RelaxLoss: Defending Membership Inference Attacks without Losing Utility

Dingfan Chen[1], Ning Yu[2,3,4], Mario Fritz[1]

[1] CISPA Helmholtz Center for Information Security, [2] Salesforce Research,
[3] University of Maryland, [4] Max Planck Institute for Informatics

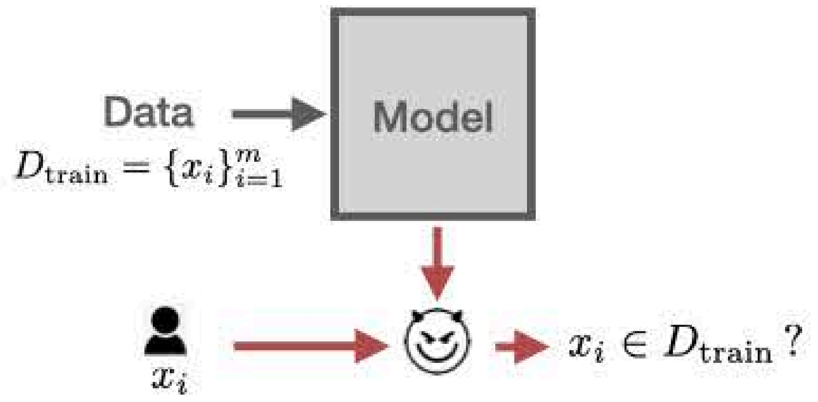
https://github.com/DingfanChen/RelaxLoss

Motivation

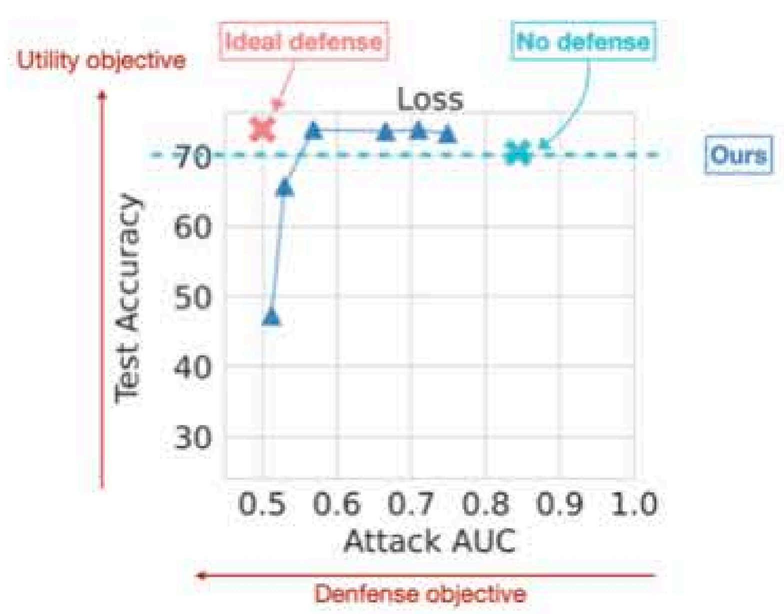
- Privacy issues when deploying ML models in many sensitive domains (e.g., healthcare, financial)
- In particular, modern deep neural networks (NN) are prone to memorize training data due to their high capacity, making them vulnerable to privacy attacks

Problem

- **Membership inference attacks (MIAs)** are pervasive in various data domains (e.g., images, medical data, transaction records)



- **Existing Approach:**
  - Regularization methods (designed for mitigating overfitting):
    - Generally unable to mitigate MIA[1]
  - Adversarial training[2,3]:
    - Hard to generalize to novel attacks unanticipated by the defender (e.g., a simple metric-based attack)
  - Differentially private (DP) training[4]:
    - Inevitably compromises model utility and increases computation cost
- **Our work:**
  - Defense objective:
    - Addresses a **wide range of** attacks
  - Utility objective:
    - **Preserve (or even improve)** the model utility.

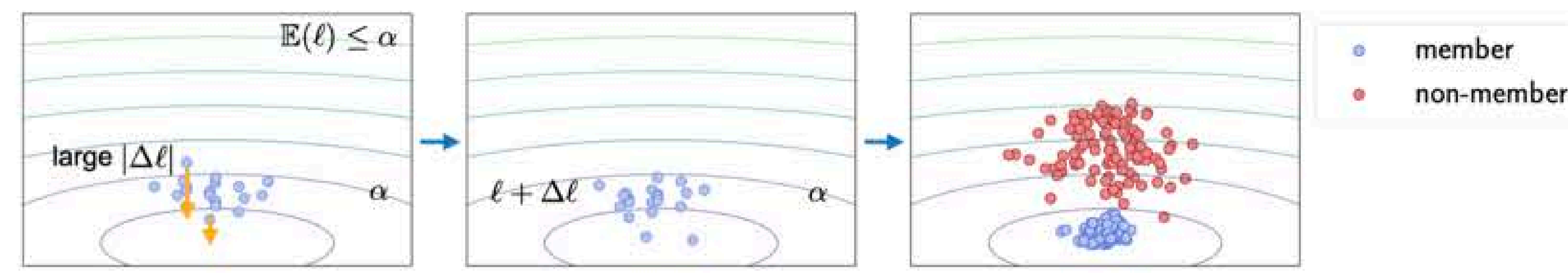


References

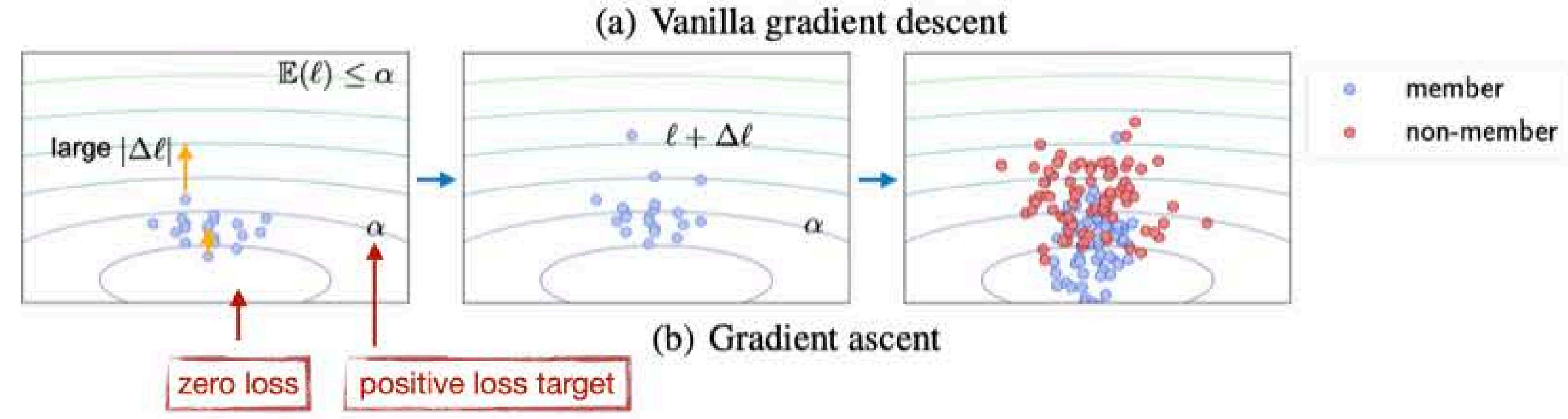
[1] Kaya et al., "When does data augmentation help with membership inference attacks?", ICML 2021
[2] Jia et al., "Memguard: Defending against black-box membership inference attacks via adversarial examples", CCS 2019
[3] Nasr et al., "Machine learning with membership privacy using adversarial regularization", CCS 2018
[4] Abadi et al., "Deep learning with differential privacy", CCS 2016
[5] Yeom et al., "Privacy risk in machine learning: Analyzing the connection to overfitting", CSF 2018
[6] Sablayrolles et al., "White-box vs black-box: Bayes optimal strategies for membership inference", ICML 2019

Approach: RelaxLoss

- **Existing theoretical results**
  - A large gap in the losses, i.e., $\mathbb{E}[\ell]_{\text{non}} - \mathbb{E}[\ell]_{\text{mem}}$, is sufficient for conducting membership inference attacks[5]
  - The Bayes optimal attack only depends on the sample loss[6]
- **Approach:**
  - **Relaxing loss target** with gradient ascent



(a) Vanilla gradient descent



(b) Gradient ascent

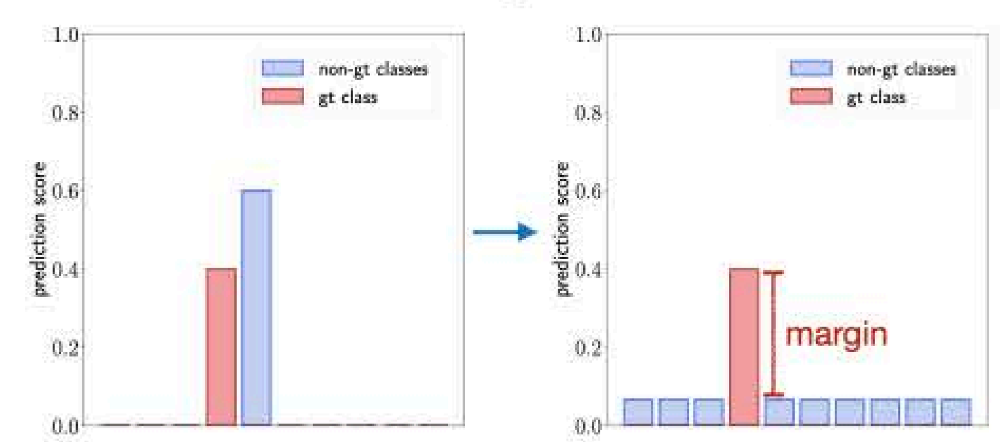
  - **Flattening** the target **posterior scores** for non-ground-truth classes

Construct softlabel t_i with

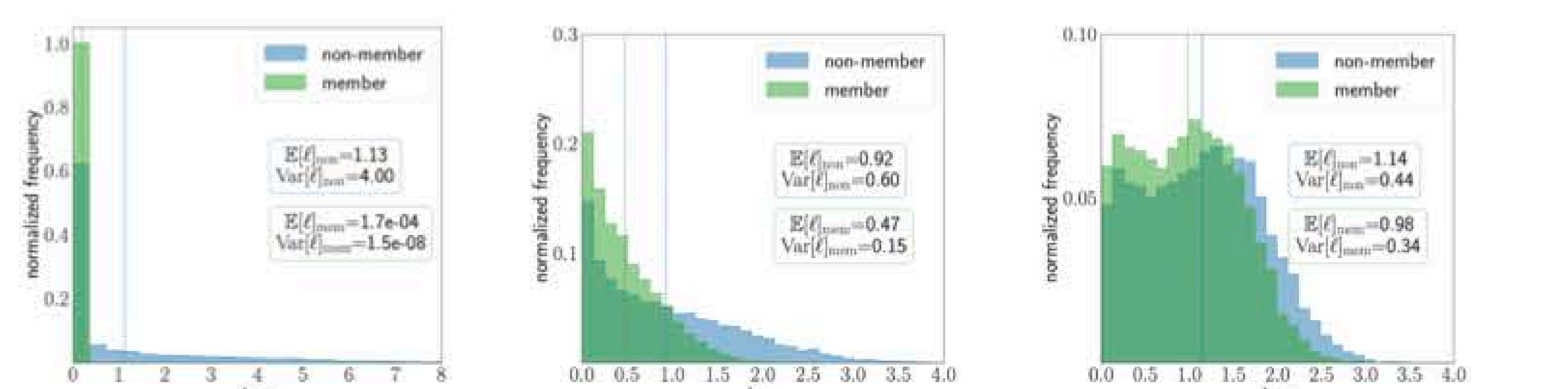
$$t_i^c = \begin{cases} p_i^c & \text{if } y_i^c = 1 \\ (1-p_i^c)/(C-1) & \text{otherwise} \end{cases}$$

Compute cross entropy loss with the softlabel: $^{a)}$

$$\ell(\theta, z_i) = -\sum_{c=1}^{C} \mathrm{sg}[t_i^c] \log p_i^c$$



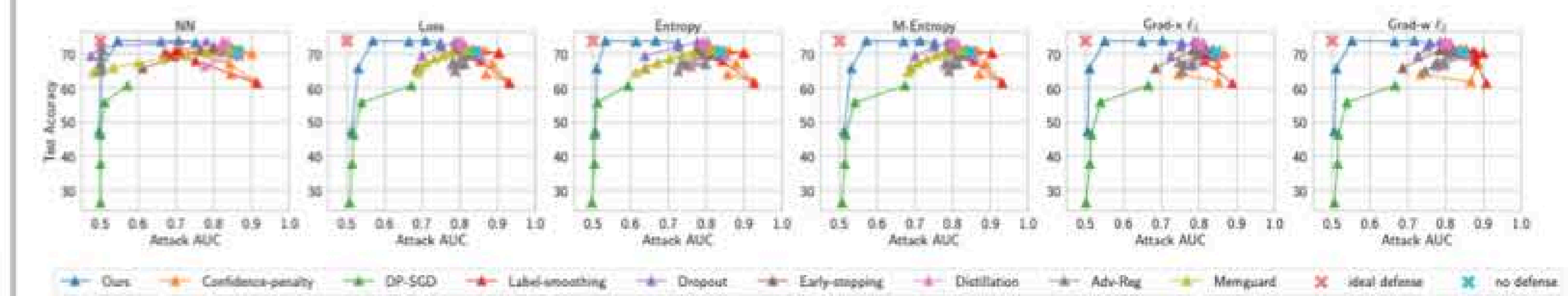
- **Properties**
  - Reduces generalization gap
  - Increase variance of training loss distributions



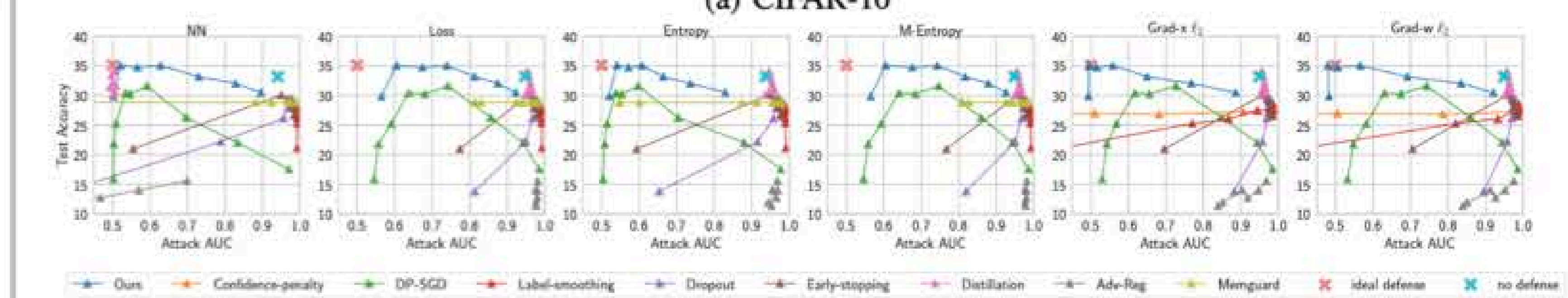
(a) Vanilla (b) Ours ($\alpha = 0.5$) (c) Ours ($\alpha = 1.0$)

Evaluation

- **Comparison to existing defense methods**
  - Test accuracy **(Utility)** vs. Attack AUC **(Effectiveness)**
  - **Baselines:** Memguard, Adv-Reg, Early-stopping, Dropout, Label-smoothing, Confidence-penalty, Distillation, DP-SGD



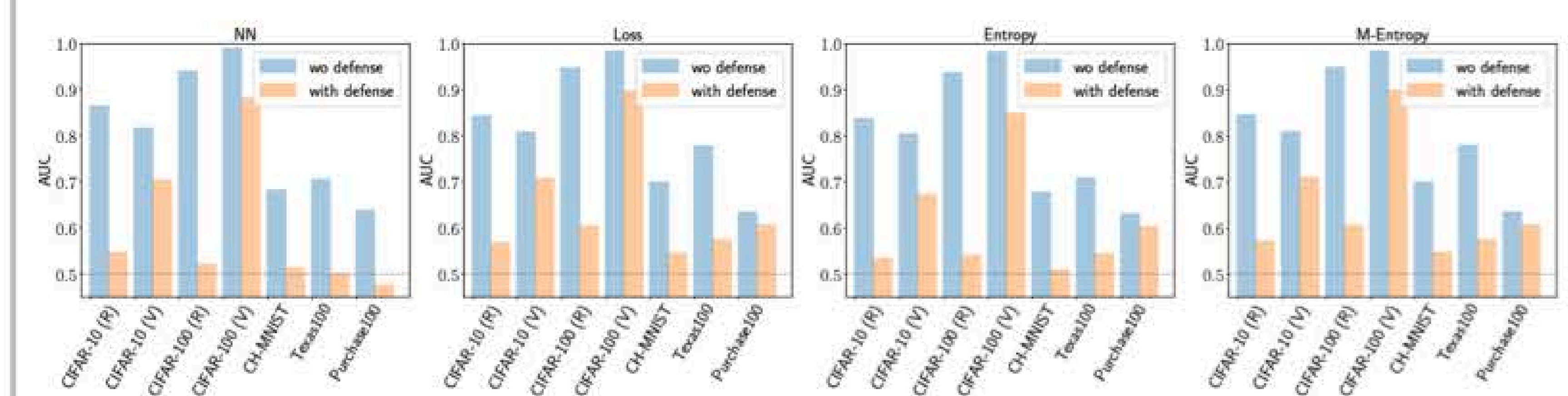
(a) CIFAR-10



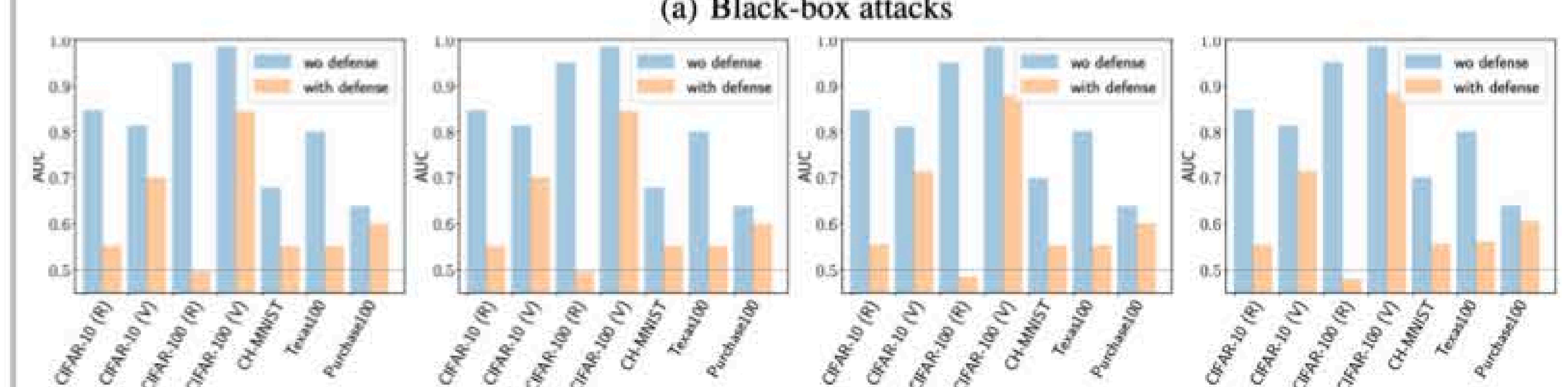
(b) CIFAR-100

- **Defense effectiveness without losing utility**

	CIFAR10 (ResNet20)		CIFAR10 (VGG11)		CIFAR100 (ResNet20)		CIFAR100 (VGG11)		CH-MNIST		Texas100		Purchase100	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
wo defense	70.5	96.6	73.8	97.0	33.2	63.0	41.4	67.5	77.1	99.6	52.3	82.6	89.1	99.8
with defense	73.8	98.2	74.4	97.8	35.1	67.7	41.4	69.9	78.4	99.7	55.3	86.8	89.1	99.6
Δ	4.68	1.66	0.81	0.82	5.72	7.46	0.00	3.56	1.69	0.10	5.74	5.08	0.00	-0.20



(a) Black-box attacks



(b) White-box attacks

- **Adaptive attack**

	CIFAR10 (ResNet20)	CIFAR10 (VGG11)	CIFAR100 (ResNet20)	CIFAR100 (VGG11)	CH-MNIST	Texas100	Purchase100
w/o defense	87.3	80.7	92.6	97.5	67.1	79.0	65.7
w/ defense (non-adaptive)	50.0	50.0	50.0	50.0	50.7	50.0	50.1
Δ (non-adaptive)	-42.7	-38.0	-46.0	-48.7	-24.4	-36.7	-23.9
w/ defense (adaptive)	56.0	68.2	57.8	84.2	56.6	53.8	56.0
Δ (adaptive)	-35.9	-15.5	-37.6	-13.6	-15.6	-31.9	-14.8