vector score s and a loss $\ell(y, -)$ convex in the score, define $\overline{s}(v) = |N|^{-1} \sum_n s(nv)$. For measurable scores and a nonnegative loss with finite risk, Jensen's inequality gives

$$\begin{aligned} \mathbb{E}_\mu[\ell(y, \overline{s}(v))] &\le \frac{1}{|N|} \sum_{n \in N} \mathbb{E}_\mu[\ell(y, s(nv))] \\ &= \mathbb{E}_\mu[\ell(y, s(v))]. \end{aligned}$$

The averaged predictor descends to V/N by invariance. If s had an additional H-equivariance with N acting trivially on the output, the factor is K-equivariant as in Proposition 6.1. This is an exact-invariance example of orbit averaging in data augmentation [8]. The inequality compares realized predictors. An architecture must contain the average, for example by parallel evaluation, before it gives a comparison within that architecture.

6.2 A filter with a twisted coefficient

Let the deployment base be $B = S^1 = \mathbb{R}/(2\pi\mathbb{Z})$. In $\mathrm{Sh}(B)$ let G be the locally constant group sheaf with fiber $C_3 = \mathbb{Z}/3$ and monodromy $j \mapsto -j$. Only 0 is fixed by inversion, so $\Gamma(B, G)$ is the trivial group. The local groups are nevertheless nontrivial.

Use real regular-representation feature fibers $V_t = \mathbb{R}[C_3]$ with basis $(e_j)_{j \in C_3}$. The monodromy on the bundle V is the orthogonal permutation $Se_j = e_{-j}$. Let $Re_j = e_{j+1}$ be the group shift in a local lift of the angular coordinate. Then $SRS^{-1} = R^{-1}$, which makes the regular action compatible with the group sheaf's transition. The sheaf of smooth sections of V is an internal G-object. Its fiberwise G-linear maps define equivariant linear layers in the relative classifying topos.

Proposition 6.2 (The coefficient bundle of equivariant filters)**.** *The bundle of fiberwise linear intertwiners from V to itself is $\mathbb{R}^2 \oplus L$ over S^1, where L is the real Möbius line and $\mathbb{R}^2$ denotes a trivial rank-two bundle. A flat filter, meaning locally constant coefficients in the given flat frames, has two free real coefficients. Smooth filters can also use the third coefficient as a section of L.*

Proof. A linear map commuting with R is determined by its value w on e_0, since its values on the other basis vectors must be Rw and $R^2 w$. Every w gives a map by this rule. Hence the commutant consists precisely of $c_0 I + c_1 R + c_2 R^2$. Changing frame by S exchanges c_1, c_2. In the basis $I, R + R^2, R - R^2$ it is the identity on the first two directions and multiplication by -1 on the third. This gives the bundle splitting and the flat coefficient count. □

The distinction affects model capacity. Using the global section group alone would impose no commutation conditions and would allow all nine entries of a local matrix. Requiring internal equivariance leaves three local coefficients. Choosing flat global coefficients ties the two shifted terms together, leaving two. For smooth models the third coefficient $b(t)$ can be any smooth function on the lifted angle satisfying $b(t + 2\pi) = -b(t)$. For example $b(t) = \cos(t/2)$ gives a well-defined global filter $b(t)(R - R^2)$. Every continuous such b has a zero: if $b(t)$ is nonzero, the values at t and $t + 2\pi$ have opposite signs. Hence a model that requires this coefficient to stay nonzero can work over the orientation double cover, where the

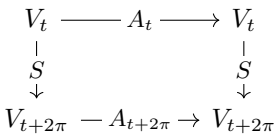


Figure 1: A filter descends when it intertwines the change of frame. Conjugation by the transition S reverses the coefficient of $R - R^2$.

pulled-back line is trivial. Learning this coefficient in the bundle and learning one untwisted scalar are different parameterization problems.

A coordinatewise activation commutes with the local shift and the frame permutation. Consequently one may stack these linear filters with such activations to obtain a global nonlinear layer. A sum of the three output coordinates then gives a scalar readout, invariant under both the local group and the frame transition.

A simple regression calculation gives a training interpretation with a specified observation law. In a local frame draw $x = e_j$ uniformly over j, draw a hidden sign $\varepsilon \in \{\pm 1\}$ fairly and independently, and observe the target $y = R^{\varepsilon}x$. Inverting the frame sends (x, y, ε) to $(Sx, Sy, -\varepsilon)$, so the law descends even though a deterministic forward shift R does not define a flat global filter. The least-squares estimator conditional on x is $(R + R^2)x/2$, an admissible flat filter. The two conditional target vectors are distinct basis vectors, so each lies squared distance $1/2$ from their mean. The best expected squared error using only x is therefore $1/2$. The rule $(x, \varepsilon) \mapsto R^{\varepsilon}x$ has zero error if the sign is supplied with its frame transition. The ambiguity here is a stated property of the observation law. Monodromy supplies the transformation of the hidden sign and the constraint on flat filters.

7 Conclusion

The literature provides several operational cases of learning with symmetry-sensitive structure. Nominal automata support active queries over infinitely many names with finitely described orbits. Sheaf neural networks use local incidence geometry, and groupoid-equivariant networks use transformations that are defined only on parts of the input space. Topos and stack semantics also supply formal descriptions of network behavior and symbolic reasoning. The type of action differs across these examples.

The two models in Section 6 give concrete consequences of following the type of action through to the parameters and observations. A central quotient identifies the input orbit as the information available to every nuisance-invariant predictor. The circle group sheaf imposes a transition law on one learnable coefficient, even though its global group of sections is trivial. An equivariant optimizer requires compatible metric, loss, and data transport in addition to compatible layers. The topological statements describe admissible realizations. Prediction error and optimization depend on the observation law and update rule specified for the model.

References

[1] S. Abramsky and A. Brandenburger. The sheaf-theoretic structure of non-locality and contextuality. *New J. Phys.* 13 (2011), 113036. DOI.

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

[3] Anonymous. *Actions of Toposes on Toposes: Geometric Actions, Internal Symmetry, and Quotient Constructions*. Companion manuscript, submitted to Higher Categorical Semantics, 2026.

[4] K. Barbé. *Learning Functorial Automata in Topoi*. Research internship report, IRIF, 2024. Report.

[5] J.-C. Belfiore and D. Bennequin. Topos and stacks of deep neural networks. Preprint, version 3, 2022. arXiv:2106.14587.

[6] C. Bodnar, F. Di Giovanni, B. P. Chamberlain, P. Liò, and M. M. Bronstein. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in GNNs. In *Advances in Neural Information Processing Systems* 35, 2022. Proceedings.

[7] O. Caramello. *Relative Toposes for Artificial General Intelligence*, part 2. CentraleSupélec, July 2025. Course slides.

[8] S. Chen, E. Dobriban, and J. H. Lee. A group-theoretic framework for data augmentation. *J. Mach. Learn. Res.* 21(245) (2020), 1–71. Journal article.

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

[10] T. Cohen and M. Welling. Group equivariant convolutional networks. In *Proceedings of ICML*, PMLR 48 (2016), 2990–2999. Proceedings.

[11] 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, 1–28. DOI.

[12] P. de Haan, M. Weiler, T. Cohen, and M. Welling. Gauge equivariant mesh CNNs: Anisotropic convolutions on geometric graphs. In *Proceedings of ICLR*, 2021. Proceedings.

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

[14] T. Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Adv. Math.* 370 (2020), 107239. DOI.

[15] J. Hansen and T. Gebhart. Sheaf neural networks. Preprint, 2020. arXiv:2012.06333.

[16] J. Hansen and R. Ghrist. Toward a spectral theory of cellular sheaves. *J. Appl. Comput. Topol.* 3 (2019), 315–358. DOI.

[17] C. Heunen, O. Kammar, S. Staton, and H. Yang. A convenient category for higher-order probability theory. In *Proceedings of LICS*, 2017. DOI.

[18] A. Ibort, M. Jimenez-Vazquez, and J. M. Perez-Pardo. Theory for groupoid equivariant neural networks: An approach for steerable CNNs on bounded domains. Preprint, 2026. arXiv:2609.25987.

[19] V. Iwaniack. Automata in toposes, and general Myhill–Nerode theorems. Preprint, 2023. arXiv:2307.14855.

[20] Y. Jia, G. Peng, Z. Yang, and T. Chen. Category-theoretical and topos-theoretical frameworks in machine learning: A survey. *Axioms* 14(3) (2025), 204. DOI.

[21] L. Lafforgue. *Some Possible Roles for AI of Grothendieck Topos Theory*. ETH Zürich, 2022. Talk slides.

[22] J. M. Lee. *Introduction to Riemannian Manifolds*. Second edition. Springer, 2018. DOI.

[23] S. Mac Lane and I. Moerdijk. *Sheaves in Geometry and Logic: A First Introduction to Topos Theory*. Springer, 1992. DOI.

[24] J. Moerman, M. Sammartino, A. Silva, B. Klin, and M. Szynwelski. Learning nominal automata. In *Proceedings of POPL*, 2017, 613–625. DOI.

[25] M. Simonovsky and N. Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *Proceedings of CVPR*, 2017, 3693–3702. Proceedings.

[26] A. Singer and H.-T. Wu. Vector diffusion maps and the connection Laplacian. *Comm. Pure Appl. Math.* 65(8) (2012), 1067–1144. DOI.

[27] T. St. Clere Smithe. Bayesian updates compose optically. Preprint, 2020. arXiv:2006.01631.

[28] D. I. Spivak. Learners' languages. *Electron. Proc. Theor. Comput. Sci.* 372 (2022), 14–28. DOI.

[29] J. van Oostrum, J. Müller, and N. Ay. Invariance properties of the natural gradient in overparametrised systems. *Information Geometry* 6 (2023), 51–67. DOI.

[30] M. J. Villani and P. McBurney. The topos of transformer networks. Withdrawn preprint, 2024. arXiv record.