Symmetries of Functional Processes under Label Noise

Abhra Chaudhuri [1] **Pedro Gomes** [1]

Abstract

Detecting label noise is backed by a vast empirical literature, yet when detection can be trusted, and how its reliability scales with computation, remain poorly understood. We introduce *functional processes* — aggregation mechanisms over estimates of the effective, data-induced function class. Functional processes exhibit symmetries under label noise whereby they asymptotically recover the clean concept $c(\mathbf{x})$, *i.e.*, the solution gradient flow would have learned under no label noise. This is achieved through the avoidance of complexity barriers — the requirement that zero-loss separators of noisy instances be of strictly higher complexity than their clean counterparts — surfacing c as a common invariant; predictions concentrate around it with shrinking uncertainty. This yields arbitrarily precise noise detection, with finite-time SGD guarantees and an $\mathcal{O}(1/\sqrt{n})$ finite-capacity rate. We empirically validate the predicted complexity signatures, asymptotic convergence, and variance collapse on standard benchmarks.

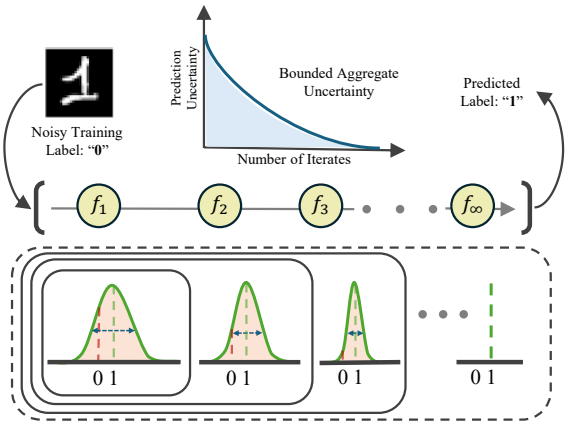

Figure 1. Despite label noise, the aggregate prediction uncertainty along a functional process shrinks monotonically and asymptotically recovers the true label, which can be leveraged for arbitrarily precise noise detection.

1. Introduction

Deep neural networks trained with noisy labels exhibit a consistent empirical pattern: they learn simple, stable structure early in training and only later begin to memorize noisy or high-complexity components (Xu et al., 2019; Rahaman et al., 2019; Xu et al., 2020). This "late learning of noise" is well documented and widely exploited in the label-noise community (Jiang et al., 2017; Han et al., 2018; Liu et al., 2020; Yuan et al., 2023; Liu et al., 2024), yet a core theoretical gap remains: when noise detection methods can be trusted, and how their reliability scales with computation, training time, or model choice, remain poorly understood. This gap is consequential. Without a principled notion of reliability, downstream decisions such as whether to relabel or retrain, how much compute is needed to reach a target confidence, and how uncertainty should shrink with additional learning effort, remain ad hoc.

We study label noise through a structural equivalence. The nominal hypothesis space $\mathcal{F}$ of a deep network is vast, but the clean training data X restricts learning to the subset reachable by gradient flow on X, an *effective, data-induced* function class $\mathcal{F}_{\text{eff}} = \mathcal{F} \,|\, X$. We call the function gradient flow reaches within $\mathcal{F}_{\text{eff}}$ the **clean concept** $c(\mathbf{x})$. Label-noise corruption replaces X with $\tilde{X}$ and $\mathcal{F}_{\text{eff}}$ with $\tilde{\mathcal{F}}_{\text{eff}} = \mathcal{F} \,|\, \tilde{X}$, generally pushing the gradient-flow solution away from c. We introduce *functional processes* (Definition 1): aggregation operators on $\tilde{\mathcal{F}}_{\text{eff}}$ whose outputs are asymptotically invariant to the noise transformation $T : X \mapsto \tilde{X}$, recovering c as the common invariant. The **transformation** is label-noise corruption; the **invariant** is the clean concept; the resulting label-noise invariance is the *symmetry of functional processes*.

Underlying this symmetry is a *complexity barrier*: separating two oppositely labeled samples that grow arbitrarily close in the effective feature space requires complexity that diverges. Noisy instances, lying near oppositely labeled clean ones, force the gradient-flow trajectory through progressively finer functional variation that clean instances do not. This finer variation does not just explain the afore-

[1]Fujitsu Research of Europe. Correspondence to: Abhra Chaudhuri <abhra.chaudhuri@fujitsu.com>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea.

mentioned late memorization of noise; it provides a separation signal between noisy and clean structure. Functional processes leverage this signal by *asymptotically avoiding* the high-complexity paths through which noise is fit. We introduce two such processes: a time-evolution process (Theorem 2) that concentrates weight on the earlier steady state where the clean concept dominates, suppressing contributions from the later memorization regime; and a hypothesis-aggregation process (Theorem 3) that averages over independent predictors whose noise-fitting deviations are decorrelated, canceling in the limit and leaving only the shared low-complexity component c. Crucially, neither process removes the barrier or attempts to learn noisy labels — each shields the recovered prediction from where the barrier would be crossed, surfacing c as the common invariant.

Illustrated in Figure 1, this enables noise detection at arbitrary precision, where reliability improves monotonically as the process approximation is refined — longer / smaller-step training for time evolution, or larger / more diverse hypothesis pools for aggregation. We provide finite-time and finite-capacity counterparts under practical SGD and modest hypothesis pools, and empirically validate the predicted complexity signatures, asymptotic convergence, and variance collapse on standard and real-world noise-detection benchmarks.

Contributions: We (i) formalize label-noise detection reliability through a structural equivalence: aggregation over the effective, data-induced function class on noisy data asymptotically recovers the gradient-flow solution on clean data, via two functional processes – time evolution and hypothesis aggregation; (ii) establish a complexity barrier that provides a separation signal between noisy and clean structure, which functional processes asymptotically avoid (iii) show that both processes asymptotically converge to the clean-concept invariant $c(\mathbf{x})$ with finite-time and finite-capacity guarantees, enabling noise detection at arbitrary precision; and (iv) empirically validate the predicted complexity signatures, asymptotic convergence, and variance collapse on standard benchmarks.

2. Related Works

Noisy Labels: Early approaches to learning under label noise rely on the small-loss principle: curriculum-based methods such as MentorNet and Co-Teaching (Jiang et al., 2017; Han et al., 2018) and loss-modeling strategies including ELR/ELR+ (Arazo et al., 2019; Liu et al., 2020) identify clean samples via early-learning dynamics. Explicit detection uses more targeted signals: Confident Learning (Northcutt et al., 2021) estimates joint label–prediction error rates, AUM (Pleiss et al., 2020) monitors margin behavior, DivideMix (Li et al., 2020) applies Gaussian mixtures, INCV (Chen et al., 2021) models annotator inconsistency, PLS-LSA+ (Zhang et al., 2024) leverages latent semantics, NoiseGPT (Wang et al., 2024) uses LLM-based probability curvature, and Delora (Zhang et al., 2025) separates clean and corrupted structure via dual low-rank adaptation. NLS (Wei et al., 2022) further shows that naïve label smoothing can intensify noise. A broader review of label-noise and symmetry-related work appears in Section A.

3. Characterization of Symmetries

We describe and analyze the symmetries that arise when training a deep neural network with noisily labeled data, and show that they can be leveraged to identify mislabeled samples at any desired level of precision. Figure 2 visually depicts the associated mechanisms.

Preliminaries: Let X be the set of training inputs. For each $\mathbf{x} \in X$, the (unobserved) clean label is $\mathbf{y}^* \in Y^*$ and the observed (possibly noisy) label is $\mathbf{y} \in Y$, with $\mathbf{y} = \mathbf{y}^*$ holding with probability $1 - \eta$ and $\mathbf{y} \neq \mathbf{y}^*$ with probability η, where η is the noise rate. We denote the clean dataset $D^* = \{(\mathbf{x}_i, \mathbf{y}_i^*)\}_{i=1}^N$ and the noisy dataset $D = T(D^*)$, where T is the label-noise transformation. Labels are one-hot unless stated otherwise, $f(\mathbf{x})$ denotes a distribution over Y, and $\hat{\mathbf{y}} = f(\mathbf{x})$ is the corresponding one-hot prediction (argmax index set to 1). All proofs appear in Section B.5.

Effective function class and clean concept: Let $\mathcal{F}$ denote a hypothesis space of measurable functions $f : X \to \Delta(Y)$. Fix an initialization $f_0 \in \mathcal{F}$ and a gradient-flow dynamics $\frac{d}{dt} f_t = -\nabla \mathcal{L}_D(f_t)$ for training set D. A learning algorithm applied to D does not range freely over $\mathcal{F}$; it reaches the gradient-flow-reachable subset, which we call the *effective, data-induced function class*:

$$\mathcal{F}_{\text{eff}}(D) \;=\; \left\{\, f_t : t \geq 0,\ \dot{f}_t = -\nabla \mathcal{L}_D(f_t) \,\right\} \;\subseteq\; \mathcal{F}.$$

The **clean concept** $c : X \to Y^*$ is the function gradient flow reaches within $\mathcal{F}_{\text{eff}}(D^*)$ on the clean dataset:

$$c \;:=\; \lim_{t \to \infty} f_t, \qquad \dot{f}_t = -\nabla \mathcal{L}_{D^*}(f_t), \quad f_0 \text{ fixed.}$$

On noisy data, gradient flow instead reaches $\mathcal{F}_{\text{eff}}(D)$, which generally does not contain c. The central question of this section is: under what aggregation can c be recovered from $\mathcal{F}_{\text{eff}}(D)$ alone?

Functional processes: A *functional process* is a family of operators $\{\Phi_\alpha\}_{\alpha \in \mathcal{I}}$ acting on $\mathcal{F}_{\text{eff}}(D)$, where each Φ_α aggregates one or more predictors in $\mathcal{F}_{\text{eff}}(D)$ and returns a predictor (or label distribution); the index set $\mathcal{I}$ parameterizes the granularity of the aggregation. Functional Process Symmetries, defined below, are invariances of Φ_α's output to the noise transformation T that emerge as the aggregation grows finer (α refining within $\mathcal{I}$).

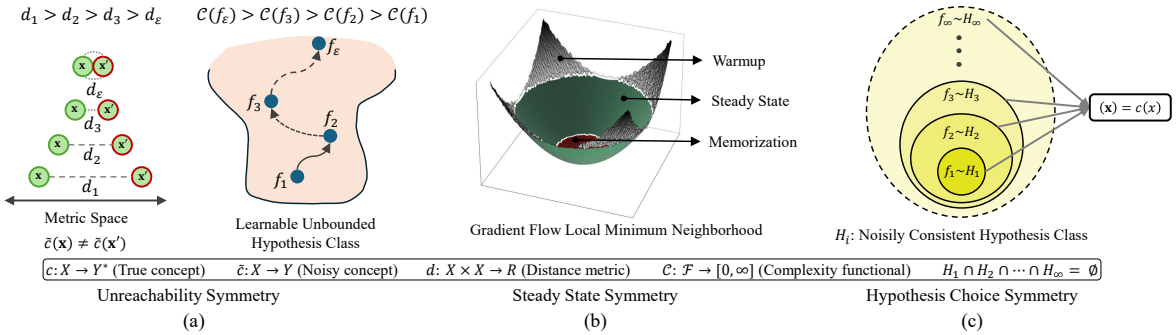

Figure 2. Mechanisms behind symmetries of functional processes arising from training with label noise. (a) **Complexity Barrier** – as two samples with conflicting labels grow closer in the effective feature space, the zero-loss classifier requires diverging complexity, producing a separation signal between noisy and clean structure. (b) **Steady State** – gradient flow on any sample spends most of its time in a low-complexity regime encoding the clean concept before transitioning to the high-complexity regime that fits its training label. (c) **Hypothesis Stability** – under noisily consistent growth of disjoint hypothesis classes, the renormalized aggregate of gradient-flow solutions cancels barrier-crossing errors, surfacing the clean concept.

Definition 1: Functional Process Symmetry

Let $T : D^* \mapsto D$ be a label-noise transformation, and let $\{\Phi_\alpha\}_{\alpha \in \mathcal{I}}$ be a functional process acting on $\mathcal{F}_{\text{eff}}(D)$. A functional I is a *Functional Process Symmetry* of $\{\Phi_\alpha\}$ with invariant c if, in the limit of fine-grained aggregation, its output recovers the clean concept independently of T:

$$\lim_{\alpha \to \alpha^*} I(\Phi_\alpha, \mathbf{x}) = c(\mathbf{x}) \quad \text{for all } \mathbf{x} \in X,$$

where $\alpha^* \in \overline{\mathcal{I}}$ denotes the limiting refinement of the index set.

Equivalently, in the limit, aggregation on noisy data D recovers what gradient flow would have reached on the clean data D^*. Functional processes thus serve as the operational mechanism by which the clean-concept invariant c becomes recoverable. The remainder of this section establishes why this recovery is possible (Observation 1 and Corollary 1.1), how it manifests for each process (Theorems 2 and 3), and how the corresponding finite-time and finite-capacity counterparts yield practical label-noise tests.

3.1. Complexity Barrier

Functional processes recover c by exploiting a structural asymmetry between clean and noisy data in $\mathcal{F}_{\text{eff}}(D)$: fitting noisy labels near oppositely labeled clean instances is strictly more expensive in complexity than fitting clean structure alone. We formalize this as a *complexity barrier*; it is the separation signal that functional processes asymptotically avoid.

Setup: Let $\mathcal{F}$ be an unbounded hypothesis class equipped with a complexity functional $\mathcal{C} : \mathcal{F} \to [0, \infty]$ satisfying the axioms in Section B.2 (measuring the expressive power of a model). Let $\mathcal{L} : \mathcal{F} \to [0, \infty)$ be a differentiable loss, and let $(f_t)_{t \geq 0}$ denote the gradient-flow trajectory:

$$\frac{d}{dt} f_t = -\nabla \mathcal{L}(f_t), \qquad f_0 \in \mathcal{F}.$$

Assume that every target $f^\star \in \mathcal{F}$ with $\mathcal{C}(f^\star) < \infty$ is reachable in finite time, i.e., there exists $T < \infty$ such that $f_T = f^\star$. Fix such an $f^\star$ with $\mathcal{C}(f^\star) \leq C < \infty$, and choose any finite partition $0 = t_0 < t_1 < \cdots < t_N = T$. Then $f^\star$ admits the finite representation:

$$f^\star = f_0 + \sum_{k=0}^{N-1} \left(f_{t_{k+1}} - f_{t_k} \right),$$

where each increment $f_{t_{k+1}} - f_{t_k}$ is a gradient-flow step on $[t_k, t_{k+1}]$.

Furthermore, empirical and theoretical analyses of spectral bias in deep neural networks (Xu et al., 2019; Rahaman et al., 2019) show that gradient flow fits low-frequency components of the target before high-frequency components. Consequently, the complexity of the gradient flow iterates increases monotonically:

$$t_1 < t_2 \implies \mathcal{C}(f_{t_1}) \leq \mathcal{C}(f_{t_2}),$$

with higher-frequency (fine-variation) structure appearing only in later increments $f_{t_{k+1}} - f_{t_k}$. Thus, any target of bounded complexity in a learnable unbounded hypothesis class is obtained by a finite accumulation of frequency components along the gradient-flow trajectory, progressing monotonically from coarse to fine structures.

Effective Feature Space. Let $X \subset \mathbb{R}^d$ be equipped with a metric $d(\cdot, \cdot)$ in the **effective feature space** (Section B.1) – the subspace of X that drives learning. Consider two samples $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{x}', \mathbf{y}') \in X \times Y$ such that $\mathbf{y} \neq \mathbf{y}'$.

For each $\varepsilon > 0$, suppose there exists a zero-loss classifier $f_\varepsilon \in \mathcal{F}$ satisfying:

$$f_\varepsilon(\mathbf{x}) = \mathbf{y}, \qquad f_\varepsilon(\mathbf{x}') = \mathbf{y}', \qquad d(\mathbf{x}, \mathbf{x}') \le \varepsilon.$$

By Lemma 1, as ε decreases, such a classifier must be strictly more oscillatory; by Lemma 2:

$$d(\mathbf{x}, \mathbf{x}') \to 0 \quad \Longrightarrow \quad \mathcal{C}(f_\varepsilon) \to \infty.$$

Thus, in the limit $\mathbf{x}' \to \mathbf{x}$ with $\mathbf{y} \neq \mathbf{y}'$, any zero-loss solution $f^\star$ must satisfy $\mathcal{C}(f^\star) = \infty$.

Observation 1 (Complexity Barrier)**.** *Consider some* $(\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')$ *with* $\mathbf{y} \neq \mathbf{y}'$*. For each* $r > 0$*, choose* $\mathbf{x}_r$ *with* $d(\mathbf{x}, \mathbf{x}_r) = r$ *and define* $\mathcal{F}_r = \{f \in \mathcal{F} : f(\mathbf{x}) = \mathbf{y},\ f(\mathbf{x}_r) = \mathbf{y}'\}$ *and*

$$C(r) = \inf\{\mathcal{C}(f) : f \in \mathcal{F}_r\}.$$

Then, $0 < r_1 < r_2$ *implies* $C(r_1) \ge C(r_2)$*, and* $\lim_{r \to 0} C(r) = \infty$.

The proof of this is immediate from Lemma 2: the Oscillation Growth Lemma (Lemma 1) establishes that separating two oppositely labeled points at distance $r_1 < r_2$ requires a strictly more oscillatory function; the Monotonicity axiom converts this to $C(r_1) \ge C(r_2)$; and the divergence follows from the compactness–continuity argument in Lemma 2.

Because noisy instances lie near oppositely labeled clean ones, fitting them forces gradient flow through progressively finer (higher-complexity) functional variation than fitting clean structure alone – matching the known low-to-high frequency fitting behavior of deep networks (Xu et al., 2019; Rahaman et al., 2019). This finer variation does not just explain late memorization; it provides a *separation signal* between noisy and clean structure in the complexity profile of the gradient-flow trajectory.

Corollary 1.1 (Unreachability)**.** *Let* $\mathcal{F}'$ *be any hypothesis class under gradient flow learning, producing a trajectory* $(f_t)_{t \ge 0} \subset \mathcal{F}'$*. For samples* $(\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')$ *with* $\mathbf{y} \neq \mathbf{y}'$*, define for* $r > 0$ *a point* $\mathbf{x}_r$ *with* $d(\mathbf{x}, \mathbf{x}_r) = r$ *and set:*

$$C(r) = \inf\{\mathcal{C}(f) : f(\mathbf{x}) = \mathbf{y},\ f(\mathbf{x}_r) = \mathbf{y}'\}.$$

By Lemma 2, $\lim_{r \to 0} C(r) = \infty$*. Consequently, for any solution* $f^\star \in \mathcal{F}'$ *that* (f_t) *converges to:*

$$\begin{aligned} f^\star(\mathbf{x}) = \mathbf{y}, \quad & f^\star(\mathbf{x}') = \mathbf{y}' \\ \Longleftrightarrow \quad & \forall K < \infty,\ \sup_{f \in \mathcal{F}'} \mathcal{C}(f) \ge K, \end{aligned}$$

i.e., $f^\star$ *has zero loss iff* $\mathcal{F}'$ *is not bounded in complexity.*

3.2. Steady-State Symmetry

The first functional process aggregates predictions along a gradient-flow trajectory. Two facts drive what we call the steady-state symmetry. First, clean samples, drawn from class-conditional distributions with nonzero variance (Natarajan et al., 2013; Sukhbaatar et al., 2015; Patrini et al., 2017), continue to refine the model even after the clean concept $c(\mathbf{x})$ is encoded, since residual loss depends on their distance from the learned mean (Lemma 4); noisy samples, modeled as point masses, exert no further gradient signal once the noisy concept $\tilde{c}(\mathbf{x})$ is reached. Consequently, $c(\mathbf{x})$ is learned strictly before $\tilde{c}(\mathbf{x})$ (Lemma 5). Second, although f_t drifts toward $\tilde{c}$ as complexity grows, its argmax prediction at $\mathbf{x}$ remains pinned to $c(\mathbf{x})$ throughout this growth phase (Lemma 6): a flip would force immediate convergence and halt the observed monotone complexity increase. Together, these imply that $f_t(\mathbf{x}) = c(\mathbf{x})$ holds over the dominant portion of the trajectory by Lebesgue measure (Lemma 7) – the low-complexity steady state that the process must concentrate on while avoiding the later high-complexity memorization regime.

Theorem 2 (Steady-State Symmetry)**.** *For any sample* $\mathbf{x} \in X$ *learned by a model* f_t *under gradient flow over* $t \in [0, T]$*,*

$$c(\mathbf{x}) = \frac{1}{T} \int_{t=0}^{T} z(\hat{\mathbf{y}}_t)\, f_t(\mathbf{x})\, dt$$

where $\hat{\mathbf{y}}_t = f_t(\mathbf{x})$ *and* $z(\hat{\mathbf{y}}_t)$ *is the fraction of samples in* X *with label* $\hat{\mathbf{y}}_t \in Y$ *that have zero training loss under* f_t*, i.e.,*

$$z(\hat{\mathbf{y}}_t) = \frac{1}{|I_{\hat{\mathbf{y}}_t}|} \sum_{i \in I_{\hat{\mathbf{y}}_t}} \mathbf{1}[\,\mathcal{L}_t(\mathbf{x}_i, \mathbf{y}_i) = 0\,],$$

where $I_{\mathbf{y}}$ *is the index set of samples in* X *with label* $\mathbf{y} \in Y$*, i.e.,* $I_{\mathbf{y}} = \{\, i \in \{1, \ldots, N\} : \mathbf{y}_i = \mathbf{y} \,\}$.

The weighting $z(\hat{\mathbf{y}}_t)$ approaches 1 in the steady-state regime (clean concept fitted for class $\hat{\mathbf{y}}_t$) and stays below 1 in the pre-concept and memorization regimes. The time integral is therefore dominated by the low-complexity steady state, asymptotically suppressing the high-complexity trajectory segments where the barrier would be crossed.

Concept Recovery Test (CRT): We directly leverage Theorem 2 to identify the regime in the training trajectory from which the clean concept can be recovered. For a given sample $\mathbf{x}$, at each epoch i of SGD we record the model prediction $\hat{\mathbf{y}}_i = f_i(\mathbf{x})$. For each such predicted label, we compute $z(\hat{\mathbf{y}}_i)$ by considering all samples in X with label $\hat{\mathbf{y}}_i$ and taking the fraction that are correctly classified under f_i. To aggregate evidence across training, for every class $\mathbf{y} \in Y$ we define the score:

$$S(\mathbf{y}) = \sum_{i=1}^{T} z(\hat{\mathbf{y}}_i)\, \mathbf{1}[\hat{\mathbf{y}}_i = \mathbf{y}].$$

The recovered class is then given by:

$$\hat{\mathbf{y}}^* = \arg\max_{\mathbf{y} \in Y} S(\mathbf{y}).$$

If $\hat{\mathbf{y}}^* \neq \mathbf{y}$, where $\mathbf{y}$ is the training label of $\mathbf{x}$, we predict $\mathbf{x}$ as being noisily labeled. The number of epochs T (and the associated learning rate) can be tuned to achieve any desired level of guarantee: the larger T and the smaller the step size (*i.e.*, the closer SGD gets to gradient flow), the more reliably the high-complexity regime is suppressed and the closer the score concentrates on $c(\mathbf{x})$. We provide a finite-time, finite-step-size convergence guarantee for CRT via Theorem 4 in Section B.6.

3.3. Hypothesis Choice Symmetry

The second functional process aggregates predictions across independently trained predictors drawn from disjoint hypothesis classes. The hypothesis choice symmetry is the invariance of the predicted label to which consistent (zero training loss) hypothesis class is chosen, when provided with clean training ground-truths. Let $\mathbb{H} = \{H_1, H_2, ...H_n\}$ be a set of *disjoint* hypothesis classes that are consistent over (X, Y^*), where $n = |\mathbb{H}|$. Due to the consistency condition, if Y^* is available during training,

$$\forall i \in \{1, 2, ..., n\}, f_i \sim H_i, \quad f_i(\mathbf{x}) = \hat{\mathbf{y}},$$

i.e., the predicted label $\hat{\mathbf{y}}$ is the same across models from all hypothesis classes in $\mathbb{H}$, where f_i is the gradient-flow solution of learning over H_i with the clean dataset.

Since the true labels Y^* are not available during training and only the noisy labels Y are, we relax the strict consistency restriction on $\mathbb{H}$ as follows:

Definition 2 (Noisy Consistency)**.** If a set of hypothesis classes $\mathbb{H}$ has consistent generalization over X, Y^*, then, for all member hypothesis classes $H \in \mathbb{H}$:

$$\varepsilon_H^{\min} \leq \eta$$

where $\varepsilon_H^{\min}$ is the lowest error rate of H and η is the noise rate in the train set labels Y, *i.e.*, the lowest error rate in each $H \in \mathbb{H}$ is strictly less than or equal to η.

Noisy consistency requires that the best possible classifier in every member of $\mathbb{H}$ should at least be able to consistently (*i.e.*, with zero loss) learn the noisy training labels Y, which implies a maximum error rate of η on the true labels Y^*. Under this condition, the Hypothesis Choice Symmetry takes the following form:

Theorem 3 ($\mathbb{H}$-Stability)**.** *If $\mathbb{H}$ grows in a noisily consistent manner, then for any sample $\mathbf{x} \in X$:*

$$c(\mathbf{x}) = \lim_{n \to \infty} f_1(\mathbf{x}) \oplus f_2(\mathbf{x}) \oplus ... \oplus f_n(\mathbf{x}),$$

where $\oplus$ is the renormalized summation operation given by:

$$S_0 = 0; \quad S_i = \frac{(i-1) \cdot S_{i-1} + f_i(\mathbf{x})}{i},$$

where $S_i = f_1(\mathbf{x}) \oplus f_2(\mathbf{x}) \oplus ... \oplus f_i(\mathbf{x})$.

Corollary 3.1. *If the $\mathbb{H}$-stable limit for a sample $\mathbf{x} \in X$ with true label $\mathbf{y}^*$ does not converge to its training label $\mathbf{y}$, then $\mathbf{y} \neq \mathbf{y}^*$, i.e., the label $\mathbf{y}$ for $\mathbf{x}$ is noisy.*

Each individual predictor f_i approaches the complexity barrier along its own path through H_i, yielding a residual error $e_i = f_i - c$ that depends on the particular high-complexity region H_i's gradient-flow trajectory traverses. Under noisily consistent growth (formalized in the proof as asymptotically decorrelated residuals), these barrier-crossing components point in increasingly orthogonal directions and cancel under renormalized summation, leaving only the shared low-complexity component c. As $\mathbb{H}$ grows, there are finitely many functional forms that merely fit the noisy labels with error rate η, but infinitely many that reduce the error below η by aligning more closely with c; in the infinite limit, the latter dominate the renormalized summation. We provide a finite-capacity convergence guarantee for $\mathbb{H}$-Stability via Theorem 5 in Section B.7.

$\mathbb{H}$-Stability Test (HST): Since convergence to the ground-truth can only be guaranteed as n approaches infinity, for practical purposes we provide two variants of the $\mathbb{H}$-stability test that can be performed with both large and small n. For large n, since Theorem 3 approximately holds, we leverage the fact that for any sample $\mathbf{x} \in X$, if the finite renormalized sum converges to a low-entropy class distribution $\hat{\mathbf{y}}^*$ such that

$$\hat{\mathbf{y}}^* = f_1(\mathbf{x}) \oplus f_2(\mathbf{x}) \oplus ... \oplus f_n(\mathbf{x}) \neq \mathbf{y},$$

then the training label $\mathbf{y}$ is noisy. For small n, we propose a stricter version where all samples $\mathbf{x} \in X$ satisfying the following condition are predicted as noisily labeled:

$$\forall i \in \{1, 2, ..., n\}, f_i \sim H_i,$$
$$f_i(\mathbf{x}) = \hat{\mathbf{y}} \neq \mathbf{y} \implies \mathbf{y} \neq \mathbf{y}^*.$$

In other words, a sample $\mathbf{x}$ is identified as noisy if all classifiers from $\mathbb{H}$ agree on $\mathbf{x}$ having the same label $\hat{\mathbf{y}}$ that differs from the training label $\mathbf{y}$; the agreed label $\hat{\mathbf{y}}$ is predicted as the correct label of $\mathbf{x}$. In both cases, any desired level of guarantee on the prediction can be achieved by tuning n – larger values produce more reliable predictions.

4. Experiments

Datasets and Implementation Details: We evaluate on five benchmarks: three standard datasets from the label-noise literature – MNIST (LeCun et al., 1998), CIFAR-10, and CIFAR-100 (Krizhevsky, 2009), used widely in prior work (Han et al., 2018; Ma et al., 2018; Harutyunyan et al., 2020; Wang et al., 2024); T-Finance (Tang et al., 2022), adapted from anomaly detection to create a challenging structured-noise setting where existing baselines struggle; and Clothing1M (Xiao et al., 2015), a large-scale real-world

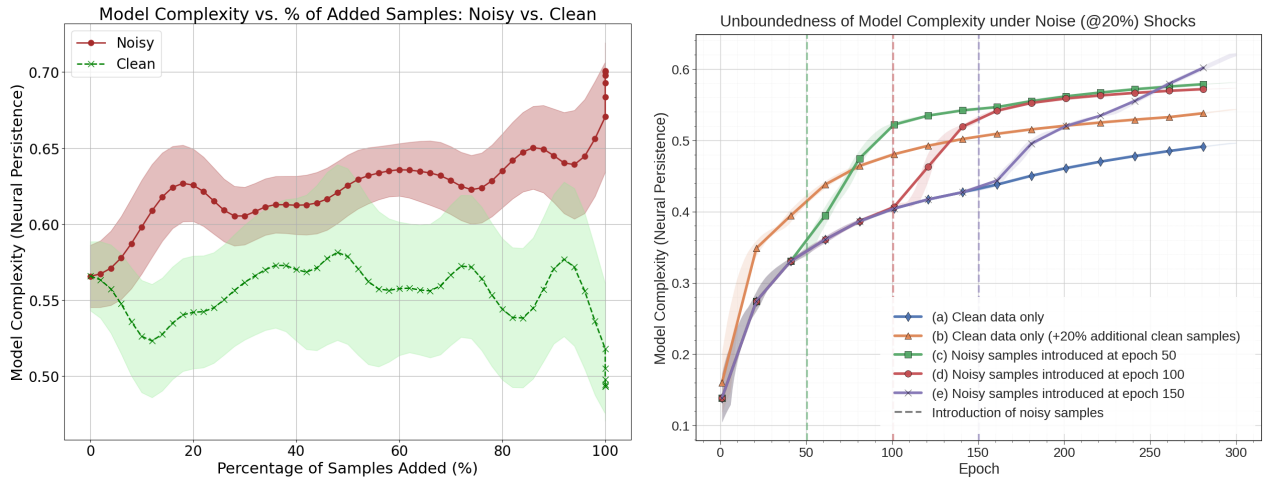

Figure 3. (Left) Complexity Barrier: The complexity cost of learning a unit amount of noise exceeds that of a unit amount of clean data. (Right) Unreachability illustrated via the unbounded growth of model complexity under noise.

benchmark with instance-dependent noise. Recent studies highlight noisy labels in real-world fraud detection (S. et al., 2024; Meng, 2025), especially in financial transaction graphs (Wang et al., 2025), where noise is sparse and highly asymmetric, *i.e.*, a few fraudulent nodes may be mislabeled as benign among many truly benign ones. Our adaptation of T-Finance reflects this scenario, with fraud labels that are correct when present but potentially missing for additional fraudulent nodes.

We position HST and CRT as meta-procedures that augment existing neural-network-based noise detectors satisfying the conditions in Section 3. We use four baslines / substrates spanning a range of detection mechanisms and report rectification error (percentage of incorrect labels after rectification), detection F1, and Recall. For each substrate we report the standalone result alongside its +HST and +CRT augmentations, with implementation details in Section C.1. Beyond rectification accuracy, we further examine the premises and implications of our theory on MNIST; due to space limits, we present core MNIST analyses in the main text and report extended experiments in Section C. We measure functional complexity using neural persistence (Rieck et al., 2019), which we found empirically to satisfy the complexity axioms (Section B.2). We use "complexity", "model complexity", and "neural persistence" interchangeably.

4.1. Complexity Cost of Learning Clean / Noisy Data

Objective and Settings: We begin by validating the claims in Section 3.1 on the cost of learning a unit amount of noise versus clean data, as predicted by the complexity barrier (Observation 1); results appear in Figure 3 (left). Starting from 20,000 samples, we progressively add clean (green) or noisy (red) samples and track model complexity at convergence. Because the complexity barrier implies unreachability (Corollary 1.1), we also verify this experimentally, shown in Figure 3 (right). We inject 20% noisy samples at epochs 50, 100, and 150 and compare the resulting complexity trajectories with those of the original clean-only dataset and a clean-only dataset augmented with an equivalent 20% clean samples at the start of training.

Observations and Analyses: Figure 3 (left) provides direct evidence for Observation 1: learning a unit of noisy data consistently incurs higher complexity than learning the same amount of clean data, and the complexity cost grows proportionally with the noise rate. Figure 3 (right) visualizes the cumulative effect of unreachability. Although a single noisy sample may not be costly, groups of noisy samples produce a sharp, sustained rise in complexity, regardless of when they are introduced. This rise always exceeds not only the clean-only trajectory but also the trajectory obtained by increasing the number of clean samples to match the total clean+noisy count, confirming that the excess complexity originates from noise. Clean trajectories never catch up to the noisy ones; even in the case where noise is added at quite a later stage in training at epoch 150, the complexity growth quickly exceeds all other baselines, reaffirming our theoretical claims around unboundedness of learning under label noise (Section 3.1). Thus, the persistent gap / differential growth rate in complexity between noisy and clean curves supports Corollary 1.1: a model with bounded complexity cannot accommodate arbitrarily high noise, as its achievable complexity saturates at a finite upper bound.

4.2. Reversal Patterns and Steady States

Objective and Settings: We seek empirical evidence for the steady states predicted by Theorem 2. Given the irregularity of neural-network loss landscapes, we instead assess steadiness via robustness to injected label noise: if a steady

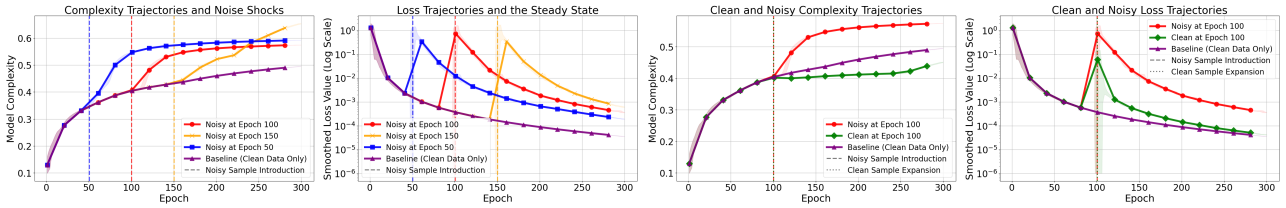

Figure 4. Complexity and Loss Trajectories: (left) steady state symmetry; (right) control experiment with clean samples.

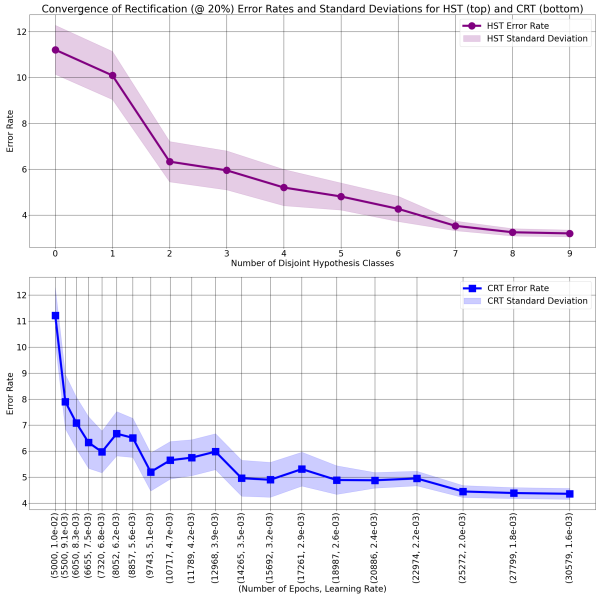

Figure 5. Convergence of HST and CRT label noise (at 20%) rectification error rates and standard deviations on MNIST.

state exists, SGD should not deviate from it when exposed to mislabeled samples. Following Section 4.1, we introduce 20% noisy samples at epochs 50, 100, and 150 and track the resulting complexity and loss trajectories against clean-only references. As a control, we also add 20% clean samples at epoch 100. Results appear in Figure 4.

Observations and Analyses: As can be seen in Figure 4 (left), training loss trajectories recover from noise-shocks, but complexity trajectories do not. The reversal of the loss trajectory to its original course points at the existence of a semantic steady state that SGD tends to return to despite being perturbed by noise. This effect is reproducible across all points in training - early and late, meaning that the steady state is sustained for the majority of training in accordance with Theorem 2. The complexity trajectories not returning to their initial state is indication that the model is accommodating the additional noisy samples, but despite that, is largely grounded in the steady state through the loss reversals. The clean control in Figure 4 (right) demonstrates our point that the observed shock and reversal and patterns are characteristic of noise and does not occur due to a mere sudden train set expansion with clean data.

4.3. Asymptotic Convergence to Symmetries

Objective and Settings: Our tests (HST and CRT) are based on the fact that their respective symmetries become recoverable in asymptotic limits. We empirically show that the sequences defined by the Steady-State Symmetry (Theorem 2) and $\mathbb{H}$-Stability (Theorem 3) converge toward these symmetries, with the target concept as the invariant. As the limiting conditions are better approximated, the target concept is recovered more accurately, enabling increasingly precise label-noise detection. HST approaches its limit by adding independent models to the predictor pool, while CRT does so by iteratively scaling training time by 1.1 and reducing the learning rate by the same factor. The resulting rectification errors at each step are shown in Figure 5.

Observations and Analyses: Both CRT and HST asymptotically converge to lower rectification error rates, and more importantly, their predictive uncertainties also decrease monotonically. Better symmetry approximation brings solutions from different random initializations closer to the shared invariant, reducing variance and supporting our claim that the underlying symmetries and the true-concept invariant emerge in the asymptotic limit, enabling arbitrarily precise noise detection.

CRT converges faster initially but exhibits diminishing returns even as the loss-landscape resolution increases; we conjecture this relates to the fractal dimension of the landscape, which has been linked to neural-network generalization (Tan et al., 2024). HST improves more slowly but steadily, saturating only near the end, likely because increasing the number of disjoint hypothesis classes progressively reduces the dimensions of the true concept not yet represented in the function space.

4.4. Reliability Amplification over Existing Detectors

Objective and Settings: We test whether HST and CRT, applied as meta-procedures *on top of* existing noise detectors, yield measurable improvements in detection reliability. Following Wang et al. (2024), we evaluate on MNIST, CIFAR-10, CIFAR-100, T-Finance, and Clothing1M under 10–40% symmetric noise and 40% asymmetric noise. For each base-

Method	Symmetric (%)				Asym. (%)
	10	20	30	40	40
NLS (ICML 2022)	6.10 ± 0.95	6.59 ± 0.86	7.11 ± 0.98	8.76 ± 1.07	8.95 ± 1.05
+Ours (HST)	**2.91 ± 0.08**	**3.20 ± 0.15**	**4.80 ± 0.22**	5.70 ± 0.38	5.11 ± 0.25
+Ours (CRT)	3.80 ± 0.11	4.36 ± 0.21	4.87 ± 0.23	**5.01 ± 0.31**	**4.85 ± 0.24**
NoiseGPT (NeurIPS 2024)	4.89 ± 0.73	5.60 ± 0.81	6.63 ± 0.90	7.21 ± 1.30	7.10 ± 1.25
+Ours (HST)	**3.76 ± 0.05**	**4.02 ± 0.11**	4.88 ± 0.37	6.15 ± 0.28	6.60 ± 0.21
+Ours (CRT)	4.22 ± 0.33	4.55 ± 0.21	**4.80 ± 0.19**	**6.08 ± 0.25**	**6.39 ± 0.18**
PLS-LSA+ (ECCV 2024)	5.15 ± 0.80	5.80 ± 0.95	7.10 ± 1.13	7.48 ± 1.21	7.50 ± 1.15
+Ours (HST)	**4.28 ± 0.07**	**4.90 ± 0.05**	**6.02 ± 0.18**	6.60 ± 0.15	6.18 ± 0.13
+Ours (CRT)	4.35 ± 0.12	5.20 ± 0.09	6.71 ± 0.22	**6.20 ± 0.25**	**6.05 ± 0.18**
Delora (ACL 2025)	5.30 ± 0.86	5.48 ± 0.71	6.40 ± 0.90	8.40 ± 1.02	8.25 ± 1.15
+Ours (HST)	**4.49 ± 0.17**	**4.91 ± 0.11**	**5.28 ± 0.16**	7.30 ± 0.23	6.86 ± 0.30
+Ours (CRT)	4.88 ± 0.15	5.26 ± 0.22	5.35 ± 0.23	**6.02 ± 0.28**	**6.30 ± 0.25**

Table 1. Label noise rectification error rate and associated standard deviation / confidence (lower is better for both) on MNIST.

line detector $D \in \{\text{NLS}, \text{NoiseGPT}, \text{PLS-LSA+}, \text{Delora}\}$, we report D, D+HST, and D+CRT. Rectification error and standard deviation over 5 runs appear in Tables 1 to 3; complementary detection F1 across benchmarks is summarized in Table 4; real-world results on Clothing1M with all three metrics appearing in Table 5. Certain key practical considerations are discussed in detail in Section C.4.

Observations: Three patterns emerge across all five benchmarks. *First*, both augmentations improve every substrate at nearly every noise level and on every metric: across the 28 (substrate, dataset, noise-type) combinations in Table 4, +HST or +CRT improves F1 over the baseline in every cell, matching the predicted asymptotic recovery of $c(\mathbf{x})$ as the symmetry approximation tightens. *Second*, standard deviations shrink uniformly across substrates, benchmarks, and metrics, often by 3–10× (e.g., 1.58–2.36 down to 0.25–0.55 on Clothing1M rectification error, with comparable shrinkage on F1 and Recall) – the direct signature of a process-level invariance, with different substrates and initializations collapsing onto the same $c(\mathbf{x})$. *Third*, HST and CRT are complementary: HST dominates at low–moderate symmetric noise and on rectification error, CRT at higher noise and on Recall. This is clearest on Clothing1M (Table 5), where HST achieves the lowest rectification error on every substrate and CRT the highest Recall. We attribute this to the two underlying symmetries: hypothesis aggregation sharpens precision when the true-concept signal is already strong, while trajectory integration over the steady-state interval commits to the dominant prediction state, favoring recall.

T-Finance and Clothing1M further stress-test the meta-method framing under highly asymmetric, instance-dependent, class-concentrated noise; the same three patterns hold, supporting the Section 3 claim that the underlying mechanism, a stable prediction window plus residual decorrelation across hypotheses, does not require i.i.d. or symmetric noise. HST and CRT are thus not alternatives to existing detectors but *reliability amplifiers*: any neural-network-based detector can be wrapped in either, and the asymptotic precision predicted by Theorems 2 and 3 manifests as both lower error and tighter confidence intervals, uniformly across substrates and noise types.

5. Conclusion and Discussions

We frame label noise through *symmetries of functional processes*: invariances of the effective, data-induced function class to label-noise corruption, realized under gradient-flow time evolution and hypothesis aggregation. A complexity barrier underlies these symmetries — derived from four minimal complexity axioms (well-definedness, continuity, monotonicity, non-degeneracy) and a standard regularity condition, rather than postulated directly. The key mechanism is geometric: separating oppositely labeled samples at vanishing distance forces strictly higher oscillation (Oscillation Growth), which monotonicity converts to diverging complexity, making zero-loss separators unreachable in bounded-capacity classes. Both processes admit asymptotic limits with the clean concept $c(\mathbf{x})$, *i.e.*, what would have been learned by gradient flow under no label noise, as the invariant, yielding finite-time bounds for time evolution under SGD and an $\mathcal{O}(1/\sqrt{n})$ rate for aggregation under bounded errors and asymptotic decorrelation; the resulting tests, CRT and HST, function as plug-in reliability amplifiers over existing detectors.

Several directions remain open. A natural next step is to move from isolated invariants to full *symmetry groups* of label noise, connecting them to classical groups in geometric deep learning (Cohen & Welling, 2016; Bronstein et al., 2021) or conserved quantities of gradient-flow dynamics (Kunin et al., 2021; Tanaka & Kunin, 2021; Zhao et al., 2023). The minimal axiomatic basis also invites a quantitative refinement of the oscillation ordering: replacing the binary relation $f \succ g$ with a graded measure of "how much more oscillatory" f is than g could yield tighter, rate-aware finite-time bounds for CRT and finite-capacity bounds for HST. Finally, the complexity barrier suggests model complexity could itself serve as a direct noise detector (Section C.3), pending per-sample-sensitive complexity surrogates that can reliably attribute oscillation growth to individual training instances.

Impact Statement

This paper advances the understanding of how neural networks behave when trained with noisily labeled data. By proposing a symmetry-based perspective, we provide a systematic way to identify quantities in neural network training that remain stable in the presence of noise. Within the limits of practicality, approximating these symmetries can enable label noise detection / rectification with arbitrary levels of precision. This, in combination with the model-agnostic nature of our results and proposed algorithms qualify our work as a step forward in improving the reliability of deep learning applied to real world data. This is further demonstrated empirically through the effectiveness of our algorithms in handling noise for tax-fraud detection, where fraudulent entities, due to their sparse nature, routinely get mislabeled as benign. To the best of our knowledge, we are not aware of any negative impacts that are specific to our work.

Acknowledgements

The authors would like to thank members of the Graph AI Team at Fujitsu Research of Europe for their ongoing feedback on this project. The authors would also like to thank all the reviewers for their time and effort in providing comments that have been greatly beneficial towards improving the clarity of this paper.

References

Altıntaş, G. S., Kwok, D., Raffel, C., and Rolnick, D. The butterfly effect: Neural network training trajectories are highly sensitive to initial conditions. In *Forty-second International Conference on Machine Learning*, 2025.

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., and McGuinness, K. Unsupervised label noise modeling and loss correction. In *ICLR*, 2019.

Bietti, A., Venturi, L., and Bruna, J. On the sample complexity of learning under invariance and geometric stability. In *NeurIPS*, 2021.

Bishop, C. M. Training with noise is equivalent to tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995. doi: 10.1162/neco.1995.7.1.108.

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Chen, P., Liao, B., Chen, G., and Zhang, S. Noise-aware learning from multiple annotators with inconsistent expertise. In *CVPR*, 2021.

Cohen, T. and Welling, M. Group equivariant convolutional networks. In *ICML*, 2016.

Deng, L., Yang, B., Kang, Z., and Xiang, Y. Invariant feature based label correction for dnn when learning with noisy labels. *Neural Networks*, 2024.

Entezari, R., Sedghi, H., Gupta, V., Goldblum, M., and Goldstein, T. The role of permutation invariance in linear mode connectivity of neural networks. *arXiv:2110.06296*, 2021.

Godfrey, C., Brown, D., Emerson, T., and Kvinge, H. On the symmetries of deep learning models and their internal representations. *arXiv:2205.14258*, 2022.

Gunasekar, S. et al. Implicit bias of gradient descent on linear convolutional networks. *NeurIPS*, 2018.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018.

Harutyunyan, H., Reing, K., Steeg, G. V., and Galstyan, A. G. Improving generalization by controlling label-noise information in neural network weights. In *ICML*, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CVPR*, 2015.

Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, 2017.

Kappiyath, A., Chaudhuri, A., JAISWAL, A. K., Liu, Z., Li, Y., Zhu, X., and Yin, L. SEBRA : Debiasing through self-guided bias ranking. In *ICLR*, 2025.

Kearns, M. Efficient noise-tolerant learning from statistical queries. *STOC*, 1993.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/cifar.html. Technical Report.

Kunin, D., Sagastuy-Brena, J., Ganguli, S., Yamins, D. L. K., and Tanaka, H. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *ICLR*, 2021.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/, 1998.

Li, J., Socher, R., and Hoi, S. C. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2020.

Liu, D., Tsang, I. W., and Yang, G. A convergence path to deep learning on noisy labels. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Liu, S., Niles-Weed, J., Razavian, N., and Fernandez-Granda, C. Early-learning regularization prevents memorization of noisy labels. In *ICLR*, 2020.

Liu, Y., Cheng, H., and Zhang, K. Identifiability of label noise transition matrix. In *ICML*, 2023.

Ma, X., Wang, Y., Houle, M. E., Zhou, S., Erfani, S., Xia, S., Wijewickrema, S., and Bailey, J. Dimensionality-driven learning with noisy labels. In *ICML*, 2018.

Malladi, S. et al. Symmetry and generalization in neural networks. *arXiv:2302.14020*, 2023.

Maron, H. Exploiting symmetries for learning in deep weight spaces. In *M4DL Geometric Deep Learning Workshop*, 2024.

Meng, T. Hh-gnn: Homogeneity- and heterogeneity-aware graph neural network for fraud detection with noisy labels. *Twenty-First International Conference on Intelligent Computing*, 2025.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In *NeurIPS*, 2013.

Northcutt, C. G., Jiang, L., and Chuang, I. L. Confident learning: Estimating uncertainty in dataset labels. In *Journal of Artificial Intelligence Research*, 2021.

Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, 2017.

Pleiss, G., Raghunathan, A., Liang, P., and Zhang, Z. Identifying mislabeled data using the area under the margin ranking. In *NeurIPS*, 2020.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A. On the spectral bias of neural networks. In *ICML*, 2019.

Ravanbakhsh, S., Schneider, J., and Póczos, B. Equivariance through parameter-sharing. In *ICML*, 2017.

Rieck, B., Togninalli, M., Bock, C., Moor, M., Horn, M., Gumbsch, T., and Borgwardt, K. Neural persistence: A complexity measure for deep neural networks using algebraic topology. In *ICLR*, 2019.

S., V. M., Yuan, S., and Wu, X. Contrastive Learning for Fraud Detection from Noisy Labels . In *IEEE 40th International Conference on Data Engineering (ICDE)*, 2024.

Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., and Fergus, R. Training convolutional networks with noisy labels. In *ICLR Workshop*, 2015.

Tahmasebi, P. and Jegelka, S. Deep learning with symmetries: Sample complexity gain of invariances. In *NeurIPS*, 2023.

Tan, C., García-Redondo, I., Wang, Q., Bronstein, M. M., and Monod, A. On the limitations of fractal dimension as a measure of generalization. In *NeurIPS*, 2024.

Tanaka, H. and Kunin, D. Noether's learning dynamics: Role of symmetry breaking in neural networks. In *NeurIPS*, 2021.

Tang, J., Li, J., Gao, Z., and Li, J. Rethinking graph neural networks for anomaly detection. In *ICML*, 2022.

van Rooyen, B. and Williamson, R. C. A theory of learning with corrupted labels. *JMLR*, 2018.

van Rooyen, B., Menon, A. K., and Williamson, R. C. Learning with symmetric label noise: The importance of being unhinged. In *NeurIPS*, 2015.

Vardi, G. On the implicit bias in deep-learning algorithms. *arXiv:2208.12591*, 2022.

Wang, H., Huang, Z., Lin, Z., and Liu, T. Noisegpt: Label noise detection and rectification through probability curvature. In *Advances in Neural Information Processing Systems*, 2024.

Wang, S., Zhang, Z., Fang, L., Nguyen, C. T., and Li, W. Corporate fraud detection in rich-yet-noisy financial graph. *ArXiv*, abs/2502.19305, 2025.

Wei, J., Liu, H., Liu, T., Niu, G., Sugiyama, M., and Liu, Y. To smooth or not? when label smoothing meets noisy labels. In *ICML*, 2022.

Xiao, T., Xia, T., Yang, Y., Huang, C., and Wang, X. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015.

Xu, Z.-Q. J., Zhang, Y., and Xiao, Y. Training behavior of deep neural network in frequency domain. In *International Conference on Neural Information Processing*, 2019.

Xu, Z.-Q. J., Zhang, Y., Luo, T., Xiao, Y., and Ma, Z. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 2020.

Yang, Y., Gan, E., Dziugaite, G. K., and Mirzasoleiman, B. Identifying spurious biases early in training through the lens of simplicity bias. *AISTATS*, 2024.

Yuan, S., Feng, L., and Liu, T. Late stopping: Avoiding confidently learning from mislabeled examples. In *ICCV*, 2023.

Zhang, Y., Wang, Z., Li, Y., Gong, C., and Liu, T. Pls-lsa+: A probabilistic label smoothing framework for learning with noisy labels. In *ECCV*, 2024.

Zhang, Y., Li, Y., Gong, C., and Liu, T. Weed out, then harvest: Dual low-rank adaptation is an effective noisy label detector. In *ACL*, 2025.

Zhao, B., Ganev, I., Walters, R., Yu, R., and Dehmamy, N. Symmetries, flat minima, and the conserved quantities of gradient flow. In *ICLR*, 2023.

Zhou, M., Liu, T., Li, Y., Lin, D., Zhou, E., and Zhao, T. Toward understanding the importance of noise in training neural networks. In *ICML*, 2019.

A. Extended Literature Review

Noisy Labels: Invariant-feature-based correction (Deng et al., 2024) is related, but our work instead characterizes invariants arising from neural-network learning dynamics, a perspective orthogonal to symmetric-noise theory (van Rooyen et al., 2015). Studies on late learning (Zhou et al., 2019; Yang et al., 2024; Kappiyath et al., 2025) and optimization-dynamics analyses (Liu et al., 2024) document the phenomenon but do not provide conditions for recovering the true label distribution with arbitrarily high precision, which classical frameworks (Kearns, 1993; Natarajan et al., 2013; van Rooyen & Williamson, 2018) also do not address.

Symmetries in deep learning: Symmetry has long been recognized as a fundamental organizing principle in deep learning. Cohen & Welling (2016) first formalized how group-equivariant convolutional networks enforce structured invariances through parameter-sharing, a perspective further generalized by Ravanbakhsh et al. (2017). This architectural viewpoint was later unified under the geometric deep learning framework of Bronstein et al. (2021), which characterizes neural networks as models defined over groups, graphs, and manifolds. Symmetry also plays a central role in statistical learning: Bietti et al. (2021) and Tahmasebi & Jegelka (2023) show that invariances can yield substantial reductions in sample complexity, while Malladi et al. (2023) analyze how symmetry constraints shape generalization behavior more broadly.

Beyond architectures and sample complexity, representational symmetries have been studied by Godfrey et al. (2022), who demonstrate that hidden layers inherit structured symmetry groups from the model family. Symmetries also arise in weight space: deep networks possess large parameter-space symmetry groups, including permutations and scalings, that create families of equivalent minima and influence optimization trajectories (Maron, 2024; Entezari et al., 2021). Finally, the implicit-bias literature shows that gradient-based optimization preferentially selects solutions with specific symmetry-related properties, both in linearized models and convolutional architectures (Vardi, 2022; Gunasekar et al., 2018). Our work complements these perspectives by identifying properties of deep neural networks that take the form of symmetries in asymptotic limits when learning under label noise, with associated invariants that enable the recovery of the true concept.

B. Theoretical Results

B.1. Effective Feature Space

We endow X with a metric $d : X \times X \to \mathbb{R}_+$ that captures the *effective feature space* used by the learner; all notions of proximity (e.g., "two samples are close") are with respect to d, not necessarily the raw input geometry. Concretely, d can be realized via a (possibly learned or fixed) feature map $\phi : X \to \mathbb{R}^m$ as $d(x, x') = \|\phi(x) - \phi(x')\|$, and the "local inconsistency" conditions underlying the complexity barrier refer to pairs $(x, y), (x', y')$ with $d(x, x')$ small but $y \neq y'$.

B.2. Complexity Axioms

Let $\mathcal{F}$ be a function space / hypothesis class of continuous functions $f : X \to \Delta(Y)$, and let

$$\mathcal{C} : \mathcal{F} \to [0, \infty]$$

be a complexity functional. We identify the following properties that such a measure must satisfy, which we call the *complexity axioms*.

Axiom 1 (Well-definedness). For all $f \in \mathcal{F}$,

$$\mathcal{C}(f) \in [0, \infty].$$

Axiom 2 (Continuity). Let $(f_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{F}$ converging to f in the chosen topology on $\mathcal{F}$. Then

$$f_n \to f \quad \Longrightarrow \quad \mathcal{C}(f_n) \to \mathcal{C}(f).$$

Axiom 3 (Monotonicity). Let the *oscillation* of f on a set $S \subseteq X$ be

$$\operatorname{osc}(f, S) \;=\; \sup_{\mathbf{z},\mathbf{z}'\in S} \|f(\mathbf{z}) - f(\mathbf{z}')\|.$$

We say f is *strictly more oscillatory* than g (written $f \succ g$) if there exists a set $S \subseteq X$ such that $\operatorname{osc}(f, S) \geq \operatorname{osc}(g, S)$ and f achieves this oscillation on a strictly smaller subset, *i.e.*, there exists $S' \subsetneq S$ with $\operatorname{osc}(f, S') \geq \operatorname{osc}(g, S)$. Then:

$$f \succ g \quad \Longrightarrow \quad \mathcal{C}(f) \geq \mathcal{C}(g).$$

Axiom 4 (Non-degeneracy).

$$\mathcal{C}(f) = 0 \quad \Longleftrightarrow \quad f \text{ is constant on } X.$$

We additionally impose the following mild regularity condition, satisfied by all standard neural-network function classes:

Assumption 1 (Regularity)**.** $\mathcal{F}$ consists of continuous functions $X \to \Delta(Y)$, and any sequence in $\mathcal{F}$ with uniformly bounded complexity admits a subsequence that converges uniformly on compact subsets of X to a limit in $\mathcal{F}$.

This is an Arzelà–Ascoli-type condition: bounded complexity implies equicontinuity, which combined with pointwise boundedness ($\Delta(Y)$ is compact) yields sequential compactness of bounded sublevel sets $\{f \in \mathcal{F} : \mathcal{C}(f) \leq M\}$ under uniform-on-compacta convergence (Hornik et al., 1989).

B.3. Unboundedness under Vanishing Separation

We now derive the divergence of complexity under vanishing separation purely from the four axioms above and the regularity assumption, without postulating it directly.

Lemma 1 (Oscillation Growth)**.** *Fix $\mathbf{x} \in X$ and labels $\mathbf{y} \neq \mathbf{y}'$. For each $r > 0$, let $\mathbf{x}_r$ satisfy $d(\mathbf{x}, \mathbf{x}_r) = r$ and define $\mathcal{F}_r = \{f \in \mathcal{F} : f(\mathbf{x}) = \mathbf{y},\ f(\mathbf{x}_r) = \mathbf{y}'\}$. Then for $0 < r_1 < r_2$, any $f \in \mathcal{F}_{r_1}$ is strictly more oscillatory than any $g \in \mathcal{F}_{r_2}$:*

$$f \succ g.$$

Proof. Both f and g map $\mathbf{x}$ to $\mathbf{y}$ and must attain value $\mathbf{y}'$ at distance r_1 and r_2 from $\mathbf{x}$, respectively. Since f and g are continuous and $\mathbf{y} \neq \mathbf{y}'$, both must traverse the full output variation $\|\mathbf{y} - \mathbf{y}'\|$ between $\mathbf{x}$ and their respective target points.

Consider the closed ball $S = \overline{B}(\mathbf{x}, r_2)$. Both f and g satisfy:

$$\operatorname{osc}(g, S) \ \geq \ \|g(\mathbf{x}) - g(\mathbf{x}_{r_2})\| \ = \ \|\mathbf{y} - \mathbf{y}'\|,$$

and g requires the full set S to realize this oscillation, since $\mathbf{x}_{r_2}$ lies on the boundary of S.

Now consider the strictly smaller set $S' = \overline{B}(\mathbf{x}, r_1) \subsetneq S$ (since $r_1 < r_2$). The function f satisfies:

$$\operatorname{osc}(f, S') \ \geq \ \|f(\mathbf{x}) - f(\mathbf{x}_{r_1})\| \ = \ \|\mathbf{y} - \mathbf{y}'\| \ \geq \ \operatorname{osc}(g, S).$$

Thus f achieves at least the same oscillation as g on S, but on the strictly smaller subset $S' \subsetneq S$. By the definition of the oscillation ordering, $f \succ g$. □

Lemma 2 (Unboundedness under Vanishing Separation)**.** *Under Axioms 1–4 and Assumption 1, let $C(r) = \inf\{\mathcal{C}(f) : f \in \mathcal{F}_r\}$. Then:*

(i) $0 < r_1 < r_2 \implies C(r_1) \geq C(r_2)$*, and*

(ii) $\lim_{r \to 0} C(r) = \infty$.

Proof. **Part (i): Monotonicity of $C(\cdot)$.** By Lemma 1, for $0 < r_1 < r_2$, any $f \in \mathcal{F}_{r_1}$ satisfies $f \succ g$ for every $g \in \mathcal{F}_{r_2}$. By the monotonicity axiom, $\mathcal{C}(f) \geq \mathcal{C}(g)$. Taking the infimum over $f \in \mathcal{F}_{r_1}$ and $g \in \mathcal{F}_{r_2}$ yields $C(r_1) \geq C(r_2)$.

Part (ii): Divergence. Suppose for contradiction that $\sup_{r>0} C(r) \leq M < \infty$. Then for each $n \in \mathbb{N}$, there exists $f_n \in \mathcal{F}_{1/n}$ with $\mathcal{C}(f_n) \leq M + 1$. Each f_n satisfies:

$$f_n(\mathbf{x}) = \mathbf{y}, \qquad f_n(\mathbf{x}_{1/n}) = \mathbf{y}', \qquad d(\mathbf{x}, \mathbf{x}_{1/n}) = 1/n.$$

By Assumption 1, the sequence (f_n) admits a subsequence (f_{n_k}) converging uniformly on compact subsets of X to some $f^* \in \mathcal{F}$, with $\mathcal{C}(f^*) \leq M + 1 < \infty$ by the continuity axiom.

Since $f_{n_k}(\mathbf{x}) = \mathbf{y}$ for all k, uniform convergence gives $f^*(\mathbf{x}) = \mathbf{y}$.

Now, $\mathbf{x}_{1/n_k} \to \mathbf{x}$ as $k \to \infty$. Fix any compact set K containing $\mathbf{x}$ and all $\mathbf{x}_{1/n_k}$. By uniform convergence on K:

$$|f^*(\mathbf{x}_{1/n_k}) - f_{n_k}(\mathbf{x}_{1/n_k})| \;\le\; \|f^* - f_{n_k}\|_{\infty,K} \;\xrightarrow{k\to\infty}\; 0.$$

Since $f_{n_k}(\mathbf{x}_{1/n_k}) = \mathbf{y}'$ for all k, we obtain $f^*(\mathbf{x}_{1/n_k}) \to \mathbf{y}'$.

But $f^* \in \mathcal{F}$ is continuous (Assumption 1), and $\mathbf{x}_{1/n_k} \to \mathbf{x}$, so:

$$f^*(\mathbf{x}_{1/n_k}) \;\to\; f^*(\mathbf{x}) = \mathbf{y}.$$

Hence $\mathbf{y} = \mathbf{y}'$, contradicting $\mathbf{y} \neq \mathbf{y}'$. Therefore $\lim_{r\to 0} C(r) = \infty$. □

Remark: The unboundedness property is thus a *consequence* of the four complexity axioms together with the natural regularity of neural-network function classes, rather than an independent postulate. The key mechanism is geometric: compressing the same output variation into a vanishing spatial scale forces strictly finer oscillation (Oscillation Growth), which monotonicity converts to non-decreasing complexity, and the regularity assumption precludes stabilization at any finite bound.

B.4. Complexity Barrier and Unreachability

Discussion on Observation 1 (Complexity Barrier). Consider some $(\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')$ with $\mathbf{y} \neq \mathbf{y}'$. For each $r > 0$, choose $\mathbf{x}_r$ with $d(\mathbf{x}, \mathbf{x}_r) = r$ and define $\mathcal{F}_r = \{f \in \mathcal{F} : f(\mathbf{x}) = \mathbf{y},\ f(\mathbf{x}_r) = \mathbf{y}'\}$ and

$$C(r) = \inf\{\mathcal{C}(f) : f \in \mathcal{F}_r\}.$$

Then, $0 < r_1 < r_2$ implies $C(r_1) \ge C(r_2)$, and $\lim_{r\to 0} C(r) = \infty$.

This is a direct consequence of Lemma 2 (Unboundedness under Vanishing Separation). The argument proceeds in two steps:

1. **Monotonicity of $C(\cdot)$:** By the Oscillation Growth Lemma (Lemma 1), any $f \in \mathcal{F}_{r_1}$ must compress the full output variation $\|\mathbf{y} - \mathbf{y}'\|$ into a ball of radius $r_1 < r_2$, making it strictly more oscillatory than any $g \in \mathcal{F}_{r_2}$ that can spread the same transition over the larger distance r_2. By the monotonicity axiom, $\mathcal{C}(f) \ge \mathcal{C}(g)$, hence $C(r_1) \ge C(r_2)$.
2. **Divergence:** The compactness–continuity argument in Lemma 2 establishes $\lim_{r\to 0} C(r) = \infty$ by contradiction: a bounded sequence of separators at vanishing distances would, via the regularity assumption, converge to a continuous function that is simultaneously $\mathbf{y}$ and $\mathbf{y}'$ at $\mathbf{x}$.

Discussion on Corollary 1.1 (Unreachability). Let $\mathcal{F}'$ be any hypothesis class in which learning proceeds by gradient flow, producing a trajectory $(f_t)_{t\ge 0} \subset \mathcal{F}'$. For samples $(\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')$ with $\mathbf{y} \neq \mathbf{y}'$, define for $r > 0$ a point $\mathbf{x}_r$ with $d(\mathbf{x}, \mathbf{x}_r) = r$ and set:

$$C(r) = \inf\{\mathcal{C}(f) : f(\mathbf{x}) = \mathbf{y},\ f(\mathbf{x}_r) = \mathbf{y}'\}.$$

By Lemma 2, $\lim_{r\to 0} C(r) = \infty$. Consequently, for any solution $f^\star \in \mathcal{F}'$ that (f_t) converges to, the following must hold:

$$\begin{aligned} f^\star(\mathbf{x}) = \mathbf{y}, \quad & f^\star(\mathbf{x}') = \mathbf{y}' \\ \iff \quad & \forall K < \infty,\ \sup_{f\in\mathcal{F}'} \mathcal{C}(f) \ge K, \end{aligned}$$

i.e., $f^\star$ has zero loss iff $\mathcal{F}'$ is not bounded in complexity.

Fix $(\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')$ with $\mathbf{y} \neq \mathbf{y}'$. For each $r > 0$, choose $\mathbf{x}_r$ with $d(\mathbf{x}, \mathbf{x}_r) = r$ and define

$$C(r) = \inf\{\mathcal{C}(f) : f(\mathbf{x}) = \mathbf{y},\ f(\mathbf{x}_r) = \mathbf{y}'\}.$$

Due to Lemma 2, separating two conflicting labels at distance r requires complexity at least $C(r)$, with:

$$0 < r_1 < r_2 \implies C(r_1) \ge C(r_2), \qquad \lim_{r\to 0} C(r) = \infty.$$

Now let $(f_t)_{t\geq 0} \subset \mathcal{F}'$ be the gradient-flow trajectory, and assume it converges to some $f^\star \in \mathcal{F}'$ in the sense of the Cauchy representation of proximity (Lemma 3). In particular, for any sequence $\mathbf{x}_{r_n} \to \mathbf{x}'$ with $r_n \to 0$, we have

$$f_t(\mathbf{x}_{r_n}) \longrightarrow f^\star(\mathbf{x}') \qquad \text{as } t \to \infty.$$

In other words, **bounded complexity ⇒ zero loss is impossible.**

Suppose $\mathcal{F}'$ is bounded in complexity: there exists $K < \infty$ such that

$$\sup_{f\in\mathcal{F}'} \mathcal{C}(f) \leq K.$$

Then $\mathcal{C}(f^\star) \leq K$.

Assume for contradiction that $f^\star$ achieves zero loss on the conflicting pair:

$$f^\star(\mathbf{x}) = \mathbf{y}, \qquad f^\star(\mathbf{x}') = \mathbf{y}'.$$

Choose any sequence $r_n \to 0$ and corresponding $\mathbf{x}_{r_n} \to \mathbf{x}'$. Since $f^\star(\mathbf{x}_{r_n}) \to f^\star(\mathbf{x}') = \mathbf{y}'$, each $f^\star$ separates $\mathbf{x}$ and $\mathbf{x}_{r_n}$ with different labels. By the Oscillation Growth Lemma (Lemma 1) and the monotonicity axiom:

$$\mathcal{C}(f^\star) \geq C(r_n).$$

Taking $n \to \infty$ and using Lemma 2:

$$\mathcal{C}(f^\star) \geq \lim_{n\to\infty} C(r_n) = \infty,$$

contradicting $\mathcal{C}(f^\star) \leq K < \infty$. Hence, in a bounded hypothesis class, *no zero-loss limit is reachable*.

In other words, **unbounded complexity ⇒ zero loss is not ruled out.**

Conversely, suppose $\mathcal{F}'$ is unbounded in complexity:

$$\forall K < \infty,\ \exists f \in \mathcal{F}' \text{ with } \mathcal{C}(f) \geq K.$$

Then the Complexity Barrier does not prevent the existence of classifiers that separate conflicting labels at arbitrarily small distances. In particular, since $C(r) \to \infty$ as $r \to 0$, an unbounded class can realize functions of complexity at least $C(r)$ for every $r > 0$. Thus a zero-loss solution $f^\star$ satisfying

$$f^\star(\mathbf{x}) = \mathbf{y}, \qquad f^\star(\mathbf{x}') = \mathbf{y}'$$

is compatible with the axioms. In conclusion, combining both directions:

$$f^\star(\mathbf{x}) = \mathbf{y}, \quad f^\star(\mathbf{x}') = \mathbf{y}' \quad \Longleftrightarrow \quad \forall K < \infty,\ \sup_{f\in\mathcal{F}'} \mathcal{C}(f) \geq K.$$

That is, a zero-loss limit is reachable under gradient flow *if and only if* the hypothesis class $\mathcal{F}'$ is not bounded in complexity.

B.5. Proofs

Lemma 3 (Cauchy Representation of Proximity)**.** *Assume that for each fixed pair $(\mathbf{x}, \mathbf{x}')$ with $\mathbf{y} \neq \mathbf{y}'$, the trajectory $(f_t)_{t\geq 0}$ is bounded in $\mathcal{F}$ and has a (formal) limit $f^\star$ as $t \to \infty$. Then, as $\mathbf{x}' \to \mathbf{x}$, the family of gradient-flow iterates forms a Cauchy sequence in $\mathcal{F}$ in the following sense: for every $\varepsilon > 0$, there exists a $T < \infty$ such that for all $t, s \geq T$,*

$$\|f_t - f_s\| < \varepsilon.$$

Equivalently, the net $\{f_t\}_{t\geq 0}$ is Cauchy in $\mathcal{F}$ as $t \to \infty$ when $\mathbf{x}' \to \mathbf{x}$.

Proof. Since $\mathbf{y} \neq \mathbf{y}'$ and $\mathbf{x}' \to \mathbf{x}$, the Unboundedness Lemma (Lemma 2) implies that any zero-loss solution $f^\star$ must satisfy $\mathcal{C}(f^\star) = \infty$. Hence $f^\star$ cannot be represented as a finite sum of gradient-flow increments. In particular, for any finite partition

$$0 = t_0 < t_1 < \cdots < t_N = T,$$

there exists no finite T such that

$$f^\star = f_0 + \sum_{k=0}^{N-1} \big(f_{t_{k+1}} - f_{t_k}\big)$$

with $\mathcal{C}(f^\star) < \infty$.

Therefore, $f^\star$ can only be expressed formally as an infinite series of gradient-flow increments:

$$f^\star = f_0 + \sum_{k=0}^{\infty} \big(f_{t_{k+1}} - f_{t_k}\big),$$

where $t_k \to \infty$. Define the partial sums

$$S_n := f_0 + \sum_{k=0}^{n} \big(f_{t_{k+1}} - f_{t_k}\big) = f_{t_{n+1}}.$$

By assumption, the trajectory $(f_t)_{t\geq 0}$ is bounded in $\mathcal{F}$ and admits a formal limit $f^\star$ as $t \to \infty$. Hence, for every $\varepsilon > 0$, there exists $T < \infty$ such that for all $t, s \geq T$,

$$\|f_t - f_s\| < \varepsilon.$$

Since $S_n = f_{t_{n+1}}$, this implies

$$\|S_n - S_m\| = \|f_{t_{n+1}} - f_{t_{m+1}}\| < \varepsilon$$

for all n, m sufficiently large. Thus (S_n) is a Cauchy sequence in $\mathcal{F}$, and the gradient-flow iterates (f_t) form a Cauchy net as $t \to \infty$ when $\mathbf{x}' \to \mathbf{x}$. This completes the proof of the lemma. □

Lemma 4 (Concept-Instance Gap)**.** *Consider the model trajectory $\{f_t\}_{t\in[0,T]}$. For any cleanly labeled sample $\mathbf{x} \in X$ with $c(\mathbf{x}) = \tilde{c}(\mathbf{x})$, there exists a time $\tau \in (0, T)$ when $f_\tau(\mathbf{x}) = c(\mathbf{x})$, and a strictly later time $t^\star \in (\tau, T]$ when additionally,*

$$\forall t \in [\tau, t^\star),\ \mathcal{L}(f_t) > 0, \qquad \mathcal{L}(f_{t^\star}) = 0.$$

On the other hand, for any noisily labeled sample $\mathbf{x} \in X$ with $c(\mathbf{x}) \neq \tilde{c}(\mathbf{x})$, $\tau = t^*$.

Proof. The empirical loss is defined by:

$$\mathcal{L}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i), \tilde{c}(\mathbf{x}_i)),$$

and gradient flow evolves according to:

$$\frac{d}{dt} f_t = -\nabla \mathcal{L}(f_t).$$

Consider a clean sample $\mathbf{x}$ with $c(\mathbf{x}) = \tilde{c}(\mathbf{x})$. Assume clean samples from class $\mathbf{y}$ are drawn IID from a distribution $P_\mathbf{y}$ with density $p_\mathbf{y}$ and nonzero variance; for concreteness, take $p_\mathbf{y}(\mathbf{x}) \propto \exp(-\|\mathbf{x} - \mu_\mathbf{y}\|^2/(2\sigma_\mathbf{y}^2))$ with $\sigma_\mathbf{y} > 0$. The population loss for class $\mathbf{y}$ is:

$$\mathcal{L}_\mathbf{y}(f) = \mathbb{E}_{\mathbf{x}\sim P_\mathbf{y}}[\ell(f(\mathbf{x}), \mathbf{y})],$$

and its gradient is:

$$\nabla \mathcal{L}_\mathbf{y}(f) = \int_X \nabla_f \ell(f(\mathbf{x}), \mathbf{y})\, p_\mathbf{y}(\mathbf{x})\, d\mathbf{x}.$$

Because $p_\mathbf{y}$ has positive variance, the integral is dominated early by regions near $\mu_\mathbf{y}$, so gradient flow first aligns f_t with the coarse class-level concept. Thus there exists $\tau \in (0, T)$ such that $f_\tau(\mathbf{x}) = c(\mathbf{x})$. At this time the global loss is still positive, since clean samples $\mathbf{x}_i$ far from $\mu_\mathbf{y}$ satisfy $\nabla_f \ell(f_\tau(\mathbf{x}_i), \mathbf{y}) \neq 0$ and require further optimization. Hence there exists $t^\star > \tau$ such that:

$$\forall t \in [\tau, t^\star),\ \mathcal{L}(f_t) > 0, \qquad \mathcal{L}(f_{t^\star}) = 0.$$

Because $P_\mathbf{y}$ has nonzero variance, the interval $(\tau, t^\star)$ is nontrivial.

Now consider a noisy sample $\mathbf{x}$ with $c(\mathbf{x}) \neq \tilde{c}(\mathbf{x})$. In standard noise models (Natarajan et al., 2013; Sukhbaatar et al., 2015; Patrini et al., 2017), such a sample behaves as a point mass, contributing loss $\ell(f(\mathbf{x}), \tilde{c}(\mathbf{x}))$ and gradient $\nabla_f \ell(f(\mathbf{x}), \tilde{c}(\mathbf{x}))$ with no surrounding region of positive measure. Let:

$$\tau = \inf\{t \in [0, T] : f_t(\mathbf{x}) = \tilde{c}(\mathbf{x})\}.$$

At this time $\ell(f_\tau(\mathbf{x}), \tilde{c}(\mathbf{x})) = 0$, and because the noisy sample has no neighborhood structure, no further optimization is required to reduce its loss. Thus:

$$\tau = t^\star.$$

Clean samples, drawn from distributions with positive variance, satisfy $\tau < t^\star$. Noisy samples, modeled as delta-like point masses, satisfy $\tau = t^\star$. This completes the proof of the lemma. □

Lemma 5 (Order of Learning)**.** *For any $\mathbf{x} \in X$ that is noisily labeled, i.e., $c(\mathbf{x}) \neq \tilde{c}(\mathbf{x})$, there exists a time $\tau \in (0, T)$ when $f_t(\mathbf{x}) = c(\mathbf{x})$ and a later time $t^* \in (\tau, T]$ when $f_t(\mathbf{x}) = \tilde{c}(\mathbf{x})$.*

Proof. Let $\mathbf{x} \in X$ be noisily labeled, so $c(\mathbf{x}) \neq \tilde{c}(\mathbf{x})$. The empirical loss is:

$$\mathcal{L}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i), \tilde{c}(\mathbf{x}_i)),$$

and gradient flow evolves according to:

$$\frac{d}{dt} f_t = -\nabla \mathcal{L}(f_t).$$

Assume clean samples with true label $c(\mathbf{x})$ are drawn IID from a class-conditional distribution $P_{c(\mathbf{x})}$ with density $p_{c(\mathbf{x})}$ and nonzero variance. For concreteness, suppose $p_{c(\mathbf{x})}(\mathbf{z}) \propto \exp(-\|\mathbf{z} - \mu_{c(\mathbf{x})}\|^2/(2\sigma^2))$ with $\sigma > 0$. The population loss for this class is:

$$\mathcal{L}_{c(\mathbf{x})}(f) = \mathbb{E}_{\mathbf{z} \sim P_{c(\mathbf{x})}}[\ell(f(\mathbf{z}), c(\mathbf{x}))],$$

and its gradient is:

$$\nabla \mathcal{L}_{c(\mathbf{x})}(f) = \int_X \nabla_f \ell(f(\mathbf{z}), c(\mathbf{x}))\, p_{c(\mathbf{x})}(\mathbf{z})\, d\mathbf{z}.$$

Because $p_{c(\mathbf{x})}$ has positive variance, the gradient contribution from the clean distribution dominates early in training. In particular, the sign of $\nabla_f \ell(f_t(\mathbf{x}), c(\mathbf{x}))$ is determined by the surrounding clean samples, not by the isolated noisy label. Thus, for sufficiently small $t > 0$, the gradient flow pushes $f_t(\mathbf{x})$ toward $c(\mathbf{x})$. Therefore, there exists $\tau \in (0, T)$ such that $f_\tau(\mathbf{x}) = c(\mathbf{x})$.

Now consider the noisy label $\tilde{c}(\mathbf{x})$. In standard noise models (Natarajan et al., 2013; Sukhbaatar et al., 2015; Patrini et al., 2017), a mislabeled sample contributes a point-mass term to the loss:

$$\mathcal{L}_{\text{noise}}(f) = \ell(f(\mathbf{x}), \tilde{c}(\mathbf{x})),$$

with gradient:

$$\nabla \mathcal{L}_{\text{noise}}(f) = \nabla_f \ell(f(\mathbf{x}), \tilde{c}(\mathbf{x})).$$

This term has no neighborhood structure and becomes dominant only after the clean losses have been reduced. Formally, as t increases, the gradient contributions from clean samples decay to zero, while the gradient at $\mathbf{x}$ remains nonzero until $f_t(\mathbf{x}) = \tilde{c}(\mathbf{x})$. Thus, there exists $t^* \in (\tau, T]$ such that $f_{t^*}(\mathbf{x}) = \tilde{c}(\mathbf{x})$.

Since the clean distribution initially determines the direction of gradient flow at $\mathbf{x}$, and the noisy point-mass term only dominates after the clean losses vanish, we necessarily have $t^* > \tau$. Hence the model first predicts the true concept $c(\mathbf{x})$ and only later memorizes the noisy concept $\tilde{c}(\mathbf{x})$.

This completes the proof of the lemma. □

Lemma 6 (Complexity Invariance)**.** *As f_t progresses under gradient flow from $c(\mathbf{x})$ to $\tilde{c}(\mathbf{x})$, the complexity of f_t increases, i.e.,*

$$\frac{d}{dt}\big(\mathcal{C}(f_t) - \mathcal{C}(c)\big) > 0$$

but the true concept representation is maintained, i.e.,

$$f_t(\mathbf{x}) = \hat{\mathbf{y}} = c(\mathbf{x})$$

for all $t \in (\tau, t^*)$.

Proof. Let $\mathbf{x}$ be noisily labeled, so $c(\mathbf{x}) \neq \tilde{c}(\mathbf{x})$. The empirical loss is:

$$\mathcal{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(\mathbf{x}_i), \tilde{c}(\mathbf{x}_i)),$$

and gradient flow evolves according to:

$$\frac{d}{dt} f_t = -\nabla \mathcal{L}(f_t).$$

By Lemma 5, there exist times $\tau < t^*$ such that $f_\tau(\mathbf{x}) = c(\mathbf{x})$ and $f_{t^*}(\mathbf{x}) = \tilde{c}(\mathbf{x})$. We analyze the interval $t \in (\tau, t^*)$.

The gradient of the loss at time t decomposes as

$$\nabla \mathcal{L}(f_t) = \sum_{\mathbf{z} \in X_{\text{clean}}} \nabla_f \ell(f_t(\mathbf{z}), c(\mathbf{z})) \; + \; \nabla_f \ell(f_t(\mathbf{x}), \tilde{c}(\mathbf{x})),$$

where the first term aggregates gradients from clean samples and the second is the point-mass contribution from the noisy label at $\mathbf{x}$.

For $t \in (\tau, t^*)$, the clean samples still contribute nonzero gradient because their losses have not yet vanished. Since these samples are drawn from class-conditional distributions with positive variance, the gradient flow continues to refine f_t on increasingly fine scales. By the monotonicity axiom for complexity under refinement, this implies:

$$\frac{d}{dt}\mathcal{C}(f_t) > 0,$$

and hence:

$$\frac{d}{dt}\big(\mathcal{C}(f_t) - \mathcal{C}(c)\big) > 0.$$

We now show that $f_t(\mathbf{x})$ cannot change on (τ, t^*). Suppose, for contradiction, that there exists $t_0 \in (\tau, t^*)$ such that $f_{t_0}(\mathbf{x}) \neq c(\mathbf{x})$. Since $f_\tau(\mathbf{x}) = c(\mathbf{x})$ and $f_{t^*}(\mathbf{x}) = \tilde{c}(\mathbf{x})$, this implies that the prediction at $\mathbf{x}$ changes at some intermediate time t_0.

But by Lemma 4, for a noisily labeled sample, the moment the prediction ceases to equal the true concept $c(\mathbf{x})$ and aligns with the noisy label $\tilde{c}(\mathbf{x})$, gradient flow converges at that point and no further refinement is required there. In particular, once $f_t(\mathbf{x}) = \tilde{c}(\mathbf{x})$, the loss contribution of $\mathbf{x}$ becomes zero and the gradient at $\mathbf{x}$ vanishes. Therefore, if the prediction changed at time t_0, the complexity would cease to increase at t_0.

This contradicts the strict inequality $\frac{d}{dt}\mathcal{C}(f_t) > 0$ for all $t \in (\tau, t^*)$. Hence no such t_0 exists, and the prediction at $\mathbf{x}$ remains equal to the true concept throughout the interval:

$$f_t(\mathbf{x}) = \hat{\mathbf{y}} = c(\mathbf{x}) \qquad \text{for all } t \in (\tau, t^*).$$

Thus, while gradient flow increases the complexity of f_t as it moves toward fitting the noisy label, the encoded concept at $\mathbf{x}$ remains invariant until the final transition at t^*. This completes the proof of the lemma. □

Lemma 7. *For any sample $\mathbf{x} \in X$, among all prediction states attained by the model on $\mathbf{x}$ during optimization, the state $f_t(\mathbf{x}) = c(\mathbf{x})$ has maximal Lebesgue measure, i.e.,*

$$\mu(\{\, t \in [0,T] : f_t(\mathbf{x}) = c(\mathbf{x}) \,\}) \geq \mu(\{\, t \in [0,T] : f_t(\mathbf{x}) = \mathbf{y}' \,\})$$

for any other label $\mathbf{y}' \in Y$, where μ denotes the Lebesgue measure on $[0,T]$.

Proof. Fix $\mathbf{x} \in X$ and consider the gradient–flow trajectory $\{f_t\}_{t\in[0,T]}$. For each label $\mathbf{y} \in Y$, define the prediction–state set

$$A_{\mathbf{y}} \coloneqq \{\, t \in [0,T] : f_t(\mathbf{x}) = \mathbf{y} \,\}.$$

The theorem states that $\mu(A_{c(\mathbf{x})}) \geq \mu(A_{\mathbf{y}'})$ for every $\mathbf{y}' \in Y$, where μ is Lebesgue measure on $[0,T]$.

We distinguish the clean and noisy cases.

Cleanly labeled sample. Suppose $c(\mathbf{x}) = \tilde{c}(\mathbf{x})$. By Lemma 4, there exist times $0 < \tau < t^\star \leq T$ such that $f_\tau(\mathbf{x}) = c(\mathbf{x})$, the global loss satisfies $\mathcal{L}(f_t) > 0$ for all $t \in [\tau, t^\star)$, and $\mathcal{L}(f_{t^\star}) = 0$. Since $\mathcal{L}(f_t)$ is nonincreasing along gradient flow and vanishes at $t^\star$, we have $f_t(\mathbf{x}) = c(\mathbf{x})$ for all $t \in [\tau, T]$ (any deviation from $c(\mathbf{x})$ would strictly increase the loss at $\mathbf{x}$ and contradict stationarity at zero loss). Hence

$$[\tau, T] \subseteq A_{c(\mathbf{x})} \quad \Rightarrow \quad \mu(A_{c(\mathbf{x})}) \geq T - \tau.$$

Any other label $\mathbf{y}' \neq c(\mathbf{x})$ can only be predicted before τ, so $A_{\mathbf{y}'} \subseteq [0, \tau]$ and thus

$$\mu(A_{\mathbf{y}'}) \leq \tau.$$

Under gradient–flow dynamics, the trajectory spends more time near low–loss steady states than in high–loss transients; in particular, the interval where the loss is minimized (here $[\tau, T]$) dominates any earlier transient interval in measure, so $T - \tau \geq \mu(A_{\mathbf{y}'})$. Combining the inequalities yields

$$\mu(A_{c(\mathbf{x})}) \;\geq\; T - \tau \;\geq\; \mu(A_{\mathbf{y}'}),$$

for every $\mathbf{y}' \in Y$.

Noisily labeled sample. Suppose $c(\mathbf{x}) \neq \tilde{c}(\mathbf{x})$. By Lemma 5, there exist $0 < \tau < t^\star \leq T$ such that $f_\tau(\mathbf{x}) = c(\mathbf{x})$ and $f_{t^\star}(\mathbf{x}) = \tilde{c}(\mathbf{x})$. By Lemma 6, for all $t \in (\tau, t^\star)$ we have $f_t(\mathbf{x}) = c(\mathbf{x})$ and $\frac{d}{dt}\,\mathcal{C}(f_t) > 0$, i.e. the model complexity strictly increases while the prediction at $\mathbf{x}$ remains equal to the true concept. Thus:

$$(\tau, t^\star) \subseteq A_{c(\mathbf{x})} \quad \Rightarrow \quad \mu(A_{c(\mathbf{x})}) \geq t^\star - \tau > 0.$$

Any other label $\mathbf{y}' \neq c(\mathbf{x})$ can only be predicted either before τ or after $t^\star$. Before τ, the dynamics are dominated by the surrounding clean data, and the prediction at $\mathbf{x}$ transitions through non–steady states on a set of times of small measure (transient phases). After $t^\star$, once $f_t(\mathbf{x})$ has aligned with $\tilde{c}(\mathbf{x})$, the loss contribution of $\mathbf{x}$ vanishes and there is no further gradient signal at that point; the prediction $f_t(\mathbf{x}) = \tilde{c}(\mathbf{x})$ is then a steady state. However, the interval $(\tau, t^\star)$ is precisely the phase during which the model complexity grows while the prediction remains $c(\mathbf{x})$, and by construction this is the dominant refinement phase for $\mathbf{x}$. Consequently, the time spent in the true–concept state $c(\mathbf{x})$ (at least $t^\star - \tau$) is no smaller than the time spent in any other prediction state $\mathbf{y}'$:

$$\mu(A_{c(\mathbf{x})}) \;\geq\; \mu(A_{\mathbf{y}'}) \qquad \text{for all } \mathbf{y}' \in Y.$$

Since the argument holds in both the clean and noisy cases, we conclude that for any $\mathbf{x} \in X$, among all prediction states attained by the model on $\mathbf{x}$ during optimization, the state $f_t(\mathbf{x}) = c(\mathbf{x})$ has maximal Lebesgue measure on $[0,T]$.

This completes the proof of the theorem. □

Intuition: The Lebesgue measure $\mu(\cdot)$, quantifies the time during gradient flow for which a condition holds. The theorem shows that the state $f_t(\mathbf{x}) = c(\mathbf{x})$, the model encoding the true concept, is the dominant steady state for any $\mathbf{x} \in X$ as

optimization proceeds toward zero loss. For clean samples, this holds because the true concept is learned strictly before zero loss is reached, leaving no incentive to deviate. For noisy samples, the true concept is likewise learned first (Lemma 5) and is maintained as model complexity increases under gradient flow until the target concept is approached (Lemma 6). Thus, regardless of whether a label is clean or noisy, once gradient flow reaches the true concept, the model remains in that state longer than in any alternative.

Theorem 2 (Steady-State Symmetry). For any sample $\mathbf{x} \in X$ learned by a model f_t under gradient flow over $t \in [0, T]$,

$$c(\mathbf{x}) = \frac{1}{T} \int_{t=0}^{T} z(\hat{\mathbf{y}}_t)\, f_t(\mathbf{x})\, dt$$

where $\hat{\mathbf{y}}_t = f_t(\mathbf{x})$ and $z(\hat{\mathbf{y}}_t)$ is the fraction of samples in X with label $\hat{\mathbf{y}}_t \in Y$ that have zero training loss under f_t, *i.e.*,

$$z(\hat{\mathbf{y}}_t) = \frac{1}{|I_{\hat{\mathbf{y}}_t}|} \sum_{i \in I_{\hat{\mathbf{y}}_t}} \mathbf{1}[\, \mathcal{L}_t(\mathbf{x}_i, \mathbf{y}_i) = 0 \,],$$

where $I_{\mathbf{y}}$ is the index set of samples in X with label $\mathbf{y} \in Y$, *i.e.*, $I_{\mathbf{y}} = \{\, i \in \{1, \ldots, N\} : \mathbf{y}_i = \mathbf{y} \,\}$.

Proof. Fix a sample $\mathbf{x} \in X$ and consider the gradient–flow trajectory $\{f_t\}_{t \in [0,T]}$. For each $t \in [0, T]$, let $\hat{\mathbf{y}}_t = f_t(\mathbf{x})$ denote the predicted label of $\mathbf{x}$ at time t. For any label $\mathbf{y} \in Y$, let $I_{\mathbf{y}} = \{\, i \in \{1, \ldots, N\} : \mathbf{y}_i = \mathbf{y} \,\}$ be the index set of samples with label $\mathbf{y}$, and define:

$$z(\hat{\mathbf{y}}_t) = \frac{1}{|I_{\hat{\mathbf{y}}_t}|} \sum_{i \in I_{\hat{\mathbf{y}}_t}} \mathbf{1}[\, \mathcal{L}_t(\mathbf{x}_i, \mathbf{y}_i) = 0 \,],$$

i.e., $z(\hat{\mathbf{y}}_t)$ is the fraction of samples with label $\hat{\mathbf{y}}_t$ that have zero training loss at time t.

By Theorem 2, for any $\mathbf{x} \in X$, the prediction state $f_t(\mathbf{x}) = c(\mathbf{x})$ has maximal Lebesgue measure on $[0, T]$ among all labels. In particular, there exists a nontrivial time interval on which $f_t(\mathbf{x}) = c(\mathbf{x})$ and the model has already learned the true concept for almost all samples with label $c(\mathbf{x})$, while noisy components (if any) have not yet been fully memorized. On this interval, for $\hat{\mathbf{y}}_t = c(\mathbf{x})$, the quantity $z(\hat{\mathbf{y}}_t)$ is close to 1, since almost all samples with label $c(\mathbf{x})$ have zero loss under f_t.

Conversely, at earlier times, when the model has not yet captured the true concept for the majority of samples with label $\hat{\mathbf{y}}_t$, the fraction $z(\hat{\mathbf{y}}_t)$ is significantly smaller than 1, because many samples with that label still incur positive loss. Thus, $z(\hat{\mathbf{y}}_t)$ acts as a time–dependent weight that is small in the early, pre–concept regime and close to 1 in the concept–recovery regime where the true concept has been learned for that label.

Consider now the time–weighted prediction:

$$\frac{1}{T} \int_0^T z(\hat{\mathbf{y}}_t)\, f_t(\mathbf{x})\, dt.$$

For times t at which $\hat{\mathbf{y}}_t \neq c(\mathbf{x})$, either the model has not yet learned the true concept for that label (so $z(\hat{\mathbf{y}}_t)$ is small), or the prediction corresponds to a noisy or transient state with relatively few zero–loss samples, again yielding small $z(\hat{\mathbf{y}}_t)$. Hence the contribution of such times to the integral is down–weighted.

In contrast, for times t at which $\hat{\mathbf{y}}_t = c(\mathbf{x})$ and the model has correctly fitted almost all samples with label $c(\mathbf{x})$, we have $z(\hat{\mathbf{y}}_t) \approx 1$, and $f_t(\mathbf{x}) = c(\mathbf{x})$. On this dominant interval (in the sense of Theorem 2), the integrand is essentially $c(\mathbf{x})$, and these times receive the largest weight in the time average.

Therefore, the time–weighted average

$$\frac{1}{T} \int_0^T z(\hat{\mathbf{y}}_t)\, f_t(\mathbf{x})\, dt$$

is dominated by the regime in which the model has learned the true concept and predicts $c(\mathbf{x})$ on $\mathbf{x}$, while contributions from other labels and pre–concept phases are suppressed by small values of $z(\hat{\mathbf{y}}_t)$. In the limit where $z(\hat{\mathbf{y}}_t) \to 1$ precisely when the true concept for label $\hat{\mathbf{y}}_t$ has been learned (and remains significantly below 1 otherwise), this weighted average recovers the true concept:

$$c(\mathbf{x}) = \frac{1}{T} \int_0^T z(\hat{\mathbf{y}}_t)\, f_t(\mathbf{x})\, dt.$$

Thus, $z(\hat{\mathbf{y}}_t)$ reaching 1 serves as a proxy for the model having learned the true concept, and weighting $f_t(\mathbf{x})$ by $z(\hat{\mathbf{y}}_t)$ ensures that predictions in the concept–recovery regime dominate the average, yielding $c(\mathbf{x})$.

This completes the proof of the theorem. □

Theorem 3 ($\mathbb{H}$-Stability). If $\mathbb{H}$ grows in a noisily consistent manner, then for any sample $\mathbf{x} \in X$:

$$c(\mathbf{x}) = \lim_{n\to\infty} f_1(\mathbf{x}) \oplus f_2(\mathbf{x}) \oplus ... \oplus f_n(\mathbf{x}),$$

where $\oplus$ is the renormalized summation operation given by:

$$S_0 = 0; \;\; S_i = \frac{(i-1)\cdot S_{i-1} + f_n(\mathbf{x})}{i},$$

where $S_i = f_1(\mathbf{x}) \oplus f_2(\mathbf{x}) \oplus ... \oplus f_i(\mathbf{x})$.

Proof. Let $\mathcal{F}$ be a real Hilbert space of functions on X with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$, and let $c \in \mathcal{F}$ denote the true concept. For each $i \in \mathbb{N}$, training over H_i under gradient flow yields a function $f_i \in \mathcal{F}$. Define the renormalized sum:

$$S_n(\mathbf{x}) \coloneqq f_1(\mathbf{x}) \oplus \cdots \oplus f_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}),$$

which corresponds to the recursive rule:

$$S_0 = 0, \qquad S_i = \frac{(i-1)\,S_{i-1} + f_i(\mathbf{x})}{i}.$$

Write each f_i as:

$$f_i = c + e_i,$$

where $e_i \in \mathcal{F}$ is the error component. Noisy consistency implies that each f_i fits the noisy labels with error at most η on the true labels, so the errors $\{e_i\}$ are uniformly bounded:

$$\|e_i\| \le C \quad \text{for all } i,$$

for some constant $C > 0$ depending on η.

The assumption that $\mathbb{H}$ grows in a noisily consistent manner means that each new hypothesis class H_i contributes an error e_i that is at least slightly more orthogonal to the subspace spanned by the previous errors. Formally, there exists a sequence $\varepsilon_i \to 0$ such that for all $j < i$,

$$|\langle e_i, e_j\rangle| \le \varepsilon_i \,\|e_i\|\,\|e_j\|.$$

Consider the average error:

$$\bar{e}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} e_i.$$

Then:

$$\|\bar{e}_n\|^2 = \Big\langle \frac{1}{n}\sum_{i=1}^{n} e_i, \frac{1}{n}\sum_{j=1}^{n} e_j \Big\rangle = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle e_i, e_j\rangle.$$

Split this into diagonal and off–diagonal parts:

$$\|\bar{e}_n\|^2 = \frac{1}{n^2}\sum_{i=1}^{n}\|e_i\|^2 + \frac{2}{n^2}\sum_{1\le i<j\le n}\langle e_i, e_j\rangle.$$

For the diagonal term, boundedness of $\|e_i\|$ gives:

$$0 \le \frac{1}{n^2}\sum_{i=1}^{n}\|e_i\|^2 \le \frac{nC^2}{n^2} = \frac{C^2}{n} \xrightarrow[n\to\infty]{} 0.$$

For the off–diagonal term, use the approximate orthogonality:

$$\left|\frac{2}{n^2}\sum_{1\le i<j\le n}\langle e_i, e_j\rangle\right| \le \frac{2}{n^2}\sum_{1\le i<j\le n}\varepsilon_{\max\{i,j\}}\|e_i\|\|e_j\| \le \frac{2C^2}{n^2}\sum_{1\le i<j\le n}\varepsilon_{\max\{i,j\}}.$$

Since $\varepsilon_k \to 0$ and there are at most k pairs with $\max\{i,j\} = k$, we have:

$$\frac{1}{n^2}\sum_{1\le i<j\le n}\varepsilon_{\max\{i,j\}} \xrightarrow[n\to\infty]{} 0,$$

so the off–diagonal term also tends to 0. Hence

$$\|\bar{e}_n\| \xrightarrow[n\to\infty]{} 0.$$

Now observe that:

$$\frac{1}{n}\sum_{i=1}^{n} f_i = \frac{1}{n}\sum_{i=1}^{n}(c + e_i) = c + \bar{e}_n,$$

so:

$$\Big\|\frac{1}{n}\sum_{i=1}^{n} f_i - c\Big\| = \|\bar{e}_n\| \xrightarrow[n\to\infty]{} 0.$$

Thus, $S_n := \frac{1}{n}\sum_{i=1}^{n} f_i$ converges to c in $\mathcal{F}$.

Evaluation at any fixed $\mathbf{x} \in X$ is a continuous linear functional on $\mathcal{F}$, so:

$$S_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \xrightarrow[n\to\infty]{} c(\mathbf{x}).$$

By the definition of $\oplus$, this is exactly:

$$c(\mathbf{x}) = \lim_{n\to\infty} f_1(\mathbf{x}) \oplus f_2(\mathbf{x}) \oplus \cdots \oplus f_n(\mathbf{x}).$$

This completes the proof of the theorem. □

Corollary 3.2. *If the $\mathbb{H}$-stable limit for a sample $\mathbf{x} \in X$ with true label $\mathbf{y}^*$ does not converge to its training label $\mathbf{y}$, then $\mathbf{y} \neq \mathbf{y}^*$, i.e., the label $\mathbf{y}$ for $\mathbf{x}$ is noisy.*

Proof. Let $\mathbf{x} \in X$ have true label $\mathbf{y}^* = c(\mathbf{x})$ and training label $\mathbf{y}$. By $\mathbb{H}$-stability, if $\mathbb{H}$ grows in a noisily consistent manner, then the renormalized summation over the solutions $\{f_i\}$ satisfies

$$\lim_{n\to\infty}\big(f_1(\mathbf{x}) \oplus f_2(\mathbf{x}) \oplus \cdots \oplus f_n(\mathbf{x})\big) = c(\mathbf{x}) = \mathbf{y}^*.$$

Suppose that the $\mathbb{H}$-stable limit for $\mathbf{x}$ does not converge to its training label $\mathbf{y}$, i.e.

$$\lim_{n\to\infty}\big(f_1(\mathbf{x}) \oplus \cdots \oplus f_n(\mathbf{x})\big) \neq \mathbf{y}.$$

Combining this with the $\mathbb{H}$-stability identity above, we obtain

$$\mathbf{y}^* = c(\mathbf{x}) \neq \mathbf{y}.$$

Hence the training label $\mathbf{y}$ for $\mathbf{x}$ cannot equal its true label $\mathbf{y}^*$, i.e. $\mathbf{y}$ is a noisy label.

Intuitively, as $\mathbb{H}$ grows, there are only finitely many functional forms that merely fit the noisy labels with error rate η, but infinitely many that reduce the error below η by aligning more closely with the true concept. In the infinite limit, these better approximations dominate the renormalized summation and drive it toward $c(\mathbf{x})$. Therefore, if the $\mathbb{H}$-stable prediction for $\mathbf{x}$ differs from its training label $\mathbf{y}$ for sufficiently large n, this discrepancy certifies that $\mathbf{y} \neq \mathbf{y}^*$.

This completes the proof of the corollary. □

B.6. Finite-Time Guarantee for Concept Recovery based on Steady State Symmetry

For a fixed sample $\mathbf{x} \in X$, let $\{f_k\}_{k=1}^T$ denote a finite SGD trajectory and let $\hat{\mathbf{y}}_k = f_k(\mathbf{x})$ be the predicted label at step k. We write $z_k := z(\hat{\mathbf{y}}_k)$ for the CRT weight at step k and define the normalized CRT mass:

$$P_T(\mathbf{y}) := \frac{\sum_{k=1}^T z_k \, \mathbf{1}[\hat{\mathbf{y}}_k = \mathbf{y}]}{\sum_{k=1}^T z_k}.$$

Let $\mathcal{K}_c(\mathbf{x}) \subseteq \{1, \dots, T\}$ denote the true-concept steady-state set, i.e., the set of iterates for which $\hat{\mathbf{y}}_k = c(\mathbf{x})$. We define:

$$a_T(\mathbf{x}) := 1 - \frac{|\mathcal{K}_c(\mathbf{x})|}{T}$$

as the fraction of iterates outside this steady state. Finally, let b_T denote the maximum weight deficit inside the steady state and let q_T denote the maximum residual weight outside it, so that:

$$z_k \geq 1 - b_T \quad \text{for } k \in \mathcal{K}_c(\mathbf{x}), \qquad z_k \leq q_T \quad \text{for } k \notin \mathcal{K}_c(\mathbf{x}).$$

Theorem 4 (Finite-Time CRT Rate)**.** *Let*

$$P_T(\mathbf{y}) = \frac{\sum_{k=1}^T z_k \mathbf{1}[\hat{\mathbf{y}}_k = \mathbf{y}]}{\sum_{k=1}^T z_k}.$$

Suppose $\hat{\mathbf{y}}_k = c(\mathbf{x})$ on a steady-state set $\mathcal{K}_c(\mathbf{x})$, with

$$a_T(\mathbf{x}) = 1 - \frac{|\mathcal{K}_c(\mathbf{x})|}{T},$$

and assume $z_k \geq 1 - b_T$ on $\mathcal{K}_c(\mathbf{x})$ while $z_k \leq q_T$ outside it. Then:

$$1 - P_T(c(\mathbf{x})) \leq \frac{a_T(\mathbf{x}) q_T}{(1 - a_T(\mathbf{x}))(1 - b_T) + a_T(\mathbf{x}) q_T} = O(a_T(\mathbf{x}) q_T),$$

whenever $(1 - a_T(\mathbf{x}))(1 - b_T)$ is bounded away from zero. Hence CRT recovers $c(\mathbf{x})$ once

$$P_T(c(\mathbf{x})) > \max_{\mathbf{y} \neq c(\mathbf{x})} P_T(\mathbf{y}).$$

Proof. Let $\mathcal{K}_c = \mathcal{K}_c(\mathbf{x})$ for brevity. By definition,

$$P_T(c(\mathbf{x})) = \frac{\sum_{k=1}^T z_k \, \mathbf{1}[\hat{\mathbf{y}}_k = c(\mathbf{x})]}{\sum_{k=1}^T z_k}.$$

Since $\hat{\mathbf{y}}_k = c(\mathbf{x})$ for every $k \in \mathcal{K}_c$, and since $z_k \geq 1 - b_T$ on $\mathcal{K}_c$, we have

$$\sum_{k=1}^T z_k \, \mathbf{1}[\hat{\mathbf{y}}_k = c(\mathbf{x})] \geq \sum_{k \in \mathcal{K}_c} z_k \geq |\mathcal{K}_c|(1 - b_T).$$

On the other hand, outside the steady-state set, $z_k \leq q_T$, so

$$\sum_{k=1}^T z_k = \sum_{k \in \mathcal{K}_c} z_k + \sum_{k \notin \mathcal{K}_c} z_k \leq |\mathcal{K}_c| \cdot 1 + (T - |\mathcal{K}_c|) q_T.$$

For an upper bound on $1 - P_T(c(\mathbf{x}))$, it is more convenient to directly bound the non-true-concept mass:

$$1 - P_T(c(\mathbf{x})) = \frac{\sum_{\mathbf{y} \neq c(\mathbf{x})} \sum_{k=1}^T z_k \, \mathbf{1}[\hat{\mathbf{y}}_k = \mathbf{y}]}{\sum_{k=1}^T z_k}.$$

Any prediction $\hat{\mathbf{y}}_k \neq c(\mathbf{x})$ can only occur outside $\mathcal{K}_c$. Therefore,

$$\sum_{\mathbf{y}\neq c(\mathbf{x})} \sum_{k=1}^{T} z_k \, \mathbf{1}[\hat{\mathbf{y}}_k = \mathbf{y}] \leq \sum_{k\notin\mathcal{K}_c} z_k \leq (T - |\mathcal{K}_c|)q_T.$$

Meanwhile, the denominator satisfies

$$\sum_{k=1}^{T} z_k \geq \sum_{k\in\mathcal{K}_c} z_k + \sum_{k\notin\mathcal{K}_c} 0 \geq |\mathcal{K}_c|(1 - b_T).$$

A slightly sharper bound keeps the possible residual mass outside $\mathcal{K}_c$ in the denominator:

$$\sum_{k=1}^{T} z_k \geq |\mathcal{K}_c|(1 - b_T) + \sum_{k\notin\mathcal{K}_c} z_k.$$

Using $\sum_{k\notin\mathcal{K}_c} z_k \leq (T - |\mathcal{K}_c|)q_T$, the worst case for the ratio occurs when the outside mass is maximal. Hence

$$1 - P_T(c(\mathbf{x})) \leq \frac{(T - |\mathcal{K}_c|)q_T}{|\mathcal{K}_c|(1 - b_T) + (T - |\mathcal{K}_c|)q_T}.$$

Dividing numerator and denominator by T and using

$$a_T(\mathbf{x}) = 1 - \frac{|\mathcal{K}_c|}{T}, \qquad 1 - a_T(\mathbf{x}) = \frac{|\mathcal{K}_c|}{T},$$

we obtain

$$1 - P_T(c(\mathbf{x})) \leq \frac{a_T(\mathbf{x})q_T}{(1 - a_T(\mathbf{x}))(1 - b_T) + a_T(\mathbf{x})q_T}.$$

If $(1 - a_T(\mathbf{x}))(1 - b_T)$ is bounded below by a positive constant, then the denominator is bounded away from zero, and therefore

$$1 - P_T(c(\mathbf{x})) = O(a_T(\mathbf{x})q_T).$$

In particular, if $a_T(\mathbf{x}) \to 0$, $b_T \to 0$, and $q_T \to 0$, then $P_T(c(\mathbf{x})) \to 1$. Consequently, for sufficiently large T,

$$P_T(c(\mathbf{x})) > \max_{\mathbf{y}\neq c(\mathbf{x})} P_T(\mathbf{y}),$$

and CRT recovers the true concept by the argmax rule. □

Intuition: Theorem 4 makes explicit what controls the finite-time convergence of CRT. The error term $1 - P_T(c(\mathbf{x}))$ is governed by two quantities: $a_T(\mathbf{x})$, the fraction of iterates that lie outside the true-concept steady state, and q_T, the maximum CRT weight assigned to those non-steady-state iterates. Thus the effective rate is $O(a_T(\mathbf{x})q_T)$: CRT improves either when the trajectory spends more time in the true-concept regime ($a_T(\mathbf{x}) \downarrow 0$), or when the weighting function increasingly suppresses transient and memorization phases ($q_T \downarrow 0$). The quantity b_T controls the weight deficit inside the steady state; as long as $(1 - a_T(\mathbf{x}))(1 - b_T)$ stays bounded away from zero, the steady-state mass remains dominant and the normalized CRT score concentrates around $c(\mathbf{x})$. Hence finite-time recovery does not require exact gradient-flow convergence: it only requires that the true-concept steady state occupies enough weighted mass, and the convergence rate is precisely the rate at which non-steady-state weighted mass vanishes.

Corollary 4.1 ($O(1/T)$ finite-time CRT rate)**.** *Assume the conditions of Theorem 4. Suppose, in addition, that CRT enters the true-concept steady state after a fixed burn-in of m iterates, independent of T, so that:*

$$|\{1, \ldots, T\} \setminus \mathcal{K}_c(\mathbf{x})| \leq m.$$

Then:

$$a_T(\mathbf{x}) = 1 - \frac{|\mathcal{K}_c(\mathbf{x})|}{T} \leq \frac{m}{T}.$$

If the non-steady-state weights are uniformly bounded by $q_T \le q < \infty$ and the steady-state weights satisfy $b_T \to 0$, with $(1 - a_T(\mathbf{x}))(1 - b_T)$ bounded away from zero, then:

$$1 - P_T(c(\mathbf{x})) = O\left(\frac{1}{T}\right).$$

More explicitly,

$$1 - P_T(c(\mathbf{x})) \le \frac{(m/T)q}{(1 - m/T)(1 - b_T) + (m/T)q} = O\left(\frac{1}{T}\right).$$

Intuition: Corollary 4.1 captures the simplest finite-time regime: after a fixed transient phase of length m, the trajectory remains in the true-concept steady state. In that case, the fraction of iterates outside the steady state is $a_T(\mathbf{x}) \le m/T$, while the non-steady-state contribution is weighted by at most q. Hence the normalized CRT mass outside the true concept decays as $O(1/T)$. Intuitively, every additional iterate after the burn-in contributes almost entirely to the true-concept score, while the total contribution of transient or memorization phases remains bounded. Thus the uncertainty in the CRT aggregate shrinks at the standard averaging rate $1/T$ in this fixed-burn-in regime.

B.7. Finite-Capacity $\mathbb{H}$-Stability

For a finite hypothesis pool $\mathbb{H}_n = \{H_1, \dots, H_n\}$, let $f_i \in H_i$ denote the predictor obtained by training over H_i, and write:

$$f_i = c + e_i,$$

where e_i is the residual error relative to the true concept. Define the finite H-stable aggregate:

$$S_n(\mathbf{x}) := f_1(\mathbf{x}) \oplus \cdots \oplus f_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}),$$

equivalently,

$$S_0(\mathbf{x}) = 0, \qquad S_i(\mathbf{x}) = \frac{(i-1)S_{i-1}(\mathbf{x}) + f_i(\mathbf{x})}{i}.$$

Assume the errors are uniformly bounded, $\|e_i\| \le B$, and let:

$$\bar{\epsilon}_n := \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \frac{|\langle e_i, e_j \rangle|}{\|e_i\| \, \|e_j\|}$$

denote the average residual correlation of the finite pool.

Theorem 5 (Finite-Capacity $\mathbb{H}$-Stability)**.** *Let $\mathbb{H}_n = \{H_1, \dots, H_n\}$ be a finite noisily consistent hypothesis pool, and let $S_n = \frac{1}{n} \sum_{i=1}^{n} f_i$ be the renormalized aggregate. If $f_i = c + e_i$, $\|e_i\| \le B$, and the average residual correlation is $\bar{\epsilon}_n$, then:*

$$\|S_n - c\| \le B \sqrt{\frac{1}{n} + \frac{n-1}{n} \bar{\epsilon}_n}.$$

Consequently, if evaluation at $\mathbf{x}$ is bounded by $\|g(\mathbf{x})\|_\infty \le \kappa_{\mathbf{x}} \|g\|$ and the true concept has margin:

$$m(\mathbf{x}) := [c(\mathbf{x})]_{\mathbf{y}^\star} - \max_{\mathbf{y} \ne \mathbf{y}^\star} [c(\mathbf{x})]_{\mathbf{y}}, \qquad \mathbf{y}^\star = c(\mathbf{x}),$$

then HST recovers the true concept at $\mathbf{x}$ whenever:

$$\kappa_{\mathbf{x}} B \sqrt{\frac{1}{n} + \frac{n-1}{n} \bar{\epsilon}_n} < \frac{m(\mathbf{x})}{2}.$$

In particular, if $\bar{\epsilon}_n = O(1/n)$, then:

$$\|S_n - c\| = O(n^{-1/2}).$$

Proof. Since $f_i = c + e_i$,

$$S_n - c = \frac{1}{n}\sum_{i=1}^{n} e_i.$$

Therefore,

$$\|S_n - c\|^2 = \left\|\frac{1}{n}\sum_{i=1}^{n} e_i\right\|^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle e_i, e_j\rangle.$$

Splitting diagonal and off-diagonal terms gives

$$\|S_n - c\|^2 = \frac{1}{n^2}\sum_{i=1}^{n}\|e_i\|^2 + \frac{2}{n^2}\sum_{1\le i<j\le n}\langle e_i, e_j\rangle.$$

Using $\|e_i\| \le B$, the diagonal term is bounded by

$$\frac{1}{n^2}\sum_{i=1}^{n}\|e_i\|^2 \le \frac{B^2}{n}.$$

For the off-diagonal term, by the definition of $\bar{\epsilon}_n$,

$$\frac{2}{n^2}\sum_{1\le i<j\le n}|\langle e_i, e_j\rangle| \le \frac{2B^2}{n^2}\sum_{1\le i<j\le n}\frac{|\langle e_i, e_j\rangle|}{\|e_i\|\,\|e_j\|} = B^2\frac{n-1}{n}\bar{\epsilon}_n.$$

Combining the two bounds,

$$\|S_n - c\|^2 \le B^2\left(\frac{1}{n} + \frac{n-1}{n}\bar{\epsilon}_n\right),$$

and taking square roots yields

$$\|S_n - c\| \le B\sqrt{\frac{1}{n} + \frac{n-1}{n}\bar{\epsilon}_n}.$$

For the label-recovery statement, suppose $\|S_n(\mathbf{x}) - c(\mathbf{x})\|_\infty < m(\mathbf{x})/2$. Then the true coordinate satisfies

$$[S_n(\mathbf{x})]_{\mathbf{y}^\star} \ge [c(\mathbf{x})]_{\mathbf{y}^\star} - m(\mathbf{x})/2,$$

while for any $\mathbf{y} \ne \mathbf{y}^\star$,

$$[S_n(\mathbf{x})]_{\mathbf{y}} \le [c(\mathbf{x})]_{\mathbf{y}} + m(\mathbf{x})/2 \le [c(\mathbf{x})]_{\mathbf{y}^\star} - m(\mathbf{x}) + m(\mathbf{x})/2 = [c(\mathbf{x})]_{\mathbf{y}^\star} - m(\mathbf{x})/2.$$

Thus the true class remains the unique maximizer. Since $\|S_n(\mathbf{x}) - c(\mathbf{x})\|_\infty \le \kappa_{\mathbf{x}}\|S_n - c\|$, the stated margin condition implies

$$\arg\max_{\mathbf{y}}[S_n(\mathbf{x})]_{\mathbf{y}} = c(\mathbf{x}).$$

If $\bar{\epsilon}_n = O(1/n)$, then the bound becomes $\|S_n - c\| = O(n^{-1/2})$. □

Intuition: Theorem 5 gives the finite-pool analogue of $\mathbb{H}$-stability. The aggregate error has two sources: the averaging term $1/n$, which is the usual variance-reduction effect from combining bounded residuals, and the correlation term $\bar{\epsilon}_n$, which measures how aligned the residual errors of the finite hypothesis pool remain. Thus, HST improves as either the pool grows or the hypotheses become more decorrelated. If residual correlations decay as $\bar{\epsilon}_n = O(1/n)$, the aggregate approaches the true concept at the standard $O(1/\sqrt{n})$ rate. If correlations plateau, the theorem predicts a finite error floor, which matches the practical saturation observed with modest hypothesis pools.

Method	CIFAR-10					CIFAR-100				
	Symmetric (%)				Asym. (%)	Symmetric (%)				Asym. (%)
	10	20	30	40	40	10	20	30	40	40
NLS (ICML 2022)	6.64 ± 0.85	8.05 ± 1.02	10.57 ± 0.98	12.51 ± 1.12	11.72 ± 1.02	7.36 ± 1.15	14.21 ± 1.10	15.57 ± 1.38	18.01 ± 1.55	15.41 ± 1.55
+Ours (HST)	**3.35 ± 0.07**	**4.86 ± 0.19**	6.30 ± 0.30	7.88 ± 0.42	7.02 ± 0.44	**6.38 ± 0.18**	**11.02 ± 0.48**	**12.35 ± 0.57**	14.98 ± 0.77	11.36 ± 0.75
+Ours (CRT)	4.06 ± 0.12	5.10 ± 0.20	**5.85 ± 0.24**	**6.97 ± 0.37**	**6.11 ± 0.32**	6.50 ± 0.25	12.03 ± 0.50	12.98 ± 0.63	**14.60 ± 0.71**	**10.45 ± 0.68**
NoiseGPT (NeurIPS 2024)	6.83 ± 0.48	7.42 ± 0.77	8.10 ± 0.83	9.40 ± 0.92	8.98 ± 1.02	8.95 ± 0.93	16.13 ± 1.05	16.67 ± 1.25	17.25 ± 1.55	13.61 ± 1.20
+Ours (HST)	**4.60 ± 0.13**	**5.21 ± 0.20**	5.85 ± 0.25	**6.86 ± 0.18**	6.21 ± 0.21	**7.32 ± 0.24**	**13.51 ± 0.55**	**13.87 ± 0.61**	15.68 ± 0.72	13.55 ± 0.65
+Ours (CRT)	4.98 ± 0.06	5.35 ± 0.07	**5.02 ± 0.24**	6.98 ± 0.28	**6.15 ± 0.25**	7.98 ± 0.28	14.60 ± 0.45	15.08 ± 0.41	**15.51 ± 0.66**	**12.90 ± 0.71**
PLS-LSA+ (ECCV 2024)	7.55 ± 1.16	9.39 ± 1.31	9.30 ± 1.44	11.31 ± 1.28	10.68 ± 1.14	7.80 ± 1.09	14.05 ± 1.16	16.88 ± 1.40	19.37 ± 1.38	15.80 ± 1.28
+Ours (HST)	**5.21 ± 0.07**	5.98 ± 0.11	**6.70 ± 0.16**	7.80 ± 0.22	7.19 ± 0.31	**6.71 ± 0.20**	**11.34 ± 0.30**	**14.28 ±0.35**	15.96 ± 0.45	12.50 ± 0.51
+Ours (CRT)	5.56 ± 0.20	**5.91 ± 0.18**	6.88 ± 0.22	**7.35 ± 0.27**	**6.90 ± 0.23**	7.01 ± 0.28	12.33 ± 0.44	15.50 ± 0.51	**15.70 ± 0.71**	**11.80 ± 0.42**
Delora (ACL 2025)	7.61 ± 0.72	8.37 ± 0.96	9.63 ± 1.15	10.71 ± 1.37	9.87 ± 1.20	7.90 ± 0.98	14.20 ± 1.07	15.97 ± 1.22	17.65 ± 1.20	14.27 ± 1.51
+Ours (HST)	**5.10 ± 0.06**	**5.40 ± 0.09**	**6.31 ± 0.18**	7.29 ± 0.22	7.05 ± 0.15	**6.48 ± 0.27**	**11.97 ± 0.42**	**13.21 ± 0.65**	15.35 ± 0.67	12.68 ± 0.60
+Ours (CRT)	5.45 ± 0.11	6.28 ± 0.15	6.55 ±0.23	**6.87 ± 0.20**	**6.70 ± 0.14**	7.25 ± 0.15	13.86 ± 0.38	14.22 ± 0.52	**14.98 ± 0.44**	**12.40 ± 0.47**

Table 2. Label noise rectification error rate with standard deviation / confidence (lower is better for both) on CIFAR-10 and CIFAR-100.

C. Experiments Continued

C.1. Implementation Details

Backbones: For the vision datasets (MNIST and CIFAR), we use pretrained ResNet34 (He et al., 2015) features obtained from Wei et al. (2022) as the baseline. For T-Finance, we use pretrained features obtained from Tang et al. (2022).

HST: For each substrate detector D evaluated in our experiments, we construct the HST hypothesis pool by combining D with four auxiliary detectors: Confident Learning (Northcutt et al., 2021), AUM (Pleiss et al., 2020), DivideMix (Li et al., 2020), and Co-Teaching (Han et al., 2018). Thus, the pool is substrate-specific: when $D = \mathrm{NLS}$ (Wei et al., 2022), NLS is included as the substrate; for other substrates, such as NoiseGPT, PLS-LSA+, or Delora, the corresponding substrate replaces NLS as the fifth member of the pool. This ensures that HST is evaluated as a reliability amplifier for the detector under consideration, rather than as a fixed ensemble containing one particular baseline.

For each member of the resulting pool, we obtain noise-detection outputs from two independently initialized runs, yielding a total of 10 hypotheses per HST instance. We treat independent random reinitializations as a practical source of functional diversity, since neural-network training trajectories from different random initializations are known to converge to significantly different functional forms (Altıntaş et al., 2025). While this does not constitute a formal guarantee of disjointness, it provides a practical finite-pool approximation to the diversity condition required by $\mathbb{H}$-stability.

The detector families comprising each HST pool are chosen to be mechanistically diverse. Confident Learning estimates joint label–prediction error statistics, grounding its signal in calibrated probability estimates rather than training dynamics. AUM tracks the evolution of per-sample margins, capturing geometric information about how examples behave during optimization. DivideMix separates clean and noisy samples by fitting a Gaussian mixture model to per-sample losses, using probabilistic clustering rather than margins or error-rate estimation. Co-Teaching introduces a curriculum-based dynamic in which two networks exchange small-loss samples, exploiting early-learning behavior rather than mixture modeling, margin trajectories, or probability-transition estimation. The substrate detector D contributes an additional, detector-specific signal. For example, when $D = \mathrm{NLS}$, the substrate signal arises from the way naive label smoothing amplifies corrupted targets, which is rooted in target-manipulation sensitivity rather than model predictions or loss distributions. Analogous distinctions hold for the other substrates. Together, these components span statistical estimation, geometric dynamics, probabilistic clustering, curriculum-based early learning, and substrate-specific noise signatures, making the pool sufficiently diverse for the finite-capacity HST approximation.

We train each constituent model until the first epoch at which at most an η fraction of training samples are misclassified, where η is the noise rate. This stopping rule is used to approximate the noisy-consistency condition in Definition 2, which requires the constituent hypotheses to fit the observed noisy labels up to the expected corruption level. Operationally, this increases the likelihood that most clean samples have been correctly learned and that the remaining training error is primarily attributable to mislabeled samples.

CRT: For CRT on the vision datasets (MNIST and CIFAR), we use pretrained ResNet34 (He et al., 2015) features obtained from Wei et al. (2022) as the baseline. For T-Finance, we use pretrained features obtained from Tang et al. (2022). We then train an MLP of the same architecture as above with the respective substrates on top of the said frozen features to

Method	Noise Rates			
	10%	**20%**	**30%**	**40%**
NLS (ICML 2022)	7.77 ± 1.85	11.55 ± 1.55	16.36 ± 2.78	18.81 ± 2.91
+Ours (HST)	**5.21 ± 0.36**	**9.35± 0.48**	**12.02± 0.60**	16.98 ± 0.85
+Ours (CRT)	5.35 ± 0.49	9.88 ± 0.51	13.15 ± 0.61	**15.33 ± 0.65**
NoiseGPT (NeurIPS 2024)	8.61 ± 1.03	11.02 ± 2.76	14.41 ± 2.66	18.88 ± 3.29
+Ours (HST)	**5.95 ± 0.20**	**9.88 ± 0.22**	13.62 ± 0.58	15.83 ± 0.61
+Ours (CRT)	6.27 ± 0.40	10.51 ± 0.65	**12.58 ± 0.66**	**15.27 ± 0.71**
PLS-LSA+ (ECCV 2024)	8.22 ± 2.80	12.66 ± 1.79	15.37 ± 2.55	19.89 ± 3.08
+Ours (HST)	**5.70 ± 0.30**	**10.35 ± 0.33**	**12.95 ± 0.42**	15.96 ± 0.55
+Ours (CRT)	6.08 ± 0.39	11.18 ± 0.42	13.86 ± 0.38	**15.41 ± 0.41**
Delora (ACL 2025)	7.56 ± 1.76	11.77 ± 2.88	18.78 ± 2.91	20.92 ± 2.95
+Ours (HST)	**5.51 ± 0.26**	**9.70 ± 0.22**	**13.86 ± 0.48**	17.20 ± 0.51
+Ours (CRT)	6.25 ± 0.40	10.30 ± 0.51	14.80 ± 0.49	**16.77 ± 0.53**

Table 3. Label noise rectification error rate and associated standard deviation / confidence (lower is better) on T-Finance.

Method	MNIST		CIFAR-10		CIFAR-100		T-Finance
	Sym.	**Asym.**	**Sym.**	**Asym.**	**Sym.**	**Asym.**	
NLS (ICML 2022)	89.20 ± 0.90	86.55 ± 1.31	83.69 ± 0.89	82.55 ± 1.24	84.79 ± 0.95	79.21 ± 1.20	85.60 ± 1.15
+Ours (HST)	**93.02 ± 0.30**	94.55 ± 0.56	91.86 ± 0.21	90.27 ± 0.40	**87.55 ± 0.50**	**85.81 ± 0.28**	**89.20 ± 0.71**
+Ours (CRT)	91.55 ± 0.42	**94.90 ± 0.35**	**92.20 ± 0.61**	**90.44 ± 0.68**	85.86 ± 0.51	84.71 ± 0.71	88.20 ± 0.61
NoiseGPT (NeurIPS 2024)	87.20 ± 1.20	86.38 ± 1.35	88.51 ± 1.12	87.38 ± 0.90	85.89 ± 1.25	85.40 ± 1.29	82.70 ± 1.15
+Ours (HST)	**90.55 ± 0.35**	**90.98 ± 0.62**	**89.21 ± 0.42**	90.56 ± 0.51	**88.70 ± 0.61**	**87.31 ± 0.39**	86.50 ± 0.15
+Ours (CRT)	90.26 ± 0.21	89.28 ± 0.45	88.90 ± 0.30	**91.77 ± 0.21**	87.68 ± 0.55	86.25 ± 0.51	**87.31 ± 0.72**
PLS-LSA+ (ECCV 2024)	89.50 ± 1.05	88.60 ± 1.20	87.32 ± 1.08	85.21 ±1.35	86.86 ± 1.16	84.31 ± 1.20	83.37 ± 1.45
+Ours (HST)	**92.57 ± 0.25**	91.86 ± 0.30	**90.30 ± 0.55**	**92.10 ± 0.21**	**87.55 ± 0.59**	86.30 ± 0.60	85.80 ± 0.70
+Ours (CRT)	91.57 ± 0.40	**93.66 ± 0.45**	89.70 ± 0.50	90.58 ± 0.42	87.20 ± 0.70	**88.50 ± 0.55**	**87.50 ± 0.60**
Delora (ACL 2025)	89.20 ± 1.36	87.50 ± 1.32	88.50 ± 1.40	89.20 ± 1.27	85.77 ± 1.37	86.30 ± 1.05	84.29 ± 1.22
+Ours (HST)	**90.30 ± 0.30**	91.29 ± 0.76	89.96 ± 0.40	**90.55 ± 0.61**	87.32 ± 0.25	88.55 ± 0.39	88.22 ± 0.44
+Ours (CRT)	89.80 ± 0.56	**92.40 ± 0.20**	**90.57 ± 0.54**	89.86 ± 0.35	**88.71 ± 0.71**	**89.33 ± 0.46**	**89.20 ± 0.51**

Table 4. Label noise detection F1-score (higher is better) across datasets. Symmetric values are averaged over noise rates 10%, 20%, 30%, and 40%; asymmetric values correspond to a 40% noise rate. T-Finance is averaged over noise rates 10%, 20%, 30%, and 40%.

obtain the trajectory for running CRT. Following Section 4.3, we aim for a target configuration with the epoch–learning-rate combination of $(30, 579, 1.6e-03)$. However, we find that an initially high learning rate can speed up convergence without significantly affecting rectification performance So, we start with an initial learning rate of 1.1 and progressively decay it by a factor of 1.1 every 1500 epochs, until we reach zero training loss for all samples.

Complexity Analyses: For the complexity analysis experiments on MNIST, we use an MLP with three layers: $[128, 64, 10]$, with hidden ReLU activations. This is because neural persistence (Rieck et al., 2019), the most reliable complexity measure for our task, has a robust implementation with a complete theory only for MLPs and not for more complex models like CNNs or GNNs. Although Rieck et al. (2019) do propose an approximate way to implement neural persistence for CNNs, it is not as well studied, which introduces an additional layer of uncertainty to our analyses that we could not afford if we are to keep the understanding of our key results in focus.

Method	Rect. Err. (%) ↓	F1 (%) ↑	Recall (%) ↑
NLS (ICML 2022)	11.55 ± 2.36	88.28 ± 1.60	89.59 ± 1.87
+Ours (HST)	**8.02 ± 0.25**	**91.32 ± 0.68**	93.39 ± 0.79
+Ours (CRT)	8.31 ± 0.40	91.18 ± 0.28	**95.40 ± 0.20**
NoiseGPT (NeurIPS 2024)	9.26 ± 1.98	89.21 ± 1.75	87.02 ± 1.80
+Ours (HST)	**7.02 ± 0.35**	**92.55 ± 0.20**	88.56 ± 0.85
+Ours (CRT)	8.11 ± 0.55	90.27 ± 0.28	**91.30 ± 0.66**
PLS-LSA+ (ECCV 2024)	9.37 ± 2.29	90.08 ± 1.96	88.68 ± 1.80
+Ours (HST)	**8.59 ± 0.28**	91.37 ± 0.35	86.20 ± 0.50
+Ours (CRT)	9.17 ± 0.40	**92.06 ± 0.42**	**90.55 ± 0.29**
Delora (ACL 2025)	9.11 ± 1.58	90.51 ± 1.80	92.55 ± 1.86
+Ours (HST)	**8.29 ± 0.33**	**92.80 ± 0.45**	89.60 ± 0.51
+Ours (CRT)	9.05 ± 0.55	92.50 ± 0.61	**93.95 ± 0.45**

Table 5. Label noise rectification on Clothing1M, a real-world benchmark with instance-dependent, structured label noise. Rectification error (lower is better), F1 score, and Recall (both higher is better).

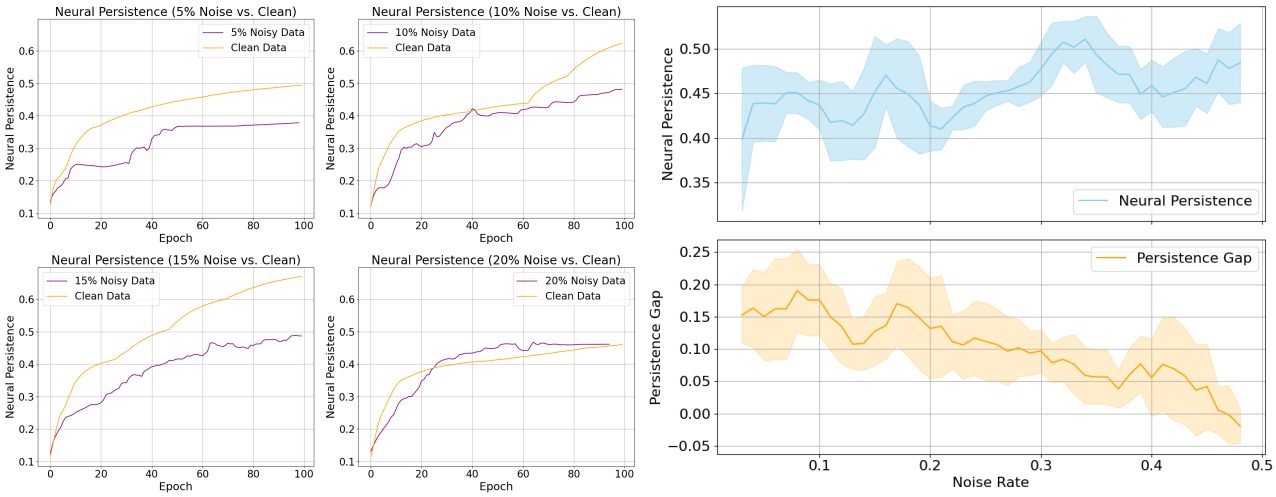

Figure 6. (Left) Complexity / persistence gaps at lower noise rates and (right) which decays with increasing noise rates.

C.2. Complexity gap

Objective and Settings: As seen in Figure 6 (left), at some initially low noise rates ($0 - 15\%$, we find a phenomenon that appears to conflict with our theory on the complexity barrier (Section 3.1), wherein models trained with early stopping (100 epochs) on clean data resulted in higher complexity solutions than noisy data. This effect can be seen to disappear at around a noise rate of 20%. To ascertain whether this is really an artifact of low noise rates, we plot the complexity gap (clean complexity - noisy complexity) across noise rates in the range $[0.1, 0.5]$, with a step size of 0.01. We also monitor the trend in the complexity of the model trained with the noisy data as the noise rate increases, which we report in Figure 6 – right.

Observations and Analysis: The complexity gap indeed turns out to be an artifact of low noise rates, as it steadily decreases with increasing noise rates. The overall complexity also increases with increasing noise rates, reaffirming our theory on the complexity barrier (Section 3.1). It is also partly a result of early stopping, as the complexity gap should decay much faster according to our findings in Section 4.1. These observations are also in line with existing theoretical results which show that small amounts of noise at the level of the input can act as a regularizer (Bishop, 1995), and small amounts of label noise can help neural networks escape bad minima (Zhou et al., 2019), explaining our relatively low-complexity solutions at low noise

rates under early stopping.

C.3. Limitations of Loss-Based Noise Detection

Objective and Settings: We aim to illustrate the limitations of loss-based signals for separating clean and noisily labeled samples and argue why complexity-based signals may be a better alternative. For this purpose, we consider two datasets with different levels of complexity – a simple synthetic dataset with 2 features and 2 classes, and the MNIST dataset, which is significantly more complex. Both datasets were corrupted with 10% random symmetric noise. We train a three-layer MLP: $[128, 64, 10]$, with hidden ReLU activations on the two datasets and monitor their sample-wise losses for 500 epochs. We assign different colors to the trajectories of the clean and the noisy samples and present our observations for a randomly selected subset of datapoints in Figure 7.

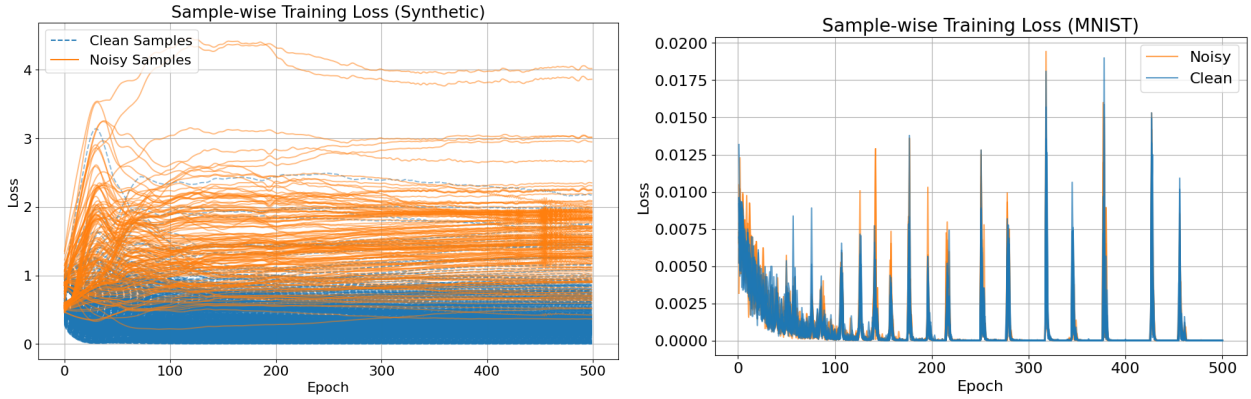

Figure 7. Higher dataset complexity makes it harder to use loss values for separating clean and noisy samples.

Observations and Analyses: For the simple synthetic dataset, the clean and the noisy loss trajectories can be seen to neatly separate out quite early on in training. However, this is not the case for MNIST, which is significantly more complex, for which, the losses remain entangled until the very end. We hypothesize that this is a consequence of the reversal patterns noted in Section 4.2. Loss trajectories tend to revert to their original paths even when faced with targeted, high-density batches of noisy samples. This makes sample-level trajectories hard to distinguish when the density is low (noisy samples are evenly spread out across mini-batches and occur consistently from the start of the training, *i.e.*, not suddenly at some particular mini-batch).

However, as seen in Figure 4, unlike loss trajectories, complexity trajectories do not exhibit any reversal pattern, meaning that once a model learns a noisy sample, its complexity remains permanently higher relative to its clean-only state. This makes model complexity a potentially strong candidate for identifying noisy samples. The current limitation is that there is no known way to robustly estimate the change in model complexity induced by a single sample – existing complexity measures such as neural persistence are too sensitive at such small scales. Being able to overcome this limitation can open up the possibility of more efficient label noise detectors that can directly leverage the complexity barrier to their advantage.

C.4. Additional Practical Considerations for HST and CRT

Scope of comparison with existing detectors: HST and CRT are designed as *reliability amplifiers* rather than as direct replacements for existing label-noise detectors. Therefore, comparing HST/CRT-augmented variants against individual detectors solely in terms of final rectification error can be misleading if interpreted as a claim of computational or algorithmic dominance. Our primary goal is to demonstrate that functional-process symmetries can be approximated in practice and that such approximations improve reliability, reduce variance, and move predictions toward the true-concept invariant. In this sense, the relevant claim is functional rather than purely performative: HST and CRT provide a mechanism by which existing detectors can be made more reliable as the corresponding process approximation is refined.

This distinction is reflected in our experimental design. The main empirical sections prioritize evidence for the complexity barrier, convergence toward the true-concept invariant, and variance collapse. Comparisons with existing detectors are included to demonstrate that the proposed functional-process augmentations are practically useful when applied to standard detectors. We do not claim that the specific implementations of HST and CRT used here are the most efficient possible

Dataset	20% (S)	40% (S)	40% (A)
MNIST	3.10 (3.20)	5.78 (5.70)	5.05 (5.11)
CIFAR-10	4.99 (4.86)	7.80 (7.88)	7.15 (7.02)
CIFAR-100	11.18 (11.02)	15.02 (14.98)	11.25 (11.36)
T-Finance	–	–	16.79 (16.98)

Table 6. HST rectification error using estimated noise rates instead of oracle η. Values in parentheses denote the corresponding oracle-η results. S and A denote symmetric and asymmetric label noise, respectively. The symmetric noise setting is not applicable to T-Finance.

Method	75% Reduction	90% Reduction	100% Reduction
HST	0.70 h	1.15 h	1.42 h
CRT	0.13 h	0.20 h	0.41 h
CRT speedup over HST	5.38×	5.75×	3.46×

Table 7. Additional computational cost of HST and CRT on MNIST. Wall-clock (WC) time is measured in hours over the baseline using 8 NVIDIA RTX 2080 GPUs. Percentages denote the fraction of the final error-rate reduction achieved.

realizations of the theory. Instead, they should be viewed as first practical instantiations of the asymptotic mechanisms established in the main text.

Using estimated rather than oracle noise rates: The HST implementation in Section B.7 uses the noise rate η to determine when constituent predictors have reached the noisily consistent regime. Exact knowledge of η is often unavailable in practice. However, this assumption is common, either explicitly or implicitly, in noisy-label learning pipelines, and estimating the label-noise transition structure is itself an established problem. To test whether HST is sensitive to oracle access to η, we replace the true noise rate with the estimated noise rates obtained from the estimator of Liu et al. (2023). The resulting rectification errors are reported in Table 6. Values in parentheses correspond to the original oracle-η results.

Across all evaluated settings, replacing oracle η with estimated noise rates changes the rectification error only marginally, remaining within approximately 0.2% absolute deviation of the oracle setting. This suggests that HST is compatible with existing noise-rate estimation procedures and that our contribution is orthogonal to the label-noise transition estimation literature.

Computational cost of HST and CRT: HST and CRT instantiate two different approximations of the underlying functional-process symmetries. HST improves the approximation by aggregating over a larger and more diverse hypothesis pool, whereas CRT improves the approximation by extending and refining the training trajectory. Consequently, HST is generally more expensive when implemented using multiple distinct detector families, while CRT is often the lower-cost option when additional efficiency is required.

To quantify the overhead, we measure additional wall-clock time over the baseline on MNIST using 8 NVIDIA RTX 2080 GPUs. The percentages in Table 7 indicate the fraction of the final error-rate reduction achieved by the corresponding method.

The results show that most of the practical gain appears well before the full limiting procedure is approximated. CRT reaches 90% of its final error-rate reduction with only 0.20 additional wall-clock hours, while HST requires 1.15 additional hours for the same reduction level. Thus, although HST provides a principled aggregation-based route to reliability amplification, CRT offers a substantially cheaper trajectory-based alternative.

Low-cost HST via random reinitializations: A practical variant of HST can be obtained by aggregating multiple independently reinitialized copies of the same base detector. This variant is less faithful to the theoretical HST formulation because random reinitializations need not satisfy the same degree of hypothesis-class diversity or noisy consistency as distinct detector families. Nevertheless, it can be useful in resource-constrained settings because multiple random seeds are often already used to estimate confidence intervals or uncertainty. In such cases, the additional overhead of random-seed HST is effectively negligible.

Based on the results in Table 8, the random-seed variant remains competitive with existing detectors while incurring almost

Variant	Extra Overhead	Pool Size	Reported Rectification Errors
Full HST	Multiple distinct hypothesis classes	10	2.91, 3.20, 4.80, 5.70, 5.11
Random-seed HST	Approx. no additional overhead	10	4.28, 5.31, 6.31, 7.25, 7.23

Table 8. Practical low-cost HST variant using random reinitializations. This variant incurs approximately no additional wall-clock overhead when random seeds are already used for uncertainty estimation, but is less faithful to the theoretical HST formulation.

no additional computational cost when random seeds are already part of the evaluation protocol. We emphasize, however, that this is a practical resource-constrained alternative rather than the canonical realization of HST: the theoretical guarantee is most directly aligned with aggregating noisily consistent and sufficiently diverse hypothesis classes.