Bounding Training Data Reconstruction in Private (Deep) Learning

Chuan Guo, Brian Karrer, Kamalika Chaudhuri, Laurens van der Maaten

FAIR

Meta AI

Motivation

Differential privacy has been the de facto standard for data privacy

$$\frac{P(\mathcal{A}(\mathbf{x} \cup \mathcal{D}) \in S)}{P(\mathcal{A}(\mathbf{x}' \cup \mathcal{D}) \in S)} \leq e^{\epsilon} \text{ for all } \mathbf{x}, \mathbf{x}', \mathcal{D} \text{ and } S \subseteq \mathcal{H}$$

What it says:

$$\underbrace{P(\mathcal{A}(\mathbf{x} \cup \mathcal{D}) \in S)}_{\text{Likelihood of observing model trained on } \mathbf{x} \cup \mathcal{D}} \approx \underbrace{P(\mathcal{A}(\mathbf{x}' \cup \mathcal{D}) \in S)}_{\text{Likelihood of observing model trained on } \mathbf{x}' \cup \mathcal{D}}$$

In general this difference can be measured using any statistical divergence

$$D(P(\mathcal{A}(\mathbf{x} \cup \mathcal{D}) \in S) || P(\mathcal{A}(\mathbf{x}' \cup \mathcal{D}) \in S))$$

Motivation

Semantic guarantee: An observer can't tell whether your data is $\mathbf{x}$ or $\mathbf{x}'$

- This is captured formally by a membership inference attack game
- If $\mathcal{A}$ is ϵ-DP then advantage $\leq (e^\epsilon - 1)/(e^\epsilon + 1)$ [Humphries et al., 2020]

Private Learner

Adversary

$$\mathcal{D} \in \mathcal{Z}^{n-1},\ \mathbf{z}_0, \mathbf{z}_1 \in \mathcal{Z}$$
$$b \sim \text{Bernoulli}(1/2)$$
$$\mathcal{D}_{\text{train}} \leftarrow \mathcal{D} \cup \{\mathbf{z}_b\}$$
$$h \leftarrow \mathcal{A}(\mathcal{D}_{\text{train}})$$

$$\xrightarrow{h, \mathcal{D}, \mathbf{z}_0, \mathbf{z}_1}$$

$$\hat{b} \leftarrow \texttt{Att}(h, \mathcal{D}, \mathbf{z}_0, \mathbf{z}_1)$$

$$\texttt{Adv} = \mathbb{P}(\hat{b} = 0 \mid b = 0) - \mathbb{P}(\hat{b} = 1 \mid b = 0)$$

Motivation

Membership privacy is not enough

- Membership status is not always sensitive

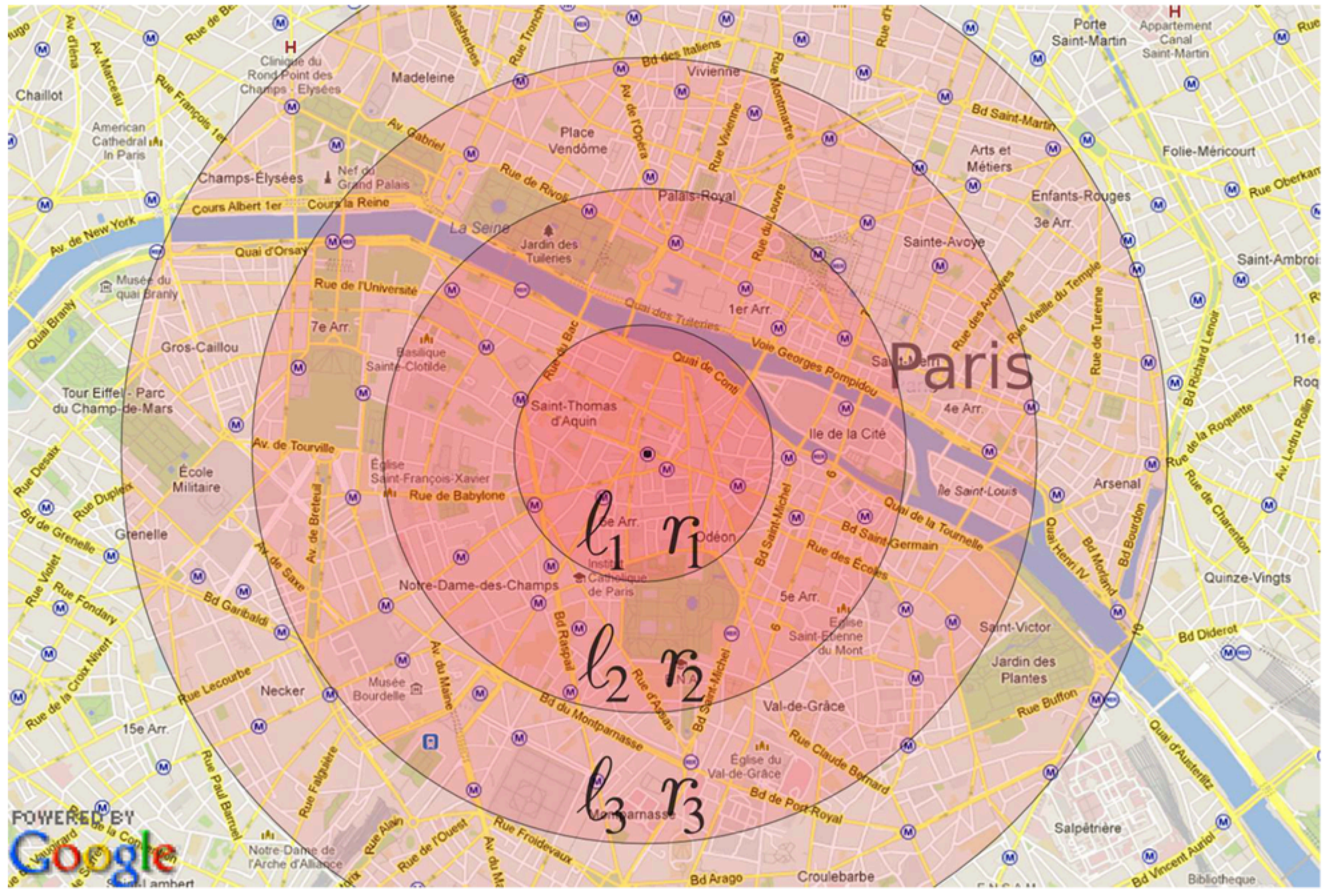


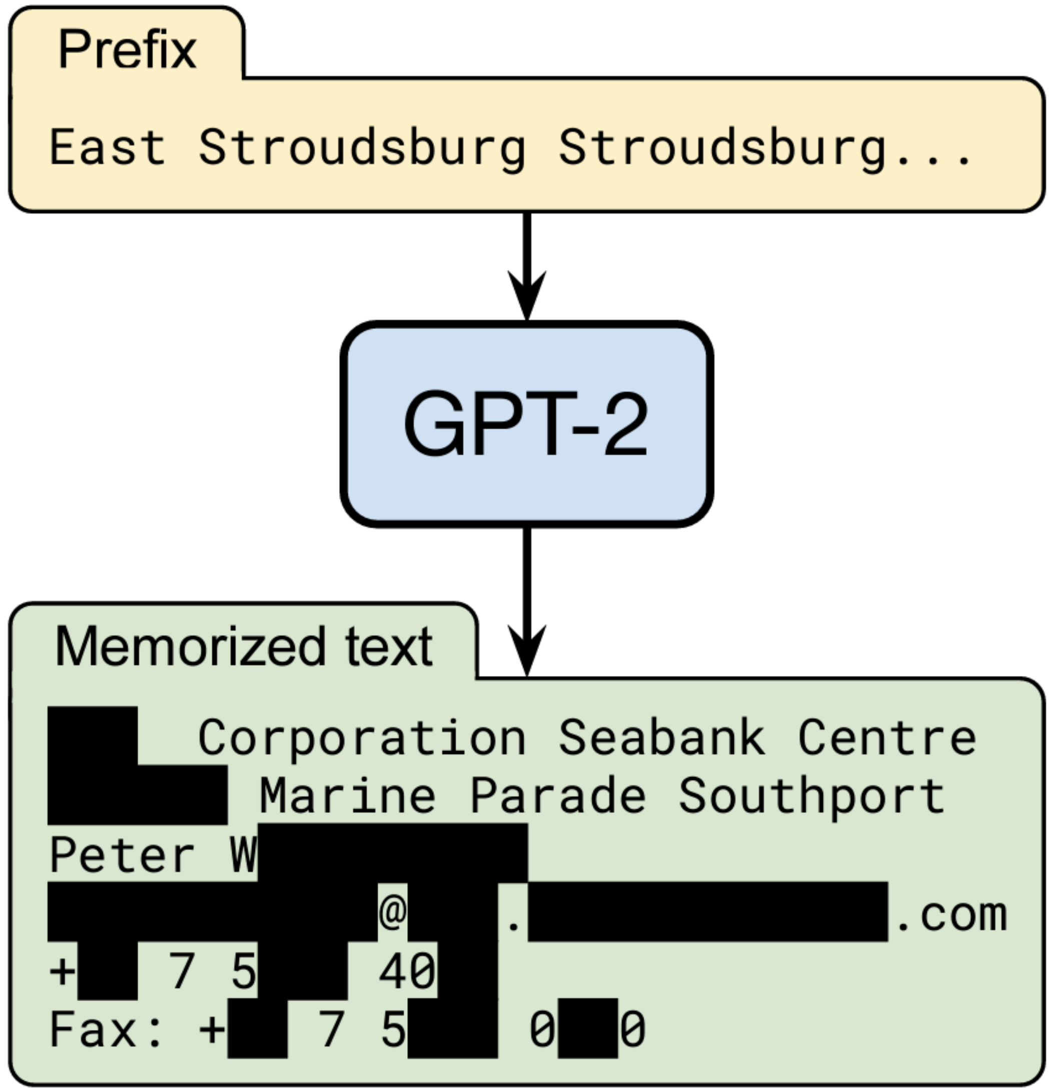
Figure 1: Geo-indistinguishability: privacy varying with r.

[Andrés et al., 2014]

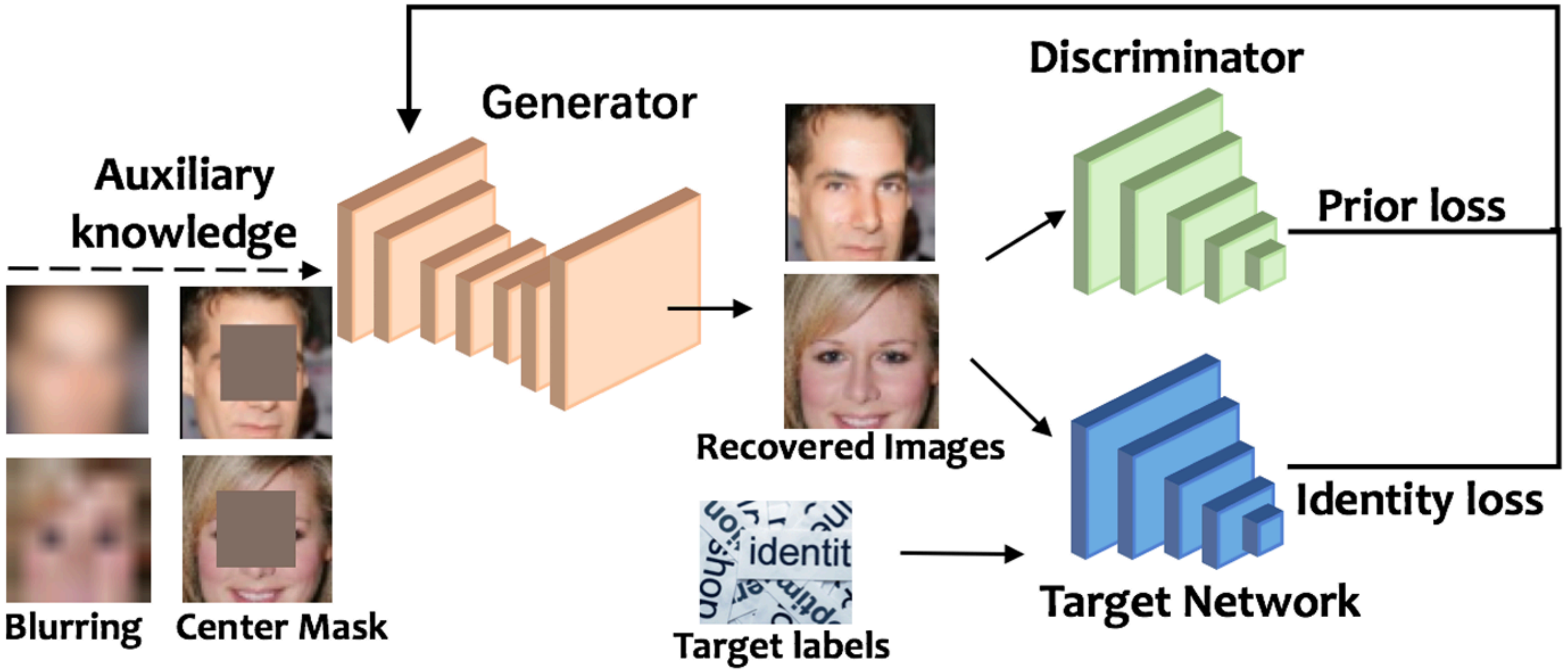
Motivation

Membership privacy is not enough

- Data reconstruction attack is much more concerning



[Carlini et al., 2020]



[Zhang et al., 2019]

Data Reconstruction Privacy

Can we give semantic guarantees in terms of data reconstruction privacy?

Data Reconstruction Privacy

Can we give semantic guarantees in terms of data reconstruction privacy?

Membership inference attack (MIA) game

- Goal: Test $\mathbf{z}_0$ vs. $\mathbf{z}_1$
- Success metric: Advantage
- Analogous to hypothesis testing

Private Learner

Adversary

$$\mathcal{D} \in \mathcal{Z}^{n-1},\ \mathbf{z}_0, \mathbf{z}_1 \in \mathcal{Z}$$

$$b \sim \text{Bernoulli}(1/2)$$

$$\mathcal{D}_{\text{train}} \leftarrow \mathcal{D} \cup \{\mathbf{z}_b\}$$

$$h \leftarrow \mathcal{A}(\mathcal{D}_{\text{train}})$$

$$\xrightarrow{h, \mathcal{D}, \mathbf{z}_0, \mathbf{z}_1}$$

$$\hat{b} \leftarrow \texttt{Att}(h, \mathcal{D}, \mathbf{z}_0, \mathbf{z}_1)$$

$$\texttt{Adv} = \mathbb{P}(\hat{b} = 0 \mid b = 0) - \mathbb{P}(\hat{b} = 1 \mid b = 0)$$

Data Reconstruction Privacy

Can we give semantic guarantees in terms of data reconstruction privacy?

Data reconstruction attack (DRA) game

- Goal: Infer $\mathbf{z}$
- Success metric: MSE
- Analogous to parameter estimation

Private Learner

$$\mathcal{D} \in \mathcal{Z}^{n-1}, \ \mathbf{z} \in \mathcal{Z}$$

$$\mathcal{D}_{\text{train}} \leftarrow \mathcal{D} \cup \{\mathbf{z}\}$$

$$h \leftarrow \mathcal{A}(\mathcal{D}_{\text{train}})$$

$$\xrightarrow{h, \mathcal{D}}$$

Adversary

$$\hat{\mathbf{z}} \leftarrow \texttt{Att}(h, \mathcal{D})$$

$$\texttt{MSE} = \mathbb{E}[\|\hat{\mathbf{z}} - \mathbf{z}\|_2^2 / d]$$

Data Reconstruction Privacy

Data reconstruction is “inevitable”

- For any $\mathbf{z}$, there exists a reconstruction attack $\mathcal{R}_{\mathbf{z}}$ that perfectly recovers $\mathbf{z}$

$$\mathcal{R}_{\mathbf{z}}(h, \mathcal{D}) = \mathbf{z}$$

A reasonable reconstruction attack should change with $\mathbf{z}$.

- We focus on unbiased estimators, i.e. $\mathbb{E}_{h \leftarrow \mathcal{A}(\mathcal{D}_{\text{train}})}[\texttt{Att}(h, \mathcal{D})] = \mathbf{z}$

Bound for Rényi DP

Lower bound for RDP using the Hammersley-Chapman-Robbins Bound (HCRB)

Theorem 1. *Let* $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^d$ *be a sample in the data space* $\mathcal{Z}$*, and let* $\texttt{Att}$ *be a reconstruction attack that outputs* $\hat{\mathbf{z}}(h)$ *upon observing the trained model* $h \leftarrow \mathcal{A}(\mathcal{D}_{train})$*, with expectation* $\mathbb{E}_{\mathcal{A}(\mathcal{D}_{train})}[\hat{\mathbf{z}}(h)] = \mathbf{z}$*. If* $\mathcal{A}$ *is a* $(2, \epsilon)$*-RDP learning algorithm then:*

$$\mathbb{E}[\|\hat{\mathbf{z}}(h) - \mathbf{z}\|_2^2/d] \geq \frac{\sum_{i=1}^d \mathrm{diam}_i(\mathcal{Z})^2/4d}{e^\epsilon - 1}.$$

where $\mathrm{diam}_i(\mathcal{Z}) = \sup_{\mathbf{z},\mathbf{z}' \in \mathcal{Z}: \mathbf{z}_j = \mathbf{z}'_j \forall j \neq i} |\mathbf{z}_i - \mathbf{z}'_i|$

Bound for Rényi DP

Lower bound for RDP using the Hammersley-Chapman-Robbins Bound (HCRB)

Theorem 1. *Let* $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^d$ *be a sample in the data space* $\mathcal{Z}$*, and let* $\texttt{Att}$ *be a reconstruction attack that outputs* $\hat{\mathbf{z}}(h)$ *upon observing the trained model* $h \leftarrow \mathcal{A}(\mathcal{D}_{train})$*, with expectation* $\mathbb{E}_{\mathcal{A}(\mathcal{D}_{train})}[\hat{\mathbf{z}}(h)] = \mathbf{z}$*. If* $\mathcal{A}$ *is a* $(2, \epsilon)$*-RDP learning algorithm then:*

$$\mathbb{E}[\|\hat{\mathbf{z}}(h) - \mathbf{z}\|_2^2/d] \geq \frac{\sum_{i=1}^{d} \mathrm{diam}_i(\mathcal{Z})^2/4d}{e^{\epsilon} - 1}.$$

where $\mathrm{diam}_i(\mathcal{Z}) = \sup_{\mathbf{z},\mathbf{z}' \in \mathcal{Z}:\mathbf{z}_j = \mathbf{z}'_j \forall j \neq i} |\mathbf{z}_i - \mathbf{z}'_i|$

Observation: Lower bound increases as $\epsilon \to 0$

- Low ϵ implies less information leakage from training data
- Without information about $\mathbf{z}$, the attacker's estimate has high variance

Bound for Fisher Information Loss

Fisher information loss (FIL) [Hannun et al., 2021] is a more natural framework for reconstruction privacy

- Privacy accounting using Fisher information matrix $\mathcal{I}_h(\mathbf{z}) = -\mathbb{E}[\nabla^2_{\mathbf{z}} \log p_{\mathcal{A}}(h; \mathbf{z})]$
- Measures the rate of change of $\mathcal{A}$ with respect to $\mathbf{z}$

Lower bound for FIL using the Cramér-Rao Bound (CRB)

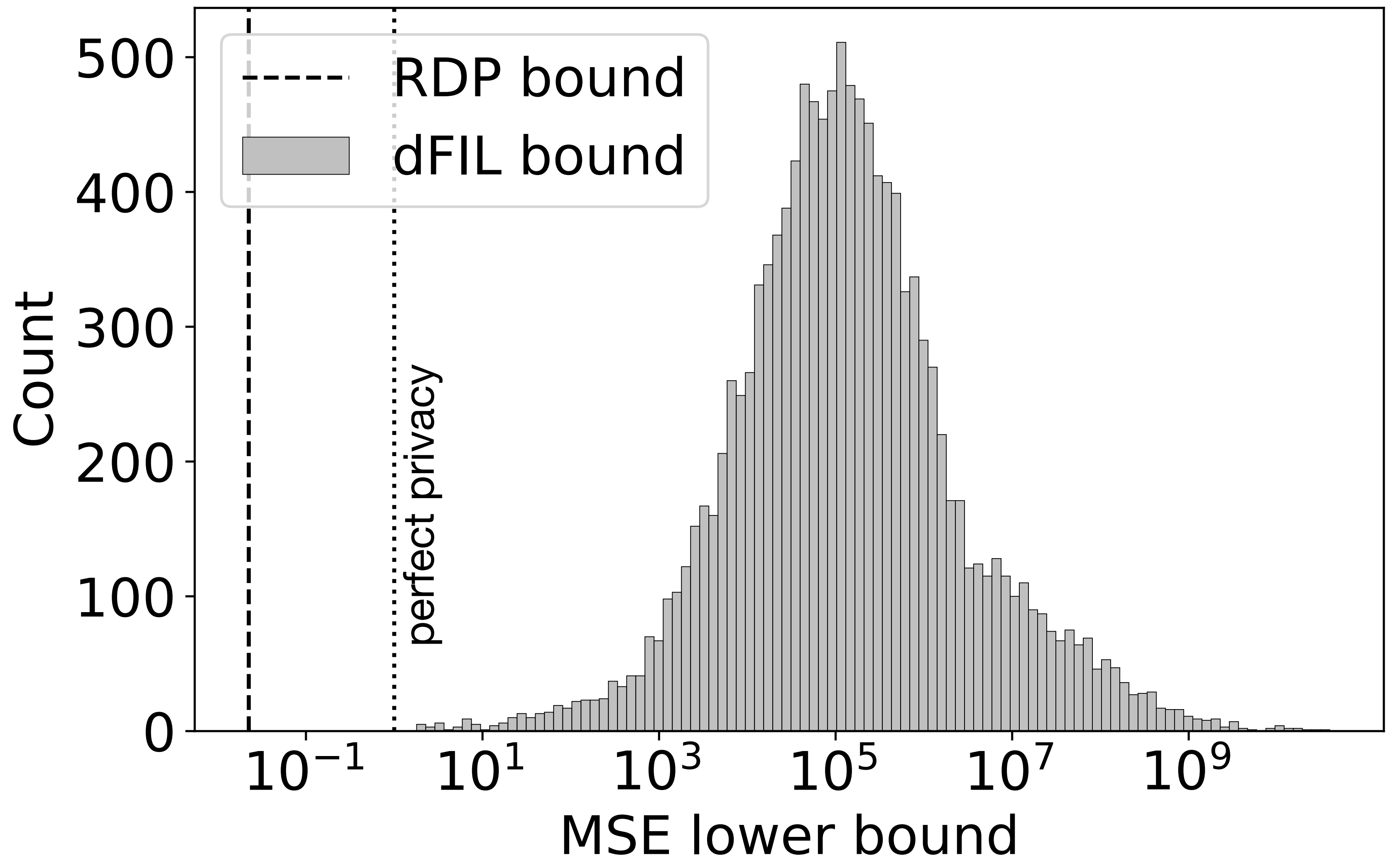
Theorem 2. *Assume the setup of Theorem 1, and additionally that the log density function* $\log p_{\mathcal{A}}(h|\zeta)$ *satisfies some wellness conditions. Then for any unbiased estimator* $\hat{\mathbf{z}}(h)$*:*

$$\mathbb{E}[\|\hat{\mathbf{z}}(h) - \mathbf{z}\|_2^2/d] \geq d/\mathrm{Tr}(\mathcal{I}_h(\mathbf{z})).$$

Experiment

Setup: MNIST 0 vs. 1 training using linear logistic regression

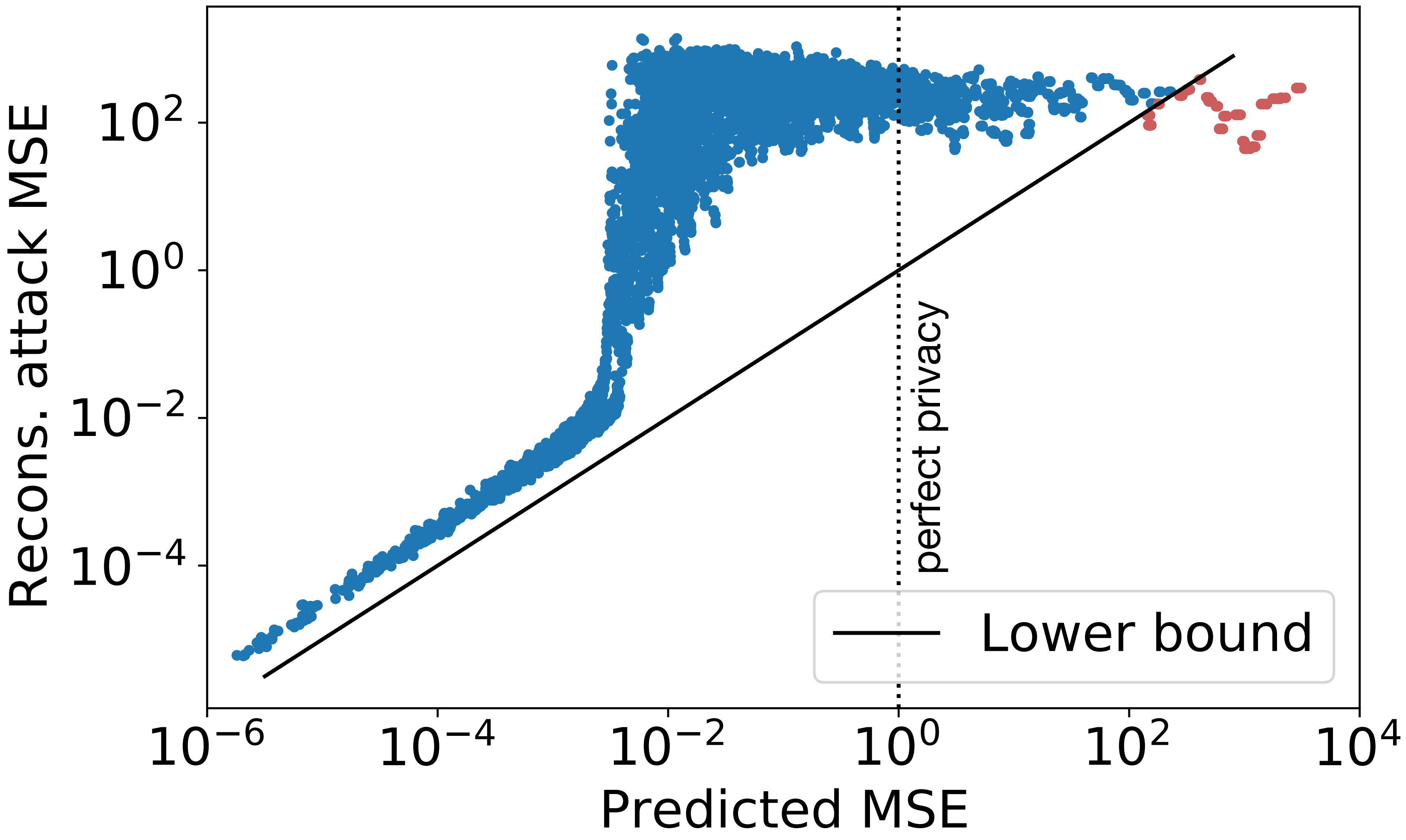
- Output perturbation satisfies $(2, 4/(n\lambda\sigma)^2)$-RDP and Theorem 1 applies
- dFIL denotes $\bar{\eta}^2 = \mathrm{Tr}(\mathcal{I}_h(\mathbf{z}))/d$ so Theorem 2 gives MSE $\geq 1/\bar{\eta}^2$



Experiment

Evaluation of reconstruction attack against GLMs [Balle et al., 2022]

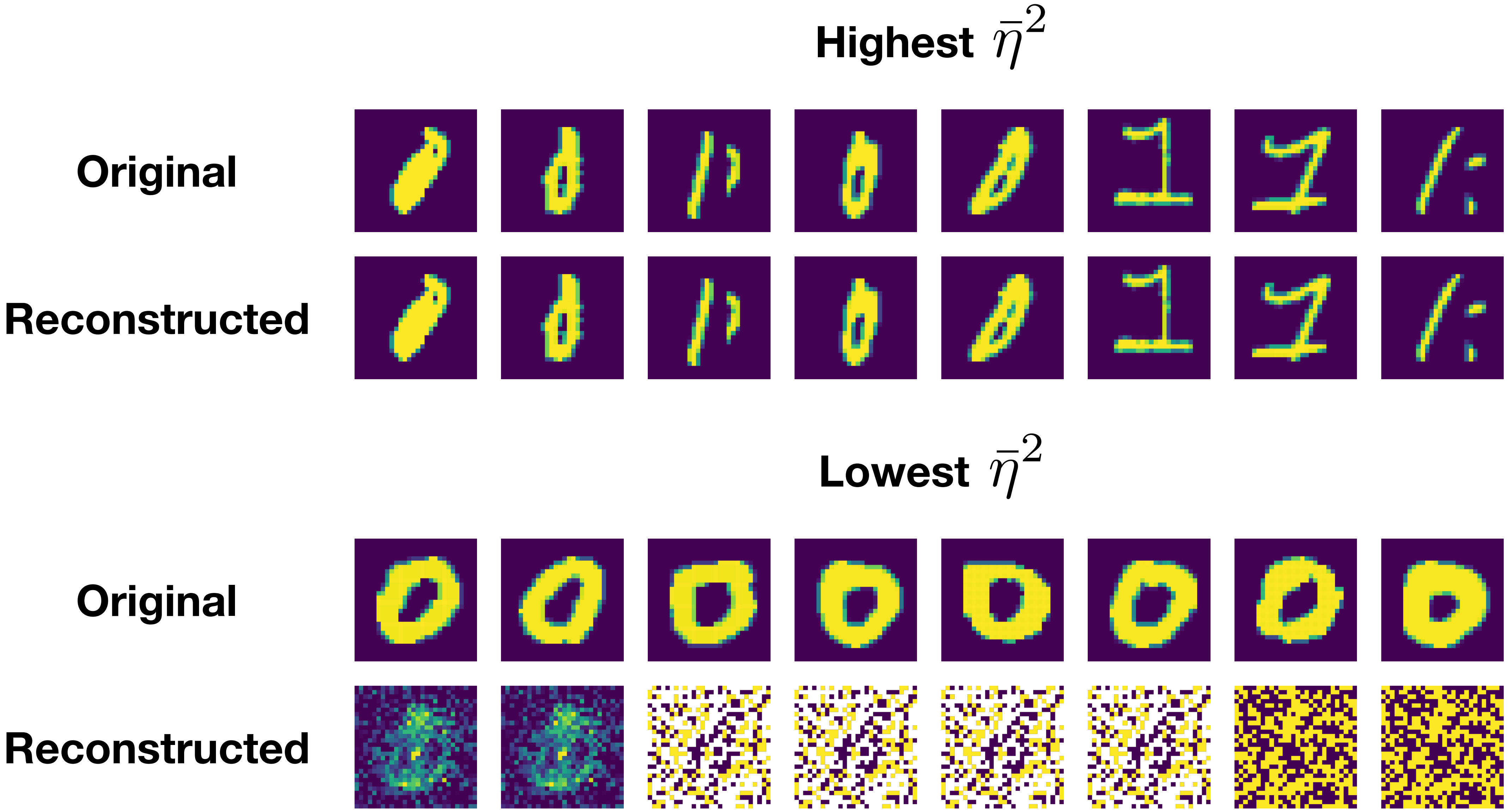
- FIL tightly captures per-sample reconstruction vulnerability



Experiment

Evaluation of reconstruction attack against GLMs [Balle et al., 2022]

- FIL tightly captures per-sample reconstruction vulnerability



Fisher Information For Private SGD

We can also compute Fisher information of gradient w.r.t. training data

- Private SGD [Abadi et al., 2016] adds Gaussian noise to clipped gradient

$$
\begin{aligned}
\mathbf{g}_t(\mathbf{z}) &\leftarrow \nabla_{\mathbf{w}} \ell(\mathbf{z}; \mathbf{w})|_{\mathbf{w}=\mathbf{w}_{t-1}} \quad \forall \mathbf{z} \in \mathcal{B}_t \\
\tilde{\mathbf{g}}_t(\mathbf{z}) &\leftarrow \mathbf{g}_t(\mathbf{z}) / \max(1, \|\mathbf{g}_t(\mathbf{z})\|_2 / C) \\
\bar{\mathbf{g}}_t &\leftarrow \frac{1}{|\mathcal{B}_t|} \left(\sum_{\mathbf{z} \in \mathcal{B}_t} \tilde{\mathbf{g}}_t(\mathbf{z}) + \mathcal{N}(\mathbf{0}, \sigma^2 C^2 \mathbf{I}) \right) \\
\mathbf{w}_t &\leftarrow \mathbf{w}_{t-1} - \rho \bar{\mathbf{g}}_t
\end{aligned}
$$

Fisher Information For Private SGD

We can also compute Fisher information of gradient w.r.t. training data

- Private SGD [Abadi et al., 2016] adds Gaussian noise to clipped gradient

$$
\begin{aligned}
\mathbf{g}_t(\mathbf{z}) &\leftarrow \nabla_{\mathbf{w}} \ell(\mathbf{z};\mathbf{w})\big|_{\mathbf{w}=\mathbf{w}_{t-1}} \quad \forall \mathbf{z} \in \mathcal{B}_t \\
\tilde{\mathbf{g}}_t(\mathbf{z}) &\leftarrow \mathbf{g}_t(\mathbf{z}) / \max(1, \|\mathbf{g}_t(\mathbf{z})\|_2 / C) \\
\bar{\mathbf{g}}_t &\leftarrow \frac{1}{|\mathcal{B}_t|}\left(\sum_{\mathbf{z}\in\mathcal{B}_t} \tilde{\mathbf{g}}_t(\mathbf{z}) + \mathcal{N}(\mathbf{0}, \sigma^2 C^2 \mathbf{I})\right) \\
\mathbf{w}_t &\leftarrow \mathbf{w}_{t-1} - \rho \bar{\mathbf{g}}_t
\end{aligned}
$$

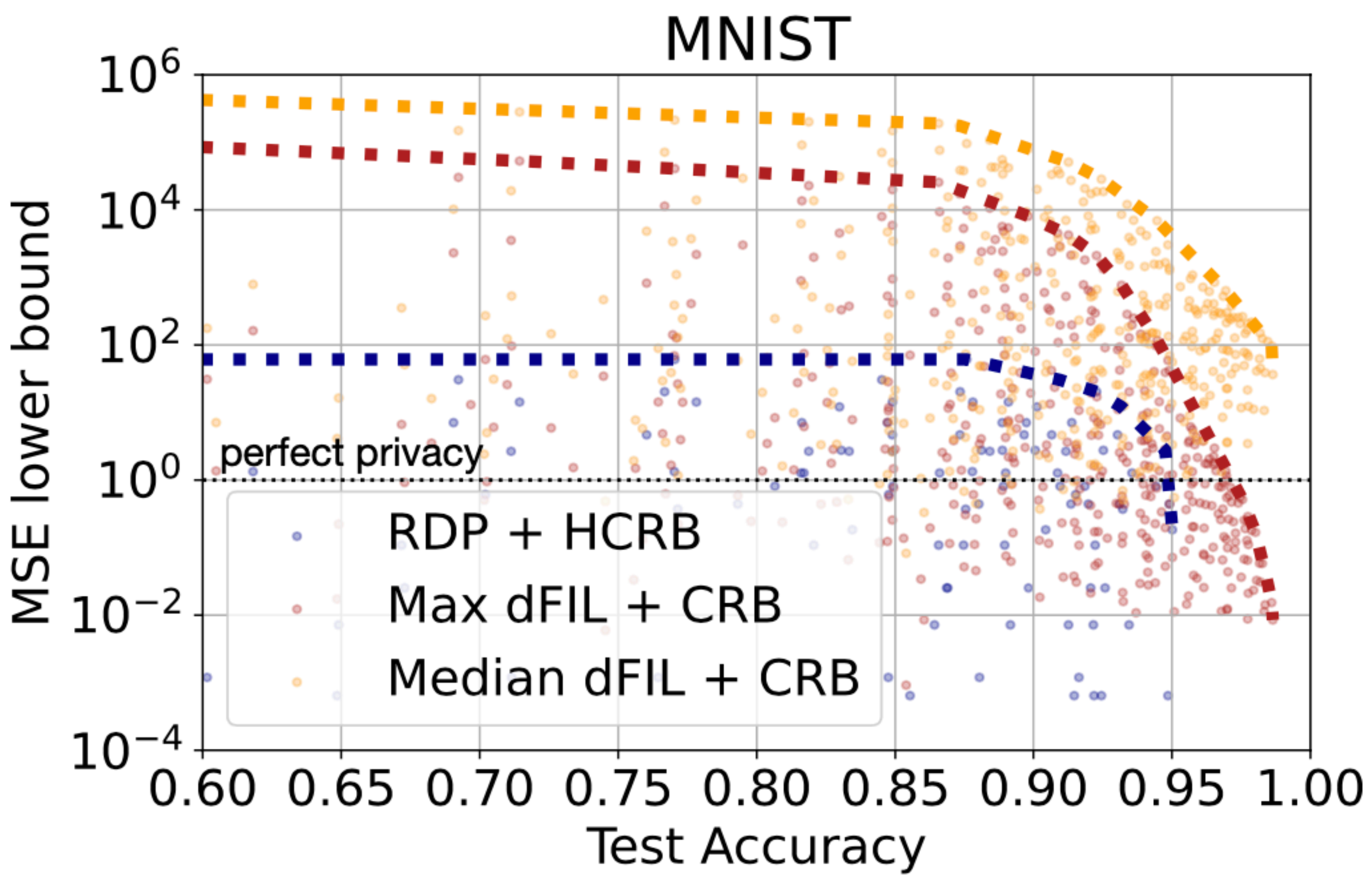
- Fisher information is a second-order derivative (easily computable using JAX!)

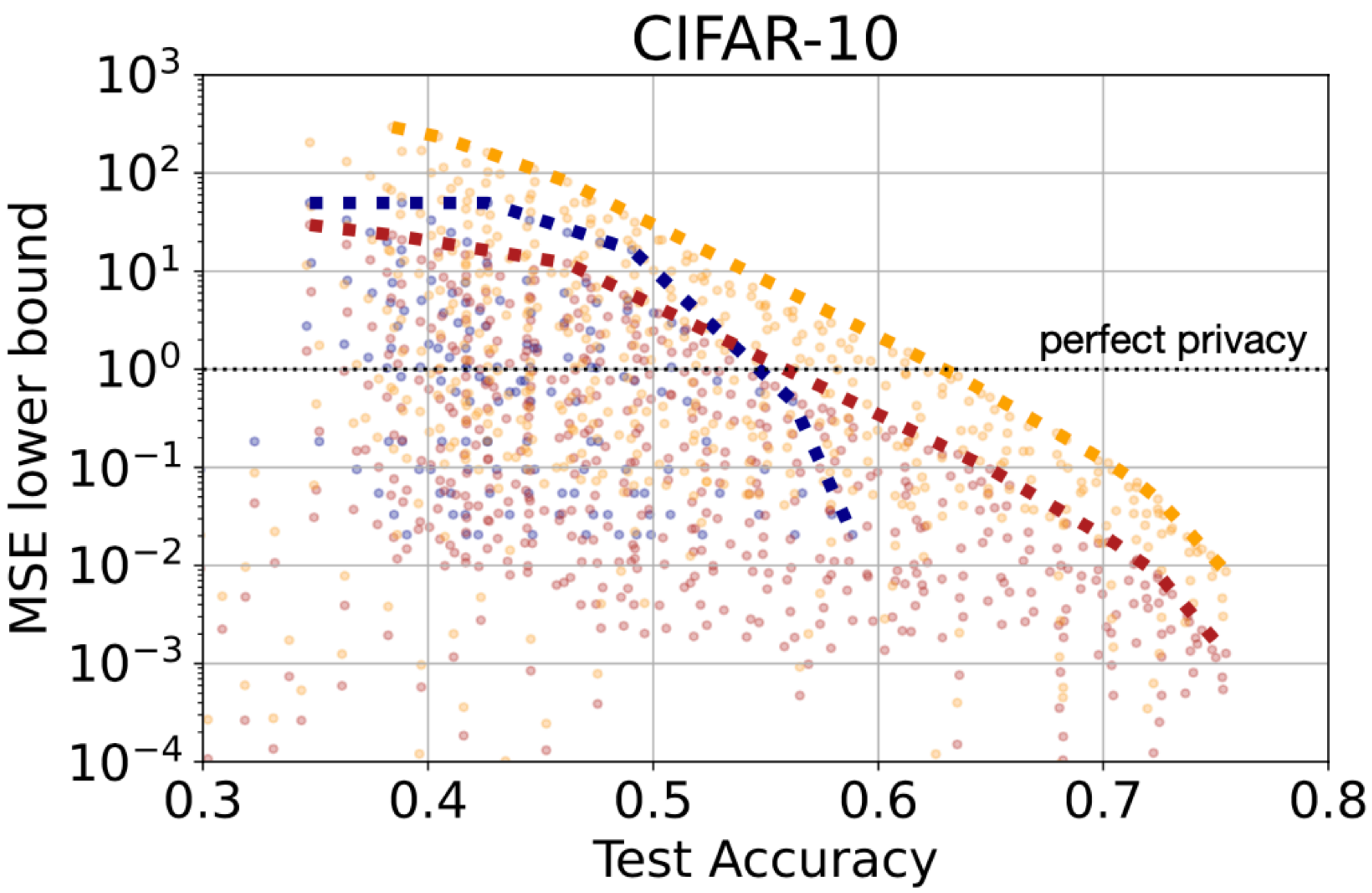
$$
\mathcal{I}_{\bar{\mathbf{g}}_t}(\mathbf{z}) = \frac{1}{\sigma^2} \nabla_\zeta \tilde{\mathbf{g}}_t(\zeta)^\top \nabla_\zeta \tilde{\mathbf{g}}_t(\zeta)\Big|_{\zeta=\mathbf{z}}
$$

Experiment

Setup: 10-class MNIST and CIFAR-10

- Convolutional networks with tanh activation [Papernot et al., 2020]





Conclusion

We derive semantic privacy guarantees against data reconstruction attacks

- By connecting DRA to parameter estimation in statistics
- Using both RDP and FIL accounting
- FIL closely captures per-sample vulnerability to DRA and yields better privacy-utility trade-off

Bounding Training Data Reconstruction in Private (Deep) Learning
Poster: Tuesday July 19, 6:30-8:30pm