Journal of Privacy and Confidentiality
Vol. 12 (1) 2022

Submitted Dec 2021
Published Jul 2022

A UNIFIED INTERPRETATION OF THE GAUSSIAN MECHANISM FOR DIFFERENTIAL PRIVACY THROUGH THE SENSITIVITY INDEX

GEORGIOS KAISSIS, MORITZ KNOLLE, FRIEDERIKE JUNGMANN, ALEXANDER ZILLER, DMITRII USYNIN, AND DANIEL RUECKERT

GK, DU: Institute for Artificial Intelligence in Medicine and Healthcare and Institute of Radiology, Technical University of Munich; Department of Computing, Imperial College London; OpenMined

MK, DR: Institute for Artificial Intelligence in Medicine and Healthcare and Institute of Radiology, Technical University of Munich; Department of Computing, Imperial College London

FJ, AZ: Institute for Artificial Intelligence in Medicine and Healthcare and Institute of Radiology, Technical University of Munich

Abstract. The Gaussian mechanism (GM) represents a universally employed tool for achieving differential privacy (DP), and a large body of work has been devoted to its analysis. We argue that the three prevailing interpretations of the GM, namely (ε, δ)-DP, f-DP and Rényi DP can be expressed by using a single parameter ψ, which we term the *sensitivity index*. ψ uniquely characterises the GM and its properties by encapsulating its two fundamental quantities: the sensitivity of the query and the magnitude of the noise perturbation. With strong links to the ROC curve and the hypothesis-testing interpretation of DP, ψ offers the practitioner a powerful method for interpreting, comparing and communicating the privacy guarantees of Gaussian mechanisms.

1. Introduction

Differential Privacy (DP) is the gold standard technique for providing quantifiable privacy guarantees to individuals whose sensitive data is subjected to algorithmic processing (Dwork and Roth, 2014). Since its inception, its applicability to real-world machine learning and statistics tasks has continuously increased, as witnessed by its utilisation in the US census (Abowd, 2018; Dwork, 2019) and recent large-scale deployments in industry (Apple, 2017;

Key words and phrases: Differential Privacy, Gaussian Mechanism, Rényi Differential Privacy, Gaussian Differential Privacy, Receiver-Operator Characteristic, Hypothesis Testing.

Correspondence should be addressed to Georgios Kaissis, g.kaissis@tum.de.

These authors contributed equally: Georgios Kaissis and Moritz Knolle.

JPrivacy Confidentiality www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.807

Erlingsson et al., 2014). DP is typically realised through noise perturbation of query outputs over sensitive databases.

The Gaussian mechanism (GM) is the prototypical mechanism for obtaining (ε, δ)-DP, especially in the setting of high-dimensional queries. A large body of research has been devoted to better characterising the GM and has proposed relaxations with desirable properties, such as facilitated composition (Mironov, 2017). However, to practitioners without expert-level theoretical knowledge, translating the concepts presented in these works to tangible, real-world applications is arguably difficult, which may ultimately hinder DP's broad deployment and democratisation. Moreover, many predominantly theoretical works are also highly topical and can thus be hard to contextualise within the broader field of DP. This bears the risk of incorrectly choosing, constructing or applying DP mechanisms (Lyu et al., 2016), potentially resulting in the unintended disclosure of sensitive information, such as medical or financial data. Another layer of complexity emerges upon deployment of DP tools to user-facing products: For example, *Apple*'s use of DP was criticised as insufficient due to its resetting of users' privacy budgets on a frequent basis and poor communication (Tang et al., 2017), further aggravated by the high complexity of the subject matter.

Promoting trust and acceptance of DP and reinforcing its meaningful application will thus be contingent on a clear understanding of the underlying mechanisms, such as the GM (Dwork et al., 2019; Cummings et al., 2021). This will not only facilitate clear communication to users and developers, but also help bridge the gap between foundational DP research and its translation into task-specific tools, e.g. software libraries. Here, we present a unifying perspective on the GM which encapsulates its fundamental properties into an interpretable quantity and intended as a reference to practitioners. Our contributions towards achieving the above-mentioned goal are as follows:

- We demonstrate that the comprehensive characterisation of the GM requires knowledge of only a single parameter, termed the *sensitivity index* ψ, which captures the two fundamental properties of the mechanism, namely the sensitivity of the query function and the magnitude of the noise perturbation;
- We prove that ψ fundamentally links the (ε, δ)-interpretation, the hypothesis-testing interpretation of f-DP and the divergence-based interpretation of Rényi DP and can be used to translate between them;
- Due to its link to the *receiver-operator characteristic* (ROC) curve, a universally applied tool in machine learning and statistics, the sensitivity index can be easily understood and communicated by and to practitioners;
- ψ leads to a intuitive geometric interpretation of the GM, as it can be used to express privacy loss in terms of the area under the ROC curve (AUC). This facilitates comparisons among mechanisms;
- Finally, we theoretically and empirically demonstrate the optimal conversion strategy between ψ-DP and (ε, δ)-DP.

2. Prior Work

DP and the GM were originally introduced by Dwork and Roth (2014). In its original form, the GM was constrained to the *high privacy* regime, that is, values of $\varepsilon \in (0, 1)$. This constraint was lifted in the later work by Balle and Wang (2018), who extended the analysis of the GM to arbitrary ε values by employing the cumulative distribution function of the normal distribution instead of the previously used tail bounds. Rényi DP (Mironov, 2017) was proposed as a natural relaxation of DP, based on the divergence of the same name. It provides favourable properties under composition, however suffers from a lossy conversion to (ε, δ)-DP, to which improvements have only recently been proposed (Balle et al., 2020; Asoodeh et al., 2021). f-DP was proposed by Dong et al. (2019) and is conceptually the closest to our work in that it expresses the properties of the DP mechanism in terms of a *trade-off function*, which is an affine transformation of the ROC curve used by us. The ROC curve was used to quantify privacy guarantees in Matthews et al. (2010) and Matthews and Harel (2012), however these works pre-date more modern developments in DP, notably Rényi DP and f-DP, and do not describe machine learning applications. We moreover rely on the *hypothesis testing interpretation* of DP and on results due to Wasserman and Zhou (2010) and Kairouz et al. (2015), who introduce the notion of a *privacy region*, which is conceptually related to the AUC utilised in our work. The term *sensitivity index* is folklore and was first utilised as early as 1965 (Dempster and Schatzoff, 1965). It (and the synonymous term *discriminability index*) is nowadays typically used to express the ratio between the mean separation and the standard deviation of two Gaussian distributions with equal variance (Das and Geisler, 2021) (their Mahalanobis distance (Mahalanobis, 1936)), which motivates its utilisation in our work. We are not aware of any prior use of this term in DP literature.

3. Preliminaries

We begin by introducing key terminology used in the remainder of the work. We will consider the case in which a trusted curator in possession of a sensitive database wants to employ the GM to privatise the outputs of some function (synonymously, *query*) applied to the database in order to privately publish the results while being able to offer the individuals whose data is contained in the database a quantifiable privacy guarantee.

We will refer to the sensitive database (synonymously, *dataset*, to designate that individuals are present in the database only once) as D. Its adjacent database, designated as D' can be constructed from D by adding or removing a single individual's data, and we will use the symbol $\simeq$ to denote adjacency.

We assume that a (query) function q is applied to the database. An example of such a query is counting how many individuals in the database exhibit a certain attribute. Moreover, the execution of a training step of e.g. a neural network resulting in the publication of a gradient update is also considered a query.

We will use a similar formalism as Laud et al. (2020) to describe the global sensitivity of q. We consider the set X of all possible databases a (Banach) space equipped with a metric d_X which expresses the distance between them (here, the Hamming metric). The query function q maps an element of X to an element of an output space Y. We define the *global sensitivity* Δ of q as follows:

Definition 1 (Global sensitivity Δ of q)**.** The global sensitivity Δ of q is:

$$\Delta(q) = \sup_{D \simeq D'} \frac{d_Y\left(q(D), q\left(D'\right)\right)}{d_X\left(D, D'\right)}. \tag{3.1}$$

Here, the sup is taken over all adjacent dataset pairs whose Hamming distance equals 1. We note that alternative definitions of adjacency may use a different metric or result in a different distance between adjacent databases (such as when an individual is replaced by another), but we only use the adjacency definition described above in this work. Thus, when Y is the Euclidean space equipped with the L_2 norm, we can equivalently write:

$$\Delta_2(q) = \sup_{D \simeq D'} \|q(D) - q(D')\|_2. \tag{3.2}$$

Equation 3.2 describes the global L_2-sensitivity of q, which forms the basis of our discussion. We will henceforth omit the subscript and argument and refer to the global L_2-sensitivity as Δ.

A DP *mechanism* is a higher-order function which takes as an input a database, the query function, its sensitivity and one or more hyperparameters which control how much noise to apply to the output of the query before publishing the privatised result. We will limit the scope of our discussion to the *Gaussian* mechanism on real-valued queries (Dwork and Roth, 2014):

Definition 2 (Gaussian mechanism)**.** The Gaussian mechanism $\mathcal{M}$ on the query function $q : X \to \mathbb{R}^d$ with sensitivity Δ applied over a database $D \in X$ outputs:

$$\mathcal{M}(q(D)) = q(D) + \xi,\ \xi \sim \mathcal{N}(0, \sigma\, \mathbf{I}_d), \tag{3.3}$$

where σ denotes the standard deviation of the normal distribution $\mathcal{N}$ and is calibrated to the sensitivity Δ such that $\sigma = \underline{\mathrm{O}}(\frac{\Delta}{\varepsilon})$, where $\underline{\mathrm{O}}$ denotes a lower bound due to an additional parameter δ, described below. $\mathbf{I}_d$ denotes the identity matrix with d diagonal elements.

The GM can be used to render the publication of q's output differentially private with respect to the individuals in the database. The publication of $\mathcal{M}$'s outputs results in *privacy loss*, which affects the individuals whose records are contained in q's inputs. The exact quantification of this privacy loss (for example using the parameters (ε, δ)) and constraining it using specified mechanisms is central to the study of DP.

Definition 3 ((ε, δ)-DP, Dwork and Roth (2014))**.** We say that the randomised mechanism $\mathcal{M}$ preserves (ε, δ)-DP if, for all pairs of adjacent databases D and D' and all subsets $\mathcal{S}$ of $\mathcal{M}$'s range:

$$\mathbb{P}(\mathcal{M}(q(D) \in \mathcal{S})) \leq e^{\varepsilon}\, \mathbb{P}(\mathcal{M}(q(D') \in \mathcal{S})) + \delta. \tag{3.4}$$

We note that the definition is symmetric. At a global L_2-sensitivity of Δ and a noise standard deviation of σ, the GM satisfies (ε, δ)-DP with:

$$\varepsilon(\delta) = \frac{\Delta}{\sigma} \sqrt{2 \log\left(\frac{1.25}{\delta}\right)}, \tag{3.5}$$

for $\varepsilon \in (0, 1)$. The $\varepsilon(\delta)$ notation denotes a relationship between the two parameters which will be explained below. In brief, a single (ε, δ)-tuple is insufficient to comprehensively describe the privacy loss attributes of the GM. An extension to values of $\varepsilon \geq 1$ is presented below.

(ε, δ)-DP has several beneficial properties, among others *group privacy*, *closure under post-processing* and *composition*, which are described in detail in Dwork and Roth (2014).

Rényi DP (RDP) was introduced by Mironov (2017) and interprets the DP guarantees in an information theoretic way. Intuitively, two distributions P and Q are *close* to each other when the Rényi divergence from one to the other is small. RDP utilises this fact to specify DP guarantees. The Rényi divergence is a parameterised divergence whose parameter α can be used to bound some moment of the ratio between two distributions and given by:

$$D_\alpha(P \parallel Q) = \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim Q} \left(\frac{P(x)}{Q(x)} \right)^\alpha . \tag{3.6}$$

Using the Rényi divergence, RDP is defined as follows:

Definition 4 (Rényi-DP)**.** $\mathcal{M}$ preserves (α, ρ)-Rényi-DP (RDP) (Mironov, 2017) if, for all pairs of adjacent databases D and D':

$$D_\alpha \left(\mathcal{M}(q(D)) \parallel \mathcal{M}(q(D')) \right) \leq \rho, \tag{3.7}$$

where D_α denotes the Rényi divergence of order $\alpha > 1$ described above. At $\alpha = 1$, defined by continuity, this corresponds to bounding the Kullback-Leibler-divergence. Of note, D_∞-RDP is equivalent to $(\varepsilon, 0)$-DP (which the GM never achieves). This definition is asymmetric, as is the Rényi divergence (but this detail is often insignificant in practice). We will abuse notation and write $D_\alpha(\mathcal{M}_1 \parallel \mathcal{M}_2)$ to denote $\max\{D_\alpha(\mathcal{M}_1 \parallel \mathcal{M}_2), D_\alpha(\mathcal{M}_2 \parallel \mathcal{M}_1)\}$.

Beyond RDP and the broader class of *divergence-based* DP definitions, DP interpretations relying on the techniques of *statistical hypothesis testing* have also been introduced. Here, a hypothetical adversary is assumed to conduct a statistical test to distinguish between the two databases and the strength of the DP guarantee is intuitively measured by how (un-)successful such a statistical test is. The most notable hypothesis testing interpretation is f-DP:

Definition 5 (f-DP)**.** $\mathcal{M}$ preserves f-DP if, for all pairs of adjacent databases D and D':

$$T(\mathcal{M}(q(D)), \mathcal{M}(q(D'))) \geq f \tag{3.8}$$

where T denotes a trade-off function of the form:

$$T(P, Q)(\alpha) = \inf \{\beta_\phi : \alpha_\phi \leq \alpha\} . \tag{3.9}$$

Here, ϕ is a rejection rule such that $0 \leq \phi \leq 1$, P and Q are two probability distributions on the same space and α, β are the Type I and Type II statistical errors of a hypothesis test. This test is assumed to be *optimal* in the sense that it has the highest power among all tests for distinguishing between P and Q (or, in the case of DP, between D and D'). The Neyman-Pearson lemma (Neyman and Pearson, 1933) motivates constructing such a test via the log-likelihood ratio, as will be discussed below. Gaussian DP (GDP) represents a specialisation of f-DP to the GM. Here, the trade-off function has a closed-form representation:

$$G_\mu(\alpha) = \Phi \left(\Phi^{-1}(1 - \alpha) - \mu \right) , \tag{3.10}$$

where Φ is the cumulative distribution of the standard normal distribution. A mechanism which fulfils $T(\mathcal{M}(q(D)), \mathcal{M}(q(D'))) \geq G_\mu$ is said to preserve μ-GDP where $\mu = \frac{\Delta}{\sigma}$. This fact establishes a strong link between μ-GDP and ψ-DP, which will be explored below.

We note that, whereas RDP conveys the properties of the GM as the value of a divergence function from the output distribution of $\mathcal{M}$ applied to $q(D)$ to the output distribution of $\mathcal{M}$ applied to $q(D')$, f-DP utilises the notion of a *trade-off* between the *Type I* and *Type II* statistical errors the adversary is facing when trying to distinguish between the aforementioned outputs.

Expressing the similarities between RDP, f-DP and the (ε, δ)-interpretation through the sensitivity index ψ represents the core of our work. We therefore define the sensitivity index as follows:

Definition 6 (Sensitivity index ψ)**.** The sensitivity index ψ is defined as:

$$\psi = \frac{\Delta}{\sigma}. \tag{3.11}$$

The aforementioned introductory remarks allow us to reason about the GM using the sensitivity index ψ and to define ψ-DP:

Definition 7 (ψ-DP)**.** Let q be a query function with global L_2-sensitivity Δ and let $\mathcal{M}$ be an instance of the Gaussian Mechanism with noise standard deviation σ. Then, the application of $\mathcal{M}$ to an output of q satisfies ψ-DP with $\psi = \frac{\Delta}{\sigma}$.

As is apparent from Equations 3.2 and 3.3, ψ encapsulates the two parameters which completely determine the privacy properties of the GM in a single parameter, and can be interpreted as the *signal-to-noise ratio* (SNR) of the mechanism, where Δ represents the signal and σ the noise. Intuitively, a ψ-DP mechanism preserves strong privacy if it has low SNR, since the distinguishing details of the individuals in the database are "drowned out" by the noise. Moreover, as the GM relies on a random noise draw and its behaviour is otherwise independent of the query function and the input database, one may state that ψ is a *singular* parameter of the GM. Furthermore, ψ-DP (like other DP definitions), offers a *worst-case interpretation*: As discussed above, ψ expresses the ratio between the mean separation and standard deviation of two normal distributions with equal variance (equivalently, the Mahalanobis distance between them). In our work, the invocation of the GM results in an output which can be considered a random variable drawn from a normal distribution induced by the randomness of the mechanism. The worst-case interpretation assumes that the means of these distributions are separated by Δ (i.e. maximally separated) when the mechanism is executed over adjacent databases. We note that ψ-DP is a formulation *specific* to the GM in the sense that it is designed to hold for the global L_2 sensitivity and for Gaussian noise. However, a similar quantity to the sensitivity index arises e.g. in the Laplace mechanism under the (ε, δ)-DP interpretation, where $\varepsilon = \frac{S}{b}$ with S denoting the global L_1-sensitivity and b denoting the scale parameter of the Laplace distribution from which noise is added. We will not expand on this tangent further in this work, but note it to be of interest as it shows the SNR analogy to be useful beyond the GM and a similar concept as the sensitivity index to exist.

4. Bridging the gap between DP definitions

Throughout this and the following sections, we will utilise the terminology and notation introduced above. We furthermore introduce an adversary $\mathcal{A}$, who has black-box access to a single output of $\mathcal{M}$ and attempts to determine whether the observed output originated from q being executed over D vs. D'. We grant $\mathcal{A}$ access to arbitrary side-information including e.g. an understanding of the inner workings of $\mathcal{M}$, the characteristics of the noise distribution, the query function q, as well as the ability to arbitrarily post-process the output of $\mathcal{M}$.

4.1. Relating the sensitivity index to (ε, δ)-DP.

We assume that $\mathcal{A}$ observes an output O from $\mathcal{M}$. The *privacy loss random variable on* O, Ω is then defined as:

$$\Omega_{\mathcal{M}(q(D))} = \log\left(\frac{\mathbb{P}(\mathcal{M}(q(D)) = O)}{\mathbb{P}(\mathcal{M}(q(D')) = O)}\right). \tag{4.1}$$

where log is the natural logarithm. As is common in literature, we will use the notation $\Omega_{\mathcal{M}(q(D))}$ to specify that Ω quantifies the privacy loss when $\mathcal{M}$ is executed whilst considering D (rather than D') the "base" dataset (Dwork and Rothblum, 2016). Under the (ε, δ)-DP definition, the magnitude of the privacy loss random variable is bounded by ε with probability $1-\delta$. We are therefore interested in the probability δ that the magnitude of $\Omega_{\mathcal{M}(q(D))}$ *exceeds* a certain value ε. By Definition 3:

$$\mathbb{P}(\Omega_{\mathcal{M}(q(D))} \geq \varepsilon) \leq \delta. \tag{4.2}$$

We begin by noting that $\Omega_{\mathcal{M}(q(D))}$ is also normally distributed with mean $\frac{\Delta^2}{2\sigma^2}$ and standard deviation $\frac{\Delta}{\sigma}$ (Dwork and Rothblum, 2016):

$$\Omega_{\mathcal{M}(q(D))} \sim \mathcal{N}\left(\frac{\Delta^2}{2\sigma^2}, \frac{\Delta}{\sigma}\right) \overset{\text{Def.6}}{=} \mathcal{N}\left(\frac{1}{2}\psi^2, \psi\right). \tag{4.3}$$

where we have substituted $\|q(D) - q(D')\|_2 = \Delta$, assuming the worst-case distribution without loss of generality. The distributions of $\Omega_{\mathcal{M}(q(D))}$ and of the outputs of $\mathcal{M}$ are exemplified in Figure 1. Here and below we will slightly abuse notation by conceptually equating distributions to their density functions.

We will bound δ using the cumulative distribution function of $\Omega_{\mathcal{M}(q(D))}$ to relate $\Omega_{\mathcal{M}(q(D))}$ to ψ:

Lemma 1. If $\mathcal{M}$ is $(\varepsilon, \delta(\varepsilon))$-DP $\forall\ \varepsilon \geq 0, \delta \in [0, 1]$, the following inequality holds:

$$\Phi\left(\frac{1}{2}\psi - \frac{1}{\psi}\varepsilon\right) \leq \delta. \tag{4.4}$$

We recall that Φ represents the cumulative distribution function of the standard normal distribution.

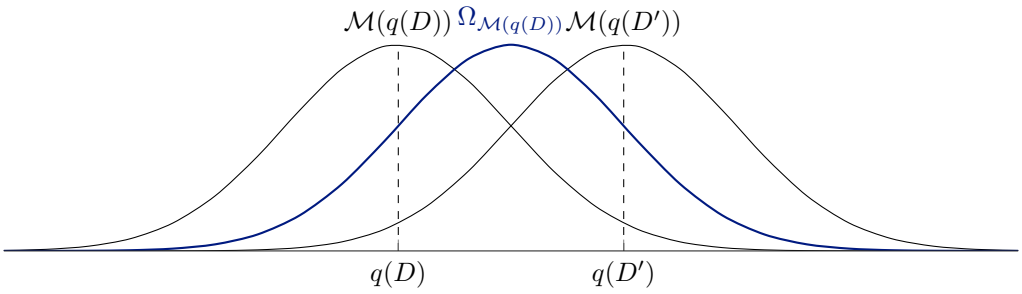

FIGURE 1. Exemplary plot showing the distributions of the Gaussian mechanism $\mathcal{M}$'s outputs given input databases D and D' (black curves). The *privacy loss random variable* $\Omega_{\mathcal{M}(q(D))}$ (blue curve) is also normally distributed, a particular property of the Gaussian mechanism.

Proof.

$$\mathbb{P}\left(\Omega_{\mathcal{M}(q(D))} \geq \varepsilon\right) = \mathbb{P}\left(\mathcal{N}\left(\frac{1}{2}\psi^2, \psi\right) \geq \varepsilon\right) =$$
$$= \mathbb{P}\left(\mathcal{N}\left(0,1\right) \geq \frac{\varepsilon - \frac{1}{2}\psi^2}{\psi}\right).$$

By symmetry of the standard normal distribution:

$$\mathbb{P}\left(\mathcal{N}\left(0,1\right) \geq \frac{\varepsilon - \frac{1}{2}\psi^2}{\psi}\right) = \mathbb{P}\left(\mathcal{N}\left(0,1\right) \leq \frac{\frac{1}{2}\psi^2 - \varepsilon}{\psi}\right) =$$
$$= \Phi\left(\frac{\psi^2}{2\psi} - \frac{\varepsilon}{\psi}\right) = \Phi\left(\frac{1}{2}\psi - \frac{1}{\psi}\varepsilon\right).$$

The inverse case, $\mathbb{P}\left(\Omega_{\mathcal{M}(q(D'))} \leq -\varepsilon\right)$, required by the symmetry of Definition 3, follows by the same argument. □

Notably, the $(\varepsilon, \delta(\varepsilon))$ notation in Lemma 1 implies the existence of infinitely many valid (ε, δ) pairs, corresponding to the *privacy profile* of $\mathcal{M}$ (Balle et al., 2018). Moreover, it forms the foundation of the *Analytic Gaussian Mechanism* (Balle and Wang, 2018), which we will utilise later. Also of note, this formulation holds for any $\varepsilon > 0$. Building upon this relationship between the sensitivity index ψ and the (ε, δ)-DP definition, we now link ψ to the hypothesis-testing DP interpretation, which we express via the ROC curve and its AUC.

4.2. **Relating the sensitivity index to the ROC curve.** The *probabilistic* definition of DP we investigated above is centred around quantifying $\mathcal{A}$'s (posterior) information gain as a function of their prior beliefs and the information disclosed by publishing the output of $\mathcal{M}$. Orthogonal to this approach, we now investigate the ability of $\mathcal{A}$ to distinguish between these outputs using a statistical hypothesis test. Formally, the adversary posits a *null hypothesis* (e.g. H_0: *the underlying database is* D) and an *alternative hypothesis* (H_1*: the underlying database is* D') and then tries to determine whether to reject or fail to reject the null hypothesis. This interpretation is closely related to the probabilistic definition above, as the *Neyman-Pearson-Lemma* (Neyman and Pearson, 1933) motivates constructing

the hypothesis test by relying on the *likelihood ratio* to achieve optimal statistical power. This links it to the privacy loss random variable via the right hand side of Equation (4.1), itself a (log-) likelihood ratio. In intuitive terms, $\mathcal{A}$'s task can be summarised as follows:

(1) $\mathcal{A}$ formulates a binary classification problem concerned with discriminating whether the observed output O of $\mathcal{M}$ is more likely given input D or D'.

(2) $\mathcal{A}$ develops a classification algorithm $\mathcal{H}$ which takes as its input O and a parameter c, termed the *cut-off-value* or *decision threshold*. $\mathcal{H}$ deterministically outputs a decision $\chi \in \{\text{“}D\text{”}, \text{“}D'\text{”}\}$ depending on the observed values of O and the value of c.

Our further analysis will rely on the following fundamental question: *what is the optimal true-positive rate (TPR) $\mathcal{H}$ can afford $\mathcal{A}$ for any given false-positive rate (FPR) at any threshold c?* We will use Figure 2 to illustrate our argument.

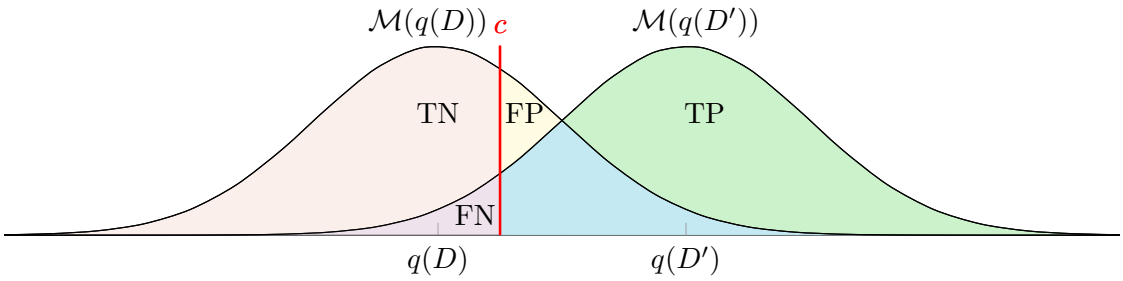

FIGURE 2. Illustration of the distributions of $\mathcal{M}$'s possible outputs and their relation to $\mathcal{A}$'s adversarial task of binary classification between D and D'. $\mathcal{A}$'s classification algorithm $\mathcal{H}$ outputs a decision d based on the *cut-off* or *decision threshold* c (red line). At any given value of c, $\mathcal{H}$ achieves a classification performance which is fully described by the rates of *true positive* (TP, green and blue areas), *false positive* (FP, yellow and blue areas), *true negative* (TN, pink area) and *false negative* (FN, light purple area) classifications.

The classification performance of $\mathcal{H}$ can be fully described by its *true positive rate* (TPR) and its *false positive rate* (FPR) (the *true negative* and *false negative* rates follow by the fact that $\text{TNR} + \text{FPR} = \text{FNR} + \text{TPR} = 1$. Intuitively, if $\mathcal{A}$ cannot simultaneously achieve a high TPR *and* a low FPR through the use of $\mathcal{H}$, then $\mathcal{M}$ preserves privacy. Evidently, the TPR and FPR are immediately dependent on the choice of c. The ROC curve represents the performance of $\mathcal{H}$ as c is varied as a plot of TPR against FPR.

To create the ROC curve from the TPR and FPR, we will rely on the following property:

Corollary 1. Let $P_{q(D)}(x)$ be the density function of $\mathcal{M}$ when $x \sim q(D)$ and $P_{q(D')}(x)$ be the density function of $\mathcal{M}$ when $x \sim q(D')$. Then:

$$\text{TPR}(c) = \int_{c}^{+\infty} P_{q(D)}(t)dt = 1 - \Phi_{q(D'),\sigma}(c) \text{ and}$$
$$\text{FPR}(c) = \int_{-\infty}^{c} P_{q(D')}(t)dt = 1 - \Phi_{q(D),\sigma}(c),$$

where $c \in (-\infty, +\infty)$.

Proof. (Pictorial) Notice in Figure 2 that $\mathcal{M}$'s outputs follow a normal distribution. Then, $\text{TPR}(c)$ is the sum of the green and blue shaded areas and $\text{FPR}(c)$ is the sum of the yellow and the blue shaded areas. □

We are now able to relate the sensitivity index to the classification performance of $\mathcal{H}$ via the ROC curve. We will use the following facts:

Corollary 2 (Adapted from (Gonçalves et al., 2014), (3.1))**.** Let X and Y be independent normal variables with means $\boldsymbol{\mu_0}$ and $\boldsymbol{\mu_1}$ and covariance matrices $\sigma_0^2\mathbf{I}$ and $\sigma_1^2\mathbf{I}$. Then, the *binormal* ROC curve (whose construction assumes that the two populations, in this case, the random variables, are normally distributed) is given by:

$$R(x) = \Phi\left(a + b\,\Phi^{-1}(x)\right), x \in (0,1). \tag{4.5}$$

where $a = \frac{\|\boldsymbol{\mu_1} - \boldsymbol{\mu_0}\|_2}{\sigma_1}$ and $b = \frac{\sigma_0}{\sigma_1}$.

Corollary 3 (Adapted from (Gonçalves et al., 2014), (3.2))**.** The area (AUC) under the *binormal* ROC curve is then given by:

$$\text{AUC} = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right). \tag{4.6}$$

where a and b are defined as above.

Lemma 2. Let $R : (0,1) \to [0,1]$ be the function of the ROC curve describing the classification performance of $\mathcal{H}$. Then:

$$R(x) = \Phi\left(\psi + \Phi^{-1}(x)\right). \tag{4.7}$$

Proof. The ROC curve is constructed by parametrically plotting:

$$(\text{FPR}(c), \text{TPR}(c)),\ c \in (-\infty, +\infty) \tag{4.8}$$

on the unit square and is strictly monotonically increasing (Gonçalves et al., 2014). Recognising that $P_{q(D)}(x)$ and $P_{q(D')}(x)$ follow a normal distribution with means $q(D)$ and $q(D')$ and common standard deviation σ, we can leverage the properties of the binormal ROC curve and use equation (4.5) to re-write Equation 4.8 as:

$$
\begin{aligned}
R(x) &= \Phi\left(\frac{\|q(D') - q(D)\|_2 + \sigma\Phi^{-1}(x)}{\sigma}\right) = \\
&= \Phi\left(\frac{\Delta + \sigma\Phi^{-1}(x)}{\sigma}\right) = \\
&= \Phi\left(\frac{\Delta}{\sigma} + \Phi^{-1}(x)\right) = \\
&= \Phi\left(\psi + \Phi^{-1}(x)\right), x \in (0,1).
\end{aligned}
$$

□

Lemma 3. The AUC of R is given by:

$$
\mathrm{AUC}_R = \Phi\left(\frac{\psi}{\sqrt{2}}\right). \tag{4.9}
$$

Proof. Substitute $a = \frac{\|q(D)-q(D')\|_2}{\sigma} = \frac{\Delta}{\sigma} = \psi$ and $b = \frac{\sigma}{\sigma} = 1$ in Equation 4.6. □

Note that in both cases, we have substituted $\|q(D) - q(D')\|_2 = \Delta$, as we are interested in the *worst-case* ROC curve and AUC, respectively.

From Equations 4.7 and 4.9, the following relationships are observed:

$$
\lim_{\psi\to+\infty} \mathrm{AUC}_R = 1. \tag{4.10}
$$

This case corresponds to a mechanism which offers *no privacy*, as $\mathcal{A}$ is able to distinguish between D and D' with a TPR $= 1$ and a FPR $= 0$ and the vertex of the graph of $R(x)$ approaches the point $(0, 1)$. This situation arises when $\Delta \to +\infty$ and/or when $\sigma \to 0$. Conversely,

$$
\lim_{\psi\to 0} \mathrm{AUC}_R = 0.5. \tag{4.11}
$$

In this case, $\mathcal{A}$ is unable to distinguish between D and D' and graph of $R(x)$ approaches the graph of TPR(FPR) = FPR, which is the line through the origin with unit slope. This situation arises when $\Delta \to 0$ and/or when $\sigma \to +\infty$ and corresponds to *perfect privacy*.

Figure 3 visually demonstrates the ROC curve and the AUC. As the AUC depends only on the value of ψ, which is sufficient to fully characterise the behaviour of the GM, the results above and the AUC can be utilised to (visually) compare the privacy guarantees between GMs. This represents an advantage over (ε, δ)-DP whose characterisation and comparison would require an infinite collection of $(\varepsilon, \delta(\varepsilon))$ tuples. Moreover, the two components of these tuples represent fundamentally different quantities, further encumbering the mental model.

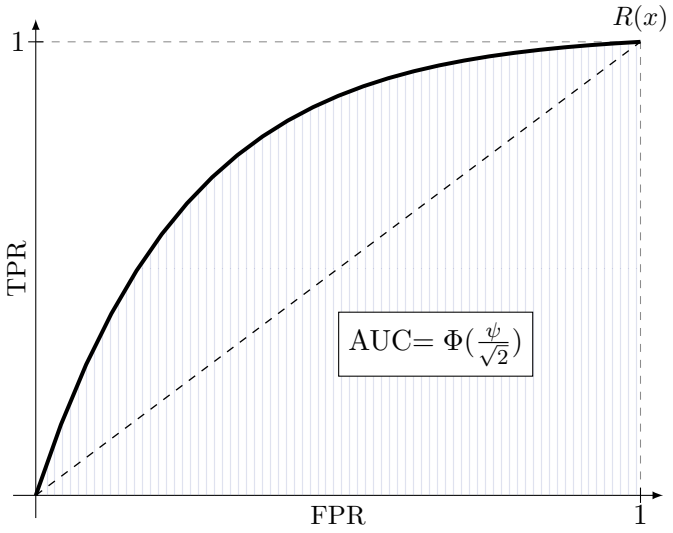

FIGURE 3. Exemplary illustration of the ROC curve $R(x)$ and of AUC_R. The ROC curve describes the performance of a classification algorithm at all decision thresholds by plotting its TPR against its FPR on the unit square.

4.3. **Recovering (ε, δ)-DP.** The ROC curve comprehensively characterises $\mathcal{M}$ at *all* possible threshold values. We now turn to the question of how to losslessly convert between the *privacy profile* of $\mathcal{M}$ and $R(x)$. From the works of Wasserman and Zhou (2010) and Kairouz et al. (2015), the following relationship is known:

Corollary 4 (Wasserman and Zhou (2010) and Kairouz et al. (2015), Theorem (2.1))**.** For a classification algorithm $\mathcal{H}$ executed on the outputs of an (ε, δ)-DP mechanism, the following holds:

$$\mathrm{TPR}_{\mathcal{H}} \leq e^{\varepsilon}\ \mathrm{FPR}_{\mathcal{H}} + \delta. \tag{4.12}$$

This implies a linear relationship between $\mathrm{TPR}_{\mathcal{H}}$ and $\mathrm{FPR}_{\mathcal{H}}$. Moreover, the *optimal* classifier's graph in this setting must also be tangent to $R(x)$, which represents the worst case in terms of privacy (Metz, 1978). This also follows intuitively by the fact that the vertex of such a classifier's graph lies closest to the point of perfect discrimination $(\mathrm{FPR}, \mathrm{TPR}) = (0, 1)$.

Based on these assumptions, we can formulate the following relationship:

Lemma 4. The tangent with slope e^{ε} to the graph of a classifier $\mathcal{H}$ with ROC curve $R(x)$ has intercept at most δ.

Proof. To derive the tangent's intercept, we rely on the *Legendre-Fenchel*-transform of R (which, as strictly monotonically increasing, is amenable to such transformation). Let $R^{\star}$ be the convex conjugate of R. Then:

$$R^{\star}(t) = ty - R(y). \tag{4.13}$$

The condition $\frac{dR^\star}{dt} = 0$ is satisfied when:

$$t = \Phi(m), m = -\frac{\Delta^2 + 2\sigma^2 \log(y)}{2\Delta\sigma}. \tag{4.14}$$

Then, substituting t into Equation (4.7):

$$\begin{aligned} R(t) &= \Phi\left(\frac{\Delta}{\sigma} + \Phi^{-1}(\Phi(m))\right) = \\ &= \Phi\left(\frac{\Delta}{\sigma} + m\right) = \\ &= \Phi\left(\frac{\Delta}{\sigma} - \frac{\Delta^2 + 2\sigma^2 \log(y)}{2\Delta\sigma}\right) = \\ &= \Phi\left(\frac{\Delta}{\sigma} - \frac{\Delta^2}{2\Delta\sigma} - \frac{2\sigma^2 \log(y)}{2\Delta\sigma}\right) = \\ &= \Phi\left(\frac{\Delta}{2\sigma} - \frac{\sigma \log(y)}{\Delta}\right). \end{aligned}$$

When $t = e^\varepsilon$:

$$\begin{aligned} R(t)\big|_{t=e^\varepsilon} &= \Phi\left(\frac{\Delta}{2\sigma} - \frac{\sigma \log(e^\varepsilon)}{\Delta}\right) = \\ &= \Phi\left(\frac{\Delta}{2\sigma} - \frac{\sigma\varepsilon}{\Delta}\right) \overset{\text{Lemma 1}}{=} \\ &= \Phi\left(\frac{1}{2}\psi - \frac{1}{\psi}\varepsilon\right) \leq \delta. \end{aligned}$$

□

This relationship is visualised in Figure 4. We note that a similar case is observed when the slope of the line is $e^{-\varepsilon}$ (required by the symmetric definition of (ε, δ)-DP), whereby the tangent in question is reflected about the diagonal $y = -x + 1$. The resulting set of infinitely many symmetric line pairs draw the boundary of the ROC curve and are conceptually equivalent to the lines bounding the *privacy region* in Kairouz et al. (2015) and the *trade-off* function in Dong et al. (2019).

Of note, this fact also admits an interesting geometric interpretation: The (fictitious) ROC curve of a "failed" privacy mechanism, that is, a mechanism for which the δ-probability event of catastrophic failure has occurred, is shown in Figure 5. The curve intercepts the y-axis at a point $(0, \kappa), \kappa > 0$, indicating that $\mathcal{A}$ is able to discriminate between D and D' with a FPR of 0 while still achieving a TPR of κ. A similar result can be found in Figure 3 of Pless et al. (2003).

5. Relating the sensitivity index to f-DP and Rényi DP

In the previous sections, we introduced the sensitivity index ψ as a *singular* value comprehensively characterising the GM and related it to the mechanism's $(\varepsilon, \delta(\varepsilon))$ privacy profile.

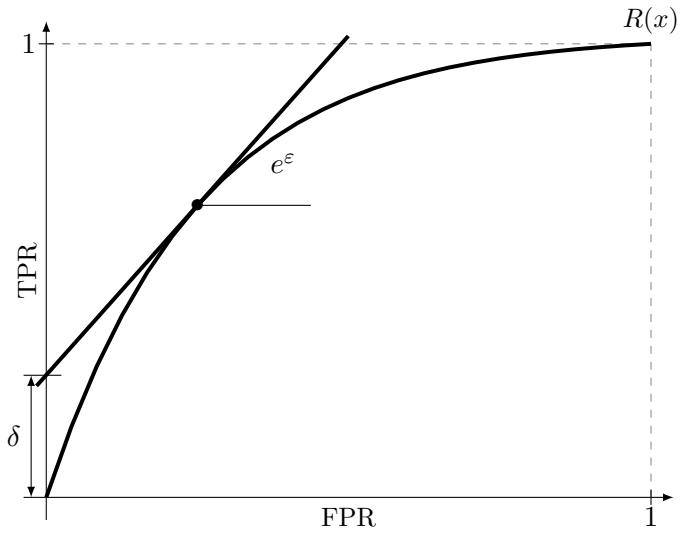

FIGURE 4. Relationship between the tangent with slope e^ε to the ROC curve $R(x)$ and its intercept δ. We note that, although we have plotted the probability δ and the TPR (which also represents a probability) on the same axis, the two are distinct (and usually on very different scales).

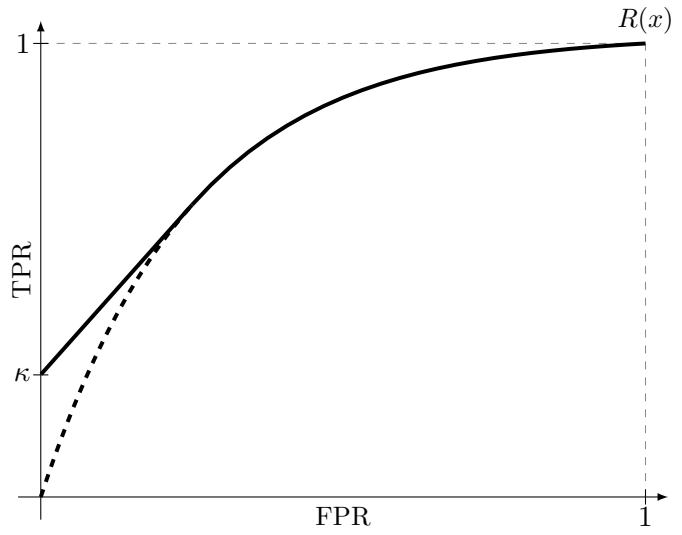

FIGURE 5. Fictitious ROC curve of a failed privacy mechanism $\mathcal{M}$ (solid line) compared to its actual ROC curve (dashed continuation). The adversary $\mathcal{A}$ is able to distinguish between D and D' with a TPR of $\kappa > 0$ while maintaining an FPR of 0.

We now study the relationship between ψ and two newer DP interpretations, f-DP and Rényi DP.

5.1. **Converting between ψ-DP and f-DP.** We begin by recalling that the f-DP framework (Dong et al., 2019) is based on the *hypothesis testing* interpretation of DP and employs *trade-off* functions between $\mathcal{A}$'s *Type I* (α) and *Type II* (β) errors to characterise the behaviour of mechanisms. When the mechanism analysed is the GM, this strongly links the ROC curve used in our work to a particular sub-type of f-DP, namely Gaussian DP (GDP), which is introduced above. Indeed, it is easily shown that the trade-off function used in GDP is an affine transformation of a specific instance of the ROC curve function:

Corollary 5 (Equation 6 in Dong et al. (2019))**.** The trade-off function G_μ is given by:

$$G_\mu \coloneqq T\left(\mathcal{N}(0,1), \mathcal{N}(\mu,1)\right), \tag{5.1}$$

and

$$G_\mu(\alpha) = \Phi\left(\Phi^{-1}(1-\alpha) - \mu\right). \tag{5.2}$$

Lemma 5. Let $R(x)$ be the ROC curve of $\mathcal{M}$ defined above and let $\Delta = \mu$ and $\sigma = 1$. Then:

$$R(x) = 1 - G_\mu(x). \tag{5.3}$$

Proof. By substituting in Equation 4.7 and using the fact that $\psi = \frac{\Delta}{\sigma} = \frac{\mu}{1} = \mu$:

$$\begin{aligned} R(x) &= \Phi\left(\mu + \Phi^{-1}(x)\right) = 1 - \left(-\Phi^{-1}(x) - \mu\right) = \\ &= 1 - \left(\Phi^{-1}(1-x) - \mu\right) = 1 - G_\mu(x), \end{aligned}$$

where we have used the facts that $\Phi(x) = 1 - \Phi(-x)$ and $-\Phi^{-1}(x) = \Phi^{-1}(1-x)$. This operation corresponds to reflecting the graph of $R(x)$ about the x-axis and then transposing its x-coordinates by $+1$. □

We will also use the following property of G_μ:

Corollary 6. G_μ is strictly monotonically decreasing in μ such that if $\mu_1 \geq \mu_2$, then $G_{\mu_1} \leq G_{\mu_2}$.

Proof. The claim follows from the fact that $R(x)$ is strictly monotonically increasing and from the reflection operation. □

We can now show the following connection between f-DP and the ψ-based DP interpretation:

Lemma 6. Let $\mathcal{M}$ be a GM with sensitivity index ψ on the query function q over adjacent databases D and D'. Then, $\mathcal{M}$ is ψ-GDP if and only if $\psi \leq \mu$.

Proof. The trade-off function between $\mathcal{M}_{q(D)}$ and $\mathcal{M}_{q(D')}$ is:

$$T\left(\mathcal{N}(q(D),\sigma), \mathcal{N}(q(D'),\sigma)\right) = G_{\frac{\|q(D)-q(D')\|_2}{\sigma}}. \tag{5.4}$$

As

$$\frac{\|q(D) - q(D')\|_2}{\sigma} \leq \frac{\Delta}{\sigma} = \psi, \tag{5.5}$$

we obtain:

$$T\left(\mathcal{M}_{q(D)}, \mathcal{M}_{q(D')}\right) \geq G_\psi. \tag{5.6}$$

As G_μ is strictly monotonically decreasing in μ, $G_\psi \geq G_\mu \Leftrightarrow \psi \leq \mu$. However, if $G_\psi \geq G_\mu$, $\mathcal{M}$ is ψ-GDP. Thus, $\psi \leq \mu$ is necessary and sufficient for $\mathcal{M}$ to be ψ-GDP. □

This relationship between f-DP and ψ-based interpretations of the GM is noteworthy, as it endows them with the attractive properties of GDP, while maintaining the advantages of the (in our opinion) more intuitive ROC-curve-based GM characterisation. In particular, the following properties are a direct consequence of Lemma 6:

Corollary 7 (Group Privacy, Theorem 2.14 in Dong et al. (2019))**.** If $\mathcal{M}$ is ψ-GDP, then it is $k\psi$-GDP for groups of size k.

Corollary 8 (Composition, Corollary 3.3 in Dong et al. (2019))**.** Let $\mathcal{M}_i$ be a sequence of ψ_i-GDP GMs, $i \in [1 \dots n]$. Then their n-fold composition is $\sqrt{\psi_1^2 + \dots + \psi_n^2}$-GDP.

Moreover, ψ-based GM interpretations are amenable to the *subsampling amplification* and *central-limit-theorem-type* phenomenon arising in f-DP. This allows the privacy analysis of GMs iterated over many steps on (secret) subsamples of a database, such as DP stochastic gradient descent (DP-SGD) DP-SGD (Abadi et al., 2016; Bu et al., 2020):

Corollary 9 (DP-SGD analysis, Corollary 5.4 in Dong et al. (2019))**.** Let $\mathcal{D}$ be an instance of the DP-SGD algorithm with Gaussian noise magnitude σ, executed on a database of cardinality n for T iterations, where $n \to +\infty$ and $T \to +\infty$. If secret subsamples of the database are drawn uniformly at random with sampling rate r at every iteration, then $\mathcal{D}$ is ψ-GDP with ψ given by:

$$\psi = s\sqrt{2}\sqrt{\exp\left(\frac{1}{\sigma^2}\right)\Phi\left(\frac{3}{2\sigma}\right) + 3\Phi\left(-\frac{1}{2\sigma}\right) - 2}, \tag{5.7}$$

where $r\sqrt{T} \to s$.

Of note, this formulation is asymptotic, yet reasonable for large databases and large numbers of iterations, as are common in deep learning, and allows one to avoid cumbersome composition computations otherwise required for the analysis of subsampled, iterated GMs by f-DP. Recently, Asoodeh et al. (2021) showed that for a specific range of σ values, RDP composition can be tigher than f-DP while Gopi et al. (2021) have indicated that such asymptotic privacy analysis could under-estimate the true privacy cost of composition. Moreover, newer techniques (Zhu et al., 2022) utilising the *characteristic function* for facilitated accounting have been proposed and await investigation beyond the proof-of-principle, as well as integration into DP machine learning libraries. We note that the aforementioned works on composition and privacy accounting are orthogonal to ours in the sense that our findings are completely independent of the choice to technique used to conduct the privacy accounting and ψ can be used to interpret, compare and communicate the privacy parameters of e.g. trained machine learning models using the ROC curve and the ROC-AUC.

Despite the definitions of ψ and μ being nominally identical, we remark one –purely conceptual– difference: μ-GDP intuitively states that telling apart the distributions of $\mathcal{M}(q(D))$ and $\mathcal{M}(q(D'))$ is at least as hard as telling apart $\mathcal{N}(0,1)$ and $\mathcal{N}(\mu,1)$ based on a

single draw. ψ-DP states that that telling apart $\mathcal{M}(q(D))$ and $\mathcal{M}(q(D'))$ is at least as hard as telling apart $\mathcal{N}(\mu_0, \sigma)$ and $\mathcal{N}(\mu_1, \sigma)$ such that $|\mu_1 - \mu_0| = \Delta$. This change is of no formal consequence as the two parameters can be converted into each other through a change of variables procedure, but we choose to make the dependence on σ explicit in ψ-DP for clarity.

5.2. **Relating ψ-DP and RDP.** RDP (Mironov, 2017) was proposed as a natural relaxation of DP, tailored to the specific properties of the GM. We choose to study the relationship between the sensitivity index and RDP over other divergence-based DP relaxation such as concentrated DP (Dwork and Rothblum, 2016) or zero-concentrated DP (Bun and Steinke, 2016), as it enjoys the greatest popularity among machine learning practitioners, representing the basis of privacy accounting of both major DP machine learning libraries (*Opacus* and *TensorFlow Privacy*). Fortunately, RDP can easily be expressed using the sensitivity index ψ, facilitating the conversion between our ψ-based characterisation of the GM and RDP:

Lemma 7. Let $\mathcal{M}$ be a GM with sensitivity index ψ on the query function q over adjacent databases D and D'. Then, $\mathcal{M}$ satisfies (α, ρ)-RDP, $\alpha \geq 1$ if and only if it also satisfies $(\alpha, \frac{\alpha}{2}\psi^2)$-RDP.

We will rely on the following fact about the Rényi divergence from one Gaussian probability distribution to another:

Corollary 10 (See Van Erven and Harremoës (2014).)**.** Let $P_i \coloneqq \mathcal{N}(\mu_i, \sigma_i \mathbf{I})$ and $P_j \coloneqq \mathcal{N}(\mu_j, \sigma_j \mathbf{I})$ be two Gaussian probability distributions. Then, the Rényi divergence of order α from P_i to P_j is given by:

$$
\begin{aligned}
&D_\alpha(P_i \parallel P_j) = \\
&= \log\left(\frac{\sigma_j}{\sigma_i}\right) + \frac{1}{2(\alpha-1)} \log\left(\frac{\sigma_j^2}{\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2}\right) + \\
&+ \frac{1}{2} \frac{\alpha(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T}{\alpha\sigma_j^2 + (1-\alpha)\sigma_i^2}.
\end{aligned}
$$

We can now prove Lemma 7:

Proof. We recall that the density functions of $\mathcal{M}$ over $q(D)$ and $q(D')$ follow Gaussian distributions with means $q(D)$ and $q(D')$, respectively and common covariance matrices $\sigma\mathbf{I}$.

Substituting in Corollary 10 yields:

$$
\begin{aligned}
&D_\alpha\left(\mathcal{N}(q(D),\sigma\mathbf{I}) \parallel \mathcal{N}(q(D'),\sigma\mathbf{I})\right) = \log\left(\frac{\sigma}{\sigma}\right) + \\
&+ \frac{1}{2(\alpha-1)}\log\left(\frac{\sigma^2}{\alpha\sigma^2+(1-\alpha)\sigma^2}\right) + \\
&+ \frac{1}{2}\frac{\alpha(q(D)-q(D'))(q(D)-q(D'))^T}{\alpha\sigma^2+(1-\alpha)\sigma^2} = \\
&= \frac{1}{2}\frac{\alpha(q(D)-q(D'))(q(D)-q(D'))^T}{\sigma^2} = \\
&= \frac{\alpha}{2}\frac{\|q(D)-q(D')\|_2^2}{\sigma^2} \leq \frac{\alpha}{2}\frac{\Delta^2}{\sigma^2} = \frac{\alpha}{2}\psi^2.
\end{aligned}
$$

Thus, for $\mathcal{M}$ to satisfy (α,ρ)-RDP it is sufficient for $\rho = \frac{\alpha}{2}\psi^2$. The inverse condition follows by the same argument. □

We note that this result represents a generalisation of Proposition 7 and Corollary 3 of Mironov (2017). Evidently, the translation between ψ and ρ also allows to make use of the RDP composition theorem, including its subsampled variants, for which we refer to Zhu and Wang (2019); Mironov et al. (2019).

6. Optimal conversion

In this section, we investigate the optimal conversion strategy between ψ-DP and (ε,δ)-DP under a fixed value of δ. This question bears elaboration seeing as (ε,δ)-DP is –in our view– widely regarded as the *canonical* version of DP and stakeholders may be more accustomed to it. Moreover, δ is not typically treated as a free variable but chosen depending on the size of the database in question $\left(\bar{O}\left(\frac{1}{\text{poly}(|D|)}\right)\right)$, where $\bar{O}$ denotes an upper bound and poly$(|D|)$ denotes a polynomial in the size of the dataset (e.g. $|D|^2$), or chosen to be *negligibly small* (e.g. 10^{-128}). We note that the exact choice of δ is of no consequence to the following experiments. Our results admit two different strategies to achieve conversion: (1) Directly converting from ψ-DP using Lemma 4 or (2) converting to (ε,δ)-DP by taking a "detour" via f-DP or RDP. We note that method (1) and the conversion via f-DP are equivalent in the sense that the conversion between ψ-DP and (μ/ψ)-GDP is immediate and relies only on a single parameter. The conversion via RDP is more involved, as it relies on the optimal choice of a second parameter (α). One might –at this point– wonder why a conversion through RDP is even necessary. However, RDP is widely used in deep learning applications with DP, as it has elegant properties under composition. As deep learning with DP requires composing over many applications of the GM, practitioners often require this type of conversion to express the guarantees of models in terms of (ε,δ)-DP, motivating its inclusion in our work. We thus investigate all four techniques, and begin by briefly introducing the conversion rules used in our experiments. The following three conversion techniques are used to convert between RDP and (ε,δ)-DP. We note that it is currently impossible to convert *perfectly* between these interpretations because such a conversion would attempt to create a correspondence between infinitely many (α,ρ)-RDP and infinitely many (ε,δ)-DP pairs. Thus, even for a given δ, one would have to report the conversion for all values of α to not discard any

information. It is thus often said that the RDP to (ε, δ)-DP conversion is *lossy*. The first conversion technique was proposed in Mironov (2017), who introduced RDP:

Corollary 11 (Standard RDP conversion (Mironov, 2017))**.** If a mechanism $\mathcal{M}$ satisfies $(\alpha, \rho) - RDP$, it also satisfies $\left(\rho + \frac{\log\left(\frac{1}{\delta}\right)}{\alpha - 1}, \delta\right)$-DP for $\delta \in (0, 1)$. Equivalently, $\mathcal{M}$ satisfies:

$$\left(\frac{\alpha}{2}\psi^2 + \frac{\log\left(\frac{1}{\delta}\right)}{\alpha - 1}, \delta\right)\text{-DP} \tag{6.1}$$

Follow-up work from Mironov (2017), namely Asoodeh et al. (2021); Balle et al. (2020), has attempted to *tighten* the conversion, that is, introduce improved conversion techniques to reduce the loss of conversion information.

Corollary 12 (Improved RDP conversion (A) (Asoodeh et al., 2021))**.** Let $\alpha > 1$ and $\rho \geq 0$ characterise a mechanism $\mathcal{M}$ satisfying (α, ρ)-RDP and let $\zeta_\alpha \coloneqq \frac{1}{\alpha}\left(1 - \frac{1}{\alpha}\right)^{\alpha - 1}$. Then, $\mathcal{M}$ satisfies:

$$\varepsilon(\rho, \alpha, \delta) = (\rho + \log(1 - \delta)) \tag{6.2}$$

if $\alpha\delta \geq 1$ and

$$\varepsilon(\rho, \alpha, \delta) \leq \frac{1}{\alpha - 1}\min\begin{cases}\left((\alpha - 1)\rho - \log\frac{\delta}{\zeta_\alpha}\right) \\ \log\left(\frac{\exp((\alpha-1)\rho)-1}{\alpha\delta} + 1\right)\end{cases} \tag{6.3}$$

if $0 < \alpha\delta < 1$. Equivalently, $\mathcal{M}$ satisfies:

$$\varepsilon(\psi, \alpha, \delta) = \left(\frac{\alpha}{2}\psi^2 + \log(1 - \delta)\right) \tag{6.4}$$

if $\alpha\delta \geq 1$ and

$$\varepsilon(\psi, \alpha, \delta) \leq \frac{1}{\alpha - 1}\min\begin{cases}\left(\frac{\alpha}{2}\psi^2(\alpha - 1) - \log\frac{\delta}{\zeta_\alpha}\right) \\ \log\left(1 + \frac{1}{\alpha\delta}(\exp(\frac{\alpha}{2}\psi^2(\alpha - 1)) - 1)\right)\end{cases} \tag{6.5}$$

if $\alpha\delta \geq 1$.

Corollary 13 (Improved RDP conversion (B) (Balle et al., 2020))**.** If a mechanism $\mathcal{M}$ satisfies (α, ρ)-RDP, it also satisfies $\left(\rho + \log\left(\frac{\alpha-1}{\alpha}\right) - \frac{\log(\delta)+\log(\alpha)}{\alpha-1}, \delta\right)$-DP for $\delta \in (0, 1)$. Equivalently, $\mathcal{M}$ satisfies:

$$\left(\frac{\alpha}{2}\psi^2 + \log\left(\frac{\alpha - 1}{\alpha}\right) - \frac{\log(\delta) + \log(\alpha)}{\alpha - 1}, \delta\right)\text{-DP} \tag{6.6}$$

We note that Zhu et al. (2022) also discuss the topic of conversion (cf. Appendix F), conjecturing that a truly *optimal* conversion technique exists, but leave a proof to future work.

Lastly, due the relationship between ψ and δ proven in Lemma 4, we can leverage $\mathcal{M}$'s *privacy profile* to convert between ψ-based descriptions of the GM and (ε, δ)-DP. We will use the following fact from Balle and Wang (2018), which was foreshadowed in Lemma 1:

Corollary 14 (Analytic Gaussian Mechanism, (Balle and Wang, 2018))**.** $\mathcal{M}$ preserves $(\varepsilon, \delta(\varepsilon))$-DP if and only if the following holds $\forall\, \varepsilon > 0, \delta \in [0, 1]$:

$$\delta(\varepsilon) \geq \Phi\left(\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) - e^{\varepsilon}\Phi\left(-\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) =$$
$$= \Phi\left(\frac{1}{2}\psi - \frac{1}{\psi}\varepsilon\right) - e^{\varepsilon}\Phi\left(-\frac{1}{2}\psi - \frac{1}{\psi}\varepsilon\right)$$

This conversion method is *lossless*, as ψ-DP is a single-parameter interpretation in the sense that a single value of ψ corresponds to infinitely many (ε, δ) pairs, hence the conversion discards no additional information and, given ψ_g and δ_g, ε can be found by numerically solving:

$$\varepsilon_{\psi_g, \delta_g} = \arg\min_x \{\delta(x, \psi_g) - \delta_g\} \tag{6.7}$$

which we will refer to as the *privacy profile*-based conversion.

We selected the following conditions for our numerical experiments: $\delta = 10^{-5}$ (a value often used in literature, e.g. Abadi et al. (2016); Tramèr and Boneh (2020)), $\psi \in [0.1, 6]$, $\alpha \in (1, 64]$. For solving Equation 6.7, we used a numerical solver employing *Brent's* method (Brent, 2013). Expectedly, using the privacy profile resulted in the best conversion, including when $\psi \to 0$ since the conversion is lossless up to numerical stability. Also the improved RDP conversion methods expectedly resulted in tighter conversions compared to the standard RDP conversion. Interestingly, the two techniques performed almost on par in the higher α regime. Neither RDP-based technique applies when $\delta \to 0$ due to the logarithm and/or a δ term in the denominator, and thus both experience an increase in conversion error near zero. These results corroborate our claim that the single-parameter characterisation provided by our technique facilitates the interpretation of the GM's behaviour: The RDP-based methods obviously exhibit α-dependent behaviour. For values of $\psi > 3$, the privacy profile-based direct conversion was better approximated by RDP-based conversion using low α-values. Nevertheless, the *lossiness* of the conversion was more evident in this region. Conversely, for low values of ψ around 1, the RDP-based conversion performed better at high α-values and also matched the privacy-profile based conversion almost perfectly, albeit only within a narrow interval. Of note, improved conversion (A) introduced in Asoodeh et al. (2021) outperformed improved conversion (B) (Balle et al., 2020) for low values of ψ, in line with the results of Asoodeh et al. (2021). All results are visualised in Figure 6.

Our experiments support that, whereas RDP-based conversions are in principle able to very closely match the privacy profile-based conversion shown in Equation (6.7), especially in the low-ψ-regime, they introduce an –in our view– undesirable additional consideration via the α parameter. We therefore recommend the privacy profile-based strategy for converting between ψ-based DP interpretations (or, equivalently, f-DP) and $(\varepsilon, \delta(\varepsilon))$-DP.

7. Conclusion

We conclude our systematisation by summarising our findings on the properties of the Gaussian Mechanism, viewed through the lens of the sensitivity index ψ:

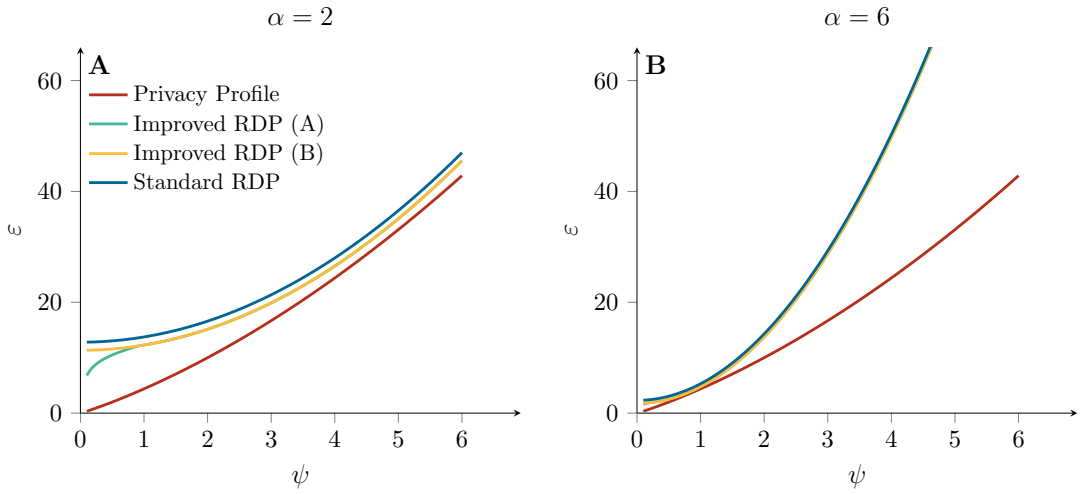

FIGURE 6. Exemplary numerical experiment results showcasing three strategies for converting between ψ and $(\varepsilon, \delta(\varepsilon))$-DP. The privacy profile-based conversion consistently provides the tightest conversion compared to the Standard RDP or either Improved RDP method. Moreover, all two RDP-based conversion techniques are α-dependent. In the high-ψ regime (**A**), RDP conversions utilising lower α-values (here: optimal $\alpha = 2$) more closely approximate the privacy profile-based conversion. In the low-ψ regime (**B**), higher α values (here: $\alpha = 6$) lead to a better approximation, which –in this case –almost perfectly matches the privacy profile-based conversion within a narrow interval. Observe also the increase in error for the Standard RDP and Improved RDP (B) conversion methods when $\psi \to 0$, which is –in part– remedied by the Improved RDP (A) conversion.

Theorem. Let $\mathcal{M}$ be an instance of the Gaussian Mechanism with noise magnitude σ on a query function with global L_2-sensitivity Δ. We then say that $\mathcal{M}$ has sensitivity index $\psi = \frac{\Delta}{\sigma}$, and all of the following statements hold and are are equivalent:

- $\mathcal{M}$ is $(\varepsilon, \delta(\varepsilon))$-DP with:
$$\Phi\left(\frac{1}{2}\psi - \frac{1}{\psi}\varepsilon\right) - e^{\varepsilon}\Phi\left(-\frac{1}{2}\psi - \frac{1}{\psi}\varepsilon\right) \leq \delta(\varepsilon) \tag{7.1}$$
- $\mathcal{M}$ is ψ-GDP
- $\mathcal{M}$ is $\left(\alpha, \frac{\alpha}{2}\psi^2\right)$-RDP
- The ROC curve fully characterising the properties of $\mathcal{M}$ under the hypothesis testing DP interpretation is given by:
$$R(x)_{\mathcal{M}} = \Phi\left(\psi + \Phi^{-1}(x)\right) \tag{7.2}$$
- The area under $R(x)_{\mathcal{M}}$ quantifying the privacy guaranteed by the application of $\mathcal{M}$ is given by:
$$\mathrm{AUC}_{R_{\mathcal{M}}} = \Phi\left(\frac{\psi}{\sqrt{2}}\right) \tag{7.3}$$

The broad utilisation of privacy technologies such as DP can herald an era of large dataset availability, empowered by the trust granted through the provision of objectifiable privacy guarantees. Since the inception of DP, the importance of practitioner and user participation has been posited as central to the design of trustworthy data processing systems relying on its utilisation (Dwork et al., 2019; Cummings et al., 2021). Our work aims to bridge a gap between DP definitions which, to non-DP-experts, can seem disparate, abstract and hard to relate to tangible concepts. The sensitivity index provides a parsimonious, yet comprehensive description of the GM's properties by encapsulating the quintessential properties of the query function (sensitivity) and the mechanism used to privatise its output (noise magnitude) in a "signal-to-noise ratio"-type argument. Moreover, it translates the threat model DP aims to avert into quantities familiar to machine learning practitioners and statisticians (the ROC curve and errors of hypothesis tests). Finally, it allows one to effortlessly compare the guarantees offered by different realisations of the GM through a single calculation and subsequent visual comparison of areas under the ROC curve. In future work, we intend to expand our formalism to encompass the interpretation of mechanism composition, and to propose similar analyses of other DP mechanisms.

References

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*, pages 308–318, 2016. https://doi.org/10.1145/2976749.2978318.

J. M. Abowd. The US Census Bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2867, 2018. https://doi.org/10.1145/3219819.3226070.

Apple. Apple Differential Privacy Technical Overview, 2017. www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf. Accessed: 2021-13-08.

S. Asoodeh, J. Liao, F. P. Calmon, O. Kosut, and L. Sankar. Three Variants of Differential Privacy: Lossless Conversion and Applications. *IEEE Journal on Selected Areas in Information Theory*, 2(1):208–222, 2021. https://doi.org/10.1109/JSAIT.2021.3054692.

B. Balle and Y.-X. Wang. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 394–403. PMLR, 10–15 Jul 2018. https://proceedings.mlr.press/v80/balle18a.html.

B. Balle, G. Barthe, and M. Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 6280–6290, Red Hook, NY, USA, 2018. Curran Associates Inc. https://doi.org/10.5555/3327345.3327525.

B. Balle, G. Barthe, M. Gaboardi, J. Hsu, and T. Sato. Hypothesis testing interpretations and Rényi differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 2496–2506. PMLR, 2020. https://proceedings.mlr.press/v108/balle20a.html.

R. P. Brent. *Algorithms for minimization without derivatives*. Courier Corporation, 2013. https://maths-people.anu.edu.au/~brent/pub/pub011.html.

Z. Bu, J. Dong, Q. Long, and S. Weijie. Deep Learning with Gaussian Differential Privacy. *Harvard Data Science Review*, 9 2020. https://doi.org/10.1162/99608f92.cfc5dd25.

M. Bun and T. Steinke. Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds. *CoRR*, abs/1605.02065, 2016. https://arxiv.org/abs/1605.02065.

R. Cummings, G. Kaptchuk, and E. M. Redmiles. "I need a better description": An Investigation Into User Expectations For Differential Privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. ACM, Nov. 2021. https://doi.org/10.1145/3460120.3485252.

A. Das and W. S. Geisler. A method to integrate and classify normal distributions. *Journal of Vision*, 21(10):1–1, 09 2021. ISSN 1534-7362. https://doi.org/10.1167/jov.21.10.1.

A. P. Dempster and M. Schatzoff. Expected significance level as a sensitivity index for test statistics. *Journal of the American Statistical Association*, 60(310):420–436, 1965. https://doi.org/10.2307/2282680.

J. Dong, A. Roth, and W. J. Su. Gaussian Differential Privacy, 2019. https://arxiv.org/abs/1905.02383.

C. Dwork. Differential Privacy and the US Census. In *Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 1–1, 2019. https://doi.org/10.1145/3294052.3322188.

C. Dwork and A. Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. ISSN 1551-305X. https://doi.org/10.1561/0400000042.

C. Dwork and G. N. Rothblum. Concentrated Differential Privacy. *arXiv preprint arXiv:1603.01887*, 2016. http://arxiv.org/abs/1603.01887.

C. Dwork, N. Kohli, and D. Mulligan. Differential Privacy in Practice: Expose your Epsilons! *Journal of Privacy and Confidentiality*, 9(2), Oct. 2019. https://doi.org/10.29012/jpc.689.

Ú. Erlingsson, V. Pihur, and A. Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014. https://doi.org/10.1145/2660267.2660348.

L. Gonçalves, A. Subtil, M. R. Oliveira, and P. d. Bermudez. ROC curve estimation: An overview. *REVSTAT–Statistical Journal*, 12(1):1–20, 2014. https://doi.org/10.57805/revstat.v12i1.141.

S. Gopi, Y. T. Lee, and L. Wutschitz. Numerical composition of differential privacy. *Advances in Neural Information Processing Systems*, 34, 2021. https://proceedings.neurips.cc/paper/2021.

P. Kairouz, S. Oh, and P. Viswanath. The Composition Theorem for Differential Privacy. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1376–1385, Lille, France, 07–09 Jul 2015. PMLR. http://proceedings.mlr.press/v37/kairouz15.pdf.

P. Laud, A. Pankova, and M. Pettai. A Framework of Metrics for Differential Privacy from Local Sensitivity. *Proc. Priv. Enhancing Technol.*, 2020(2):175–208, 2020. https://doi.org/10.2478/popets-2020-0023.

M. Lyu, D. Su, and N. Li. Understanding the sparse vector technique for differential privacy. *arXiv preprint arXiv:1603.01699*, 2016. https://doi.org/10.14778/3055330.3055331.

P. C. Mahalanobis. On the generalized distance in statistics. In *Proceedings of the National Institute of Science of India*. National Institute of Science of India, 1936. http://library.isical.ac.in:8080/xmlui/bitstream/handle/10263/6765/Vol02_1936_1_Art05-pcm.pdf.

G. J. Matthews and O. Harel. Assessing the privacy of randomized vector-valued queries to a database using the area under the Receiver-Operator Characteristic curve. *Health Services and Outcomes Research Methodology*, 12(2-3):141–155, May 2012. https://doi.org/10.1007/s10742-012-0093-y.

G. J. Matthews, O. Harel, and R. H. Aseltine. Assessing database privacy using the area under the Receiver-Operator Characteristic curve. *Health Services and Outcomes Research Methodology*, 10(1-2):1–15, June 2010. https://doi.org/10.1007/s10742-010-0061-3.

C. E. Metz. Basic principles of ROC analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier, 1978. https://doi.org/10.1016/s0001-2998(78)80014-2.

I. Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017. https://doi.org/10.1109/CSF.2017.11.

I. Mironov, K. Talwar, and L. Zhang. Rényi Differential Privacy of the Sampled Gaussian Mechanism. *arXiv e-prints*, art. arXiv:1908.10530, Aug. 2019. http://arxiv.org/abs/1908.10530.

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933. https://doi.org/10.1098/rsta.1933.0009.

R. Pless, J. Larson, S. Siebers, and B. Westover. Evaluation of local models of dynamic backgrounds. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.* IEEE Comput. Soc, 2003. https://doi.org/10.1109/cvpr.2003.1211454.

J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang. Privacy loss in Apple's implementation of Differential Privacy on macOS 10.12. *arXiv preprint arXiv:1709.02753*, 2017. http://arxiv.org/abs/1709.02753.

F. Tramèr and D. Boneh. Differentially private learning needs better features (or much more data). *arXiv preprint arXiv:2011.11660*, 2020. https://arxiv.org/abs/2011.11660.

T. Van Erven and P. Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. https://doi.org/10.1109/TIT.2014.2320500.

L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010. https://doi.org/10.1198/jasa.2009.tm08651.

Y. Zhu and Y.-X. Wang. Poisson subsampled Rényi differential privacy. In *International Conference on Machine Learning*, pages 7634–7642. PMLR, 2019. https://proceedings.mlr.press/v97/zhu19c.html.

Y. Zhu, J. Dong, and Y.-X. Wang. Optimal accounting of differential privacy via characteristic function. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 4782–4817. PMLR, 28–30 Mar 2022. https://proceedings.mlr.press/v151/zhu22c.html.