RényiTester: A Variational Approach to Testing Differential Privacy

William Kong
Google Research
NYC
weiweikong@google.com

Andrés Muñoz Medina
Google Research
NYC
ammedina@google.com

Mónica Ribero
Google Research
NYC
mribero@google.com

Abstract

Governments and industries have widely adopted differential privacy as a measure to protect users' sensitive data, creating the need for new implementations of differentially private algorithms. In order to properly test and audit these algorithms, a suite of tools for testing the property of differential privacy is needed. In this work we expand this testing suite and introduce RENYITESTER , an algorithm that can reject a mechanism is not Rényi differentially private. Our algorithm computes computes a lower bound of the Rényi divergence between the distributions of a mechanism on neighboring datasets, only requiring black-box access to samples from the audited mechanism. We test this approach on a variety of pure and Rényi differentially private mechanisms with diverse output spaces and show that RENYITESTER detects bugs in mechanisms' implementations and design flaws. While detecting that a general mechanism is differentially private is known to be NP hard, we empirically show that tools like RENYITESTER provide a way for researchers and engineers to decrease the risk of deploying mechanisms that expose users' privacy.

1 Introduction

In the past decade, there has been an explosion of data driven technologies such as automated chat bots, medical image classifiers and face recognition systems. As these technologies become more ingrained in our everyday lives, society is realizing that sharing data with these technologies, even in aggregate, may pose privacy risks. With this realization, regulators and tech companies have had to update their systems to handle data in a privacy safe manner. At the same time, users expect technology to be automated and frictionless. This automation is generally data-driven, putting both goals of *usability* and *privacy* seemingly at odds.

Luckily, the concept of differential privacy [13] has demonstrated that high quality statistical information or machine learning models can still be generated without compromising the privacy of any individual user. At the heart of differential privacy is the concept of a *mechanism*. A mechanism $\mathcal{M}$ is a randomized function that maps a dataset D to an object, such as a set of statistics or a machine learning model. Differential privacy quantifies how much any individual user in the dataset affects the output of a mechanism, and this quantification is measured by the *privacy budget* ϵ. The smaller ϵ is the less each user affects the outcome of the mechanism and, hence, the less information about

Workshop on Regulatable Machine Learning at the 37th Conference on Neural Information Processing Systems (RegML @ NeurIPS 2023).

specific users may be leaked from the output of the mechanism. This intuition is formalized by bounding the distance between the distributions of the output of $\mathcal{M}$ on two *neighboring* datasets D and D'. More formally, this is the distance between the distribution of random variables $\mathcal{M}(D)$ and $\mathcal{M}(D')$, where D' is a dataset obtained from D by adding or subtracting a single record.

The introduction of differential privacy to the research community has revolutionized the world of statistics and machine learning. Research in this field has been prolific and the community has developed differentially private algorithms for a variety of learning tasks. More importantly, mechanisms for these tasks are continuously being improved to extract the most utility, without compromising any privacy. It is in these improvements that one of the issues of differential privacy is observed. Unlike other privacy notions, like k-anonymity, one cannot verify if a mechanism is differentially private based only on a single output of a mechanism. Indeed, differential privacy is an information theoretical property of the mechanism that can only be verified by understanding the probability distribution over the space of outputs of a mechanism. This is straightforward when the mechanism is the well-known Laplace or Gaussian mechanism (albeit there are known errors in the implementation of even these mechanisms). However, as mechanisms become more accurate, the distributions generally become more complex. Fully understanding the distributions of such mechanisms becomes harder and errors on the analysis of such distributions (or errors in the implementations of such mechanisms) have occurred in the past [23]. In some of these scenarios, mechanisms that were asserted to be differentially private at a certain privacy budget level ϵ turned out to be either private at a different level or not private at all. As these mechanisms get deployed into real-world systems, it is important for researchers and regulators to verify the privacy claims of their mechanisms.

Ideally, given a privacy budget ϵ, there would be a system that takes, as input, the implementation of a mechanism and validates that the mechanism is differentially private at the asserted level of ϵ. The stochastic nature of differential privacy makes this difficult, since verifying differential privacy requires bounding the distance between two distributions, which is generally hard to estimate.

In this paper we propose a tester for detecting if a mechanism satisfies so called Rényi differential privacy (RDP) guarantees [24]. RDP provides some advantages over approximate (ϵ, δ)-differentialy privacy. For one, it provides a better understanding of the privacy properties of the Gaussian mechanism by smoothly quantifying the probability of failing to achieve privacy. Moreover, its composability properties makes it a great tool for calculating overall privacy budgets of iterative algorithms such as the celebrated differentially private stochastic gradient descent (DP-SGD). Indeed, popular open source privacy accounting libraries [2, 34] are implemented with RDP as their backbone. For this reason we believe that Rény DP tester would be of the utmost importance to the privacy community and to the best of our knowledge, this is the first proposed such tester. As an added benefit, we show how a Rényi differential privacy tester can be used to test ϵ-differential privacy. Finally, we believe that estimating lower bounds of the Rényi divergence is of independent interest to the statistics community [6, 21].

Another contribution of our work comes from the use of Bayesian optimization methods to find neighboring datasets D and D' for which the privacy guarantee is violated. This approach allows a user to not only discover whether a mechanism is private, but also provides information about the type of datasets for which the mechanism leaks the most information. Previous work either ignores this [16] or tests only on grids containing extremal datasets [9, 5]. Our experiments show that in some cases the privacy violation does not occur in an extremal dataset.

The rest of the paper is organized as follows. First, we introduce the necessary concepts to derive our statistical test, then we discuss previous work on testing of differentially private mechanisms. We then proceed to introduce our test and its theoretical guarantees. Finally, we conduct extensive empirical evaluation to demonstrate that a) our distance estimator performs very well in practice and b) known privacy bugs can easily be detected using our tester.

2 Preliminaries

Notation. $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ denotes a mechanism that receives an input dataset $D \subseteq \mathcal{X}^n$ with n records and domain $\mathcal{X} \subseteq \mathbb{R}^p$ and outputs a statistic $y \in \mathcal{Y} \subseteq \mathbb{R}^d$.

2.1 Differential privacy and Renyi divergence

Differential privacy [13] quantifies the level of risk that a user is exposed to when they contribute their data to a randomized mechanism. We formalize this concept in definition 2.1

Definition 2.1. Datasets D, D' are called neighbors, denoted by $D \sim D'$, if D can be obtained from D' by adding or removing one record from D. A randomized mechanism $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ satisfies (ϵ, δ)–approximate differential privacy, or is (ϵ, δ)–differentially private ((ϵ, δ)–DP), if for every pair of neighboring datasets D and D' and every set $O \subseteq \mathcal{Y}$ in the output space, we have

$$P(\mathcal{M}(D) \in O) \leq e^{\epsilon} P(\mathcal{M}(D') \in O) + \delta \tag{1}$$

We say $\mathcal{M}$ satisfies pure differential privacy, or is ϵ–differentially private (ϵ–DP), when $\delta = 0$.

An interpretation of differential privacy suggests that a mechanism is private if the distance between the distributions of $\mathcal{M}(D)$ and $\mathcal{M}(D')$ is small (relative to ϵ and δ). Under this interpretation, novel notions of privacy have emerged by introducing different ways of measuring divergences between distributions. Notably, the Rényi divergence [31, 15] (which we define below) has become a popular choice when analyzing the privacy properties of mechanisms such as DP-SGD [14].

Definition 2.2. Let $(\Omega, \mathcal{F})$ be an arbitrary measurable space. Let P and Q denote two probabilities in $(\Omega, \mathcal{F})$. We assume that P is absolutely continuous with respect to Q [1] and let $\frac{dP}{dQ}$ denote the Radon-Nykodym derivative of P with respect to Q. For $\alpha > 0$, the Rényi divergence of order α between P and Q is given by

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \ln \int \left(\frac{dP}{dQ} \right)^\alpha dQ \tag{2}$$

We now make two remarks about the above definition. First, as $\alpha \downarrow 0$, the quantity $D_\alpha(P||Q)$ tends to the well-known Kullback–Leibler (KL) divergence. Second, when P and Q admit density functions p, q respectively, the above expression is equivalent

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \ln \int \left(\frac{p(x)^\alpha}{q(x)^{\alpha-1}} \right) dx$$

We will abuse the notation sometimes for random variables $X \sim P$ and $Y \sim Q$ we will denote $D_\alpha(X||Y) = D_\alpha(P||Q)$. Using this divergence, we can introduce the notion of Rényi differential privacy [25].

Definition 2.3. A randomized mechanism $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ satisfies (α, ϵ)–Rényi differential privacy if for every pair of neighboring datasets D and D', we have

$$D_\alpha(\mathcal{M}(D)||\mathcal{M}(D')) \leq \epsilon \tag{3}$$

The next two results present some important properties about $D_\alpha(P||Q)$.

Lemma 2.4 (Proposition 9 in [25])**.** *Let $1 < \alpha_1 < \alpha_2$ and P and Q be probability measures. Then $D_{\alpha_1}(P||Q) < D_{\alpha_2}(P||Q)$*

Lemma 2.5 (Lemma 1 in [25])**.** *Let $\mathcal{M}$ be an ϵ–differentially private mechanism and $\alpha > 1$. Then $D_\alpha(\mathcal{M}(D)||\mathcal{M}(D')) \leq \min\{\epsilon, 2\alpha\epsilon^2\}$.*

Fact 2.6. $\alpha = \infty$ corresponds to pure-DP, i.e., $\mathcal{M}$ is an ϵ–DP mechanism if and only if for any $D \sim D'$, we have $D_\infty(\mathcal{A}(D)||\mathcal{A}(D') \leq \epsilon$.

3 Related work

There are generally two kinds of approaches used in differential privacy testing. The first approach uses adversarial attacks that try to break the privacy definition, like membership inference attacks [20, 30, 7] and data reconstruction attacks [18, 3] of deep learning models trained with DP-SGD. Hence, the validation of whether a mechanism satisfies privacy is linked to the ability of the attack to succeed. The tests generated by these approaches are very valuable when trying to understand

[1] A measure P is absolutely continuous with respect to Q if for every set $A \subset \Omega$ such that $Q(A) = 0$ then $P(A) = 0$.

potential privacy risks on a single data set, by manually designing canaries that are expected to have highest sensitivity. However, they do not attempt to understand the worst case (unknown) scenario that differential privacy tries to protect. Running these tests generally requires white box access to the trained model and, more importantly, requires access to large portions of the training data, making auditing of a privately trained model impossible for someone who is not the data curator. Consequently, the resulting lower bounds from these approaches tend to be loose [28, 27]. Moreover, the budget ϵ predicted by these experiments is generally much smaller than the theoretical budget. For example, some authors assert that their proposed models were private with an $\epsilon = 10^{-3}$ when these models were trained without privacy.

The second approach, that contains our proposed method, attempts to directly estimate the *effective* privacy parameters from black-box access to the tested mechanism and compare these effective privacy parameters with the ones stated by the privacy guarantee. This approach focuses on estimating the distance between the distribution induced by the mechanism in two different datasets. However, two key challenges arise: 1) how do we estimate the distance between distributions given two fixed neighboring datasets? and 2) how do we find the pair of neighboring datasets that maximize the distance between these distributions?

The problem of estimating distance between distributions has been thoroughly studied in the statistics and hypothesis testing community. While providing a full overview of the literature in this space is beyond the scope of this work, we do highlight [29, 35, 33] which consider estimating probability distances through optimization methods over function spaces. Their work provided asymptotic guarantees while we provide strict finite sample complexities to obtain a lower bound on the Rényi divergence between two distributions.

For the specific task of estimating the Rényi divergence, our estimator is inspired by the work of [6] which considers using neural networks to estimate Rényi divergence. The finite sample complexity bounds provided in that work, however, depend on the structure of the neural network and can rapidly become vacuous for the purpose of testing differential privacy. In contrast, our complexity bounds are independent of the network structure as we are primarily concerned with lower bounds on the Rényi divergence. In a related approach, [11] proposes to estimate the *regularized kernel Rényi divergence*, a lower bound on the Rényi divergence between distributions of a randomized mechanism. However, this approach requires knowledge about the covariance matrix of the underlying distributions, which is impractical for most mechanisms other than the Gaussian and Laplace mechanisms. Recent work on tight estimation of the privacy loss distribution [12] provides techniques for lower-bounding ϵ, and in some cases it can be tighter. Unfortunately, the previous method needs access to the cumulative distribution function of the distribution of the privacy loss random variable, which is precisely unknown in our considered setting.

There is also a large body of literature pertaining to the testing of a mechanism's privacy, which we briefly go through here. [10] proposes a differential privacy tester for mechanisms with discrete and finite output, requiring access to the distribution over datasets and the probability measure over outputs induced by the tested algorithm. Instead of testing privacy in the worst case setting, they test if the mechanism satisfy the guarantee over datasets with high probability. More importantly the tester does not work for continuous output spaces. StatDP [9] proposes a system for detecting differential privacy violations by post-processing the output of the mechanisms through different statistics. The tester requires semi-black box access to the mechanisms (as one of the post processing techniques requires running the mechanism without privacy), which is infeasiable for auditing certain systems. [16] presents a test for discrete (ϵ, δ)-DP mechanisms but omits the problem of finding the worst case pair of neighboring datasets. DP-Sniper [5] provides an ϵ-DP tester that tries to explicitly find a set in the output space that maximizes the difference in probability for the output of the mechanism. The choice of neighboring datasets, however, is done using some hard-coded rules that may hinder the ability to detect violations on new tasks, and under non-classic neighboring relations llike the ℓ_∞ relation instead of the classic swap or add/remove definition of neighboring. Their framework is also specific to detecting ϵ-DP, as low probability events are hard to estimate. In contrast, our mechanism estimates RDP, which *averages out* low probability events. Moreover, we use our estimates to inform the search of worst case datasets through a Bayesian optimizer. [19] proposes a similar approach but targets specifially auditing the privacy of DP-SGD. [22] extends the work of [5] by developing data poisoning attacks to explore the space of datasets, focusing on machine learning predictive models learning algorithms rather than arbitrary statistical tasks.

Algorithm 1 RenyiTester

```
1: Input: Mechanism (black-box access) $\mathcal{M}$, probability of failure $\beta$, estimation error $\kappa$, $(\alpha, \epsilon)$–
   Rényi differential privacy parameters, number of trials $N$, family of bounded functions $\mathcal{H}$.
2: Initialize the sample complexity $n(\epsilon, \alpha, \beta, \kappa)$ (see theorem 4.3).
3: Compute the test divergence threshold
   $$\tau(n) := \begin{cases} \epsilon + \kappa & \text{if testing Rényi DP,} \\ \min(\epsilon, 2\alpha\epsilon^2) + \kappa & \text{if testing pure DP.} \end{cases}$$
4: for $i = 1$ to $N$ do
5:     Generate datasets $D, D'$ (see section 4.4).
6:     Generate $2n$ samples $X \sim \mathcal{M}(D)$, and $2n$ samples from $Y \sim \mathcal{M}(D')$ that are evenly split
       as $X_{\text{train}}, X_{\text{test}}, Y_{\text{train}}, Y_{\text{test}}$.
7:     $h_1 = \arg\max_{h \in \mathcal{H}} R_\alpha^{h,n}(X_{\text{train}}||Y_{\text{train}})$.
8:     $h_2 = \arg\max_{h \in H} R_\alpha^{h,n}(Y_{\text{train}}||X_{\text{train}})$.
9:     if $R^{h_1,n}(X_{\text{test}}||Y_{\text{test}}) \geq \tau$ or $R^{h_2,n}(Y_{\text{test}}||X_{\text{test}}) \geq \tau$ then
10:        Return False (w.p. $1 - \beta$, $\mathcal{M}$ is not private)
11:    end if
12: end for
13: Return Passed (no proof of privacy violation)
```

4 Rényi Tester

In this section, we propose RenyiTester , an RDP and ϵ-DP (or pure DP) tester that is able to find instances where non-private mechanisms violate the privacy guarantee that they claim to have. While the sample complexity to prove that a mechanism satisfies pure ϵ-DP can be exponentially large [16], we use several heuristics that help detect mechanisms that are not private. We start by providing an overview of RenyiTester followed by a derivation of the algorithm's subroutines. We finish by proving a sample complexity bound that ensures the test results hold with high probability.

We introduce RenyiTester in algorithm 1. The tester receives, as input, (i) black-box access to the tested mechanism $\mathcal{M}$, (ii) a value ϵ if validating ϵ-DP or a tuple (α, ϵ) if validating RDP, and (iii) a probability of failure β. It then proceeds as follows:

1. **Generate neighboring datasets (line 5).** This is done according to the process discussed in section 4.4.
2. **Generate samples from mechanism.** Given the datasets, the tester generates samples for the mechanism for each dataset.
3. **Obtain a lower bound for the Rényi between both samples.** The details of the estimation process are described in Section 4.2 and through Corollary 4.5.
4. **Detect if the mechanism violates privacy.** Specifically, use the bound in Lemma 2.5 with Corollary 4.5.

4.1 Variational formulations

We now present estimator for a lower bound on the Rényi divergence of two distributions. Our estimator relies on a variational formulation of the Rényi divergence. The first such formulation is a special case of the problem of calculating f divergences via convex optimization [29], and the formulation that we use (described below) is the one recently proposed by [6].

Theorem 4.1. *Let $\alpha > 1$, and P and Q be probability measures on $(\Omega, \mathcal{F})$. Let Γ be any function space such that $M_b(\Omega) \subseteq \Gamma \subseteq M(\Omega)$ where $M_b(\Omega)$ and $M(\Omega)$ are the sets of measurable bounded and measurable functions on Ω respectively. Then,*

$$D_\alpha(P||Q) = \sup_{g \in \Gamma} \tfrac{\alpha}{\alpha - 1} \log \left(\mathbb{E}_P \left[e^{(\alpha-1)g(X)} \right] \right) - \log \left(\mathbb{E}_Q \left[e^{\alpha g(X)} \right] \right) \tag{4}$$

Exact computation of the supremum in eq. (4) is generally hard, given that the complexity of the function space can be arbitrarily large for general distributions. We propose to relax this definition in two ways that allow us to derive a lower bound on the Rényi divergence. First, we fix a space of functions $\Phi \subseteq \Gamma$. By restricting the search space for the supremum, the obtained value will be a lower bound on the real divergence. For example, one can define Φ as the set of functions generated by dense neural networks with bounded outputs. Second, we estimate the expectations using approximate (empirical) measures from samples, P_n, Q_n. While this last step introduces estimation error, this error can be bounded with high probability, thus allowing us to find a confidence interval for the lower bound.

Definition 4.2. Let $h : \mathcal{Y} \subseteq \Omega \to \mathbb{R}$ be a function in Φ on Ω and $\alpha > 1$. Define

$$R_\alpha^h(P||Q) := \tfrac{\alpha}{\alpha-1} \log\left(\int e^{(\alpha-1)h(x)} dP\right) - \log\left(\int e^{\alpha h(x)} dQ\right) \tag{5}$$

and, given samples $X_1, ..., X_n \sim P, Y_1, ..., Y_n \sim Q$, define its empirical counterpart

$$R_\alpha^{h,n}(X||Y) := \tfrac{\alpha}{\alpha-1} \log\left(\frac{1}{n}\sum_{i=1}^{n} e^{(\alpha-1)h(X_i)}\right) - \log\left(\frac{1}{n}\sum_{i=1}^{n} e^{\alpha h(Y_i)}\right). \tag{6}$$

The next section derives a sample complexity bound to quantify the estimation error $\text{err}(n, \delta) := |R_\alpha^{h,n}(X||Y) - R_\alpha^h(P||Q)|$ with probability $1 - \delta$.

Note that with the error function, we can provide a lower bound to the true Rényi divergence between P and Q as follows: for $h_0 \in \Phi$ and $M_b(\Omega) \subseteq \Gamma \subseteq M(\Omega)$, we have

$$R_\alpha(P||Q) = \sup_{h\in\Gamma} R_\alpha^h(P||Q) \geq \sup_{h\in\Phi} R_\alpha^h(P||Q) \geq R_\alpha^{h_0}(P||Q) \geq R_\alpha^{h,n}(X||Y) - \text{err}(n, \delta).$$

4.2 Sample complexity

The following theorem derives an inequality satisfied by every private mechanism with high probability for all neighboring datasets (cf. line 9). We provide a proof in the supplementary material.

Theorem 4.3. *Let P and Q be two distributions. Let $h\colon \Omega \subseteq \mathcal{Y} \to \mathbb{R}$ be a function such that $\sup_{x\in\Omega} h(x) < C$, $\mathbf{x} = (x_1, ..., x_n)$ and $\mathbf{y} = (y_1, ..., y_n)$ be n realizations of P and Q, respectively, $\mu_1 = \mathbb{E}_P\left[e^{(\alpha-1)h(x)}\right]$, and $\mu_2 = \mathbb{E}_Q\left[e^{\alpha h(x)}\right]$. Define also $M_1 = e^{(\alpha-1)C}$ and $M_2 = e^{\alpha C}$. Then, if $\gamma \in [0, \min(\frac{M_1}{\mu_1}, \frac{M_2}{\mu_2})]$, and $n \geq \max\left(\frac{3M_1 \log(2/\beta)}{\mu_1\gamma^2}, \frac{2M_2 \log(2/\beta)}{\mu_2\gamma^2}\right)$, with probability at least $1 - \beta$, we have*

$$R_\alpha^h(P||Q) \geq R_\alpha^{h,n}(\mathbf{x}||\mathbf{y}) - \log\left(\frac{1+\gamma}{1-\gamma}\right) \tag{7}$$

Remark 4.4. Our sample complexity is dimension independent. On the other hand, there are results showing that sample complexity of estimating the Rényi divergence from samples is lower bounded by e^d, where d is the dimension of the distribution output space. Our result does not contradict this fact because we are not estimating the true Rényi divergence, but a lower bound of the divergence. As the dimensions of the mechanism increases, one could expect that a more complex space of functions is required for in the definition of the lower bound.

The next result shows how our estimate $R_\alpha^{h,n}(\mathbf{x}\|\mathbf{y})$ is used as a lower bound for the true Rényi divergence.

Corollary 4.5. *Let $h\colon \Omega \subset \mathcal{Y} \to \mathbb{R}$ be a function such that $\sup_{x\in\Omega} |h(x)| \leq C$, $\mathcal{M}$ denote a mechanism and D, D' be two neighboring databases., $\mathbf{x} = (x_1, \ldots, x_n)$ be a sample from $\mathcal{M}(D)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ be a sample from $\mathcal{M}(D')$, and $\beta > 0$ and γ be defined as in theorem 4.3. If n is chosen according to theorem 4.3, then with probability at least $1 - \beta$, we have*

$$D_\alpha(\mathcal{M}(D)||\mathcal{M}(D')) \geq R_\alpha^{h,n}(\mathbf{x}\|\mathbf{y}) - \log\left(\frac{1+\gamma}{1-\gamma}\right).$$

4.3 Selection of function h

The previous section showed that we can choose a function h to lower bound the Rényi divergence between the output of a mechanism in two neighboring datasets. It remains to show how to select the function that obtains the tightest lower bound. In this section we provide a natural heuristic for choosing h. Fix $C > 0$ and let Φ denote a collection of functions bounded by C. We propose the following two step approach. First, sample $\mathbf{x} = (x_1, \ldots, x_n)$ from $\mathcal{M}(D)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ from $\mathcal{M}(D')$. Let h^* be defined by

$$h^* = \arg\max_{h \in \Phi} R_\alpha^h(\mathbf{x}||\mathbf{y}).$$

Second, given h^*, generate a new sample $\mathbf{x}' = (x'_1, \ldots, x'_n)$ from $\mathcal{M}(D)$ and $\mathbf{y}' = (y'_1, \ldots, y'_n)$ from $\mathcal{M}(D')$, and use theorem 4.3 on this sample to obtain a lower bound on the true Rényi divergence. The process just described corresponds to lines 6–8 in algorithm 1. It is also worth mentioning that the above approach is somewhat similar to DP-Sniper [5]. Specifically, the latter approach uses a *training* sample to find a set where the DP guarantee can fail and then use a *test* sample to estimate the actual privacy violation.

Model considerations. Even though the model complexity does not appear in the sample complexity of our mechanism, it is important to constrain the model class as our heuristic only makes sense when $R_\alpha^{h.n}(\mathbf{x}||\mathbf{y})$ and $R_\alpha^{h.n}(\mathbf{x}'||\mathbf{y}')$ are close.

4.4 Dataset generation

One of the main difficulties of testing for differential privacy is the worst-case nature of differential privacy guarantees. Namely, to prove a mechanism is not private, one has to find a dataset where inequality (1) or (3) fails to hold. We propose to use black-box optimization to find datasets that maximize $R^{h,n}(X||Y)$. Specifically, assuming that we have access to $R_\alpha^{h,n} : (D, D') \subseteq \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, our goal is to produce a sequence $(D_t, D'_t)_t$ that approaches the optimum. In our case, we only need to generate a point (D, D') where line 9 does not hold. Available techniques include pure exploration methods, such as grid search, and techniques that use prior information to trade between exploration and exploitation that can accelerate the optimization, such as evolutionary methods. We refer the reader to [4] for an overview. In our experiments we will use an open-sourced implementation of the well known Bayesian optimization software Vizier.

5 Experiments

This section presents numerical experiments for RenyiTester . We first demonstrate how RenyiTester can be used to detect pure differential privacy guarantees. We then focus on RDP violations and specifically look into two common errors in DP-SGD implementations. We include in the supplementary an analysis on the accuracy of estimating Rényi divergence. Throughout our exposition, we let $\varepsilon > 0$ and $n \geq 1$ be fixed and $X \in \mathbb{R}^n$ denote the input dataset.

Pure DP mean mechanisms. The first three mechanisms attempt to privately compute the mean by generating the random estimates

$$\text{DPMean}(X) := \frac{\sum_{i=1}^n X_i}{\tilde{n}} + \rho_1, \quad \text{NonDPMean1}(X) := \frac{\sum_{i=1}^n X_i}{n} + \rho_2$$
$$\text{NonDPMean2}(X) := \frac{\sum_{i=1}^n X_i}{n} + \rho_1$$

where $\tilde{n} = \max\{10^{-12}, n + \tau\}$, $\tau \sim \text{Laplace}(0, 2/\varepsilon)$, $\rho_1 \sim \text{Laplace}(0, 2/[\tilde{n}\varepsilon])$, and $\rho_2 \sim \text{Laplace}(0, 2/[n\varepsilon])$. The first estimate satisfies ϵ-DP, the second one violates the guarantee because it has access to the private number of points, and the third one privatizes the number of points to estimate the scale of the noise added to the mean statistic but the mean itself is computed using the non-private number of points.

Sparse vector technique mechanisms. The next six mechanisms address different private and non-private implementations of the sparse vector technique (SVT), a mechanism for releasing a stream of c queries on a fixed dataset. SVT mechanisms compare each query value against a threshold and the given algorithm returns certain outputs for a maximum number of queries c. We denote these by SVT1 –SVT6 and they correspond to Algorithms 1-6 in [23]. SVT1 and SVT2 satisfy ϵ-DP. SVT4 satisfies $(\frac{1+6c}{4})$-DP, and SVT3, SVT5, and SVT6 do not satisfy ϵ-DP for any finite ϵ.

Rényi DP mean mechanisms. To verify the ability of our tester to detect violations of Rényi differential privacy we first instantiate NonADPMean, a non-private Gaussian mean analog of

Non-Private-Mean1 that uses the true number of points to compute the mean and noise scale, but adds Gaussian noise instead of Laplace noise.

DP-SGD mechanisms. We also include two flawed DP-SGD's [1] implementations.

Recall that DP-SGD is parametrized by a clip norm G (which clips individual per-example gradients to have ℓ_2 norm G) and a noise multiplier σ, and that a single iteration of DP-SGD is guaranteed to be (α, ϵ)-RDP for $\epsilon = \frac{2\alpha}{\sigma^2 G^2}$. The first implementation simulates a scenario where a developer assumes they are using a noise multiplier σ_{theory} but in reality uses a noise multiplier $\sigma_{\text{effective}}$. We dub this scenario SCALEDSGD .

For the second implementation, we consider an accounting error when using batch or micro-batch clipping instead of per-example clipping in DP-SGD. Per-example clipping is memory and computationally expensive when training high-dimensional models. To address these constraints at the cost of utility, practitioners split a batches of size n into m microbatches of size n/m, compute average gradients over each micro-batch, clip and noise the per-microbatch gradient, and finally average the resulting noisy micro-batch gradients. It sometimes goes unnoticed but the sensitivity of per-microbatch gradients is $2G$ instead of G. WRONGSENSITIVITYSGD below refers to an implementation of a DP-SGD optimizer that receives a model f_θ, learning rate, noise multiplier σ, clip norm value G, number of micro-batches, and takes a DP-SGD with noise scaled by σG respect to the parameters θ, and does privacy accounting using a library that receives batch size, number of epochs, noise multiplier, assuming per-example clipping and ignoring the of microbatch clipping. The final budget should be $\epsilon = \frac{2\alpha}{\sigma^2}$ but by ignoring microbatching results in the misleadingly stricter guarantee of $\epsilon = \frac{\alpha}{2\sigma^2}$.

Baselines. We compare RENYITESTER 's auditing capacity first with the the approximate differential privacy tester (ADPTESTER) presented in [16]. For completeness we introduce this algorithm as algorithm 2 in the supplement. For a fixed pair of neighboring datasets, the algorithm estimates from samples the probability z of the algorithm violating a pure ϵ-differential privacy guarantee (line 10). If the mechanism is (ϵ, δ)–differentially private, then $z < \delta$ up to estimation error η (line 11).

We also compare our method with DP-Sniper [5]. Recall that the original DP-Sniper paper uses different neighboring relationships for different mechanisms. Below we compare the methods under the same neighboring relationships to elucidate the power of these testers under similar conditions. DP-Sniper is generally unsuited for RDP, hence we do not include a comparison in the experimental section for non-pure DP mechanims.

The RKRDTESTER introduced in [11] is similar to RENYITESTER but requires knowledge of certain covariance matrices that are generally not known *a priori*. Consequently, we do not compare with this test in our auditing experiments, but do compare it with RENYITESTER in the estimation of Rényi divergence between Gaussian distributions in appendix B.6.

Table 1: Sample complexity for RENYITESTER , ADPTESTER with different universe sizes m, and DP-Sniper.

ϵ	α	RENYITESTER	ADPTESTER $m = 10$	 $m = 100$	 $m = 1000$	DP-Sniper
0.01	1.5	520K	810K	8.1M	81M	10M
0.1	1.5	360K	810K	8.1M	81M	10M
0.01	2.	610K	890K	8.9M	89M	10M
0.1	2.	1.8M	890K	8.9M	89M	10M

Methodology. We run RENYITESTER tester with Φ being the class of functions generated by a two-layer dense neural networks consisting of 100 units for each hidden layer. To ensure the output of the network is bounded we use a scaled hyperbolic tangent loss activation scaled to $C = 16\epsilon$ for the last layer. DPSNIPER proposes its own grid of test cases to generate pairs of datasets. RENYITESTER and ADPTESTER are run on trials by generating pairs of neighboring datasets using an open sourced version of Vizier [32, 17], with an underlying NSGA-II evolutionary algorithm [8]. This method performed slighly better than a a random search algorithm, but obtaining similar speed of detection, or no detection at all (see appendix B.4).

We test each mechanism for different values of ϵ and α, and test 5 times for each mechanisms. We found that both RENYITESTER and ADPTESTER had different estimator values over the five runs but the outcome (`False` or `Passed`) was consistent across runs.

Pure DP results. The results of our experiments are summarized in Table 2. RENYITESTER detects all one-dimensional non-private mechanisms while the ADPTESTER fails to detect NONDPMEAN1,

Table 2: Non-Privacy detection with aDPTester and RenyiTester and add/remove neighboring relation. We report `false` if a mechanism fails the privacy test. Otherwise we report `passed` . Experiments were run 5 times always with the same outcome. aDPTester was run with universe sizes 10,100, and 1000. . ScaledSGD uses $\sigma_{\text{theory}} = 10$ and $\sigma_{\text{effective}} = 1.5$

	Guarantee	RenyiTester	aDPTester	DPSniper
DPMean	ϵ	passed	passed	passed
NonDPMean1	ϵ	false	passed	passed
NonDPMean2	ϵ	false	false	passed
SVT1	ϵ	passed	-	passed
SVT2	ϵ	passed	-	passed
SVT3	ϵ	passed	-	passed
SVT4	ϵ	passed	-	passed
SVT5	ϵ	false	-	false
SVT6	ϵ	passed	-	passed
Joint	ϵ	passed	-	passed
WrongSensitivitySGD	(α, ϵ)	false	passed	-
ScaledSGD	(α, ϵ)	false	passed	-
NonADPMean	(α, ϵ)	false	passed	-

and is not defined for high dimensional output spaces, and cannot apply it to sparse vector technique algorithms. RenyiTester misses SVT3 and SVT6 but catches all the errors for at least a pair of parameters (α, ϵ). DP-Sniper suceeds at detecting the same mechanisms that RenyiTester . However, it requires $10M$ samples while RenyiTester only needs 400K samples.

Rényi DP results. RenyiTester detects all errors while the aDPTester does not, even when varying the outcome's space discretization size. It does so by evaluating less than 10 pairs of neighboring datasets (we present average number of trials in the appendix). DP-Sniper does not apply in this setting. In the appendix we further investigate the potential of RenyiTester to detect ScaledSGD 's implementation for different values of $\sigma_{\text{effective}}$. WrongSensitivitySGD presents an example where exploring extremal datasets is not useful for catching privacy violations but our dataset generation technique can find pairs of datasets violating the privacy constraint on an average of 5 trials. In this case, assuming gradients are in the $[-2, 2]$ interval, and assuming a clip norm of $G = 1$, the privacy violation occurs at datasets neighboring datasets $D = \{-1\}$ and $D' = \{-1, 2\}$, where the sensitivity of the clipped averaged gradient is 2 and not at neighboring datasets $D = \{-2\}$ and $D = \{-2, 2\}$ where the sensitivity is 1.

It is important to highlight that our implementation for detecting errors for higher values of $\sigma_{\text{effective}}$ is mostly limited due to the cap C used to define the space Φ. This capping parameter noticeably delivers smaller divergence estimates making it harder to find privacy leaks. Unfortunately, increasing this constant substantially increases the required $\Omega(e^{\alpha C})$ sample size. In the following section we find that removing this cap provides very accurate estimation for Gaussian distributions. We leave tightening the sample complexity as future work.

The high sample complexity for measuring divergence distribution seems to be universal. In table 1 we add the number of samples for RenyiTester , aDPTester , and DPSniper . RenyiTester requires at least one order of magnitude less than baselines and does not need a discretization parameter m.

6 Discussion

We presented a new test for detecting privacy violations that is suited to pure and Rényi differential privacy and, hence, is able to detect flaws in non-private mechanisms. While failing to detect a few pure differential privacy leaks, it appears to be the first one to test Rényi differential privacy guarantees with only black-box access to the mechanism. We highlight that our tester is particularly flexible and that it can easily be improved as we derive better sample complexity bounds for variational approaches of Rényi divergence estimators. As demonstrated in appendix B.6, there is still a noticeable gap between the theoretical and practical error bounds on these estimates. We leave possible theoretical improvements as a future area of research.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM SIGSAC conference on computer and communications security*, 2016.

[2] Andrew, Galen and Chien, Steve and Papernot, Nicolas . Tensorflow privacy.

[3] Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*, pages 1138–1156. IEEE, 2022.

[4] Zahra Beheshti and Siti Mariyam Hj Shamsuddin. A review of population-based meta-heuristic algorithms. *International Journal of Advances in Soft Computing and its Applications*, 5(1):1–35, 2013.

[5] Benjamin Bichsel, Samuel Steffen, Ilija Bogunovic, and Martin Vechev. Dp-sniper: Black-box discovery of differential privacy violations using classifiers. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 391–409. IEEE, 2021.

[6] Jeremiah Birrell, Paul Dupuis, Markos A Katsoulakis, Luc Rey-Bellet, and Jie Wang. Variational representations and neural network estimation of Rényi divergences. *SIAM Journal on Mathematics of Data Science*, 2021.

[7] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *ACM SIGSAC Conference on Computer and Communications Security*, 2020.

[8] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 2002.

[9] Zeyu Ding, Yuxin Wang, Guanhong Wang, Danfeng Zhang, and Daniel Kifer. Detecting violations of differential privacy. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, page 475–489. Association for Computing Machinery, 2018.

[10] Kashyap Dixit, Madhav Jha, Sofya Raskhodnikova, and Abhradeep Thakurta. Testing the Lipschitz property over product distributions with applications to data privacy. In *Theory of Cryptography Conference (TCC*, 2013.

[11] Carles Domingo-Enrich and Youssef Mroueh. Auditing differential privacy in high dimensions with the kernel quantum Rényi divergence. *arXiv preprint arXiv:2205.13941*, 2022.

[12] Vadym Doroshenko, Badih Ghazi, Pritish Kamath, Ravi Kumar, and Pasin Manurangsi. Connect the dots: Tighter discrete approximations of privacy loss distributions. *Proceedings on Privacy Enhancing Technologies*, 2022.

[13] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, 2006.

[14] Badih Ghazi, Pasin Manurangsi, Pritish Kamath, Ravi Kumar Ravikumar, and Vadym Doroshenko. Connect the dots: Tighter discrete approximations of privacy loss distributions. *arXiv preprint arXiv:2207.04380*, 2022.

[15] Manuel Gil, Fady Alajaji, and Tamas Linder. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 2013.

[16] Anna C Gilbert and Audra McMillan. Property testing for differential privacy. In *Allerton Conference on Communication, Control, and Computing*, 2018.

[17] Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and D. Sculley. Google vizier: A service for black-box optimization. In *International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*, 2017.

[18] Chuan Guo, Brian Karrer, Kamalika Chaudhuri, and Laurens van der Maaten. Bounding training data reconstruction in private (deep) learning. In *International Conference on Machine Learning (ICML)*. PMLR, 2022.

[19] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 2020.

[20] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *USENIX Security Symposium*, 2019.

[21] Akshay Krishnamurthy, Kirthevasan Kandasamy, Barnabas Poczos, and Larry Wasserman. Nonparametric estimation of renyi divergence and friends. In *International Conference on Machine Learning*. PMLR, 2014.

[22] Fred Lu, Joseph Munoz, Maya Fuchs, Tyler LeBlond, Elliott V Zaresky-Williams, Edward Raff, Francis Ferraro, and Brian Testa. A general framework for auditing differentially private machine learning. In *Advances in Neural Information Processing Systems*.

[23] Min Lyu, Dong Su, and Ninghui Li. Understanding the sparse vector technique for differential privacy. *arXiv preprint arXiv:1603.01699*, 2016.

[24] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

[25] Ilya Mironov. Rényi differential privacy. In *IEEE computer security foundations symposium (CSF)*, 2017.

[26] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[27] Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. *arXiv preprint arXiv:2302.07956*, 2023.

[28] Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on security and privacy (SP)*, 2021.

[29] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theory*, 56(11):5847–5861, 2010.

[30] Shadi Rahimian, Tribhuvanesh Orekondy, and Mario Fritz. Differential privacy defenses and sampling attacks for membership inference. In *ACM Workshop on Artificial Intelligence and Security*, 2021.

[31] Alfréd Rényi. On measures of entropy and information. In *Berkeley symposium on mathematical statistics and probability*, 1961.

[32] Xingyou Song, Sagi Perel, Chansoo Lee, Greg Kochanski, and Daniel Golovin. Open source vizier: Distributed infrastructure and api for reliable and flexible black-box optimization. In *Automated Machine Learning Conference, Systems Track (AutoML-Conf Systems)*, 2022.

[33] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[34] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

[35] Puning Zhao and Lifeng Lai. Minimax optimal estimation of kl divergence for continuous distributions. *IEEE Trans. Inf. Theor.*, 66(12):7787–7811, dec 2020.