Analyzing Privacy Leakage in Machine Learning via Multiple Hypothesis Testing: A Lesson From Fano

Chuan Guo [1] **Alexandre Sablayrolles** [1] **Maziar Sanjabi** [1]

Abstract

Differential privacy (DP) is by far the most widely accepted framework for mitigating privacy risks in machine learning. However, exactly how small the privacy parameter ϵ needs to be to protect against certain privacy risks in practice is still not well-understood. In this work, we study data reconstruction attacks for discrete data and analyze it under the framework of multiple hypothesis testing. For a learning algorithm satisfying (α, ϵ)-Rényi DP, we utilize different variants of the celebrated Fano's inequality to upper bound the attack advantage of a data reconstruction adversary. Our bound can be numerically computed to calibrate the privacy parameter ϵ to the desired level of privacy protection in practice, and complements the empirical evidence for the effectiveness of DP against data reconstruction attacks even at relatively large values of ϵ.

1. Introduction

As machine learning becomes increasingly ubiquitous in the real world, proper understanding of the privacy risks of ML also becomes a crucial aspect for its safe adoption. Numerous prior works have demonstrated privacy vulnerabilities throughout the ML training pipeline (Shokri et al., 2017; Song et al., 2017; Nasr et al., 2019; Zhu et al., 2019; Carlini et al., 2021). So far the only comprehensive defense against privacy attacks is differential privacy (DP; Dwork et al. (2006)), which has been successfully adapted for training private ML models (Chaudhuri et al., 2011; Shokri & Shmatikov, 2015; Abadi et al., 2016).

Unfortunately, differentially private training also comes at a huge cost to model accuracy if a small privacy parameter ϵ is desired. In contrast, in terms of the level of empirical protection conferred by DP against privacy attacks, the picture is much more optimistic: Across a wide range of attacks including membership inference, attribute inference and data reconstruction, even a small amount of DP noise is sufficient for thwarting most attacks (Jayaraman & Evans, 2019; Zhu et al., 2019; Carlini et al., 2019; Hannun et al., 2021). However, there is very little theoretical understanding of this phenomenon.

In this paper, we analyze privacy leakage in connection to the multiple hypothesis testing problem in information theory, and show that the empirical privacy protection conferred by DP with high ϵ may be more than previously thought. To this end, we first define a game (see Figure 1) between a private learner and an adversary that tries to perform a *data reconstruction attack* against the learned model. We analyze this game using the celebrated Fano's inequality to derive upper bounds on the adversary's attack advantage when the model is trained differentially privately.

Our analysis reveals an interesting and practically important insight that the DP parameter ϵ should scale with the number of possible values M that the private data can take on. When M is large, *e.g.*, $M = 10^{10}$ when extracting social security numbers from a trained language model (Carlini et al., 2019), even a relatively large ϵ has sufficient protection against data reconstruction attacks. More generally, given an input data distribution and a DP parameter ϵ, we give a numerical method for deriving an upper bound on the advantage of an arbitrary data reconstruction adversary. We empirically validate our bound against several existing attacks and show that it can provide useful guidance for selecting the appropriate value of ϵ in practice.

Contributions. Our main contributions are the following:

1. We formalize data reconstruction attacks for discrete data as an attack game (section 3).
2. We use Fano's inequality to derive a numerical method that upper bounds the adversary's advantage for (α, ϵ)-Rényi DP mechanisms (section 4 and section 5).
3. We experimentally validate our advantage bound against existing attack and show that it can be used to guide the selection of ϵ in practice (section 6).

[1]Meta AI. Correspondence to: Chuan Guo <chuanguo@meta.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

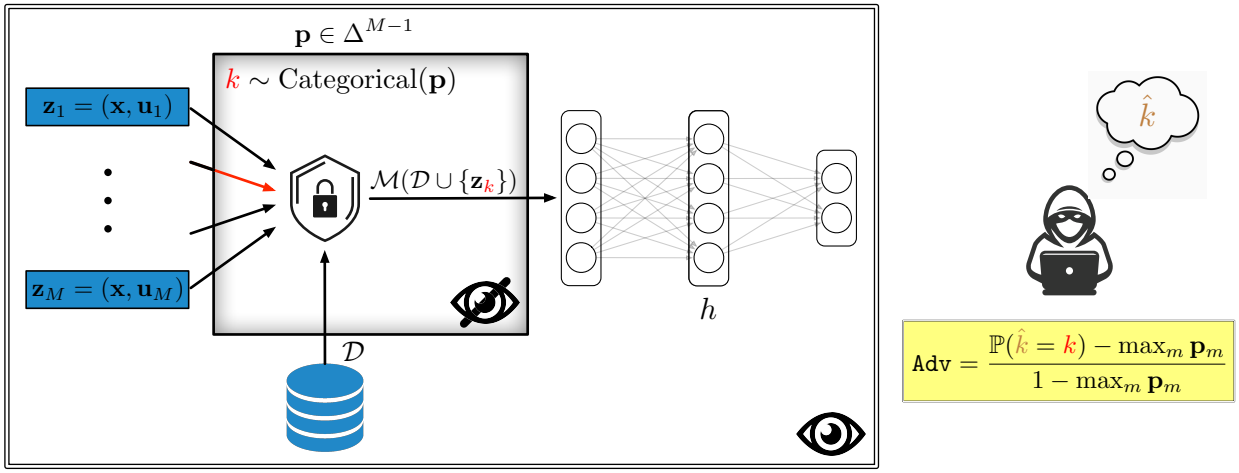


Figure 1: Illustration of the data reconstruction attack game. The target data takes on M possible values $\mathbf{u}_1, \ldots, \mathbf{u}_M$ with conditional probabilities $\mathbb{P}(\mathbf{u}_m|\mathbf{x}) = \mathbf{p}_m$. The game begins with k drawn from Categorical$(\mathbf{p})$, and the private learner trains a model $h \leftarrow \mathcal{M}(\mathcal{D} \cup \{\mathbf{z}_k\})$. The adversary guesses $\hat{k}$ and wins if $\hat{k} = k$. Advantage is defined so that $\texttt{Adv} \leq 1$ and guessing $\hat{k} = \arg\max_m \mathbf{p}_m$ achieves zero advantage. Processes inside the box marked with are unobserved by the adversary, while everything else is observable.

2. Background and Motivation

Privacy attacks. The machine learning pipeline exposes training samples to the outside world through the training procedure and/or trained model. Prior works showed that adversaries can exploit this exposure to compromise the privacy of training samples. The most well-studied type of privacy attack is *membership inference attack* (Shokri et al., 2017; Salem et al., 2018; Yeom et al., 2018; Sablayrolles et al., 2019), which aims to infer whether a sample $\mathbf{z}$ corresponding to an individual's data was part of the model's training set. This membership status can be a very sensitive attribute, *e.g.*, whether or not an individual participated in a cancer study indicates their disease status. Most state-of-the-art attacks (Ye et al., 2021; Watson et al., 2021; Carlini et al., 2022) follow a common strategy of measuring the model's loss compared to that of a random model trained without the target sample $\mathbf{z}$, with a large difference indicating that the sample was seen during training (*i.e.*, a member).

Other privacy attacks can extract more detailed information about a training sample beyond membership status. *Attribute inference attacks* (Fredrikson et al., 2014; Yeom et al., 2018) aim to reconstruct a training sample given access to the trained model and partial knowledge of the sample. *Data reconstruction attacks* (Fredrikson et al., 2015; Carlini et al., 2019; 2021; Balle et al., 2022) relax the partial knowledge assumption of attribute inference attacks and can recover training samples given only the trained model. In federated learning (McMahan et al., 2017), adversaries that observe the gradient updates can reconstruct private training samples using a process called *gradient inversion attack* (Zhu et al., 2019; Geiping et al., 2020). The existence of these privacy attacks calls for countermeasures that can preserve the utility of ML models while preventing unintended leakage of private information.

Differential privacy (Dwork et al., 2006) is a mathematical definition of privacy that upper bounds the amount of information leakage through a private mechanism. In the context of ML, the private mechanism $\mathcal{M}$ is a learning algorithm that, given any pair of datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ in a single training sample, ascertains that $\mathcal{M}(\mathcal{D})$ and $\mathcal{M}(\mathcal{D}')$ are ϵ-close in distribution for some chosen privacy parameter $\epsilon > 0$. In the classical definition of differential privacy, ϵ-closeness is defined in terms of *max divergence*: $\mathcal{M}$ is ϵ-differentially private (denoted ϵ-DP) if:

$$D_\infty(\mathcal{M}(\mathcal{D})||\mathcal{M}(\mathcal{D}')) := \sup_O \left[\log \mathbb{P}(\mathcal{M}(\mathcal{D}) \in O) - \log \mathbb{P}(\mathcal{M}(\mathcal{D}') \in O)\right] \leq \epsilon,$$

where O denotes a subset of the model space. One variant of DP that uses Rényi divergence to quantify closeness is *Rényi DP* (RDP; Mironov (2017)). For a given $\alpha \geq 1$, we say that $\mathcal{M}$ is (α, ϵ)-RDP if:

$$D_\alpha(\mathcal{M}(\mathcal{D})||\mathcal{M}(\mathcal{D}')) := \frac{1}{\alpha - 1} \log \mathbb{E}_{h \sim \mathcal{M}(\mathcal{D}')} \left[\frac{\mathbb{P}(\mathcal{M}(\mathcal{D}))^\alpha}{\mathbb{P}(\mathcal{M}(\mathcal{D}'))^\alpha}\right] \leq \epsilon.$$

Notably, as $\alpha \to \infty$, (α, ϵ)-RDP coincides with ϵ-DP.

Semantic privacy. Differential privacy has been shown to effectively prevent all of the aforementioned privacy attacks

when the privacy parameter ϵ is small enough (Jayaraman & Evans, 2019; Zhu et al., 2019; Carlini et al., 2019; Hannun et al., 2021). Conceptually, when $\epsilon \approx 0$, the learning algorithm outputs roughly the same distribution of models when a single training sample $\mathbf{z}$ is removed/replaced, hence an adversary cannot accurately infer any private information about $\mathbf{z}$. However, this reasoning does not quantify *how small ϵ needs to be to prevent a certain class of privacy attacks to a certain degree*. In practice, this form of *semantic guarantee* is arguably more meaningful, as it may inform policy decisions regarding the suitable range of ϵ to provide sufficient privacy protection, and enables privacy auditing (Jagielski et al., 2020) to verify that the learning algorithm's implementation is compliant.

Several existing works made partial progress towards answering this question. Yeom et al. (2018) formalized membership inference attacks by defining a game between the private learner and an adversary, and showed that the adversary's *advantage*—how well the adversary can infer a particular sample's membership status—is bounded by $e^{\epsilon} - 1$ when the learning algorithm is ϵ-DP. This bound has been tightened significantly in subsequent works (Erlingsson et al., 2019; Humphries et al., 2020; Mahloujifar et al., 2022; Thudi et al., 2022). Similarly, Bhowmick et al. (2018); Balle et al. (2022); Guo et al. (2022) formalized data reconstruction attacks and showed that for DP learning algorithm, the adversary's expected reconstruction error can be lower bounded using DP, Rényi-DP (Mironov, 2017), and Fisher information leakage (Hannun et al., 2021) privacy accounting. Our work makes further progress in this direction by analyzing data reconstruction attacks using tools from the multiple hypothesis testing literature, which we show is well-suited for discrete data.

3. Formalizing Data Reconstruction

To understand the semantic privacy guarantee for DP mechanisms against data reconstruction attacks, we first formally define a data reconstruction game for discrete data. Our formulation generalizes the membership inference game in existing literature (Yeom et al., 2018; Humphries et al., 2020), while specializing the formulation of Balle et al. (2022) to discrete data.

Data reconstruction game. Let $\mathcal{D}_{\text{train}} = \mathcal{D} \cup \{\mathbf{z}\}$ be the training set consisting of a *public set* $\mathcal{D}$ and a *private record* $\mathbf{z} = (\mathbf{x}, \mathbf{u})$, where $\mathbf{x}$ are attributes known to the adversary and $\mathbf{u}$ is unknown. Let $\mathcal{M}$ be the learning algorithm. We consider a white-box adversary with full knowledge of the public set $\mathcal{D}$ and the trained model $h = \mathcal{M}(\mathcal{D}_{\text{train}})$ whose objective is to infer the unknown attributes $\mathbf{u}$. Importantly, we assume that the unknown attribute is discrete (*e.g.*, gender, race, marital status) and can take on M values $\mathbf{u}_1, \ldots, \mathbf{u}_M$. For example, the experiment setting of Carlini et al. (2019) can be stated as $\mathbf{x}$ = "`My social security number is`" and $\mathbf{u} \in \{0, 1, \ldots, 9\}^{10}$ is the SSN number.

The attack game (see Figure 1 for an illustration) begins by drawing a random index k from a categorical distribution defined by the probability vector $\mathbf{p}$ and setting the unknown attribute $\mathbf{u} = \mathbf{u}_k$. The private learner $\mathcal{M}$ then trains a model h with $\mathbf{z} = \mathbf{z}_k = (\mathbf{x}, \mathbf{u}_k)$ and gives it to the adversary, who then outputs a guess $\hat{k}$ of the underlying index k. Note that both the random index k and any randomness in the learning algorithm $\mathcal{M}$ are unobserved by the adversary, but the learning algorithm itself is known.

Success metric. We generalize the *advantage* metric (Yeom et al., 2018) used in membership inference attack games to multiple categories. Here, the (Bayes) optimal guessing strategy without observing h is to simply guess $\hat{k} = \arg\max_m \mathbf{p}_m$ with success rate $\max_m \mathbf{p}_m$. The probability of successfully guessing k upon observing h, *i.e.*, $\mathbb{P}(\hat{k} = k)$, must be at least $\max_m \mathbf{p}_m$ in order to meaningfully leverage the private information contained in h about $\mathbf{u}$. Thus, we define advantage as the (normalized) difference between $\mathbb{P}(\hat{k} = k)$ and the *baseline success rate* $p^* := \max_m \mathbf{p}_m$, *i.e.*,

$$\texttt{Adv} := \frac{\mathbb{P}(\hat{k} = k) - p^*}{1 - p^*} \in [0, 1]. \tag{1}$$

Interpretation. Our data reconstruction game has the following important implications for privacy semantics.

1. *The private attribute* $\mathbf{u}$ *is considered leaked if and only if it is guessed exactly.* This is a direct consequence of defining adversary success as $\hat{k} = k$. For example, if the attribute is a person's age, then guessing $\hat{k} = 50$ when the ground truth is $k = 49$ should be considered more successful than when $k = 40$. In such settings, it may be more suitable to partition the input space into broader categories, *e.g.*, age ranges 0-9, 10-19, *etc.*, to allow inexact guesses.

2. *The attack game subsumes attribute inference attacks.* This can be done by setting the known attribute $\mathbf{x}$ accordingly. When $\mathbf{x} = \emptyset$, our game corresponds to the scenario commonly referred to as *data reconstruction* in existing literature (Carlini et al., 2019; 2021; Balle et al., 2022).

3. *Prior information is captured through the sampling probability* $\mathbf{p}$. Success rate of the Bayes optimal strategy is $p^* = \max_m \mathbf{p}_m$, which depends on the sampling probability vector $\mathbf{p}$. In the extreme case where $\mathbf{p}$ is a delta distribution on some k^*, which corresponds to the adversary having perfect knowledge of the private attribute $\mathbf{u}$, the model h provides *no additional information* about $\mathbf{u}$. This is in accordance with the "no free lunch theorem" in

differential privacy, which says that inference of private information is not preventable when the adversary is given arbitrary prior information (Dwork & Naor, 2010).

4. *Advantage captures attacker uncertainty.* One can also interpret advantage as a measure of uncertainty for the recovered value $\hat{k}$. Even when advantage is low, it is possible that the adversary recovers k by chance. For instance, an attack that always guesses a fixed $\hat{k}$ can correctly recover the true input k with non-trivial probability, but this should not be considered a successful recovery. By defining advantage in this manner, we consider an attack successful only when it can reduce its uncertainty sufficiently below the inherent randomness in k.

Relation to prior work. Our attack game is a direct generalization of the membership inference attack game in Yeom et al. (2018). In particular, the "strong adversary" game defined in Humphries et al. (2020); Nasr et al. (2021) is a special case of our data reconstruction game with $M = 2$. Our game formulation also complements Guo et al. (2022), which bounds the inferential power of a data reconstruction attack under the assumption that semantic similarity between private records is measured by mean squared error (MSE), which is better suited to continuous-valued data. In contrast, our formulation considers discrete data with the zero-one similarity measure, which is better suited for categorical data such as gender, race and marital status.

Our definition can also be viewed as a specialization of the data reconstruction attack game defined in Balle et al. (2022), where the learning algorithm $\mathcal{M}$ is said to be (η, γ)-ReRo (reconstruction robust) if:

$$\mathbb{P}_{k\sim\text{Categorical}(\mathbf{p}),h\sim\mathcal{M}(\mathcal{D}\cup\{\mathbf{z}_k\})}(\ell(\mathbf{z}_k, \hat{\mathbf{z}}) \leq \eta) \leq \gamma, \quad (2)$$

where $\hat{\mathbf{z}}$ is the adversary's reconstruction of $\mathbf{z}_k$ and ℓ is an error function that measures how faithful the reconstruction $\hat{\mathbf{z}}$ is to $\mathbf{z}_k$. By setting $\ell(\mathbf{z}_k, \hat{\mathbf{z}}) = \mathbb{1}(\hat{\mathbf{z}} = \mathbf{z}_k)$ and $\eta = 0$, one can show that Equation 2 is satisfied if and only if $\mathbb{P}(\hat{k} = k) \leq \gamma$, thus if $\mathcal{M}$ is (η, γ)-ReRo with $\eta < 1$ then $\texttt{Adv} \leq (\gamma - p^*)/(1 - p^*)$.

4. Connecting Data Reconstruction to Multiple Hypothesis Testing

In this section, we analyze the data reconstruction game defined in section 3 and derive a numerical method for the advantage upper bound when $\mathcal{M}$ is differentially private.

Fano's inequality. The central tool in our analysis is Fano's inequality—a celebrated result in information theory that lower bounds the error rate in multiple hypothesis testing. Suppose that X is a discrete random variable taking on M different values, and Y is an observed random variable determined by X, *e.g.*, Y is the output of the private learning algorithm $\mathcal{M}$. Let $\hat{X}$ be an estimate of X, and define the error random variable $E = (\hat{X} \neq X) \in \{0, 1\}$. Then:

$$H(X|Y) \leq H(E) + \mathbb{P}(E = 1)\log(M - 1), \quad (3)$$

where $H(E)$ denotes the entropy of E and $H(X|Y)$ denotes the conditional entropy of X given Y.

From Fano to advantage bound. The attack game defined in Figure 1 essentially tasks the adversary with a *multiple hypothesis test*, where the hypotheses are $H_k : h \sim \mathcal{M}(\mathcal{D} \cup \{\mathbf{z}_k\})$ for $k = 1, \ldots, M$. When $M = 2$, we recover the binary hypothesis testing interpretation of membership inference attack (Yeom et al., 2018; Humphries et al., 2020). Consequently, we can apply Fano's inequality to bound the adversary's advantage (*cf.* Equation 1) by setting $X = \text{Categorical}(\mathbf{p})$ and $Y = \mathcal{M}(\mathcal{D}_{\text{train}})$. In effect, X encodes all the information contained in the sensitive attribute $\mathbf{u}$, and Y is the output of a private mechanism that the adversary observes with partial information about X. For any adversary, let $t = \mathbb{P}(E = 1) = \mathbb{P}(\hat{k} \neq k)$ be the error probability, and define:

$$f(t) = H(X|Y) + t\log t + (1-t)\log(1-t) - t\log(M-1). \quad (4)$$

Since $t = \mathbb{P}(E = 1)$, Fano's inequality (*cf.* Equation 3) defines a constraint $f(t) \leq 0$ on t. In other words, the error probability of *any* adversary must satisfy $f(t) \leq 0$, and thus we can upper bound advantage (or equivalently, lower bound error probability) by solving

$$t^* = \min\{t \in [0, 1] : f(t) \leq 0\}, \quad (5)$$

and the advantage bound is $\texttt{Adv} \leq (1 - t^* - p^*)/(1 - p^*)$.

Mutual information privacy. In order to compute the advantage bound, we need to first evaluate the conditional entropy term $H(X|Y)$ in Equation 4. An alternative expression for conditional entropy is

$$H(X|Y) = H(X) - I(X; Y),$$

where $I(X; Y)$ is the mutual information between X and Y. Since $H(X)$ depends only on the sampling probability vector $\mathbf{p}$, the only factor that the DP mechanism controls is $I(X; Y)$. Intuitively, $I(X; Y)$ measures the amount of information leaked about the data $\mathbf{z}_k$ through the trained model, and a DP learning algorithm can limit this leakage to reduce the advantage of a data reconstruction adversary. It is easy to show that using mutual information to quantify privacy satisfies key properties of DP such as adaptive composition and post-processing.

The connection between mutual information and differential privacy has been observed in prior work on mutual information DP (MI-DP; Mir (2013); Cuff & Yu (2016); Wang et al.

(2016)), quantitative information flow (Alvim et al., 2011), and privacy funnel (Makhdoumi et al., 2014; Salamatian et al., 2020). Some works also empirically control mutual information in order to reduce privacy leakage (Bertran et al., 2019; Wang et al., 2021). Similar to our approach, Salamatian et al. (2020) also used Fano's inequality to derive a closed-form lower bound on the error rate $\mathbb{P}(\hat{k} \neq k)$, but the bound is too loose for practical purposes. In contrast, we give a *numerical method* below for upper bounding the adversary's advantage that is much tighter in practice.

Computing advantage bound. The advantage bound is a function of t^* in Equation 5, which we can obtain by solving a one-dimensional minimization problem as follows. Assume that the mechanism $\mathcal{M}$ satisfies $I(X;Y) \leq \mu$ for some $\mu > 0$, and define:

$$f_\mu(t) = H(X) - \mu + t\log t + (1-t)\log(1-t) - t\log(M-1). \tag{6}$$

Since $H(X|Y) = H(X) - I(X;Y) \geq H(X) - \mu$, we have that $f_\mu(t^*) \leq f(t^*) \leq 0$. Thus, we can instead minimize t subject to $f_\mu(t) \leq 0$ to obtain a lower bound $t^* \leq \min\{t \in [0,1] : f_\mu(t) \leq 0\}$ for the error probability of *any* data reconstruction adversary. Taking derivative gives:

$$f'_\mu(t) = \log t - \log(1-t) - \log(M-1). \tag{7}$$

It is easy to see that $f'_\mu(t) \geq 0$ if and only if $t \geq 1 - 1/M$, so f_μ is decreasing for $t \in [0, 1-1/M)$ and increasing for $t \in (1-1/M, 1]$. Moreover, we know that $f_\mu(0) > 0$ if $\mu < H(X)$, and $f_\mu(1-1/M) = H(X) - \mu - \log(M) \leq 0$. Hence we can use binary search over the interval $[0, 1-1/M]$ to find $\min\{t \in [0,1] : f_\mu(t) \leq 0\}$. Algorithm 1 summarizes the computation in pseudo-code.

Algorithm 1 Computing the advantage bound using Fano's inequality.

Input: Upper bound $I(X;Y) \leq \mu$, number of distinct attributes M, sampling probability vector $\mathbf{p}$.
$p^* \leftarrow \max_m \mathbf{p}_m$
if $\mu \geq H(\mathbf{p})$ **then**
  return 1
else
  Define $f_\mu(t) = H(\mathbf{p}) - \mu + t\log t + (1-t)\log(1-t) - t\log(M-1)$.
  assert $f_\mu(0) > 0$ and $f_\mu(1-1/M) \leq 0$
  Use binary search to find $t^* = \min\{t \in [0, 1-1/M] : f_\mu(t) \leq 0\}$.
  return $(1 - t^* - p^*)/(1 - p^*)$
end if

5. Advantage Bounds for DP Mechanisms

The advantage bound in section 4 depends crucially on the upper bound $I(X;Y) \leq \mu$ for the mutual information between the private attribute and the trained model. We first present a general bound when the mechanism is (α, ϵ)-RDP. For commonly used DP primitives such as randomized response and Gaussian mechanism, we show that $I(X;Y)$ can be exactly computed or estimated, leading to even tighter advantage bounds in practice.

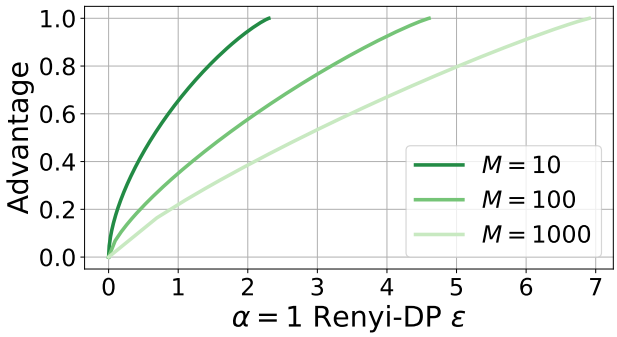


Figure 2: Reconstruction advantage vs. $\alpha = 1$ Rényi DP parameter ϵ for different values of M. As the number of possibilities M for the private attribute increases, the advantage curve becomes flatter, suggesting that a high ϵ still gives strong protection against data reconstruction.

5.1. Bound from RDP

The general bound for (α, ϵ)-RDP mechanisms involves a generalization of mutual information known as *Arimoto information* (Arimoto, 1977; Verdú, 2015), denoted $I_\alpha(X;Y)$. Notably, Arimoto information has natural connections to the order-α Rényi divergence used to define RDP, and coincides with mutual information when $\alpha = 1$. Theorem 1 below shows that if the private mechanism $\mathcal{M}$ is (α, ϵ)-RDP then $I_\alpha(X;Y) \leq \epsilon$. The special case of $\alpha = 1$ corresponds to $I(X;Y) = I_\alpha(X;Y) \leq \epsilon$, which we can use in Algorithm 1 to derive an advantage bound. Proof is given in Appendix B.

Theorem 1. *For any $\alpha \geq 1$, if $\mathcal{M}$ is (α, ϵ)-RDP then $I_\alpha(X;Y) \leq \epsilon$.*

Generalized Fano's inequality. For (α, ϵ)-RDP mechanisms with $\alpha > 1$, one can apply Theorem 1 by simply leveraging the monotonicity of Rényi divergence with respect to the order α. That is, if $\mathcal{M}$ is (α, ϵ)-RDP then it is also $(1, \epsilon)$-RDP, so $I(X;Y) \leq \epsilon$. However, this will give a loose advantage bound since the full range of α is not being effectively utilized. Fortunately, Fano's inequality can be generalized accordingly using Arimoto information for any order α, which enables us to generalize the advantage bound in section 4 with improved tightness. We give details for this generalized analysis in Appendix A.

Example. We showcase the power of the Fano's inequality analysis in combination with Theorem 1 in the following example. Here, we consider M training samples $\mathbf{z}_1, \ldots, \mathbf{z}_M$ with uniform sampling probability $\mathbf{p}_m = 1/M$ for $m = 1, \ldots, M$, hence $H(X) = \log M$. For a $(1, \epsilon)$-RDP learning algorithm $\mathcal{M}$, we can use Theorem 1 to upper bound $I(X;Y)$ and use Algorithm 1 to compute $t^* = \mathbb{P}(E = 1)$ and the advantage bound accordingly. Figure 2 shows the advantage bound as a function of ϵ for different values of M. As expected, all curves are monotonically increasing in ϵ, *i.e.*, more privacy leakage results in higher advantage for the adversary.

More interestingly, for the same value of ϵ, having a higher M reduces advantage, which reflects the fact that there is more entropy in the private attribute and thus the adversary's success probability decreases with M. In fact, it is easy to show that when M is uniformly distributed, Fano's inequality simplifies to:

$$\begin{aligned} & H(X) - I(X;Y) \leq H(E) + \mathbb{P}(E=1)\log(M-1) \\ \Leftrightarrow\ & I(X;Y) \geq (1 - \mathbb{P}(E=1))\log M - \log 2. \end{aligned}$$

Thus, for a given $c > 0$, the adversary's advantage at $\epsilon \approx c \log M$ is the same for all values of M, suggesting that ϵ should scale with $\log M$ to target a specific level of privacy risk. This theoretical analysis confirms the empirical observation that even a large ϵ confers non-trivial protection against data reconstruction attacks (Carlini et al., 2019). For instance, the membership inference advantage at $\epsilon = 1$ is the same as the data reconstruction advantage at $\epsilon = \log M \approx 23$ when the private attribute is uniformly randomly sampled from a set of size $M = 10^{10}$.

5.2. Exact Computation of $I(X;Y)$

For certain DP primitives, the mutual information $I(X;Y)$ can be computed either exactly or estimated numerically.

Randomized response. The (generalized) randomized response (RR) mechanism is defined as follows:

$$Y|X = \begin{cases} X & \text{w.p. } 1-q, \\ \text{Unif}(\{1, \ldots, M\}) & \text{w.p. } q \end{cases} \tag{8}$$

where $q \in [0, 1]$. To derive $I(X;Y)$, first note that $\mathbb{P}(X = x, Y = y) = P(X = x)P(Y = y|X = x)$ and $\mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x, Y = y)$ can be easily computed given q, M and the sampling probability vector $\mathbf{p}$. We can then numerically compute:

$$I(X;Y) = \sum_x \sum_y \mathbb{P}(X = x, Y = y) \log \frac{\mathbb{P}(X=x, Y=y)}{\mathbb{P}(X=x)\mathbb{P}(Y=y)}. \tag{9}$$

Gaussian mechanism. The Gaussian mechanism is a common primitive in the design of more complex private mechanisms such as DP-SGD (Abadi et al., 2016). Here, the mechanism first maps each private record $\mathbf{z}_m$ to some encoding $\mathbf{e}_m \in \mathbb{R}^d$ and outputs:

$$(Y|X = m) = \mathbf{e}_m + \mathcal{N}(0, \sigma^2 \mathbf{I}_{d\times d}). \tag{10}$$

As a result, Y is a Gaussian mixture with mixture probabilities $\mathbf{p}$. For example, in output perturbation (Chaudhuri et al., 2011), $\mathbf{e}_m$ is the model obtained when training a convex model on $\mathcal{D} \cup \{\mathbf{z}_m\}$; in DP-SGD, $\mathbf{e}_m$ is the gradient of the training loss for sample $\mathbf{z}_m$. For our purpose, it is equivalent to treat $\mathbf{e}_m$ as the private record instead of $\mathbf{z}_m$ since the mapping can be easily inverted (Zhu et al., 2019).

One can show that (see Equation 18 in Appendix B):

$$I(X;Y) = \sum_x \mathbb{P}(X = x) D_1((Y|X = x)||Y),$$

where D_1 denotes KL divergence. For the Gaussian mechanism, each term in the sum is the KL divergence between a Gaussian distribution and a mixture of Gaussians. Although this quantity is difficult to bound in closed form, it can be approximated via Monte-Carlo simulation by drawing $k \sim \text{Categorical}(\mathbf{p})$ and computing the KL divergence $D_1((Y|X = k)||Y)$. Taking expectation over multiple draws gives an unbiased estimate of $I(X;Y)$.

6. Experiment

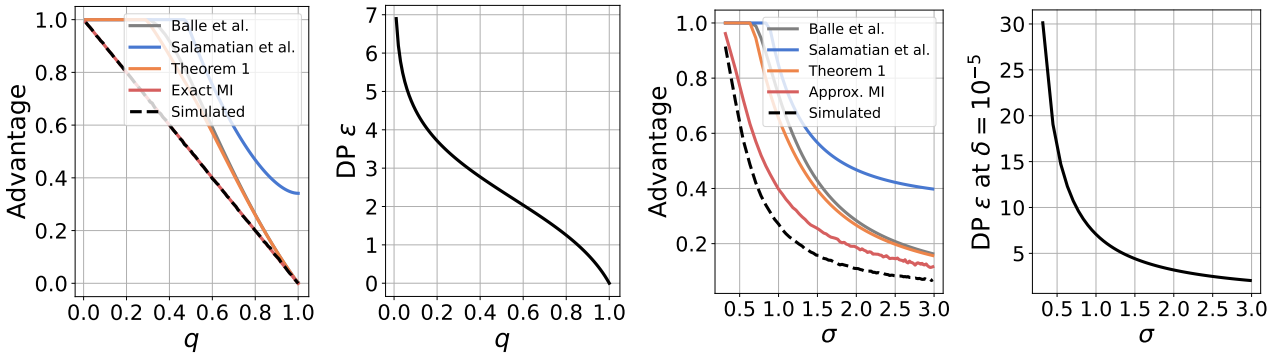
We perform a series of experiments to show that our Fano's inequality bound provides meaningful semantic guarantees against attribute inference and data reconstruction attacks.

6.1. Synthetic Data Experiments

We first experiment on synthetic data to numerically evaluate our Fano's inequality bound for both randomized response and the Gaussian mechanism.

Data generation. We generate data with $M = 10$ possible values sampled uniformly randomly, *i.e.*, $\mathbf{p}_m = 1/M$ for $m = 1, \ldots, M$. For randomized response, we sample the output value Y according to Equation 8 with different values of $q \in [0, 1]$. For the Gaussian mechanism, we encode the data as one-hot vectors $\mathbf{e}_1, \ldots, \mathbf{e}_M$ and output Y according to Equation 10 with $\sigma \in (0, 3]$.

Advantage bounds. We evaluate our Fano's inequality bound using both the general mutual information bound from RDP (Theorem 1) as well as either exact or approximate computation in subsection 5.2. As baselines, we evaluate the Fano's inequality bound from Salamatian et al. (2020) and the (η, γ)-ReRo bound from Balle et al. (2022), which upper bounds the adversary's success probability as: $\mathbb{P}(E = 0) \leq (e^\epsilon/M)^{\frac{\alpha-1}{\alpha}}$ if $\mathcal{M}$ is (α, ϵ)-RDP. We minimize this upper bound over α and convert it to an advantage bound for comparison.



(a) Randomized response

(b) Gaussian mechanism

Figure 3: Advantage plot for randomized response and the Gaussian mechanism. The advantage bound using Fano's inequality and exact/approximate mutual information closely matches the empirical lower bound. The advantage can be much lower than what the corresponding DP ϵ suggests in terms of privacy leakage.

Empirical lower bound. To compute an empirical lower bound for the advantage, we simulate a data reconstruction adversary using the *maximum a posteriori* (MAP) estimator $\hat{X}(Y)$, which returns the most likely value for X given Y using Bayes' rule. Each experiment then consists of:

1. Sampling X from $1, \ldots, M$ uniformly; **2.** Using either randomized response or the Gaussian mechanism to sample Y; **3.** Computing the MAP estimate $\hat{X}(Y)$ and comparing it to X. We repeat this experiment $100,000$ times to compute an empirical lower bound for the advantage.

Result. Figure 3 plots advantage as a function of the sampling probability q for randomized response and noise standard deviation σ for the Gaussian mechanism. For randomized response (Figure 3a), a higher q increases privacy and is reflected in the empirical lower bound (dashed line) as the adversary's advantage decreases to 0 when $q = 1$. The corresponding DP ϵ for different values of q are shown in the right plot. The bound from Theorem 1 is consistently lower than both baseline bounds, although all three bounds are relatively loose compared to the empirical lower bound. In contrast, using Fano's inequality with exact mutual information gives a *tight* advantage upper bound that matches the empirical lower bound. This is to be expected as it can be shown that the MAP estimator attains equality in Fano's inequality under the uniform input distribution; see Appendix A.

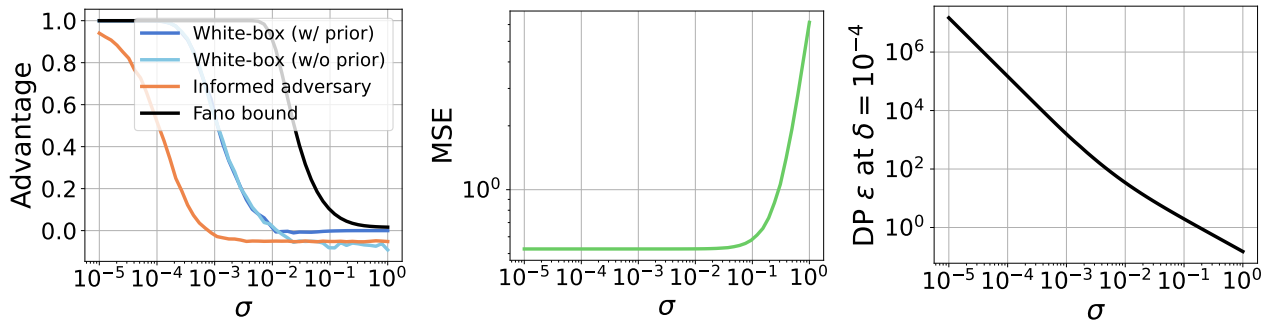
For the Gaussian mechanism (Figure 3b), a higher σ introduces more randomness and increases privacy, and thus both the empirical lower bound as well as the upper bounds decrease as σ increases. Similar to the randomized response result, the baseline bounds are strictly dominated by Theorem 1. Using Monte-Carlo to approximate mutual information gives the tightest advantage bound that closely matches the behavior of the empirical lower bound. These results demonstrate the power and tightness of the Fano's inequality analysis, and suggests that tight mutual information privacy accounting can grant strong semantic privacy guarantees for relatively low amounts of DP noise.

6.2. Attribute Inference for Pharmacogenetics Modeling

Next, we evaluate our advantage bound for privately trained pharmacogenetics models studied in Fredrikson et al. (2014) and show that it gives meaningful protection against attribute inference attacks.

Dataset and model. The IWPC dataset (Klein et al., 2009) contains data for clinical trial subjects, with the goal of training an ML model to predict the stable dosage of warfarin given the subjects' attributes. One particularly privacy-sensitive attribute is the VKORC1 gene type, which can be one of three values: `CC`, `CT` or `TT`. We train an L_2-regularized linear regression model on the attributes to predict the warfarin dosage (*i.e.*, target label) according to (Hannun et al., 2021).

DP mechanism. We employ the output perturbation mechanism (Chaudhuri et al., 2011), which adds Gaussian noise to the model parameters to guarantee differential privacy. For each sample j, let $\mathcal{D}_{-j}$ be the data subset containing all but the j-th sample $\mathbf{z}^{(j)}$. For the privacy analysis, we train three different models on datasets $\mathcal{D}_{-j} \cup \{\mathbf{z}_m^{(j)}\}$ for $m = 1, 2, 3$, where $\mathbf{z}_m^{(j)}$ represents the j-th sample with its VKORC1 gene set to one of the three values `CC`, `CT` or `TT`. We can then treat the three models' parameters as encodings of the private gene types under our data reconstruction game formulation, and the output perturbation mechanism is equivalent to the Gaussian mechanism applied to these encodings. To compute the advantage bound, we use a specialized mutual information bound in Theorem 2 in Ap-



(a) Advantage vs. σ (b) Test MSE vs. σ (c) DP ϵ vs. σ

Figure 4: Plots for the IWPC attribute inference attack experiment. The model is trained using output perturbation by adding Gaussian noise with standard deviation σ to the loss minimizer. At $\sigma = 0.1$, advantage (left) is relatively low when measured according to both the Fano bound and the empirical attacks, while the model's test MSE (middle) increases only moderately. In comparison, the DP ϵ (right) remains relatively high, overestimating privacy leakage.

pendix B instead of the generic RDP bound in Theorem 1.

Attacks. We consider attribute inference attacks for inferring the VKORC1 gene type from the trained model. Following Hannun et al. (2021), we implement a white-box attack with full knowledge of the data subset $\mathcal{D}_{-j}$ and all attributes of sample $\mathbf{z}^{(j)}$ except for the VKORC1 gene type. The attack essentially implements a maximum likelihood estimator, and we augment it with prior knowledge of the marginal distribution of the three gene types to produce a Bayesian variant. We also implement the Informed Adversary attack from Balle et al. (2022), which only has access to $\mathcal{D}_{-j}$ for predicting the VKORC1 gene type of $\mathbf{z}^{(j)}$. Details for the attacks are given in Appendix A.

Result. Figure 4a shows the adversary's advantage as a function of σ, which shows a decreasing trend as expected. The white-box attack with prior knowledge of the marginal VKORC1 gene distribution (dark blue) attains the highest advantage. Notably, its advantage converges to 0 as $\sigma \to \infty$, whereas the white-box attack without prior (light blue) and the Informed Adversary attack (orange) have negative advantage as $\sigma \to \infty$. This is because the VKORC1 gene's marginal distribution is non-uniform, with probabilities $0.367, 0.339, 0.294$ for `CC`, `CT` and `TT`, respectively. Since advantage as defined in Equation 1 uses $p^* = 0.367$ (corresponding to always predicting `CC`) as the baseline, it is sub-optimal to predict the uniform distribution when the model contains close to no information about $\mathbf{z}^{(j)}$.

The advantage bound using Fano's inequality (black line) strictly dominates the advantage of the three attacks. In particular, at $\sigma = 0.1$ the advantage bound predicts close to 0.1 advantage, which can be considered relatively low risk. Moreover, the test mean squared error (MSE) in Figure 4b shows that model utility remains satisfactory when $\sigma = 0.1$. At the same time, the corresponding DP ϵ in Figure 4c paints a pessimistic picture, with $\epsilon \approx 2$ at $\sigma = 0.1$. Our analysis suggests that $\sigma = 0.1$ can be a reasonable value for the output perturbation mechanism applied to this dataset for optimal privacy-utility trade-off.

6.3. Language Model Training with Canaries

Finally, we consider a language model canary extraction attack similar to the one performed in Carlini et al. (2019).

Setup. We experiment using a pre-trained GPT-2 (Radford et al., 2019) model—a causal language model for predicting the next token given a context token sequence. The vocabulary consists of $M = 50,256$ tokens. We fine-tune the model using DP-SGD (Abadi et al., 2016) on a single *canary sequence* of the form "`John Smith's credit card number is X`", with `X` being a token chosen uniformly randomly from the vocabulary. For the DP-SGD hyperparameters, we use a clipping norm of $C = 1$ and vary the number of fine-tuning steps T and the noise multiplier σ. For any $\alpha \geq 1$, one can show that DP-SGD is (α, ϵ)-RDP with $\epsilon = \alpha T / 2\sigma^2$, and then use Fano's inequality with Theorem 1 to obtain the advantage upper bound.

Attack. Since the model is fine-tuned to predict a single secret token `X` in the canary sequence, we can perform an extraction attack that compares the log-likelihood[1] of every token `X` before and after fine-tuning. The adversary then predicts the token with the largest difference. Note that such a differencing attack has been used in state-of-the-art membership inference attacks such as Ye et al. (2021); Watson et al. (2021); Carlini et al. (2022). Thus, each experiment

[1] We also experimented using the likelihood and found that the attack performs significantly worse.

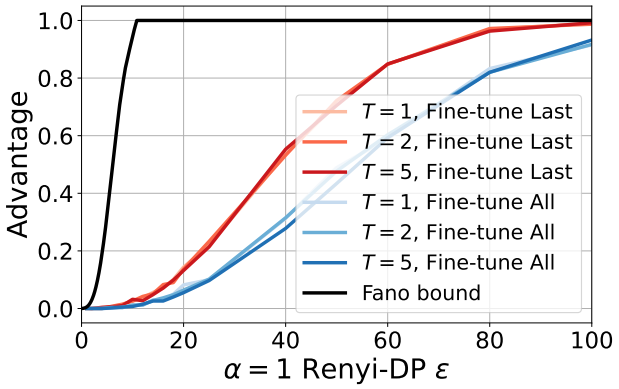


Figure 5: Result for the language model canary extraction attack. The model is fine-tuned on the canary sequence for T steps using DP-SGD. See text for details.

run consists of sampling X uniformly from $M = 50,256$ tokens, fine-tuning the GPT-2 model using DP-SGD on the canary sequence, and then running the extraction attack to recover X. We perform this experiment 1000 times to compute the adversary's advantage empirically.

Result. Figure 5 shows the adversary's advantage as a function of the $\alpha = 1$ RDP parameter $\epsilon = 2T/\sigma^2$, where $T \in \{1, 2, 5\}$ is the number of fine-tuning steps. We include two variants of the fine-tuning method, either fine-tune only the last layer (red) or fine-tune all layers (blue). As expected, the adversary's advantage increases monotonically as ϵ increases, while the advantage bound using Fano's inequality strictly dominates the advantage of simulated extraction attacks. Although the advantage bound overestimates the empirically computed advantage, it can still serve as a useful guideline for selecting the privacy parameter ϵ. For instance, at $\epsilon = 6$, the advantage upper bound is close to 0.5, which is certainly a non-trivial level of privacy protection for a relatively large value of ϵ. We suspect that the large gap is a result of upper bounding mutual information using the generic RDP bound in Theorem 1 and is an inherent limitation of using RDP accounting. A closer privacy analysis of the Gaussian mechanism directly in terms of mutual information may be able to close this gap significantly.

7. Conclusion

We presented a rigorous formulation of data reconstruction attacks and analyzed the privacy leakage of DP mechanisms using Fano's inequality. Our analysis gives a numerical method for upper bounding the advantage of a reconstruction adversary, which we show empirically can be a strong indicator for the mechanism's actual privacy protection against privacy attacks. Consequently, we advocate for the use of our bounds both for interpreting the DP parameter ϵ and as a guideline for selecting it in practice.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Alvim, M. S., Andrés, M. E., Chatzikokolakis, K., and Palamidessi, C. On the relation between differential privacy and quantitative information flow. In *International Colloquium on Automata, Languages, and Programming*, pp. 60–76. Springer, 2011.

Arimoto, S. Information measures and capacity of order α for discrete memoryless channels. *Topics in information theory*, 1977.

Balle, B., Cherubin, G., and Hayes, J. Reconstructing training data with informed adversaries. *arXiv preprint arXiv:2201.04845*, 2022.

Bertran, M., Martinez, N., Papadaki, A., Qiu, Q., Rodrigues, M., Reeves, G., and Sapiro, G. Adversarially learned representations for information obfuscation and inference. In *International Conference on Machine Learning*, pp. 614–623. PMLR, 2019.

Bhowmick, A., Duchi, J., Freudiger, J., Kapoor, G., and Rogers, R. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, 2019.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Cuff, P. and Yu, L. Differential privacy as a mutual information constraint. In *Proceedings of the 2016 ACM SIGSAC*

Conference on Computer and Communications Security, pp. 43–54, 2016.

Dwork, C. and Naor, M. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1), 2010.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Erlingsson, Ú., Mironov, I., Raghunathan, A., and Song, S. That which we call private. *arXiv preprint arXiv:1908.03566*, 2019.

Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., and Ristenpart, T. Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In *23rd USENIX Security Symposium (USENIX Security 14)*, pp. 17–32, 2014.

Fredrikson, M., Jha, S., and Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1322–1333, 2015.

Geiping, J., Bauermeister, H., Dröge, H., and Moeller, M. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020.

Guo, C., Karrer, B., Chaudhuri, K., and van der Maaten, L. Bounding training data reconstruction in private (deep) learning. *arXiv preprint arXiv:2201.12383*, 2022.

Hannun, A., Guo, C., and van der Maaten, L. Measuring data leakage in machine-learning models with fisher information. In *Uncertainty in Artificial Intelligence*, pp. 760–770. PMLR, 2021.

Humphries, T., Rafuse, M., Tulloch, L., Oya, S., Goldberg, I., Hengartner, U., and Kerschbaum, F. Differentially private learning does not bound membership inference. *arXiv preprint arXiv:2010.12112*, 2020.

Jagielski, M., Ullman, J., and Oprea, A. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33:22205–22216, 2020.

Jayaraman, B. and Evans, D. Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 1895–1912, 2019.

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

Mahloujifar, S., Sablayrolles, A., Cormode, G., and Jha, S. Optimal membership inference bounds for adaptive composition of sampled gaussian mechanisms. *arXiv preprint arXiv:2204.06106*, 2022.

Makhdoumi, A., Salamatian, S., Fawaz, N., and Médard, M. From the information bottleneck to the privacy funnel. In *2014 IEEE Information Theory Workshop (ITW 2014)*, pp. 501–505. IEEE, 2014.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Mir, D. J. Information-theoretic foundations of differential privacy. In *Foundations and Practice of Security: 5th International Symposium, FPS 2012, Montreal, QC, Canada, October 25-26, 2012, Revised Selected Papers 5*, pp. 374–381. Springer, 2013.

Mironov, I. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Nasr, M., Shokri, R., and Houmansadr, A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 739–753. IEEE, 2019.

Nasr, M., Songi, S., Thakurta, A., Papernot, N., and Carlin, N. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 866–882. IEEE, 2021.

Nielsen, F. and Nock, R. On w-mixtures: Finite convex combinations of prescribed component distributions. *arXiv preprint arXiv:1708.00568*, 2017.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rényi, A. et al. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1. Berkeley, California, USA, 1961.

Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., and Jégou, H. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pp. 5558–5567. PMLR, 2019.

Salamatian, S., Calmon, F. P., Fawaz, N., Makhdoumi, A., and Médard, M. Privacy-utility tradeoff and privacy funnel. *Unpublished preprint, http://www. mit. edu/ salmansa/files/privacy TIFS. pdf*, 2020.

Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., and Backes, M. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

Sason, I. and Verdú, S. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Shokri, R. and Shmatikov, V. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321, 2015.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Song, C., Ristenpart, T., and Shmatikov, V. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*, pp. 587–601, 2017.

Thudi, A., Shumailov, I., Boenisch, F., and Papernot, N. Bounding membership inference. *arXiv preprint arXiv:2202.12232*, 2022.

Verdú, S. α-mutual information. In *2015 Information Theory and Applications Workshop (ITA)*, pp. 1–6. IEEE, 2015.

Wang, T., Zhang, Y., and Jia, R. Improving robustness to model inversion attacks via mutual information regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11666–11673, 2021.

Wang, W., Ying, L., and Zhang, J. On the relation between identifiability, differential privacy, and mutual-information privacy. *IEEE Transactions on Information Theory*, 62(9):5018–5029, 2016.

Watson, L., Guo, C., Cormode, G., and Sablayrolles, A. On the importance of difficulty calibration in membership inference attacks. *arXiv preprint arXiv:2111.08440*, 2021.

Ye, J., Maddi, A., Murakonda, S. K., and Shokri, R. Enhanced membership inference attacks against machine learning models. *arXiv preprint arXiv:2111.09679*, 2021.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018.

Zhu, L., Liu, Z., and Han, S. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

A. Method Details

A.1. Generalized Fano's Inequality Analysis

We derive a generalized version of the Fano's inequality analysis in section 4 using Arimoto information. We first define a few key quantities in information theory below.

Measures of information. In information theory, *entropy* measures the amount of information contained in a random variable. For a discrete random variable X, the entropy of X is defined as

$$H(X) = -\sum_x \mathbb{P}(X = x) \log \mathbb{P}(X = x). \tag{11}$$

A closely related notion is *conditional entropy* $H(X|Y)$. For two random variables X and Y, $H(X|Y)$ measures the amount of information contained in X after observing Y, and is defined as

$$H(X|Y) = \sum_y \mathbb{P}(Y = y) H(X|Y = y), \tag{12}$$

where $H(X|Y = y)$ denotes entropy of the conditional distribution $X|Y = y$. In essence, if X and Y are highly dependent then $H(X|Y) \approx 0$, which suggests that there is very little entropy left in X after observing Y. On the other hand, if X and Y are independent then $H(X|Y) = H(X)$, meaning that observing Y gives no additional information about Y. Thus, the difference

$$I(X; Y) := H(X) - H(X|Y), \tag{13}$$

known as *mutual information*, is a non-negative quantity that measures the amount of dependence between X and Y.

Generalization using Rényi entropy. The three measures of information defined above can be generalized. We start from the generalization of entropy to *Rényi entropy of order* α (Rényi et al., 1961), denoted H_α, given by:

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \sum_x \mathbb{P}(X = x)^\alpha \tag{14}$$

for $\alpha > 1$. Notably, it can be shown that this definition recovers the standard entropy $H(X)$ when $\alpha \to 1$. The corresponding generalization of conditional entropy is the so-called *Arimoto-Rényi conditional entropy* (Arimoto, 1977), defined as

$$H_\alpha(X|Y) = \frac{\alpha}{1-\alpha} \log \mathbb{P}(Y = y) \exp\left(\frac{1-\alpha}{\alpha} H_\alpha(X|Y = y)\right). \tag{15}$$

The generalization of mutual information $I_\alpha(X; Y)$, known as *Arimoto information* (Verdú, 2015), is the difference $H_\alpha(X) - H_\alpha(X|Y)$, and has the following form:

$$I_\alpha(X; Y) = \frac{\alpha}{\alpha - 1} \log \sum_y \left(\sum_x \mathbb{P}(X_\alpha = x) \mathbb{P}(Y = y|X = x)^\alpha \right)^{1/\alpha}, \tag{16}$$

where X_α denotes the α-scaled distribution with $\mathbb{P}(X_\alpha = x) \propto \mathbb{P}(X = x)^\alpha$. Similar to Rényi entropy, Arimoto-Rényi conditional entropy and Arimoto information both converge to the corresponding conditional entropy and mutual information as $\alpha \to 1$.

Generalized Fano's inequality. Fano's inequality can be generalized using an analogue of conditional entropy known as *Arimoto-Rényi conditional entropy* (Arimoto, 1977). The generalized version is given in (Sason & Verdú, 2016), which states:

$$H_\alpha(X|Y) \leq \log M - D_\alpha(\text{Bernoulli}(t) || \text{Bernoulli}(1 - 1/M)), \tag{17}$$

where Bernoulli(t) denotes the Bernoulli distribution with success probability t. It can be shown that $H_\alpha(X|Y) = H_\alpha(X) - I_\alpha(X; Y)$. $H_\alpha(X)$ can be computed directly using Equation 14, while $I_\alpha(X; Y)$ can be upper bounded using Theorem 1. Thus, we can lower bound $H_\alpha(X|Y)$ and establish an inequality for t using the generalized Fano's inequality. We can then apply the binary search method in Algorithm 1 to upper bound the advantage for a given α. Algorithm 2 details the computation in pseudo-code.

Algorithm 2 Computing the advantage bound using generalized Fano's inequality.

Input: $\alpha \geq 1$, upper bound $I_\alpha(X;Y) \leq \mu$, number of distinct attributes M, sampling probability vector $\mathbf{p}$.
$p^* \leftarrow \max_m \mathbf{p}_m$
if $\mu \geq H_\alpha(\mathbf{p})$ **then**
  return 1
else
  Define $f_\mu(t) = H_\alpha(\mathbf{p}) - \mu - \log M + D_\alpha(\text{Bernoulli}(t)||\text{Bernoulli}(1-1/M))$.
  assert $f_\mu(0) > 0$ and $f_\mu(1-1/M) \leq 0$
  Use binary search to find $t^* = \min\{t \in [0, 1-1/M] : f_\mu(t) \leq 0\}$.
  return $(1 - t^* - p^*)/(1-p^*)$
end if

A.2. Tightness of Fano's Inequality for Randomized Response

Recall the generalized randomized response mechanism in subsection 5.2. We show that the Fano's inequality analysis is tight when the input is sampled uniformly randomly. For any $Y = y$, we can derive the posterior distribution for X using Bayes rule:

$$\mathbb{P}(X = x|Y = y) = \mathbb{P}(Y = y|X = x)\mathbb{P}(X = i)/\mathbb{P}(Y = y) = \begin{cases} 1 - q + q/M & \text{if } x = y, \\ q/M & \text{otherwise.} \end{cases}$$

Suppose $q < 1$ so that $1 - q + q/M > q/M$, then the MAP adversary's estimate is $\hat{X}(Y) = Y$, with error probability $\mathbb{P}(E = 1) = q - q/M$. For any y, we can derive the conditional entropy $H(X|Y = y)$ in closed form:

$$\begin{aligned} H(X|Y = y) &= \left(-(1 - q + q/M)\log(1 - q + q/M) - \sum_{x \neq y} \frac{q}{M} \log \frac{q}{M} \right) \\ &= \left(-(1 - q + q/M)\log(1 - q + q/M) - (q - q/M)\log \frac{q(M-1)}{M(M-1)} \right) \\ &= (-(1 - q + q/M)\log(1 - q + q/M) - (q - q/M)\log(q - q/M) + (q - q/M)\log(M - 1)) \\ &= H(E) + \mathbb{P}(E = 1)\log(M - 1). \end{aligned}$$

Multiplying both sides by $\mathbb{P}(Y = y)$ and taking sum over y gives $H(X|Y) = H(E) + \mathbb{P}(E = 1)\log(M - 1)$, which attains equality in Equation 3.

A.3. Attribute Inference Attack Details

We give details for the white-box attribute inference attack in subsection 6.2. Fix sample j, and for $m = 1, 2, 3$ let $\mathbf{w}_m$ be the parameter vector of the model trained on $\mathcal{D}_{-j} \cup \{\mathbf{z}_m^{(j)}\}$. Suppose the output perturbation mechanism returns the parameter vector $\mathbf{w}$. Since the mechanism obtained $\mathbf{w}$ by perturbing one of $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$ with Gaussian noise, we can evaluate the likelihood of $\mathbf{w}$ under all three choices and maximize it to determine X:

$$\log \mathbb{P}(X = m|Y = \mathbf{w}) = -\|\mathbf{w} - \mathbf{w}_m\|_2^2/2\sigma^2 + \text{constant}.$$

This is possible since the adversary has access to $\mathcal{D}_{-j}$ and all attributes of $\mathbf{z}^{(j)}$ except for the VKORC1 gene, and hence can enumerate all possible values for $\mathbf{z}_i^{(j)}$ to obtain $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$. To derive the MAP estimator, we apply Bayes' rule:

$$\begin{aligned} \mathbb{P}(X = m|Y = \mathbf{w}) &\propto \mathbb{P}(Y = \mathbf{w}|X = m)\mathbb{P}(X = m) = \mathbb{P}(Y = \mathbf{w}|X = m)\mathbf{p}_m \\ \Rightarrow \log \mathbb{P}(X = m|Y = \mathbf{w}) &= -\|\mathbf{w} - \mathbf{w}_m\|_2^2/2\sigma^2 + \log \mathbf{p}_m + \text{constant}, \end{aligned}$$

where $\mathbf{p} = (0.367, 0.339, 0.294)$ is the marginal distribution of the VKORC1 gene.

B. Proofs

Theorem 1. *For any $\alpha \geq 1$, if $\mathcal{M}$ is (α, ϵ)-RDP then $I_\alpha(X;Y) \leq \epsilon$.*

Proof. We first prove this for $\alpha = 1$. By definition:

$$\begin{aligned} I(X;Y) &= \sum_x \sum_y \mathbb{P}(X=x, Y=y) \log \frac{\mathbb{P}(X=x, Y=y)}{\mathbb{P}(X=x)\mathbb{P}(Y=y)} \\ &= \sum_x \mathbb{P}(X=x) \sum_y \mathbb{P}(Y=y|X=x) \log \frac{\mathbb{P}(Y=y|X=x)}{\mathbb{P}(Y=y)} \\ &= \sum_x \mathbb{P}(X=x) D_1((Y|X=x)||Y), \end{aligned} \tag{18}$$

where D_1 denotes KL divergence, which is equivalent to the order $\alpha = 1$ Rényi divergence. By convexity,

$$\begin{aligned} D_1((Y|X=x)||Y) &\leq \sum_{x'} \mathbb{P}(X=x') D_1((Y|X=x)||(Y|X=x')) \\ &\leq \sum_{x'} \mathbb{P}(X=x')\epsilon \qquad \text{since } \mathcal{M} \text{ is } (1,\epsilon)\text{-RDP} \\ &= \epsilon. \end{aligned}$$

Plugging back into Equation 18 gives the desired result. For $\alpha > 1$, let X_α denote the α-scaled distribution in the definition of Arimoto information, *i.e.*, $\mathbb{P}(X_\alpha = x) \propto \mathbb{P}(X=x)^\alpha$. Then:

$$\begin{aligned} I_\alpha(X;Y) &= \frac{\alpha}{\alpha-1} \log \sum_y \left(\sum_x \mathbb{P}(X_\alpha = x)\mathbb{P}(Y=y|X=x)^\alpha \right)^{1/\alpha} \\ &= \frac{\alpha}{\alpha-1} \log \sum_y \mathbb{P}(Y=y) \left(\sum_x \mathbb{P}(X_\alpha = x) \frac{\mathbb{P}(Y=y|X=x)^\alpha}{\mathbb{P}(Y=y)^\alpha} \right)^{1/\alpha} \\ &\leq \frac{1}{\alpha-1} \log \sum_y \mathbb{P}(Y=y) \left(\sum_x \mathbb{P}(X_\alpha = x) \frac{\mathbb{P}(Y=y|X=x)^\alpha}{\mathbb{P}(Y=y)^\alpha} \right) \quad \text{by convexity of } z \mapsto z^\alpha \\ &\leq \max_x \frac{1}{\alpha-1} \log \left(\sum_y \mathbb{P}(Y=y) \frac{\mathbb{P}(Y=y|X=x)^\alpha}{\mathbb{P}(Y=y)^\alpha} \right) \quad \text{by monotonicity of } z \mapsto \frac{1}{\alpha-1}\log z \\ &= \max_x D_\alpha((Y|X=x)||Y). \end{aligned}$$

The proof follows by quasi-convexity of Rényi divergence and the fact that $\mathcal{M}$ is (α, ϵ)-RDP. □

Mutual information bound for Gaussian mechanism. The Gaussian mechanism satisfies (α, ϵ)-RDP with $\epsilon = \alpha\Delta^2/2\sigma^2$ when the encoding function has L_2-sensitivity Δ (Mironov, 2017). This readily gives a bound for $I(X;Y)$ using Theorem 1 by setting $\alpha = 1$. Below we present a tighter bound for the mutual information of Gaussian mechanisms.

Theorem 2. *Let $\mathbf{e}_m$ be the encoding for the private record $\mathbf{z}_m$ for $m = 1, \ldots, M$, and let $\Delta = \max_{m,l} \|\mathbf{e}_m - \mathbf{e}_l\|_2$ be the L_2-sensitivity of the encoding function. Then the Gaussian mechanism (*cf. *Equation 10) satisfies:*

$$I(X;Y) \leq -\sum_m \mathbf{p}_m \log \left(\mathbf{p}_m + (1-\mathbf{p}_m) \exp\left(\frac{-\Delta^2}{2\sigma^2} \right) \right). \tag{19}$$

Proof. First note that $I(X;Y) = H(Y) - H(Y|X)$ with $H(Y|X) = \frac{M}{2}\log(2\pi e\sigma^2)$. By Section 3.1 of (Nielsen & Nock, 2017), because Y is a mixture of Gaussians we have the following bound for $H(Y)$:

$$\begin{aligned} H(Y) &\leq \sum_m \mathbf{p}_m H\Big(\mathcal{N}(\mathbf{e}_m, \sigma^2\mathbf{I}_{dxd})\Big) - \sum_m \mathbf{p}_m \log \left(\sum_l \mathbf{p}_l \exp\left(-D_1\Big(\mathcal{N}(\mathbf{e}_m, \sigma^2\mathbf{I}_{dxd}), \mathcal{N}(\mathbf{e}_l, \sigma^2\mathbf{I}_{dxd})\Big)\right)\right) \\ &\leq \frac{M}{2}\log\left(2\pi e\sigma^2\right) - \sum_m \mathbf{p}_m \log\left(\mathbf{p}_m + (1-\mathbf{p}_m)\exp\left(\frac{-\Delta^2}{2\sigma^2}\right)\right). \end{aligned}$$

Combining the two quantities gives the desired result. □