Auditing Differential Privacy Guarantees Using Density Estimation

Antti Koskela
Nokia Bell Labs

Jafar Mohammadi (Aco)
Nokia

***Abstract*—We present a novel method for accurately auditing the differential privacy (DP) guarantees of DP mechanisms. In particular, our solution is applicable to auditing DP guarantees of machine learning (ML) models. Previous auditing methods tightly capture the privacy guarantees of DP-SGD trained models in the white-box setting where the auditor has access to all intermediate models. However, the success of these methods depends on prior information about the parametric form of the noise and the subsampling ratio used for sampling the gradients. We present a method that does not require such information and is agnostic to the randomization used for the underlying mechanism. Similar to a large majority of previous DP auditing methods, we assume that the auditor has access to a set of independent observations from two one-dimensional distributions corresponding to outputs from two neighboring datasets. Our solution is based on a simple histogram-based density estimation technique to find lower bounds for the statistical distance between these distributions when measured using the hockey-stick divergence. We show that our approach also naturally generalizes the previously considered class of threshold membership inference auditing methods. We improve upon the state-of-the-art accurate auditing methods, specifically f-DP auditing. We also address an open problem on how to accurately audit the subsampled Gaussian mechanism without any knowledge of the parameters of the underlying mechanism.**

***Index Terms*—Differential Privacy, Auditing, Machine Learning, DP-SGD**

I. Introduction

Differential Privacy (DP) [1] limits the disclosure of membership information about individuals in statistical data analyses. It has also been successfully applied to the training of machine learning (ML) models, in which the de facto standard is the DP stochastic gradient descent (DP-SGD) [2], [3]. DP-SGD enables the analysis of formal (ε, δ)-DP guarantees via *composition analysis* in a threat model in which the guarantees hold against an adversary with access to the whole history of models. Using modern numerical accounting tools [4], [5], [6], it is also possible to obtain accurate (ε, δ)-DP guarantees for DP-SGD within this threat model.

We motivate the privacy auditing problem with the following scenario: Consider a federated learning (FL) setup, where a partially trusted server enhances the DP protection by aggregating the local model updates and adding noise to the global updates. Achieving the theoretical privacy guarantees of DP-SGD requires a notoriously complex implementation setup [7], [8]. Since parts of the model updates are performed by an external entity, a data owner cannot fully ensure that the DP guarantees hold [9], [10]. This raises the question: how can the data owner conduct privacy auditing to ensure at least a certain amount of DP protection? Therefore, establishing some certainty about lower bounds for the DP parameters ε and δ is essential.

The problem of DP auditing has gained increasing attention in recent years. Many existing works on DP-SGD auditing focus on inserting well-designed data elements or gradients into the training dataset, known *canaries*. By observing their effect in the trained model, one can infer about the DP guarantees [11], [12], [13], [8]. We note that these methods often require training several models in order to obtain the estimates of the DP guarantees, even up to thousands [13]. To address the computational burden of training the model multiple times, recently [14] and [10] have proposed different approaches where auditing can be carried out in a single DP-SGD training iteration.

Most of the methods such as [11], [12], [8], [13], [10] are ultimately based on estimating the (ε, δ)-DP distance between two distributions corresponding to the outcomes of DP mechanisms evaluated on datasets differing only by one data element. For instance, in black-box auditing, one distribution would correspond to loss function values evaluated on a dataset including a given data sample z while the other corresponds to the same dataset with z excluded [11], [12], [8], [15]. The (ε, δ)-guarantees are then commonly estimated using threshold membership inference attacks [16], [17], where a model is deemed to contain the given sample if its loss function value for that sample falls below a certain threshold. By training multiple models, once with and once without the differing sample, and by measuring the false positive rates (FPRs) and false negative rates (FNRs) of the membership inferences, empirical ε-estimates can be derived for a given value of δ [18]. Our work can be seen as a generalization of this approach such that we estimate the two neighboring distributions using histograms. As we show, the empirical ε-values obtained via threshold membership inference attacks are equivalent to measuring the hockey-stick divergence between the two discrete distribution obtained by histogram estimation with two bins determined by the threshold.

One drawback of the threshold membership inference based

This work is Co-funded by the European Union under Grant Agreement 101191936. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of all SUSTAIN-6G consortium parties nor those of the European Union or the SNS JU (granting authority). Neither the European Union nor the granting authority can be held responsible for them.

auditing methods is that they tend to underestimate the ε-values. To address this, [8] proposes f-DP auditing, where a certain trade-off curve f is fitted so that the FNRs and FPRs of the membership inference attacks stay below f with high confidence, thereby providing high-confidence lower bounds for the DP parameters ε and δ. This approach, however, also has its drawbacks, as its success depends on selecting suitable f for the trade-off curve which generally requires prior knowledge about the DP randomization mechanism. Additionally, it often involves a complicated numerical integration procedure, which may lead to instabilities, as we also demonstrate in this paper. Our approach is similar in the sense that we also aim to accurately approximate the trade-off function. Our approach differs in that its success does not depend on any prior knowledge about the DP mechanism and it has a simple and robust implementation.

In the FL setting, two notable works are related to ours, both based on estimating the (ε, δ)-distance between two Gaussian distributions [10], [9]. The work [10] proposes inserting randomly sampled canaries in the model updates and the method given in [9] involves carefully crafting canary gradients. We show that our approach generalizes the auditing method of [10] by analytically showing that our histogram-based method gives asymptotically the correct guarantees as the model dimension increases, similarly to the method of [10] which requires prior knowledge about the noise.

Our paper is organized as follows. After presenting the necessary definitions and results on DP, in Section IV, we describe the idea of obtaining (ε, δ)-DP lower bounds through the hockey-stick divergence between certain histogram estimates. In Section V, we outline how our approach generalizes the threshold inference auditing, and in Section VI we provide numerical toy examples to illustrate the advantages of our approach. Finally, in Section VII, experiments on small neural networks in both black-box and white-box settings further confirm the benefits of our auditing method.

Our main contributions can be summarized as follows:

- We introduce a novel method of auditing the DP guarantees of a mechanism $\mathcal{M}$ using samples from two distributions that are known to be post-processed outputs of $\mathcal{M}$ evaluated on neighboring datasets. This fits perfectly to several previously considered black-box and white-box auditing scenarios [11], [12], [8], [15]. Our method extends the class of threshold membership inference methods by simultaneously considering multiple membership regions and by using histogram estimation of the distributions to obtain a more accurate estimate of the (ε, δ)-distance.
- We solve an open problem posed in [8] regarding how to tightly audit the subsampled Gaussian mechanism. As demonstrated in [8] shows, a single threshold membership inference is insufficient to capture the accurate trade-off curve of the subsampled Gaussian mechanism. We show, both theoretically and empirically, that the trade-off curve estimated using our method converges to the true trade-off curve. In addition to addressing the implementation bugs of DP-SGD discussed in the works [7] and [8], our method also allows auditing errors related to the implementation of subsampling.
- We propose a heuristic algorithm for estimating the privacy loss distribution of the underlying mechanism in the white-box auditing setting. This allows for accurate estimates for a given number of compositions of the DP mechanisms to be audited.
- We conduct numerical experiments on neural network training on two datasets to demonstrate the benefits of our approach.
- Additionally, in Appendix Section H, we analytically illustrate that the total variation distance provides a robust estimator when the privacy profiles depend on a single parameter, and in Appendix Section G we analytically show that our method generalizes the one-shot auditing method of [10] that uses random gradient canaries.

II. Background

A. Differential Privacy

We denote the space of possible data points by X. We denote a dataset containing n data points as $D = (x_1, \ldots, x_n) \in X^n$, and the space of all possible datasets (of all sizes) by $\mathcal{X}$. We say D and D' are neighboring datasets if we get one by substituting one element in the other (denoted $D \simeq D'$). Let $\mathcal{P}(\mathcal{O})$ denote the set of probability distributions with a support in an output space $\mathcal{O}$. We say that a mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{P}(\mathcal{O})$ (or, a randomized function) is (ε, δ)-DP if it satisfies the condition of the following definition.

Definition 1. *Let $\varepsilon \geq 0$ and $\delta \in [0, 1]$. Mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{P}(\mathcal{O})$ is (ε, δ)-DP if for every pair of neighboring datasets $D, D' \in \mathcal{X}$ and every measurable set $E \subset \mathcal{O}$,*

$$\mathbb{P}(\mathcal{M}(D) \in E) \leq \mathrm{e}^{\varepsilon} \mathbb{P}(\mathcal{M}(D') \in E) + \delta.$$

For two distributions P and Q with support in the output space $\mathcal{O}$, we say that they are (ε, δ)-indistinguishable in case for every measurable set $E \subset \mathcal{O}$, $\mathbb{P}(P \in E) \leq \mathrm{e}^{\varepsilon} \mathbb{P}(Q \in E) + \delta$ and $\mathbb{P}(Q \in E) \leq \mathrm{e}^{\varepsilon} \mathbb{P}(P \in E) + \delta$. Thus, a mechanism $\mathcal{M}$ is (ε, δ)-DP if the output distributions for neighboring datasets are always (ε, δ)-indistinguishable.

The tight (ε, δ)-guarantees for a mechanism $\mathcal{M}$ can be stated using the hockey-stick divergence. For $\alpha \geq 0$ the hockey-stick divergence H_α from a distribution P to a distribution Q is defined as

$$H_\alpha(P||Q) = \int [P(t) - \alpha \cdot Q(t)]_+ \, \mathrm{d}t, \tag{II.1}$$

where $[t]_+ = \max\{0, t\}$. The (ε, δ)-DP guarantees can be characterized using the hockey-stick divergence as follows (see Theorem 1 in [19]).

Lemma 2. *For a given $\varepsilon \in \mathbb{R}$, a mechanism $\mathcal{M}$ satisfies (ϵ, δ)-DP* if and only if*, for all neighboring datasets D, D',*

$$H_{\mathrm{e}^{\varepsilon}}(\mathcal{M}(D)||\mathcal{M}(D')) \leq \delta.$$

We also refer to $\delta_{\mathcal{M}}(\varepsilon) := \max_{D \simeq D'} H_{e^\varepsilon}(\mathcal{M}(D)||\mathcal{M}(D'))$ as the *privacy profile* of mechanism $\mathcal{M}$.

By Lemma 2, if we can bound $H_{\mathrm{e}^\varepsilon}(\mathcal{M}(D)||\mathcal{M}(D'))$ accurately for all neighboring datasets D, D', we also obtain accurate (ε, δ)-DP bounds. For compositions of general DP mechanisms, this can be carried out by using so-called dominating pairs of distributions [5] and numerical techniques [20], [6]. In some cases, such as for the Gaussian mechanism, the hockey-stick divergence (II.1) leads to analytical expressions for tight (ε, δ)-DP guarantees [21].

Lemma 3. *Let $d_0, d_1 \in \mathbb{R}^d$, $\sigma \geq 0$, and let P be the density function of $\mathcal{N}(d_0, \sigma^2 I_d)$ and Q the density function of $\mathcal{N}(d_1, \sigma^2 I_d)$. Then, for all $\varepsilon \in \mathbb{R}$, the divergence $H_{\mathrm{e}^\varepsilon}(P||Q)$ is given by the expression*

$$\delta(\varepsilon) = \Phi\left(-\frac{\varepsilon\sigma}{\Delta} + \frac{\Delta}{2\sigma}\right) - e^{\varepsilon}\Phi\left(-\frac{\varepsilon\sigma}{\Delta} - \frac{\Delta}{2\sigma}\right), \qquad \text{(II.2)}$$

where Φ denotes the CDF of the standard univariate Gaussian distribution and $\Delta = \|d_0 - d_1\|_2$.

Setting $\alpha = 1$ in Eq. (II.1), we get the total variation (TV) distance between the probability distributions P and Q (see, e.g., [22]),

$$\begin{aligned} \mathrm{TV}(P, Q) &= \frac{1}{2}\int |P(x) - Q(x)| \, \mathrm{d}x \\ &= \int [P(x) - Q(x)]_+ \, \mathrm{d}x. \end{aligned} \qquad \text{(II.3)}$$

When P and Q are discrete, defined by probabilities p_k and q_k, $k \in \mathbb{Z}$, respectively, we have the important special case of discrete TV distance defined by

$$\mathrm{TV}(P, Q) = \sum\nolimits_{k \in \mathbb{Z}} \max\{p_k - q_k, 0\}.$$

B. Trade-Off Functions and Functional DP

DP can also be understood from a hypothesis testing perspective [23]. In the context of ML model auditing, this can be formulated as follows [8]. Consider the hypothesis testing problem

$$\begin{aligned} H_0 &: \quad \text{the model } \theta \text{ is drawn from } P \\ H_1 &: \quad \text{the model } \theta \text{ is drawn from } Q, \end{aligned}$$

where P and Q are obtained via some post-processing of the probability distributions of $\mathcal{M}(D)$ and $\mathcal{M}(D')$, respectively. This ensures, in particular, by the post-processing property of DP, that if $\mathcal{M}$ is (ε, δ)-DP, then P and Q are (ε, δ)-indistinguishable. Furthermore, this allows obtaining lower bounds for the DP guarantees of $\mathcal{M}$ by obtaining lower bounds for the (ε, δ)-parameters of P and Q.

The trade-off function, as defined in [24], captures the difficulty of distinguishing the hypotheses H_0 and H_1. Given a rejection rule $0 \leq \phi(\theta) \leq 1$ that takes as an input the model θ trained by the mechanism $\mathcal{M}$, the type I error is defined as $\alpha_\phi = \mathbb{E}_P[\phi]$ and the type II error as $\beta_\phi = 1 - \mathbb{E}_Q[\phi]$. Then, the trade-off function that describes the upper bound for the distinguishability is given as follows.

Definition 4. *Define the trade-off function $T(P, Q) : [0, 1] \to [0, 1]$ for two probability distributions P and Q as*

$$T(P, Q)(\alpha) = \inf\{\beta_\phi \, : \, \alpha_\phi \leq \alpha\}.$$

For an arbitrary function $f : [0, 1] \to [0, 1]$, the following properties characterize whether it is a trade-off function (see Prop. 2.2 in [24]).

Lemma 5. *A function $f : [0, 1] \to [0, 1]$ is a trade-off function if and only if f is convex, continuous, non-increasing, and $f(x) \leq 1 - x$ for all $x \in [0, 1]$.*

The f-DP can be then defined as follows.

Definition 6. *Let f be a trade-off function. A mechanism $\mathcal{M}$ is f-DP if*

$$T\big(\mathcal{M}(D), \mathcal{M}(D')\big) \geq f$$

for all neighboring datasets D and D'.

As shown in [24], (ε, δ)-DP is equivalent to f-DP for the following trade-off function:

$$f_{\varepsilon,\delta}(\alpha) = \max\{0, 1 - \delta - \mathrm{e}^{\varepsilon}\alpha, \mathrm{e}^{-\varepsilon}(1 - \delta - \alpha)\}.$$

From this, we directly get the following accurate characterization of the trade-off function for a given mechanism $\mathcal{M}$.

Lemma 7. *Suppose we have a privacy profile $h(\alpha)$ of the mechanism $\mathcal{M}$. Then, the function given by*

$$f(x) = \max_{\alpha \geq 0} \max\{0, 1 - h(\alpha) - \alpha x, \alpha^{-1}(1 - h(\alpha) - x)\}$$

is a trade-off function of $\mathcal{M}$.

This also allows approximating the trade-off function using numerical integrators given a set of points $(\varepsilon_1, \delta_1), \ldots, (\varepsilon_m, \delta_m)$. From Lemma 7, we directly have the following approximation algorithm which is essentially the one given if Appendix A of [8].

Algorithm 1 Estimation of the trade-off function f using a privacy profile $\delta(\varepsilon)$

$F_{\mathcal{M}}$ privacy analysis function that gives that outputs ε for a given δ, n number of discretization points, δ target delta in the DP analysis.
$\Delta \leftarrow n$ linearly spaced points on the interval $[\delta, 1 - \delta]$.
for $\delta' \in \Delta$: **do**
  $\widehat{\varepsilon} \leftarrow F_{\mathcal{M}}(\delta')$
  $f_{\delta'}(x) := \max\{0, 1 - \delta' - x\mathrm{e}^{\widehat{\varepsilon}}, \mathrm{e}^{-\widehat{\varepsilon}}(1 - \delta' - x)\}$
end for
$f(x) := \max_{\delta' \in \Delta} f_{\delta'}(x)$

We may also utilize Lemma 7 and use the procedure of Alg. 1 to estimate the trade-off function when we are given a point-wise estimation of the privacy profile represented by a discrete set $\{(\varepsilon_1, \delta_1), \ldots, (\varepsilon_m, \delta_m)\}$,. This is shown in Alg. 1.

Informally speaking, a mechanism is μ-GDP if the outcomes from two neighboring distributions are not more distinguishable than two unit variance Gaussians μ apart from each other.

Algorithm 2 Estimation of the trade-off function f using a set of points $\Delta = \{(\varepsilon_1, \delta_1), \ldots, (\varepsilon_m, \delta_m)\}$

Set of points $\Delta = \{(\varepsilon_1, \delta_1), \ldots, (\varepsilon_m, \delta_m)\}$, n number of discretization points, δ target delta in the DP analysis.
for $(\varepsilon', \delta') \in \Delta$: **do**
$f_{\delta'}(x) := \max\{0, 1 - \delta' - x\mathrm{e}^{\varepsilon'}, \mathrm{e}^{-\varepsilon'}(1 - \delta' - x)\}$
end for
$f(x) := \max_{\delta' \in \Delta} f_{\delta'}(x)$

Using a trade-off function determined by $\mathcal{N}(0,1)$ and $\mathcal{N}(\mu,1)$, we have the following charaterization [24].

Definition 8. *A mechanism $\mathcal{M}$ is μ-GDP if for all $\alpha \in [0,1]$,*

$$T\big(\mathcal{M}(D), \mathcal{M}(D')\big)(\alpha) \geq \Phi(\Phi^{-1}(1-\alpha) - \mu)$$

for all neighboring datasets D, D' where Φ is the standard normal CDF.

C. Confidence Intervals for f-DP

By using empirical upper bounds for α and β, obtained using, e.g., the Clopper–Pearson intervals or Jeffreys intervals, and Def. 8, we may obtain an empirical lower bound for the GDP parameter μ as

$$\mu_{\text{emp}}^{\text{lower}} = \Phi^{-1}(1-\bar{\alpha}) - \Phi^{-1}(\bar{\beta}). \tag{II.4}$$

We remark, however, that this estimation of μ using a point-wise estimate $(\bar{\alpha}, \bar{\beta})$ may lead to false (ε, δ)-lower bounds in case the privacy profile of the mechanism deviates significantly from that of a Gaussian mechanism.

The work [8] proposes also to use the *credible intervals* for ε as a basis for the confidence interval estimation in f-DP. This approach is based on a certain Bayesian estimation of ε-values proposed in [25]. Therein, given the estimated FP and FN-values of the attack, a posterior distribution $u_{(\text{FPR},\text{FNR})}(\alpha, \beta)$ is defined as

$$\begin{aligned} u_{(\text{FPR},\text{FNR})}(\alpha,\beta) =& \text{Beta}(\alpha; 0.5 + \text{FN}, 0.5 + N - \text{FN}) \cdot \\ & \text{Beta}(\beta; 0.5 + \text{FP}, 0.5 + N - \text{FP}) \end{aligned}$$

A trade-off curve f is then determined to give an f-DP guarantee with confidence c, where c is the probability mass of the posterior distribution $u_{(\text{FPR},\text{FNR})}(\alpha, \beta)$ in the privacy region determined by f, i.e., between the curves $f(\alpha)$ and $1 - f(1-\alpha)$, $\alpha \in [0,1]$. The confidence value c is then determined by the cumulative distribution function

$$P_{\cdot}(f) = \int_0^1 \int_{f(\alpha)}^{1-f(1-\alpha)} u_{(\text{FPR},\text{FNR})}(\alpha,\beta)\, \mathrm{d}\beta\, \mathrm{d}\alpha \tag{II.5}$$

which gives the mass of the distribution $u_{(\text{FPR},\text{FNR})}(\alpha, \beta)$ in the privacy region determined by the trade-off function f. Finding a suitable trade-off curve using the integral (II.5) is difficult for several reasons and we remark that the work [8] mostly uses in its experiments the GDP estimate II.4, where $\bar{\alpha}$ and $\bar{\beta}$ are obtained either using the Clopper–Pearson estimates or Bayesian estimates using the approach of [25].

III. Difficulties in Auditing with f-DP

As shown in [8], the success of threshold inference based μ-GDP auditing does not depend on the value of the threshold in case $P \sim \mathcal{N}(1, \sigma^2)$ and $Q \sim \mathcal{N}(0, \sigma^2)$. Asymptotically, we then have that for a threshold value $z \in \mathbb{R}$, $\alpha = 1 - \Phi\left(\frac{z}{\sigma}\right)$ and $\beta = \Phi\left(\frac{z-1}{\sigma}\right)$, and for all $z \in \mathbb{R}$,

$$\mu = \Phi^{-1}(1-\alpha) - \Phi^{-1}(\beta). \tag{III.1}$$

However, while the relation (III.1) and the threshold independence hold for P and Q that are exactly Gaussians with an equal variance, they do not hold for general distributions and in general finding the GDP parameter accurately requires tuning of the threshold parameter z. To illustrate this, consider the example given in [8]: let $P \sim q \cdot \mathcal{N}(1, \sigma^2) + (1-q) \cdot \mathcal{N}(0, \sigma^2)$ and $Q \sim \cdot \mathcal{N}(0, \sigma^2)$, where $\sigma = 0.3$ and $q = 0.25$. Using accurate numerical calculation of the privacy profile $\delta(\varepsilon) = \max\{H_{\mathrm{e}^\varepsilon}(P||Q), H_{\mathrm{e}^\varepsilon}(Q||P)\}$ and numerical optimization, we find that the pair (P, Q) is μ-GDP for $\mu \approx 1/0.404$ (see Appendix Figure 15).

Figure 1 shows the μ-value estimated using equation (III.1). Clearly, the threshold independence of the μ-GDP auditing does not hold for non-Gaussian distributions. Also, we experimentally find that the largest μ-estimate require large threshold values (very small FPRs) so that the confidence intervals easily become large and we are not able to get close to the accurate μ-values even when using $n = 10^5$ samples. Also, as we see, finding a suitable value for the threshold value z requires careful tuning as the μ-estimation is z-independent only for a pair of Gaussian with equal variance.

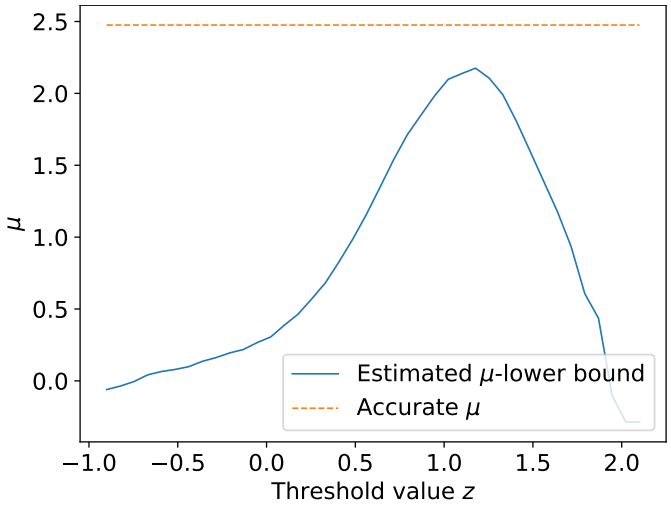

Fig. 1. Adjusting the μ-GDP parameter for the pair of distributions $P \sim q \cdot \mathcal{N}(1, \sigma^2) + (1-q) \cdot \mathcal{N}(0, \sigma^2)$ and $Q \sim \cdot \mathcal{N}(0, \sigma^2)$ using a threshold attack with threshold value $z \in \mathbb{R}$. The figure shows the estimated μ-value as a function of z. Each μ-lower bound value is estimated using the Clopper–Pearson confidence intervals and $n = 10^4$ samples from both P and Q.

In the general case, such as when carrying our f-DP auditing of the subsampled Gaussian mechanism, one has to use the formula (II.5). The first major difficulty with using the integral (II.5) one encounters is in the case when auditing

mechanisms determined by more than one parameter. For example, when auditing the subsampled Gaussian mechanism, the potential f-curves are parameterized by two parameters, q and σ. Thus, given only the observations, it is not obvious how to adapt q and σ to obtain high-confidence privacy regions for the posterior distribution $u_{(\mathrm{FPR},\mathrm{FNR})}(\alpha,\beta)$ as both of these parameters will affect the shape of the trade-off function f. If one is focused on point-wise (ε,δ)-DP estimates one may end up with wildly different f-DP guarantees: as demonstrated recently in [26], two mechanisms can have wildly different privacy profiles while having the same point-wise (ε,δ)-DP guarantees.

The second difficulty one quickly encounters with the formula (II.5) is the numerical approximation. The formula (II.5) which does not seem to exhibit analytical solutions even in the simplest cases (e.g., μ-GDP estimation). Therein, one specific issue that requires careful attention is that even the $f(\alpha)$-curve that determines the boundary of the privacy region may not have analytical expression but has to be approximated numerically. This is the case, e.g., in case f is a trade-off curve of the subsampled Gaussian mechanism, where we approximate it using Algorithm 1. However, the biggest difficulty seems to arise from the numerical stability of the integration.

We demonstrate the difficulty of the numerical f-DP auditing with an example where we are auditing the one-dimensional distributions $P\sim q\cdot\mathcal{N}(1,\sigma^2)+(1-q)\cdot\mathcal{N}(0,\sigma^2)$ and $Q\sim\mathcal{N}(0,\sigma^2)$ where $q=0.25$ and $\sigma=0.3$. We consider a situation where the auditor is given the value of q and is trying to determine the upper bound f-trade-off curve by scaling σ and by using a numerical approximation of the integral (II.5). The posterior distribution $u_{(\mathrm{FPR},\mathrm{FNR})}$ is constructed using threshold membership inference and $n=10^5$ samples from both P and Q. To reduce the influence of the numerical integrator on our conclusions, we use two different numerical integration methods: we use the dblquad-integrator included in scipy.integrate library [27] and a simple two-dimensional Euler method. For a given value of σ, we compute an approximation of the accurate f-DP curve of the subsampled Gaussian mechanism using Alg. 1with 200 points and estimate the true f-function by a piece-wise linear function constructed using these points. The privacy region estimated using this approximated f-curve is given as an integral region for the dblquad-integrator. When $\sigma=0.35$, both integrators correctly indicate that the f-curve is an upper bound for the privacy region (Fig. 2). However, when $\sigma=0.29$, we can still find threshold values for which both integrators would deem the privacy region to be under the f-curve, which is clearly a wrong conclusion (Fig. 3).

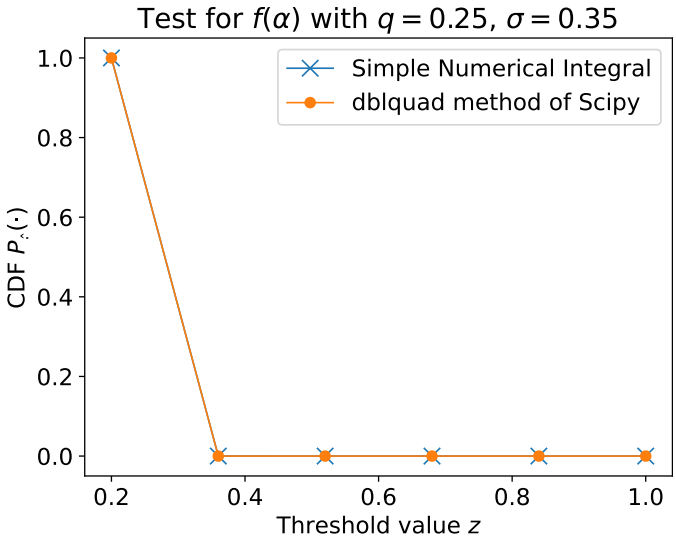

Fig. 2. Estimate of the cumulative density function $P_z(\cdot)$, i.e., the probability mass of the posterior distribution $u_{(\mathrm{FPR},\mathrm{FNR})}(\alpha,\beta)$ inside the privacy region determined by the trade-off function f of the subsampled Gaussian mechanism with sampling ratio $q=0.25$ and noise parameter $\sigma=0.35$. Using a threshold value between -0.75 and 0.2 would lead us to conclude with high confidence that the mass of $u_{(\mathrm{FPR},\mathrm{FNR})}(\alpha,\beta)$ is inside the privacy region. While this would lead to a correct lower bound for the DP parameter ε, it would give an inaccurate approximation of the true trade-off function.

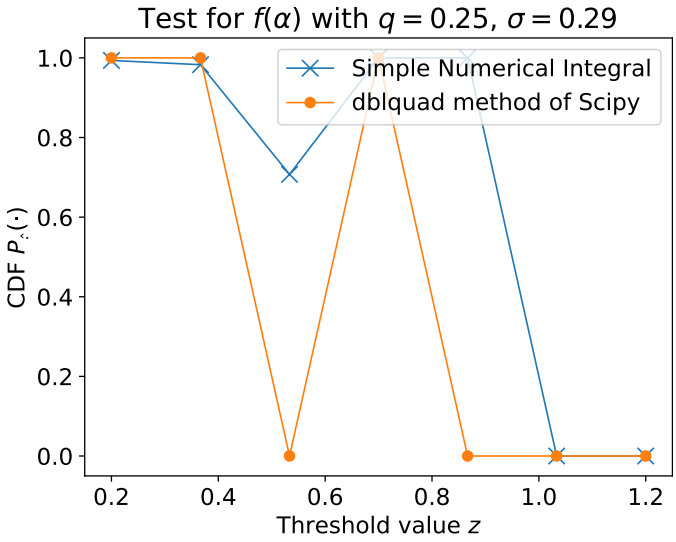

Fig. 3. Estimate of the cumulative density function $P_z(\cdot)$, i.e., the probability mass of the posterior distribution $u_{(\mathrm{FPR},\mathrm{FNR})}(\alpha,\beta)$ inside the privacy region determined by the trade-off function f of the subsampled Gaussian mechanism with sampling ratio $q=0.25$ and noise parameter $\sigma=0.35$. Using a threshold value between -0.75 and 0.2 would lead us to conclude with high confidence that the mass of $u_{(\mathrm{FPR},\mathrm{FNR})}(\alpha,\beta)$ is inside the privacy region. While this would lead to a correct lower bound for the DP parameter ε, it would give an inaccurate approximation of the true trade-off function.

IV. Histogram-Based Auditing of DP Guarantees

We next present our histogram-based DP auditing method that does not require any a priori information about the underlying DP mechanism.

A. Problem Formulation

Similarly to the hypothesis testing formulation of DP presented in Section II, our method is based on a general problem formulation, where the privacy profile of the underlying DP mechanism $\mathcal{M}$ dominates the privacy profile

$h(\alpha) = H_\alpha(P, Q)$ determined by some distributions P and Q and we have a number of independent samples from both P and Q. Then, having an estimate (or high-confidence lower bound) for $h(\alpha)$ will also give a lower bound for the privacy profile of $\mathcal{M}$.

We can motivate this formulation for example via black-box auditing of an ML model training algorithm $\mathcal{M}$ as follows. Let $\theta \in \mathbb{R}^d$ denote the ML model parameters, $F(\theta, x)$ the forward mapping for the feature x of a data element $z = (x, y)$, where y denotes the label, and let $\ell\big(F(\theta, z), y\big)$ be some loss function. Then, in case the mechanism $\mathcal{M}$ is (ε, δ)-DP, by the post-processing property of DP, the distributions

$$\begin{aligned} P &= \ell\big(F(\theta, x), y\big), \quad \theta \sim \mathcal{M}(D \cup z), \\ Q &= \ell\big(F(\theta, x), y\big), \quad \theta \sim \mathcal{M}(D) \end{aligned} \tag{IV.1}$$

are (ε, δ)-close to each other, i.e. $H_{\mathrm{e}^\varepsilon}(P, Q) \leq \delta$.

We next show how to lower bound the privacy profile $h(\alpha)$ using histogram density estimates of the distributions P and Q. Notice that in particular, P and Q are not dominating pairs of distributions for the mechanism $\mathcal{M}$; on the contrary, the privacy profile of $\mathcal{M}$ dominates the privacy profile $h(\alpha) = H_\alpha(P, Q)$, $\alpha \geq 0$.

B. Estimating Hockey-Stick Divergence Using Histograms

We estimate the distributions P and Q by first sampling n samples from P and n samples from Q, and then using binning such that we place the score values into k bins, each of given width $h > 0$. Denote these samples by $P_S = \{P_1, \ldots, P_n\}$ and $Q_S = \{Q_1, \ldots, Q_n\}$. Notice that we could use an adaptive division of the real line to generate the bin, however we here focus on equidistant bins for simplicity. Also, we could consider drawing a different amount of samples from P and Q. Given left and right end points a and b, respectively, we define the bin j, $j \in \{2, \ldots, k-1\}$, as

$$\mathrm{Bin}_j = [a + (j-1) \cdot h, a + j \cdot h)$$

and

$$\mathrm{Bin}_1 = (\infty, a + h), \quad \mathrm{Bin}_k = [b - h, \infty).$$

We define the probabilities p_j and q_j, $j \in [k]$, by the relative frequencies of P's and Q's samples hitting bin j:

$$\begin{aligned} p_j &\leftarrow \frac{1}{n} \left| \{x \in P_S \,:\, x \in \mathrm{Bin}_j\} \right|, \\ q_j &\leftarrow \frac{1}{n} \left| \{x \in Q_S \,:\, x \in \mathrm{Bin}_j\} \right|. \end{aligned}$$

Denote these estimated discrete distributions with probabilities p_j and q_j, $j \in [k]$, by $\widehat{P}$ and $\widehat{Q}$, respectively. Then, we estimate the parameters of the mechanism $\mathcal{M}$ by using the hockey-stick divergence $H_{\mathrm{e}^\varepsilon}(\widehat{P} || \widehat{Q})$, $\varepsilon \in \mathbb{R}$. This is motivated by the following observation.

Lemma 9. *Denote the limiting distributions by $\widetilde{P}$ and $\widetilde{Q}$, i.e.,*

$$\widetilde{P}_j = \int_{Bin_j} P(t)\,\mathrm{d}t, \quad \textit{and} \quad \widetilde{Q}_j = \int_{Bin_j} Q(t)\,\mathrm{d}t$$

for all $j \in [k]$, where $P(t)$ and $Q(t)$ denote the density functions of P and Q, respectively. Then, for all $\varepsilon \in \mathbb{R}$:

$$H_{\mathrm{e}^\varepsilon}(\widetilde{P}, \widetilde{Q}) \leq H_{\mathrm{e}^\varepsilon}(P, Q).$$

Proof. The distributions $\widetilde{P}$ and $\widetilde{Q}$ are obtained by applying the same post-processing function to P and Q and the claim follows from the data processing inequality. □

To obtain a high-confidence lower bound for the hockey-stick divergence $H_{\mathrm{e}^\varepsilon}(\widetilde{P}, \widetilde{Q})$, the challenge is then how bound the error in the estimate $H_{\mathrm{e}^\varepsilon}(\widehat{P} || \widehat{Q})$. We next show how to obtain frequentist confidence intervals for this estimate.

C. Confidence Intervals for Histogram-Based ε-Estimates

We consider frequentist confidence intervals, and thus by definition a $(1-\alpha)$-confidence interval will contain the true parameter with $(1-\alpha)\%$ of the time the estimation is carried out.

An important observation here is that the counts of samples hitting bins,

$$\left| \{x \in P_S \,:\, x \in \mathrm{Bin}_j\} \right| \quad \text{and} \quad \left| \{x \in Q_S \,:\, x \in \mathrm{Bin}_j\} \right|,$$

$j \in [k]$, are independent draws from multinomial distributions with k events and event probabilities $\widetilde{P}$ and $\widetilde{Q}$, respectively. Denote the set of possible multinomial probabilities for the discrete set X by

$$\Delta(X) = \{p \in \mathbb{R}_{\geq 0}^{|X|} \,:\, \|p\|_1 = 1\}.$$

To obtain confidence intervals, we use the following high-probability bound for the total variation distance given in [28].

Lemma 10. *Consider the empirical distribution $\widetilde{p}$ obtained by drawing n independent samples $s_1, \ldots, s_n$ from the underlying distribution $p \in \Delta([k])$:*

$$\widetilde{p}_i = \frac{1}{n} \left| \{s \in \{s_1, \ldots, s_n\} \,:\, s = i\} \right|, \quad i \in [k].$$

Then, as long as

$$n \geq \max\left\{ \frac{k}{\tau^2}, \frac{2}{\tau^2} \log \frac{2}{\gamma} \right\},$$

we have that with probability at least $1 - \gamma$,

$$\mathrm{TV}(p, \widetilde{p}) \leq \tau.$$

It is evident that by choosing n as guided by Lemma 10, the interval $[\widetilde{p} - \tau, \widetilde{p} + \tau]$ will be a $100\% \cdot (1-\gamma)$ - confidence interval for the TV distance estimate. Notice that for fixed k, Lemma 10 implies that the error of the estimate in TV-distance is essentially upper bounded $\sqrt{\frac{k}{n}}$. We remark that by results of [29], in expectation, the TV-distance error of the estimate behaves as $\Omega\left(\sqrt{\frac{k}{n}}\right)$ so the bound is optimal in this sense.

We can use the confidence intervals for TV distance also to obtain high-confidence lower bounds for other parts of the privacy profile using Alg. 3 via the following result which is also given in [30].

Lemma 11. *Denote P, Q probability distributions on the same probability space. Suppose*

$$\mathrm{TV}(P, \widetilde{P}) \leq \tau$$

and

$$\mathrm{TV}(Q, \widetilde{Q}) \leq \tau$$

for some $\tau \geq 0$. Then, for all $\varepsilon \in \mathbb{R}$,

$$H_{\mathrm{e}^\varepsilon}(P, Q) \leq H_{\mathrm{e}^\varepsilon}(\widetilde{P}, \widetilde{Q}) + (1 + \mathrm{e}^\varepsilon) \cdot \tau.$$

For obtaining high-confidence f-DP upper bounds, our strategy is to determine the high-confidence lower bounds for the privacy profile using Lemma 11 and then convert these lower bounds to trade-off functions using Lemma 7. We remark that rigorously, this approach does not give a high-confidence f-DP upper bound. Due to the convexity of the trade-off functions, using point-wise upper bounds for the privacy profile would give a high-confidence lower bound for the trade-off function, however for the upper bound we would need to use an approach similar to that of [31], where they give an optimistic numerical approximation of the privacy profile that strictly lower bounds the true privacy profile. We believe however that the effect would be small and in experiments we simply use as a high-confidence upper bound the trade-off function approximated using Lemma 7.

D. Convergence Result for the Hockey-Stick Divergence Estimate

The density estimation using histograms is a classical problem in statistics, and existing results such as those of [32] can be used to derive suitable bin widths for the histograms. We also mention the work [33] which gives methods based on kernel estimation theory and the work [34] which gives binning based on a Bayesian procedure.

Consider the approach and notation of Section IV-B, except that for the theoretical analysis we consider an infinite number of bins and focus on find the optimal bin width h. I.e., we define the bins such that for $j \in \mathbb{Z}$,

$$\mathrm{Bin}_j = [j \cdot h, (j+1) \cdot h),$$

and place the n randomly drawn samples from P and Q into these bins to estimate the probabilities $\int_{\mathrm{Bin}_j} P(x)\,\mathrm{d}x$ and $\int_{\mathrm{Bin}_j} Q(x)\,\mathrm{d}x$ using the bin-wise frequencies of the histograms. If we denote the piece-wise continuous density function as

$$\widehat{P}(x) = \widehat{P}_j / h, \quad \text{when} \quad x \in \mathrm{Bin}_j,$$

then the analysis of [32] gives an optimal bin width for minimizing the mean-square error $\mathbb{E}(P(x) - \widehat{P}(x))^2$ for a density function $P(x)$ with bounded and continuous derivatives up to second order (and similarly for Q). We can directly use this result for analysing the convergence of the numerical hockey-stick divergence $H_{\mathrm{e}^\varepsilon}(\widehat{P}||\widehat{Q})$, $\varepsilon \in \mathbb{R}$, as a function of the number of samples n.

Theorem 12. *Let P and Q be one-dimensional probability distributions with differentiable density functions $P(x)$ and $Q(x)$, respectively, and consider the histogram-based density estimation described above. Draw n samples both from P and Q, giving density estimators $\widehat{P} = (\widehat{P}_1, \ldots, \widehat{P}_k)$ and $\widehat{Q} = (\widehat{Q}_1, \ldots, \widehat{Q}_k)$, respectively. Let the bin width be chosen as*

$$h_n = \left(\frac{12}{\int_{\mathbb{R}} P'(x)^2\,\mathrm{d}x + \int_{\mathbb{R}} Q'(x)^2\,\mathrm{d}x} \right)^{\frac{1}{3}} n^{-\frac{1}{3}}, \qquad \text{(IV.2)}$$

where $P'(x)$ denotes the derivative of the density function $P(x)$ and similarly for Q. Then, for any $\alpha \geq 0$, the numerical hockey-stick divergence $H_\alpha(\widehat{P}||\widehat{Q})$ convergences in expectation to $H_\alpha(P||Q)$ with rate $\mathcal{O}(n^{-1/3})$, i.e.,

$$\mathbb{E}\left| H_\alpha(\widehat{P}||\widehat{Q}) - H_\alpha(P||Q) \right| = \mathcal{O}(n^{-1/3}),$$

where the expectation is taken over the random draws for constructing $\widehat{P}$ and $\widehat{Q}$.

***Remark* 13.** *In the proof of Thm.* (12) *the number of bins k is chosen to minimize the approximation error and is $\mathcal{O}(n^{1/3})$. In that case, Lemmas 11 and 10 gives the same order $\mathcal{O}(n^{-1/3})$.*

In case P and Q are Gaussians with an equal variance, we directly get the following from the expression (IV.2).

Corollary 14. *Suppose P and Q are one-dimensional normal distributions both with variance σ^2. Then, the bin width h_k of Eq.* (IV.2) *is given by*

$$h_n = 2 \cdot 3^{1/3} \cdot \pi^{1/6} \cdot \sigma \cdot n^{-1/3}. \qquad \text{(IV.3)}$$

We may use the expression of Eq. (IV.3) for Gaussians as a rule of thumb also for other distributions with σ denoting the standard deviation.

E. Pseudocode for Histogram-Based Estimation of DP-Guarantees

The pseudocode for our (ε, δ)-DP auditing method is given in Alg. 3. Notice that in in order to find a suitable bin width h, we may also estimate the standard deviations of the samples P_S and Q_S. This is also motivated by the experimental observation that the variances of the score values for auditing training and test sets are similar. Then, having an std estimate σ, we could set the bin width $h = 3.5 \cdot n^{-1/3} \widehat{\sigma}$ which approximately equals the expression (IV.3).

F. Pseudocode for Histogram-Based Estimation of Trade-Off Functions

We next show how to obtain high-confidence lower bounds for the DP guarantees using Alg. 3. We first obtain point-wise high-confidence lower bounds for the privacy privacy profile, we use Alg. 3 combined with Lemmas 11 and 10. Notice that given the number of samples n, number of bins k and $\gamma > 0$, the error estimate τ is given by Lemma 10 as

$$\tau = \max\left\{ \sqrt{\frac{k}{n}}, \sqrt{\frac{2 \log 2/\gamma}{n}} \right\}.$$

Algorithm 3 Estimation of (ε, δ)-DP parameters Using Histogram Density Estimation

Input: n indepedent samples from the distributions P and Q: $P_S = \{P_1, \ldots, P_n\}$ and $Q_S = \{Q_1, \ldots, Q_n\}$, DP parameter $\delta \in (0, 1)$. Number of Bins k, end points $a, b \in \mathbb{R}$.

Set the bin width $h = \frac{b-a}{k}$.

Divide the real line into k disjoint intervals such that for $j \in \{2, \ldots, k-1\}$,

$$\text{Bin}_j = [a + (j-1) \cdot h, a + j \cdot h)$$

and $\text{Bin}_1 = (\infty, a + h)$ and $\text{Bin}_k = [b - h, \infty)$.

Estimate the probabilities p_j and q_j, $j \in [k]$, by the relative frequencies of hitting bin j as

$$p_j \leftarrow \frac{1}{n} \left| \{x \in P_S \,:\, x \in \text{Bin}_j\} \right|,$$
$$q_j \leftarrow \frac{1}{n} \left| \{x \in Q_S \,:\, x \in \text{Bin}_j\} \right|$$

giving the discrete-valued distributions $\widehat{P} = \{p_i\}_{i=1}^k$ and $\widehat{Q} = \{q_i\}_{i=1}^k$.

Set: $\delta \leftarrow H_{\mathrm{e}^\varepsilon}\big(\widehat{P} || \widehat{Q}\big)$.

return δ.

Algorithm 4 Computation of High-Confidence Lower Bounds for (ε, δ)-DP parameters Using Histogram Density Estimation

Input: n independent samples from the distributions P and Q: $P_S = \{P_1, \ldots, P_n\}$ and $Q_S = \{Q_1, \ldots, Q_n\}$, grid of ε-values $\{\varepsilon_1, \ldots, \varepsilon_m\}$ to construct the trade-off function, confidence parameter $\gamma > 0$.

Compute TV distance error: $\tau = \max\left\{\sqrt{\frac{k}{n}}, \sqrt{\frac{2\log 2/\gamma}{n}}\right\}$

for $i \in [m]$: **do**

Estimate $H_{\mathrm{e}^{\varepsilon_i}}\big(\widehat{P}_S || \widehat{Q}_S\big)$ using Alg. 3, giving an estimate δ_i.

Subtract τ from δ_i to get $(1-\gamma)$-lower bound:

$$\delta_i \leftarrow \max\{0, \delta_i - (1 + \mathrm{e}^\varepsilon) \cdot \tau\}.$$

end for

Using the points $\{(\varepsilon_1, \delta_1), \ldots, (\varepsilon_m, \delta_m)\}$, determine the trade-off function using Alg. 2.

return $f(x)$.

Thus obtained δ-lower bounds are then further converted to a trade-off function using Alg. 2. The pseudocode for this procedure is shown in the pseudocode of Alg. 3.

***Remark* 15.** *Algorithm 4 is also related to the property testing algorithm by [30] (their Algorithm 2) which test whether a discrete-output mechanism is at least (ε, δ)-DP for pre-defined ε and δ. Their algorithm rejects at least with probability 2/3 in case the δ-estimate for given ε is more than α-far from the true estimate where α is a pre-defined parameter.*

The sample complexity of Algorithm 2 of [30] as a function of k and α is asymptotically the same as in our method, as implied by our Lemmas 10 and 11. However, the sample complexity given by their Theorem 14 is stated only as a big-$\mathcal{O}$ result. We provide an explicit bound for the required sample size which leads to high-probability confidence intervals for the estimates, providing a practical method for evaluating an estimate and a confidence interval for the hockey-stick divergence. This allows obtaining a high-confidence lower bound for the privacy profile and an upper bound for the trade-off functions. Also, one clear difference is that we introduce the binning for continuous-output distributions which allows auditing, e.g., ML model training algorithms.

V. Relation to Existing Work on Threshold Membership Auditing

As we next show, the commonly considered membership inference attacks can be seen as a special case of our auditing method. Suppose we have two distributions P and Q which give distributions to some score values originating from a dataset with and without a given sample z (e.g., as in the black-box setting described in Eq. (IV.1)), and suppose that we are given some a fixed threshold τ. We infer that a sample originates from a dataset with the sample z in case it is below τ. This gives the true positive ratio (TPR) and false positive ratio (FPR)

$$\begin{aligned} \text{TPR} &= \mathbb{P}_{z\sim P}\left(z \le \tau\right), \\ \text{FPR} &= \mathbb{P}_{z\sim Q}\left(z \le \tau\right). \end{aligned} \tag{V.1}$$

We can interpret the (ε, δ)-estimates given by this threshold membership inference as the (ε, δ)-distance between two-bin approximations (bins defined by the parameter $\tau \in \mathbb{R}$ dividing the real line into two bins) of the distributions P and Q.

Let $\widehat{P}$ and $\widehat{Q}$ denote the two-bin histogram approximations of P and Q, respectively, where the bins are determined by the threshold parameter τ. The following lemma shows that the (ε, δ)-distance between $\widehat{P}$ and $\widehat{Q}$ exactly matches with the expression commonly used for the empirical ε-values.

Lemma 16. *Consider the discrete-valued distributions $\widehat{P} =: (\text{TPR}, \text{FNR})$ and $\widehat{Q} =: (\text{FPR}, \text{TNR})$. Assume $\text{TPR} \ge \text{FPR}$. Let $\delta \in [0, 1]$ such that*

$$\max\{H_{\mathrm{e}^{\widehat{\varepsilon}}}\big(\widehat{P} || \widehat{Q}\big), H_{\mathrm{e}^{\widehat{\varepsilon}}}\big(\widehat{Q} || \widehat{P}\big)\} = \delta$$

for some $\widehat{\varepsilon} \ge 0$. Then,

$$\widehat{\varepsilon} = \max\left\{\log\frac{\text{TPR} - \delta}{\text{FPR}}, \log\frac{\text{TNR} - \delta}{\text{FNR}}\right\}. \tag{V.2}$$

By the post-processing property of DP, we directly get the following corollary.

Corollary 17. *Suppose the underlying mechanism $\mathcal{M}$ is (ε, δ)-DP for some $\varepsilon \ge 0$ and $\delta \in [0, 1]$. Then, asymptotically, $\widehat{\varepsilon} = \max\left\{\log\frac{\text{TPR}-\delta}{\text{FPR}}, \log\frac{\text{TNR}-\delta}{\text{FNR}}, 0\right\}$ gives a lower bound for ε.*

The ε-estimate of Eq.(V.2) was originally considered in [12] and it is the commonly used formula for obtaining (ε, δ)-DP lower bounds via success rates of membership inference attacks. It follows also directly from the characterization given

in [18]. Our novelty is to generalize the auditing based on Eq.(V.2) such that we consider histograms with more than two bins, and instead of estimating TPRs and FPRs, we estimate the relative frequencies of the scores hitting each of the bins and then measure the (ε, δ)-distance between the approximated distributions corresponding to the score distributions of the two auditing sets.

As an example, suppose that we have a division of an interval into 2^k bins, $k \in \mathbb{N}$, denoted D_k, such that half of the bins are right to the threshold τ and half of them are left to τ and suppose the division D_{k+1} is obtained by dividing each interval of D_k in half. Then, by the post-processing property, the asymptotic distributions P_k and Q_k obtained using the histogram D_k can be seen as a post-processing of the distributions P_{k+1} and Q_{k+1} (simply sum up the probabilities of adjacent bins) and therefore the finer the division the closer the (ε, δ)-estimates get to the actual (ε, δ)-distance between the distributions of the scores.

Following the discussion of [35], we see that our approach is also related to the exposure metric defined in [36]. Given n auditing training samples $\{c_i\}_{i=1}^n$ and n auditing test samples $\{r_i\}_{i=1}^n$, [36] defines the exposure of a sample c_i via its rank

$$\text{Exposure}(c_i) = \log_2 n - \log_2 \text{rank}(c_i, \{r_i\}_{i=1}^n),$$

where $\text{rank}(c_i, \{r_i\}_{i=1}^n)$ equals the number of auditing test samples with loss smaller then the loss of c_i. As shown in [35], a reasonable approximation for the the expected exposure is given by the threshold membership inference (i.e., a two-bin histogram approximation desribed above) with threshold parameter $\tau = \ell_{\text{median}}$, where ℓ_{median} is the median value of the losses of the auditing training samples, i.e., the median of $\{\ell(c_i)\}_{i=1}^n$. This leads to the ε-estimate given by Eq. (V.2) with

$$\text{TPR} = \mathbb{P}_{x \sim \{c_i\}_{i=1}^n} \left(\ell(x) < \ell_{\text{median}}\right)$$

and

$$\text{FPR} = \mathbb{P}_{x \sim \{r_i\}_{i=1}^n} \left(\ell(x) < \ell_{\text{median}}\right).$$

We remark that the auditing training and test sample scores would generally need to be independent to conclude that the estimate of Eq.(V.2) gives a lower bound for the actual (ε, δ)-DP guarantees.

VI. Numerical Toy Examples

Next, we give numerical examples to illustrate the histogram-based estimation presented in Section IV.

A. Numerical Example: Estimating TV Distance Between Two Gaussians

We illustrate our approach for estimating the (ε, δ)-distance between two one-dimensional Gaussians. The example also illustrates the effect of the bin size. Let $\sigma > 0$. We draw n random samples $x_1, \ldots, x_n$ from the distribution $P \sim \mathcal{N}(0, \sigma^2)$ and n samples $y_1, \ldots, y_n$ from the distribution $Q \sim \mathcal{N}(z, \sigma^2)$. We know that P and Q are $\big(\varepsilon, \delta(\varepsilon)\big)$-distinguishable, where $\delta(\varepsilon)$ denotes the privacy profile of the Gaussian mechanism with noise scale σ and sensitivity 1 and in particular we know by Lemma 3 that the total variation distance $\text{TV}(P, Q)$ is given by

$$\delta(0) = 2 \cdot \left(1 - \Phi\left(\frac{1}{2\sigma}\right)\right), \tag{VI.1}$$

where Φ denotes the CDF of the standard univariate Gaussian distribution. We determine a and b such that x_i's and y_i's are inside the interval $[a, b]$ with high probability and fix the number of bins $N \in \mathbb{N}$, and carry out the TV distance estimation using Algorithm 3 (i.e., using $\varepsilon = 0$). Figure 4 illustrates the accuracy of the TV distance estimation as the number of bins N varies.

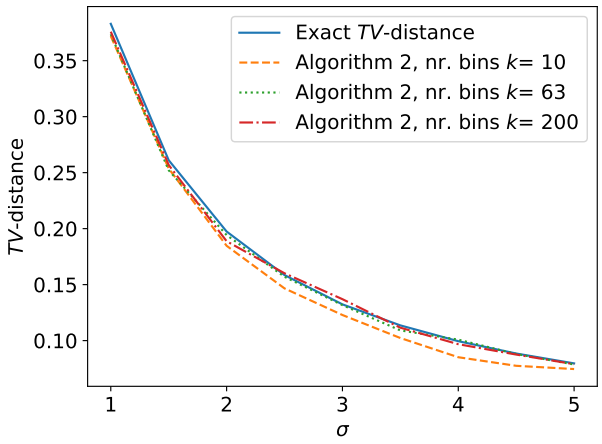

Fig. 4. Exact TV distance $TV(P, Q)$ and the TV distance approximated using Alg. 3 for different values of σ, when $n = 50000$. The bin width h_n set using Eq. (IV.3) gives $k = 63$ bins.

As also suggested by the upper bound of Lemma 11, Alg. 3 is most accurate for the hockey-stick divergence with $\varepsilon = 0$, i.e., for the TV-distance. We show in Appendix J also results for estimating the hockey-stick divergence with $\varepsilon = 1$. Those results also illustrate that there is possible room for improvement in the bound of Lemma 11. Improving the bound of Lemma 11 would directly improve the sample complexity of obtaining high-confidence lower bounds with Alg. 3.

B. Numerical Example: Auditing the Subsampled Gaussian Mechanism

In [8] an open problem of how to accurately audit the subsampled Gaussian mechanism is posed. The concrete example of [8] is given by the pair of distributions $P \sim q \cdot \mathcal{N}(1, \sigma^2) + (1 - q) \cdot \mathcal{N}(0, \sigma^2)$ and $Q \sim \mathcal{N}(0, \sigma^2)$ with the parameter values $q = 1/4$ and $\sigma = 0.3$. Figure 5 replicates the experimental results given in [8], however, it includes the trade-off function estimated using Alg. 3. The accurate trade-off curve is computed using numerical privacy accounting method of [20] and Alg. 1. We see that the histogram-based method is able to accurately estimate this trade-off curve.

C. Numerical Example: Auditing the Laplace Mechanism

The Laplace mechanism adds Laplace distributed noise to a function with limited L_1-sensitivity, and the (ε, δ)-DP privacy guarantees are determined by a dominating pair of

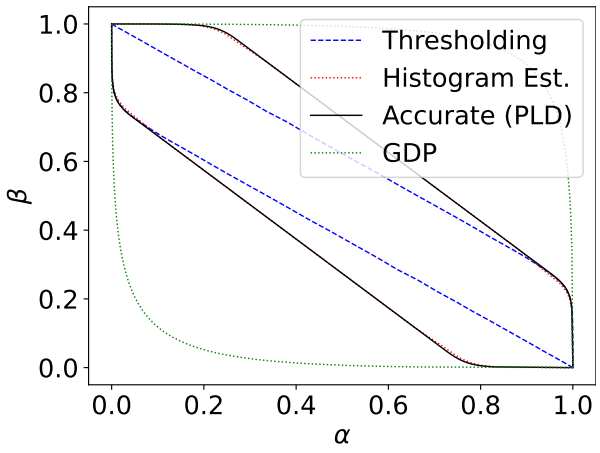

Fig. 5. Estimating the trade-off function of the subsampled Gaussian mechanism with $q = \frac{1}{4}$ and $\sigma = 0.3$. The histogram-based auditing method is able to accurately estimate the trade-off function without any information about P and Q. We sample $n = 10^6$ samples from both P and Q.

distributions $P \sim \mathrm{Lap}(0, \lambda)$ and $Q \sim \mathrm{Lap}(\Delta_1, \lambda)$, where Δ_1 is the L_1-norm sensitivity of the underlying function and λ denotes the noise scale. In [24] it is shows that the accurate trade-off function of the Laplace mechanism is given by

$$T\big(\mathrm{Lap}(0,\lambda), \mathrm{Lap}(\Delta_1,\lambda)\big)(\alpha) = F\big(F^{-1}(1-\alpha) - \mu\big)$$

where $\mu = \lambda/\Delta_1$ and F denotes the CDF of $\mathrm{Lap}(0, 1)$ (see Appendix for the exact analytical form).

Figure 6 shows that the binning-based method is able to accurately estimate the exact trade-off function without any information about the underlying distributions. To compute the trade-off functions, we use $k = 10^5$ samples and 100 bins.

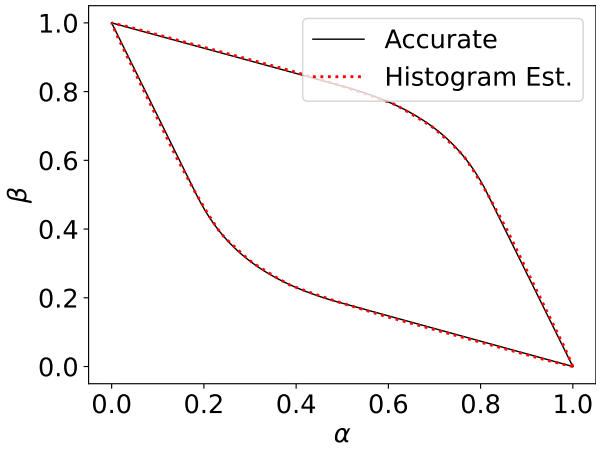

Fig. 6. Estimation of the trade-off function of the Laplace mechanism with noise scale $\lambda = 1.0$ and sensitivity $\Delta_1 = 1.0$. The histogram-based auditing method is able to accurately estimate the trade-off function without any information about P and Q. We sample $n = 10^5$ samples from both P and Q.

VII. Experiments on Neural Networks

We illustrate the effectiveness of our method in ML model auditing in both black-box and white-box settings.

A. Experiments on Black-Box Auditing

To obtain the black-box auditing samples, we consider the method used in [8] and depicted in Alg. 5. By using an auditing sample z', we draw n samples of the loss function evaluated on a model that is trained on a dataset D that does not include z' and n samples on dataset $D' = D \cup z'$. In all experiments, we draw $n = 10^5$ samples for both D and D'.

Algorithm 5 Black-box auditing method for DP-SGD.

Input: Training dataset D, loss function ℓ, canary input (x', y'), number of observations T.
Observations: $O \to [], O' \to []$.
Set: $D' = D \cup \{(x', y')\}$.
for $t \in [n]$: **do**
  $\theta \to \mathcal{M}(D)$ (DP-SGD on the dataset D).
  $\theta' \to \mathcal{M}(D')$ (DP-SGD on the dataset D').
  $O[t] \to \ell\big(\theta, (x', y')\big)$.
  $O'[t] \to \ell\big(\theta', (x', y')\big)$.
end for
return O, O'.

We first consider a one hidden-layer feedforward network for MNIST classification [37], with hidder-layer width 200. We minimize the cross-entropy loss, and the clipping constant C is set to 1.0. We train the models with a random subset of 1000 samples from the training split of MNIST. As a baseline method we consider the μ-GDP auditing method of [8] that uses threshold inference and Clopper–Pearson confidence intervals for the FPR and FNR estimates. We choose the threshold parameter as the mean of all loss values. We remark the μ-GDP auditing method gives rigorous (ε, δ)-bounds only in case the two distributions P and Q are equal-variance Gaussians, and in that case the μ-estimate is independent of the threshold (in the limit as the number of samples grows). We also use our histogram-based method of Alg. 4 and set the number of bins $k = 15$ which, by Lemma. 10, gives approximately 99.9% confidence intervals for the TV distance.

Figures 7 and 8 show the histograms of the losses and the resulting trade-off functions, when the additional sample z' is chosen randomly from the rest of the MNIST training data and the models are trained using the Adam optimizer [38] with the initial learning rate 0.001 for 50 epochs. Here the μ-GDP accounting is justified and actually gives similar estimates and high-confidence upper bounds for the trade-off functions as Alg. 4. Notice that the reported trade-off curve upper bounds correspond to high-confidence lower bounds for the DP parameters which explains the legends.

Figures 9 and 10 show the histograms of the losses and the resulting trade-off functions, when $z' = (x', y')$ is chosen such that the feature x' is a random vector scaled to have similar norm as other samples and the models are trained using DP-SGD for 5 epochs with the learning rate 0.1. The histograms are far from being equal-variance Gaussians, and thus μ-GDP auditing with a single FNR-FPR estimate does not necessarily lead to valid (ε, δ)-DP lower bounds. However, we

show the resulting trade-off functions for comparison, and see that Alg. 4 gives much lower trade-off curves (meaning higher (ε, δ)-lower bounds).

We consider a similar setting for the binary classification dataset "Synthetic H" considered in [39] which has feature dimension 5000. We consider a one hidden-layer feedforward network with hidder-layer width 80 and minimize the cross-entropy loss, and the clipping constant C is set to 1.0. We choose as training data random 1000 samples and $z' = (x', y')$ such that the feature x' is a random vector scaled to have a similar norm as rest of the samples. Figures 11 and 12 show that the histograms are very far from being Gaussians and that Alg. 4 gives much more accurate trade-off functions.

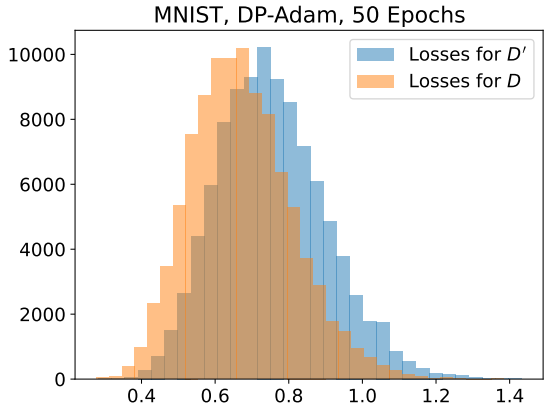

Fig. 7. Histograms of the loss function values $\ell(\theta, x)$ at the end of the training, when the model θ is trained using a) a dataset D and b) dataset $D' = D \cup \{(x', y')\}$. The empirical distributions look like Gaussians with approximately an equal variance.

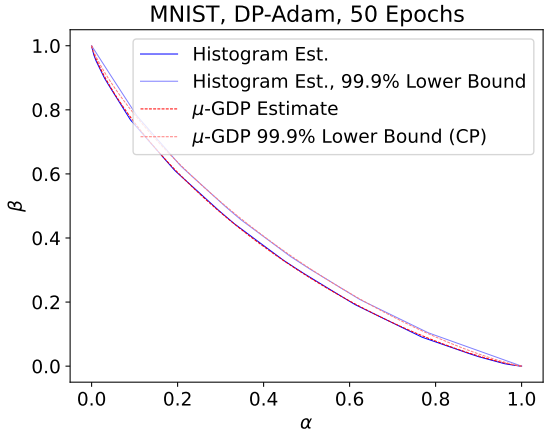

Fig. 8. Estimated 99.9 % - upper bound trade-off curves obtained using a) the thresholding and μ-GDP and b) the histogram-based method using Alg. 3. The empirical distributions of the losses are almost like Gaussians, which explains the fact that μ-GDP auditing gives almost equally good estimates.

B. Experiments on White-Box Auditing

Lastly, we propose a heuristic white-box auditing method to estimate the privacy loss distributions of the underlying model training mechanism that can also be used to obtain estimates

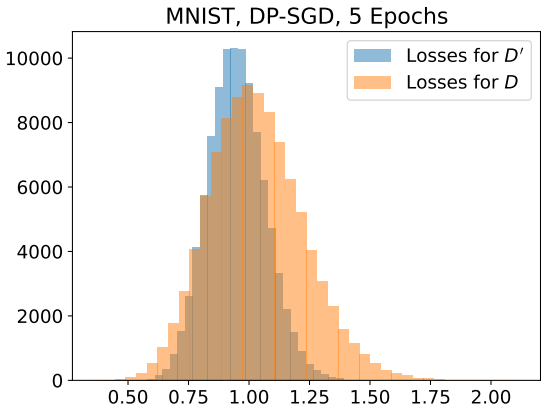

Fig. 9. Histograms of the loss function values $\ell(\theta, x)$ at the end of the training, when the model θ is trained using a) a dataset D and b) dataset $D' = D \cup \{(x', y')\}$. The empirical distribution deviate markedly from Gaussians.

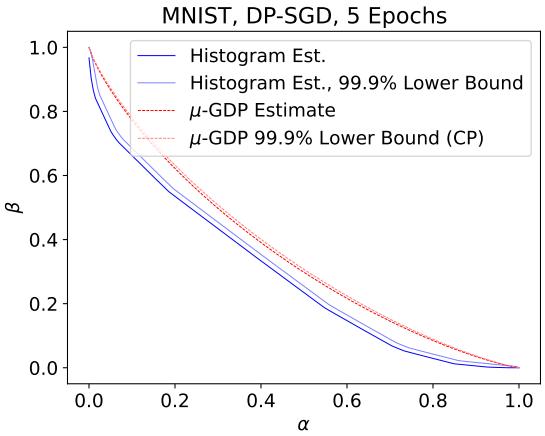

Fig. 10. Estimated 99.9 % - upper bound trade-off curves obtained using a) the thresholding and μ-GDP and b) the histogram-based method using Alg. 3. The empirical distribution deviate markedly from Gaussians, which explains the fact that μ-GDP auditing fails to capture the accurate lower bound.

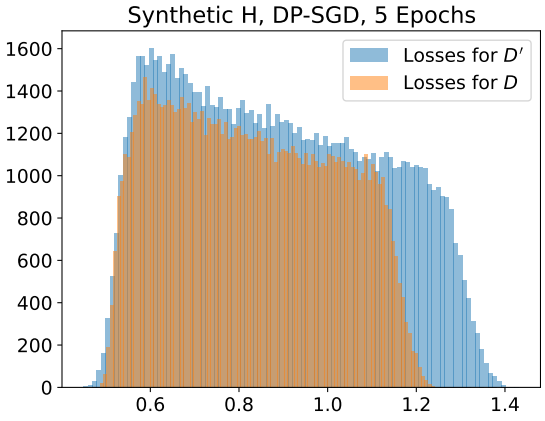

Fig. 11. Histograms of the loss function values $\ell(\theta, x)$ at the end of the training, when the model θ is trained using a) a dataset D and b) dataset $D' = D \cup \{(x', y')\}$. The empirical distribution deviate markedly from Gaussians.

for compositions, without any a priori information about the

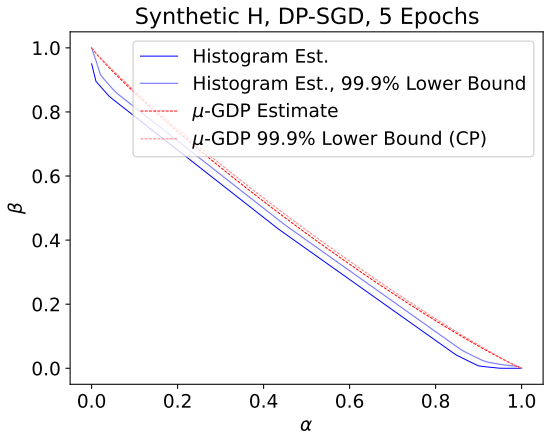

Fig. 12. Estimated 99.9 % - upper bound trade-off curves obtained using a) the thresholding and μ-GDP and b) the histogram-based method using Alg. 3. The empirical distribution deviate markedly from Gaussians, which explains the fact that μ-GDP auditing fails to capture the accurate lower bound.

parameters of the training algorithm.

In [8] white-box auditing is carried out using Alg. I given in Appendix, such that the canary gradient is added with probability 1 at each iteration, i.e., the auditing neglects the effect of subsampling. Having an estimate of the μ-GDP parameter gives then estimates of the so-called dominating pairs of distributions for the subsampled Gaussian mechanism in case the subsampling ratio q is known. Using these, one can construct numerical privacy loss distributions (PLDs) and using FFT-based accounting methods [20], [6] furthermore compute empirical $\delta(\varepsilon)$-bounds also for compositions.

We consider the same setting of white-box auditing, however, we include the canaries with the same probability as other gradients, and we carry out numerical estimation of the PLDs by estimating the distributions of inner product values in Alg. I using histograms, i.e., we calculate the discrete probabilities $\widehat{P}$ and $\widehat{Q}$ as in Alg. 3, and then get the discrete-valued PLDs $\omega_{\widehat{P}/\widehat{Q}}$ and $\omega_{\widehat{Q}/\widehat{P}}$, such that for $j \in [k]$,

$$\mathbb{P}\left(\omega_{\widehat{P}/\widehat{Q}} = \log \frac{\widehat{P}_j}{\widehat{Q}_j}\right) = \widehat{P}_j$$

and

$$\mathbb{P}\left(\omega_{\widehat{Q}/\widehat{P}} = \log \frac{\widehat{Q}_j}{\widehat{P}_j}\right) = \widehat{Q}_j.$$

We approximate the PLDs of a c-fold composition of the mechanism then by PLDs $\omega^c_{\widehat{P}/\widehat{Q}}$ and $\omega^c_{\widehat{Q}/\widehat{P}}$ that are given by c-fold self-convolutions of distributions $\omega_{\widehat{P}/\widehat{Q}}$ and $\omega_{\widehat{Q}/\widehat{P}}$, respectively, and obtain an estimate $\widetilde{\delta}(\varepsilon)$ of the privacy profile $\delta(\varepsilon)$ of the c-fold composition of the mechanism as

$$\begin{aligned}\widetilde{\delta}(\varepsilon) = \max\{&\mathbb{E}_{s\sim\omega^c_{\widehat{P}/\widehat{Q}}}[1-\mathrm{e}^{\varepsilon-s}]_+, \\ &\mathbb{E}_{s\sim\widetilde{\omega}^c_{\widehat{Q}/\widehat{P}}}[1-\mathrm{e}^{\varepsilon-s}]_+\}.\end{aligned} \tag{VII.1}$$

The convolutions and the integrals (VII.1) are evaluated using the numerical method of [20].

Figure 13 shows results for a one-dimensional toy problem, where $P \sim q\cdot\mathcal{N}(1,\sigma^2)+(1-q)\cdot\mathcal{N}(0,\sigma^2)$ and $Q \sim \mathcal{N}(0,\sigma^2)$ with the parameter values $q = 1/2$ and $\sigma = 2.0$. We draw $n = 10^6$ random samples from both P and Q. We compute the $\delta(\varepsilon)$-bounds for $c = 10$ compositions and the accurate bounds are computed using the method of [6].

Similarly, Fig. 14 shows results for white-box auditing using Alg. I for the feedforward neural network, using a random subset of 1000 samples from the training split of the MNIST dataset. We use random normally distributed canaries and draw a new random canary vector at each step. We train 10^5 models, each for 10 epochs, with a batch size of 500 and noise scale $\sigma = 2.0$. We concatenate all the scores, giving in total $n = 10^6$ samples from both P_S and Q_S from which the histogram-estimates $\widehat{P}$ and $\widehat{Q}$ are constructed.

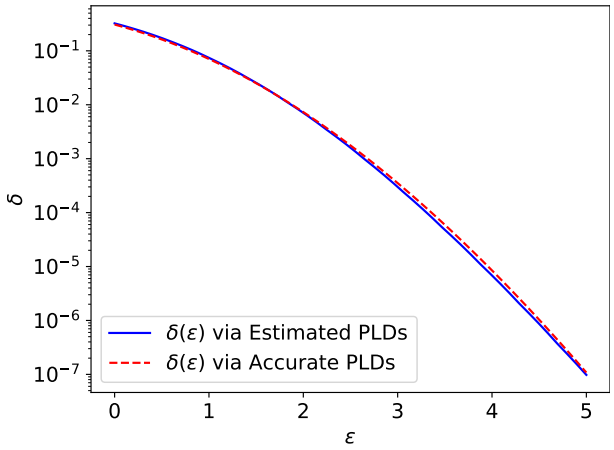

Fig. 13. Comparison of the accurate privacy profile $\delta(\varepsilon)$ and the estimated privacy profile that is computed using the discrete distributions $\widetilde{P}$ and $\widetilde{Q}$ obtained from the histogram estimates of P and Q. Here $P \sim q\cdot\mathcal{N}(1,\sigma^2)+(1-q)\cdot\mathcal{N}(0,\sigma^2)$ and $Q \sim \cdot\mathcal{N}(0,\sigma^2)$, where $\sigma = 2.0$ and $q = 0.5$.

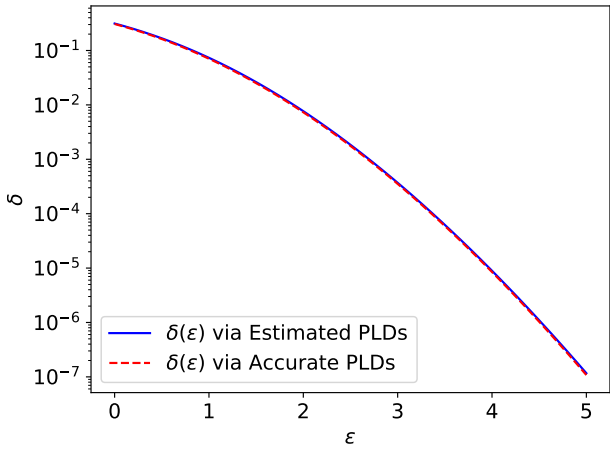

Fig. 14. Comparison of the accurate privacy profile $\delta(\varepsilon)$ and the estimated privacy profile that is computed using the discrete distributions $\widetilde{P}$ and $\widetilde{Q}$ obtained from the histogram estimates of P and Q. Here samples from P and Q are obtained using inner products with random canaries (Alg. I).

VIII. Conclusions

We have proposed a simple and practical technique to compute empirical estimates of DP privacy guarantees that

does not require any a priori information about the underlying mechanism. We have shown that our method can be seen as a generalization of the existing threshold membership inference auditing methods. One limitation of our method is that the reported ε-estimates in the white-box setting are heuristic and we do not provide confidence intervals for them. To improve our methods, it would be important to find tighter confidence intervals for estimates of multinomial distributions (see, e.g., [40]). We leave this however for future work. To increase the computational efficiency, it will also be interesting to find conditions under which we can circumvent the assumption of the independence of the auditing score values when carrying out one-shot estimation and possibly give confidence intervals for ε-lower bounds in that case.

REFERENCES

[1] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. TCC 2006*, 2006, pp. 265–284. [Online]. Available: http://dx.doi.org/10.1007/11681878_14

[2] S. Song, K. Chaudhuri, and A. D. Sarwate, "Stochastic gradient descent with differentially private updates," in *2013 IEEE global conference on signal and information processing*. IEEE, 2013, pp. 245–248.

[3] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318.

[4] A. Koskela, J. Jälkö, and A. Honkela, "Computing tight differential privacy guarantees using FFT," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2560–2569.

[5] Y. Zhu, J. Dong, and Y.-X. Wang, "Optimal accounting of differential privacy via characteristic function," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 4782–4817.

[6] S. Gopi, Y. T. Lee, and L. Wutschitz, "Numerical composition of differential privacy," in *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[7] F. Tramer, A. Terzis, T. Steinke, S. Song, M. Jagielski, and N. Carlini, "Debugging differential privacy: A case study for privacy auditing," *arXiv preprint arXiv:2202.12219*, 2022.

[8] M. Nasr, J. Hayes, T. Steinke, B. Balle, F. Tramèr, M. Jagielski, N. Carlini, and A. Terzis, "Tight auditing of differentially private machine learning," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 1631–1648.

[9] S. Maddock, A. Sablayrolles, and P. Stock, "Canife: Crafting canaries for empirical privacy measurement in federated learning," in *The Eleventh International Conference on Learning Representations*, 2023.

[10] G. Andrew, P. Kairouz, S. Oh, A. Oprea, H. B. McMahan, and V. M. Suriyakumar, "One-shot empirical privacy estimation for federated learning," in *The Twelfth International Conference on Learning Representations*, 2024.

[11] M. Jagielski, J. Ullman, and A. Oprea, "Auditing differentially private machine learning: How private is private SGD?" *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 205–22 216, 2020.

[12] M. Nasr, S. Songi, A. Thakurta, N. Papernot, and N. Carlin, "Adversary instantiation: Lower bounds for differentially private machine learning," in *2021 IEEE Symposium on security and privacy (SP)*. IEEE, 2021, pp. 866–882.

[13] K. Pillutla, G. Andrew, P. Kairouz, H. B. McMahan, A. Oprea, and S. Oh, "Unleashing the power of randomization in auditing differentially private ml," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[14] T. Steinke, M. Nasr, and M. Jagielski, "Privacy auditing with one (1) training run," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[15] M. S. M. S. Annamalai and E. De Cristofaro, "Nearly tight black-box auditing of differentially private machine learning," *arXiv preprint arXiv:2405.14106*, 2024.

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

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

[18] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," in *International conference on machine learning*. PMLR, 2015, pp. 1376–1385.

[19] B. Balle, G. Barthe, and M. Gaboardi, "Privacy amplification by subsampling: Tight analyses via couplings and divergences," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[20] A. Koskela, J. Jälkö, L. Prediger, and A. Honkela, "Tight differential privacy for discrete-valued mechanisms and for the subsampled gaussian mechanism using FFT," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 3358–3366.

[21] B. Balle and Y.-X. Wang, "Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising," in *International Conference on Machine Learning*, 2018, pp. 394–403.

[22] B. Balle, G. Barthe, and M. Gaboardi, "Privacy profiles and amplification by subsampling," *Journal of Privacy and Confidentiality*, vol. 10, no. 1, 2020.

[23] L. Wasserman and S. Zhou, "A statistical framework for differential privacy," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 375–389, 2010.

[24] J. Dong, A. Roth, W. J. Su *et al.*, "Gaussian differential privacy," *Journal of the Royal Statistical Society Series B*, vol. 84, no. 1, pp. 3–37, 2022.

[25] S. Zanella-Béguelin, L. Wutschitz, S. Tople, A. Salem, V. Rühle, A. Paverd, M. Naseri, B. Köpf, and D. Jones, "Bayesian estimation of differential privacy," in *International Conference on Machine Learning*. PMLR, 2023, pp. 40 624–40 636.

[26] G. Kaissis, S. Kolek, B. Balle, J. Hayes, and D. Rueckert, "Beyond the calibration point: Mechanism comparison in differential privacy," in *Forty-first International Conference on Machine Learning*, 2024.

[27] P. Virtanen, R. Gommers, T. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright *et al.*, "Fundamental algorithms for scientific computing in python and scipy 1.0 contributors. scipy 1.0," *Nat. Methods*, vol. 17, pp. 261–272, 2020.

[28] C. L. Canonne, "A short note on learning discrete distributions," *arXiv preprint arXiv:2002.11457*, 2020.

[29] S. Kamath, A. Orlitsky, D. Pichapati, and A. T. Suresh, "On learning distributions from their samples," in *Conference on Learning Theory*. PMLR, 2015, pp. 1066–1100.

[30] A. C. Gilbert and A. McMillan, "Property testing for differential privacy," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2018, pp. 249–258.

[31] V. Doroshenko, B. Ghazi, P. Kamath, R. Kumar, and P. Manurangsi, "Connect the dots: Tighter discrete approximations of privacy loss distributions," *Proceedings on Privacy Enhancing Technologies*, 2022.

[32] D. W. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605–610, 1979.

[33] M. Wand, "Data-based choice of histogram bin width," *The American Statistician*, vol. 51, no. 1, pp. 59–64, 1997.

[34] K. H. Knuth, "Optimal data-based binning for histograms," *arXiv preprint physics/0605197*, 2013.

[35] M. Jagielski, "A note on interpreting canary exposure," *arXiv preprint arXiv:2306.00133*, 2023.

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

[37] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[39] R. Iyengar, J. P. Near, D. Song, O. Thakkar, A. Thakurta, and L. Wang, "Towards practical differentially private convex optimization," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 299–316.

[40] D. Chafai and D. Concordet, "Confidence regions for the multinomial parameter with small sample size," *Journal of the American Statistical Association*, vol. 104, no. 487, pp. 1071–1079, 2009.

[41] T. Cai, J. Fan, and T. Jiang, "Distributions of angles in random packing on spheres," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1837–1864, 2013.
[42] T. T. Cai and T. Jiang, "Phase transition in limiting distributions of coherence of high-dimensional random matrices," *Journal of Multivariate Analysis*, vol. 107, pp. 24–39, 2012.
[43] L. Devroye, A. Mehrabian, and T. Reddad, "The total variation distance between high-dimensional gaussians with the same mean," *arXiv preprint arXiv:1810.08693*, 2018.
[44] S. Asoodeh, M. Aliakbarpour, and F. P. Calmon, "Local differential privacy is equivalent to contraction of an f-divergence," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021, pp. 545–550.

APPENDIX

A. Numerical Optimization to Find Accurate μ-GDP parameter

We can numerically compute the μ-GDP parameter such that using the privacy profile

$$\delta(\varepsilon) = \max\{H_{\mathrm{e}^\varepsilon}\big(P||Q\big), H_{\mathrm{e}^\varepsilon}\big(Q||P\big)\}$$

where $P \sim q\cdot\mathcal{N}(1,\sigma^2)+(1-q)\cdot\mathcal{N}(0,\sigma^2)$ and $Q \sim \cdot\mathcal{N}(0,\sigma^2)$, we find a value of σ for the Gaussian mechanism such that we search for a point, where the tangent and value of the privacy profiles $\delta(\varepsilon)$ and $\delta_{\mathrm{Gauss},\mu}(\varepsilon)$ are equal. To this end, we solve numerically the problem

$$\mathrm{argmin}_\sigma \, \min_\varepsilon \left\| \begin{bmatrix} \delta(\varepsilon) \\ \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\delta(\varepsilon) \end{bmatrix} - \begin{bmatrix} \delta_{\mathrm{Gauss},\sigma}(\varepsilon) \\ \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\delta_{\mathrm{Gauss},\sigma}(\varepsilon) \end{bmatrix} \right\|. \tag{A.1}$$

Given a numerical solution $\widehat{\sigma}$, the μ-parameter is then given by $\mu = 1/\widehat{\sigma}$. Figure 15 illustrates the result of this optimization.

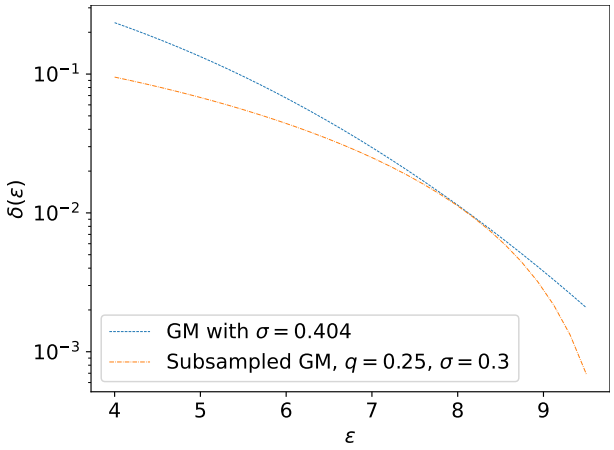

Fig. 15. Adjusting the μ-GDP parameter for the pair of distributions $P \sim q\cdot\mathcal{N}(1,\sigma^2)+(1-q)\cdot\mathcal{N}(0,\sigma^2)$ and $Q \sim \mathcal{N}(0,\sigma^2)$, where $\sigma = 0.3$ and $q = 0.25$. The tight μ-GDP is given by $\mu = 1/\sigma$, where σ is the minimal value such that the the privacy profile of the Gaussian mechanism with noise scale σ is under the privacy profile $h(\alpha) = H_\alpha(P,Q)$. This value can be found, e.g., by solving the problem (A.1).

B. Proof of Lemma 11

Lemma A.1. *Denote P, Q probability distributions on the same probability space. Suppose*

$$\mathrm{TV}(P,\widetilde{P}) \leq \tau$$

and

$$\mathrm{TV}(Q,\widetilde{Q}) \leq \tau$$

for some $\tau \geq 0$. Then, for all $\varepsilon \in \mathbb{R}$,

$$H_{\mathrm{e}^\varepsilon}(P,Q) \leq H_{\mathrm{e}^\varepsilon}(\widetilde{P},\widetilde{Q}) + (1+\mathrm{e}^\varepsilon)\cdot\tau.$$

Proof. Using the inequality $\leq [a]_+ + [b]_+$ that holds for all $a, b \in \mathbb{R}$, we have that

$$\begin{aligned}
H_{\mathrm{e}^\varepsilon}(P,Q) &= \int [P(t) - \mathrm{e}^\varepsilon Q(t)]_+ \,\mathrm{d}t \\
&= \int [P(t) - \widetilde{P}(t) + \mathrm{e}^\varepsilon(\widetilde{Q}(t) - Q(t)) \\
&\quad + \widetilde{P}(t) - \mathrm{e}^\varepsilon \widetilde{Q}(t)]_+ \,\mathrm{d}t \\
&\leq \int [P(t) - \widetilde{P}(t)]_+ \,\mathrm{d}t \\
&\quad + \mathrm{e}^\varepsilon \int [\widetilde{Q}(t) - Q(t)]_+ \,\mathrm{d}t \\
&\quad + \int [\widetilde{P}(t) - \mathrm{e}^\varepsilon \widetilde{Q}(t)]_+ \,\mathrm{d}t \\
&= TV(P,\widetilde{P}) + \mathrm{e}^\varepsilon TV(\widetilde{Q},Q) + H_{\mathrm{e}^\varepsilon}(\widetilde{P},\widetilde{Q}) \\
&\leq (1+\mathrm{e}^\varepsilon)\cdot\tau + H_{\mathrm{e}^\varepsilon}(\widetilde{P},\widetilde{Q}).
\end{aligned}$$

□

C. Trade-Off Function for the Laplace Mechanism

The accurate trade-off function of the Laplace mechanism is given in Lemma A.6 of [24]

$$T\big(\mathrm{Lap}(0,1), \mathrm{Lap}(\mu,1)\big)(\alpha) = \begin{cases} 1 - \mathrm{e}^\mu \alpha, & \alpha < \mathrm{e}^{-\mu}/2. \\ \mathrm{e}^{-\mu}/4\alpha, & \mathrm{e}^{-\mu}/2 \leq \alpha \leq 1/2, \\ \mathrm{e}^{-\mu}(1-\alpha), & \alpha \geq 1/2, \end{cases}$$

D. Illustration: Conversion Between the Noise Parameter and TV Distance for the Gaussian Mechanism

Figure 16 shows the TV distance $TV(P,Q)$ when $P \sim q\cdot\mathcal{N}(1,\sigma^2) + (1-q)\cdot\mathcal{N}(0,\sigma^2)$ and $Q \sim \mathcal{N}(0,\sigma^2)$ for three different values of q and for varying values of σ.

Using the conversion from the TV distance to σ, we can also convert the confidence intervals for the confidence interval of $TV(P,Q)$ in case we know the subsampling parameter q.

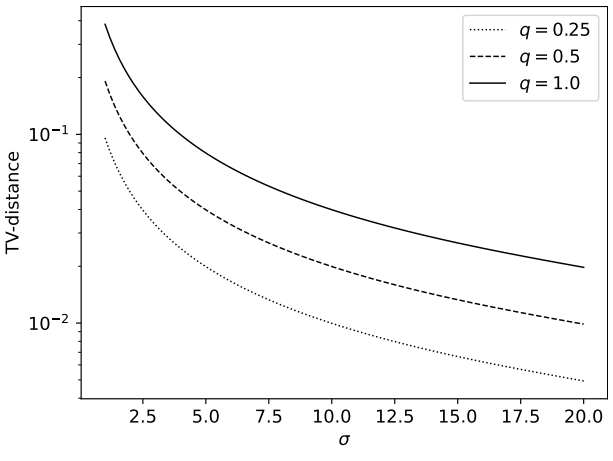

Fig. 16. Relationship between TV distance $TV(P,Q)$ and ε when $\delta = 10^{-5}$ for the Gaussian mechanism.

E. Proof of Lemma 16 and Cor. 17

Lemma A.2. *Consider the discrete-valued distributions* $\widehat{P} =: (\mathrm{TPR}, \mathrm{FNR})$ *and* $\widehat{Q} =: (\mathrm{FPR}, \mathrm{TNR})$. *Assume* $\mathrm{TPR} \geq \mathrm{FPR}$. *Let* $\delta \in [0,1]$ *such that*

$$\max\{H_{\mathrm{e}^{\widehat{\varepsilon}}}(\widehat{P}||\widehat{Q}), H_{\mathrm{e}^{\widehat{\varepsilon}}}(\widehat{Q}||\widehat{P})\} = \delta$$

for some $\widehat{\varepsilon} \geq 0$. *Then,*

$$\widehat{\varepsilon} = \max\left\{\log\frac{\mathrm{TPR}-\delta}{\mathrm{FPR}}, \log\frac{\mathrm{TNR}-\delta}{\mathrm{FNR}}\right\}.$$

Proof. Using also the assumptions $\widehat{\varepsilon} \geq 0$ and $\mathrm{TPR} \geq \mathrm{FPR}$, i.e., $\mathrm{TNR} \geq \mathrm{FNR}$, we have that

$$\begin{aligned} H_{\mathrm{e}^{\widehat{\varepsilon}}}(\widehat{P}||\widehat{Q}) &= [\mathrm{TPR} - \mathrm{e}^{\widehat{\varepsilon}}\mathrm{FPR}]_+ + [\mathrm{FNR} - \mathrm{e}^{\widehat{\varepsilon}}\mathrm{TNR}]_+ \\ &= [\mathrm{TPR} - \mathrm{e}^{\widehat{\varepsilon}}\mathrm{FPR}]_+ \end{aligned}$$

and

$$\begin{aligned} H_{\mathrm{e}^{\widehat{\varepsilon}}}(\widehat{Q}||\widehat{P}) &= [\mathrm{FPR} - \mathrm{e}^{\widehat{\varepsilon}}\mathrm{TPR}]_+ + [\mathrm{TNR} - \mathrm{e}^{\widehat{\varepsilon}}\mathrm{FNR}]_+ \\ &= [\mathrm{TNR} - \mathrm{e}^{\widehat{\varepsilon}}\mathrm{FNR}]_+. \end{aligned}$$

Therefore,

$$\begin{aligned} &\max\{H_{\mathrm{e}^{\widehat{\varepsilon}}}(\widehat{P}||\widehat{Q}), H_{\mathrm{e}^{\widehat{\varepsilon}}}(\widehat{Q}||\widehat{P})\} \\ &\quad = \max\{\mathrm{TPR} - \mathrm{e}^{\widehat{\varepsilon}}\mathrm{FPR}, \mathrm{TNR} - \mathrm{e}^{\widehat{\varepsilon}}\mathrm{FNR}\}. \end{aligned} \tag{A.2}$$

Setting the right-hand side of Eq. (A.2) equal to δ and solving for $\widehat{\varepsilon}$ shows the claim.

□

Corollary A.3. *Suppose the underlying mechanism* $\mathcal{M}$ *is* (ε,δ)*-DP for some* $\varepsilon \geq 0$ *and* $\delta \in [0,1]$. *Then, asymptotically,*

$$\widehat{\varepsilon} = \max\left\{\log\frac{\mathrm{TPR}-\delta}{\mathrm{FPR}}, \log\frac{\mathrm{TNR}-\delta}{\mathrm{FNR}}, 0\right\}$$

gives a lower bound for ε.

Proof. Since $\widehat{P}$ and $\widehat{Q}$ are obtained by applying the same post-processing to P and Q, respectively, by the post-processing property of DP, we have that in case

$$\max\{H_{\mathrm{e}^{\widehat{\varepsilon}}}(\widehat{P}||\widehat{Q}), H_{\mathrm{e}^{\widehat{\varepsilon}}}(\widehat{Q}||\widehat{P})\} = \delta$$

for some $\widehat{\varepsilon} \geq 0$, $\widehat{\varepsilon} \leq \varepsilon$ for

$$\widehat{\varepsilon} = \max\left\{\log\frac{\mathrm{TPR}-\delta}{\mathrm{FPR}}, \log\frac{\mathrm{TNR}-\delta}{\mathrm{FNR}}\right\}.$$

Otherwise, $\widehat{\varepsilon} < 0$ and $\widehat{\varepsilon} = 0$ gives a lower bound for ε. □

F. Proof of Theorem 12

Theorem A.4. *Let* P *and* Q *be one-dimensional probability distributions with differentiable density functions* $P(x)$ *and* $Q(x)$, *respectively, and consider the histogram-based density estimation described above. Draw* n *samples both from* P *and* Q, *giving density estimators* $\widehat{P} = (\widehat{P}_1, \ldots, \widehat{P}_k)$ *and* $\widehat{Q} = (\widehat{Q}_1, \ldots, \widehat{Q}_k)$, *respectively. Let the bin width be chosen as*

$$h_n = \left(\frac{12}{\int_{\mathbb{R}} P'(x)^2 \,\mathrm{d}x + \int_{\mathbb{R}} Q'(x)^2 \,\mathrm{d}x}\right)^{\frac{1}{3}} n^{-\frac{1}{3}},$$

where $P'(x)$ denotes the derivative of the density function $P(x)$ and similarly for Q. Then, for any $\alpha \geq 0$, the numerical hockey-stick divergence $H_\alpha(\widehat{P}||\widehat{Q})$ convergences in expectation to $H_\alpha(P||Q)$ with rate $\mathcal{O}(n^{-1/3})$, i.e.,

$$\mathbb{E}\left|H_\alpha(\widehat{P}||\widehat{Q}) - H_\alpha(P||Q)\right| = \mathcal{O}(n^{-1/3}),$$

where the expectation is taken over the random draws for constructing $\widehat{P}$ and $\widehat{Q}$.

Proof. Define the piece-wise continuous functions $\widehat{P}(x)$ and $\widehat{Q}(x)$ such that $\widehat{P}(x) = \widehat{P}_\ell/h$, if $x \in \text{Bin}_\ell$ and similarly for $\widehat{Q}(x)$. To analyse the error in the hockey-stick divergence estimate we can use $\widehat{P}(x)$ and $\widehat{Q}(x)$ since

$$\begin{aligned} H_\alpha(\widehat{P}||\widehat{Q}) &= \sum_{\ell=1}^{N}[\widehat{P}_\ell - \alpha\cdot\widehat{Q}_\ell]_+ \\ &= \int_{\mathbb{R}}[\widehat{P}(x) - \alpha\cdot\widehat{Q}(x)]_+\,\mathrm{d}x \\ &= H_\alpha(\widehat{P}(x)||\widehat{Q}(x)). \end{aligned}$$

We can bound the divergence $H_\alpha(\widehat{P}||\widehat{Q})$ as follows (all the integrals are over the whole $\mathbb{R}$):

$$\begin{aligned} &H_\alpha(\widehat{P}||\widehat{Q}) \\ &\quad= \int[\widehat{P}(x) - \alpha\cdot\widehat{Q}(x)]_+\,\mathrm{d}x \\ &\quad= \int[\widehat{P}(x) - P(x) \\ &\qquad- \alpha\cdot(\widehat{Q}(x) - Q(x)) + P(x) - \alpha\cdot Q(x)]_+\,\mathrm{d}x \\ &\quad\leq \int\left|\widehat{P}(x) - P(x)\right|\,\mathrm{d}x \\ &\qquad+ \alpha\int\left|\widehat{Q}(x) - Q(x)\right|\,\mathrm{d}x \\ &\qquad+ \int[P(x) - \alpha\cdot Q(x)]_+\,\mathrm{d}x \\ &\quad\leq \sqrt{\int(\widehat{P}(x) - P(x))^2\,\mathrm{d}x} \\ &\qquad+ \alpha\sqrt{\int(\widehat{Q}(x) - Q(x))^2\,\mathrm{d}x} \\ &\qquad+ \int[P(x) - \alpha\cdot Q(x)]_+\,\mathrm{d}x \\ &\quad= \sqrt{\int(\widehat{P}(x) - P(x))^2\,\mathrm{d}x} \\ &\qquad+ \alpha\sqrt{\int(\widehat{Q}(x) - Q(x))^2\,\mathrm{d}x} + H_\alpha(P||Q) \\ &\quad\leq \max\{1,\alpha\}\cdot\left(\sqrt{\int(\widehat{P}(x) - P(x))^2\,\mathrm{d}x}\right. \\ &\qquad\left.+ \sqrt{\int(\widehat{Q}(x) - Q(x))^2\,\mathrm{d}x}\right) + H_\alpha(P||Q), \end{aligned} \tag{A.3}$$

where the first inequality follows from the fact that $[a+b]_+ \leq |a| + [b]_+$ for all $a, b \in \mathbb{R}$, and the second inequality follows from the Hölder inequality.

Similarly, carrying out the same calculation starting from $H_\alpha(P||Q)$, we have

$$\begin{aligned} &H_\alpha(P||Q) \\ &\quad= \int[P(x) - \alpha\cdot Q(x)]_+\,\mathrm{d}x \\ &\quad\leq \int\left|\widehat{P}(x) - P(x)\right|\,\mathrm{d}x \\ &\qquad+ \alpha\int\left|\widehat{Q}(x) - Q(x)\right|\,\mathrm{d}x \\ &\qquad+ \int[\widehat{P}(x) - \alpha\cdot\widehat{Q}(x)]_+\,\mathrm{d}x \\ &\quad= \int\left|\widehat{P}(x) - P(x)\right|\,\mathrm{d}x \\ &\qquad+ \alpha\int\left|\widehat{Q}(x) - Q(x)\right|\,\mathrm{d}x \\ &\qquad+ H_\alpha(\widehat{P}||\widehat{Q}) \\ &\quad\leq \max\{1,\alpha\}\cdot\left(\sqrt{\int(\widehat{P}(x) - P(x))^2\,\mathrm{d}x}\right. \\ &\qquad\left.+ \sqrt{\int(\widehat{Q}(x) - Q(x))^2\,\mathrm{d}x}\right) \\ &\qquad+ H_\alpha(\widehat{P}||\widehat{Q}) \end{aligned} \tag{A.4}$$

From the inequalities (A.3) and (A.4) it follows that

$$\begin{aligned} &\left|H_\alpha(\widehat{P}||\widehat{Q}) - H_\alpha(P||Q)\right| \\ &\leq \max\{1,\alpha\}\cdot\left(\left(\int(\widehat{P}(x) - P(x))^2\,\mathrm{d}x\right)^{\frac{1}{2}}\right. \\ &\left.+ \left(\int(\widehat{Q}(x) - Q(x))^2\,\mathrm{d}x\right)^{\frac{1}{2}}\right). \end{aligned} \tag{A.5}$$

Taking the expectation over the random draws from P and Q and applying Jensen's inequality to the the square root function, we get

$$\begin{aligned} &\mathbb{E}\left|H_\alpha(\widehat{P}||\widehat{Q}) - H_\alpha(P||Q)\right| \\ &\leq \max\{1,\alpha\}\cdot\left(\left(\int\mathbb{E}(\widehat{P}(x) - P(x))^2\,\mathrm{d}x\right)^{\frac{1}{2}}\right. \\ &\left.+ \left(\int\mathbb{E}(\widehat{Q}(x) - Q(x))^2\,\mathrm{d}x\right)^{\frac{1}{2}}\right) \\ &\leq \sqrt{2}\max\{1,\alpha\}\left(\int\mathbb{E}(\widehat{P}(x) - P(x))^2\,\mathrm{d}x\right. \\ &\left.+ \int\mathbb{E}(\widehat{Q}(x) - Q(x))^2\,\mathrm{d}x\right)^{\frac{1}{2}}, \end{aligned} \tag{A.6}$$

where the second inequality follows from the inequality $\sqrt{a} + \sqrt{b} \leq \sqrt{2}\sqrt{a+b}$ which holds for any $a, b \geq 0$. From the

derivations of Sec. 3 of [32] we have that

$$\begin{aligned}&\int \mathbb{E}(\widehat{P}(x)-P(x))^2\,\mathrm{d}x+\int \mathbb{E}(\widehat{Q}(x)-Q(x))^2\,\mathrm{d}x\\&=\frac{2}{n\cdot h}+\frac{1}{12}h^2\left[\int P'(x)^2\,\mathrm{d}x+\int Q'(x)^2\,\mathrm{d}x\right]\\&\quad+\mathcal{O}\left(\frac{1}{n}+h^3\right)\end{aligned}\tag{A.7}$$

Minimizing the first two terms on the right-hand side of (A.7) with respect to h gives the expression of h_n and furthermore, with this choice h_n, we have that

$$\int \mathbb{E}(\widehat{P}(x)-P(x))^2\,\mathrm{d}x+\int \mathbb{E}(\widehat{Q}(x)-Q(x))^2\,\mathrm{d}x=\mathcal{O}\left(n^{-\frac{2}{3}}\right).$$

which together with the inequality (A.6) shows that

$$\mathbb{E}\left|H_\alpha(\widehat{P}||\widehat{Q})-H_\alpha(P||Q)\right|=\mathcal{O}(n^{-\frac{1}{3}}).$$

□

G. One-Shot Estimation Using Random Canaries

We next show results for one-shot estimation of DP guarantees using random canary gradients. In the white-box setting, auditing of DP-SGD is often based on the assumption that the inserted canary gradient is (approximately) orthogonal against the rest of the per sample gradients [8]. The approach of [10] leverages the fact that this is approximately obtained by sampling the gradients randomly since the inner products between random unit vectors diminish as the dimension increases. In [10] it is shown that by taking the mean and variance of the inner products between the random canaries and model parameters, one can infer the (ε,δ)-DP guarantees and show that under mild assumptions the guarantees converge to the correct ones as the dimension $d\to\infty$. As we show, we can obtain a similar asymptotic result by applying Algorithm 3 directly to the samples to estimate the (ε,δ)-guarantees instead of using means and variances and by assuming the Gaussian parametric form for the underlying DP noise. To put our approach into perspective, we consider the same setting as in Thm. 3.3 of [10].

We first state some required auxiliary results needed for the main result. Recall first the following result by [41] which states that maximal angle between n random unit vectors goes to $\frac{\pi}{2}$ in probability as the dimension d grows, in case $\frac{\log n}{d}\to 0$ (see Thm. 5 in [41]).

Lemma A.5. *Let $x_1,\dots,x_n$ be independently uniformly chosen random vectors from the unit sphere $\mathbb{S}^{d-1}$. Let $d=d_n\to\infty$ satisfy $\frac{\log n}{d}\to 0$ as $n\to\infty$. Denote θ_{ij} the angle between the vectors x_i and x_j. Then,*

$$\max_{1\le i<j\le n}\left|\theta_{ij}-\frac{\pi}{2}\right|\to 0$$

in probability as $n\to\infty$.

Looking at the proof of Thm. 5 of [41], we obtain the following convergence speed for $\max_{1\le i<j\le n}\left|\theta_{ij}-\frac{\pi}{2}\right|$.

Lemma A.6. *Let the assumptions of Lemma A.5 hold. Then,*

$$\sqrt{\frac{d}{\log n}}\cdot\max_{1\le i<j\le n}\left|\theta_{ij}-\frac{\pi}{2}\right|\to 4.$$

Proof. This result corresponds to Corollary 2.1 of [41]. It can be shown similarly as in Thm. 5 of [41], i.e., by replacing in the proof of Thm. 1 of [42] L_n and $|\rho_{ij}|$ by $\max_{1\le i<j\le n}\left|\theta_{ij}-\frac{\pi}{2}\right|$ and ρ_{ij}, respectively. □

The convergence of the cosine angles trivially follows from the Lipschitz continuity of the cosine function.

Corollary A.7. *Let $x_1,\dots,x_n$ be independently uniformly chosen random vectors from the unit sphere $\mathbb{S}^{d-1}$. Let $d=d_n\to\infty$ satisfy $\frac{\log n}{d}\to 0$ as $n\to\infty$. Denote ρ_{ij} the cosine angle between the vectors x_i and x_j. Then,*

$$\max_{1\le 1<j\le n}|\rho_{ij}|\to 0$$

in probability as $n\to\infty$.

Proof. The results follows from Lemma A.5 and from the fact that cosine function is 1-Lipschitz:

$$|\rho_{ij}|=|\langle x_i,x_j\rangle|=|\cos\theta_{ij}|=\left|\cos\theta_{ij}-\cos\frac{\pi}{2}\right|\le\left|\theta_{ij}-\frac{\pi}{2}\right|.$$

□

We will also need the following result by [43] for the TV distance between two Gaussians with equal means (see Thm. 1.1 in [43]).

Lemma A.8. *Let $\mu\in\mathbb{R}^d$, Σ_1 and Σ_2 be positive definite $d\times d$ matrices, and $\lambda_1,\dots,\lambda_d$ denote the eigenvalues of $\Sigma_2^{-1}\Sigma_1-I$. Then,*

$$\mathrm{TV}\big(\mathcal{N}(\mu,\Sigma_1),\mathcal{N}(\mu,\Sigma_2)\big)\le\frac{3}{2}\min\left\{1,\sqrt{\sum\nolimits_{i=1}^{d}\lambda_i^2}\right\}.$$

We are now ready to state a result which shows that the sample of auditing scores converge in TV-distance to a set of i.i.d. from one-dimensional Gaussian distributions.

Theorem A.9. *Denote the auditing training canaries $A_{\mathrm{train}}=\{x_1,\dots,x_n\}$ and the auditing test canaries $A_{\mathrm{test}}=\{z_1,\dots,z_n\}$, where x_i's and z_i's are i.i.d. uniformly sampled from the unit sphere $\mathbb{S}^{d-1}$. Let $n=\omega(1)$ (as a function of d) and $d=\omega(n^3\log n)$. Suppose $\mathcal{M}$ is such that for any dataset D consisting of vectors in $\mathbb{R}^d$, $X\in\mathbb{R}^d$ denotes the sum of the vectors in D, and*

$$\mathcal{M}(D)=X+\sum_{x\in A_{\mathrm{train}}}x+Z,\quad Z\sim\mathcal{N}(0,\sigma^2 I_d).$$

Let $\theta\sim\mathcal{M}(D)$ and $\|X\|_2=o\left(\sqrt{\frac{d}{n\log n}}\right)$. Denote the training and test scores by

$$\widetilde{P}=\begin{bmatrix}\langle x_1,\theta\rangle\\ \vdots\\ \langle x_n,\theta\rangle\end{bmatrix},\quad \widetilde{Q}=\begin{bmatrix}\langle z_1,\theta\rangle\\ \vdots\\ \langle z_n,\theta\rangle\end{bmatrix}.$$

Then, denoting $\mathbb{1}_n = \begin{bmatrix}1 & \dots & 1\end{bmatrix}^T \in \mathbb{R}^n$*, we have that*

$$\mathrm{TV}\left(\begin{bmatrix}\widetilde{P}\\ \widetilde{Q}\end{bmatrix}, \mathcal{N}\left(\begin{bmatrix}\mathbb{1}_n\\ 0\end{bmatrix}, \sigma^2 I_{2n}\right)\right) \to 0,$$

as $d \to \infty$ *in probability.*

Proof. Denote $\theta = X + \sum_{x\in A_{\mathrm{train}}} x + Z$, where $Z \sim \mathcal{N}(0, \sigma^2 I_d)$. We see that for any $x_i \in A_{\mathrm{train}}$,

$$\begin{aligned} S\big(x_i, \theta\big) &= x_i^T\left(X + \sum_{x\in A_{\mathrm{train}}} x + Z\right) \\ &= x_i^T X + 1 + \sum_{x\in A_{\mathrm{train}},\, x\neq x_i} x_i^T x + x_i^T Z, \end{aligned} \tag{A.8}$$

and for any $z_i \in A_{\mathrm{test}}$,

$$\begin{aligned} S\big(z_i, \theta\big) &= z_i^T\left(X + \sum_{x\in A_{\mathrm{train}}} x + Z\right) \\ &= z_i^T X + \sum_{x\in A_{\mathrm{train}}} z_i^T x + z_i^T Z. \end{aligned} \tag{A.9}$$

From Eq. (A.8) an A.9 we see that

$$\begin{bmatrix}\widetilde{P}\\ \widetilde{Q}\end{bmatrix} = C^T X + \begin{bmatrix}\mathbb{1}_n\\ 0\end{bmatrix} + \tau + C^T Z,$$

where

$$C = \begin{bmatrix}x_1 & \dots & x_n & z_1 & \dots & z_n\end{bmatrix}$$

and

$$\tau_i = \begin{cases} \sum\limits_{x\in A_{\mathrm{train}},\, x\neq x_i} x_i^T x, & 1 \le i \le n \\ \sum\limits_{x\in A_{\mathrm{train}}} z_i^T x, & n < i \le 2n \end{cases}$$

Denote the maximum absolute cosine angle between the vectors $\frac{X}{\|X\|}, x_1, \dots, x_n, z_1, \dots, z_n$ by $\rho_{\max}$. We easily see that $\rho_{\max}$ has the same distribution as the maximum absolute cosine angle between $2n+1$ vectors uniformly sampled from the unit sphere $\mathbb{S}^{d-1}$. Also, we have that for all $x_i \in A_{\mathrm{train}}$,

$$\left|\sum\nolimits_{x\in A_{\mathrm{train}},\, x\neq x_i} x_i^T x\right| \le n\cdot\rho_{\max}$$

and for all $z_i \in A_{\mathrm{test}}$,

$$\left|\sum\nolimits_{x\in A_{\mathrm{train}}} \widehat{z}_i^T x\right| \le n\cdot\rho_{\max}.$$

Thus,

$$\|\tau\|_2 \le \sqrt{2} n^{3/2}\rho_{\max}. \tag{A.10}$$

Moreover, we have that

$$\left\|C^T X\right\|_2 \le \|X\|_2 \sqrt{2n}\cdot\rho_{\max}. \tag{A.11}$$

Moreover, by Lemma A.6, we have that

$$\sqrt{\frac{d}{\log 2n+1}}\cdot\rho_{\max} \to 4 \tag{A.12}$$

as $d\to\infty$ in probability (since $n = \omega(1)$ as a function of d). Combining Eq. (A.12) with the bounds (A.10) and (A.11), we see that $\|\tau\|_2 \to 0$ as $d\to\infty$ in probability in case $d = \omega(n^3\log n)$ and $\left\|C^T X\right\|_2 \to 0$ as $d\to\infty$ in probability in case $\|X\|_2 = o\left(\sqrt{\frac{d}{n\log n}}\right)$.

We then bound using the triangle inequality

$$\begin{aligned} &\mathrm{TV}\left(\begin{bmatrix}\widetilde{P}\\ \widetilde{Q}\end{bmatrix}, \mathcal{N}\left(\begin{bmatrix}\mathbb{1}_n\\ 0\end{bmatrix}, \sigma^2 I_{2n}\right)\right) \\ =&\mathrm{TV}\left(C^T X + \begin{bmatrix}\mathbb{1}_n\\ 0\end{bmatrix} + \tau + C^T Z, \mathcal{N}\left(\begin{bmatrix}\mathbb{1}_n\\ 0\end{bmatrix}, \sigma^2 I_{2n}\right)\right) \\ \le&\mathrm{TV}\left(\begin{bmatrix}\mathbb{1}_n\\ 0\end{bmatrix} + C^T Z, \mathcal{N}\left(\begin{bmatrix}\mathbb{1}_n\\ 0\end{bmatrix}, \sigma^2 I_{2n}\right)\right) \\ &+ \mathrm{TV}\left(C^T X + \tau + \begin{bmatrix}\mathbb{1}_n\\ 0\end{bmatrix} + C^T Z, \begin{bmatrix}\mathbb{1}_n\\ 0\end{bmatrix} + C^T Z\right) \end{aligned} \tag{A.13}$$

We next use Lemma A.8 to show the convergence of the first term on the right hand side of the inequality (A.13). Clearly, since $Z \sim \mathcal{N}(0, \sigma^2 I_d)$, we have that $C^T Z \sim \mathcal{N}(0, \sigma^2 C^T C)$. We next use Lemma A.8 with $\mu = \left[\begin{smallmatrix}\mathbb{1}_n\\ 0\end{smallmatrix}\right]$, $\Sigma_1 = C^T C$ and $\Sigma_2 = \sigma^2 I_{2n}$. Denoting $\lambda_1, \dots, \lambda_{2n}$ the eigenvalues of the matrix $\Sigma_2^{-1}\Sigma_1 = (\sigma^2 I)^{-1}\sigma^2 C^T C - I = C^T C - I$, we have that

$$\begin{aligned} \sum_{i=1}^{2n}\lambda_i^2 &= \left\|C^T C - I\right\|_F^2 \\ &= \sum_{a,b\in\{x_1,\dots,x_n,z_1,\dots,z_n\},\, a\neq b} (a^T b)^2 \\ &\le (2n)^2\rho_{\max}^2. \end{aligned}$$

By the assumption $d = \omega(n^3\log n)$ and Lemma A.6 and Eq. (A.12) we have that $\sum_{i=1}^{2n}\lambda_i^2 \to 0$ as $d\to\infty$ in probability, and therefore by Lemma A.8,

$$\mathrm{TV}\left(\begin{bmatrix}\mathbb{1}_n\\ 0\end{bmatrix} + C^T Z, \mathcal{N}\left(\begin{bmatrix}\mathbb{1}_n\\ 0\end{bmatrix}, \sigma^2 I_{2n}\right)\right) \to 0$$

as $d\to\infty$ in probability.

To show the convergence of the second term on the right hand side of the inequality (A.13), we again use the fact that $\left\|C^T C - I\right\|_F^2 \to 0$ as $d\to\infty$ in probability, the unitary invariance of the total variation distance and the fact that $\|\tau\|_2 \to 0$ and $\left\|C^T X\right\|_2 \to 0$ as $d\to\infty$ in probability. □

Combining Theorem A.9 with the convergence result of Theorem 12 we find that the (ε,δ)-distance between the histogram estimates of $\widetilde{P}$ and $\widetilde{Q}$ also converge to the DP guarantees of the Gaussian mechanism with noise scale σ.

Corollary A.10. *Suppose the assumptions of Theorem A.9 hold. Denote by* $\widehat{P}$ *and* $\widehat{Q}$ *the histogram estimates obtained from the samples* $\widetilde{P}$ *and* $\widetilde{Q}$*, respectively, for some division of the real line. Then, for all* $\varepsilon\in\mathbb{R}$*, with an appropriate division of real line, we have that*

$$H_{\mathrm{e}^\varepsilon}(\widehat{P},\widehat{Q}) \to H_{\mathrm{e}^\varepsilon}\big(\mathcal{N}(1,\sigma^2),\mathcal{N}(1,\sigma^2)\big)$$

as $d\to\infty$ *in probability.*

Proof. Consider an equidistant division of the real line into intervals, with some bin width h, and suppose the probability estimates $\widehat{P}$ and $\widehat{Q}$ are obtained from the histogram estimates

of the samples $\widetilde{P}$ and $\widetilde{Q}$, respectively. Denote by N_1 and N_0 the histogram estimates from using n samples from $\mathcal{N}(0,\sigma^2)$ and $\mathcal{N}(1,\sigma^2)$, respectively. Similarly to the proof of Lemma 11, we have that

$$H_{\mathrm{e}^\varepsilon}(\widehat{P},\widehat{Q}) \leq H_{\mathrm{e}^\varepsilon}(N_0,N_1) + (1+\mathrm{e}^\varepsilon)\big(\mathrm{TV}(\widehat{P},N_1)+\mathrm{TV}(\widehat{Q},N_0)\big). \tag{A.14}$$

We obtain the sequences of masses $(\widehat{P},\widehat{Q})$ and (N_0,N_1) by applying the same post-processing to the vectors $\begin{bmatrix}\widetilde{P}\\ \widetilde{Q}\end{bmatrix}$ and $\mathcal{N}\left(\begin{bmatrix}\mathbb{1}_n\\ 0\end{bmatrix},\sigma^2 I_{2n}\right)$, respectively.

Therefore

$$\mathrm{TV}(\widehat{P},N_1)+\mathrm{TV}(\widehat{Q},N_0) \leq \mathrm{TV}\left(\begin{bmatrix}\widetilde{P}\\ \widetilde{Q}\end{bmatrix},\mathcal{N}\left(\begin{bmatrix}\mathbb{1}_n\\ 0\end{bmatrix},\sigma^2 I_{2n}\right)\right)$$

and also, by Thm. A.9, we have that

$$\mathrm{TV}(\widehat{P},N_1)+\mathrm{TV}(\widehat{Q},N_0)\to 0$$

as $d\to\infty$. Moreover, by Thm. 12 and the assumption that $n=\omega(1)$ as a function of d, we have that $H_{\mathrm{e}^\varepsilon}(N_0,N_1)\to 0$ as $n\to\infty$, for an appropriate choice of the bid width h. Thus, the claim follows from the inequality (A.14). □

H. Lower Bound for a Single Parameter Using TV Distance

We could in principle use any hockey-stick divergence to estimate the privacy profile of a mechanism $\mathcal{M}$ in case we can parameterize the privacy profile with a single real-valued parameter in a way that the privacy guarantees depend monotonically on that parameter. Consider, for example, the noise level σ for the Gaussian mechanism with sensitivity 1, where finding the δ-value for any $\varepsilon\in\mathbb{R}$ will also give a unique value for σ. This kind of single-parameter dependence serves as a good heuristics for analyzing DP-SGD trained models, as the privacy profiles for large compositions are commonly very close to those of a Gaussian mechanism with a given noise scale [24]. We can give an intuitive explanation for this as follows.

The privacy loss random variable (PRLV) for compositions is the sum of the PLRVs of the individual mechanisms in the composition. By the central limit theorem, this sum converges in distribution to a Gaussian. On the other hand, the PRLV for a Gaussian mechanism (GM) is also Gaussian where the mean μ and variance σ^2 are determined by the sensitivity and the noise scale of the GM, and satisfy a relation $\sigma^2=2\mu$. We can write the privacy profile $\delta(\varepsilon)$ as a certain expectation over the PLRV ω, namely

$$\delta(\varepsilon)=\mathbb{E}_\omega[1-\mathrm{e}^{\varepsilon-\omega}]_+, \tag{A.15}$$

see [6] for more details. We see that plugging in to (A.15) any ω that is a Gaussian gives a shifted privacy profile of some GM. In [24] it is additionally proven that actually for the sum of PLRVs of the subsampled Gaussian mechanisms, if the product $\sqrt{T}q$ is constant, where T is the number of compositions and q the subsampling probability, then in the limit $T\to\infty$, also for the resulting distribution the variance equals two times the mean. Thus, in the limit the PLRV is a Gaussian that is also a PLRV of some GM.

Thereby, given an estimate of any hockey-stick divergence between the frequency estimates $\widehat{P}$ and $\widehat{Q}$ for an DP-SGD trained model, we get an estimate of the whole privacy profile and in particular get an estimate of an ε-value for a fixed δ-value. Figure 16 (Appendix) illustrates this by showing the relationship between the TV distances and ε-values for a fixed $\delta>0$ for the Gaussian mechanism, obtained by varying the noise parameter σ. I.e., the parameter σ is first numerically determined using the TV distance and the analytical expression of Eq. VI.1, and then the ε-value is numerically determined using the analytical expression of Eq. (II.2).

We next analytically show that the choice $\alpha=1$, i.e., the TV distance, in fact gives an estimator that is not far from optimal among all hockey-stick divergences for estimating the distance between two Gaussians.

1) Optimal Choice of Hockey-Stick: Total Variation Distance: In principle, we could use any hockey-stick divergence to estimate the statistical distance between the frequency estimates $\widehat{P}$ and $\widehat{Q}$ and to subsequently deduce the parameter of the underlying mechanism $\mathcal{M}$. However, experiments indicate that the TV distance is generally not far from optimum for this procedure. This is analytically explained by the following example.

Consider two one-dimensional Gaussians $P_\sigma\sim\mathcal{N}(0,\sigma^2)$ and $Q_\sigma\sim\mathcal{N}(1,\sigma^2)$. We first rigorously show that there is a one-to-one relationship between the hockey-stick divergence values and σ, i.e., that the hockey-stick divergence $H_\alpha(P_\sigma||Q_\sigma)$ is an invertible function of σ for all $\sigma\in(0,\infty)$ for all $\alpha>0$.

Lemma A.11. *The hockey-stick divergence $H_\alpha(P_\sigma||Q_\sigma)$ as a function of σ is invertible for all $\sigma\in(0,\infty)$.*

Proof. From Eq. (II.2) we know that

$$F'_\alpha(\sigma)=\left(-\log\alpha-\frac{1}{2\sigma^2}\right)\cdot f\left(-\sigma\log\alpha+\frac{1}{2\sigma}\right) - \alpha\cdot\left(-\log\alpha+\frac{1}{2\sigma^2}\right)\cdot f\left(-\sigma\log\alpha-\frac{1}{2\sigma}\right), \tag{A.16}$$

where f denotes the density function of the standard univariate Gaussian distribution. We see from Eq. (A.16) that for $\alpha=1$, the value of $F'_\alpha(\sigma)$ is strictly negative for all $\sigma>0$. We also know that if a post-processing function reduces the total variation distance, it reduces then all other hockey-stick divergences, since the contraction constant of all hockey-stick divergences is bounded by the contraction constant of the total variation distance. This follows from the fact that for any Markov kernel K, and for any pair of distributions (P,Q) and for any f-divergence $D_f(\cdot||\cdot)$, we have that $D_f(P\mathrm{K}||Q\mathrm{K})\leq\eta_{\mathrm{TV}}(K)\cdot D_f(P||Q)$ (see, e.g., Lemma 1 and Thm. 1 in [44]), where $\eta_{\mathrm{TV}}(K)$ denotes the contraction constant of K for the TV distance, i.e., $\eta_{\mathrm{TV}}(K)=\sup_{P,Q,\,\mathrm{TV}(P,Q)\neq 0}\frac{\mathrm{TV}(P\mathrm{K},Q\mathrm{K})}{\mathrm{TV}(P,Q)}$. Therefore, $F'_\alpha(\sigma)$ is strictly negative for all $\alpha\geq 0$. □

Denote $F_\alpha(\sigma) := H_\alpha(P_\sigma||Q_\sigma)$. To find a robust estimator, we would like to find an order $\alpha > 0$ such that the σ-value that we obtain using the numerical approach would be least sensitive to errors in the evaluated α-divergence. If we have an error $\approx \Delta H$ in the estimated α-divergence, we would approximately have an error $\Delta\sigma = \left|\frac{\mathrm{d}}{\mathrm{d}H} F_\alpha^{-1}(H)\right| \cdot \Delta H$ in the estimated σ-value. Thus, we want to solve

$$\mathrm{argmin}_{\alpha>0} \left|\frac{\mathrm{d}}{\mathrm{d}H} F_\alpha^{-1}(H)\right|.$$

By the inverse function rule, if $H = F_\alpha(\sigma)$, we have that

$$\begin{aligned}\mathrm{argmin}_{\alpha>0} \left|\frac{\mathrm{d}}{\mathrm{d}H} F_\alpha^{-1}(H)\right| &= \mathrm{argmin}_{\alpha>0} \left|\frac{1}{F'_\alpha(\sigma)}\right| \\ &= \mathrm{argmax}_{\alpha>0} |F'_\alpha(\sigma)|\end{aligned} \tag{A.17}$$

Using the relation (A.17), we can show that the optimal hockey-stick divergence estimator is always near $\alpha = 1$ which corresponds to the TV distance.

Lemma A.12. *For any $\sigma > 1$, as a function of α, $|F'_\alpha(\sigma)|$ has its maximum on the interval $[1, \mathrm{e}^{\frac{1}{2\sigma}}]$.*

Proof. The proof goes by looking at the expression $\frac{\mathrm{d}}{\mathrm{d}\alpha}F'_\alpha(\sigma)$. Clearly $F'_\alpha(\sigma)$ is negative for all $\sigma > 0$ and for all $\alpha \geq 0$. Thus $|F'_\alpha(\sigma)| = -F'_\alpha(\sigma)$.

Using the expression (A.16), a lengthy calculation shows that

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\alpha}F'_\alpha(\sigma) \\ &= \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\frac{1}{2\sigma}+\log\alpha)^2}\left(\left(\frac{1}{2\sigma}+\log\alpha\right)\left(\frac{1}{2\sigma^2}-\log\alpha\right)\right. \\ &\left.-\left(\frac{1}{2\sigma^2}-\log\alpha\right)+1\right) \\ &+ \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\frac{1}{2\sigma}-\log\alpha)^2}\left(\left(-\frac{1}{2\sigma^2}-\log\alpha\right)\cdot\right. \\ &\left.\left(\frac{1}{2\sigma}-\log\alpha\right)-\frac{1}{\alpha}\right).\end{aligned} \tag{A.18}$$

When $\alpha = 1$, i.e., $\log\alpha = 0$, we find from Eq. (A.18) that

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}F'_\alpha(\sigma)|_{\alpha=1} = -\frac{1}{2\sqrt{2\pi}\sigma^2}\mathrm{e}^{-\frac{1}{8\sigma^2}} < 0.$$

On the other hand, when $\log\alpha = \frac{1}{2\sigma}$, we see from Eq. (A.18) that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\alpha}F'_\alpha(\sigma)|_\alpha &= \exp\left(\frac{1}{2\sigma}\right) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2\sigma^2}}\left(\frac{1}{\sigma}\left(\frac{1}{2\sigma^2}-\frac{1}{\sigma}\right)\right. \\ &\left.-\left(\frac{1}{2\sigma^2}-\frac{1}{\sigma}\right)\right)+\frac{1}{\sqrt{2\pi}}\left(\mathrm{e}^{-\frac{1}{2\sigma^2}}-\mathrm{e}^{-\frac{1}{2\sigma}}\right) \\ &= \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2\sigma^2}}\left(1-\frac{1}{\sigma}\right)\left(\frac{1}{\sigma}-\frac{1}{2\sigma^2}\right) \\ &+\frac{1}{\sqrt{2\pi}}\left(\mathrm{e}^{-\frac{1}{2\sigma^2}}-\mathrm{e}^{-\frac{1}{2\sigma}}\right)\end{aligned}$$

which shows that $\frac{\mathrm{d}}{\mathrm{d}\alpha}F'_\alpha(\sigma)|_{\alpha=\exp\left(\frac{1}{2\sigma}\right)} > 0$ when $\sigma > 1$.

Moreover, we can infer from Eq. (A.18) that $\frac{\mathrm{d}}{\mathrm{d}\alpha}F'_\alpha(\sigma)$ is negative when $0 \leq \alpha < 1$ and positive for $\alpha > \mathrm{e}^{\frac{1}{2\sigma}}$. Thus $\frac{\mathrm{d}}{\mathrm{d}\alpha}|F'_\alpha(\sigma)| = -\frac{\mathrm{d}}{\mathrm{d}\alpha}F'_\alpha(\sigma)$ has its maximum on the interval $[1, \mathrm{e}^{\frac{1}{2\sigma}}]$.

□

Figure 17 illustrates numerically that the optimal value of α is not commonly far from 1.

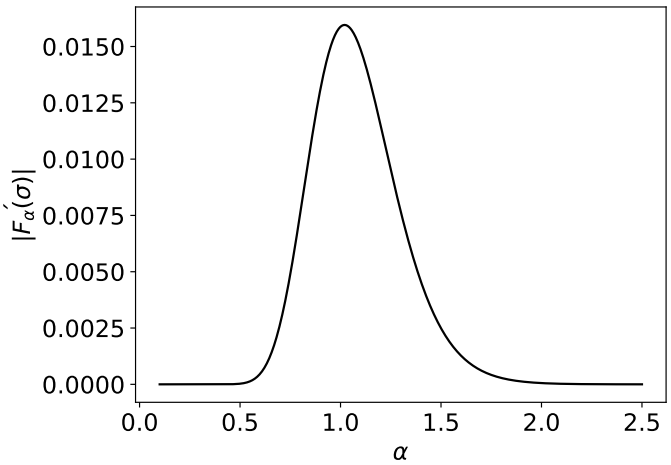

Fig. 17. The value of $|F'_\alpha(\sigma)|$ as a function of α, when $\sigma = 5$. We see that the optimal value is not far from $\alpha = 1$, indicating that the choice $\alpha = 1$ gives an estimate of σ that is robust to errors.

Numerical Example: Consider the two distributions $P \sim q \cdot \mathcal{N}(1, \sigma^2) + (1-q) \cdot \mathcal{N}(0, \sigma^2)$ and $Q \sim \mathcal{N}(0, \sigma^2)$ with the parameter values $q = 1/4$ and $\sigma = 0.3$. Estimating the TV-distance using $k = 10^6$ samples from both P and Q and 20 bins we get the estimate 0.2256. Using the fixed value $q = 1/4$, this translates to a σ-estimate of 0.302. Using Lemma 10 we get for the TV distance a 99.99 % - confidence interval $[0.302 - 0.005, 0.302 + 0.005]$ which translates to a σ-interval $[0.285, 0.32]$ so that 0.285 would be a 99.99 %-confidence lower bound for σ.

I. Algorithm for White-Box Auditing

The following algorithm is considered in the white-box auditing experiments of [8] and also in the experiments in our Section VII-B.

Input: Training dataset D, sampling rate q, learning rate η, noise scale σ, gradient clipping constant C, loss function ℓ, canary gradient g', canary sampling rate q_c, function clip$(\cdot$ that clips vectors to max 2-norm $\widetilde{C}$, number of observations T, number of training iterations τ.
Observations: $O \to [], O' \to []$.
Observations: $O \to [], O' \to []$.
Set: $D' = D \bigcup \{(x', y')\}$.
Initialize: $\theta = \theta_0$.
for $t \in [T]:$ **do**
$B_t \to$ Poisson subsample instances from D, each with probability q.
$B'_t \to$ Poisson subsample instances from D, each with probability q.
$\nabla[t] \to \sum_{(x,y)\in B_t} \text{clip}\big(\nabla_\theta \ell\big(\theta, (x, y)\big)\big)$.
$\nabla[t] \to \nabla[t] + \mathcal{N}(0, C^2\sigma^2)$.
$\nabla[t]' \to \sum_{(x,y)\in B'_t} \text{clip}\big(\nabla_\theta \ell\big(\theta, (x, y)\big)\big)$.
$\nabla[t]' \to \nabla[t]' + \mathcal{N}(0, C^2\sigma^2)$.
With probability q_c: $\nabla[t]' \to \nabla[t]' + g'$ (add canary with probability q_c).
$O[t] \to \langle \nabla[t], g' \rangle$.
$O[t]' \to \langle \nabla[t]', g' \rangle$.
$\theta \to \theta - \eta \nabla[t]$.
end for
return O, O'.

J. Further Experimental Results: HS-Divergence Between $1d$ *Gaussians*

Figures 18 and 19 we show the results for $n = 5 \cdot 10^5$ and $n = 5 \cdot 10^6$, respectively. In case $n = 5 \cdot 10^5$, Lemma 11 would give for $k = 10$ the error estimate $\approx 2 \cdot 10^{-2}$ whereas the actual error seems to be $\approx 10^{-4}$. In case $n = 5 \cdot 10^5$, Lemma 11 would give for $k = 10$ the error estimate $\approx 6 \cdot 10^{-3}$ where as the actual error seems to be $\approx 10^{-5}$. Improving the bound of Lemma 11 would directly improve the sample complexity of obtaining high-confidence lower bounds for the DP parameters.

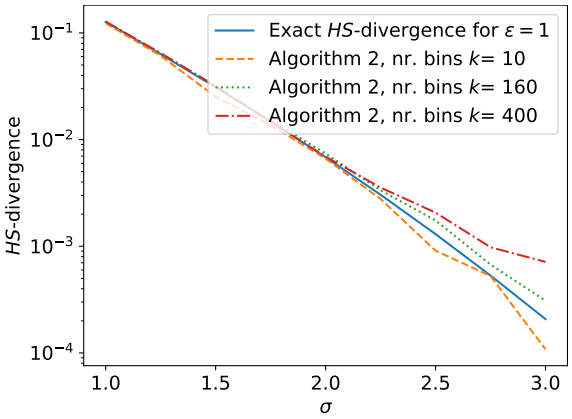

Fig. 18. Exact hockey-stick divergence $H_\varepsilon(P, Q)$ and the approximation obtained using Alg. 3 for different values of σ, when $k = 5 \cdot 10^5$. The bin width h_n set using Eq. (IV.3) gives here $k \approx 160$ bins.

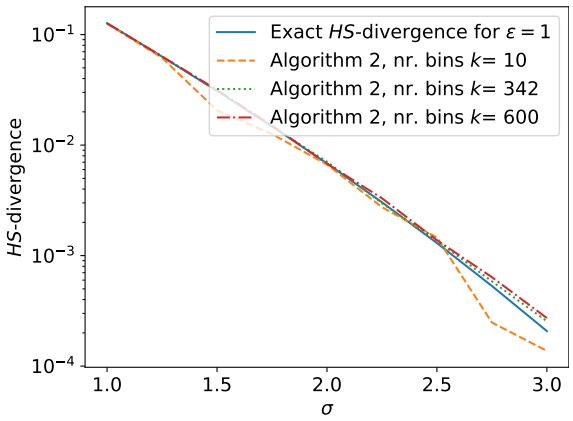

Fig. 19. Exact TV distance $TV(P, Q)$ and the TV distance approximated using Alg. 3 for different values of σ, when $k = 5 \cdot 10^6$. The bin width h_n set using Eq. (IV.3) gives $k \approx 342$ bins.