Optimal Membership Inference Bounds for Adaptive Composition of Sampled Gaussian Mechanisms

Anonymous Author(s)
Affiliation
Address
email

Abstract

Given a trained model and a data sample, membership-inference (MI) attacks predict whether the sample was in the model's training set. A common countermeasure against MI attacks is to utilize differential privacy (DP) during model training to mask the presence of individual examples. While this use of DP is a principled approach to limit the efficacy of MI attacks, there is a gap between the bounds provided by DP and the empirical performance of MI attacks. In this paper, we derive bounds for the *advantage* of an adversary mounting a MI attack, and demonstrate tightness for the widely-used Gaussian mechanism. Our bounds are much stronger than those obtained by DP analysis. Our analysis shows that although differential privacy guarantees might sometimes lead to large/vacuous membership inference guarantees, mechanisms that obtain differential privacy can be in fact much more secure against membership inference attacks. Previously, this phenomenon was observed for DP-SGD and here, for the first time, we prove it. Finally, using our analysis, we provide MI metrics for models trained on CIFAR10 dataset. To the best of our knowledge, our analysis provides the state-of-the-art membership inference bounds.

1 Introduction

The recent success of machine learning models makes them the go-to approach to solve a variety of problems, ranging from computer vision [Krizhevsky et al., 2012] to NLP [Sutskever et al., 2014], including applications to sensitive data such as health records or chatbots. Access to a trained machine learning model, through a black-box API or a white-box access to a published model, can leak traces of information [Dwork et al., 2015] from the training data. Researchers have tried to measure this information leakage through metrics such as membership inference [Shokri et al., 2017]. Membership inference is the task of guessing, from a trained model, whether it includes a given sample or not. This task is interesting in its own right, as the participation of an individual in a data collection can be sensitive information. Furthermore, it also serves as the "most significant bit" of information: if membership inference fails, attacks revealing more information such as reconstruction attacks [Fredrikson et al., 2014, Carlini et al., 2020] will also fail. In other words, defending against membership inference attacks also defends against more advanced attacks.

The standard approach to provably defeat these membership privacy attacks is differential privacy [Dwork et al., 2006]. The traditional variant of differential privacy defines a class of algorithms that apply on a database D and respect a privacy budget ϵ and a probability of failure δ. These parameters control the divergence between the distribution of outcome of the algorithm when applied on two "neighboring" databases that only differ in one location. In particular, our mechanism of interest, the Sampled Gaussian Mechanism (SGM), makes an arbitrary function $f : X^* \to R^d$ differentially private by first sub-sampling a database of smaller size from the original database,

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

using Poisson sampling, and then calculating f on the sub-sampled database, and finally adding Gaussian noise to the outcome. Increasing the amount of injected noise will provide stronger privacy guarantees but it often lead to a trade-off between privacy utility of the trained model. To measure the privacy of a given algorithm, researchers have developed advanced mathematical tools and notions such as Renyi differential privacy [Mironov, 2017, Abadi et al., 2016] and advanced composition theorems [Dwork et al., 2010, Kairouz et al., 2015]. These tools allow us to calculate (ϵ, δ) values for carefully designed algorithms.

Previous work has shown that any differentially private algorithm will provably bound the accuracy of any membership inference adversary. Specifically, starting with a given (ϵ, δ), Humphries et al. [2020] prove that any model trained with (ϵ, δ) differential privacy will induce an upper bound on the accuracy of membership adversary and this upper bound depends only on ϵ and δ. These upper bounds enable us to obtain provable defenses against membership inference attacks by using differentially private algorithms. In fact, there is a large gap between upper bounds proved for the power of adversaries in performing membership inference, and the power of real adversaries that try to attack differentially private models. In this work, we develop a new proof technique to bypass DP analysis and directly bound the advantage of MI attacks.

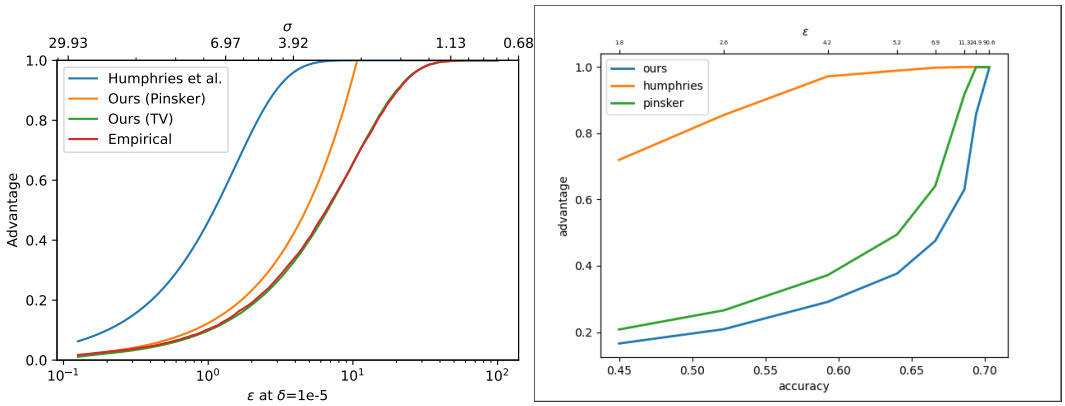
The most widely-used DP algorithm in machine learning applications is DP-SGD: it is a small change of the classical stochastic gradient descent algorithm that only requires to clip per-sample gradients, average them and add Gaussian noise. DP-SGD has been shown to be more accurate than other differentially private training algorithms in the case of linear and convex models [van der Maaten and Hannun, 2020]. Each iteration of DP-SGD is an instance of the sampled Gaussian mechanism, which chooses a fraction q of a dataset and outputs a noisy sum of the desired quantity. Our analysis mirrors the recent shift in the field of empirical membership inference, from advantage (or accuracy) metrics [Shokri et al., 2017, Yeom et al., 2018, Sablayrolles et al., 2019] to precision/recall measures [Watson et al., 2022, Carlini et al., 2020]. Differential privacy guarantees are known to yield tight true positive and false positive rates [Nasr et al., 2021]. Our paper is the first to show equivalent results in the case of advantage. Our analysis explains, from a theoretical point of view, why the precision and recall of membership inference attacks is stronger than the accuracy.

Our contributions in this work are as follows:

- Our main theorem in Section 4 bounds the membership inference advantage of any adversary for the composition of an arbitrary set of sampled Gaussian mechanisms that could adaptively depend on each other. Specifically, for any adaptive series of sampled Gaussian mechanism $(M_1, \ldots, M_T)$ where M_i has sensitivity 1.0 and standard deviation σ_i and sub-sampling rate q_i, we show that the membership inference advantage of any adversary is bounded by the total variation distance between the mixture of two Gaussians defined by $\sigma = (\sigma_1, \ldots, \sigma_T)$ and $q = (q_1, \ldots, q_T)$. This bound is optimal as it reflects the membership inference advantage for a real adversary on a particular series of Gaussian mechanisms.
- We propose a numerical way to calculate the total variation distance between mixtures of Gaussians (Section 4.3). Our algorithm is computationally tractable and works in linear time with respect to number of mechanisms and is independent from the dimension. We use our numerical approach to obtain concrete bounds on membership inference for Gaussian mechanisms and compare our bounds to that of Humphries et al. [2020].
- Finally, to understand the practical implication of our bound for mainstream datasets, we use DP-SGD to train models on CIFAR10 and MNIST and calculate the membership inference bounds using our techniques. Our approach allows us to achieve state-of-the-art provable membership inference privacy for any given accuracy.

2 Background

In this section, we briefly recall the main properties of KL divergence and total variation (TV). We consider a random variable X (resp. Y) with probability distribution μ (resp. ν) over a measureable space Ω. With a slight abuse of notation, we consider quantities as function of either the random variables X, Y or their probability distributions μ, ν.



(a) Bounds on membership advantage. (b) Experiments on CIFAR-10.

Figure 1: Empirical results on membership advantage

2.1 Total variation, KL divergence, and Pinsker inequality

Total variation distance (TV) $\mathbf{TV}(X, Y) = \sup_{\mathcal{A}} \Pr[X \in \mathcal{A}] - \Pr[Y \in \mathcal{A}]$.

TV properties. We recall briefly some properties of TV that will be useful in the remainder of this paper. Given a function $g : E \to F$, applying g to X and Y can only decrease $\mathbf{TV}$:

$$\mathbf{TV}(g(X), g(Y)) \leq \mathbf{TV}(X, Y).$$

In particular if g is a bijective transform, we have $\mathbf{TV}(g(X), g(Y)) = \mathbf{TV}(X, Y)$.

Kullback-Leibler (KL) divergence. The KL divergence between X and Y is defined as $\mathrm{KL}(X \parallel Y) = \int_{\Omega} \log\left(\frac{d\mu}{d\nu}\right) d\mu$.

Now we are ready to state Pinsker's inequality that connects total variation to KL divergence:

Pinsker's Inequality. For any two probability distributions X and Y we have

$$\mathbf{TV}(X, Y) \leq \sqrt{\frac{\mathrm{KL}(X \parallel Y)}{2}}. \tag{1}$$

2.2 Differential Privacy

Differential privacy. A randomized algorithm $\mathcal{M}$ satisfies (ϵ, δ)-differential privacy if, for any two datasets D and D' that differ in at most one sample, and any subset R of the output space, we have:

$$\mathbb{P}(\mathcal{M}(D) \in R) \leq \exp(\epsilon)\mathbb{P}(\mathcal{M}(D') \in R) + \delta. \tag{2}$$

While the probability of failure δ is often chosen to be inversely proportional to the number of samples, there is no consensus in the literature over desirable values of ϵ.

DP-SGD with the subsampled Gaussian mechanism. DP-SGD [Abadi et al., 2016] is a modification of Stochastic Gradient Descent that makes it differentially private. The core mechanism behind it is the subsampled Gaussian mechanism. Given a function ϕ that operates on sets S with sensitivity 1 (i.e. $\|\phi(S) - \phi(S')\|_2 \leq 1$), the subsampled Gaussian mechanism selects sets S at random and adds noise to the output of ϕ. Note that the mechanism is differentially private *even* if all intermediate steps of the training process are revealed.

2.3 Membership inference attacks

Membership Inference is the task of predicting whether a given sample was in the training set of a given model. Homer et al. [2008] showed the first proof of concept, and Shokri et al. [2017]

showed that a wide variety of machine learning models are vulnerable to such attacks. Shokri et al. [2017] train neural networks to attack machine learning models, and measure the success of the attack by the percentage of correctly predicted (train/test) samples, or equivalently the advantage [Yeom et al., 2018]. While Shokri et al. [2017] trained neural networks to attack machine learning models, it was shown later that simple heuristics such as the loss [Yeom et al., 2018, Sablayrolles et al., 2019] provide a more accurate and robust measure of membership inference.

Recent works [Watson et al., 2022, Carlini et al., 2021, Rezaei and Liu, 2021] have proposed to evaluate membership inference by the precision/recall trade-off [Watson et al., 2022] or the precision at low levels of recall [Carlini et al., 2021]. In particular, such works show that some setups which were thought to be private because the membership accuracy is significantly less than 100% can actually reveal membership of a small group of samples with very high precision.

There is also a line of work developed to design algorithms to specifically defend against membership inference attacks. Although differential privacy would bound the membership inference, these empirical defenses are potentially able to achieve better utility while withstanding existing membership inference attacks.

3 Membership inference and total variation

In this section, we define our security game for membership inference and show its connection to the notion of total variation distance (or statistical distance) between probability distributions.

3.1 Security Game

We adopt the classical assumptions of membership inference [Yeom et al., 2018, Sablayrolles et al., 2019, Humphries et al., 2020]. In particular, for the remainder of this paper, we assume that the datasets differ by adding/removing one element. Saeed: Talk about federated learning and "transcript" + fact that the learner is adversarial

1. Adversary picks a datasets $D = \{z_1, \ldots, z_n\}$ and a data point z'
2. Challenger samples a bit b uniformly at random and creates

$$D' = \begin{cases} D \cup \{z'\} & \text{if } b = 1 \\ D & \text{if } b = 0 \end{cases}$$

3. Challenger runs the SGM on D' for T steps with transcript $s = (s_1, \ldots, s_T)$
4. Adversary observes the transcript s and guesses a bit b'. Adversary wins if $b' = b$.

We define the advantage of adversary A on learning algorithm L as $\mathbf{Adv}(L, A) = 2 \cdot \Pr[b = b'] - 1$. We also use $\mathbf{Adv}(L) = \sup_A \mathbf{Adv}(A, L)$ to denote the advantage of any adversary against L.

Remark 1. *In Section 4 we prove bounds on the membership advantage for composition of Gaussian mechanisms with and without sub-sampling. Both of these bounds are tight for the advantage defined based on addition/removal. However, for the notion of advantage defined based on replacement, if we double the bounds, only the bound without replacement will remain tight. We leave the question of obtaining tight bounds on the advantage based on replacement for the sub-sampled Gaussian as an open question.*

Let $X := s \mid (b = 0)$ and $Y := s \mid (b = 1)$. A deterministic adversary A defines a region $\mathcal{A}$ and predicts that θ is sampled from μ if $s \in \mathcal{A}$ and from Y if $s \notin \mathcal{A}$. We use $\mu(\mathcal{A})$ and $\nu(\mathcal{A})$ to denote $\Pr[X \in \mathcal{A}]$ and $\Pr[Y \in \mathcal{A}]$. For such an adversary we have

$$\mathbf{Adv}(L, A) = |\Pr[X \in \mathcal{A}] - \Pr[Y \in \mathcal{A}]|.$$

Note that with a simple averaging argument we can show that the best adversarial strategy in membership security game is a deterministic strategy. Therefore, the advantage for the learning algorithm L is then defined as

$$\mathbf{Adv}(L) = \sup_{\mathcal{A}} \mu(\mathcal{A}) - \nu(\mathcal{A}) = \mathbf{TV}(X, Y), \tag{3}$$

where $\mathbf{TV}$ is the total variation distance. Therefore, total variation distance gives us an upper bound on the advantage of any adversary. However, it is not clear how to calculate the total variation distance in general. In next subsection we discuss an approximation of total variation distance that can actually be calculated using existing techniques for RDP accounting.

3.2 Bounding membership inference using Pinsker's inequality

Using Equation 3 and directly applying Pinsker's inequality (1), we have:

$$\mathbf{Adv}(L) \leq \mathbf{TV}(X, Y) \leq \sqrt{\frac{\mathrm{KL}(X \parallel Y)}{2}} = \lim_{\alpha \to 1} \sqrt{\frac{D_\alpha(X \parallel Y)}{2}} \tag{4}$$

where D_α is the Renyi divergence at α. Now one might ask why this bound is preferable to our bound using total variation distance. The reason we state this bound is that we have techniques for calculating the Renyi divergence of composition of adaptive and sampled Gaussian mechanisms. This enables us to calculate numerical upper bounds on the membership inference advantage of any adversary against adaptive composition of sampled Gaussian mechanisms, e.g. DP-SGD.

To this end, we use RDP accounting to calculate the D_α for an $\alpha > 1$. We know that for any $\alpha > 1$, $D_\alpha(X, Y)$ is greater than $\mathrm{KL}(X \parallel Y)$ because D_α is increasing in α. This means that for any $\alpha > 1$ D_α will be a valid upper bound on the membership inference advantage and the bound becomes better as we decrease α.

In Figures 1(a) and 1(b) we calculate numerical upper bounds for membership inference for DP-SGD using typical parameters. We refer to this bound as the Pinsker bound. The figure shows that this bound obtains better numerical values compared to the bound of Humphries et al. [2020]. However, there are two main limitations with our Pinsker bound: 1) The bound is not tight. We are applying Pinsker's inequality which is not optimal for the Gaussian mechanism. 2) It provides vacuous bounds in cases where $\mathrm{KL}(X \parallel Y) > 2$. In next section, we will optimally bound the membership inference for composition of adaptive and sampled gaussian mechanisms.

4 Membership inference bounds for composition of Gaussian mechanisms

In this section, we show a tighter upper-bound for the adversary's advantage. Our main results are Theorem 4 and Theorem 5. In particular, we will upper-bound the total variation between the *transcripts* of the (subsampled) Gaussian mechanism, specifically the noisy gradients produced by the DP-SGD algorithm. The upper bound on the result of the DP-SGD algorithm follows by application of post-processing.

The proof technique is the following: we show that the entire process of DP-SGD can be replaced by a process where each step is replaced by either $\mathcal{N}(0, \sigma^2)$ or $\mathcal{N}(r, \sigma^2)$ for the basic Gaussian mechanism, and replaced by $(1-q)\mathcal{N}(0, \sigma^2) + q\mathcal{N}(r, \sigma^2)$ for the subsampled Gaussian mechanism. Our proof technique relies on a modified version of $\mathbf{TV}$, called $\mathbf{TV}_a$.

Definition 2. *For $a > 0$ define $\mathbf{TV}_a(X, Y) = \frac{1}{2} \int_\Omega |P(x) - a \cdot Y(x)| dx$.*

While we do not know an explicit form of the $\mathbf{TV}_a$ divergence between Gaussians, we can still prove that it increases with $\|u_1 - u_2\|_2$, as formalized in the following lemma.

Lemma 3. *Let $X \equiv \mathcal{N}(u_1, \sigma \cdot I_d)$ and $Y \equiv \mathcal{N}(u_2, \sigma \cdot I_d)$. Then, for any $a \in \mathbb{R}^+$, $\mathbf{TV}_a(X, Y)$ is only a function of $\|u_1 - u_2\|_2$ and σ. Moreover this function is monotonically increasing with respect to $\|u_1 - u_2\|_2$. That is,*

$$\mathbf{TV}_a(X, Y) \leq \mathbf{TV}_a(\mathcal{N}(0, \sigma), \mathcal{N}(\|u_1 - u_2\|_2, \sigma)).$$

The proof is deferred to Appendix A.

4.1 Warm up: Without sampling

Notation. We use $s = (s_1, \ldots, s_T)$ to denote the output (or *transcript*) of a random process that consists of T adaptive steps (typically the subsampled Gaussian mechanism). We use $s_{\leq t}$ to denote the first t steps of the transcript of the random process. The sampling rate q is the probability of

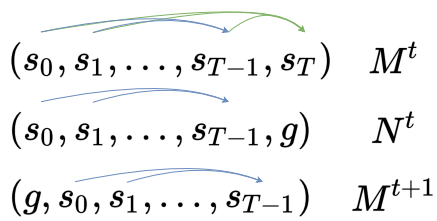


Figure 2: Main component of the proof of Theorem 4. The last step of DP-SGD is replaced by Gaussian noise (independent of previous steps). This step is then put at the start of the transcript.

including any sample in the random set S. We use S_i to denote the union of the support set of the ith step of the mechanism M on all possible datasets. That is $S_i = \{s_i; \exists D, s_i \in \mathsf{Supp}(M(D)_i)\}$. We also define $S_{\leq i} = S_1 \times \ldots, S_i$.

Theorem 4 (Gaussian Composition without sub-sampling)**.** *Let $M_1, \ldots, M_T$ be a series of adaptive Gaussian Mechanisms with L_2 sensitivity r and Gaussian noise with standard deviation σ. The membership inference risk of the composition of M_i's is at most as much as a single Gaussian mechanism with sensitivity $\sqrt{T} \cdot r$ and standard deviation σ.*

$$\mathbf{TV}(X, Y) \leq \mathbf{TV}(\mathcal{N}(0, \sigma I_T), \mathcal{N}(r1_T, \sigma I_T)) \tag{5}$$

Proof. Let $M^i(D)$ be the mechanism that works on D (resp. D') and consists of i Gaussians $\mathcal{N}(0, \sigma^2)$ (resp. $\mathcal{N}(r, \sigma^2)$) steps followed by $T - i$ steps of DP-SGD applied to D (resp. D'). Specifically, the mechanism M^0 corresponds to DP-SGD, and M^T to pure Gaussians. We will show in this proof that

$$\mathbf{TV}(M^i(D), M^i(D')) \leq \mathbf{TV}(M^{i+1}(D), M^{i+1}(D')) \tag{6}$$

and hence

$$\mathbf{TV}(X, Y) = \mathbf{TV}(M^0(D), M^0(D')) \leq \mathbf{TV}(M^T(D), M^T(D')) = \mathbf{TV}(\mathcal{N}(0, \sigma^2 I_T), \mathcal{N}(r1_T, \sigma^2 I_T).$$

To this end, we will first argue that the very final step of the mechanism M^i can be replaced with a Gaussian step without increasing the total variation distance. Then we can move this noise to the start of the process without affecting the result, obtaining M^{i+1}. Figure 2 illustrates this procedure.

Let us fix a step i and let $s = (s_1, \ldots, s_T)$ be a transcript. Denoting $X \sim M^i(D)$ and $Y \sim M^i(D')$, we have

$$\begin{aligned}
2\mathbf{TV}(M^i(D), M^i(D')) &= \sum_{s_{\leq T} \in S_{\leq T}} |\Pr[X = s_{\leq T}] - \Pr[Y = s_{\leq T}]| \\
&= \sum_{s_{\leq T-1} \in S_{\leq T-1}} \sum_{s_T \in S_T} |\Pr[X = s_{\leq T}] - \Pr[Y = s_{\leq T}]| \\
&= \sum_{s_{\leq T-1} \in S_{\leq T-1}} \sum_{s_T \in S_T} |\Pr[X_T = s_T \mid s_{\leq T-1}] \cdot \Pr[X_{\leq T-1} = s_{\leq T-1}] \\
&\qquad - \Pr[Y_T = s_T \mid s_{\leq T-1}] \Pr[Y_{\leq T-1} = s_{\leq T-1}]| \,. \\
&= \sum_{s_{\leq T-1} \in S_{\leq T-1}} \Pr[X_{\leq T-1} = s_{\leq T-1}] \sum_{s_T \in S_T} \Big| \Pr[X_T = s_T \mid s_{\leq T-1}] \\
&\qquad - \Pr[Y_T = s_T \mid s_{\leq T-1}] \frac{\Pr[Y_{\leq T-1} = s_{\leq T-1}]}{\Pr[X_{\leq T-1} = s_{\leq T-1}]} \Big| .
\end{aligned} \tag{7}$$

We know that the last (T'th) step of $M^i(D)$ (resp. $M^i(D')$) has two isotropic Gaussian distributions centered around two points u_1 and u_2 such that $\|u_1 - u_2\|_2 \leq r$ and with standard deviation σ. These centers could be chosen adaptively according to the history of the mechanism. We use $a(s_{\leq T-1})$ to denote $\frac{\Pr[Y_{\leq T-1} = s_{\leq T-1}]}{\Pr[X_{\leq T-1} = s_{\leq T-1}]}$. By Lemma 3 we have

$$\begin{aligned}
&\sum_{s_T \in S_T} \Big| \Pr[X_T = s_T \mid s_{\leq T-1}] - \Pr[Y_T = s_T \mid s_{\leq T-1}] a(s_{\leq T-1}) \Big| \\
&\qquad\qquad \leq \; 2\mathbf{TV}_{a(s_{\leq T-1})}(\mathcal{N}(0, \sigma^2), \mathcal{N}(r, \sigma^2)).
\end{aligned}$$

Denoting by N^i the mechanism that coincides with M^i for the first $T-1$ steps and is replaced by a Gaussian at the last (T'th) step, we thus have, by following Equation (7) in the reverse direction

$$2\mathbf{TV}(M^i(D), M^i(D')) \leq 2\mathbf{TV}(N^i(D), N^i(D')). \tag{8}$$

Given that the last step of N^i does not depend on the first $T-1$ steps, we can permute to put it in first position (see Figure 2), which shows that $\mathbf{TV}(N^i(D), N^i(D')) = \mathbf{TV}(M^{i+1}(D), M^{i+1}(D'))$. □

4.2 With sampling

Notation. We use $B(q)$ to denote a Bernoulli random variable that is equal to 1 with probability q and 0 with probability $1-q$. We use $B(q)^n$ to denote an n dimensional random variable where each coordinate is independent and distributed as $B(q)$. For $r \in \mathbb{R}$, we use $r \cdot B(q)^n$ to denote the random variable that is sampled by the following process: first sample from $B(q)^n$ and then multiply by r. We also use $\mathcal{N}(r \cdot B(q)^n, \sigma)$ to denote an n dimensional random variable that is distributed according to a mixture of Gaussians which all have standard deviation σ and centers are chosen at random from $r \cdot B(q)^n$.

Theorem 5 (Gaussian Composition with sub-sampling)**.** *Let $M_1, \ldots, M_T$ be a series of adaptive Gaussian Mechanisms with L_2 sensitivity r and Gaussian noise with standard deviation σ and sub-sampling rate q. The membership inference risk of the composition of M_i's is at most*

$$\mathbf{TV}\big(\mathcal{N}(0^T, \sigma), \mathcal{N}(r \cdot B(q)^T, \sigma)\big)$$

Lemma 6. *Let $X' \equiv (1-q) \cdot Y + q \cdot X$. Then we have*

$$\mathbf{TV}_a(X', Y) = q\mathbf{TV}_{\frac{a+q-1}{q}}(X, Y)$$

Proof. We have

$$\begin{aligned}
2\mathbf{TV}_a(X', Y) = \int_\Omega |d\mu'(x) - a d\nu(x)| &= \int_\Omega |q d\mu - (q+a-1) d\nu| \\
&= q \int_\Omega \left| d\mu - \frac{(q+a-1)}{q} d\nu \right| \\
&= 2q\mathbf{TV}_{\frac{a+q-1}{q}}(X, Y).
\end{aligned}$$

□

Proof of Theorem 5. The proof steps are similar to Theorem 4. First, we have

$$\begin{aligned}
2\mathbf{TV}(X, Y) = &\sum_{s_{\leq T-1} \in S_{\leq T-1}} \Pr[X_{\leq T-1} = s_{\leq T-1}] \cdot \\
&\left(\sum_{s_T} \left| \Pr[X_T = s_T \mid s_{\leq T-1}] - \Pr[Y_T = s_T \mid s_{\leq T-1}] \frac{\Pr[Y_{\leq T-1} = s_{\leq T-1}]}{\Pr[X_{\leq T-1} = s_{\leq T-1}]} \right| \right) \\
= 2 &\sum_{s_{\leq T-1} \in S_{\leq T-1}} \Pr[X_{\leq T-1} = s_{\leq T-1}] \mathbf{TV}_{a(s_{\leq T-1})}(X_T \mid s_{\leq T-1}, Y_T \mid s_{\leq T-1}).
\end{aligned}$$

But since X_T and Y_T are subsampled Gaussian mechanisms we have $X_T \equiv (1-q)Y_T + qX'_T$ where Y and X' are mixtures of Gaussians. Therefore, by Lemma 3 and Lemma 6 we have

$$\begin{aligned}
&\mathbf{TV}(X, Y) \\
&= \sum_{s_{\leq T-1} \in S_{\leq T-1}} \Pr[X_{\leq T-1} = s_{\leq T-1} \in S_{\leq T-1}] q\mathbf{TV}_{\frac{a(s_{\leq T-1})+q-1}{q}}(X'_T \mid s_{\leq T-1}, Y_T \mid s_{\leq T-1}) \quad \text{(By Lemma 6)} \\
&\leq \sum_{s_{\leq T-1} \in S_{\leq T-1}} \Pr[X_{\leq T-1} = s_{\leq T-1} \in S_{\leq T-1}] q\mathbf{TV}_{\frac{a(s_{\leq T-1})+q-1}{q}}(\mathcal{N}(0, \sigma), \mathcal{N}(r, \sigma)) \quad \text{(By Lemma 3)} \\
&= \sum_{s_{\leq T-1} \in S_{\leq T-1}} \Pr[X_{\leq T-1} = s_{\leq T-1}] \mathbf{TV}_{a(s_{\leq T-1})}(\mathcal{N}(0, \sigma), (1-q)\mathcal{N}(0, \sigma) + q\mathcal{N}(r, \sigma)) \quad \text{(By Lemma 6)} \\
&= \sum_{s_{\leq T-1} \in S_{\leq T-1}} \Pr[X_{\leq T-1} = s_{\leq T-1}] \mathbf{TV}_{a(s_{\leq T-1})}(\mathcal{N}(0, \sigma), \mathcal{N}(r \cdot B(q), \sigma)).
\end{aligned}$$

Therefore, we can replace X_T with a mixture of two Gaussians centered at 0 and r and Y_T with a single Gaussian centered at 0. Now we can use the same technique used in proof of Theorem 4 and move X_T and Y_T to the first round and repeat this process. At the end, Y is replaced by a n-dimensional Gaussian centered at 0 and standard deviation σ, and X by a mixture of Gaussians with center randomly selected according to a n-dimensional Bernoulli distribution with probability q. That is, the advantage is bounded by

$$\mathbf{TV}(\mathcal{N}(0^n, \sigma), \mathcal{N}(rB(q)^n, \sigma))$$

□

Remark 7. *Theorem 4 could be simply extended to composition of Gaussian mechanisms with varying noise levels. However, if the noise levels are selected adaptively, the proof is not clear. We leave the composition of Gaussians with adaptive noise selection as an open question.*

Remark 8. *Although Theorem 4 is stated only for Gaussian mechanism, the Theorem extends to any mechanism that satisfies monotonicity under $\mathbf{TV}_a$ according to some notion of sensitivity. For example, if one can show that $\mathbf{TV}_a(\mathcal{L}(0, \sigma), \mathcal{L}(u, \sigma))$ is monotonically increasing with respect to $|u|_1$, then Theorem 4 extends to composition of Laplace mechanisms with bounded ℓ_1 sensitivity.*

4.3 Numerical computation

In order to numerically approximate the upper-bound, we first convert the notion of $\mathbf{TV}$ into an expectation formulation as follows:

$$\mathbf{TV}(X, Y) = \int_\Omega (d\mu(t) - d\nu(t))\, \mathbb{1}\,(d\nu(t) \leq d\mu(t)) \tag{9}$$

$$= \int_\Omega \left(1 - \frac{d\nu(t)}{d\mu(t)}\right) \mathbb{1}\,(d\nu(t) \leq d\mu(t))\, d\mu(t) \tag{10}$$

$$= \mathbb{E}_{t\sim X}\left(\left(1 - \frac{d\nu(t)}{d\mu(t)}\right) \mathbb{1}\,(d\nu(t) \leq d\mu(t))\right) \tag{11}$$

Note that this expectation is over distribution X. So we can sample a dataset from X and approximate this expectation using empirical averaging (or Monte-Carlo sampling):

$$\mathbf{TV}(X, Y) \approx \frac{1}{m}\sum_{i=1}^{m}\left(1 - \frac{d\nu(t_i)}{d\mu(t_i)}\right) \mathbb{1}\,(d\nu(t_i) \leq d\mu(t_i)) \tag{12}$$

Where $t_i \sim_{\text{i.i.d.}} \mu$. We know that Monte-Carlo estimation of this quantity can be very accurate, because it is bounded between 0 and 1. Note that in-order to calculate this, we need to calculate $Y(t_i)/X(t_i)$ and that is possible because we have the mathematical form of the probability distribution function of for Y and X. In Figures 1(a) and 1(b) we calculate the upper bound using this Monte-Carlo simulation for typical settings in DP-SGD.

Proposition 9. *Algorithm 1 return a value η that with high probability is close to the optimal membership inference bound η^* of Theorem 4:*

$$\Pr[|\eta - \eta^*| \geq \delta] \leq e^{-S\cdot\delta^2/3}.$$

5 Evaluation

In this section we perform experiments to calculate concrete trade-off between accuracy and membership inference on mainstream datasets.

6 Conclusion

In this paper, we directly analyzed membership inference bounds for adaptive composition of sampled Gaussian mechanisms. Our analysis enables us to obtain bounds that are better than bounds from prior analyses of differential privacy. Our analysis shows that although differential privacy guarantees might sometimes lead to large/vacuous membership inference guarantees, the mechanisms that obtain

Algorithm 1 Optimal MI

Require: sample rate vector q,
number of iterations T,
learning rate vector η,
noise scale vector σ,
gradient norm clip vector r,
sample size S

for all $j \in [S]$ **do**
  $\eta_j \leftarrow 0$
for $i \in [T]$ **do**
  for $j \in [S]$ **do**
    sample $t_{i,j} \sim \mathcal{N}(0, \sigma[i]^2 \cdot r[i]^2 \cdot \mathbb{I})$
    $\eta_j \leftarrow \eta_j + \ln\left(1 + q[i](e^{\frac{2t_{i,j}/r[i]-1}{2\sigma[i]^2}} - 1)\right)$
  $\eta \leftarrow 0$
for $j \in [S]$ **do**
  $\eta_j = \max(\eta_j, 0)$.
  $\eta \leftarrow \eta + \frac{1-e^{\eta_j}}{S}$
return η

Algorithm 2 DP-SGD Abadi et al. [2016]

Require: training dataset D, sample rate vector q,
number of iterations T, learning rate vector η,
noise scale vector σ, gradient norm clip vector
r, loss function L

Initiate θ randomly
for $i \in \{T\}$ **do**
  $B_i \leftarrow$ Sample batch via Poisson sampling
  with rate $q[i]$
  $\nabla[t] \leftarrow \vec{0}$
  for all $(x, y) \in B_t$ **do**
    $\nabla^{(x,y)} \leftarrow$ gradient of $L(x, y)$
    $\overline{\nabla^{(x,y)}} \leftarrow r[t] \cdot \frac{(\nabla^{(x,y)})}{\max(r[t], \|\nabla^{(x,y)})\|_2)}$
    $\nabla[t] \leftarrow \nabla[t] + \overline{\nabla^{(x,y)}}$
  $\widetilde{\nabla[t]} \leftarrow \nabla[t] + \mathcal{N}(0, \sigma[t]^2 r[t]^2 \mathbb{I})$
  $\theta \leftarrow \theta - \eta \widetilde{\nabla[t]}$.
return θ

differential privacy can be in fact much more secure against membership inference attacks. Previously, this phenomenon was observed for DP-SGD and here, for the first time, we prove it.

Our analysis is the first to directly analyze membersihp inference bounds. We limited our study to membership inference attacks against sampled Gaussian mechanisms as DP-SGD is the most widely used differentially private learning algorithm. Direct study of membership inference could be potentially done for other mechanisms and algorithms as well. We leave this for future work.

We also note that the hyper-parameters for DP-SGD that achieve optimal membership privacy versus utility might be different than that of differential privacy. Our new analysis opens up the possibility of a systematic search for the hyper-parameters and architectures (using recent advancements in privacy preserving hyperparameter tuning and architecture search Papernot et al. [2020], Papernot and Steinke [2021], Chaudhuri and Vinterbo [2013]) that obtain optimal utility for a given upper bound on membership inference.

References

Martin Abadi, Andy Chu, Ian Goodfellow, Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *CCS*, 2016.

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.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. *arXiv preprint arXiv:2112.03570*, 2021.

Kamalika Chaudhuri and Staal A Vinterbo. A stability-based validation procedure for differentially private machine learning. *Advances in Neural Information Processing Systems*, 26, 2013.

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.

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60. IEEE, 2010.

Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 650–669. IEEE, 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*, pages 17–32, 2014.

Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.

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

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

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

Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papemoti, and Nicholas Carlin. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 866–882. IEEE, 2021.

Opacus. Opacus PyTorch library. Available from opacus.ai.

Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with renyi differential privacy. *arXiv preprint arXiv:2110.03620*, 2021.

Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, and Ulfar Erlingsson. Tempered sigmoid activations for deep learning with differential privacy. *arXiv preprint arXiv:2007.14191*, page 10, 2020.

Shahbaz Rezaei and Xin Liu. On the difficulty of membership inference attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7892–7900, 2021.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pages 5558–5567. PMLR, 2019.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Laurens van der Maaten and Awni Hannun. The trade-offs of private prediction. *arXiv preprint arXiv:2007.05089*, 2020.

Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations*, 2022.

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.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **[TODO]**
   (b) Did you describe the limitations of your work? **[TODO]**
   (c) Did you discuss any potential negative societal impacts of your work? **[TODO]**
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? **[TODO]**
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? **[TODO]**
   (b) Did you include complete proofs of all theoretical results? **[TODO]**
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[TODO]**
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **[TODO]**
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **[TODO]**
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **[TODO]**
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? **[TODO]**

(b) Did you mention the license of the assets? **[TODO]**

(c) Did you include any new assets either in the supplemental material or as a URL? **[TODO]**

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **[TODO]**

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[TODO]**

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? **[TODO]**

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **[TODO]**

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **[TODO]**

A Proof of Lemma 3

Proof. The first part follows by the symmetry of isotropic Gaussian. For the second part (monotonicity) we use the definition of $\mathbf{TV}_a$. Without loss of generality we can assume $a \in [0,1]$ as otherwise we can work with $\mathbf{TV}_a(P,Q)/a = \mathbf{TV}_{1/a}(Q,P)$. Let $r = \|u_1 - u_2\|_2$. We can show that the derivative of the integral is always positive. In the following calculations, we use c_1, c_2, c_3 and c_4 to denote positive constants that are independent of r.

First note that $x^* = \frac{r^2 - 2\sigma^2 \ln(a)}{2r}$ is a middle point where $e^{-\frac{x^2}{2\sigma^2}} - ae^{-\frac{(x-r)^2}{2\sigma^2}}$ goes from positive to negative as x increases. By our assumption that $a \in [0,1]$, we have that $x^* > 0$. Recalling that $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp(-t^2)dt$, and that $\mathrm{erf}(\infty) = 1$ so that (by symmetry) $\frac{2}{\sqrt{\pi}} \int_{-\infty}^0 \exp(-t^2)dt = 1$, we can write

$$
\begin{aligned}
\mathbf{TV}_a(P,Q) &= c_1 \left(\int_{-\infty}^{\infty} \left| e^{-\frac{x^2}{2\sigma^2}} - ae^{-\frac{(x-r)^2}{2\sigma^2}} \right| dx \right) \\
&= c_1 \left(\int_{-\infty}^{x^*} e^{-\frac{x^2}{2\sigma^2}} - ae^{-\frac{(x-r)^2}{2\sigma^2}} + \int_{x^*}^{\infty} ae^{-\frac{(x-r)^2}{2\sigma^2}} - e^{-\frac{x^2}{2\sigma^2}} \right) \\
&= c_1 \left(1 + \mathrm{erf}\left(x^*/\sqrt{2}\sigma \right) - a\mathrm{erf}((x^* - r)/\sqrt{2}\sigma) \right) \\
&\quad + \left(a(1 - \mathrm{erf}\left((x^* - r)/\sqrt{2}\sigma \right) + (1 - \mathrm{erf}\left(x^*/\sqrt{2}\sigma \right) \right) \\
&= c_2 \left(\mathrm{erf}\left(\frac{r^2 - \ln(a)\sigma^2}{2\sqrt{2}\sigma r} \right) + 1 - a\mathrm{erf}\left(\frac{-r^2 - \ln(a)\sigma^2}{2\sqrt{2}\sigma r} \right) - a \right).
\end{aligned}
$$

Now, let $f_1(r) = \mathrm{erf}\left(\frac{r^2 - \ln(a)\sigma^2}{2\sqrt{2}\sigma r} \right)$ and $f_2(r) = -a\mathrm{erf}\left(\frac{-r^2 - \ln(a)\sigma^2}{2\sqrt{2}\sigma r} \right)$. Taking the derivative with respect to r we have

$$
\begin{aligned}
\frac{\partial f_1}{\partial r} &= c_3 \left(\frac{1}{2\sqrt{2}\sigma} + \frac{\ln(a)\sigma}{2\sqrt{2}r^2} \right) e^{-\left(\frac{r^2 - \ln(a)\sigma^2}{2\sqrt{2}\sigma r} \right)^2} \\
\frac{\partial f_2}{\partial r} &= c_3 a \left(\frac{1}{2\sqrt{2}\sigma} - \frac{\ln(a)\sigma}{2\sqrt{2}r^2} \right) e^{-\left(\frac{-r^2 - \ln(a)\sigma^2}{2\sqrt{2}\sigma r} \right)^2}
\end{aligned}
$$

Now note that we have $e^{-\left(\frac{r^2 - \ln(a)\sigma^2}{2\sqrt{2}\sigma r} \right)^2} = a^{1/2} \cdot e^{-\left(\frac{-r^2 - \ln(a)\sigma^2}{2\sqrt{2}\sigma r} \right)^2}$. Therefore, we have

$$
c_4 \frac{\partial \mathbf{TV}_a}{\partial r} = e^{-\left(\frac{-r^2 - \ln(a)\sigma^2}{2\sqrt{2}\sigma r} \right)^2} \cdot \left(\frac{1 + \sqrt{a}}{2\sqrt{2}\sigma} + \frac{\ln(a)\left(\sqrt{a} - 1\right)\sigma}{2\sqrt{2}r^2} \right).
$$

Now since $a \in [0,1]$, we have $\ln(a) \leq 0$ and $\sqrt{a} - 1 < 0$, which means the term $\frac{1+\sqrt{a}}{2\sqrt{2}\sigma} + \frac{\ln(a)(\sqrt{a}-1)\sigma}{2\sqrt{2}r^2}$ is positive. This implies that the whole gradient is positive.

□

B Estimating KL with Rényi divergence

For two probability distributions μ and ν, the Rényi divergence of order $\alpha > 1$ is

$$
D_\alpha(\mu \parallel \nu) \triangleq \frac{1}{\alpha - 1} \log \mathbb{E}_{t \sim \nu} \left(\frac{d\mu}{d\nu}(t) \right)^\alpha, \tag{13}
$$

Rényi divergence is non-decreasing in α, and $\lim_{\alpha \to 1} D_\alpha(P \parallel Q) = \mathrm{KL}(P \parallel Q)$.

RDP accounting for DP-SGD. [Abadi et al., 2016, Mironov, 2017] propose methods to account for RDP for the Gaussian mechanism. Implementations of DP-SGD such as Opacus make use of these accounting procedures. This is important as we use these accounting methods to calculate one of our bounds for membership inference for composition of sub-sampled Gaussian mechanims.

C Membership inference precision

In this section, we refine the analysis of Sablayrolles et al. [2019] for the accuracy of a membership attack.

Upper-bound on precision. Let us first derive a bound on the precision of membership inference. We assume that there are two datasets D and D' and that a differentially-private mechanism $\mathcal{M}$ trains a model represented by θ.

With probability $(1-\delta)$ over the choice of θ, we have:

$$-\epsilon \leq \log\left(\frac{\mathbb{P}(M(D)=\theta)}{\mathbb{P}(M(D')=\theta)}\right) \leq \epsilon \tag{14}$$

Given that there is a balanced prior $\mathbb{P}(D)=\mathbb{P}(D')$, using Bayes rule, we have:

$$\mathbb{P}(D \mid \theta) = \frac{\mathbb{P}(M(D)=\theta)\mathbb{P}(D)}{\mathbb{P}(M(D)=\theta)\mathbb{P}(D)+\mathbb{P}(M(D')=\theta)\mathbb{P}(D')} \tag{15}$$

$$= \frac{\mathbb{P}(M(D)=\theta)}{\mathbb{P}(M(D)=\theta)+\mathbb{P}(M(D')=\theta)} \tag{16}$$

$$= \sigma\left(\log\left(\frac{\mathbb{P}(M(D)=\theta)}{\mathbb{P}(M(D')=\theta)}\right)\right), \tag{17}$$

with $\sigma(u) = 1/(1+\exp(-u))$ the sigmoid function.

Hence the precision $\mathbb{P}(D \mid \theta)$ is bounded between $\sigma(-\epsilon)$ and $\sigma(\epsilon)$, as $\sigma(\cdot)$ is non decreasing.

Upper-bound on attack accuracy. The accuracy of the Bayes classifier is

$$\mathrm{Acc} = \max(\mathbb{P}(D \mid \theta), 1-\mathbb{P}(D \mid \theta)), \tag{18}$$

and thus

$$\mathrm{Acc} \leq \max(\sigma(\epsilon), \sigma(-\epsilon)) \tag{19}$$

$$= \sigma(\epsilon) \tag{20}$$

This means that the attack accuracy is bounded by $\sigma(\epsilon)$ with probability $1-\delta$. Empirically, we see that the sigmoid function closely matches the bound given by Humphries et al. [2020]. Simply stated, this derivation shows that the bound proven by Humphries et al. [2020] actually holds with probability $1-\delta$ instead of on average.