Curvature Clues: Decoding Deep Learning Privacy with Input Loss Curvature

Deepak Ravikumar, Efstathia Soufleri, Kaushik Roy

Nano(Neuro) Electronics Research Laboratory

NEURAL INFORMATION PROCESSING SYSTEMS

2024 Spotlight Paper

CoCoSys
CENTER FOR THE
CO-DESIGN OF COGNITIVE SYSTEMS

Semiconductor
Research
Corporation

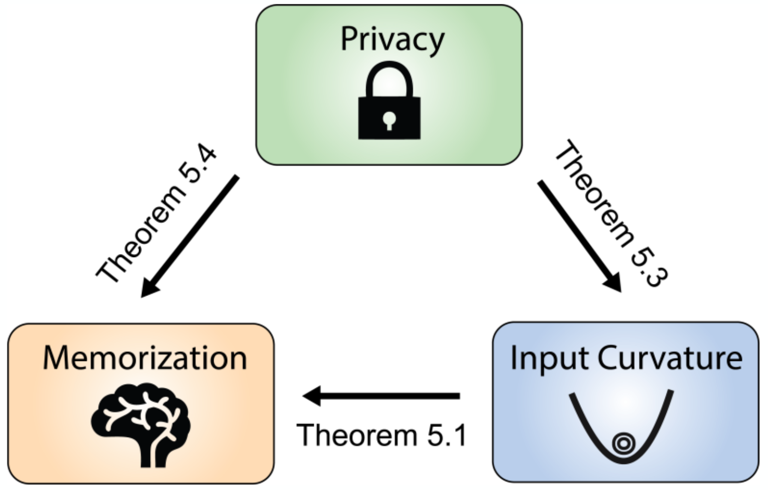
NSF
U.S. National
Science
Foundation

PURDUE
UNIVERSITY

Problem Motivation

Why Study Curvature?

- Deep neural networks often process sensitive data, and privacy is crucial.
- Recent research [1, 2] has shown the link between privacy, input curvature and memorization.
- Existing studies focus on loss curvature w.r.t training data.
- No/limited understanding of input loss curvature on unseen/test data.
- Question:
  - What does curvature mean for unseen test data?
  - What privacy properties does it capture?
  - What can it do?



Result from [1] showing link on training data

1. Deepak Ravikumar, Efstathia Soufleri, Abolfazl Hashemi, and Kaushik Roy. “Unveiling Privacy, Memorization, and Input Curvature Links", In ICML 2024
2. Isha Garg, Deepak Ravikumar, and Kaushik Roy. “Memorization through the lens of curvature of loss function around samples.” In 2024 ICML Spotlight.

Input Loss Curvature

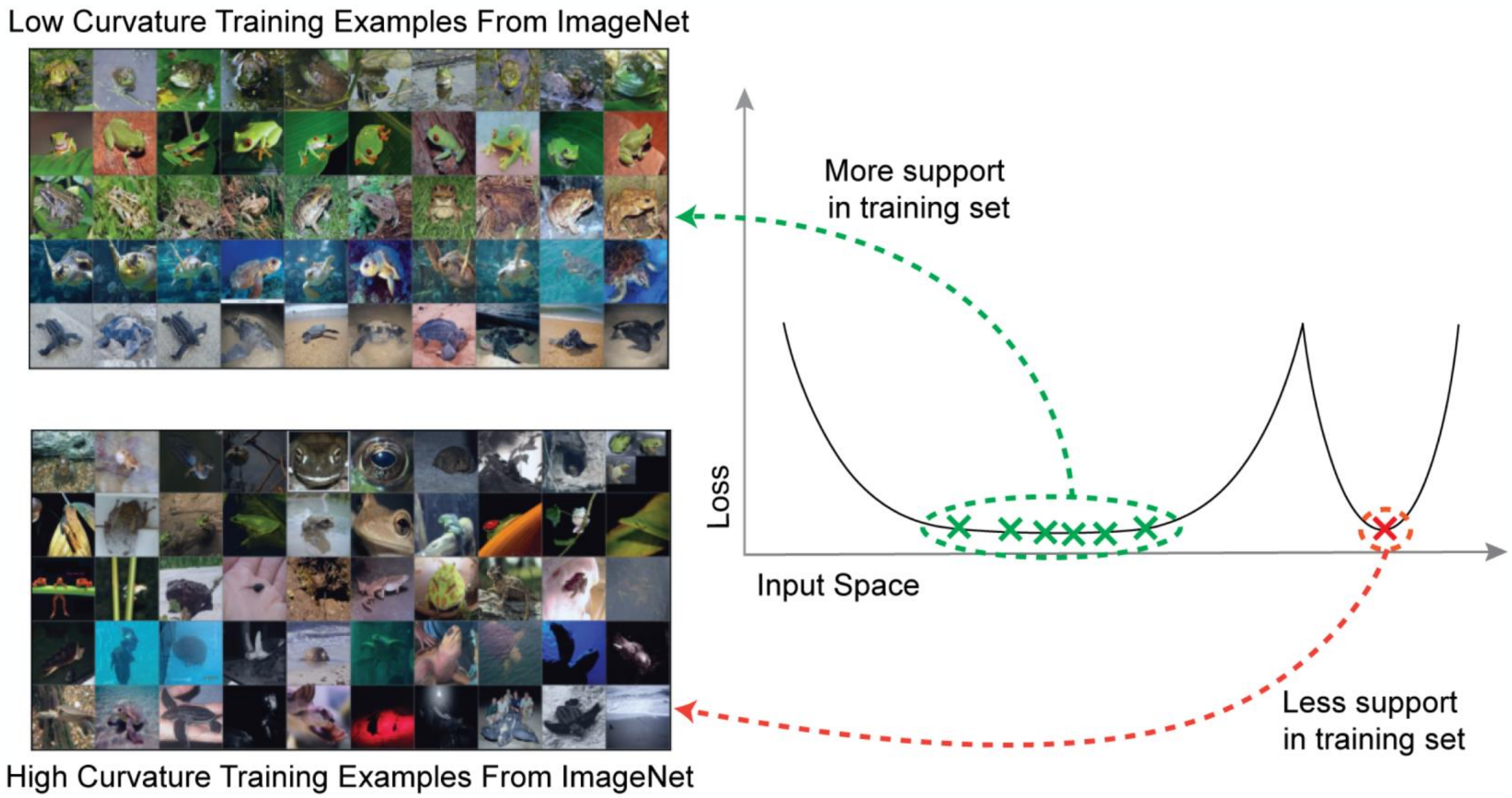
What is input loss curvature?

Definition: Input loss curvature is the trace of the Hessian of the loss with respect to input, representing model sensitivity to specific inputs.

Figure 1 Overview: Visualization of low and high curvature samples in ImageNet.

Low Curvature: Easy, prototypical, well-represented in the dataset.

High Curvature: Hard, atypical images with limited representation in the dataset.



Low Curvature Training Examples From ImageNet
More support
in training set
Loss
Input Space
Less support
in training set
High Curvature Training Examples From ImageNet

Contributions

Curvature, Test sets, MIA, Zero Order Estimation

Key Insight: Test data has higher input loss curvature than training data due to unoptimized regions in the loss landscape.

Theoretical Framework: We develop an upper bound on the KL divergence for train-test distinguishability in membership inference attacks.

Proposed Privacy Test:

- A new MIA using input loss curvature, aiming to surpass existing black-box methods.
- Propose Zero-order estimation to estimate curvature in a black-box setting.

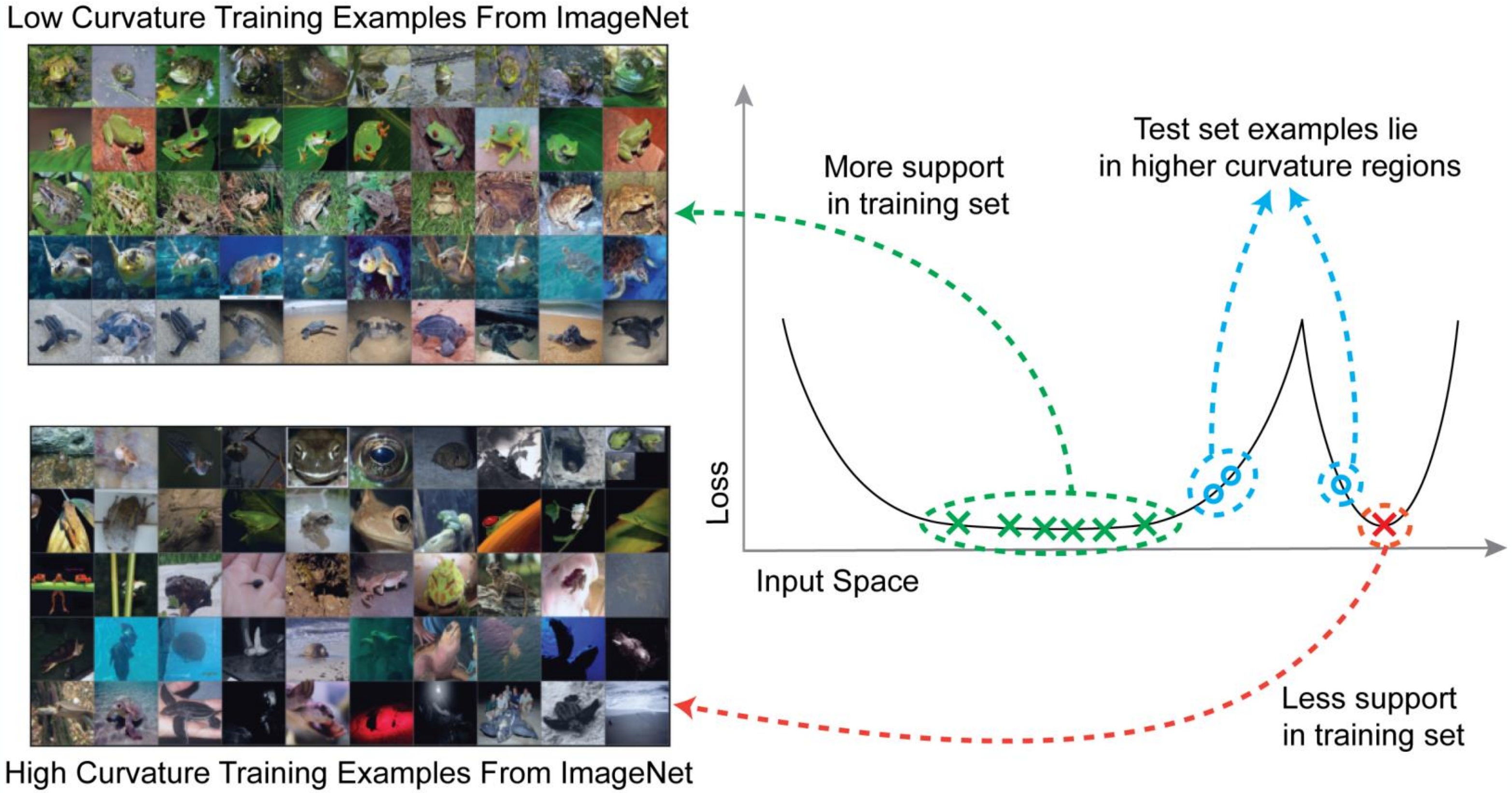

Low Curvature Training Examples From ImageNet
More support
in training set
Test set examples lie
in higher curvature regions
Loss
Input Space
Less support
in training set
High Curvature Training Examples From ImageNet

Figure 1: Visualizing low and high input curvature samples from a ResNet50 trained on ImageNet. Low input curvature training set images are prototypical and have lots of support in the trainset, while high input curvature train set examples have less support and are atypical. Test set examples lie around the training set images in higher curvature regions.

Theoretical Bounds for MIA Performance

Performance bounds using KL divergence

Theorem 4.1 Privacy bounds Train-Test KL Divergence

Theorem 4.2 Dataset Size and Privacy bound Curvature KL Divergence

$$D_{KL}\left(p_c(\phi, S, z_i) \,\|\, p_c\left(\phi, S^{\setminus i}, z_i\right)\right) \le \frac{[Lm(1 - e^{-\epsilon}) + c\,]^2}{2\,\sigma^2}$$

$$c = (4m - 1)\gamma + 2\,(m-1)\Delta + \frac{\rho}{6}\mathbb{E}[\|\alpha\|^3] + L$$

Theorem 4.3 Dataset Size and Curvature MIA Performance

The above two theorems (4.1, 4.2) give the best case performance bound, using the upper bounds we can say that the performance of MIA using curvature scores exceeds that of confidence scores with a probability at least 1 − δ when $m > \frac{(\sqrt{2\sigma^2\epsilon}) - c}{L(1 - e^{-\epsilon})}$

Theory to Practice

Zero-order estimation

Challenge: Direct calculation of input loss curvature requires model parameters, which are unavailable in black-box settings.

Solution: Use Zero-order (zo) estimation to approximate curvature scores without parameter access.

Implementation: To compute zo-curvature scores for MIA, we propose using the finite differences method. The Hessian can be estimated using the following formula:

$$\nabla^2 f(z_i) = n^2 \frac{f(z_i + h\,v + h\,u) - f(z_i - h\,v + h\,u) - f(z_i + h\,v - h\,u) + f(z_i - h\,v - h\,u)}{4\,h^2} uv^\top$$

ZO Curvature MIA Results

ZO MIA

Method	ImageNet		CIFAR100		CIFAR10	
	Bal Acc.	AUROC	Bal Acc.	AUROC	Bal Acc.	AUROC
Curv ZO NLL (Ours)	**69.16 ± 0.08**	**77.45 ± 0.09**	**84.47 ± 0.21**	**93.49 ± 0.18**	**61.92 ± 0.87**	**68.82 ± 1.30**
Curv ZO LR (Ours)	68.76 ± 0.04	72.28 ± 0.04	80.48 ± 0.10	90.15 ± 0.04	55.00 ± 0.17	58.89 ± 0.38
Carlini et al. (2022)	66.14 ± 0.01	73.46 ± 0.02	81.55 ± 0.13	88.89 ± 0.16	58.23 ± 0.29	61.73 ± 0.32
Yeom et al. (2018)	58.50 ± 0.02	63.23 ± 0.03	76.29 ± 0.39	82.11 ± 0.31	55.57 ± 0.52	60.44 ± 0.75
Sablayrolles et al. (2019)	66.93 ± 0.05	76.50 ± 0.04	70.22 ± 0.41	81.11 ± 0.39	56.65 ± 0.56	61.50 ± 0.79
Watson et al. (2021)	61.40 ± 0.06	69.44 ± 0.05	62.71 ± 0.31	71.66 ± 0.50	54.86 ± 0.59	58.58 ± 0.86
Ye et al. (2022)	66.16 ± 0.02	75.79 ± 0.05	80.73 ± 0.24	90.88 ± 0.19	59.62 ± 0.84	67.30 ± 1.25
Song et al. (2021)	57.88 ± 0.03	63.29 ± 0.03	75.58 ± 0.29	82.28 ± 0.27	55.63 ± 0.61	60.42 ± 0.85

Table 1: Comparison of the proposed curvature score based MIA with prior methods tested on ImageNet, CIFAR100, and CIFAR10 datasets. Results reported are the mean ± std obtained over 3 seeds. For CIFAR10 and CIFAR100 64 shadow models were used and 52 for ImagNet.

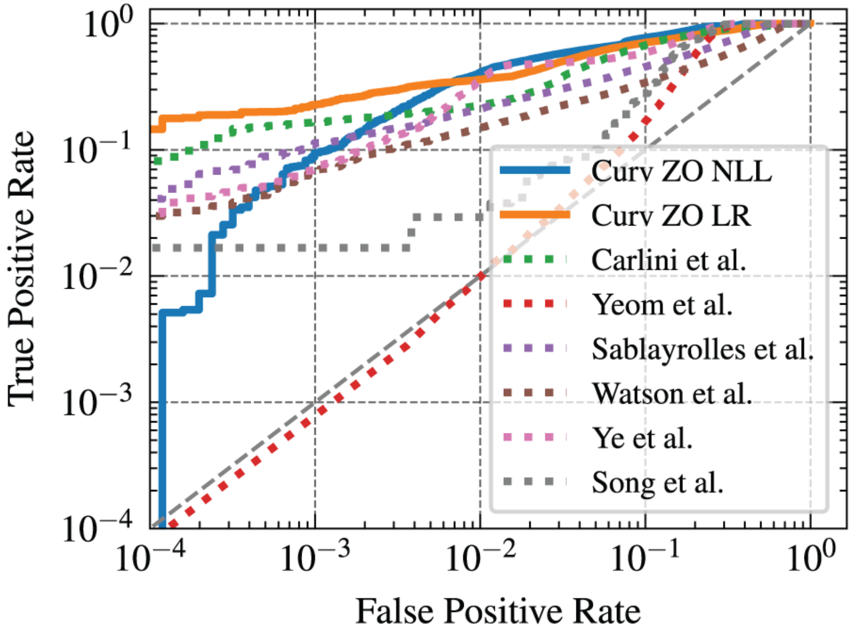


Figure 2: Comparing our method against existing techniques at low FPR. The proposed parametric Curv LR technique has the highest TPR at very low FPR.

Validating Theory

(a) Validating the effect of Privacy

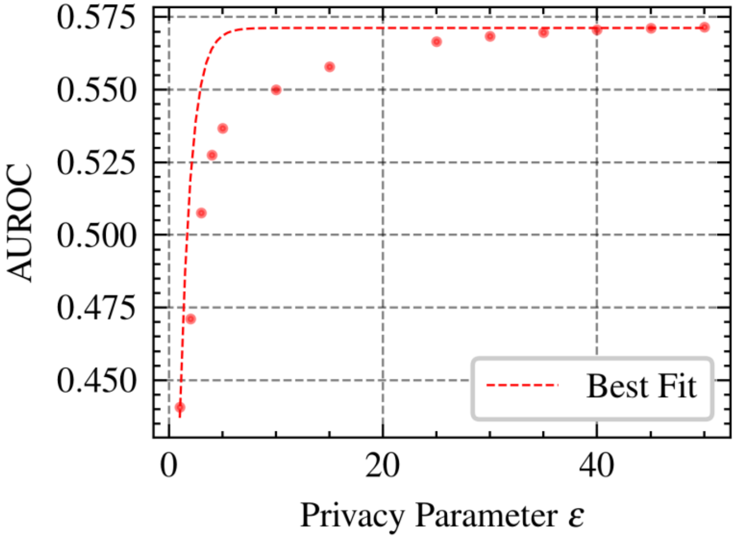


Figure 3: Validating the upper bound from Theorem 4.2 by fitting the MIA performance (AUROC of Curv LR) on DP-SGD trained models for various privacy parameters ϵ values.

Takeaways: We see that the theoretical prediction from Theorem 4.2 about MIA performance is well matched. Since Theorem 4.2 provides an upper bound, the results validate the theory.

(b) Validating the effect of Dataset Size

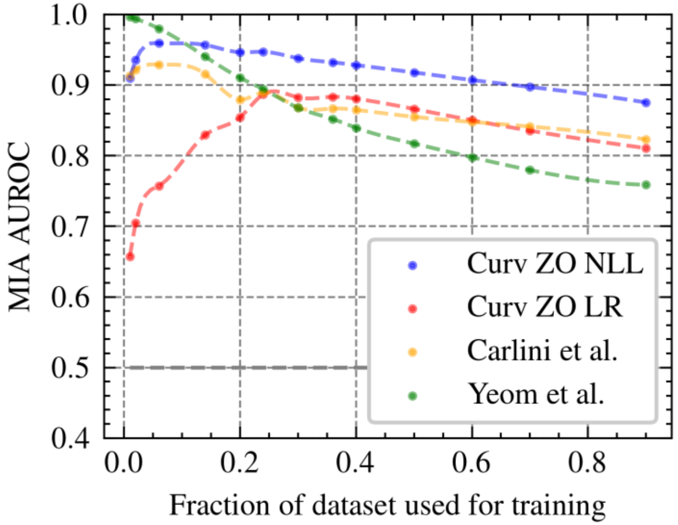


Figure 4: Visualizing MIA performance as a function of the size of the train set, which is randomly sampled.

Takeaways: We validate Theorem 4.3, noting that curvature-based MIA methods outperform probability-based ones beyond certain dataset sizes: Curv ZO LR excels with subsets of 30-40% of the training set, and Curv ZO NLL outperforms prior methods beyond 10%.

Conclusion

This study explores input loss curvature and its role in improving membership inference attacks (MIA). Theoretical analysis and experiments confirm that curvature-based MIA surpasses probability-based methods, particularly on large datasets. A novel zero-order estimation method enables efficient black-box MIA, achieving state-of-the-art results on CIFAR and ImageNet.

Contributions

- **Theoretical Insights**: Demonstrates that curvature scores enhance train-test distinguishability, improving MIA performance.
- **Practical Implementation**: Introduced zero-order curvature estimation for black-box settings and new MIA methods (Curv ZO NLL/LR).
- **Experimental Validation**: Validates theoretical predictions and shows superior performance of curvature-based MIA across datasets.

QR Code for code and paper

Thank You

Deepak Ravikumar, dravikum@purdue.edu

PURDUE UNIVERSITY®