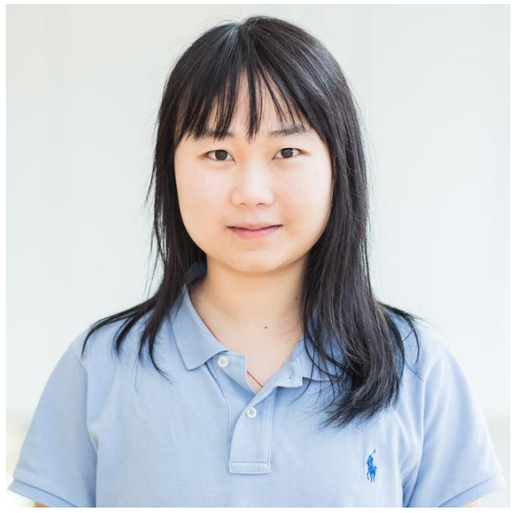
RelaxLoss: Defending Membership Inference Attacks without Losing Utility



Dingfan Chen[1]

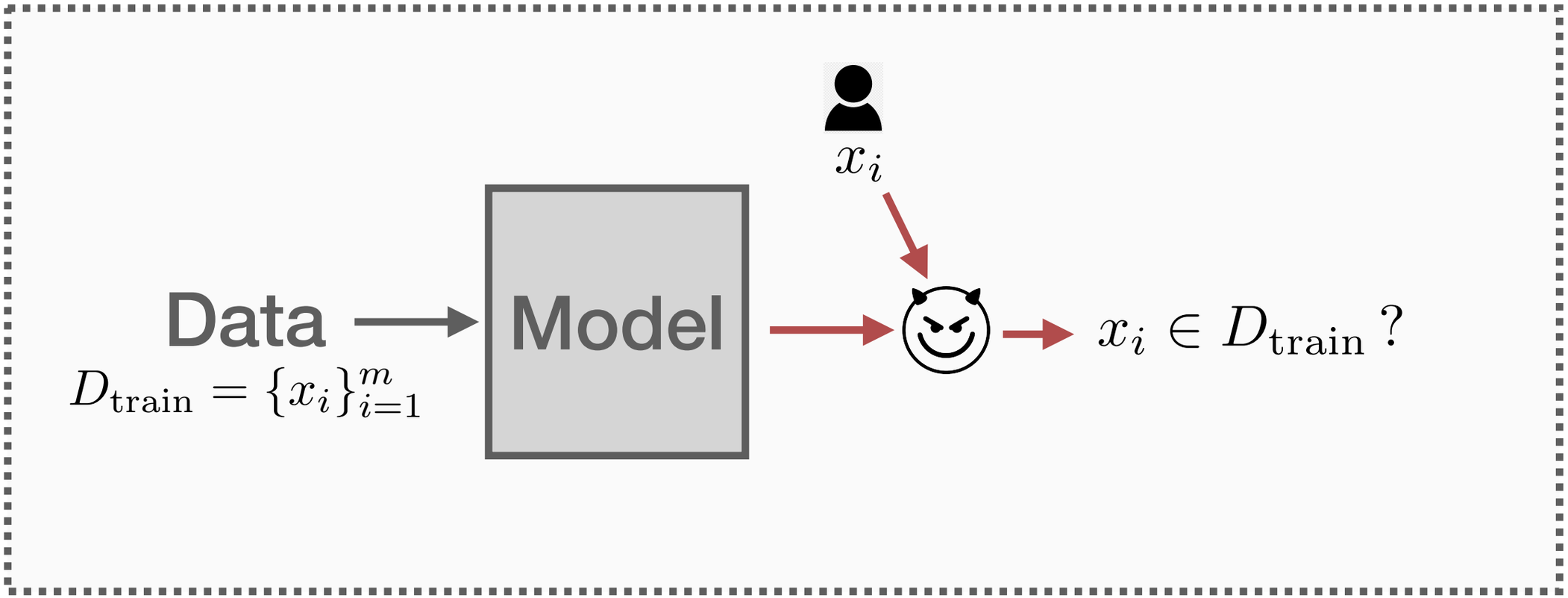
Ning Yu[2,3,4]

Mario Fritz[1]

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

Problem

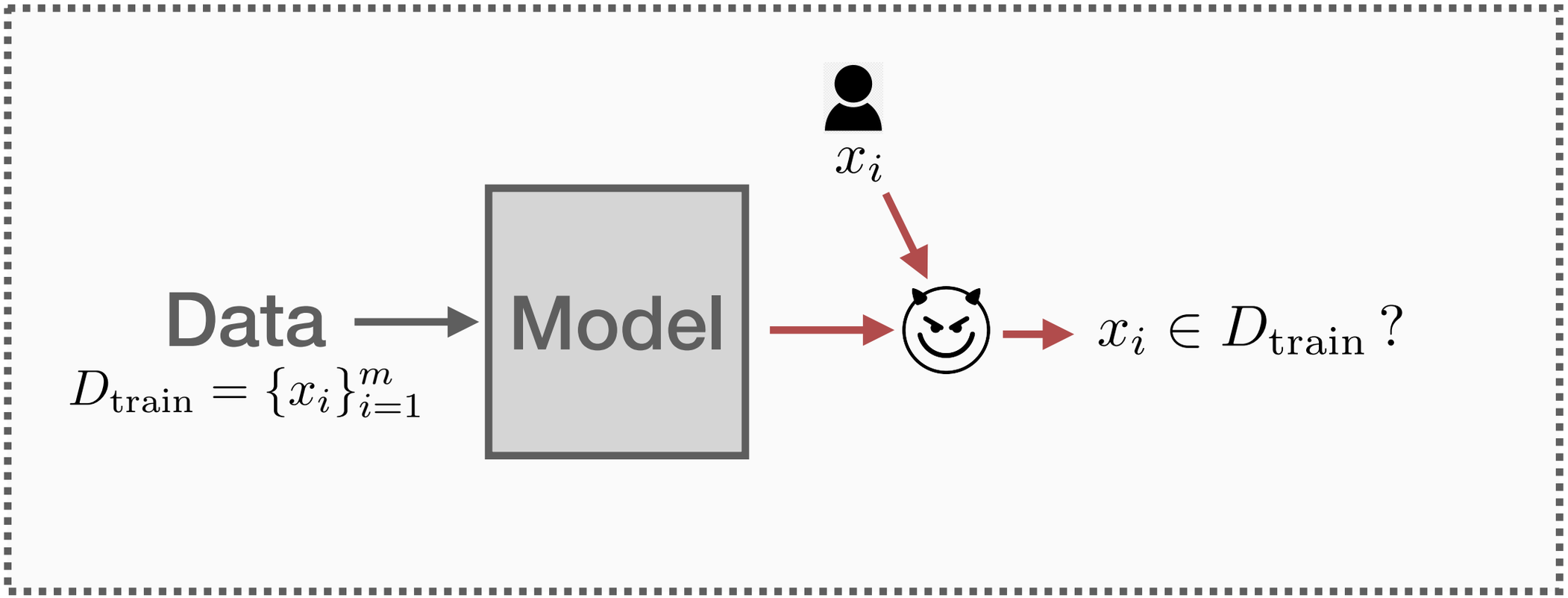
- **Membership inference attack (MIA)** [1] — an adversary tries to identify whether a given sample was included in the target model's training set



[1] Shokri, et al., "Membership inference attacks against machine learning models", S&P 2017

Problem

- **Membership inference attack (MIA)** [1] — an adversary tries to identify whether a given sample was included in the target model's training set



[1] Shokri, et al., "Membership inference attacks against machine learning models", S&P 2017

Problem

Shokri et al. 2017
Membership inference attacks against machine learning models

Nasr et al. 2018
Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning.

Yeom et al. 2018
Privacy risk in machine learning: Analyzing the connection to overfitting

Salem et al. 2019
Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models.

Sablayrolles et al. 2019
White-box vs black-box: Bayes optimal strategies for membership inference.

Song et al. 2020
Systematic evaluation of privacy risks of machine learning models

Rezae et al. 2021
On the Difficulty of Membership Inference Attacks

Choo et al. 2021
Label-only membership inference attacks.

Problem

- **Membership inference attacks are effective**
  - Given only black-box access
  - Or even partially observed output predictions

- **Such attacks are pervasive in various data domains, posing privacy threats to individuals**
  - E.g., images, medical data, transaction records

Shokri et al. 2017
Membership inference attacks against machine learning models

Nasr et al. 2018
Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning.

Yeom et al. 2018
Privacy risk in machine learning: Analyzing the connection to overfitting

Salem et al. 2019
Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models.

Sablayrolles et al. 2019
White-box vs black-box: Bayes optimal strategies for membership inference.

Song et al. 2020
Systematic evaluation of privacy risks of machine learning models

Rezae et al. 2021
On the Difficulty of Membership Inference Attacks

Choo et al. 2021
Label-only membership inference attacks.

Problem

- **Membership inference attacks are effective**
  - Given only black-box access
  - Or even partially observed output predictions

- **Such attacks are pervasive in various data domains, posing privacy threats to individuals**
  - E.g., images, medical data, transaction records

- **Existing defenses inevitably <u>compromise model utility</u> for a reasonable level of defense effectiveness**

- **Our work for the first time addresses <u>a wide range of attacks</u> while <u>preserving (or even improving)</u> the model utility.**

Approach

ICLR
International Conference On
Learning Representations

[1] Yeom et al., “Privacy risk in machine learning: Analyzing the connection to overfitting”, CSF 2018
[2] Sablayrolles, et al., “White-box vs black-box: Bayes optimal strategies for membership inference”, ICML 2019

Approach

- **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[1]
  - The Bayes optimal attack only depends on the sample loss[2]

[1] Yeom et al., “Privacy risk in machine learning: Analyzing the connection to overfitting”, CSF 2018
[2] Sablayrolles, et al., “White-box vs black-box: Bayes optimal strategies for membership inference”, ICML 2019

Approach

- **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[1]
  - The Bayes optimal attack only depends on the sample loss[2]

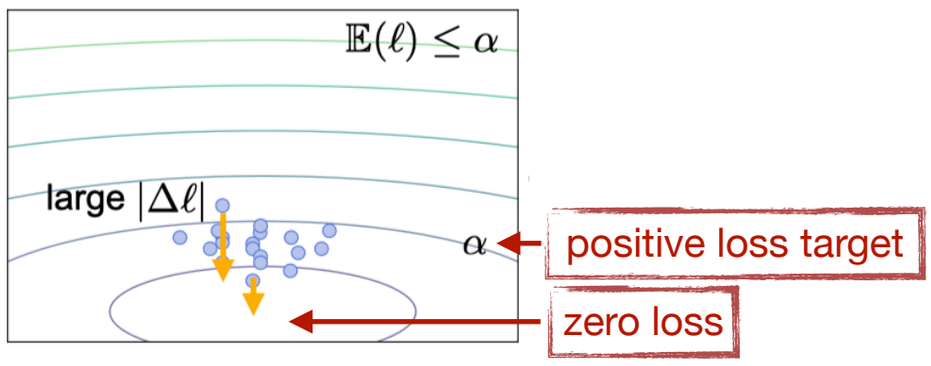
- **Approach (RelaxLoss):**
  - Relaxing loss target with gradient ascent

[1] Yeom et al., “Privacy risk in machine learning: Analyzing the connection to overfitting”, CSF 2018
[2] Sablayrolles, et al., “White-box vs black-box: Bayes optimal strategies for membership inference”, ICML 2019

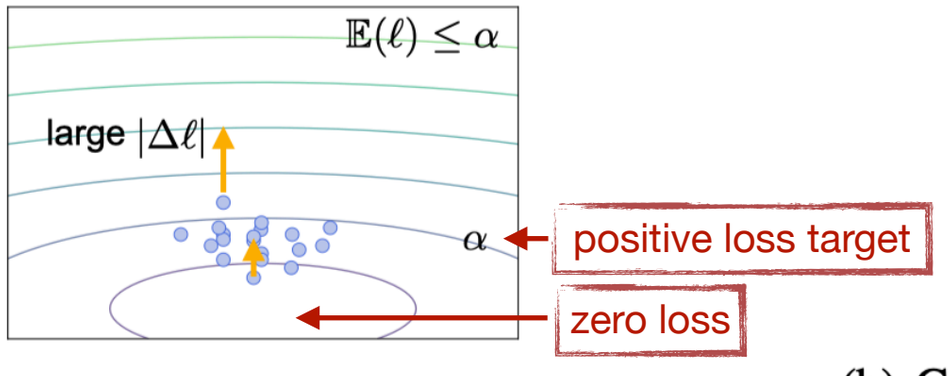
Approach

- **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[1]
  - The Bayes optimal attack only depends on the sample loss[2]

- **Approach (RelaxLoss):**
  - Relaxing loss target with gradient ascent



(a) Vanilla gradient descent



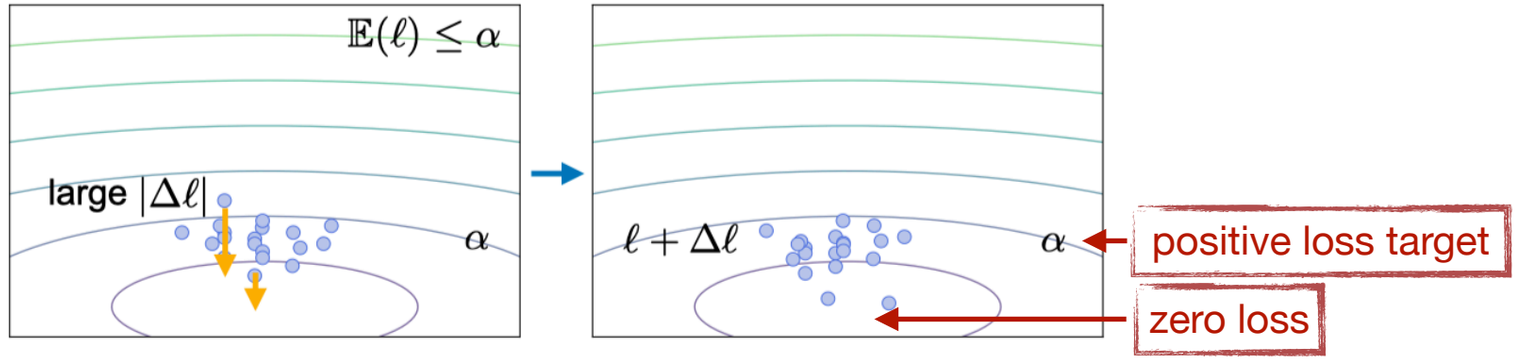
(b) Gradient ascent

[1] Yeom et al., "Privacy risk in machine learning: Analyzing the connection to overfitting", CSF 2018
[2] Sablayrolles, et al., "White-box vs black-box: Bayes optimal strategies for membership inference", ICML 2019

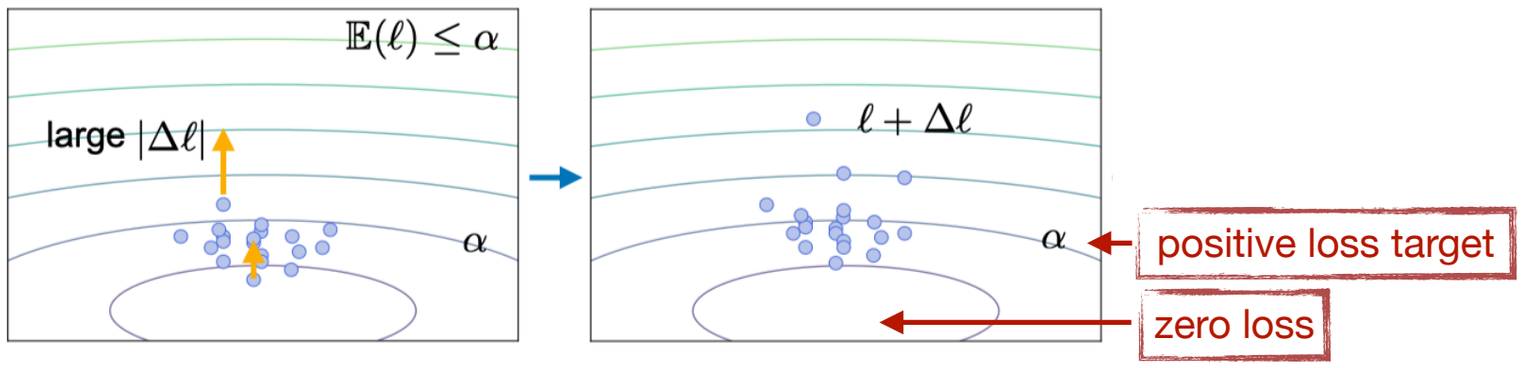
Approach

- **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[1]
  - The Bayes optimal attack only depends on the sample loss[2]

- **Approach (RelaxLoss):**
  - Relaxing loss target with gradient ascent



(a) Vanilla gradient descent



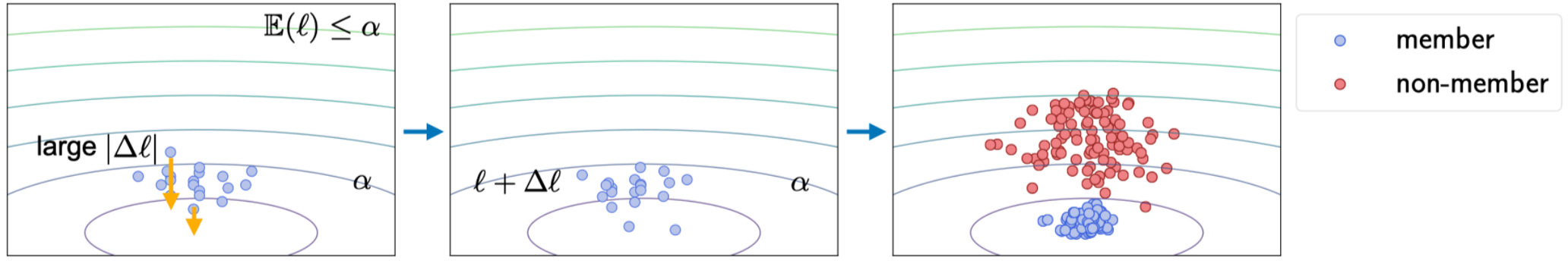
(b) Gradient ascent

[1] Yeom et al., "Privacy risk in machine learning: Analyzing the connection to overfitting", CSF 2018
[2] Sablayrolles, et al., "White-box vs black-box: Bayes optimal strategies for membership inference", ICML 2019

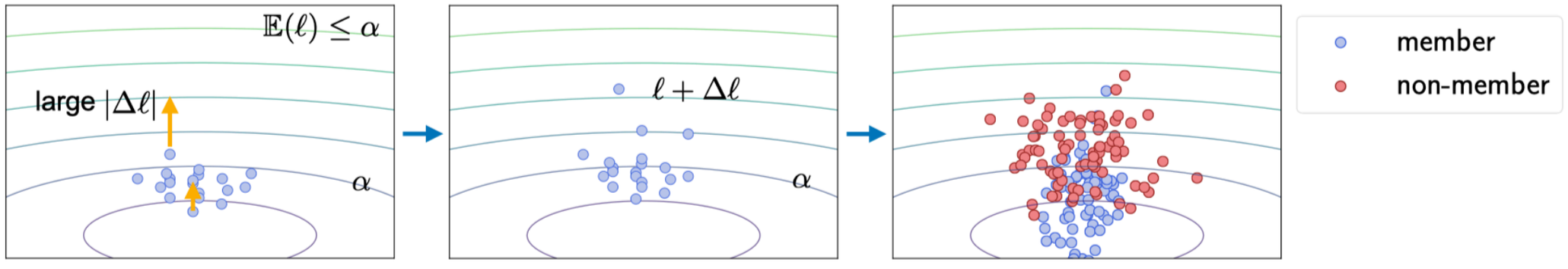
Approach

- **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[1]
  - The Bayes optimal attack only depends on the sample loss[2]

- **Approach (RelaxLoss):**
  - Relaxing loss target with gradient ascent



(a) Vanilla gradient descent



(b) Gradient ascent

[1] Yeom et al., "Privacy risk in machine learning: Analyzing the connection to overfitting", CSF 2018
[2] Sablayrolles, et al., "White-box vs black-box: Bayes optimal strategies for membership inference", ICML 2019

Approach

- **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[1]
  - The Bayes optimal attack only depends on the sample loss[2]

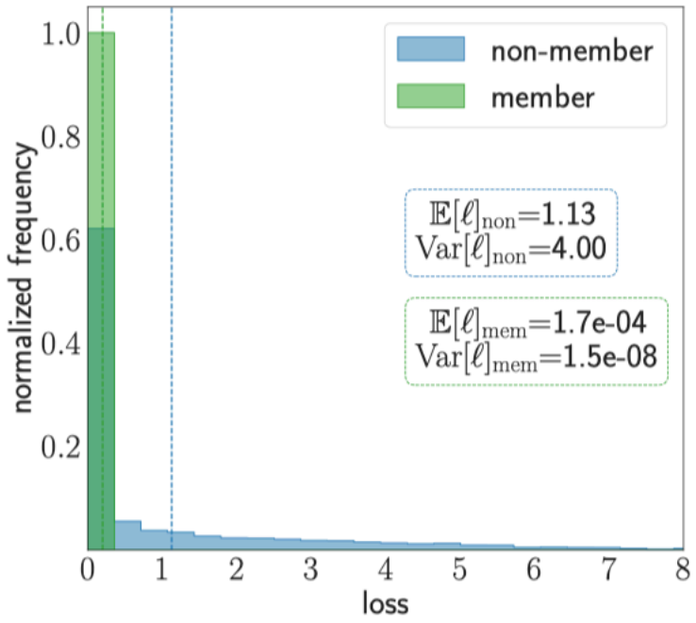
- **Approach (RelaxLoss):**
  - Relaxing loss target with gradient ascent

[1] Yeom et al., “Privacy risk in machine learning: Analyzing the connection to overfitting”, CSF 2018
[2] Sablayrolles, et al., “White-box vs black-box: Bayes optimal strategies for membership inference”, ICML 2019

Approach

- **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[1]
  - The Bayes optimal attack only depends on the sample loss[2]

- **Approach (RelaxLoss):**
  - Relaxing loss target with gradient ascent



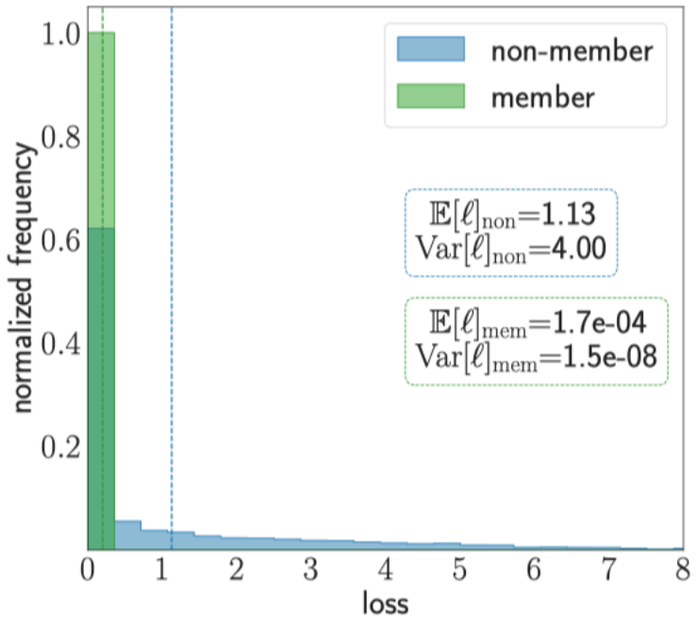
(a) Vanilla

[1] Yeom et al., “Privacy risk in machine learning: Analyzing the connection to overfitting”, CSF 2018
[2] Sablayrolles, et al., “White-box vs black-box: Bayes optimal strategies for membership inference”, ICML 2019

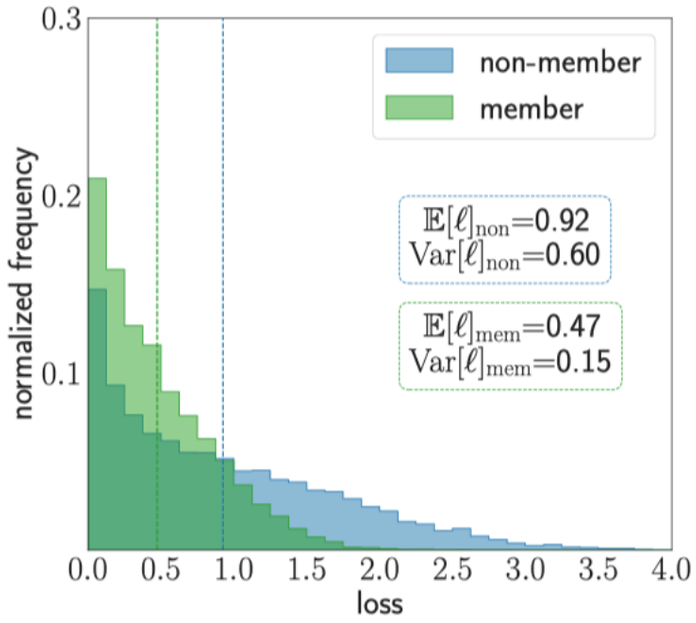
Approach

- **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[1]
  - The Bayes optimal attack only depends on the sample loss[2]

- **Approach (RelaxLoss):**
  - Relaxing loss target with gradient ascent



(a) Vanilla



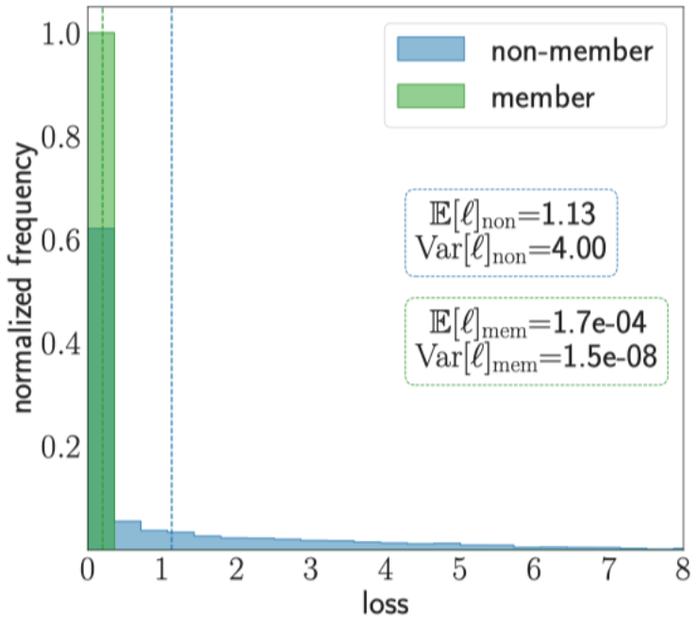
(b) Ours ($\alpha = 0.5$)

[1] Yeom et al., “Privacy risk in machine learning: Analyzing the connection to overfitting”, CSF 2018
[2] Sablayrolles, et al., “White-box vs black-box: Bayes optimal strategies for membership inference”, ICML 2019

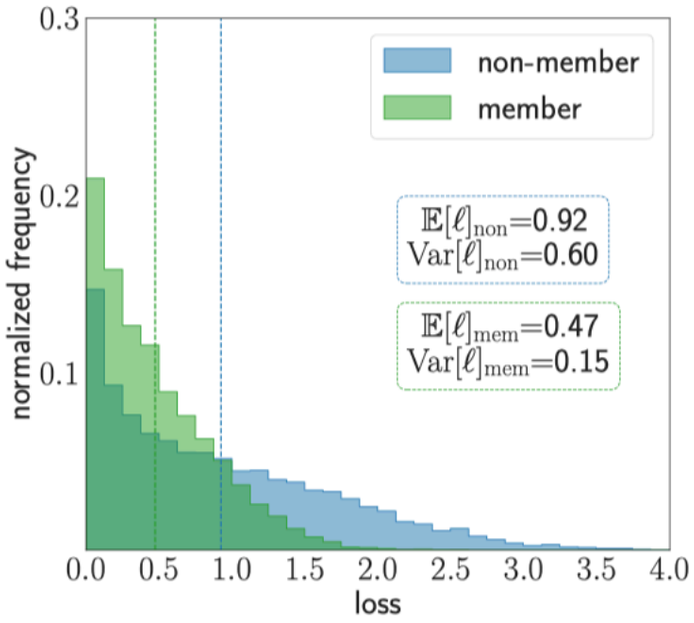
Approach

- **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[1]
  - The Bayes optimal attack only depends on the sample loss[2]

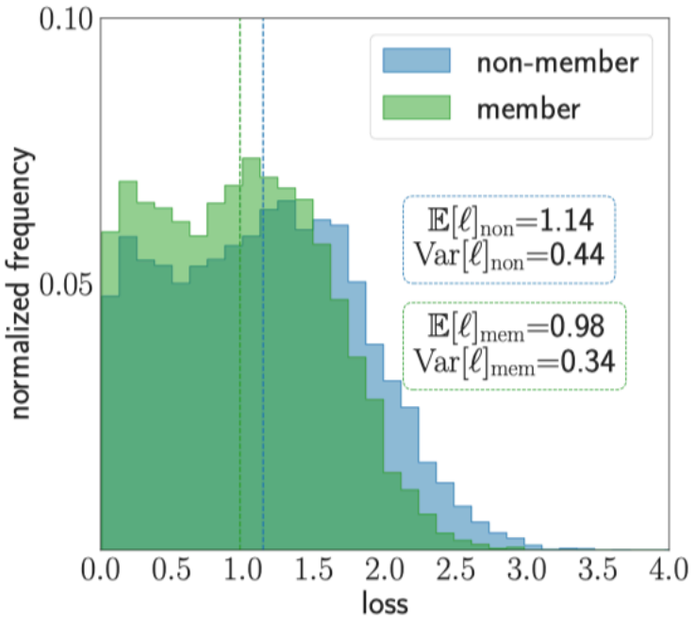
- **Approach (RelaxLoss):**
  - Relaxing loss target with gradient ascent



(a) Vanilla



(b) Ours ($\alpha = 0.5$)



(c) Ours ($\alpha = 1.0$)

[1] Yeom et al., "Privacy risk in machine learning: Analyzing the connection to overfitting", CSF 2018
[2] Sablayrolles, et al., "White-box vs black-box: Bayes optimal strategies for membership inference", ICML 2019

Approach

- **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[1]
  - The Bayes optimal attack only depends on the sample loss[2]

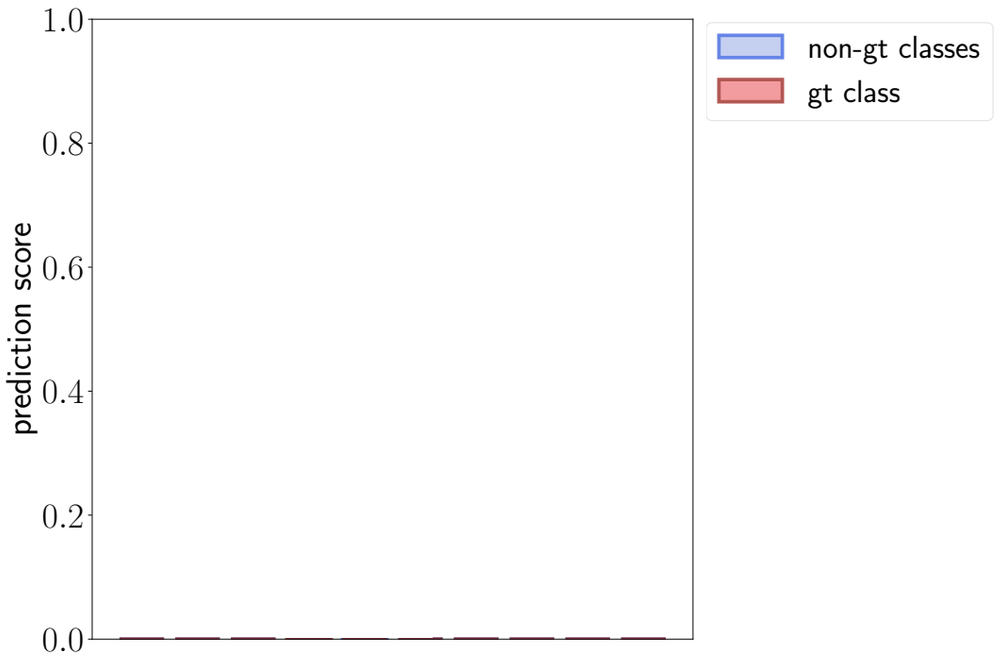
- **Approach (RelaxLoss):**
  - Relaxing loss target with gradient ascent
  - Flattening the target posterior scores for non-ground-truth classes

[1] Yeom et al., “Privacy risk in machine learning: Analyzing the connection to overfitting”, CSF 2018
[2] Sablayrolles, et al., “White-box vs black-box: Bayes optimal strategies for membership inference”, ICML 2019

Approach

- **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[1]
  - The Bayes optimal attack only depends on the sample loss[2]

- **Approach (RelaxLoss):**
  - Relaxing loss target with gradient ascent
  - Flattening the target posterior scores for non-ground-truth classes

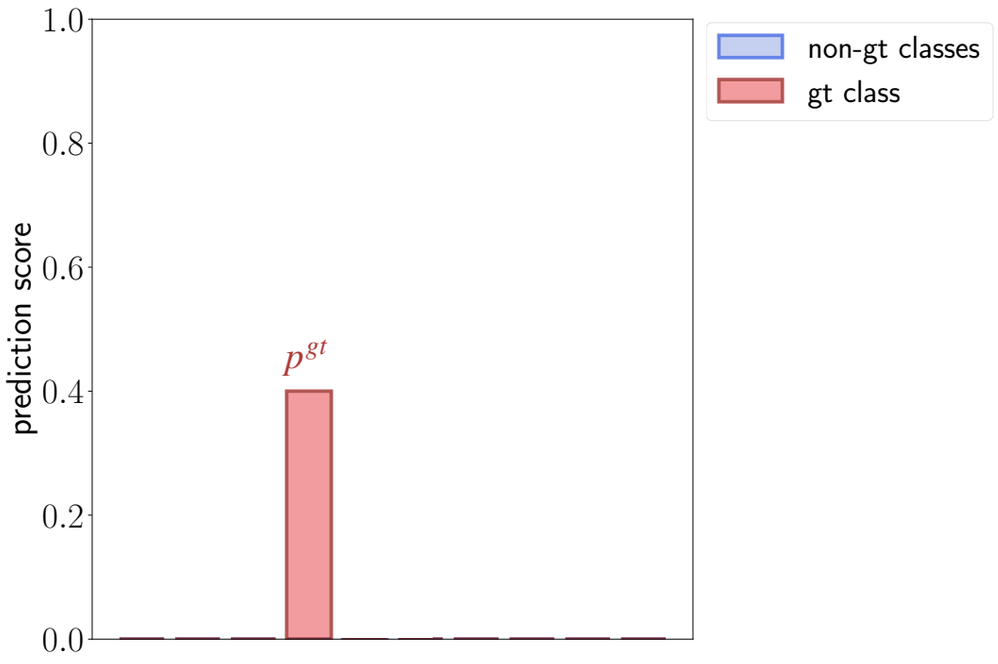


[1] Yeom et al., “Privacy risk in machine learning: Analyzing the connection to overfitting”, CSF 2018
[2] Sablayrolles, et al., “White-box vs black-box: Bayes optimal strategies for membership inference”, ICML 2019

Approach

- **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[1]
  - The Bayes optimal attack only depends on the sample loss[2]

- **Approach (RelaxLoss):**
  - Relaxing loss target with gradient ascent
  - Flattening the target posterior scores for non-ground-truth classes

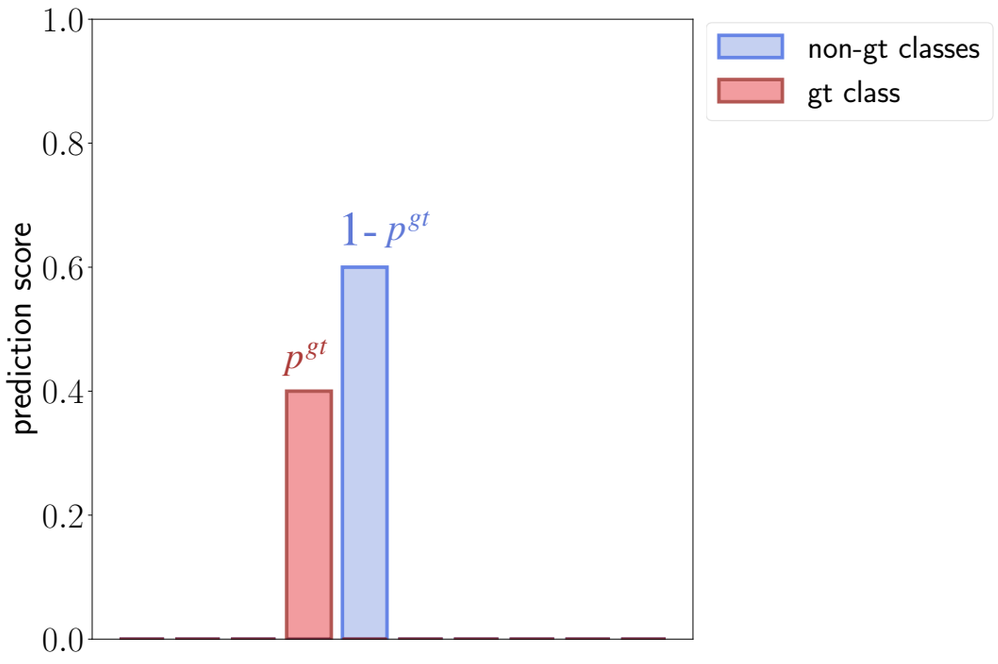


[1] Yeom et al., “Privacy risk in machine learning: Analyzing the connection to overfitting”, CSF 2018
[2] Sablayrolles, et al., “White-box vs black-box: Bayes optimal strategies for membership inference”, ICML 2019

Approach

- **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[1]
  - The Bayes optimal attack only depends on the sample loss[2]

- **Approach (RelaxLoss):**
  - Relaxing loss target with gradient ascent
  - Flattening the target posterior scores for non-ground-truth classes

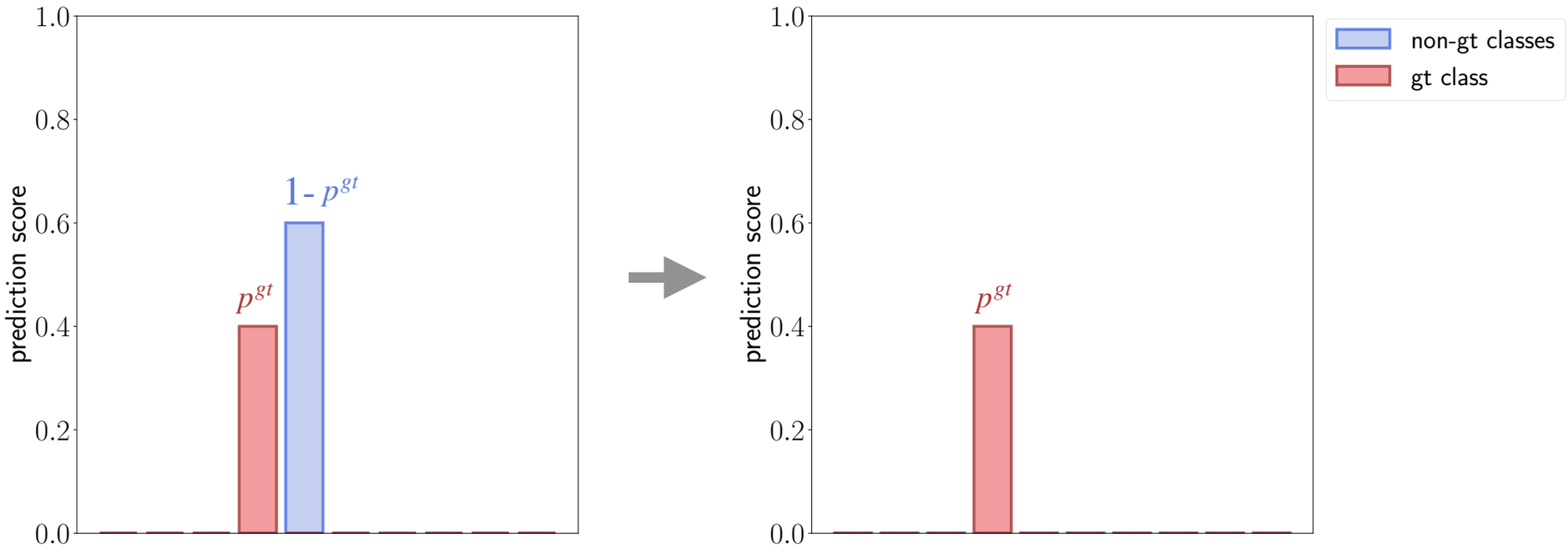


[1] Yeom et al., "Privacy risk in machine learning: Analyzing the connection to overfitting", CSF 2018
[2] Sablayrolles, et al., "White-box vs black-box: Bayes optimal strategies for membership inference", ICML 2019

Approach

- **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[1]
  - The Bayes optimal attack only depends on the sample loss[2]

- **Approach (RelaxLoss):**
  - Relaxing loss target with gradient ascent
  - Flattening the target posterior scores for non-ground-truth classes

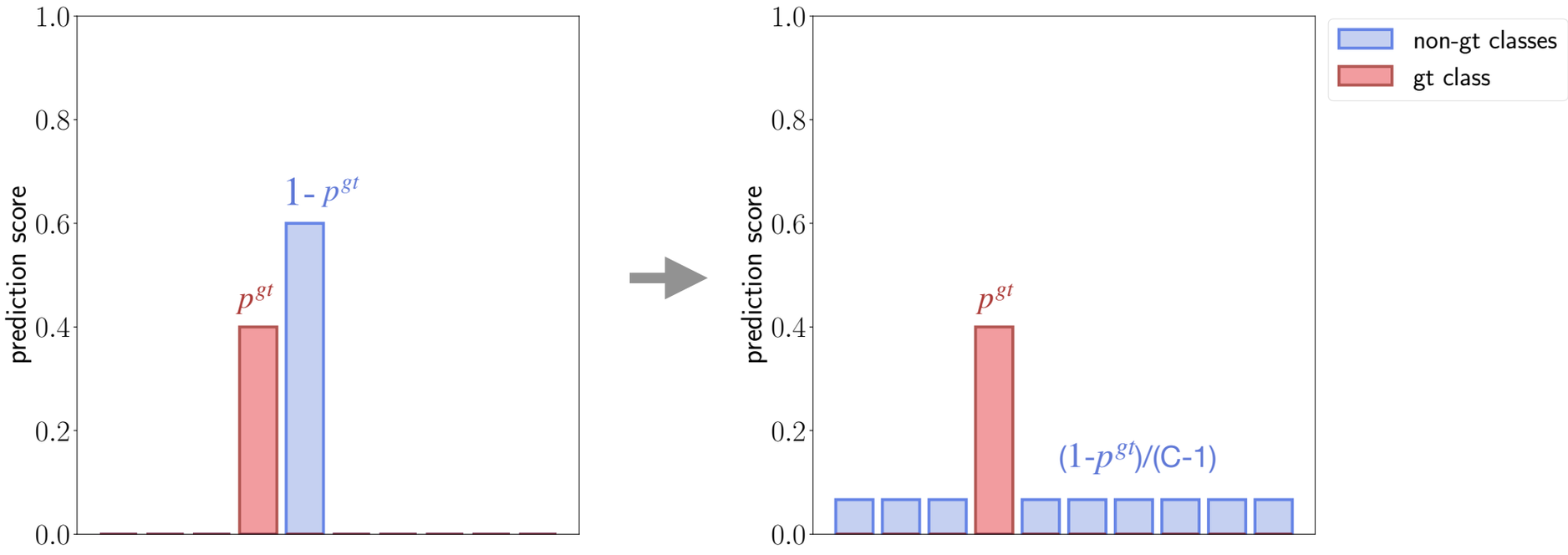


[1] Yeom et al., “Privacy risk in machine learning: Analyzing the connection to overfitting”, CSF 2018
[2] Sablayrolles, et al., “White-box vs black-box: Bayes optimal strategies for membership inference”, ICML 2019

Approach

- **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[1]
  - The Bayes optimal attack only depends on the sample loss[2]

- **Approach (RelaxLoss):**
  - Relaxing loss target with gradient ascent
  - Flattening the target posterior scores for non-ground-truth classes

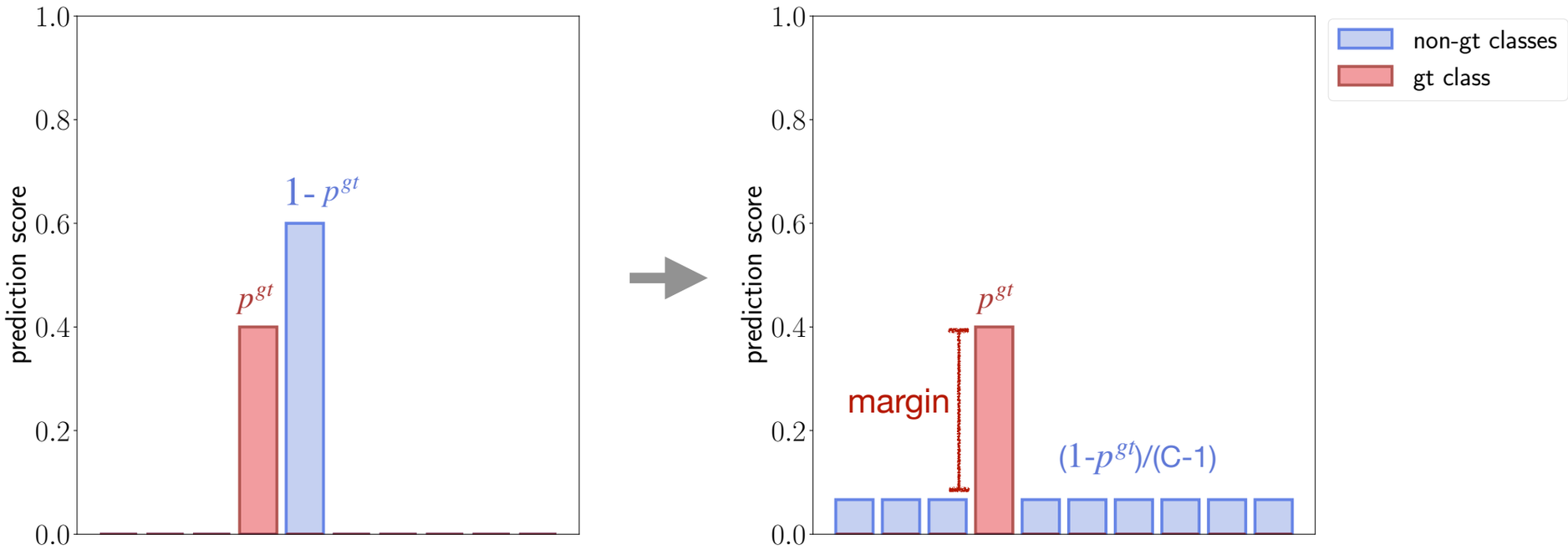


[1] Yeom et al., "Privacy risk in machine learning: Analyzing the connection to overfitting", CSF 2018
[2] Sablayrolles, et al., "White-box vs black-box: Bayes optimal strategies for membership inference", ICML 2019

Approach

- **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[1]
  - The Bayes optimal attack only depends on the sample loss[2]

- **Approach (RelaxLoss):**
  - Relaxing loss target with gradient ascent
  - Flattening the target posterior scores for non-ground-truth classes



[1] Yeom et al., "Privacy risk in machine learning: Analyzing the connection to overfitting", CSF 2018
[2] Sablayrolles, et al., "White-box vs black-box: Bayes optimal strategies for membership inference", ICML 2019

Evaluation

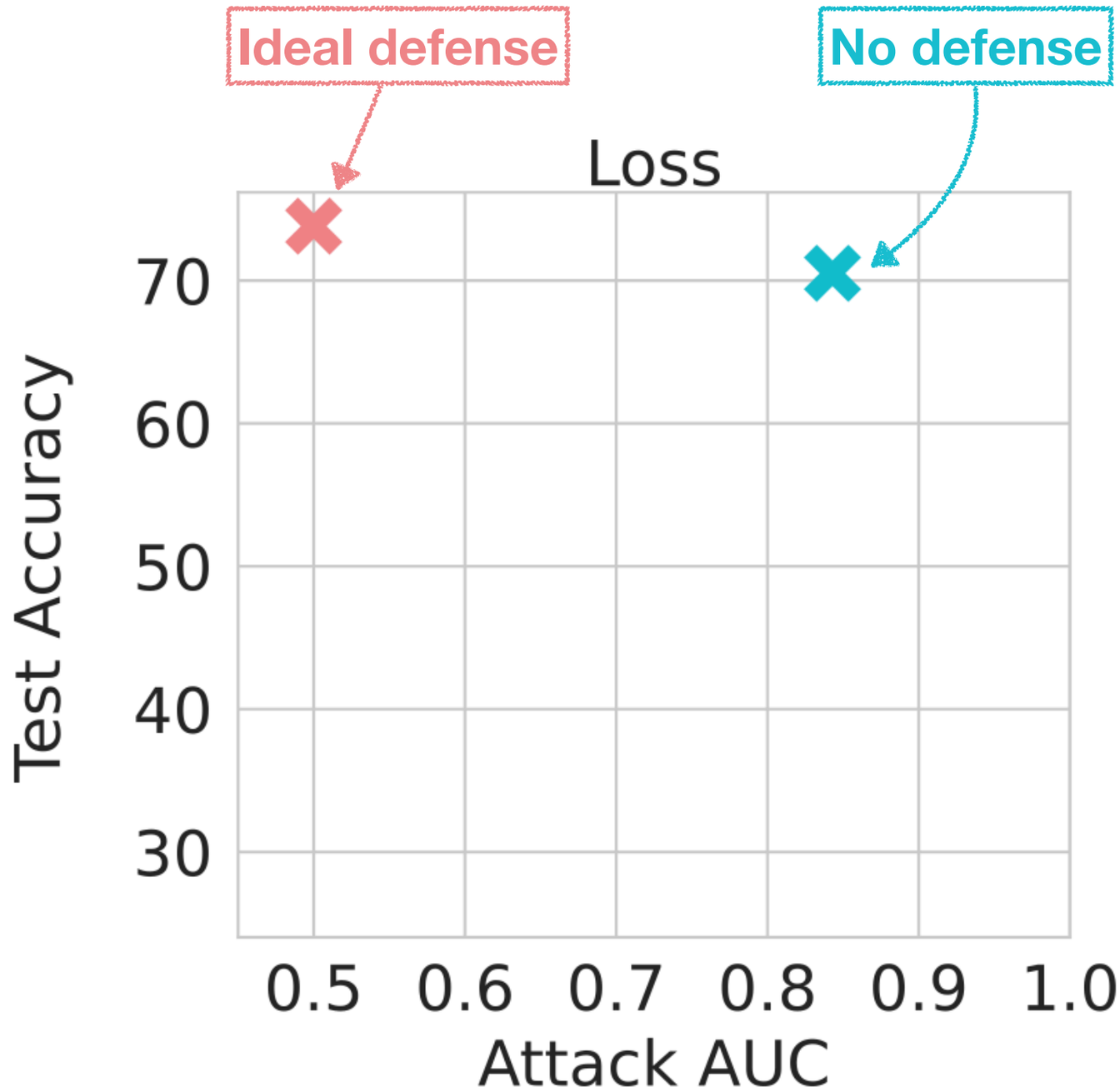
- **5 Datasets with diverse modalities**
  - CIFAR-10, CIFAR-100, CH-MNIST, Texas100, Purchase100

- **6 Attack methods**
  - **White-box:** Grad-x, Grad-w
  - **Black-box:** NN, Loss, Entropy, M-Entropy

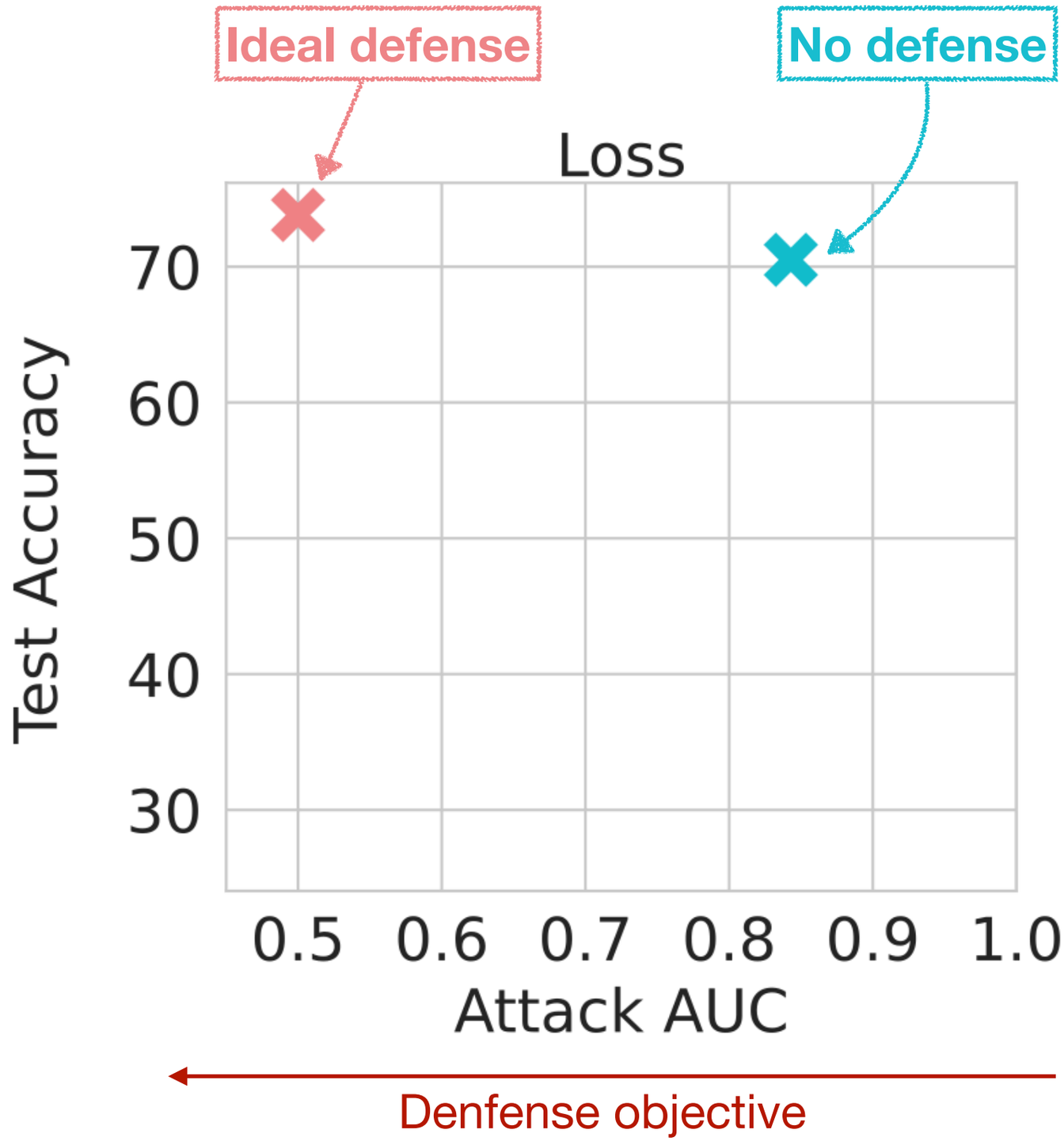
- **8 Defense baselines**
  - Memguard, Adv-Reg, Early-stopping, Dropout, Label-smoothing, Confidence-penalty, (Self-)Distillation, DP-SGD

- **Evaluation metrics**
  - **Utility**: Test accuracy of target models
  - **Defense effectiveness**: Attack accuracy; Attack AUC

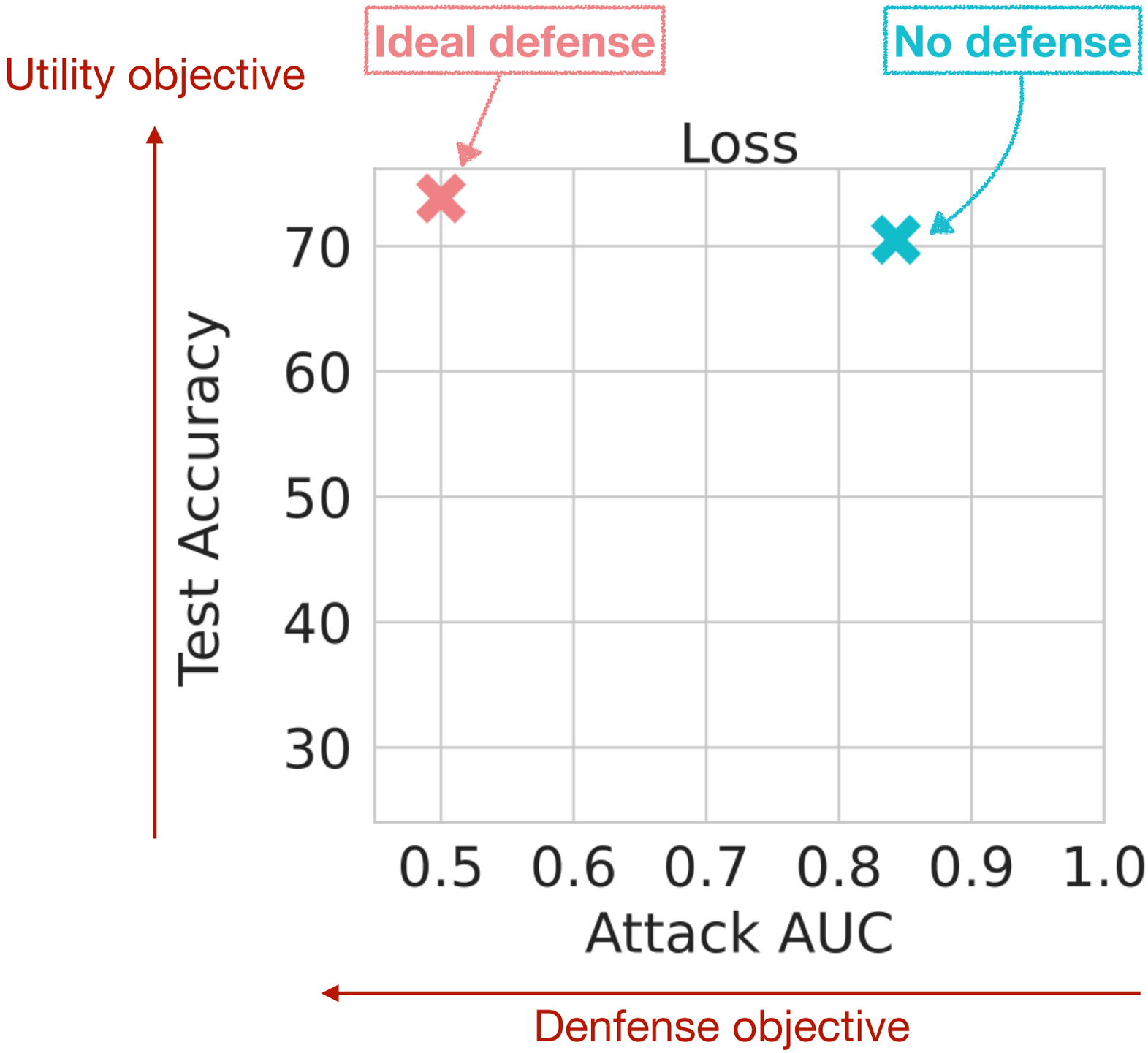
Results


Ideal defense
No defense
Loss
70
60
50
40
30
Test Accuracy
0.5
0.6
0.7
0.8
0.9
1.0
Attack AUC

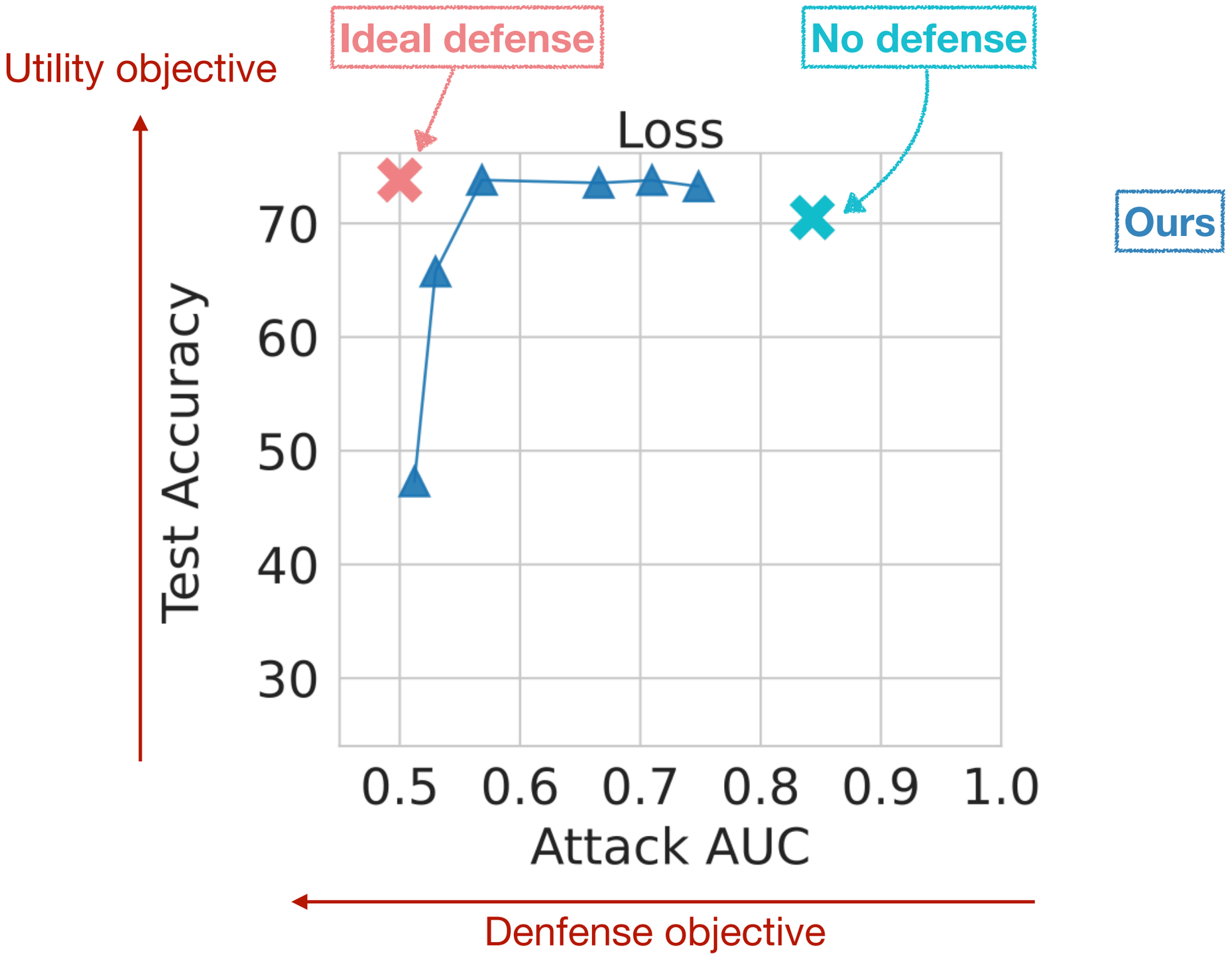
Results


Ideal defense
No defense
Loss
Test Accuracy
70
60
50
40
30
0.5
0.6
0.7
0.8
0.9
1.0
Attack AUC
Denfense objective

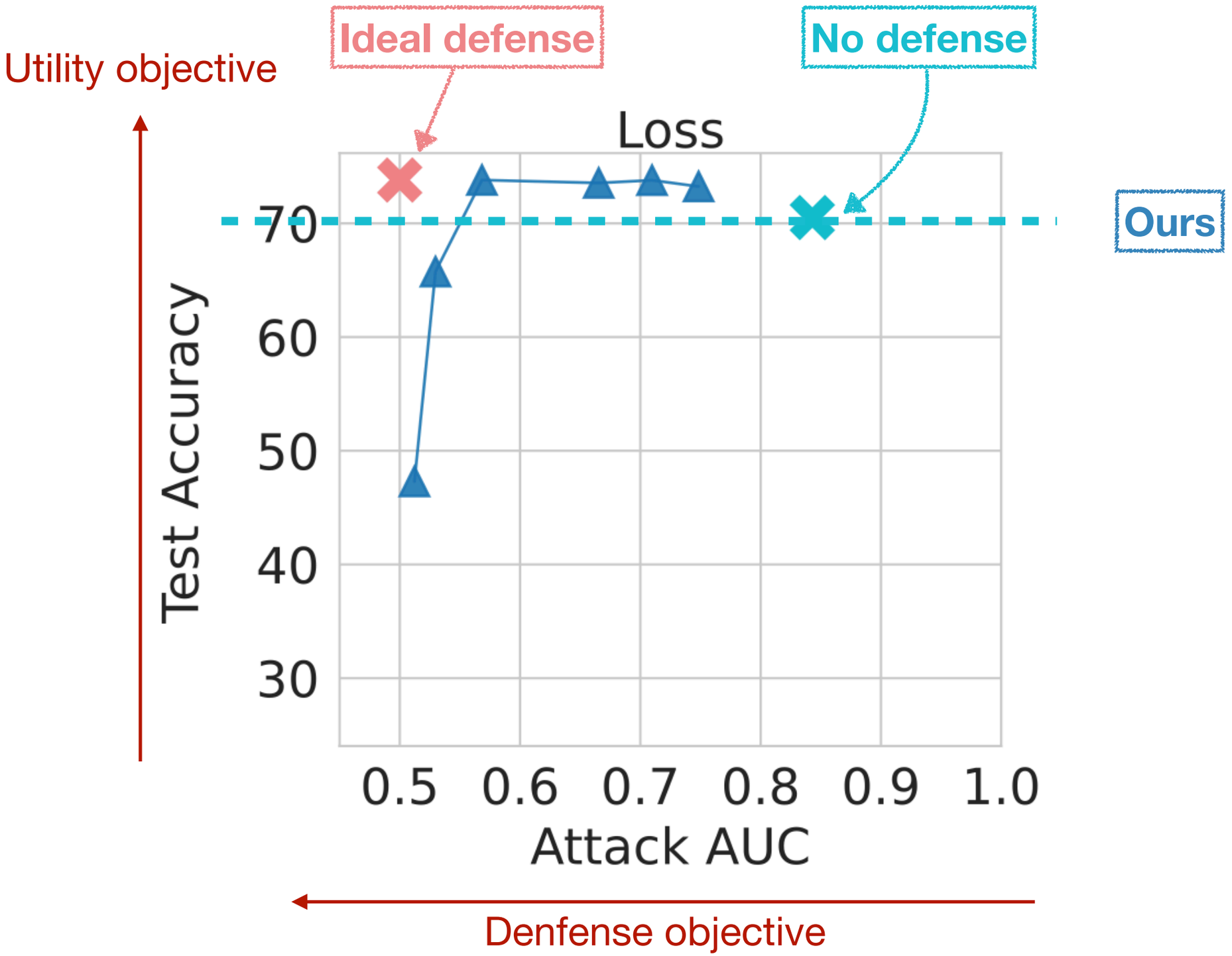
Results


Ideal defense
No defense
Utility objective
Loss
Test Accuracy
70
60
50
40
30
0.5
0.6
0.7
0.8
0.9
1.0
Attack AUC
Denfense objective

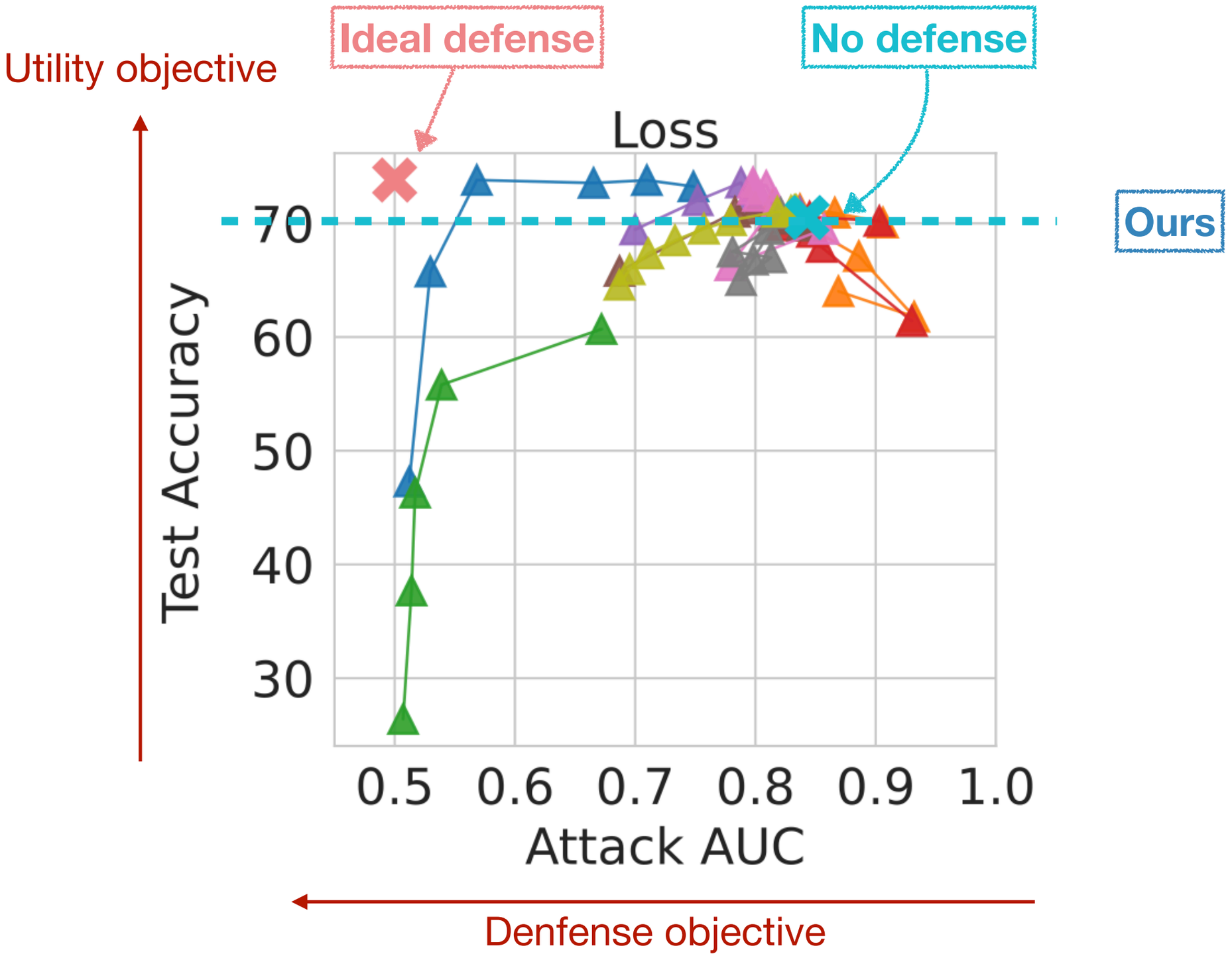
Results


Ideal defense
No defense
Ours
Utility objective
Loss
Test Accuracy
70
60
50
40
30
0.5
0.6
0.7
0.8
0.9
1.0
Attack AUC
Denfense objective

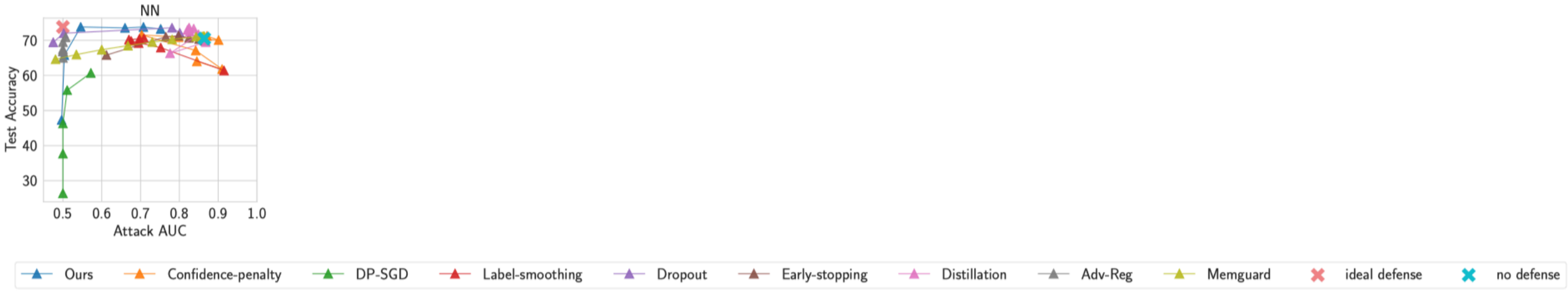
Results


Utility objective
Ideal defense
No defense
Loss
Ours
Test Accuracy
70
60
50
40
30
0.5
0.6
0.7
0.8
0.9
1.0
Attack AUC
Denfense objective

Results

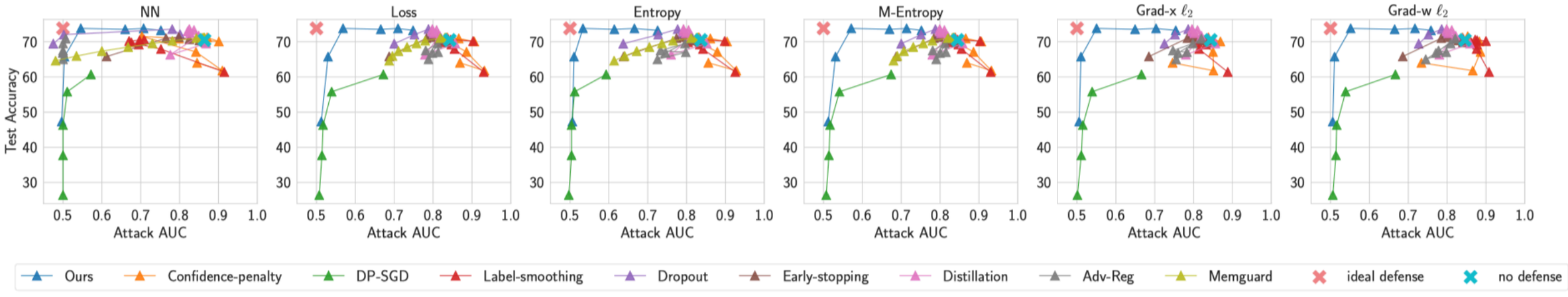

Ideal defense
No defense
Utility objective
Loss
Ours
70
60
50
40
30
Test Accuracy
0.5
0.6
0.7
0.8
0.9
1.0
Attack AUC
Denfense objective

Results



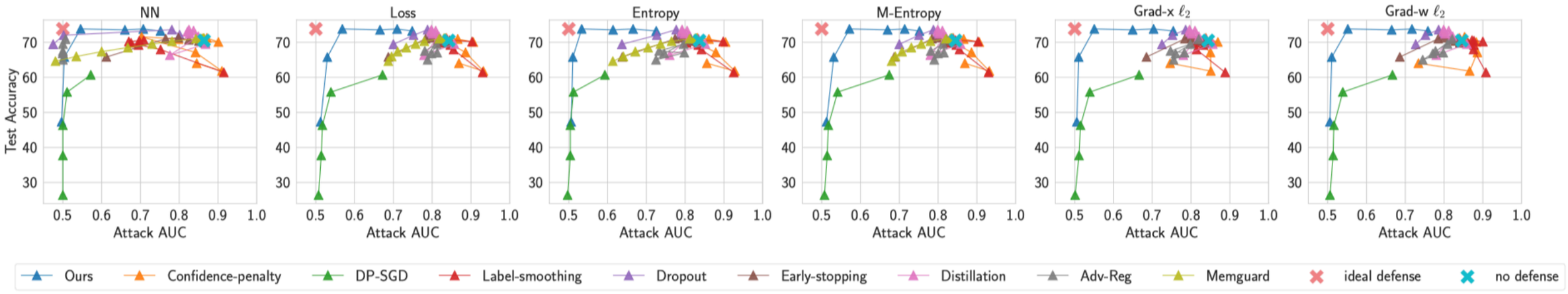
(a) CIFAR-10 (ResNet20)

Results

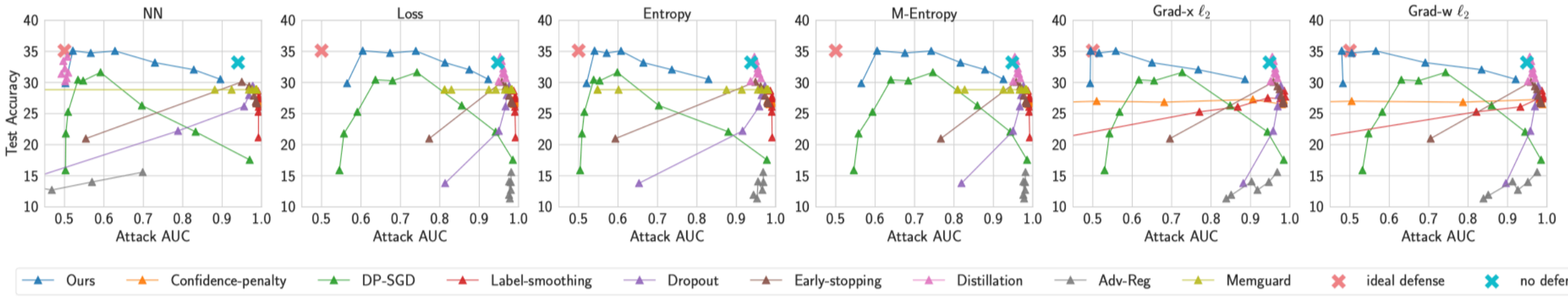


(a) CIFAR-10 (ResNet20)

Results

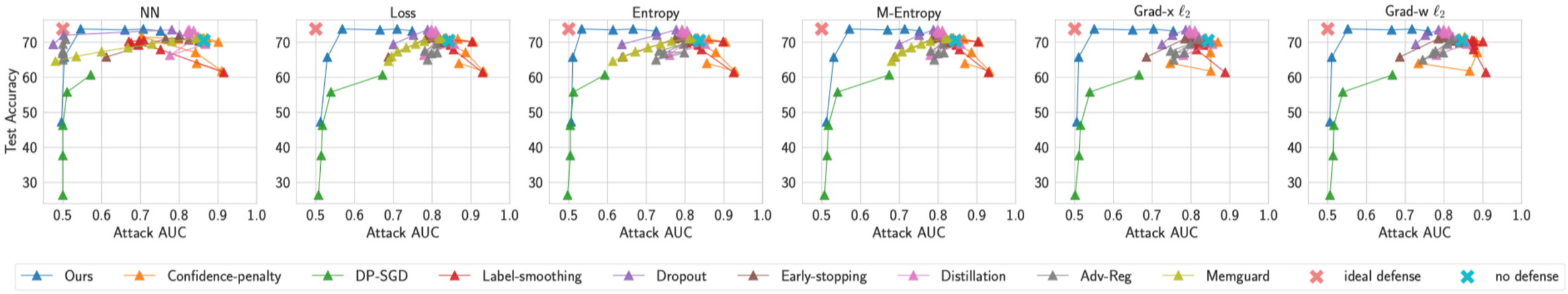


(a) CIFAR-10 (ResNet20)

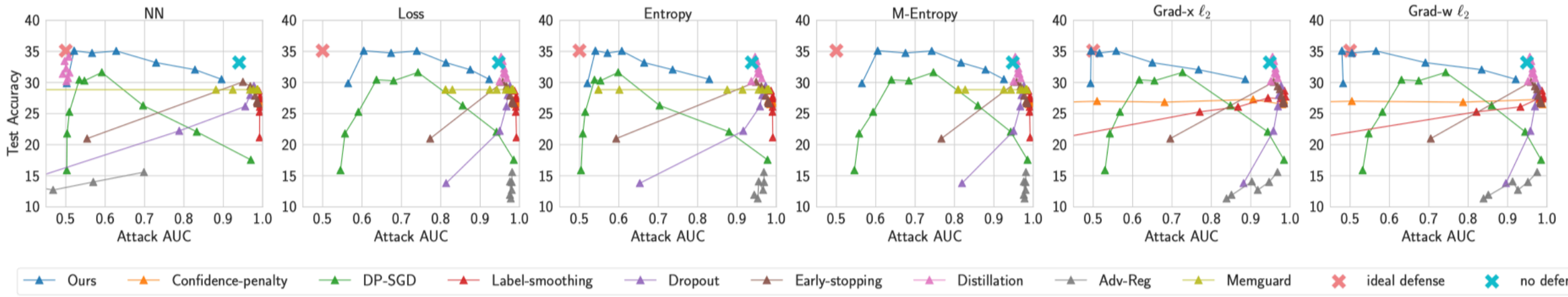


(b) CIFAR-100 (ResNet20)

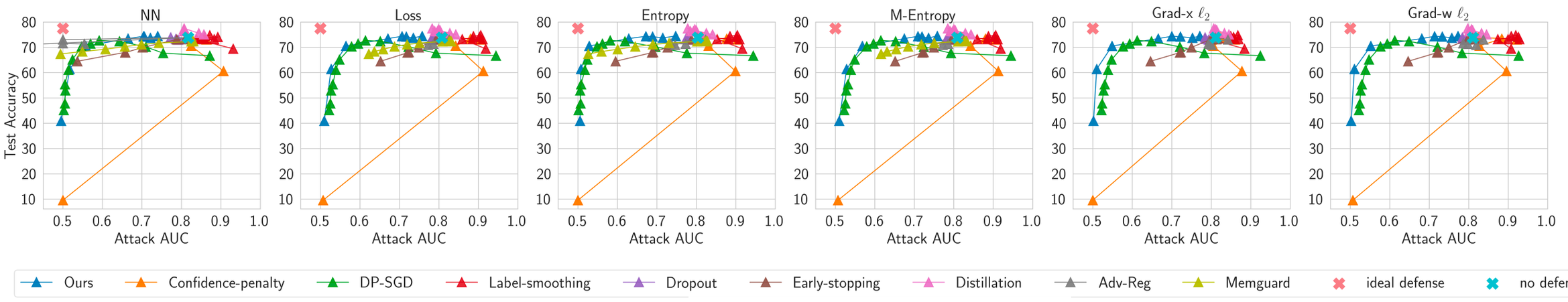
Results



(a) CIFAR-10 (ResNet20)



(b) CIFAR-100 (ResNet20)



(c) CIFAR-10 (VGG11)

More details in the paper

RelaxLoss: Defending Membership Inference Attacks without Losing Utility

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

Please visit our github repository for source code:

https://github.com/DingfanChen/RelaxLoss

Contact:
Dingfan Chen, dingfan.chen@cispa.de