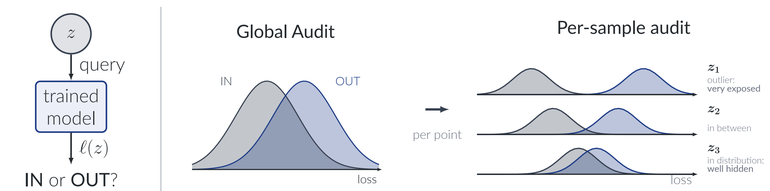
Assessing Per-Sample Membership Inference Vulnerability without Retraining

Valentin Dorseuil [1] Jamal Atif [2] Olivier Cappé [1]

[1]DI ENS, École normale supérieure, Université PSL, CNRS
[2]CMAP, École polytechnique, Institut Polytechnique de Paris

Motivation

- A **Membership Inference Attack (MIA)** [5] aims to determine whether a specific data point z was used to train a model.
- In the black-box setting, MIAs often use the point's **loss** $\ell(z)$ to distinguish training (IN) from non-training (OUT) points (Fig. 1a).
- MIAs are used to audit empirical privacy leakage, specifically in differentially private (DP) models [2, 3].



(a) Black-box MIA scheme.

(b) Aggregated IN and OUT loss distributions of all points, estimated with a single model.

(c) IN and OUT loss distributions of each point, estimated with multiple shadow models.

Figure 1. Membership inference: black-box attack scheme (a); global audit (b); per-sample audit (c).

- Privacy risk is **highly heterogeneous**: a few outliers are far more exposed than typical samples (Figs. 1b-1c).
- Per-sample audits, using attacks like LiRA [1] or RMIA [6], typically require **hundreds of shadow models**.

Question

Can the privacy vulnerability of individual training points be assessed without training shadow models?

Toy model: Gaussian Linear Regression

Which points does a membership attack expose most, and why?

We study a Gaussian linear regression model, in which per-sample vulnerability can be analyzed explicitly.

- Fix a **target point** $z_i = (x_i, y_i)$.
- To simulate the shadow model setting, draw the remaining $n-1$ samples from:

$$X_j \sim \mathcal{N}(0, \Sigma), \qquad Y_j \mid X_j \sim \mathcal{N}(X_j^\top \theta, \sigma^2).$$

- Fit OLS models with and without z_i, and denote their losses on the target by ℓ_i (IN) and ℓ_{-i} (OUT).
- The quantity of interest is the **per-sample loss gap** $\ell_{-i} - \ell_i$, exploited by most MIAs.

Proposition

The expected loss gap between the OUT and IN models satisfies

$$\mathbb{E}[\ell_{-i} - \ell_i] = \frac{2\varepsilon_i^2 \bar{h}_{ii}}{n-p} + \mathcal{O}\left(\frac{1}{(n-p)^2}\right).$$

Result: the loss gap is governed by leverage $\times$ error^2

- The structural **error** $\varepsilon_i = y_i - x_i^\top \theta$ measures the target's deviation from the true regression mean.
- The population **leverage score** $\bar{h}_{ii} = x_i^\top \Sigma^{-1} x_i$ measures how unusual its features are.
- The expected loss gap is largest when **both quantities are large**, and decreases as $n - p$ grows.

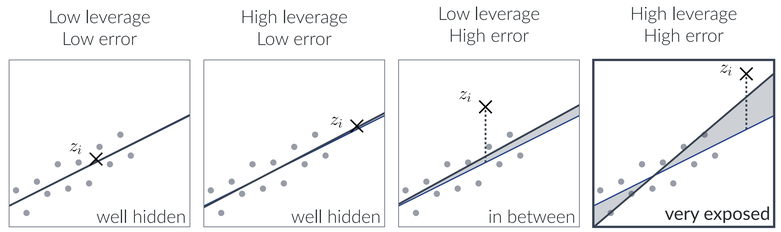


Figure 2. Effect of a target point's leverage and structural error on the fitted regression line. The largest shift occurs when both are high.

- In this model, population leverage $\bar{h}_{ii}$ and squared structural error ε_i^2 govern the **per-sample vulnerability** to loss-based MIAs.
- The most exposed points are **outliers in the feature space** and have **large structural errors**.

From linear models to deep networks

Can we estimate this gap in deep networks without retraining?

- **Two estimators.** We use influence functions [4] and a Newton-step approximation to estimate the loss gap without retraining.
- **Last-layer approximation.** Treating the final network layer as a linear model allows us to derive computationally cheap vulnerability scores.

We derive closed-form loss-gap estimates for regression (with direct extensions to classification):

$$\widehat{\Delta\ell}_i^{\mathrm{IF}} = 2\, e_i^2\, h_{ii} \qquad \text{and} \qquad \widehat{\Delta\ell}_i^{\mathrm{NS}} = \frac{2\, e_i^2\, h_{ii}}{1 - h_{ii}}$$

where:

- $h_{ii} = x_i^\top (X^\top X)^{-1} x_i$ (**leverage**): Distance of x_i from the rest of the data.
- $e_i = y_i - x_i^\top \hat{\theta}$ (**residual**): The model's prediction error.

These quantities are the **observable counterparts** of the theory: $e_i \approx \varepsilon_i$ and $h_{ii} \approx \bar{h}_{ii}/n$, so $\widehat{\Delta\ell}_i^{\mathrm{IF}} \approx 2\varepsilon_i^2 \bar{h}_{ii}$ matches the theoretical model's proposition.

Our approach

Evaluate $\widehat{\Delta\ell}_i^{\mathrm{IF}}$ and $\widehat{\Delta\ell}_i^{\mathrm{NS}}$ on the final layer of a single model to generate computationally cheap, per-sample vulnerability scores for membership inference.

Experiments: recovering LiRA's most vulnerable points

Comparison across **8 dataset–architecture pairs.** LiRA uses **200 shadow models**; our scores use **one**.

- Measure the overlap between LiRA's and each score's top-$k\%$ most vulnerable samples, as k varies.
- Higher recall means better recovery. The diagonal represents random ranking.
- Comparison with two other cheap baselines: the loss ℓ_i and the gradient norm $\|\nabla_\theta \ell_i\|$.

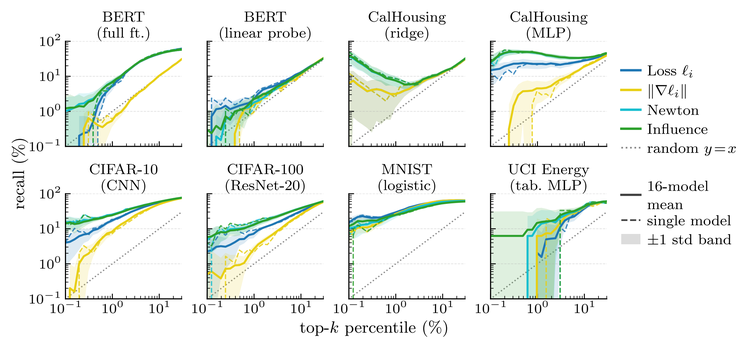


Figure 3. Recall of LiRA's top-$k\%$ most vulnerable samples recovered by each surrogate (log–log scale). Dotted diagonal: random ranking.

Our scores outperformed other cheap baselines on 7 out of 8 models, recovering up to 90% of the top 1% most LiRA-vulnerable points (when retaining the top 5% of points), using only **one model**.

Key findings

$7 / 8$	10^3–$10^4\times$	1
settings beating the loss baseline	faster than LiRA	trained model, no retraining

Conclusion

- **Main idea:** Per-sample MIA risk is well approximated by **loss $\times$ leverage**.
- **Limitation:** The theory is derived from linear models, and the deep-network score is only a surrogate.
- **Uses:** Canary selection, first-pass audits, and per-sample privacy training.

References

[1] N. Carlini et al. "Membership Inference Attacks From First Principles". In: *IEEE Symposium on Security and Privacy (SP)*. 2022, pp. 1897–1914.

[2] C. Dwork. "Differential Privacy". In: *Automata, Languages and Programming*. Springer, 2006, pp. 1–12.

[3] M. Jagielski et al. "Auditing Differentially Private Machine Learning: How Private Is Private SGD?" In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 22205–22216.

[4] P. W. Koh et al. "Understanding Black-box Predictions via Influence Functions". In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 2017, pp. 1885–1894.

[5] R. Shokri et al. "Membership Inference Attacks Against Machine Learning Models". In: *2017 IEEE Symposium on Security and Privacy (SP)*. 2017, pp. 3–18.

[6] S. Zarifzadeh et al. "Low-Cost High-Power Membership Inference Attacks". In: *Proceedings of the 41st International Conference on Machine Learning*. 2024, pp. 58244–58282.

Paper

Scan to read the paper.

Project funded as part of France 2030 by the ANR under reference ANR-22-PESN-0014.