

Homotopy Fibers and Observation-Preserving Semantics for Finite-Precision Training

Anonymous

21 October 2026

Abstract

Finite-precision evaluation and automatic pullbacks need distinct semantics. We prove an obstruction to recovering pullbacks from forward values, model hidden optimizer dynamics by homotopy fibers, and give a directed semantics preserving protected traces. A presheaf infinity-topos supports a policy-descent criterion. Universality claims are explicitly resource-relative.

1. Question, provenance, and scope

Litman [1] reports zero-valued losses with nonzero automatic pullbacks and finite-state optimizer dynamics. We ask: **which equivalences preserve training behavior?** We do not reproduce the reported training experiments.

We apply standard localization and descent [5,6] to operational pullbacks and observed training systems.

Fix a numerical configuration q : representation, rounding mode, device semantics, compiler configuration, accumulation order, and optimizer. Executions are deterministic with fixed tensor shapes and explicit state, including generators and data schedules. Admitted exceptional outcomes require distinguished tags. Our numerical witness uses separate binary32 operations with round-to-nearest, ties-to-even. Other execution configurations require validation. PyTorch documents both graph-based reverse differentiation [2] and configuration-dependent floating-point behavior [3].

2. Operational pullbacks are additional structure

Definition 2.1 (Pullback-program category). An object is a pair (X, C_X) of a value set and a cotangent-carrier set. An arrow to (Y, C_Y) is

$$p = (f_p, b_p), \quad f_p : X \rightarrow Y, \quad b_p : X \times C_Y \rightarrow C_X.$$

Here b_p is specified operationally, without a derivative claim. Define

$$(g, c) \circ (f, b) = (g \circ f, (x, \zeta) \mapsto b(x, c(f(x), \zeta))),$$

with identity $(\text{id}, (x, \xi) \mapsto \xi)$. Denote this category by PB and its forward-forgetting functor by $U : \text{PB} \rightarrow \text{Set}$, $U(f, b) = f$.

Lemma 2.2. These operations define a category.

Proof. Both parenthesizations of three backward compositions yield

$$b_p(x, b_r(f_p(x), b_t(f_r(f_p(x)), \omega))).$$

Forward associativity and both identity laws are ordinary substitution laws. $\square$

This lens-shaped structure assumes no linearity or differential axioms. With real vector spaces, smooth f , and $b(x, \xi) = Df(x)^T \xi$, the reverse chain rule gives a familiar special case; reverse derivative categories supply stronger axiomatizations [4].

For a deterministic, terminating tensor expression, evaluate its tape under q and then execute its prescribed reverse rules. Saved intermediates are recovered by that forward execution, giving (f_p, b_p) . Serial composition obeys Definition 2.1 when tape evaluation and reverse accumulation order are preserved. This serial-fragment interpretation does not cover aliasing, parallel accumulation, effects, or higher-order differentiation.

Lemma 2.3 (A numerical witness). Let $C = 2^{30}$ in binary32 and

$$H(u) = \text{fl}(\text{fl}(C + u) - C).$$

For representable $-32 < u < 64$, $H(u) = 0$. Under the usual reverse rules for addition and subtraction of constants, its operational pullback is $b_H(u, \xi) = \xi$, whereas a constant-zero expression has pullback zero.

Proof. The adjacent representable numbers below and above C are $C - 64$ and $C + 128$. Their midpoint bounds are $C - 32$ and $C + 64$. The stated interval therefore rounds $C + u$ to C , and subtraction gives zero. Each prescribed reverse rule passes its incoming cotangent unchanged, so their composite is the identity. The reverse rule for a constant is zero. $\square$

The asymmetric spacing makes $|u| < 32$ a sufficient symmetric bound; $|u| < 64$ is insufficient. This calculation uses ordinary numerical analysis [7]. The operational pullback is not the derivative of the locally constant rounded map.

Theorem 2.4 (Failure of forward extensionality). On the common restricted input domain, the prescribed pullback is not determined by the rounded forward map.

Proof. $UH = UZ$, but $b_H(0, 1) = 1 \neq 0 = b_Z(0, 1)$. A recovery function would assign both answers to the same forward map. $\square$

Corollary 2.5 (A distinguishing training context). Starting at parameter zero, one plain SGD step of learning rate one distinguishes the two expressions: the resulting parameter is -1 for H and 0 for Z , although both initial forward results are zero.

Proof. Substitute their pullbacks with seed one in $\theta' = \theta - b(\theta, 1)$. These small integer operations are exact in binary32. $\square$

This one-step witness identifies information lost by the forward quotient; it neither invalidates syntactic AD nor establishes indefinite repeated execution.

3. Complete dynamics and protected observations

A complete checkpoint has the form

$$s = (\theta, m, d, r, k, \dots), \quad \Phi_p : S_p \rightarrow S_p,$$

where m is optimizer memory, d the data/schedule state, r generator state, and k control state. Interior checkpoints also retain tape and phase. Termination and errors have tagged absorbing outcomes. Unmodeled divergence is outside this total-step model.

Definition 3.1. Fix a reported-observation alphabet O and a protected alphabet P . A system is

$$X = (S_X, I_X, \Phi_X, o_X, \pi_X),$$

with initial-state subset I_X , total deterministic transition Φ_X , reported observation $o_X : S_X \rightarrow O$, and protected observation $\pi_X : S_X \rightarrow P$. Its protected trace language is

$$\text{Tr}(X) = \{((o_X, \pi_X)(\Phi_X^n(s)))_{n \geq 0} : s \in I_X\}.$$

Choose P to retain policy-relevant parameters, writes, memory, and outcomes. A loss-only language $\text{Tr}_o(X)$ forgets P .

Definition 3.2 (Relative category of systems). A morphism $h : X \rightarrow Y$ is a state map with

$$h(I_X) \subseteq I_Y, \quad h\Phi_X = \Phi_Y h, \quad o_Y h = o_X, \quad \pi_Y h = \pi_X.$$

Let $\mathcal{D}$ be the resulting category, in a fixed universe. Let W consist of those morphisms satisfying $h(I_X) = I_Y$. Thus W preserves the chosen labels and covers all initial behaviors; it includes identities and is closed under composition.

Theorem 3.3 (Trace preservation and localization). Every $h \in W$ gives $\text{Tr}(X) = \text{Tr}(Y)$. The trace-language functor factors through the infinity-categorical localization

$$\mathcal{L} = \mathcal{D}[W^{-1}].$$

Equivalent objects of $\mathcal{L}$ satisfy the same predicates of protected trace languages, including safety and liveness.

Proof. Induction gives $h\Phi_X^n = \Phi_Y^n h$. Label preservation matches the trace of each initial state; the equality $h(I_X) = I_Y$ supplies a preimage for each initial state of Y . For a general morphism, the same argument gives inclusion of languages. Hence traces define a functor to the poset of subsets of $(O \times P)^{\mathbb{N}}$, ordered by inclusion, taking W to equalities. The localization universal property [5] gives the factorization. Equivalences then map to equal languages. $\square$

System comparisons, not physical time steps, are inverted.

Proposition 3.4 (The coarse model is truncated). If $\mathcal{D}$ contains all systems of Definition 3.1, $\mathcal{L}$ is equivalent to the nerve of the trace-language poset. Its core is therefore zero-truncated.

Proof. Put $\Lambda = O \times P$. For a language L , define $B(L)$ with state set $\Lambda^{\mathbb{N}}$, shift transition, first-letter labels, and initial set L . Inclusions act by the identity state map. Then $\text{Tr } B = \text{id}$. The behavior map sending a state to its infinite labelled trace is a natural map $X \rightarrow B \text{Tr}(X)$ belonging to W . It becomes a natural equivalence after localization, giving inverse equivalences. $\square$

4. The homotopy fiber of an innocent observation

Let $\mathcal{M} = \mathcal{L}^{\simeq}$ be the maximal infinity-groupoid of systems and their equivalences. Loss-language observation defines

$$\text{Obs} : \mathcal{M} \rightarrow \mathcal{O},$$

where $\mathcal{O}$ is the discrete space of subsets of $O^{\mathbb{N}}$. Theorem 3.3 and trace projection define this map. For a reported language ℓ , its homotopy fiber is

$$\mathcal{F}_{\ell} = \mathcal{M} \times_{\mathcal{O}}^h \{\ell\} \simeq \sum_{X:\mathcal{M}} (\text{Obs}(X) = \ell).$$

The identity-type witness records observational equality [8]. Proposition 3.4 makes this coarse fiber zero-truncated; a finer model retains implementation symmetries below.

Example 4.1 (Two components in one fiber). Take $S = \{0, 1\}$, $I = \{0\}$, $o = 0$, and $\pi(s) = s$. Let A have identity transition and B have transition $s \mapsto 1 - s$. Both have loss language $\{0^{\omega}\}$. Their protected traces are respectively 0^{ω} and $0101 \dots$. They are inequivalent in $\mathcal{L}$ by Theorem 3.3, and hence lie in different components of $\mathcal{F}_{\{0^{\omega}\}}$.

These abstract machines require separate PyTorch realizability proofs.

Proposition 4.2 (Obstruction to observation-only classification). Let $Q : \mathcal{M} \rightarrow \{0, 1\}$ be an equivalence-invariant security predicate. If $X, Y \in \mathcal{F}_{\ell}$ have $Q(X) \neq Q(Y)$, there is no map $\bar{Q} : \mathcal{O} \rightarrow \{0, 1\}$ with $Q \simeq \bar{Q} \text{Obs}$.

Proof. Such a factorization would give $Q(X) = \bar{Q}(\ell) = Q(Y)$. $\square$

For Example 4.1, choose $Q(X) = 1$ exactly when no trace reaches protected label one. It separates A and B . Even the entire reported loss history therefore need not decide this policy. Observing a finite prefix provides still less information.

Proposition 4.3 (A finer fiber has genuine loops). Instead invert only system isomorphisms, giving the core groupoid $\mathcal{M}_{\text{iso}}$. Let X have two states, both initial, identity transition, and constant reported and protected labels. Its component in the constant-observation fiber is BC_2 , with $\pi_1 = C_2$.

Proof. Both permutations of the two states preserve every structure map, and these are all its automorphisms. The connected component of an object in a groupoid is equivalent to the one-object groupoid of its automorphisms. Since observation is discrete and constant on this component, the fiber retains it. $\square$

This loop is a hidden-state relabelling symmetry, not a time step or evidence of maliciousness. The coarse map collapses X to a one-state machine; the finer semantics retains its symmetry.

5. Directed executions and a genuine infinity-topos

Lemma 5.1 (Why ordinary homotopy type can erase the question). The category of finite execution prefixes of a system, including the empty prefix and ordered by extension, has contractible classifying space.

Proof. The empty prefix is initial. Its unique maps to every object give a natural transformation from the constant functor to the identity, contracting the nerve after realization. $\square$

Both safe and unsafe prefix trees can therefore have contractible classifying spaces. Retain direction, labels, and observation maps: homology of the unlabelled spaces is insufficient.

Definition 5.2 (Prefix semantics). Let $\mathcal{C}$ be the small prefix poset of finite Λ -words, including the empty word. Define

$$\mathcal{E} = \text{Fun}(\mathcal{C}^{\text{op}}, \mathcal{S}),$$

where $\mathcal{S}$ is the infinity-category of spaces. This presheaf infinity-category is an infinity-topos [6]. The prefix poset itself is not asserted to be a topos.

For a system X , define the discrete-valued presheaf

$$\mathcal{B}_X(w) = \{s \in I_X : w \text{ is a prefix of the trace from } s\}.$$

For $v \preceq w$, restriction is the inclusion $\mathcal{B}_X(w) \hookrightarrow \mathcal{B}_X(v)$. The root retains initial states and restrictions retain temporal compatibility. These objects are zero-truncated.

Proposition 5.3 (A safety subobject). Let $D \subseteq P$ be forbidden labels. Set $\mathcal{A}_D(w) = *$ when no protected letter of w belongs to D , and $\mathcal{A}_D(w) = \emptyset$ otherwise. This is a subobject of the terminal presheaf. A map $\mathcal{B}_X \rightarrow \mathcal{A}_D$ exists exactly when no execution from I_X reaches a label in D .

Proof. A good word has only good prefixes, so restriction maps exist. A map to $\mathcal{A}_D$ exists precisely when every inhabited $\mathcal{B}_X(w)$ has a good word w . Every bad occurrence has a finite witnessing prefix; conversely, a bad prefix is part of an execution by the definition of $\mathcal{B}_X$. $\square$

Forgetting P gives a functor $\rho : \mathcal{C} \rightarrow \mathcal{C}_o$ and a precomposition functor ρ^* between the associated presheaf infinity-topoi. Precomposition preserves limits and colimits and has a right Kan extension right adjoint, so it is an inverse-image functor of a geometric morphism. Policy descent requires an additional condition.

Theorem 5.4 (Policy descent). In an infinity-topos, let $e : \mathcal{B} \rightarrow \mathcal{V}$ be an effective epimorphism representing an observation, replacing its codomain by its image if necessary. A subobject $A \hookrightarrow \mathcal{B}$ descends to $\mathcal{V}$ exactly when

$$\text{pr}_1^* A = \text{pr}_2^* A \quad \text{in } \text{Sub}(\mathcal{B} \times_{\mathcal{V}}^h \mathcal{B}).$$

When it exists, the descended subobject is unique.

Proof. Necessity follows by pulling back along $e \text{pr}_1 \simeq e \text{pr}_2$. For sufficiency, classify A by $\chi_A : \mathcal{B} \rightarrow \Omega$. The equality says that its two pullbacks to the kernel pair agree. Since Ω is zero-truncated, the further equalities needed on the Čech nerve have unique coherence when they exist. Effective epimorphisms

are effective descent morphisms in an infinity-topos [6]; therefore χ_A descends uniquely to $\mathcal{V} \rightarrow \Omega$. Pulling back the universal subobject gives the required policy. $\square$

At the set level this says exactly that the policy is constant on each observation fiber. In a context containing the two runs in Example 4.1, the good-run subobject fails the kernel-pair equality for loss-only observation. Relational policies can additionally use hyperproperties [9]; bad-label reachability already supplies our counterexample.

6. Compiler refinement and numerical margins

Proposition 6.1 (An exact validation contract). Suppose a translation from X to Y supplies a state map α satisfying all equations of Definition 3.2 and $\alpha(I_X) = I_Y$. Then the two systems become equivalent in $\mathcal{L}$ and have identical protected trace languages.

Proof. These conditions state that $\alpha \in W$. Apply localization and Theorem 3.3. $\square$

An equality of scalar forward outputs proves only one component of this contract. Memory, saved values, accumulation order, control, and policy labels must also be respected; no existing compiler is certified here.

For approximate maps F, G , assuming F is K -Lipschitz and $\|F(y) - G(y)\| \leq \delta$ on an invariant domain, induction gives

$$\|x_n - y_n\| \leq K^n \varepsilon_0 + \delta \sum_{j=0}^{n-1} K^j.$$

Preserving Boolean labels additionally requires balls of these radii around x_n to remain in its policy region. A distance tolerance alone gives no such guarantee. Respecting floating-point order also does not force an unavoidable vulnerability: richer observations, restricted fragments, and checked transitions remain possible.

7. Resource-relative universality and decidability

Proposition 7.1 (Fixed finite systems). If the complete state space is finite, transitions are total and effectively executable, and forbidden labels are decidable, bad-label reachability from a given initial state is decidable after at most $|S|$ distinct states have been examined.

Proof. A forbidden state is a witness. Otherwise a state repeats within $|S| + 1$ visits. Determinism makes every subsequent state repeat the same cycle, so no new forbidden state can occur. $\square$

All storage, control counters, generator states, and environmental inputs must satisfy the finiteness assumption. Neither a finite tensor nor a finite toroidal automaton establishes unbounded universality.

Theorem 7.2 (Conditional unbounded obstruction). Suppose a class of training systems has a uniform effective encoding $(M, x) \mapsto X_{M,x}$ of arbitrary Turing-machine computations, with unbounded accessible storage or an effective unbounded family of compatible simulations. Suppose a decidable protected label is reached exactly when M halts on x . Then no total algorithm decides bad-label reachability throughout that class.

Proof. Composing a proposed decider with the encoding decides the halting problem, contradicting Turing's undecidability result [10]. $\square$

The encoding and halting correspondence are hypotheses. Compatible simulations must continue effectively as resources grow. One automaton step is not an entire unbounded computation. Changing optimizers changes Φ and requires a new realization proof; SGD and Adam are not identified.

8. Reproducibility and limits

The artifact provides `verify_semantics.py` and `verification.json`. Python 3.12.14 with NumPy 2.3.5 on x86_64 CPU passed all checks:

- 262,144 pullback-composition associativity triples on the complete 64-arrow carrier;
- 17,066 finite-state reachability instances;
- 67,332 observation-policy configurations, including 15,658 failing descent;
- 1,752 commuting surjections with protected Boolean labels;
- the binary32 boundary witness and two-state constant-report counterexample.

The numerical checks use an independently specified reverse rule, not PyTorch or GPU execution. Enumeration verifies finite instances, not localization or effective descent. The mathematical proofs have not been checked in a proof assistant. No deployment, accuracy, stability, or complete-detector claim is made.

Authorship and automation. The owner authorized anonymous preparation and submission. OpenAI ChatGPT/Codex assisted with drafting, mathematical analysis, reference verification, implementation, and artifact checks. The manuscript is submitted for editorial evaluation; no independent referee report or acceptance is asserted.

References

1. Elon Litman. *How to Run Turing-Complete Malware in PyTorch’s Backward Pass*. 20 October 2026. [Original article](#). Motivating technical report.
2. PyTorch contributors. *Autograd mechanics*. [Official documentation](#). Accessed 21 October 2026.
3. PyTorch contributors. *Numerical accuracy*. [Official documentation](#). Accessed 21 October 2026.
4. Robin Cockett, Geoffrey Cruttwell, Jonathan Gallagher, Jean-Simon Pacaud Lemay, Benjamin MacAdam, Gordon Plotkin, and Dorette Pronk. *Reverse derivative categories*. CSL 2020; [arXiv:1910.07065](#).
5. Vladimir Hinich. *Dwyer-Kan localization revisited*. Homology, Homotopy and Applications 18(1), 27-48 (2016). [arXiv:1311.4128](#).
6. Jacob Lurie. *Higher Topos Theory*. Annals of Mathematics Studies 170, Princeton University Press (2009), Chapters 5-6. [Author’s text](#).
7. David Goldberg. *What every computer scientist should know about floating-point arithmetic*. ACM Computing Surveys 23(1), 5-48 (1991). [doi:10.1145/103162.103163](#).
8. The Univalent Foundations Program. *Homotopy Type Theory: Univalent Foundations of Mathematics*. Institute for Advanced Study (2013), Chapters 2 and 4. [Official book](#).
9. Michael R. Clarkson and Fred B. Schneider. *Hyperproperties*. Journal of Computer Security 18(6), 1157-1210 (2010). [Authors’ preprint](#).
10. A. M. Turing. *On Computable Numbers, with an Application to the Entscheidungsproblem*. Proceedings of the London Mathematical Society s2-42(1), 230-265 (1936-37). [doi:10.1112/plms/s2-42.1.230](#).