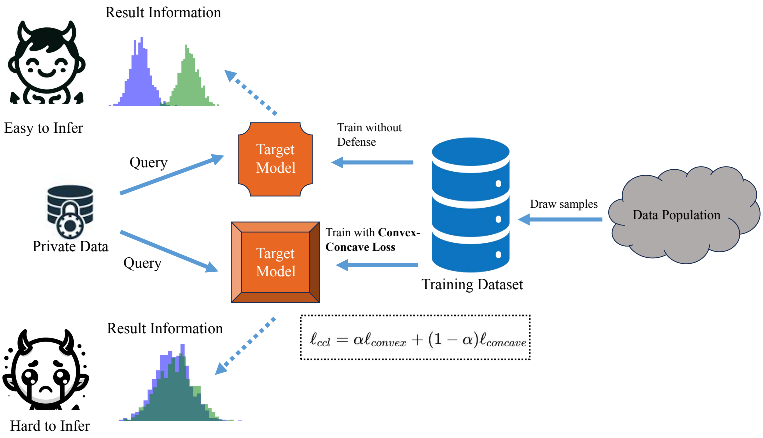
Mitigating Privacy Risk in Membership Inference by Convex-Concave Loss

Zhenlong Liu, Lei Feng, Huiping Zhuang, Xiaofeng Cao, Hongxin Wei

ICML 2024

Membership Inference Attack

- The goal of membership inference attack(MIA) is to identify whether a data point was in a model's training set.



Attack Advantage

- An attacker infers an input record $\boldsymbol{x}$ as a member if its prediction loss is smaller than a threshold τ.

$$\mathcal{A}_{\text{loss}} = \mathbb{I}(\mathcal{L}(h_S(\boldsymbol{x}), y) \leqslant \tau)$$

Evaluation Metric

To quantify the performance of the attack model $\mathcal{A}$, we use the *membership advantage*

$$Adv = \Pr(\mathcal{A} = 1 | m = 1) - \Pr(\mathcal{A} = 1 | m = 0)$$

Membership Advantage by Metric-based Attack

Suppose ϵ is a random variable denoting loss, such that $\epsilon \sim N(\mu_S, \sigma_S^2)$ when $m = 1$ and $\epsilon \sim N(\mu_D, \sigma_D^2)$ when $m = 0$. Then the membership advantage of $\mathcal{A}_{\text{loss}}$ is:

$$
\begin{aligned}
Adv &= \Pr(\mathcal{A} = 1 | m = 1) - \Pr(\mathcal{A} = 1 | m = 0) \\
&= \Pr(\epsilon \leqslant \tau | m = 1) - \Pr(\epsilon \leqslant \tau | m = 0) \\
&= \Phi(\frac{\tau - \mu_S}{\sigma_S}) - \Phi(\frac{\tau - \mu_D}{\sigma_D})
\end{aligned}
$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution.
Assume τ is chosen such that $\Phi(\frac{\tau - \mu_D}{\sigma_D}) = \alpha$,

$$
Adv = \Phi\{\frac{\Phi^{-1}(\alpha)\sigma_D + \mu_D - \mu_S}{\sigma_S}\} - \alpha
$$

Prior SOTA Defend Method - RelaxLoss [Chen et al. 2022]

- Perform **gradient ascent** to promote a high variance of the training loss distribution.

Achieve Success in Privacy Defense, but...

- Optimizing toward a reverse direction leads to suboptimal performance.

Can we achieve a comparable defense effect using gradient descent?

Setup

- We study the problem of membership inference attacks in K-class classification tasks.
- For a sample $\boldsymbol{x} \in \mathcal{X}$, we denote the distribution over different labels by $q(k|\boldsymbol{x})$, the output probability of $h_S(\boldsymbol{x})$ by $p(k|\boldsymbol{x})$.
- In particular, the confidence in the true label $p(y|\boldsymbol{x})$ is abbreviated as p_y.
- The most commonly used Cross Entropy loss function: $\ell_{ce} = -\log p_y$

Assume that p_y is a random variable with mean $1-\epsilon$ and variance σ^2, where $\epsilon > 0$

For cross entropy loss ℓ_{ce}, by taylor expansion, we have

$$E\ell_{ce} = E(-\log p_y) > E[(1-p_y) + \frac{1}{2}(1-p_y)^2] = \epsilon + \frac{1}{2}(\sigma^2 + \epsilon^2)$$

Training loss can be optimized toward a smaller value of variance σ^2

Alternatively, we could interpret σ^2 as a penalty term.

Given a twice continuously differentiable function $\ell \in C^2(0,1]$ such that $\ell(1) = 0$ and $\ell'(x) < 0, \forall x \in (0,1]$. If ℓ is strictly **convex**, then

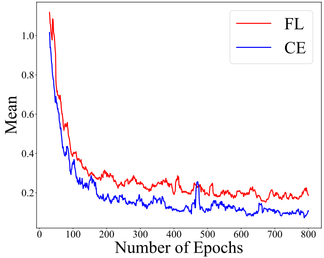
$$\mathbb{E}_{\mathcal{D}}[\ell(p_y)] \geqslant A\epsilon + \frac{B}{2}(\epsilon^2 + \sigma^2)$$

where $A = -\ell'(1) > 0$, $B \geqslant 0$ is a non-negative lower bound of $\ell''(x)$.

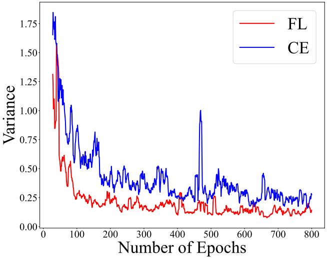
Focal Loss

$$\ell_{\mathrm{fl}} = -(1-p_y)^{\gamma}\log(1-p_y)$$

$$\ell''_{\mathrm{fl}} \geqslant \ell''_{\mathrm{ce}}, \forall x \in (0,1]$$



(a) Mean of Loss



(b) Variance of Loss

Figure 1: Models are trained on CIFAR-10 with Resnet-34 using Cross-entropy loss (CE) and Foc al loss (FL).

Can concave functions increase the loss variance?

Given a twice continuously differentiable function $\ell \in C^2[0,1]$ such that $\ell(1) = 0$ and $\ell'(x) < 0, \forall x \in [0,1]$. If ℓ is strictly **concave**, there must exist a **negative** constant $B \leqslant 0$ such that

$$\mathbb{E}_{\mathcal{D}}[\ell(p_y)] = A\epsilon + B(\sigma^2 + \epsilon^2) \tag{1}$$

where $A = -\ell'(1) > 0$.

Add concave term

Since concave functions can be leveraged to design loss functions, we propose to add a concave term into the original loss function (e.g., cross-entropy loss), which is called *Convex-Concave Loss* (CCL).
We define a concave function set as:

$$\mathcal{F} = \{f \in C^2[0,1] \mid f'(x) < 0, f''(x) < 0, \forall x \in [0,1]\}$$

Convex-Concave Loss

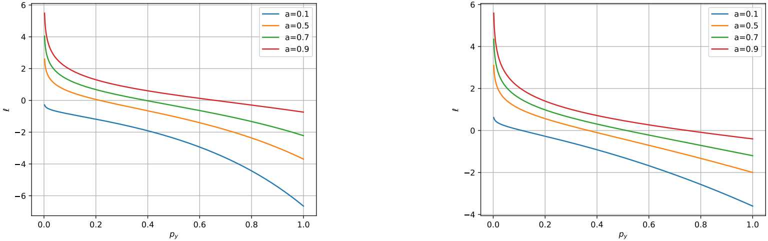
$$\ell_{\text{ccl}} = \alpha\hat{\ell} + (1-\alpha)\tilde{\ell}$$

where $\hat{\ell}$ is the origin convex function, $\tilde{\ell} \in \mathcal{F}$ is a concave term

Our Proposed Method

Concave Exponential Loss (CEL) and Concave Quadratic Loss (CQL)

$$\tilde{\ell}_{\textsf{exp}} = -\exp(p_y), \quad \tilde{\ell}_{\textsf{qua}} = -p_y - \frac{1}{2}p_y^2$$



(a) CCEL (b) CCQL

Figure 2: ℓ with different parameters α

RelaxLoss:

$$\mathrm{Var}(\ell + \Delta\ell) = \mathrm{Var}(\ell) + \mathrm{Var}(\Delta\ell) + 2\mathrm{Cov}(\ell, \Delta\ell)$$

With the convex function, the larger the loss value is, the faster it changes.

$$\mathrm{Cov}(\ell, \Delta\ell) > 0$$

ConcaveLoss:

$$\mathrm{Var}(\ell - \Delta\ell) = \mathrm{Var}(\ell) + \mathrm{Var}(\Delta\ell) - 2\mathrm{Cov}(\ell, \Delta\ell)$$

As for the concave terms, $\mathrm{Cov}(\ell, \Delta\ell) < 0$

Experimental Results

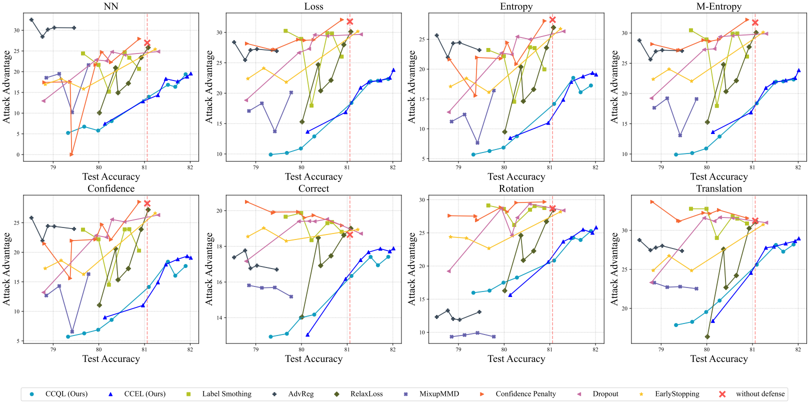


Figure 3: CIFAR10 Resnet34

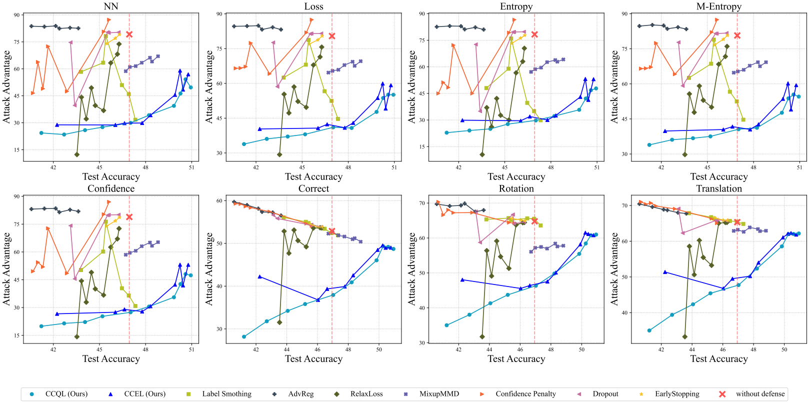


Figure 4: CIFAR100 Desnet121

- **Analysis**: We provide rigorous theoretical analyses to establish a key insight: convex loss functions tend to decrease the loss variance
- **Method**: We introduce the concept of Convex-Concave Loss (CCL), a generalized loss function that incorporates a concave term into the original convex loss, i.e., Cross-Entropy (CE) loss.
- **Results**: We establish that CCL offers a state-of-the-art balance in the privacy-utility trade-off, with extensive experiments

Thanks!