LEARNING DYNAMICS REVEAL ADVERSARIAL DISTANCE AND MEMBERSHIP RISK

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks memorize individual training samples to different degrees, with consequences for both adversarial robustness and privacy. Yet sample-level adversarial distance and membership vulnerability are expensive to measure at scale because they typically require iterative attacks, auxiliary models, or repeated retraining. We show that both quantities are encoded in the learning trajectory itself. We study Cumulative Sample Loss (CSL) and Cumulative Sample Gradient (CSG), which summarize how a sample's loss and input gradient evolve during training and can be collected with negligible overhead. Theoretically, we first establish a direct relation between CSL and CSG. We then derive two-sided bounds on adversarial distance from sample loss and input-gradient statistics and extend these results to trajectory-level quantities. Under stability, generalization, and model-bias assumptions, we yield a formal connection between memorization, adversarial vulnerability, and privacy leakage. These results motivate efficient sample-level vulnerability estimators from training dynamics. Our adversarial-distance proxy improves mislabeled-sample detection and achieves near-perfect duplicate detection on CIFAR-10 and CIFAR-100. For privacy auditing, we introduce CSL-Diff, which identifies samples vulnerable to LiRA, AttackR, and RMIA without auxiliary models at deployment and consistently outperforms prior artifact-based predictors. Together, our results turn learning dynamics into a reusable signal for auditing robustness, privacy, and dataset quality at the cost of standard training.

1 INTRODUCTION

Deep neural networks do not learn every training example in the same way. Rare, mislabeled, and conflicting samples can retain high loss or require substantially more training iterations to fit than typical examples, forcing the model to encode sample-specific information (Feldman, 2020; Feldman & Zhang, 2020). This heterogeneity matters beyond training accuracy. Memorized samples can expose whether they appeared in the training set (Yeom et al., 2018; Carlini et al., 2022). Adversarial distance measures the minimum perturbation required to move a sample across the learned decision boundary (Moosavi-Dezfooli et al., 2016; Carlini & Wagner, 2016). This motivates a sample-level view that connects how an example is learned, how locally vulnerable it is to adversarial perturbation, and how much membership information the model reveals about it.

Existing tools make these quantities expensive to obtain together. Stability-based memorization estimates the effect of including a sample by comparing models trained on neighboring datasets (Feldman & Zhang, 2020). Adversarial distance requires a per-sample optimization that searches for a decision-boundary crossing (Moosavi-Dezfooli et al., 2016; Carlini & Wagner, 2016). Membership inference attacks distinguish member behavior from non-member behavior, with stronger attacks often relying on reference or shadow models (Shokri et al., 2017; Carlini et al., 2022). These procedures answer complementary questions after training, but none use the original learning trajectory to connect all three quantities at the level of individual samples.

Learning dynamics provide a cheaper source of sample-specific information. Learning and forgetting events identify examples that remain difficult throughout optimization (Toneva et al., 2019; Maini et al., 2022). Cumulative Sample Loss (CSL) (Ravikumar et al., 2025) and Cumulative Sample Gradient (CSG) (Ravikumar et al., 2026) summarize per-sample loss and input-gradient dynamics during training, providing efficient proxies for sample memorization. They are available directly

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

from the training trajectory, but no formal connection currently establishes how these learning dynamics constrain a sample's local decision geometry and its vulnerability to membership inference.

We study this connection through untargeted adversarial distance, the infimum ℓ_2 perturbation required to cause misclassification. Under input-smoothness and positive-margin assumptions, we bound this distance using loss and input-gradient norm at a fixed checkpoint, and provide bounds for trajectory minimum and maximum adversarial distances with CSL and CSG.

The same trajectory statistics also connect local robustness to privacy leakage. Under error-stability, generalization, and model-bias assumptions, we relate a sample's CSL to its membership-inference advantage. Combining this result with the adversarial-distance bound yields an explicit relation between sample-wise membership leakage and local adversarial distance. The resulting proxies are obtained from quantities accumulated during training, avoiding per-sample adversarial optimization and auxiliary-model retraining.

Together, these results provide a sample-level view of memorization, robustness, privacy, and data quality. Learning dynamics describe how an example is fit, while our analysis reveals what that trajectory implies about the model's local geometry and its exposure to membership inference. This turns training dynamics into a reusable diagnostic signal for both model behavior and dataset quality.

Our approach makes the following contributions:

- We establish a formal connection between sample-level learning dynamics and local adversarial robustness. Specifically, we bound CSG through CSL over the training trajectory and derive two-sided bounds on untargeted adversarial distance from sample loss and input-gradient statistics, connecting learning dynamics to local decision geometry.
- We connect the same learning dynamics to membership inference. Under stability, generalization, and model-bias assumptions, we relate CSL to membership-inference advantage, yielding a theoretical link between a sample's training trajectory, its distance to the decision boundary, and its privacy leakage.
- We turn these results into low-overhead sample-level auditing tools that reuse statistics collected during training. Across CIFAR-10 and CIFAR-100 with multiple ResNet architectures, the resulting proxies track empirical adversarial distance and membership vulnerability while supporting mislabeled-sample detection and duplicate detection without per-sample adversarial optimization or auxiliary-model retraining.

2 NOTATION

Throughout this paper, vectors are denoted by overhead arrows (e.g., $\vec{v}$), and matrices by uppercase italics (e.g., H). Distributions are represented using bold capital letters (e.g., $\mathbf{D}$).
We formulate the learning problem as discovering a mapping $f : \vec{x} \mapsto y$, where the input feature vector is $\vec{x} \in \mathbb{R}^d$ and the corresponding categorical label is $y \in \{1, \ldots, K\}$, with K representing the total number of classes, and d representing the input dimension. We assume the joint tuple is drawn from a data distribution, $(\vec{x}, y) \sim \mathbf{D}$. A training dataset $S = (\vec{z}_1, \ldots, \vec{z}_{m-1}) \sim \mathbf{D}^{m-1}$ consists of $m-1$ independent and identically distributed (iid) samples, where each sample is $\vec{z}_i = (\vec{x}_i, y_i)$. To facilitate our analysis of membership inference, we construct two adjacent datasets of size m: an inclusive dataset $S_i = S \cup \{\vec{z}_i\}$ containing a specific target sample $\vec{z}_i = (\vec{x}_i, y_i)$, and an exclusive dataset $S_j = S \cup \{\vec{z}_j\}$ where $\vec{z}_j \sim \mathbf{D}$ is an independently drawn out-of-sample instance. Dataset $S^{\setminus i}$ is constructed by removing the i-th element $\vec{z}_i$ from S, and hence has one less element than S.

A deep neural network, parameterized by a weight vector $\vec{w} \in \mathbb{R}^p$, is defined as a mapping $\vec{g}_{\vec{w}} : \mathbb{R}^d \to \mathbb{R}^K$. For a given input $\vec{x}_i$, the model outputs a K-dimensional vector $\vec{g}_{\vec{w}}(\vec{x}_i)$, representing the model's scores across all classes. Let $\phi \sim \Phi$ denote the intrinsic randomness of the training algorithm $\mathcal{A}$. We evaluate the model's performance using a cost function $c : \mathbb{R}^K \times \{1, \ldots, K\} \to \mathbb{R}^+$. The cost evaluated at a given sample $\vec{z}_i$ is denoted as the loss l, compactly written as $l(\vec{w}, \vec{z}_i) = c(\vec{g}_{\vec{w}^\phi}(\vec{x}_i), y_i)$. We use $\vec{w}^\phi_{S,t}$ to explicitly denote the model weights obtained after training using algorithm $\mathcal{A}$ on a specific dataset S for optimization iteration t, bounded by a maximum of T_{max} iterations. Furthermore, we denote the untargeted adversarial distance of a given input sample $\vec{x}_i$ as $d_{adv}(\vec{x}_i)$.

Finally, the model's generalization capability over the underlying total data distribution $\mathbf{D}$ is quantified by the population risk, defined as $R(\vec{w}) = \mathbb{E}_{\vec{z}\sim\mathbf{D}}[l(\vec{w}, \vec{z})]$. Because the true distribution $\mathbf{D}$ is generally unknown and hence inaccessible, training relies on the empirical risk, computed over the dataset S as $R_{emp}(\vec{w}, S) = \frac{1}{m-1}\sum_{i=1}^{m-1} l(\vec{w}, \vec{z}_i)$, where $\vec{z}_i \in S$.

3 Background

Error stability. Error stability of a randomized algorithm $\mathcal{A}$ for some $\beta > 0$, where S is a dataset of size m, and where $\vec{z} \sim \mathbf{D}$ and $\phi \sim \mathbf{\Phi}$ is defined as (Kearns & Ron, 1997):

$$\forall i \in \{1, \ldots, m\}, \quad |\mathbb{E}_{\phi,\vec{z}}[l(\vec{w}^{\phi}_{S,t}, \vec{z})] - \mathbb{E}_{\phi,\vec{z}}[l(\vec{w}^{\phi}_{S^{\setminus i},t}, \vec{z})]| \leq \beta \tag{1}$$

Generalization. A randomized algorithm $\mathcal{A}$ is said to generalize if there exists a rate $\gamma > 0$ such that with probability at least $1 - \xi$ the following holds (Feldman & Vondrak, 2019):

$$\mathbb{E}_{\phi}[|R_{emp}(\vec{w}^{\phi}_{S,t}, S) - R(\vec{w}^{\phi}_{S,t})|] \leq \gamma \tag{2}$$

Uniform model bias. The hypothesis parameterized by weights $\vec{w}^{\phi}_{S,t}$, resulting from the application of algorithm $\mathcal{A}$ to learn the theoretically optimal true conditional mapping $\vec{h}^* = \mathbb{E}[y|\vec{x}]$ from a dataset $S \sim \mathbf{D}^m$, has a uniform bound on model bias Δ (Ravikumar et al., 2024) if:

$$\forall S \sim \mathbf{D}^m, \quad |\mathbb{E}_{\phi}[R(\vec{w}^{\phi}_{S,t}) - R(\vec{h}^*)]| \leq \Delta \tag{3}$$

ρ-Lipschitz Hessian. The Hessian of the loss l with respect to the input is Lipschitz continuous if there exists $\rho > 0$ such that (Ravikumar et al., 2024):

$$\forall \vec{z}_1, \vec{z}_2, \quad ||\nabla^2_{\vec{x}_1} l(\vec{w}^{\phi}_{S,t}, \vec{z}_1) - \nabla^2_{\vec{x}_2} l(\vec{w}^{\phi}_{S,t}, \vec{z}_2)|| \leq \rho||\vec{x}_1 - \vec{x}_2|| \tag{4}$$

L-bounded loss. A loss function is L-bounded if it satisfies (Ravikumar et al., 2025; 2026; 2024):

$$\forall \vec{w}, \vec{z}, \quad 0 \leq l(\vec{w}, \vec{z}) \leq L \tag{5}$$

ρ_x-Lipschitz input gradients. A loss function l is ρ_x-smooth with respect to the input space if its gradient is Lipschitz continuous (Moosavi-Dezfooli et al., 2019). Formally, there exists a constant $\rho_x > 0$ such that:

$$\forall \vec{z}_1, \vec{z}_2, \quad ||\nabla_{\vec{x}_1} l(\vec{w}^{\phi}_{S,t}, \vec{z}_1) - \nabla_{\vec{x}_2} l(\vec{w}^{\phi}_{S,t}, \vec{z}_2)||_2 \leq \rho_x||\vec{x}_1 - \vec{x}_2||_2 \tag{6}$$

Memorization. For sample $\vec{z}_i \in S$ it is defined by Feldman (2020) for an algorithm $\mathcal{A}$ as:

$$\text{mem}(\mathcal{A}, S, i) = \Pr_{\phi}[\arg\max_{k\in\{1,\ldots,K\}} [\vec{g}_{\vec{w}^{\phi}_{S}}(\vec{x}_i)]_k = y_i] - \Pr_{\phi}[\arg\max_{k\in\{1,\ldots,K\}} [\vec{g}_{\vec{w}^{\phi}_{S^{\setminus i}}}(\vec{x}_i)]_k = y_i]] \tag{7}$$

Untargeted Adversarial distance. The untargeted adversarial distance $d_{adv}(\vec{x}_i)$ is defined as the minimum perturbation $\vec{\delta} \in \mathbb{R}^d$ required in the input space to force the model to misclassify:

$$d_{adv}(\vec{x}_i) \triangleq \inf_{\vec{\delta}\in\mathbb{R}^d} \left\{ \|\vec{\delta}\|_2 \;\middle|\; \arg\max_{k\in\{1,\ldots,K\}} [\vec{g}_{\vec{w}^{\phi}}(\vec{x}_i + \vec{\delta})]_k \neq y_i \right\} \tag{8}$$

MIA advantage. For an adversary $\mathcal{B} : \mathbb{R}^p \times \mathcal{Z} \to \{0, 1\}$ tasked with predicting whether $\vec{z}_i$ belongs to the training set of a target model, the advantage is the difference between the true positive rate and false positive rate over $\phi \sim \mathbf{\Phi}$, $S \sim \mathbf{D}^{m-1}$ and $\vec{z}_j \sim \mathbf{D}$ (Yeom et al., 2018) :

$$Adv_{MIA}(\vec{z}_i) \triangleq Pr_{\phi}[\mathcal{B}(\vec{w}^{\phi}_{S_i,t}, \vec{z}_i) = 1] - Pr_{\phi}[\mathcal{B}(\vec{w}^{\phi}_{S_j,t}, \vec{z}_i) = 1] \tag{9}$$

4 Related Works

Memorization and learning dynamics. Sample-level analysis requires both a definition of memorization and an estimator that scales beyond repeated retraining. Feldman (2020) studies why memorization can be necessary for learning long-tailed distributions. Feldman & Zhang (2020) quantify sample memorization through the effect of including individual examples in training. Their formulation provides a direct sample-level notion of memorization, but estimating it requires training over many dataset subsets.

Work on learning dynamics replaces repeated retraining with signals collected during optimization. Toneva et al. (2019) characterize samples through forgetting events. Maini et al. (2022) use second-split forgetting to distinguish mislabeled, rare, and complex examples. Curvature-based statistics provide another sample-level view of memorization Garg et al. (2024). CSL and CSG move this line toward training-efficient memorization estimation by accumulating per-sample loss and input-gradient information Ravikumar et al. (2025; 2026). We use these learning dynamics for a different target. Our analysis derives what they imply about adversarial distance and membership leakage.

Adversarial distance and local margins. Adversarial distance measures the minimum perturbation required to cross a model's decision boundary. DeepFool (Moosavi-Dezfooli et al., 2016) iteratively approximates a nearby boundary crossing, while optimization-based attacks such as Carlini & Wagner (2016) and minimum-norm attack procedures estimate this distance through repeated input optimization. These methods provide direct post-training measurements but require attack computation for every audited sample. Our analysis takes the complementary direction. We derive adversarial-distance bounds from sample loss and input gradients, then use their trajectory-level counterparts to identify vulnerability from statistics collected during training. This replaces repeated attack optimization with a single reusable training-time signal.

Membership inference and record-level privacy risk. Membership inference attacks determine whether a specific example was used during training. Shokri et al. (2017) establish the shadow-model paradigm, while Yeom et al. (2018) connect membership leakage to overfitting via loss-based attacks. More recent methods such as LiRA (Carlini et al., 2022), AttackR (Ye et al., 2022), and RMIA (Zarifzadeh et al., 2023) estimate record-level membership evidence by comparing target behavior against auxiliary models or reference distributions. These methods measure vulnerability directly, but their computational cost makes exhaustive record-level auditing expensive.

Artifact-based methods instead predict privacy risk from quantities already produced by the target model. Pollock et al. (2025) empirically show that statistics of the loss trajectory can identify high-risk records without executing a full reference attack. Our work connects this empirical direction to learning-dynamics theory. We derive a relation between CSL and optimal membership-inference advantage (Yeom et al., 2018), then use the temporal change in cumulative loss to construct CSL-Diff. Unlike the reference attacks used as evaluation targets, CSL-Diff requires no shadow or reference models at deployment and predicts vulnerability across LiRA, AttackR, and RMIA operating points.

5 Learning Dynamics, Adversarial Distance and Membership Inference Attack

5.1 Linking CSL and CSG

To analyze the learning dynamics of individual samples, we evaluate the optimization trajectory over T_{max} iterations. Following the SGD convergence framework established by Ghadimi & Lan (2013) and utilized in Ravikumar et al. (2025; 2026), we evaluate the model weights $\vec{w}^{\phi}_{S,t}$ at a randomly stopped optimization iteration $t \sim \mathbf{T}$. Here, $\mathbf{T}$ is a discrete random variable sampled non-uniformly from the set $\{1, \ldots, T_{max}\}$. We define Cumulative Sample Loss (CSL) and Cumulative Sample Gradient (CSG) as expectations over this non-uniformly sampled iteration following the definitions of Ravikumar et al. (2025; 2026):

$$CSL(\vec{z}_i) \triangleq T_{max}\mathbb{E}_{t\sim\mathbf{T}}[l(\vec{w}^{\phi}_{S,t}, \vec{z}_i)] \quad \text{and} \quad CSG(\vec{z}_i) \triangleq T_{max}\mathbb{E}_{t\sim\mathbf{T}}[\|\nabla_{\vec{x}_i} l(\vec{w}^{\phi}_{S,t}, \vec{z}_i)\|_2] \tag{10}$$

Theorem 1. *Assume that the loss function is strictly non-negative ($l \geq 0$), and the assumptions of ρ_x-Lipschitz input gradients 6 and ρ-Lipschitz Hessian 4 hold. Then for an arbitrary perturbation scalar $\epsilon > 0$, the expected Cumulative Sample Gradient (CSG) of an input sample $\vec{z}_i$ and its expected Cumulative Sample Loss (CSL) satisfies:*

$$\epsilon \frac{CSG(\vec{z}_i)}{T_{max}} \leq \frac{CSL(\vec{z}_i)}{T_{max}} + \frac{\rho_x}{2}\epsilon^2 + \frac{\rho}{6}\epsilon^3 \tag{11}$$

Sketch of Proof. The proof relies on the ρ-Lipschitz Hessian assumption, directly adapting the formulation of Nesterov & Polyak (2006) utilized in Lemma A.3 of Ravikumar et al. (2024). By evaluating this polynomial bound for a specifically aligned perturbation of magnitude ϵ, we isolate the input gradient norm. Taking the expectation over $\mathbf{T}$ bridges these per-iteration bounds to their cumulative trajectory metrics. The full proof is provided in Appendix B.1.

Interpreting the Theory. Theorem 1 establishes a direct upper bound on CSG using CSL. Prior literature introduces CSL (Ravikumar et al., 2025) and CSG (Ravikumar et al., 2026) as independent proxies for memorization (Feldman & Zhang, 2020). Despite measuring fundamentally different properties of learning dynamics and relying on distinct proof techniques, both metrics demonstrate remarkably similar empirical performance in estimating memorization and facilitating dataset diagnostics. Theorem 1 provides the theoretical justification for this phenomenon.

5.2 Bounds on Untargeted Adversarial Distance

For an adversarial perturbation $\vec{\delta}$ to successfully force a misclassification on a sample $\vec{z}_i$, the perturbed loss must cross a decision boundary. We define L_c as the critical threshold necessary to cross the decision boundary, and L_u as the threshold sufficient to guarantee misclassification for a non-negative loss with K classes. To avoid ties, we fix $\tau > 0$ and set $L_u = \ln(K) + \tau$, and $L_c = \ln(2)$ for Cross-Entropy, we formally prove this in Lemma 1 in Appendix B.2.

Theorem 2. *Assume a strictly non-negative loss ($l \geq 0$) and smoothness assumptions 6, 4. Let the necessary and sufficient misclassification thresholds be L_c and L_u.*

Single Checkpoint: *For a single iteration t, let $C_t = \|\nabla_{\vec{x}_i} l(\vec{w}^{\phi}_{S,t}, \vec{z}_i)\|_2$, margins $\Delta^{(c)}_t = L_c - l(\vec{w}^{\phi}_{S,t}, \vec{z}_i)$ and $\Delta^{(u)}_t = L_u - l(\vec{w}^{\phi}_{S,t}, \vec{z}_i)$, and $\tilde{B}_t = \frac{\rho_x}{2} + \frac{\rho}{6}(\frac{\Delta^{(c)}_t}{C_t})$. The adversarial distance δ_t is unconditionally bounded from below, and if $C_t^2 \geq 2\rho_x \Delta^{(u)}_t$, bounded from above:*

$$\frac{2\Delta^{(u)}_t}{C_t + \sqrt{C_t^2 - 2\rho_x \Delta^{(u)}_t}} \geq \delta_t \geq \frac{2\Delta^{(c)}_t}{\sqrt{C_t^2 + 4\tilde{B}_t \Delta^{(c)}_t} + C_t} \tag{12}$$

Trajectory-Based: *For the trajectory $\mathbf{T}$, let $C = \frac{CSG(\vec{z}_i)}{T_{max}}$, $\Delta_{min} = L_c - \frac{CSL(\vec{z}_i)}{T_{max}}$, and $\tilde{B} = \frac{\rho_x}{2} + \frac{\rho}{6}(\frac{\Delta_{min}}{C})$. The minimum and maximum adversarial distances, $\delta_{min} = \min_t \delta_t$ and $\delta_{max} = \max_t \delta_t$, are strictly bounded by the learning dynamics (provided $C_t^2 \geq 2\rho_x \Delta^{(u)}_t, \forall t \in \mathbf{T}$ for the upper bound):*

$$\delta_{min} \leq \frac{2\left(L_u T_{max} - CSL(\vec{z}_i)\right)}{CSG(\vec{z}_i)}, \quad \delta_{max} \geq \frac{2\Delta_{min}}{\sqrt{C^2 + 4\tilde{B}\Delta_{min}} + C} \tag{13}$$

Sketch of Proof. For the lower bounds, we use the ρ-Lipschitz Hessian assumption 4 via Nesterov & Polyak (2006) to construct a monotonic cubic inequality modeling the maximum theoretical loss growth. Enforcing the necessary boundary crossing condition $l \geq L_c$ and extracting the positive root yields the local lower bound δ_t. Taking the expectation over $\mathbf{T}$ and factoring out δ_{max} extends this identical cubic structure to CSL and CSG. For the upper bounds, the ρ_x-Lipschitz gradient assumption 6 establishes a guaranteed minimum loss growth for a gradient-aligned perturbation. Enforcing the sufficient condition $l \geq L_u$ bounds δ_t from above via a quadratic root. Taking the expectation over $\mathbf{T}$, and substituting δ_{min} isolates it within the expected gradient product, yielding the trajectory-driven upper bound. The full proof is provided in Appendix B.2.

Interpreting the Theory. Theorem 2 introduces two-sided bounds on a sample's adversarial distance, explicitly linking it to learning dynamics over the optimization trajectory. Our bounds clearly demonstrate that memorization proxies (CSL and CSG) directly dictate the geometry of the decision boundary. From a Defender's perspective, the lower bound guarantees that no attack with a perturbation less than that radius can be successful. From an Attacker's perspective, the upper bound guarantees the existence of a successful attack within that perturbation budget.

5.3 Linking Learning Dynamics to Membership Inference

To formally link a sample's learning dynamics to privacy leakage, we model a membership inference adversary tasked with identifying if a target sample $\vec{z}_i$ belongs to the training set. We utilize adjacent training datasets of size m: the inclusive dataset $S_i = S \cup \{\vec{z}_i\}$ and exclusive dataset $S_j = S \cup \{\vec{z}_j\}$, where S is a dataset of size $m-1$ and $\vec{z}_j \sim \mathcal{D}$ is a randomly drawn out-sample.

Theorem 3. *Assume an L-bounded loss function and that the assumptions of error stability β (1), generalization γ (2), and uniform model bias Δ (3) hold. The expected optimal Membership Inference Attack advantage $Adv^*_{MIA}(\vec{w}_t, \vec{z}_i)$ is strictly lower-bounded by the expected CSL:*

$$\frac{m-2}{m^2}\mathbb{E}_{\phi, S\sim\mathcal{D}^{m-1}}\left[\frac{CSL(\vec{z}_i)}{T_{max}}\right] \leq \frac{L}{m}\mathbb{E}_{t\sim\mathbf{T}}[Adv^*_{MIA}(\vec{w}_t, \vec{z}_i)] + \kappa \tag{14}$$

where κ is a constant made of the constants of uniform bias, generalization, and error stability.

Sketch of Proof. We bound the empirical risk discrepancies between models trained on S_i and S_j. Decomposing risks over the base dataset S and sequentially applying the stability, generalization, and uniform bias bounds analytically isolates the target sample's expected loss. We relate this loss

to privacy using a randomized loss based attack following Theorem 2 of Yeom et al. (2018), which is always upper-bounded by the optimal adversary (Choi et al., 2023). Taking the expectation over $t \sim \mathbf{T}$ and substituting the CSL definition yields the final constraint. Derivation is in Appendix B.3.

Interpreting the Theory. Theorem 3 formally links the temporal accumulation of loss to privacy leakage, rigorously justifying empirical observations that models leak more membership information for highly memorized samples (Choi et al., 2023). It establishes that a sample's resistance to fit which is quantified by its CSL forces a theoretical lower bound on the optimal MIA advantage. This indicates that only tracking the sample loss enables identifying vulnerable samples in the dataset at zero-computational overhead instead of training shadow models for quick privacy auditing.

5.4 Unified Vulnerability Bounds: Privacy, Memorization, and Robustness

By utilizing Theorems 1, 2, and 3, we relate the trajectory extremes of adversarial distance directly to membership inference advantage and sample-level memorization.

Corollary 1. *Assume the conditions for Theorems 1 and 2 hold, with* $K_\epsilon = \frac{\rho_x}{2}\epsilon + \frac{\rho}{6}\epsilon^2$. *Let* L_u *and* L_c *be the sufficient and necessary misclassification thresholds, and let* ϕ *denote the realization of algorithmic randomness.*
(i) For strictly non-negative losses under the memorization framework of Ravikumar et al. (2025), expected memorization is bounded by $\mathbb{E}_\phi[mem(\vec{z}_i)] \leq \mathbb{E}_\phi[CSL(\vec{z}_i)] + \beta - \xi$. *Provided* $C_t^2 \geq 2\rho_x \Delta_t^{(u)}$ *for all* t *and* ϕ, *the absolute minimum adversarial distance* $\delta_{\inf}(\vec{x}_i) = \inf_\phi \delta_{min}(\vec{x}_i)$ *is upper-bounded by expected memorization:*

$$\delta_{\inf}(\vec{x}_i) \leq \frac{2\left(L_u T_{max} - \mathbb{E}_\phi[mem(\vec{z}_i)] + \beta - \xi\right)}{\mathbb{E}_\phi[CSG(\vec{z}_i)]} \tag{15}$$

(ii) For L-bounded losses satisfying Theorem 3, let $\alpha = \frac{mL}{m-2}$, $\kappa = \frac{m^2 K}{m-2}$, *and* $\overline{Adv} = \mathbb{E}_{t\sim\mathbf{T}}[Adv^*_{MIA}(\vec{w}_t, \vec{z}_i)]$. *The absolute maximum adversarial distance* $\delta_{\sup}(\vec{x}_i) = \sup_\phi \delta_{max}(\vec{x}_i)$ *is unconditionally lower bounded by the expected optimal MIA advantage:*

$$\delta_{\sup}(\vec{x}_i) \geq \frac{2\Delta_{mia}}{\sqrt{C_{mia}^2 + 4\tilde{B}_{mia}\Delta_{mia}} + C_{mia}} \tag{16}$$

where $\Delta_{mia} = L_c - (\alpha\overline{Adv} + \kappa)$, $C_{mia} = \frac{1}{\epsilon}(\alpha\overline{Adv} + \kappa) + K_\epsilon$, *and* $\tilde{B}_{mia} = \frac{\rho}{6}\left(\frac{\Delta_{mia}}{C_{mia}}\right) + \frac{\rho_x}{2}$.

Takeaways. Corollary 1 mathematically unifies adversarial robustness, membership privacy, and memorization (proofs in Appendix B.4). As memorization increases, the upper bound strictly decreases, forcing memorized samples into a tighter boundary radius at their most vulnerable point, entirely independent of algorithmic randomness ($\delta_{\inf}$). Conversely, as the MIA advantage increases, the lower bound shrinks, compressing the absolute worst-case safety margin ($\delta_{\sup}$). Thus, severe privacy leakage or high memorization fundamentally limits the possible adversarial distance a sample can achieve under any seed, making the Defender's task strictly tougher.

6 Experiments

For Experiment 6.1 we use the trajectory proxy $\bar{d}_{adv,T} = \frac{2(L_u T - CSL(\vec{z}_i))}{CSG(\vec{z}_i)}$. For Experiments 6.2 and 6.3, we set $T = 1$, we use the single-checkpoint proxy: $\bar{d}_{adv,1} = \frac{2(L_u - l(\vec{w}^\phi_{S,t}, \vec{z}_i))}{\|\nabla_{\vec{x}_i} l(\vec{w}^\phi_{S,t}, \vec{z}_i)\|_2}$. While these proxies mirror Theorem 2's upper bounds (for δ_{min} and δ_t); rather than being certified bounds, these are empirical proxies because we omit computing ρ_x empirically to verify the prerequisite condition $C_t^2 \geq 2\rho_x \Delta_t^{(u)}$. The $T = 1$ proxy successfully isolates mislabeled and duplicate samples because their persistently high gradients tend to satisfy this condition even at the final checkpoint. Conversely, global dataset auditing (Appendix section C.5) requires the trajectory proxy because normal samples' gradients tend to vanish by the end of training, causing the condition to fail locally.

6.1 Validating Theory

Experiment. We empirically evaluate our Theorems 1, 2, 3 and Corollary 1 linking Cumulative Sample Loss (CSL), Cumulative Sample Gradient (CSG), adversarial distance (d_{adv}), and MIA advantage. We evaluate ResNet-18, 34 and 50 (He et al., 2016) trained via SGD on CIFAR-10 and

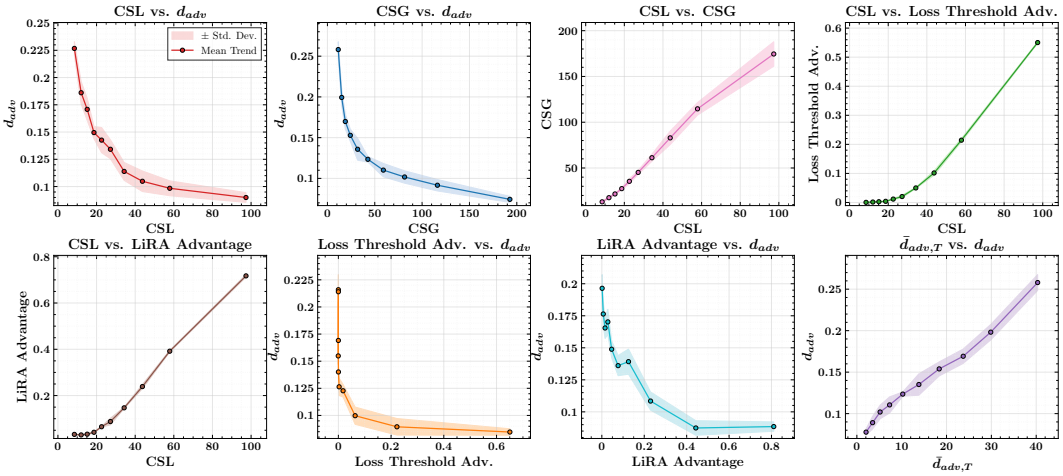


Figure 1: Empirical evaluation of privacy, adversarial vulnerability, and learning dynamics metrics for ResNet-18 on CIFAR-10. Clear monotonic relations are observed in $\bar{d}_{adv,T}$ vs d_{adv}, CSL vs CSG, and CSL against Loss threshold Advantage (Yeom et al., 2018), and LiRA Advantage (Carlini et al., 2022). Variance shading represents standard deviation across 3 training seeds.

Table 1: Alignment metrics across ResNet variants evaluated over 3 independent seeds. C-10 and C-100 stand for CIFAR-10 and CIFAR-100 respectively, and R-18, 34, 50 stand for ResNet-18, 34, 50 respectively. We report Spearman rank (ρ) and Cosine Similarity (cos) to capture alignment

		d_{adv} vs. $\bar{d}_{adv,T}$ (Theoretical)		CSL vs. CSG		CSL vs. LiRA Adv.		CSL vs. Yeom Loss Adv.	
Dataset	**Model**	ρ	cos	ρ	cos	ρ	cos	ρ	cos
C-10	**R-18**	0.720 ± 0.007	0.906 ± 0.004	0.901 ± 0.007	0.935 ± 0.006	0.705 ± 0.004	0.872 ± 0.009	0.741 ± 0.009	0.781 ± 0.012
	R-34	0.703 ± 0.016	0.898 ± 0.001	0.886 ± 0.008	0.914 ± 0.012	0.666 ± 0.021	0.828 ± 0.013	0.704 ± 0.022	0.735 ± 0.013
	R-50	0.750 ± 0.017	0.909 ± 0.001	0.897 ± 0.007	0.905 ± 0.012	0.659 ± 0.004	0.795 ± 0.032	0.714 ± 0.006	0.737 ± 0.035
C-100	**R-18**	0.507 ± 0.023	0.857 ± 0.007	0.830 ± 0.006	0.944 ± 0.001	0.855 ± 0.016	0.945 ± 0.002	0.855 ± 0.020	0.904 ± 0.007
	R-34	0.480 ± 0.029	0.853 ± 0.005	0.812 ± 0.006	0.943 ± 0.002	0.849 ± 0.009	0.946 ± 0.002	0.852 ± 0.009	0.902 ± 0.004
	R-50	0.558 ± 0.014	0.857 ± 0.003	0.793 ± 0.011	0.935 ± 0.007	0.832 ± 0.009	0.930 ± 0.005	0.831 ± 0.010	0.887 ± 0.010

100 (Krizhevsky, 2009). Theorems 1 and 2 are validated on full datasets ($N = 50,000$), whereas Theorem 3 and Corollary 1 utilize random half-subsets ($N = 25,000$) to enable Yeom's loss threshold (Yeom et al., 2018) and LiRA (Carlini et al., 2022) MIA evaluations via 64 shadow models per configuration. Final-checkpoint adversarial distance d_{adv} is computed with Carlini and Wagner attack(C&W attack) (Carlini & Wagner, 2016). Appendix C.1 has details about the experiment.

Results. Figure 1 represents the results for CIFAR-10 and ResNet-18. Figure 5 to Figure 9 in the Appendix represent the results for the other configurations. For visual clarity, data points are partitioned into 10 equal-frequency bins (deciles), plotted at their mean, with solid lines and shaded regions representing the mean and standard deviation across 3 seeds. Table 1 reports the results for all configurations for Spearman's rank correlation coefficient (ρ) and Cosine Similarity (cos), and Appendix Table 4 reports the results for the Pearson's correlation coefficient (r) scores.

Takeaways. As shown in Figure 1 and Table 1 (ρ and cos scores), CSL and CSG exhibit a near-linear relationship, validating Theorem 1. Furthermore, d_{adv} inversely correlates with both CSL and CSG which aligns with the bounds of Theorem 2, while its trajectory proxy $\bar{d}_{adv,T}$ scales linearly. Across both calculation techniques, MIA advantage relates approximately linearly with CSL (validating Theorem 3) and inversely with d_{adv}, hence is related to Corollary 1. Additionally we note that Ravikumar et al. (2025; 2026) observed an inverse relationship between memorization and adversarial distance which relates to Corollary 1. Collectively, these results confirm that our proposed training-dynamics quantities effectively capture both adversarial vulnerability and privacy leakage.

6.2 Mislabeled Sample Detection

Experiment. We detect mislabeled samples by thresholding our adversarial distance proxy $\bar{d}_{\text{adv},1}$ for a single checkpoint, as defined in Theorem 2, and follow the experimental setup of Ravikumar et al. (2026). We train ResNet-18 on CIFAR-10 and CIFAR-100 under 5% to 30% symmetric label noise (uniform flipping of labels to other classes) and evaluate AUROC (higher indicates better). Refer to Appendix C.2 for more details on the setup and the baselines.

Table 2: AUROC (mean ± std over 3 seeds of ResNet-18) for duplicate and mislabeled-sample detection on CIFAR-10 and CIFAR-100. $\bar{d}_{adv,1}$ is the final checkpoint proxy. Second best results are underlined; best overall results are in **bold**.

	Method	Duplicate	5% Noise	10% Noise	20% Noise	25% Noise	30% Noise
CIFAR-10	CL (Northcutt et al., 2021)	0.5533 ± 0.0031	0.8874 ± 0.0019	0.8551 ± 0.0030	0.7169 ± 0.1539	0.6960 ± 0.1387	0.6794 ± 0.1264
	In Conf. (Carlini et al., 2019)	0.9237 ± 0.0114	0.7254 ± 0.0214	0.6528 ± 0.0042	0.5978 ± 0.0131	0.5800 ± 0.0051	0.5669 ± 0.0106
	SSFT (Maini et al., 2022)	0.8490 ± 0.0034	0.9498 ± 0.0042	0.9360 ± 0.0020	0.9077 ± 0.0023	0.8910 ± 0.0050	0.8710 ± 0.0071
	Curv. (Garg et al., 2024)	0.9536 ± 0.0030	0.9800 ± 0.0003	0.9819 ± 0.0006	0.9934 ± 0.0002	0.9932 ± 0.0001	0.9932 ± 0.0006
	CSL (Ravikumar et al., 2025)	0.9822 ± 0.0011	0.9845 ± 0.0026	0.9864 ± 0.0004	0.9870 ± 0.0003	0.9904 ± 0.0005	0.9906 ± 0.0003
	CSG (Ravikumar et al., 2026)	0.9525 ± 0.0015	0.9783 ± 0.0009	0.9809 ± 0.0011	0.9936 ± 0.0002	0.9934 ± 0.0001	0.9935 ± 0.0006
	SGAL (Ravikumar et al., 2026)	0.9255 ± 0.0019	0.9115 ± 0.0057	0.8835 ± 0.0128	0.8664 ± 0.0070	0.8553 ± 0.0145	0.8470 ± 0.0283
	$\bar{d}_{adv,1}$ (Ours)	**0.9998 ± 0.0001**	**0.9963 ± 0.0006**	**0.9955 ± 0.0002**	**0.9953 ± 0.0002**	**0.9946 ± 0.0003**	**0.9947 ± 0.0004**
CIFAR-100	CL (Northcutt et al., 2021)	0.5873 ± 0.0090	0.8733 ± 0.0010	0.8536 ± 0.0006	0.7030 ± 0.1565	0.6833 ± 0.1427	0.6662 ± 0.1289
	In Conf. (Carlini et al., 2019)	0.8623 ± 0.0131	0.7069 ± 0.0069	0.6884 ± 0.0053	0.6493 ± 0.0075	0.6324 ± 0.0051	0.6257 ± 0.0044
	SSFT (Maini et al., 2022)	0.7938 ± 0.0045	0.8784 ± 0.0030	0.8664 ± 0.0024	0.8358 ± 0.0008	0.8203 ± 0.0016	0.8043 ± 0.0061
	Curv. (Garg et al., 2024)	0.9639 ± 0.0030	0.9876 ± 0.0021	0.9892 ± 0.0012	0.9931 ± 0.0004	0.9931 ± 0.0004	0.9932 ± 0.0002
	CSL (Ravikumar et al., 2025)	0.9894 ± 0.0012	0.9891 ± 0.0003	0.9895 ± 0.0002	0.9902 ± 0.0002	0.9904 ± 0.0002	0.9903 ± 0.0002
	CSG (Ravikumar et al., 2026)	0.9743 ± 0.0022	0.9896 ± 0.0008	0.9904 ± 0.0006	0.9934 ± 0.0003	0.9936 ± 0.0002	0.9936 ± 0.0001
	SGAL (Ravikumar et al., 2026)	0.9537 ± 0.0017	0.9895 ± 0.0006	0.9897 ± 0.0008	0.9856 ± 0.0008	0.9863 ± 0.0009	0.9861 ± 0.0004
	$\bar{d}_{adv,1}$ (Ours)	**0.9994 ± 0.0001**	**0.9942 ± 0.0001**	**0.9944 ± 0.0001**	**0.9946 ± 0.0001**	**0.9947 ± 0.0001**	**0.9947 ± 0.0002**

Table 3: Precision@k on LiRA vulnerability (FPR=10^{-3}) across 3 seeds on CIFAR-10 and CIFAR-100 (256 shadows) on WideResNet28-2. Second best results are underlined; best results are in **bold**.

	CIFAR-10			CIFAR-100		
Method	**Prec@1%**	**Prec@3%**	**Prec@5%**	**Prec@1%**	**Prec@3%**	**Prec@5%**
Final Loss	0.055 ± 0.012	0.080 ± 0.011	0.095 ± 0.016	0.003 ± 0.002	0.010 ± 0.012	0.018 ± 0.014
CSG (Ravikumar et al., 2026)	0.277 ± 0.033	0.325 ± 0.021	0.323 ± 0.007	0.175 ± 0.041	0.212 ± 0.025	0.239 ± 0.033
LT-Mean (Pollock et al., 2025)	0.471 ± 0.065	0.467 ± 0.022	0.445 ± 0.010	0.072 ± 0.042	0.116 ± 0.050	0.179 ± 0.066
LT-Std (Pollock et al., 2025)	0.729 ± 0.044	0.590 ± 0.028	0.517 ± 0.027	0.897 ± 0.008	0.826 ± 0.021	0.762 ± 0.019
LT-IQR (Pollock et al., 2025)	0.793 ± 0.027	0.636 ± 0.018	0.540 ± 0.005	0.927 ± 0.024	0.856 ± 0.023	0.797 ± 0.021
CSL-Diff-40 (Ours)	**0.839 ± 0.023**	**0.685 ± 0.033**	0.587 ± 0.024	**0.947 ± 0.002**	0.878 ± 0.014	0.826 ± 0.013
CSL-Diff-50 (Ours)	0.829 ± 0.040	0.680 ± 0.037	**0.591 ± 0.025**	0.947 ± 0.010	**0.884 ± 0.014**	**0.830 ± 0.022**

Results. Table 2 presents the performance of our method in comparison to the baselines across all noise levels (5%, 10%, 20%, 25% and 30%) for both CIFAR-10 and CIFAR-100.

Takeaways. Last checkpoint $\bar{d}_{\text{adv},1}$ consistently performs better than the previous state-of-the-art methods CSL (Ravikumar et al., 2025) and CSG (Ravikumar et al., 2026) across all noise levels for both the datasets, while being computationally cheaper than CSG as sample gradients are necessary for only the final checkpoint unlike the whole trajectory.

6.3 Duplicate Sample Detection

Experiment. For Figure 2 we train on the unmodified CIFAR-100 dataset on ResNet-18 to identify duplicate samples in the dataset. For comparison to baselines, following the experiment setup of Ravikumar et al. (2025), we modify the CIFAR-10 and CIFAR-100 datasets to introduce 250 duplicate samples into them. We train the datasets on ResNet-18 for this experiment. We used our proxy $\bar{d}_{\text{adv},1}$ to identify duplicates. Refer to Appendix C.3 for more details.

Results. Table 2 presents the performance of our method in comparison to the baselines for both CIFAR-10 and CIFAR-100. AUROC scores are used for evaluation of the methods.

Takaways. Figure 2 clearly indicates the existence of multiple duplicate samples in the unmodified CIFAR-100 dataset which have been identified by our proxy. Table 2 shows that $\bar{d}_{\text{adv},1}$ outperforms all the baselines by a significant margin and is a very reliable method for duplicate sample detection.

6.4 Zero-Cost Privacy Auditing

Experiment. We evaluated the efficacy of our proposed metric, CSL-Diff, in identifying training samples most vulnerable to membership inference attacks (MIAs). We trained WideResNet28-2 models on the CIFAR-10 and CIFAR-100 datasets across 3 target model seeds(trained on same subsets). To assess robustness against varying attacker capabilities, we defined vulnerability targets using LiRA (Carlini et al., 2022), AttackR (Ye et al., 2022), and RMIA (Zarifzadeh et al., 2023) across multiple shadow model budgets (16, 64, and 256) and treated them as ground truth. We compared CSL-Diff against established artifact-based predictors like final loss (Yeom et al., 2018), LT-Std (Pollock et al., 2025), LT-IQR (Pollock et al., 2025)(previous state-of-the art method), CSG (Ravikumar et al., 2026) and others. Please refer to Appendix C.4 for more details.

Results. Our zero-cost metric, CSL-Diff-n, computes the CSL difference between the first and last n% of training checkpoints. Table 3 reports Precision@k, evaluated on the top k% highest-scoring samples against ground-truth vulnerability for identifying LiRA-vulnerable points at a strict False

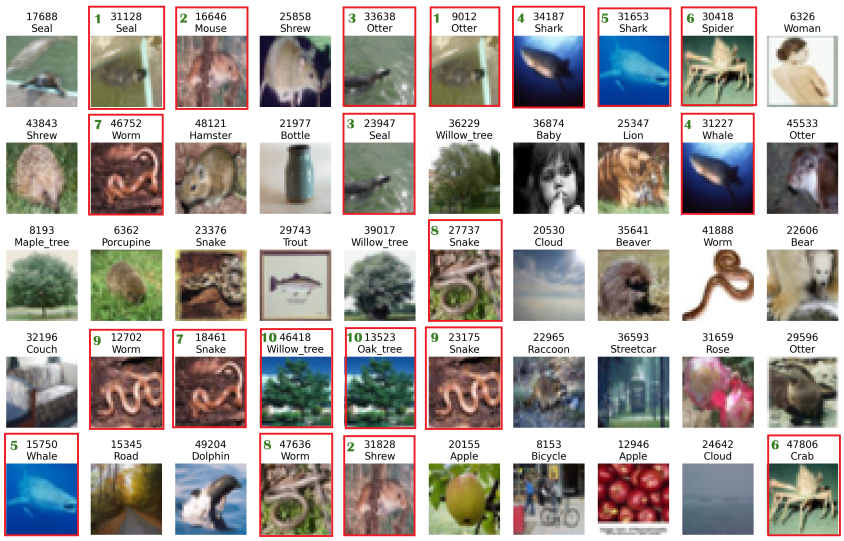


Figure 2: 50 lowest $\bar{d}_{adv,1}$ samples in unmodified CIFAR-100 (on ResNet-18). Visually identical images with conflicting labels, like PyTorch indices 31128 (Seal) and 9012 (Otter), are marked.

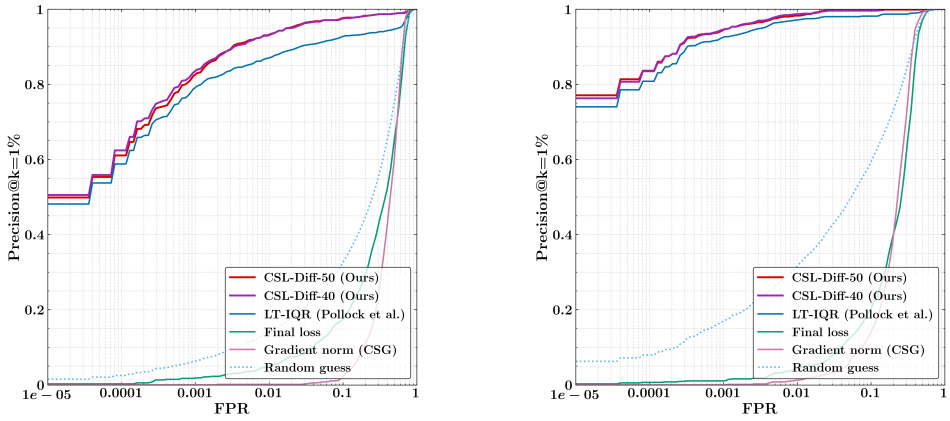


Figure 3: Precision@1% across FPRs for CIFAR-10 (left) & CIFAR-100 (right). Evaluated against LiRA ground truth using 256 shadows on WideResNet28-2 (mean of 3 seeds).

Positive Rate (FPR) of 10^{-3} on CIFAR-10/100 (256 shadows), while Figure 3 plots Precision@1% across an FPR sweep.The appendix details full Precision@k and Recall@k metrics for LiRA and an attack union (LiRA, AttackR, RMIA) in Tables 7–10. Tables 11 and 12 summarize robustness under relaxed FPR bounds (up to 10^{-2}) and restricted shadow budgets (16 and 64), with comprehensive breakdowns supplied in Tables 13–18. Furthermore, FPR sweeps for constrained shadow budgets comparing our method with the baselines, along with detailed precision and recall curves for our CSL-Diff metric across $k \in \{0.1\%, 1\%, 2\%, 3\%, 5\%\}$, are plotted in Appendix Figures 10–17.

Takeaways. CSL-Diff consistently outperforms existing baselines, including CSG and the previous state-of-the-art, LT-IQR. It maintains superior predictive accuracy across diverse threat models (isolated attacks versus a union of attacks), and dynamic FPR security thresholds. Crucially, CSL-Diff demonstrates robustness under restricted shadow model budgets, establishing it as a highly reliable proxy for evaluating privacy vulnerabilities while being zero-cost and aligning with Theorem 3.

7 Conclusion

In this paper we show that sample-level learning dynamics captured by Cumulative Sample Loss (CSL) and Cumulative Sample Gradient (CSG), connect training behavior to adversarial distance and membership-inference vulnerability. Our theoretical analysis establishes bounds under explicit assumptions, while experiments on CIFAR-10 and CIFAR-100 show that our metrics can identify vulnerable samples and support mislabeled-sample and duplicate detection. These audits reuse statistics collected during training, avoiding per-sample attack optimization and auxiliary-model retraining at deployment.

AI USE STATEMENT

In this work, we have not used generative AI tools for any tasks with required disclosure, such as helping develop theoretical models, formulating mathematical claims, providing critical ingredients for proving mathematical claims, or assisting in the writing of proofs. The rest of the required disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools for tasks with recommended disclosure; specifically, to edit the research paper to improve readability (grammar and spelling). We have reviewed all AI-assisted work. All AI-suggested grammar and spelling edits were manually reviewed by the authors to ensure the original scientific meaning was preserved. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our paper submission aligns with the Code of Ethics and does not raise any concerns.

REPRODUCIBILITY STATEMENT

All the assumptions we made for the theory have been defined and stated in the Section 3. The proofs for Theorems 1, 2, 3 and Corollary 1 have been provided in Appendix Section B. We provide detailed setup and configurations for all the experiments performed in Appendix Section C. We also provide code in the supplementary material via an anonymous repository. The computational resources necessary for reproduction has been provided in Appendix C.6.

REFERENCES

Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2016. URL https://api.semanticscholar.org/CorpusID:2893830.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy*, pp. 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649.

Jihye Choi, Shruti Tople, Varun Chandrasekaran, and Somesh Jha. Why train more? effective and efficient membership inference via memorization, 2023. URL https://arxiv.org/abs/2310.08015.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *ACM SIGACT Symposium on Theory of Computing*, 2020.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In Alina Beygelzimer and Daniel Hsu (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1270–1279. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/feldman19a.html.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems*, 2020.

Isha Garg, Deepak Ravikumar, and Kaushik Roy. Memorization through the lens of curvature of loss function around samples. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=WQbDS9RydY.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optim.*, 23:2341–2368, 2013. URL https://api.semanticscholar.org/CorpusID:14112046.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Michael J. Kearns and Dana Ron. Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation. In *Annual Conference on Computational Learning Theory*, pp. 152–162, 1997. doi: 10.1145/267460.267491. URL https://mlanthology.org/colt/1997/kearns1997colt-algorithmic/.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Pratyush Maini, Saurabh Garg, Zachary Lipton, and J Zico Kolter. Characterizing datapoints via second-split forgetting. *Advances in Neural Information Processing Systems*, 35:30044–30057, 2022.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *Conference on Computer Vision and Pattern Recognition*, 2016.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via Curvature Regularization, and Vice Versa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. doi: 10.1109/CVPR.2019.00929. URL https://mlanthology.org/cvpr/2019/moosavidezfooli2019cvpr-robustness/.

Yurii Nesterov and Boris Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108:177–205, 2006. URL https://api.semanticscholar.org/CorpusID:7964929.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *J. Artif. Int. Res.*, 70:1373–1411, May 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12125. URL https://doi.org/10.1613/jair.1.12125.

Maura Pintor, Fabio Roli, Wieland Brendel, and Battista Biggio. Fast minimum-norm adversarial attacks through adaptive norm constraints, 2021. URL https://arxiv.org/abs/2102.12827.

Joseph Pollock, Igor Shilov, Euodia Dodd, and Yves-Alexandre de Montjoye. Free {Record-Level} privacy risk evaluation through {Artifact-Based} methods. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 5525–5544, 2025.

Deepak Ravikumar, Efstathia Soufleri, Abolfazl Hashemi, and Kaushik Roy. Unveiling Privacy, Memorization, and Input Curvature Links. In *International Conference on Machine Learning*, volume 235, pp. 42192–42212, 2024. URL https://mlanthology.org/icml/2024/ravikumar2024icml-unveiling/.

Deepak Ravikumar, Efstathia Soufleri, Abolfazl Hashemi, and Kaushik Roy. Towards memorization estimation: Fast, formal and free. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=KZlQEoEtiu.

Deepak Ravikumar, Efstathia Soufleri, Abolfazl Hashemi, and Kaushik Roy. Memorization through the lens of sample gradients. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jeTiBeW3iZ.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, 2017.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses, 2020. URL https://arxiv.org/abs/2002.08347.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*, pp. 3093–3106, 2022.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *Proceedings of the IEEE 31st Computer Security Foundations Symposium*, pp. 268–282, 2018. doi: 10.1109/CSF.2018.00027.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. *arXiv preprint arXiv:2312.03262*, 2023.

APPENDIX

TABLE OF CONTENTS

A Definitions and Adversarial Distance **13**

B Discussion on Theory **14**
B.1 Proof of Theorem 1 14
B.2 Proof of Theorem 2 15
B.2.1 Preliminary: Necessary and Sufficient Thresholds for Misclassification 15
B.2.2 Proof of the Bounds 15
B.3 Proof of Theorem 3: Related CSL and MIA 19
B.3.1 Dataset Definitions and Goal 19
B.3.2 Preliminary Generalization and Risk Bounds 19
B.3.3 Bounding the Empirical Risk Differences 20
B.3.4 Isolating the Target Sample Loss 20
B.3.5 Relating to Membership Inference Attacks 22
B.4 Proof of Corollary 1: Unified Vulnerability Bounds 22
B.4.1 Preliminaries: Memorization Bound of Ravikumar et al. (2025) 22
B.4.2 Derivation of the Memorization Upper Bound (Non-negative Loss) . . 23
B.4.3 Derivation of the MIA Lower Bound (L-Bounded Loss) 23

C Experiments **24**
C.1 Validation Experiment Details 24
C.2 Mislabeled Detection Setup Details 28
C.2.1 Baseline Methods 29
C.2.2 Compute Cost Analysis 29
C.3 Duplicate Detection: Setup and Additional Analysis 31
C.4 Zero-Cost Privacy Auditing 33
C.4.1 Experimental Setup 33
C.5 Zero-overhead Adversarial Vulnerability Auditing: Setup and Analysis 42
C.6 Compute Resources: 43

A DEFINITIONS AND ADVERSARIAL DISTANCE

In this section, we establish the definitions for adversarial distance and the local classification margin. Recall that $\vec{g}_{\vec{w}^\phi}(\vec{x}_i)$ denotes the K-dimensional vector of scores output by the model parameterized by weights $\vec{w}$ for an input $\vec{x}_i$.

Adversarial Distance:
The untargeted adversarial distance $d_{adv}(\vec{x}_i)$ is defined as the minimum perturbation $\vec{\delta}$ in the input space $\mathbb{R}^d$ that causes a misclassification:

$$d_{adv}(\vec{x}_i) \triangleq \inf_{\vec{\delta}\in\mathbb{R}^d} \left\{ \|\vec{\delta}\|_2 \;\middle|\; \arg\max_{k\in\{1,\dots,K\}} [\vec{g}_{\vec{w}^\phi}(\vec{x}_i+\vec{\delta})]_k \neq y_i \right\} \tag{17}$$

Alternatively, we can define a targeted adversarial distance to a specific class t:

$$d^{targ}_{adv}(\vec{x}_i, t) \triangleq \inf_{\vec{\delta}\in\mathbb{R}^d} \left\{ \|\vec{\delta}\|_2 \;\middle|\; \arg\max_k [\vec{g}_{\vec{w}^\phi}(\vec{x}_i+\vec{\delta})]_k = t \right\} \tag{18}$$

The untargeted and targeted adversarial distances are directly related by finding the minimum targeted distance across all incorrect classes:

$$d_{adv}(\vec{x}_i) = \min_{t\neq y_i} d^{targ}_{adv}(\vec{x}_i, t) \tag{19}$$

B Discussion on Theory

B.1 Proof of Theorem 1

Proof. We begin by relating the loss at a perturbed input $\vec{x}_i + \vec{\delta}$ to the loss at the original input $\vec{x}_i$.The proof relies on the ρ-Lipschitz Hessian assumption, directly adapting the Nesterov-Polyak formulation (Nesterov & Polyak, 2006) utilized in Lemma A.3 of Ravikumar et al. (2024):

$$-\frac{\rho}{6}\|\vec{\delta}\|_2^3 \leq l(\vec{w}_{S,t}^{\phi}, \vec{z}_i+\vec{\delta}) - l(\vec{w}_{S,t}^{\phi}, \vec{z}_i) - \langle \nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i), \vec{\delta}\rangle - \frac{1}{2}\vec{\delta}^T \nabla_{\vec{x}_i}^2 l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)\vec{\delta} \leq \frac{\rho}{6}\|\vec{\delta}\|_2^3 \quad (20)$$

The upper bound on the difference in loss is:

$$l(\vec{w}_{S,t}^{\phi}, \vec{z}_i + \vec{\delta}) - l(\vec{w}_{S,t}^{\phi}, \vec{z}_i) \leq \langle \nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i), \vec{\delta}\rangle + \frac{1}{2}\vec{\delta}^T \nabla_{\vec{x}_i}^2 l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)\vec{\delta} + \frac{\rho}{6}\|\vec{\delta}\|_2^3 \quad (21)$$

For the second-order term, let $H = \nabla_{\vec{x}_i}^2 l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)$. Since the Hessian H is a real and symmetric matrix, Rayleigh's theorem states that:

$$\vec{\delta}^T H \vec{\delta} \leq \lambda_{max}(H)\|\vec{\delta}\|_2^2 \quad (22)$$

Under the assumption of ρ_x-Lipschitz input gradients as descibed in Equation 6, the largest eigenvalue of the Hessian is strictly bounded by ρ_x, meaning $\lambda_{max}(H) \leq \rho_x$.

Substituting Equation 22 back into Equation 21 yields:

$$l(\vec{w}_{S,t}^{\phi}, \vec{x}_i + \vec{\delta}) \leq l(\vec{w}_{S,t}^{\phi}, \vec{z}_i) + \langle \nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i), \vec{\delta}\rangle + \frac{\rho_x}{2}\|\vec{\delta}\|_2^2 + \frac{\rho}{6}\|\vec{\delta}\|_2^3 \quad (23)$$

By definition, the loss function is strictly non-negative ($l \geq 0$), which implies $0 \leq l(\vec{w}_{S,t}^{\phi}, \vec{x}_i + \vec{\delta})$ must hold for any valid perturbation $\vec{\delta}$. Substituting zero into the left-hand side of Eq. 23 yields:

$$0 \leq l(\vec{w}_{S,t}^{\phi}, \vec{z}_i) + \langle \nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i), \vec{\delta}\rangle + \frac{\rho_x}{2}\|\vec{\delta}\|_2^2 + \frac{\rho}{6}\|\vec{\delta}\|_2^3 \quad (24)$$

Because Inequality 24 holds universally for any arbitrary perturbation, we can construct a specific adversarial perturbation $\vec{\delta}$ that moves a distance $\epsilon > 0$ in the opposite direction of the gradient with respect to the input space:

$$\vec{\delta} = -\epsilon \frac{\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)}{\|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)\|_2} \quad (25)$$

Note that the L_2 norm of this specific perturbation is exactly $\|\vec{\delta}\|_2 = \epsilon$. Substituting the structured $\vec{\delta}$ from Eq. 25 back into the inner product term of Eq. 24 yields:

$$0 \leq l(\vec{w}_{S,t}^{\phi}, \vec{z}_i) - \epsilon\|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)\|_2 + \frac{\rho_x}{2}\epsilon^2 + \frac{\rho}{6}\epsilon^3 \quad (26)$$

Rearranging the terms to isolate the input gradient magnitude on the left side of the inequality provides a per-iteration bound:

$$\epsilon\|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)\|_2 \leq l(\vec{w}_{S,t}^{\phi}, \vec{z}_i) + \frac{\rho_x}{2}\epsilon^2 + \frac{\rho}{6}\epsilon^3 \quad (27)$$

To capture the optimization dynamics over the full trajectory, we take the expectation $\mathbb{E}_{t\sim\mathbf{T}}[\cdot]$ over the randomly stopped iterations on both sides of Eq. 27. Here, $\mathbf{T}$ represents the non-uniform sampling distribution required for randomized SGD convergence:

$$\epsilon\mathbb{E}_{t\sim\mathbf{T}}\left[\|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)\|_2\right] \leq \mathbb{E}_{t\sim\mathbf{T}}\left[l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)\right] + \frac{\rho_x}{2}\epsilon^2 + \frac{\rho}{6}\epsilon^3 \quad (28)$$

Finally, we apply our formal trajectory definitions from Section 5.1. We substitute the expectations with the cumulative trajectory metrics ($\mathbb{E}_{t\sim\mathbf{T}}[l(\vec{w}^{\phi}_{S,t}, \vec{z}_i)] = \frac{CSL(\vec{z}_i)}{T_{max}}$ and $\mathbb{E}_{t\sim\mathbf{T}}[\|\nabla_{\vec{x}_i} l(\vec{w}^{\phi}_{S,t}, \vec{z}_i)\|_2] = \frac{CSG(\vec{z}_i)}{T_{max}}$):

$$\epsilon \frac{CSG(\vec{z}_i)}{T_{max}} \leq \frac{CSL(\vec{z}_i)}{T_{max}} + \frac{\rho_x}{2}\epsilon^2 + \frac{\rho}{6}\epsilon^3 \tag{29}$$

This concludes the proof, formally establishing that the accumulated gradient norm CSG is bounded by the accumulated loss CSL. □

B.2 Proof of Theorem 2

B.2.1 Preliminary: Necessary and Sufficient Thresholds for Misclassification

Before deriving the two-sided adversarial distance bounds, we must formally establish the critical thresholds of misclassification. We define L_c as the absolute minimum loss necessary to cross the decision boundary, and L_u as the loss sufficient to mathematically guarantee misclassification.

Lemma 1. *For a multi-class classification task over K classes utilizing Cross-Entropy (CE) loss, the necessary threshold for misclassification is $L_c = \ln(2)$, and the sufficient threshold for misclassification is $L_u = \ln(K) + \tau$, for some $\tau > 0$.*

Proof. Let $\vec{g}(\vec{x}) \in \mathbb{R}^K$ represent the continuous prediction scores (logits) of the model. The softmax probability assigned to class k is $p_k = \frac{\exp(\vec{g}_k)}{\sum_{j=1}^{K}\exp(\vec{g}_j)}$. The CE loss for a sample with true class label y is computed as $l = -\ln(p_y)$.

Necessary Condition (L_c): To establish the lower bound on adversarial distance, we must find the absolute minimum loss that exists anywhere on the decision boundary. For an untargeted misclassification to occur, the predicted probability of the true class p_y must be less than or equal to the predicted probability of at least one incorrect class $k \neq y$. Therefore, misclassification strictly requires $\exists k \neq y$ such that $p_k \geq p_y$.

Because the output probabilities must form a valid distribution ($\sum_{j=1}^{K} p_j = 1$), the sum of the true class probability and the most competitive adversarial class probability is bounded by 1:

$$p_y + p_k \leq 1 \tag{30}$$

Substituting the boundary crossing condition $p_k \geq p_y$ into this probability bound yields the maximum possible value p_y can hold while still being misclassified:

$$p_y + p_y \leq 1 \implies 2p_y \leq 1 \implies p_y \leq 0.5 \tag{31}$$

Since the CE loss is a monotonically decreasing function of p_y, evaluating this maximum probability establishes the strict lower bound on the misclassified loss:

$$l \geq -\ln(0.5) = \ln(2) \tag{32}$$

Thus, any perturbation $\vec{\delta}$ that successfully causes misclassification must push the loss to at least the necessary threshold $L_c = \ln(2)$.

Sufficient Condition (L_u): Conversely, to establish an upper bound on adversarial distance, we require a condition that guarantees misclassification regardless of the attack direction. If the loss strictly exceeds $\ln(K)$, then $p_y < \frac{1}{K}$. By the Pigeonhole Principle, since $\sum_{j=1}^{K} p_j = 1$, at least one other class must hold $p_k \geq \frac{1}{K} > p_y$. Thus, pushing the loss to $L_u = \ln(K) + \tau$, where $\tau > 0$ ensures no ties is a strictly sufficient condition to guarantee misclassification. □

B.2.2 Proof of the Bounds

With the necessary (L_c) and sufficient (L_u) thresholds established, we proceed to prove Theorem 2 by bounding the learning dynamics over the optimization trajectory.

Proof. **Part 1: The Lower Bounds**

We relate the loss at an original input $\vec{x}_i$ to a perturbed input $\vec{x}_i + \vec{\delta}_t$ at a specific optimization iteration t. Assuming the Hessian of the loss is ρ-Lipschitz (as in Lemma A.3 of Ravikumar et al. (2024)) which is adapted from Nesterov & Polyak (2006) we get:

$$-\frac{\rho}{6}\|\vec{\delta}_t\|_2^3 \leq l(\vec{w}_{S,t}^{\phi}, \vec{z}_i + \vec{\delta}_t) - l(\vec{w}_{S,t}^{\phi}, \vec{z}_i) - \langle \nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i), \vec{\delta}_t \rangle - \frac{1}{2}\vec{\delta}_t^T \nabla_{\vec{x}_i}^2 l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)\vec{\delta}_t \leq \frac{\rho}{6}\|\vec{\delta}_t\|_2^3 \tag{33}$$

From this, we extract the upper bound on the difference in loss:

$$l(\vec{w}_{S,t}^{\phi}, \vec{z}_i + \vec{\delta}_t) - l(\vec{w}_{S,t}^{\phi}, \vec{z}_i) \leq \langle \nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i), \vec{\delta}_t \rangle + \frac{1}{2}\vec{\delta}_t^T \nabla_{\vec{x}_i}^2 l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)\vec{\delta}_t + \frac{\rho}{6}\|\vec{\delta}_t\|_2^3 \tag{34}$$

By definition, the adversarial distance at iteration t, denoted $\delta_t = \|\vec{\delta}_t\|_2$, is the minimum perturbation that successfully causes misclassification. Based on Lemma 1, this mathematically requires the perturbed loss to exceed the necessary threshold:

$$l(\vec{w}_{S,t}^{\phi}, \vec{z}_i + \vec{\delta}_t) \geq L_c \tag{35}$$

We analyze the terms in the upper bound of Equation 34. Applying the Cauchy-Schwarz inequality to the first-order term yields:

$$\langle \nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i), \vec{\delta}_t \rangle \leq \|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)\|_2 \|\vec{\delta}_t\|_2 \tag{36}$$

For the second-order term, let $H = \nabla_{\vec{x}_i}^2 l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)$. Since the Hessian H is a real and symmetric matrix, Rayleigh's theorem states that:

$$\vec{\delta}_t^T H \vec{\delta}_t \leq \lambda_{max}(H)\|\vec{\delta}_t\|_2^2 \tag{37}$$

Under the assumption of ρ_x-Lipschitz input gradients as described in Equation 6, the largest eigenvalue of the Hessian is strictly bounded by ρ_x, meaning $\lambda_{max}(H) \leq \rho_x$.

Substituting these components and Equation 35 back into Equation 34 yields the guaranteed minimum distance relationship strictly for iteration t:

$$L_c - l(\vec{w}_{S,t}^{\phi}, \vec{z}_i) \leq \|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)\|_2 \delta_t + \frac{\rho_x}{2}\delta_t^2 + \frac{\rho}{6}\delta_t^3 \tag{38}$$

To simplify the analysis of this monotonic cubic function, we define the following strictly positive geometric constants for iteration t:

$$\Delta_t = L_c - l(\vec{w}_{S,t}^{\phi}, \vec{z}_i) \tag{39}$$

$$C_t = \|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)\|_2 \tag{40}$$

$$B = \frac{\rho_x}{2} \tag{41}$$

$$A = \frac{\rho}{6} \tag{42}$$

The inequality becomes a monotonic cubic function of δ_t since $A, B, C_t > 0$:

$$\Delta_t \leq A\delta_t^3 + B\delta_t^2 + C_t\delta_t \tag{43}$$

We seek to find the smallest valid perturbation magnitude δ_t^* that satisfies this necessary condition, since the true iteration adversarial distance $\delta_t \geq \delta_t^*$. The exact solution lies at the root $\Delta_t = A(\delta_t^*)^3 + B(\delta_t^*)^2 + C_t\delta_t^*$. From this relation, we can establish the linear bound $C_t\delta_t^* \leq \Delta_t \implies \delta_t^* \leq \frac{\Delta_t}{C_t}$.

We substitute this linear limit for one of the δ_t^* terms in the cubic component $\left(A(\delta_t^*)^3 = A\delta_t^*(\delta_t^*)^2\right)$ to safely reduce the order of the polynomial without violating the inequality:

$$\Delta_t \leq A\left(\frac{\Delta_t}{C_t}\right)(\delta_t^*)^2 + B(\delta_t^*)^2 + C_t\delta_t^* \tag{44}$$

$$\Delta_t \leq \left(A\frac{\Delta_t}{C_t} + B\right)(\delta_t^*)^2 + C_t\delta_t^* \tag{45}$$

Letting $\tilde{B}_t = A\frac{\Delta_t}{C_t} + B$, we yield the quadratic inequality:

$$\tilde{B}_t(\delta_t^*)^2 + C_t\delta_t^* - \Delta_t \geq 0 \tag{46}$$

Solving for the strictly positive root provides:

$$\delta_t^* = \frac{-C_t + \sqrt{C_t^2 + 4\tilde{B}_t\Delta_t}}{2\tilde{B}_t} \tag{47}$$

To improve theoretical interpretability, we multiply the numerator and denominator by the conjugate $(C_t + \sqrt{C_t^2 + 4\tilde{B}_t\Delta_t})$:

$$\delta_t \geq \frac{2\Delta_t}{\sqrt{C_t^2 + 4\tilde{B}_t\Delta_t} + C_t} \tag{48}$$

Because the adversarial distance at iteration t must be at least this exact minimum magnitude ($\delta_t \geq \delta_t^*$), this establishes the strict geometric lower bound for a single iteration.

Trajectory Maximum Lower Bound

To capture the learning dynamics, we evaluate the model at a randomly stopped iteration $t \sim \mathbf{T}$ drawn from the training process. Taking the expectation $\mathbb{E}_{t\sim\mathbf{T}}[\cdot]$ of both sides of Equation 38 provides:

$$\mathbb{E}_{t\sim\mathbf{T}}[L_c - l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)] \leq \mathbb{E}_{t\sim\mathbf{T}}[\|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)\|_2\delta_t] + \frac{\rho_x}{2}\mathbb{E}_{t\sim\mathbf{T}}[\delta_t^2] + \frac{\rho}{6}\mathbb{E}_{t\sim\mathbf{T}}[\delta_t^3] \tag{49}$$

We define $\delta_{max} = \max_{t\in\mathbf{T}} \delta_t$ as the maximum adversarial distance across all iterations. Because $\delta_t \leq \delta_{max}$ for all t, we can substitute δ_{max} into the right-hand side to establish a strict upper bound on the expected terms, allowing us to pull the perturbation out of the expectations:

$$L_c - \mathbb{E}_{t\sim\mathbf{T}}[l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)] \leq \delta_{max}\mathbb{E}_{t\sim\mathbf{T}}[\|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)\|_2] + \frac{\rho_x}{2}\delta_{max}^2 + \frac{\rho}{6}\delta_{max}^3 \tag{50}$$

We now substitute our previously established definitions for Cumulative Sample Loss (CSL) and Cumulative Sample Gradient (CSG):

$$L_c - \frac{CSL(\vec{z}_i)}{T_{max}} \leq \frac{CSG(\vec{z}_i)}{T_{max}}\delta_{max} + \frac{\rho_x}{2}\delta_{max}^2 + \frac{\rho}{6}\delta_{max}^3 \tag{51}$$

Notice that this expected inequality shares the exact same monotonic cubic structure as Equation 43. By defining the expected constants $\Delta_{min} = L_c - \frac{CSL(\vec{z}_i)}{T_{max}}$ and $C = \frac{CSG(\vec{z}_i)}{T_{max}}$, we bypass repeating the polynomial reduction process. Following the exact same logical steps that produced Equation 48, we directly obtain the lower bound for the maximum perturbation:

$$\delta_{max} \geq \frac{2\Delta_{min}}{\sqrt{C^2 + 4\tilde{B}\Delta_{min}} + C} \tag{52}$$

where $\tilde{B} = \frac{\rho_x}{2} + \frac{\rho}{6}\left(\frac{\Delta_{min}}{C}\right)$.

Part 2: The Upper Bounds

By the standard descent lemma for ρ_x-smooth functions (derived from Equation 6), the change in loss for an arbitrary perturbation $\vec{\delta}_t$ at iteration t is strictly lower-bounded by the first-order Taylor expansion minus the maximum penalty from the curvature:

$$l(\vec{w}_{S,t}^{\phi}, \vec{x}_i + \vec{\delta}_t) - l(\vec{w}_{S,t}^{\phi}, \vec{x}_i) \geq \langle \nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i), \vec{\delta}_t \rangle - \frac{\rho_x}{2}\|\vec{\delta}_t\|_2^2 \tag{53}$$

We act as the adversary and deploy a specific perturbation vector $\vec{d}_t$ aligned perfectly with the input gradient to maximize the first-order growth: $\vec{d}_t = d_t \frac{\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)}{\|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\|_2}$, where $d_t = \|\vec{d}_t\|_2$. Substituting this specific direction yields the guaranteed minimum growth strictly for iteration t:

$$l(\vec{w}_{S,t}^{\phi}, \vec{x}_i + \vec{d}_t) \geq \|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\|_2 d_t + l(\vec{w}_{S,t}^{\phi}, \vec{x}_i) - \frac{\rho_x}{2}d_t^2 \tag{54}$$

Single Iteration Upper Bound

To mathematically guarantee misclassification at this specific iteration t, we must find a magnitude d_t such that the guaranteed lower bound on the perturbed loss reaches the sufficient threshold L_u. Setting the right-hand side of Equation 54 exactly to L_u, we define the required margin at iteration t as $\Delta_t^{(u)} = L_u - l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)$ and the local gradient norm as $C_t = \|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\|_2$. This establishes the equality:

$$\Delta_t^{(u)} = C_t d_t - \frac{\rho_x}{2} d_t^2 \implies \frac{\rho_x}{2} d_t^2 - C_t d_t + \Delta_t^{(u)} = 0 \tag{55}$$

Assuming the local gradient dominates the curvature ($C_t^2 \geq 2\rho_x \Delta_t^{(u)}$), we solve for the smallest positive root d_t^*. Multiplying the numerator and denominator by the conjugate yields the required perturbation magnitude:

$$d_t^* = \frac{2\Delta_t^{(u)}}{C_t + \sqrt{C_t^2 - 2\rho_x \Delta_t^{(u)}}} \tag{56}$$

Because our constructed vector $\vec{d}_t$ mathematically guarantees misclassification at a distance of d_t^*, the true adversarial distance δ_t (which represents the absolute minimum perturbation required to cross the boundary in the optimal direction) must be less than or equal to this working radius. Substituting back the full definitions, the true adversarial distance at iteration t is bounded by:

$$\delta_t \leq \frac{2\left(L_u - l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\right)}{\|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\|_2 + \sqrt{\|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\|_2^2 - 2\rho_x \left(L_u - l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\right)}} \tag{57}$$

Trajectory Minimum Upper Bound

To establish a trajectory-level bound using learning dynamics, we start from a simpler, more interpretable form of the single-iteration upper bound. Because the square root term in the denominator is non-negative, dropping it strictly decreases the denominator and thereby safely yields a looser, purely linear geometric proxy for the upper bound:

$$\delta_t \leq \frac{2\left(L_u - l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\right)}{\|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\|_2} \tag{58}$$

Because the gradient norm is strictly non-negative, we can safely multiply both sides by the denominator to isolate the loss margin:

$$\|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\|_2 \delta_t \leq 2\left(L_u - l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\right) \tag{59}$$

Taking the expectation $\mathbb{E}_{t\sim\mathbf{T}}$ of both sides over the randomly stopped optimization trajectory yields:

$$\mathbb{E}_{t\sim\mathbf{T}}[\|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\|_2 \delta_t] \leq 2\left(L_u - \mathbb{E}_{t\sim\mathbf{T}}[l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)]\right) \tag{60}$$

We define $\delta_{min} = \min_{t\in\mathbf{T}} \delta_t$ as the absolute minimum adversarial distance across the entire training trajectory, and assuming that $C_t^2 \geq 2\rho_x \Delta_t^{(u)}$ is true across all t. Because $\delta_t \geq \delta_{min}$ for all iterations t, and the gradient norm is always non-negative ($\|\nabla_{\vec{x}_i} l\|_2 \geq 0$), substituting δ_{min} provides a strict lower bound for the term inside the expectation:

$$\|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\|_2 \delta_t \geq \|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\|_2 \delta_{min} \tag{61}$$

Taking the expectation of this lower bound and chaining it with Equation 60 allows us to safely pull δ_{min} out of the expectation product:

$$\delta_{min}\mathbb{E}_{t\sim\mathbf{T}}[\|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\|_2] \leq \mathbb{E}_{t\sim\mathbf{T}}[\|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\|_2 \delta_t] \leq 2\left(L_u - \mathbb{E}_{t\sim\mathbf{T}}[l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)]\right) \tag{62}$$

We can now substitute the formal definitions of Cumulative Sample Gradient (CSG) and Cumulative Sample Loss (CSL):

$$\delta_{min} \frac{CSG(\vec{z}_i)}{T_{max}} \leq 2 \left(L_u - \frac{CSL(\vec{z}_i)}{T_{max}} \right) \tag{63}$$

By rearranging to isolate δ_{min}, we arrive at the final trajectory-driven upper bound. This strictly limits the minimum adversarial vulnerability encountered during training to a zero-overhead function of the sample's learning dynamics:

$$\delta_{min} \leq \frac{2 \left(L_u T_{max} - CSL(\vec{z}_i) \right)}{CSG(\vec{z}_i)} \tag{64}$$

□

B.3 Proof of Theorem 3: Related CSL and MIA

B.3.1 Dataset Definitions and Goal

To evaluate the vulnerability of the model to Membership Inference Attacks (MIA), we define our training sets. Let S be a dataset of size $m-1$. We define two adjacent datasets of size m:

- $S_i = S \cup \{z_i\}$, where z_i is the target sample (the "in" sample).
- $S_j = S \cup \{z_j\}$, where $z_j \sim \mathcal{D}$ is a sample drawn from the data distribution (the "out" sample).

B.3.2 Preliminary Generalization and Risk Bounds

Before deriving the main inequality, we establish several preliminary bounds based on the model's generalization capability and uniform model bias at epoch t. Expectation over ϕ is implied for the equations.

We use Lemma A.1 of Ravikumar et al. (2024) and Assumption 2 to bound the generalization gap over the algorithm's randomness ϕ:

$$\mathbb{E}_\phi |R_{emp}(\vec{w}^\phi_{S_i,t}, S) - R(\vec{w}^\phi_{S_i,t})| \leq \gamma \tag{65}$$

$$\mathbb{E}_\phi |R_{emp}(\vec{w}^\phi_{S_i,t}, S_i) - R(\vec{w}^\phi_{S_i,t})| \leq \gamma \tag{66}$$

$$\mathbb{E}_\phi |R_{emp}(\vec{w}^\phi_{S_j,t}, S_j) - R(\vec{w}^\phi_{S_j,t})| \leq \gamma \tag{67}$$

Additionally, we use Assumption 1 on the dataset$S$:

$$\mathbb{E}_\phi |R(\vec{w}^\phi_{S_i,t}) - R(\vec{w}^\phi_{S,t})| \leq \beta \tag{68}$$

$$\mathbb{E}_\phi |R(\vec{w}^\phi_{S_j,t}) - R(\vec{w}^\phi_{S,t})| \leq \beta \tag{69}$$

Using Equation 68 and Equation 69, we can directly relate the population risks of the models trained on S_i and S_j:

$$R(\vec{w}^\phi_{S_j,t}) \leq R(\vec{w}^\phi_{S,t}) + \beta \leq R(\vec{w}^\phi_{S_i,t}) + 2\beta \tag{70}$$

Next, we establish the algebraic expansions for the empirical risks over datasets S_i and S_j by decomposing them into the risk over the base dataset S and the individual risk of the differing sample:

$$R_{emp}(\vec{w}^\phi_{S_i,t}, S_i) = \frac{m-1}{m} R_{emp}(\vec{w}^\phi_{S_i,t}, S) + \frac{1}{m} l(\vec{w}^\phi_{S_i,t}, z_i) \tag{71}$$

$$R_{emp}(\vec{w}^\phi_{S_j,t}, S_i) = \frac{m-1}{m} R_{emp}(\vec{w}^\phi_{S_j,t}, S) + \frac{1}{m} l(\vec{w}^\phi_{S_j,t}, z_i) \tag{72}$$

Utilizing the decomposition in Equation 71, we derive a critical lower bound for the empirical risk on the base dataset. From the generalization limits established in Equation 65, we know:

$$R_{emp}(\vec{w}^\phi_{S_i,t}, S) \geq R(\vec{w}^\phi_{S_i,t}) - \gamma \tag{73}$$

Further expanding the population risk utilizing Equation 66, we have:

$$R(\vec{w}^{\phi}_{S_i,t}) \geq R_{emp}(\vec{w}^{\phi}_{S_i,t}, S_i) - \gamma \tag{74}$$

Combining these expansions yields the lower bound:

$$\frac{m-1}{m} R_{emp}(\vec{w}^{\phi}_{S_i,t}, S) \geq \frac{m-1}{m} \left(R_{emp}(\vec{w}^{\phi}_{S_i,t}, S_i) - 2\gamma \right) \tag{75}$$

Finally, we upper-bound the discrepancy between the population risk and the empirical risk evaluated on S_i by substituting Equation 72 and sequentially applying the bounds from Equation 65 and the triangle inequality in Equation 70:

$$\begin{aligned}
R(\vec{w}^{\phi}_{S_j,t}) - R_{emp}(\vec{w}^{\phi}_{S_j,t}, S_i) &= R(\vec{w}^{\phi}_{S_j,t}) - \frac{m-1}{m} R_{emp}(\vec{w}^{\phi}_{S_j,t}, S) - \frac{1}{m} l(\vec{w}^{\phi}_{S_j,t}, z_i) \\
&= \frac{m-1}{m} (R(\vec{w}^{\phi}_{S_j,t}) - R_{emp}(\vec{w}^{\phi}_{S_j,t}, S)) + \frac{1}{m} R(\vec{w}^{\phi}_{S_j,t}) - \frac{1}{m} l(\vec{w}^{\phi}_{S_j,t}, z_i) \\
&\leq \frac{m-1}{m}\gamma + \frac{R(\vec{w}^{\phi}_{S_i,t}) + 2\beta}{m} - \frac{1}{m} l(\vec{w}^{\phi}_{S_j,t}, z_i) \\
&\leq \frac{m-1}{m}\gamma + \frac{2\beta}{m} + \frac{\gamma}{m} + \frac{1}{m} R_{emp}(\vec{w}^{\phi}_{S_i,t}, S_i) - \frac{1}{m} l(\vec{w}^{\phi}_{S_j,t}, z_i) \quad (76) \\
&\leq \gamma + \frac{2\beta}{m} + \frac{1}{m} R_{emp}(\vec{w}^{\phi}_{S_i,t}, S_i) - \frac{1}{m} l(\vec{w}^{\phi}_{S_j,t}, z_i) \quad (77)
\end{aligned}$$

B.3.3 Bounding the Empirical Risk Differences

With our preliminaries established, we form a continuous chain of four inequalities to rigorously bound the difference in empirical risks between training on S_i and S_j and evaluating on S_i:

$$R_{emp}(\vec{w}^{\phi}_{S_i,t}, S_i) - R(\vec{w}^{\phi}_{S_i,t}) \leq \gamma \tag{78}$$

$$R(\vec{w}^{\phi}_{S_i,t}) - R(\vec{w}^{\phi}_{S,t}) \leq \beta \tag{79}$$

$$R(\vec{w}^{\phi}_{S,t}) - R(\vec{w}^{\phi}_{S_j,t}) \leq \beta \tag{80}$$

$$R(\vec{w}^{\phi}_{S_j,t}) - R_{emp}(\vec{w}^{\phi}_{S_j,t}, S_i) \leq \gamma + \frac{2\beta}{m} + \frac{1}{m} R_{emp}(\vec{w}^{\phi}_{S_i,t}, S_i) - \frac{1}{m} l(\vec{w}^{\phi}_{S_j,t}, z_i) \tag{81}$$

Summing these four inequalities strictly collapses the intermediate population risks, yielding our main bounding equation for the empirical risk discrepancy:

$$R_{emp}(\vec{w}^{\phi}_{S_i,t}, S_i) - R_{emp}(\vec{w}^{\phi}_{S_j,t}, S_i) \leq 2\gamma + 2\beta + \frac{2\beta}{m} + \frac{1}{m} R_{emp}(\vec{w}^{\phi}_{S_i,t}, S_i) - \frac{1}{m} l(\vec{w}^{\phi}_{S_j,t}, z_i) \tag{82}$$

B.3.4 Isolating the Target Sample Loss

We now manipulate Equation 82 to algebraically isolate the loss of the target sample $\vec{z}_i$. First, we expand the empirical risk difference $R_{emp}(\vec{w}^{\phi}_{S_i,t}, S_i) - R_{emp}(\vec{w}^{\phi}_{S_j,t}, S_i)$ using the foundational decompositions from Equation 71 and Equation 72:

$$\begin{aligned}
R_{emp}(\vec{w}^{\phi}_{S_i,t}, S_i) - R_{emp}(\vec{w}^{\phi}_{S_j,t}, S_i) = &\left(\frac{m-1}{m} R_{emp}(\vec{w}^{\phi}_{S_i,t}, S) + \frac{1}{m} l(\vec{w}^{\phi}_{S_i,t}, \vec{z}_i) \right) \\
&- \left(\frac{m-1}{m} R_{emp}(\vec{w}^{\phi}_{S_j,t}, S) + \frac{1}{m} l(\vec{w}^{\phi}_{S_j,t}, \vec{z}_i) \right)
\end{aligned} \tag{83}$$

Next, we apply the lower bound derived in Equation 75 to this expansion, providing:

$$\begin{aligned}
R_{emp}(\vec{w}^{\phi}_{S_i,t}, S_i) - R_{emp}(\vec{w}^{\phi}_{S_j,t}, S_i) \geq &\frac{m-1}{m} R_{emp}(\vec{w}^{\phi}_{S_i,t}, S_i) - \frac{2(m-1)}{m}\gamma \\
&+ \frac{1}{m} l(\vec{w}^{\phi}_{S_i,t}, \vec{z}_i) - \frac{m-1}{m} R_{emp}(\vec{w}^{\phi}_{S_j,t}, S) - \frac{1}{m} l(\vec{w}^{\phi}_{S_j,t}, \vec{z}_i)
\end{aligned} \tag{84}$$

We substitute this strictly expanded lower bound back into the empirical risk difference defined in Equation 82. Rearranging the fractional R_{emp} terms consolidates them as follows:

$$\left(\frac{m-1}{m}-\frac{1}{m}\right)R_{emp}(\vec{w}^{\phi}_{S_i,t},S_i)=\frac{m-2}{m}R_{emp}(\vec{w}^{\phi}_{S_i,t},S_i) \tag{85}$$

Moving all γ and β stability constants to the opposing side of the inequality yields:

$$\frac{m-2}{m}R_{emp}(\vec{w}^{\phi}_{S_i,t},S_i)-\frac{m-1}{m}R_{emp}(\vec{w}^{\phi}_{S_j,t},S)\leq\frac{1}{m}\left(l(\vec{w}^{\phi}_{S_j,t},\vec{z}_i)-l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)\right)+K_{const} \tag{86}$$

where the combined generalization and stability constant is defined as:

$$K_{const}=2\gamma\left(\frac{2m-1}{m}\right)+2\beta\left(\frac{m+1}{m}\right) \tag{87}$$

We expand $R_{emp}(\vec{w}^{\phi}_{S_i,t},S_i)$ one final time using Equation 71:

$$\begin{aligned}\frac{(m-2)(m-1)}{m^2}R_{emp}(\vec{w}^{\phi}_{S_i,t},S)+\frac{m-2}{m^2}l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)-\frac{m-1}{m}R_{emp}(\vec{w}^{\phi}_{S_j,t},S)\\ \leq\frac{1}{m}[l(\vec{w}^{\phi}_{S_j,t},\vec{z}_i)-l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)]+K_{const}\end{aligned} \tag{88}$$

Isolating the target sample loss term $l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)$ provides:

$$\begin{aligned}\frac{m-2}{m^2}l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)\leq&\frac{m-1}{m}R_{emp}(\vec{w}^{\phi}_{S_j,t},S)-\frac{(m-2)(m-1)}{m^2}R_{emp}(\vec{w}^{\phi}_{S_i,t},S)\\ &+\frac{1}{m}[l(\vec{w}^{\phi}_{S_j,t},\vec{z}_i)-l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)]+K_{const}\end{aligned} \tag{89}$$

Applying the generalization limits directly to the empirical risk terms yields the upper bound:

$$\begin{aligned}\frac{m-2}{m^2}l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)\leq&\frac{m-1}{m}(\gamma+R(\vec{w}^{\phi}_{S_j,t}))+\frac{(m-2)(m-1)}{m^2}(\gamma-R(\vec{w}^{\phi}_{S_i,t}))\\ &+\frac{1}{m}[l(\vec{w}^{\phi}_{S_j,t},\vec{z}_i)-l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)]+K_{const}\end{aligned} \tag{90}$$

We now apply the Uniform Model Bound assumption 3, which states:

$$-\Delta\leq R(\vec{w})-R(\vec{h}^*)\leq\Delta \tag{91}$$

For a fixed dataset $\mathcal{D}$, $R(\vec{h}^*)$ is treated as R^*a stationary constant, and hence is written as . Substituting the explicit bounds:

$$R(\vec{w}^{\phi}_{S_j,t})\leq R^*+\Delta \tag{92}$$

$$-R(\vec{w}^{\phi}_{S_i,t})\leq\Delta-R^* \tag{93}$$

into our target equation yields:

$$\begin{aligned}\frac{m-2}{m^2}l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)\leq&\gamma\left(\frac{m-1}{m}+\frac{(m-2)(m-1)}{m^2}\right)\\ &+\frac{m-1}{m}(R^*+\Delta)+\frac{(m-2)(m-1)}{m^2}(\Delta-R^*)\\ &+\frac{1}{m}[l(\vec{w}^{\phi}_{S_j,t},\vec{z}_i)-l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)]+K_{const}\end{aligned} \tag{94}$$

We group all data-independent constants into a single constant K:

$$\begin{aligned}K=&2\gamma\left(\frac{2m-1}{m}\right)+2\beta\left(\frac{m+1}{m}\right)+(\Delta+\gamma)\left(\frac{m-1}{m}+\frac{(m-2)(m-1)}{m^2}\right)\\ &+R^*\left(\frac{m-1}{m}-\frac{(m-2)(m-1)}{m^2}\right)\end{aligned} \tag{95}$$

This strictly simplifies our main inequality to:

$$\frac{m-2}{m^2}l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)\leq\frac{1}{m}[l(\vec{w}^{\phi}_{S_j,t},\vec{z}_i)-l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)]+K \tag{96}$$

B.3.5 Relating to Membership Inference Attacks

Now clearly writing Equation 96 which had implicit Expectation notations with explicit notations where $\mathbb{E}_\phi$ is over the algorithm's randomness, and $\mathbb{E}_{S,\vec{z}_j}$ over the dataset sampling $S \sim \mathcal{D}^{m-1}$ and the out-sample $\vec{z}_j \sim \mathcal{D}$, we establish:

$$\frac{m-2}{m^2}\mathbb{E}_{\phi,S\sim\mathcal{D}^{m-1}}[l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)] \leq \frac{1}{m}\mathbb{E}_\phi\left[\mathbb{E}_{S,\vec{z}_j}[l(\vec{w}^{\phi}_{S_j,t},\vec{z}_i)] - \mathbb{E}_S[l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)]\right] + K \tag{97}$$

To relate this to MIA, we define a specific attack strategy for an L-bounded loss function as described in Theorem 2 of Yeom et al. (2018) that predicts a sample as a member ("in") with probability:

$$P(\text{in}) = 1 - \frac{l(\vec{w},\vec{z})}{L} \tag{98}$$

The expected advantage of this specific linear strategy evaluates directly to:

$$\begin{aligned} Adv(\vec{w}_t,\vec{z}_i) &= \mathbb{E}\left[\left(1 - \frac{l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)}{L}\right) - \left(1 - \frac{l(\vec{w}^{\phi}_{S_j,t},\vec{z}_i)}{L}\right)\right] \\ &= \frac{1}{L}\mathbb{E}\left[l(\vec{w}^{\phi}_{S_j,t},\vec{z}_i) - l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)\right] \end{aligned} \tag{99}$$

Since this represents merely one specific strategy, its advantage is upper-bounded by the optimal adversarial advantage which is the sup Advantage of all possible attack strategies (Choi et al., 2023):

$$Adv(\vec{w}_t,\vec{z}_i) \leq Adv^*(\vec{w}_t,\vec{z}_i) \tag{100}$$

Substituting the term $L \cdot Adv^*(\vec{w}_t,\vec{z}_i)$ back into Equation 97 and taking the expectation $\mathbb{E}_{t\sim\mathbf{T}}$ over the randomized SGD stopping iterations yields:

$$\frac{m-2}{m^2}\mathbb{E}_{\phi,S\sim\mathcal{D}^{m-1}}\left[\mathbb{E}_{t\sim\mathbf{T}}[l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)]\right] \leq \frac{L}{m}\mathbb{E}_{t\sim\mathbf{T}}[Adv^*(\vec{w}_t,\vec{z}_i)] + \mathbb{E}_{t\sim\mathbf{T}}[K] \tag{101}$$

Let κ be defined as the maximum K across all t. And by definition, the Cumulative Sample Loss is:

$$CSL(\vec{z}_i) = T_{max}\mathbb{E}_{t\sim\mathbf{T}}[l(\vec{w}^{\phi}_{S_i,t},\vec{z}_i)] \tag{102}$$

We substitute CSL into the final bound:

$$\frac{m-2}{m^2}\mathbb{E}_{\phi,S\sim\mathcal{D}^{m-1}}\left[\frac{CSL(\vec{z}_i)}{T_{max}}\right] \leq \frac{L}{m}\mathbb{E}_{t\sim\mathbf{T}}[Adv^*(\vec{w}_t,\vec{z}_i)] + \kappa \tag{103}$$

B.4 Proof of Corollary 1: Unified Vulnerability Bounds

B.4.1 Preliminaries: Memorization Bound of Ravikumar et al. (2025)

For completeness, we restate the theoretical connection between sample memorization and Cumulative Sample Loss for strictly non-negative losses, as derived in Theorem 5.5 of Ravikumar et al. (2025):

Theorem 4 (Ravikumar et al. (2025))**.** *Assume the loss function is strictly non-negative and the algorithm satisfies error stability β. For a target sample $\vec{z}_i$, let ξ denote a constant offset. The memorization score $mem(\vec{z}_i)$, evaluated over the algorithm's intrinsic randomness ϕ, is bounded by the expected Cumulative Sample Loss (CSL):*

$$\mathbb{E}_\phi[mem(\vec{z}_i)] \leq \mathbb{E}_\phi[CSL(\vec{z}_i)] + \beta - \xi \tag{104}$$

Rearranging Equation 104 establishes an explicit lower bound on the expected CSL, which we define as M_{lb}:

$$\mathbb{E}_\phi[CSL(\vec{z}_i)] \geq M_{lb} \triangleq \mathbb{E}_\phi[mem(\vec{z}_i)] - \beta + \xi \tag{105}$$

B.4.2 Derivation of the Memorization Upper Bound (Non-negative Loss)

From the derivation of the trajectory upper bound in Theorem 2, we established the linear geometric proxy for the adversarial distance δ_t at any iteration t:

$$\|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\|_2 \delta_t \leq 2\left(L_u - l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\right) \tag{106}$$

We define $\delta_{\inf} = \inf_{\phi} \min_{t \in \mathbf{T}} \delta_t$ as the absolute minimum adversarial distance encountered across the entire optimization trajectory, universally across all intrinsic stochasticity ϕ, and assuming $C_t^2 \geq 2\rho_x \Delta_t^{(u)}$ is true for all t and ϕ. Because $\delta_t \geq \delta_{\inf}$ for all iterations and algorithmic randomness, and the gradient norm is strictly non-negative, substituting $\delta_{\inf}$ provides a strict deterministic lower bound for the left-hand side:

$$\delta_{\inf} \|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\|_2 \leq \|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\|_2 \delta_t \leq 2\left(L_u - l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\right) \tag{107}$$

Because $\delta_{\inf}$ is a deterministic constant, taking the expectation $\mathbb{E}_{\phi,\mathbf{T}}$ over both the algorithm's intrinsic randomness and the optimization trajectory safely factors it out:

$$\delta_{\inf} \mathbb{E}_{\phi,\mathbf{T}}[\|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)\|_2] \leq 2\left(L_u - \mathbb{E}_{\phi,\mathbf{T}}[l(\vec{w}_{S,t}^{\phi}, \vec{x}_i)]\right) \tag{108}$$

Substituting the formal definitions of Expected Cumulative Sample Gradient ($\mathbb{E}_{\phi}[CSG]$) and Cumulative Sample Loss ($\mathbb{E}_{\phi}[CSL]$) yields:

$$\delta_{\inf} \frac{\mathbb{E}_{\phi}[CSG(\vec{z}_i)]}{T_{max}} \leq 2\left(L_u - \frac{\mathbb{E}_{\phi}[CSL(\vec{z}_i)]}{T_{max}}\right) \tag{109}$$

By rearranging to isolate $\delta_{\inf}$, we obtain:

$$\delta_{\inf} \leq \frac{2\left(L_u T_{max} - \mathbb{E}_{\phi}[CSL(\vec{z}_i)]\right)}{\mathbb{E}_{\phi}[CSG(\vec{z}_i)]} \tag{110}$$

From the memorization framework of Ravikumar et al. (2025), we established the lower bound $\mathbb{E}_{\phi}[CSL(\vec{z}_i)] \geq M_{lb} \triangleq \mathbb{E}_{\phi}[\text{mem}(\vec{z}_i)] - \beta + \xi$. Substituting M_{lb} preserves the downward inequality because replacing a subtracted term with its strict lower bound strictly increases the numerator:

$$\delta_{\inf} \leq \frac{2\left(L_u T_{max} - M_{lb}\right)}{\mathbb{E}_{\phi}[CSG(\vec{z}_i)]} \tag{111}$$

Substituting the definition of M_{lb} directly yields the final memorization upper bound defined in Corollary 1(i).

B.4.3 Derivation of the MIA Lower Bound (L-Bounded Loss)

The minimum distance constraint is governed by the monotonic cubic inequality derived from the necessary misclassification threshold $L_c = \ln(2)$ in Theorem 2:

$$L_c - l(\vec{w}_{S,t}^{\phi}, \vec{z}_i) \leq \|\nabla_{\vec{x}_i} l(\vec{w}_{S,t}^{\phi}, \vec{z}_i)\|_2 \delta_t + \frac{\rho_x}{2}\delta_t^2 + \frac{\rho}{6}\delta_t^3 \tag{112}$$

To avoid violating Jensen's Inequality when passing expectations through convex polynomial terms, we define a strict deterministic absolute maximum $\delta_{\sup} = \sup_{\phi} \max_{t \in \mathbf{T}} \delta_t = \sup_{\phi} \delta_{max}$. Because $\delta_t \leq \delta_{\sup}$ universally across all iterations and all algorithmic randomness, we strictly have $\delta_t^2 \leq \delta_{\sup}^2$ and $\delta_t^3 \leq \delta_{\sup}^3$.

Substituting $\delta_{\sup}$ bounds the right-hand side from above. Taking the expectation $\mathbb{E}_{\phi,\mathbf{T}}$ safely factors out the deterministic scalar powers:

$$L_c - \frac{\mathbb{E}_{\phi}[CSL(\vec{z}_i)]}{T_{max}} \leq \delta_{\sup} \frac{\mathbb{E}_{\phi}[CSG(\vec{z}_i)]}{T_{max}} + \frac{\rho_x}{2}\delta_{\sup}^2 + \frac{\rho}{6}\delta_{\sup}^3 \tag{113}$$

Theorem 3 establishes an upper bound on the expected loss via the optimal MIA advantage:

$$\frac{\mathbb{E}_\phi[CSL(\vec{z}_i)]}{T_{max}} \leq M_{ub} \triangleq \alpha \mathbb{E}_{t\sim\mathbf{T}}[Adv^*_{MIA}(\vec{w}_t, \vec{z}_i)] + \kappa \tag{114}$$

Substituting M_{ub} in place of the expected CSL strictly decreases the unperturbed margin on the left-hand side. Simultaneously, leveraging Theorem 1 and transitivity, substituting M_{ub} strictly increases the expected gradient bound on the right-hand side:

$$\frac{\mathbb{E}_\phi[CSG(\vec{z}_i)]}{T_{max}} \leq \frac{1}{\epsilon}\frac{\mathbb{E}_\phi[CSL(\vec{z}_i)]}{T_{max}} + K_\epsilon \leq \frac{1}{\epsilon}M_{ub} + K_\epsilon \tag{115}$$

Applying these relaxed expected limits constructs a universally valid generalized cubic inequality governed entirely by the expected MIA advantage:

$$L_c - M_{ub} \leq \left(\frac{1}{\epsilon}M_{ub} + K_\epsilon\right)\delta_{\text{sup}} + \frac{\rho_x}{2}\delta^2_{\text{sup}} + \frac{\rho}{6}\delta^3_{\text{sup}} \tag{116}$$

By defining the MIA-exclusive constants $\Delta_{mia} = L_c - M_{ub}$ and $C_{mia} = \frac{1}{\epsilon}M_{ub} + K_\epsilon$, we arrive at the standard monotonic form $\Delta_{mia} \leq C_{mia}\delta_{\text{sup}} + B\delta^2_{\text{sup}} + A\delta^3_{\text{sup}}$.

Isolating the linear term establishes a valid substitution limit $\delta_{\text{sup}} \leq \frac{\Delta_{mia}}{C_{mia}}$. Following the exact quadratic reduction detailed in Theorem 2, we substitute this linear limit into one degree of the cubic component $\left(A\delta^3_{\text{sup}} = A\delta_{\text{sup}}(\delta_{\text{sup}})^2\right)$ to safely reduce the polynomial order.

This yields the quadratic form $\tilde{B}_{mia}\delta^2_{\text{sup}} + C_{mia}\delta_{\text{sup}} - \Delta_{mia} \geq 0$. Solving for the strictly positive root yields the explicit lower bound $\delta_{\text{sup}} \geq \frac{2\Delta_{mia}}{\sqrt{C^2_{mia}+4\tilde{B}_{mia}\Delta_{mia}}+C_{mia}}$. This directly provides the lower bound on the worst-case safety margin defined in Corollary 1(ii), completing the proof.

C Experiments

C.1 Validation Experiment Details

Datasets and Splits: We empirically validate our theoretical bounds on the CIFAR-10 and CIFAR-100 (Krizhevsky, 2009) datasets. Both datasets consist of color images partitioned into 50,000 training examples and 10,000 test examples. For validating the learning dynamics and adversarial distance relationships, we evaluate models over the complete dataset of $N = 50,000$ training samples. For the membership privacy evaluations, we construct adjacent training splits of size $N = 25,000$ sampled without replacement from the training pool, reserving the remaining examples as the hold-out non-member pool.

Model Architectures: We evaluate our framework across three standard residual network depths: ResNet-18, ResNet-34, and ResNet-50 (He et al., 2016). All networks are trained from scratch using the standard Cross-Entropy loss which is supported by Theorem 2.

Training: Models are optimized using Stochastic Gradient Descent (SGD) with a momentum coefficient of 0.9, a mini-batch size of 128, and a weight decay penalty of 1×10^{-4}. Training runs for a total of 200 epochs. The learning rate is initialized at 0.1 and follows a multi-step decay schedule, dropping by a factor of 10 at epoch 100 to 0.01, and again at epoch 150 to 0.001. Data augmentation during training follows standard vision pipelines: a random horizontal flip and a random crop with 4 pixels of padding, followed by channel normalization. During inference, augmentations are disabled. All experiments are conducted across 3 independent random seeds to evaluate variance against algorithmic stochasticity.

Adversarial Distance Computation: The ground-truth adversarial distance at the final converged checkpoint is measured as the minimum l_2 perturbation required to cross the learned decision boundary. If a sample is already misclassified by the clean model, its adversarial distance is assigned a value of zero. For correctly classified samples, we employ an untargeted Carlini & Wagner (C&W)

l_2 optimization attack (Carlini & Wagner, 2016). To strictly enforce valid image bounding boxes without inducing gradient masking, inputs are first inverse-normalized to absolute pixel space $[0, 1]$, and the perturbation is optimized in the inverse hyperbolic tangent (arctanh) space. The perturbed image is parameterized as $x' = \frac{1}{2}(\tanh(w) + 1)$. The objective function minimizes the l_2 norm of the perturbation alongside a margin-based classification loss, balanced by a trade-off constant c. Optimization is executed using Adam with a learning rate of 0.01 for 200 iterations. To isolate the absolute minimal perturbation budget required to force a misclassification, we conduct a 5-step binary search over the constant c, initializing the search at $c = 0.01$.

Adversarial distance proxy: We use the trajectory based proxy $\bar{d}_{adv,T} = \frac{2(L_u T - CSL(\vec{z}_i))}{CSG(\vec{z}_i)}$, where we set $T = T_{max}$ to indicate that we evaluate this for the full trajectory. Since this proxy is only dependent on CSL and CSG scores of a sample, which require zero, and very negligible computational (one extra matrix multiplication per training epoch) overhead respectively, it can perform as a cheap and quick proxy for true adversarial distance.

Results: The table 4 lists the results of correlation metrics for all of the experiment configurations for comparing the relationship between CSL, CSG, adversarial distance, the proxy, and MIA advantages. Figure 4 to Figure 9 represent the results in form of grid plots. It is observed that increase in either of CSL, CSG or MIA advantage results in decrease of the sample's adversarial distance which aligns with our Theorem 2 and Corollary 1.

Membership Inference Attacks (MIA): To evaluate sample-level privacy leakage, we construct ground-truth MIA references using 64 independent shadow models trained on the random $N = 25,000$ half-splits. We audit the models using two complementary attack frameworks from the literature. First, we employ a loss thresholding approach for the Yeom loss threshold attack (Yeom et al., 2018), which infers membership by evaluating whether a sample's loss falls below a specific threshold compared to the expected loss distribution. Second, we utilize Carlini et al. (2022)'s state-of-the-art Likelihood Ratio Attack (LiRA) method, which models the sample-level distribution of log-transformed losses by fitting parametric Gaussian distributions over the shadow models' member and non-member evaluations to compute a highly calibrated, per-example membership score.

Evaluation Metrics and Binning: Alignment between learning dynamics proxies, true adversarial distance, and privacy advantages is evaluated using Spearman's rank correlation coefficient, Cosine Similarity and Pearson's coefficient. Because individual sample trajectories contain inherent optimization noise, we smooth these fluctuations for visualization by sorting the audited samples by the horizontal axis metric and partitioning them into 10 equal-frequency bins. We then compute and plot the mean metric values within each bin and shade the standard deviations, highlighting the underlying geometric trends predicted by our theoretical bounds.

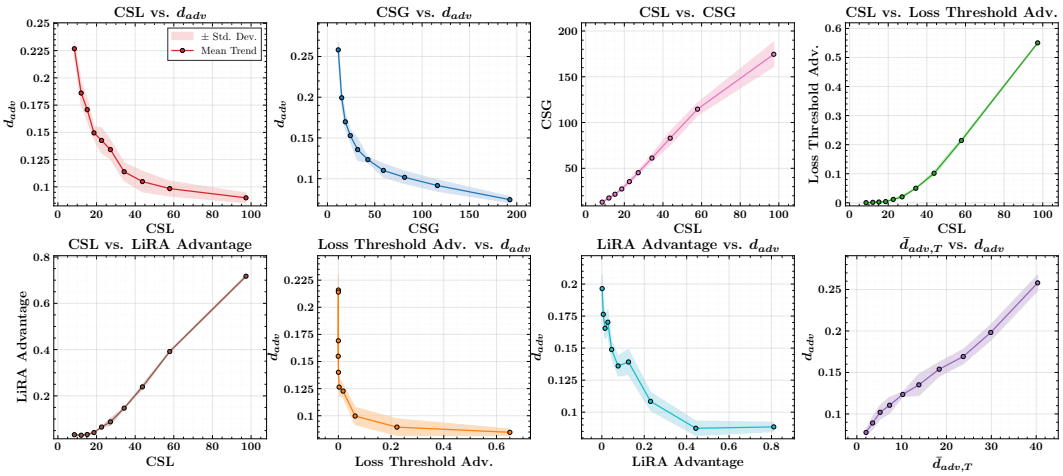


Figure 4: Metric evaluations for ResNet-18 on CIFAR-10. Shaded regions denote standard deviation across independent seeds (10 bins).

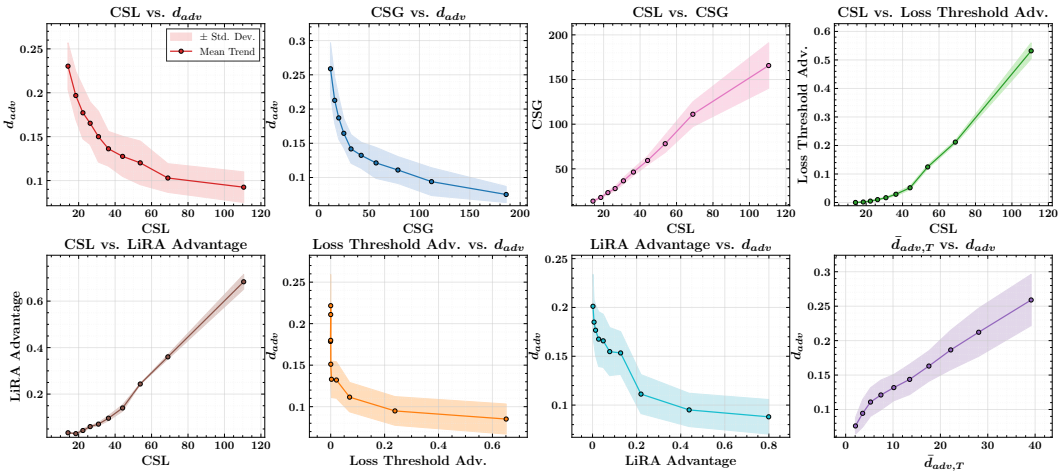


Figure 5: Metric evaluations for ResNet-34 on CIFAR-10. Shaded regions denote standard deviation across independent seeds (10 bins).

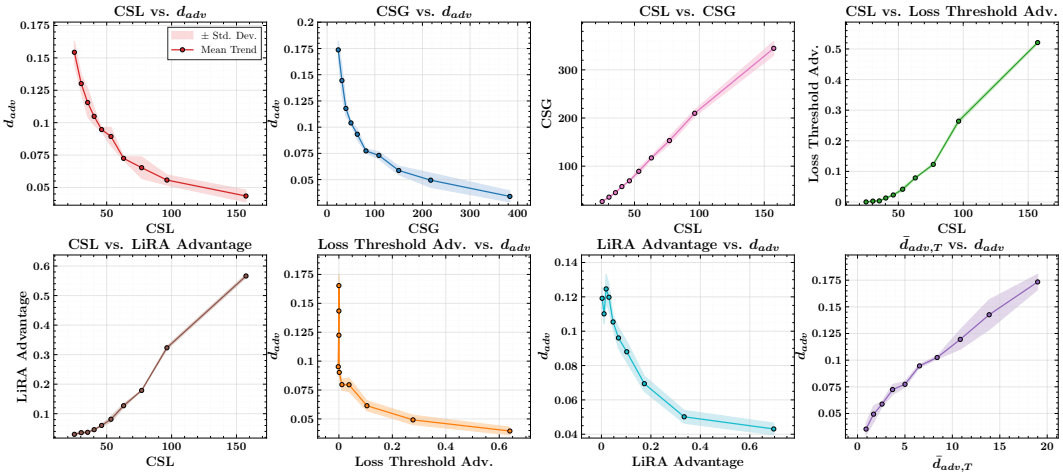


Figure 6: Metric evaluations for ResNet-50 on CIFAR-10. Evaluated on the non-divergent seed pair. Shaded regions denote standard deviation (10 bins).

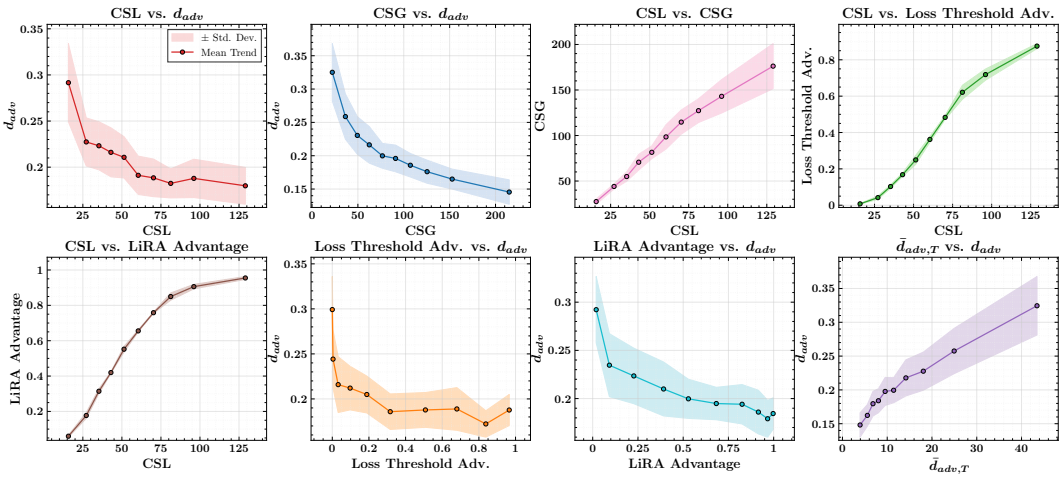


Figure 7: Metric evaluations for ResNet-18 on CIFAR-100. Shaded regions denote standard deviation across independent seeds (10 bins).

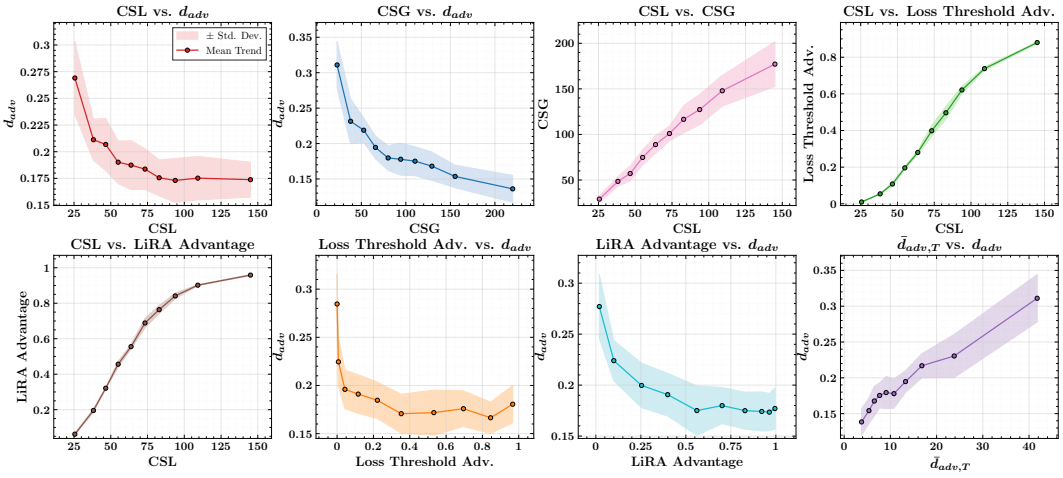


Figure 8: Metric evaluations for ResNet-34 on CIFAR-100. Shaded regions denote standard deviation across independent seeds (10 bins).

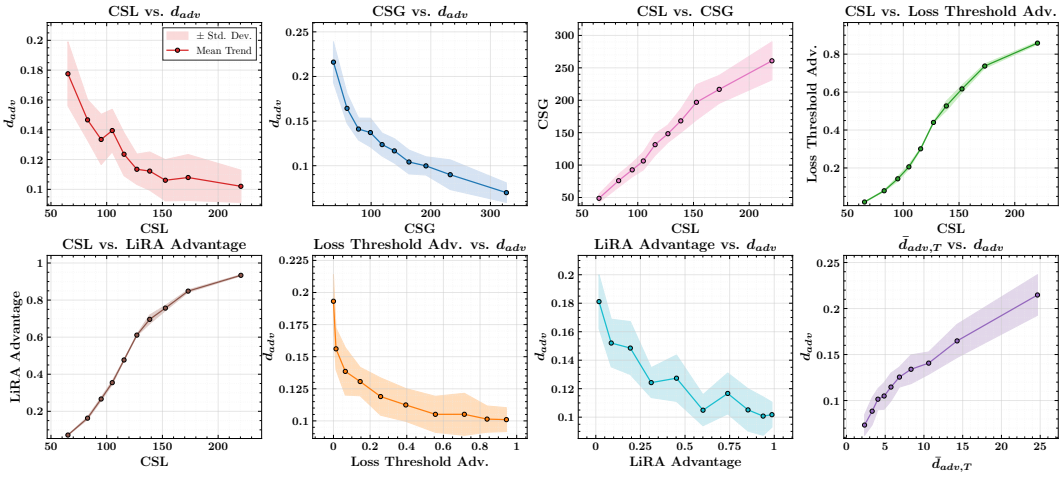


Figure 9: Metric evaluations for ResNet-50 on CIFAR-100. Shaded regions denote standard deviation across independent seeds (10 bins).

Table 4: Comprehensive alignment metrics across all model architectures and datasets. Evaluated over 3 independent seeds. Reporting Pearson correlation (r), Spearman rank correlation (ρ), and Cosine Similarity (Mean $\pm$ Std).

Dataset	Architecture	Metric Relationship	Pearson (r)	Spearman (ρ)	Cosine Similarity
CIFAR-10	ResNet-18	d_{adv} vs. $\bar{d}_{adv,T}$	0.735 ± 0.015	0.720 ± 0.007	0.906 ± 0.004
		CSL vs. CSG	0.858 ± 0.011	0.901 ± 0.007	0.935 ± 0.006
		CSL vs. LiRA Adv.	0.826 ± 0.006	0.705 ± 0.004	0.872 ± 0.009
		CSL vs. Yeom Adv.	0.797 ± 0.005	0.741 ± 0.009	0.781 ± 0.012
	ResNet-34	d_{adv} vs. $\bar{d}_{adv,T}$	0.698 ± 0.010	0.703 ± 0.016	0.898 ± 0.001
		CSL vs. CSG	0.810 ± 0.028	0.886 ± 0.008	0.914 ± 0.012
		CSL vs. LiRA Adv.	0.759 ± 0.040	0.666 ± 0.021	0.828 ± 0.013
		CSL vs. Yeom Adv.	0.739 ± 0.040	0.704 ± 0.022	0.735 ± 0.013
	ResNet-50	d_{adv} vs. $\bar{d}_{adv,T}$	0.735 ± 0.002	0.750 ± 0.017	0.909 ± 0.001
		CSL vs. CSG	0.795 ± 0.026	0.897 ± 0.007	0.905 ± 0.012
		CSL vs. LiRA Adv.	0.692 ± 0.069	0.659 ± 0.004	0.795 ± 0.032
		CSL vs. Yeom Adv.	0.698 ± 0.070	0.714 ± 0.006	0.737 ± 0.035
CIFAR-100	ResNet-18	d_{adv} vs. $\bar{d}_{adv,T}$	0.584 ± 0.024	0.507 ± 0.023	0.857 ± 0.007
		CSL vs. CSG	0.781 ± 0.009	0.830 ± 0.006	0.944 ± 0.001
		CSL vs. LiRA Adv.	0.789 ± 0.010	0.855 ± 0.016	0.945 ± 0.002
		CSL vs. Yeom Adv.	0.811 ± 0.019	0.855 ± 0.020	0.904 ± 0.007
	ResNet-34	d_{adv} vs. $\bar{d}_{adv,T}$	0.575 ± 0.025	0.480 ± 0.029	0.853 ± 0.005
		CSL vs. CSG	0.762 ± 0.011	0.812 ± 0.006	0.943 ± 0.002
		CSL vs. LiRA Adv.	0.782 ± 0.007	0.849 ± 0.009	0.946 ± 0.002
		CSL vs. Yeom Adv.	0.810 ± 0.009	0.852 ± 0.009	0.902 ± 0.004
	ResNet-50	d_{adv} vs. $\bar{d}_{adv,T}$	0.569 ± 0.008	0.558 ± 0.014	0.857 ± 0.003
		CSL vs. CSG	0.737 ± 0.018	0.793 ± 0.011	0.935 ± 0.007
		CSL vs. LiRA Adv.	0.780 ± 0.007	0.832 ± 0.009	0.930 ± 0.005
		CSL vs. Yeom Adv.	0.787 ± 0.008	0.831 ± 0.010	0.887 ± 0.010

C.2 Mislabeled Detection Setup Details

Setup. All experiments are conducted using standard ResNet-18 backbones trained on CIFAR-10 and CIFAR-100, following the setup of Ravikumar et al. (2026). Optimization is performed over 300 epochs using Stochastic Gradient Descent (SGD) with an initial learning rate $\eta = 0.1$, decayed by a factor of 10 at epochs 180 and 240. We employ a mini-batch size of 128, a momentum parameter of 0.9, and a weight decay coefficient of 1×10^{-4}. Symmetric label noise is synthetically introduced by randomly selecting a fraction $\eta_{\text{noise}} \in \{5\%, 10\%, 20\%, 25\%, 30\%\}$ of the 50,000 training examples and reassigning their ground-truth labels uniformly at random to an alternative class. In accordance with realistic diagnostic scenarios, no clean validation set is used for tuning or early stopping.

Evaluation Metric: Following the previous methods (Ravikumar et al., 2025; 2026) we assess mislabeled sample detection using the Area Under the Receiver Operating Characteristic curve (AUROC). AUROC provides a threshold-independent measure of how effectively a scoring metric separates clean instances from corrupted labels. Formally, it integrates the True Positive Rate (TPR) against the False Positive Rate (FPR) across all possible classification cutoffs:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{117}$$

where TP, FN, FP, and TN represent the counts of true positives (correctly flagged mislabeled samples), false negatives, false positives (clean samples erroneously flagged), and true negatives, respectively. An AUROC of 0.5 corresponds to an uninformative baseline equivalent to random guessing, whereas an AUROC of 1.0 denotes ideal discrimination, wherein every corrupted sample is assigned an anomaly score strictly higher than that of any cleanly labeled instance.

Adversarial Distance Proxy. Our diagnostic score is a proxy (not a certified bound) which mirrors the formulation of the theoretical adversarial distance upper bound derived in Theorem 2:

$$\bar{d}_{\text{adv},T}(\vec{z}_i) = \frac{2\left(L_u T - \text{CSL}(\vec{z}_i)\right)}{\text{CSG}(\vec{z}_i)} \tag{118}$$

where L_u denotes the sufficient cross-entropy misclassification threshold, $\vec{z}_i = (\vec{x}_i, y_i)$, and T represents the effective accumulation horizon. Because mislabeled instances are forced into unnatural

decision basins during late-stage memorization, their local margins to alternative class boundaries collapse, resulting in exceptionally small values of $\bar{d}_{\text{adv}}$ due to high gradients. Consequently, we use $-\bar{d}_{\text{adv}}$ as the anomaly ranking statistic.

We evaluate this proxy across ultra-late training windows as well as the terminal checkpoint, with the results of the terminal checkpoint being reported in Table 2, and all the other horizons in Table 5. Specifically, for the Last epoch/checkpoint setting, we instantiate the bound at $T = 1$ using strictly the instantaneous loss and input-gradient norm at the final training epoch:

$$\bar{d}_{\text{adv},1}(\vec{z}_i) = \frac{2\,(L_u - \ell(\vec{x}_i, y_i))}{\|\nabla_{\vec{x}}\ell(\vec{x}_i, y_i)\|_2} \tag{119}$$

For trailing-window horizons (e.g., Last 5%, 3%, 2%, and 1%), CSL and CSG are summed exclusively over the final $T_w = \lceil \alpha T_{\max} \rceil$ epochs, setting $T = T_w$ in Equation equation 118.

C.2.1 Baseline Methods

We benchmark our theoretical proxy against established data-cleansing and memorization metrics:

In Confidence. Following Carlini et al. (2019), the posterior deficit $1 - p(y \mid \vec{x})$ is computed at the terminal checkpoint, where $p(y \mid \vec{x})$ denotes the model's softmax probability assigned to the observed training label y.

Input Curvature. As introduced by Garg et al. (2024), input loss curvature estimates the local geometric sharpness via the trace of the input Hessian matrix, $\text{Tr}(\nabla^2_{\vec{x}}\ell)$. This quantity is approximated via Hutchinson's stochastic trace estimator with $n = 10$ Rademacher random vector projections and a finite-difference step size of $h = 0.001$.

Confident Learning (CL). Following Northcutt et al. (2021), confident learning is implemented through the open-source `cleanlab` package. Out-of-sample predicted class probability distributions are obtained via 3-fold cross-validation to estimate the joint distribution of noisy and latent true labels.

Second Split Forgetting Time (SSFT). Following Maini et al. (2022), SSFT tracks the speed with which a learned instance is forgotten when the model is continuously fine-tuned on a secondary, non-overlapping partition of the dataset.

Cumulative Sample Loss (CSL). Following Ravikumar et al. (2025), CSL records the cumulative trajectory loss $\sum_{t=1}^{T} \ell_t(\vec{x}, y)$ of each sample over optimization, capturing instances that persistently resist optimization.

Cumulative Sample Gradient (CSG). Following Ravikumar et al. (2026), CSG integrates the ℓ_2 norm of the input-space loss gradient, $\sum_{t=1}^{T} \|\nabla_{\vec{x}}\ell_t(\vec{x}, y)\|_2$, throughout training, leveraging the observation that memorized points require steep local gradient signals.

Sample Gradient Assisted Loss (SGAL). Proposed by Ravikumar et al. (2026), SGAL forms a combined ranking score by modulating the loss dynamics with input-gradient magnitudes to enhance alignment with sample-level memorization.

C.2.2 Compute Cost Analysis

Let $\mathcal{F}$ denote the computational FLOPs required for a single network forward pass on a mini-batch of inputs. In standard backpropagation, evaluating the parameter gradients during the backward pass requires approximately $2\mathcal{F}$, giving a baseline per-epoch training cost of $3\mathcal{F}$ per sample. For a training dataset containing N examples optimized over E epochs, we evaluate the algorithmic complexity of each diagnostic approach:

Input Curvature. Computing Hessian-vector products via finite differences (Garg et al., 2024) requires two forward evaluations and one backward evaluation per stochastic vector probe. Across H Hutchinson iterations, this adds an auxiliary overhead of $4\mathcal{F}\cdot H$ per sample per epoch. Combining this with baseline training yields an aggregate computational budget of:

$$3\mathcal{F} \cdot E \cdot N + 4\mathcal{F} \cdot H \cdot E \cdot N \tag{120}$$

Under the standard setting of $H = 10$, this elevates the total cost to $43\mathcal{F} \cdot E \cdot N$ ($\approx 14.3\times$ vanilla training). Even with a minimal setting of $H = 2$, the cost remains $11\mathcal{F} \cdot E \cdot N$ ($\approx 3.7\times$ vanilla training).

Confident Learning (CL). Generating out-of-sample probability estimates via K-fold cross-validation requires training K entirely separate models on disjoint splits. With $K = 3$, the cumulative overhead scales linearly to:

$$3 \times 3\mathcal{F} \cdot E \cdot N = 9\mathcal{F} \cdot E \cdot N \quad (\approx 3.0 \times \text{ vanilla training}) \tag{121}$$

Second Split Forgetting Time (SSFT). SSFT requires an initial training run on the primary dataset split followed by a secondary training pass on a held-out split, resulting in:

$$2 \times 3\mathcal{F} \cdot E \cdot N = 6\mathcal{F} \cdot E \cdot N \quad (\approx 2.0 \times \text{ vanilla training}) \tag{122}$$

CSL and CSG. Ravikumar et al. (2025) extract CSL directly from forward evaluations that occur naturally during optimization, incurring no additional floating-point operations ($\approx 3\mathcal{F} \cdot E \cdot N$). For CSG (Ravikumar et al., 2026), the input gradient $\nabla_{\vec{x}}\ell$ is obtained during the backward step as a byproduct of backpropagation, adding negligible memory overhead (only one extra matrix multiplication per epoch) and preserving a training cost of $\approx 3\mathcal{F} \cdot E \cdot N$.

Adversarial Distance Proxy (Ours). Our proxy $\bar{d}_{\mathrm{adv,T}}$ is evaluated analytically from the accumulated loss and input-gradient norms according to Equation equation 118. It requires no supplementary model training, no auxiliary forward-backward passes, and no cross-validation folds. The total cost remains strictly:

$$\approx 3\mathcal{F} \cdot E \cdot N \quad (1.0 \times \text{ vanilla training}) \tag{123}$$

Furthermore, because the proxy can be computed with state-of-the-art accuracy using only ultra-late checkpoints (down to the single final epoch $E = 1$), diagnostic logging can be restricted exclusively to the terminal phase of training, minimizing tracking overhead while outperforming compute-intensive alternatives.

Table 5: AUROC (mean ± std over 3 seeds) for mislabeled-sample detection on CIFAR-10 and CIFAR-100 at five noise levels. The table compares established baselines against our Adversarial Distance proxy ($\bar{d}_{\mathrm{adv,T}}$, evaluated across ultra-late training horizons (e.g., "Last 5%" and "Last Epoch" denote evaluating metrics using only the final 5% of training epochs and the absolute final training epoch, respectively). Previous state-of-the-art values among the baselines are underlined, and the best overall results are in **bold**.

Dataset	Method	5% Noise	10% Noise	20% Noise	25% Noise	30% Noise
CIFAR-10	CL (Northcutt et al., 2021)	0.8874 ± 0.0019	0.8551 ± 0.0030	0.7169 ± 0.1539	0.6960 ± 0.1387	0.6794 ± 0.1264
	In Conf.(Carlini et al., 2019)	0.7254 ± 0.0214	0.6528 ± 0.0042	0.5978 ± 0.0131	0.5800 ± 0.0051	0.5669 ± 0.0106
	SSFT (Maini et al., 2022)	0.9498 ± 0.0042	0.9360 ± 0.0020	0.9077 ± 0.0023	0.8910 ± 0.0050	0.8710 ± 0.0071
	Curv. (Garg et al., 2024)	0.9800 ± 0.0003	0.9819 ± 0.0006	0.9934 ± 0.0002	0.9932 ± 0.0001	0.9932 ± 0.0006
	CSL (Ravikumar et al., 2025)	0.9845 ± 0.0026	0.9864 ± 0.0004	0.9870 ± 0.0003	0.9904 ± 0.0005	0.9906 ± 0.0003
	CSG (Ravikumar et al., 2026)	0.9783 ± 0.0009	0.9809 ± 0.0011	0.9936 ± 0.0002	0.9934 ± 0.0001	0.9935 ± 0.0006
	SGAL (Ravikumar et al., 2026)	0.9115 ± 0.0057	0.8835 ± 0.0128	0.8664 ± 0.0070	0.8553 ± 0.0145	0.8470 ± 0.0283
	$\bar{d}_{\mathrm{adv}}$ (Last 5%)	**0.9963 ± 0.0006**	**0.9955 ± 0.0002**	**0.9953 ± 0.0001**	**0.9946 ± 0.0003**	0.9946 ± 0.0004
	$\bar{d}_{\mathrm{adv}}$ (Last 3%)	**0.9963 ± 0.0006**	**0.9955 ± 0.0002**	**0.9953 ± 0.0001**	**0.9946 ± 0.0003**	0.9946 ± 0.0004
	$\bar{d}_{\mathrm{adv}}$ (Last 2%)	**0.9963 ± 0.0006**	**0.9955 ± 0.0002**	**0.9953 ± 0.0002**	**0.9946 ± 0.0003**	**0.9947 ± 0.0004**
	$\bar{d}_{\mathrm{adv}}$ (Last 1%)	**0.9963 ± 0.0006**	**0.9955 ± 0.0002**	**0.9953 ± 0.0001**	**0.9946 ± 0.0003**	**0.9947 ± 0.0004**
	$\bar{d}_{\mathrm{adv}}$ (Last Epoch)	**0.9963 ± 0.0006**	**0.9955 ± 0.0002**	**0.9953 ± 0.0002**	**0.9946 ± 0.0003**	**0.9947 ± 0.0004**
CIFAR-100	CL (Northcutt et al., 2021)	0.8733 ± 0.0010	0.8536 ± 0.0006	0.7030 ± 0.1565	0.6833 ± 0.1427	0.6662 ± 0.1289
	In Conf. (Carlini et al., 2019)	0.7069 ± 0.0069	0.6884 ± 0.0053	0.6493 ± 0.0075	0.6324 ± 0.0051	0.6257 ± 0.0044
	SSFT (Maini et al., 2022)	0.8784 ± 0.0030	0.8664 ± 0.0024	0.8358 ± 0.0008	0.8203 ± 0.0016	0.8043 ± 0.0061
	Curv. (Garg et al., 2024)	0.9876 ± 0.0021	0.9892 ± 0.0012	0.9931 ± 0.0004	0.9931 ± 0.0004	0.9932 ± 0.0002
	CSL (Ravikumar et al., 2025)	0.9891 ± 0.0003	0.9895 ± 0.0002	0.9902 ± 0.0002	0.9904 ± 0.0002	0.9903 ± 0.0002
	CSG (Ravikumar et al., 2026)	0.9896 ± 0.0008	0.9904 ± 0.0006	0.9934 ± 0.0003	0.9936 ± 0.0002	0.9936 ± 0.0001
	SGAL (Ravikumar et al., 2026)	0.9895 ± 0.0006	0.9897 ± 0.0008	0.9856 ± 0.0008	0.9863 ± 0.0009	0.9861 ± 0.0004
	$\bar{d}_{\mathrm{adv}}$ (Last 5%)	**0.9942 ± 0.0001**	**0.9944 ± 0.0001**	**0.9946 ± 0.0001**	**0.9947 ± 0.0001**	**0.9947 ± 0.0002**
	$\bar{d}_{\mathrm{adv}}$ (Last 3%)	**0.9942 ± 0.0001**	**0.9944 ± 0.0001**	**0.9946 ± 0.0001**	**0.9947 ± 0.0001**	**0.9947 ± 0.0002**
	$\bar{d}_{\mathrm{adv}}$ (Last 2%)	**0.9942 ± 0.0001**	**0.9944 ± 0.0001**	**0.9946 ± 0.0001**	**0.9947 ± 0.0001**	**0.9947 ± 0.0002**
	$\bar{d}_{\mathrm{adv}}$ (Last 1%)	**0.9942 ± 0.0001**	**0.9944 ± 0.0001**	**0.9946 ± 0.0001**	**0.9947 ± 0.0001**	**0.9947 ± 0.0002**
	$\bar{d}_{\mathrm{adv}}$ (Last Epoch)	**0.9942 ± 0.0001**	**0.9944 ± 0.0001**	**0.9946 ± 0.0001**	**0.9947 ± 0.0001**	**0.9947 ± 0.0002**

C.3 Duplicate Detection: Setup and Additional Analysis

Evaluation Metric. Similar to the mislabeled detection setup, we assess duplicate sample identification using the Area Under the Receiver Operating Characteristic curve (AUROC).

Diagnostic Rationale and Injection Protocol. For comparison with the baselines we intentionally contaminate the datasets with exact synthetic duplicates. When a neural network is forced to train on identical input instances that possess conflicting labels, it must aggressively distort its local decision boundary to satisfy the optimization objective. This severe boundary distortion inherently collapses the local adversarial distance (d_{adv}). Our proxy formulation $\bar{d}_{\mathrm{adv},T}$ is designed to expose these highly vulnerable, contorted data points.

Following the same method of Ravikumar et al. (2025) we procedurally generate the corrupted training sets by injecting exactly 250 duplicates into both the CIFAR-10 and CIFAR-100 training pools. For each dataset, we randomly sample 250 unique index pairs from the 50,000 available training examples. The image data at the target index is entirely overwritten by the image data from the source index, resulting in an exact pixel-level clone. To maximize the optimization difficulty and ensure extreme memorization behavior, the newly overwritten image is uniformly assigned a random, conflicting class label from the remaining incorrect categories.

Training Setup. We optimize ResNet-18 architectures using Stochastic Gradient Descent (SGD) configured with a momentum coefficient of 0.9, a weight decay penalty of 1×10^{-4}, and a mini-batch size of 128. The models are trained from scratch for a total of 200 epochs. The learning rate is initialized at 0.1 and undergoes a step-wise reduction by a factor of 10 at 60% of training (epoch 120) and 80% of training (epoch 160). Inputs are augmented dynamically during training via 32×32 random crops (with 4 pixels of reflection padding) and random horizontal flips. To ensure statistical reliability, all evaluations and diagnostic scores are averaged across 3 independently initialized random seeds.

Adversarial Distance Proxy Formulation. We use the proxy $\bar{d}_{adv,T}$ as explained in the mislabeled experiment setup here. We extract the cumulative loss (CSL) and cumulative input-gradient norms (CSG) over varying temporal windows T, allowing us to evaluate the proxy's sensitivity from

Table 6: AUROC evaluation (mean ± std) for duplicate and near-duplicate training sample detection on CIFAR-10 and CIFAR-100. We benchmark our theoretical Adversarial Distance proxy $\bar{d}_{\text{adv},T}$ against established trajectory and confidence metrics. The proxy is evaluated across multiple temporal integration horizons, ranging from the full trajectory to strictly the final checkpoint ("Last Epoch"). Previous state-of-the-art performances among the baselines are underlined, while the best overall diagnostic results are highlighted in **bold**.

Method	CIFAR-10	CIFAR-100
In Conf. (Carlini et al., 2019)	0.9237 ± 0.0114	0.8623 ± 0.0131
CL (Northcutt et al., 2021)	0.5533 ± 0.0031	0.5873 ± 0.0090
SSFT (Maini et al., 2022)	0.8490 ± 0.0034	0.7938 ± 0.0045
Curv. (Garg et al., 2024)	0.9536 ± 0.0030	0.9639 ± 0.0030
CSL (Ravikumar et al., 2025)	$\underline{0.9822 \pm 0.0011}$	$\underline{0.9894 \pm 0.0012}$
CSG (Ravikumar et al., 2026)	0.9525 ± 0.0015	0.9743 ± 0.0022
SGAL (Ravikumar et al., 2026)	0.9255 ± 0.0019	0.9537 ± 0.0017
$\bar{d}_{\text{adv},T}$ (Full 100%)	0.9712 ± 0.0018	0.9833 ± 0.0014
$\bar{d}_{\text{adv},T}$ (Last 50%)	0.9991 ± 0.0002	0.9986 ± 0.0003
$\bar{d}_{\text{adv},T}$ (Last 10%)	$\mathbf{0.9998 \pm 0.0001}$	$\mathbf{0.9994 \pm 0.0001}$
$\bar{d}_{\text{adv},T}$ (Last 5%)	$\mathbf{0.9998 \pm 0.0001}$	$\mathbf{0.9994 \pm 0.0001}$
$\bar{d}_{\text{adv},T}$ (Last 3%)	$\mathbf{0.9998 \pm 0.0001}$	$\mathbf{0.9994 \pm 0.0001}$
$\bar{d}_{\text{adv},T}$ (Last 2%)	$\mathbf{0.9998 \pm 0.0001}$	$\mathbf{0.9994 \pm 0.0001}$
$\bar{d}_{\text{adv},T}$ (Last 1%)	$\mathbf{0.9998 \pm 0.0001}$	$\mathbf{0.9994 \pm 0.0001}$
$\bar{d}_{\text{adv},T}$ (Last Epoch)	$\mathbf{0.9998 \pm 0.0001}$	$\mathbf{0.9994 \pm 0.0001}$

the complete 200-epoch trajectory down to the instantaneous final epoch ($T = 1$). We utilize the inverse negative score $-\bar{d}_{\text{adv},T}$ to rank the samples, operating on the premise that duplicates exhibit the smallest local margins.

Baselines. We benchmark against Threshold Learning Time (Thr. LT), which isolates the earliest epoch a sample achieves correct classification. We additionally evaluate In Confidence (In Conf.), Confident Learning (CL) via cross-validation, Second Split Forgetting Time (SSFT), Hutchinson Input Curvature (Curv.), Cumulative Sample Loss (CSL) and Cumulative sample Gradients(CSG). Baseline implementation details align precisely with the formal definitions provided in Appendix C.2.1.

Results and Temporal Ablation. The diagnostic results detailed in Table 6 demonstrate the definitive superiority of the proposed proxy. When restricted to late-training intervals, $\bar{d}_{\text{adv},T}$ delivers near-perfect discrimination, yielding AUROC scores of 0.9998 on CIFAR-10 and 0.9994 on CIFAR-100. This substantially outperforms the strongest baseline metric (CSL), which tops out at 0.9821 and 0.9886, respectively.

The temporal ablation highlights the proxy's reliance on late-stage optimization dynamics. Accumulating the proxy over the full 200-epoch horizon actually suppresses performance (e.g., dropping to 0.9712 on CIFAR-10). During the initial phases of optimization, the network attempts to construct broad feature representations, causing elevated loss and gradient signals across both clean and corrupted data uniformly. Once the core data distribution is absorbed, the clean samples sink into flat, zero-gradient minima. The duplicated samples, however, sustain massive localized gradient norms and persistent loss as the optimizer battles the conflicting labels. By intentionally restricting the proxy accumulation to the terminal 10% of epochs or strictly the final epoch we isolate this precise memorization conflict, stripping away the initial underfitting noise to achieve state-of-the-art detection with zero auxiliary training cost.

C.4 Zero-Cost Privacy Auditing

C.4.1 Experimental Setup

Diagnostic Rationale and Threat Model. Empirical evaluation of privacy risk in deep learning models is predominantly conducted via Membership Inference Attacks (MIAs), which measure the likelihood that an adversary can discern whether a specific record was included in the training set. In high-assurance security auditing, privacy risk is not uniformly distributed across the training distribution; rather, it is heavily concentrated within a small subset of vulnerable outliers and atypical instances. Auditing against worst-case adversaries requires evaluating attack success at extremely conservative False Positive Rate (FPR) thresholds (e.g., $\alpha = 10^{-3}$), ensuring that identified points reflect genuine privacy leakage rather than statistical false alarms.

However, state-of-the-art shadow-model-based attacks (such as LiRA, RMIA, and AttackR) require training tens or hundreds of auxiliary shadow models to calibrate reference non-member loss distributions. This imposes a prohibitive computational burden that prevents iterative privacy auditing during model development. To resolve this bottleneck, we formulate CSL-Diff as a zero-cost dataset privacy auditing metric. Rather than relying on external shadow ensembles, CSL-Diff extracts the distinctive memorization signature left by vulnerable samples directly from the native loss dynamics of a single training run, enabling instantaneous record-level privacy auditing at zero marginal compute cost.

Dataset Partitioning and Training Configuration. Following standard auditing benchmarks established in the privacy literature (Pollock et al., 2025), all experiments are conducted on CIFAR-10 and CIFAR-100 using WideResNet28-2 (WRN28-2) architectures. The full 50,000-sample training dataset is randomly partitioned into two equal, disjoint subsets of 25,000 examples: a member split (D_{train}) used to train the target model, and a non-member holdout split (D_{test}) used for shadow calibration and empirical FPR estimation.

Target models are trained from scratch for 100 epochs using Stochastic Gradient Descent (SGD) with a mini-batch size of 256, a momentum parameter of 0.9, and a weight decay coefficient of 5×10^{-4}. Optimization is governed by a Cosine Annealing learning rate schedule initialized at $\eta_0 = 0.1$. Standard vision augmentations (random horizontal flips and random crops with 4 pixels of reflection padding) are applied during training. To eliminate stochastic evaluation noise, all reported metrics represent the mean and standard deviation across 3 independently initialized target model training runs, trained on the same subset.

Ground-Truth Membership Vulnerability. To establish rigorous ground-truth evaluation targets, we execute full-scale shadow-model-based MIAs. A training example $(\vec{x}_i, y_i) \in D_{\text{train}}$ is defined as *vulnerable* if an attacker's confidence score exceeds the classification threshold corresponding to a strict false positive constraint $\alpha = 10^{-3}$:

$$\mathcal{V}_\alpha = \{(\vec{x}_i, y_i) \in D_{\text{train}} \mid \mathcal{A}(\vec{x}_i, y_i; \theta) \geq \tau_\alpha\}, \quad \text{where } \text{FPR}(\tau_\alpha) = \alpha \tag{124}$$

We benchmark against three state-of-the-art shadow attack methodologies, training full shadow model ensembles under hyperparameter regimes identical to the target model:

1. **LiRA** (Carlini et al., 2022): Implemented with 256 shadow models (128 member models and 128 non-member models per audited sample) using Gaussian mixture modeling over log-transformed losses with two query augmentations per instance.
2. **RMIA** (Zarifzadeh et al., 2023): Instantiated with 256 shadow models to compute population-conditioned likelihood ratios.
3. **AttackR** (Ye et al., 2022): Instantiated with 128 non-member shadow models to evaluate sample-level rank statistics.

Because an auditor cannot predict the precise mathematical formulation an external adversary will deploy, we additionally construct a comprehensive **Union of Attacks** ground truth:

$$\mathcal{V}_\alpha^{\text{Union}} = \mathcal{V}_\alpha^{\text{LiRA}} \cup \mathcal{V}_\alpha^{\text{RMIA}} \cup \mathcal{V}_\alpha^{\text{AttackR}} \tag{125}$$

An example is designated as vulnerable under the Union target if it is successfully breached by at least one of the three reference attacks at $\text{FPR} = 10^{-3}$.

Zero-Cost Privacy Metric Formulation (CSL-Diff). The mathematical foundation of CSL-Diff stems from the divergent learning dynamics of prototypical versus memorized data. Typical, easily generalized instances are learned in the earliest epochs; their loss decays rapidly and remains near zero, yielding negligible cumulative loss across both early and late training windows. Conversely, vulnerable instances are atypical outliers that strongly resist optimization. In early training, they exhibit substantial loss; as the model overfits, the optimization forces their loss toward zero by the terminal epochs.

To isolate this specific memorization trajectory without auxiliary training runs, we define the Cumulative Sample Loss Difference (CSL-Diff-n) as the contrast between the cumulative loss accrued over the initial $n\%$ of training epochs and the final $n\%$ of training epochs:

$$\text{CSL-Diff-}n(\vec{z}_i) = \sum_{t=1}^{\lfloor \frac{n}{100} T \rfloor} \ell(\vec{w}_t, \vec{z}_i) - \sum_{t=T-\lfloor \frac{n}{100} T \rfloor + 1}^{T} \ell(\vec{w}_t, \vec{z}_i) \tag{126}$$

where $\vec{z}_i = (\vec{x}_i, y_i)$, T is the total epoch budget (here $T = 100$), and $\ell(\vec{w}_t, \vec{z}_i)$ denotes the instantaneous cross-entropy loss at epoch t. We evaluate two specific temporal horizons: CSL-Diff-40 ($n = 40\%$) and CSL-Diff-50 ($n = 50\%$). Samples are ranked in descending order according to Equation equation 126, such that candidates exhibiting the steepest loss differential between early underfitting and late-stage memorization are audited as the most vulnerable.

Comparative Baseline Predictors. We benchmark CSL-Diff against established artifact-based heuristics that require zero shadow models:

- **Loss** (Yeom et al., 2018): Ranks samples using the instantaneous cross-entropy loss evaluated at the final converged checkpoint $\vec{w}_T$.
- **CSG** (Ravikumar et al., 2026): Integrates the ℓ_2 norm of the input-space loss gradient across the full training trajectory: $\sum_{t=1}^{T} \|\nabla_{\vec{x}} \ell(\vec{w}_t, \vec{z}_i)\|_2$.
- **LT-Mean** (Pollock et al., 2025): Computes the arithmetic mean of the per-epoch loss trace: $\frac{1}{T} \sum_{t=1}^{T} \ell(\vec{w}_t, \vec{z}_i)$.
- **LT-Std** (Pollock et al., 2025): Computes the sample standard deviation of the per-epoch loss trace across the optimization path.
- **LT-IQR** (Pollock et al., 2025): Computes the interquartile range ($Q_{0.75} - Q_{0.25}$) of the per-epoch loss trace, capturing intermediate loss volatility.

Evaluation Metrics. Model developers primarily require high precision when flagging the top fraction of at-risk candidates for targeted mitigations (such as selective unlearning or data deduplication). We therefore evaluate detection performance using Precision@k and Recall@k, defined over the top $k\%$ highest-scoring samples:

$$\text{Precision@}k = \frac{|\text{top}_k(\phi) \cap \mathcal{V}_\alpha|}{|\text{top}_k(\phi)|}, \qquad \text{Recall@}k = \frac{|\text{top}_k(\phi) \cap \mathcal{V}_\alpha|}{|\mathcal{V}_\alpha|} \tag{127}$$

where $\text{top}_k(\phi)$ denotes the top $k\%$ highest-ranked samples according to predictor ϕ, evaluated at $k \in \{1\%, 3\%, 5\%\}$.

Primary Auditing Performance (LiRA and Attack Union). As shown in Table 7 and Table 9, our zero-cost metric consistently achieves state-of-the-art predictive performance when identifying points vulnerable to LiRA at FPR $= 10^{-3}$. On CIFAR-10, CSL-Diff-40 achieves a Precision@1% of 0.839, outperforming the previous state-of-the-art metric LT-IQR (0.793) while drastically exceeding conventional metrics such as CSG (0.277) and final Loss (0.055). On CIFAR-100, where fine-grained semantics exacerbate privacy leakage, CSL-Diff-40 achieves an exceptional Precision@1% of 0.947, capturing over 94% truly vulnerable records within its top 250 predictions.

Importantly, this predictive precision is not confined to the LiRA formulation alone. When evaluated against the multi-threat Union of Attacks target (Table 8 and Table 10), CSL-Diff-40 attains Precision@1% scores of 0.848 on CIFAR-10 and 0.957 on CIFAR-100. This confirms that

CSL-Diff isolates a fundamental property of sample memorization rather than overfitting to the statistical quirks of a single attack algorithm. Across all evaluations, CSL-Diff-40 and CSL-Diff-50 exhibit comparable, highly stable performance, demonstrating robustness to the specific choice of temporal window size n.

Sensitivity and Robustness to Security Thresholds (FPR Sweeps). In practical deployment, auditors may operate under varying risk tolerances. We evaluate the sensitivity of CSL-Diff across a sweep of FPR constraints: $\alpha \in \{10^{-2}, 5\times10^{-3}, 10^{-3}\}$. As documented in the summary evaluations of Table 11 and the extended breakdowns in Table 17 and Table 18, relaxing the false positive threshold naturally increases the total volume of ground-truth vulnerable samples.

Across every evaluated threshold, CSL-Diff maintains a commanding advantage over all baselines. On CIFAR-10 at $\text{FPR} = 1.0\%$, CSL-Diff-40 achieves 0.933 $\text{Precision@}1\%$ (compared to 0.869 for LT-IQR and 0.384 for CSG). On CIFAR-100 at the same threshold, its precision reaches 0.987. The consistent superiority of CSL-Diff across two orders of magnitude of FPR demonstrates that the differential loss trajectory remains a reliable indicator of individual privacy risk regardless of the strictness of the threat model. This continuous dominance is visually corroborated in Figure 3, which plots the precision curves across the entire FPR spectrum. Furthermore, Figures 12 and 13 confirm the metric's stability across different top-k selection fractions ($k \in \{0.1\%, 1\%, 2\%, 3\%, 5\%\}$), showing that CSL-Diff scales gracefully and maintains high precision even as the audited fraction of the dataset expands.

Auditing Stability Under Constrained Attacker Budgets (Shadow Scalability). While benchmarks deploy hundreds of shadow models, practical adversaries might be constrained by compute and may execute attacks with far fewer shadow models (e.g., 64 or 16). In such resource-constrained settings, the empirical shadow reference distribution becomes noisy, degrading attack calibration. We investigate how effectively our zero-cost metric predicts vulnerable points defined by these lower-capacity attacks.

Table 12, along with the extended results in Tables 13–16, presents auditing performance across 256, 64, and 16 shadow models. Even when the ground-truth attack signal is degraded by small shadow ensembles, CSL-Diff consistently outperforms all alternative predictors. On CIFAR-10 with only 16 shadow models, CSL-Diff-40 achieves 0.372 precision at $k = 1\%$, outperforming LT-IQR (0.336) and nearly quadrupling the precision of CSG (0.095). On CIFAR-100 with 64 shadow models, CSL-Diff-50 reaches 0.897 precision at $k = 1\%$. This resilience confirms that CSL-Diff captures the primary underlying drivers of privacy vulnerability, allowing model developers to accurately anticipate membership leakage even when reference attack signals are noisy. These trends are visually reinforced in Figures 10 and 11, which illustrate the baseline comparisons across the continuous FPR sweep for the 64 and 16 shadow settings. Additionally, Figures 14 through 17 provide the extended varying-k analysis for these restricted budgets, demonstrating that our metric provides a predictable and stable lower bound on privacy risks even under significant signal degradation.

Computational Complexity and Diagnostic Utility. The overriding advantage of CSL-Diff lies in its computational efficiency. Evaluating reference shadow-model attacks like LiRA or RMIA with 256 models requires $256 \times 3\mathcal{F} \cdot T \cdot N$ operations, scaling the auditing footprint by over two orders of magnitude relative to standard training. In contrast, CSL-Diff requires no auxiliary models, no gradient clipping, and no secondary backpropagation passes. Because per-sample cross-entropy losses are already computed during the forward pass of normal training, tracking CSL-Diff introduces zero floating-point overhead ($\approx 1.0\times$ baseline training cost).

By providing a shadow-free ranking of record-level privacy exposure, CSL-Diff equips practitioners with an actionable diagnostic tool. It enables developers to audit training privacy continuously throughout model training, compare hyperparameter choices, and execute targeted dataset pruning or localized unlearning on the most exposed samples prior to production deployment.

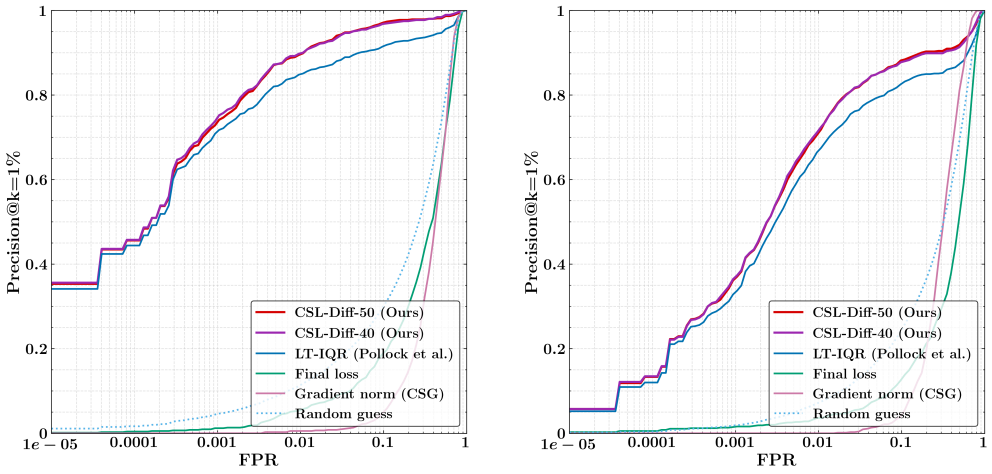


Figure 10: Precision@1% under restricted attacker budgets for CIFAR-10: evaluated with 64 shadow models (left) and 16 shadow models (right).

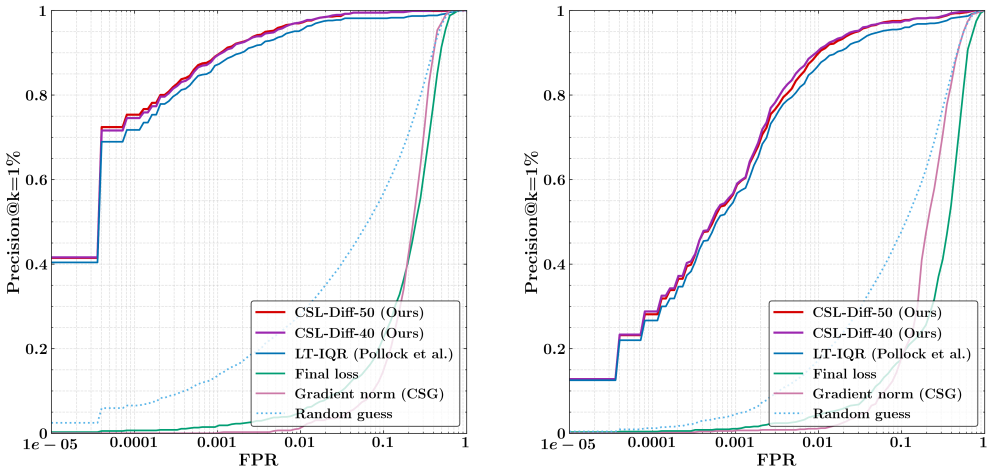


Figure 11: Precision@1% under highly restricted attacker budgets for CIFAR-100: evaluated with 64 shadow models (left) and 16 shadow models (right).

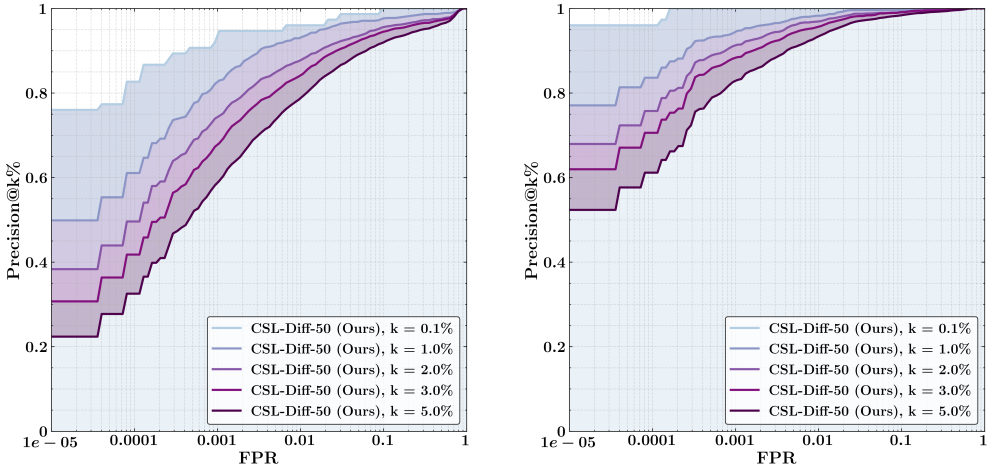


Figure 12: Precision@k% of the CSL-Diff-50 method across variable k fractions and FPR thresholds. Reported on CIFAR-10 (left) and CIFAR-100 (right) using 256 shadow models.

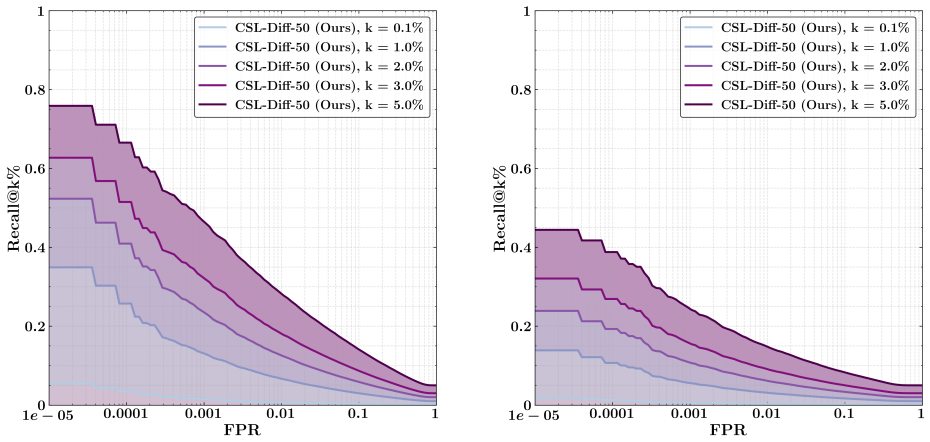


Figure 13: Recall@k% of the CSL-Diff-50 method across variable k fractions and FPR thresholds. Reported on CIFAR-10 (left) and CIFAR-100 (right) using 256 shadow models.

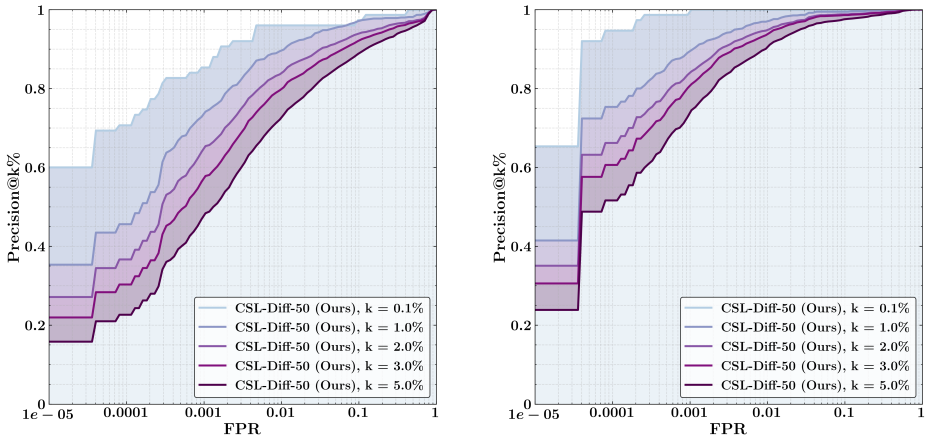


Figure 14: Precision@k% of the CSL-Diff-50 method across variable k fractions and FPR thresholds. Reported on CIFAR-10 (left) and CIFAR-100 (right) under a restricted attacker budget of 64 shadow models.

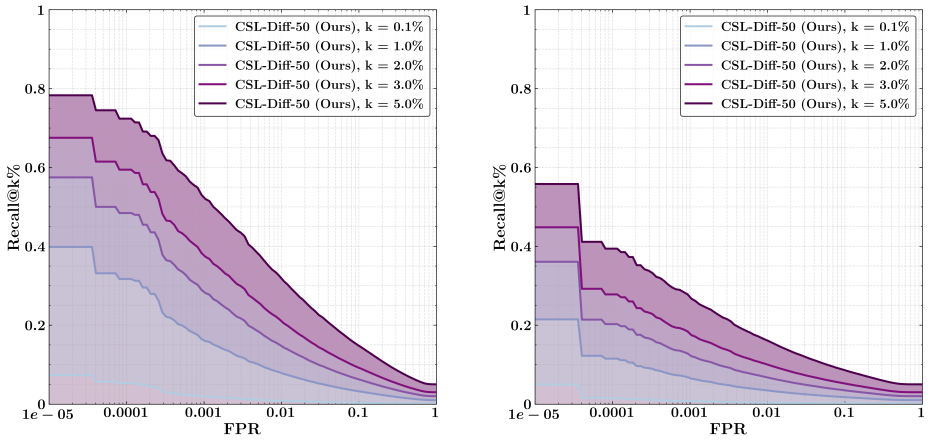


Figure 15: Recall@k% of the CSL-Diff-50 method across variable k fractions and FPR thresholds. Reported on CIFAR-10 (left) and CIFAR-100 (right) under a restricted attacker budget of 64 shadow models.

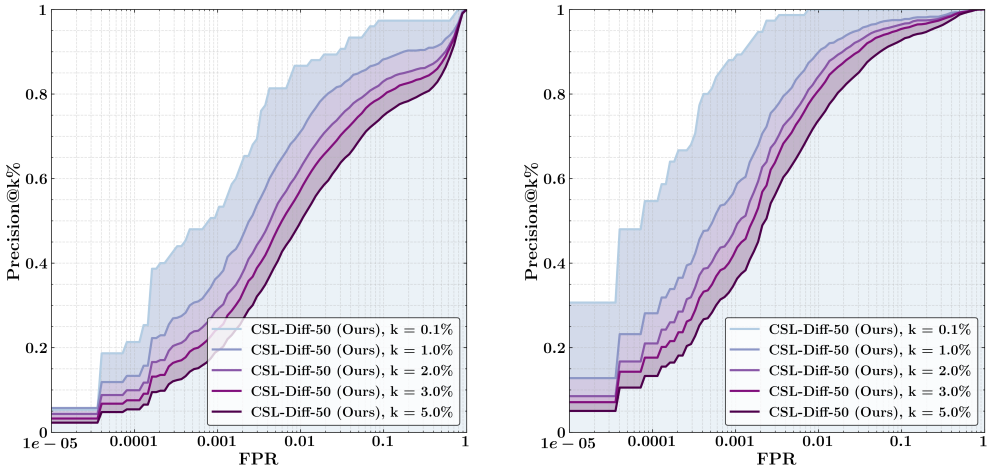


Figure 16: Precision@k% of the CSL-Diff-50 method across variable k fractions and FPR thresholds. Reported on CIFAR-10 (left) and CIFAR-100 (right) under a highly restricted attacker budget of 16 shadow models.

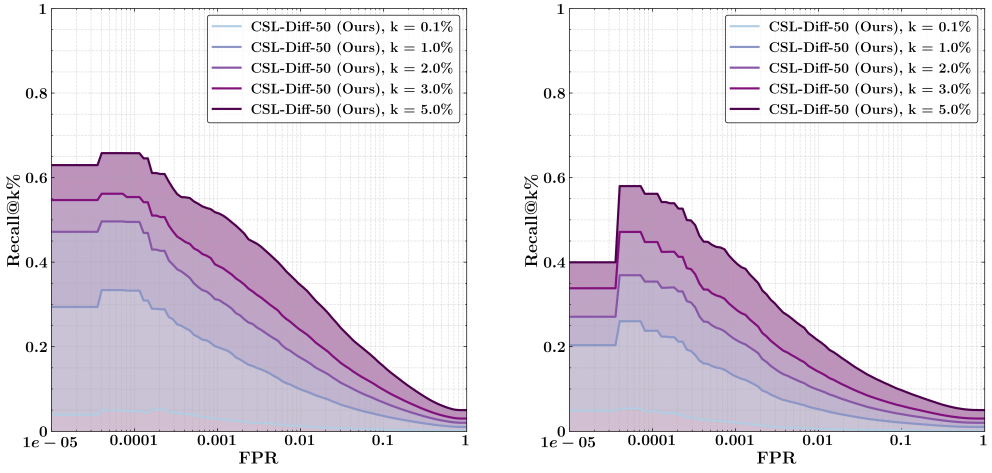


Figure 17: Recall@k% of the CSL-Diff-50 method across variable k fractions and FPR thresholds. Reported on CIFAR-10 (left) and CIFAR-100 (right) under a highly restricted attacker budget of 16 shadow models.

Table 7: Precision@k and Recall@k identifying points vulnerable to the LiRA attack. Reported at strict security bounds (FPR=10^{-3}) across 3 target model seeds for WideResNet28-2 on CIFAR-10 (256 shadows).

Method	Prec@1%	Rec@1%	Prec@3%	Rec@3%	Prec@5%	Rec@5%
Final Loss	0.055 ± 0.012	0.009 ± 0.002	0.080 ± 0.011	0.037 ± 0.005	0.095 ± 0.016	0.075 ± 0.015
CSG (Ravikumar et al., 2026)	0.277 ± 0.033	0.044 ± 0.007	0.325 ± 0.021	0.153 ± 0.018	0.323 ± 0.007	0.253 ± 0.020
LT-Mean (Pollock et al., 2025)	0.471 ± 0.065	0.074 ± 0.014	0.467 ± 0.022	0.220 ± 0.023	0.445 ± 0.010	0.349 ± 0.029
LT-Std (Pollock et al., 2025)	0.729 ± 0.044	0.114 ± 0.006	0.590 ± 0.028	0.277 ± 0.017	0.517 ± 0.027	0.404 ± 0.017
LT-IQR (Pollock et al., 2025)	0.793 ± 0.027	0.124 ± 0.009	0.636 ± 0.018	0.299 ± 0.020	0.540 ± 0.005	0.423 ± 0.022
CSL-Diff-40 (Ours)	$\mathbf{0.839 \pm 0.023}$	$\mathbf{0.131 \pm 0.008}$	$\mathbf{0.685 \pm 0.033}$	$\mathbf{0.322 \pm 0.016}$	0.587 ± 0.024	0.459 ± 0.012
CSL-Diff-50 (Ours)	0.829 ± 0.040	0.130 ± 0.010	0.680 ± 0.037	0.319 ± 0.018	$\mathbf{0.591 \pm 0.025}$	$\mathbf{0.462 \pm 0.015}$

Table 8: Precision@k and Recall@k predicting vulnerable points defined by the union of LiRA, AttackR, and RMIA. Reported at FPR=10^{-3} on CIFAR-10 (256 shadows).

Method	Prec@1%	Rec@1%	Prec@3%	Rec@3%	Prec@5%	Rec@5%
Final Loss	0.092 ± 0.021	0.013 ± 0.004	0.107 ± 0.020	0.045 ± 0.007	0.119 ± 0.013	0.084 ± 0.013
CSG (Ravikumar et al., 2026)	0.288 ± 0.034	0.041 ± 0.008	0.336 ± 0.016	0.142 ± 0.018	0.335 ± 0.001	0.235 ± 0.019
LT-Mean (Pollock et al., 2025)	0.503 ± 0.062	0.071 ± 0.014	0.495 ± 0.014	0.209 ± 0.022	0.470 ± 0.003	0.330 ± 0.026
LT-Std (Pollock et al., 2025)	0.740 ± 0.050	0.104 ± 0.004	0.604 ± 0.027	0.254 ± 0.013	0.533 ± 0.032	0.373 ± 0.012
LT-IQR (Pollock et al., 2025)	0.801 ± 0.026	0.113 ± 0.009	0.651 ± 0.016	0.274 ± 0.018	0.562 ± 0.014	0.394 ± 0.023
CSL-Diff-40 (Ours)	$\mathbf{0.848 \pm 0.025}$	$\mathbf{0.119 \pm 0.008}$	$\mathbf{0.705 \pm 0.038}$	$\mathbf{0.296 \pm 0.011}$	0.611 ± 0.031	0.428 ± 0.014
CSL-Diff-50 (Ours)	0.847 ± 0.034	0.119 ± 0.008	0.701 ± 0.043	0.295 ± 0.012	$\mathbf{0.614 \pm 0.032}$	$\mathbf{0.430 \pm 0.014}$

Table 9: Precision@k and Recall@k identifying points vulnerable to the LiRA attack. Reported at strict security bounds (FPR=10^{-3}) across 3 target model seeds for WideResNet28-2 on CIFAR-100 (256 shadows).

Method	Prec@1%	Rec@1%	Prec@3%	Rec@3%	Prec@5%	Rec@5%
Final Loss	0.003 ± 0.002	0.000 ± 0.000	0.010 ± 0.012	0.002 ± 0.002	0.018 ± 0.014	0.005 ± 0.004
CSG (Ravikumar et al., 2026)	0.175 ± 0.041	0.010 ± 0.002	0.212 ± 0.025	0.037 ± 0.003	0.239 ± 0.033	0.070 ± 0.006
LT-Mean (Pollock et al., 2025)	0.072 ± 0.042	0.004 ± 0.002	0.116 ± 0.050	0.020 ± 0.008	0.179 ± 0.066	0.052 ± 0.017
LT-Std (Pollock et al., 2025)	0.897 ± 0.008	0.053 ± 0.002	0.826 ± 0.021	0.145 ± 0.008	0.762 ± 0.019	0.223 ± 0.009
LT-IQR (Pollock et al., 2025)	0.927 ± 0.024	0.054 ± 0.002	0.856 ± 0.023	0.150 ± 0.008	0.797 ± 0.021	0.233 ± 0.009
CSL-Diff-40 (Ours)	$\mathbf{0.947 \pm 0.002}$	0.055 ± 0.003	0.878 ± 0.014	0.154 ± 0.008	0.826 ± 0.013	0.242 ± 0.009
CSL-Diff-50 (Ours)	0.947 ± 0.010	$\mathbf{0.055 \pm 0.003}$	$\mathbf{0.884 \pm 0.014}$	$\mathbf{0.155 \pm 0.009}$	$\mathbf{0.830 \pm 0.022}$	$\mathbf{0.243 \pm 0.007}$

Table 10: Precision@k and Recall@k predicting vulnerable points defined by the union of LiRA, AttackR, and RMIA. Reported at FPR=10^{-3} on CIFAR-100 (256 shadows).

Method	Prec@1%	Rec@1%	Prec@3%	Rec@3%	Prec@5%	Rec@5%
Final Loss	0.003 ± 0.002	0.000 ± 0.000	0.010 ± 0.012	0.002 ± 0.002	0.018 ± 0.014	0.005 ± 0.004
CSG (Ravikumar et al., 2026)	0.181 ± 0.040	0.010 ± 0.002	0.219 ± 0.023	0.035 ± 0.002	0.248 ± 0.033	0.067 ± 0.006
LT-Mean (Pollock et al., 2025)	0.128 ± 0.056	0.007 ± 0.003	0.161 ± 0.057	0.026 ± 0.008	0.214 ± 0.070	0.057 ± 0.016
LT-Std (Pollock et al., 2025)	0.905 ± 0.014	0.049 ± 0.002	0.837 ± 0.025	0.136 ± 0.008	0.779 ± 0.021	0.210 ± 0.009
LT-IQR (Pollock et al., 2025)	0.933 ± 0.030	0.050 ± 0.002	0.868 ± 0.026	0.141 ± 0.007	0.813 ± 0.025	0.220 ± 0.008
CSL-Diff-40 (Ours)	$\mathbf{0.957 \pm 0.006}$	$\mathbf{0.052 \pm 0.003}$	0.891 ± 0.018	0.144 ± 0.007	0.842 ± 0.015	0.228 ± 0.008
CSL-Diff-50 (Ours)	0.953 ± 0.016	0.052 ± 0.002	$\mathbf{0.895 \pm 0.015}$	$\mathbf{0.145 \pm 0.008}$	$\mathbf{0.846 \pm 0.024}$	$\mathbf{0.229 \pm 0.007}$

Table 11: Precision@1.0% predicting LiRA vulnerability under varying FPR security thresholds across CIFAR-10 and CIFAR-100 (256 shadows). Mean $\pm$ standard deviation across 3 seeds.

Dataset	Evaluation FPR	LT-Std	LT-IQR	CSL-Diff-40	CSL-Diff-50
CIFAR-10	FPR $= 10^{-2}$ (1.0%)	0.833 ± 0.012	0.869 ± 0.024	$\mathbf{0.933 \pm 0.008}$	0.932 ± 0.022
CIFAR-10	FPR $= 5 \times 10^{-3}$ (0.5%)	0.808 ± 0.024	0.851 ± 0.016	0.911 ± 0.016	$\mathbf{0.913 \pm 0.015}$
CIFAR-10	FPR $= 10^{-3}$ (0.1%)	0.729 ± 0.044	0.793 ± 0.027	$\mathbf{0.839 \pm 0.023}$	0.829 ± 0.040
CIFAR-100	FPR $= 10^{-2}$ (1.0%)	0.949 ± 0.019	0.952 ± 0.030	$\mathbf{0.972 \pm 0.007}$	0.971 ± 0.016
CIFAR-100	FPR $= 5 \times 10^{-3}$ (0.5%)	0.937 ± 0.018	0.963 ± 0.021	$\mathbf{0.977 \pm 0.006}$	0.975 ± 0.006
CIFAR-100	FPR $= 10^{-3}$ (0.1%)	0.897 ± 0.008	0.927 ± 0.024	$\mathbf{0.947 \pm 0.002}$	0.947 ± 0.010

Table 12: Precision@1.0% predicting LiRA vulnerability (FPR=10^{-3}) under restricted attacker shadow budgets across CIFAR-10 and CIFAR-100. Results averaged over 3 seeds.

Dataset	Shadows	CSG	LT-IQR	CSL-Diff-40	CSL-Diff-50
CIFAR-10	256 Shadows	0.277 ± 0.033	0.793 ± 0.027	$\mathbf{0.839 \pm 0.023}$	0.829 ± 0.040
CIFAR-10	64 Shadows	0.232 ± 0.045	0.713 ± 0.024	$\mathbf{0.748 \pm 0.021}$	0.737 ± 0.024
CIFAR-10	16 Shadows	0.095 ± 0.021	0.336 ± 0.069	$\mathbf{0.372 \pm 0.069}$	0.369 ± 0.074
CIFAR-100	256 Shadows	0.175 ± 0.041	0.927 ± 0.024	$\mathbf{0.947 \pm 0.002}$	0.947 ± 0.010
CIFAR-100	64 Shadows	0.121 ± 0.027	0.873 ± 0.043	0.896 ± 0.029	$\mathbf{0.897 \pm 0.040}$
CIFAR-100	16 Shadows	0.028 ± 0.007	0.557 ± 0.021	$\mathbf{0.579 \pm 0.036}$	0.576 ± 0.011

Table 13: Extended Results: Precision@k and Recall@k for CIFAR-10 at FPR=10^{-3} across restricted shadow model regimes (64 and 16 shadows). LiRA vulnerability target.

Method	Prec@1%	Rec@1%	Prec@3%	Rec@3%	Prec@5%	Rec@5%
CIFAR-10 — 64 Shadow Models						
Final Loss	0.040 ± 0.014	0.009 ± 0.003	0.063 ± 0.017	0.041 ± 0.009	0.076 ± 0.017	0.083 ± 0.015
CSG (Ravikumar et al., 2026)	0.232 ± 0.045	0.051 ± 0.011	0.268 ± 0.018	0.176 ± 0.012	0.261 ± 0.013	0.285 ± 0.009
LT-Std (Pollock et al., 2025)	0.637 ± 0.038	0.139 ± 0.013	0.496 ± 0.021	0.325 ± 0.027	0.413 ± 0.013	0.452 ± 0.030
LT-IQR (Pollock et al., 2025)	0.713 ± 0.024	0.156 ± 0.012	0.543 ± 0.010	0.356 ± 0.025	0.443 ± 0.007	0.484 ± 0.022
CSL-Diff-40 (Ours)	$\mathbf{0.748 \pm 0.021}$	$\mathbf{0.164 \pm 0.011}$	$\mathbf{0.576 \pm 0.022}$	$\mathbf{0.378 \pm 0.029}$	0.476 ± 0.015	0.520 ± 0.026
CSL-Diff-50 (Ours)	0.737 ± 0.024	0.161 ± 0.011	0.574 ± 0.030	0.377 ± 0.033	$\mathbf{0.479 \pm 0.012}$	$\mathbf{0.523 \pm 0.029}$
CIFAR-10 — 16 Shadow Models						
Final Loss	0.001 ± 0.002	0.001 ± 0.001	0.008 ± 0.004	0.013 ± 0.008	0.012 ± 0.005	0.036 ± 0.024
CSG (Ravikumar et al., 2026)	0.095 ± 0.021	0.055 ± 0.026	0.121 ± 0.027	0.195 ± 0.011	0.119 ± 0.028	0.318 ± 0.007
LT-Std (Pollock et al., 2025)	0.335 ± 0.066	0.181 ± 0.015	0.225 ± 0.054	0.362 ± 0.012	0.174 ± 0.046	0.465 ± 0.006
LT-IQR (Pollock et al., 2025)	0.336 ± 0.069	0.182 ± 0.016	0.220 ± 0.056	0.354 ± 0.011	0.173 ± 0.044	0.462 ± 0.013
CSL-Diff-40 (Ours)	$\mathbf{0.372 \pm 0.069}$	$\mathbf{0.202 \pm 0.019}$	$\mathbf{0.248 \pm 0.066}$	$\mathbf{0.396 \pm 0.011}$	$\mathbf{0.192 \pm 0.049}$	$\mathbf{0.515 \pm 0.008}$
CSL-Diff-50 (Ours)	0.369 ± 0.074	0.200 ± 0.018	0.244 ± 0.063	0.392 ± 0.012	0.192 ± 0.048	0.514 ± 0.009

Table 14: Extended Results: Precision@k and Recall@k predicting LiRA vulnerability for CIFAR-100 at FPR=10^{-3} across 64 and 16 shadow models. Averaged across 3 seeds.

Method	Prec@1%	Rec@1%	Prec@3%	Rec@3%	Prec@5%	Rec@5%
CIFAR-100 — 64 Shadow Models						
Final Loss	0.003 ± 0.002	0.000 ± 0.000	0.008 ± 0.006	0.002 ± 0.001	0.014 ± 0.009	0.005 ± 0.003
CSG (Ravikumar et al., 2026)	0.121 ± 0.027	0.009 ± 0.001	0.164 ± 0.021	0.036 ± 0.001	0.191 ± 0.032	0.071 ± 0.005
LT-Std (Pollock et al., 2025)	0.835 ± 0.034	0.062 ± 0.006	0.737 ± 0.029	0.165 ± 0.013	0.669 ± 0.031	0.249 ± 0.018
LT-IQR (Pollock et al., 2025)	0.873 ± 0.043	0.065 ± 0.006	0.780 ± 0.033	0.175 ± 0.015	0.708 ± 0.037	0.264 ± 0.017
CSL-Diff-40 (Ours)	0.896 ± 0.029	0.067 ± 0.005	0.801 ± 0.026	0.179 ± 0.015	0.734 ± 0.040	0.273 ± 0.017
CSL-Diff-50 (Ours)	$\mathbf{0.897 \pm 0.040}$	$\mathbf{0.067 \pm 0.005}$	$\mathbf{0.807 \pm 0.025}$	$\mathbf{0.181 \pm 0.016}$	$\mathbf{0.737 \pm 0.041}$	$\mathbf{0.275 \pm 0.017}$
CIFAR-100 — 16 Shadow Models						
Final Loss	0.000 ± 0.000	0.000 ± 0.000	0.001 ± 0.001	0.001 ± 0.001	0.003 ± 0.001	0.004 ± 0.002
CSG (Ravikumar et al., 2026)	0.028 ± 0.007	0.007 ± 0.002	0.064 ± 0.010	0.044 ± 0.009	0.075 ± 0.004	0.085 ± 0.009
LT-Std (Pollock et al., 2025)	0.541 ± 0.050	0.123 ± 0.008	0.398 ± 0.018	0.271 ± 0.014	0.331 ± 0.015	0.377 ± 0.019
LT-IQR (Pollock et al., 2025)	0.557 ± 0.021	0.127 ± 0.008	0.405 ± 0.011	0.277 ± 0.019	0.335 ± 0.011	0.381 ± 0.024
CSL-Diff-40 (Ours)	$\mathbf{0.579 \pm 0.036}$	$\mathbf{0.132 \pm 0.005}$	0.418 ± 0.014	0.285 ± 0.018	$\mathbf{0.355 \pm 0.015}$	$\mathbf{0.404 \pm 0.022}$
CSL-Diff-50 (Ours)	0.576 ± 0.011	0.131 ± 0.010	$\mathbf{0.426 \pm 0.014}$	$\mathbf{0.291 \pm 0.018}$	0.352 ± 0.012	0.401 ± 0.025

Table 15: Extended Results: Precision@k and Recall@k for CIFAR-10 at FPR=10^{-3} predicting vulnerable points defined by the Union of Attacks (64 and 16 shadows).

Method	Prec@1%	Rec@1%	Prec@3%	Rec@3%	Prec@5%	Rec@5%
CIFAR-10 — 64 Shadow Models						
Final Loss	0.089 ± 0.021	0.016 ± 0.004	0.101 ± 0.019	0.055 ± 0.009	0.110 ± 0.016	0.100 ± 0.016
CSG (Ravikumar et al., 2026)	0.253 ± 0.044	0.046 ± 0.010	0.294 ± 0.013	0.160 ± 0.015	0.284 ± 0.008	0.257 ± 0.015
LT-Std (Pollock et al., 2025)	0.667 ± 0.051	0.121 ± 0.010	0.526 ± 0.029	0.286 ± 0.020	0.442 ± 0.023	0.401 ± 0.023
LT-IQR (Pollock et al., 2025)	0.739 ± 0.034	0.134 ± 0.011	0.573 ± 0.017	0.312 ± 0.021	0.481 ± 0.010	0.436 ± 0.020
CSL-Diff-40 (Ours)	$\mathbf{0.772 \pm 0.026}$	$\mathbf{0.140 \pm 0.009}$	$\mathbf{0.611 \pm 0.031}$	$\mathbf{0.333 \pm 0.021}$	0.515 ± 0.022	0.467 ± 0.019
CSL-Diff-50 (Ours)	0.767 ± 0.028	0.139 ± 0.009	0.608 ± 0.041	0.331 ± 0.022	$\mathbf{0.518 \pm 0.021}$	$\mathbf{0.470 \pm 0.020}$
CIFAR-10 — 16 Shadow Models						
Final Loss	0.043 ± 0.006	0.014 ± 0.005	0.051 ± 0.011	0.048 ± 0.009	0.053 ± 0.005	0.085 ± 0.018
CSG (Ravikumar et al., 2026)	0.133 ± 0.019	0.044 ± 0.013	0.164 ± 0.032	0.156 ± 0.021	0.168 ± 0.032	0.265 ± 0.024
LT-Std (Pollock et al., 2025)	0.453 ± 0.102	0.142 ± 0.010	0.316 ± 0.071	0.297 ± 0.029	0.254 ± 0.058	0.398 ± 0.039
LT-IQR (Pollock et al., 2025)	0.472 ± 0.090	0.149 ± 0.022	0.335 ± 0.069	0.317 ± 0.041	0.278 ± 0.050	0.438 ± 0.036
CSL-Diff-40 (Ours)	0.520 ± 0.092	0.165 ± 0.020	$\mathbf{0.365 \pm 0.082}$	$\mathbf{0.344 \pm 0.039}$	0.296 ± 0.060	0.465 ± 0.031
CSL-Diff-50 (Ours)	$\mathbf{0.521 \pm 0.088}$	$\mathbf{0.165 \pm 0.022}$	0.362 ± 0.080	0.341 ± 0.035	$\mathbf{0.299 \pm 0.056}$	$\mathbf{0.471 \pm 0.036}$

Table 16: Extended Results: Precision@k and Recall@k for CIFAR-100 at FPR=10^{-3} predicting vulnerable points defined by the Union of Attacks (64 and 16 shadows).

Method	Prec@1%	Rec@1%	Prec@3%	Rec@3%	Prec@5%	Rec@5%
CIFAR-100 — 64 Shadow Models						
Final Loss	0.040 ± 0.017	0.003 ± 0.001	0.048 ± 0.018	0.010 ± 0.003	0.048 ± 0.018	0.016 ± 0.005
CSG (Ravikumar et al., 2026)	0.131 ± 0.024	0.009 ± 0.001	0.172 ± 0.018	0.035 ± 0.001	0.200 ± 0.030	0.067 ± 0.006
LT-Std (Pollock et al., 2025)	0.852 ± 0.042	0.058 ± 0.005	0.762 ± 0.025	0.155 ± 0.010	0.696 ± 0.028	0.235 ± 0.014
LT-IQR (Pollock et al., 2025)	0.889 ± 0.037	0.060 ± 0.004	0.801 ± 0.036	0.162 ± 0.011	0.730 ± 0.034	0.246 ± 0.013
CSL-Diff-40 (Ours)	0.911 ± 0.023	0.062 ± 0.004	0.823 ± 0.025	0.167 ± 0.011	0.759 ± 0.033	0.256 ± 0.013
CSL-Diff-50 (Ours)	$\mathbf{0.913 \pm 0.033}$	$\mathbf{0.062 \pm 0.004}$	$\mathbf{0.827 \pm 0.026}$	$\mathbf{0.168 \pm 0.012}$	$\mathbf{0.764 \pm 0.035}$	$\mathbf{0.258 \pm 0.013}$
CIFAR-100 — 16 Shadow Models						
Final Loss	0.035 ± 0.019	0.005 ± 0.002	0.042 ± 0.013	0.018 ± 0.008	0.038 ± 0.010	0.027 ± 0.011
CSG (Ravikumar et al., 2026)	0.044 ± 0.011	0.006 ± 0.001	0.082 ± 0.006	0.035 ± 0.006	0.094 ± 0.007	0.066 ± 0.008
LT-Std (Pollock et al., 2025)	0.608 ± 0.060	0.085 ± 0.004	0.481 ± 0.030	0.202 ± 0.014	0.407 ± 0.035	0.284 ± 0.015
LT-IQR (Pollock et al., 2025)	0.629 ± 0.026	0.088 ± 0.008	0.484 ± 0.022	0.203 ± 0.020	0.407 ± 0.028	0.284 ± 0.021
CSL-Diff-40 (Ours)	$\mathbf{0.660 \pm 0.042}$	0.092 ± 0.006	0.504 ± 0.026	0.211 ± 0.017	$\mathbf{0.436 \pm 0.034}$	$\mathbf{0.304 \pm 0.016}$
CSL-Diff-50 (Ours)	$\mathbf{0.660 \pm 0.045}$	$\mathbf{0.092 \pm 0.009}$	$\mathbf{0.513 \pm 0.030}$	$\mathbf{0.215 \pm 0.016}$	0.434 ± 0.032	0.303 ± 0.020

Table 17: Extended Results: Evaluation of artifact metrics at relaxed security thresholds (FPR=0.5% and 1.0%) on CIFAR-10 (256 shadows). LiRA vulnerability target.

Method	Prec@1%	Rec@1%	Prec@3%	Rec@3%	Prec@5%	Rec@5%
CIFAR-10 — Evaluated at FPR = 0.5%						
Final Loss	0.088 ± 0.014	0.008 ± 0.001	0.124 ± 0.021	0.034 ± 0.006	0.149 ± 0.026	0.067 ± 0.014
CSG (Ravikumar et al., 2026)	0.353 ± 0.021	0.032 ± 0.003	0.421 ± 0.021	0.115 ± 0.010	0.441 ± 0.011	0.200 ± 0.012
LT-Std (Pollock et al., 2025)	0.808 ± 0.024	0.073 ± 0.001	0.713 ± 0.011	0.194 ± 0.004	0.649 ± 0.011	0.293 ± 0.007
LT-IQR (Pollock et al., 2025)	0.851 ± 0.016	0.077 ± 0.004	0.743 ± 0.006	0.202 ± 0.009	0.673 ± 0.006	0.305 ± 0.013
CSL-Diff-40 (Ours)	0.911 ± 0.016	0.082 ± 0.002	$\mathbf{0.806 \pm 0.017}$	$\mathbf{0.219 \pm 0.004}$	0.736 ± 0.008	0.333 ± 0.009
CSL-Diff-50 (Ours)	$\mathbf{0.913 \pm 0.015}$	$\mathbf{0.083 \pm 0.002}$	0.804 ± 0.020	0.218 ± 0.004	$\mathbf{0.738 \pm 0.010}$	$\mathbf{0.334 \pm 0.008}$
CIFAR-10 — Evaluated at FPR = 1.0%						
Final Loss	0.103 ± 0.018	0.007 ± 0.002	0.152 ± 0.020	0.033 ± 0.005	0.177 ± 0.031	0.063 ± 0.013
CSG (Ravikumar et al., 2026)	0.384 ± 0.032	0.027 ± 0.003	0.466 ± 0.030	0.100 ± 0.009	0.490 ± 0.021	0.175 ± 0.012
LT-Std (Pollock et al., 2025)	0.833 ± 0.012	0.059 ± 0.001	0.755 ± 0.013	0.162 ± 0.003	0.700 ± 0.014	0.250 ± 0.004
LT-IQR (Pollock et al., 2025)	0.869 ± 0.024	0.062 ± 0.003	0.784 ± 0.014	0.168 ± 0.007	0.722 ± 0.010	0.258 ± 0.011
CSL-Diff-40 (Ours)	$\mathbf{0.933 \pm 0.008}$	$\mathbf{0.067 \pm 0.002}$	0.840 ± 0.014	0.180 ± 0.003	0.788 ± 0.013	0.281 ± 0.005
CSL-Diff-50 (Ours)	0.932 ± 0.022	0.067 ± 0.001	$\mathbf{0.844 \pm 0.018}$	$\mathbf{0.181 \pm 0.002}$	$\mathbf{0.789 \pm 0.015}$	$\mathbf{0.281 \pm 0.004}$

Table 18: Extended Results: Evaluation of artifact metrics at relaxed security thresholds (FPR=0.5% and 1.0%) on CIFAR-100 (256 shadows). LiRA vulnerability target.

Method	Prec@1%	Rec@1%	Prec@3%	Rec@3%	Prec@5%	Rec@5%
CIFAR-100 — Evaluated at FPR = 0.5%						
Final Loss	0.008 ± 0.007	0.000 ± 0.000	0.018 ± 0.017	0.002 ± 0.002	0.037 ± 0.024	0.007 ± 0.004
CSG (Ravikumar et al., 2026)	0.245 ± 0.037	0.009 ± 0.001	0.287 ± 0.026	0.033 ± 0.001	0.317 ± 0.041	0.060 ± 0.004
LT-Std (Pollock et al., 2025)	0.937 ± 0.018	0.036 ± 0.002	0.895 ± 0.016	0.102 ± 0.005	0.851 ± 0.028	0.162 ± 0.005
LT-IQR (Pollock et al., 2025)	0.963 ± 0.021	0.037 ± 0.002	0.916 ± 0.019	0.105 ± 0.004	0.880 ± 0.027	0.168 ± 0.005
CSL-Diff-40 (Ours)	$\mathbf{0.977 \pm 0.006}$	$\mathbf{0.037 \pm 0.002}$	0.940 ± 0.011	0.108 ± 0.005	0.907 ± 0.020	0.173 ± 0.007
CSL-Diff-50 (Ours)	0.975 ± 0.006	0.037 ± 0.002	$\mathbf{0.942 \pm 0.007}$	$\mathbf{0.108 \pm 0.006}$	$\mathbf{0.908 \pm 0.024}$	$\mathbf{0.173 \pm 0.006}$
CIFAR-100 — Evaluated at FPR = 1.0%						
Final Loss	0.013 ± 0.013	0.000 ± 0.000	0.028 ± 0.021	0.003 ± 0.002	0.049 ± 0.029	0.008 ± 0.004
CSG (Ravikumar et al., 2026)	0.285 ± 0.038	0.009 ± 0.001	0.331 ± 0.028	0.031 ± 0.002	0.366 ± 0.040	0.058 ± 0.004
LT-Std (Pollock et al., 2025)	0.949 ± 0.019	0.030 ± 0.001	0.917 ± 0.015	0.087 ± 0.002	0.883 ± 0.019	0.140 ± 0.003
LT-IQR (Pollock et al., 2025)	0.971 ± 0.021	0.031 ± 0.001	0.934 ± 0.016	0.089 ± 0.002	0.905 ± 0.020	0.144 ± 0.003
CSL-Diff-40 (Ours)	$\mathbf{0.987 \pm 0.009}$	$\mathbf{0.031 \pm 0.001}$	0.955 ± 0.009	0.091 ± 0.003	$\mathbf{0.935 \pm 0.013}$	$\mathbf{0.148 \pm 0.003}$
CSL-Diff-50 (Ours)	0.983 ± 0.009	0.031 ± 0.001	$\mathbf{0.956 \pm 0.007}$	$\mathbf{0.091 \pm 0.003}$	0.934 ± 0.017	0.148 ± 0.003

C.5 Zero-overhead Adversarial Vulnerability Auditing: Setup and Analysis

Table 19: Spearman rank correlation (ρ)(mean $\pm$ std over 3 seeds) predicting C&W attack (Carlini & Wagner, 2016) across CIFAR-10 and CIFAR-100 on ResNet-18. Best results are in **bold**; second best are underlined.

Dataset	PGD-3	PGD-5	PGD-10	PGD-20	$\bar{d}_{\text{adv,T}}$
CIFAR-10	0.5073 ± 0.0166	0.5473 ± 0.0851	0.6074 ± 0.0123	$\underline{0.6470 \pm 0.0151}$	$\mathbf{0.7204 \pm 0.0073}$
CIFAR-100	0.4460 ± 0.0233	0.4798 ± 0.0252	$\underline{0.5046 \pm 0.0092}$	0.4933 ± 0.0189	$\mathbf{0.5063 \pm 0.0232}$

Diagnostic Rationale. We evaluate whether trajectory-based training signals can accurately rank individual training samples by their vulnerability to adversarial perturbations without executing computationally prohibitive per-sample attack optimizations. The central hypothesis, grounded in Theorem 2, posits that a sample's local adversarial distance can be bounded and approximated by a geometric proxy derived from its Cumulative Sample Loss (CSL) and Cumulative Sample Gradient (CSG). If this zero-overhead proxy correlates strongly with the true adversarial margin, practitioners can preemptively identify vulnerable samples as a natural byproduct of standard training, bypassing the massive $O(N \times \text{PGD steps})$ cost of per-sample boundary searches when using PGD based methods as metrics.

Experimental Setup and Proxy Formulation. We maintain the core experimental setup, architectures, and optimization schedules detailed in Appendix C.1. Specifically, we evaluate ResNet-18 models (He et al., 2016) trained on the full CIFAR-10 and CIFAR-100 datasets (Krizhevsky, 2009) over 200 epochs. For each training instance $\vec{z}_i$, we rank its adversarial vulnerability using the proxy:

$$\bar{d}_{\text{adv},T}(\vec{z}_i) = \frac{2\left(L_u T - \text{CSL}(\vec{z}_i)\right)}{\text{CSG}(\vec{z}_i)} \tag{128}$$

where L_u represents the cross-entropy misclassification threshold and T indicates the specific evaluation horizon. We set $T = 200$ here.

Ground Truth and Empirical Baselines. The ground-truth adversarial distance targets are established on the final converged models. If a sample is already misclassified by the clean model, its adversarial distance is assigned a value of zero. For correctly classified samples, we employ an untargeted Carlini & Wagner (C&W) l_2 optimization attack (Carlini & Wagner, 2016) using the same configurations as described in the Validation experiments in C.1. For the baselines we are using Projected Gradient Descent (PGD) (Madry et al., 2017) combined with a 25-step binary search over the perturbation radius ε (Pintor et al., 2021; Tramer et al., 2020). We evaluate our proxy against baselines with PGD step sizes varying as 3, 5, 10, 20.

Main Results and Temporal Ablation. As reported in Table 19, our trajectory proxy successfully outperforms even the highly computationally expensive PGD-20 baselines for the most cases while being a fraction of the cost. The PGD-20 using 25 step binary search requires additional 500 backward passes in the worst case, while our metric doesn't even need an additional backwards pass.

Computational Efficiency. The practical utility of $\bar{d}_{\text{adv},T}$ lies in its computational asymmetry compared to traditional auditing. CSL is extracted entirely for free from the standard forward pass, while CSG requires only one supplementary matrix multiplication per sample to extract the input-space gradient norm during backpropagation (Ravikumar et al., 2026). The proxy ranking is then computed instantaneously post-hoc. By contrast, computing the true per-sample margins via PGD requires iteratively computing full backward passes over 25 binary search steps $\times$ the number of PGD steps, rendering it orders of magnitude more expensive. Our results demonstrate that tracking loss and gradient dynamics provides a highly scalable, zero-overhead alternative to adversarial optimization.

C.6 COMPUTE RESOURCES:

All of the experiments were performed on a single A40 NVIDIA GPU with 48GB of VRAM. However, the results can be replicated with a single GPU with as little as 8GB of VRAM.