Measuring Data Leakage in Machine-Learning Models with Fisher Information

Awni Hannun, Chuan Guo, and Laurens van der Maaten

Fisher Information Loss (FIL)

- $I_h(D)$ is the *Fisher information matrix* of model h for dataset D
- h has *Fisher information loss* of η with respect to D if:

$$\|I_h(D)\|_2 \leq \eta^2$$

- The largest singular value of $I_h(D)$ is bounded by η^2

Output Perturbation and FIL

- The *Gaussian mechanism* adds noise $b \sim \mathcal{N}(0, \sigma^2 I)$ to model w^*:

$$w_{\text{priv}} = w^* + b$$

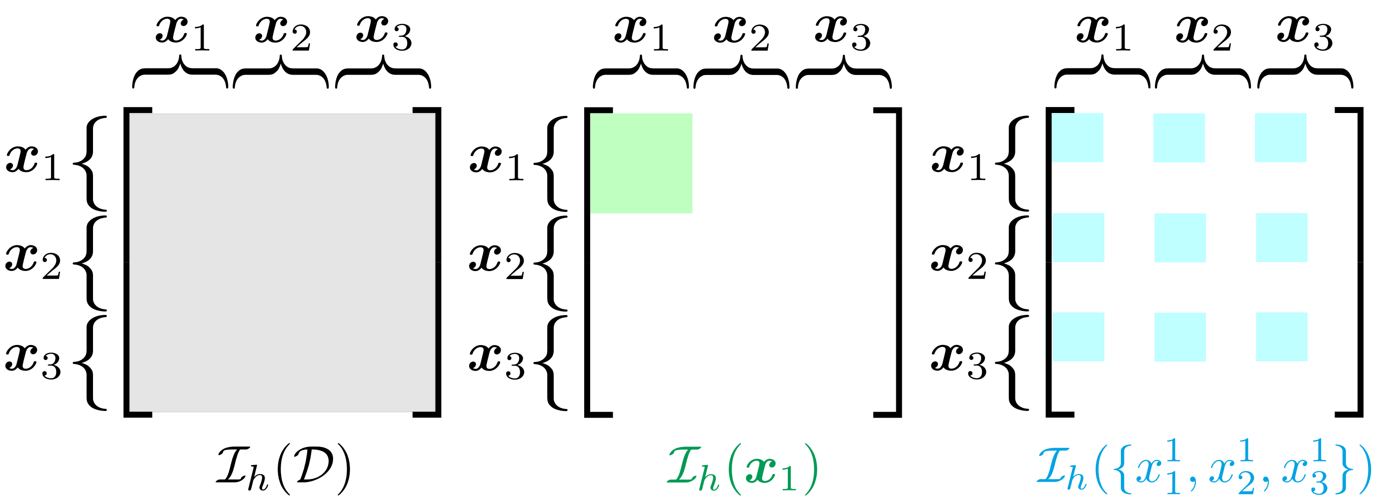
- The FIL of the Gaussian mechanism with standard deviation σ is:

$$\eta = \frac{1}{\sigma} \left\| J_f \right\|_2$$

- J_f is the Jacobian of the minimizer with respect to the data

Properties of FIL

- Compute FIL for different subsets of the training set
  - Individual attributes
  - Individual examples
  - Groups of examples
  - The full dataset


$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_3$
$\boldsymbol{x}_1$
$\boldsymbol{x}_2$
$\boldsymbol{x}_3$
$\mathcal{I}_h(\mathcal{D})$
$\mathcal{I}_h(\boldsymbol{x}_1)$
$\mathcal{I}_h(\{x_1^1, x_2^1, x_3^1\})$

(a) Full FIM

(b) Example-level FIM

(c) Attribute-level FIM

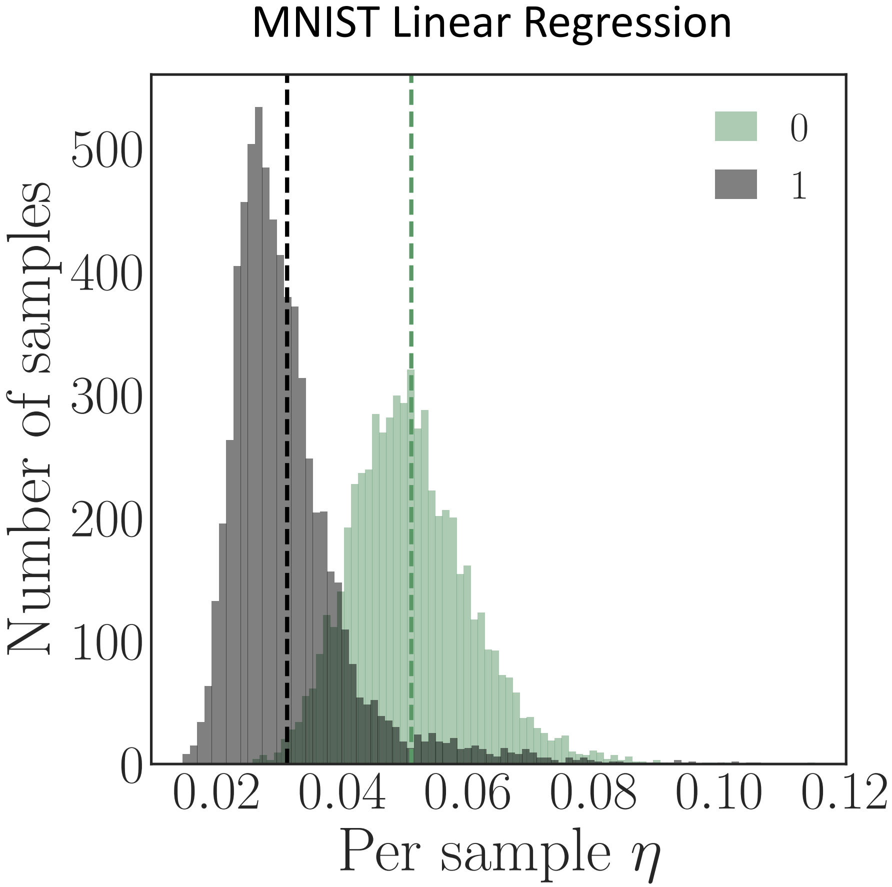
Properties of FIL

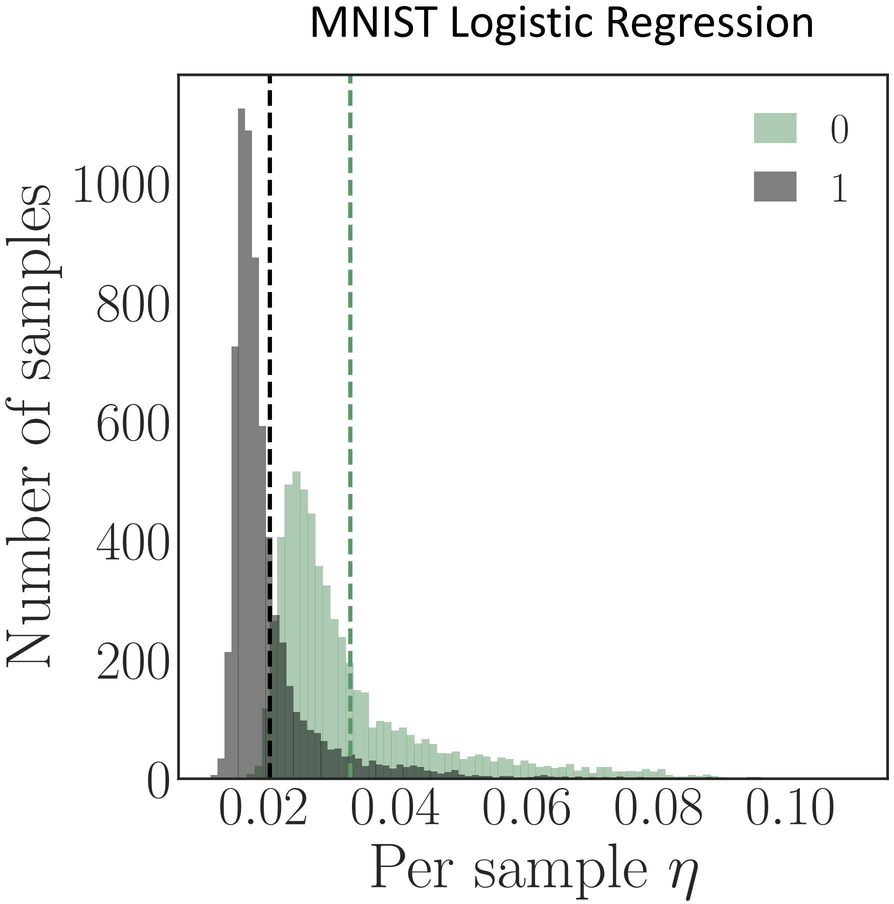
- By the Cramér-Rao bound, for any unbiased estimator $\hat{x}$ of x:

$$\mathrm{Var}(\hat{x}) \ \geq \frac{1}{\eta^2}$$

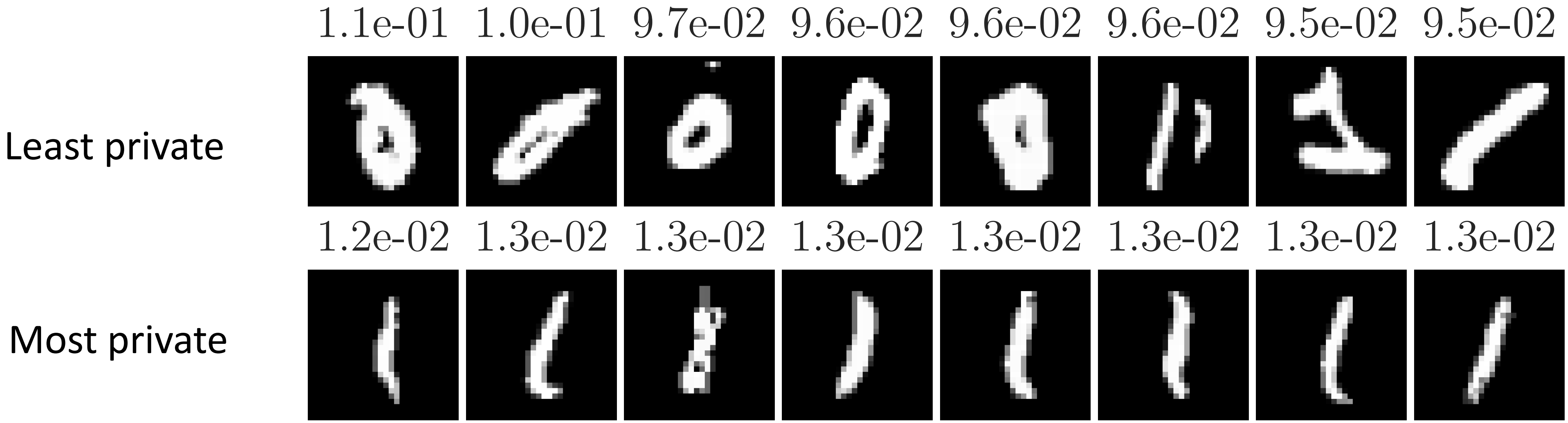
- FIL provides security even with intra-dataset correlations
- Composes additively and closed under post-processing

Fisher Information Loss: MNIST

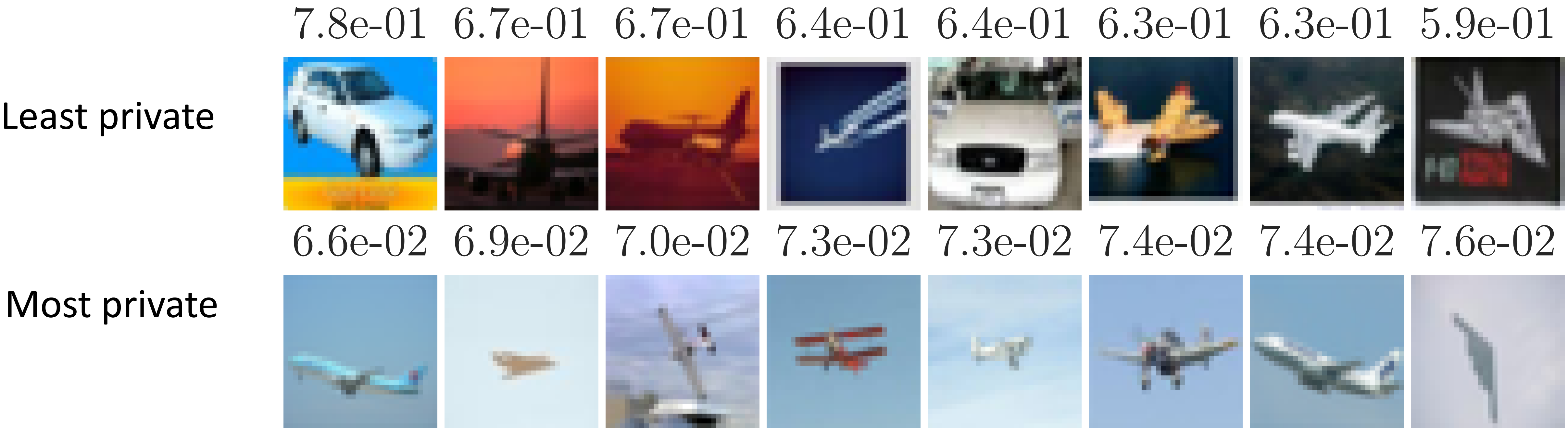




Fisher Information Loss: MNIST



Fisher Information Loss: CIFAR-10

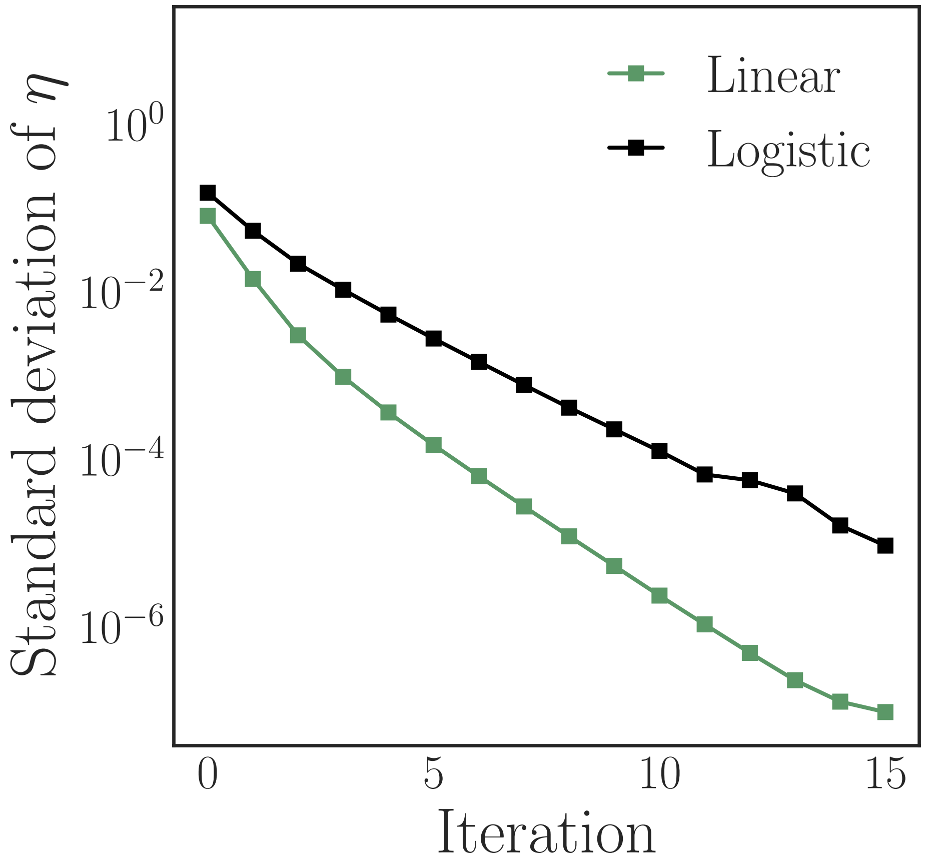


Iteratively Reweighted FIL

Goal: Equitably distribute privacy loss for individuals in the data

Algorithm: Iteratively Reweighted FIL (IRFIL)

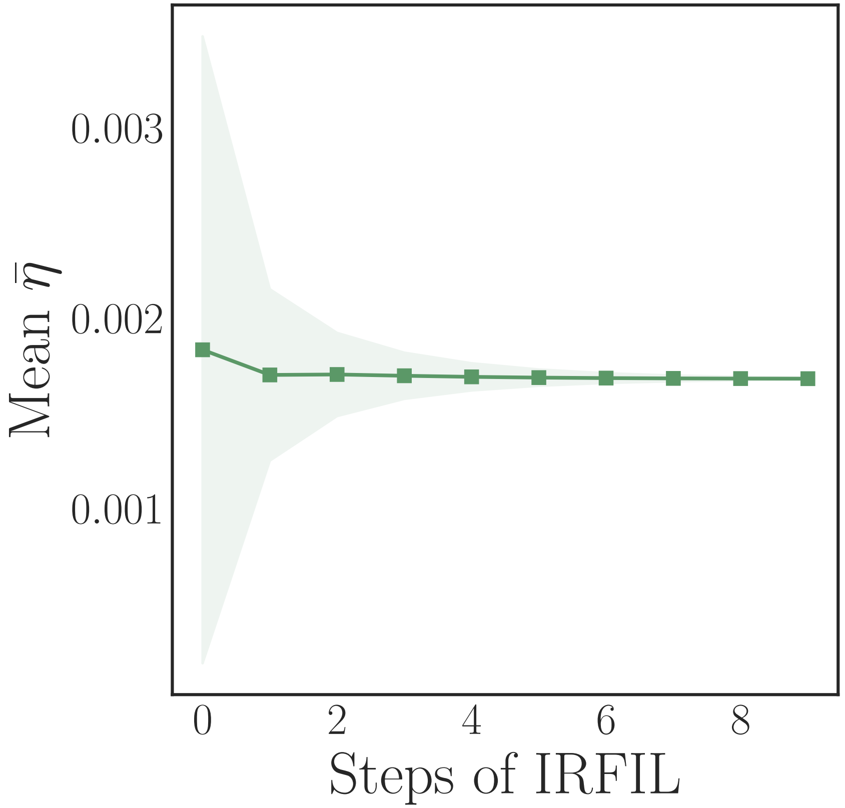
```
Iterate 1 . . . T
    1: Train model
    2: Compute example-level FIL ($\eta_i$)
    3: New loss with weights $\propto {}^{1}/_{\eta_i}$
```

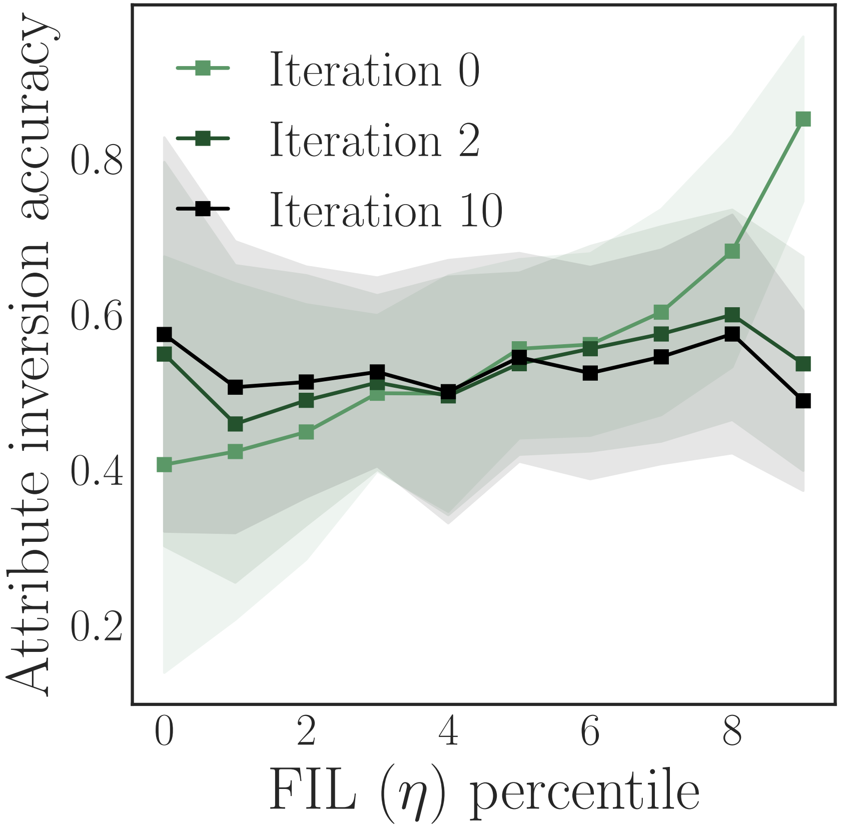

MNIST

Fairness Under Adversarial Attacks

IWPC dataset: classify patients by medical dosage

- Target feature is one of three possible alleles of a gene

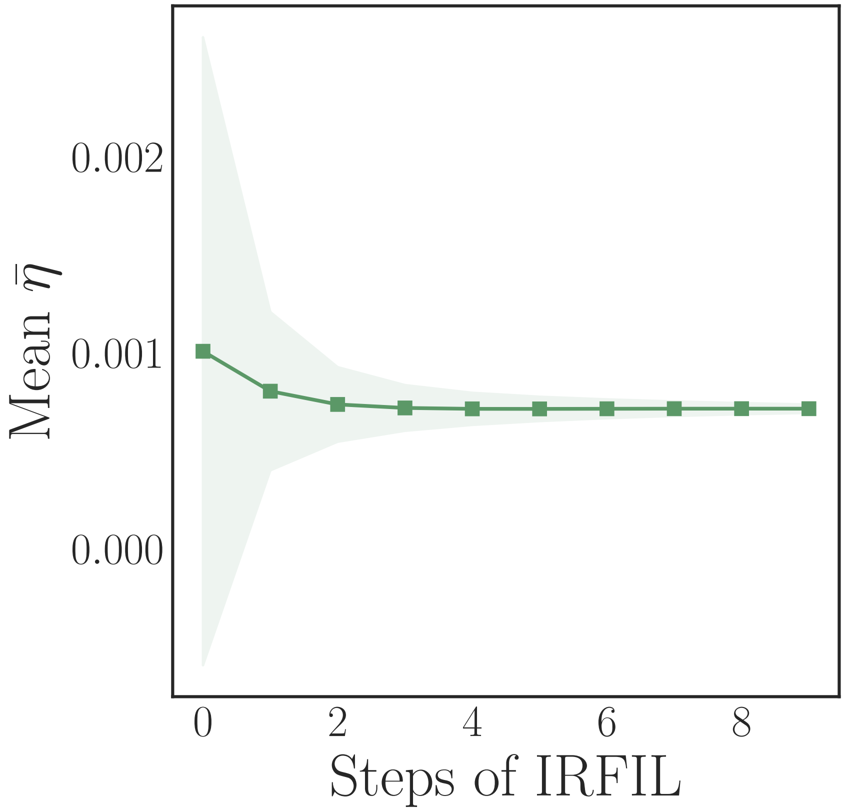


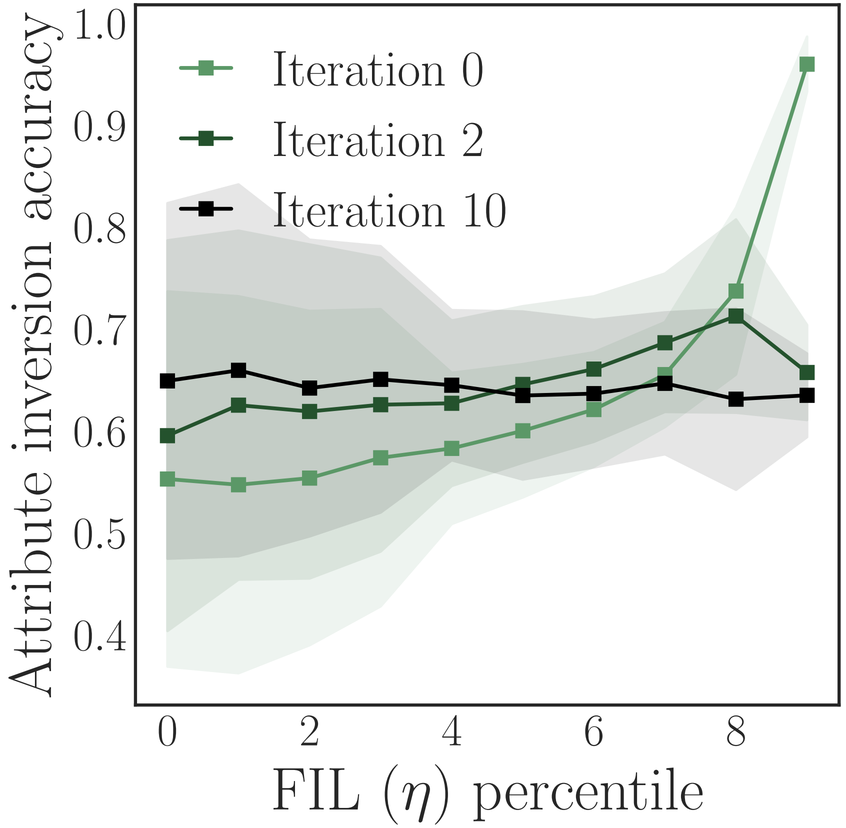


Fairness Under Adversarial Attacks

UCI adult dataset: classify individuals by salary given demographic features

- Target feature is marital status


Mean $\bar{\eta}$
Steps of IRFIL


Attribute inversion accuracy
Iteration 0
Iteration 2
Iteration 10
FIL (η) percentile