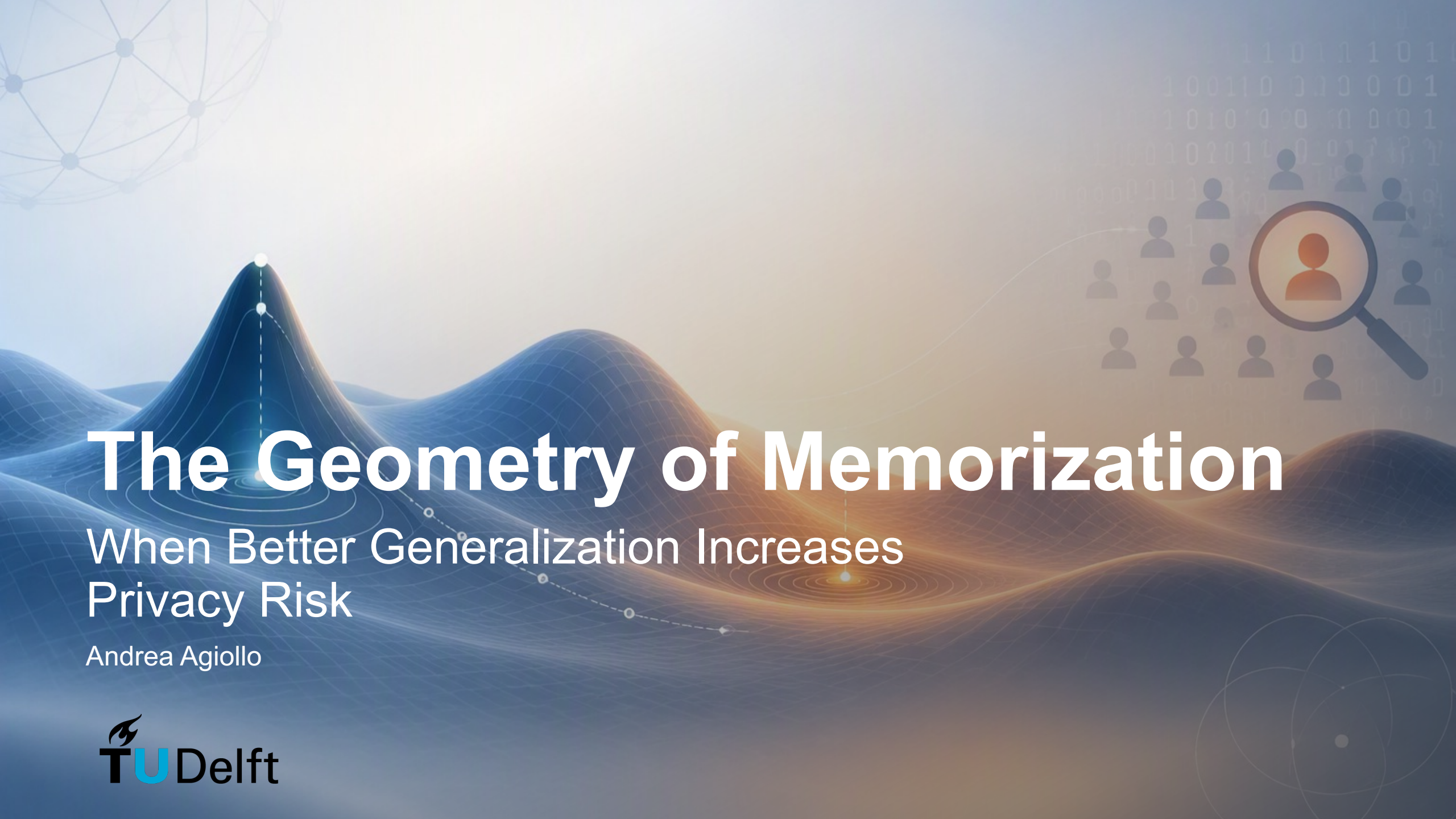


The Geometry of Memorization

When Better Generalization Increases Privacy Risk

Andrea Agiollo

TU Delft

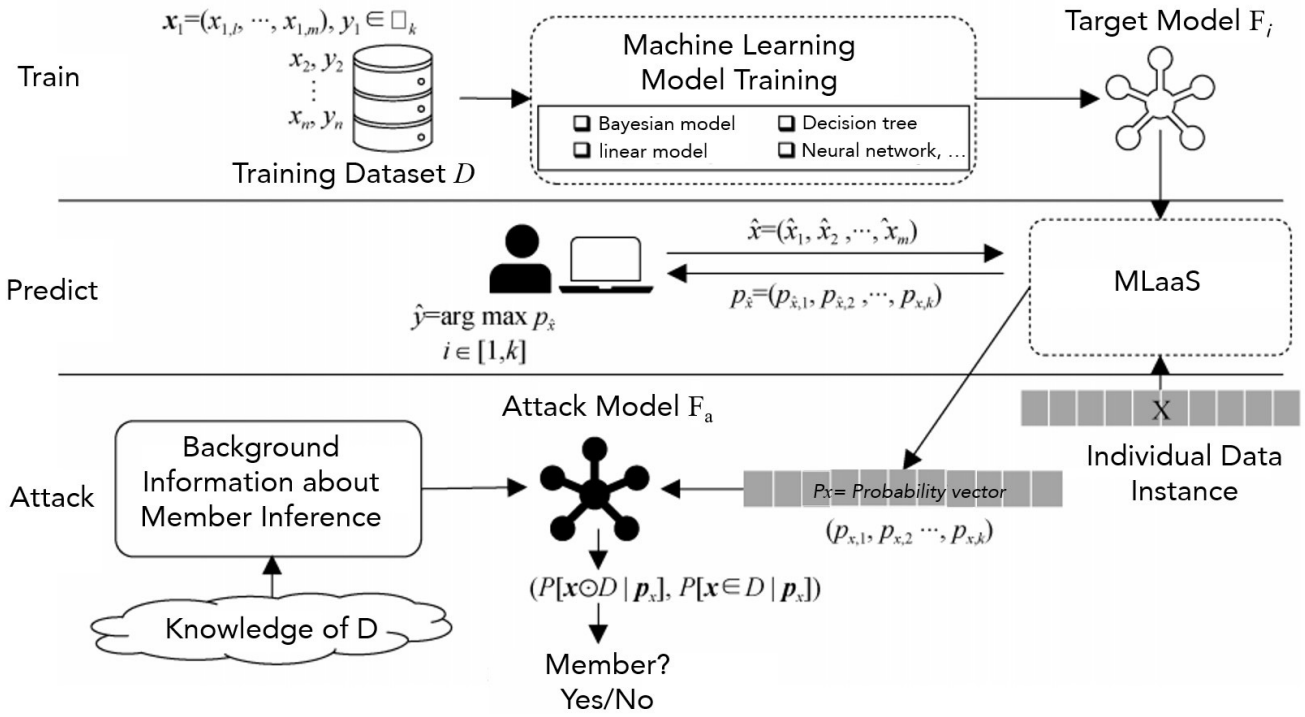
Membership Inference Attacks

The context

- MLaaS scenario
- Adversary plan to find out if data x was used in training the model
- Adversary queries model with data x and other data to decide if x is in training set or not

Why MIAs?

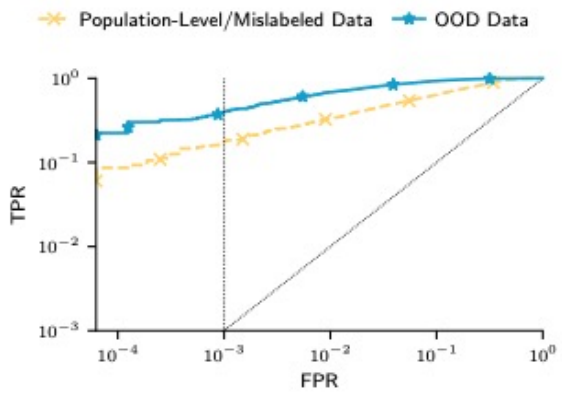
- Clinical data: was a patient treated for a specific condition?
- Behavioral data: did a user visit a sensitive place or use a particular service?
- Preliminary step for several other privacy leaks



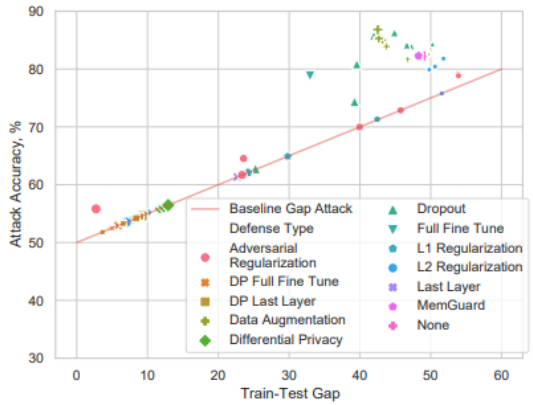
Niu et al. "A survey on membership inference attacks and defenses in machine learning." Journal of Information and Intelligence

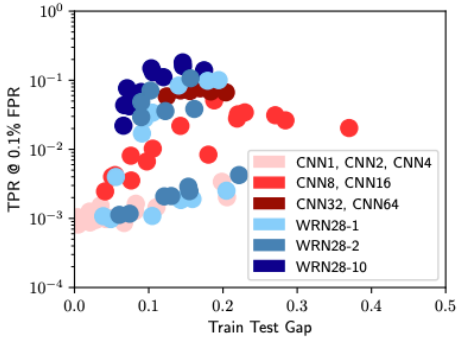
Membership Inference Attacks (Part II)

- Naïve membership attacker query the model using the target sample itself (loss, confidence, entropy, likelihood ratio, etc.)
- Depending on attacker information some data might be easier to infer
- Informative membership evidence may be found in the model's behavior around the target sample
- *Underlying question:* what properties of the ML training process make MIA evidence stronger in the first place?
- Common intuition connects privacy with generalization: reducing generalization gap should reduce risk
- Two models with similar generalization performance depend on individual training records in different ways
- Need to understand sample-level dependence of training









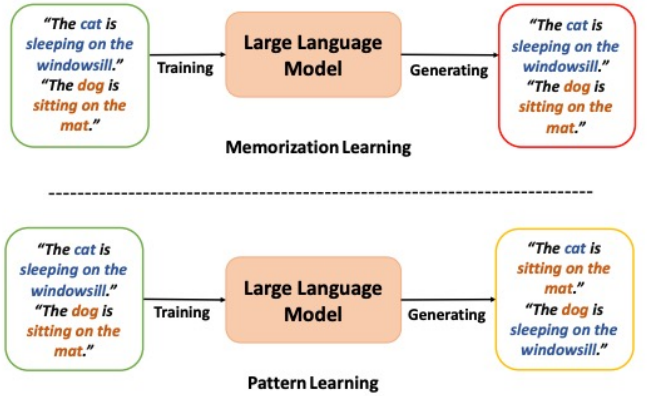
Aerni et al. "Evaluations of machine learning privacy defenses are misleading." CCS 2024.
Carlini et al. "Membership inference attacks from first principles." IEEE S&P 2022.
Choquette-Choo et al. "Label-only membership inference attacks." ICML 2021.

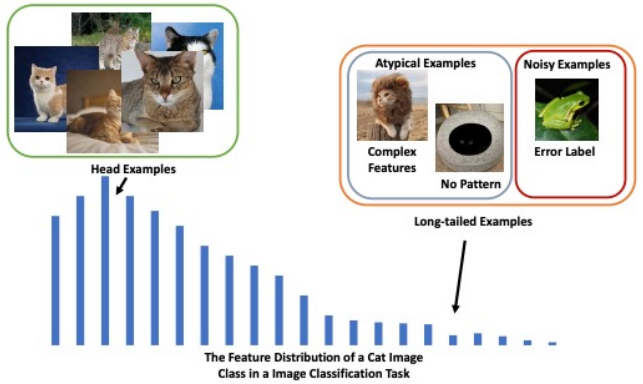
MIAs and Memorization

- Start investigation of training from memorization
- Stability-based definition of memorization: how much including sample x in training change model's behavior on that same record?

$$m(\mathbf{x}_i) = P\left[h^{\mathcal{A}}_{\mathcal{T}}(\mathbf{x}_i) = y_i\right] - P\left[h^{\mathcal{A}}_{\mathcal{T}\setminus i}(\mathbf{x}_i) = y_i\right]$$

- *Important nuance:* memorization support generalization when it corresponds to rare but reusable patterns
- Dependence on particular record leave detectable trace of whether that record was used in training
- Memorization and membership leakage are related
- Measuring memorization is impractical: thousands of model trainings to get a sufficient estimate
- Is it possible to generalize on memorization samples and keep model private (generalization vs privacy gap)?





Theorem 5.4 (Privacy $\Longrightarrow$ Less Memorization). *Let $\mathcal{A}$ be an ϵ-differentially private algorithm and z_i be the i^{th} element of $S \in \mathcal{Z}^m$. Then, we have*

$$\forall i \in \{1, \cdots, m\}, \quad \text{mem}(\mathcal{A}, S, i) \leq 1 - e^{-\epsilon}. \qquad (12)$$

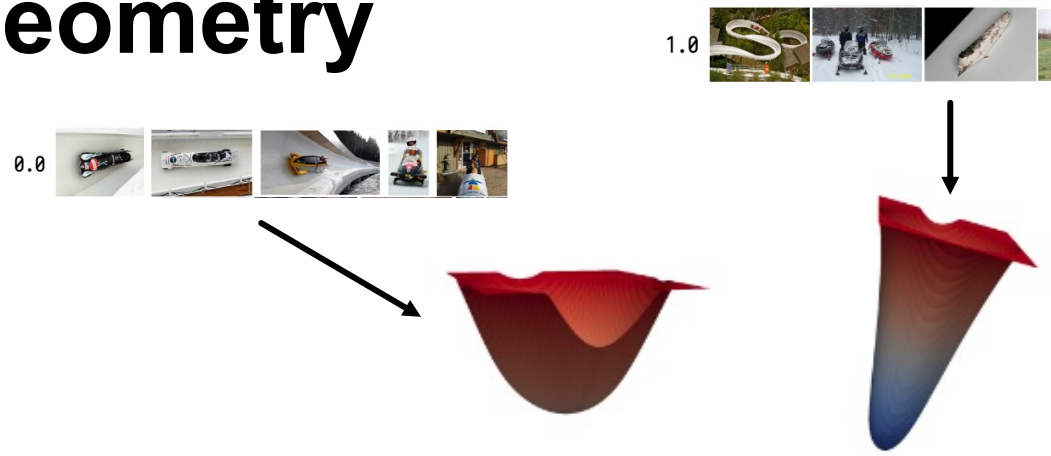
Feldman et al. "What neural networks memorize and why: Discovering the long tail via influence estimation." NeurIPS 2020.
Wei et al. "Memorization in deep learning: A survey" ACM Computing Surveys 2025.
Ravikumar et al. "Unveiling Privacy, Memorization, and Input Curvature Links" ICML 2024.

Memorization and Loss Geometry

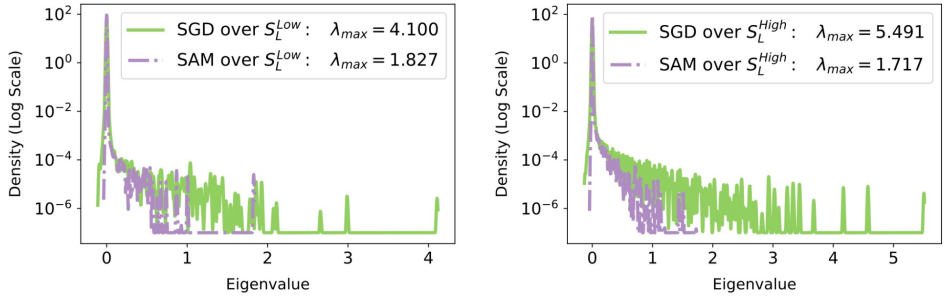
- Low memorization data points implicitly trigger flatter minima in SGD
- Sharpness Aware Minimization (SAM):

$$\min_{\boldsymbol{w}} L_S^{SAM}(\boldsymbol{w}) + \lambda ||\boldsymbol{w}||_2^2 \quad \text{where} \quad L_S^{SAM}(\boldsymbol{w}) \triangleq \max_{||\boldsymbol{\epsilon}||_p \leq \rho} L_S(\boldsymbol{w} + \boldsymbol{\epsilon}),$$

- SAM flattens high-memorization samples more strongly
- Important observation: SAM changes the geometry associated with particular training samples
- *Intuition 1:* build scalable memorization proxy measuring loss-surface geometry
- *Intuition 2:* can this dependence be exploited by a membership attacker?



(a) Low mem (b) High mem



(a) Low memorization (b) High memorization

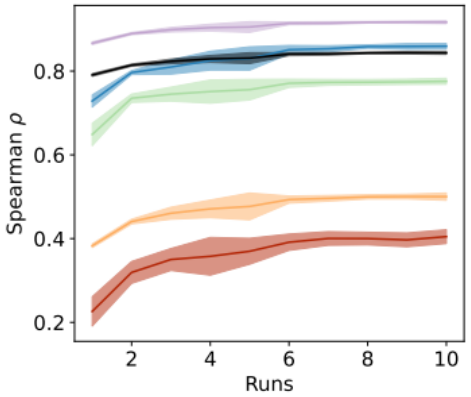
Agiollo et al. "Approximating memorization using loss surface geometry for dataset pruning and summarization." KDD 2024.
Foret et al. "Sharpness-aware minimization for efficiently improving generalization." arXiv 2020.

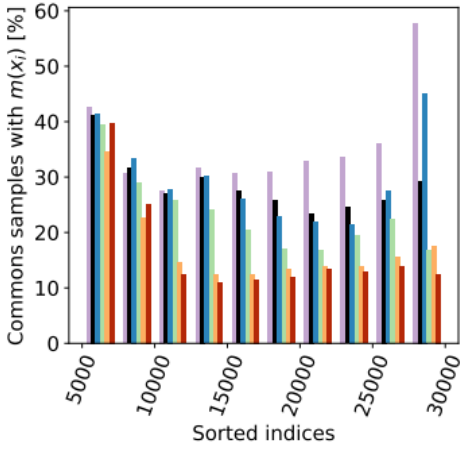
Geometry as a Proxy for Memorization

- SAMIS: absolute difference in the loss between the SAM model and SGD model to approximate memorization

$$S_L(\mathbf{x}_i) = |\mathcal{L}\left(h^{SAM}_{S^{\setminus i,..,z}}(\mathbf{x}_i), y_i\right) - \mathcal{L}\left(h^{SGD}_{S^{\setminus i,..,z}}(\mathbf{x}_i), y_i\right)|$$

- SAMIS reaches highest spearman correlation
- SAMIS correlates well both on low and high memorization extremes
- After visualization, several images are matching between memorization scores and SAMIS
- Similar results with other non flatness-aware optimizers (Adam, RMSProp, etc.)







(a) Low memorization/proxy

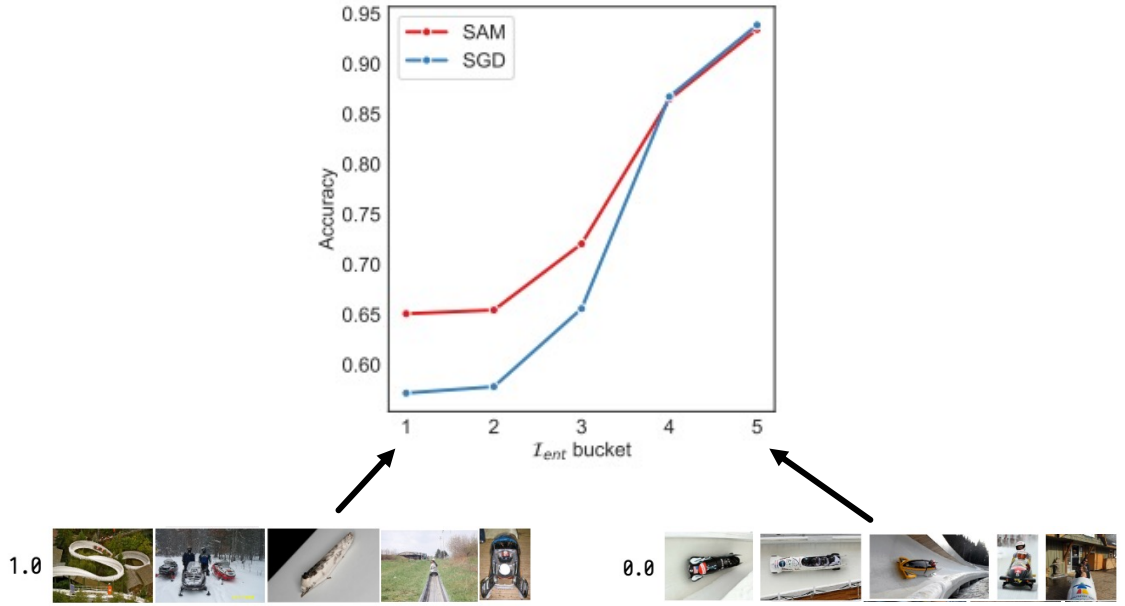


(b) High memorization/proxy

Agiollo et al. "Approximating memorization using loss surface geometry for dataset pruning and summarization." KDD 2024.

Loss Geometry Impact on Privacy

- SAM generalizes more on atypical samples
- SAM reduces the generalization gap over samples that can incur in memorization
- Expected outcome: Better generalization gap over memorized samples should reduce the member/non-member behavioral gap
- Observed outcome: SAM improves test accuracy and increases membership attack success



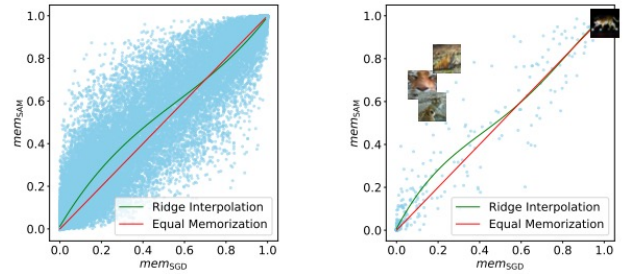
Dataset	Attack	SGD				SAM			
		Test Acc	AUC	Attack Acc	TPR@.1	Test Acc	AUC	Attack Acc	TPR@.1
CIFAR-100	RMIA	67.7%	90.4%	80.8%	21.0%	69.1%	91.6%	82.2%	23.4%
	LiRA		92.6%	82.9%	27.0%		93.7%	84.1%	31.0%
CIFAR-10	RMIA	92.3%	71.4%	63.5%	4.8%	93.1%	74.9%	65.9%	6.7%
	LiRA		73.0%	64.2%	8.8%		76.4%	66.7%	12.5%
Purchase100	RMIA	76.5%	68.8%	62.7%	1.5%	77.4%	70.2%	63.7%	1.7%
	LiRA		68.9%	62.6%	1.4%		70.2%	63.4%	1.6%
Texas100	RMIA	46.9%	79.8%	70.8%	2.9%	49.2%	80.6%	71.5%	2.8%
	LiRA		80.8%	71.3%	6.9%		81.6%	72.0%	8.2%

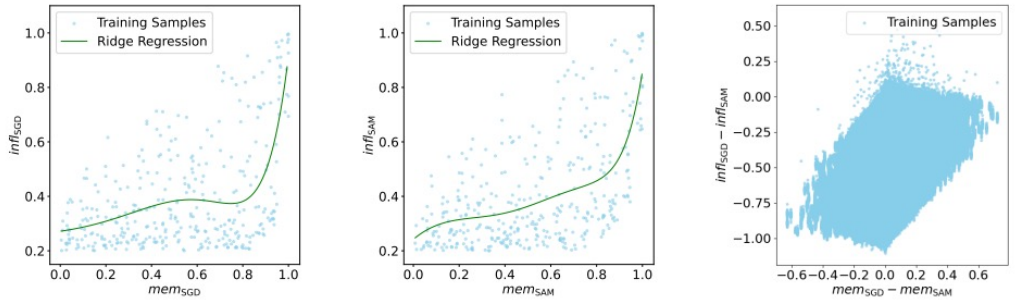
Dataset	Optimizer	NN	Confidence	Entropy	M-entropy	Test Acc
CIFAR-100	SGD	76.62%	77.19%	76.61%	77.30%	80.30%
	SAM	77.99%	79.10%	78.66%	79.25%	81.60%
CIFAR-10	SGD	50%	59.37%	59.09%	59.51%	96.00%
	SAM	50.08%	61.64%	61.64%	61.70%	96.72%
Purchase-100	SGD	66.00%	66.76%	64.78%	67.13%	85.50%
	SAM	66.62%	67.30%	65.35%	67.54%	85.54%
Texas-100	SGD	59.81%	65.20%	55.74%	65.13%	50.83%
	SAM	59.56%	66.59%	57.14%	65.42%	51.34%
EyePacs	SGD	73.62%	73.40%	68.50%	73.40%	73.67%
	SAM	77.73%	77.07%	73.37%	77.36%	75.41%

Kim et al. "Membership Privacy Risks of Sharpness Aware Minimization." ICLR 2026.

Loss Geometry Impact on Privacy (Part II)

- Measure exact memorization over SAM
- SGD rely heavily on patterns shared by many observations
- SAM is more capable of capturing atypical but reusable subpatterns
- Because relatively few training records provide the rare feature, learning it increase the model's dependence on those particular records
- Theoretical analysis: sharpness regularization reduce variance of confidence on unseen samples
- Lower variance guarantees larger attack advantage for the confidence and likelihood-ratio membership attacks





Confidence-threshold attack. The attacker uses the confidence score $\mathrm{Conf}_G(x) := |f_G(x)|$ and predicts "member" iff $\mathrm{Conf}_G(x) \geq \tau$. Let $(X_{\mathrm{in}}, Y_{\mathrm{in}})$ be a random training pair and $X_{\mathrm{out}} \sim \mathcal{N}(0, \Sigma)$ an independent non-member. Define attack advantage

$$\mathrm{Adv}_G^{\mathrm{conf}} := \sup_{\tau \geq 0} \big(\mathrm{TPR}_G(\tau) - \mathrm{FPR}_G(\tau)\big).$$

Theorem 2 (SAM strictly increases confidence-based MIA advantage)**.** *Under Assumption 1, for all sufficiently small* $\eta > 0$,

$$\mathrm{Adv}_{G_\eta}^{\mathrm{conf}} > \mathrm{Adv}_{G_0}^{\mathrm{conf}} \qquad \textit{with probability } 1 - o(1).$$

Kim et al. "Membership Privacy Risks of Sharpness Aware Minimization." ICLR 2026.

Conclusions

- Membership leakage is heterogeneous: it depends on individual record, training procedure, attack statistic, and attacker's strategy
- Loss-surface geometry provides a sample-level view of model dependence: difference between SAM and SGD can be used to identify highly memorized and atypical samples, and reveals that different records interact differently with the optimization process
- Better generalization can coexist with stronger membership leakage: learning atypical subpatterns more effectively result in better test performance, but learning those patterns increases dependence on some training records
- Security question becomes *"what did the model have to remember in order to generalize?"*



Thank you for your attention

Questions?