

Higher Categorical Semantics for PAC Learning

A summary survey of foundations, recent results, and remaining gaps

Anonymous author

Literature checked 20 October 2026 (UTC)

1. Assessment and scope

The available literature supports a developing interface between categorical semantics and PAC learning, rather than one established higher categorical PAC theory. Categorical learning describes composition, parameterization, differentiation, invariance, and structural transfer. PAC theory supplies quantitative statements about learning from finite random samples. A theorem connecting them must specify observations, probability, loss, statistical complexity, and an actual learner.

This survey prioritizes primary papers and author-maintained sources available by 20 October 2026 UTC, emphasizing 2025-2026 developments. It distinguishes three levels of evidence: **reported results** in the cited literature; **elementary consequences** proved here; and **proposed synthesis**, which is a research framework rather than a theorem attributed to those sources. It is a selective survey, not a claim that every relevant work has been found or independently re-proved.

Three uses of “higher” require separation. A **2-category** has morphisms between morphisms. An **infinity-category** retains coherent higher homotopies. **Higher arity PAC learning** studies relational observations and structured sampling; its arity does not make it higher category theory. An ordinary Grothendieck topos is not, merely by that name, an infinity-topos of space-valued sheaves.

The clearest recent connection to higher categorical semantics is the 2-categorical representation of error minimization by Pugh and collaborators [1]. The most recent verified revision considered here is Melodia’s 9 October 2026 work on Kan extensions of learned topological invariants [2]. Neither result alone is a distribution-free PAC sample-complexity theorem.

2. The statistical statement that semantics must preserve

Let $Z = X \times Y$ be a measurable example space, H a measurable hypothesis space, and $\ell : H \times Z \rightarrow [0, 1]$ a jointly measurable loss. For a probability law D on Z , write

$$R_D(h) = \mathbb{E}_{z \sim D} \ell(h, z), \quad \widehat{R}_S(h) = \frac{1}{m} \sum_{i=1}^m \ell(h, z_i).$$

A possibly randomized learner is a Markov kernel $A_m : Z^m \rightsquigarrow H$. An agnostic PAC guarantee has the quantifier order

$$\forall \varepsilon, \delta \in (0, 1) \ \exists m_0 \ \forall D \ \forall m \geq m_0 : \Pr_{S \sim D^m, h \sim A_m(S)} \left[R_D(h) \leq \inf_{g \in H} R_D(g) + \varepsilon \right] \geq 1 - \delta.$$

Statistical learnability does not assert computability or polynomial runtime. The realizable binary case assumes a target in H with zero risk; agnostic learning does not. For ordinary binary classes, finite VC dimension characterizes distribution-free PAC learnability under the usual measurability conventions [3,4]. This does not extend unchanged to every multiclass, partial, relational, or computational model.

For finite H , a directly useful benchmark is

$$\Pr \left[\sup_{h \in H} |R_D(h) - \widehat{R}_S(h)| > t \right] \leq 2|H|e^{-2mt^2}.$$

Indeed, apply Hoeffding to each bounded loss and take a union bound. On the complementary event, an empirical risk minimizer $\widehat{h}$ satisfies $R_D(\widehat{h}) \leq \min_h R_D(h) + 2t$. Thus

$$m \geq \frac{2}{\varepsilon^2} \log \frac{2|H|}{\delta}$$

is sufficient in this finite agnostic setting. If the binary class is realizable, a consistent learner instead has failure probability at most $|H|e^{-m\varepsilon}$, giving the sufficient bound $m \geq (\log |H| + \log(1/\delta))/\varepsilon$. These are elementary finite-class bounds, not claims of sharp constants for every PAC model.

3. Recent developments and their precise reach

Work and verified date	Mathematical contribution	Relation to PAC learning
Pugh et al. [1], 2025; TMLR publication record dated 10 November 2025	An error-minimization problem can be represented in a constructed 2-category so that its global minimizers are left Kan extensions.	A representation theorem for optimization; no automatic generalization bound.
Chernikov and Towsner [5], v2, 15 October 2025	Higher arity VC dimension, a packing theorem, and proper PAC_k learnability in the paper’s product-measure and fiber-observation model.	A genuine statistical characterization in a specified higher arity model.
Kattermann and Krapp [4], COLT 2026	Computable PAC learning related to recursively enumerable representable classes and effective VC dimension.	Shows why semantic existence and finite ordinary VC dimension do not alone produce an executable learner.
Melodia [2], v4, 9 October 2026	Kan extensions of representation invariants, discrepancy, persistent-homology stability, and exact finite-sample structural results.	Relates task changes to measurable structural discrepancies; does not identify discrepancy with predictive risk.
Maeda, Seiller, and Oseki [6], public 2026 manuscript	Structural and lexical identifications induce Kan extensions diagnosing admissibility in COGS.	Studies what a corpus licenses structurally, without training a predictor or establishing a PAC rate.

Chernikov and Towsner’s Corollary 6.6 identifies finite VC_k dimension, the appropriate packing property, and PAC_k learnability; Theorem 6.5 obtains a proper learner. Their Definition 4.3 allows restrictions to coordinate fibers through sampled tuples. Replacing that access with an ordinary list of independently labeled tuples changes the learning problem. Coregliano and Malliaris [7] separately develop high arity learning through induced substructures and exchangeable distributions. These models should not be conflated.

Melodia’s latest abstract explicitly reports that discrepancy across pretrained models does not track linear-probe accuracy [2]. This is a useful limit: a rigorous structural invariant can detect specified changes while remaining insufficient to predict downstream statistical performance. Maeda et al. similarly separate corpus-level admissibility from model accuracy [6].

4. What the categorical frameworks contribute

4.1 Compositional learners, lenses, and architecture

Fong, Spivak, and Tuyeras organize supervised learning compositionally [8]. Cruttwell and collaborators use parametrized maps, lenses, and reverse derivative categories to describe gradient-based learning and several optimizer families [9]. A parametrized implementation has the form

$$f : P \times X \longrightarrow Y;$$

the parameter space and its update rules are part of the structure. This exposes how a composite model receives and propagates training information. It does not imply that arbitrary optimizers converge, have equivalent computational behavior, or generalize equally.

Gavranovic and collaborators’ ICML 2024 position paper formulates architecture constraints and implementations using monads in a 2-category of parametrized maps [10]. This offers a language for specifying a hypothesis family. A PAC analysis must then study the induced family of measurable predictors or loss functions: an algebraic presentation is not itself a complexity bound.

4.2 Kan extensions and optimization

For $i : C \rightarrow D$ and $F : C \rightarrow E$, a left Kan extension has a unit $F \Rightarrow (\text{Lan}_i F)i$ and, when it exists, the universal property

$$\text{Nat}(\text{Lan}_i F, G) \cong \text{Nat}(F, Gi).$$

Pugh et al.’s Theorem 4.1 constructs the surrounding 2-category from the error-minimization problem [1]. The construction uses a selection of existing minimizers. It therefore supplies neither a general algorithm for finding them nor an existence theorem when the infimum is unattained. Its error comparison uses lax 2-functorial structure. This is genuinely 2-categorical, but it is not by itself a full infinity-topos semantics or a proof of uniform convergence.

4.3 Categorical probability and statistical information

Markov categories describe stochastic channels compositionally. Fritz’s synthetic treatment includes conditional independence and sufficient statistics [11]. Fritz, Gonda, Perrone, and Rischel prove a categorical Blackwell-Sherman-Stein theorem under their hypotheses, comparing statistical experiments by garbling [12]. Perrone equips categorical probability with quantitative divergence and information-theoretic structure [13].

This is a strong foundation for the stochastic part of PAC semantics: sampling, randomization, postprocessing, and information loss. Nevertheless, bare Markov-category axioms do not contain Hoeffding’s inequality, measurable optimization, or finite VC dimension. To obtain a numerical PAC theorem one must supply a suitable probabilistic realization and the relevant concentration and complexity assumptions.

5. PAC-Bayes as a quantitative interface

For independent examples, bounded loss, and a prior π on H chosen independently of the sample, a standard PAC-Bayes bound states that, with probability at least $1 - \delta$, simultaneously for all posterior laws ρ ,

$$\text{kl}(\mathbb{E}_\rho \hat{R}_S \parallel \mathbb{E}_\rho R_D) \leq \frac{\text{KL}(\rho \parallel \pi) + \log(2\sqrt{m}/\delta)}{m}.$$

Here $m \geq 2$, logarithms are natural, and $\text{kl}(a \parallel b) = a \log(a/b) + (1-a) \log((1-a)/(1-b))$, with the usual endpoint conventions. See Maurer [14] and Alquier, Theorem 3.2.2 [15]. The posterior can depend on the sample and need not be a Bayesian posterior. A sample-dependent prior needs a separate justification.

A categorical interpretation uses a measurable semantics map $q : H \rightarrow \overline{H}$. If the loss factors through q , both empirical and population risks depend only on the pushforward posterior. Data processing gives

$$\text{KL}(q_*\rho\|q_*\pi) \leq \text{KL}(\rho\|\pi).$$

Applying the ordinary PAC-Bayes theorem directly on $\overline{H}$ can therefore remove a penalty for distinctions invisible to the chosen loss. This is an elementary consequence of loss factorization and data processing; it is not a new higher categorical concentration theorem. The quotient must be measurable. A smaller KL term also does not prove that a representative can be computed or efficiently sampled.

6. A precise higher categorical synthesis

The following construction is a **proposed organizing framework**, with two elementary consequences proved explicitly. Let $\mathcal{H}$ be a small infinity-category of structured hypotheses and W a chosen collection of morphisms designated as semantic weak equivalences. Consider its localized core

$$\mathcal{M} = (\mathcal{H}[W^{-1}])^{\simeq}, \quad H_0 = \pi_0(\mathcal{M}).$$

The core retains objects and invertible higher comparisons; H_0 records their equivalence classes. Equip H_0 with a sigma-algebra and supply jointly measurable observations $\ell : H_0 \times Z \rightarrow [0, 1]$. A learner is now a measurable kernel $Z^m \rightsquigarrow H_0$, or a kernel into a measurable space of presentations whose loss descends to H_0 . None of these measurable structures is supplied automatically by localization.

Proposition 1 (transport of a PAC guarantee). Let $q : H \rightarrow K$ be a measurable surjection of ordinary hypothesis spaces. Suppose

$$\ell_H(h, z) = \ell_K(q(h), z)$$

for all h, z , and define $A_m^K(S) = q_* A_m^H(S)$. Then the excess-risk random variables of the two learners have the same law. In particular, every agnostic PAC bound for A^H transfers with the same m, ε, δ .

Proof. Loss factorization gives $R_D^H = R_D^K q$. Surjectivity gives $\inf_H R_D^H = \inf_K R_D^K$. The output law for A^K is exactly the pushforward law, so the probabilities of corresponding excess-risk events agree. $\square$

The converse needs a measurable lift of learner outputs; computable or efficient transport needs corresponding effective lifts. An equivalence of higher categorical presentations that preserves the complete sampling-and-loss experiment can be analyzed this way. An abstract equivalence of categories, without those data, is insufficient.

Proposition 2 (finite observational quotient). Suppose all losses factor through a fixed finite discrete measurable quotient Q of cardinality $N \geq 1$. An empirical minimizer over Q , with any chosen representatives, satisfies

$$R_D(\hat{q}) \leq \min_{q \in Q} R_D(q) + \sqrt{\frac{2 \log(2N/\delta)}{m}}$$

with probability at least $1 - \delta$.

Proof. Apply the finite-class bound of Section 2 to the N quotient losses. A minimizer exists because Q is finite. The same empirical and population values hold for its representative. $\square$

This is a concrete statistical payoff of semantic identification: the bound counts distinct observable loss profiles. It does not count objects, morphisms, or homotopy groups without a theorem relating those quantities to the loss class. The quotient and observations must be fixed independently of the sample for this direct argument.

An invariant scalar loss on a space $\mathcal{M}$ factors through $\pi_0(\mathcal{M})$ because $[0, 1]$, viewed here as a set, is 0-truncated. Thus scalar risk alone cannot recover automorphism groups or higher coherence. Learning those structures requires observations sensitive to them and an identification theorem for that richer experiment. Univalence identifies equivalent structures within an appropriate type theory; it does not identify arbitrary statistically similar predictors.

7. Topoi, descent, and contextual learning

Belfiore and Bennequin use topoi, stacks, and homotopical structures to describe neural networks, invariance, and semantic information [16]. Their framework motivates contextual descriptions of learned systems. Its proposed connection between invariance and generalization should be separated from a proved, quantified PAC theorem.

For a genuine infinity-topos $\mathcal{E}$, Lurie’s descent theory [17] supplies a precise way to compare local objects. If $U \rightarrow T$ is an effective epimorphism with Čech nerve $U_\bullet$, there is an equivalence

$$\mathcal{E}_{/T} \simeq \lim_{[n] \in \Delta} \mathcal{E}_{/U_n}.$$

Applying $\text{Fun}(K, -)$ transports this equivalence to K -shaped diagrams. Additional loss or probability structures must themselves be specified compatibly. Descent requires objects, overlap equivalences, and all coherence data. The existence of some local learner on each patch does not provide those data.

There is a separate elementary statistical gluing argument. Suppose $D = \sum_{i=1}^r w_i D_i$, $w_i \geq 0$, $\sum_i w_i = 1$, and a *single* learned hypothesis h has local risks $R_i(h) \leq b_i$ on events of probability at least $1 - \delta_i$, all under the same training experiment. Then

$$\Pr \left[R_D(h) \leq \sum_i w_i b_i \right] \geq 1 - \sum_i \delta_i.$$

This follows from linearity of expectation and a union bound; independence between the local success events is unnecessary. Distinct local hypotheses first need a valid global construction. If the b_i are local optima plus errors, the weighted local optima can be a stronger benchmark than any globally compatible model can attain. Sheaf descent and statistical confidence budgeting are therefore complementary obligations.

8. What remains open in this synthesis

The reviewed sources do not provide one general theorem characterizing PAC learnability of arbitrary infinity-categorical hypotheses. This is a scoped finding about this survey, not a proof that no such paper exists. A useful research program would address the following concrete questions.

1. **Observable complexity.** Relate a specified family of categorical queries to VC dimension, pseudodimension, covering numbers, or another statistical complexity. Raw numbers of higher cells need not control any of these.
2. **Quantitative coherence.** Bound global predictive error by local error, overlap disagreement, and composition depth under explicit Lipschitz or stability assumptions. Pairwise approximate agreement alone is not homotopy-coherent descent.
3. **Effective realization.** Turn categorical minimizer existence into computable learners with effective sample bounds. The 2026 computable-PAC results [4] make this distinction mathematically substantive.
4. **Structured sampling.** Connect categorical relational semantics to a precise exchangeable or fiber-observation model [5,7], proving the required concentration statement rather than treating correlated tuples as independent examples.
5. **Identification beyond scalar risk.** Determine which experiments can distinguish equivalence classes, automorphisms, and higher coherence, and whether those distinctions matter to the downstream task.

The present mathematical interface is thus explicit: categorical structure specifies the models and comparisons; measurable observations define the statistical experiment; established PAC tools bound its loss class. A new higher categorical PAC theorem must prove a nontrivial connection between those layers.

9. A compact reading route

For the statistical baseline, read [3], [14], and [15]. For learning composition and architecture, read [8]-[10], then [1]. For stochastic semantics, use [11]-[13]. For actual higher arity statistical theorems and effective constraints, read [5], [7], and [4]. For the contextual and homotopical direction, pair [16]-[17] with the recent structural-transfer

work [2,6]. The author-hosted PACC terminology in [18] is an additional exploratory program; the glossary alone should not be treated as a verified sample-complexity theorem.

References

All links below were checked against primary publication records, primary preprints, or the named authors' sites during preparation. Dates distinguish a publication year from a later preprint revision. A publication record is not an independent certification of every theorem in a paper.

- [1] Matthew Pugh, Nick Harris, Corina Cirstea, and Jo Grundy. *Learning Is a Kan Extension*. Transactions on Machine Learning Research, 2025. [Publication record](#); [preprint and full text](#), Theorems 3.15 and 4.1. The preprint has a different author order.
- [2] Luciano Melodia. *Learning Transfers: Kan Extensions of Topological Invariants of Learned Representations*. arXiv:2606.07627v4, revised 9 October 2026. [Primary record](#). The survey uses the current abstract's explicit statements and limits; it does not claim an independent audit of all proofs or experiments.
- [3] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. [Authors' book page](#). Background reference; the finite-class estimates used here are derived in Section 2.
- [4] David Kattermann and Lothar Sebastian Krapp. *Recursively Enumerably Representable Classes and Computable Versions of the Fundamental Theorem of Statistical Learning*. COLT 2026, PMLR 336:3950-3969. [Proceedings and paper](#).
- [5] Artem Chernikov and Henry Towsner. *Higher-arity PAC learning, VC dimension and packing lemma*. arXiv:2510.02420v2, 15 October 2025. [Primary record](#); [full text](#). See Definition 4.3, Theorem 6.5, and Corollary 6.6.
- [6] Akihiro Maeda, Thomas Seiller, and Yohei Oseki. *Compositional Generalization via Structural Identification in a Category-Theoretic Framework*. Public manuscript, 2026, arXiv:2608.26465. [Author's research page and abstract](#); [preprint identifier](#). The author's page lists a forthcoming December 2026 workshop; this survey does not describe that event as already held. The abstract was verified on the author's page; direct arXiv retrieval failed.
- [7] Leonardo N. Coregliano and Maryanthe Malliaris. *High-arity PAC learning via exchangeability*. Preprint, 16 September 2024 (v3). [Primary preprint](#). Cited for its distinct exchangeable sampling model, without asserting a later journal status.
- [8] Brendan Fong, David I. Spivak, and Remy Tuyeras. *Backprop as Functor: A compositional perspective on supervised learning*. arXiv:1711.10455, 2017; revised 2019. [Primary record](#).
- [9] G. S. H. Cruttwell, Bruno Gavranovic, Neil Ghani, Paul Wilson, and Fabio Zanasi. *Categorical Foundations of Gradient-Based Learning*. arXiv:2103.01931v2, 2021. [Primary preprint](#).
- [10] Bruno Gavranovic, Paul Lessard, Andrew Joseph Dudzik, Tamara von Glehn, Joao Guilherme Madeira Araujo, and Petar Velickovic. *Position: Categorical Deep Learning is an Algebraic Theory of All Architectures*. ICML 2024, PMLR 235:15209-15241. [Proceedings and paper](#).
- [11] Tobias Fritz. *A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics*. Advances in Mathematics 370:107239, 2020. [Primary preprint and publication metadata](#). DOI: [10.1016/j.aim.2020.107239](#).
- [12] Tobias Fritz, Tomas Gonda, Paolo Perrone, and Eigil Fjeldgren Rischel. *Representable Markov Categories and Comparison of Statistical Experiments in Categorical Probability*. Theoretical Computer Science 961:113896, 2023. [Primary preprint](#). DOI: [10.1016/j.tcs.2023.113896](#).
- [13] Paolo Perrone. *Markov Categories and Entropy*. IEEE Transactions on Information Theory 70(3), 2024. [Primary preprint](#). DOI: [10.1109/TIT.2023.3328825](#).
- [14] Andreas Maurer. *A Note on the PAC Bayesian Theorem*. 2004. [Primary preprint](#).
- [15] Pierre Alquier. *User-friendly Introduction to PAC-Bayes Bounds*. Foundations and Trends in Machine Learning 17(2):174-303, 2024; arXiv revision v6, 28 February 2025. [Primary preprint](#); [full text](#). DOI: [10.1561/2200000100](#).

[16] Jean-Claude Belfiore and Daniel Bennequin. *Topos and Stacks of Deep Neural Networks*. arXiv:2106.14587v3, 16 June 2022. [Primary preprint](#). Cited for its proposed semantic framework, not as an independently checked PAC characterization.

[17] Jacob Lurie. *Higher Topos Theory*. Annals of Mathematics Studies 170, Princeton University Press, 2009; author’s maintained manuscript. [Author’s full text](#), especially Chapter 6 on descent and infinity-topoi.

[18] Sridhar Mahadevan. *A Categorical Research Program for AI: Shared Glossary*. Author-hosted research program, accessed 20 October 2026 UTC. [PACC learning and related terminology](#). No peer-review or finite-sample theorem status is inferred from this glossary.

Preparation note. This survey was prepared with AI assistance. It is a literature synthesis with short mathematical derivations, not a proof-assistant formalization or an empirical study. Primary-source abstracts were used for some broad descriptions, as identified above. The propositions in Section 6 and the mixture estimate in Section 7 are elementary consequences of their explicitly stated assumptions.