Data Reconstruction Attacks and Defenses: From Theory to Practice

Qi Lei, Courant Math and CDS

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

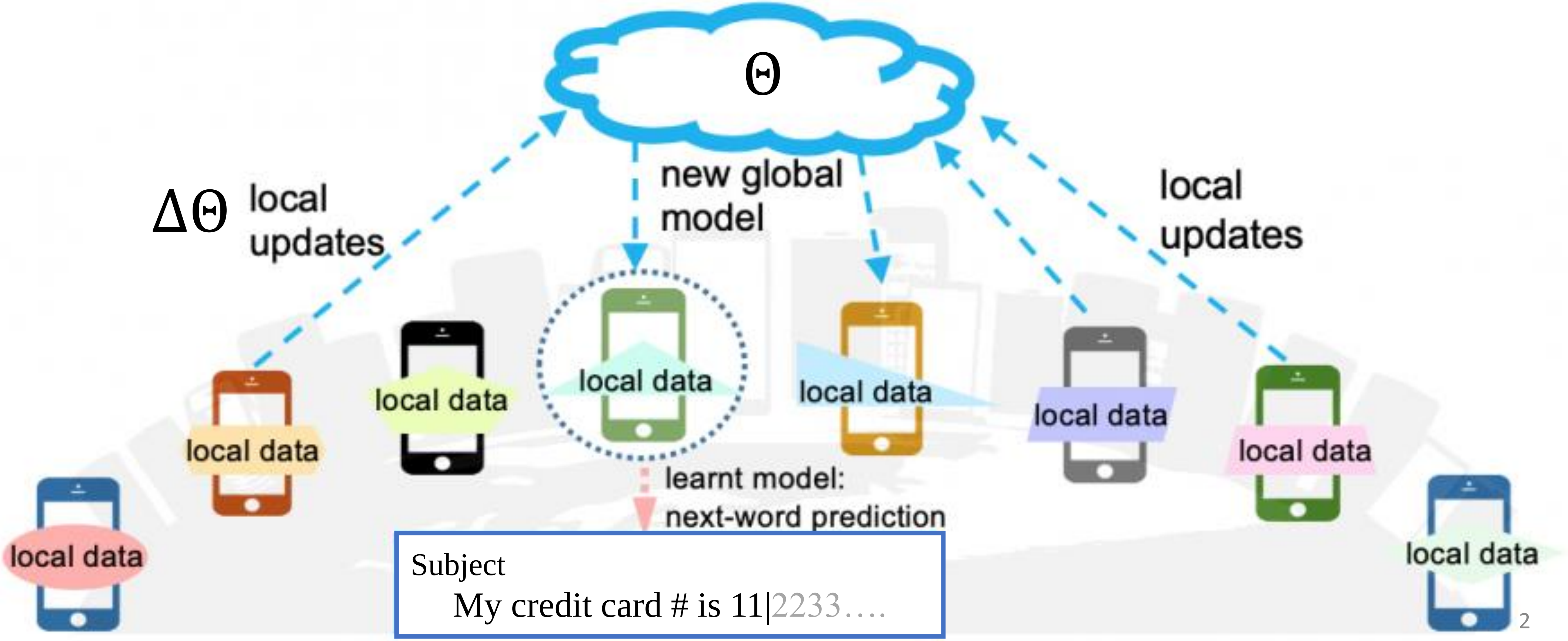
58th Annual Conference on Information Sciences and Systems

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

Federated learning

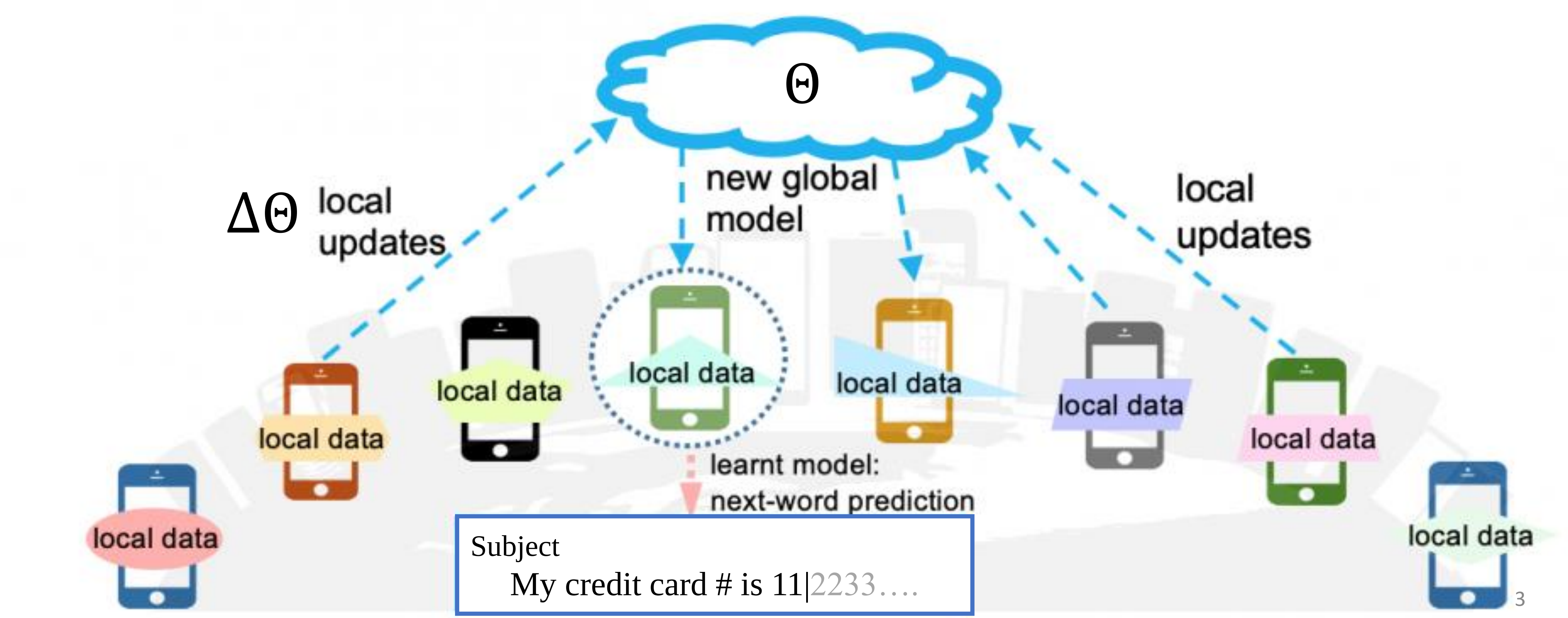
[Konečný et al. 2016, McMahan et al. 2017]

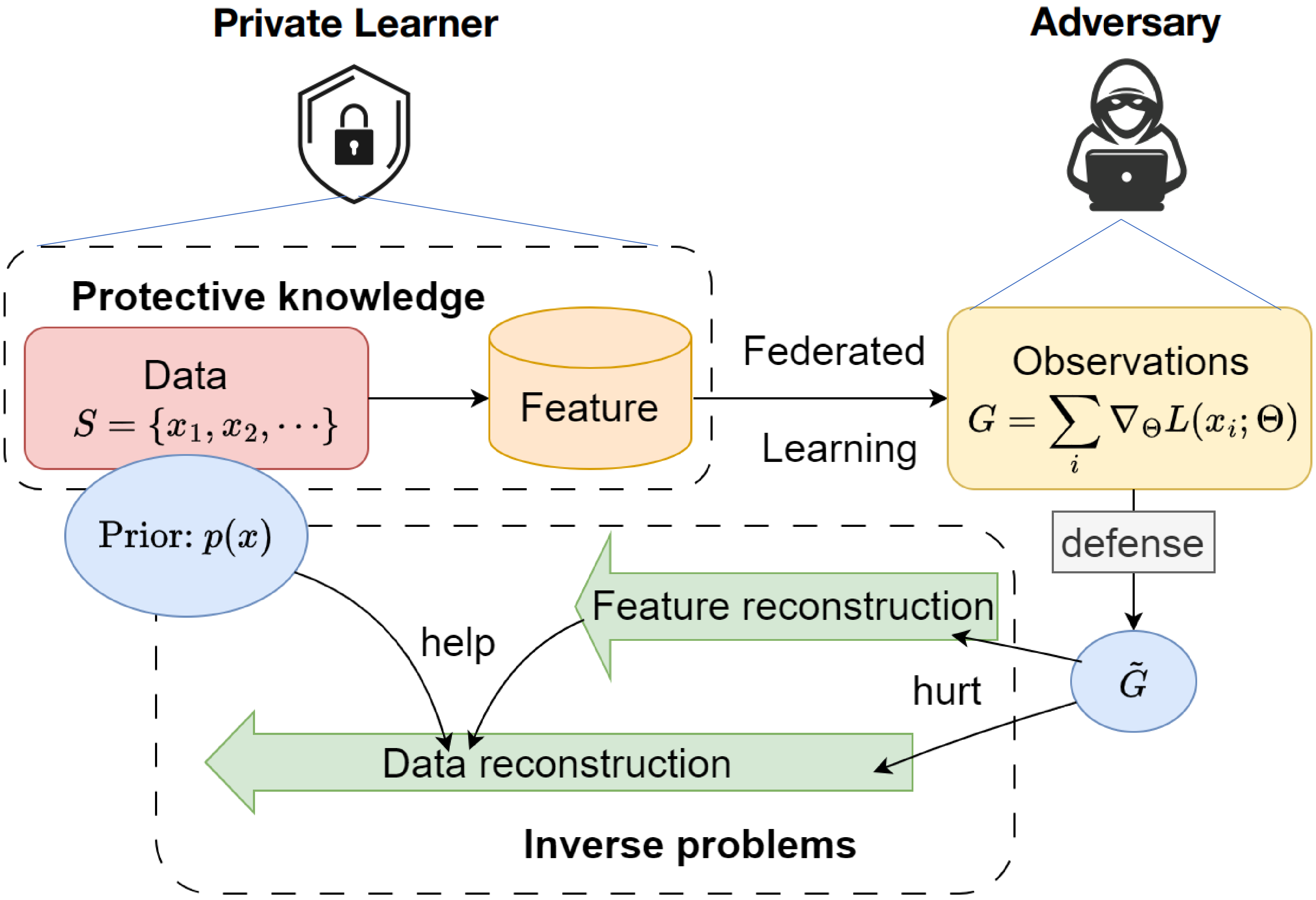
- Privacy leakage in distributed learning - Data and model not co-located



Privacy leakage in distributed learning

- Does local update reveal the training data?





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

Threat model more formally:

Private learner

- Local batch of data:
  - $S = \{(x_1, y_1), (x_2, y_2), \cdots, (x_B, y_B)\}$

Adversary

- Prediction function:
  - $x \to f(x; \Theta)$
- Local update:
  - $\mathrm{G} := \frac{1}{B} \nabla_\Theta \sum_{i=1}^{B} \ell(f(x_i, \Theta), y_i)$

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

Fundamental questions

- When is the model gradient G sufficient to identify the training samples?

- If so, is there an efficient algorithm to recover the samples?

Fundamental questions

- When is the model gradient G sufficient to identify the training samples?

  ➔ defending with information-theoretic bottleneck guarantees

- If so, is there an efficient algorithm to recover the samples?

  ➔ defending with computational barriers

Prior work

- Attacking methods
  - Learn to generate the training samples from a local user
  - Match the gradient: $\min_{S=\{(x_i, y_i)\}} \left|\left| G - \sum_{i=1}^{B} \nabla \ell(f(x_i; \Theta), y_i) \right|\right|^2$

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

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

Prior work

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

Illustrating example:

- $S = \{x_1, x_2, \cdots, x_B\}$, G $= x_1 + x_2 + \cdots + x_B$
- No DP guarantee, not possible to reconstruct (unless with prior information)

[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

- A more common trajectory in security:

➔ stronger attack ➔ stronger defense➔…

[Guo et al. 2022]

Warm-up:

- Two-layer neural network

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

- Choose proper $\Theta = \{w_j, a_j\}$ to query the gradient at

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

Our findings: 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 $\widehat{T_3} := \frac{1}{m} \sum_{j=1}^{m} g(w_j) H_3(w_j),\ g(w_j) := \nabla_{a_j} L$

[Wang et al. 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: $H_2(w) = ww^\top - I, H_3(w) = w^{\otimes 3} - w \,\widetilde{\otimes}\, I$.

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

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

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

[Wang et al. 2023] https://arxiv.org/abs/2212.03714

Theoretical analysis on attack

- 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 $\leq \tilde{O}\left(\sqrt{d/m}\right)$.

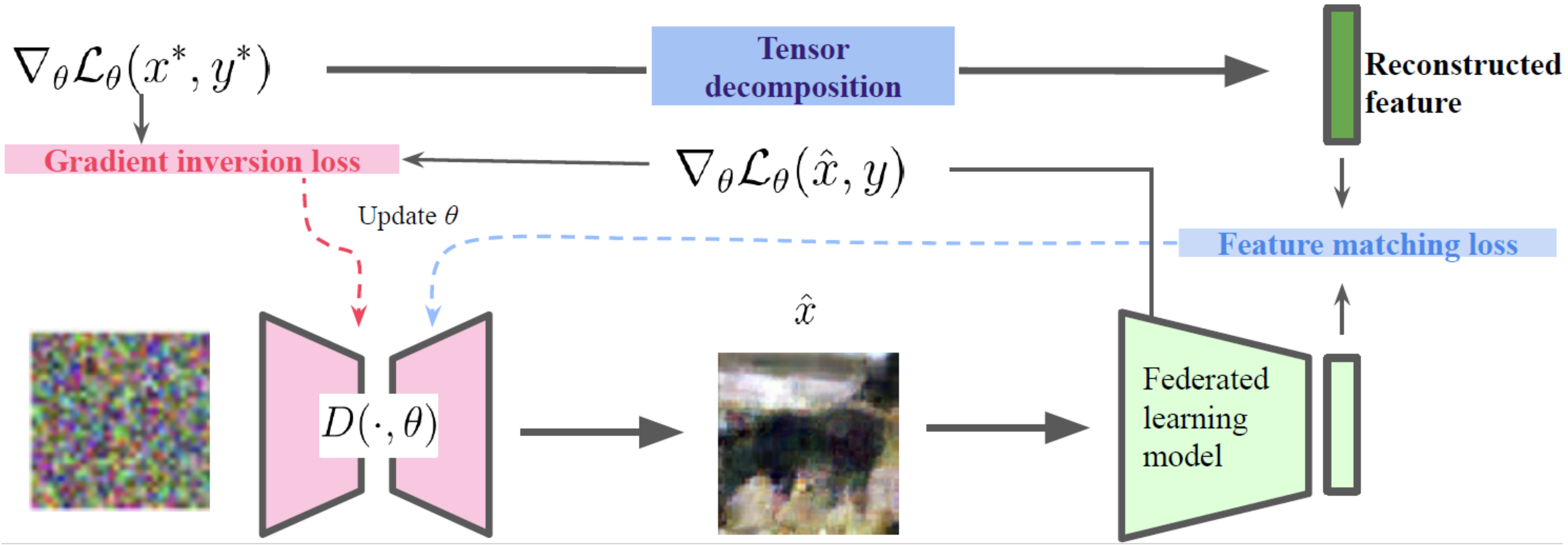
[Wang et al. 2023] https://arxiv.org/abs/2212.03714

Theoretical analysis on defense

- No defense: error $\leq \tilde{O}\left(\sqrt{d/m}\right)$
- K-Local aggregation: error $\leq \tilde{O}\left(K\sqrt{d/m}\right)$
- Add σ^2-noise: error $\leq \tilde{O}\left(\sqrt{(1+\sigma^2)d/m}\right)$
- K-Secure aggregation: error $\leq \tilde{O}\left(K\sqrt{d/m}\right)$
- p-Dropout: error $\leq \tilde{O}\left(\sqrt{d/pm}\right)$
- Gradient pruning: not applicable

[Liu et al. 2024] https://arxiv.org/abs/2402.09478

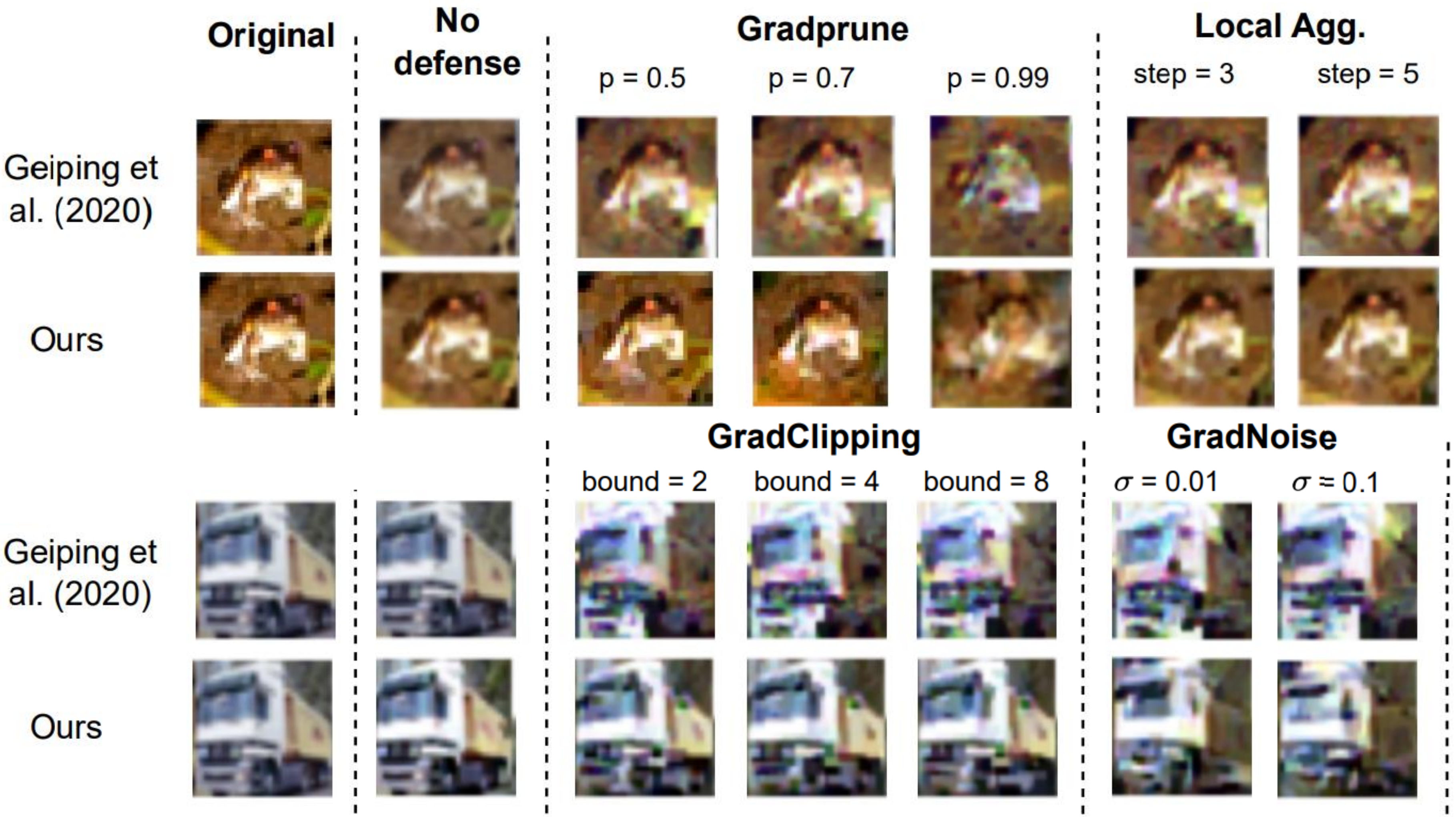
Beyond two-layer networks



- 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

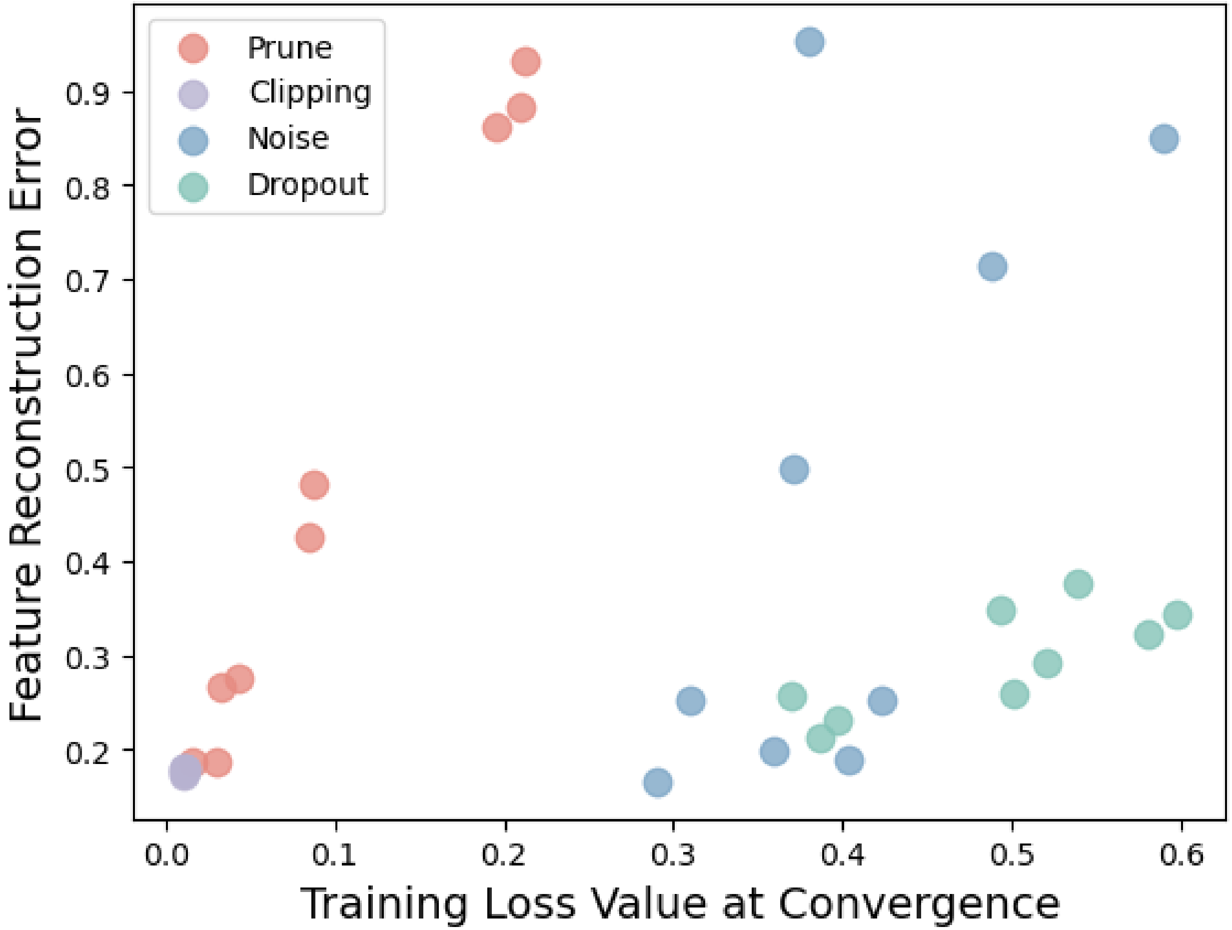
[Liu et al. 2024] https://arxiv.org/abs/2402.09478

Empirical results:



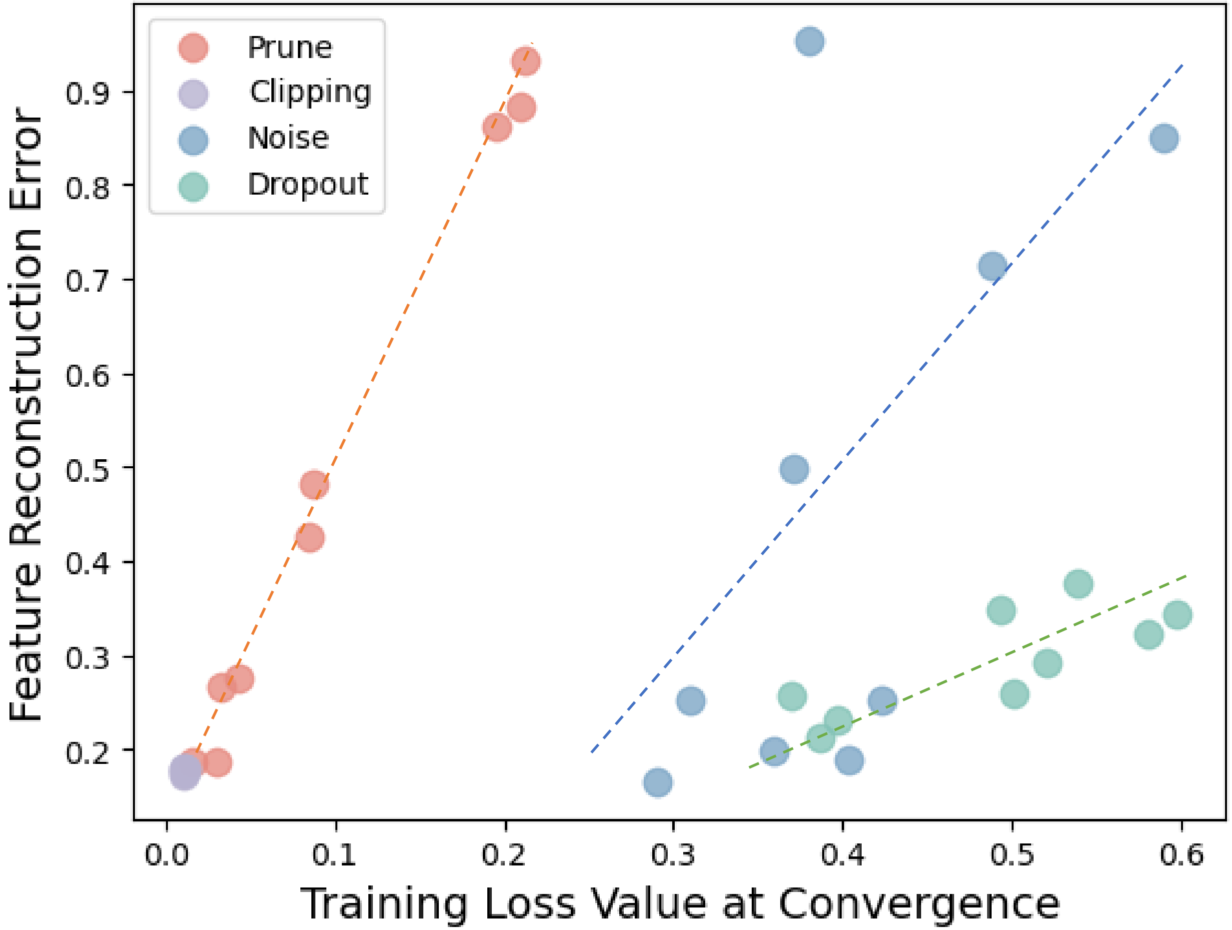
[Liu et al. 2024] https://arxiv.org/abs/2402.09478

Privacy-utility trade-offs



[Liu et al. 2024] https://arxiv.org/abs/2402.09478

Privacy-utility trade-offs



[Liu et al. 2024] https://arxiv.org/abs/2402.09478

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 et al. 2024] https://arxiv.org/abs/2312.05720

Discussions

- Gradient pruning is the most effective defending method (theoretically, empirically: better privacy-utility trade-off)

- Call for more tailored concept of privacy in reconstruction attack in federated learning

- Information-theoretical or computational lower bound in reconstruction attack

Thank you