Beyond the Calibration Point: Mechanism Comparison in Differential Privacy

Georgios Kaissis [*1] **Stefan Kolek** [*2] **Borja Balle** [3] **Jamie Hayes** [3] **Daniel Rueckert** [1]

Abstract

In differentially private (DP) machine learning, the privacy guarantees of DP mechanisms are often reported and compared on the basis of a single (ε, δ)-pair. This practice overlooks that DP guarantees can vary substantially *even between mechanisms sharing a given* (ε, δ), and potentially introduces privacy vulnerabilities which can remain undetected. This motivates the need for robust, rigorous methods for comparing DP guarantees in such cases. Here, we introduce the Δ-divergence between mechanisms which quantifies the worst-case excess privacy vulnerability of choosing one mechanism over another in terms of (ε, δ), f-DP and in terms of a newly presented Bayesian interpretation. Moreover, as a generalisation of the Blackwell theorem, it is endowed with strong decision-theoretic foundations. Through application examples, we show that our techniques can facilitate informed decision-making and reveal gaps in the current understanding of privacy risks, as current practices in DP-SGD often result in choosing mechanisms with high excess privacy vulnerabilities.

1. Introduction

Protecting private information in machine learning (ML) workflows involving sensitive data is of paramount importance. Differential Privacy (DP) has emerged as the preferred method for providing rigorous and verifiable privacy guarantees, quantifiable by a *privacy budget*. This represents the privacy loss incurred by publicly releasing data that has been processed by a system using DP, e.g. when a deep learning model is trained on sensitive data using DP stochastic gradient descent (DP-SGD, (Abadi et al., 2016)). In principle, workflows utilising DP can offer strong protection against specific attacks, such as membership inference (MIA) and data reconstruction attacks. However, the proper application of DP to defend against such threats relies on a correct understanding of the quantitative aspects of privacy protection, which are expressed differently under the various DP interpretations. For instance, in approximate DP, the privacy budget is quantified using two parameters (ε, δ). Most relevant DP mechanisms, e.g. the subsampled Gaussian mechanism (SGM) typically used in DP-SGD, satisfy DP across a *continuum* of $(\varepsilon, \delta(\varepsilon))$-values rather than a single (ε, δ) tuple. For these mechanisms, δ is a function of ε, represented as the *privacy profile* (Balle et al., 2020a). An equivalent (*dual*) functional view is expressed by the trade-off function in f-DP (Dong et al., 2022).

However, despite the fact that the DP guarantee of such mechanisms can only be characterised by a collection of (ε, δ)-values, it is common practice in literature to calibrate against and report *a single* (ε, δ)-pair to express the privacy guarantee of a DP mechanism (Abadi et al., 2016; Papernot et al., 2021; De et al., 2022). This highlights a potential misconception that such a single pair is sufficient to fully characterise or compare DP guarantees. This assumption is not generally true, as mechanisms can conform to the same (ε, δ)-values but still differ significantly, as seen in Figure 1. In other words: *two DP mechanisms can be calibrated to share an* (ε, δ)*-guarantee while offering substantially different privacy protections*.

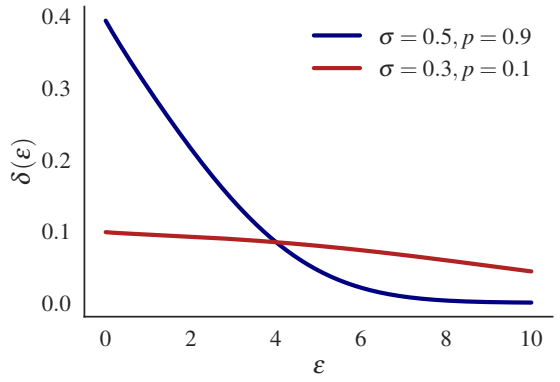


Figure 1: Privacy profiles of two SGMs with different noise scales σ and sampling rates p. Both satisfy $(4.00, 0.08)$-DP, but offer otherwise different levels of privacy protection.

[*]Equal contribution [1]AI in Healthcare and Medicine and Institute of Radiology, Technical University of Munich, Germany [2]Mathematical Foundations of AI, LMU Munich [3]Google DeepMind. Correspondence to: Georgios Kaissis <g.kaissis@tum.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

This leads us to ask whether interpreting and/or comparing the privacy guarantees of DP mechanisms based on their behaviours at a single (ε, δ)-tuple can lead to privacy vulnerabilities. An affirmative answer is suggested by the recent work of Hayes et al. (2023) on reconstruction attacks. Therein, the authors demonstrate that calibrating two SGMs with different parameters to meet the same (ε, δ)-guarantee as shown above results in disparate effectiveness against reconstruction attacks. In practice, this can occur when the user simultaneously increases the sampling rate (e.g. to utilise all available GPU memory) and the noise scale in an attempt to maintain the same (ε, δ)-DP guarantee. In reality, the privacy guarantee has been changed *everywhere except the calibration point* (i.e. the (ε, δ)-tuple in question), weakening the model's protection against data reconstruction attacks. Similar evidence was presented by Lokna et al. (2023), where it was shown that a single (ε, δ)-pair is insufficient to fully characterise a mechanism's protection against MIA. Both examples illustrate that differences between DP guarantees which remain undetected by only considering a single (ε, δ)-pair can lead to privacy hazards.

This reflects an unmet requirement for tools to quantitatively compare the privacy guarantees offered by DP mechanisms in a principled manner. Most existing techniques for comparing DP guarantees either rely on summarisation into a single scalar (which can discard information), on average-case metrics or on assumptions, thus lacking the required generality. The arguably most theoretically rigorous mechanism comparison technique relies on the so-called *Blackwell theorem*, which allows for comparing the privacy guarantees in a strong, decision-theoretic sense. However, the Blackwell theorem is exclusively applicable to the special case in which the privacy guarantees of two mechanisms coincide nowhere, i.e. when their trade-off functions/privacy profiles *never cross*, excluding, among others, DP-SGD, as shown above. To thus extend rigorous mechanism comparisons to this important setting, a set of novel techniques is required, which our work introduces through the following contributions.

Contributions To enable principled comparisons between mechanism whose privacy guarantees coincide at a single point but differ elsewhere, we generalise Blackwell's theorem by introducing an *approximate* ordering between DP mechanisms. This ordering, which we express through the newly presented Δ-*divergence* between mechanisms, quantifies the worst-case increase in privacy vulnerability incurred by choosing one mechanism over another in terms of hypothesis testing errors, $\delta(\varepsilon)$, and in terms of a novel *Bayes error interpretation*. The latter is a probabilistic extension of the hypothesis testing interpretation of DP and allows for principled reasoning over the capabilities of DP adversaries. In addition, we analyse the evolution of approximate comparisons into universal comparisons under composition, yielding insights into the privacy dynamics of algorithms like DP-SGD. Finally, we experimentally show how our techniques can facilitate a more granular privacy analysis of private ML workflows, and pinpoint vulnerabilities which remain undetected by only focusing on a single (ε, δ)-pair.

Related Work Blackwell's theorem (Blackwell, 1953) originates in the theory of comparisons between information structures called *statistical experiments*, and describes conditions under which one statistical experiment is universally more informative than another. Blackwell's framework was later expanded by LeCam (1964); Torgersen (1991), and we refer to the latter for a comprehensive overview of the field. The equivalence between a subclass of statistical experiments (binary experiments) and the decision problem faced by the MIA adversary led Dong et al. (2022) to leverage the Blackwell theorem to provide conditions under which one DP mechanism is *universally* more private than another. This limits mechanism comparisons to the special case when the mechanisms' trade-off functions (or privacy profiles (Balle et al., 2020a)) never cross. However, as demonstrated above, crossing trade-off functions or privacy profiles are *not the exception but the norm*; however, no specific tools to compare privacy guarantees in this case are introduced by Dong et al. (2022).

As discussed above, privacy guarantees have so far often be compared using metrics like attack accuracy or area under the trade-off curve (see Carlini et al. (2022) for a list of works). Besides summarising the privacy guarantee into a single scalar (thus discarding much of the information about the DP mechanism contained in the privacy profile or trade-off function), such metrics model the average case instead of the desirable worst case, rendering them sub-optimal for DP applications. To remedy this, Carlini et al. (2022) proposed comparing attack performance at a "low" Type-I error. However, this method requires an arbitrary assumption about the correct choice of a "low" Type-I error rather than considering the entire potential operating range of an adversary, thereby also discarding information. Moreover, absent a universally agreed upon standard of what a correct choice of Type-I error is, this could incentivise the reporting of research results at a Type-I error which is "cherry-picked" to e.g. emphasise the benefits of a newly introduced MIA, i.e. *p-hacking* (Wasserstein & Lazar, 2016).

Notation and Background Here, we briefly introduce the notation and relevant concepts used throughout the paper for readers with technical familiarity with DP terminology. A detailed background discussion introducing all following concepts can be found in Appendix A. We will denote DP mechanisms by $\mathcal{M} : (P, Q)$, where (P, Q) denote the *tightly dominating pair* of probability distributions which characterise the mechanism as described in Zhu et al. (2022), and will assume that P and Q are mutually abso-

lutely continuous. The Likelihood Ratios (LRs) will be denoted $\bar{X} = {}^{Q(\omega)}/_{P(\omega)}, \omega \sim P$ and $\bar{Y} = {}^{Q(\omega)}/_{P(\omega)}, \omega \sim Q$ for a mechanism outcome ω, where $\sim$ denotes sampling, and the Privacy Loss Random Variables (PLRVs) will be denoted $X = \log(\bar{X})$ and $Y = \log(\bar{Y})$. We will denote the trade-off function (Dong et al., 2022) corresponding to $\mathcal{M}$ by $f : \alpha \mapsto \beta(\alpha)$, where $(\alpha, \beta(\alpha))$ are the Type-I/II errors of the most powerful test between P and Q with null hypothesis $H_0 : \omega \sim P$ and alternative hypothesis $H_1 : \omega \sim Q$, and α is fixed by the adversary. We will assume without loss of generality that f is symmetric (thereby omitting the dominating pair (Q, P)), and defined on $\mathbb{R}$ with $f(x) = 1, x < 0$ and $f(x) = 0, x > 1$.

The privacy profile (Balle et al., 2020a; Gopi et al., 2021) of $\mathcal{M}$ will be denoted by $\delta(\varepsilon)$, while the N-fold self-composition of $\mathcal{M}$ (as is usually practised in DP-SGD (Abadi et al., 2016)) will be denoted by $\mathcal{M}^{\otimes N}$. We will moreover denote the total variation distance between P and Q by $\mathsf{TV}(P, Q) = \min_\alpha(1-\alpha-f(\alpha)) = \mathsf{Adv}$, where Adv is the MIA advantage (Yeom et al., 2018), and the Rényi divergence of order t of P to Q by $\mathsf{D}_t(P \,\|\, Q)$ (Mironov, 2017). The party employing a DP mechanism to protect privacy will be referred to as the *analyst* or *defender*.

2. A Bayesian Interpretation of f-DP

We begin by introducing a novel interpretation of f-DP based on the *minimum Bayes error* of a MIA adversary. While f-DP characterises mechanisms through their trade-off between hypothesis testing errors, our interpretation enriches this characterisation by incorporating the adversary's *prior knowledge* (i.e. auxiliary information). As will become evident below, this allows for incorporating probabilistic reasoning over the adversary and facilitates intuitive operational interpretations of mechanism comparisons, while preserving the same information as f-DP.

Suppose that a Bayesian adversary assigns a prior probability π to the decision "reject H_0". Considering that the adversary's goal is a successful MIA on a specific *challenge example*, H_0 is synonymous with the hypothesis "the mechanism outcome was generated from the database which does not contain the challenge example". Thus, the prior on rejecting H_0 expresses the prior belief that the challenge example is actually part of the database (i.e. a prior probability of positive membership). For example, in privacy auditing (where the analyst assumes the role of the adversary), π corresponds to the probability of including the challenge example (also called "canary") in the database which is attacked (Carlini et al., 2022; Nasr et al., 2023).

From the trade-off function, the Bayes error R at a prior π can be obtained as follows:

$$R(\pi) = \pi\alpha + (1-\pi)f(\alpha), \tag{1}$$

where it is implied that the adversary fixes a level of Type I error α. The *minimum Bayes error function* is derived from the above by minimising over the trade-off between Type I and Type II errors:

$$R_{\min}(\pi) = \min_\alpha \left(\pi\alpha + (1-\pi)f(\alpha)\right). \tag{2}$$

We will refer to $R_{\min}$ as just the *Bayes error function* for short. $R_{\min}$ is continuous, concave, maps $[0,1] \to [0, 1/2]$, satisfies $R_{\min}(0) = R_{\min}(1) = 0$, and $R_{\min}(\pi) \leqslant \min\{\pi, 1-\pi\}$. The *minimax Bayes error* R^* is the maximum of $R_{\min}$ over all values of $\pi \in [0,1]$:

$$R^* = \max_\pi R_{\min}(\pi). \tag{3}$$

R^* is realised at $\pi = 1/2$ since f is assumed symmetric.

$R_{\min}$ is a lossless representation of the mechanism's privacy properties as f can be reconstructed from $R_{\min}$ as follows:

$$f(\alpha) = \max_{0 \leqslant \pi < 1} \left(-\frac{\pi}{1-\pi}\alpha + \frac{R_{\min}(\pi)}{1-\pi} \right). \tag{4}$$

For examples of $R_{\min}$, see Figure 3 and Figure 8 in the Appendix.

3. Blackwell Comparisons

3.1. Universal Blackwell Dominance

As stated above, the Blackwell theorem states equivalent conditions under which a mechanism $\mathcal{M}$ is *universally more informative/less private* than a mechanism $\widetilde{\mathcal{M}}$, denoted $\mathcal{M} \geq \widetilde{\mathcal{M}}$ from now on. For completeness, we briefly re-state these conditions here, and extend them to include our novel Bayes error interpretation.

Theorem 1. The following statements are equivalent:

1. $\forall \alpha \in [0,1] : f(\alpha) \leqslant \widetilde{f}(\alpha)$;
2. $\forall \varepsilon \in \mathbb{R} : \delta(\varepsilon) \geqslant \widetilde{\delta}(\varepsilon)$;
3. $\forall \pi \in [0,1] : R_{\min}(\pi) \leqslant \widetilde{R}_{\min}(\pi)$.

The proofs of clause (1) and (2) can be found in Sections 2.3 and 2.4 of Dong et al. (2022), while the proof of (3) and all following theoretical results can be found in Appendix B.4.

If any of the above conditions hold, we write $\mathcal{M} \geq \widetilde{\mathcal{M}}$ and say that $\mathcal{M}$ *Blackwell dominates* $\widetilde{\mathcal{M}}$. Note the lack of a clause related to Rényi DP (RDP), which is a consequence of the fact that, while $\mathcal{M} \geq \widetilde{\mathcal{M}}$ implies that $\mathsf{D}_t(P\|Q) \geqslant \mathsf{D}_t(\widetilde{P}\|\widetilde{Q})$, for all $t \geqslant 1$, the reverse does not hold in general (Dong et al., 2022). RDP is thus a *generally weaker basis of comparison* between DP mechanisms.

The relation $\geq$ induces a partial order on the space of DP mechanisms and expresses a strong condition, as it implies

that the dominating mechanism is more useful for *any* downstream task, benign (e.g. training an ML model) or malicious (e.g. privacy attacks) (Dong et al., 2022; Torgersen, 1991). Note that Theorem 1 is inapplicable when the trade-off functions, privacy profiles or Bayes error functions cross. Addressing this issue is the topic of the rest of the paper.

3.2. Approximate Blackwell Dominance

As discussed above, Blackwell dominance expresses that choosing the dominated mechanism is, in a universal sense, a better choice in terms of privacy protection. In other words, an analyst choosing the dominated mechanism would *never regret* this choice from a privacy perspective. However, more frequently, the choice between mechanisms is equivocal because their privacy guarantees coincide at the calibration point, but differ elsewhere. They thus offer disparate protection against different adversaries, meaning that no choice fully eliminates potential regret in terms of privacy vulnerability. A natural decision strategy under the principle of DP to protect against the worst case is to choose the mechanism which minimises the *worst-case regret* in terms of privacy vulnerability. To formalise this strategy, we next introduce a relaxation of the Blackwell theorem. Similar to how approximate DP relaxes pure DP, we term comparisons using this relaxation *approximate Blackwell comparisons*.[1]

To motivate this formalisation within the DP threat model, suppose that an analyst must choose between $\mathcal{M}$ and $\widetilde{\mathcal{M}}$, however they cannot unequivocally decide between them because *neither mechanism is universally more or less vulnerable to MIA*. To express "how close" the analyst is to being able to choose unequivocally between the mechanisms (i.e. to Blackwell dominance being restored), we determine the smallest shift $\kappa \geqslant 0$ which suffices to move f below and to the left of $\widetilde{f}$ such that Theorem 1 kicks in and $\mathcal{M} \succeq \widetilde{\mathcal{M}}$, as shown in Figure 2.

Definition 1. The Δ-*divergence* of $\mathcal{M}$ to $\widetilde{\mathcal{M}}$ is given by

$$\Delta(\mathcal{M} \| \widetilde{\mathcal{M}}) = \inf\{\kappa \geqslant 0 \mid \forall \alpha : \; f(\alpha + \kappa) - \kappa \leqslant \widetilde{f}(\alpha)\}.$$

This allows us to define approximate Blackwell dominance:

Definition 2. If $\Delta(\mathcal{M} \parallel \widetilde{\mathcal{M}}) \leqslant \mathfrak{D}$, we say that $\mathcal{M}$ $\mathfrak{D}$-approximately dominates $\widetilde{\mathcal{M}}$, denoted $\mathcal{M} \succeq_{\mathfrak{D}} \widetilde{\mathcal{M}}$.

The next theorem formally states equivalent criteria for approximate Blackwell dominance:

Theorem 2. The following are equivalent to $\mathcal{M} \succeq_{\mathfrak{D}} \widetilde{\mathcal{M}}$:

1. $\forall \alpha \in [0,1] : f(\alpha + \mathfrak{D}) - \mathfrak{D} \leqslant \widetilde{f}(\alpha)$;
2. $\forall \varepsilon \in \mathbb{R} : \delta(\varepsilon) + \mathfrak{D} \cdot (1 + \mathrm{e}^{\varepsilon}) \geqslant \widetilde{\delta}(\varepsilon)$;
3. $\forall \pi \in [0,1] : R_{\min}(\pi) - \widetilde{R}_{\min}(\pi) \leqslant \mathfrak{D}$.

[1] A related term in the experimental comparisons literature is "deficiencies" (LeCam, 1964).

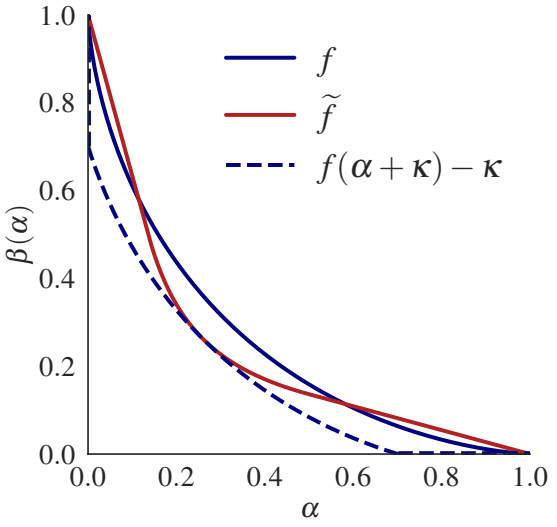


Figure 2: The trade-off functions of a Gaussian (f, blue) and a Laplace ($\widetilde{f}$, red) mechanism cross, therefore neither mechanism universally Blackwell dominates. Shifting f left and down by κ yields $f(\alpha) \leqslant \widetilde{f}(\alpha)$ for all $\alpha \in [0,1]$ (dashed blue), and restores universal Blackwell dominance.

The proof relies on fundamental properties of trade-off functions, of the convex conjugate and its order-reversing property and on the lossless conversion between trade-off function and Bayes error function.

Intuitively, when $\mathfrak{D}$ is very small, the clauses of Theorem 2 are "approximate" versions of the corresponding clauses of Theorem 1. In particular, $\mathfrak{D}$ represents an upper bound on the excess vulnerability of $\widetilde{\mathcal{M}}$ at any level α, choice of ε or prior π. The computation of $\Delta(\mathcal{M} \| \widetilde{\mathcal{M}})$ is most naturally expressed through the Bayes error functions:

Corollary 1. $\Delta(\mathcal{M} \| \widetilde{\mathcal{M}}) = \max_{\pi}(R_{\min}(\pi) - \widetilde{R}_{\min}(\pi))$.

The Δ-divergence can be computed numerically through grid discretisation with N points (i.e. to tolerance $1/N$) in $\mathcal{O}(N)$ time, and requires only oracle access to a function implementing the trade-off functions of the mechanisms. An example is provided in Appendix B.3.

Moreover, Corollary 1 admits the following interpretation:

> *$\Delta(\mathcal{M} \parallel \widetilde{\mathcal{M}})$ expresses the worst-case regret of an analyst choosing to employ $\widetilde{\mathcal{M}}$ instead of $\mathcal{M}$, whereby regret is expressed in terms of the adversary's decrease in minimum Bayes error.*

We consider this connection between Bayesian decision theory and DP the most natural interpretation of our results.

3.3. Metrising the Space of DP Mechanisms

After introducing tools for establishing a ranking between DP mechanisms in the preceding sections, we here show that the Δ-divergence can actually be used to define a *metric* on the space of DP mechanisms. In the sequel, we will say that two mechanisms are *equal* and write $\mathcal{M} = \widetilde{\mathcal{M}}$ if and only if their trade-off functions, privacy profiles and Bayes error functions are equal. For a formal discussion on this choice of terminology, see Remark 1 in the Appendix. Moreover, we define the following extension of the Δ-divergence:

Definition 3 (Symmetrised Δ-divergence $\Delta^{\leftrightarrow}$)**.**

$$\Delta^{\leftrightarrow}(\mathcal{M}\|\widetilde{\mathcal{M}}) = \max\left\{\Delta(\mathcal{M}\|\widetilde{\mathcal{M}}), \Delta(\widetilde{\mathcal{M}}\|\mathcal{M})\right\}. \quad (5)$$

Using Corollary 1, $\Delta^{\leftrightarrow}$ can be written as:

$$\Delta^{\leftrightarrow}(\mathcal{M}\|\widetilde{\mathcal{M}}) = \|R_{\min}(\pi) - \widetilde{R}_{\min}(\pi)\|_{\infty}. \quad (6)$$

In terms of the trade-off functions, the following holds:

Lemma 1. Let $\Delta^{\leftrightarrow} = \Delta^{\leftrightarrow}(\mathcal{M}\|\widetilde{\mathcal{M}})$. Then it holds that:

$$f(\alpha + \Delta^{\leftrightarrow}) - \Delta^{\leftrightarrow} \leqslant \widetilde{f}(\alpha) \leqslant f(\alpha - \Delta^{\leftrightarrow}) + \Delta^{\leftrightarrow}. \quad (7)$$

This substantiates that $\Delta^{\leftrightarrow}(\mathcal{M}\|\widetilde{\mathcal{M}}) = 0$ is equivalent to the equality of the trade-off functions, and thus of the privacy profiles and Bayes error functions.

The similarity of Equation (7) to the Lévy distance is not coincidental, and it is shown in Appendix B.2 that, by considering the trade-off function as a CDF (via $f(1-\alpha)$), $\Delta^{\leftrightarrow}$ exactly plays the role of the Lévy distance. Similarly to how the Lévy distance metrises the weak convergence of random variables, $\Delta^{\leftrightarrow}$ metrises the space of DP mechanisms:

Corollary 2. $\Delta^{\leftrightarrow}$ is a metric.

Note that this implies that $\Delta^{\leftrightarrow}(\mathcal{M} \,\|\, \widetilde{\mathcal{M}}) > 0$ unless the mechanisms have identical privacy profiles, trade-off functions or Bayes error functions, underscoring that sharing a single (ε, δ)-guarantee is an insufficient condition for stating that mechanisms provide equal protection.

3.4. Comparisons with Extremal Mechanisms

Next, we use the Δ-divergence to interpret comparisons with two "extremal" reference mechanisms: the *blatantly non-private* (totally informative) mechanism $\mathcal{M}_{\mathrm{BNP}}$ and the *perfectly private* (totally non-informative) mechanism $\mathcal{M}_{\mathrm{PP}}$. These two mechanisms represent the "extremes" of the privacy/information spectrum.

For this purpose, we define for $\mathcal{M}_{\mathrm{BNP}}$: $f_{\mathrm{BNP}}(\alpha) = 0$, $R^{\mathrm{BNP}}_{\min}(\pi) = 0$, and $\delta_{\mathrm{BNP}}(\varepsilon) = 1$. Moreover, we define for $\mathcal{M}_{\mathrm{PP}}$: $f_{\mathrm{PP}}(\alpha) = 1 - \alpha$, $R^{\mathrm{PP}}_{\min}(\pi) = \min\{\pi, 1-\pi\}$, and $\delta_{\mathrm{PP}}(\varepsilon) = 0$. The next lemma establishes the "extremeness":

Lemma 2. $\mathcal{M} \geq \mathcal{M}_{\mathrm{PP}}$ and $\mathcal{M}_{\mathrm{BNP}} \geq \mathcal{M}$ for any $\mathcal{M}$.

We can thus compute a "divergence from perfect privacy" $\Delta(\mathcal{M}_{\mathrm{PP}} \,\|\, \mathcal{M})$, and a "divergence to blatant non-privacy" $\Delta(\mathcal{M}\|\mathcal{M}_{\mathrm{BNP}})$. Both have familiar operational interpretations in terms of quantities from the field of DP:

Lemma 3. It holds that $\Delta(\mathcal{M}_{\mathrm{PP}} \,\|\, \mathcal{M}) = {}^1\!/_2\mathsf{TV}(P, Q) = {}^1\!/_2\mathsf{Adv} = {}^1\!/_2\delta(0)$.

This conforms to the intuition that, the "further" the mechanism is from perfect privacy, the higher the adversary's MIA advantage can be.

Lemma 4. It holds that $\Delta(\mathcal{M}\|\mathcal{M}_{\mathrm{BNP}}) = R^* = \alpha^*$, where R^* is the minimax Bayes error and α^* the fixed point of the trade-off function of $\mathcal{M}$.

Recall that R^* is the error rate of an "uninformed" adversary ($\pi = 0.5$, compare Figure 8a), whereas α^* is the point on the trade-off curve closest to the origin, i.e. to $(\alpha, f(\alpha)) = (0, 0)$ (see Figure 8b). When either point coincides with the origin, the mechanism is blatantly non-private. Moreover, the following holds for any mechanism:

Lemma 5. $\Delta(\mathcal{M}_{\mathrm{PP}}\|\mathcal{M}) + \Delta(\mathcal{M}\|\mathcal{M}_{\mathrm{BNP}}) = 0.5$.

The results of Section 3.3 and Section 3.4 lead to the following conclusions: On one hand, the metric $\Delta^{\leftrightarrow}$ can be used to measure a notion of "informational distance" even between *completely different* mechanisms (e.g. Randomised Response and DP-SGD). Additionally, the space of DP mechanisms is a *bounded partially ordered set* with a maximal ($\mathcal{M}_{\mathrm{BNP}}$) and a minimal ($\mathcal{M}_{\mathrm{PP}}$) bound, and *any* DP mechanism can be placed on the information spectrum between them. While not discussed in detail here, we note that this set is also a *lattice* (Blackwell, 1953, Theorem 10).

4. Emergent Blackwell Dominance

We next study the interplay of mechanism comparisons and composition. The fact that $\mathcal{M} \geq \widetilde{\mathcal{M}}$ implies $\mathcal{M}^{\otimes N} \geq \widetilde{\mathcal{M}}^{\otimes N}$ is known (Torgersen, 1991). So far however, the questions of (1) whether mechanisms which are initially *not* Blackwell ranked will eventually *become* Blackwell ranked and (2) which of their properties determine the resulting ranking have not been directly investigated.

The next result follows from the fact that –under specific preconditions– composition qualitatively transforms mechanisms towards Gaussians mechanisms (GMs) due to a central limit theorem (CLT)-like phenomenon (Dong et al., 2022; Sommer et al., 2018). Since GMs are always Blackwell ranked (see Lemma B.2 in the Appendix), we expect Blackwell dominance to emerge once mechanisms are suffi-

ciently well-approximated by GMs. We first define:

$$\eta = \frac{v_1}{\sqrt{v_2 - v_1^2}}, \tag{8}$$

which plays an important role in the analysis below. Moreover, in the sequel, v_1, v_2, v_3 and v_4 will denote the following functionals of f:

$$v_1 = -\int_0^1 \log\left|\frac{\mathrm{d}}{\mathrm{d}x} f(x)\right| \mathrm{d}x, \tag{9}$$

$$v_2 = \int_0^1 \log^2\left|\frac{\mathrm{d}}{\mathrm{d}x} f(x)\right| \mathrm{d}x, \tag{10}$$

$$v_3 = \int_0^1 \left|\log\left|\frac{\mathrm{d}}{\mathrm{d}x} f(x)\right| + v_1\right|^3 \mathrm{d}x, \text{ and} \tag{11}$$

$$v_4 = \int_0^1 \left|\log\left|\frac{\mathrm{d}}{\mathrm{d}x} f(x)\right|\right|^3 \mathrm{d}x. \tag{12}$$

Intuitively, these represent moments of the PLRV.

Lemma 6. Let $\{\mathcal{M}_{Ni} : 1 \leqslant i \leqslant N\}_{N=1}^{\infty}$ be a triangular array of mechanisms satisfying the following conditions:

1. $\lim_{N\to\infty} \sum_{i=1}^{N} v_1(f_{Ni}) = K$;
2. $\lim_{N\to\infty} \max_{1\leqslant i\leqslant N} v_1(f_{Ni}) = 0$;
3. $\lim_{N\to\infty} \sum_{i=1}^{N} v_2(f_{Ni}) = s^2$;
4. $\lim_{N\to\infty} \sum_{i=1}^{N} v_4(f_{Ni}) = 0$.

Analogously, define $\{\widetilde{\mathcal{M}}_{Ni} : 1 \leqslant i \leqslant N\}_{N=1}^{\infty}$ for constants $\widetilde{K}, \tilde{s}$. Then, if $K/s > \widetilde{K}/\tilde{s}$, there exists N^* such that, for all $N \geqslant N^*$:

$$\mathcal{M}_{N1} \otimes \cdots \otimes \mathcal{M}_{NN} \succeq \widetilde{\mathcal{M}}_{N1} \otimes \cdots \otimes \widetilde{\mathcal{M}}_{NN}, \tag{13}$$

where $\mathcal{M}_{N1} \otimes \cdots \otimes \mathcal{M}_{NN}$ denotes N-fold mechanism composition and analogously for $\widetilde{\mathcal{M}}_{Ni}$.

Our proof strategy relies on first showing that, under the stated preconditions, mechanisms asymptotically converge to Gaussian mechanisms under composition and combining this fact with the property that Gaussian mechanisms are always either equal, or one Blackwell dominates the other.

The conditions above are also used in Dong et al. (2022) to prove the CLT-like convergence of the trade-off functions of composed mechanisms to that of a GM, which we adapt here to show conditions for the emergence of Blackwell dominance between compositions of mechanisms in the limit. Concretely, $\{\mathcal{M}_{Ni}\}_{i=1}^{N}$ is a collection of mechanisms calibrated to provide a certain level of privacy after composition, and the mechanisms in the sequence change (become progressively more private) as N grows to ∞ to maintain that level of privacy as more mechanisms are composed. However, from the more practical standpoint of comparing instances of DP-SGD with different parameters, we are rather interested in the question of approximate Blackwell dominance after a *finite* number of self-compositions of *fixed* parameter mechanisms. This is shown next.

Theorem 3. Let $\mathcal{M}, \widetilde{\mathcal{M}}$ be two mechanisms with $v_4, \tilde{v}_4 < \infty$ and denote by $\mathcal{M}^{\otimes N}, \widetilde{\mathcal{M}}^{\otimes \widetilde{N}}$ their N- and $\widetilde{N}$-fold self-compositions. Then, $N/\widetilde{N} \geqslant \tilde{\eta}^2/\eta^2$ implies:

$$\Delta(\mathcal{M}^{\otimes N} \| \widetilde{\mathcal{M}}^{\otimes \widetilde{N}}) \leqslant 0.56 \left(\frac{\eta^3 v_3}{\sqrt{N} v_1^3} + \frac{\tilde{\eta}^3 \tilde{v}_3}{\sqrt{\widetilde{N}} \tilde{v}_1^3}\right) \tag{14}$$

In particular, if $N = \widetilde{N}$, $\eta \geqslant \tilde{\eta}$ implies:

$$\Delta(\mathcal{M}^{\otimes N} \| \widetilde{\mathcal{M}}^{\otimes \widetilde{N}}) \leqslant \frac{0.56}{\sqrt{N}} \left(\frac{\eta^3 v_3}{v_1^3} + \frac{\tilde{\eta}^3 \tilde{v}_3}{\tilde{v}_1^3}\right). \tag{15}$$

The proof relies on the aforementioned Blackwell dominance properties between Gaussian mechanisms combined with the triangle inequality property of the Δ-divergence and a judicious application of the Berry-Esséen-Theorem.

Theorem 3 intuitively states that the Δ-divergence will approach zero not asymptotically as in Lemma 6, but within a specific number of update steps and allows for choosing $N, \widetilde{N}$ differently. Seeing as the number of update steps is a crucial hyper-parameter in DP-SGD (De et al., 2022), this is required for practical usefulness. In addition, it pinpoints the exact relationship between the mechanisms ($\tilde{\eta}^2/\eta^2$) that determines which mechanism will eventually dominate. In particular, if $N/\widetilde{N} \geqslant \tilde{\eta}^2/\eta^2$, then $\Delta(\mathcal{M}^{\otimes N} \| \widetilde{\mathcal{M}}^{\otimes \widetilde{N}})$ will vanish at least as fast as $\min\{N, \widetilde{N}\}^{-1/2}$, and if $N = \widetilde{N}$, then the emergence of Blackwell dominance depends only on the parameters $\eta, \tilde{\eta}$, i.e. on the PLRV moments. Moreover, this result does *not* require scaling the mechanism parameters at every step to prevent them from becoming blatantly non-private, even at very large numbers of compositions.

5. Experiments

Approximate Comparisons in Practice Figure 3 demonstrates a "canonical" example of an approximate comparison between the GM ($\sigma = 1$) and the Laplace mechanism ($b = 1$) on a function with unit global sensitivity. Observe that the Bayes error functions cross at $\pi \approx 0.4$, and that $\Delta(\mathrm{Gauss} \,\|\, \mathrm{Lap}) = 0.005 < \Delta(\mathrm{Lap} \,\|\, \mathrm{Gauss}) = 0.034$, as seen by the length of the black "rulers" in the figure. Therefore, the worst-case regret in terms of privacy vulnerability of choosing the Laplace mechanism is smaller than for the GM. Moreover, the Gaussian mechanism offers only marginally stronger protection for a narrow range of π around $1/2$ corresponding to the prior of an "uninformed" adversary. This allows for much more granular insights

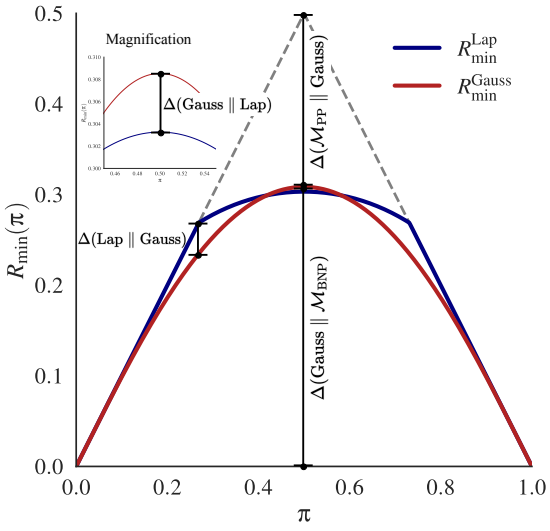


Figure 3: Approximate comparison between the Laplace ($b = 1$) and the Gaussian ($\sigma = 1$) mechanism via their Bayes error representations. Note that all divergence values of interest are visually depicted in this representation.

into the mechanisms' privacy properties beyond the folklore statement that *pure DP mechanisms (Laplace) offer "stronger privacy" than approximate DP mechanisms (GM)*.

Tightness of the Bound in Theorem 3 To evaluate the bound, we compare two SGMs $\mathcal{M}, \widetilde{\mathcal{M}}$ with $\sigma = 2, \tilde{\sigma} = 3$, $p = \tilde{p} = 9 \cdot 10^{-4}$, $N = 1.4 \cdot 10^6$ and $\widetilde{N} = 3.4 \cdot 10^6$. The predicted bound is $\Delta(\mathcal{M}^{\otimes N} \,\|\, \widetilde{\mathcal{M}}^{\otimes \widetilde{N}}) < 10^{-3}$, while the empirically computed bound is $8 \cdot 10^{-4}$. The parameter choices in this example mirror those used in De et al. (2022) for fine-tuning on the JFT-300M dataset to $(8, 5 \cdot 10^{-7})$-DP, underscoring the applicability of our bound to large-scale ML workflows.

Bayesian Mechanism Selection An additional benefit of our Bayes error interpretation is that it facilitates principled reasoning about the adversary's auxiliary information. Recall that π expresses the adversary's "informedness", i.e. the strength of their prior belief about the challenge example's membership. This allows for introducing hierarchical Bayesian modelling techniques to mechanism comparisons by introducing *hyper-priors*, i.e. probability distributions over the adversary's values of π. For example, if the defender is very uncertain about the anticipated adversary's prior, they can use an "uninformative" hyper-prior such as the Jeffreys prior (Jeffreys, 1946) (here: $\mathrm{Beta}(0.5, 0.5)$). Alternatively, a more "informed adversary" with stronger prior beliefs (i.e. low or high values or π) could be modelled by e.g. the $\mathrm{UQuadratic}[0, 1]$ distribution. Then, denoting by $\Psi(\pi)$ the hyper-prior, one can obtain the *weighted minimum Bayes error* $R^{\Psi}_{\min}(\pi) = R_{\min}(\pi)\Psi(\pi)$.

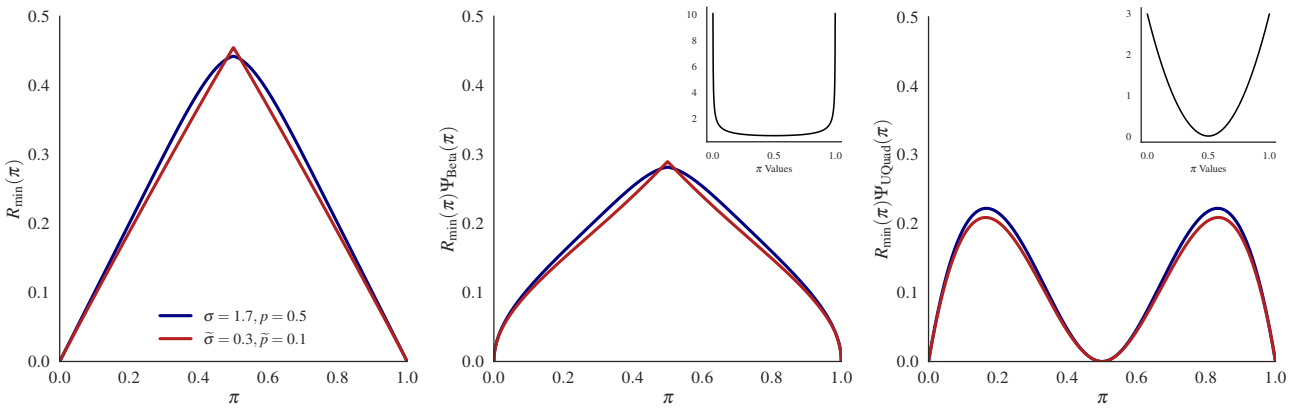
Similarly, a weighted Δ-divergence $\Delta^{\Psi}(\mathcal{M} \parallel \widetilde{\mathcal{M}}) = \max_{\pi}(R^{\Psi}_{\min}(\pi) - \widetilde{R}^{\Psi}_{\min}(\pi))$ can be computed, which expresses the excess regret of choosing $\widetilde{\mathcal{M}}$ over $\mathcal{M}$ *modulated by the defender's beliefs* about the adversary's prior. Incorporating such adversarial priors has recently witnessed growing interest (Balle et al., 2022; Jayaraman et al., 2021).

Our method is a principled probabilistic extension of the recommendation by Carlini et al. (2022) to choose the mechanism whose trade-off function offers higher Type-II errors at "low α". This recommendation requires a (more or less arbitrary) choice of a "low" α; as discussed above, no standardised recommendation on this choice exists, leading to poor comparability of results, and potentially skewed reporting. Moreover, the technique does not take all possible adversaries into account.

These shortcomings are addressed by our proposed technique, as shown in Figure 4, which compares two SGMs: $\mathcal{M}$ (blue) and $\widetilde{\mathcal{M}}$ (red). Without any hyper-prior (Figure 4a), $\Delta(\mathcal{M} \| \widetilde{\mathcal{M}}) = 0.01 < 0.02 = \Delta(\widetilde{\mathcal{M}} \| \mathcal{M})$, indicating that choosing $\mathcal{M}$ is slightly riskier in the worst-case. Applying the Jeffreys hyper-prior (Figure 4b), which expresses a minimal set of assumptions about the adversary, yields $\Delta^{\mathrm{Beta}}(\mathcal{M} \parallel \widetilde{\mathcal{M}}) = 0.007 < 0.015 = \Delta^{\mathrm{Beta}}(\widetilde{\mathcal{M}} \parallel \mathcal{M})$, expectedly not changing the ranking. However, when the more pessimistic $\mathrm{UQuadratic}[0, 1]$ hyper-prior is applied (Figure 4c), which models an adversary with strong prior beliefs, we obtain $\Delta^{\mathrm{UQuad}}(\mathcal{M} \parallel \widetilde{\mathcal{M}}) = 0.014 > 0 = \Delta^{\mathrm{UQuad}}(\widetilde{\mathcal{M}} \parallel \mathcal{M})$, indicating that –against an informed adversary– one would consistently prefer $\mathcal{M}$.

Pareto-Efficient Choice of Noise Multipliers Deep learning with DP-SGD presents a trilemma between model accuracy, privacy protection and resource efficiency. The privacy-accuracy trade-off is well-known in the community, whereas the efficiency trade-off is more apparent when training deep learning models on large-scale datasets. In the recent work of De et al. (2022, Section 5), the authors posit that there exists an "optimal" combination of noise multiplier and number of update steps to achieve the best possible accuracy. Concretely, the authors calibrate seven CIFAR-10 training runs with different noise multipliers and numbers of steps while fixing the sampling rate to obtain models which all satisfy $(8, 10^{-5})$-DP. Subsequently, they determine that the "optimal" noise multiplier for their application is $\tilde{\sigma} = 3.0$, whereas both higher and lower noise multipliers deteriorate the training and validation accuracy.

Here, we re-assess the authors' results using the novel techniques introduced in this paper. For readability, we will from now equate mechanisms with their noise multipliers, writing e.g. $\Delta(\sigma \| \tilde{\sigma} = 3.0)$ to denote the Δ-divergence from the baseline mechanism $\mathcal{M}$ with noise multiplier $\sigma = 2.0$ and a validation accuracy of 72.6%, to $\widetilde{\mathcal{M}}$ with $\tilde{\sigma} = 3.0$. The



(a) No hyper-prior (b) $\text{Beta}(0.5, 0.5)$ hyper-prior (c) $\text{UQuadratic}[0, 1]$ hyper-prior

Figure 4: Hierarchical Bayesian modelling of adversarial strategies. Hyper-prior densities are plotted in the insets.

number of steps and sub-sampling rate are chosen exactly as in De et al. (2022, Figure 6). Figure 5 summarises our observations. First, note that, even though the mechanisms are all nominally calibrated to $(8, 10^{-5})$-DP, they are not equal in the sense of Lemma 1. This is not unexpected, as it the same phenomenon observed in Figure 1, and it once again underscores the pitfalls of relying on a single (ε, δ)-pair to calibrate the SGM. More interestingly, the Δ-divergence from the baseline increases monotonically with increasing noise multipliers. This introduces an additional "dimension" to the result of De et al. (2022): Choosing $\widetilde{\mathcal{M}}$ to have $\tilde{\sigma} = 3.0$ is not actually an "optimal" choice but –at best– a *Pareto efficient* choice in terms of balancing accuracy and excess vulnerability over $\mathcal{M}$. In particular, choosing $\tilde{\sigma}$ to be larger or smaller than $\tilde{\sigma} = 3.0$ cannot simultaneously increase accuracy and decrease excess vulnerability over $\mathcal{M}$. Thus, in this case, all mechanisms with $\tilde{\sigma} > 3$ are *Pareto inefficient* choices, since one could simultaneously increase accuracy and decrease excess vulnerability over $\mathcal{M}$ by choosing $\tilde{\sigma} = 3.0$.

Effect of DP-SGD Parameters on the Δ-Divergence To further examine the effect of mechanism parameter choices on the Δ-divergence, Figure 6 investigates switching from a base SGM $\mathcal{M}$ with $p = 0.01, N = 500$ and $\sigma = 0.54$ to $\widetilde{\mathcal{M}}$, where $\tilde{p} \in [0.04, 0.9], \tilde{N} \in [534, 1500]$ and the resulting $\tilde{\sigma} \in [0.55, 21]$. All mechanisms are calibrated to $(8, 10^{-5})$-DP using the numerical system by Doroshenko et al. (2022) and the absolute calibration error in terms of ε is $\leqslant 0.00042$. A monotonic increase in the Δ-divergence with the noise multiplier is observed, culminating in a maximum divergence value of around 0.12. In particular, increases in $\tilde{p}$ and $\tilde{N}$ are associated with an increase in the Δ-divergence. Moreover, the Δ-divergence exhibits greater sensitivity to variations in $\tilde{p}$ compared to changes in $\tilde{N}$.

Figure 6 suggests that the maximal excess vulnerabilities are

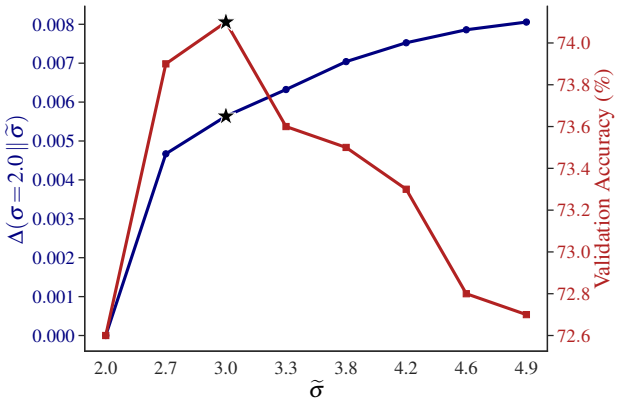


Figure 5: Validation accuracy (red) and Δ-divergence values (blue) for mechanisms satisfying $(8, 10^{-5})$-DP. The combinations to the right of $\star$ are not Pareto efficient in terms of balancing accuracy and excess vulnerability over $\mathcal{M}$.

realised by large $\tilde{p}$ and $\tilde{N}$. This once again highlights not just that these vulnerabilities remain completely undetected when only reporting that the mechanisms "satisfy $(8, 10^{-5})$-DP", but also that the current best practices in selecting SGM parameters for training large-scale ML models with DP, i.e. large sampling rates and many steps (De et al., 2022; Berrada et al., 2023) unfortunately correspond to the most vulnerable regime.

From Δ-Divergences to Attack Vulnerability To provide a practical understanding of what an excess vulnerability of 0.12 (i.e. the maximum attained in Figure 6) means in practice, we revisit the example by Hayes et al. (2023) discussed in the introduction. Recall that the authors empirically demonstrated that calibrating different SGMs to a constant (ε, δ)-guarantee while changing the underlying noise multiplier and sampling rate leads to mechanism with

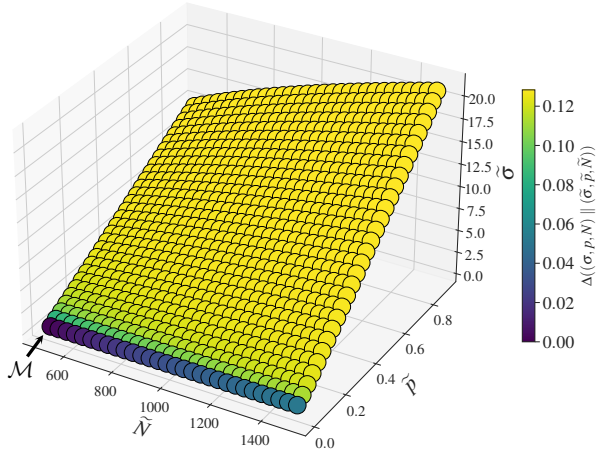


Figure 6: Association between the choice of SGM parameters $(\widetilde{\sigma}, \widetilde{p}, \widetilde{N})$ and the Δ-divergence compared to a baseline mechanism $\mathcal{M}$ (marked with an arrow).

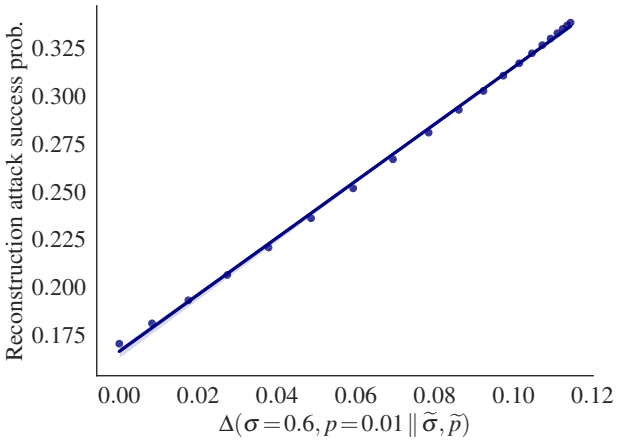


Figure 7: Δ-divergence values from a baseline mechanism compared to upper bounds on the success rate of data reconstruction attacks against DP-SGD.

disparate vulnerability against data reconstruction attacks.

Using our newly introduced techniques, we can now formally substantiate this finding, shown in Figure 7. The horizontal axis shows the Δ-divergence value from $\mathcal{M}$ with $\sigma = 0.6, p = 0.01$) to a series of mechanisms $\widetilde{\mathcal{M}}$ with increasing values of $\widetilde{p}$ and $\widetilde{\sigma}$, where all $\widetilde{\mathcal{M}}$ are calibrated to $(4, 10^{-5})$-DP as previously described. The vertical axis shows the theoretical upper bound on a successful data reconstruction attack against the model (called *Reconstruction Robustness* by Hayes et al. (2023)). We note that these theoretical upper bounds are matched almost exactly by actual attacks, so the bounds are almost tight in practice. These mechanism settings and resulting reconstruction attack bounds are identical to Hayes et al. (2023, Figure 5).

Observe that the probability of a successful data reconstruction attack increases almost exactly linearly with the Δ-divergence of the mechanisms from the baseline. This lends the notion of "excess regret" a concrete quantitative interpretation in terms of attack vulnerability, as in this example, an increase of the Δ-divergence from 0 to 0.12 corresponds to a 15% (!) vulnerability increase to data reconstruction attacks compared to the baseline.

6. Discussion and Conclusion

In this work, we established novel mechanism comparison techniques based on the rigorous foundations of the Blackwell theorem. Our results extend previous works by allowing for principled comparisons between DP mechanisms whose privacy guarantees coincide at the calibration point but differ elsewhere. Operationally, this enables expressing the regret of switching from one mechanism to another in terms of excess privacy vulnerability in the worst case.

Our results are supported by a novel Bayesian interpretation, which allows for modelling adversarial auxiliary information. Such adversarial modelling is currently witnessing increasing interest, as it enables a principled reasoning about adversarial capabilities both in and beyond the worst case (Balle et al., 2022). Moreover, our analysis characterises the properties of mechanisms that determine the order of universal Blackwell dominance that inevitably emerges under sufficiently many compositions, which facilitates the application of our results to DP-SGD. Employing our results to large-scale DP-SGD workflows reveals that calibrating mechanism parameters to attain optimal accuracy must be mindful of associated privacy vulnerabilities, emphasising the risks of the common practice of reporting privacy guarantees in terms of a single (ε, δ)-pair. Thus, while approximate mechanism comparisons quantify differences between mechanism in terms of privacy vulnerability, we have shown that they can be integrated with considerations of model utility in private ML. In future work, we aim to additionally incorporate factors such as the cost of training models, into our framework.

In conclusion, the widespread adoption of privacy-enhancing technologies like DP relies heavily on a correct and transparent understanding of privacy guarantees. Our findings further this understanding, and offer tools to aid informed decision-making in privacy-preserving ML.

Impact Statement

We improve the granularity of DP analyses by introducing a novel method to compare privacy guarantees, which can be applied to enhance the security properties of sensitive data processing systems, benefiting individuals. We foresee no specific negative social consequences of our work.

References

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

Balle, B., Barthe, G., and Gaboardi, M. Privacy profiles and amplification by subsampling. *Journal of Privacy and Confidentiality*, 10(1), 2020a.

Balle, B., Barthe, G., Gaboardi, M., Hsu, J., and Sato, T. Hypothesis testing interpretations and Rényi differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pp. 2496–2506. PMLR, 2020b.

Balle, B., Cherubin, G., and Hayes, J. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy*, pp. 1138–1156. IEEE, 2022.

Berrada, L., De, S., Shen, J. H., Hayes, J., Stanforth, R., Stutz, D., Kohli, P., Smith, S. L., and Balle, B. Unlocking accuracy and fairness in differentially private image classification. *arXiv preprint arXiv:2308.10888*, 2023.

Blackwell, D. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, pp. 265–272, 1953.

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

De, S., Berrada, L., Hayes, J., Smith, S. L., and Balle, B. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

Dong, J., Roth, A., and Su, W. J. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 2022. URL https://academic.oup.com/jrsssb/article/84/1/3/7056089.

Doroshenko, V., Ghazi, B., Kamath, P., Kumar, R., and Manurangsi, P. Connect the dots: Tighter discrete approximations of privacy loss distributions. *Proceedings on Privacy Enhancing Technologies*, 4:552–570, 2022.

Gopi, S., Lee, Y. T., and Wutschitz, L. Numerical composition of differential privacy. *Advances in Neural Information Processing Systems*, 34:11631–11642, 2021.

Hayes, J., Mahloujifar, S., and Balle, B. Bounding Training Data Reconstruction in DP-SGD. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Jayaraman, B., Wang, L., Knipmeyer, K., Gu, Q., and Evans, D. Revisiting membership inference under realistic assumptions. *Proceedings on Privacy Enhancing Technologies*, 2021(2), 2021.

Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

LeCam, L. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, pp. 1419–1455, 1964.

Lokna, J., Paradis, A., Dimitrov, D. I., and Vechev, M. Group and attack: Auditing differential privacy. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1905–1918. Association for Computing Machinery, 2023. ISBN 9798400700507.

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

Nasr, M., Hayes, J., Steinke, T., Balle, B., Tramèr, F., Jagielski, M., Carlini, N., and Terzis, A. Tight auditing of differentially private machine learning. In *Proceedings of the 32nd USENIX Conference on Security Symposium*. USENIX Association, 2023.

Neyman, J. and Pearson, E. S. 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.

Papernot, N., Thakurta, A., Song, S., Chien, S., and Erlingsson, Ú. Tempered sigmoid activations for deep learning with differential privacy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9312–9321, 2021.

Sommer, D., Meiser, S., and Mohammadi, E. Privacy loss classes: The central limit theorem in differential privacy. *Cryptology ePrint Archive*, 2018.

Torgersen, E. *Comparison of statistical experiments; Encyclopaedia of Mathematics and its Applications*, volume 36. Cambridge University Press, 1991.

Wasserman, L. and Zhou, S. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

Wasserstein, R. L. and Lazar, N. A. The asa statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, April 2016. ISSN 1537-2731. doi: 10.1080/00031305.2016.

1154108. URL http://dx.doi.org/10.1080/00031305.2016.1154108.

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

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

Appendix

A. Extended Background

In this section, we provide an extended introduction to the fundamental concepts used in our work for the purpose of self-containedness and for readers without extensive background knowledge of DP.

(ε, δ)-DP A randomised mechanism $\mathcal{M}$ satisfies (ε, δ)-DP if, for all adjacent pairs of databases D, D' (i.e. differing in the data of a single individual), and all $S \subseteq \text{Range}(\mathcal{M})$:

$$\Pr\left(\mathcal{M}(D) \in S\right) \leqslant \mathrm{e}^{\varepsilon} \Pr\left(\mathcal{M}(D') \in S\right) + \delta. \tag{16}$$

We will denote adjacent D, D' by $D \simeq D'$.

(Log-) Likelihood Ratios The likelihood ratios (LRs) are defined as:

$$\bar{X} = \frac{\mathcal{L}(\omega \mid \mathcal{M}(D'))}{\mathcal{L}(\omega \mid \mathcal{M}(D))}, \omega \sim \mathcal{M}(D) \text{ and} \tag{17}$$

$$\bar{Y} = \frac{\mathcal{L}(\omega \mid \mathcal{M}(D'))}{\mathcal{L}(\omega \mid \mathcal{M}(D))}, \omega \sim \mathcal{M}(D'), \tag{18}$$

for arbitrary $D \simeq D'$, where $\mathcal{L}(\omega \mid \cdot)$ denotes the likelihood of ω and $\sim$ denotes "is sampled from". Moreover, the log LRs (LLRs) are defined as $X = \log(\bar{X})$ and $Y = \log(\bar{Y})$.

The LLRs are customarily called the *privacy loss random variables* (PLRVs), and their densities, denoted p_X, p_Y, are called the privacy loss distributions (PLDs). We will make no other assumptions about (P, Q) other than that they are mutually absolutely continuous for all $D \simeq D'$. This only excludes mechanisms whose PLDs have non-zero probability mass at $\pm\infty$ e.g. mechanisms which can fail catastrophically, but allows us to study almost all mechanisms commonly used in private statistics/ML.

Hypothesis Testing and f-DP In the hypothesis testing interpretation (Wasserman & Zhou, 2010; Dong et al., 2022), a MIA adversary observes a mechanism outcome ω and establishes the following hypotheses:

$$H_0 : \omega \sim \mathcal{M}(D) \text{ and } H_1 : \omega \sim \mathcal{M}(D') \tag{19}$$

for arbitrary $D \simeq D'$. H_0 is called the null hypothesis and H_1 the alternative hypothesis and H_0 is tested against H_1 using a randomised rejection rule (i.e. test) $\phi : \omega \mapsto \phi(\omega) \in [0, 1]$, where 0 encodes "reject H_0" and 1 "fail to reject H_0". We then denote the Type-I error of ϕ by $\alpha_\phi = \mathbb{E}_{\omega \sim \mathcal{M}(D)}[\phi(\omega)]$ and its Type-II error by $\beta_\phi = 1 - \mathbb{E}_{\omega \sim \mathcal{M}(D')}[\phi(\omega)]$, where the expectation is over the joint randomness of ϕ and $\mathcal{M}$.

The Neyman-Pearson lemma (Neyman & Pearson, 1933) states that the test with the lowest Type-II error at a given level of Type-I error (called the most powerful test) is constructed by thresholding the (L)LR test statistic; therefore the PLRVs serve as the test statistics for the adversary's hypothesis test. At a level α fixed by the adversary, the trade-off function T of the most powerful test is given by:

$$T(\mathcal{M}(D), \mathcal{M}(D'))(\alpha) = \inf_{\phi} \{\beta_\phi \mid \alpha_\phi \leqslant \alpha\}. \tag{20}$$

f-DP (Dong et al., 2022) is defined by comparing T to a "reference" trade-off function. Formally, $\mathcal{M}$ satisfies f-DP if, for a trade-off function f and for all $D \simeq D'$:

$$\forall \alpha \in [0, 1] : \sup_{D \simeq D'} T(\mathcal{M}(D), \mathcal{M}(D'))(\alpha) \geqslant f(\alpha). \tag{21}$$

Trade-off functions are convex, continuous and weakly decreasing with $f(0) = 1$ and $f(1) = 0$. We will, without loss of generality, extend any trade-off function f to $\mathbb{R} \to [0, 1]$ and set $f(x) = 1, x < 0$ and $f(x) = 0, x > 1$.

Dominating Pairs Working with pairs of adjacent databases is not desirable, and not even always feasible when studying general DP mechanisms. As shown by Zhu et al. (2022), it is instead possible to fully characterise the properties of DP mechanisms by a pair of distributions, called the mechanism's *dominating pair*. Formally, a pair of distributions (P, Q) is called a dominating pair for mechanism $\mathcal{M}$ if, for all $\alpha \in [0, 1]$ it satisfies:

$$\sup_{D,D'} T(P, Q)(\alpha) \leqslant T(\mathcal{M}(D), \mathcal{M}(D'))(\alpha). \tag{22}$$

In particular, when for all $\alpha \in [0, 1]$ it holds that:

$$\sup_{D,D'} T(P, Q)(\alpha) = T(\mathcal{M}(D), \mathcal{M}(D'))(\alpha), \tag{23}$$

(P, Q) is called a *tightly dominating pair*.

As noted by Zhu et al. (2022), a tightly dominating pair which encapsulates the worst-case properties of the mechanism, exists or can always be constructed. Therefore, we will from now on write $\mathcal{M} : (P, Q)$ to indicate that (P, Q) is a tightly dominating pair of $\mathcal{M}$, denote the trade-off function corresponding to the most powerful test between P and Q by f, its Type-I and Type-II errors by α, $\beta(\alpha)$ and the LLRs/PLRVs corresponding to P and Q by $\bar{X}, X$ and $\bar{Y}, Y$. The trade-off function f can be constructed from X and Y as follows. Denoting the CDF by F:

$$f(\alpha) = F_Y(F_X^{-1}(1 - \alpha)). \tag{24}$$

Privacy Profile As shown by Dong et al. (2022); Gopi et al. (2021), the privacy profile of $\mathcal{M}$ can be constructed as:

$$\delta(\varepsilon) = 1 + f^*(P, Q)(-\mathrm{e}^{\varepsilon}) = \bar{F}_Y(\varepsilon) - \mathrm{e}^{\varepsilon}\bar{F}_X(\varepsilon), \tag{25}$$

where T^* is the convex conjugate and $\bar{F}$ the survival function. The privacy profile can also be defined through the *hockey-stick divergence* of order e^{ε} of P to Q:

$$\mathsf{H}_{\mathrm{e}^{\varepsilon}}(P \parallel Q) = \int \max\{P(x) - \mathrm{e}^{\varepsilon}Q(x), 0\}\,\mathrm{d}x = \delta(\varepsilon). \tag{26}$$

Note that, for $\varepsilon = 0$, $\mathsf{H}_1(P \parallel Q) = \delta(0) = \mathsf{TV}(P, Q)$, where

$$\mathsf{TV}(P, Q) = {}^1\!/_2 \int |P(x) - Q(x)|_1\,\mathrm{d}x \tag{27}$$

is the total variation distance.

Additionally, the following property holds:

$$\min_{\alpha\in[0,1]} (\alpha + \beta(\alpha)) = 1 - \mathsf{TV}(P, Q), \tag{28}$$

which links the properties of the privacy profile and the trade-off function. This also allows us to define the *MIA advantage* (Yeom et al., 2018) of the adversary as follows:

$$\mathsf{Adv} = 1 - \min_{\alpha\in[0,1]} (\alpha + \beta(\alpha)) = \mathsf{TV}(P, Q). \tag{29}$$

Rényi-DP Rényi DP (RDP) (Mironov, 2017) is a DP interpretation with beneficial composition properties. A mechanism $\mathcal{M} : (P, Q)$ satisfies $(t, \rho(t))$-RDP if it holds that:

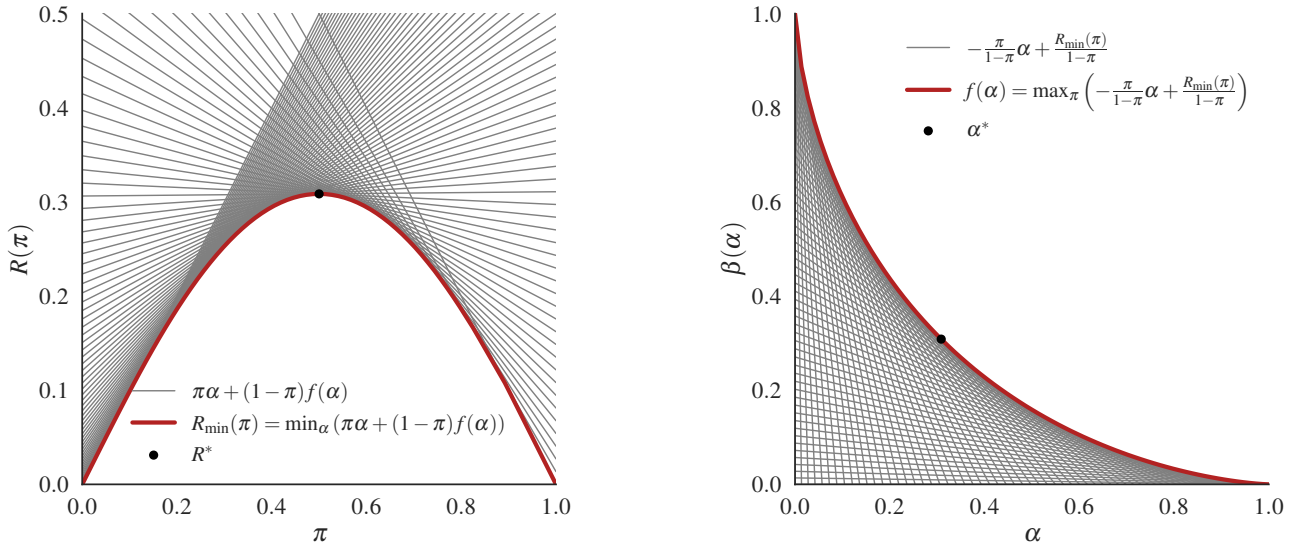
$$\mathsf{D}_t(P\|Q) \leqslant \rho(t)\ \forall t \geqslant 1 \tag{30}$$

for all adjacent (D, D'), where D_t is the Rényi divergence of order t. The conversion between f-DP and the privacy profile is exact, but conversions from RDP to either of the aforementioned are not, as RDP lacks a hypothesis testing interpretation (Balle et al., 2020b; Zhu et al., 2022).

B. Additional Results

B.1. Bayes Error Functions

Here, we demonstrate the construction of the minimum Bayes error function $R_{\min}$ from the trade-off function f and *vice versa* using the example of a Gaussian mechanism with $\sigma^2 = 1$ on a function with unit global sensitivity. Figure 8a shows the construction of $R_{\min}$ from f, while Figure 8b shows the construction of f from $R_{\min}$. Both directions incur no loss of information, and thus the minimum Bayes error is equivalent to the trade-off function in terms of fully characterising the mechanism.



(a) $f \to R_{\min}$ (b) $R_{\min} \to f$

Figure 8: Lossless conversion between $R_{\min}$ and f. R^* and α^* are the minimax Bayes error and fixed point of f, respectively.

B.2. Interpreting $\Delta^{\leftrightarrow}$ as the Lévy Distance

Following the discussion in Section 3.3 regarding the conceptual equivalence of the Δ-divergence and the Lévy distance between random variables, we here formally introduce and prove the statement.

Lemma B.1. *For any mechanism with trade-off function f, (U, W) are a tightly dominating pair, where U is the continuous uniform distribution on the unit interval and W has CDF: $f(1-\alpha)$. Moreover, for two mechanisms $\mathcal{M} : (U, W)$ and $\widetilde{\mathcal{M}} : (U, W')$, it holds that $\Delta^{\leftrightarrow}(\mathcal{M} \| \widetilde{\mathcal{M}}) = \Lambda(W, W')$, where Λ denotes the Lévy distance.*

Proof. We will first show that, if $\mathcal{M}$ is tightly dominated by (P, Q) and has trade-off function f, it is also tightly dominated by (U, W). From Equation (24), $T(U, W)$ is constructed as follows:

$$T(U, W)(\alpha) = F_W(F_U^{-1}(1-\alpha)) = F_W(1-\alpha) = f(1-(1-\alpha)) = f(\alpha), \tag{31}$$

which follows since the inverse CDF (quantile function) of the continuous uniform distribution is the identity function. Therefore, $T(U, W)(\alpha) = f(\alpha)$, for all $\alpha \in [0, 1]$, hence (U, W) is a tightly dominating pair for $\mathcal{M}$. Next, recall the definition of the Lévy distance:

$$\Lambda(W, W') = \inf \{\lambda \geqslant 0 \mid \forall x \in \mathbb{R} : \; F_W(x-\lambda) - \lambda \leqslant F_{W'}(x) \leqslant F_W(x+\lambda) + \lambda)\}. \tag{32}$$

Denoting $f, \widetilde{f}$ the trade-off functions of $\mathcal{M}, \widetilde{\mathcal{M}}$ respectively and inserting the respective CDFs of W, W', we obtain:

$$\begin{aligned} \Lambda(W, W') &= \inf\{\lambda \geqslant 0 \mid \forall \alpha \in \mathbb{R} : \; f(1-(\alpha-\lambda)) - \lambda \leqslant \widetilde{f}(1-\alpha) \leqslant f(1-(\alpha+\lambda)) + \lambda\} && (33) \\ &= \inf\{\lambda \geqslant 0 \mid \forall \alpha \in \mathbb{R} : \; f(1-\alpha+\lambda) - \lambda \leqslant \widetilde{f}(1-\alpha) \leqslant f(1-\alpha-\lambda) + \lambda\}. && (34) \end{aligned}$$

We can reparameterise the inequality chain from $1-\alpha$ to α to obtain:

$$\Lambda(W, W') = \inf\{\lambda \geqslant 0 \mid \forall \alpha \in \mathbb{R} : \ f(\alpha+\lambda) - \lambda \leqslant \widetilde{f}(\alpha) \leqslant f(\alpha-\lambda) + \lambda\}. \tag{35}$$

Noticing that the result is identical to the definition of $\Delta^{\leftrightarrow}(\mathcal{M} \| \widetilde{\mathcal{M}})$ completes the proof. □

B.3. Δ-Divergence Implementation

The following code listing implements the Δ-divergence computation corresponding to the mechanisms in Figure 3 in Python. As seen, the algorithm only requires oracle access to a function implementing the trade-off function of the mechanism.

```python
from scipy.stats import norm, laplace
import numpy as np
from functools import partial
from scipy.optimize import minimize_scalar
from multiprocessing import Pool
from os import cpu_count
from typing import Callable, Sequence, Union

def f_gauss(alpha: Union[Sequence[float], float], mu: float) -> float:
  "Gaussian mechanism trade-off function at alpha with parameter mu."
  assert (alpha >= np.zeros_like(alpha)).all() and (
    alpha <= np.ones_like(alpha)
  ).all(), "alpha must be in [0, 1]"
  assert mu >= 0, "mu must be non-negative"
  return norm.cdf(norm.isf(alpha) - mu)

def f_lap(alpha: Union[Sequence[float], float], mu: float) -> float:
  "Laplace mechanism trade-off function at alpha with parameter mu."
  assert (alpha >= np.zeros_like(alpha)).all() and (
    alpha <= np.ones_like(alpha)
  ).all(), "alpha must be in [0, 1]"
  assert mu >= 0, "mu must be non-negative"
  return laplace.cdf(laplace.isf(alpha) - mu)

def _compute_one_rmin(
  pi: float,
  f: Callable[[Union[Sequence[float], float]], float],
) -> float:
  assert 0 <= pi <= 1, "pi must be in [0, 1]"

  def func(alpha: float) -> float:
    assert 0 <= alpha <= 1, "alpha must be in [0, 1]"
    return pi * alpha + (1 - pi) * f(alpha)

  return minimize_scalar(func, bounds=(0, 1)).fun

def rmin(
  *,
  f: Callable[[Union[Sequence[float], float]], float],
  tol: float = 1e-4,
  n_jobs: int = -1,
) -> Union[Sequence[float], float]:
  "Bayes error function corresponding to f computed with tolerance tol."
  assert tol > 0, "tol must be positive"
  assert n_jobs == -1 or n_jobs > 0, "n_jobs must be positive or -1"
  N: int = int(np.ceil(1 / tol))
  pis: Sequence[float] = np.linspace(0, 1, N)
  if n_jobs == -1:
    processes = cpu_count()
  else:
    processes = n_jobs
  with Pool(processes) as pool:
    result = np.array(pool.map(partial(_compute_one_rmin, f=f), pis))
  return result
```

```python
if __name__ == "__main__":
  tol: float = 1e-4
  mu: float = 1.0
  rmin_lap: Sequence[float] = rmin(f=partial(f_lap, mu=mu), tol=tol, n_jobs=-1)
  rmin_gauss: Sequence[float] = rmin(f=partial(f_gauss, mu=mu), tol=tol, n_jobs=-1)
  divergence_gauss_lap: float = max(rmin_gauss - rmin_lap)
  divergence_lap_gauss: float = max(rmin_lap - rmin_gauss)
  print(f"Delta(Gauss || Lap): {divergence_gauss_lap:.3f}") #prints 0.005
  print(f"Delta(Lap || Gauss): {divergence_lap_gauss:.3f}") #prints 0.034
```

B.4. Proofs

Theorem 1. The following statements are equivalent:

1. $\forall \alpha \in [0,1] : f(\alpha) \leqslant \widetilde{f}(\alpha)$;
2. $\forall \varepsilon \in \mathbb{R} : \delta(\varepsilon) \geqslant \widetilde{\delta}(\varepsilon)$;
3. $\forall \pi \in [0,1] : R_{\min}(\pi) \leqslant \widetilde{R}_{\min}(\pi)$.

Proof. For a full proof, see the proof of Theorem 2, which recovers Theorem 1 for $\mathfrak{D} = 0$. □

Theorem 2. The following are equivalent to $\mathcal{M} \geq_{\mathfrak{D}} \widetilde{\mathcal{M}}$:

1. $\forall \alpha \in [0,1] : f(\alpha + \mathfrak{D}) - \mathfrak{D} \leqslant \widetilde{f}(\alpha)$;
2. $\forall \varepsilon \in \mathbb{R} : \delta(\varepsilon) + \mathfrak{D} \cdot (1 + \mathrm{e}^{\varepsilon}) \geqslant \widetilde{\delta}(\varepsilon)$;
3. $\forall \pi \in [0,1] : R_{\min}(\pi) - \widetilde{R}_{\min}(\pi) \leqslant \mathfrak{D}$.

Proof.
(1) : Suppose $\Delta = \Delta(\mathcal{M} \| \widetilde{\mathcal{M}}) \leqslant \mathfrak{D}$. Since trade-off functions are weakly decreasing, we have:

$$f(\alpha + \mathfrak{D}) - \mathfrak{D} \leqslant f(\alpha + \Delta) - \Delta \leqslant \widetilde{f}(\alpha). \tag{36}$$

Conversely, if $f(\alpha + \mathfrak{D}) - \mathfrak{D} \leqslant \widetilde{f}(\alpha)$, then we have $\Delta(\mathcal{M} \| \widetilde{\mathcal{M}}) \leqslant \mathfrak{D}$ due to the infimum definition of the Δ-Divergence.
(1) $\Rightarrow$ **(2)** : Suppose for all $-\infty < \alpha < \infty$, we have $f(\alpha + \mathfrak{D}) - \mathfrak{D} \leqslant \widetilde{f}(\alpha)$. From Dong et al. (2022), we know that $\delta(\varepsilon) = 1 + f^*(-\mathrm{e}^{\varepsilon})$, where:

$$f^*(x) = \sup_{-\infty < \alpha < \infty} x\alpha - f(\alpha). \tag{37}$$

denotes the convex conjugate. By direct computation of the convex conjugate we obtain:

$$\delta(\varepsilon) - 1 = f^*(-\mathrm{e}^{\varepsilon}) = \sup_{-\infty < \alpha < \infty} (-\mathrm{e}^{\varepsilon}\alpha - (f(\alpha - \mathfrak{D} + \mathfrak{D}) - \mathfrak{D}) - \mathfrak{D}) \geqslant \sup_{-\infty < \alpha < \infty} (-\mathrm{e}^{\varepsilon}\alpha - \widetilde{f}(\alpha - \mathfrak{D}) - \mathfrak{D}) \tag{38}$$

$$= \sup_{-\infty < \alpha < \infty} (-\mathrm{e}^{\varepsilon}(\alpha + \mathfrak{D}) - \widetilde{f}(\alpha)) - \mathfrak{D} = \widetilde{f}^*(-\mathrm{e}^{\varepsilon}) - \mathfrak{D} \cdot (1 + \mathrm{e}^{\varepsilon}) = \widetilde{\delta}(\varepsilon) - 1 - \mathfrak{D} \cdot (1 + \mathrm{e}^{\varepsilon}), \tag{39}$$

which yields the desired inequality.

(2) $\Rightarrow$ **(1)** : Suppose that, for all $0 \leqslant \varepsilon < \infty$, we have $\delta(\varepsilon) + \mathfrak{D} \cdot (1 + e^{\varepsilon}) \geqslant \widetilde{\delta}(\varepsilon)$. Define the function $\widetilde{f}(\alpha) = \widetilde{f}(\alpha - \mathfrak{D}) + \mathfrak{D}$. We then have for all $\varepsilon \geqslant 0$:

$$f^*(-\mathrm{e}^{\varepsilon}) = \delta(\varepsilon) - 1 \geqslant \widetilde{\delta}(\varepsilon) - 1 - \mathfrak{D} \cdot (1 + \mathrm{e}^{\varepsilon}) = \widetilde{f}^*(-\mathrm{e}^{\varepsilon}) - \mathfrak{D} \cdot (1 + \mathrm{e}^{\varepsilon}) \tag{40}$$

$$= \sup_{-\infty < \alpha < \infty} (-\mathrm{e}^{\varepsilon}\alpha - \widetilde{f}(\alpha)) - \mathfrak{D} \cdot (1 + \mathrm{e}^{\varepsilon}) = \sup_{-\infty < \alpha < \infty} (-\mathrm{e}^{\varepsilon}(\alpha + \mathfrak{D}) - (\widetilde{f}(\alpha) + \mathfrak{D})) \tag{41}$$

$$= \sup_{-\infty < \alpha < \infty} (-\mathrm{e}^{\varepsilon}\alpha - (\widetilde{f}(\alpha - \mathfrak{D}) + \mathfrak{D})) = \sup_{-\infty < \alpha < \infty} (-\mathrm{e}^{\varepsilon}\alpha - \widetilde{f}(\alpha)) = \widetilde{f}^*(-\mathrm{e}^{\varepsilon}). \tag{42}$$

This shows $f^* \geqslant \widetilde{f}^*$, which implies $f \leqslant \widetilde{f}$ since the convex conjugate is order-reversing. By definition of $\widetilde{f}$, we showed for all α:

$$f(\alpha) \leqslant \widetilde{f}(\alpha - \mathfrak{D}) + \mathfrak{D}. \tag{43}$$

(1) ⇒ (3) : Suppose for all $\alpha \in [0,1]$ we have $f(\alpha+\mathfrak{D})-\mathfrak{D} \leqslant \widetilde{f}(\alpha)$. Let $\alpha \in [0,1]$, such that $\widetilde{R}_{\min}(\pi) = \pi\alpha+(1-\pi)\widetilde{f}(\alpha)$. If $\alpha + \mathfrak{D} \in [0,1]$, then we have:

$$R_{\min}(\pi) \leqslant \pi(\alpha+\mathfrak{D}) + (1-\pi)f(\alpha+\mathfrak{D}) \leqslant \pi(\alpha+\mathfrak{D}) + (1-\pi)(\widetilde{f}(\alpha)+\mathfrak{D}) \tag{44}$$
$$= R_{\min}(\pi) + \mathfrak{D}. \tag{45}$$

In the other, case, we have $\alpha + \mathfrak{D} > 1$. But then, $\alpha - \mathfrak{D} \in [0,1]$ since $\alpha \in [0,1]$. Using $f(1) = 0 = f(\alpha+\mathfrak{D})$, we also obtain the desired bound:

$$R_{\min}(\pi) \leqslant \pi + (1-\pi)f(1) \leqslant \pi(\alpha+\mathfrak{D}) + (1-\pi)f(\alpha+\mathfrak{D}) \leqslant \pi(\alpha+\mathfrak{D}) + (1-\pi)(\widetilde{f}(\alpha)+\mathfrak{D}) \tag{46}$$
$$= R_{\min}(\pi) + \mathfrak{D}. \tag{47}$$

(3) ⇒ (1) : Suppose $\max_\pi R_{\min}(\pi) - \widetilde{R}_{\min}(\pi) \leqslant \mathfrak{D}$. Let $\alpha \in [0,1]$. If $\alpha + \mathfrak{D} > 1$, then trivially $f(\alpha+\mathfrak{D}) - \mathfrak{D} = -\mathfrak{D} \leqslant 0 = f(\alpha)$ holds. Thus, assume $\alpha + \mathfrak{D} \in [0,1]$. Then, there exists a $\pi \in [0,1]$ such that:

$$R_{\min}(\pi) = \pi(\alpha+\mathfrak{D}) + (1-\pi)f(\alpha+\mathfrak{D}). \tag{48}$$

We use the fact that $\widetilde{R}_{\min}(\pi) \leqslant \pi\alpha + (1-\pi)\widetilde{f}(\alpha)$ and obtain:

$$\mathfrak{D} \geqslant R_{\min}(\pi) - \widetilde{R}_{\min}(\pi) \geqslant \pi(\alpha+\mathfrak{D}) + (1-\pi)f(\alpha+\mathfrak{D}) - (\pi\alpha + (1-\pi)\widetilde{f}(\alpha)) \tag{49}$$
$$= \pi\mathfrak{D} + (1-\pi)(f(\alpha+\mathfrak{D}) - \widetilde{f}(\alpha)). \tag{50}$$

Subtracting $\pi\mathfrak{D}$ from both sides and subsequently dividing by $1-\pi$ yields the desired inequality:

$$\mathfrak{D} \geqslant f(\alpha+\mathfrak{D}) - \widetilde{f}(\alpha). \tag{51}$$

□

Corollary 1. $\Delta(\mathcal{M}\|\widetilde{\mathcal{M}}) = \max_\pi(R_{\min}(\pi) - \widetilde{R}_{\min}(\pi))$.

Proof. By definition, we have

$$\Delta(\mathcal{M}\|\widetilde{\mathcal{M}}) = \inf\{\kappa \geqslant 0 \mid f(\alpha+\kappa) - \kappa \leqslant \widetilde{f}(\alpha)\}.$$

Applying clause (3) in Theorem 2, we immediately obtain:

$$\Delta(\mathcal{M}\|\widetilde{\mathcal{M}}) = \inf\{\kappa \geqslant 0 \mid \max_\pi(R_{\min}(\pi) - \widetilde{R}_{\min}(\pi)) \leqslant \kappa\}.$$

The $\inf$ is attained at the largest difference in the Bayes error functions, thus:

$$\Delta(\mathcal{M}\|\widetilde{\mathcal{M}}) = \max_\pi(R_{\min}(\pi) - \widetilde{R}_{\min}(\pi)). \tag{52}$$

□

Lemma 1. Let $\Delta^\leftrightarrow = \Delta^\leftrightarrow(\mathcal{M}\|\widetilde{\mathcal{M}})$. Then it holds that:

$$f(\alpha+\Delta^\leftrightarrow) - \Delta^\leftrightarrow \leqslant \widetilde{f}(\alpha) \leqslant f(\alpha-\Delta^\leftrightarrow) + \Delta^\leftrightarrow. \tag{7}$$

Proof. Let $\mathfrak{D} = \Delta(\mathcal{M}\,\|\,\widetilde{\mathcal{M}})$ and $\mathfrak{F} = \Delta(\widetilde{\mathcal{M}}\,\|\,\mathcal{M})$, i.e. we have $\mathcal{M} \succeq_\mathfrak{D} \widetilde{\mathcal{M}}$ and $\widetilde{\mathcal{M}} \succeq_\mathfrak{F} \mathcal{M}$. By Theorem 2, we have that $f(\alpha+\mathfrak{D}) - \mathfrak{D} \leqslant \widetilde{f}(\alpha)$ and $\widetilde{f}(\alpha) \leqslant f(\alpha-\mathfrak{F}) + \mathfrak{F}$, for all α. Since trade-off functions are weakly decreasing and $\mathfrak{D}, \mathfrak{F} \leqslant \Delta^\leftrightarrow$, we have:

$$f(\alpha+\Delta^\leftrightarrow) - \Delta^\leftrightarrow \leqslant f(\alpha+\mathfrak{D}) - \mathfrak{D} \leqslant \widetilde{f}(\alpha) \leqslant f(\alpha-\mathfrak{F}) + \mathfrak{F} \leqslant f(\alpha-\Delta^\leftrightarrow) + \Delta^\leftrightarrow. \tag{53}$$

□

Corollary 2. $\Delta^\leftrightarrow$ is a metric.

Proof. We need to show that $\Delta^{\leftrightarrow}(\mathcal{M}\|\widetilde{\mathcal{M}}) = 0 \Leftrightarrow \mathcal{M} = \widetilde{\mathcal{M}}$ and that $\Delta^{\leftrightarrow}$ is symmetric and satisfies the triangle inequality. Applying Corollary 1 we obtain:

$$\Delta^{\leftrightarrow}(\mathcal{M}\|\widetilde{\mathcal{M}}) = \max\left\{\Delta(\mathcal{M}\|\widetilde{\mathcal{M}}), \Delta(\widetilde{\mathcal{M}}\|\mathcal{M})\right\} \tag{54}$$

$$= \max\left\{\max_{\pi}\left(R_{\min}(\pi) - \widetilde{R}_{\min}(\pi)\right), \max_{\pi}\left(\widetilde{R}_{\min}(\pi) - R_{\min}(\pi)\right)\right\} = \|R_{\min} - \widetilde{R}_{\min}\|_{\infty}. \tag{55}$$

Symmetry, triangle inequality, and $\Delta^{\leftrightarrow}(\mathcal{M}\|\mathcal{M}) = 0 \Leftrightarrow \mathcal{M} = \widetilde{\mathcal{M}}$ follow from the fact that $\|\cdot\|_{\infty}$ is a norm. □

Remark 1. Note that we introduced the order relation $\geq$ which is implied by the Blackwell theorem as a *partial order*, and refer to mechanisms as *equal* ($\mathcal{M} = \widetilde{\mathcal{M}}$) if and only if they offer identical privacy guarantees. Moreover, we refer to $\Delta^{\leftrightarrow}$ as a *metric* on the space of DP mechanisms. This choice is motivated by an operational interpretation: *For all practical intents and purposes, mechanisms which provide identical guarantees are the same mechanism*. It is however also possible to subject the aforementioned statements to a more formal order-theoretic treatment, where the symbol "=" is reserved for objects which satisfy *identity*. Since conferring identical privacy guarantees is not sufficient for being identical, it can be argued that it is more appropriate to refer to distinct mechanisms with identical privacy guarantees as being *equivalent*, and writing $\mathcal{M} \equiv \widetilde{\mathcal{M}}$. For example, the mechanisms $\mathcal{M} : (\mathcal{N}(0,1), \mathcal{N}(1,1))$ and $\widetilde{\mathcal{M}} : (\mathcal{N}(0,2), \mathcal{N}(2,2))$ have identical trade-off functions, privacy profiles and Bayes error functions and are thus *equivalent*, but they have different dominating pairs, and are therefore not *identical*. Under this perspective, the order relation $\geq$ formally loses its antisymmetry property, since $\mathcal{M} \geq \widetilde{\mathcal{M}}$ and $\widetilde{\mathcal{M}} \geq \mathcal{M}$ no longer implies that $\mathcal{M} = \widetilde{\mathcal{M}}$ but rather $\mathcal{M} \equiv \widetilde{\mathcal{M}}$, and thus should be referred to as a *preorder*. Moreover, since under this treatment, $\Delta^{\leftrightarrow}(\mathcal{M}\,\|\,\mathcal{M}) = 0$ implies $\Leftrightarrow \mathcal{M} \equiv \widetilde{\mathcal{M}}$ rather than $\Leftrightarrow \mathcal{M} = \widetilde{\mathcal{M}}$, $\Delta^{\leftrightarrow}$ should be referred to as a *pseudometric* (which assigns zero value to *non-identical* (but equivalent) elements). We stress that the discussed distinction is largely terminological and does not change any of the results of the paper.

Lemma 2. $\mathcal{M} \geq \mathcal{M}_{\mathrm{PP}}$ and $\mathcal{M}_{\mathrm{BNP}} \geq \mathcal{M}$ for any $\mathcal{M}$.

Proof. ($\mathcal{M} \geq \mathcal{M}_{\mathrm{PP}}$): We have $R^{\mathrm{PP}}_{\min} \geqslant R_{\min}$, since, by definition, $R^{\mathrm{PP}}_{\min}(\pi) = \min\{\pi, 1-\pi\}$, and the Bayes error function of any mechanism satisfies $R_{\min}(\pi) \leqslant \min\{\pi, 1-\pi\}$, for all $\pi \in [0,1]$. Thus, by Theorem 1, $\mathcal{M}_{\mathrm{PP}}$ is Blackwell dominated by any mechanism.

($\mathcal{M}_{\mathrm{BNP}} \geq \mathcal{M}$): By definition, $R^{\mathrm{BNP}}_{\min}(\pi) = 0$ and thus $R^{\mathrm{BNP}}_{\min}(\pi) \leqslant R_{\min}(\pi)$, for all $\pi \in [0,1]$. Thus, by Theorem 1, $\mathcal{M}_{\mathrm{BNP}}$ Blackwell dominates any mechanism. □

Lemma 3. It holds that $\Delta(\mathcal{M}_{\mathrm{PP}}\|\mathcal{M}) = 1/2\mathsf{TV}(P,Q) = 1/2\mathsf{Adv} = 1/2\delta(0)$.

Proof. We denote the Bayes error functions of $\mathcal{M}, \mathcal{M}_{\mathrm{PP}}$ as $R_{\min}, R^{\mathrm{PP}}_{\min}$ respectively. Note that $R^{\mathrm{PP}}_{\min}(\pi) = \min\{\pi, 1-\pi\} \geqslant R_{\min}(\pi)$, for all $\pi \in [0,1]$. Using Corollary 1 we obtain:

$$\Delta(\mathcal{M}_{\mathrm{PP}}\|\mathcal{M}) = \max_{\pi}(R^{\mathrm{PP}}_{\min}(\pi) - R_{\min}(\pi)) = \max_{\pi}\left(\min\{\pi, 1-\pi\} - R_{\min}(\pi)\right). \tag{56}$$

Next, note that the maximum of $\min\{\pi, 1-\pi\}$ is at $\pi = 1/2$ and that all Bayes error functions are concave by definition and their maximum is also realised at $\pi = 1/2$. Hence, the largest difference between the perfectly private mechanism and any Bayes risk function must also be at $\pi = 1/2$. We have:

$$\Delta(\mathcal{M}_{\mathrm{PP}}\|\mathcal{M}) = 1/2 - R_{\min}(1/2) \tag{57}$$

$$= 1/2 - \min_{\alpha\in[0,1]}(\alpha 1/2 + f(\alpha)1/2) \tag{58}$$

$$= 1/2 \min_{\alpha\in[0,1]}(1 - \alpha - f(\alpha)) \tag{59}$$

$$= 1/2\mathsf{Adv} = 1/2\mathsf{TV}(P,Q) = 1/2\delta(0). \tag{60}$$

□

Lemma 4. It holds that $\Delta(\mathcal{M}\|\mathcal{M}_{\mathrm{BNP}}) = R^* = \alpha^*$, where R^* is the minimax Bayes error and α^* the fixed point of the trade-off function of $\mathcal{M}$.

Proof. Since the Bayes risk function of $\mathcal{M}_{\mathrm{BNP}}$ is 0 on the unit interval, the Δ-divergence becomes:

$$\Delta(\mathcal{M}\|\mathcal{M}_{\mathrm{BNP}}) = \max_{\pi}(R_{\min}(\pi) - R^{\mathrm{BNP}}_{\min}(\pi)) = \max_{\pi\in[0,1]}(R_{\min}(\pi) - 0) = \max_{\pi\in[0,1]} R_{\min}(\pi) = R^*, \tag{61}$$

where we used Corollary 1 for the first equality. It remains to show that $R^* = \alpha^*$. Recall that $R_{\min}$ is concave and symmetric around $\pi = 1/2$ and assumes its maximum at $\pi = 1/2$. To compute $R_{\min}(1/2)$, we set the following derivative equal to 0:

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\left[1/2\pi\alpha + 1/2(1 - \pi f(\alpha))\right] = 0 \iff \frac{\mathrm{d}}{\mathrm{d}\alpha} f(\alpha) = -1 \iff \alpha = f(\alpha). \tag{62}$$

The last equivalence follows from the fact that f is a symmetric trade-off function. Denote by α^* the unique point in $[0, 1]$ such that $\alpha^* = f(\alpha^*)$. Then, we have:

$$\Delta(\mathcal{M}\|\mathcal{M}_{\mathrm{BNP}}) = R^*_{\min} = R_{\min}(1/2) = 1/2\alpha^* + 1/2 f(\alpha^*) = 1/2\alpha^* + 1/2\alpha^* = \alpha^*. \tag{63}$$

□

Lemma 5. $\Delta(\mathcal{M}_{\mathrm{PP}}\|\mathcal{M}) + \Delta(\mathcal{M}\|\mathcal{M}_{\mathrm{BNP}}) = 0.5$.

Proof. Since $R^{\mathrm{PP}}_{\min}(\pi) = \min\{\pi, 1 - \pi\}$ which has a maximum at $\pi = 1/2$, we have from Equation (57) that $\Delta(\mathcal{M}_{\mathrm{PP}}\|\mathcal{M}) = 1/2 - R_{\min}(1/2)$. Moreover, by Equation (63), we have $\Delta(\mathcal{M}\|\mathcal{M}_{\mathrm{BNP}}) = R_{\min}(1/2)$. Therefore, we obtain:

$$\Delta(\mathcal{M}_{\mathrm{PP}}\|\mathcal{M}) + \Delta(\mathcal{M}\|\mathcal{M}_{\mathrm{BNP}}) = 1/2 - R_{\min}(1/2) + R_{\min}(1/2) = 1/2. \tag{64}$$

□

Before proceeding with Lemma 6 and Theorem 3, we prove the following statements, which will be used below:

Lemma B.2. *If $G_\mu, G_{\tilde{\mu}}$ are two Gaussian trade-off functions with $\mu \leqslant \tilde{\mu}$, then $G_\mu \geqslant G_{\tilde{\mu}}$.*

Proof. We will prove that the trade-off function of the Gaussian mechanism is decreasing in μ for any fixed α. To show this, we take the first derivative of the trade-off function of the Gaussian mechanism with respect to μ:

$$\frac{\partial}{\partial\mu} G_\mu(\alpha) = \frac{\partial}{\partial\mu}\Phi(\Phi^{-1}(1-\alpha) - \mu) = -\frac{\sqrt{2}\mathrm{e}^{-\frac{\left(\mu - \sqrt{2}\,\mathrm{erfinv}\,(1-2\alpha)\right)^2}{2}}}{2\sqrt{\pi}}, \tag{65}$$

where erfinv denotes the inverse error function of the normal distribution. Since the exponential is always non-negative, the right hand side is always negative. Hence:

$$\mu \geqslant \tilde{\mu} \iff \forall\alpha \in [0,1]:\ G_\mu(\alpha) \leqslant G_{\tilde{\mu}}(\alpha). \tag{66}$$

□

Lemma B.3. *Let $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ be three mechanisms. Then, $\Delta(\mathcal{M}_1\|\mathcal{M}_3) \leqslant \Delta(\mathcal{M}_1\|\mathcal{M}_2) + \Delta(\mathcal{M}_2\|\mathcal{M}_3)$.*

Proof. Let $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ be three mechanisms and $R^1_{\min}, R^2_{\min}, R^3_{\min}$ their respective Bayes error functions. Using Corollary 1, we have:

$$\begin{aligned}
\Delta(\mathcal{M}_1\|\mathcal{M}_3) &= \max_{\pi}(R^1_{\min}(\pi) - R^3_{\min}(\pi)) && (67)\\
&= \max_{\pi}(R^1_{\min}(\pi) - R^2_{\min}(\pi) + R^2_{\min}(\pi) - R^3_{\min}(\pi)) && (68)\\
&\leqslant \max_{\pi}(R^1_{\min}(\pi) - R^2_{\min}(\pi)) + \max_{\pi}(R^2_{\min}(\pi) - R^3_{\min}(\pi)) && (69)\\
&\leqslant \max_{\pi}(R^1_{\min}(\pi) - R^2_{\min}(\pi)) + \max_{\pi}(R^2_{\min}(\pi) - R^3_{\min}(\pi)) && (70)\\
&= \Delta(\mathcal{M}_1\|\mathcal{M}_2) + \Delta(\mathcal{M}_2\|\mathcal{M}_3). && (71)
\end{aligned}$$

□

We now proceed with the proofs of Lemma 6 and Theorem 3 in the main manuscript.

Lemma 6. Let $\{\mathcal{M}_{Ni} : 1 \leqslant i \leqslant N\}_{N=1}^{\infty}$ be a triangular array of mechanisms satisfying the following conditions:

1. $\lim_{N\to\infty} \sum_{i=1}^{N} v_1(f_{Ni}) = K$;
2. $\lim_{N\to\infty} \max_{1\leqslant i\leqslant N} v_1(f_{Ni}) = 0$;
3. $\lim_{N\to\infty} \sum_{i=1}^{N} v_2(f_{Ni}) = s^2$;
4. $\lim_{N\to\infty} \sum_{i=1}^{N} v_4(f_{Ni}) = 0$.

Analogously, define $\{\widetilde{\mathcal{M}}_{Ni} : 1 \leqslant i \leqslant N\}_{N=1}^{\infty}$ for constants $\widetilde{K}, \widetilde{s}$. Then, if $K/s > \widetilde{K}/\widetilde{s}$, there exists N^* such that, for all $N \geqslant N^*$:

$$\mathcal{M}_{N1} \otimes \cdots \otimes \mathcal{M}_{NN} \geq \widetilde{\mathcal{M}}_{N1} \otimes \cdots \otimes \widetilde{\mathcal{M}}_{NN}, \tag{13}$$

where $\mathcal{M}_{N1} \otimes \cdots \otimes \mathcal{M}_{NN}$ denotes N-fold mechanism composition and analogously for $\widetilde{\mathcal{M}}_{Ni}$.

Proof. Denote by $f_{N1} \otimes \cdots \otimes f_{NN}$ and $\widetilde{f}_{N1} \otimes \cdots \otimes \widetilde{f}_{NN}$ the trade-off functions of the compositions $\mathcal{M}_{N1} \otimes \cdots \otimes \mathcal{M}_{NN}$ and $\widetilde{\mathcal{M}}_{N1} \otimes \cdots \otimes \widetilde{\mathcal{M}}_{NN}$ respectively. Next, we apply Theorem 6 in Dong et al. (2022), which states that these trade-off functions uniformly converge to the Gaussian trade-off functions $G_{2K/s}$ and $G_{2\widetilde{K}/\widetilde{s}}$ respectively, i.e.

$$\lim_{N\to\infty} f_{N1} \otimes \cdots \otimes f_{NN} = G_{2K/s}, \tag{72}$$

$$\lim_{N\to\infty} \widetilde{f}_{N1} \otimes \cdots \otimes \widetilde{f}_{NN} = G_{2\widetilde{K}/\widetilde{s}}. \tag{73}$$

Suppose $2K/s > 2\widetilde{K}/\widetilde{s}$ holds. By Lemma B.2, we then have $G_{2K/s} < G_{2\widetilde{K}/\widetilde{s}}$. Moreover, we have:

$$\lim_{N\to\infty} f_{N1} \otimes \cdots \otimes f_{NN} = G_{2K/s}(\alpha) < G_{2\widetilde{K}/\widetilde{s}}(\alpha) = \lim_{N\to\infty} \widetilde{f}_{N1} \otimes \cdots \otimes \widetilde{f}_{NN}(\alpha), \tag{74}$$

where the limits converge uniformly in α. In particular, since the limits converge uniformly and are strictly ordered, there must exist N^* such that for all $N \geqslant N^*$:

$$f_{N1} \otimes \cdots \otimes f_{NN} \leqslant \widetilde{f}_{N1} \otimes \cdots \otimes \widetilde{f}_{NN}. \tag{75}$$

This shows if $2K/s > 2\widetilde{K}/\widetilde{s}$, then there must exist N^* such that for all $N \geqslant N^*$:

$$\mathcal{M}_{N1} \otimes \cdots \otimes \mathcal{M}_{NN} \geq \widetilde{\mathcal{M}}_{N1} \otimes \cdots \otimes \widetilde{\mathcal{M}}_{NN} \tag{76}$$

□

Theorem 3. Let $\mathcal{M}, \widetilde{\mathcal{M}}$ be two mechanisms with $v_4, \widetilde{v}_4 < \infty$ and denote by $\mathcal{M}^{\otimes N}, \widetilde{\mathcal{M}}^{\otimes \widetilde{N}}$ their N- and $\widetilde{N}$-fold self-compositions. Then, $N/\widetilde{N} \geqslant \widetilde{\eta}^2/\eta^2$ implies:

$$\Delta(\mathcal{M}^{\otimes N} \| \widetilde{\mathcal{M}}^{\otimes \widetilde{N}}) \leqslant 0.56 \left(\frac{\eta^3 v_3}{\sqrt{N} v_1^3} + \frac{\widetilde{\eta}^3 \widetilde{v}_3}{\sqrt{\widetilde{N}} \widetilde{v}_1^3} \right) \tag{14}$$

In particular, if $N = \widetilde{N}$, $\eta \geqslant \widetilde{\eta}$ implies:

$$\Delta(\mathcal{M}^{\otimes N} \| \widetilde{\mathcal{M}}^{\otimes \widetilde{N}}) \leqslant \frac{0.56}{\sqrt{N}} \left(\frac{\eta^3 v_3}{v_1^3} + \frac{\widetilde{\eta}^3 \widetilde{v}_3}{\widetilde{v}_1^3} \right). \tag{15}$$

Proof. Assume $N/\widetilde{N} \geqslant \eta/\widetilde{\eta}$. Let $\mathcal{M}_G, \widetilde{\mathcal{M}}_G$ be two Gaussian mechanisms with trade-off functions $G_\mu, G_{\widetilde{\mu}}$ respectively, where:

$$\mu = \frac{\sqrt{N} 2 v_1}{\sqrt{v_2 - v_1^2}} = \sqrt{N} 2\eta \tag{77}$$

$$\widetilde{\mu} = \frac{\sqrt{\widetilde{N}} 2 \widetilde{v}_1}{\sqrt{\widetilde{v}_2 - \widetilde{v}_1^2}} = \sqrt{\widetilde{N}} 2\widetilde{\eta}. \tag{78}$$

Between two Gaussian trade-off functions the one with the smaller mean parameter has a larger trade-off function:

$$G_\mu \leqslant G_{\tilde{\mu}} \iff \mu \geqslant \tilde{\mu} \iff \sqrt{N}\eta \geqslant \sqrt{\widetilde{N}}\tilde{\eta} \iff {}^{N}\!/_{\widetilde{N}} \geqslant {}^{\eta}\!/_{\tilde{\eta}} \tag{79}$$

Since we assumed ${}^{N}\!/_{\widetilde{N}} \geqslant {}^{\eta}\!/_{\tilde{\eta}}$, we also have $G_\mu \leqslant G_{\tilde{\mu}}$. In particular, this implies $\Delta(\mathcal{M}_G \| \widetilde{\mathcal{M}}_G) = 0$. Next, we apply the triangle inequality from Lemma B.3 and obtain:

$$\begin{aligned}
\Delta(\mathcal{M}^{\otimes N} \| \widetilde{\mathcal{M}}^{\otimes \widetilde{N}}) &\leqslant \Delta(\mathcal{M}^{\otimes N} \| \mathcal{M}_G) + \Delta(\mathcal{M}_G \| \widetilde{\mathcal{M}}^{\otimes \widetilde{N}}) && (80)\\
&\leqslant \Delta(\mathcal{M}^{\otimes N} \| \mathcal{M}_G) + \Delta(\mathcal{M}_G \| \widetilde{\mathcal{M}}_G) + \Delta(\widetilde{\mathcal{M}}_G \| \widetilde{\mathcal{M}}^{\otimes \widetilde{N}}) && (81)\\
&= \Delta(\mathcal{M}^{\otimes N} \| \mathcal{M}_G) + \Delta(\widetilde{\mathcal{M}}_G \| \widetilde{\mathcal{M}}^{\otimes \widetilde{N}}). && (82)
\end{aligned}$$

To bound the last two summands, we apply Theorem 5 in (Dong et al., 2022), which gives that for all $\alpha \in [0, 1]$:

$$\begin{aligned}
G_\mu(\alpha + \gamma) - \gamma \leqslant f^{\otimes N}(\alpha) \geqslant G_\mu(\alpha - \gamma) + \gamma, && (83)\\
G_{\tilde{\mu}}(\alpha + \tilde{\gamma}) - \tilde{\gamma} \leqslant \tilde{f}^{\otimes \widetilde{N}}(\alpha) \leqslant G_{\tilde{\mu}}(\alpha - \tilde{\gamma}) + \tilde{\gamma}, && (84)
\end{aligned}$$

where

$$\begin{aligned}
\gamma &= \frac{0.56 v_3}{\sqrt{N}(v_2 - v_1^2)^{3/2}}, && (85)\\
\tilde{\gamma} &= \frac{0.56 \tilde{v}_3}{\sqrt{\widetilde{N}}(\tilde{v}_2 - \tilde{v}_1^2)^{3/2}}. && (86)
\end{aligned}$$

In particular, applying a shift by $\tilde{\gamma}$ in the last inequality in Equation (84) gives for all $\alpha \in [0, 1]$:

$$G_\mu(\alpha + \gamma) - \gamma \leqslant f^{\otimes N}(\alpha) \;\text{ and }\; \tilde{f}^{\otimes \widetilde{N}}(\alpha + \tilde{\gamma}) - \tilde{\gamma} \leqslant G_{\tilde{\mu}}(\alpha). \tag{87}$$

Next, note the definition of the Δ-divergence via the infimum to see that the above implies:

$$\Delta(\mathcal{M}^{\otimes N} \| \mathcal{M}_G) \leqslant \gamma \;\text{ and }\; \Delta(\widetilde{\mathcal{M}}_G \| \widetilde{\mathcal{M}}^{\otimes \widetilde{N}}) \leqslant \tilde{\gamma}. \tag{88}$$

Moreover, we can write $\gamma, \tilde{\gamma}$ in terms of $\eta, \tilde{\eta}$ respectively:

$$\gamma = \frac{0.56 \eta^3 v_3}{\sqrt{N} v_1^3} \;\text{ and }\; \tilde{\gamma} = \frac{0.56 \tilde{\eta}^3 \tilde{v}_3}{\sqrt{\widetilde{N}} \tilde{v}_1^3}. \tag{89}$$

Thus, we have:

$$\Delta(\mathcal{M}^{\otimes N} \| \widetilde{\mathcal{M}}^{\otimes \widetilde{N}}) \leqslant \gamma + \tilde{\gamma} = 0.56 \left(\frac{\eta^3 v_3}{\sqrt{N} v_1^3} + \frac{\tilde{\eta}^3 \tilde{v}_3}{\sqrt{\widetilde{N}} \tilde{v}_1^3} \right). \tag{90}$$

For $N = \widetilde{N}$ our result above becomes:

$$\Delta(\mathcal{M}^{\otimes N} \| \widetilde{\mathcal{M}}^{\otimes \widetilde{N}}) \leqslant \frac{0.56}{\sqrt{N}} \left(\frac{\eta^3 v_3}{v_1^3} + \frac{\tilde{\eta}^3 \tilde{v}_3}{\tilde{v}_1^3} \right). \tag{91}$$

□