Measuring Data Leakage in Machine-Learning Models with Fisher Information

Awni Hannun[1]   **Chuan Guo**[1]   **Laurens van der Maaten**[1]

[1] Facebook AI Research

Abstract

Machine-learning models contain information about the data they were trained on. This information leaks either through the model itself or through predictions made by the model. Consequently, when the training data contains sensitive attributes, assessing the amount of information leakage is paramount. We propose a method to quantify this leakage using the Fisher information of the model about the data. Unlike the worst-case *a priori* guarantees of differential privacy, *Fisher information loss* measures leakage with respect to specific examples, attributes, or sub-populations within the dataset. We motivate Fisher information loss through the Cramér-Rao bound and delineate the implied threat model. We provide efficient methods to compute Fisher information loss for output-perturbed generalized linear models. Finally, we empirically validate Fisher information loss as a useful measure of information leakage.

1 INTRODUCTION

Machine-learning models trained on sensitive data are often made public. Even when the models are not explicitly released, they may be implicitly leaked from their predictions [Papernot et al., 2017, Tramèr et al., 2016]. Undeniably, these models contain information about the data they were trained on. Without mitigating measures, training set membership can often be inferred [Shokri et al., 2017, Yeom et al., 2018], and, in some cases, sensitive attributes or even whole examples can be extracted [Carlini et al., 2019, 2020, Fredrikson et al., 2014, 2015].

Assessing the information leaked from models about their training data is commonly done with techniques such as differential privacy [Dwork et al., 2006]. Differential privacy successfully avoids the "just a few" failure mode of more heuristic privacy assessments, in which privacy is protected for many but not all individuals [Dwork and Roth, 2014]. This comes at the cost of a worst-case assessment, leading to large differences in the vulnerability of individuals to privacy attacks. Also, a mismatch exists between the protection of differential privacy, which is relative to participation in a dataset, and privacy attacks, which can take advantage of the absolute information leaked from a trained model [Carlini et al., 2019, Long et al., 2018]. Furthermore, differential privacy implicitly degrades when correlations exist in the dataset [Ghosh and Kleinberg, 2016, Humphries et al., 2020, Kasiviswanathan and Smith, 2008, Liu et al., 2016].

We propose an example-specific and correlation-aware measure of data leakage using Fisher information. The quantity, which we term *Fisher information loss*, can assess the leakage of a model about various subsets of the full dataset. This includes, for example, assessments at the granularity of individual attributes, individual examples, groups of examples, or the full training set. We show, via the Cramér-Rao bound, that under specific assumptions the ability of an adversary to estimate the underlying data from a model with bounded Fisher information loss is limited. We also demonstrate that, unlike differential privacy, Fisher information loss does not implicitly degrade when data is correlated.

We derive tractable computations of Fisher information loss in the case of Gaussian noise perturbation with generalized linear models. We empirically validate Fisher information loss with experiments on four datasets. We further demonstrate that Fisher information loss accurately captures examples susceptible to attribute inversion attacks.

The ability of Fisher information loss to measure per-example privacy loss means it can be used as the basis for algorithms that aim to achieve fairness in privacy [Cummings et al., 2019, Ekstrand et al., 2018]. We demonstrate this by developing an algorithm that balances Fisher information loss for individuals in the training set, thereby resolving the problem that subgroups may have "disparate vulnerability" to privacy attacks [Yaghini et al., 2019].

2 RELATED WORK

This work builds on and complements a significant body of prior work in assessing the privacy of a model with respect to the data it was trained on. This includes approaches which obfuscate the original data such as k-anonymity [Samarati and Sweeney, 1998] and l-diversity [Machanavajjhala et al., 2007], information theoretic criteria [Agrawal and Aggarwal, 2001], and, perhaps the most commonly studied, differential privacy [Dwork et al., 2006] and its more recent variations [Dong et al., 2019, Mironov, 2017]. While these methods often provide rigorous privacy guarantees, most of them do not explicitly bound the *inferential power* of an adversary, and correlations in the training dataset can be exploited by an adversary to mount a successful inference attack [Machanavajjhala et al., 2007, Li et al., 2007, Ghosh and Kleinberg, 2016, Liu et al., 2016]. Moreover, these definitions do not offer simple and flexible approaches to measure the information leaked by a model about varying subsets of the training data. This can lead to unequal vulnerability to privacy attacks despite attempts to protect privacy [Yaghini et al., 2019].

Because these privacy assessment techniques do not identify vulnerability at the level of the individual or sub-population, prior work exists to quantify the susceptibility of such subgroups to privacy attacks. Farokhi and Kaafar [2020] propose an information theoretic measure of the vulnerability of individual examples to membership inference attacks. Less rigorous heuristics have also been studied [Long et al., 2017, 2018]. Carlini et al. [2019] propose a heuristic which can infer the susceptibility of data to model inversion attacks.

This work also builds upon prior studies of Fisher information as a measure of privacy. An early study analyzed loss of Fisher information in a general randomized response framework [Anderson, 1977]. More recently, the relationship of privacy with Fisher information to estimator error through the use of the Cramér-Rao bound was investigated [Farokhi and Sandberg, 2017]. In particular, Farokhi and Sandberg [2017] turn to Fisher information as a practical alternative to differential privacy in protecting data collected from smart power meters. We build from these works in several directions, including a broader application to generalized linear models as well as the use of Fisher information loss in an algorithm which can provide fairness in data leakage.

3 FISHER INFORMATION LOSS

Let $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ be a training dataset of n examples with $\boldsymbol{x}_i \in \mathbb{R}^d$ and scalar target y_i. We denote by $\mathcal{A}(\mathcal{D})$ a randomized learning algorithm which outputs a hypothesis h from a predefined hypothesis space $\mathcal{H}$. Treating the hypothesis as a random variable, $p_{\mathcal{A}}(h \mid \mathcal{D})$ is the probability density of h given $\mathcal{D}$ for the randomized algorithm $\mathcal{A}(\mathcal{D})$. We denote by $\mathcal{I}_h(\mathcal{D}) \in \mathbb{R}^{n(d+1) \times n(d+1)}$ the Fisher information matrix (FIM) defined by

$$\mathcal{I}_h(\mathcal{D}) = -\mathbb{E}_h \left[\nabla_{\mathcal{D}}^2 \log p_{\mathcal{A}}(h \mid \mathcal{D}) \right] \quad (1)$$

where $\nabla_{\mathcal{D}}^2$ yields the matrix of second derivatives of $\log p_{\mathcal{A}}(h \mid \mathcal{D})$ with respect to the values in $\mathcal{D}$, and the expectation is taken over the randomness in $\mathcal{A}(\mathcal{D})$. The FIM, $\mathcal{I}_h(\mathcal{D})$, is a measure of the information that the hypothesis h contains about the training data $\mathcal{D}$. Hence, we use the FIM to measure the information loss from releasing the output h of a single evaluation of $\mathcal{A}(\mathcal{D})$.

Definition (Fisher information loss). *We say that $h \sim \mathcal{A}(\mathcal{D})$ has Fisher information loss (FIL) of η with respect to $\mathcal{D}$ if*

$$\|\mathcal{I}_h(\mathcal{D})\|_2 \leq \eta^2, \quad (2)$$

where $\|\mathcal{I}_h(\mathcal{D})\|_2$ denotes the 2-norm, or largest singular value, of the FIM. A smaller η means h contains less Fisher information about the training data, $\mathcal{D}$.

Motivation. Fisher information is a classical tool used in statistics for lower bounding the variance of an estimator [Lehmann and Casella, 2006]. We utilize this property to demonstrate that a small FIL implies a large variance for any unbiased estimate of the data.

Let $\boldsymbol{z} = [\boldsymbol{x}_1^\top, y_1, \ldots, \boldsymbol{x}_n^\top, y_n] \in \mathbb{R}^{n(d+1)}$ be the vector formed by concatenating the examples in $\mathcal{D}$, and $z \in \boldsymbol{z}$ an arbitrary element. If the FIL of h with respect to $\mathcal{D}$ is bounded by η, then for any unbiased estimator $\hat{z}$ of z, we have:

$$\mathrm{Var}(\hat{z}) \geq \frac{1}{\eta^2}. \quad (3)$$

Hence, a smaller FIL implies a larger variance in any unbiased attempt to infer z. Equation 3 follows directly from the Cramér-Rao bound. Indeed, under a fairly relaxed regularity condition [Kay, 1993], for any unbiased estimator $\hat{\boldsymbol{z}}$ of the data $\boldsymbol{z}$, the Cramér-Rao bound states that:

$$\mathbb{E}\left[(\hat{\boldsymbol{z}} - \boldsymbol{z})(\hat{\boldsymbol{z}} - \boldsymbol{z})^\top\right] \succeq \mathcal{I}_h(\boldsymbol{z})^{-1}, \quad (4)$$

where $A \succeq B$ if the matrix $A - B$ is positive semidefinite. Since the estimator is unbiased, this implies the covariance of $\hat{\boldsymbol{z}}$ is similarly bounded:

$$\mathrm{Cov}(\hat{\boldsymbol{z}}) \succeq \mathcal{I}_h(\boldsymbol{z})^{-1}. \quad (5)$$

If $\mathcal{A}(\mathcal{D})$ has an FIL of η then equation 2 implies $\mathcal{I}_h(\boldsymbol{z})^{-1}_{ii} \geq 1/\eta^2$ for all i, and from equation 5, $\mathrm{Var}(\hat{z}) \geq 1/\eta^2$ follows.

Correlated data. Fisher information loss also provides some security in the presence of intra-dataset correlations. If a model has an FIL bounded by η with respect to the training data (in vector form $\boldsymbol{z}$), then the covariance matrix for any unbiased estimator $\hat{\boldsymbol{z}}$ of that data is bounded by $\|\mathrm{Cov}(\hat{\boldsymbol{z}})\|_2 \geq 1/\eta^2$. This limits the ability of an adversary using an unbiased estimator to infer relative differences between elements in the dataset.
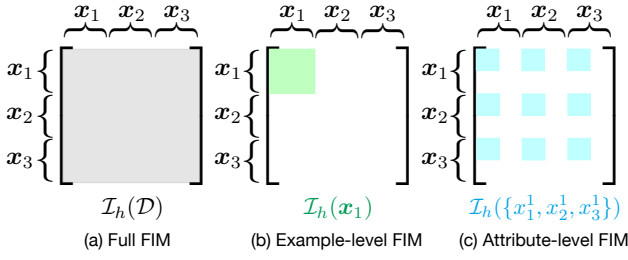
Figure 1: The FIM for different subsets of the training data.

3.1 PROPERTIES OF FIL

Subsets. In many cases we are interested in measuring FIL with respect to a single example $(\boldsymbol{x}_i, y_i) \in \mathcal{D}$. In this case, we can compute the example-specific FIM by selecting the corresponding entries of the full FIM. Given the FIM for the vector represented data $\boldsymbol{z}$, computing the FIM for the i-th example amounts to selecting the submatrix of size $(d+1) \times (d+1)$ with upper left corner $\mathcal{I}_h(\boldsymbol{z})_{i(d+1),i(d+1)}$. Similarly, computing the FIM for a specific attribute over all examples, $\{x_i^j\}_{i=1}^n$, amounts to constructing the submatrix by selecting the corresponding entries from the full FIM. See figure 1 for an illustration. In general, we can compute the FIM for any subset of elements of $\mathcal{D}$ by selecting their corresponding entries from the full FIM.

Composition. The Fisher information, and hence FIL, compose additively. Given k independent evaluations of $\mathcal{A}(\mathcal{D})$ each with an FIL of η, the combined FIL is at most $k\eta$. More generally, given an evaluation of k unique but independent randomized algorithms $\{h_i \sim \mathcal{A}_i(\mathcal{D}) \mid i = 1, \ldots, k\}$ each with an FIM of $\mathcal{I}_{h_i}(\mathcal{D})$, the FIM with respect to all of the h_i's is given by $\mathcal{I}_{h_1,\ldots,h_k}(\mathcal{D}) = \sum_{i=1}^k \mathcal{I}_{h_i}(\mathcal{D})$ [Lehmann and Casella, 2006]. Hence, if $\|\mathcal{I}_{h_i}(\mathcal{D})\|_2 \leq \eta_i^2$, then the combined FIL for all h_i about $\mathcal{D}$ is at most $(\sum_{i=1}^k \eta_i^2)^{1/2}$. This follows from the triangle inequality applied to the matrix 2-norm.

Closed under post-processing. As in differential privacy, FIL is closed under post-processing. If $\|\mathcal{I}_h(\mathcal{D})\|_2 \leq \eta^2$, then for any function of the hypothesis, $g(h)$, $\|\mathcal{I}_{g(h)}(\mathcal{D})\|_2 \leq \eta^2$. This follows since the matrix $\mathcal{I}_h(\mathcal{D}) - \mathcal{I}_{g(h)}(\mathcal{D})$ is positive semidefinite for any function $g(h)$ [Schervish, 2012].

Threat model. The bounds on an adversary's ability to estimate the data apply under several assumptions about the problem setting and the adversary:

- The required regularity condition is satisfied so that the FIM exists and the Cramér-Rao bound applies [Kay, 1993]. Namely, the expected value of the score function of the density $p_{\mathcal{A}}(h \mid \mathcal{D})$ should be 0 for all $\mathcal{D}$:

$$\mathbb{E}_h \left[\frac{\partial \log p_{\mathcal{A}}(h \mid \mathcal{D})}{\partial \mathcal{D}} \right] = 0, \qquad (6)$$

which essentially states that the derivative and integral commute:

$$\mathbb{E}_h \left[\frac{\partial \log p_{\mathcal{A}}(h \mid \mathcal{D})}{\partial \mathcal{D}} \right] = \int_h \frac{p_{\mathcal{A}}(h \mid \mathcal{D})}{p_{\mathcal{A}}(h \mid \mathcal{D})} \frac{\partial p_{\mathcal{A}}(h \mid \mathcal{D})}{\partial \mathcal{D}} dh$$
$$= \frac{\partial}{\partial \mathcal{D}} \int_h p_{\mathcal{A}}(h \mid \mathcal{D}) dh = 0.$$

This condition is satisfied for the models and privacy mechanisms we consider.

- We assume the adversary is limited to unbiased estimators of the unknown data. While achieving lower variance estimators is possible, this means the estimator will have to incur bias. Denoting by $\psi(\boldsymbol{z}) = \mathbb{E}[\hat{\boldsymbol{z}}]$ the expected value of the estimator as a function of the unknown data, the Cramér-Rao bound for biased estimators generalizes to:

$$\mathrm{Cov}(\hat{\boldsymbol{z}}) \succeq \boldsymbol{J}_\psi \mathcal{I}_h(\boldsymbol{z})^{-1} \boldsymbol{J}_\psi^\top, \qquad (7)$$

where $\boldsymbol{J}_\psi$ is the Jacobian of ψ with respect to $\boldsymbol{z}$. This gives a bound on the variance:

$$\mathrm{Var}(\hat{z}_i) \geq \boldsymbol{J}_{\psi,i}^\top \mathcal{I}_h(\boldsymbol{z})^{-1} \boldsymbol{J}_{\psi,i}, \qquad (8)$$

where $\boldsymbol{J}_{\psi,i}$ is the i-th row of $\boldsymbol{J}_\psi$ [Lehmann and Casella, 2006]. Because mean squared error decomposes into a variance and a bias term, we observe:

$$\mathbb{E}\left[\|\hat{\boldsymbol{z}} - \boldsymbol{z}\|_2^2 \right] \geq \boldsymbol{J}_{\psi,i}^\top \mathcal{I}_h(\boldsymbol{z})^{-1} \boldsymbol{J}_{\psi,i} + \|\mathbb{E}\left[\hat{\boldsymbol{z}}\right] - \boldsymbol{z}\|_2^2. \qquad (9)$$

When estimating individual elements $z_i \in \boldsymbol{z}$, equation 9 reduces to:

$$\mathbb{E}\left[(\hat{z}_i - z_i)^2 \right] \geq \left(\frac{\partial \psi}{\partial z_i} \right)^2 \mathcal{I}_h(\boldsymbol{z})_{ii}^{-1} + (\mathbb{E}\left[\hat{z}_i\right] - z_i)^2. \qquad (10)$$

We observe from equation 8 that there are two ways to reduce the variance of the estimator: 1) increase the Fisher information or 2) reduce the sensitivity of the estimator to the value being estimated. If the Fisher information is constant, then the reduction in variance must come from the second option, which incurs bias. In the scalar case, per equation 10, any attempt to reduce the variance of the estimator will result in higher bias and will not yield a smaller mean squared error. However, in the multivariate case bounded FIL alone does not guarantee a bounded mean squared error. James-Stein type shrinkage can result in estimators with a smaller increase in bias than the corresponding decrease in variance relative to an unbiased estimator [Lehmann and Casella, 2006].

- When estimating FIL for subsets of $\mathcal{D}$, we assume the remainder of the data is *known* by the adversary. This assumption arises from the fact that selecting elements of the full FIM corresponding to the subset is equivalent to computing the FIM for that subset directly.

Using the FIM of the subset to compute FIL makes no assumptions about the remainder of $\mathcal{D}$. In fact, if the remainder of the data is assumed to be unknown or partially known, FIL is still a valid upper bound on information leakage. Intuitively, the adversary *gains at least as much information* about the target subset with knowledge of the remaining data. We derive a mathematically precise statement of this claim below.

Suppose the adversary aims to infer example $(\boldsymbol{x}_1, y_1)$ without knowledge of the remaining data. Let $\boldsymbol{z}$ be the full data vector and let $g(\boldsymbol{z})$ be a function that selects the first $d+1$ dimensions of $\boldsymbol{z}$ corresponding to $(\boldsymbol{x}_1, y_1)$. For an unbiased estimator $(\hat{\boldsymbol{x}}_1, \hat{y}_1)$ of $(\boldsymbol{x}_1, y_1)$, the Cramér-Rao bound under the parameter transformation g is given by:

$$\mathrm{Cov}((\hat{\boldsymbol{x}}_1, \hat{y}_1)) \succeq \boldsymbol{J}_g^\top \mathcal{I}_h(\boldsymbol{z})^{-1} \boldsymbol{J}_g, \qquad (11)$$

where $\boldsymbol{J}_g$ is the Jacobian of g. Simplifying this expression gives $\mathrm{Cov}((\hat{\boldsymbol{x}}_1, \hat{y}_1)) \succeq [\mathcal{I}_h(\boldsymbol{z})^{-1}]_{0:d,0:d}$.

To finish the derivation, one can use the matrix block-inversion formula [Petersen and Pedersen, 2007] to obtain:

$$[\mathcal{I}_h(\boldsymbol{z})^{-1}]_{0:d,0:d} \succeq [\mathcal{I}_h(\boldsymbol{z})_{0:d,0:d}]^{-1}, \qquad (12)$$

the latter of which is the FIM for $(\boldsymbol{x}_1, y_1)$ under the assumption that the remaining data is known. Thus $\mathrm{Cov}((\hat{\boldsymbol{x}}_1, \hat{y}_1)) \succeq \eta_1^2$ where η_1 is the FIL for $(\boldsymbol{x}_1, y_1)$. This derivation can be easily generalized to any subset of the training data.

4 COMPUTING FIL

Learning setting. We assume a linear model with parameters $\boldsymbol{w} \in \mathbb{R}^d$ and minimize the regularized empirical risk:

$$\boldsymbol{w}^* = f(\mathcal{D}) \stackrel{\text{def}}{=} \arg\min_{\boldsymbol{w}} \sum_{i=1}^{n} \ell(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i) + \frac{n\lambda}{2} \|\boldsymbol{w}\|_2^2. \tag{13}$$

Furthermore, we assume that the loss $\ell(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i)$ is convex and twice differentiable. We denote by $f(\mathcal{D})$ the minimizer, $\boldsymbol{w}^*$, of equation 13 as a function of the dataset $\mathcal{D}$. When computing FIL at the example level, we let $f_i(\boldsymbol{x}, y)$ be the minimizer of equation 13 as a function of the i-th data point:

$$f_i(\boldsymbol{x}, y) \stackrel{\text{def}}{=} \arg\min_{\boldsymbol{w}} \sum_{j \neq i} \ell(\boldsymbol{w}^\top \boldsymbol{x}_j, y_j) + \ell(\boldsymbol{w}^\top \boldsymbol{x}, y) + \frac{n\lambda}{2} \|\boldsymbol{w}\|_2^2. \tag{14}$$

Output perturbation. The definition of FIL in equation 3 only applies to a randomized learning algorithm $\mathcal{A}$ with a differentiable density function. To obtain such a randomized learning algorithm from the minimizer $\boldsymbol{w}^*$ in equation

13, we adopt the Gaussian mechanism from differential privacy. A randomized algorithm $\mathcal{A}$ satisfies (ϵ, δ)-differential privacy with respect to dataset $\mathcal{D}$ if:

$$p(\mathcal{A}(\mathcal{D}) = h) \leq e^\epsilon p(\mathcal{A}(\mathcal{D}') = h) + \delta \qquad (15)$$

for all $\mathcal{D}$ and $\mathcal{D}'$ which differ by one example and all hypothesis $h \in \mathcal{H}$ [Dwork et al., 2006]. The Gaussian mechanism adds zero-mean isotropic Gaussian noise to the parameters $\boldsymbol{w}^*$. For a given ϵ and δ, the standard deviation σ can be chosen such that the Gaussian mechanism satisfies (ϵ, δ)-differential privacy [Dwork and Roth, 2014].

Fisher information loss. Let $\mathcal{A}(\mathcal{D}) = f(\mathcal{D}) + \boldsymbol{b}$ be the output-perturbed learning algorithm, where $\boldsymbol{b} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$. The FIM of $\boldsymbol{w}' \sim \mathcal{A}(\mathcal{D})$ is given by:

$$\mathcal{I}_{\boldsymbol{w}'}(\mathcal{D}) = \frac{1}{\sigma^2} \boldsymbol{J}_f^\top \boldsymbol{J}_f, \qquad (16)$$

where $\boldsymbol{J}_f \in \mathbb{R}^{d \times n(d+1)}$ is the Jacobian of $f(\mathcal{D})$ with respect to $\mathcal{D}$. The Jacobian $\boldsymbol{J}_f$ captures the sensitivity of the minimizer $\boldsymbol{w}^*$ with respect to the training dataset $\mathcal{D}$. See Appendix A for a simplified derivation of equation 16 and Kay [1993] for a more rigorous and general treatment. The FIL of the Gaussian mechanism with scale σ is then:

$$\eta = \frac{1}{\sigma} \|\boldsymbol{J}_f\|_2. \qquad (17)$$

Jacobian of the minimizer. We aim to compute the Fisher information loss of the Gaussian mechanism in equation 17 at the example level. This requires the Jacobian $\boldsymbol{J}_{f_i}$ of $f_i(\boldsymbol{x}, y)$, equation 14, with respect to $(\boldsymbol{x}, y)$ evaluated at $(\boldsymbol{x}_i, y_i)$. For a convex, twice-differentiable loss function $\ell(\boldsymbol{w}^\top \boldsymbol{x}, y)$, the Jacobian $\boldsymbol{J}_{f_i}$ evaluated at $(\boldsymbol{x}_i, y_i)$ is given by:

$$\boldsymbol{J}_{f_i}\Big|_{\boldsymbol{x}_i, y_i} = -\boldsymbol{H}_{\boldsymbol{w}^*}^{-1} \nabla_{\boldsymbol{x}, y} \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}^{*\top} \boldsymbol{x}_i, y_i). \qquad (18)$$

The Hessian is computed over the full dataset $\boldsymbol{H}_{\boldsymbol{w}^*} = \sum_{i=1}^{n} \nabla_{\boldsymbol{w}}^2 \ell(\boldsymbol{w}^{*\top} \boldsymbol{x}_i, y_i) + n\lambda \boldsymbol{I}$. The term $\nabla_{\boldsymbol{x}, y} \nabla_{\boldsymbol{w}} \ell = [\nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{w}} \ell, \nabla_y \nabla_{\boldsymbol{w}} \ell] \in \mathbb{R}^{d \times (d+1)}$ is the Jacobian of $\nabla_{\boldsymbol{w}} \ell$ with respect to $(\boldsymbol{x}, y)$ with entries given by:

$$(\nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{w}} \ell)_{ij} = \frac{\partial (\nabla_{\boldsymbol{w}} \ell)_i}{\partial \boldsymbol{x}_j} \quad \text{and} \quad (\nabla_y \nabla_{\boldsymbol{w}} \ell)_i = \frac{\partial (\nabla_{\boldsymbol{w}} \ell)_i}{\partial y}. \tag{19}$$

A derivation is given in Appendix B.

4.1 MODEL-SPECIFIC DERIVATIONS

Linear regression. In linear regression, the loss function is:

$$\ell(\boldsymbol{w}^\top \boldsymbol{x}, y) = \frac{1}{2} \left(\boldsymbol{w}^\top \boldsymbol{x} - y \right)^2. \qquad (20)$$

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top$ be the design matrix and $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$ be the vector of labels. The gradient of ℓ with respect to $\boldsymbol{w}$ is:

$$\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}^\top \boldsymbol{x}, y) = (\boldsymbol{w}^\top \boldsymbol{x} - y)\boldsymbol{x}, \qquad (21)$$

Taking Jacobian of equation 21 with respect to $(\boldsymbol{x}, y)$ gives:

$$\nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{w}} \ell = \boldsymbol{x}\boldsymbol{w}^\top + \left(\boldsymbol{w}^\top \boldsymbol{x} - y\right)\boldsymbol{I}, \text{ and } \nabla_y \nabla_{\boldsymbol{w}} \ell = -\boldsymbol{x}. \qquad (22)$$

Using $\boldsymbol{H}_{\boldsymbol{w}^*} = \boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{I}$, we can then combine the above with equation 18 to obtain the partial Jacobians $\boldsymbol{J}_{f_i, \boldsymbol{x}}$ and $\boldsymbol{J}_{f_i, y}$. Finally, using equation 17, the FIL η_i for example $(\boldsymbol{x}_i, y_i)$ with the Gaussian mechanism is $\eta_i = \frac{1}{\sigma}\|\boldsymbol{J}_{f_i}\|_2$ where $\boldsymbol{J}_{f_i} = [\boldsymbol{J}_{f_i, \boldsymbol{x}}, \ \boldsymbol{J}_{f_i, y}]$.

Logistic regression. The binary logistic regression loss is:

$$\ell(\boldsymbol{w}^\top \boldsymbol{x}, y) = -y \log s(\boldsymbol{w}^\top \boldsymbol{x}) - (1 - y) \log(1 - s(\boldsymbol{w}^\top \boldsymbol{x})) \qquad (23)$$

where $y \in \{0, 1\}$ and $s(a) = 1/(1 + \exp(-a))$. The gradient of the loss with respect to $\boldsymbol{w}$ is:

$$\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}^\top \boldsymbol{x}, y) = (s(\boldsymbol{w}^\top \boldsymbol{x}) - y)\boldsymbol{x}, \qquad (24)$$

and the Hessian is:

$$\nabla_{\boldsymbol{w}}^2 \ell(\boldsymbol{w}^\top \boldsymbol{x}, y) = s(\boldsymbol{w}^\top \boldsymbol{x})(1 - s(\boldsymbol{w}^\top \boldsymbol{x}))\boldsymbol{x}\boldsymbol{x}^\top. \qquad (25)$$

The partial Jacobians of the gradient in equation 24 with respect to $(\boldsymbol{x}, y)$ are:

$$\nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{w}} \ell = s(\boldsymbol{w}^\top \boldsymbol{x})(1 - s(\boldsymbol{w}^\top \boldsymbol{x}))\boldsymbol{x}\boldsymbol{w}^\top + (s(\boldsymbol{w}^\top \boldsymbol{x}) - y)\boldsymbol{I}, \qquad (26)$$

and

$$\nabla_y \nabla_{\boldsymbol{w}} \ell = -\boldsymbol{x}. \qquad (27)$$

We can combine the partial Jacobians with the Hessian in equation 25 to form the full Jacobian $\boldsymbol{J}_{f_i}$ in equation 18.

Computational considerations. Computing the FIM for a given example $(\boldsymbol{x}_i, y_i)$ requires computing the Jacobian $\boldsymbol{J}_{f_i}$ in equation 18. For both linear and logistic regression, computing the inverse Hessian, $\boldsymbol{H}_{\boldsymbol{w}^*}^{-1}$ requires $O(d^2 n)$ operations, assuming $n \geq d$. Since the Hessian is the same for all examples in $\mathcal{D}$, the cost may be amortized over n evaluations of per-example FIL. Constructing the per-example FIM requires $O(d^3)$ operations, and the largest singular value can computed efficiently by the power method. In the more general case, with $\boldsymbol{z} \in \mathbb{R}^m$ computing $\mathcal{I}_{\boldsymbol{w}'}(\boldsymbol{z})$ requires $O(dm^2)$ operations, which is quadratic in the size of $\boldsymbol{z}$. If $\boldsymbol{z}$ is the full dataset, for example, then $m = (d+1)n$ which results in $O(d^3 n^2)$ operations to compute the FIM.

Algorithm 1 Iteratively reweighted Fisher information loss.

1: **Input**: Data set $\mathcal{D}$, loss function $\ell(\cdot)$, number of iterations T, and noise scale σ.
2: Initialize sample weights $\omega_i^0 \leftarrow 1$.
3: **for** $t \leftarrow 1$ to T **do**
4: $\quad \boldsymbol{w}^* \leftarrow \arg\min_{\boldsymbol{w}} \sum_{i=1}^n \omega_i^{t-1} \ell(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i) + \frac{n\lambda}{2}\|\boldsymbol{w}\|_2^2$.
5: $\quad \boldsymbol{w}' \leftarrow \boldsymbol{w}^* + \boldsymbol{b}$ where $\boldsymbol{b} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$.
6: $\quad \eta_i \leftarrow (\|\mathcal{I}_{\boldsymbol{w}'}(\boldsymbol{x}_i, y_i)\|_2)^{1/2}$.
7: $\quad \omega_i^t \leftarrow \frac{n\omega_i^{t-1}/\eta_i}{\sum_{i=1}^n \omega_i^{t-1}/\eta_i}$.
8: **end for**
9: **Return**: The private weights $\boldsymbol{w}'$.

---

5 ITERATIVELY REWEIGHTED FIL

Prior work has shown that existing privacy mechanisms fail to provide equitable protection against privacy attacks for different subgroups [Yaghini et al., 2019]. We address this issue with iteratively reweighted Fisher information loss (IRFIL; Algorithm 1), which yields a model with an equal per-example FIL across all examples in $\mathcal{D}$. This is done by re-weighting the per-example surrogate loss $\ell(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i)$ over repeated computations of the minimizer $\boldsymbol{w}^*$.

After the first iteration in Algorithm 1, the weight for the i-th example is inversely proportional to the initial η_i. At successive iterations, the per-example weight ω_i^t is multiplicatively updated by a value inversely proportional to the current model's FIL. The normalization on line 7 is primarily for numerical stability, keeping the weights from shrinking too rapidly. Without L_2 regularization, the resulting η_i are invariant to the norm of the weights. With L_2 regularization, the normalization helps to keep the ratio of the primary objective to the regularization term constant.

If $\eta_1 = \ldots = \eta_n = \eta$ are constant, then Algorithm 1 has converged since the update on line 7 yields a fixed point:

$$\omega_i^t = \frac{n\omega_i^{t-1}/\eta}{\sum_{i=1}^n \omega_i^{t-1}/\eta} = \frac{n\omega_i^{t-1}}{\sum_{i=1}^n \omega_i^{t-1}} = \frac{n\omega_i^{t-1}}{n} = \omega_i^{t-1}, \qquad (28)$$

where the second-to-last equality follows since the ω_i^{t-1} were normalized to sum to n at the previous iteration.

Including the weights in computing $\mathcal{I}_{\boldsymbol{w}'}(\boldsymbol{x}_i, y_i)$ via equation 18 is a straightforward application of the chain rule:

$$\boldsymbol{J}_{f_i} = -\omega_i (\text{diag}(\omega_1, \ldots, \omega_n)\boldsymbol{H}_{\boldsymbol{w}^*})^{-1} \nabla_{\boldsymbol{x}, y} \nabla_{\boldsymbol{w}} \ell. \qquad (29)$$

6 EXPERIMENTS

We perform experiments with linear and logistic regression models on MNIST [LeCun and Cortes, 1998] and CIFAR-10 [Krizhevsky, 2009]. For attribute inversion attacks, we use the "IWPC" dataset from the
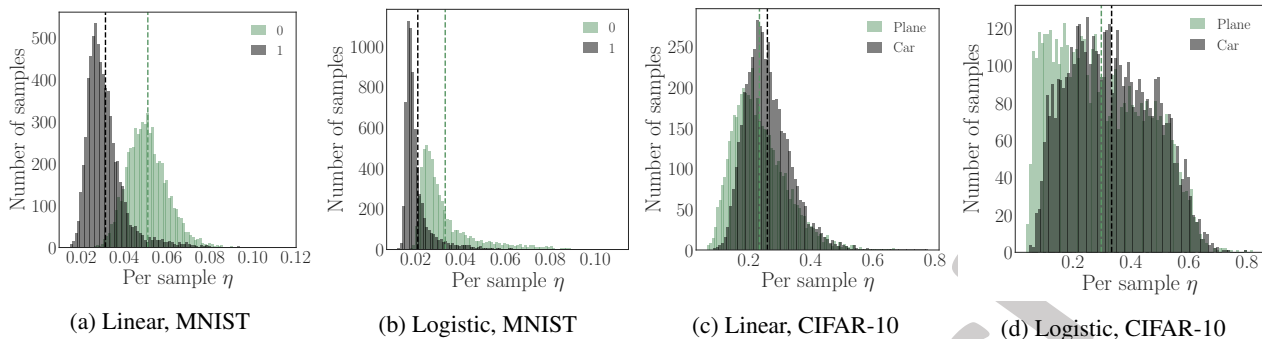
Figure 2: Histograms of per-example η separated by class label for the MNIST and CIFAR-10 training sets for linear and logistic regression. Each class label's mean η is denoted by the dashed vertical line.
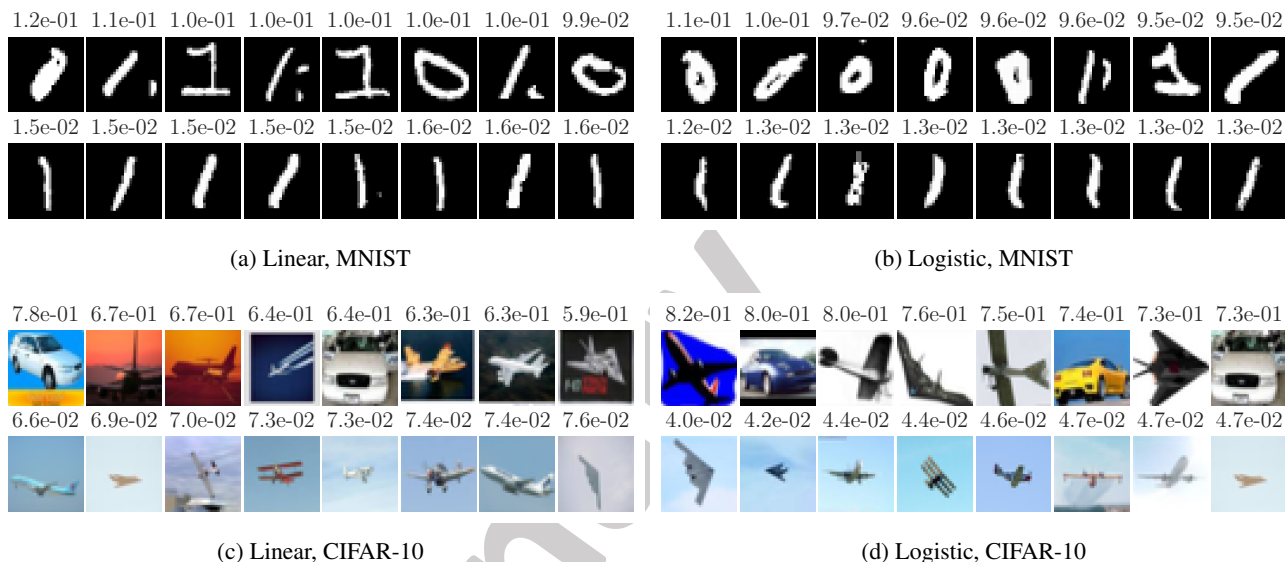

Figure 3: The eight images with the smallest and largest η over the MNIST and CIFAR-10 training sets for linear and logistic regression. The number above each individual image is the corresponding η.

Pharmacogenetics and Pharmacogenomics Knowledge Base [Klein et al., 2009] and the "Adult" dataset from the UC Irvine data repository [Dua and Graff, 2017]. Code to reproduce our results is available at `https://github.com/facebookresearch/fisher_information_loss`.

6.1 EXPERIMENTAL SETUP

Datasets. On MNIST we perform binary classification of the digits 0 and 1 using a training dataset of 12,665 examples. On CIFAR-10, we classify images of planes from cars using a training dataset with 10,000 examples. We normalize all inputs to lie in the unit ball, $\max_i \|\boldsymbol{x}_i\|_2 \leq 1$, and then project each input using PCA onto the top twenty principal components for the corresponding dataset.

The IWPC dataset requires estimating the stable dose of warfarin (an oral anticoagulant) based on clinical and ge-

netic traits. We use the same preprocessed dataset as Yeom et al. [2018]. Of the 4,819 examples, we randomly split 20% into a test set and use the remaining 80% for the training set. The UCI Adult dataset requires classifying individuals with income above or below $50,000. We remove examples with missing features, leaving 30,162 examples in the training set and 15,060 in the test set. For the purpose of attribute inference, the marital status features are combined into a single category of married or unmarried and relationship features are removed [Mehnaz et al., 2020]. For both IWPC and UCI Adult, nominal features are converted to one-hot vectors with the last value dropped to avoid perfect collinearity in the encoding. Numerical features are centered to zero mean and unit variance. We do not otherwise preprocess the datasets, yielding a total of 14 features per example for IWPC and 86 features per example for UCI Adult.

Hyperparameters. Unless otherwise stated, for linear regression we do not apply L_2 regularization. For logistic
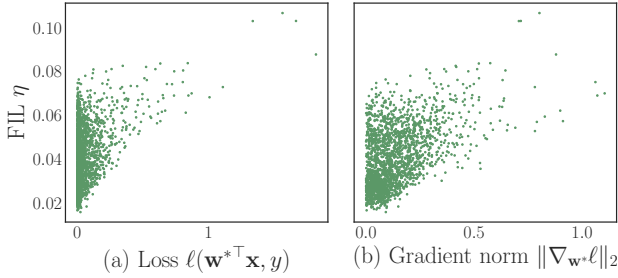
Figure 4: Scatter plots of the FIL η with (a) the loss, and (b) the norm of the gradient on MNIST with linear regression and 2,000 randomly selected examples.

(a) MNIST   (b) CIFAR-10

Figure 5: The standard deviation of the example-level η over iterations of the IRFIL algorithm.

Table 1: The mean $\bar{\eta}$ ($\pm$ standard deviation) of the example-level η and test accuracy before and after IRFIL.

Model	MNIST		CIFAR-10	
	$\bar{\eta}$	Accuracy	$\bar{\eta}$	Accuracy
Linear	0.040 ± 0.014	100	0.26 ± 0.08	79.8
+IRFIL	0.047 ± 0.000	99.8	0.30 ± 0.00	80.7
Logistic	0.027 ± 0.012	99.8	0.32 ± 0.15	79.9
+IRFIL	0.024 ± 0.000	99.7	0.25 ± 0.00	79.1

regression, λ is set to the largest value such that the training set accuracy is the same up to two significant digits as that of linear regression. For linear regression, the targets are $y_i \in \{-1, 1\}$, and we compute the exact minimizer. For logistic regression, we use limited-memory BFGS to compute the minimizer. We typically report η assuming a Gaussian noise scale of $\sigma = 1$; hence $\eta = \|\boldsymbol{J}_f\|_2$.

6.2 VALIDATING FISHER INFORMATION LOSS

Histograms of the per-example η for both MNIST and CIFAR-10 are shown in figure 2, separated by output class label. On MNIST the histograms have distinct modes. For both linear and logistic regression, the mode at the larger value is for the digit 0, implying that the model in general contains more information about images of 0 than of 1. While the modes are not as distinct on CIFAR-10, the class-specific means are still separate, with class label "car" having a larger η in general.

Figure 3 shows the eight images with the largest and smallest η in each of the four settings. The images with the smallest η are consistent with the class means in table 1. For MNIST, these correspond to the digit 1 written in a very typical manner. For CIFAR-10, the smallest η images are small planes on a blue-sky background. As expected, the images with the largest η are much more idiosyncratic. Of the 100 largest η examples for logistic and linear regression, 24 overlap for MNIST and 34 overlap for CIFAR-10.

We compare η to alternative heuristics that could correlate to the information a model contains about an example. Specifically, we measure the per-example loss $\ell(\boldsymbol{w}^{*\top} \boldsymbol{x}_i, y_i)$, and the norm of the gradient $\|\nabla_{\boldsymbol{w}^*} \ell(\boldsymbol{w}^{*\top} \boldsymbol{x}_i, y_i)\|_2$. Figure 4 shows scatter plots of η against these alternatives. The FIL η correlates with both metrics; however, high values of η exist throughout the range of both alternatives. On the other hand, a low η invariably implies a low loss. This shows that while loss-based heuristics may enjoy high precision in assessing privacy vulnerability, such heuristics should be used with caution, since they could suffer from low recall.
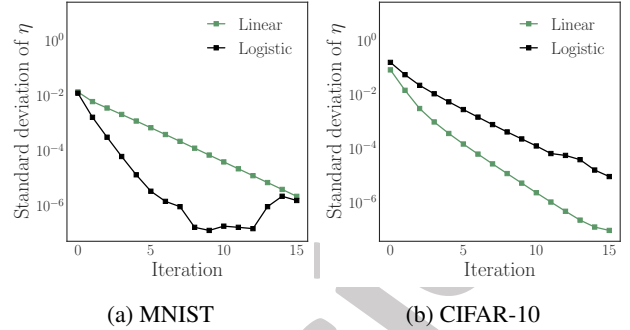
6.3 REWEIGHTED FIL

We empirically evaluate the IRFIL algorithm in figure 5, which plots the standard deviation of the per-example η against the number of re-weighting iterations. The per-example η converge to the same value after only a few iterations for both linear and logistic regression on MNIST and CIFAR-10. Table 1 shows the mean and standard deviation of η, as well as the test accuracy for models trained with and without IRFIL. Neither the average FIL $\bar{\eta}$ nor the test accuracy are especially sensitive to the IRFIL algorithm. However, without IRFIL the standard deviation in η is significantly higher, implying that the initial information leakage varies substantially across training examples. Overall, IR-FIL achieves fairness in privacy loss with little change in accuracy or average privacy loss.

6.4 ATTRIBUTE INFERENCE ATTACKS

We investigate how the success of attribute inference attacks varies with different levels of Fisher information loss. To do so, we use two attribute inversion attacks: 1) a white-box attack based on the FIL threat model, and 2) a black-box attack following the method described in Fredrikson et al. [2014]. The goal of the adversary is to infer the value of a nominal target attribute $x^{\text{tgt}} \in \boldsymbol{x}$ for a given example.

White-box attack. The white-box setting assumes the adversary has access to the complete training dataset $\mathcal{D} \setminus x_i^{\text{tgt}}$ for all but the target attribute of the example under attack.
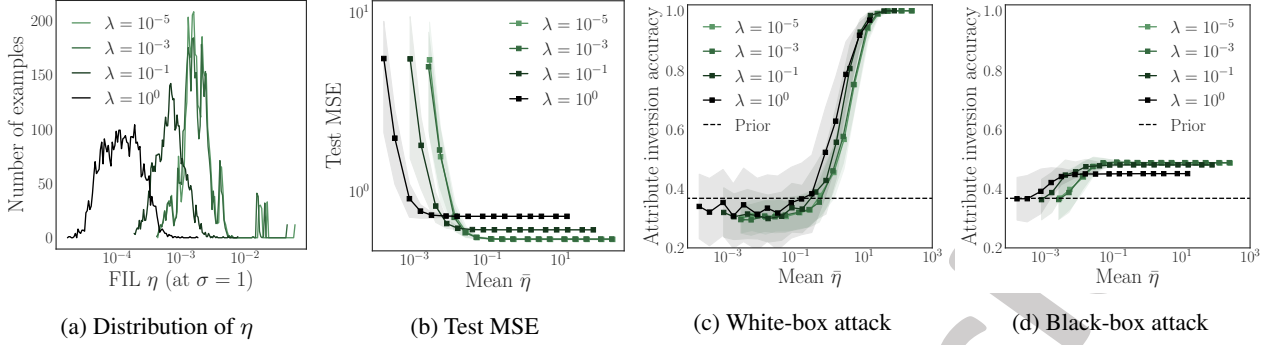
(a) Distribution of η (b) Test MSE   (c) White-box attack   (d) Black-box attack

Figure 6: The effect of (a) the L_2 regularization parameter λ on η for IWPC, and the effect of the mean $\bar{\eta}$ varying λ on (b) test MSE, (c) white-box accuracy, and (d) black-box accuracy. Standard deviations are computed over 100 trials.
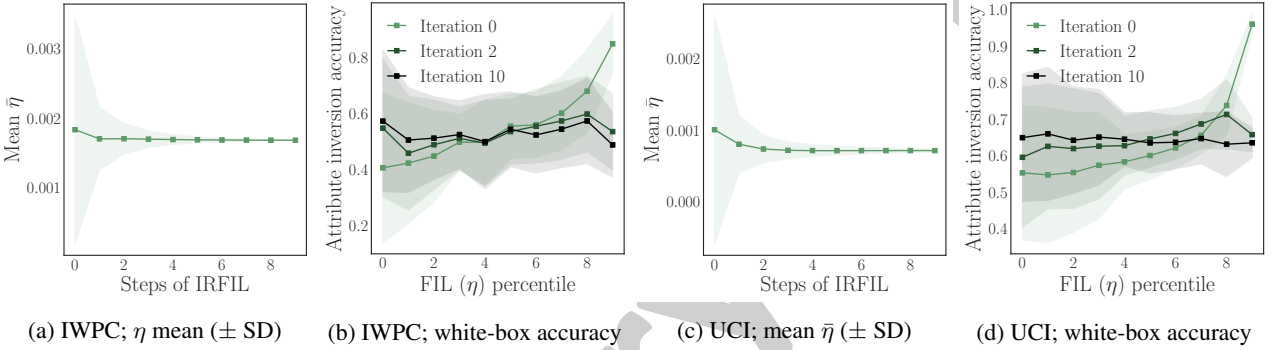

(a) IWPC; η mean ($\pm$ SD)   (b) IWPC; white-box accuracy   (c) UCI; mean $\bar{\eta}$ ($\pm$ SD)   (d) UCI; white-box accuracy

Figure 7: We show how the per-example attribute-level η converge over iterations of IRFIL for (a) IWPC and (c) UCI Adult. The mean accuracy of the white-box attack over increasing η-deciles also converges to a similar value over iterations of IRFIL for (b) IWPC and (d) UCI Adult. Standard deviations are computed over 100 trials.

The adversary also has access to the perturbed model parameters $\boldsymbol{w}'$. We also assume the adversary has complete knowledge of the model training details. Only the value of x_i^{tgt} for the example under attack is opaque to the adversary.

The adversary infers the hidden attribute of the i-th example by estimating $\hat{x}_i^{\text{tgt}}$ to minimize the distance of the derived model from the given model $\boldsymbol{w}'$:

$$\hat{x}_i^{\text{tgt}} = \arg\min_{x^{\text{tgt}}} \left\| \boldsymbol{w}' - f(\mathcal{D} \setminus x_i^{\text{tgt}} \cup x^{\text{tgt}}) \right\|_2, \qquad (30)$$

where $f(\cdot)$ yields the minimizer (equation 13).

Black-box attack. In the black-box setting the adversary has access to the target example except the value of the target attribute $\boldsymbol{x}_i \setminus x_i^{\text{tgt}}$, the label y_i, the prior distribution $p(x^{\text{tgt}})$, model predictions $\phi(\boldsymbol{w}', \boldsymbol{x})$ via black-box queries, and model performance statistics $\pi(\phi(\boldsymbol{w}', \boldsymbol{x}), y)$. For a given model the attack infers the target attribute value by maximizing:

$$\hat{x}_i^{\text{tgt}} = \arg\max_{x^{\text{tgt}}} p(x^{\text{tgt}}) \pi(\phi(\boldsymbol{w}', \boldsymbol{x}_i \setminus x_i^{\text{tgt}} \cup x^{\text{tgt}}), y_i). \quad (31)$$

We compare both attacks to a simple baseline adversary which infers the target attribute $\hat{x}_i^{\text{tgt}} = \arg\max_{x^{\text{tgt}}} p(x^{\text{tgt}})$.

For patients in IWPC, we infer the allele of the VKORC1 gene, which has three possible values. We use linear regression with $\lambda = 10^{-2}$. For the black-box adversary, the performance metric $\pi(\phi(\boldsymbol{w}', \boldsymbol{x}), y) = p(\boldsymbol{w}'^\top \boldsymbol{x}; y, s^2)$ is a Gaussian distribution with mean y and variance given by the standard error s^2 on the training set.

Result. Figure 6 shows the effect of L_2 regularization and the mean $\bar{\eta}$ on the accuracy of the white-box and black-box attacks. For larger λ, the MSE degrades at smaller $\bar{\eta}$ (6b), suggesting more L_2 regularization requires less noise (smaller σ) to achieve the same $\bar{\eta}$. The white-box (6c) and black-box (6d) accuracies degrade as η decreases. However, accross λ the white-box accuracy follows $\bar{\eta}$ more closely than the black-box accuracy (6d), which tends to track with MSE. By choosing σ such that $\bar{\eta} \approx 10^{-3}$, we can protect against both adversaries with little degradation in MSE.

We show the effect of IRFIL on the white-box attack using both the IWPC and UCI Adult datasets. For UCI Adult, we infer the binary marital status attribute. For both datasets, we compute η specific to the attribute under attack for each example and use $\sigma = 10^{-3}$. Figure 7 shows that the η for each example rapidly converge to the same value over iterations of IRFIL. We sort the examples into deciles based on

their initial η values. We then compute the average attribute inversion accuracy for each decile at iterations of IRFIL. For both datasets, the highest η decile is initially much more susceptible to attribute inversion than the lowest η decile. However, after only two iterations of IRFIL, the inversion accuracies flatten substantially across η deciles, and after ten iterations they are nearly constant. Hence, IRFIL can equalize vulnerability to privacy attacks across individuals.

7 DISCUSSION AND FUTURE WORK

We demonstrated that Fisher information loss can be used to assess the information leaked by a model about its training data. A primary benefit of FIL over *a priori* guarantees like differential privacy is the ability to measure information leakage at various granularities with respect to the data at hand. This also allows FIL to be used to construct models with equi-distributed leakage, which can be done with iteratively reweighted FIL. Furthermore, FIL explicitly measures the inferential power of an adversary, and we validated that it correctly captures vulnerability to privacy attacks. As a result, FIL can be used by practitioners to tailor the resulting privacy to the desired granularity and to the adversary's knowledge and capabilities.

We motivated the use of FIL via the Cramér-Rao bound and delineated the corresponding threat model. However, the assumption that the adversary is limited to unbiased estimators may not hold in the presence of auxiliary information. The implications of this should be further investigated. Furthermore, unlike differential privacy, FIL does not implicitly degrade with correlated data. This property of FIL should also be further studied.

The IRFIL algorithm closely resembles iteratively reweighted least squares (IRLS), which has been widely studied for ℓ_p-norm regression [Green, 1984, Burrus et al., 1994] and sparse recovery [Daubechies et al., 2010]. While IRLS often converges rapidly in practice, the theoretical convergence rates are difficult to derive and do not reflect the empirical results [Ene and Vladu, 2019]. We also observed rapid and robust convergence with the IRFIL algorithm without any hyper-parameter tuning. Future work may help understand the convergence of IRFIL from a theoretical standpoint.

Finally, we considered FIL in the common setting of output-perturbed generalized linear models with Gaussian noise. However, many possible extensions exist in the randomization used including alternative noise distributions such as the Laplace distribution [Dwork et al., 2006], objective perturbation [Chaudhuri et al., 2011], or quantifying leakage via predictions directly using, for example, the exponential mechanism [McSherry and Talwar, 2007]. Furthermore, extending FIL to the setting of non-linear and non-convex models will facilitate its utility and broader adoption.

References

Dakshi Agrawal and Charu C Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 247–255, 2001.

Harald Anderson. Efficiency versus protection in a general randomized response model. *Scandinavian Journal of Statistics*, pages 11–19, 1977.

C Sidney Burrus, JA Barreto, and Ivan W Selesnick. Iterative reweighted least-squares design of fir filters. *IEEE Transactions on Signal Processing*, 42(11):2926–2936, 1994.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, 2019.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*, 2020.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 309–315, 2019.

Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 63(1):1–38, 2010.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL http://dx.doi.org/10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

Michael D Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on Fairness, Accountability and Transparency*, pages 35–47, 2018.

Alina Ene and Adrian Vladu. Improved convergence for ℓ_1 and ℓ_∞ regression via iteratively reweighted least squares. In *International Conference on Machine Learning*, pages 1794–1801. PMLR, 2019.

Farhad Farokhi and Mohamed Ali Kaafar. Modelling and quantifying membership information leakage in machine learning. *arXiv preprint arXiv:2001.10648*, 2020.

Farhad Farokhi and Henrik Sandberg. Fisher information as a measure of privacy: Preserving privacy of households with smart meters using batteries. *IEEE Transactions on Smart Grid*, 9(5):4726–4734, 2017.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 17–32, 2014.

Arpita Ghosh and Robert Kleinberg. Inferential privacy guarantees for differentially private mechanisms. *arXiv preprint arXiv:1603.01508*, 2016.

Peter J Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(2):149–170, 1984.

Thomas Humphries, Matthew Rafuse, Lindsey Tulloch, Simon Oya, Ian Goldberg, and Florian Kerschbaum. Differentially private learning does not bound membership inference. *arXiv preprint arXiv:2010.12112*, 2020.

Shiva Prasad Kasiviswanathan and Adam Smith. On thesemantics' of differential privacy: A bayesian formulation. *arXiv preprint arXiv:0803.3946*, 2008.

Steven M Kay. *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993.

T.E. Klein, R.B. Altman, N. Eriksson, B.F. Gage, S.E. Kimmel, M.-T.M. Lee, N.A. Limdi, D. Page, D.M. Roden, M.J. Wagner, M.D. Caldwell, and J.A. Johnson. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8): 753–764, 2009.

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

Y. LeCun and C. Cortes. The MNIST database of handwritten digits, 1998.

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE, 2007.

Changchang Liu, Supriyo Chakraborty, and Prateek Mittal. Dependence makes you vulnberable: Differential privacy under dependent tuples. In *NDSS*, volume 16, pages 21–24, 2016.

Yunhui Long, Vincent Bindschaedler, and Carl A Gunter. Towards measuring membership privacy. *arXiv preprint arXiv:1712.09136*, 2017.

Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. Understanding membership inferences on well-generalized learning models. *arXiv preprint arXiv:1802.04889*, 2018.

Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1): 3–es, 2007.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

Shagufta Mehnaz, Ninghui Li, and Elisa Bertino. Black-box model inversion attribute inference attacks on classification models. *arXiv preprint arXiv:2012.03404*, 2020.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

Kaare Brandt Petersen and Michael Syskind Pedersen. The matrix cookbook, 2007.

P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, 1998. URL http://www.csl.sri.com/papers/ sritr-98-04/.

Mark J Schervish. *Theory of statistics*. Springer Science & Business Media, 2012.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 601–618, 2016.

Mohammad Yaghini, Bogdan Kulynych, and Carmela Troncoso. Disparate vulnerability: On the unfairness of privacy attacks against machine learning. *arXiv preprint arXiv:1906.00389*, 2019.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE, 2018.