Data Reconstruction Attacks and Defenses: From Theory to Practice

Qi Lei, Courant Math and CDS

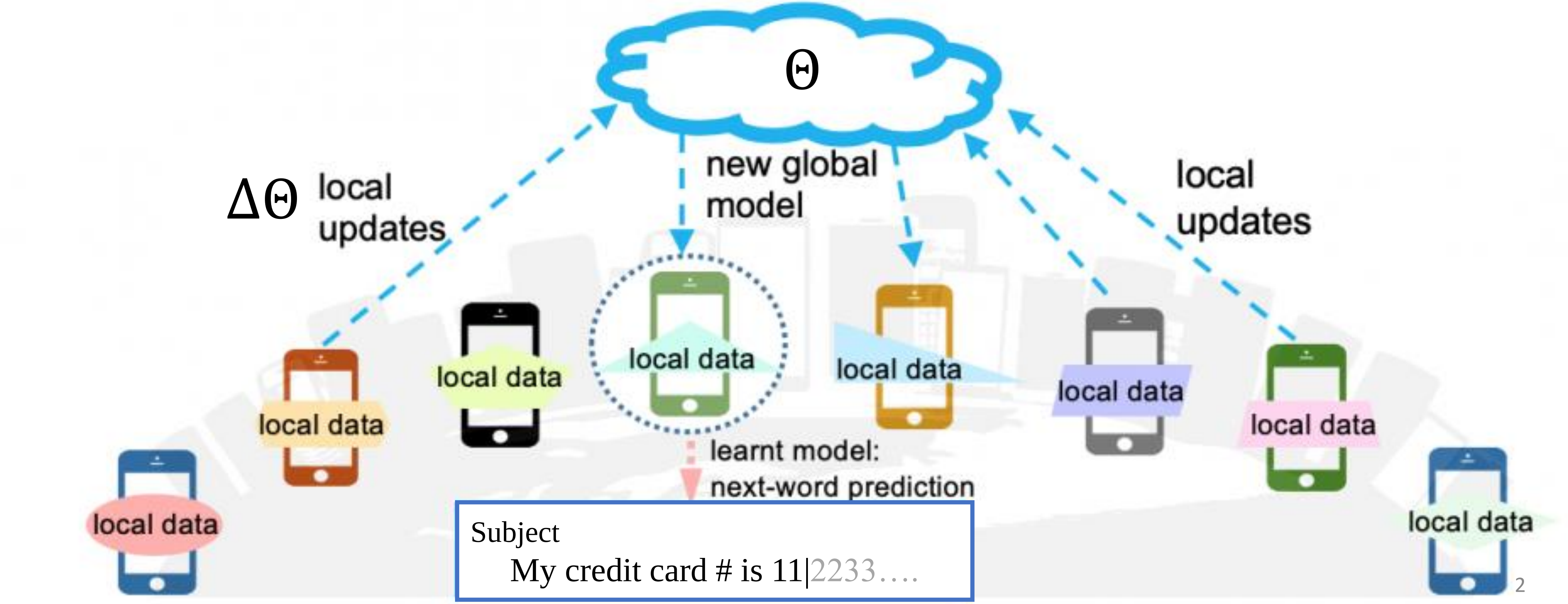
With Zihan Wang, Sheng Liu, Yuxiao Chen, Jianwei Li, Jason Lee

https://arxiv.org/abs/2212.03714
https://arxiv.org/abs/2402.09478
https://arxiv.org/abs/2312.05720

Privacy leakage

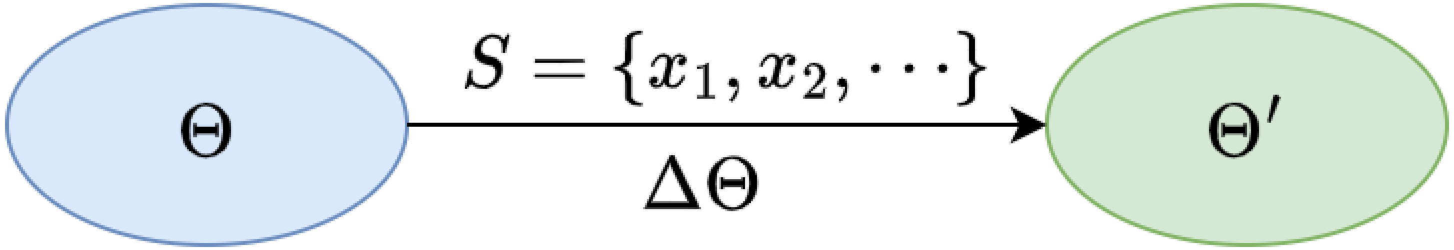
- Privacy leakage in distributed learning - data and model not co-located

  [Konečný et al. 2016, McMahan et al. 2017]

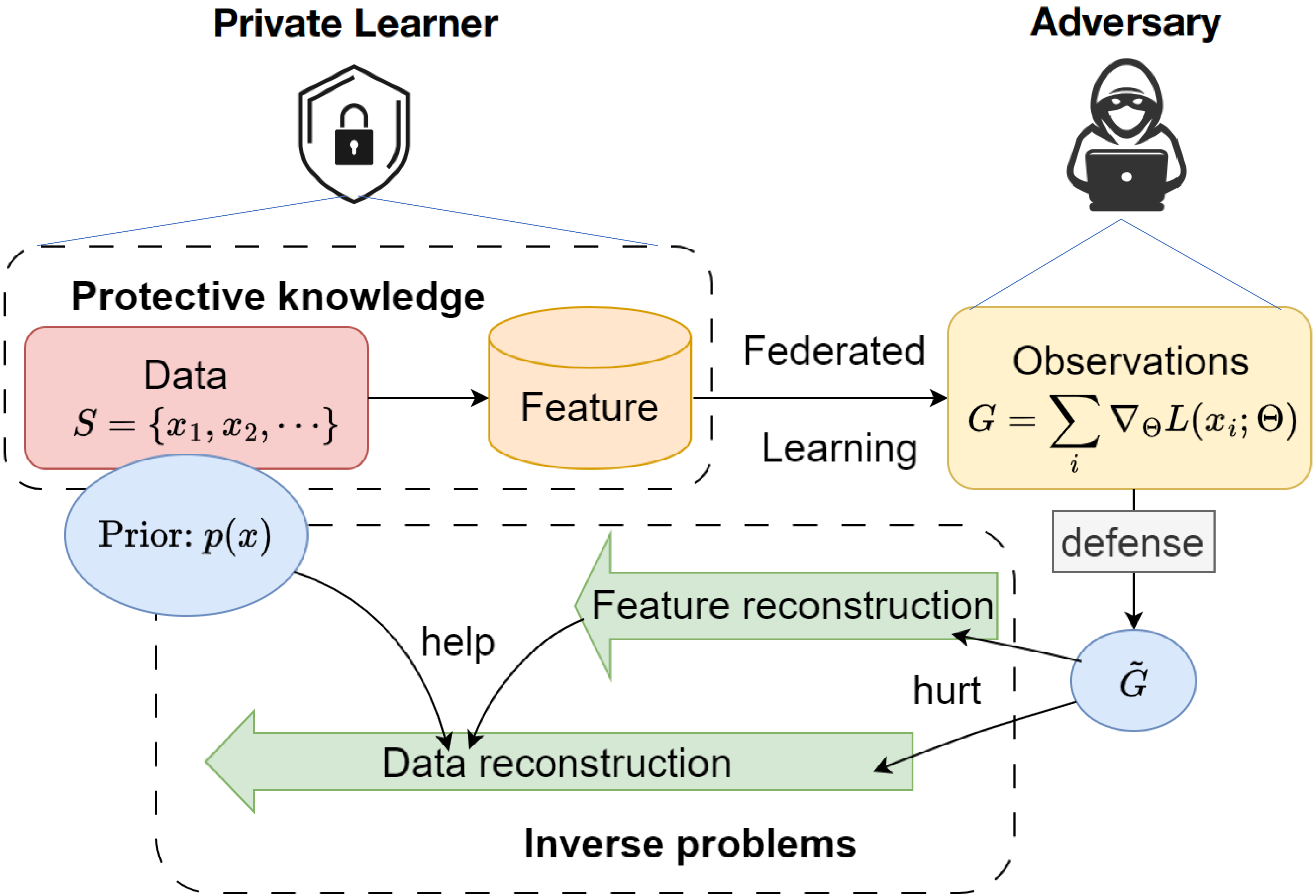


Privacy leakage

- Privacy leakage in fine-tuned model – trained with licensed/private data



- Question: When and how does our observation reveal the training data?


Private Learner
Adversary
Protective knowledge
Data
$S = \{x_1, x_2, \cdots\}$
Feature
Federated
Learning
Observations
$G = \sum_i \nabla_\Theta L(x_i; \Theta)$
Prior: $p(x)$
defense
Feature reconstruction
help
hurt
$\tilde{G}$
Data reconstruction
Inverse problems

Threat model more formally:

- Batch of data:
  - $S = \{(x_1, y_1), (x_2, y_2), \cdots, (x_B, y_B)\}$
- Prediction function:
  - $x \to f(x; \Theta)$
- Model update:
  - $G := \frac{1}{B} \nabla_\Theta \sum_{i=1}^{B} \ell(f(x_i, \Theta), y_i)$

Private learner

Adversary

- Inverse problem:
  - Recover S from G, Θ is known

Prior work

- Attacking methods
  - Gradient matching (gradient inversion):

$$\min_{S=\{(x_i, y_i)\}} \left|\left|G - \sum_{i=1}^{B} \nabla \ell(f(x_i; \Theta), y_i)\right|\right|^2$$

[Zhu et al., 2019; Yin et al., 2021; Jeon et al., 2021]

Prior work

- Attacking methods
  - Gradient matching (gradient inversion):

$$\min_{S=\{(x_i,y_i)\}} \left|\left|G - \sum_{i=1}^{B} \nabla \ell(f(x_i;\Theta), y_i)\right|\right|^2$$

  - Feature reconstruction through linear algebra techniques

[Wang, Lee, L, 2023, Kariyappa et al., 2023]

Prior work

- Attacking methods
  - Gradient matching (gradient inversion):

$$\min_{S=\{(x_i,y_i)\}} \left|\left|G - \sum_{i=1}^{B} \nabla \ell(f(x_i;\Theta), y_i)\right|\right|^2$$

  - Feature reconstruction through linear algebra techniques
  - Partial data reconstruction through fishing parameters

[Wen et al., 2022, Boenisch et al., 2023, Fowl et al., 2021]

Prior work

- Defending methods
  - Quantizing/pruning the gradient
  - Dropout
  - Secure aggregation
  - Multiple local aggregation

Quantizing/pruning the gradient, Dropout: Reduce observation's dimension

Secure aggregation, Multiple local aggregation: Increase unknown signal's dimension

Prior work

- Defending methods
  - Quantizing/pruning the gradient
  - Dropout
  - Secure aggregation
  - Multiple local aggregation
  - Add noise

Reduce observation to signal ratio (Quantizing/pruning the gradient, Dropout, Secure aggregation, Multiple local aggregation)

Add noise → Reduce observation to noise ratio

Prior work

- Theoretical analysis
  - Differential Privacy: more tailored for membership inference attack
    - Definition of (ϵ)-DP: can not distinguish any two neighboring datasets well (not much better than random guessing)
  - Renyi-DP: reconstructing last sample with other samples known
    - Distance measured in max divergence (DP) => in more relaxed choice of divergence
  - However: they only have constant conversion rate

[Dwork. 2006] [Guo et al 2022]

Prior work

- Theoretical analysis
  - Differential Privacy: more tailored for membership inference attack
  - Renyi-DP: reconstructing last sample with other samples known

Problems:

1. Not practical: For a model f with S_f sensitivity, adding Gaussian noise with variance $\frac{S_f^2}{\epsilon^2}$ will satisfy (ϵ) -DP
   - But in a 2-layer m-width neural network, $S_f \propto m$
2. Too strong: Not necessary in some scenarios:
   - $S = \{x_1, x_2, \cdots, x_B\}$, $G = x_1 + x_2 + \cdots + x_B$
   - No DP guarantee, but not possible to reconstruct (unless with prior information)

[Abadi et al. 2016] [Liu, Wang, Chen, L, 2024]

Prior work

- Theoretical analysis
  - Differential Privacy: more tailored for membership inference attack
  - Renyi-DP: reconstructing last sample with other samples known

- Instead, we want to achieve:
  - A more common trajectory in security:

    ➔ stronger attack ➔ stronger defense ➔…
  - algorithmic upper bound for the reconstruction error

Prior work

- Theoretical analysis
  - Differential Privacy: more tailored for membership inference attack
  - Renyi-DP: reconstructing last sample with other samples known

- Instead, we want to achieve:
  - A more common trajectory in security:

    ➔ stronger attack ➔ stronger defense ➔…
  - algorithmic upper bound for the reconstruction error

Prior work

- Theoretical analysis
  - Differential Privacy: more tailored for membership inference attack
  - Renyi-DP: reconstructing last sample with other samples known

- Instead, we want to achieve:
  - A more common trajectory in security:

    ➔ stronger attack ➔ stronger defense ➔…
  - algorithmic upper bound for the reconstruction error

Prior work

- Theoretical analysis
  - Differential Privacy: more tailored for membership inference attack
  - Renyi-DP: reconstructing last sample with other samples known

- Instead, we want to achieve:
  - A more common trajectory in security:

    ➔ stronger attack ➔ stronger defense ➔…
  - algorithmic upper bound for the reconstruction error
  - (hopefully matching) information theoretic lower bound

Prior work

- Theoretical analysis
  - Differential Privacy: more tailored for membership inference attack
  - Renyi-DP: reconstructing last sample with other samples known

- Instead, we want to achieve:
  - A more common trajectory in security:

    ➔ stronger attack ➔ stronger defense ➔…
  - algorithmic upper bound for the reconstruction error
  - (hopefully matching) information theoretic lower bound

Part I: Theoretical analysis under two-layer neural networks

Warm-up:

- Two-layer neural network

$$f(x; \{W, a\}) = \sum_{j=1}^{m} a_j \sigma\left(w_j^\top x\right) = a^\top \sigma(W^\top x)$$

- Observations G:

$$\nabla_{a_j} L = \sum_{i=1}^{B} l_i' \sigma\left(w_j^\top x_i\right), \nabla_{w_j} L = \sum_{i=1}^{B} l_i' \sigma'\left(w_j^\top x_i\right) x_i$$

(Bad) Examples:

- Linear activation:
- $\nabla_a L = W\left(\sum_{i=1}^{B} l_i' x_i\right); \nabla_W L = a\left(\sum_{i=1}^{B} l_i' x_i\right)^\top$
- Can only identify a linear combination of X

- Quadratic activation:
- $\nabla_{a_j} L = w_j^\top \bar{\bar{\Sigma}} w_j; \nabla_{w_j} L = 2\bar{\bar{\Sigma}} w_j,$ here $\bar{\bar{\Sigma}} = \sum_{i=1}^{B} l_i' x_i x_i^\top$
- Can only identify the span of X

Our goal:

- Upper bound:
  - $R_U(A) \coloneqq \max_S d\big(S, A(O)\big)$,
  - Distance metric: $d\big(S, \hat{S}\big) \coloneqq \min_\pi \sqrt{\frac{1}{B}\sum_i ||S_i - \widehat{S_{\pi(i)}}||^2}$ (up to permutation)
  - No defense: O=G, with defense: O=D(G)
- Lower bound:
  - $R_L = \min_{\hat{S}=A(O)} \max_S d\big(S, \hat{S}\big)$
  - No defense: O=G+$\epsilon, \epsilon \sim N(0, \sigma^2)$, with defense: O=D(G) +$\epsilon$
- Remark: our focus is on properties of model architecture/weight + defense method (not on data)

Algorithmic upper bound on defenses

Defense	Upper bound
No defense	$\tilde{O}\left(B\sqrt{d/m}\right)$
Local aggregation	$\tilde{O}\left(KB\sqrt{d/m}\right)$
σ^2-gradient noise	$\tilde{O}\left((B+\sigma)\sqrt{d/m}\right)$
DP-SGD	$\tilde{O}\left((B+\sigma\max\{1,\|G\|/\mathrm{C}\})\sqrt{d/m}\right)$
p-Dropout	$\tilde{O}\left(B\sqrt{d/(1-p)m}\right)$
Gradient pruning:	unknown

[Liu, Wang, Chen, L, 2024] https://arxiv.org/abs/2402.09478

How: recover third moment of data

- We want to estimate $T_p := \sum_{i=1}^{B} E_w\left[\sigma^{(p)}(w^\top x_i)\right] x_i^{\otimes p}$
- Uniquely identify $\{x_1, x_2, \cdots, x_B\}$ through tensor decomposition when data is linearly independent for p>=3. [Kuleshov et al. 2015]

- Our strategy: choose $a_j = \frac{1}{m}, w_j \sim N(0, I)$, estimate T by $\widehat{T_3} := \frac{1}{m}\sum_{j=1}^{m} g(w_j) H_3(w_j),\ g(w_j) := \nabla_{a_j} L = \sum_{i=1}^{B} l_i' \sigma(w_j^\top x_i)$

[Wang, Lee, L, 2023] https://arxiv.org/abs/2212.03714

Tensor decomposition

- Stein's lemma: $E_{w\sim N(\mathbb{0},I)}\left[g(a^\top w)H_p(w)\right] = E\left[g^{(p)}a^{\otimes p}\right]$.
- Hermite function: $H_2(w) = ww^\top - I, H_3(w) = w^{\otimes 3} - w \widetilde{\otimes} I$.

- $\widehat{T_p} := \frac{1}{m}\sum_{j=1}^{m} g(w_j)H_p(w_j) \approx E_{w\sim N(\mathbb{0},I)}\left[g(w)H_p(w)\right]$

$$\equiv \sum_{i=1}^{B} E\left[\sigma^{(p)}(w^\top x_i)x_i^{\otimes p}\right] =: T_p$$

- $g(w_j) := \nabla_{a_j} L = \sum_{i=1}^{B} l_i' \sigma(w_j^\top x_i)$ is our observation from the model gradient

[Wang, Lee, L, 2023] https://arxiv.org/abs/2212.03714

Algorithmic upper bound on attacks

- Applies when $E\left[\sigma^{(3)}(w)\right]$ or $E\left[\sigma^{(4)}(w)\right] \neq 0$. Applies to sigmoid, tanh, ReLU, leaky ReLU, GeLU, SELU, ELU etc.
- Reconstruction error $\leq \tilde{O}\left(\sqrt{d/m}\right)$.

[Wang, Lee, L, 2023] https://arxiv.org/abs/2212.03714

Our goal:

- Upper bound:
  - $R_U(A) \coloneqq \max_S d(S, A(O))$,
  - Distance metric: $d(S, \hat{S}) \coloneqq \min_\pi \frac{1}{B} \sqrt{\sum_i ||S_i - \widehat{S_{\pi(i)}}||^2}$ (up to permutation)
  - No defense: O=G, with defense: O=D(G)
- Lower bound:
  - $R_L = \min_{\hat{S}=A(O)} \max_S d(S, \hat{S})$
  - No defense: O=G+$\epsilon, \epsilon \sim N(0, \sigma^2)$, with defense: O=D(G) +$\epsilon$
- Remark: our focus is on properties of model architecture/weight + defense method (not on data)

Comparisons with information-theoretic lower bound on defenses

defense	Upper bound	Lower bound
No defense	$\tilde{O}\left(B\sqrt{d/m}\right)$	$\Omega\left(\sigma\sqrt{d/m}\right)$
Local aggregation	$\tilde{O}\left(KB\sqrt{d/m}\right)$	$\Omega\left(\sigma\sqrt{d/m}\right)$
σ^2-gradient noise	$\tilde{O}\left((B+\sigma)\sqrt{d/m}\right)$	$\Omega\left(\sigma\sqrt{d/m}\right)$
DP-SGD	$\tilde{O}\left((B+\sigma\max\{1,\|G\|/C\})\sqrt{d/m}\right)$	$\Omega\left(\sigma\max\{1,\|G\|/C\}\sqrt{d/m}\right)$
p-Dropout	$\tilde{O}\left(B\sqrt{d/(1-p)m}\right)$	$\Omega\left(\sqrt{d/(1-p)m}\right)$
Gradient pruning:	unknown	$\Omega\left(\sqrt{d/(1-\hat{p})m}\right)$

[Liu, Wang, Chen, L, 2024] https://arxiv.org/abs/2402.09478

Lower bound analysis

- (Bayesian) Cramer-rao: $R_L^2 \geq \sigma^2 \mathrm{Tr}((JJ^T)^{-1})$
  - J is Jacobian of the forward function (after defense): $F: S \rightarrow D\big(\nabla L(S; \Theta)\big)$
  - Key factor: how is J modified, ill-conditioned

- Connection to the linear and quadratic examples:
  - When Jacobian is singular, generally hard to reconstruct.

Take-away on the theoretical results:

- This is a promising framework (with matched dependence on d,m,p,C)

- The analysis is focused on properties of model architectures/weights, defense strength, not data (worst case of data, no prior info).

- Lower bound analysis is general, upper bound is more restrictive. (Need new tools to go beyond two-layer networks)

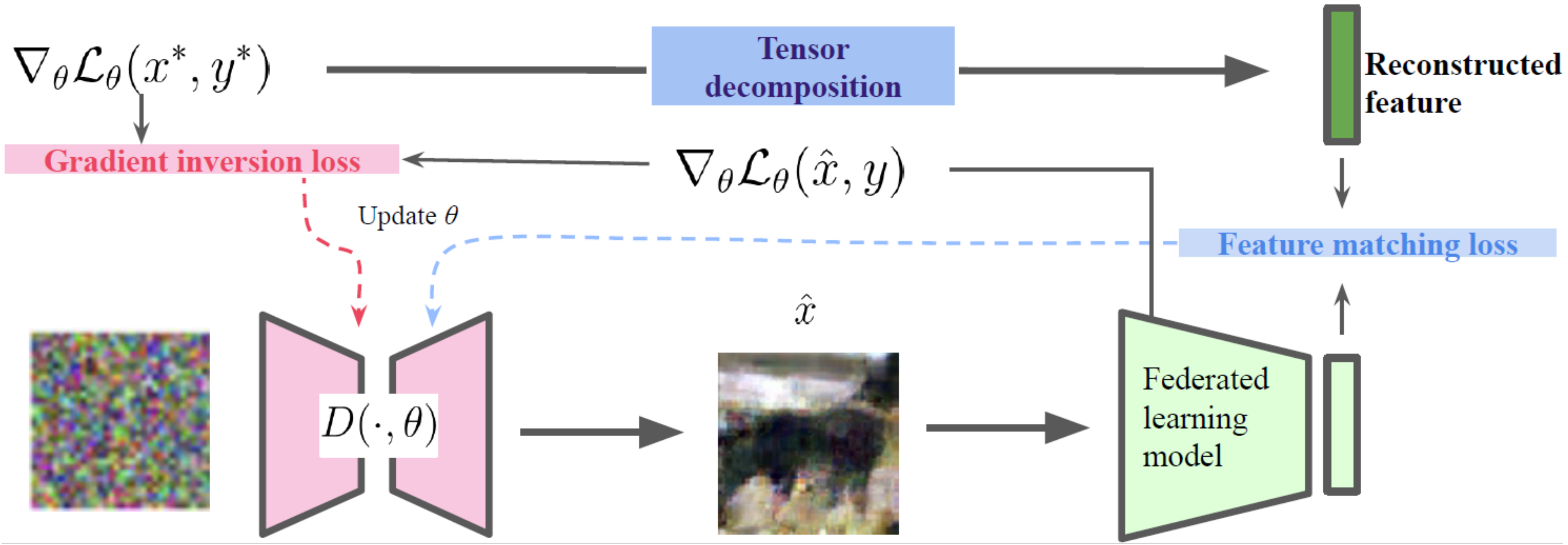
- Did not analyze utility-privacy trade-off

Part II: To go beyond

To go beyond

- Beyond two-layer networks
  - Empirical studies on general architectures
- Comparisons across various defense types
  - Exploit utility-privacy trade-off
  - $\text{Strength}(D) = \max_{A} d\left(S, A\big(D(G)\big)\right)$. Compare D with similar utility loss
- Beyond images
  - Exploit discrete data like text, or time series

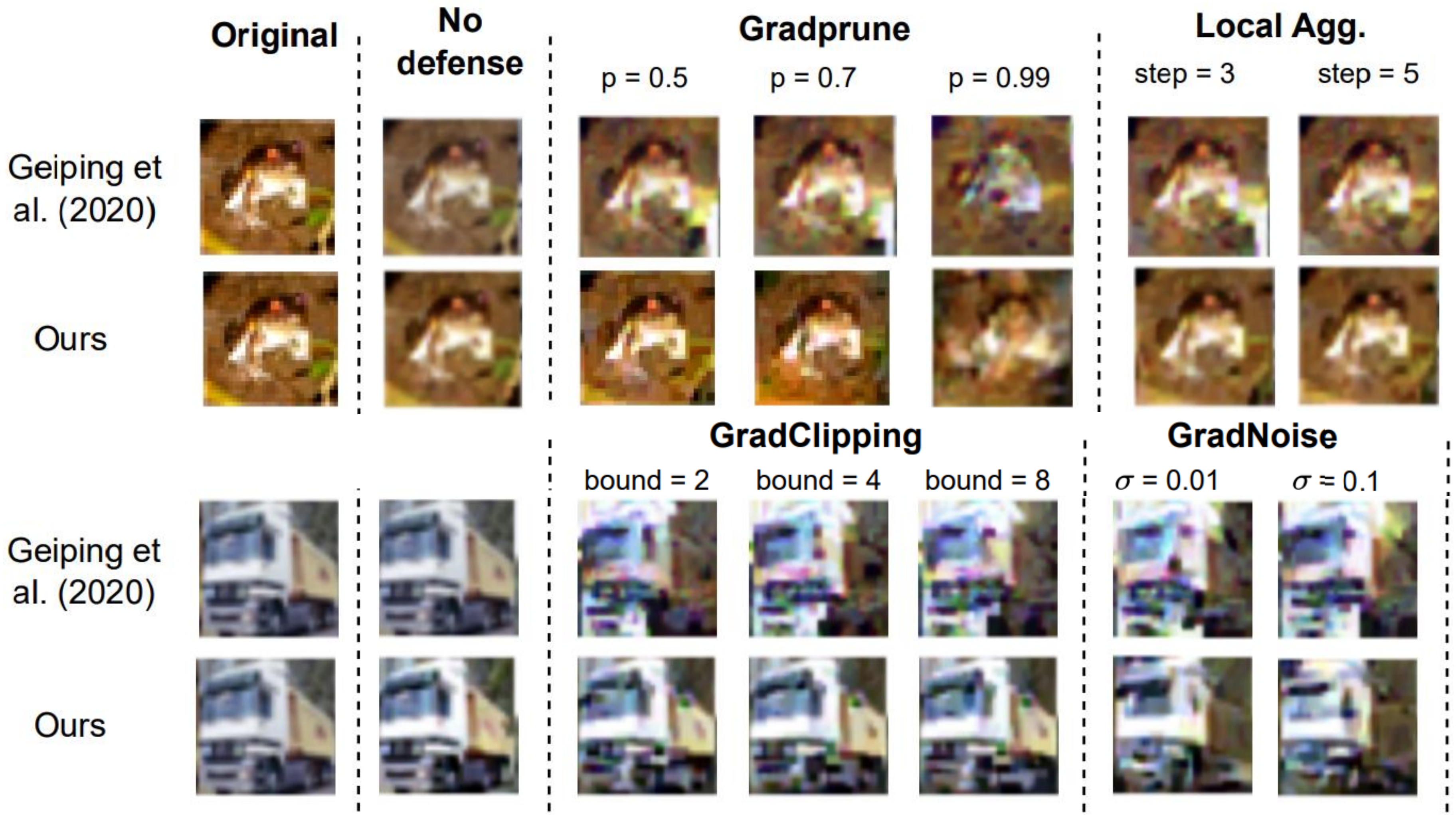
Beyond two-layer networks



- Previous findings: if last two layers are fully connected, can recover the features from the $(l-2)$-th layer
- Other structured data modalities: recover the embeddings first

[Liu, Wang, Chen, L, 2024] https://arxiv.org/abs/2402.09478

Empirical results:

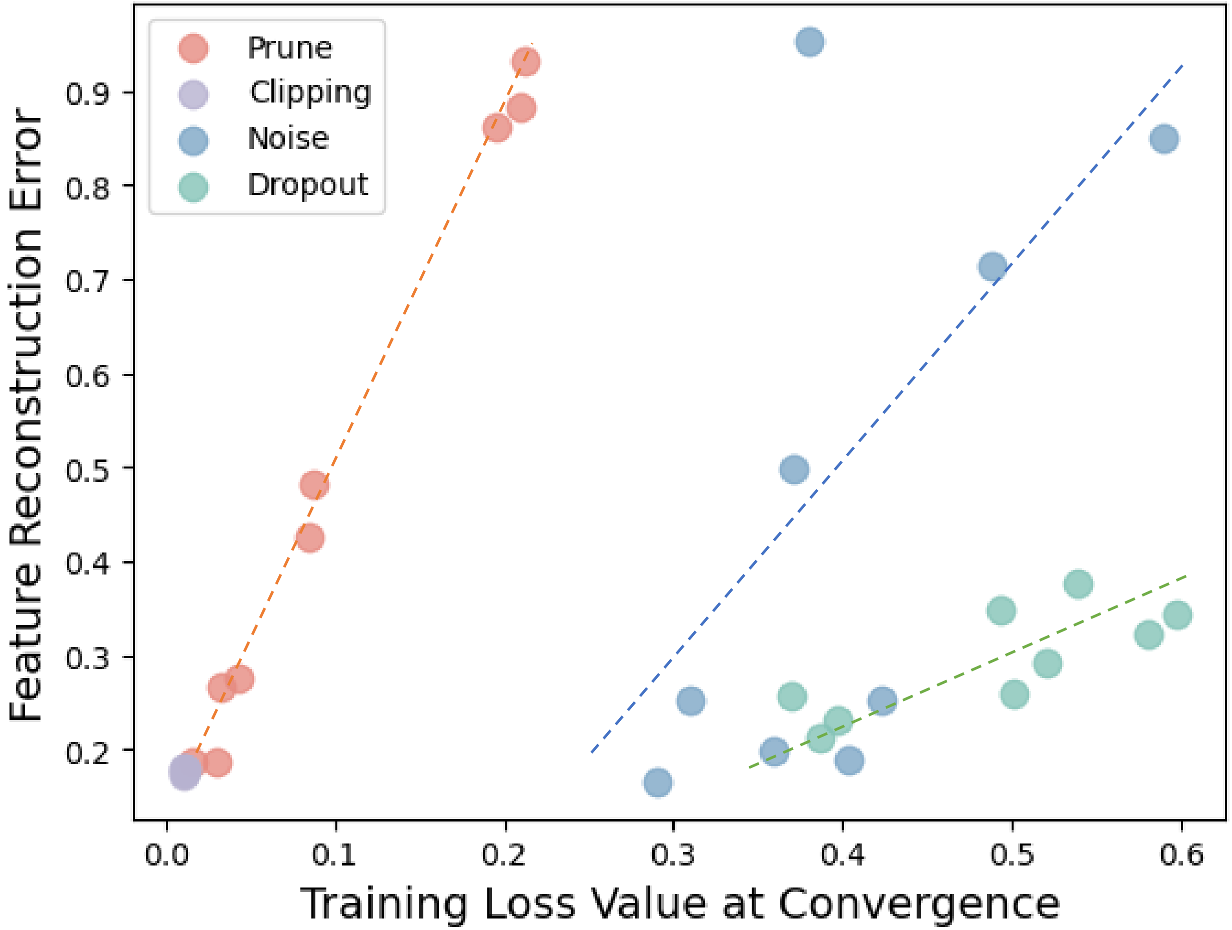


[Liu, Wang, Chen, L, 2024] https://arxiv.org/abs/2402.09478

To go beyond

- Beyond two-layer networks
  - Empirical studies on general architectures
- Comparisons across various defense types
  - Exploit utility-privacy trade-off
  - $\text{Strength}(D) = \max_{A} d\left(S, A\big(D(G)\big)\right)$. Compare D with similar utility loss
- Beyond images
  - Exploit discrete data like text, or time series

Privacy-utility trade-offs



[Liu, Wang, Chen, L, 2024] https://arxiv.org/abs/2402.09478

To go beyond

- Beyond two-layer networks
  - Empirical studies on general architectures
- Comparisons across various defense types
  - Exploit utility-privacy trade-off
  - $\text{Strength}(D) = \max_{A} d\left(S, A\big(D(G)\big)\right).$ Compare D with similar utility loss
- Beyond images
  - Exploit discrete data like text, or time series

Beyond computer vision tasks…

Dataset	Method	R-1	R-2	R-L	Coss	Recovered Samples
CoLA	reference sample: The box contains the ball					
	LAMP	15.5	2.6	14.4	0.36	likeTHETw box contains divPORa
	Ours	**17.4**	**3.8**	**15.9**	**0.41**	like Mess box contains contains balls
SST2	reference sample: slightly disappointed					
	LAMP	**20.1**	**2.2**	15.9	0.56	likesmlightly disappointed a
	Ours	19.7	2.1	**16.8**	**0.59**	like lightly disappointed a
Toma	reference sample: vaguely interesting, but it's just too too much					
	LAMP	19.9	1.6	15.1	0.48	vagueLY', interestingtooMuchbuttoojusta
	Ours	**21.5**	**1.8**	**16.0**	**0.51**	vagueLY, interestingBut seemsMuch Toolaughs

More results in: [Li, Liu, L, 2024] https://arxiv.org/abs/2312.05720

Discussions

- Call for more theoretical analysis under the inverse problem framework
  - Computational barrier for lower bound result
  - Need new tools to go beyond two-layer networks for upper bound

- Study how data properties (ill-conditioned, prior knowledge) affect the vulnerability to privacy attacks

- Based on $\text{Strength}(D) = \max_{A} d\left(S, A\big(D(G)\big)\right)$, gradient pruning is the strongest. Call for more evaluations when stronger attacks are proposed.

Thank you