

USENIX

THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

Certification of Machine Learning Models via Directional Sharpness

Gefei Tan, *Northwestern University & Google;* Adrià Gascón, *Google;* Sarah Meiklejohn, *University College London;* Mariana Raykova, *Google*

https://www.usenix.org/conference/usenixsecurity26/presentation/tan

This paper is included in the Proceedings of the 35th USENIX Security Symposium.

August 12–14, 2026 • Baltimore, MD, USA

ISBN 978-1-939133-58-8

Open access to the Proceedings of the 35th USENIX Security Symposium is sponsored by



جامعة الملك عبدالله للعلوم والتقنية

King Abdullah University of Science and Technology

ARTIFACT EVALUATED
USENIX
AVAILABLE

Certification of Machine Learning Models via Directional Sharpness

Gefei Tan
Northwestern University & Google
gefeitan@u.northwestern.edu

Adrià Gascón
Google
adriag@google.com

Sarah Meiklejohn*
University College London
s.meiklejohn@ucl.ac.uk

Mariana Raykova
Google
marianar@google.com

Abstract

In machine learning, model certification has been identified as an important method for gaining assurance about a model's trustworthiness and quality. A model's quality is largely determined by its ability to generalize, i.e., to perform well on data beyond what it was trained on. It is not possible to certify generalization directly, however, as it depends on unknown data and is not directly measurable. Proxies such as test accuracy can be misleading when the training process is perturbed (intentionally or accidentally), and metrics such as *sharpness*—which has an empirically supported link to generalization—are computationally expensive and can also serve as unreliable signals when training deviates from a prescribed procedure. In this work, we propose *directional sharpness*, a metric designed to efficiently and reliably indicate generalization despite potential training deviations. We provide empirical and analytical evidence that directional sharpness (1) correlates more strongly with generalization than existing metrics and (2) identifies models with poor generalization more reliably than existing metrics. Furthermore, directional sharpness is efficiently computable in model auditing settings, where the verifier has access to training data, and via zero-knowledge proofs that certify quality without revealing training data.

1 Introduction

There are many properties of machine learning (ML) models that model providers and external stakeholders may wish to verify before those models are deployed. Arguably, the most universal property for such *model certification* is the *quality* of the model, in terms of its performance on downstream tasks. A natural and widely accepted notion of model quality is *generalization* [2, 40]: the ability of the model to perform well on new, unseen data.

In practice, model quality is typically estimated using a holdout test set, and a model is said to "generalize well" if it achieves high test accuracy. Test accuracy is not a reliable metric, however, as test sets can contain errors [64], fail to detect overfitting [19, 31, 35], or miss real-world distribution shifts [8, 48, 68, 78]. In adversarial settings, research has shown that a model can achieve high accuracy on test data while being vulnerable to adversarial inputs [32] or containing backdoors that activate only on specific triggers [28, 34, 54].

In statistical learning theory, generalization is captured by the gap between the empirical risk on the training set and population risk on fresh samples from the same (unknown) data distribution. Ideally, a certificate of generalization would show that this gap is small. However, the true data distribution is unknown, and thus generalization cannot be measured directly. Researchers have instead focused on generalization bounds: upper bounds on this gap computable from the training set and model [2, 40, 58, 59, 62, 83]. While theoretically appealing, existing bounds are often numerically vacuous or too loose to track actual performance for modern, over-parameterized neural networks, rendering them unsuitable for certification.

These gaps have motivated empirically grounded *generalization metrics*: quantities computable from the learned model and training data that reliably predict generalization. Such metrics can serve as a principled proxy for generalization without relying on a specific holdout test set.

Recently, the *sharpness* of the loss landscape has emerged as a compelling generalization metric, with empirical and theoretical evidence that models in wide "flat" basins generalize better than those in narrow "sharp" minima [14, 36, 43, 45, 67]. One particularly effective approach measures the *sensitivity* of training loss to small parameter perturbations [21, 25, 45, 50, 67]. The seminal work of Foret et al. [25] introduces sharpness-aware minimization (SAM), an optimizer that explicitly minimizes the model's sharpness. SAM has been shown to significantly improve generalization [14, 25, 50], robustness to noisy labels [6], and adversarial robustness [90] across various benchmarks and model architectures. SAM training is also associated with improved resistance to membership inference attacks [56], backdoor defense [100], and improvements in machine unlearning [22, 80]. These results

*Portions of this work were done while the author was at Google.

provide strong evidence that perturbation-based sharpness is a promising proxy for model generalization.

Despite this explanatory power, existing sharpness metrics have not been explored in the context of certification, likely due to two fundamental barriers. First, these metrics have been validated only via correlation with test accuracy on models trained honestly on clean data. Given that test accuracy itself can fail to detect poor generalization, it remains unclear whether or not sharpness can reliably predict generalization under (intentional or unintentional) training deviations. Second, existing metrics incur prohibitive computational cost even with direct access to the model and training data, let alone the overhead required in a setting where an external stakeholder would like to verify (without access to the training data) a cryptographic proof computed by the model provider.

This motivates our goal: to *design a generalization metric that is strongly predictive of generalization, remains informative under training deviations, and is efficient enough to support privacy-preserving verification.*

Our results. In this paper, we introduce *directional sharpness*, a generalization metric designed specifically for model certification. Unlike static sharpness, which measures the loss landscape geometry at a single point, directional sharpness measures *dynamic stability*: the extent to which a model remains in a flat region under continuous, stochastic perturbation. We show that directional sharpness (1) outperforms existing metrics in terms of empirical correlation with generalization; (2) remains predictive of generalization even under training deviations, making it more sensitive in detecting poor generalization than existing measures; and (3) is significantly more efficient to compute. Directional sharpness thus provides an efficient and robust metric for evaluating generalization when the verifier has access to the training data. We also show how to extend this metric to the setting where the verifier does not have access to the training dataset, via zero-knowledge proofs for directional sharpness.

1.1 Technical Overview

We now outline the key ideas behind directional sharpness. We first explain the shift from static to dynamic measurement, then discuss its theoretical grounding, and finally present empirical validation showing the metric is efficient and enables practical certification.

From static sharpness to dynamic stability. Our starting point is perturbation-based sharpness, which shows strong empirical correlation with generalization in honestly trained models [43]. However, certification demands more than correlation in benign settings: a metric must detect low-quality models that are adversarially chosen to evade detection.

We identify a fundamental limitation of existing sharpness metrics: they are *static*, and evaluate a single, averaged perturbation around the final model to measure sharpness. This one-shot view can miss two failure modes that matter for certification: (i) *incoherent per-example geometry*, where loss landscapes are sharp only on data subsets but appear flat when averaged over the full dataset, and (ii) *unstable flat minima*, where a model sits in an apparently flat, narrow pocket adjacent to sharp regions that a single perturbation may not reliably detect. Both failure modes are linked to models with poor generalization [4, 38], overfitting, and backdoors [29, 95].

Our core insight is that a reliable metric should *verify sharpness dynamically*: a model in a wide, flat basin will remain well-behaved under a *sequence* of small, stochastic perturbations, while models with incoherent or unstable geometry will eventually be pushed into sharp regions. We call this property *dynamic stability*. Convergent evidence from sharpness-aware optimization supports this view: the effectiveness of SAM depends critically on *continuous, stochastic* perturbations computed on small mini-batches, while perturbations using the full dataset are substantially less effective [4, 6, 53].

Measuring dynamic stability via directional sharpness. Building on this insight, we design directional sharpness, a generalization metric that measures the dynamic stability of a model. Given a trained model, we apply a short sequence of SAM-style, stochastic perturbations and record the resulting sharpness after each step. A model in a genuinely flat basin absorbs these perturbations and maintains stable sharpness values; a model in a narrow, unstable region will be pushed toward sharp neighbors, producing detectable fluctuations. As we show in Sections 4.5 and 5, this shift from static to dynamic measurement yields significantly more reliable predictions of generalization, even under training deviations.

Theoretical grounding. We provide theoretical analysis linking directional sharpness to a formal notion of stability. Under standard assumptions, we show that if the initial model $\boldsymbol{w}_0$ is near a stable minimum, then the expected squared sharpness at step t under continuous perturbation is uniformly bounded:

$$\mathbb{E}[s_t^2] \leq \overline{\kappa}_\rho \cdot C \cdot \mathcal{L}_D(\boldsymbol{w}_0) \qquad \text{for all } t,$$

where $C > 0$ and $\overline{\kappa}_\rho$ are constants and $\mathcal{L}_D(\boldsymbol{w}_0)$ is the training loss of the initial model. Conversely, if the expected squared sharpness grows exponentially, then $\boldsymbol{w}_0$ resides in an unstable region, which makes directional sharpness a good indicator of apparently flat but dynamically unstable regions.

We further analyze the robustness of directional sharpness under training deviations. We formalize a class of deviations that induce low gradient coherence, a phenomenon linked to poor generalization, overfitting, and backdoors [12, 13, 70]. For this class of deviations, we show that directional sharpness can detect instability signals that static measures may miss.

We complement our analysis with empirical evaluations showing that directional sharpness (i) correlates with generalization more strongly than static sharpness metrics, and (ii) more consistently detects low-quality models under training deviations like overfitting and backdoors.

Empirical validation and practical certification. We conduct extensive experiments validating directional sharpness for practical model certification. We consider two scenarios: *auditing*, where a trusted verifier has access to training data and computes directional sharpness directly; and *certification*, where an external verifier, without accessing training data, checks a zero-knowledge proof that a potentially adversarial provider's claimed directional sharpness value was computed correctly. We validate three properties that make directional sharpness suitable for both scenarios:

- *strong correlation with generalization*: In Section 4.5, we show that directional sharpness correlates significantly more strongly with generalization than even the best-performing static metrics.
- *reliability under training deviations*: Existing generalization metrics have been evaluated only in the context of honestly trained models. However, for many certification settings, it is important that directional sharpness remain correlated with generalization even under training deviations. In Section 5, we evaluate directional sharpness on models with both intentional deviations (e.g., injected backdoors) and unintentional ones (e.g., overfitting). Directional sharpness reliably flags these models even when both test accuracy and static sharpness fail, confirming that dynamic stability measurement is more sensitive to subtle forms of poor generalization.
- *efficiency*: In the auditing setting, directional sharpness is 4× faster to compute than test accuracy. In certification, its zero-knowledge proof is up to 80,000× faster than proving training correctness step-by-step (88.87 minutes versus an estimated 118,671 hours for VGG-11 over 150 epochs).

1.2 Related Work

ML model auditing and certification. We distinguish between the terms *model auditing* and *model certification* based on the verifier's ability to access the training data. Specifically, we define *model auditing* as a setting where the verifier has access to the training data. Many existing techniques for checking specific trustworthy properties of ML models fall under this category: *fairness* auditing ensures models do not discriminate against particular groups [37]; *differential privacy auditing* [5] empirically tests the theoretical guarantee that a model's output hides the presence of individual training examples; and *robustness* auditing [16, 44, 52, 89] aims to provide provable guarantees against adversarial examples.

Conversely, we define *model certification* as a setting where the verifier has no access to the training data. Many works in privacy-preserving ML adopt this setting, where cryptographic techniques like zero-knowledge proofs allow the verifier to *certify* trustworthy properties while preserving the privacy of the training data. Existing works include certification of properties such as fairness [26, 47, 65, 72, 81, 92, 97], differential-privacy guarantees [71], and vicinity to an optimum [79]. Beyond property-specific guarantees, zero-knowledge proofs of training (zkPoT) allow the verifier to certify that a model was obtained by faithfully running a prescribed training algorithm [1, 30, 77].

Generalization and sharpness. While existing methods can certify the *integrity* of the training process or specific properties, they do not directly certify the most fundamental measure of model quality: its ability to generalize to unseen data. Theoretically, a generalization certificate is a tight, computable upper bound on the population risk [66]. Many works have focused on deriving such generalization bounds for modern ML models [2, 20, 21, 24, 36, 58, 59, 61, 63, 83]. However, existing generalization bounds are often vacuous or too loose when applied to practical neural networks, and thus cannot certify generalization in practice.

A parallel line of research has studied empirical proxies for generalization. Among these, the geometry of the loss landscape, specifically its *sharpness*, has long been linked to generalization, with flatter minima generalizing better [43, 45]. This connection has motivated optimization techniques, such as SAM, that explicitly seek flatter minima to improve generalization [25, 50]. While sharpness has been studied extensively as a generalization proxy, to the best of our knowledge, it has not been explored as a practical certificate of generalization.

2 Preliminaries

2.1 Machine Learning

In this paper, we consider a generalized framework for supervised learning where a model is parameterized by $\boldsymbol{w} \in \mathbb{R}^p$. Let $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ denote the training dataset with inputs $\boldsymbol{x}_i \in \mathbb{R}^d$ and labels $y_i \in \mathbb{R}$. A loss function $\ell(\boldsymbol{w}, (\boldsymbol{x}, y))$ measures the quality of the model's prediction on a data point. The training loss is $\mathcal{L}_D(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^N \ell(\boldsymbol{w}, (\boldsymbol{x}_i, y_i))$, which we abbreviate as $\mathcal{L}(\boldsymbol{w}) := \mathcal{L}_D(\boldsymbol{w})$ and $\ell_i(\boldsymbol{w}) := \ell(\boldsymbol{w}, (\boldsymbol{x}_i, y_i))$ when the dataset is clear from context. A training algorithm $\mathsf{Train}_{\mathcal{L}_D} : \mathcal{R} \to \mathbb{R}^p$ takes a random seed $r \in \mathcal{R}$ and outputs a model $\boldsymbol{w}$ that approximately minimizes $\mathcal{L}_D(\boldsymbol{w})$. We write $g(\boldsymbol{w}) := \nabla \mathcal{L}_D(\boldsymbol{w})$ for the full-dataset gradient and $H(\boldsymbol{w})$ for the Hessian, dropping $\boldsymbol{w}$ when clear from context.

In practice, optimization uses mini-batches $\xi \subset D$ of size $B = |\xi|$ sampled uniformly at random. We write $\mathcal{L}_\xi(\boldsymbol{w})$ for the loss on ξ, and define the stochastic gradient as $g_\xi(\boldsymbol{w}) := \nabla \mathcal{L}_\xi(\boldsymbol{w})$. The gradient noise is $\zeta(\boldsymbol{w}, \xi) := g_\xi(\boldsymbol{w}) - \nabla \mathcal{L}_D(\boldsymbol{w})$.

2.2 Generalization and Sharpness

Following the standard statistical learning framework [2, 62], we assume training examples $z_i := (\boldsymbol{x}_i, y_i)$ are drawn i.i.d. from a fixed but unknown distribution $\mathcal{P}$ over an instance space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The *population risk* of a model $\boldsymbol{w}$ is $R(\boldsymbol{w}) = \mathbb{E}_{z \sim \mathcal{P}}[\ell(\boldsymbol{w}, z)]$, which cannot be computed directly

since $\mathcal{P}$ is unknown. Instead, we estimate it using the *empirical risk*: $\hat{R}(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} \ell(\boldsymbol{w}, z_i)$, which can be computed directly on the training set.

The core challenge in learning is that a model with low empirical risk might not necessarily have a low population risk. The *generalization gap*, $\mathrm{Gap}(\boldsymbol{w}) = R(\boldsymbol{w}) - \hat{R}(\boldsymbol{w})$, quantifies how well training performance predicts test performance. A model generalizes well when this gap is small. In Section 3, we discuss how to certify generalization given that the true gap is unobservable.

Generalization metrics and sharpness. A generalization metric, or *complexity measure* [20, 36, 43, 84], is a quantity that monotonically relates to generalization and is computable without access to a test set. Among the many proposed generalization metrics, one of the most prominent and empirically successful is the *sharpness* of the model [25, 45, 50]. In this paper, we focus on perturbation-based (worst-case) sharpness:

Definition 1 (Worst-Case Sharpness)**.** *Given a loss function $\mathcal{L}_D$ parameterized by the training data D, a radius $\rho > 0$, and norm $\|\cdot\|_p$, the **worst-case sharpness** of parameters $\boldsymbol{w}$ is defined as:* $\mathsf{S}_{\rho,p}(\boldsymbol{w}, D) := \max_{\|\boldsymbol{\varepsilon}\|_p \le \rho} \left(\mathcal{L}_D(\boldsymbol{w}+\boldsymbol{\varepsilon}) - \mathcal{L}_D(\boldsymbol{w})\right)$.

This definition, however, is sensitive to parameter re-scaling [17, 50], which can be exploited to make a sharp minimum appear artificially flat. This led to a magnitude-aware worst-case version invariant to such re-scaling:

Definition 2 (Magnitude-Aware Worst-Case Sharpness [50])**.** *Given a loss function $\mathcal{L}_D$ parameterized by the training data D, a radius $\rho > 0$, and norm $\|\cdot\|_p$, denote $\mathbf{T}_{\boldsymbol{w}} := \mathit{diag}(|\boldsymbol{w}|)$. The **magnitude-aware worst-case sharpness** of parameters $\boldsymbol{w}$ is:* $\mathsf{S}^{\mathsf{mag}}_{\rho,p}(\boldsymbol{w}, D) := \max_{\|\mathbf{T}_{\boldsymbol{w}}^{-1}\boldsymbol{\varepsilon}\|_p \le \rho} \left(\mathcal{L}_D(\boldsymbol{w}+\boldsymbol{\varepsilon}) - \mathcal{L}_D(\boldsymbol{w})\right)$.

2.3 Zero-Knowledge Proofs

A zero-knowledge proof (ZKP) of circuit satisfiability allows a prover holding a witness w to prove to a verifier that $\mathcal{C}(w) = 1$ for some public circuit $\mathcal{C}$ without revealing any information about w. Many existing ZKP protocols can be used for proving directional sharpness, including zk-SNARKs [15, 27, 33, 51] and proofs based on vector oblivious linear evaluation (VOLE) [9–11, 18, 87, 93].

Zero-knowledge proofs of training. Zero-knowledge proofs of training (zkPoTs) [1, 30, 77, 79] allow a prover to prove that a model is trained correctly on a committed dataset without revealing any additional information about the model or the dataset. ZkPoTs can be used for distributed model training, privacy auditing, and proving ownership [1, 71, 91].

3 Certification of Model Generalization

3.1 Certification Threat Model

We now motivate and formalize what it means to certify generalization. The goal of certification is to generate a (short) certificate that enables efficient verification of the certified property. Thus, there are two main steps in the certification process: computation of the certificate (proof) and verification of the certificate. The roles, the threat models, and the available inputs for the parties that execute these steps may differ in different applications. We distinguish between two main settings: *model auditing* and *model certification*.

In model auditing, the verifier is trusted by the model provider to have access to the training data. This includes settings where the model provider uses generalization certificates as a quality metric for candidate model checkpoints and to detect unintentional training issues such as overfitting or memorization. Since the model provider is the natural consumer of the generalization certificates, the threat model for the soundness of the certificates assumes that training has been done honestly.

In external auditing applications, the verifier is a third party, but the trust framework for the auditor provides access to training data as well. In this case, the model provider may or may not be trusted to execute the training honestly.

On the other hand, in model certification, the verifier and the model provider—acting now as a cryptographic *prover*—are mutually distrustful. Thus, the verifier is not given access to the training data, and the prover is not trusted to follow any prescribed steps for the training. In this setting, the prover certifies a claim about the model relative to a committed training dataset D. We assume that D is authenticated, so the prover cannot arbitrarily craft, replace, or modify the dataset used for certification. This assumption is consistent with existing proof-of-training literature [1, 30, 77, 79] and is natural in practical certification settings, where the authenticity of D can be enforced by external authentication, trusted data sources, or additional proofs. Attacks in which the provider can freely choose or forge the committed dataset are therefore outside our threat model.

Bearing in mind the above certification applications, we proceed to describe our approach to certifying generalization and our security goals. The first required step towards computing a generalization certificate is the ability to measure this property. Next, we discuss the challenges to quantifying generalization.

3.2 Quantifying Generalization

As discussed in Section 1, for "certifying generalization" to have practical meaning, it cannot mean proving the exact generalization gap or a loose theoretical bound. Instead, we seek to certify an empirically motivated *generalization metric* that is computable and has direct, actionable implications in practice. We now formalize what it means to certify generalization via such metrics.

Metric-based certification. In our solution, we adopt a metric-based approach in which the generalization certificate for the model is based on a generalization metric computed

on the final model. Specifically, a *generalization metric* is an efficiently computable (randomized) function $\gamma_{\mathcal{L}}$ mapping $(\boldsymbol{w}, D, r)$ to a value $\gamma_{\mathcal{L}}(\boldsymbol{w}, D, r) \in \mathbb{R}_{\geq 0}$, where D and $\boldsymbol{w}$ denote the training set and trained model held by the model provider (prover), $\mathcal{L}$ is a public loss function, and r is a public random seed. Smaller values of $\gamma_{\mathcal{L}}(\boldsymbol{w}, D, r)$ are interpreted as stronger evidence of generalization.

Therefore, by *certificate of generalization* we mean a certificate (proof) of a predicate on this metric: $\mathsf{Pred}_{\tau}(\boldsymbol{w}, D, r) \equiv [\gamma_{\mathcal{L}}(\boldsymbol{w}, D, r) \leq \tau]$, where the threshold τ and seed r are public. When instantiated with a ZKP with knowledge soundness, this means that any computationally bounded prover who convinces the verifier must know a model $\boldsymbol{w}$ and dataset D satisfying $\mathsf{Pred}_{\tau}(\boldsymbol{w}, D, r)$. Thus, the prover cannot falsely claim a lower metric value for the model $\boldsymbol{w}$ and its committed dataset D. Importantly, the role of certification here is to verify the value of $\gamma_{\mathcal{L}}(\boldsymbol{w}, D, r)$ and not to prove the link between $\gamma_{\mathcal{L}}$ and the true generalization gap.

3.3 Desiderata of Generalization Metrics in Certification

Having defined metric-based certification, we now ask what properties a generalization metric must satisfy to serve as the basis of a certificate.

Existing evaluation method. Because generalization does not have an observable "ground truth", evaluating how well a metric predicts generalization must be done through some proxy. Most empirical work evaluates a candidate metric by training a large grid of models, computing the metric on each model, and measuring its correlation with an empirical proxy for generalization—most commonly the *empirical generalization gap* measured on a holdout test set (e.g., [43]). Metrics with stronger and more consistent correlation are viewed as more predictive of generalization.

Evaluating metrics for certification. However, existing metrics are primarily proposed as *explanatory* tools for understanding generalization in settings where models are trained honestly and evaluation data is benign. In certification, the training procedure may deviate from the nominal one, either accidentally (e.g., label noise, shortcut learning, or distributional artifacts) or intentionally (e.g., data poisoning or backdoors). Thus, a metric used for certification requires stronger evidence than correlation on honestly trained models alone.

Specifically, we propose the following desiderata that a metric should satisfy for the resulting certificate to be reliable.

1. **Correlated with generalization.** As an empirical generalization metric, it must correlate strongly with the empirical generalization gap across a diverse set of models. This is the established baseline for any generalization metric.

2. **Predictive despite training deviations.** The metric should continue to track generalization even when the training process deviates from the nominal procedure (intentionally or accidentally). Concretely, this means it should be able to identify training deviations that preserve low training loss but degrade generalization, such as overfitting, memorization, spurious shortcuts, or adversarial behaviors that are only triggered under specific conditions.

3. **Efficiently computable.** As a practical certification tool, it must be computationally efficient, requiring at most a handful of mini-batch operations and avoiding full-dataset or Hessian-level costs.

On robustness to training deviations. While a generalization metric used for certification should ideally provide guarantees against any training deviations, we note that formalizing such guarantees faces fundamental challenges. First, the relationship between training deviations and generalization is not well defined. Generalization, in the strict learning-theoretic sense, is defined with respect to a fixed data distribution $\mathcal{P}$. For example, a backdoor model might behave maliciously only on attacker-chosen trigger inputs that lie outside the support of $\mathcal{P}$; such behavior may not increase the generalization gap as classically defined. Second, ruling out *all* strategies that fool a metric would require quantifying over unconstrained adversarial objectives. Some objectives may be statistically independent of the training set and therefore cannot be detected from the model and training data alone; without additional assumptions, universal detection can be impossible in a strong sense.

Given these challenges, we instead take a principled approach to validate our metric. In Section 4.4, we provide analysis of a class of deviations characterized by *low gradient coherence*—a property empirically linked to poor generalization and known deviations—and show that such deviations can remain invisible to existing metrics yet can be detected by directional sharpness. In Section 5, we complement our analysis with extensive experimental evidence that directional sharpness reliably flags models with known deviations even when test accuracy and static sharpness do not.

4 Defining and Validating Directional Sharpness

Section 3 formalized certifying generalization and identified three desiderata for generalization metrics: strong correlation, reliability under training deviations, and computational efficiency. In this section, we develop such a metric based on the dynamic stability of the loss landscape. We begin by reviewing the stochastic behavior of the SAM optimizer and deriving our metric from its dynamics (Sections 4.1 and 4.2), then provide theoretical grounding (Sections 4.3 and 4.4), and present empirical validation (Section 4.5).

4.1 From SAM to Dynamic Stability

The starting point of our generalization metric is Sharpness-Aware Minimization (SAM) [25]. SAM seeks flat minima by minimizing both the loss and its *worst-case sharpness* (Def. 1), yielding the min-max objective:

$$\min_{\boldsymbol{w}} \max_{\|\boldsymbol{\varepsilon}\|_p \le \rho} \mathcal{L}_D(\boldsymbol{w}+\boldsymbol{\varepsilon}). \tag{1}$$

Solving the inner maximization exactly is intractable, so SAM approximates it via a first-order Taylor expansion: $\mathcal{L}_D(\boldsymbol{w}+\boldsymbol{\varepsilon}) \approx \mathcal{L}_D(\boldsymbol{w}) + \boldsymbol{\varepsilon}^\top \nabla \mathcal{L}_D(\boldsymbol{w})$. The inner problem then reduces to $\max_{\|\boldsymbol{\varepsilon}\|_p \le \rho} \boldsymbol{\varepsilon}^\top \nabla \mathcal{L}_D(\boldsymbol{w})$, which is a dual norm problem with the closed-form solution (for $1 < p < \infty$ and $\nabla \mathcal{L}_D(\boldsymbol{w}) \neq 0$):

$$\hat{\boldsymbol{\varepsilon}}(\boldsymbol{w}) = \rho \frac{\operatorname{sign}(\nabla \mathcal{L}_D(\boldsymbol{w})) \, |\nabla \mathcal{L}_D(\boldsymbol{w})|^{q-1}}{\|\nabla \mathcal{L}_D(\boldsymbol{w})\|_q^{q-1}}, \tag{2}$$

where q satisfies $\frac{1}{p} + \frac{1}{q} = 1$. Substituting this back yields:

Definition 3 (Approximated Sharpness [25])**.** *Let $\mathcal{L}_D$ be the loss, and let $\nabla \mathcal{L}_D(\boldsymbol{w})$ be its gradient. Given a radius $\rho > 0$ and a pair of dual p, q norms (where $\frac{1}{p} + \frac{1}{q} = 1$), the approximated sharpness or **SAM sharpness** of model $\boldsymbol{w}$ is:*

$$\mathsf{S}^{\mathsf{sam}}_{\rho,p}(\boldsymbol{w}, D) := \nabla \mathcal{L}_D(\boldsymbol{w})^\top \hat{\boldsymbol{\varepsilon}}(\boldsymbol{w}) = \rho \|\nabla \mathcal{L}_D(\boldsymbol{w})\|_q.$$

ASAM [50] proposes using the magnitude-aware worst-case sharpness (Def. 2) as a more robust objective; this yields:

Definition 4 (Approximated Magnitude-Aware Sharpness [50])**.** *Let $\mathbf{T}_{\boldsymbol{w}}$ denote $diag(|\boldsymbol{w}|)$. Given a radius $\rho > 0$ and a pair of dual p, q norms (where $\frac{1}{p} + \frac{1}{q} = 1$), the magnitude-aware approximated sharpness or **ASAM sharpness** of $\boldsymbol{w}$ is:*

$$\mathsf{S}^{\mathsf{asam}}_{\rho,p}(\boldsymbol{w}, D) := \max_{\|\mathbf{T}_{\boldsymbol{w}}^{-1}\boldsymbol{\varepsilon}\|_p \le \rho} \left(\nabla \mathcal{L}_D(\boldsymbol{w})^\top \boldsymbol{\varepsilon} \right) = \rho \|\mathbf{T}_{\boldsymbol{w}} \nabla \mathcal{L}_D(\boldsymbol{w})\|_q.$$

SAM effectively makes two approximations: (i) replacing the worst-case perturbation with the gradient direction, and (ii) replacing full-batch loss with mini-batch loss. Recent work reveals that these approximations are not merely computational shortcuts but are *essential* for improving generalization: using full-batch sharpness significantly degrades performance [6, 25]. This suggests that SAM succeeds not merely because it finds points of low sharpness, but because it finds minima that remain stable under *continuous*, *stochastic* perturbation—two properties that existing static measures lack:

- **Continuous perturbation.** Static sharpness measures evaluate the loss increase from a single, bounded perturbation, which can be misleading when a minimum is locally flat but adjacent to sharp regions. A sequence of perturbations can incrementally push parameters out of such narrow pockets, revealing the adjacent sharpness through detectable increases in sharpness. This is confirmed by [99]: applying SAM to an SGD-trained model pushes it from the SGD minimum to a much flatter SAM minimum.
- **Stochastic perturbation.** Static measures compute sharpness on the full training set and can potentially mask high sharpness on specific subsets. A model that appears flat on the full dataset but exhibits high sharpness on particular data points would likely lead to poor generalization.

These findings motivate our central proposal: we define a generalization metric that measures sharpness not as a static property at a single point, but as a measure of *dynamic stability*—how consistently flat a model remains under continuous, sharpness-aware perturbation. We detail our approach next.

Input: Initial model parameters $\boldsymbol{w}_0$, dataset D, public random seed r, configuration $\mathcal{C} = (T, B, \Phi, \mathsf{S}, \mathsf{fluc}, \mathcal{L}, \Theta)$.
Output: Directional Sharpness measure $\mathcal{F}$.
$\boldsymbol{w} \leftarrow \boldsymbol{w}_0$
$\mathcal{H} \leftarrow []$
for $t \leftarrow 0$ **to** $T-1$ **do**
  Sample a mini-batch $\xi_t \subset D$ of size B via (r, t)
  $\mathcal{L}_{\xi_t} \leftarrow \mathcal{L}_{\xi_t}(\boldsymbol{w})$
  $s_t \leftarrow \mathsf{S}(\boldsymbol{w}, \xi_t, \Theta)$
  Append $(\boldsymbol{w}, \xi_t, s_t, \mathcal{L}_{\xi_t})$ to $\mathcal{H}$
  $\boldsymbol{w} \leftarrow \Phi(\boldsymbol{w}, \xi_t, \Theta)$
end
$\mathcal{F} \leftarrow \mathsf{fluc}(\mathcal{H})$
return $\mathcal{F}$

Algorithm 1: Computing directional sharpness.

4.2 Our Approach: Dynamic Stability

We define our generalization metric, directional sharpness, as a measure of a model's geometric stability under a series of sharpness-aware perturbations.

Definition 5 (Directional Sharpness)**.** *Given a model $\boldsymbol{w}_0$, a dataset D, a public random seed r, batch size B, number of steps T, a loss function $\mathcal{L}(\cdot)$, a sharpness function $\mathsf{S}(\cdot)$, a SAM update operator $\Phi_\Theta(\cdot)$ with hyperparameters Θ, and a fluctuation statistic $\mathsf{fluc}(\cdot)$, for $t \in \{0, \ldots, T-1\}$, we first generate a probe history $\mathcal{H} = \{h_0, \ldots, h_{T-1}\}$ by iteratively computing: $h_t := (\boldsymbol{w}_t, \xi_t, s_t, \mathcal{L}_{\xi_t})$ where ξ_t is a mini-batch of size B drawn uniformly from D using seed r, $s_t = \mathsf{S}(\boldsymbol{w}_t, \xi_t, \Theta)$, $\mathcal{L}_{\xi_t} = \mathcal{L}_{\xi_t}(\boldsymbol{w}_t)$, and $\boldsymbol{w}_{t+1} := \Phi_\Theta(\boldsymbol{w}_t, \xi_t)$. The **directional sharpness** of model $\boldsymbol{w}_0$ parameterized by $\mathcal{C} = (T, B, \Phi, \mathsf{S}, \mathsf{fluc}, \mathcal{L}, \Theta)$ is defined as:*

$$\mathsf{S}^{\mathsf{d}}_{\mathcal{C}}(\boldsymbol{w}_0, D, r) := \mathsf{fluc}(\mathcal{H}).$$

Algorithm 1 shows the procedure for computing directional sharpness. Given a trained model $\boldsymbol{w}_0$, we apply T steps of

SAM-style updates using freshly sampled mini-batches. At each step t, we record a per-step sharpness $s_t = \mathsf{S}^{\mathsf{sam}}_{\rho,p}(\boldsymbol{w}_t, \xi_t)$ (e.g., SAM sharpness or ASAM sharpness) with mini-batch ξ_t. This probing process yields a time series of SAM sharpness values $\{s_0, s_1, \ldots, s_{T-1}\}$. We define the directional sharpness metric using a statistic that quantifies the stability of this sequence.

Choice of fluctuation statistic. Here fluc is a function that outputs a final non-negative number quantifying the amount of fluctuation in $\mathcal{H}$. In practice, we found that tracking the raw sharpness values $\{s_t\}$ alone can be sensitive to the overall loss scale across models and datasets. We therefore use an *empirical* normalization strategy via $r_t := s_t^2 / \mathcal{L}_{\xi_t}$ and set $\mathsf{fluc}(\mathcal{H}) := \mathrm{Std}(\{\log(r_t + \varepsilon)\}_{t=0}^{T-1})$ for a small $\varepsilon > 0$. Empirically, this ratio-based statistic is slightly more correlated with generalization than using $\{s_t\}$ alone, so we adopt it by default.

The key semantic difference from static sharpness is that directional sharpness tests how a model *responds* to SAM dynamics rather than measuring geometry at a single point. If $\boldsymbol{w}_0$ resides in a truly flat basin—the type SAM converges to—then applying SAM perturbations will produce stable sharpness values. Conversely, if $\boldsymbol{w}_0$ resides in a sharp or unstable region, SAM will push the model toward neighboring sharp areas, producing an unstable series. Thus, a low directional sharpness value suggests that the model remains geometrically stable under the SAM probing dynamics, while a high value indicates instability along this trajectory. We validate empirically that this dynamic-stability signal correlates with generalization and provide a theoretical analysis of this connection in the next section.

4.3 A Theoretical Foundation for Directional Sharpness

In this section, we provide a theoretical analysis for our directional sharpness metric. We proceed by defining a notion of stability based on recent results on SAM dynamics [99], then deriving bounds that connect our sharpness measurements to this stability.

Formalizing stability. Recall that the goal of directional sharpness is to verify whether a trained model exhibits stability properties that are predictive of good generalization. We formalize this stability property using the concept of SAM-stability from recent optimization literature, as recent work suggests that SAM's dynamics capture a generalization-relevant bias stronger than static sharpness measures: among multiple minima of equal static sharpness, SAM exhibits an implicit bias favoring those that generalize better [3, 76, 85, 99].

Concretely, we consider the model we are certifying to have a small training loss and be near some global minimum $\boldsymbol{w}^*$. In this regime, the model can be approximated by the local linearization of its prediction function f: $f_{\mathrm{lin}}(\boldsymbol{x}; \boldsymbol{w}) = f(\boldsymbol{x}; \boldsymbol{w}^*) + \langle \nabla_{\boldsymbol{w}} f(\boldsymbol{x}; \boldsymbol{w}^*), \boldsymbol{w} - \boldsymbol{w}^* \rangle$. The corresponding linearized loss is:

$$\mathcal{L}_{\mathrm{lin}}(\boldsymbol{w}) = \frac{1}{N} \sum_{i=1}^{N} \ell\big(f_{\mathrm{lin}}(\boldsymbol{x}_i; \boldsymbol{w}), y_i\big).$$

This linearization allows us to analyze the local geometry of the loss landscape near convergence. Throughout our analysis, we work in this linearized model and write $\mathcal{L}(\boldsymbol{w}) := \mathcal{L}_{\mathrm{lin}}(\boldsymbol{w})$. We define stability in terms of SAM's behavior on $\mathcal{L}$:

Definition 6 (SAM-Stability)**.** *A minimum $\boldsymbol{w}^*$ is (linearly)* SAM-stable *if there exists $C > 0$ such that for any starting point $\boldsymbol{w}_0$ near $\boldsymbol{w}^*$ where the linearization is valid, applying the SAM optimizer to the loss $\mathcal{L}$ yields a trajectory $\{\boldsymbol{w}_t\}$ satisfying $\mathbb{E}[\mathcal{L}(\boldsymbol{w}_t)] \le C \cdot \mathbb{E}[\mathcal{L}(\boldsymbol{w}_0)]$ for all $t \ge 0$. A minimum that is not SAM-stable is called* SAM-unstable.

Intuitively, a SAM-stable minimum is one where SAM's perturbation-based dynamics remain bounded—the optimizer cannot escape, and the loss stays controlled. Our central claim is that directional sharpness provides an efficient, empirical test for this stability. We show that bounded SAM loss implies bounded expected sharpness, while exponential sharpness growth certifies instability. Thus, a stable basin can yield controlled sharpness along the probing trajectory, and exponential growth rules out a stable basin.

Linking directional sharpness to stability. We establish this claim in two steps. First, we derive a "sandwich" bound showing that the sharpness s_t is tightly coupled to the loss $\mathcal{L}(\boldsymbol{w}_t)$ at each step t. Second, we use this bound to characterize how sharpness evolves at stable versus unstable minima.

We first adopt the following assumptions from [99] and establish that per-step SAM sharpness measurements are coupled to the loss:

Assumption 1 (Smoothness and Polyak-Łojasiewicz Condition [99])**.** *There exist $\mu > 0, L_s > 0$ such that $\|H(\boldsymbol{w})\|_2 \le L_s$ and $\|g(\boldsymbol{w})\|_2^2 \ge 2\mu\,\mathcal{L}(\boldsymbol{w})$ for all $\boldsymbol{w}$.*

Assumption 2 (Bounded Gradient Noise [99])**.** *The gradient noise variance $\zeta(\boldsymbol{w}, \xi) = g_\xi(\boldsymbol{w}) - g(\boldsymbol{w})$ is bounded by the full-dataset loss. That is there exists a constant $\sigma > 0$ such that for all $\boldsymbol{w}$, $\mathbb{E}[\|\zeta(\boldsymbol{w}, \xi)\|_2^2] \le \sigma^2 \mathcal{L}(\boldsymbol{w})$.*

Assumption 1 is standard and frequently used in non-convex optimization. Assumption 2 states that mini-batch noise is controlled by the loss value, a phenomenon that has been empirically observed [23, 60, 88].

Lemma 1 (Linear Sandwich Bound)**.** *Let $\{\boldsymbol{w}_t\}_{t \ge 0}$ be a sequence of models where Assumptions 1 and 2 hold. Let s_t denote the SAM sharpness (Def. 3) of model $\boldsymbol{w}_t$ with $p = q = 2$ and radius ρ, evaluated over a mini-batch of size B sampled uniformly without replacement from D. Then for all t:*

$$\underline{\kappa}_\rho \cdot \mathcal{L}(\boldsymbol{w}_t) \le \mathbb{E}\big[s_t^2 \big| \boldsymbol{w}_t\big] \le \overline{\kappa}_\rho \cdot \mathcal{L}(\boldsymbol{w}_t),$$

where $\underline{\kappa}_\rho = 2\rho^2 \mu$ and $\overline{\kappa}_\rho = \rho^2 \left(2L_s + \frac{\sigma^2}{B}\right)$.

Proof. See Appendix B.1. □

This lemma shows that the expected squared sharpness value is bounded between constant multiples of the loss, with constants depending on gradient noise and local curvature. We now present our main theorem showing the link between directional sharpness and SAM-stability.

Theorem 1 (Sharpness Dynamics and SAM-Stability)**.** *Let $\{\boldsymbol{w}_t\}_{t\geq 0}$ be the trajectory generated by applying SAM to $\mathcal{L}$, starting from $\boldsymbol{w}_0$ near a global minimum $\boldsymbol{w}^*$. Suppose Assumptions 1 and 2 hold, and let s_t denote the mini-batch SAM sharpness at step t. Then the following statements hold:*
(a) Stability implies bounded sharpness. *If $\boldsymbol{w}^*$ is SAM-stable in the sense of Def. 6 with constant $C > 0$, then*

$$\mathbb{E}[s_t^2] \leq \overline{\kappa}_\rho C \mathcal{L}(\boldsymbol{w}_0) \qquad \text{for all } t \geq 0.$$

(b) Exponential sharpness growth certifies instability. *Assume $\mathcal{L}(\boldsymbol{w}_0) > 0$. If there exist $K > 0$, $\alpha > 1$, and $t_0 \geq 0$ such that for all $t \geq t_0$, $\mathbb{E}[s_t^2] \geq K\alpha^t \mathcal{L}(\boldsymbol{w}_0)$, then $\boldsymbol{w}^*$ is SAM-unstable.*

Proof. See Appendix B.2. □

Interpretation. Theorem 1 provides theoretical support for our claim that directional sharpness is directly linked to SAM optimizer dynamics. If the iterate lies in a SAM-stable basin, then the expected sharpness is uniformly bounded (Part (a)). Conversely, exponential growth of the expected sharpness is a sufficient certificate that the iterate is not SAM-stable (Part (b)). While the theorem is one-way, empirically we find that poor generalization and training deviations consistently lead to large directional sharpness scores (Section 5).

Remark (Extension to ASAM). For clarity, our analysis is stated for SAM sharpness (Def. 3). The same argument extends to ASAM (Def. 4), where the perturbation is constrained by $\|(\mathbf{T}_{\boldsymbol{w}})^{-1}\boldsymbol{\varepsilon}\|_p \leq \rho$ for a scaling operator $\mathbf{T}_{\boldsymbol{w}} := \mathrm{diag}(|\boldsymbol{w}|)$. In particular, ASAM uses $\mathbf{T}'_{\boldsymbol{w}} = \mathbf{T}_{\boldsymbol{w}} + \varepsilon_0 I$ with $\varepsilon_0 > 0$ [50], so the per-coordinate scaling is bounded away from 0. Assume each weight magnitude is upper bounded, i.e., $\max_i(|(\boldsymbol{w}_t)_i| + \varepsilon_0) \leq M$ for all $t \in \{0, \ldots, T-1\}$. Then Lemma 1 holds for ASAM with $\underline{\kappa}_\rho = 2\varepsilon_0^2\rho^2\mu$ and $\overline{\kappa}_\rho = M^2\rho^2\left(2L_s + \frac{\sigma^2}{B}\right)$.

4.4 Detecting Training Deviations

We now analyze why directional sharpness can detect poor generalization under training deviations that full-batch sharpness measurements miss. Specifically, we formalize a class of training deviations with low *gradient coherence*, which has been empirically linked to poor generalization behavior and observed in known deviations; we then show that these deviations can be invisible to the full-batch sharpness metric but detectable by directional sharpness.

Gradient coherence measure. Recent work [12] links flat minima and generalizable solutions to the notion of *data coherence*: how consistently different training examples "agree" on the local geometry. Given per-example Hessians $H_i := \nabla^2 \ell_i(\boldsymbol{w})$, the authors define a coherence score over the matrix $\mathbf{S} \in \mathbb{R}^{N\times N}$ with entries $\mathbf{S}_{ij} := \sqrt{\mathrm{Tr}(H_i H_j)}$. Intuitively, $\mathbf{S}_{ij}$ is large when examples i and j have aligned curvature.

Data coherence has also been studied in terms of gradient alignment [13, 70], which can be captured using the gradient Gram matrix G with entries $G_{ij} := \langle g_i(\boldsymbol{w}), g_j(\boldsymbol{w})\rangle$ where $g_i(\boldsymbol{w}) := \nabla \ell_i(\boldsymbol{w})$ is the per-example gradient. Intuitively, large off-diagonal entries in G indicate that two examples have aligned gradients. We can define a coherence score on G by measuring how strongly per-example gradients agree on the average direction:

Definition 7 (Gradient Coherence)**.** *Let $g_i(\boldsymbol{w}) := \nabla \ell_i(\boldsymbol{w})$ and $g(\boldsymbol{w}) := \frac{1}{N}\sum_{i=1}^N g_i(\boldsymbol{w})$. The* gradient coherence *of model $\boldsymbol{w}$ on dataset D is:*

$$c_g(\boldsymbol{w}) := \frac{\frac{1}{N^2}\sum_{ij} G_{ij}}{\frac{1}{N}\mathrm{Tr}(G)} = \frac{\|g(\boldsymbol{w})\|_2^2}{\frac{1}{N}\sum_{i=1}^N \|g_i(\boldsymbol{w})\|_2^2} \in [0,1].$$

Intuitively, high c_g means per-example gradients are largely aligned, while low c_g means substantial cancellation happens across gradients.

Training deviations induce low coherence. Empirically, training deviations often manifest as poorly aligned (incoherent) gradients/curvature across examples. In backdoor attacks, Yuan et al. [95] observe a substantially more dispersed distribution of activation gradients within the backdoor target class, indicating reduced within-class alignment of per-example geometry. Separately, Garg et al. [29] show that memorized examples correlate with high per-example local curvature, suggesting strong per-example heterogeneity rather than shared curvature structure. Motivated by these observations, we formalize the class of training deviations using gradient coherence:

Assumption 3 (Low-Coherence Deviations)**.** *A training deviation produces a model $\boldsymbol{w}$ that achieves low training loss on a clean dataset D of size N with small gradient $\|g(\boldsymbol{w})\|_2 \leq \varepsilon$ and low gradient coherence $c_g(\boldsymbol{w}) \leq c_0 \ll 1$.*

Our assumption says deviations produce models with low training loss and small gradient, yet these models have hidden poor generalization behavior manifested through low gradient coherence.

Directional sharpness detects low gradient coherence. We now demonstrate that directional sharpness can detect low-coherence deviations in cases where static full-batch sharpness cannot. Consider a single step of directional sharpness measuring the mini-batch SAM sharpness $\mathsf{S}_\xi = \rho\|g_\xi(\boldsymbol{w})\|_2$ and a static, full-batch SAM sharpness metric $\mathsf{S} = \rho\|g(\boldsymbol{w})\|_2$. To compare the scale of the random mini-batch signal with the

deterministic full-batch signal, define the root-mean-square (RMS) mini-batch sharpness $\mathsf{S}_{\text{rms}}(\boldsymbol{w}) := \sqrt{\mathbb{E}_\xi[\mathsf{S}_\xi^2]}$. The next theorem shows that under low-coherence deviations, the full-batch sharpness S can be small while the RMS mini-batch sharpness remains large. Moreover, the squared RMS gap $\mathsf{S}_{\text{rms}}^2 - \mathsf{S}^2$ grows as the batch size B decreases.

Theorem 2 (Detection Gap)**.** *Let $\boldsymbol{w}$ be a model produced by a training deviation satisfying Assumption 3. Let ξ be a mini-batch of size B sampled uniformly from $[N]$ without replacement and $g_i := \nabla \ell_i(\boldsymbol{w})$. Then the squared RMS gap between mini-batch sharpness and full-batch sharpness is:*

$$\mathsf{S}_{\text{rms}}^2 - \mathsf{S}^2 \;=\; \Big(\rho^2 \cdot \frac{N-B}{B(N-1)} \cdot \frac{1}{N}\sum_{i=1}^{N} \|g_i\|_2^2\Big) \;\cdot\; \big(1 - c_g(\boldsymbol{w})\big). \tag{3}$$

In particular, when $B \ll N$ and $c_g(\boldsymbol{w}) \ll 1$, this gap is dominated by

$$\mathsf{S}_{\text{rms}}^2 - \mathsf{S}^2 \;\approx\; \rho^2 \cdot \frac{1}{B} \cdot \frac{1}{N}\sum_{i=1}^{N} \|g_i\|_2^2. \tag{4}$$

Moreover, if $\|g(\boldsymbol{w})\|_2 \le \varepsilon$ and $\varepsilon^2 \ll \frac{1}{B} \cdot \frac{1}{N}\sum_{i=1}^{N} \|g_i\|_2^2$, then the full-batch sharpness is negligible relative to the RMS mini-batch sharpness, i.e., $\mathsf{S} \ll \mathsf{S}_{\text{rms}}$.

The proof appears in Appendix B.3.

Implications for directional sharpness. Intuitively, Theorem 2 shows that full-batch sharpness only sees the *average* gradient, while mini-batch sharpness is influenced by the *typical per-example* gradient magnitude. Thus, if per-example gradients disagree, they can cancel in the full-batch average, making $\|g(\boldsymbol{w})\|_2$ small even when the per-example gradients themselves have large norm. In other words, full-batch sharpness can average away high per-example gradient signals, making low-coherence deviations difficult to detect.

In contrast, these deviations can be detected by mini-batch sharpness. Eq. 4 shows that under low-coherence deviations, the squared RMS gap is controlled by the average per-example gradient scale $\frac{1}{N}\sum_i \|g_i\|_2^2$ and grows as $\approx 1/B$. Equivalently, the RMS mini-batch sharpness scale grows as $\approx 1/\sqrt{B}$. Thus, even when S is small, the mini-batch second-moment signal $\mathbb{E}_\xi[\mathsf{S}_\xi^2]$ can remain large, and using a smaller batch size makes this gap easier to detect.

Finally, directional sharpness can amplify this one-step RMS gap by aggregating mini-batch sharpness over T steps. Results from [12] suggest that low-coherence solutions are less stable under SAM-like dynamics; since directional sharpness includes a SAM-style probing trajectory, low-coherence models are more likely to exhibit larger fluctuation or escape behavior than coherent ones, making them more detectable.

Sharpness Metric	Correlation with Generalization Gap		
	Spearman ρ	Kendall τ	Kendall Ψ
Ours ($B=16, T=5$)	**0.902**	**0.702**	**0.694**
Ours ($B=8, T=5$)	**0.811**	**0.712**	**0.676**
ASAM ($B=256$)	0.581	0.264	0.194
ASAM ($B=\lvert D\rvert$)	0.639	0.433	0.411
Worst-Case ($B=\lvert D\rvert$)	0.601	0.577	0.552

Table 1: Correlation between different sharpness measures and generalization gap on a grid of 1,152 models with different architectures, optimizers, and hyperparameters. B denotes the batch size used for the sharpness computation. Our directional sharpness measure has a significantly stronger correlation than the static magnitude-aware worst-case sharpness (**Worst-Case**) and ASAM sharpness (**ASAM**) baselines.

4.5 Empirical Analysis of Directional Sharpness

Our theoretical analysis provides evidence that by measuring *dynamic stability* rather than a *static snapshot*, directional sharpness could serve as a more robust and reliable predictor of generalization. We now validate directional sharpness against Desideratum 1 by evaluating its empirical correlation with the generalization gap.

Setup. We mimic the experimental setup of the large-scale experiment by Jiang et al. [43] and train a large grid of models on CIFAR-10 and CIFAR-100 [49] with different optimizers, model architectures, and hyperparameters. For CIFAR-10, we use VGG with different sizes: VGG-13-BN, VGG-16-BN, and VGG-19-BN [74]; for CIFAR-100, we use WideResNet28-10 [96]. We train these models using the optimizers SGD, Adam, SAM [25], ASAM [50]. For Adam, we choose learning rates $\{0.001, 0.0005, 0.00032, 0.0001\}$, while for the other optimizers we use $\{0.1, 0.032, 0.01, 0.05\}$. We sweep weight decay in $\{0.0001, 0.00005\}$, batch size in $\{32, 64, 128\}$, and dropout rate in $\{0.0, 0.25, 0.5\}$. All models are trained to a 0.01 cross-entropy loss (estimated on 100 mini-batches), and we discard any that fail to converge within 200 epochs. For each converged model, we estimate the generalization gap using the difference between the test and training losses $\text{Gap}(\boldsymbol{w}) \approx \mathcal{L}_{D_{test}}(\boldsymbol{w}) - \mathcal{L}_{D_{train}}(\boldsymbol{w})$ and compute its correlation with directional sharpness, ASAM sharpness (**ASAM**), and the best-performing magnitude-aware worst-case sharpness (**Worst-Case**) reported by Jiang et al. [43]. Following [43], we report Spearman's ρ, Kendall's τ, and the granulated Kendall's coefficient Ψ.

Results and discussion. Table 1 shows all three correlation coefficients for our directional sharpness measure (**Ours**) compared against several existing sharpness measures. Our directional sharpness measure shows a significantly stronger correlation with the generalization gap than all other sharp-

ness measures. These findings provide empirical support for our theoretical analysis and validate our core insight: reframing sharpness from a static geometric property to a measure of dynamic stability yields a metric that is more predictive of generalization than existing approaches.

This section confirms that directional sharpness meets Desideratum 1, showing a stronger correlation with generalization than any static baseline. We next proceed to validate it against Desideratum 2 (robustness under training deviations).

5 Directional Sharpness for Model Certification

The previous section demonstrated the high correlation between directional sharpness and generalization, but a certification metric must also satisfy our two remaining desiderata: it must be robust against training deviations (Desideratum 2) and efficiently computable (Desideratum 3). In this section, we explore these two properties. First, in Section 5.1, we address Desideratum 2 by showing that directional sharpness consistently detects model failures that result from training deviations, including ones that cannot be detected by test accuracy or existing sharpness measures.

Next, in Sections 5.2 and 5.3, we explore efficiency in the context of two related applications: *model auditing* and *model certification*. We define model auditing using a single algorithm, Audit, that takes as input a model and its training data and outputs a binary decision indicating whether or not the model is satisfactory. (Alternatively, we could imagine Audit outputting a confidence score or other numeric value. In Section 5.3, we discuss a methodology for identifying thresholds for directional sharpness that verifiers could then use in predicates for producing a binary decision.) In our experiments, we show that this algorithm can be run up to 4× faster with directional sharpness than with test-accuracy evaluation, while also obtaining better correlation with generalization and robustness.

The Audit algorithm can be run by a model provider wanting to audit its own model before releasing it, but it is not appropriate for an external verifier to whom the model provider might not want to reveal the training data. Instead, we capture this setting of model *certification* as a pair of algorithms, Prove and Verify, where Prove takes as input the model and training data and outputs a proof, and Verify takes as input the model and the proof and again outputs either a binary decision (in keeping with the traditional model for zero knowledge and other cryptographic primitives) or a numeric score. Here, we demonstrate that proving directional sharpness in zero knowledge is up to 80,000× more efficient than proving the entire training process.

Testbed. All of our experiments are implemented in Python using PyTorch and run on a Google Colab A100 runtime with 83.5 GB of RAM and a single NVIDIA A100 GPU. For SAM training, we adopt the code provided by [99]. For backdoor attacks, we use the toolbox from Li et al. [55].

Implementation details of sharpness measures. Unless specified otherwise, we set $B = 16$, $T = 10$, $lr = 0.0001$, $\rho = 0.05$, and use ASAM sharpness as the per-step sharpness function in directional sharpness. The static sharpness baselines we consider are also magnitude-aware. Specifically, we consider magnitude-aware worst-case sharpness, which we compute using the gradient-ascent algorithm from Jiang et al. [43, Algorithm 3]; we also consider ASAM sharpness (Def. 4). We sometimes refer to these sharpness measures as "static" to contrast them with the dynamic nature of our directional sharpness measure.

5.1 Directional Sharpness is More Reliable under Training Deviations

To evaluate reliability under training deviations, we run a set of *distinguishability* experiments that mimic common auditing and certification settings. Each experiment involves a benign model M_b paired with a faulty model M_f, matched to (nearly) the same *observed accuracy*—the accuracy signal the verifier would naturally inspect in that scenario (e.g., standard test accuracy, pre-quantization accuracy, or accuracy on a small test set). We then ask whether sharpness metrics can flag the faulty model when observed accuracy fails. For each pair, we evaluate observed accuracy together with two metrics computed on the *clean* training set: full-dataset ASAM sharpness (Def. 4) and directional sharpness. We summarize each metric's distinguishability by its benign/faulty ratio (its value on M_b divided by its value on M_f); a smaller ratio indicates better separation. The failure settings we consider are:

Noisy labels. This setting mimics an auditor who observes similar holdout accuracy, yet one model is trained on mislabeled data. We train a VGG-16-BN M_b on clean CIFAR-10 and a faulty M_f on CIFAR-10 with 10% random label corruption (each corrupted label replaced by a uniformly sampled incorrect class), stopping both runs at 90% training accuracy. Observed accuracy is measured on the standard test set.

Overfitting to spurious features. This setting mimics a scenario where two models appear equally accurate on the standard test set, but one has learned a non-robust shortcut that may fail under distribution shift. We train a VGG-16-BN M_b on clean CIFAR-10 and a faulty M_f on a modified training set where, with probability 0.5, we overlay a top-left square whose color is deterministically tied to the label. This spurious feature [31, 42, 98] is highly predictive during training but non-causal. Both runs are stopped at 90% training accuracy, and observed accuracy is standard test accuracy.

Backdoors. A backdoor attack [34, 54, 75, 82] implants a hidden trigger that causes the model to behave normally on benign inputs but switch to a malicious target behavior when the trigger is present. We train a VGG-16-BN M_b for 200

epochs on the clean CIFAR-10 dataset and a faulty M_f on CIFAR-10 with 1% poisoned images, following BadNets [34]. Observed accuracy is clean test accuracy.

Post-quantization model quality. This setting mimics a scenario where two models have similar pre-quantization accuracy, but one is more robust to quantization-induced utility degradation [57]. To construct two models with similar pre-quantization accuracy but different post-quantization performance, we train two VGG-16-BN models, M_b with SAM and M_f with standard SGD, and then perform 4-bit quantization on both models; M_b sees a minimal accuracy drop ($0.920 - 0.914 = 0.006$), while M_f suffers a much larger degradation ($0.918 - 0.893 = 0.025$). We report the pre-quantization accuracy as observed accuracy.

Small test sets. In many real-world scenarios, test sets can be unreliable, unrepresentative, or maliciously crafted. We simulate a scenario where the available test set is too small to reliably reflect true performance. We train a VGG-16-BN M_b on clean CIFAR-10 and a faulty M_f with 10% label corruption, both for 200 epochs, and craft a 100-sample test set on which both models achieve 90% accuracy. On the full CIFAR-10 test set, M_f performs substantially worse (75% vs. 88%). Observed accuracy is the small test set accuracy.

The results for all five categories are in Table 2, and clearly demonstrate the limitations of static metrics. In all five scenarios, observed accuracy is (nearly) identical for the benign and faulty models, and the static sharpness ratios are much closer to 1 than the directional-sharpness ratios, indicating weaker separation.

Experiments on subtler and stronger deviations. To further test whether directional sharpness remains informative beyond the settings in Table 2, we run additional distinguishability experiments in Appendix A covering more diverse and challenging scenarios: (i) finer-grained label-noise rates; (ii) an adaptive, sharpness-aware backdoor attack [39]; and (iii) memorization/backdoor failures on language models. Across these settings, directional sharpness remains the strongest signal for benign/faulty separation.

5.2 Model Quality Auditing

We now turn to the problem of model auditing, in which a verifier wants to measure the quality of a model given access to both the model and its training data. We show that an auditing algorithm based on standard test accuracy fails to detect certain types of flaws that directional sharpness detects, and that prior sharpness metrics are too inefficient for this use case.

Efficiency comparison with existing sharpness metrics. We measure the runtime of computing directional sharpness and compare it against test accuracy and magnitude-aware worst-case sharpness (Def. 2) computed using the gradient-ascent method described by Jiang et al. [43]. Table 3 shows the runtime and correlation for different metrics. When computed on a small batch ($B = 8$, $T = 5$), directional sharpness is over $1{,}000\times$ faster than static sharpness and even $4\times$ faster than computing accuracy over the full test set (0.21 vs. 0.92 seconds). Consistent with Section 4.5, directional sharpness also provides the strongest correlation with generalization.

Experiment	Sharpness Ratio		Observed Acc. Ratio
	Directional	Static	
Noisy labels	**0.177**	0.991	1.00
Overfitting	**0.084**	0.963	1.00
Backdoors	**0.061**	0.587	1.00
Quantization	**0.240**	0.812	1.00
Small test set	**0.289**	0.560	1.00

Table 2: Distinguishability of faulty models under observed accuracy, static ASAM sharpness, and directional sharpness. Each entry reports the benign/faulty ratio. Lower ratios indicate stronger separation. Observed accuracy denotes the natural accuracy signal available to the verifier and is experiment-specific.

Metric	Runtime (in s)	Correlation
Test Accuracy	0.92	—
Worst-Case Sharpness [43]	210.32	$\tau = 0.577$
Ours ($B = 16, T = 10$)	**0.45**	$\tau = 0.708$
Ours ($B = 16, T = 5$)	**0.29**	$\tau = 0.702$
Ours ($B = 8, T = 10$)	**0.39**	$\tau = 0.691$
Ours ($B = 8, T = 5$)	**0.21**	$\tau = 0.712$

Table 3: Runtime of evaluating different metrics and their correlation with the generalization gap. Directional sharpness is not only the most efficient metric, faster than even plain test accuracy, but also offers the strongest Kendall's τ correlation with generalization.

Auditing model privacy via membership inference. Membership inference attacks (MIAs) [73] are a standard auditing tool for assessing privacy leakage: if an attacker can reliably distinguish training points ("members") from unseen points ("non-members"), then the model reveals information about its training data. Prior work links MIA vulnerability to poor generalization and overfitting [7, 94], and shows that training deviations can further amplify membership leakage [41, 86]. Motivated by our finding that directional sharpness provides a more reliable signal of poor generalization under deviations, we ask whether it can also serve as a stronger membership score—i.e., yield better separability between members and non-members—and thus improve privacy auditing.

Most black-box MIAs use per-sample loss or confidence as the decision signal, based on the observation that memorized training examples tend to incur lower loss. We evaluate three per-example scores—loss, static ASAM sharpness, and

MIA Metric	AUC	ACC
Loss	0.490	0.530
ASAM Sharpness	0.532	0.539
Directional Sharpness (Ours)	**0.556**	**0.570**

Table 4: MIA performance when using different per-sample scores as the attack signal. Higher AUC indicates better separability.

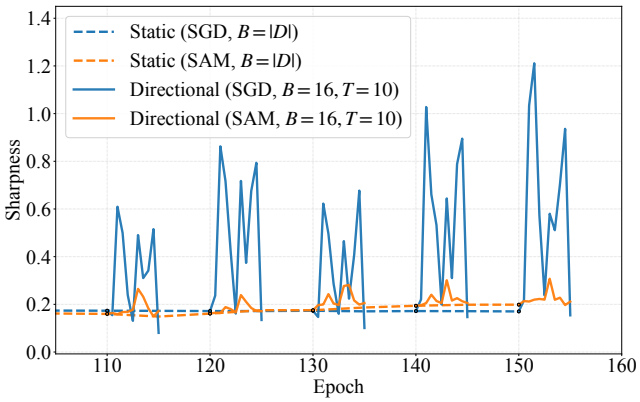


Figure 1: Static ASAM sharpness vs. directional sharpness probe trajectories on models trained with SGD and SAM. Static sharpness incorrectly suggests the SGD model generalizes better at epochs 130, 140, and 150, while directional sharpness trajectories reveal the SAM model is more stable under perturbations.

directional sharpness—on a VGG-16-BN model trained on CIFAR-10, using a balanced set of 1,000 members (training points) and 1,000 non-members (test points). Per-example directional sharpness is computed by treating the example as the only training data and setting $B = 1$. We compare attack performance using threshold-independent AUC and best-threshold accuracy (ACC). As shown in Table 4, directional sharpness achieves the highest AUC and ACC, indicating that our dynamic stability test strengthens loss-based MIA privacy auditing by improving member/non-member separability.

5.3 Model Quality Certification

We finally turn our attention to the problem of model certification, in which a prover wants to convince a verifier of the quality of a model to which the verifier has access (but, crucially, the verifier does not have access to the training data). Given that we have demonstrated the strong correlation between our directional sharpness metric and model generalization (and thus quality), we show how to achieve this by generating a zero-knowledge proof (ZKP) of directional sharpness.

Effectiveness of directional sharpness. We first demonstrate the effectiveness of directional sharpness compared to existing static sharpness metrics in model certification. In particular, we build on our previous results by considering the specific

Models	Sharpness		Generalization Gap
	Ours	Static ASAM	
SAM	**1.123**	**0.195**	**0.205**
SGD	5.711	0.170	0.273

Table 5: Static ASAM sharpness vs. directional sharpness values on two models trained with SAM and SGD. Static sharpness incorrectly suggests the SGD model has better generalization, while directional sharpness correctly identifies the SAM model as the better-generalizing one.

use case in which a malicious model trainer wants to convince a verifier that a model was trained using SAM (which is expensive) but, in reality, trained the model using standard SGD (which is much cheaper). To simulate this scenario, we train two VGG-16-BN models on CIFAR-10 for 150 epochs. One model was trained with standard SGD, representing the malicious trainer's model, while the other was honestly trained with SAM. After training, the SAM model has a lower generalization gap (0.205) than the SGD model (0.273) and thus better generalization.

Figure 1 plots the trajectory of both sharpness metrics over the last 50 epochs of training. The static ASAM sharpness metric (Def. 4, dashed lines) proves to be an unreliable signal, showing both models converging to similar values and incorrectly suggesting the SGD model has better generalization (Table 5). In contrast, directional sharpness (solid lines) is effective at distinguishing SAM and SGD models at all epochs. The SAM model (orange lines) consistently displays low and stable sharpness, indicating it resides in a uniformly flat basin. The SGD model (blue lines) exhibits high and erratic sharpness spikes. This shows that while its full-dataset sharpness may appear low, it is stuck in a sharp basin that is immediately revealed by directional sharpness.

Efficiency of directional sharpness. Now that we have established directional sharpness as more effective for model certification, we begin evaluating the end-to-end performance of proving directional sharpness in zero knowledge to demonstrate its efficiency. We compare this cost against a *zero-knowledge proof of training (zkPoT)* [1, 30].

To obtain realistic and concrete cost estimates, we extrapolate from the reported results of Kaizen [1]. Kaizen proves mini-batch gradient-descent training using an optimized GKR-style proof for each gradient update. Its instantiation uses 64-bit fixed-point encoding with 32 fractional bits over $\mathbb{F}_{p^2}$, where $p = 2^{61} - 1$. Mini-batches are sampled by applying a public pseudorandom permutation to the committed dataset indices and taking consecutive blocks of size B. Since both the zkPoT baseline and our ZKP of directional sharpness use the same Kaizen instantiation with parameters chosen for $\lambda = 100$-bit security, both inherit a knowledge-soundness error of at most 2^{-100} for the corresponding statements. We

Models	ZKP of Directional Sharpness ($B=8, T=5$)			ZKP of Training ($B=16$)	
	Time (min.)	Comm. (MB)	Speedup ($\times$)	Time (hr.)	Comm. (GB)
LeNet	23.63	9.85	**66,063**	26,016	471.64
AlexNet	55.43	11.59	**69,091**	63,830	579.72
VGG-11	88.87	15.28	**80,119**	118,671	751.59

Table 6: Cost estimates for ZKPs of directional sharpness vs. ZKPs of LATE-SAM training. For the zkPoT baseline, all models are trained for 145 epochs using SGD and switched to SAM for the final 5 epochs. $|D| = 50{,}000$.

refer to [1] for their full protocol and implementation details.

Let P_B^{SGD} be the cost of proving one SGD iteration on batch size B with prover runtime and proof size taken from [1]. Since SAM requires twice the computation per step versus SGD [25], we set $P_B^{\text{SAM}} \approx 2 \cdot P_B^{\text{SGD}}$. To avoid over-penalizing our baseline, we adopt the finding that 1–5 epochs of SAM at the late stage achieve comparable generalization to full SAM training [99] and consider a "LATE-SAM" baseline that proves 145 SGD epochs followed by 5 SAM epochs with batch size $B = 16$. The total cost for the LATE-SAM baseline is therefore $\frac{155|D|}{B} \cdot P_B^{\text{SGD}}$. Our ZKP of directional sharpness proves T SAM steps plus a small circuit for the fluctuation function. Since the cost of the fluctuation function is small compared with P_B^{SAM}, we estimate our total cost as $(2T) \cdot P_B^{\text{SGD}}$.

Table 6 reports these cost estimates. A ZKP of directional sharpness (running for $T = 5$ steps with a $B = 8$ batch) is up to 80,000$\times$ faster than proving the full 150-epoch LATE-SAM training run: a ZKP of LATE-SAM training for VGG-11 would take an estimated 118,671 hours (over 13 years), while our ZKP of directional sharpness takes under 90 minutes.

Certification decision based on directional sharpness. Our experiment shows that directional sharpness provides a robust signal for distinguishing high-quality models from low-quality ones. In practice, however, a verifier must still choose a threshold or decision rule to convert this signal into a binary "accept/reject" decision.

We note that this threshold selection is a limitation of any metric-based certification approach, since the true generalization gap is unobservable. Thus, the goal is not to find a universal constant, but rather a metric reliable enough that practical calibration procedures can be effective. Our theoretical and empirical results show that directional sharpness is well suited for this purpose: theoretically, it provides a principled measure of instability that is robust under training deviations (Section 4.3 and Section 4.4); empirically, it exhibits stronger correlation with generalization (Section 4.5) and wider separation margins between benign and faulty models (Table 2) compared to existing metrics.

These properties enable a straightforward calibration procedure: (i) construct a small calibration set of models of known quality for the task family (e.g., a verifier may already have baseline models trained with SGD or SAM); (ii) select a threshold τ that separates these models based on an acceptable risk level; (iii) apply τ to evaluate new models within the same task family. In certification settings where a binary decision is not required, verifiers can also treat the directional sharpness value directly as a continuous confidence score, with lower values indicating higher model quality.

6 Conclusion

Model quality certification is essential for providing transparency and accountability in machine learning systems. In this work, we use generalization as a proxy for overall model quality and develop a new, efficient methodology for its certification. Our central contribution is a metric we call directional sharpness. This measure strongly correlates with generalization, is reliable under existing training deviations, and is efficiently computable. These properties make directional sharpness a suitable ingredient in ML model certification. Therefore, directional sharpness can play an important role in mitigating attacks in adversarial machine learning. Using directional sharpness to devise rigorous defenses against adversarial models, and incorporating these defenses into existing systems, is a natural continuation of this research.

Ethical Considerations

Stakeholders. The main stakeholders are *model providers*, who audit candidate models and provide evidence of quality; *external verifiers*, who want such evidence without access to sensitive training data; *end users*, who may be harmed by poorly generalizing or backdoored models; the *research community*, which may build on our metric; and *broader society*, which is affected by the reliability and accountability of deployed ML systems.

Impacts. Our work benefits model providers by providing a practical tool for detecting low-quality models, and external verifiers by enabling privacy-preserving model certification. It also benefits end users and broader society by helping identify training deviations, such as backdoors, missed by standard evaluation. For the research community, our work provides a concrete metric that can be studied and improved.

Potential negative implications. One risk is that malicious model providers may adapt stronger adaptive attacks to evade our certification once it is public. Another risk is over-reliance: providers and end users may misinterpret our method as certification of the true generalization gap or unconditional model safety. More broadly, the research community and broader society could also misinterpret the scope of our certification guarantee, which could negatively impact deployment decisions or future research.

Mitigation. We mitigate these risks by explicitly scoping our claims, clarifying that our method does not certify the true generalization gap or the absence of all possible failure modes. We also evaluate our method against an adaptive backdoor attack. While these experiments are not exhaustive, we believe the risk of evasion is limited by the fact that our metric targets generalization, a property that many harmful behaviors inherently conflict with, and our results suggest it is harder to evade than existing static metrics.

Decision to publish. We believe publication is justified because the benefits outweigh the potential risks. Our work advances model auditing and privacy-preserving model certification, and provides the community with a concrete method that can be scrutinized and improved. More broadly, publishing practical tools for model quality certification is in the public interest, especially when the alternative is continued reliance on weaker signals.

Open Science

A well-documented implementation is available under the Apache 2.0 license at https://zenodo.org/records/20385823. Directory `df_utils` contains the code for measuring directional sharpness (Algorithm 1) and directory `experiments` contains implementations of the experiments, including (i) correlation experiments (Section 4.5), (ii) distinguishability experiments (Section 5.1) with label-noise, backdoor, SAPA, small test set, overfitting, and quantization, and (iii) MIA experiments (Section 5.2).

Acknowledgments

Gefei Tan is supported by a Google PhD Fellowship.

References

[1] Kasra Abbaszadeh, Christodoulos Pappas, Jonathan Katz, and Dimitrios Papadopoulos. Zero-knowledge proofs of training for deep neural networks. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 4316–4330, 2024.

[2] Pierre Alquier. User-friendly introduction to pac-bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303, 2024. doi:10.1561/2200000100.

[3] Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-aware minimization leads to low-rank features. *Advances in Neural Information Processing Systems*, 36:47032–47051, 2023.

[4] Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *International conference on machine learning*, pages 639–668. PMLR, 2022.

[5] Meenatchi Sundaram Muthu Selva Annamalai, Borja Balle, Jamie Hayes, Georgios Kaissis, and Emiliano De Cristofaro. The hitchhiker's guide to efficient, end-to-end, and tight dp auditing. *arXiv preprint arXiv:2506.16666*, 2025.

[6] Christina Baek, J Zico Kolter, and Aditi Raghunathan. Why is sam robust to label noise? In *The Twelfth International Conference on Learning Representations*, 2024.

[7] Teodora Baluta, Shiqi Shen, S Hitarth, Shruti Tople, and Prateek Saxena. Membership inference attacks and generalization: A causal perspective. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 249–262, 2022.

[8] Björn Barz and Joachim Denzler. Do we train on test data? purging cifar of near-duplicates. *Journal of Imaging*, 6(6):41, 2020.

[9] Carsten Baum, Lennart Braun, Alexander Munch-Hansen, Benoît Razet, and Peter Scholl. Appenzeller to brie: Efficient zero-knowledge proofs for mixed-mode arithmetic and Z2k. In Giovanni Vigna and Elaine Shi, editors, *ACM CCS 2021*, pages 192–211. ACM Press, November 2021.

[10] Carsten Baum, Lennart Braun, Alexander Munch-Hansen, and Peter Scholl. Moz$\mathbb{Z}_{2^k}$arella: Efficient vector-OLE and zero-knowledge proofs over $\mathbb{Z}_{2^k}$. In Yevgeniy Dodis and Thomas Shrimpton, editors, *CRYPTO 2022, Part IV*, volume 13510 of *LNCS*, pages 329–358. Springer, Cham, August 2022.

[11] Carsten Baum, Alex J Malozemoff, Marc B Rosen, and Peter Scholl. Mac'n'cheese: zero-knowledge proofs for boolean and arithmetic circuits with nested disjunctions. In *Annual International Cryptology Conference*, pages 92–122. Springer, 2021.

[12] Wei-Kai Chang and Rajiv Khanna. A unified stability analysis of SAM vs SGD: Role of data coherence and emergence of simplicity bias. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[13] Satrajit Chatterjee. Coherent gradients: An approach to understanding generalization in gradient descent-based optimization. In *International Conference on Learning Representations*, 2020.

[14] Zixiang Chen, Junkai Zhang, Yiwen Kou, Xiangning Chen, Cho-Jui Hsieh, and Quanquan Gu. Why does sharpness-aware minimization generalize better than SGD? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[15] Alessandro Chiesa, Yuncong Hu, Mary Maller, Pratyush Mishra, Psi Vesely, and Nicholas P. Ward. Marlin: Preprocessing zkSNARKs with universal and updatable SRS. In Anne Canteaut and Yuval Ishai, editors, *EUROCRYPT 2020, Part I*, volume 12105 of *LNCS*, pages 738–768. Springer, Cham, May 2020.

[16] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

[17] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1019–1028. JMLR.org, 2017.

[18] Samuel Dittmer, Yuval Ishai, Steve Lu, and Rafail Ostrovsky. Improving line-point zero knowledge: Two multiplications for the price of one. In Heng Yin, Angelos Stavrou, Cas Cremers, and Elaine Shi, editors, *ACM CCS 2022*, pages 829–841. ACM Press, November 2022.

[19] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126, 2015.

[20] Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M Roy. In search of robust measures of generalization. *Advances in Neural Information Processing Systems*, 33:11723–11733, 2020.

[21] Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2017.

[22] Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. Towards llm unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. In *International Conference on Machine Learning*, pages 15762–15778. PMLR, 2025.

[23] Yu Feng and Yuhai Tu. The inverse variance–flatness relation in stochastic gradient descent is critical for finding flat minima. *Proceedings of the National Academy of Sciences*, 118(9):e2015617118, 2021.

[24] Andrew Foong, Wessel Bruinsma, David Burt, and Richard Turner. How tight can pac-bayes be in the small data regime? *Advances in Neural Information Processing Systems*, 34:4093–4105, 2021.

[25] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

[26] Olive Franzese, Ali Shahin Shamsabadi, Carter Luck, and Hamed Haddadi. Secure and confidential certificates of online fairness. *Advances in Neural Information Processing Systems*, 38:40077–40107, 2025.

[27] Ariel Gabizon, Zachary J. Williamson, and Oana Ciobotaru. PLONK: Permutations over Lagrange-bases for oecumenical noninteractive arguments of knowledge. Cryptology ePrint Archive, Report 2019/953, 2019. URL: https://eprint.iacr.org/2019/953.

[28] Yansong Gao, Bao Gia Doan, Zhi Zhang, Siqi Ma, Jiliang Zhang, Anmin Fu, Surya Nepal, and Hyoungshick Kim. Backdoor attacks and countermeasures on deep learning: A comprehensive review. *arXiv preprint arXiv:2007.10760*, 2020.

[29] Isha Garg, Deepak Ravikumar, and Kaushik Roy. Memorization through the lens of curvature of loss function around samples. In *Proceedings of the 41st International Conference on Machine Learning*, pages 15083–15101, 2024.

[30] Sanjam Garg, Aarushi Goel, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Guru-Vamsi Policharla, and Mingyuan Wang. Experimenting with zero-knowledge proofs of training. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, pages 1880–1894, 2023.

[31] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[32] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[33] Jens Groth. On the size of pairing-based non-interactive arguments. In Marc Fischlin and Jean-Sébastien Coron, editors, *EUROCRYPT 2016, Part II*, volume 9666 of *LNCS*, pages 305–326. Springer, Berlin, Heidelberg, May 2016.

[34] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *Ieee Access*, 7:47230–47244, 2019.

[35] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT*, pages 107–112, 2018.

[36] Maxime Haddouche, Paul Viallard, Umut Şimşekli, and Benjamin Guedj. A pac-bayesian link between generalisation and flat minima. In *ALT 2025-36th International Conference on Algorithmic Learning Theory*, pages 1–31, 2025.

[37] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc.

[38] Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: Beyond sharp and flat local minima. *Advances in neural information processing systems*, 32, 2019.

[39] Pengfei He, Han Xu, Jie Ren, Yingqian Cui, Shenglai Zeng, Hui Liu, Charu C. Aggarwal, and Jiliang Tang. Sharpness-aware data poisoning attack. In *The Twelfth International Conference on Learning Representations*, 2024.

[40] Fredrik Hellström, Giuseppe Durisi, Benjamin Guedj, and Maxim Raginsky. Generalization bounds: Perspectives from information theory and pac-bayes. *Foundations and Trends in Machine Learning*, 18(1):1–223, 2025.

[41] Hongsheng Hu, Zoran Salcic, Gillian Dobbie, Jinjun Chen, Lichao Sun, and Xuyun Zhang. Membership inference via backdooring. In *International Joint Conference on Artificial Intelligence (31st: 2022)*, pages 3832–3838. International Joint Conferences on Artificial Intelligence Organization, 2022.

[42] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[43] Yiding Jiang, Behnam Neyshabur*, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020.

[44] Matt Jordan, Justin Lewis, and Alexandros G Dimakis. Provable certificates for adversarial examples: Fitting a ball in the union of polytopes. *Advances in neural information processing systems*, 32, 2019.

[45] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

[46] Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023.

[47] Niki Kilbertus, Adrià Gascón, Matt Kusner, Michael Veale, Krishna Gummadi, and Adrian Weller. Blind justice: Fairness with encrypted sensitive attributes. In *International Conference on Machine Learning*, pages 2630–2639. PMLR, 2018.

[48] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.

[49] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[50] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International conference on machine learning*, pages 5905–5914. PMLR, 2021.

[51] SCIPR Lab. libsnark: a C++ library for zk-SNARK proofs, 2017. URL: https://github.com/scipr-lab/libsnark.

[52] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE symposium on security and privacy (SP)*, pages 656–672. IEEE, 2019.

[53] Tao Li, Pan Zhou, Zhengbao He, Xinwen Cheng, and Xiaolin Huang. Friendly sharpness-aware minimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5631–5640, 2024.

[54] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):5–22, 2024.

[55] Yiming Li, Mengxi Ya, Yang Bai, Yong Jiang, and Shu-Tao Xia. BackdoorBox: A python toolbox for backdoor learning. In *ICLR Workshop*, 2023.

[56] Jiaxiang Liu, Simon Oya, and Florian Kerschbaum. Generalization techniques empirically outperform differential privacy against membership inference. *arXiv preprint arXiv:2110.05524*, 2021.

[57] Jing Liu, Jianfei Cai, and Bohan Zhuang. Sharpness-aware quantization for deep neural networks. *arXiv preprint arXiv:2111.12273*, 2021.

[58] David A. McAllester. Some pac-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 230–234, New York, NY, USA, 1998. Association for Computing Machinery.

[59] David A McAllester. Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170, 1999.

[60] Takashi Mori, Liu Ziyin, Kangqiao Liu, and Masahito Ueda. Power-law escape rate of sgd. In *International Conference on Machine Learning*, pages 15959–15975. PMLR, 2022.

[61] Vaishnavh Nagarajan and Zico Kolter. Deterministic PAC-bayesian generalization bounds for deep networks via generalizing noise-resilience. In *International Conference on Learning Representations*, 2019.

[62] Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. *Advances in Neural Information Processing Systems*, 32, 2019.

[63] Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*, pages 3526–3545. PMLR, 2021.

[64] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[65] Saerom Park, Seongmin Kim, and Yeon-sup Lim. Fairness audit of machine learning models with confidential computing. In *Proceedings of the ACM Web Conference 2022*, pages 3488–3499, 2022.

[66] María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22(227):1–40, 2021.

[67] Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[68] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[69] Mansi Sakarvadia, Aswathy Ajith, Arham Mushtaq Khan, Nathaniel C Hudson, Caleb Geniesse, Kyle Chard, Yaoqing Yang, Ian Foster, and Michael W. Mahoney. Mitigating memorization in language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[70] Karthik Abinav Sankararaman, Soham De, Zheng Xu, W Ronny Huang, and Tom Goldstein. The impact of neural network overparameterization on gradient confusion and stochastic gradient descent. In *International conference on machine learning*, pages 8469–8479. PMLR, 2020.

[71] Ali Shahin Shamsabadi, Gefei Tan, Tudor Ioan Cebere, Aurélien Bellet, Hamed Haddadi, Nicolas Papernot, Xiao Wang, and Adrian Weller. Confidential-dpproof: Confidential proof of differentially private training. In *International Conference on Learning Representations (ICLR)*, 2024.

[72] Ali Shahin Shamsabadi, Sierra Calanda Wyllie, Nicholas Franzese, Natalie Dullerud, Sébastien Gambs, Nicolas Papernot, Xiao Wang, and Adrian Weller. Confidential-profitt: confidential proof of fair training of trees. In *The Eleventh International Conference on Learning Representations*, 2023.

[73] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pages 3–18. IEEE Computer Society, 2017.

[74] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[75] Hossein Souri, Liam Fowl, Rama Chellappa, Micah Goldblum, and Tom Goldstein. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. *Advances in Neural Information Processing Systems*, 35:19165–19178, 2022.

[76] Jacob Mitchell Springer, Vaishnavh Nagarajan, and Aditi Raghunathan. Sharpness-aware minimization enhances feature quality via balanced learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[77] Haochen Sun, Tonghe Bai, Jason Li, and Hongyang Zhang. Zkdl: Efficient zero-knowledge proofs of deep learning training. *IEEE Transactions on Information Forensics and Security*, 2024.

[78] Iulian Emil Tampu, Anders Eklund, and Neda Haj-Hosseini. Inflation of test accuracy due to data leakage in deep learning-based classification of oct images. *Scientific Data*, 9(1):580, 2022.

[79] Gefei Tan, Adrià Gascón, Sarah Meiklejohn, Mariana Raykova, Xiao Wang, and Ning Luo. Founding zero-knowledge proof of training on optimum vicinity. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pages 1173–1187, 2025.

[80] Haoran Tang and Rajiv Khanna. Sharpness-aware machine unlearning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[81] Ehsan Toreini, Maryam Mehrnezhad, and Aad Van Moorsel. Verifiable fairness: privacy–preserving computation of fairness for machine learning systems. In *European symposium on research in computer security*, pages 569–584. Springer, 2023.

[82] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019.

[83] Guillermo Valle-Pérez and Ard A. Louis. Generalization bounds for deep learning, 2020. arXiv:2012.04115.

[84] Paul Viallard, Rémi Emonet, Amaury Habrard, Emilie Morvant, and Valentina Zantedeschi. Leveraging pac-bayes theory and gibbs distributions for generalization bounds with complexity measures. In *International conference on artificial intelligence and statistics*, pages 3007–3015. PMLR, 2024.

[85] Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. *Advances in Neural Information Processing Systems*, 36:1024–1035, 2023.

[86] Yuxin Wen, Leo Marchyok, Sanghyun Hong, Jonas Geiping, Tom Goldstein, and Nicholas Carlini. Privacy backdoors: Enhancing membership inference through poisoning pre-trained models. *Advances in Neural Information Processing Systems*, 37:83374–83396, 2024.

[87] Chenkai Weng, Kang Yang, Zhaomin Yang, Xiang Xie, and Xiao Wang. AntMan: Interactive zero-knowledge proofs with sublinear communication. In Heng Yin, Angelos Stavrou, Cas Cremers, and Elaine Shi, editors, *ACM CCS 2022*, pages 2901–2914. ACM Press, November 2022.

[88] Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type part ii: Continuous time analysis. *Journal of Nonlinear Science*, 34(1):16, 2024.

[89] Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. *Advances in Neural Information Processing Systems*, 31, 2018.

[90] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *NeurIPS*, 2020.

[91] Zhibo Xing, Zijian Zhang, Ziang Zhang, Zhen Li, Meng Li, Jiamou Liu, Zongyang Zhang, Yi Zhao, Qi Sun, Liehuang Zhu, et al. Zero-knowledge proof-based verifiable decentralized machine learning in communication network: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 2025.

[92] Chhavi Yadav, Amrita Roy-Chowdhury, Dan Boneh, and Kamalika Chaudhuri. Fairproof: Confidential and certifiable fairness for neural networks. In *Proceedings of the 41st International Conference on Machine Learning*. Journal of Machine Learning Research, 2024.

[93] Kang Yang, Pratik Sarkar, Chenkai Weng, and Xiao Wang. QuickSilver: Efficient and affordable zero-knowledge proofs for circuits and polynomials over any field. In Giovanni Vigna and Elaine Shi, editors, *ACM CCS 2021*, pages 2986–3001. ACM Press, November 2021.

[94] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE*

31st computer security foundations symposium (CSF), pages 268–282. IEEE, 2018.

[95] Danni Yuan, Mingda Zhang, Shaokui Wei, Li Liu, and Baoyuan Wu. Activation gradient based poisoned sample detection against backdoor attacks. In *The Thirteenth International Conference on Learning Representations*, 2025.

[96] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, 2016. doi:10.5244/C.30.87.

[97] Tianyu Zhang, Shen Dong, O. Deniz Kose, Yanning Shen, and Yupeng Zhang. Fairzk: A scalable system to prove machine learning fairness in zero-knowledge. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 3460–3478, 2025. doi:10.1109/SP61157.2025.00205.

[98] Lili Zhao, Qi Liu, Linan Yue, Wei Chen, Liyi Chen, Ruijun Sun, and Chao Song. Comi: Correct and mitigate shortcut learning behavior in deep neural networks. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 218–228, 2024.

[99] Zhanpeng Zhou, Mingze Wang, Yuchen Mao, Bingrui Li, and Junchi Yan. Sharpness-aware minimization efficiently selects flatter minima late in training. In *International Conference on Learning Representations*, 2025.

[100] Mingli Zhu, Shaokui Wei, Li Shen, Yanbo Fan, and Baoyuan Wu. Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In *ICCV*, 2023.

Appendix

A Additional Experiments

We provide additional distinguishability experiments extending Section 5.1, evaluating whether directional sharpness remains informative under subtler, more challenging deviations and on non-vision models. We follow the same evaluation protocol as in Section 5.1: for each setting, we train benign/faulty model pairs and ask whether sharpness metrics can separate them when observed accuracy fails. Table 7 reports the benign/faulty ratios for the same metrics as in Section 5.1.

Backdoor on ResNet-18 (ResNet-BD). We repeat the BadNets backdoor experiment on ResNet-18. The faulty model is trained on CIFAR-10 with 1% poisoned images. Both benign and faulty models are trained for 160 epochs. The observed accuracy reported is the clean test accuracy.

Experiment	Sharpness Ratio		Observed Acc. Ratio
	Directional	Static	
SAPA	**0.016**	0.217	1.000
LM-Mem	**0.083**	0.664	1.003
LM-BD	**0.053**	0.630	1.005
Noise (1%)	**0.122**	0.676	0.995
Noise (2.5%)	**0.086**	0.833	0.994
Noise (5%)	**0.079**	0.546	0.998
Noise (7.5%)	**0.081**	0.418	0.998
ResNet-BD	**0.253**	0.972	1.000

Table 7: Results of additional distinguishability experiments. Each entry reports the benign/faulty ratio; lower values indicate better separation. Directional sharpness provides the strongest separation.

Varying label noise (Noise). We repeat the label-noise experiment with varying noise rates to evaluate the sensitivity of directional sharpness to weaker forms of data-quality degradation. The observed accuracy reported is the clean test accuracy.

Sharpness-aware poisoning attack (SAPA). SAPA [39] uses SAM-style perturbations during poison generation to optimize the adversarial objective, making the poison more robust to retraining uncertainty. We train a benign ResNet-18 on clean CIFAR-10 using standard SGD, and a faulty ResNet-18 using SAPA's hidden-trigger backdoor attack with 1% of the training set poisoned, perturbation budget $\varepsilon = 16/255$, and sharpness radius $\rho = 0.05$. Both models are trained for 160 epochs. The observed accuracy is the clean test accuracy.

Backdoor (LM-BD) and memorization (LM-Mem) on language models. We extend our distinguishability tests to transformer language models using TinyMem [69], a suite of lightweight GPT-2-style models. Using the 4-layer multiplicative architecture, we train a benign model M_b on clean data and a faulty model M_f on data poisoned with either the backdoor artifact [69, Def. 2.3] or the noise artifact [69, Def. 2.2], which inject a trigger pattern and a random perturbation, respectively, into a small fraction of the training sequences (see [69] for details). In both settings, M_f memorizes the injected artifacts while matching the observed (clean-test) accuracy of M_b: the backdoored M_f attains 99.3% backdoor memorization at 96.92% observed accuracy (vs. 97.41% for M_b), and the noise-trained M_f memorizes 37.4% of the artifacts at 97.11% observed accuracy (vs. 97.40% for M_b).

B Proofs of Lemmas and Theorems

We first establish two direct consequences of our assumptions. Under Assumption 1, we have

$$2\mu\mathcal{L}(\boldsymbol{w}) \leq \|g(\boldsymbol{w})\|_2^2 \leq 2L_s\mathcal{L}(\boldsymbol{w}), \tag{5}$$

where the upper bound is a standard result from smoothness [46, Lemma 1] with $\inf_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}) = 0$.

Additionally, under Assumption 2, the gradient noise for a single sample i satisfies $\mathbb{E}_i\|\zeta_i^1(\boldsymbol{w})\|_2^2 \le \sigma^2\mathcal{L}(\boldsymbol{w})$. For a mini-batch ξ of size B sampled uniformly without replacement, we have

$$\mathbb{E}_\xi\|\zeta^B(\boldsymbol{w},\xi)\|_2^2 = \frac{N-B}{B(N-1)}\mathbb{E}_i\|\zeta_i^1(\boldsymbol{w})\|_2^2.$$

Since $(N-B)/(N-1) \le 1$, we obtain

$$\mathbb{E}_\xi\|\zeta^B(\boldsymbol{w},\xi)\|_2^2 \le \frac{\sigma^2}{B}\mathcal{L}(\boldsymbol{w}). \tag{6}$$

B.1 Proof of Lemma 1

Proof. For $p=2$, the SAM sharpness on mini-batch ξ_t is

$$s_t = \rho\|\nabla\mathcal{L}_{\xi_t}(\boldsymbol{w}_t)\|_2 = \rho\|g(\boldsymbol{w}_t)+\zeta_t\|_2,$$

where $\zeta_t := g_{\xi_t}(\boldsymbol{w}_t) - g(\boldsymbol{w}_t)$. Conditioning on $\boldsymbol{w}_t$, and using the fact that the mini-batch gradient is unbiased, we have

$$\mathbb{E}[s_t^2 \mid \boldsymbol{w}_t] = \rho^2\left(\|g(\boldsymbol{w}_t)\|_2^2 + \mathbb{E}[\|\zeta_t\|_2^2 \mid \boldsymbol{w}_t]\right).$$

For the lower bound, drop the nonnegative noise term and apply the PL lower bound in Eq. 5:

$$\mathbb{E}[s_t^2 \mid \boldsymbol{w}_t] \ge \rho^2\|g(\boldsymbol{w}_t)\|_2^2 \ge 2\rho^2\mu\mathcal{L}(\boldsymbol{w}_t).$$

For the upper bound, apply the upper bound in Eq. 5 and the mini-batch noise bound Eq. 6:

$$\mathbb{E}[s_t^2 \mid \boldsymbol{w}_t] \le \rho^2\left(2L_s + \frac{\sigma^2}{B}\right)\mathcal{L}(\boldsymbol{w}_t).$$

This completes the proof. □

B.2 Proof of Theorem 1

Proof. Taking expectations in Lemma 1 and using the tower rule gives

$$\underline{\kappa}_\rho\,\mathbb{E}[\mathcal{L}(\boldsymbol{w}_t)] \le \mathbb{E}[s_t^2] \le \overline{\kappa}_\rho\,\mathbb{E}[\mathcal{L}(\boldsymbol{w}_t)]. \tag{7}$$

(a) Stability implies bounded sharpness. SAM-stability gives $\mathbb{E}[\mathcal{L}(\boldsymbol{w}_t)] \le C\mathcal{L}(\boldsymbol{w}_0)$ for all t. Substituting into the upper bound in Eq. 7 gives

$$\mathbb{E}[s_t^2] \le \overline{\kappa}_\rho C\mathcal{L}(\boldsymbol{w}_0).$$

(b) Exponential sharpness growth certifies instability. The upper side of the sandwich bound in Eq. 7 implies

$$\mathbb{E}[\mathcal{L}(\boldsymbol{w}_t)] \ge \frac{1}{\overline{\kappa}_\rho}\mathbb{E}[s_t^2].$$

Therefore, if $\mathbb{E}[s_t^2] \ge K\alpha^t\mathcal{L}(\boldsymbol{w}_0)$ for all $t \ge t_0$, then

$$\mathbb{E}[\mathcal{L}(\boldsymbol{w}_t)] \ge \frac{K}{\overline{\kappa}_\rho}\alpha^t\mathcal{L}(\boldsymbol{w}_0).$$

Since $\alpha > 1$ and $\mathcal{L}(\boldsymbol{w}_0) > 0$, this loss trajectory is unbounded relative to $\mathcal{L}(\boldsymbol{w}_0)$ and violates the condition in Def. 6. Hence $\boldsymbol{w}^*$ is SAM-unstable. □

B.3 Proof of Theorem 2

We first prove the following lemma for mini-batch gradients.

Lemma 2. *Let ξ be a mini-batch of size B sampled uniformly without replacement from a dataset of size N. Write $S_g^2 := \frac{1}{N}\sum_{i=1}^N\|g_i(\boldsymbol{w})\|_2^2$ and define $c_g(\boldsymbol{w}) := \|g(\boldsymbol{w})\|_2^2/S_g^2$ (Def. 7). Assuming $S_g^2 > 0$,*

$$\mathbb{E}_\xi\|g_\xi(\boldsymbol{w})\|_2^2 = \|g(\boldsymbol{w})\|_2^2 + \frac{N-B}{B(N-1)}(1-c_g(\boldsymbol{w}))S_g^2.$$

Proof. Expanding the squared norm and using $\Pr[i\in\xi] = B/N$ and $\Pr[i,j\in\xi] = B(B-1)/(N(N-1))$ for $i\ne j$,

$$\begin{aligned}\mathbb{E}_\xi\|g_\xi\|_2^2 &= \frac{1}{B^2}\left(\frac{B}{N}\sum_i\|g_i\|_2^2 + \frac{B(B-1)}{N(N-1)}\sum_{i\ne j}\langle g_i,g_j\rangle\right)\\ &= \frac{1}{B}S_g^2 + \frac{B-1}{B}\cdot\frac{1}{N(N-1)}\sum_{i\ne j}\langle g_i,g_j\rangle.\end{aligned}$$

Since

$$\sum_{i\ne j}\langle g_i,g_j\rangle = N^2\|g\|_2^2 - NS_g^2,$$

substitution gives

$$\mathbb{E}_\xi\|g_\xi\|_2^2 = \|g\|_2^2 + \frac{N-B}{B(N-1)}(S_g^2 - \|g\|_2^2).$$

Plugging in $S_g^2 - \|g\|_2^2 = (1-c_g(\boldsymbol{w}))S_g^2$ yields the lemma. □

Proof of Theorem 2. For $p=q=2$,

$$\mathsf{S} = \rho\|g(\boldsymbol{w})\|_2, \qquad \mathsf{S}_\xi = \rho\|g_\xi(\boldsymbol{w})\|_2, \qquad \mathsf{S}_{\mathrm{rms}}^2 = \mathbb{E}_\xi[\mathsf{S}_\xi^2].$$

Therefore,

$$\mathsf{S}_{\mathrm{rms}}^2 - \mathsf{S}^2 = \mathbb{E}_\xi[\mathsf{S}_\xi^2] - \mathsf{S}^2 = \rho^2\left(\mathbb{E}_\xi\|g_\xi(\boldsymbol{w})\|_2^2 - \|g(\boldsymbol{w})\|_2^2\right).$$

Applying Lemma 2 yields Eq. 3:

$$\mathsf{S}_{\mathrm{rms}}^2 - \mathsf{S}^2 = \rho^2\cdot\frac{N-B}{B(N-1)}\cdot\left(1-c_g(\boldsymbol{w})\right)\cdot S_g^2.$$

If $B \ll N$, then $\frac{N-B}{B(N-1)} \approx \frac{1}{B}$. If also $c_g(\boldsymbol{w}) \ll 1$, then $1 - c_g(\boldsymbol{w}) \approx 1$, and hence we have Eq. 4:

$$\mathsf{S}_{\mathrm{rms}}^2 - \mathsf{S}^2 \approx \rho^2\frac{1}{B}S_g^2 = \rho^2\frac{1}{B}\cdot\frac{1}{N}\sum_{i=1}^N\|g_i(\boldsymbol{w})\|_2^2.$$

Finally, if $\|g(\boldsymbol{w})\|_2 \le \varepsilon$, then $\mathsf{S}^2 \le \rho^2\varepsilon^2$. Therefore, whenever $\varepsilon^2 \ll S_g^2/B$, we have

$$\mathsf{S}^2 \ll \rho^2 S_g^2/B \approx \mathsf{S}_{\mathrm{rms}}^2 - \mathsf{S}^2.$$

Hence $\mathsf{S}^2 \ll \mathsf{S}_{\mathrm{rms}}^2$ and $\mathsf{S} \ll \mathsf{S}_{\mathrm{rms}}$. □