Gaussian Differential Privacy

Jinshuo Dong[*] Aaron Roth[†] Weijie J. Su[‡]

May 8, 2019

Abstract

Differential privacy has seen remarkable success as a rigorous and practical formalization of data privacy in the past decade. But it also has some well known weaknesses: notably, it does not tightly handle composition. This weakness has inspired several recent relaxations of differential privacy based on Renyi divergences. We propose an alternative relaxation of differential privacy, which we term "f-differential privacy", which has a number of appealing properties and avoids some of the difficulties associated with divergence based relaxations. First, it preserves the hypothesis testing interpretation of differential privacy, which makes its guarantees easily interpretable. It allows for lossless reasoning about composition and post-processing, and notably, a direct way to import existing tools from differential privacy, including privacy amplification by subsampling. We define a canonical single parameter family of definitions within our class which we call "Gaussian Differential Privacy", defined based on the hypothesis testing of two shifted Gaussian distributions. We show that this family is focal by proving a central limit theorem, which shows that the privacy guarantees of *any* hypothesis-testing based definition of privacy (including differential privacy) converges to Gaussian differential privacy in the limit under composition. We also prove a finite (Berry-Esseen style) version of the central limit theorem, which gives a useful tool for tractably analyzing the exact composition of potentially complicated expressions. We demonstrate the use of the tools we develop by giving an improved analysis of the privacy guarantees of noisy stochastic gradient descent.

[*] Department of Applied Mathematics and Computational Science, University of Pennsylvania. Email: jinshuo@sas.upenn.edu

[†] Department of Computer and Information Sciences, University of Pennsylvania. Email: aaroth@cis.upenn.edu.

[‡] Wharton Statistics Department, University of Pennsylvania. Email: suw@wharton.upenn.edu.

Contents

1 Introduction

Modern statistical analysis and machine learning are overwhelmingly applied to data concerning *people*. Valuable data sets generated from the personal devices and online behavior of billions of individuals contain data on location, web search histories, media consumption, physical activity, social networks, and more. This is on top of continuing large scale analysis of traditionally sensitive data records, including those collected by hospitals, schools, and the Census. This reality requires the development of tools to perform large scale data analysis in a way that still protects the *privacy* of individuals represented in the data.

Unfortunately, the history of data privacy for many years consisted of ad-hoc attempts at "anonymizing" personal information, followed by high profile de-anonymizations. This includes the release of AOL search logs, de-anonymized by the New York Times [BZ06], the Netflix Challenge dataset, de-anonymized by Narayanan and Shmatikov [NS08], the realization that participants in genome wide association studies could be identified from aggregate statistics such as minor allele frequencies that were publicly released [HSR$^+$08], and the reconstruction of individual level census records from aggregate statistical releases [Abo18].

Thus, a rigorous and principled privacy-preserving framework was in urgent need to prevent breaches of personal information in data analysis. In this context, *differential privacy* has put private data analysis on firm theoretical foundations [DMNS06, DKM$^+$06]. This definition has become tremendously successful: in addition to an enormous and growing academic literature, it has been adopted as a key privacy technology by Google [EPK14], Apple [App17], Microsoft [DKY17], and the US Census Bureau [Abo18]. The definition of this new concept involves privacy parameters $\varepsilon \geqslant 0$ and $0 \leqslant \delta \leqslant 1$.

Definition 1.1 ([DMNS06, DKM$^+$06])**.** *A randomized algorithm M that takes as input a dataset consisting of individuals is (ε, δ)-differentially private (DP) if for any pair of datasets S, S' that differ in the record of a single individual, and any event E,*

$$\mathbb{P}\left[M(S) \in E\right] \leqslant \mathrm{e}^{\varepsilon} \mathbb{P}\left[M(S') \in E\right] + \delta. \tag{1}$$

When $\delta = 0$, the guarantee is simply called ε-DP.

In this definition, datasets are *fixed* and the probabilities are taken *only* over the randomness of the mechanism[1]. In particular, the event E can take any measurable set in the range of M. To achieve differential privacy, a mechanism is necessarily randomized. Take as an example the problem of privately releasing the average cholesterol level of individuals in the dataset $S = (x_1, \ldots, x_n)$, each x_i corresponding to an individual. A privacy-preserving mechanism may take the form[2] $M(S) = \frac{1}{n}(x_1 + \cdots + x_n) +$ noise. The level of the noise term has to be sufficiently large to mask the *characteristics* of any individual's cholesterol level, while not being too large to distort the population average for accuracy purposes. Consequently, the probability distributions of $M(S)$ and $M(S')$ are close to each other for any datasets S, S' that differ only in one individual.

Alternatively, the definition of differential privacy can be formulated as a hypothesis testing problem. This useful perspective was first observed by [WZ10] and then further developed by [KOV17], which is a direct inspiration for our work. In short, consider the hypothesis testing problem

$$H_0 : S \ \text{ versus } \ H_1 : S'$$

[1] A randomized algorithm M is often referred to as a mechanism in the differential privacy literature

[2] Here we identify x_i with his/her cholesterol level for each individual.

and call Alice the only individual that is in S but not S'. As such, rejecting the null hypothesis corresponds to the detection of absence of Alice, whereas accepting the null hypothesis means to detect the presence of Alice in the dataset. Using the output of an (ε, δ)-DP mechanism, the power of any test at significance level $0 < \alpha < 1$ has an upper bound[3] of $\mathrm{e}^{\varepsilon}\alpha + \delta$. This bound is only slightly larger than α provided that ε, δ are small and, therefore, *any* test is essentially powerless. Put differently, differential privacy with small privacy parameters protects against any inferences of the presence of Alice, or any other individual, in the dataset.

However, there are good reasons to want to relax the original definition of differential privacy, which has lead to a long line of proposals for such relaxations. The most important shortcoming is that differential privacy does not tightly handle composition. Composition concerns how privacy guarantees degrade under repetition of mechanisms applied to the same dataset. Without composition properties, it would be near impossible to develop complex differentially private data analysis methods. Although it has been known since the original papers defining differential privacy [DMNS06, DKM+06] that the composition of an $(\varepsilon_1, \delta_1)$-DP algorithm M_1 and an $(\varepsilon_2, \delta_2)$-DP algorithm M_2 yields an $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-DP algorithm M, the corresponding upper bound $\mathrm{e}^{\varepsilon_1+\varepsilon_2}\alpha + \delta_1 + \delta_2$ no longer tightly characterizes the trade-off between significance level and type II error for the testing problem of distinguishing S from S'. In [DRV10], Dwork, Rothblum, and Vadhan gave an improved composition theorem, but one that still fails to capture the correct hypothesis testing trade-off. This is for a fundamental reason: (ε, δ)-differential privacy is misparameterized in the sense that the guarantees of the composition of $(\varepsilon_i, \delta_i)$-differentially private algorithms cannot be characterized by any pair of parameters (ε, δ). Worse, finding the parameters (ε, δ) that most tightly approximate the correct trade-off between significance level and power for a composition of a sequence of differentially private algorithms is computationally hard [MV16], and so in practice, one must resort to approximations. These are substantial drawbacks, given that composition and modularity are first order desiderata for a useful privacy definition — and ones that often continues to push practical algorithms with meaningful privacy guarantees out of reach.

In light of this, substantial recent effort has been devoted to developing relaxations of differential privacy for which composition can be handled exactly. This line of work includes several variants of "concentrated differential privacy" [DR16, BS16], "Renyi Differential Privacy" [Mir17], and "Truncated Concentrated Differential Privacy" [BDRS18]. These definitions are tailored to be able to exactly and easily track the "privacy cost" of compositions of the most basic primitive in differential privacy, which is the perturbation of a real valued statistic with Gaussian noise. While this direction has been quite fruitful, there are still several places one might wish for improvement.

First, these notions no longer have hypothesis testing interpretations, but are rather based on studying divergences that satisfy an information processing inequality. There are good reasons to prefer definitions based on hypothesis testing. Most immediately, hypothesis testing-based definitions provide an easy way to interpret the guarantees of a privacy definition. However, more fundamentally, Blackwell's theorem provides a formal sense in which a tight understanding of the trade-off between type I and type II errors for the hypothesis testing problem of distinguishing between $M(S)$ and $M(S')$ contains only more information than any divergence between the distributions $M(S)$ and $M(S')$ (so long as the divergence satisfies an information processing inequality).

Second, certain simple, fundamental primitives associated with differential privacy — most notably, *privacy amplification by subsampling* [KLN+11] — either fail to apply to these relaxations, or require a substantially complex analysis [WBK18]. This is especially problematic when analyzing

[3] A more precise bound is given in Proposition 2.5.

stochastic gradient descent — arguably the most important learning/optimization algorithm — and what necessitated Abadi et al. [ACG$^+$16] to develop the numerical *moments accountant* method to sidestep the issue.

1.1 Our Contributions

In this work, we introduce a new relaxation of differential privacy that avoids these issues, and has other attractive properties. Rather than giving a "divergence" based relaxation of differential privacy, we start from the hypothesis testing interpretation of differential privacy, and generalize it by allowing the trade-offs between type I and type II errors in the simple hypothesis testing problem to be governed by an nonlinear function f. This general class of definitions which we term f-differential privacy — and which captures differential privacy as a special case — is robust to post-processing by construction, and we develop a simple general calculus to reason about composition. Here we briefly summarize our contributions:

1. We show that our privacy definition is *closed under composition*, which means that the exact trade-off function between type I and type II errors that results from the composition of an f_1-DP mechanism with an f_2-DP mechanism can always be *exactly* described by some function f_3. This is what makes it possible to reason losslessly about composition. This is in contrast to (ε, δ)-differential privacy or any other privacy definition that artificially restricts itself to a small number of parameters: by allowing for a *function* to keep track of the privacy guarantee of the mechanism, we avoid the pitfall of premature summarization[4].

2. We show a general duality between f-DP and infinite collections of (ε, δ)-DP guarantees. This duality is useful in two ways. First, it allows one to analyze an algorithm in the framework of f-DP, and then convert back to an (ε, δ)-privacy guarantee at the end, if desired. More fundamentally, this duality provides a means to import techniques developed for (ε, δ)-DP to the framework of f-DP. As an example of the power of this method, we use this duality to show how to reason simply about privacy amplification by subsampling for f-differential privacy, by leveraging existing results for (ε, δ)-differential privacy. This is in contrast to divergence based notions of privacy, in which reasoning about amplification by subsampling is difficult.

3. We define a particular, single-parameter family of functions G_μ that have especially nice properties. G_μ characterizes the type I and type II error trade-off for the hypothesis testing problem of distinguishing $\mathcal{N}(0,1)$ from $\mathcal{N}(\mu,1)$. We term f-differentially private mechanisms for $f = G_\mu$ "μ-Gaussian Differentially Private Mechanisms", or say that they satisfy μ-GDP. Like the family of "concentrated differential privacy" relaxations, GDP admits a simple additive composition rule that is exact for methods that perturb statistics with Gaussian noise.

4. We show that Gaussian differential privacy is a "canonical" privacy guarantee in a fundamental sense: we prove a central limit theorem that shows that for *any* hypothesis testing-based definition of differential privacy (i.e. f-DP for any f), in the limit under composition, the trade-off function characterizing the privacy guarantee converges to G_μ for some μ. In other words, in the limit under composition, ε-differential privacy, (ε, δ)-differential privacy, or

[4] To quote Susan Holmes, "Premature summarization is the root of all evil in statistics".

any other variant that retains a hypothesis testing interpretation converges to the guarantees of GDP. An implication of this is that GDP is the only simply-parameterized family of hypothesis-testing based notions of privacy that can tightly track composition.

5. We also show a finite convergence version of our central limit theorem (i.e. a "Berry-Esseen" type theorem) that establishes a $1/\sqrt{k}$ rate of convergence to GDP for compositions of k f-DP mechanisms. This is useful as a practical analytical tool, since for some f, exact computation of the f-DP guarantee under composition can be intractable. In these cases, one can instead analyze the algorithm using the simple parameters of the central limit theorem, and obtain upper bounds on the error of the approximation, which rapidly tends to 0 with the number of compositions. We find that the simple bounds given by our central limit theorems approximate the exact trade-off function extremely closely: see Figure 1.

6. Finally, we use this collection of tools to give a substantially sharper analysis of the privacy guarantees of noisy stochastic gradient descent, improving on previous special-purpose analyses that reasoned about divergences rather than directly about hypothesis testing [ACG+16].

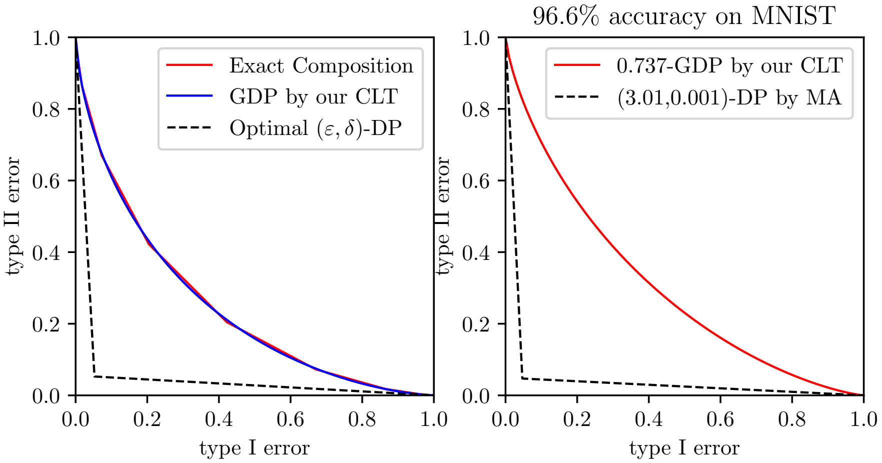

Figure 1: Left: Our central limit theorem approximation is almost perfect, even for a composition of just 10 ε-differentially private mechanisms. In contrast, the tightest possible approximation via an (ε, δ)-differential privacy is substantially looser. See Section 3 for details. Right: When we use our central limit theorem to analyze private stochastic gradient descent used to train a convolutional neural network on MNIST, we obtain a substantially tighter understanding of the privacy guarantee, compared to the best (ε, δ)-privacy guarantee that can be obtained by the "moment accountant" (MA) method developed in [ACG+16]. See Section 5 for more plots and details.

2 f-Differential Privacy and Its Basic Properties

In Section 2.1, we give a formal definition of f-differential privacy (f-DP). Section 2.2 introduces a special case of f-DP that we refer to as *Gaussian differential privacy* (GDP). In Section 2.3, we highlight some appealing properties of this new privacy notation from an information-theoretic perspective. Next, Section 2.4 offers a profound connection between f-DP and (ε, δ)-DP. Finally, we discuss the group privacy properties of f-DP.

Before moving on, we first establish several key pieces of notation from the differential privacy literature.

- **Dataset.** A dataset S is a collection of n records, each corresponding to an individual. Formally, we write the dataset as $S = (x_1, \ldots, x_n)$, and an individual $x_i \in X$ for some abstract space X. Two datasets $S' = (x'_1, \ldots, x'_n)$ and S are said to be neighbors if they differ in exactly one record, that is, there exists an index j such that $x_i = x'_i$ for all $i \neq j$ and $x_j \neq x'_j$.
- **Mechanism.** A mechanism M refers to a randomized algorithm that takes as input a dataset S and releases some (randomized) statistics $M(S)$ of the dataset in some abstract space Y. For example, a mechanism can release the average salary of individuals in the dataset plus some random noise.

2.1 Trade-off Functions and f-DP

All variants of differential privacy informally require that it be hard to *distinguish* any pairs of *neighboring* datasets based on the information released by a private a mechanism M. From the (imaginary) attacker's perspective, it is natural to formalize this notion of "indistinguishability" as a hypothesis testing problem for two neighboring datasets S and S':

$$H_0: \text{the underlying dataset is } S \quad \text{vs} \quad H_1: \text{the underlying dataset is } S'.$$

The output of the mechanism M serves as the basis for performing the hypothesis testing problem. Denote by P and Q the probability distributions of the mechanism applied to the two datasets, namely $M(S)$ and $M(S')$, respectively. The fundamental difficulty in distinguishing the two hypotheses is best delineated by the *optimal* trade-off between the achievable type I and type II errors. More precisely, consider a rejection rule $0 \leqslant \phi \leqslant 1$, with type I and type II error rates defined as[5]

$$\alpha_\phi = \mathbb{E}_P[\phi], \quad \beta_\phi = 1 - \mathbb{E}_Q[\phi],$$

respectively. The two errors satisfy, for example, the constraint

$$\alpha_\phi + \beta_\phi \geqslant 1 - \mathrm{TV}(P, Q), \tag{2}$$

where the total variation distance $\mathrm{TV}(P, Q)$ is the supremum of $|P(A) - Q(A)|$ over all measurable sets A. Instead of this rough constraint, we seek to characterize the fine-grained *trade-off* between the two errors. Explicitly, fixing the type I error at *any* level, we consider the minimal achievable type II error. This motivates the following definition.

Definition 2.1 (trade-off function)**.** *For any two probability distributions P and Q on the same space, define the trade-off function $T(P, Q) : [0, 1] \to [0, 1]$ as*

$$T(P, Q)(\alpha) = \inf \left\{ \beta_\phi : \alpha_\phi \leqslant \alpha \right\},$$

where the infimum is taken over all (measurable) rejection rules.

[5] If $\phi(\omega) = 1$, then we deterministically reject the null; if $\phi(\omega) = 0$, then we deterministically accept the null; if $\phi(\omega) = p \in (0, 1)$, then we flip a coin and reject the null with probablity p.

The trade-off function serves as a clear-cut boundary of the achievable and unachievable regions of type I and type II errors, rendering itself the *complete* characterization of the fundamental difficulty in testing between the two hypothesis. In particular, the greater this function is, the harder it is to distinguish the two distributions. For completeness, we remark that the minimal β_ϕ can be achieved by the likelihood ratio test—a fundamental result known as the Neyman-Pearson lemma, which we state in the appendix as Theorem A.1.

A function is called a trade-off function if it is equal to $T(P, Q)$ for some distributions P and Q. Below we give a necessary and sufficient condition for f to be a trade-off function. This characterization reveals, for example, that $\max\{f, g\}$ is a trade-off function if both f and g are trade-off functions.

Proposition 2.2. *A function $f : [0, 1] \to [0, 1]$ is a trade-off function if and only if f is convex, continuous*[6]*, non-increasing and $f(x) \leqslant 1 - x$ for $x \in [0, 1]$.*

Now, we propose a new generalization of differential privacy built on top of trade-off functions. Below, we write $g \geqslant f$ for two functions defined on $[0, 1]$ if $g(x) \geqslant f(x)$ for all $0 \leqslant x \leqslant 1$, and we abuse notation by identifying $M(S)$ and $M(S')$ with their corresponding probability distributions. Note that if $T(P, Q) \geqslant T(\widetilde{P}, \widetilde{Q})$, then in a very strong sense, P and Q are harder to distinguish than $\widetilde{P}$ and $\widetilde{Q}$ at *any* level of type I error.

Definition 2.3 (f-differentially privacy)**.** *Let f be a trade-off function. A mechanism M is said to be f-differentially private if*

$$T\big(M(S), M(S')\big) \geqslant f$$

for all neighboring datasets S and S'.

A graphical illustration of this definition is shown in Figure 2. Letting P and Q be the distributions such that $f = T(P, Q)$, this privacy definition amounts to saying that a mechanism is f-DP if distinguishing any two neighboring datasets based on the released information is at least as difficult as distinguishing P and Q based on a single draw. In contrast to existing definitions of differential privacy, our new definition is parameterized by a function, as opposed to several real valued parameters (e.g. ε and δ). This functional perspective offers a complete characterization of "privacy", thereby avoiding the pitfall of summarizing statistical information too early (hence losing information subsequently). This fact is crucial to the development of a composition theorem for f-DP in Section 3. Although this completeness comes at the cost of increased complexity, as we will see in Section 2.2, a simple family of trade-off functions can often closely capture privacy loss in many scenarios.

Naturally, the definition of f-DP is symmetric in the same sense as the neighboring relationship is symmetric. Observe that this privacy notion also requires

$$T\big(M(S'), M(S)\big) \geqslant f$$

for any neighboring pair S, S'. Therefore, it is desirable to restrict our attention to "symmetric" trade-off functions f. Proposition 2.4 shows that this is without loss of generality.

[6] Convexity itself implies continuity in $(0, 1)$ for f. In addition, $f(\alpha) \geqslant 0$ and $f(\alpha) \leqslant 1 - \alpha$ implies continuity at 1. Hence, the continuity condition only matters at $x = 0$.

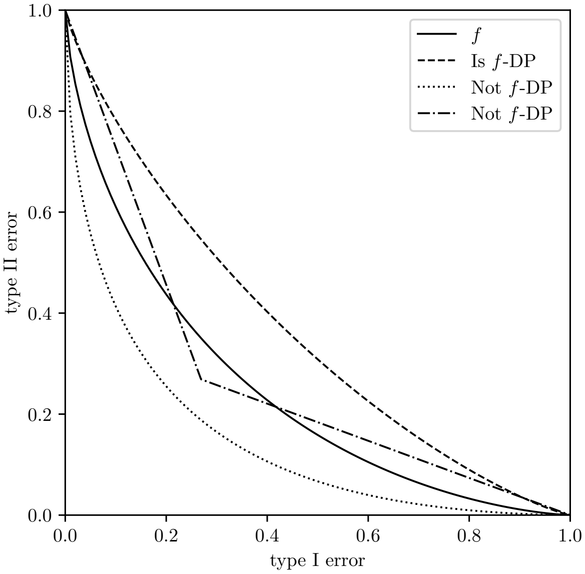

Figure 2: Three different examples of $T\big(M(S), M(S')\big)$. Only the dashed line corresponds to a trade-off function satisfying f-DP.

Proposition 2.4. *Let a mechanism M be f-DP. Then, M is f^{S}-DP with $f^{\mathrm{S}} = \max\{f, f^{-1}\}$, where the inverse function is defined as*[7]

$$f^{-1}(\alpha) := \inf\{t \in [0,1] : f(t) \leqslant \alpha\} \tag{3}$$

for $\alpha \in [0,1]$.

When we write $f = T(P, Q)$, we can express the inverse as $f^{-1} = T(Q, P)$ and hence observe that it is also a trade-off function. As a consequence of this, f^{S} continues to be a trade-off function by making use of Proposition 2.2 and, moreover, is *symmetric* in the sense that $f^{\mathrm{S}} = (f^{\mathrm{S}})^{-1}$. Importantly, this symmetrization gives a tighter bound in the privacy definition since $f^{\mathrm{S}} \geqslant f$. In the remainder of the paper, therefore, trade-off functions will always be assumed to be symmetric unless otherwise specified. We prove Proposition 2.4 in Appendix A.

We conclude this subsection by showing that f-DP is a generalization of (ε, δ)-DP. (Forshadowing, a deeper connection between f-DP and (ε, δ)-DP will be discussed in Section 2.4). Denote the trade-off function

$$f_{\varepsilon,\delta}(\alpha) = \max\big\{0, 1 - \delta - \mathrm{e}^{\varepsilon}\alpha, \mathrm{e}^{-\varepsilon}(1 - \delta - \alpha)\big\} \tag{4}$$

for $0 \leqslant \alpha \leqslant 1$. Figure 3 shows the graph of this function and its evident symmetry. The following result is adapted from [WZ10].

Proposition 2.5 ([WZ10])**.** *A mechanism M is (ε, δ)-DP if and only if M is $f_{\varepsilon,\delta}$-DP.*

[7] Equation (3) is the standard definition of the left-continuous inverse of a decreasing function. When f is strictly decreasing and $f(0) = 1$ and hence bijective as a mapping, (3) corresponds to the inverse function in the ordinary sense, i.e. $f(f^{-1}(x)) = f^{-1}(f(x)) = x$. However, this is not true in general.

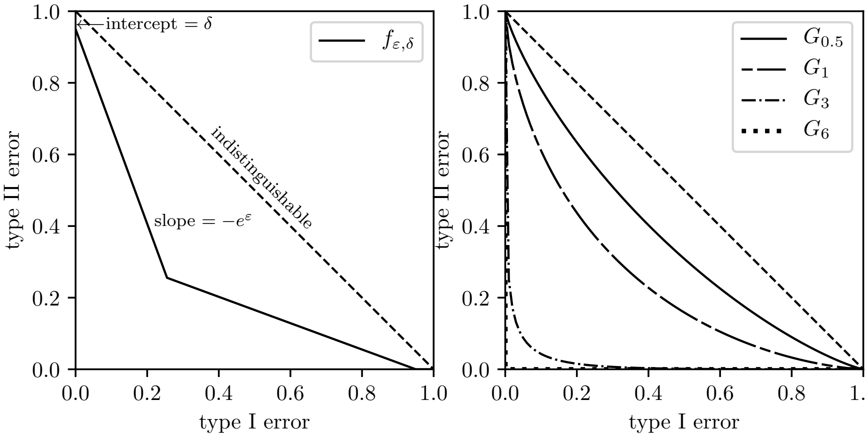

Figure 3: Left: $f_{\varepsilon,\delta}$ is a "symmetrized" linear function, with slope $-e^{\pm\varepsilon}$ and intercept $1-\delta$. Right: Trade-off functions of Gaussian distributions with different parameters. $\mu = 0.5$ is reasonably private, $\mu = 1$ is borderline private. $\mu = 3$ is only marginally private: an adversary can simultaneously obtain type I and type II errors of only 0.07. At $\mu = 6$ (hard to see on the plot) these errors are driven down to 0.001.

2.2 Gaussian Differential Privacy

This subsection introduces a parametric family of f-DP guarantees, where f is the trade-off function of two normal distributions. We refer to this specialization as Gaussian differential privacy (GDP). GDP enjoys many desirable properties that lead to its central role in this paper. Among others, we can now precisely define the trade-off function with a single parameter. To define this notion, let

$$G_\mu := T\big(\mathcal{N}(0,1), \mathcal{N}(\mu,1)\big)$$

for $\mu \geqslant 0$. An explicit expression for the trade-off function G_μ reads

$$G_\mu(\alpha) = \Phi\big(\Phi^{-1}(1-\alpha) - \mu\big), \tag{5}$$

where Φ denotes the standard normal CDF. For completeness, we provide a proof of (5) in Appendix A. This trade-off function is decreasing in μ in the sense that $G_\mu \leqslant G_{\mu'}$ if $\mu \geqslant \mu'$. We now define Gaussian Differential Privacy:

Definition 2.6. *A mechanism M is said to satisfy μ-Gaussian Differential Privacy (GDP) if it is G_μ-DP. That is,*

$$T\big(M(S), M(S')\big) \geqslant G_\mu$$

for all neighboring datasets S and S'.

GDP has several attractive properties. First, this privacy definition is fully described by the single mean parameter of a unit-variance Gaussian distribution, which makes it easy to describe and interpret. For instance, one can see from the right panel of Figure 3, $\mu \leqslant 0.5$ guarantees a reasonable amount of privacy, whereas if $\mu \geqslant 6$, almost nothing is being promised. Second, loosely speaking, GDP occupies a role among all hypothesis testing-based notions of privacy that is similar to the role that the Gaussian distribution has among general probability distributions.

We formalize this important point by proving central limit theorems for f-DP in Section 3, which, roughly speaking, says that f-DP converges to GDP under composition in the limit. Lastly, as shown in the remainder of this subsection, GDP *precisely* characterizes the Gaussian mechanism, one of the most fundamental building blocks of differential privacy.

Consider the problem of privately releasing a univariate statistic $\theta(S)$ of the dataset S. Define the sensitivity of θ as

$$\mathrm{sens}(\theta) = \sup_{S,S'} |\theta(S) - \theta(S')|,$$

where the supremum is over all neighboring datasets. The Gaussian mechanism adds Gaussian noise to the statistic θ in order to obscure whether θ is computed on S or S'. The following result shows that the Gaussian mechanism with noise properly scaled to the sensitivity of the statistic satisfies GDP.

Theorem 2.7. *Define the Gaussian mechanism that operates on a statistic θ as $M(S) = \theta(S) + \xi$, where $\xi \sim \mathcal{N}(0, \mathrm{sens}(\theta)^2/\mu^2)$. Then, M is μ-GDP.*

Proof of Theorem 2.7. Recognizing that $M(S), M(S')$ are normally distributed with means $\theta(S), \theta(S')$, respectively, and common variance $\sigma^2 = \mathrm{sens}(\theta)^2/\mu^2$, we get

$$T\big(M(S), M(S')\big) = T\big(\mathcal{N}(\theta(S), \sigma^2), \mathcal{N}(\theta(S'), \sigma^2)\big) = G_{|\theta(S)-\theta(S')|/\sigma}.$$

By the definition of sensitivity, $|\theta(S) - \theta(S')|/\sigma \leqslant \mathrm{sens}(\theta)/\sigma = \mu$. Therefore, we get

$$T\big(M(S), M(S')\big) = G_{|\theta(S)-\theta(S')|/\sigma} \geqslant G_\mu.$$

This completes the proof. □

As implied by the proof above, GDP offers the tightest possible privacy bound of the Gaussian mechanism. More precisely, the Gaussian mechanism in Theorem 2.7 satisfies

$$G_\mu(\alpha) = \inf_{\text{neighboring } S,S'} T\big(M(S), M(S')\big)(\alpha), \tag{6}$$

where the infimum is (asymptotically) achieved at the two neighboring datasets such that $|\theta(S) - \theta(S')| = \mathrm{sens}(\theta)$ *irrespective* of the type I error α. As such, the characterization by GDP is precise in the pointwise sense. In contrast, the right-hand side of (6) in general is not necessarily a convex function of α and, in such case, is not a trade-off function according to Proposition 2.2. This nice property of Gaussian mechanism is related to the log-concavity of Gaussian distributions. See Proposition A.3 for a detailed treatment of log-concave distributions.

2.3 Post-Processing and the Informativeness of Trade-off Functions

Intuitively, a data analyst cannot make a statistical analysis more disclosive only by processing the output of the mechanism M. This is called the post-processing property, a natural requirement that any notion of privacy, including our definition of f-DP, should satisfy.

To formalize this point for f-DP, denote by $\mathrm{Proc}: Y \to Z$ a (randomized) algorithm that maps the input $M(S) \in Y$ to some space Z, yielding a new mechanism that we denote by $\mathrm{Proc} \circ M$. The following result confirms the post-processing property of f-DP.

Proposition 2.8. *If a mechanism M is f-DP, then its post-processing* $\mathrm{Proc} \circ M$ *is also f-DP.*

Proposition 2.8 is a consequence of the following lemma. Let $\mathrm{Proc}(P)$ be the probability distribution of $\mathrm{Proc}(\zeta)$ with ζ drawn from P. Define $\mathrm{Proc}(Q)$ likewise.

Lemma 2.9. *For any two distributions P and Q, we have*

$$T\big(\mathrm{Proc}(P), \mathrm{Proc}(Q)\big) \geqslant T(P, Q).$$

This lemma means that post-processed distributions can only become more difficult to tell apart than the original distributions from the perspective of trade-off functions. While the same property holds for many divergence-based measures of indistinguishability such as the Rényi divergences used by the concentrated differential privacy family of definitions [DR16, BS16, Mir17, BDRS18], a consequence of the following theorem is that trade-off functions offer a finer-grained accounting. This remarkable inverse of Lemma 2.9 is due to Blackwell (see also Theorem 2.5 in [KOV17]).

Theorem 2.10 ([Bla50], Theorem 10). *Let P, Q be probability distributions on Y and P', Q' be probability distributions on Z. The following two statements are equivalent:*

(a) $T(P, Q) \leqslant T(P', Q')$.

(b) There exists a randomized algorithm $\mathrm{Proc} : Y \to Z$ *such that* $\mathrm{Proc}(P) = P', \mathrm{Proc}(Q) = Q'$.

To appreciate the consequence of this theorem, we begin by pointing out that post-processing imposes an ordering[8] over pairs of distributions called the Blackwell ordering (see e.g. [Rag11]). Specifically, we say $(P, Q) \preceq_{\mathrm{Blackwell}} (P', Q')$ if the above condition (b) holds. Taking the set $L_{\preceq_{\mathrm{Blackwell}}}$ to be the collection of all such inequalities, we get a set-theoretic description of the Blackwell ordering.

Informally, any quantitative measure of differential privacy, abstractly denoted by desc, also induces an ordering $\preceq_{\mathrm{desc}}$, and hence a set $L_{\preceq_{\mathrm{desc}}}$. For example, when we use trade-off functions as in f-DP, we say $(P, Q) \preceq_{\mathrm{tradeoff}} (P', Q')$ if $T(P, Q) \leqslant T(P', Q')$. (ε, δ)-differential privacy and Renyi-based notions of privacy similarly induce orderings on pairs of distributions. The trivial direction $(a) \Rightarrow (b)$ in Blackwell's theorem can be succinctly expressed as $L_{\preceq_{\mathrm{tradeoff}}} \supseteq L_{\preceq_{\mathrm{Blackwell}}}$. More generally, any notion of privacy that is robust to post-processing must satisfy that $L_{\preceq_{\mathrm{desc}}} \supseteq L_{\preceq_{\mathrm{Blackwell}}}$. The non-trivial direction of Blackwell's theorem tells us that f-DP offers the most refined accounting of post-processing, in the sense that $L_{\preceq_{\mathrm{tradeoff}}} = L_{\preceq_{\mathrm{Blackwell}}}$. This is not true for the ordering on distributions induced by Renyi-divergences or (ε, δ)-indistinguishability (used to define (ε, δ)-differential privacy).

2.4 A Primal-Dual Perspective

In this subsection, we show that f-DP is equivalent to an infinite *collection* of (ε, δ)-DP guarantees via the convex conjugate of the trade-off function. As a consequence of this, we can view f-DP as the *primal* privacy representation and, accordingly, its *dual* representation is the collection of (ε, δ)-DP guarantees. Taking this powerful viewpoint, many results from the large body of differential privacy work can be carried over to f-DP in a seamless fashion. In particular, this primal-dual perspective is crucial to our analysis of "privacy amplification by subsampling" in Section 4. All proofs are deferred to Appendix A.

First, we present the result that converts a collection of (ε, δ)-DP guarantees into an f-DP guarantee.

[8] Not a partial ordering.

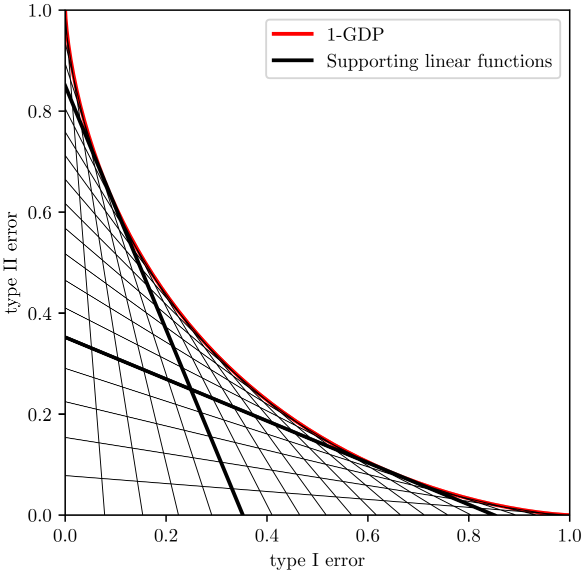

Figure 4: Each $(\varepsilon, \delta(\varepsilon))$-differential privacy guarantee corresponds to two supporting linear functions (symmetric to each other) to the trade-off function describing the complete f-DP guarantee. In general, describing a privacy guarantee using only a subset of (ε, δ)-differential privacy guarantees (for example, only those with small δ) is lossy.

Proposition 2.11 (Dual to Primal)**.** *Let I be an arbitrary index set such that each $i \in I$ is associated with $\varepsilon_i \in [0, \infty)$ and $\delta_i \in [0, 1]$. A mechanism is $(\varepsilon_i, \delta_i)$-DP for all $i \in I$ if and only if it is f-DP with*

$$f = \sup_{i \in I} f_{\varepsilon_i, \delta_i}.$$

This proposition follows easily from the equivalence of (ε, δ)-DP and $f_{\varepsilon,\delta}$-DP. We remark that the function f constructed above remains a trade-off function.

The more interesting direction is to convert f-DP into a collection of (ε, δ)-DP guarantees. Recall that the convex conjugate of a function g defined on $(-\infty, \infty)$ is defined as

$$g^*(y) = \sup_{-\infty < x < \infty} yx - g(x). \tag{7}$$

To define the conjugate of a trade-off function f, we extend its domain by setting $f(x) = \infty$ for $x < 0$ and $x > 1$. With this adjustment, the supremum is effectively taken over $0 \leqslant x \leqslant 1$.

Proposition 2.12 (Primal to Dual)**.** *For a symmetric trade-off function f, a mechanism is f-DP if and only if it is $\big(\varepsilon, \delta(\varepsilon)\big)$-DP for all $\varepsilon \geqslant 0$ with $\delta(\varepsilon) = 1 + f^*(-\mathrm{e}^{\varepsilon})$.*

For example, taking $f = G_\mu$, the following corollary provides a lossless conversion from GDP to a collection of (ε, δ)-DP guarantees. This conversion is exact and, therefore, any other (ε, δ)-DP guarantee derived for the Gaussian mechanism is implied by this corollary.

Corollary 2.13. *A mechanism is μ-GDP if and only if it is $\big(\varepsilon, \delta(\varepsilon)\big)$-DP for all $\varepsilon \geqslant 0$, where*

$$\delta(\varepsilon) = 1 - \mathrm{e}^{\varepsilon}\Phi\Big(-\frac{\varepsilon}{\mu} - \frac{\mu}{2}\Big) - \Phi\Big(\frac{\varepsilon}{\mu} - \frac{\mu}{2}\Big).$$

The primal-dual perspective provides a useful tool through which we can bridge the two privacy definitions. In some cases, it is easier to work with f-DP by leveraging the interpretation and informativeness of trade-off functions, as seen from the development of composition theorems for f-DP in Section 3. Meanwhile, (ε, δ)-DP is more convenient to work with in the cases where the lower complexity of two parameters ε, δ is helpful, for example, in the proof of the privacy amplification by subsampling theorem for f-DP. In short, our approach in Section 4 is to first work in the dual world and use existing subsampling theorems for (ε, δ)-DP, and then convert the results back to f-DP using a slightly more advanced version of Proposition 2.12.

2.5 Group Privacy

The notion of f-DP can be extended to address privacy of a *group* of individuals, and a question of interest is to quantify how privacy degrades as the group size grows. To setup the notation, we say that two datasets S, S' are k-neighbors (where $k \geqslant 2$ is an integer) if there exist datasets $S = S_0, S_1, \ldots, S_k = S'$ such that S_i and S_{i+1} are neighboring or identical for all $i = 0, \ldots, k-1$. Equivalently, S, S' are k-neighbors if they differ by at most k individuals. Accordingly, a mechanism M is said to be f-DP for *groups of size* k if

$$T\big(M(S), M(S')\big) \geqslant f$$

for all k-neighbors S and S'.

In the following theorem, we use $h^{\circ k}$ to denote the k-fold iterative composition of a function h. For example, $h^{\circ 1} = h$ and $h^{\circ 2}(x) = h(h(x))$.

Theorem 2.14. *If a mechanism is f-DP, then it is $\big[1 - (1-f)^{\circ k}\big]$-DP for groups of size k. In particular, if a mechanism is μ-GDP, then it is $k\mu$-GDP for groups of size k.*

For completeness, $1 - (1-f)^{\circ k}$ is a trade-off function and, moreover, remains symmetric if f is symmetric. These two facts and Theorem 2.14 are proved in Appendix A. As revealed in the proof, the privacy bound $1 - (1-f)^{\circ k}$ in general cannot be improved, thereby showing that the group operation in the f-DP framework is *closed* and *tight*. In addition, it is easy to see that $1 - (1-f)^{\circ k} \leqslant 1 - (1-f)^{\circ (k-1)}$ by recognizing that the trade-off function f satisfies $1 - f(x) \geqslant x$. This is consistent with the intuition that detecting changes in groups of k individuals becomes easier as the group size increases.

As an interesting consequence of Theorem 2.14, the group privacy of ε-DP in the limit corresponds to the trade-off function of two Laplace distributions. Recall that the density of $\mathrm{Lap}(\mu, b)$ is $\frac{1}{2b}\mathrm{e}^{-|x-\mu|/b}$.

Proposition 2.15. *Fix $\mu \geqslant 0$ and set $\varepsilon = \mu/k$. As $k \to \infty$, we have*

$$1 - (1 - f_{\varepsilon,0})^{\circ k} \to T\big(\mathrm{Lap}(0,1), \mathrm{Lap}(\mu, 1)\big).$$

The convergence is uniform over $[0, 1]$.

Two remarks are in order. First, $T\big(\mathrm{Lap}(0,1), \mathrm{Lap}(\mu,1)\big)$ is not equal to $f_{\varepsilon,\delta}$ for any ε, δ and. Therefore, (ε,δ)-DP is not expressive enough to measure privacy under the group operation. Second, the approximation in this theorem is very accurate even for small k. For example, for $\mu = 1, k = 4$, the function $1 - (1 - f_{\varepsilon,0})^{\circ k}$ is within 0.005 of $T\big(\mathrm{Lap}(0,1), \mathrm{Lap}(\mu,1)\big)$ uniformly over $[0,1]$. The proof of Proposition 2.15 is deferred to Appendix A.

3 Composition and Limit Theorems

Imagine that an analyst performs a sequence of analyses on a private dataset, in which each analysis is informed by prior analyses on the same dataset. Provided that every analysis alone is private, the question is whether all analyses collectively are private, and if so, how the privacy degrades as the number of analyses increases, namely under composition. It is essential for a notion of privacy to gracefully handle composition, without which the privacy analysis of complex algorithms would be almost impossible.

Now, we describe the composition of two mechanisms. In this section, the number of individuals in the dataset is irrelevant, so we drop the superscript and use X as the abstract notation for the space of datasets and abuse the notation n for the number of mechanisms in composition. As will be clear later, the use of n is consistent with the literature on central limit theorems. Let $M_1 : X \to Y_1$ be the first mechanism and $M_2 : X \times Y_1 \to Y_2$ be the second mechanism. In brief, M_2 takes as input the output of the first mechanism M_1 in addition to the dataset. With the two mechanisms in place, the joint mechanism $M : X \to Y_1 \times Y_2$ is defined as

$$M(S) = (y_1, M_2(S, y_1)), \tag{8}$$

where $y_1 = M_1(S)$.[9] Roughly speaking, the distribution of $M(S)$ is constructed from the marginal distribution of $M_1(S)$ on Y_1 and the conditional distribution of $M_2(S, y_1)$ on Y_2 given $M_1(S) = y_1$. The composition of more than two mechanisms follows recursively. In general, given a sequence of mechanisms $M_i : X \times Y_1 \times \cdots \times Y_{i-1} \to Y_i$ for $i = 1, 2, \ldots, n$, we can recursively define the joint mechanism as their composition:

$$M : X \to Y_1 \times \cdots \times Y_n.$$

Put differently, $M(x)$ can be interpreted as the trajectory of a Markov chain whose initial distribution is given by $M_1(x)$ and transition kernel $M_i(x, \cdots)$ at each step.

Using the language above, the goal of this section is to relate the privacy loss of M to that of the n mechanisms $M_1, \ldots, M_n$ in the f-DP framework. In short, Section 3.1 develops a general composition theorem for f-DP. In Sections 3.2, we identify a central limit theorem phenomenon of composition in the f-DP framework, which can be used as an approximation tool, just like we use the central limit theorem for random variables. This approximation is extended to and improved for (ε,δ)-DP in Section 3.3.

[9] Alternatively, we can write $M(S) = (M_1(S), M_2(S, M_1(S)))$, in which case it is necessary to specify that M_1 should be run only once in this expression.

3.1 A General Composition Theorem

The main thrust of this subsection is to demonstrate that the composition of private mechanisms is *closed* and *tight*[10] in the f-DP framework. This result is formally stated in Theorem 3.2, which shows that the composition mechanism remains f-DP with the trade-off function taking the form of a certain product. To define the product, consider two trade-off functions f and g that are given as $f = T(P, Q)$ and $g = T(P', Q')$ for some probability distributions P, P', Q, Q'.

Definition 3.1. *The tensor product of two trade-off functions $f = T(P, Q)$ and $g = T(P', Q')$ is defined as*

$$f \otimes g := T(P \times P', Q \times Q').$$

This term gets its name because it is defined via tensor products of distributions. Throughout the paper, write $f \otimes g(\alpha)$ for $(f \otimes g)(\alpha)$, and denote by $f^{\otimes n}$ the n-fold tensor product of f. The latter requires associativity, which we soon illustrate.

By definition, $f \otimes g$ is also a trade-off function. Nevertheless, it remains to be shown that the tensor product is well-defined: that is, that the definition is independent of the choice of distributions used to represent a trade-off function. More precisely, assuming $f = T(P, Q) = T(\tilde{P}, \tilde{Q})$ for some distributions $\tilde{P}, \tilde{Q}$, we need to ensure that

$$T(P \times P', Q \times Q') = T(\tilde{P} \times P', \tilde{Q} \times Q').$$

We defer the proof of this intuitive fact to Appendix C. Below we list some other useful properties[11] of the tensor product of trade-off functions, whose proofs are placed in Appendix D.

1. The product $\otimes$ is commutative and associative.
2. If $g_1 \geqslant g_2$, then $f \otimes g_1 \geqslant f \otimes g_2$.
3. $f \otimes \mathrm{Id} = \mathrm{Id} \otimes f = f$ where $\mathrm{Id} : [0,1] \to [0,1], \mathrm{Id}(x) = 1 - x$.
4. $(f \otimes g)^{-1} = f^{-1} \otimes g^{-1}$. See the definition of inverse in (3).

Property 3 explains why we name the function $x \mapsto 1 - x$ as Id. Property 4 implies that when f, g are symmetric trade-off functions, their product $f \otimes g$ is also symmetric.

Now we state the main theorem of this subsection. Its proof is given in Appendix C.

Theorem 3.2. *Let $M_i(\cdot, y_1, \cdots, y_{i-1})$ be f_i-DP for all $y_1 \in Y_1, \ldots, y_{i-1} \in Y_{i-1}$. Then the n-fold composition mechanism $M : X \to Y_1 \times \cdots \times Y_n$ is $f_1 \otimes \cdots \otimes f_n$-DP.*

This theorem shows that the composition of mechanisms remains f-DP or, put differently, composition is *closed* in the f-DP framework. Moreover, the privacy bound $f_1 \otimes \cdots \otimes f_n$ in Theorem 3.2 is *tight* in the sense that it cannot be improved in general. To see this point, consider the case where the second mechanism completely ignores the output of the first mechanism. In that case, the composition obeys

$$\begin{aligned} T\big(M(S), M(S')\big) &= T\big(M_1(S) \times M_2(S), M_1(S') \times M_2(S')\big) \\ &= T\big(M_1(S), M_1(S')\big) \otimes T\big(M_2(S), M_2(S')\big). \end{aligned}$$

[10] Section 2.5 showed that f-DP was “closed and tight” in a similar sense, with respect to the guarantees of group privacy.

[11] Properties 1,3 and 5 make the class of trade-off functions a *commutative monoid.* Informally, a monoid is a group without an inverse.

Next, taking neighboring datasets such that $T\big(M_1(S), M_1(S')\big) = f_1$ and $T\big(M_2(S), M_2(S')\big) = f_2$, one concludes that $f_1 \otimes f_2$ is the tightest possible bound on the two-fold composition. For comparison, the advanced composition theorem for (ε, δ)-DP [DRV10] does not admit a single pair of optimal parameters ε, δ. In particular, no pair of ε, δ can sharply capture the privacy of the composition of (ε, δ)-DP mechanisms. See Section 3.3 and Figure 5 therein.

In the case of GDP, composition enjoys a simple and convenient formulation due to the identity

$$G_{\mu_1} \otimes G_{\mu_2} \otimes \cdots \otimes G_{\mu_n} = G_\mu,$$

where $\mu = \sqrt{\mu_1^2 + \cdots + \mu_n^2}$. This identity is almost immediate once we notice the rotational invariance of Gaussian distributions. We provide the proof in Appendix D. The following corollary formally summarizes this finding.

Corollary 3.3. *The n-fold composition of μ_i-GDP mechanisms is $\sqrt{\mu_1^2 + \cdots + \mu_n^2}$-GDP.*

The pioneering work of [KOV17] was the first to take the hypothesis testing perspective in the study of privacy composition and introduced Blackwell's theorem 2.10 as an analytic tool. The authors showed an optimal composition theorem for (ε, δ)-DP. [MV16] made the proof self-contained by independently proving the "(ε, δ) special case" of Blackwell's theorem. Their arguments carry over to our general Theorem 3.2 but when going beyond (ε, δ)-differential privacy, the powerful Blackwell's theorem is unavoidable. The novel proof we provide in Appendix C is completely elementary, using only the Neyman-Pearson Lemma A.1. In our opinion, this proof illuminates the essence of the problem.

3.2 Central Limit Theorems for Composition

In this subsection, we identify a central limit theorem type phenomenon of composition in the f-DP framework. Our main results (Theorem 3.4 and Theorem 3.5), roughly speaking, show that trade-off functions corresponding to small privacy leakage accumulate to G_μ for some μ under composition. Equivalently, the privacy of the composition of many "very private" mechanisms is best measured by GDP in the limit. This identifies GDP as the focal privacy definition amongst the family of f-DP privacy guarantees, including (ε, δ)-DP. More precisely, *all* privacy definitions that are based on a hypothesis testing formulation of "indistinguishability" converge to the guarantees of GDP in the limit of composition. We remark that [SMM18] proved a conceptually related "central limit theorem" for the privacy loss random variable used to define (ε, δ)-differential privacy for the special case of non-adaptive composition. In contrast, our central limit theorem refers to a different object: the optimal hypothesis testing trade-off functions that one obtains from the composed mechanism. Our theorem also applies to arbitrary composition.

From a computational viewpoint, these limit theorems yield an efficient method of approximating the composition of general f-DP mechanisms. This is important because algorithm design is *modular*, consisting of the composition of many smaller building blocks. Optimization algorithms in particular — including the ubiqitious stochastic gradient descent — run as the composition of a very large number of simple iterates. Since exact computation of privacy guarantees under composition can be computationally hard (under complexity theoretic assumptions weaker than P≠NP [MV16]), tractable approximations are important. Using our central limit theorems, the computation of the exact overall privacy guarantee $f_1 \otimes \cdots \otimes f_n$ in Theorem 3.2 can be reduced to the evaluation of a single mean parameter μ in a GDP guarantee. We give an example of this kind of analysis in Section 5, in which we analyze stochastic gradient descent.

Specifically, the mean parameter μ in the approximation depends on certain functionals of the trade-off functions[12]:

$$\begin{aligned}
\mathrm{kl}(f) &:= -\int_0^1 \log|f'(x)|\,\mathrm{d}x \\
\kappa_2(f) &:= \int_0^1 \log^2|f'(x)|\,\mathrm{d}x \\
\kappa_3(f) &:= \int_0^1 \big|\log|f'(x)|\big|^3\,\mathrm{d}x \\
\bar{\kappa}_3(f) &:= \int_0^1 \big|\log|f'(x)| + \mathrm{kl}(f)\big|^3\,\mathrm{d}x.
\end{aligned}$$

All of these functionals take values in $[0, +\infty]$, and the last is defined for f such that $\mathrm{kl}(f) < \infty$. In essence, these functionals are calculating moments of the log-likelihood ratio of P and Q such that $f = T(P, Q)$. In particular, all of these functionals are 0 if $f(x) = \mathrm{Id}(x) = 1 - x$, which corresponds to zero privacy leakage. As its name suggests, $\mathrm{kl}(f)$ is the Kullback–Leibler (KL) divergence of P and Q and, therefore, $\mathrm{kl}(f) \geqslant 0$. Detailed elaboration on these functionals is deferred to Appendix D.

In the following theorem, $\mathbf{kl}$ denotes the vector $\big(\mathrm{kl}(f_1), \ldots, \mathrm{kl}(f_n)\big)$ and $\boldsymbol{\kappa_2}, \boldsymbol{\kappa_3}, \bar{\boldsymbol{\kappa}}_{\mathbf{3}}$ are defined similarly; in addition, $\|\cdot\|_1$ and $\|\cdot\|_2$ are the ℓ_1 and ℓ_2 norms, respectively.

Theorem 3.4. *Let $f_1, \ldots, f_n$ be symmetric trade-off functions such that $\kappa_3(f_i) < \infty$ for all $1 \leqslant i \leqslant n$. Denote*

$$\mu := \frac{2\|\mathbf{kl}\|_1}{\sqrt{\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2}} \quad \textit{and} \quad \gamma := \frac{0.56\|\bar{\boldsymbol{\kappa}}_{\mathbf{3}}\|_1}{\big(\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2\big)^{3/2}}$$

and assume $\gamma < \frac{1}{2}$. Then, for all $\alpha \in [\gamma, 1-\gamma]$, we have

$$G_\mu(\alpha + \gamma) - \gamma \leqslant f_1 \otimes f_2 \otimes \cdots \otimes f_n(\alpha) \leqslant G_\mu(\alpha - \gamma) + \gamma. \tag{9}$$

Loosely speaking, the lower bound in (9) shows that the composition of f_i-DP mechanisms for $i = 1, \ldots, n$ is approximately μ-GDP and, in addition, the upper bound demonstrates that the tightness of this approximation is specified by γ. In particular, if all f_i are equal to some $f \neq \mathrm{Id}$ and $n \to \infty$, then the theorem reveals that the composition is blatantly non-private as $\mu \to \infty$. More interesting applications of the theorem are the cases where each f_i is close to the "perfect privacy" trade-off function Id such that collectively μ is convergent and γ vanishes as $n \to \infty$ (see the example in Section 5). For completeness, the condition $\kappa_3(f_i) < \infty$ (which implies that the other three functionals are also finite) for the use of this theorem excludes the case where $f_i(0) < 1$, in particular, $f_{\varepsilon,\delta}$ in (ε, δ)-DP with $\delta > 0$. We introduce an easy and general technique in Section 3.3 to deal with this.

From a technical viewpoint, Theorem 3.4 can be thought of as a Berry–Esseen type central limit theorem. The detailed proof, as well as that of Theorem 3.5, is provided in Appendix D.

Next, we present an asymptotic version of Theorem 3.4 for composition of f-DP mechanisms. In analogue to classical central limit theorems, below we consider a triangular array of mechanisms $\{M_{n1}, \ldots, M_{nn}\}_{n=1}^\infty$, where M_{ni} is f_{ni}-DP for $1 \leqslant i \leqslant n$.

[12] Although the trade-off function satisfies $f'(x) \leqslant 0$ almost everywhere on $[0, 1]$, we prefer to use $|f'(x)|$ instead of $-f'(x)$ for aesthetic reasons.

Theorem 3.5. *Let $\{f_{ni} : 1 \leqslant i \leqslant n\}_{n=1}^{\infty}$ be a triangular array of symmetric trade-off functions and assume the following limits for some constants $K \geqslant 0$ and $s > 0$ as $n \to \infty$:*

1. $\sum_{i=1}^{n} \mathrm{kl}(f_{ni}) \to K$;
2. $\max_{1 \leqslant i \leqslant n} \mathrm{kl}(f_{ni}) \to 0$;
3. $\sum_{i=1}^{n} \kappa_2(f_{ni}) \to s^2$;
4. $\sum_{i=1}^{n} \kappa_3(f_{ni}) \to 0$.

Then, we have

$$\lim_{n\to\infty} f_{n1} \otimes f_{n2} \otimes \cdots \otimes f_{nn}(\alpha) = G_{2K/s}(\alpha)$$

uniformly for all $\alpha \in [0, 1]$.

Taken together, this theorem and Composition Theorem 3.2 amount to saying that the composition $M_{n1} \otimes \ldots \otimes M_{nn}$ is asymptotically $2K/s$-GDP. In fact, this asymptotic version is a consequence of Theorem 3.4 as one can show $\mu \to 2K/s$ and $\gamma \to 0$ for the triangular array of symmetric trade-off functions. This central limit theorem implies that GDP is the *only* parameterized family of trade-off functions that can faithfully represent the effects of composition. In contrast, neither ε- nor (ε, δ)-DP can losslessly be tracked under composition—the parameterized family of functions $f_{\varepsilon,\delta}$ cannot represent the trade-off function that results from the limit under composition.

The conditions for use of this theorem are reminiscent of Lindeberg's condition in the central limit theorem for independent random variables. The proper scaling of the trade-off functions is that both $\mathrm{kl}(f_{ni})$ and $\kappa_2(f_{ni})$ are of order $O(1/n)$ for most $1 \leqslant i \leqslant n$. As a consequence, the cumulative effects of the moment functionals are bounded. Furthermore, as with Lindeberg's condition, the second condition in Theorem 3.5 require that no single mechanism has a significant contribution to the composition in the limit.

In passing, we remark that K and s satisfy the relationship $s = \sqrt{2K}$ in all examples of the application of Theorem 3.5 in this paper, including Theorem 3.6 and Theorem 5.2 as well as their corollaries. As such, the composition is asymptotically s-GDP. An easy-to-check condition that implies this interesting observation would save us the work of explicit calculation of κ_2, and is left for future work.

3.3 Composition of (ε, δ)-DP: Beating Berry–Esseen

Now, we extend central limit theorems to (ε, δ)-DP. As shown by Proposition 2.5, (ε, δ)-DP is equivalent to $f_{\varepsilon,\delta}$-DP and, therefore, it suffices to approximate the trade-off function $f_{\varepsilon_1,\delta_1} \otimes \cdots \otimes f_{\varepsilon_n,\delta_n}$ by making use of the composition theorem for f-DP mechanisms. As pointed out in Section 3.2, however, the moment conditions required in the two central limit theorems (Theorems 3.4 and 3.5) exclude the case where $\delta_i > 0$ for some i.

To overcome the difficulty caused by a nonzero δ, we start by observing the useful fact that

$$f_{\varepsilon,\delta} = f_{\varepsilon,0} \otimes f_{0,\delta}. \tag{10}$$

This decomposition along with the commutative and associative properties of the tensor product shows

$$f_{\varepsilon_1,\delta_1} \otimes \cdots \otimes f_{\varepsilon_n,\delta_n} = \big(f_{\varepsilon_1,0} \otimes \cdots \otimes f_{\varepsilon_n,0}\big) \otimes \big(f_{0,\delta_1} \otimes \cdots \otimes f_{0,\delta_n}\big).$$

This identity allows us to work on the ε part and δ part separately. In short, the ε part $f_{\varepsilon_1,0} \otimes \cdots \otimes f_{\varepsilon_n,0}$ now can be approximated by $G_{\sqrt{\varepsilon_1^2+\cdots+\varepsilon_n^2}}$ by invoking Theorem 3.5. For the δ part, we can iteratively apply the rule

$$f_{0,\delta_1} \otimes f_{0,\delta_2} = f_{0,1-(1-\delta_1)(1-\delta_2)} \tag{11}$$

to obtain $f_{0,\delta_1} \otimes \cdots \otimes f_{0,\delta_n} = f_{0,1-(1-\delta_1)(1-\delta_2)\cdots(1-\delta_n)}$. This rule is best seen via the interesting fact that $f_{0,\delta}$ is the trade-off function of shifted uniform distributions $f_{0,\delta} = T\big(U[0,1], U[\delta, 1+\delta]\big)$.

Now, a central limit theorem for (ε,δ)-DP is just a stone's throw away. In what follows, the privacy parameters ε and δ are arranged in a triangular array $\{(\varepsilon_{ni}, \delta_{ni}) : 1 \leqslant i \leqslant n\}_{n=1}^{\infty}$.

Theorem 3.6. *Assume*

$$\sum_{i=1}^n \varepsilon_{ni}^2 \to \mu^2, \quad \max_{1\leqslant i\leqslant n} \varepsilon_{ni} \to 0, \quad \sum_{i=1}^n \delta_{ni} \to \delta, \quad \max_{1\leqslant i\leqslant n} \delta_{ni} \to 0$$

for some nonnegative constants μ, δ as $n \to \infty$. Then, we have

$$f_{\varepsilon_{n1},\delta_{n1}} \otimes \cdots \otimes f_{\varepsilon_{nn},\delta_{nn}} \to G_\mu \otimes f_{0,1-\mathrm{e}^{-\delta}}$$

uniformly over $[0,1]$ *as* $n \to \infty$.

Remark 1. A formal proof is provided in Appendix D. The assumptions concerning $\{\delta_{ni}\}$ give rise to $1-(1-\delta_{n1})(1-\delta_{n2})\cdots(1-\delta_{nn}) \to 1-\mathrm{e}^{-\delta}$. In general, tensoring with $f_{0,\delta}$ is equivalent to scaling the graph of the trade-off function f toward the origin by a factor of $1-\delta$. This property is specified by the following formula, and we leave its proof to Appendix D:

$$f \otimes f_{0,\delta}(\alpha) = \begin{cases} (1-\delta)\cdot f(\frac{\alpha}{1-\delta}), & 0 \leqslant \alpha \leqslant 1-\delta \\ 0, & 1-\delta \leqslant \alpha \leqslant 1. \end{cases} \tag{12}$$

In particular, $f \otimes f_{0,\delta}$ is symmetric if f is symmetric. Note that (10) and (11) can be deduced by the formula above.

This theorem interprets the privacy level of the composition using Gaussian and uniform distributions. Explicitly, the theorem demonstrates that, based on the released information of the composed mechanism, distinguishing between any neighboring datasets is at least as hard as distinguishing between the following two bivariate distributions:

$$\mathcal{N}(0,1) \times U[0,1] \text{ versus } \mathcal{N}(\mu,1) \times U[1-\mathrm{e}^{-\delta}, 2-\mathrm{e}^{-\delta}].$$

We note that for small δ, $\mathrm{e}^{-\delta} \approx 1-\delta$. So $U[1-\mathrm{e}^{-\delta}, 2-\mathrm{e}^{-\delta}] \approx U[\delta, 1+\delta]$.

This approximation of the tensor product $f_{\varepsilon_{n1},\delta_{n1}} \otimes \cdots \otimes f_{\varepsilon_{nn},\delta_{nn}}$ using simple distributions is important from the viewpoint of computational complexity. Murtagh and Vadhan [MV16] showed that, given a collection of $\{(\varepsilon_i,\delta_i)\}_{i=1}^n$, finding the smallest ε such that $f_{\varepsilon,\delta} \leqslant f_{\varepsilon_1,\delta_1} \otimes \cdots \otimes f_{\varepsilon_n,\delta_n}$ is #P-hard[13] for any δ. From the dual perspective (see Section 2.4), this negative result is equivalent to the #P-hardness of evaluating the convex conjugate $\big(f_{\varepsilon_1,\delta_1} \otimes \cdots \otimes f_{\varepsilon_n,\delta_n}\big)^*$ at any point. For completeness, we remark that [MV16] provided an FPTAS[14] to approximately find the smallest ε in $O(n^3)$ time for a *single* δ. In comparison, Theorem 3.6 offers a *global* approximation of the

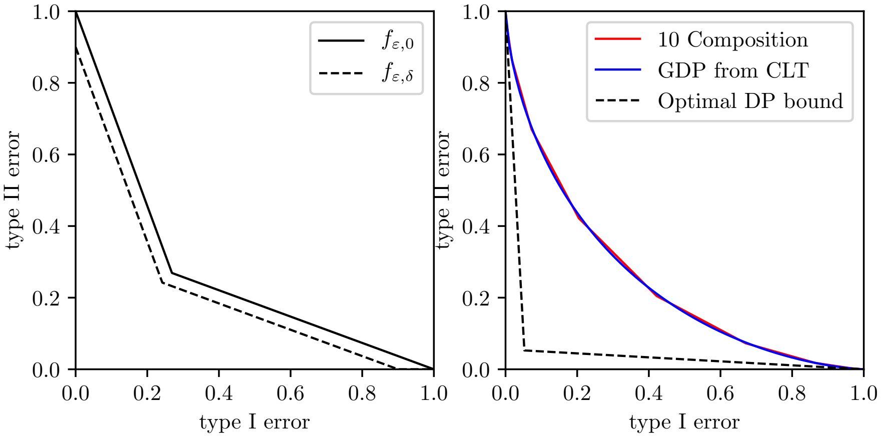

Figure 5: Left: Tensoring with $f_{0,\delta}$ scales the graph towards the origin by a factor of $1-\delta$. Right: 10-fold composition of $(1/\sqrt{10},0)$-DP mechanisms, i.e. $f_{\varepsilon,0}^{\otimes n}$ with $n=10, \varepsilon=1/\sqrt{n}$. The dashed curve corresponds to the smallest ε such that the composition is (ε,δ)-DP with $\delta=0.001$. Note that the central limit theorem approximation to the true trade-off curve is almost perfect, whereas the tightest possible approximation via (ε,δ)-differential privacy is substantially looser.

tensor product in $O(n)$ time using a closed-form expression, subsequently enabling an analytical approximation of the smallest ε for each δ.

That being said, Theorem 3.6 remains silent on the approximation error in applications with a moderately large number of (ε,δ)-DP mechanisms. Alternatively, we can apply Theorem 3.4 to obtain a non-asymptotic normal approximation to $f_{\varepsilon_1,0}\otimes\cdots\otimes f_{\varepsilon_n,0}$ and use γ to specify the approximation error. It can be shown that $\gamma=O(1/\sqrt{n})$ under mild conditions (Corollary D.7). This bound, however, is not sharp enough for tight privacy guarantees if n is not too large (note that $1/\sqrt{n}\approx 0.14$ if $n=50$, for which exact computation is already challenging, if possible at all). Surprisingly, the following theorem establishes a $O(1/n)$ bound, thereby "beating" the classical Berry–Esseen bound.

Theorem 3.7. *Fix $\mu>0$ and let $\varepsilon=\mu/\sqrt{n}$. There is a constant $c>0$ that only depends on μ satisfying*

$$G_\mu\left(\alpha+\tfrac{c}{n}\right)-\tfrac{c}{n}\leqslant f_{\varepsilon,0}^{\otimes n}(\alpha)\leqslant G_\mu\left(\alpha-\tfrac{c}{n}\right)+\tfrac{c}{n}$$

for all $n\geqslant 1$ and $c/n\leqslant\alpha\leqslant 1-c/n$.

As with Theorem 3.6, this theorem can be extended to approximate DP ($\delta\neq 0$) by making use of the decomposition (10). Our simulation studies suggest that $c\approx 0.1$ for $\mu=1$, which is best illustrated in the right panel of Figure 5. Despite a fairly small $n=10$, the difference between G_1 and its target $f_{\varepsilon,0}^{\otimes n}$ is less than 0.013 in the pointwise sense. Interestingly, numerical evidence suggests the same $O(1/n)$ rate in the inhomogeneous composition provided that $\varepsilon_1,\ldots,\varepsilon_n$ are roughly of the same size. A formal proof, or even a quantitative statement of this observation, constitutes an interesting problem for future investigation.

[13] #P is a complexity class that is "even harder than" NP (i.e. a polynomial time algorithm for any #P-hard problem would imply P=NP). See e.g. [AB09], Ch. 9.

[14] An approximation algorithm is called a fully polynomial-time approximation scheme (FPTAS) if its running time is polynomial in both the input size and the inverse of the relative approximation error. See, for example, [Vaz13], Ch. 8.

In closing this section, we highlight some novelties in the proof of Theorem 3.7. Letting $p_\varepsilon = \frac{1}{1+\mathrm{e}^\varepsilon}$ and $q_\varepsilon = \frac{\mathrm{e}^\varepsilon}{1+\mathrm{e}^\varepsilon}$, the following identity, rephrased in our framework, is already known in [KOV17]:

$$f_{\varepsilon,0}^{\otimes n} = T\big(B(n,p_\varepsilon), B(n,q_\varepsilon)\big)$$

where $B(n,p)$ denotes the binomial distribution with n trials and success probability p. At first glance, it is hard to believe any approach directly working with this expression can give an $O(1/n)$ bound because the Berry-Esseen bound is rate-optimal for binomial distributions. Our analysis, instead, rests crucially on a smooth effect that come for free in testing. It is analogous to continuity correction for normal approximations to binomial probabilities. See the technical details in Appendix D.

4 Amplifying Privacy by Subsampling

Subsampling is often used prior to a private mechanism M as a way to *amplify* privacy guarantees. Specifically, we can construct a smaller dataset $\tilde{S}$ by flipping a fair coin for each individual in the original dataset S to decide whether the individual is included in $\tilde{S}$. This subsampling scheme roughly shrinks the dataset by half and, therefore, we would expect that the induced mechanism applied to $\tilde{S}$ is about twice as private as the original mechanism M. Intuitively speaking, this privacy amplification is due to the fact that every individual enjoys perfect privacy if the individual is not included in the resulting dataset $\tilde{S}$, which happens with probability $\frac{1}{2}$.

The claim above was first formalized in [KLN$^+$11] for (ε,δ)-DP. Such a privacy amplification property is, unfortunately, no longer true for the most natural previous relaxations of differential privacy aimed at recovering precise compositions (like concentrated differential privacy [DR16, BS16]). Further modifications such as truncated CDP [BDRS18] have been introduced primarily to remedy this deficiency of CDP—but at the cost of extra complexity in the definition. Other relaxations like Rényi Differential Privacy [Mir17] can be shown to satisfy a form of privacy amplification by subsampling, but both the analysis and the statement are complex [WBK18].

In this section, we show that these obstacles can be overcome by our hypothesis testing-based relaxation of differential privacy. Explicitly, our main result is a simple, general, and easy-to-interpret subsampling theorem for f-DP. Somewhat surprisingly, our theorem significantly improves on the classical subsampling theorem for privacy amplification in the (ε,δ)-DP framework [Ull17] (by no longer expressing the resulting guarantees using (ε,δ)-differential privacy).

4.1 A Subsampling Theorem

Given an integer $1 \leqslant m \leqslant n$ and a dataset S of n individuals, let $\mathtt{Sample}_m(S)$ be a subset of S that is chosen uniformly at random among all the m-sized subsets of S. For a mechanism M defined on X^m, we call $M\big(\mathtt{Sample}_m(S)\big)$ the subsampled mechanism, which takes as input an n-sized dataset. Formally, we use $M \circ \mathtt{Sample}_m$ to denote this subsampled mechanism. To clear up any confusion, note that intermediate result $\mathtt{Sample}_m(S)$ is not released and, in particular, this is different from the composition in Section 3.

In brief, our main theorem shows that the privacy bound of the subsampled mechanism in the f-DP framework is given by an operator acting on trade-off functions. To introduce this operator, write the convex combination $f_p := pf + (1-p)\mathrm{Id}$ for $0 \leqslant p \leqslant 1$, where $\mathrm{Id}(x) = 1-x$. Note that the trade-off function f_p is asymmetric in general.

Definition 4.1. *For any $0 \leqslant p \leqslant 1$, define the operator C_p acting on trade-off functions as*

$$C_p(f) := \min\{f_p, f_p^{-1}\}^{**}.$$

We call C_p the p-sampling operator.

Above, the inverse f_p^{-1} is defined in (3). The biconjugate $\min\{f_p, f_p^{-1}\}^{**}$ is derived by applying the conjugate as defined in (7) twice to $\min\{f_p, f_p^{-1}\}$. For the moment, take for granted the fact that $C_p(f)$ is a symmetric trade-off function.

Now, we present the main theorem of this section.

Theorem 4.2. *If M is f-DP on X^m, then the subsampled mechanism $M \circ \texttt{Sample}_m$ is $C_p(f)$-DP on X^n, where the sampling ratio $p = \frac{m}{n}$.*

Appreciating this theorem calls for a better understanding of the operator C_p. In effect, C_p performs a two-step transformation: symmetrization (taking the minimum of f_p and its inverse f_p^{-1}) and convexification (taking the largest convex lower envelope of $\min\{f_p, f_p^{-1}\}$). The convexification step is seen from convex analysis that the biconjugate h^{**} of any function h is the greatest convex lower bound of h. As such, $C_p(f)$ is convex and, with a bit more analysis, Proposition 2.2 ensures that $C_p(f)$ is indeed a trade-off function. As an aside, $C_p(f) \leqslant \min\{f_p, f_p^{-1}\} \leqslant f_p$. See Figure 6 for a graphical illustration.

Next, the following facts concerning the p-sampling operator qualitatively illustrate this privacy amplification phenomenon.

1. If $0 \leqslant p \leqslant q \leqslant 1$ and f is symmetric, we have $f = C_1(f) \leqslant C_q(f) \leqslant C_p(f) \leqslant C_0(f) = \mathrm{Id}$. That is, as the sampling ratio declines from 1 to 0, the privacy guarantee interpolates monotonically between the original f and the perfect privacy guarantee Id. This monotonicity follows from the fact that $g \geqslant h$ is equivalent to $g^{-1} \geqslant h^{-1}$ for any trade-off functions g and h.

2. If two trade-off functions f and g satisfy $f \geqslant g$, then $C_p(f) \geqslant C_p(g)$. This means that if a mechanism is more private than the other, using the same sampling ratio, the subsampled mechanism of the former remains more private than that of the latter, at least in terms of lower bounds.

3. For any $0 \leqslant p \leqslant 1$, $C_p(\mathrm{Id}) = \mathrm{Id}$. That is, perfect privacy remains perfect privacy with subsampling.

Explicitly, we provide a formula to calculate $C_p(f)$ for a symmetric trade-off function f. Letting x^* be the unique fixed point of f, that is $f(x^*) = x^*$, we have

$$C_p(f)(x) = \begin{cases} f_p(x), & x \in [0, x^*] \\ x^* + f_p(x^*) - x, & x \in [x^*, f_p(x^*)] \\ f_p^{-1}(x), & x \in [f_p(x^*), 1]. \end{cases} \tag{13}$$

This expression is almost self-evident from the left panel of Figure 6. Nevertheless, a proof of this formula is given in Appendix F. This formula together with Theorem 4.2 allows us to get a closed-form characterization of the privacy amplification for (ε, δ)-DP.

Corollary 4.3. *If M is (ε,δ)-DP on X^m, then the subsampled mechanism $M\circ\texttt{Sample}_m$ is $C_p(f_{\varepsilon,\delta})$-DP on X^n, where*

$$C_p(f_{\varepsilon,\delta})(\alpha)=\max\left\{f_{\varepsilon',\delta'}(\alpha),1-p\delta-p\frac{\mathrm{e}^{\varepsilon}-1}{\mathrm{e}^{\varepsilon}+1}-\alpha\right\}. \tag{14}$$

Above, $\varepsilon'=\log(1-p+p\mathrm{e}^{\varepsilon}),\delta'=p\delta$, *and* $p=\frac{m}{n}$.

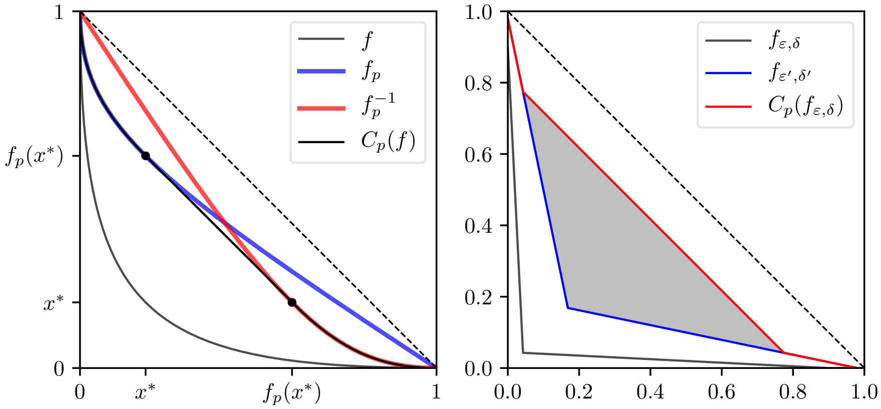

Figure 6: The action of C_p. Left panel: $f=G_{1.8},p=0.35$. Right panel: $\varepsilon=3,\delta=0.1,p=0.2$. The subsampling theorem 4.2 results in a significantly tighter trade-off function compared to the classical theorem for (ε,δ)-DP.

For comparison, we now present the existing bound on the privacy amplification by subsampling for (ε,δ)-DP. To be self-contained, Appendix F gives a proof of this result, which primarily follows [Ull17] .

Lemma 4.4 ([Ull17])**.** *If M is (ε,δ)-DP, then $M\circ\texttt{Sample}_m$ is (ε',δ')-DP with ε' and δ' defined in Corollary 4.3.*

Using the language of the f-DP framework, Lemma 4.4 states that $M\circ\texttt{Sample}_m$ is $f_{\varepsilon',\delta'}$-DP. Corollary 4.3 improves on Lemma 4.4 because, as is clear from (14),

$$C_p(f_{\varepsilon,\delta})\geqslant f_{\varepsilon',\delta'}.$$

The right panel of Figure 6 illustrates Lemma 4.4 and our Corollary 4.3 for $\varepsilon=3,\delta=0.1$, and $p=0.2$. In effect, the improvement is captured by the shaded triangle enclosed by $f_{\varepsilon,\delta,p}$ and $f_{\varepsilon',\delta'}$, revealing that the minimal sum of type I and type II errors in distinguishing two neighboring datasets with subsampling can be significantly lower than the prediction of Lemma 4.4. This gain is only made possible by the flexibility of trade-off functions in the sense that $f_{\varepsilon,\delta,p}$ *cannot* be expressed within the (ε,δ)-DP framework. The unavoidable lossiness of the (ε,δ)-DP representation of the guarantees of subsampling are compounded when analyzing the composition of many private mechanisms.

In the next section, we show a reduction style proof of Theorem 4.2, via Lemma 4.4. As a consequence, Theorem 4.2 holds for any subsampling scheme for which Lemma 4.4 is available. In particular, it holds for the subsampling scheme described at the beginning of this section, i.e. independent coin flips for every data item.

4.2 Proof of the Privacy Amplification by Subsampling Theorem

The proof strategy is as follows. First, we convert the f-DP guarantee into an infinite collection of (ε,δ)-DP guarantees by taking a dual perspective that is enabled by Proposition 2.12. Next, by applying the classical subsampling theorem (that is, Lemma 4.4) to these (ε,δ)-DP guarantees, we conclude that the subsampled mechanism satisfies a new infinite collection of (ε,δ)-DP guarantees. Finally, Proposition 2.11 allows us to convert these new privacy guarantees back into an $\tilde{f}$-DP guarantee, where $\tilde{f}$ can be shown to coincide with $C_p(f)$.

Proof of Theorem 4.2. Provided that M is f-DP, from Proposition 2.12 it follows that M is $\big(\varepsilon,\delta(\varepsilon)\big)$-DP with $\delta(\varepsilon)=1+f^*(-\mathrm{e}^{\varepsilon})$ for all $\varepsilon\geqslant 0$. Making use of Lemma 4.4, the subsampled mechanism $M\circ\mathtt{Sample}_m$ satisfies the following collection of (ε',δ')-DP guarantees for all $\varepsilon\geqslant 0$:

$$\varepsilon'=\log(1-p+p\mathrm{e}^{\varepsilon}),\quad \delta'=p\big(1+f^*(-\mathrm{e}^{\varepsilon})\big).$$

Eliminating the variable ε from the two parametric equations above, we can relate ε' to δ' using

$$\delta'=1+f_p^*(-\mathrm{e}^{\varepsilon'}),\tag{15}$$

which is proved in Appendix F. The remainder of the proof is devoted to showing that (ε',δ')-DP guarantees for all $\varepsilon'\geqslant 0$ is equivalent to the $C_p(f)$-DP guarantee.

At first glance, (15) seems to enable the use of Proposition 2.12. Unfortunately, that would be invalid because f_p is asymmetric. To this end, we need to extend Proposition 2.12 to general trade-off functions. To avoid conflicting notation, let g be a generic trade-off function, not necessarily symmetric. Denote by $\bar{x}$ be the smallest point such that $g'(x)=-1$, that is, $\bar{x}=\inf\{x\in[0,1]:g'(x)=-1\}$.[15] As a special instance of Proposition F.2 in the appendix, the following result serves our purpose.

Proposition 4.5. *If $g(\bar{x})\geqslant\bar{x}$ and a mechanism M is $(\varepsilon,1+g^*(-\mathrm{e}^{\varepsilon}))$-DP for all $\varepsilon\geqslant 0$, then M is $\min\{g,g^{-1}\}^{**}$-DP.*

The proof of the present theorem would be complete if Proposition 4.5 can be applied to the collection of privacy guarantees in (15)for f_p. To use Proposition 4.5, it suffices to verify the condition $f_p(\bar{x})\geqslant\bar{x}$ where $\bar{x}$ is the smallest point such that $f_p'(x)=-1$. Let x^* be the (unique) fixed point of f. To this end, we collect a few simple facts:

- First, $f'(x^*)=-1$. This is because the graph of f is symmetric with respect to the 45° line passing through the origin.
- Second, $\bar{x}\leqslant x^*$. This is because $f_p'(x^*)=pf'(x^*)+(1-p)\mathrm{Id}'(x^*)=-1$ and, by definition, $\bar{x}$ can only be smaller.

With these facts in place, we get

$$f_p(\bar{x})\geqslant f_p(x^*)\geqslant f(x^*)=x^*\geqslant\bar{x}$$

by recognizing that f_p is decreasing and $f_p\geqslant f$. Hence, the proof is complete.

□

[15] For simplicity, the proof assumes differentiable trade-off functions. If g is not differentiable, use the definition $\bar{x}=\inf\{x\in[0,1]:-1\in\partial g(x)\}$ instead. This adjustment applies to other parts of the proof.

5 Application: Privacy Analysis of Stochastic Gradient Descent

One of the most important algorithms in machine learning and optimization is stochastic gradient descent. It is an iterative optimization method used to train a wide variety of models — and in particular, the deep neural networks that have become tremendously popular in the last half decade. It has also served as an important benchmark in the development of private optimization: as an iterative algorithm, the tightness of its analysis crucially depends on the tightness with which composition can be accounted for. Its analysis also crucially requires a privacy amplification under sub-sampling argument. The first asymptotically optimal analysis of differentially private stochastic gradient descent was given by Bassily et al. [BST14] — however because of the inherent limits of (ε, δ)-differential privacy, this original analysis did not give meaningful privacy bounds for realistically sized datasets. This is in part what motivated the development of divergence based relaxations of differential privacy. However, because concentrated differential privacy does not admit a privacy amplification by subsampling theorem, it could not be directly applied to the analysis of stochastic gradient descent. Abadi et al. [ACG$^+$16] circumvented this challenge by developing the "moments accountant" — a numeric technique tailored specifically to repeated application of subsampling, followed by Gaussian perturbation — to give privacy bounds for stochastic gradient descent that were strong enough to give non-trivial guarantees when applying deep learning techniques to real datasets. But this analysis is ad-hoc in the sense that it uses a tool designed specifically for the analysis of stochastic gradient descent. In this section we use the general tools we have developed so far to give a simple and improved analysis of the privacy of noisy stochastic gradient descent (SGD).

5.1 NoisySGD and its Privacy Analysis

The private variant of the SGD algorithm is described in Algorithm 1. As we will see, from the perspective of its privacy analysis, it can simply be viewed as a repeated composition of Gaussian mechanisms operating on subsampled datasets.

Algorithm 1 NoisySGD

```
Input: Examples $S = (x_1, \dots, x_n)$, loss function $L(\theta, x)$.
       Parameters: initial state $\theta_0$, learning rate $\eta_t$, batch size $m$, time horizon $T$
                   noise scale $\sigma$, gradient norm bound $C$.
for $t \in [T]$ do
    Subsampling
    Take a uniformly random subsample $I_t \subseteq [n]$ with size $m$        ▷ Sample$_m$ in Section 4
    for $i \in I_t$ do
        Compute gradient
        $v_t^{(i)} \leftarrow \nabla_\theta L(\theta_t, x_i)$
        Clip gradient
        $\bar{v}_t^{(i)} \leftarrow v_t^{(i)} / \max\left\{1, \frac{\|v_t^{(i)}\|_2}{C}\right\}$
    Average, perturb and descend
    $\theta_{t+1} \leftarrow \theta_t - \eta_t \cdot \left(\frac{1}{m}\sum_i \bar{v}_t^{(i)} + \mathcal{N}(0, \sigma^2 \cdot \frac{4C^2}{m^2} I)\right)$
Output $\theta_T$
```

In order to analyze the privacy of NoisySGD, let's build up from the inner loop. Let V be the vector space where parameter θ lives in and $M : X^m \times V \to V$ be the mechanism that executes lines 4-7 in Algorithm 1. What M does in iteration t can be summarized as

$$M(S_{I_t}, \theta_t) = \theta_{t+1}$$

where S_{I_t} is the subset of S indexed by I_t. Now, let's consider the subsampling step (line 3) and use $\widetilde{M}$ to denote its composition with M, i.e. $\widetilde{M} = M \circ \mathtt{Sample}_m$. It executes 3-7 and maps from $X^n \times V$ to V.

The mechanism we are ultimately interested in is:

$$\begin{aligned} \text{NoisySGD} : X^n &\to V \times V \times \cdots \times V \\ S &\mapsto (\theta_1, \theta_2, \ldots, \theta_T) \end{aligned}$$

NoisySGD is the composition of T copies of $\widetilde{M}$. To see this, note that the trajectory $(\theta_1, \theta_2, \ldots, \theta_T)$ is obtained by the following iteration:

$$\begin{aligned} \theta_1 &= \widetilde{M}(S, \theta_0) \\ \theta_2 &= \widetilde{M}(S, \theta_1) \\ &\vdots \\ \theta_T &= \widetilde{M}(S, \theta_{T-1}) \end{aligned}$$

Straightforwardly, if M is f-DP, then by Theorem 4.2, $\widetilde{M}$ is $C_{m/n}(f)$-DP. By the composition theorem 3.2, NoisySGD is $C_{m/n}(f)^{\otimes T}$-DP.

So it suffices to give a bound on the privacy of M. Let's focus on a single time step and drop the subscript t. Changing one of the m data points only affects one $v^{(i)}$. Because of the clipping, the L^2 sensitivity of $\frac{1}{m}\sum_i \bar{v}_t^{(i)}$ is at most $\frac{2C}{m}$. By Theorem 2.7, adding Gaussian noise $N(0, \sigma^2 \cdot \frac{4C^2}{m^2} I)$ yields σ^{-1}-GDP. After perturbing the average gradient vector, the descent update of the parameter is deterministic post-processing, so we can conclude that M satisfies σ^{-1}-GDP.

In summary, these observations yield the following theorem:

Theorem 5.1. NoisySGD *is* $C_{m/n}(G_{\sigma^{-1}})^{\otimes T}$*-DP.*

Note that we analyze the privacy assuming:

1. None of the subsampled indices are released.

2. Minibatch indices are independent across iterations.

How should we interpret a theorem like this? We could try to numerically compute $C_{m/n}(G_{\sigma^{-1}})^{\otimes T}$, which would give a technique similar to the moment's accountant method, although one that is not specialized to a particular algorithm. However, our central limit theorems give us another tool: we can analytically approximate $C_{m/n}(G_{\sigma^{-1}})^{\otimes T}$ in a way that becomes exact as T grows large. In the next sections, we give two parallel results, each corresponding to one of our two central limit theorems. The asymptotic version is derived in section 5.2 by developing a general limit theorem for the composition of subsampled mechanisms. The Berry-Esseen version (with finite composition guarantees) appears in section 5.3.

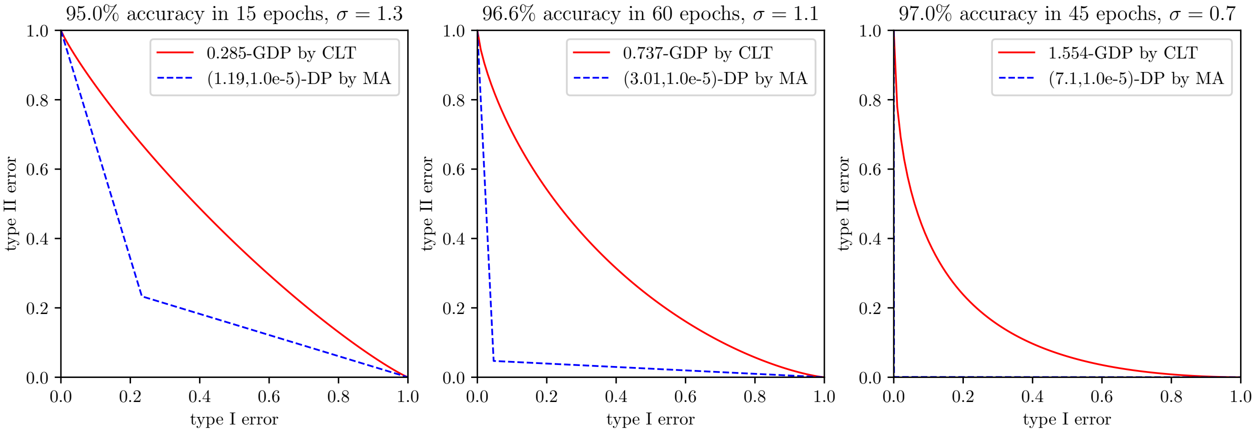

Figure 7: Comparison of the GDP bounds derived via our method, and the (ε, δ)-DP bounds derived using the moments accountant (MA) in [ACG$^+$16]. The experiments are those reported in [ACG$^+$16]. All three experiments run Algorithm 1 on the entire MNIST dataset with $n = 60,000$ data points, minibatch size $m = 256$, learning rates η_t set to 0.25, 0.15, and 0.25 respectively, and clipping thresholds C set to 1.5, 1.0, 1.5 respectively. The red lines are obtained via Theorem 5.4, while the blue dashed lines are produced by the tensorflow/privacy library. See `https://github.com/tensorflow/privacy` for the detail of the experiments.

5.2 Asymptotic Analysis of NoisySGD

In this section we first consider the limit of $C_p(f)^{\otimes T}$ for general f, then plug in $f = G_{\sigma^{-1}}$ for the analysis of SGD. The more general approach is useful for analyzing other iterative algorithms.

Recall from Section 4 that a p-subsampled f-DP mechanism is $C_p(f)$-DP, where $C_p(f)$ is defined as

$$C_p(f)(x) = \begin{cases} f_p(x), & x \in [0, x^*], \\ x^* + f_p(x^*) - x, & x \in [x^*, f_p(x^*)], \\ f_p^{-1}(x), & x \in [f_p(x^*), 1]. \end{cases}$$

x^* is the unique fixed point of f.

We will let the sampling fraction p go to 0 as T goes to infinity. In the following, a_+^2 is defined to be $(\max\{a, 0\})^2$.

Theorem 5.2. *Suppose f is a symmetric trade-off function such that $f(0) = 1$ and $\int_0^1 (f'(x) + 1)^4 \,\mathrm{d}x < +\infty$. Furthermore, $p\sqrt{T} \to p_0$ as $T \to \infty$. Then we have the uniform limit*

$$C_p(f)^{\otimes T} \to G_{p_0\sqrt{2\chi_+^2(f)}}$$

where

$$\chi_+^2(f) = \int_0^1 \big(|f'(x)| - 1\big)_+^2 \,\mathrm{d}x.$$

This theorem has implications for the design of iterative private mechanisms involving subsampling as a subroutine. One way to bound the privacy of such a mechanism is to let the sampling ratio p go to zero as the number of iteration T goes to infinity. The theorem tells us that the correct dependence is $p \sim 1/\sqrt{T}$, and explicitly determines the limit.

In order to analyze NoisySGD, we need to compute the quantity $\chi^2_+(G_\mu)$. This can be done directly from definition. In Appendix G we provide a different approach by relating $\chi^2_+(f)$ to the χ^2-divergence, which also explains the origin of this name.

Lemma 5.3.

$$\chi^2_+(G_\mu) = \mathrm{e}^{\mu^2} \cdot \Phi(3\mu/2) + 3\Phi(-\mu/2) - 2.$$

When using stochastic gradient descent to train large models, typically we have large dataset and run the algorithm for many iterations, so we consider the parameter regime in which $n \to \infty, T \to \infty$. The batch size can also vary with these quantities, but doesn't necessarily. The following theorem is a direct consequence of Theorems 5.1 and 5.2 and Lemma 5.3.

Theorem 5.4. *If $\frac{m}{n} \cdot \sqrt{T} \to c$, then* NoisySGD *is μ-GDP with*

$$\mu = \sqrt{2}c \cdot \sqrt{\mathrm{e}^{\sigma^{-2}} \cdot \Phi(1.5\sigma^{-1}) + 3\Phi(-0.5\sigma^{-1}) - 2}.$$

In many circumstances, the limit $\frac{m}{n} \cdot \sqrt{T} \to c$ holds with a small c. First, we remark that this is consistent with the analysis of private SGD in [BST14], in which $m = 1$ and $T = O(n^2)$. We also note that in the deep learning literature, the quantity $\frac{m}{n} \cdot \sqrt{T}$ is generally quite small. The convention in this literature is to reparameterize the number of gradient steps T by the number of "epochs" E, which is the number of sweeps of the entire dataset. The relationship between these parameters is that $E = Tm/n$. In this reparameterization, our assumption is that $E \cdot \frac{m}{n} \to c^2$. AlexNet [KSH12] used parameters with $m = 128, E \approx 90$ on the ILSVRC-2010 dataset with $n \approx 1.2 \times 10^6$, leading to $E \cdot \frac{m}{n} < 0.01$. Other prominent implementations lead to similarly small values[16].

5.3 Berry–Esseen Privacy Bound for NoisySGD

We can also apply the Berry–Esseen style CLT 3.4. The advantage is that it gives hard, finite composition privacy guarantees. The disadvantage is that the expressions it yields are more unwieldy: they are computer evaluatable, so usable in implementations, but don't have simple closed forms.

The individual components in Theorem 3.4 have the form $C_p(G_\mu)$ with $p = m/n, \mu = \sigma^{-1}$. It suffices to evaluate the moment functionals on $C_p(G_\mu)$.

Lemma 5.5. *Let $Z(x) = \log(p \cdot e^{\mu x - \mu^2/2} + 1 - p)$. Then*

$$\begin{aligned}
\mathrm{kl}\big(C_p(G_\mu)\big) &= p\int_{\mu/2}^{+\infty} Z(x) \cdot \big(\varphi(x-\mu) - \varphi(x)\big)\,\mathrm{d}x \\
\kappa_2\big(C_p(G_\mu)\big) &= \int_{\mu/2}^{+\infty} Z^2(x) \cdot \big(p\varphi(x-\mu) + (2-p)\varphi(x)\big)\,\mathrm{d}x \\
\bar{\kappa}_3\big(C_p(G_\mu)\big) &= \int_{\mu/2}^{+\infty} \big|Z(x) - \mathrm{kl}\big(C_p(G_\mu)\big)\big|^3 \cdot (p\varphi(x-\mu) + (1-p)\varphi(x))\,\mathrm{d}x \\
&\quad + \int_{\mu/2}^{+\infty} \big|Z(x) + \mathrm{kl}\big(C_p(G_\mu)\big)\big|^3 \cdot \varphi(x)\,\mathrm{d}x.
\end{aligned}$$

[16] See the webpage of Gluon CV Toolkit [HZZ+18, ZHZ+19] for a collection of such hyperparameters on computer vision tasks.

We can plug in these expression into Theorem 3.4 and get

Theorem 5.6. *Let $p = m/n, \mu = \sigma^{-1}$ and*

$$z = \frac{2\sqrt{T} \cdot \mathrm{kl}\big(C_p(G_\mu)\big)}{\sqrt{\kappa_2\big(C_p(G_\mu)\big) - \mathrm{kl}^2\big(C_p(G_\mu)\big)}},$$

$$\gamma = \frac{0.56}{\sqrt{T}} \cdot \frac{\bar{\kappa}_3\big(C_p(G_\mu)\big)}{\Big(\kappa_2\big(C_p(G_\mu)\big) - \mathrm{kl}^2\big(C_p(G_\mu)\big)\Big)^{\frac{3}{2}}}.$$

NoisySGD *is f-DP with*

$$f(\alpha) = \max\{G_z(\alpha+\gamma) - \gamma, 0\}.$$

6 Discussion

We have introduced a family of privacy definitions that generalizes differential privacy and has a number of attractive properties that escape the difficulties of prior work.

1. It retains an interpretable hypothesis-testing semantics.
2. It is expressive enough to losslessly reason about, composition, post-processing, and group privacy.
3. It is *dual* to differential privacy in a constructive sense, which gives the ability to import results proven for (ε, δ)-DP. This is what allows us to easily import amplification-by-subsampling theorems.
4. It admits a *central limit theorem*, which identifies a simple, single parameter family of privacy definitions as focal: Gaussian Differential Privacy. All hypothesis-testing based definitions of privacy converge to Gaussian differential privacy in the limit under composition, which implies that this is the unique such definition which can tightly handle composition.
5. The central limit theorem (and its finite Berry-Esseen variant) give a tractable analytical method to tightly analyze iterative methods — which we illustrate through our analysis of stochastic gradient descent.

Ultimately, the test of a privacy definition lies not just in its power and semantics, but in its ability as a tool to usefully analyze diverse algorithms. In this paper, we give evidence that f-DP is up to the task, but ultimately practical evaluation is left to future work.

One of the most intriguing theoretical directions suggested by our work is the possible extension of Theorem 3.7 to the inhomogeneous case. Theorem 3.7 suggests that the convergence implied by our central limit theorem to GDP might be extremely fast — faster than is established by our Berry-Esseen variant. Establishing this in the general case would make the central-limit style of analysis of algorithms available to algorithms which consist only of a much smaller number of compositions of f-DP mechanisms.

Acknowledgements

This work was supported in part by the NSF via grant CCF-1763314.

References

[AB09] Sanjeev Arora and Boaz Barak. *Computational complexity: a modern approach.* Cambridge University Press, 2009.

[Abo18] John 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. ACM, 2018.

[ACG+16] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.

[App17] Differential Privacy Team Apple. Learning with privacy at scale. Technical report, Apple, 2017.

[BDRS18] Mark Bun, Cynthia Dwork, Guy N Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated cdp. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 74–86. ACM, 2018.

[Bla50] David Blackwell. Comparison of experiments. Technical report, HOWARD UNIVERSITY Washington United States, 1950.

[BS16] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[BST14] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

[BZ06] Michael Barbaro and Tom Zeller. A face is exposed for AOL searcher no. 4417749. *The New York Times*, August 2006. `http://select.nytimes.com/gst/abstract.html?res=F10612FC345B0C7A8CDDA10894DE404482`.

[DKM+06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer, 2006.

[DKY17] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Proceedings of Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[DR16] Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

[DRV10] Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 51–60. IEEE, 2010.

[Dur19] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[E$^+$85] Shinto Eguchi et al. A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima mathematical journal*, 15(2):341–391, 1985.

[EPK14] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM, 2014.

[HSR$^+$08] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J.V. Pearson, D.A. Stephan, S.F. Nelson, and D.W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genetics*, 4(8):e1000167, 2008.

[HV11] Peter Harremoës and Igor Vajda. On pairs of f-divergences and their joint range. *IEEE Transactions on Information Theory*, 57(6):3230–3235, 2011.

[HZZ$^+$18] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. *arXiv preprint arXiv:1812.01187*, 2018.

[KLN$^+$11] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[KOV17] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049, 2017.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[Leh04] Erich Leo Lehmann. *Elements of large-sample theory.* Springer Science & Business Media, 2004.

[LR06] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses.* Springer Science & Business Media, 2006.

[LV06] Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

[Mir17] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

[MV16] Jack Murtagh and Salil Vadhan. The complexity of computing the optimal composition of differential privacy. In *Theory of Cryptography Conference*, pages 157–175. Springer, 2016.

[NS08] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 ieee symposium on security and privacy*, pages 111–125. IEEE, 2008.

[Pól20] Georg Pólya. Über den zentralen grenzwertsatz der wahrscheinlichkeitsrechnung und das momentenproblem. *Mathematische Zeitschrift*, 8(3-4):171–181, 1920.

[PW14] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6:2012–2016, 2014.

[Rag11] Maxim Raginsky. Shannon meets blackwell and le cam: Channels, codes, and statistical experiments. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 1220–1224. IEEE, 2011.

[She10] IG Shevtsova. An improvement of convergence rate estimates in the lyapunov theorem. In *Doklady Mathematics*, volume 82, pages 862–864. Springer, 2010.

[SMM18] David Sommer, Sebastian Meiser, and Esfandiar Mohammadi. Privacy loss classes: The central limit theorem in differential privacy. 2018.

[Ull17] Jonathan Ullman. Cs7880: Rigorous approaches to data privacy, spring 2017. 2017. `http://www.ccs.neu.edu/home/jullman/PrivacyS17/HW1sol.pdf`.

[Usp37] James Victor Uspensky. Introduction to mathematical probability. 1937.

[Vaz13] Vijay V Vazirani. *Approximation algorithms.* Springer Science & Business Media, 2013.

[WBK18] Yu-Xiang Wang, Borja Balle, and Shiva Kasiviswanathan. Subsampled r\'enyi differential privacy and analytical moments accountant. *arXiv preprint arXiv:1808.00087*, 2018.

[WZ10] Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[ZHZ$^+$19] Zhi Zhang, Tong He, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of freebies for training object detection neural networks. *arXiv preprint arXiv:1902.04103*, 2019.

A Technical Details in Section 2

We first state the fundamental tool of the entire paper: Neyman-Pearson Lemma.

Theorem A.1 (Neyman-Pearson Lemma. [LR06] 3.2.1)**.** *Let P and Q be probability distributions on Ω with densities p and q with respect to a measure μ, respectively. For the hypothesis testing* *problem $H_0 : P$ vs $H_1 : Q$, a test $\phi : \Omega \to [0,1]$ is the most powerful test at level α if and only if there are two constants $h \in [0,+\infty]$ and $c \in [0,1]$ such that ϕ has the form*

$$\phi(\omega) = \begin{cases} 1, & \text{if } p(\omega) > hq(\omega), \\ c, & \text{if } p(\omega) = hq(\omega), \\ 0, & \text{if } p(\omega) < hq(\omega), \end{cases}$$

and $\mathbb{E}_P[\phi] = \alpha$.

We then use it to prove the fundamental theorem of trade-off functions.

Proposition 2.2. *A function $f : [0,1] \to [0,1]$ is a trade-off function if and only if f is convex, continuous*[17]*, non-increasing and $f(x) \leqslant 1 - x$ for $x \in [0,1]$.*

In the entire appendix, from A to G, we will use $\mathscr{F}$ to denote the class of trade-off functions, and $\mathscr{F}^S$ the subclass of symmetric trade-off functions.

Proof of Proposition 2.2. "only if": Suppose $f = T(P_0, P_1)$. It is obviously non-increasing. The randomized testing rule that blindly rejects with probability p achieves $(p, 1-p)$ errors. It is suboptimal at level p, so $f(p) \leqslant 1 - p$.

Convexity follows from randomizing over two rejections rules. For given α, α', t, all in $[0,1]$, let ϕ and ϕ' be the rejection rules achieving errors $(\alpha, f(\alpha))$ and $(\alpha, f(\alpha))$ respectively. The rejection rule $\phi_t = t\phi + (1-t)\phi'$ achieves errors $t\alpha + (1-t)\alpha'$ and $tf(\alpha) + (1-t)f(\alpha')$. It is suboptimal at level $t\alpha + (1-t)\alpha'$, so we have

$$f(t\alpha + (1-t)\alpha') \leqslant t\alpha + (1-t)\alpha'.$$

As we remarked in the footnote, continuity in $(0,1]$ follows from properties we have proved. At 0 it requires a closer look at Neyman-Pearson lemma. Suppose $\alpha_n \to 0$. Without loss of generality we can assume α_n is decreasing. We want to show $f(\alpha_n) \to f(0)$. Neyman-Pearson lemma tells us the optimal test ϕ_n at level α_n must have the form

$$\phi_n(\omega) = \begin{cases} 1, & \frac{p_1(\omega)}{p_0(\omega)} > h_n \\ c_n, & \frac{p_1(\omega)}{p_0(\omega)} = h_n \\ 0, & \frac{p_1(\omega)}{p_0(\omega)} < h_n \end{cases}$$

for some $c_n \in [0,1]$ and $h_n \in [0,+\infty]$. The fact that α_n is decreasing implies that $\phi_n(\omega)$ is monotone dereasing in n except on a measure zero set, so it has a pointwise limit $\phi(\omega)$. Furthermore, $\phi(\omega)$ must be in the same form as ϕ_n. Again by Neyman-Pearson lemma, ϕ must be the optimal test at level $\mathbb{E}_{P_0}[\phi]$. By dominated convergence theorem, $\mathbb{E}_{P_i}[\phi] = \lim_{n\to\infty} \mathbb{E}_{P_i}[\phi_n]$ for $i = 0, 1$. When $i = 0$, this translates to $\mathbb{E}_{P_0}[\phi] = \lim_{n\to\infty} \alpha_n = 0$. So ϕ is at level 0. When $i = 1$, we have

$$\mathbb{E}_{P_1}[\phi] = \lim_{n\to\infty} \mathbb{E}_{P_1}[\phi_n] = \lim_{n\to\infty} 1 - f(\alpha_n)$$

where the second equality follows from the optimality of ϕ_n. So

$$\lim_{n\to\infty} f(\alpha_n) = 1 - \mathbb{E}_{P_1}[\phi] = f(0).$$

[17] Convexity itself implies continuity in $(0,1)$ for f. In addition, $f(\alpha) \geqslant 0$ and $f(\alpha) \leqslant 1 - \alpha$ implies continuity at 1. Hence, the continuity condition only matters at $x = 0$.

Here the second equality follows from the optimality of ϕ. In fact, this argument works not just for 0 but for arbitrary $\alpha \in [0,1]$.

"if": Given f, we need to find P, Q. The common measurable space is the unit interval $[0,1]$. P is the uniform distribution. Q has density $-f'(1-x)$ on $[0,1)$ and an atom at 1 with $Q[\{1\}] = 1-f(0)$. In fact, Q is constructed to have cdf $f(1-x)$, with the slight twist that the cdf is reset to be 1 at 1, because cdf has to be right continuous.

It's easy to verify that Q is indeed a probability distribution on $[0,1]$ using the properties of f.

The likelihood ratio is simply $-f'(1-x)$ when $x < 1$. At 1, it is 0 if $f(0) = 1$ and $+\infty$ if $f(0) < 1$. By convexity of f it is non-decreasing, so the likelihood ratio rejection regions have the form $[h, 1]$. Type I error is $P[h,1] = 1-h$. Type II error is $Q[0,h) = f(1-h)$.

□

An equivalent object to our trade-off function is the "testing region" used in [KOV17]. For a trade-off function f, we define a special version of epigraph of f as

$$\mathrm{epi}(f) := \{(\alpha, \beta) \mid \alpha \in [0,1], f(\alpha) \leqslant \beta \leqslant 1-\alpha\}.$$

Let P, Q be distributions on Ω. Recall that for a testing rule $\phi : \Omega \to [0,1]$, we denote its type I and type II errors by $\alpha_\phi = \mathbb{E}_P[\phi], \quad \beta_\phi = 1 - \mathbb{E}_Q[\phi]$. It is easy to see that $\mathrm{epi}(f)$ consists of all achievable type I and type II error pairs that is better than random guessing. Namely

$$\mathrm{epi}(f) = \{(\alpha_\phi, \beta_\phi) \mid \phi : \Omega \to [0,1] \text{ measurable}, \alpha_\phi + \beta_\phi \leqslant 1\}.$$

This means considering f and $\mathrm{epi}(f)$ are equivalent. For more information on the testing region, see Chapter 12 of [PW14].

Next we justify the default assumption of symmetry.

Proposition 2.4. *Let a mechanism M be f-DP. Then, M is f^{S}-DP with $f^{\mathrm{S}} = \max\{f, f^{-1}\}$, where the inverse function is defined as*[18]

$$f^{-1}(\alpha) := \inf\{t \in [0,1] : f(t) \leqslant \alpha\} \tag{3}$$

for $\alpha \in [0,1]$.

Lemma A.2. *If $f = T(P,Q)$, then $f^{-1} = T(Q,P)$.*

Proof of Lemma A.2. The lemma is best illustrated from the epigraph point of view. It is an immediate consequence of the following claim: $(\alpha, \beta) \in \mathrm{epi}(f)$ if and only if $(\beta, \alpha) \in \mathrm{epi}(f^{-1})$, which is to say, $f(\alpha) \leqslant \beta \leqslant 1-\alpha$ if and only if $f^{-1}(\beta) \leqslant \alpha \leqslant 1-\beta$. In order to show this, the definition of f^{-1}, together with continuity of f, implies that $f(\alpha) \leqslant \beta \Leftrightarrow f^{-1}(\beta) \leqslant \alpha$. This justfies the claim and hence the lemma. □

Proof of Proposition 2.4. Let S, S' be neighboring datasets. Since M is f-DP, we know that

$$T\big(M(S), M(S')\big) \geqslant f, \quad T\big(M(S'), M(S)\big) \geqslant f. \tag{16}$$

[18] Equation (3) is the standard definition of the left-continuous inverse of a decreasing function. When f is strictly decreasing and $f(0) = 1$ and hence bijective as a mapping, (3) corresponds to the inverse function in the ordinary sense, i.e. $f(f^{-1}(x)) = f^{-1}(f(x)) = x$. However, this is not true in general.

It follows easily from definition that if $f, g \in \mathscr{F}$ satisfy $g \geqslant f$, then $g^{-1} \geqslant f^{-1}$. So by Lemma A.2 and the second inequality in (16),

$$T\big(M(S), M(S')\big) = T\big(M(S'), M(S)\big)^{-1} \geqslant f^{-1}.$$

Together with the first inequality in (16), we see for all neighboring datasets we have

$$T\big(M(S), M(S')\big) \geqslant \max\{f, f^{-1}\}.$$

It is straightforward to verify that if f and g are both convex, continuous, non-increasing and below Id then $\max\{f, g\}$ also satisfy these properties. By Proposition 2.2, we have $\max\{f, g\} \in \mathscr{F}$. So $f^{\mathrm{S}} = \max\{f, f^{-1}\}$ is in $\mathscr{F}$. The proof is complete. □

Recall that Equation (5) states that

$$T\big(\mathcal{N}(0,1), \mathcal{N}(\mu,1)\big)(\alpha) = \Phi\big(\Phi^{-1}(1-\alpha) - \mu\big).$$

Proof of Equation (5). When $\mu \geqslant 0$, likelihood ratio of $\mathcal{N}(\mu,1)$ and $\mathcal{N}(0,1)$ is $\frac{\varphi(x-\mu)}{\varphi(x)} = \mathrm{e}^{\mu x - \frac{1}{2}\mu^2}$, a monotone increasing function in x. So the likelihood ratio tests must be thresholds: reject if the sample is greater than some t and accept otherwise. Assuming $X \sim \mathcal{N}(0,1)$, the corresponding type I and type II errors are

$$\alpha(t) = \mathbb{P}[X > t] = 1 - \Phi(t), \quad \beta(t) = \mathbb{P}[X + \mu \leqslant t] = \Phi(t - \mu).$$

Solving α from t, $t = \Phi^{-1}(1-\alpha)$. So

$$G_\mu(\alpha) = \beta(\alpha) = \Phi\big(\Phi^{-1}(1-\alpha) - \mu\big).$$

□

Lemma 2.9. *For any two distributions P and Q, we have*

$$T\big(\mathrm{Proc}(P), \mathrm{Proc}(Q)\big) \geqslant T(P, Q).$$

Proof of Lemma 2.9. The idea is that whatever can be done with the processed outcome can also be done with the original outcome. Formally, if an optimal test $\phi : Z \to [0,1]$ for the problem $\mathrm{Proc}(P)$ vs $\mathrm{Proc}(Q)$ at level α can achieve type II error $\beta = T\big(\mathrm{Proc}(P), \mathrm{Proc}(Q)\big)(\alpha)$, then it is easy to verify that $\phi \circ \mathrm{Proc} : Y \to [0,1]$ has the same errors α, β for the problem P vs Q. The optimal error $T(P,Q)(\alpha)$ can only be smaller than β. □

The next result is a generalization of Equation (5) and its interesting inverse.

Let P be a probability distribution on $\mathbb{R}$ with density p, cdf $F : \mathbb{R} \to [0,1]$, quantile $F^{-1} : [0,1] \to [-\infty, +\infty]$ and ξ be a random variable from the distribution P. Then we have

Proposition A.3. *$T(\xi, t+\xi)(\alpha) = F(F^{-1}(1-\alpha) - t)$ holds for every $t > 0$ if and only if the density p is log-concave.*

In particular, normal density is log-concave, so the expression of G_μ is a special case.

Proof of Proposition A.3. For convenience let

$$f_t(\alpha) := F(F^{-1}(1-\alpha) - t).$$

"if": This is the easier direction. Fix $t > 0$ and consider the log likelihood ratio of ξ and $t + \xi$:

$$\mathrm{llk} = \log p(x-t) - \log p(x).$$

llk is increasing in x because of log-concavity, so according to Neyman-Pearson lemma, the optimal rejection rule must have the form $1_{[h,+\infty)}$. Hence by a similar calculation as of Gaussian case, the trade-off function indeed has the form f_t.

"only if": We are given that $T(\xi, t+\xi) = f_t$ holds for every $t > 0$, and we want to show that p is log-concave. Now that f_t is a trade-off function for every $t > 0$, it must be convex. By chain rule

$$f_t'(\alpha) = (-1) \cdot \frac{p(F^{-1}(1-\alpha) - t)}{p(F^{-1}(1-\alpha))}.$$

Fix any $t > 0$, convexity implies $f_t'(\alpha)$ is increasing in α for any $\alpha \in [0,1]$. Setting $x = F^{-1}(1-\alpha)$, we know $\frac{p(x-t)}{p(x)}$ is increasing in x for all $x \in \mathbb{R}$, hence also $\log p(x-t) - \log p(x)$.

For convenience let $g = \log p$. We know $g(x-t) - g(x)$ is increasing in $x, \forall t > 0$. Equivalently, $g'(x-t) - g'(x) > 0, \forall x, \forall t > 0$, which means $g'(x)$ is decreasing, i.e. $g = \log p$ is concave. The proof is complete. □

Next we prove results presented in Section 2.4.

Proposition 2.12 (Primal to Dual). *For a symmetric trade-off function f, a mechanism is f-DP if and only if it is $\big(\varepsilon, \delta(\varepsilon)\big)$-DP for all $\varepsilon \geqslant 0$ with $\delta(\varepsilon) = 1 + f^*(-\mathrm{e}^{\varepsilon})$.*

Proof of Proposition 2.12. The tangent line of f with slope k has equation $y = kx - f^*(k)$, so when $k = -\mathrm{e}^{\varepsilon}$ the equation is

$$y = -\mathrm{e}^{\varepsilon} x - f^*(-\mathrm{e}^{\varepsilon}).$$

Compare it to $f_{\varepsilon,\delta}$, we see $1-\delta = -f^*(-\mathrm{e}^{\varepsilon})$. By symmetry, the collection $\{f_{\varepsilon, 1+f^*(-\mathrm{e}^{\varepsilon})}\}_{\varepsilon \geqslant 0}$ envelopes the function f. □

Corollary 2.13. *A mechanism is μ-GDP if and only if it is $\big(\varepsilon, \delta(\varepsilon)\big)$-DP for all $\varepsilon \geqslant 0$, where*

$$\delta(\varepsilon) = 1 - \mathrm{e}^{\varepsilon}\Phi\Big(-\frac{\varepsilon}{\mu} - \frac{\mu}{2}\Big) - \Phi\Big(\frac{\varepsilon}{\mu} - \frac{\mu}{2}\Big).$$

Proof of Corollary 2.13. By Proposition 2.12, μ-GDP is equivalent to $(\varepsilon, 1 + G_\mu^*(-\mathrm{e}^{\varepsilon}))$-DP, so it suffices to compute the expression of $G_\mu^*(-\mathrm{e}^{\varepsilon})$.

Recall that $G_\mu(x) = \Phi\big(\Phi^{-1}(1-x) - \mu\big)$. By definition,

$$G_\mu^*(y) = \sup_{x \in [0,1]} yx - \Phi\big(\Phi^{-1}(1-x) - \mu\big).$$

Let $t = \Phi^{-1}(1-x)$. Equivalently, $x = 1 - \Phi(t) = \Phi(-t)$. Do the change of variable and we have

$$G_\mu^*(y) = \sup_{t \in \mathbb{R}} y\Phi(-t) - \Phi(t-\mu).$$

From the shape of G_μ we know the supremum must be achieved at the unique critical point. Setting the derivative of the objective funciton to be zero yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\big[y\Phi(-t)-\Phi(t-\mu)\big]&=0\\-y\varphi(-t)-\varphi(t-\mu)&=0\\y\mathrm{e}^{-\frac{1}{2}t^2}+\mathrm{e}^{-\frac{1}{2}(t-\mu)^2}&=0\\y+\mathrm{e}^{\mu t-\frac{1}{2}\mu^2}&=0\end{aligned}$$

So $t=\frac{\mu}{2}+\frac{1}{\mu}\log(-y)$. Plug this back in the expression of G_μ^* and we have

$$G_\mu^*(y)=y\Phi\Big(-\frac{\mu}{2}-\frac{1}{\mu}\log(-y)\Big)-\Phi\Big(-\frac{\mu}{2}+\frac{1}{\mu}\log(-y)\Big).$$

When $y=-\mathrm{e}^{\varepsilon}$,

$$G_\mu^*(-\mathrm{e}^{\varepsilon})=-\mathrm{e}^{\varepsilon}\Phi\Big(-\frac{\mu}{2}-\frac{\varepsilon}{\mu}\Big)-\Phi\Big(-\frac{\mu}{2}+\frac{\varepsilon}{\mu}\Big).$$

$1+G_\mu^*(-\mathrm{e}^{\varepsilon})$ agrees with the stated formula in Corollary 2.13. The proof is complete. □

The rest of the section is devoted to group privacy results. The main theorem is

Theorem 2.14. *If a mechanism is f-DP, then it is $\big[1-(1-f)^{\circ k}\big]$-DP for groups of size k. In particular, if a mechanism is μ-GDP, then it is $k\mu$-GDP for groups of size k.*

For convenience we define an operation $\hat{\circ}$, which is function composition with a slight twist. For $f,g\in\mathscr{F}$,

$$f\mathbin{\hat{\circ}}g(x):=f\big(1-g(x)\big).$$

$f^{\hat{\circ}k}$ is defined iteratively:

$$f^{\hat{\circ}k}=\underbrace{f\mathbin{\hat{\circ}}\cdots\mathbin{\hat{\circ}}f}_{k}.$$

Notice that $f\mathbin{\hat{\circ}}g=1-(1-f)\circ(1-g)$, so $f^{\hat{\circ}k}=1-(1-f)^{\circ k}$.

Lemma A.4. *The operation $\hat{\circ}$ has the following properties for $f,g\in\mathscr{F}$:*

(a) $f\mathbin{\hat{\circ}}g\in\mathscr{F}$.

(b) $(f\mathbin{\hat{\circ}}g)^{-1}=(g^{-1})\mathbin{\hat{\circ}}(f^{-1})$. In particular, if $f\in\mathscr{F}^S$, then $f^{\hat{\circ}k}\in\mathscr{F}^S$.

Proof. (a) By Proposition 2.2, it suffices to check the four properties for $f\mathbin{\hat{\circ}}g$. Monotonicity and continuity are obvious. Convexity follows by the well-known fact that decreasing convex function composed with a concave function is convex. Finally, because $f(x)\leqslant 1-x, g(x)\leqslant 1-x$, we have

$$f\mathbin{\hat{\circ}}g(x)=f\big(1-g(x)\big)\leqslant 1-\big(1-g(x)\big)=g(x)\leqslant 1-x.$$

(b) Recall that $f^{-1}(y) = \inf\{x \in [0,1] : f(x) \leqslant y\}$. We have

$$\big[(g^{-1}) \mathbin{\hat{\circ}} (f^{-1})\big](y) = g^{-1}\big(1 - f^{-1}(y)\big) = \inf\{x \in [0,1] : g(x) \leqslant 1 - f^{-1}(y)\}.$$

For any two numbers $x, y \in [0,1]$, we have the following equivalence chain:

$$g(x) \leqslant 1 - f^{-1}(y) \Leftrightarrow f^{-1}(y) \leqslant 1 - g(x) \Leftrightarrow f(1 - g(x)) \leqslant y \Leftrightarrow f \mathbin{\hat{\circ}} g(x) \leqslant y.$$

So

$$\big[(g^{-1}) \mathbin{\hat{\circ}} (f^{-1})\big](y) = \inf\{x \in [0,1] : f \mathbin{\hat{\circ}} g(x) \leqslant y\} = (f \mathbin{\hat{\circ}} g)^{-1}(y).$$

That is, $(g^{-1}) \mathbin{\hat{\circ}} (f^{-1}) = (f \mathbin{\hat{\circ}} g)^{-1}$. The proof is complete.

□

Theorem 2.14 is an immediate consequence of the following lemma:

Lemma A.5. *Suppose $T(P,Q) \geqslant f, T(Q,R) \geqslant g$, then $T(P,R) \geqslant g \mathbin{\hat{\circ}} f$.*

Proof. Fix $\alpha \in [0,1]$. Suppose ϕ is the optimal testing rules of the problem P vs R at the level of α. Then we know the type I error $\mathbb{E}_P[\phi] = \alpha$ and the type II error achieves the optimal value, i.e.

$$1 - \mathbb{E}_R[\phi] = T(P,R)(\alpha).$$

ϕ is suboptimal as a testing rule for the problem Q vs R, so the type I and II errors must be above the trade-off function g. That is,

$$1 - \mathbb{E}_R[\phi] \geqslant T(Q,R)(\mathbb{E}_Q[\phi]) \geqslant g(\mathbb{E}_Q[\phi]).$$

Similarly, ϕ is also suboptimal for the problem P vs Q. So $1 - \mathbb{E}_Q[\phi] \geqslant f(\mathbb{E}_P[\phi]) = f(\alpha)$. Equivalently,

$$\mathbb{E}_Q[\phi] \leqslant 1 - f(\alpha).$$

Put them together

$$\begin{aligned} T(P,R)(\alpha) &= 1 - \mathbb{E}_R[\phi] \\ &\geqslant g(\mathbb{E}_Q[\phi]) \\ &\geqslant g\big(1 - f(\alpha)\big) \qquad (g \text{ is decreasing}) \\ &= g \mathbin{\hat{\circ}} f(\alpha). \end{aligned}$$

This completes the proof. □

Proof of Theorem 2.14. Suppose S and S' are k-neighbors, i.e. there exist datasets $S = S_0, S_1, \ldots, S_k = S'$ such that S_i and S_{i+1} are neighboring or identical for all $i = 0, \ldots, k-1$. By privacy of M, we know $T\big(M(S_i), M(S_{i+1})\big) \geqslant f$. Iteratively apply Lemma A.5 and we have

$$T\big(M(S), M(S_2)\big) \geqslant f \mathbin{\hat{\circ}} f, \quad T\big(M(S), M(S_3)\big) \geqslant f^{\hat{\circ} 3} \quad \ldots \quad T\big(M(S), M(S')\big) \geqslant f^{\hat{\circ} k}.$$

We know that $f^{\hat{\circ} k} = 1 - (1-f)^{\circ k}$, so the f-DP part of the claim is done.

The GDP part of the claim follows by an easy formula: $G_\mu \mathbin{\hat{\circ}} G_{\mu'} = G_{\mu+\mu'}$. To see this, recall that $G_\mu(\alpha) = \Phi(\Phi^{-1}(1-\alpha) - \mu)$.

$$G_\mu \mathbin{\hat{\circ}} G_{\mu'}(\alpha) = G_\mu\big(1 - G_{\mu'}(\alpha)\big) = \Phi\big(\Phi^{-1}(G_{\mu'}(\alpha)) - \mu\big) = \Phi\big(\Phi^{-1}(1-\alpha) - \mu - \mu'\big) = G_{\mu+\mu'}(\alpha).$$

□

In fact, similar conclusion holds for any log-concave noise. See Proposition A.3.

Proposition 2.15. *Fix $\mu \geqslant 0$ and set $\varepsilon = \mu/k$. As $k \to \infty$, we have*

$$1 - (1 - f_{\varepsilon,0})^{\circ k} \to T\big(\mathrm{Lap}(0,1), \mathrm{Lap}(\mu,1)\big).$$

The convergence is uniform over $[0,1]$.

As What makes it even more interesting is the convergence occurs with very small k. In Figure 8 we set $\varepsilon = 0.5$ and $f = 1 - (1 - f_{\varepsilon,0})^{\circ 2}$. So the blue curve in the last panel is $1 - (1 - f)^{\circ 2} = 1 - (1 - f_{\varepsilon,0})^{\circ 4}$. Next we set $\mu = k\varepsilon = 4 \cdot 0.5 = 2$. It turns out these numbers are good enough for the condition $k\varepsilon \to \mu$, because the predicted limit $T\big(\mathrm{Lap}(0,1), \mathrm{Lap}(\mu,1)\big)$ (orange curve in the last panel) is almost indistinguishable from the blue curve $1 - (1 - f_{\varepsilon,0})^{\circ 4}$.

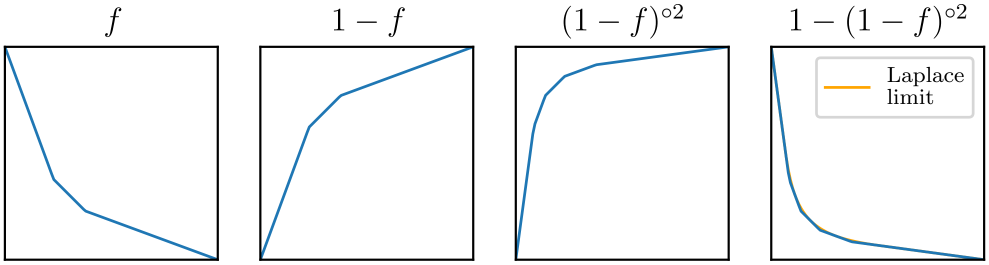

Figure 8: Group privacy corresponds to function composition. Here $f = 1 - (1 - f_{\varepsilon,0})^{\circ 2}$ with $\varepsilon = 0.5$, so the blue curve in the last panel is $1 - (1 - f)^{\circ 2} = 1 - (1 - f_{\varepsilon,0})^{\circ 4}$. Orange curve is the predicted limit $T\big(\mathrm{Lap}(0,1), \mathrm{Lap}(2,1)\big)$. The distinction is almost invisible even when k is only 4.

Lemma A.6. *The trade-off function between Laplace distributions has expression*

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

The graph of this function with $\mu = 1$ is illustrated in Figure 9. In general, it consists of two symmetric line segments: $(0,1)$ connecting $(\mathrm{e}^{-\mu}/2, 1/2)$ and $(1/2, \mathrm{e}^{-\mu}/2)$ connecting $(1,0)$. Then $(\mathrm{e}^{-\mu}/2, 1/2)$ is connected to $(1/2, \mathrm{e}^{-\mu}/2)$ by the reciprocal function. It is easy to check that this function is C^1, i.e. has continuous derivative.

Proof of Lemma A.6. Let F be the cdf of $\mathrm{Lap}(0,1)$. By Proposition A.3,

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

Easy calculation yields

$$F(x) = \begin{cases} \mathrm{e}^{x}/2, & x \leqslant 0, \\ 1 - \mathrm{e}^{-x}/2, & x > 0. \end{cases}$$

So we must expect to divide into several categories. We will refer to the above two expressions as negative and positive regimes.

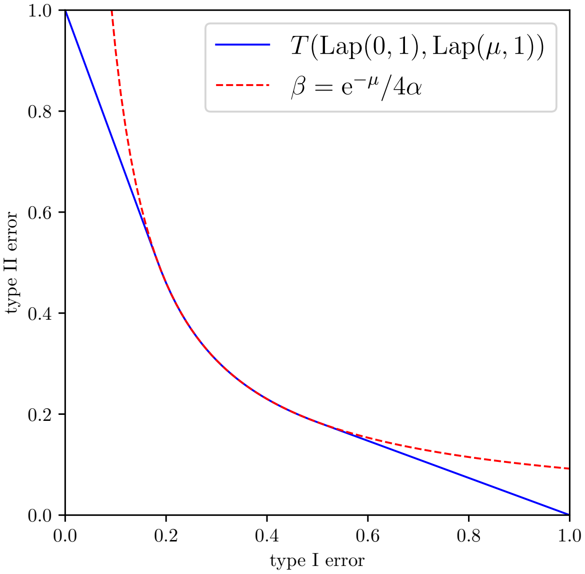

Figure 9: Graph of $T\big(\mathrm{Lap}(0,1), \mathrm{Lap}(\mu,1)\big)$ with $\mu = 1$. It agrees with the reciprocal function in the middle.

When $\alpha > 1/2$, we are in negative regime. Solving $\mathrm{e}^x/2 = 1-\alpha$ gives us $F^{-1}(1-\alpha) = \log 2(1-\alpha) < 0$. An additional $-\mu$ keeps us in negative regime, so

$$F\big(F^{-1}(1-\alpha)-\mu\big) = \exp\big(F^{-1}(1-\alpha)-\mu\big)/2 = \mathrm{e}^{\log 2(1-\alpha)-\mu}/2 = \mathrm{e}^{-\mu}(1-\alpha).$$

When $\alpha \leqslant 1/2$, solving $1-\mathrm{e}^{-x}/2 = 1-\alpha$ gives us $F^{-1}(1-\alpha) = -\log 2\alpha \geqslant 0$. If $-\log 2\alpha - \mu \leqslant 0$, i.e. $\mathrm{e}^{-\mu}/2 \leqslant \alpha$, we are in negative regime and

$$F\big(F^{-1}(1-\alpha)-\mu\big) = \mathrm{e}^{-\log 2\alpha-\mu}/2 = \mathrm{e}^{-\mu}/4\alpha.$$

If $-\log 2\alpha - \mu > 0$, i.e. $\alpha < \mathrm{e}^{-\mu}/2$, we are in positive regime and

$$F\big(F^{-1}(1-\alpha)-\mu\big) = 1-\mathrm{e}^{\log 2\alpha+\mu}/2 = 1-\mathrm{e}^{\mu}\alpha.$$

The proof is complete. □

Proof of Proposition 2.15. For simplicity assume $\mu = 1$. All arguments carry over for general μ.

Let $f_n = 1-f_{\varepsilon,0} = 1-f_{1/n,0}$. Fix x_0 and let $x_{n,k} = f_n^{\circ k}(x_0) = (1-f_{\varepsilon,0})^{\circ k}(x_0)$. We are interested in showing

$$\lim_{n\to\infty} 1-x_{n,n} = T\big(\mathrm{Lap}(0,1), \mathrm{Lap}(1,1)\big)(x_0).$$

First we make a general observation: the sequence $\{x_{n,k}\}$ is increasing in k for any n. This is because $f_{\varepsilon,0}(x) \geqslant 1-x$ and hence $f_n(x) \geqslant x$.

Let $\theta_n = \frac{1}{1+\mathrm{e}^{\frac{1}{n}}}$. By the expression of $f_{\varepsilon,0}$, we obtain the following two dynamics:

$$\begin{array}{rcll} f_n(x) &=& \mathrm{e}^{\frac{1}{n}}x, & \text{if } x \leqslant \theta_n, \\ 1-f_n(x) &=& \mathrm{e}^{-\frac{1}{n}}(1-x), & \text{if } x \geqslant \theta_n. \end{array}$$

The sequence $\{x_{n,k}\}$ evolves according to one of the two formula, potentially different for each k. We will refer to $x \leqslant \theta_n$ case as *linear dynamics* and $x \geqslant \theta_n$ case as *flip linear regime* for evident reason. For any x_0 and n, since $\{x_{n,k}\}$ is increasing in k, there exists a moment such that linear dynamics governs before and flip linear dynamics governs after. Extreme cases are one of the dynamics governs from $k=0$ to n. We divide the analysis into three cases depending on the initial location x_0:

(a) $x_0 < \frac{1}{2\mathrm{e}}$. In this case, for large enough n, the linear dynamics governs all the time. To see this, notice that θ_n increases to $\frac{1}{2}$ as $n \to \infty$. So for large enough n, $x_0 < \frac{1}{\mathrm{e}} \cdot \theta_n$. It's easy to see that $x_{n,k}$ never exceeds θ_n. Hence $x_{n,n} = \mathrm{e}x_0$.

(b) $x_0 \geqslant \frac{1}{2} = \sup_n \theta_n$. $x_{n,k}$ is born above threshold and remains above forever. Flip linear dynamics governs all the n steps, so $1 - x_{n,n} = \mathrm{e}^{-1}(1-x_0)$.

(c) $\frac{1}{2e} \leqslant x < \frac{1}{2}$. Let t be the time of dynamics change. More precisely,

$$t - 1 = \max\{k : \mathrm{e}^{\frac{k}{n}}x \leqslant \theta_n.\} \tag{17}$$

and

$$x_{n,t} = \mathrm{e}^{\frac{1}{n}}x_{n,t-1} = \mathrm{e}^{\frac{t}{n}}x_0, \quad 1 - x_{n,n} = \mathrm{e}^{-\frac{n-t}{n}}(1-x_{n,t}).$$

Taking $n \to \infty$ in (17) (using lim inf and lim sup when necessary), we know $\mathrm{e}^{\frac{t}{n}} \to \frac{1}{2x_0}$. So

$$\lim_{n\to\infty} 1 - x_{n,n} = \lim_{n\to\infty} \mathrm{e}^{\frac{t}{n}-1}(1-x_{n,t}) = \lim_{n\to\infty} \mathrm{e}^{\frac{t}{n}-1}(1-\mathrm{e}^{\frac{t}{n}}x_0) = \mathrm{e}^{-1}\cdot\frac{1}{2x_0}(1-\frac{1}{2}) = \mathrm{e}^{-1}\cdot\frac{1}{4x_0}.$$

Collecting all three cases, we have

$$\lim_{n\to\infty} 1 - x_{n,n} = \begin{cases} 1-\mathrm{e}x_0, & x_0 < \frac{1}{2\mathrm{e}}, \\ \mathrm{e}^{-1}\cdot\frac{1}{4x_0}, & \mathrm{e}^{-\mu}/2 \leqslant \alpha \leqslant 1/2, \\ \mathrm{e}^{-1}(1-x_0), & x_0 \geqslant \frac{1}{2}. \end{cases}$$

By Lemma A.6, this agrees with $T\big(\mathrm{Lap}(0,1), \mathrm{Lap}(1,1)\big)$. Uniform convergence comes for free for trade-off functions once we have pointwise convergence. This is a direct consequence of Lemma A.7 below, which will be used multiple times in this paper. □

Lemma A.7. *Let $f_n : [a,b] \to \mathbb{R}$ be a sequence of non-increasing functions. If f_n has pointwise limit $f : [a,b] \to \mathbb{R}$ where f is continuous on $[a,b]$, then the limit is uniform.*

This is an easy variant of Pólya's theorem ([Pól20]. See also Theorem 2.6.1 in [Leh04]). For completeness, we provide a proof.

Proof of Lemma A.7. We are going to show that for every $\varepsilon > 0$, there exists N such that

$$|f_n(x) - f(x)| < \varepsilon, \quad \forall x \in [a, b], \forall n \geqslant N.$$

Since f is continuous on a closed interval, it is uniformly continuous. So for a fixed $\varepsilon > 0$, we can find $\delta > 0$ such that whenever $x, y \in [a, b]$ satisfies $|x - y| < \delta$, we have $|f(x) - f(y)| < \varepsilon/2$. Then we can divide $[a, b]$ into small intervals $a = x_0 < x_1 < \cdots < x_{m-1} < x_m = b$ such that each interval is shorter than δ. For these $m + 1$ points we can find N such that

$$|f_n(x_i) - f(x_i)| < \frac{\varepsilon}{2}, \quad \forall 0 \leqslant i \leqslant m, \forall n \geqslant N. \tag{18}$$

We claim this N works for our purpose. For any $x \in [a, b]$, there exists a sub-interval that contains it, namely $[x_i, x_{i+1}]$. By monotonicity of f_n we have

$$f_n(x_{i+1}) - f(x) \leqslant f_n(x) - f(x) \leqslant f_n(x_i) - f(x). \tag{19}$$

Now for $n \geqslant N$,

$$\begin{aligned} f_n(x_i) - f(x) &= [f_n(x_i) - f(x_i)] + [f(x_i) - f(x)] \\ &< \frac{\varepsilon}{2} + [f(x_i) - f(x)] && \text{(By (18))} \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. && \text{(uniform continuity of } g\text{)} \end{aligned}$$

This shows the right hand side of (19) is less than ε. A similar argument for the left hand side yields

$$|f_n(x) - f(x)| < \varepsilon,$$

which justifies the choice of N. □

B Conversion from f-DP to divergence based DP

As the title suggests, the central question of this section is the conversion from f-DP to divergence based DP. It boils down to the conversion from trade-off functions to various divergences. We first introduce the most general tool, and then give explicit formula for a large class of divergences, including Rényi divergence. Finally we argue that f-DP is easier to use than RDP from conversion perspective.

Suppose we have a "divergence" $D(\cdot\|\cdot)$, which takes in a pair of probability distributions on a common measurable space and outputs a number. We say D satisfies data processing inequality if $D\big(\mathrm{Proc}(P)\|\mathrm{Proc}(Q)\big) \geqslant D(P\|Q)$ for any post-processing Proc.

Proposition B.1. *If $D(\cdot\|\cdot)$ satisfies data processing inequality, then there exists a functional $l_D : \mathscr{F} \to \mathbb{R}$ that computes D through the trade-off function:*

$$D(P\|Q) = l_D\big(T(P, Q)\big).$$

Proof. It's almost immediate from the following

Lemma B.2. *If $T(P', Q') \geqslant T(P, Q)$, then $D(P'\|Q') \leqslant D(P\|Q)$. In particular, $T(P, Q) = T(P', Q')$ implies $D(P\|Q) = D(P'\|Q')$.*

To see why the lemma holds, notice by Blackwell's theorem, $T(P',Q') \geqslant T(P,Q)$ implies that there is a Proc such that $P' = \mathrm{Proc}(P), Q' = \mathrm{Proc}(Q)$, and by data processing inequality, $D(P'\|Q') \leqslant D(P\|Q)$.

The lemma implies the existence of l_D because we can define $l_D(f) = D(P\|Q)$ through any pair P,Q such that $T(P,Q) = f$. This definition is independent of the choice of P and Q. □

An immediate corollary is

Corollary B.3. *If two trade-off functions f,g satisfy $f \geqslant g$, then $l_D(f) \leqslant l_D(g)$.*

Example: F-divergence Let P,Q be a pair of distributions with density p and q with respect to some common dominating measure. For a convex $F : (0,+\infty) \to \mathbb{R}$ such that $F(1) = 0$, the F-divergence $D_F(P\|Q)$ is defined as (see [LV06])

$$D_F(P\|Q) = \int_{\{pq>0\}} F\Big(\frac{p}{q}\Big)\,\mathrm{d}Q + F(0)Q[p=0] + \tau_F P[q=0]$$

where $F(0) = \lim_{t\to 0^+} F(t)$ and $\tau_F := \lim_{t\to+\infty}\frac{F(t)}{t}$. We further set the rules $F(0)\cdot 0 = \tau_F \cdot 0 = 0$ even if $F(0) = +\infty$ or $\tau_F = +\infty$.

Proposition B.4. *Let $z_f = \inf\{x \in [0,1] : f(x) = 0\}$ be the first zero of f. The functional $l_F : \mathscr{F} \to \mathbb{R}$ that computes F-divergence has expression*

$$l_F(f) = \int_0^{z_f} F\big(\big|f'(x)\big|^{-1}\big)\cdot\big|f'(x)\big|\,\mathrm{d}x + F(0)\cdot(1-f(0)) + \tau_F\cdot(1-z_f).$$

In particular, when $f \in \mathscr{F}^S$ and $f(0) = 1$, we have

$$l_F(f) = \int_0^1 F\big(\big|f'(x)\big|^{-1}\big)\cdot\big|f'(x)\big|\,\mathrm{d}x. \tag{20}$$

Proof of Proposition B.4. For a given trade-off function f, in order to determine $l_F(f)$, it suffices to find P,Q such that $f = T(P,Q)$ and then use the property $l_F(f) = D_F(P\|Q)$. Such a pair is constructed in the proof of Proposition 2.2: P is the uniform distribution on $[0,1]$ and Q has density $|f'(1-x)|$ on $[0,1)$ and an atom at 1 with $Q[\{1\}] = 1 - f(0)$. When we set the dominating measure μ to be Lebesgue in $[0,1)$ and have an atom at 1 with measure 1, the densities p and q have expressions

$$p(x) = \begin{cases} 1, & x \in [0,1), \\ 0, & x = 1. \end{cases} \qquad \text{and} \qquad q(x) = \begin{cases} |f'(1-x)|, & x \in [0,1), \\ 1-f(0), & x = 1. \end{cases}$$

Readers should keep in mind that the value at 1 matters because the base measure μ has an atom there. For a trade-off funciton f, its derivative $f'(x)$ never vanishes before f hits zero, i.e. $f'(x) > 0$ for $x < z_f$ and $f'(x) = 0$ for $x \geqslant z_f$. Equivalently, $\{q > 0\} = (1-z_f, 1]$ and $\{q = 0\} = [0, 1-z_f]$. So

$$\begin{aligned} D_F(P\|Q) &= \int_{\{pq>0\}} F\Big(\frac{p}{q}\Big)\,\mathrm{d}Q + F(0)Q[p=0] + \tau_F P[q=0] \\ &= \int_{1-z_f}^1 F(|f'(1-x)|^{-1})\cdot|f'(1-x)|\,\mathrm{d}x + F(0)\cdot(1-f(0)) + \tau_F\cdot(1-z_f) \\ &= \int_0^{z_f} F(|f'(x)|^{-1})\cdot|f'(x)|\,\mathrm{d}x + F(0)\cdot(1-f(0)) + \tau_F\cdot(1-z_f). \end{aligned}$$

Starting from the second line, the integral is Lebesgue integral. Now the proof is complete. □

Because of the generality of F-divergence, Equation (20) has broad applications. Many important divergences can be computed via a simple formula. Below are some of the examples.

- **Total variation distance** corresponds to $F(t) = \frac{1}{2}|t-1|$. Easy calculation yields

$$l_{\mathrm{TV}}(f) = \frac{1}{2}\int_0^1 \big|1+f'(x)\big|\,\mathrm{d}x.$$

- **KL divergence** corresponds to $F(t) = t\log t$. We have

$$l_{\mathrm{KL}}(f) = -\int_0^1 \log\big|f'(x)\big|\,\mathrm{d}x.$$

  This functional plays an important role in our central limit theorem. We call it $\mathrm{kl}(f)$ there.

- **Power divergence** of order α corresponds to $F_\alpha(t) = \frac{t^\alpha-\alpha(t-1)-1}{\alpha(\alpha-1)}$. The corresponding functional is

$$l_{F_\alpha}(f) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\Big(\int_0^1 \big|f'(x)\big|^{1-\alpha}\,\mathrm{d}x - 1\Big), & z_f = 1, \\ +\infty, & z_f < 1. \end{cases}$$

- **Rényi divergence** of order α is defined as

$$D_\alpha(P\|Q) = \tfrac{1}{\alpha-1}\log\big(\mathbb{E}_P(\tfrac{p}{q})^{\alpha-1}\big) = \tfrac{1}{\alpha-1}\log\int p^\alpha q^{1-\alpha}.$$

  It is related to power divergence of order α by

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\cdot\log\big(\alpha(\alpha-1)D_{F_\alpha}(P\|Q)+1\big). \tag{21}$$

  So the corredponding functional, which we denote by $l_\alpha^{\text{Rényi}}$, has expression

$$l_\alpha^{\text{Rényi}}(f) = \begin{cases} \frac{1}{\alpha-1}\log\int_0^1 \big|f'(x)\big|^{1-\alpha}\,\mathrm{d}x, & z_f = 1, \\ +\infty, & z_f < 1. \end{cases} \tag{22}$$

Proof of Equation (21).

$$\begin{aligned} D_{F_\alpha}(P\|Q) &= \int q\cdot F_\alpha\Big(\frac{p}{q}\Big) \\ &= \int q\cdot\frac{(\frac{p}{q})^\alpha - \alpha(\frac{p}{q}-1)-1}{\alpha(\alpha-1)} \\ &= \frac{1}{\alpha(\alpha-1)}\cdot\int p^\alpha q^{1-\alpha} + 0 - \frac{1}{\alpha(\alpha-1)} \\ &= \frac{1}{\alpha(\alpha-1)}\Big(e^{(\alpha-1)D_\alpha(P\|Q)}-1\Big). \end{aligned}$$

Solving for $D_\alpha(P\|Q)$ yields (21). □

Introduced in [Mir17], a mechanism M is said to be (α, ε)-Rényi differentially private (RDP) if for all neighboring pairs S, S' it holds that

$$D_\alpha(M(S)\|M(S')) \leqslant \varepsilon, \tag{23}$$

A few other DP definitions, including zero concentrated DP (zCDP) [BS16] and truncated concentrated DP (tCDP) [BDRS18], are defined through imposing bounds in the form of (23) with certain collections of α. In general, conversion from f-DP to RDP can be done via Lemma B.2 and Equation (22). Below is the corollary for the most useful case — GDP.

Proposition B.5. *If a mechanism is μ-GDP, then it is $(\alpha, \frac{1}{2}\mu^2\alpha)$-RDP for any $\alpha > 1$.*

Proof. Fix neighboring datasets S and S', if M is μ-GDP then

$$T\big(M(S), M(S')\big) \geqslant T\big(\mathcal{N}(0,1), \mathcal{N}(\mu,1)\big).$$

By Lemma B.2, this implies

$$D_\alpha\big(M(S)\|M(S')\big) \leqslant D_\alpha\big(\mathcal{N}(0,1)\|\mathcal{N}(\mu,1)\big).$$

Easy calculation shows $D_\alpha\big(\mathcal{N}(0,1)\|\mathcal{N}(\mu,1)\big) = \frac{1}{2}\mu^2\alpha$ as long as $\alpha > 0$ and $\alpha \neq 1$. Readers can refer to Proposition 7 in [Mir17] for a detailed derivation. In any case, we have

$$D_\alpha\big(M(S)\|M(S')\big) \leqslant \frac{1}{2}\mu^2\alpha,$$

which means M is $(\alpha, \frac{1}{2}\mu^2\alpha)$-RDP for any $\alpha > 1$. □

The functional l_D allows a consistent, easy conversion from an f-DP guarantee to all divergence based DP guarantees. The above GDP–RDP conversion is an example. On the other hand, conversion from divergence, either to trade-off function or to other divergences, often requires case by case analysis, sometimes significantly non-trivial. What's worse is that it is often hard to tell whether a given conversion between divergences is improvable or already lossless. For conversion between F-divergences, a systematic approach called joint range is developed in [HV11], but it is still significantly more complicated than Equation (20). On the other hand, Proposition B.1 means conversion from trade-off to divergence is lossless and unimprovable.

This root role of trade-off function (see also Section 2.3) is somewhat expected: it summarizes the distinguishability of a pair of distribution by a *function*, which is an infinite dimensional object. In contrast, divergences usually just summarize by a number, which is obviously less informative by a function.

But Rényi / power divergence is an infinite collection of divergences, indexed by the order α. What if we think of $D_\alpha(P\|Q)$ as a function of α? Is it as informative as the trade-off function $T(P, Q)$? Is it true that something like Lemma B.2 holds, i.e. whenever $D_\alpha(P\|Q) \leqslant D_\alpha(P'\|Q')$ we can conclude $D(P\|Q) \leqslant D(P'\|Q')$ for all divergences with data processing inequality? The following counterexample answers this question negatively.

Let P_ε and Q_ε denote Bernoulli distributions with success probabilities $\frac{\mathrm{e}^\varepsilon}{1+\mathrm{e}^\varepsilon}$ and $\frac{1}{1+\mathrm{e}^\varepsilon}$, respectively.

Proposition B.6. *For every $0 < \varepsilon < 4$, the following two statements are both true:*

(a) For all $\alpha > 1$, $D_\alpha(P_\varepsilon \| Q_\varepsilon) \leqslant D_\alpha\big(\mathcal{N}(0,1) \| \mathcal{N}(\varepsilon,1)\big)$;

(b) $\mathrm{TV}(P_\varepsilon, Q_\varepsilon) > \mathrm{TV}\big(\mathcal{N}(0,1), \mathcal{N}(\varepsilon,1)\big)$.

Surprisingly, although the whole collection of Rényi divergences would assert that the pair $P_\varepsilon, Q_\varepsilon$ are "harder to distinguish" than $\mathcal{N}(0,1), \mathcal{N}(\varepsilon,1)$, one nevertheless can achieve smaller summed type I and type II errors when trying to distinguish $P_\varepsilon, Q_\varepsilon$. As an aside, the example considered in Proposition B.6 is admittedly not pathological because Bernoulli and normal distributions are commonly used in randomized algorithms.

We point out that (a) in Proposition B.6 in fact holds for all $\varepsilon \geqslant 0$, which is proved in [BS16], partially based on numerical evidence. Our proof is analytical.

Proof of Proposition B.6.

$$\begin{aligned} D_\alpha(P_\varepsilon \| Q_\varepsilon) &= \frac{1}{\alpha-1} \log(p^\alpha q^{1-\alpha} + q^\alpha p^{1-\alpha}) \\ &= \frac{1}{\alpha-1} \log \frac{\mathrm{e}^{\varepsilon\alpha} + \mathrm{e}^{\varepsilon(1-\alpha)}}{1+\mathrm{e}^{\varepsilon}}. \\ &= \frac{1}{\alpha-1} \log \frac{\mathrm{e}^{\varepsilon(\alpha-\frac{1}{2})} + \mathrm{e}^{\varepsilon(\frac{1}{2}-\alpha)}}{\mathrm{e}^{-\frac{\varepsilon}{2}} + \mathrm{e}^{\frac{\varepsilon}{2}}} \\ &= \frac{1}{\alpha-1} \log \frac{\cosh \varepsilon(\alpha - \frac{1}{2})}{\cosh \frac{\varepsilon}{2}} \end{aligned}$$

Now we claim that $\cosh x \cdot \mathrm{e}^{-\frac{1}{2}x^2}$ is monotone decreasing for $x \geqslant 0$. To see this, simply take the derivative

$$\big(\cosh x \cdot \mathrm{e}^{-\frac{1}{2}x^2}\big)' = \sinh x \cdot \mathrm{e}^{-\frac{1}{2}x^2} + \cosh x \cdot (-x) \cdot \mathrm{e}^{-\frac{1}{2}x^2} = (\tanh x - x) \cdot \cosh x \cdot \mathrm{e}^{-\frac{1}{2}x^2}.$$

It is easy to show $\tanh x \leqslant x$ for $x \geqslant 0$. Hence the derivative is always non-positive, which justifies the claimed monotonicity. Since $\alpha > 1, \varepsilon \leqslant 0$, we have $\varepsilon(\alpha - \frac{1}{2}) \geqslant \frac{\varepsilon}{2}$. By the monotonicity,

$$\cosh \varepsilon(\alpha - \frac{1}{2}) \cdot \mathrm{e}^{-\frac{1}{2}\varepsilon^2(\alpha-\frac{1}{2})^2} \leqslant \cosh \frac{\varepsilon}{2} \cdot \mathrm{e}^{-\frac{1}{2}\cdot(\frac{\varepsilon}{2})^2}.$$

That is,

$$\frac{\cosh \varepsilon(\alpha - \frac{1}{2})}{\cosh \frac{\varepsilon}{2}} \leqslant \mathrm{e}^{\frac{1}{2}\varepsilon^2\alpha(\alpha-1)}.$$

So

$$D_\alpha(P_\varepsilon \| Q_\varepsilon) = \frac{1}{\alpha-1} \cdot \log \frac{\cosh \varepsilon(\alpha - \frac{1}{2})}{\cosh \frac{\varepsilon}{2}} \leqslant \frac{1}{2}\varepsilon^2\alpha = D_\alpha\big(\mathcal{N}(0,1) \| \mathcal{N}(\varepsilon,1)\big).$$

For the second part, easy calculation yields

$$\begin{aligned} \mathrm{TV}(P_\varepsilon, Q_\varepsilon) &= \frac{\mathrm{e}^\varepsilon - 1}{\mathrm{e}^\varepsilon + 1} = \tanh \frac{\varepsilon}{2}, \\ \mathrm{TV}\big(\mathcal{N}(0,1), \mathcal{N}(\varepsilon,1)\big) &= 1 - 2\Phi(-\frac{\varepsilon}{2}). \end{aligned}$$

$\mathrm{TV}(P_\varepsilon, Q_\varepsilon) \geqslant \mathrm{TV}\big(\mathcal{N}(0,1), \mathcal{N}(\varepsilon,1)\big)$ for $\varepsilon < 4$ can be verified numerically. □

C A Self-contained Proof of the Composition Theorem

In this section we prove the well-definedness of $\otimes$ and Composition Theorem 3.2.

We begin with the setting of the key lemma, which compares indistinguishability of two pairs of *randomized algorithms*. Let $K_1, K_1' : Y \to Z_1$ and $K_2, K_2' : Y \to Z_2$ be two pairs of randomized algorithms. Suppose the following is true for these four algorithms: for each fixed input $y \in Y$, testing problem $K_1(y)$ vs $K_1'(y)$ is harder than $K_2(y)$ vs $K_2'(y)$. In mathematical language, let $f_i^y = T\big(K_i(y), K_i'(y)\big)$ (See the left panel of Figure 10). The above assumption amounts to saying $f_1^y \geqslant f_2^y$. So far we have fixed the input y. In the two pairs of testing problems, if the input of the null comes from P and the input of the alternative comes from P', then intuitively both testing problems become easier than when inputs are fixed, because now the inputs also provide information. Formally, the observation comes from input-output joint distribution $\big(P, K_i(P)\big)$ or $\big(P', K_i'(P')\big)$ (with a little abuse of notation). Let $f_i = T\big((P, K_i(P)), (P', K_i'(P'))\big), i = 1, 2$ be the trade-off functions of the joint testing problems (See the right panel of Figure 10). As discussed, we expect that $f_1 \leqslant f_1^y, f_2 \leqslant f_2^y$ for all y. But what about f_1 and f_2? Which joint testing problem is harder? The following lemma answers the question.

Lemma C.1. *If $f_1^y \geqslant f_2^y$ for all $y \in Y$, then $f_1 \geqslant f_2$.*

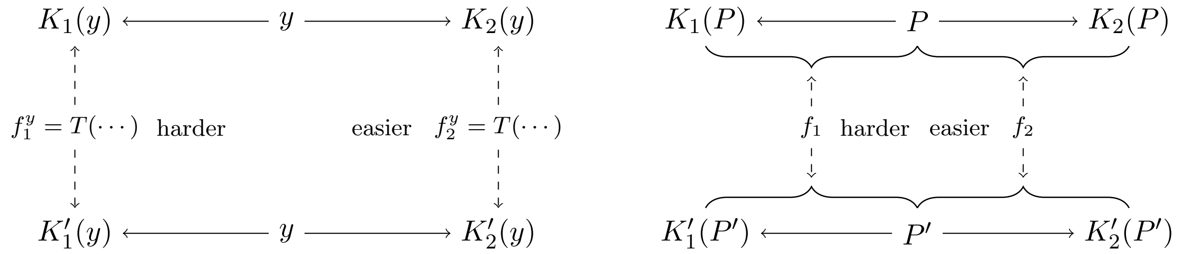

Figure 10: Assumption (left) and conclusion (right) of Lemma C.1. Solid arrows indicate (random) mapping and dashed arrows indicate the trade-off function of the two ends. For example, f_1 in the right panel is the trade-off funciton of two joint distributions: $\big(P, K_1(P)\big)$ and $\big(P', K_1(P')\big)$.

Let's first use the lemma to show the well-definedness of $\otimes$ and the composition theorem. Its own proof comes afterwards. Recall that in Definition 3.1, $f \otimes g$ is defined as $T(P \times P', Q \times Q')$ if $f = T(P, Q), g = T(P', Q')$. To show this definition does not depend on the choice of P, Q and P', Q', it suffices to verify that when $f = T(P, Q) = T(\tilde{P}, \tilde{Q})$, we have $T(P \times P', Q \times Q') = T(\tilde{P} \times P', \tilde{Q} \times Q')$. The following lemma is slightly stronger than what we need, but will be useful later.

Lemma C.2. *If $T(P, Q) \geqslant T(\tilde{P}, \tilde{Q})$, then*

$$T(P \times P', Q \times Q') \geqslant T(\tilde{P} \times P', \tilde{Q} \times Q').$$

As a consequence, if the assumption holds with an equality, then so does the conclusion.

Proof. In order to fit it into the setting of Lemma C.1, let the algorithms output a random variable independent of the input y. See Figure 11. The input-output joint distributions are just product distributions, so by the comparison lemma C.1,

$$T(P \times P', Q \times Q') \geqslant T(\tilde{P} \times P', \tilde{Q} \times Q').$$

$$\begin{tikzcd}
P \arrow[d, dashed, leftrightarrow, "T(P,Q)"] & y \arrow[l] \arrow[r] & \tilde{P} \arrow[d, dashed, leftrightarrow, "T(\tilde{P},\tilde{Q})"] \\
Q & y \arrow[l] \arrow[r] & \tilde{Q}
\end{tikzcd}$$

$$\begin{tikzcd}
P & & P' \arrow[ll] \arrow[rr] & & \tilde{P} \\
& P \times P' \arrow[d, dashed, leftrightarrow] & & \tilde{P} \times P' \arrow[d, dashed, leftrightarrow] & \\
& Q \times Q' & & \tilde{Q} \times Q' & \\
Q & & Q' \arrow[ll] \arrow[rr] & & \tilde{Q}
\end{tikzcd}$$

Figure 11: Lemma C.1 implies well-definedness of $\otimes$.

When $T(P,Q) = T(\tilde{P},\tilde{Q})$, we can apply the lemma in both directions and conclude that

$$T(P \times P', Q \times Q') = T(\tilde{P} \times P', \tilde{Q} \times Q').$$

The proof is complete. $\square$

Now that we have justified the definition of the composition tensor $\otimes$, lemma C.2 can be written in a concise way:

$$g_1 \geqslant g_2 \Rightarrow f \otimes g_1 \geqslant f \otimes g_2. \tag{24}$$

This is actually the second property we listed after the definition of $\otimes$.

For composition theorem, we prove the following two steps version:

Lemma C.3. *Suppose in a two-step composition, the two components $M_1 : X \to Y, M_2 : X \times Y \to Z$ satisfy*

1. *M_1 is f-DP;*
2. *$M_2(\cdot, y) : X \to Z$ is g-DP for each fixed $y \in Y$.*

Then the composition $M : X \to Y \times Z$ is $f \otimes g$-DP.

$$\begin{tikzcd}
M_2(S,y) \arrow[d, dashed, leftrightarrow, "T(\cdots)"] & y \arrow[l] \arrow[r] & Q \arrow[d, dashed, leftrightarrow, "g"] \\
M_2(S',y) & y \arrow[l] \arrow[r] & Q'
\end{tikzcd}$$

$$\begin{tikzcd}
M_2\big(S, M_1(S)\big) & & M_1(S) \arrow[ll] \arrow[rr] & & Q \\
& M(S) \arrow[d, dashed, leftrightarrow] & & M_1(S) \otimes Q \arrow[d, dashed, leftrightarrow] & \\
& M(S') & & M_1(S') \otimes Q' & \\
M_2\big(S', M_1(S')\big) & & M_1(S') \arrow[ll] \arrow[rr] & & Q'
\end{tikzcd}$$

Figure 12: Lemma C.1 implies Lemma C.3.

Proof of Lemma C.3. Let Q, Q' be distributions such that $g = T(Q, Q')$. Fix a pair of neighboring datasets S and S' and set everything as in Figure 12. The input y is an element in the output space of M_1. Arrows to the left correspond to the mechanism M_2, while arrows to the right ignore the input y and output Q, Q' respectively.

Here f_1^y in Lemma C.1 is $T\big(M_2(S,y), M_2(S',y)\big) \geqslant g$, so the condition in Lemma C.1 checks. Consequently,

$$\begin{aligned}
T\big(M(S), M(S')\big) &\geqslant T\big(M_1(S) \times Q, M_1(S') \times Q'\big) && \text{(Lemma C.1)} \\
&= T\big(M_1(S), M_1(S')\big) \otimes T(Q, Q') && \text{(Def. of } \otimes) \\
&= T\big(M_1(S), M_1(S')\big) \otimes g \\
&\geqslant f \otimes g && \text{(Privacy of } M_1 \text{ and (24))}
\end{aligned}$$

The proof is complete. □

Now we prove Lemma C.1. The proof is basically careful application of Neyman-Pearson Lemma A.1.

Proof of Lemma C.1. In order to further simplify the notations, for $i = 1, 2$, let μ_i and μ_i' be the joint distributions $\big(P, K_i(P)\big)$ and $\big(P', K_i'(P')\big)$ respectively. Then $f_1 = T(\mu_1, \mu_1'), f_2 = T(\mu_2, \mu_2')$ and we need to show that the testing problem μ_1 vs μ_1' is harder than μ_2 vs μ_2'.

Consider the testing problem μ_1 vs μ_1'. For $\alpha \in [0,1]$, let $\phi_1 : Y \times Z_1 \to [0,1]$ be the optimal rejection rule at level α. By definition of trade-off function, the power of this test is $1 - f_1(\alpha)$. Formally,

$$\mathbb{E}_{\mu_1}[\phi_1] = \alpha, \quad \mathbb{E}_{\mu_1'}[\phi_1] = 1 - f_1(\alpha).$$

It suffices to construct a rejection rule $\phi_2 : Y \times Z_2 \to [0,1]$ for the problem μ_2 vs μ_2', at the same level α but with greater power, i.e.

$$\mathbb{E}_{\mu_2}[\phi_2] = \alpha \quad \text{and} \quad \mathbb{E}_{\mu_2'}[\phi_2] \geqslant \mathbb{E}_{\mu_1'}[\phi_1] = 1 - f_1(\alpha).$$

If such ϕ_2 exists, then by the sub-optimality of ϕ_2 for the problem μ_2 vs μ_2',

$$1 - f_2(\alpha) \geqslant \mathbb{E}_{\mu_2'}[\phi_2] \geqslant 1 - f_1(\alpha),$$

which is what we want.

For $y \in Y$, let $\phi_1^y : Z_1 \to [0,1]$ be the slice of ϕ_1 at y, i.e. $\phi_1^y(z_1) = \phi_1(y, z_1)$. This is a rejection rule for the problem $K_1(y)$ vs $K_1'(y)$, sub-optimal in general. The type I error is

$$\alpha^y := \mathbb{E}_{z_1 \sim K_1(y)}[\phi_1^y(z_1)].$$

The power is

$$\mathbb{E}_{z_1 \sim K_1'(y)}[\phi_1^y(z_1)] \leqslant 1 - f_1^y(\alpha^y).$$

The last inequality holds because $f_1^y = T\big(K_1(y), K_1'(y)\big)$ and that ϕ_1^y is sub-optimal for this problem. Let $\phi_2^y : Z_2 \to [0,1]$ be the optimal rejection rule for the testing $K_2(y)$ vs $K_2'(y)$ at level α^y. Construction of $\phi_2 : Y \times Z_2 \to [0,1]$ is simply putting together these slices ϕ_2^y. Formally, $\phi_2(y, z_2) = \phi_2^y(z_2)$. Its level is α because α^y are averaged in terms of the same distribution P. More precisely,

$$\begin{aligned}
\mathbb{E}_{\mu_2}[\phi_2] &= \mathbb{E}_{y \sim P}\big[\mathbb{E}_{z_2 \sim K_2(y)}[\phi_2^y(z_2)]\big] && \text{(Construction of } \phi_2) \\
&= \mathbb{E}_{y \sim P}[\alpha^y] && (\phi_2^y \text{ has level } \alpha^y) \\
&= \mathbb{E}_{y \sim P}\big[\mathbb{E}_{z_1 \sim K_1(y)}[\phi_1^y(z_1)]\big] && \text{(Def. of } \alpha^y) \\
&= \mathbb{E}_{\mu_1}[\phi_1] = \alpha.
\end{aligned}$$

Let's compute its power:

$$\begin{aligned}
\mathbb{E}_{\mu_2'}[\phi_2] &= \mathbb{E}_{y\sim P}\big[\mathbb{E}_{z_2\sim K_2'(y)}[\phi_2^y(z_2)]\big] \\
&= \mathbb{E}_{y\sim P}\big[1-f_2^y(\alpha^y)\big] && (\phi_2^y \text{ is optimal}) \\
&\geqslant \mathbb{E}_{y\sim P}\big[1-f_1^y(\alpha^y)\big] && (f_1^y \geqslant f_2^y) \\
&\geqslant \mathbb{E}_{y\sim P}\big[\mathbb{E}_{z_1\sim K_1'(y)}[\phi_1^y(z_1)]\big] && (\phi_1^y \text{ is sub-optimal}) \\
&= \mathbb{E}_{\mu_1'}[\phi_1] = 1-f_1(\alpha). && (\text{Optimality of } \phi_1 \text{ for } \mu_1 \text{ vs } \mu_1')
\end{aligned}$$

So ϕ_2 constructed this way does have the desired level and power. The proof is complete. □

D Omitted Proofs in Section 3

We first collect the basic properties of $\otimes$ listed in Section 3.1.

Proposition D.1. *The product $\otimes$ defined in Definition 3.1 has the following properties:*

0. *The product $\otimes$ is well-defined.*
1. *The product $\otimes$ is commutative and associative.*
2. *If $g_1 \geqslant g_2$, then $f\otimes g_1 \geqslant f\otimes g_2$.*
3. *$f\otimes \mathrm{Id} = \mathrm{Id}\otimes f = f$.*
4. *$(f\otimes g)^{-1} = f^{-1}\otimes g^{-1}$.*
5. *For GDP, $G_{\mu_1}\otimes G_{\mu_2}\otimes\cdots\otimes G_{\mu_n} = G_\mu$, where $\mu = \sqrt{\mu_1^2+\cdots+\mu_n^2}$.*

Property 0 and 2 are already proved in Appendix C. So we only prove 1,3,4,5 here.

Proof of Properties (1,3,4,5). We will assume $f = T(P,P'), g = T(Q,Q')$ in the entire proof. The upshot is that

$$T(P,P')\otimes T(Q,Q') = T(P\times Q, P'\times Q').$$

1. Commutativity:

   $$f\otimes g = T(P,P')\otimes T(Q,Q') = T(P\times Q, P'\times Q') \overset{(a)}{=} T(Q\times P, Q'\times P') = T(Q,Q')\otimes T(P,P') = g\otimes f.$$

   In step (a), we switch the order of the components of the product, which obviously keeps the trade-off function unchanged.

   Associativity: Let $h = T(R,R')$.

   $$\begin{aligned}
   (f\otimes g)\otimes h &= T(P\times Q, P'\times Q')\otimes T(R,R') = T(P\times Q\times R, P'\times Q'\times R') \\
   f\otimes (g\otimes h) &= T(P,P')\otimes T(Q\times R, Q'\times R') = T(P\times Q\times R, P'\times Q'\times R')
   \end{aligned}$$

   So $(f\otimes g)\otimes h = f\otimes (g\otimes h)$.

2. Let R be an arbitrary degenerate distribution, i.e. R puts mass 1 on a single point. Then $\mathrm{Id} = T(R,R)$ and

   $$f\otimes \mathrm{Id} = T(P\times R, P'\times R) = T(P,P') = f.$$

3. By Lemma A.2, taking the inverse amounts to flipping the arguments of $T(\cdot,\cdot)$.
$$(f\otimes g)^{-1} = T(P'\times Q', P\times Q) = T(P',P)\otimes T(Q',Q) = f^{-1}\otimes g^{-1}.$$

4. Let $\boldsymbol{\mu}=(\mu_1,\mu_2)\in\mathbb{R}^2$ and I_2 be the 2×2 identity matrix. Then
$$\begin{aligned}G_{\mu_1}\otimes G_{\mu_2} &= T\big(\mathcal{N}(0,1),\mathcal{N}(\mu_1,1)\big)\otimes T\big(\mathcal{N}(0,1),\mathcal{N}(\mu_1,1)\big)\\ &= T\big(\mathcal{N}(0,1)\times\mathcal{N}(0,1),\mathcal{N}(\mu_1,1)\times\mathcal{N}(\mu_2,1)\big)\\ &= T\big(\mathcal{N}(0,I_2),\mathcal{N}(\boldsymbol{\mu},I_2)\big)\end{aligned}$$
Again we use the invariance of trade-off functions under invertible transformations. $\mathcal{N}(0,I_2)$ is rotation invariant, So we can rotate $\mathcal{N}(\boldsymbol{\mu},I_2)$ so that the mean is $(\sqrt{\mu_1^2+\mu_2^2},0)$. Continuing the calculation
$$\begin{aligned}G_{\mu_1}\otimes G_{\mu_2} &= T\big(\mathcal{N}(0,I_2),\mathcal{N}(\boldsymbol{\mu},I_2)\big)\\ &= T\big(\mathcal{N}(0,1)\times\mathcal{N}(0,1),\mathcal{N}(\sqrt{\mu_1^2+\mu_2^2},1)\times\mathcal{N}(0,1)\big)\\ &= T\big(\mathcal{N}(0,1),\mathcal{N}(\sqrt{\mu_1^2+\mu_2^2},1)\big)\otimes T\big(\mathcal{N}(0,1),\mathcal{N}(0,1)\big)\\ &= G_{\sqrt{\mu_1^2+\mu_2^2}}\otimes \mathrm{Id}\\ &= G_{\sqrt{\mu_1^2+\mu_2^2}}.\end{aligned}$$

□

The following proposition explains why our central limit theorems need f_n to approach Id.

Proposition D.2. *For any trade-off function f that is not* Id*,*
$$\lim_{n\to+\infty} f^{\otimes n}(\alpha)=0,\quad \forall\alpha\in(0,1].$$
In fact, the convergence is exponentially fast.

Proof of Proposition D.2. For any trade-off function f, let P,Q be probability measures such that $T(P,Q)=f$. The existence is guaranteed by Proposition 2.2. It is well-known that $1-\mathrm{TV}(P,Q)$ is the minimum sum of type I and type II error, namely,
$$1-\mathrm{TV}(P,Q)=\min_{\alpha\in[0,1]}\alpha+f(\alpha).$$
We claim that the following limit suffices to prove the theorem:
$$\lim_{n\to\infty}\mathrm{TV}(P^n,Q^n)=1. \tag{25}$$
To see why it suffices, recall that by definition $T(P^n,Q^n)=f^{\otimes n}$. Hence
$$1-\mathrm{TV}(P^n,Q^n)=\min_{\alpha\in[0,1]}\alpha+f^{\otimes n}(\alpha).$$

Let α_n be the type i error that achieves minimum in the above equation, i.e.

$$\alpha_n + f^{\otimes n}(\alpha_n) = 1 - \mathrm{TV}(P^n, Q^n).$$

The total variation limit (25) implies $\alpha_n \to 0$ and $f^{\otimes n}(\alpha_n) \to 0$. For each n, consider the piecewise linear function that interpolates $(0,1), (\alpha_n, f^{\otimes n}(\alpha_n))$ and $(1,0)$, which will be denoted by h_n. By the convexity of $f^{\otimes n}$ we know that $f^{\otimes n} \leqslant h_n$ in $[0,1]$. It suffices to show that $h_n(\alpha) \to 0, \forall \alpha \in (0,1]$. Since $\alpha_n \to 0$, for large enough n, $h_n(\alpha)$ is evaluated on the lower linear segment of h_n. So $h_n(\alpha) \leqslant h_n(\alpha_n) \leqslant f^{\otimes n}(\alpha_n) \to 0$. This yields the desired limit of $f^{\otimes n}$.
Now we use Hellinger distance $H^2(P,Q) := \mathbb{E}_Q\big[(1-\sqrt{\frac{P}{Q}})^2\big]$ to show the total variation limit (25). An elementary inequality relating total variation and Hellinger distance is

$$\frac{1}{2}H^2(P,Q) \leqslant \mathrm{TV}(P,Q) \leqslant H(P,Q).$$

Another nice property of Hellinger distance is it tensorizes in the following sense:

$$1 - \frac{H^2(P^n,Q^n)}{2} = \Big(1 - \frac{H^2(P,Q)}{2}\Big)^n.$$

f is not the diagonal $\alpha \mapsto 1-\alpha$, so $P \neq Q$. Hence $\mathrm{TV}(P,Q) > 0$. By the second inequality in the sandwich bound, $H^2(P,Q) > 0$. By the tensorization property, $H^2(P^n,Q^n) \to 2$. By the first inequality in the sandwich bound and that TV is bounded by 1 we have

$$\frac{1}{2}H^2(P^n,Q^n) \leqslant \mathrm{TV}(P^n,Q^n) \leqslant 1.$$

This shows $\mathrm{TV}(P^n,Q^n) \to 1$ and completes the proof. □

Now we set out the journey to prove the Berry-Esseen style central limit theorem 3.4. We first restate the theorem.

Theorem 3.4. *Let $f_1, \ldots, f_n$ be symmetric trade-off functions such that $\kappa_3(f_i) < \infty$ for all $1 \leqslant i \leqslant n$. Denote*

$$\mu := \frac{2\|\mathbf{kl}\|_1}{\sqrt{\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2}} \quad \text{and} \quad \gamma := \frac{0.56\|\bar{\boldsymbol{\kappa}}_{\mathbf{3}}\|_1}{\big(\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2\big)^{3/2}}$$

and assume $\gamma < \frac{1}{2}$. Then, for all $\alpha \in [\gamma, 1-\gamma]$, we have

$$G_\mu(\alpha+\gamma) - \gamma \leqslant f_1 \otimes f_2 \otimes \cdots \otimes f_n(\alpha) \leqslant G_\mu(\alpha-\gamma) + \gamma. \tag{9}$$

Our approach is to consider the log-likelihood ratio between the distributions of the composition mechanism on neighboring datasets. This log-likelihood ratio can be reduced to the sum of *independent* components that each correspond to the log-likelihood ratio of a trade-off function in the tensor product. This reduction allows us to carry over the classical Berry–Esseen bound to Theorem 3.4.

As the very first step, let's better understand the functionals kl, κ_2 and $\bar{\kappa}_3$ used in the statement of the theorem. We focus on symmetric f with $f(0) = 1$, although some of the following discussion generalizes beyond that subclass. Recall that

$$\begin{aligned}\mathrm{kl}(f) &= -\int_0^1 \log|f'(x)|\,\mathrm{d}x\\ \kappa_2(f) &= \int_0^1 \big(\log|f'(x)|\big)^2\,\mathrm{d}x\\ \bar{\kappa}_3(f) &= \int_0^1 \big|\log|f'(x)| + \mathrm{kl}(f)\big|^3\,\mathrm{d}x\end{aligned}$$

First we finish the argument mentioned in Section 3.2 that these functionals are well-defined and take values in $[0,+\infty]$. For κ_2 and $\bar{\kappa}_3$, as well as the non-central version κ_3, the argument is easy because the integrands are non-negative.

For kl, the only possible singularities of the integrand is 0 and 1. If 1 is singular then $\log|f'(x)| \to -\infty$ near 1. This is okay because the functionals are allowed to take value $+\infty$. We need to rule out the case when 0 is a singularity and $\int_0^\varepsilon \log|f'(x)|\,\mathrm{d}x = +\infty$. That cannot happen because $\log|f'(x)| \leqslant |f'(x)| - 1$ and $|f'(x)| = -f'(x)$ is integrable in $[0,1]$ as it is the derivative of $-f$, an absolute continuous function. Non-negativity of kl follows from Jensen's inequality.

In the discussion of Proposition B.4, we showed that $\mathrm{kl}(T(P,Q)) = D_{\mathrm{KL}}(P\|Q)$. This explains the name of this functional. In fact, κ_2 also corresponds to a divergence called *exponential divergence* ([E$^+$85]).

We introduce a notation that will be useful in the calculation below. For a trade-off function f, let Df be a function with the following expression:

$$Df(x) = |f'(1-x)| = -f'(1-x).$$

In fact, this is the density introduce in the proof of Proposition 2.2.

By a simple change of variable, the three functionals can be re-written as

$$\begin{aligned}\mathrm{kl}(f) &= -\int_0^1 \log Df(x)\,\mathrm{d}x\\ \kappa_2(f) &= \int_0^1 \big(\log Df(x)\big)^2\,\mathrm{d}x\\ \bar{\kappa}_3(f) &= \int_0^1 \big|\log Df(x) + \mathrm{kl}(f)\big|^3\,\mathrm{d}x\end{aligned}$$

The following "shadows" of the above functionals will appear in the proof:

$$\begin{aligned}\mathrm{lk}(f) &:= \int_0^1 Df(x)\log Df(x)\,\mathrm{d}x\\ \tilde{\kappa}_2(f) &:= \int_0^1 Df(x)\big(\log Df(x)\big)^2\,\mathrm{d}x\\ \tilde{\kappa}_3(f) &:= \int_0^1 Df(x)\big|\log Df(x) - \mathrm{lk}(f)\big|^3\,\mathrm{d}x\end{aligned}$$

These functionals are also well-defined on $\mathscr{F}$ and take values in $[0,+\infty]$. The argument is simular to that of kl, κ_2 and $\bar{\kappa}_3$.

The following calculations turn out to be useful in the proof.

Proposition D.3. *Suppose $f \in \mathscr{F}^S$ and $f(0) = 1$. Then*

$$\begin{aligned}
\mathrm{kl}(f) &= \mathrm{lk}(f)\\
\kappa_2(f) &= \tilde{\kappa}_2(f)\\
\bar{\kappa}_3(f) &= \tilde{\kappa}_3(f).
\end{aligned}$$

Proof. Our approach, taking κ_2 as example, is to show $\kappa_2(f^{-1}) = \tilde{\kappa}_2(f)$. By definition of symmetry, $f^{-1} = f$ and hence the desired result follows. First observe for $f \in \mathscr{F}^S$ with $f(0) = 1$, f^{-1} agrees with the ordinary function inverse, hence we can apply calculus rule as follows:

$$Df^{-1}(x) = -\frac{\mathrm{d}f^{-1}}{\mathrm{d}x}(1-x) = \frac{-1}{f'(f^{-1}(1-x))}.$$

We only prove $\kappa_2(f^{-1}) = \tilde{\kappa}_2(f)$ here and the other two identities can be proved similarly.

$$\begin{aligned}
\kappa_2(f^{-1}) &= \int_0^1 \big(\log Df^{-1}(x)\big)^2 \,\mathrm{d}x\\
&= \int_0^1 \big(-\log\big[-f'(f^{-1}(1-x))\big]\big)^2 \,\mathrm{d}x\\
&= \int_0^1 \log^2\big[-f'(f^{-1}(1-x))\big] \,\mathrm{d}x
\end{aligned}$$

Let $y = f^{-1}(1-x)$, then $f'(y)\,\mathrm{d}y = -\,\mathrm{d}x$, and $x = 0$ corresponds to $y = 0$, $x = 1$ corresponds to $y = 1$.

$$\begin{aligned}
\kappa_2(f^{-1}) &= \int_0^1 \log^2[-f'(y)] \cdot \big(-f'(y)\big)\,\mathrm{d}y && (y = f^{-1}(1-x))\\
&= \int_0^1 \log^2[-f'(1-z)] \cdot \big(-f'(1-z)\big)\,\mathrm{d}z && (z = 1-y)\\
&= \int_0^1 Df(z)\big(\log Df(z)\big)^2 \,\mathrm{d}z\\
&= \tilde{\kappa}_2(f).
\end{aligned}$$

□

We remark that by properly extending the definition of the shadow functionals, identities like $\mathrm{kl}(f^{-1}) = \mathrm{lk}(f)$ holds for general trade-off function f.

Before we finally start the proof, let's recall Berry-Esseen theorem for random variables. Suppose we have n independent random variables $X_1, \ldots, X_n$ with $\mathbb{E}X_i = \mu_i, \mathrm{Var}X_i = \sigma_i^2, \mathbb{E}|X_i - \mu_i|^3 = \rho_i^3$. Consider the normalized random variable

$$S_n := \frac{\sum_{i=1}^n X_i - \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}.$$

Denote its cdf by F_n. Then

Theorem D.4 (Berry-Esseen)**.** *There exists a universal constant $C > 0$ such that*

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leqslant C \cdot \frac{\sum_{i=1}^n \rho_i^3}{\big(\sum_{i=1}^n \sigma_i^2\big)^{\frac{3}{2}}}.$$

To the best of our knowledge, the best C is 0.5600 due to [She10].

Now we proceed to the proof of Theorem 3.4.

Proof of Theorem 3.4. For simplicity let

$$\boldsymbol{f} := f_1 \otimes f_2 \otimes \cdots \otimes f_n.$$

First let's find distributions P_0 and P_1 such that $T(P_0, P_1) = \boldsymbol{f}$.

First, by symmetry, if $f_i(0) < 1$, then $f_i'(x) = 0$ in some interval $[1-\varepsilon, 1]$ for some $\varepsilon > 0$, which yields $\mathrm{kl}(f_i) = +\infty$. So we can assume $f_i(0)$ for all i.

Recall that $Df_i(x) = -f_i'(1-x)$. Let P be the uniform distribution on $[0,1]$ and Q_i be the distribution supported on $[0,1]$ with density Df_i. These are the distributions constructed in the proof of Proposition 2.2. Since f_i are all symmetric and $f_i(0) = 1$, the supports of P and all Q_i are all exactly $[0,1]$, and we have $T(P, Q_i) = f_i$. Hence by definition $\boldsymbol{f} = T(P^n, Q_1 \times \cdots \times Q_n)$.

Now let's study the hypothesis testing problem P^n vs $Q_1 \times \cdots \times Q_n$. Let

$$L_i(x) := \log \frac{\mathrm{d}Q_i}{\mathrm{d}P}(x) = \log Df_i(x)$$

be the log likelihood ratio. Since both hypotheses are product distributions, Neyman-Pearson lemma implies that the optimal rejection rules of this testing problem must be a threshold function of the quantity $\sum_{i=1}^n L_i$. We need to study $\sum_{i=1}^n L_i(x_i)$ under both the null and the alternative hypothesis, i.e. when $(x_1, \ldots, x_n)$ comes from P^n and $Q_1 \times \cdots \times Q_n$. From here we implement the following plan: first find the quantities that exhibit central limit behavior, then express α and $\boldsymbol{f}(\alpha)$ in terms of these quantities.

For further simplification, let

$$T_n := \sum_{i=1}^n L_i.$$

As we turn off the x_i notation, we should bear in mind that T_n has different distributions under P^n and $Q_1 \times \cdots \times Q_n$, but it is an independent sum in both cases.

In order to find quantities with central limit behavior, it suffices to normalize T_n under both distributions. The mysterious functionals we introduced are specifically designed for this purpose.

$$\begin{aligned}
\mathbb{E}_P[L_i] &= \int_0^1 \log Df_i(x_i)\,\mathrm{d}x_i = -\mathrm{kl}(f_i),\\
\mathbb{E}_{Q_i}[L_i] &= \int_0^1 Df_i(x_i)\log Df_i(x_i)\,\mathrm{d}x_i = \mathrm{lk}(f_i) = \mathrm{kl}(f_i).
\end{aligned}$$

In the last step we used Proposition D.3. With the bold vector notation,

$$\begin{aligned}
\mathbb{E}_{P^n}[T_n] &= \sum_{i=1}^n -\mathrm{kl}(f_i) = -\|\mathbf{kl}\|_1,\\
\mathbb{E}_{Q_1\times\cdots\times Q_n}[T_n] &= \sum_{i=1}^n \mathrm{kl}(f_i) = \|\mathbf{kl}\|_1.
\end{aligned}$$

Similarly for the variances:

$$\begin{aligned}
\mathrm{Var}_P[L_i] &= \mathbb{E}_P[L_i^2] - \big(\mathbb{E}_P[L_i]\big)^2 = \kappa_2(f_i) - \mathrm{kl}^2(f_i),\\
\mathrm{Var}_{Q_i}[L_i] &= \mathbb{E}_{Q_i}[L_i^2] - \big(\mathbb{E}_{Q_i}[L_i]\big)^2 = \tilde{\kappa}_2(f_i) - \mathrm{lk}^2(f_i) = \kappa_2(f_i) - \mathrm{kl}^2(f_i).
\end{aligned}$$

$$\mathrm{Var}_{P^n}[T_n] = \mathrm{Var}_{Q_1\times\cdots\times Q_n}[T_n] = \sum_{i=1}^{n} \kappa_2(f_i) - \mathrm{kl}^2(f_i) = \|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2.$$

In order to apply Berry-Esseen theorem (for random variables) we still need the centralized third moments:

$$\begin{aligned}
\mathbb{E}_P|L_i - \mathbb{E}_P[L_i]|^3 &= \int_0^1 \big|\log Df_i(x) + \mathrm{kl}(f_i)\big|^3\,\mathrm{d}x = \bar{\kappa}_3(f_i),\\
\mathbb{E}_{Q_i}|L_i - \mathbb{E}_{Q_i}[L_i]|^3 &= \int_0^1 Df_i(x)\big|\log Df_i(x) - \mathrm{lk}(f_i)\big|^3\,\mathrm{d}x = \tilde{\kappa}_3(f_i) = \bar{\kappa}_3(f_i).
\end{aligned}$$

Let F_n be the cdf of $\frac{T_n + \|\mathbf{kl}\|_1}{\sqrt{\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2}}$ under P^n, and $\tilde{F}^{(n)}$ be the cdf of $\frac{T_n - \|\mathbf{kl}\|_1}{\sqrt{\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2}}$ under $Q_1\times\cdots\times Q_n$. By Berry-Esseen Theorem D.4,

$$\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \leqslant C \cdot \frac{\|\bar{\boldsymbol{\kappa}}_{\mathbf{3}}\|_1}{\big(\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2\big)^{\frac{3}{2}}} = \gamma \tag{26}$$

and similarly $\sup_{x\in\mathbb{R}} |\tilde{F}^{(n)}(x) - \Phi(x)| \leqslant \gamma$.

So we find the quantities that exhibit central limit behavior. Now let's relate them with $\boldsymbol{f}$. Consider the testing problem $(P^n, Q_1\times\cdots\times Q_n)$. For a fixed $\alpha\in[0,1]$, let the optimal rejection rule (potentially randomized) at level α be ϕ. By Neyman-Pearson lemma, ϕ must be a thresholding on T_n. An equivalent form that highlights the central limit behavior is the following:

$$\phi = \begin{cases} 1, & \frac{T_n + \|\mathbf{kl}\|_1}{\sqrt{\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2}} > t,\\ p, & \frac{T_n + \|\mathbf{kl}\|_1}{\sqrt{\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2}} = t,\\ 0, & \frac{T_n + \|\mathbf{kl}\|_1}{\sqrt{\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2}} < t \end{cases}$$

Here $t\in\mathbb{R}\cup\{\pm\infty\}$ and $p\in[0,1]$ are parameters uniquely determined by the condition $\mathbb{E}_{P^n}[\phi] = \alpha$. With this form $\mathbb{E}_{P^n}[\phi]$ can be easily spelled out in terms of F_n:

$$\begin{aligned}
\mathbb{E}_{P^n}[\phi] &= P^n\Big[\frac{T_n + \|\mathbf{kl}\|_1}{\sqrt{\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2}} > t\Big] + p\cdot P^n\Big[\frac{T_n + \|\mathbf{kl}\|_1}{\sqrt{\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2}} = t\Big]\\
&= 1 - F_n(t) + p\cdot[F_n(t) - F_n(t^-)].
\end{aligned}$$

Here $F_n(t^-)$ is the left limit of the function F_n at t. Simple algebra yields

$$1 - \alpha = 1 - \mathbb{E}_{P^n}[\phi] = (1-p)F_n(t) + pF_n(t^-)$$

and consequently the inequality

$$F_n(t^-) \leqslant 1 - \alpha \leqslant F_n(t).$$

For $\mathbb{E}_{Q_1\times\cdots\times Q_n}[\phi]$ it is helpful to introduce another letter $\tau := t - \mu$. In the theorem statement μ was defined to be $\frac{2\|\mathbf{kl}\|_1}{\sqrt{\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2}}$ so we have the equivalence

$$\frac{T_n + \|\mathbf{kl}\|_1}{\sqrt{\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2}} > t \Leftrightarrow \frac{T_n - \|\mathbf{kl}\|_1}{\sqrt{\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2}} > \tau. \tag{27}$$

With this extra notation we have

$$\begin{aligned}
1 - \boldsymbol{f}(\alpha) &= \mathbb{E}_{Q_1\times\cdots\times Q_n}[\phi] \\
&= Q_1 \times \cdots \times Q_n\Big[\frac{T_n + \|\mathbf{kl}\|_1}{\sqrt{\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2}} > t\Big] + \\
&\qquad p \cdot Q_1 \times \cdots \times Q_n\Big[\frac{T_n + \|\mathbf{kl}\|_1}{\sqrt{\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2}} = t\Big] && \text{(Def. of } \phi) \\
&= Q_1 \times \cdots \times Q_n\Big[\frac{T_n - \|\mathbf{kl}\|_1}{\sqrt{\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2}} > \tau\Big] + \\
&\qquad p \cdot Q_1 \times \cdots \times Q_n\Big[\frac{T_n - \|\mathbf{kl}\|_1}{\sqrt{\|\boldsymbol{\kappa_2}\|_1 - \|\mathbf{kl}\|_2^2}} = \tau\Big] && \text{(By (27))} \\
&= 1 - \tilde{F}^{(n)}(\tau) + p \cdot [\tilde{F}^{(n)}(\tau) - \tilde{F}^{(n)}(\tau^-)].
\end{aligned}$$

Similar algebra as before yields

$$\boldsymbol{f}(\alpha) = (1-p)\tilde{F}^{(n)}(\tau) + p\tilde{F}^{(n)}(\tau^-)$$

and hence

$$\tilde{F}^{(n)}(\tau^-) \leqslant \boldsymbol{f}(\alpha) \leqslant \tilde{F}^{(n)}(\tau).$$

So far we have

$$F_n(t^-) \leqslant 1 - \alpha \leqslant F_n(t), \tag{28}$$

$$\tilde{F}^{(n)}(\tau^-) \leqslant \boldsymbol{f}(\alpha) \leqslant \tilde{F}^{(n)}(\tau). \tag{29}$$

In (26) we show F_n and $\tilde{F}^{(n)}$ are within distance γ to the cdf of standard normal, so

$$\Phi(t) - \gamma \leqslant F_n(t^-) \leqslant 1 - \alpha \leqslant F_n(t) \leqslant \Phi(t) + \gamma$$

and hence

$$\Phi^{-1}(1 - \alpha - \gamma) \leqslant t \leqslant \Phi^{-1}(1 - \alpha + \gamma). \tag{30}$$

Using (29) and (30),

$$\begin{aligned}
\boldsymbol{f}(\alpha) &\leqslant \tilde{F}^{(n)}(\tau) \\
&\leqslant \Phi(\tau) + \gamma \\
&= \Phi(t - \mu) + \gamma \\
&\leqslant \Phi(\Phi^{-1}(1 - \alpha + \gamma) - \mu) + \gamma \\
&= G_\mu(\alpha - \gamma) + \gamma
\end{aligned}$$

Similarly we can show that $\boldsymbol{f}(\alpha) \geqslant G_\mu(\alpha + \gamma) - \gamma$. The proof is now complete. □

Next we prove the asymptotic version. Recall that our goal is

Theorem 3.5. *Let $\{f_{ni} : 1 \leqslant i \leqslant n\}_{n=1}^{\infty}$ be a triangular array of symmetric trade-off functions and assume the following limits for some constants $K \geqslant 0$ and $s > 0$ as $n \to \infty$:*

1. $\sum_{i=1}^{n} \mathrm{kl}(f_{ni}) \to K$;
2. $\max_{1 \leqslant i \leqslant n} \mathrm{kl}(f_{ni}) \to 0$;
3. $\sum_{i=1}^{n} \kappa_2(f_{ni}) \to s^2$;
4. $\sum_{i=1}^{n} \kappa_3(f_{ni}) \to 0$.

Then, we have

$$\lim_{n\to\infty} f_{n1} \otimes f_{n2} \otimes \cdots \otimes f_{nn}(\alpha) = G_{2K/s}(\alpha)$$

uniformly for all $\alpha \in [0, 1]$.

Proof of Theorem 3.5. We will first construct pointwise convergence $f_{n1} \otimes f_{n2} \otimes \cdots \otimes f_{nn} \to G_{2K/s}$ and then conclude uniform convergence from a general theorem.

Apply Berry-Esseen Theorem 3.4 to the n-th row of the triangular array and we have

$$G_{\mu_n}(\alpha + \gamma_n) - \gamma_n \leqslant f_{n1} \otimes f_{n2} \otimes \cdots \otimes f_{nn}(\alpha) \leqslant G_{\mu_n}(\alpha - \gamma_n) + \gamma_n.$$

Here μ_n and γ_n are the counterparts of μ and γ defined in Theorem 3.4 when applied to $f_{n1}, \dots, f_{nn}$. Namely,

$$\mu_n = \frac{2\|\mathbf{kl}^{(n)}\|_1}{\sqrt{\|\boldsymbol{\kappa}_\mathbf{2}^{(n)}\|_1 - \|\mathbf{kl}^{(n)}\|_2^2}},$$

$$\gamma_n = 0.56 \cdot \frac{\|\bar{\boldsymbol{\kappa}}_\mathbf{3}^{(n)}\|_1}{\big(\|\boldsymbol{\kappa}_\mathbf{2}^{(n)}\|_1 - \|\mathbf{kl}^{(n)}\|_2^2\big)^{\frac{3}{2}}}$$

Here the bold vector notation with a superscript (n) denotes the vector for the n-th row. For example, $\mathbf{kl}^{(n)} = \big(\mathrm{kl}(f_{n1}), \dots, \mathrm{kl}(f_{nn})\big)$.

By the sandwich inequality, pointwise convergence of $f_{n1} \otimes f_{n2} \otimes \cdots \otimes f_{nn}$ follows from the two limits

$$G_{\mu_n}(\alpha + \gamma_n) - \gamma_n \to G_{2K/s}(\alpha), \quad G_{\mu_n}(\alpha - \gamma_n) + \gamma_n \to G_{2K/s}(\alpha). \tag{31}$$

To prove these, let's first show $\gamma_n \to 0$ and $\mu_n \to 2K/s$.

Reformulating the assumptions in bold vector notations, we have

$$||\mathbf{kl}^{(n)}||_1 \to K, \quad ||\mathbf{kl}^{(n)}||_\infty \to 0, \quad ||\boldsymbol{\kappa}_\mathbf{2}^{(n)}||_1 \to s^2, \quad ||\boldsymbol{\kappa}_\mathbf{3}^{(n)}||_1 \to 0.$$

In addtion to these, it suffices to show

$$\|\mathbf{kl}^{(n)}\|_2^2 \to 0 \quad \text{and} \quad \|\bar{\boldsymbol{\kappa}}_\mathbf{3}^{(n)}\|_1 \to 0. \tag{32}$$

For the first half, notice that $\|\mathbf{kl}^{(n)}\|_2^2 = \langle \mathbf{kl}^{(n)}, \mathbf{kl}^{(n)} \rangle \leqslant \|\mathbf{kl}^{(n)}\|_\infty \cdot \|\mathbf{kl}^{(n)}\|_1 \to 0$. In fact, $\|\mathbf{kl}^{(n)}\|_\infty \to 0$ is not only sufficient but also necessary, because $\|\mathbf{kl}^{(n)}\|_\infty \leqslant \|\mathbf{kl}^{(n)}\|_2$.

Next we use the assumptions to show $\|\bar{\boldsymbol{\kappa}}_\mathbf{3}^{(n)}\|_1 \to 0$. We need a lemma

Lemma D.5. *For a trade-off function f,*

$$\bar{\kappa}_3(f) \leqslant \kappa_3(f) + 3\mathrm{kl}(f)\cdot\kappa_2(f) + 3\mathrm{kl}^2(f)\cdot\sqrt{\kappa_2(f)} + \mathrm{kl}^3(f).$$

Proof of Lemma D.5.

$$\begin{aligned}\bar{\kappa}_3(f) = &\int_0^1 \big|\log Df(x) + \mathrm{kl}(f)\big|^3\,\mathrm{d}x\\ \leqslant &\int_0^1 \big(\big|\log Df(x)\big| + \big|\mathrm{kl}(f)\big|\big)^3\,\mathrm{d}x\\ \leqslant &\int_0^1 \big|\log Df(x)\big|^3\,\mathrm{d}x + 3\mathrm{kl}(f)\cdot\int_0^1\big|\log Df(x)\big|^2\,\mathrm{d}x\\ &+ 3\mathrm{kl}^2(f)\cdot\int_0^1\big|\log Df(x)\big|\,\mathrm{d}x + \mathrm{kl}^3(f)\\ \leqslant &\ \kappa_3(f) + 3\mathrm{kl}(f)\cdot\kappa_2(f) + 3\mathrm{kl}^2(f)\cdot\sqrt{\kappa_2(f)} + \mathrm{kl}^3(f).\end{aligned}$$

In the last step we used Jensen's inequality. □

Apply Lemma D.5 to each f_{ni} and sum them up:

$$\|\bar{\boldsymbol{\kappa}}_{\mathbf{3}}{}^{(n)}\|_1 \leqslant \|\boldsymbol{\kappa}_{\mathbf{3}}{}^{(n)}\|_1 + 3\textstyle\sum_i \mathrm{kl}(f_{ni})\cdot\kappa_2(f_{ni}) + 3\sum_i \mathrm{kl}(f_{ni})\cdot\sqrt{\kappa_2(f_{ni})}\cdot\mathrm{kl}(f_{ni}) + \sum_i \mathrm{kl}(f_{ni})\cdot\mathrm{kl}^2(f_{ni}).$$

Using $|\sum a_i b_i| \leqslant |\sum a_i|\cdot\max|b_i|$ and Cauchy-Schwarz inequality yields

$$\begin{aligned}\|\bar{\boldsymbol{\kappa}}_{\mathbf{3}}{}^{(n)}\|_1 &\leqslant \|\boldsymbol{\kappa}_{\mathbf{3}}{}^{(n)}\|_1 + 3\|\mathbf{kl}^{(n)}\|_\infty\cdot\|\boldsymbol{\kappa}_{\mathbf{2}}{}^{(n)}\|_1 + 3\|\mathbf{kl}^{(n)}\|_\infty\cdot\Big(\textstyle\sum_i\sqrt{\kappa_2(f_{ni})}\cdot\mathrm{kl}(f_{ni})\Big) + \|\mathbf{kl}^{(n)}\|_\infty^2\cdot\|\mathbf{kl}^{(n)}\|_1\\ &\leqslant \|\boldsymbol{\kappa}_{\mathbf{3}}{}^{(n)}\|_1 + 3\|\mathbf{kl}^{(n)}\|_\infty\cdot\|\boldsymbol{\kappa}_{\mathbf{2}}{}^{(n)}\|_1 + 3\|\mathbf{kl}^{(n)}\|_\infty\cdot\sqrt{\|\boldsymbol{\kappa}_{\mathbf{2}}{}^{(n)}\|_1\cdot\|\mathbf{kl}^{(n)}\|_2^2} + \|\mathbf{kl}^{(n)}\|_\infty^2\cdot\|\mathbf{kl}^{(n)}\|_1.\end{aligned}$$

By the assumptions

$$||\mathbf{kl}^{(n)}||_1 \to K, \quad ||\mathbf{kl}^{(n)}||_\infty \to 0, \quad ||\boldsymbol{\kappa}_{\mathbf{2}}{}^{(n)}||_1 \to s^2, \quad ||\boldsymbol{\kappa}_{\mathbf{3}}{}^{(n)}||_1 \to 0$$

and $\|\mathbf{kl}^{(n)}\|_2^2 \to 0$ which we just proved, it's easy to see that all four terms goes to 0 as n goes to infinity.

The two limits (32) we have just proved imply $\mu_n \to 2K/s$ and $\gamma_n \to 0$. Given these, convergence (31) is easy once we notice that $G_\mu(\alpha) = \Phi(\Phi^{-1}(1-\alpha) - \mu)$ is continuous in both α and μ.

If the readers are concerned with $1-\alpha-\gamma_n$ exceeding $[0,1]$, then observe that when $\alpha \in (0,1)$, $1-\alpha-\gamma_n$ eventually ends up in $(0,1)$ where Φ^{-1} is well-defined and continuous. So the only concern is at 0 and 1. If $\alpha = 0$, $\Phi^{-1}(1-\alpha-\gamma_n) \to +\infty$ so $G_{\mu_n}(0+\gamma_n) - \gamma_n \to 1 = G_{2K/s}(0)$. A similar argument works for $\alpha = 1$.

Anyway, we have shown pointwise convergence. Uniform convergence is again a direct consequence of Lemma A.7. The proof is now complete. □

Next we explain the effect of tensoring $f_{0,\delta}$.

$$f \otimes f_{0,\delta}(\alpha) = \begin{cases} (1-\delta)\cdot f(\frac{\alpha}{1-\delta}), & 0 \leqslant \alpha \leqslant 1-\delta\\ 0, & 1-\delta \leqslant \alpha \leqslant 1.\end{cases} \tag{12}$$

Proof of Equation (12). First, $f_{0,\delta}$ is the trade-off function of two uniform distributions $f_{0,\delta} = T\big(U[0,1], U[\delta, 1+\delta]\big)$. To see this, observe that any optimal test ϕ for $U[0,1]$ vs $U[\delta, 1+\delta]$ must have the following form:

$$\phi(x) = \begin{cases} 1, & x \in (1, 1+\delta] \\ p, & x \in [\delta, 1], \\ 0, & x \in [0, \delta) \end{cases}$$

That is, we know it must be from $U[0,1]$ if we see something in $[0,\delta)$, and must be from $U[\delta, 1+\delta]$ if we see something in $(1, 1+\delta]$. Otherwise the only thing we can do is random guessing. It's easy to see that the errors of such ϕ linearly interpolates between $(0, 1-\delta)$ and $(1-\delta, 0)$, i.e. type I and type II error add up to $1-\delta$. On the other hand, by definition, $f_{0,\delta}(\alpha) = \max\{1-\delta-\alpha, 0\}$. So they indeed agree with each other.

Now suppose $f = T(P,Q)$. By definition of tensor product, $f \otimes f_{0,\delta} = T(P \times U[0,1], Q \times U[\delta, 1+\delta])$. If the optimal test for P vs Q at level α is ϕ_α, then an optimal test for $P \times U[0,1]$ vs $Q \times U[\delta, 1+\delta]$ must be of the following form:

$$\tilde{\phi}_\alpha(\omega, x) = \begin{cases} 1, & x \in (1, 1+\delta] \\ \phi_\alpha(\omega), & x \in [\delta, 1], \\ 0, & x \in [0, \delta) \end{cases}$$

The errors are

$$\begin{aligned} \mathbb{E}_{P\times U[0,1]}[\tilde{\phi}_\alpha] &= P\big[x \in (1, 1+\delta]\big] + P\big[x \in [\delta, 1]\big] \cdot \mathbb{E}_P[\phi_\alpha(\omega)] \\ &= 0 + (1-\delta)\alpha = (1-\delta)\alpha \\ 1 - \mathbb{E}_{Q\times U[\delta,1+\delta]}[\tilde{\phi}_\alpha] &= 1 - P\big[x \in (1, 1+\delta]\big] - P\big[x \in [\delta, 1]\big] \cdot \mathbb{E}_Q[\phi_\alpha(\omega)] \\ &= 1 - \delta - (1-\delta)\big(1 - f(\alpha)\big) = (1-\delta) f(\alpha) \end{aligned}$$

This completes the proof. □

Theorem 3.6. *Assume*

$$\sum_{i=1}^n \varepsilon_{ni}^2 \to \mu^2, \quad \max_{1\leqslant i \leqslant n} \varepsilon_{ni} \to 0, \quad \sum_{i=1}^n \delta_{ni} \to \delta, \quad \max_{1\leqslant i\leqslant n} \delta_{ni} \to 0$$

for some nonnegative constants μ, δ as $n \to \infty$. Then, we have

$$f_{\varepsilon_{n1}, \delta_{n1}} \otimes \cdots \otimes f_{\varepsilon_{nn}, \delta_{nn}} \to G_\mu \otimes f_{0, 1-\mathrm{e}^{-\delta}}$$

uniformly over $[0,1]$ *as* $n \to \infty$.

Proof of Theorem 3.6. As in the main body, we first apply rules $f_{\varepsilon,\delta} = f_{\varepsilon,0} \otimes f_{0,\delta}$ and $f_{0,\delta_1} \otimes f_{0,\delta_2} = f_{0,1-(1-\delta_1)(1-\delta_2)}$ to get

$$\begin{aligned} f_{\varepsilon_{n1},\delta_{n1}} \otimes \cdots \otimes f_{\varepsilon_{nn},\delta_{nn}} &= \big(f_{\varepsilon_{n1},0} \otimes \cdots \otimes f_{\varepsilon_{nn},0}\big) \otimes \big(f_{0,\delta_{n1}} \otimes \cdots \otimes f_{0,\delta_{nn}}\big) \\ &= \big(\underbrace{f_{\varepsilon_{n1},0} \otimes \cdots \otimes f_{\varepsilon_{nn},0}}_{f^{(n)}}\big) \otimes f_{0,\delta^{(n)}} \end{aligned}$$

with $\delta^{(n)} = 1 - \prod_{i=1}^n (1-\delta_{ni})$. For the second factor, let's first prove the limit $\delta^{(n)} \to 1 - \mathrm{e}^{-\delta}$. Changing the product into sum, we have

$$\log(1-\delta^{(n)}) = \textstyle\sum_{i=1}^n \log(1-\delta_{ni})$$

The limit almost follows from the Taylor expansion $\log(1+x) = x + o(x)$, but we need to be a little more careful as the number of summation terms also goes to infinity. Since $\max_{1\leqslant i\leqslant n} \delta_{ni} \to 0$, we can assume for large n, $\delta_{ni} < r$ for some r such that when $|x| < r$, the following Taylor expansion holds for some constant C:

$$|\log(1-x) + x| \leqslant Cx^2.$$

With this,

$$\big|\textstyle\sum_{i=1}^n \log(1-\delta_{ni}) + \delta_{ni}\big| \leqslant C \cdot \sum_{i=1}^n \delta_{ni}^2 \leqslant C \cdot \max_i \delta_{ni} \cdot \sum_i \delta_{ni} \to 0.$$

Therefore, $\log(1-\delta^{(n)}) = \sum_{i=1}^n \log(1-\delta_{ni})$ has the same limit as $\sum_{i=1}^n \delta_{ni}$. In other words, $\log(1-\delta^{(n)}) \to \delta$, or equivalently, $\delta^{(n)} \to 1 - \mathrm{e}^{-\delta}$.

For a fixed $x \in [0,1]$, $f_{0,\delta^{(n)}}(x) = \max\{0, 1-\delta^{(n)} - x\}$ is continuous in $\delta^{(n)}$. Hence we have the pointwise limit $f_{0,\delta^{(n)}} \to f_{0,1-\mathrm{e}^{-\delta}}$.

For the first factor $f^{(n)} = f_{\varepsilon_{n1},0} \otimes \cdots \otimes f_{\varepsilon_{nn},0}$, we will apply Theorem 3.5. Let's check the conditions.

By the continuity of the function $x \mapsto x \tanh \frac{x}{2}$ at 0, the assumption $\max_{1\leqslant i\leqslant n} \varepsilon_{ni} \to 0$ implies

$$\max_{1\leqslant i\leqslant n} \mathrm{kl}(f_{ni}) = \max_{1\leqslant i\leqslant n} \varepsilon_{ni} \tanh \tfrac{\varepsilon_{ni}}{2} \to 0.$$

Next, we show

$$\sum_{i=1}^n \mathrm{kl}(f_{ni}) = \sum_{i=1}^n \varepsilon_{ni} \tanh \frac{\varepsilon_{ni}}{2} \to K = \frac{\mu^2}{2}.$$

Preparing for the same Taylor expansion trick, let n be large enough so that Taylor expansion $|\tanh x - x| \leqslant Cx^2$ applies to all δ_{ni}.

$$\begin{aligned}\Big|\sum_{i=1}^n \varepsilon_{ni} \tanh \frac{\varepsilon_{ni}}{2} - \sum_{i=1}^n \frac{\varepsilon_{ni}^2}{2}\Big| &= \sum_{i=1}^n \varepsilon_{ni} \Big|\tanh \frac{\varepsilon_{ni}}{2} - \frac{\varepsilon_{ni}}{2}\Big| \\ &\leqslant C \cdot \sum_{i=1}^n \varepsilon_{ni} \cdot \varepsilon_{ni}^2 \\ &\leqslant C \cdot \max_{1\leqslant i\leqslant n} \varepsilon_{ni} \cdot \sum_{i=1}^n \varepsilon_{ni}^2 \to 0.\end{aligned}$$

So $\sum_{i=1}^n \varepsilon_{ni} \tanh \frac{\varepsilon_{ni}}{2}$ and $\sum_{i=1}^n \frac{\varepsilon_{ni}^2}{2}$ has the same limit, which by our assumption is $\mu^2/2$.

For second moment, $\sum_{i=1}^n \kappa_2(f_{ni}) = \sum_{i=1}^n \varepsilon_{ni}^2$ has limit μ^2. That is, s in Theorem 3.5 is equal to μ.

For third moment,

$$\sum_{i=1}^n \kappa_3(f_{ni}) = \sum_{i=1}^n \varepsilon_{ni}^3 \leqslant \big(\max_{1\leqslant i\leqslant n} \varepsilon_{ni}\big) \cdot \sum_{i=1}^n \varepsilon_{ni}^2 \to 0.$$

All four conditions of Theorem 3.5 check, so we can conclude the limit of $f^{(n)}$ is the GDP trade-off function with parameter $2K/s = s = \mu$.

The last step is to combine the two limits $f^{(n)} \to G_\mu$ and $f_{0,\delta^{(n)}} \to f_{0,1-\mathrm{e}^{-\delta}}$. By Equation (12),

$$f^{(n)} \otimes f_{0,\delta^{(n)}}(\alpha) = \begin{cases} (1-\delta^{(n)}) \cdot f^{(n)}(\frac{\alpha}{1-\delta^{(n)}}), & 0 \leqslant \alpha \leqslant 1-\delta^{(n)}, \\ 0, & 1-\delta^{(n)} \leqslant \alpha \leqslant 1. \end{cases}$$

Lemma A.7 tells us $f^{(n)}$ uniformly converges to G_μ, so we have the limit

$$f^{(n)}(\tfrac{\alpha}{1-\delta^{(n)}}) \to G_\mu(\tfrac{\alpha}{1-(1-\mathrm{e}^{-\delta})})$$

This implies the pointwise limit

$$f_{\varepsilon_{n1},\delta_{n1}} \otimes \cdots \otimes f_{\varepsilon_{nn},\delta_{nn}} = f^{(n)} \otimes f_{0,\delta^{(n)}} \to G_\mu \otimes f_{0,1-\mathrm{e}^{-\delta}}.$$

Again, uniform convergence comes for free via Lemma A.7. □

The next two corollaries are Berry-Esseen style central limit theorems for the composition of pure ε-DP. Given the existence of Theorem 3.7, these results are relatively loose, but might be good enough if we have large n. Nonzero δ is allowed following a similar argument as in Theorem 3.6.

Corollary D.6. *Set* $t_i = \tanh\frac{\varepsilon_i}{2}$ *and*

$$\mu = \frac{2\sum_{i=1}^n \varepsilon_i t_i}{\big(\sum_{i=1}^n \varepsilon_i^2(1-t_i^2)\big)^{1/2}},$$
$$\gamma = 0.56 \cdot \frac{\sum_{i=1}^n \varepsilon_i^3(1-t_i^4)}{\big(\sum_{i=1}^n \varepsilon_i^2(1-t_i^2)\big)^{3/2}}.$$

Then for any $\alpha \in [0,1]$,

$$G_\mu(\alpha+\gamma) - \gamma \leqslant f_{\varepsilon_1,0} \otimes f_{\varepsilon_2,0} \otimes \cdots \otimes f_{\varepsilon_n,0}(\alpha) \leqslant G_\mu(\alpha-\gamma) + \gamma.$$

In order to highlight the $1/\sqrt{n}$ convergence rate, we also derive an easy version in the homogeneous case.

Corollary D.7. *Let* $\mu = 2\sqrt{n}\sinh\frac{\varepsilon}{2}, \gamma = \frac{0.56}{\sqrt{n}} \cdot \frac{\cosh\varepsilon}{\cosh\frac{\varepsilon}{2}}$. *Then*

$$G_\mu(\alpha+\gamma) - \gamma \leqslant f_{\varepsilon,0}^{\otimes n}(\alpha) \leqslant G_\mu(\alpha-\gamma) + \gamma.$$

To see $\gamma = O(1/\sqrt{n})$, note that for the limit to be meaningful, ε has to be $o(1)$, which implies $\frac{\cosh\varepsilon}{\cosh\frac{\varepsilon}{2}} \approx 1$.

Both of them rely on evaluating the moment functionals on $f_{\varepsilon,0}$. We summarize the results in the following lemma:

Lemma D.8. *Let* $t = \tanh\frac{\varepsilon}{2}$. *Then*

$$\mathrm{kl}(f_{\varepsilon,0}) = \varepsilon t, \quad \kappa_2(f_{\varepsilon,0}) = \varepsilon^2, \quad \kappa_3(f_{\varepsilon,0}) = \varepsilon^3, \quad \bar\kappa_3(f_{\varepsilon,0}) = \varepsilon^3(1-t^4).$$

Proof of Lemma D.8. For convenience, let $p = \frac{\mathrm{e}^\varepsilon}{\mathrm{e}^\varepsilon+1}$ and $q = 1-p = \frac{1}{\mathrm{e}^\varepsilon+1}$. We have

$$\begin{aligned}
p &= \frac{\mathrm{e}^\varepsilon}{\mathrm{e}^\varepsilon+1} = \frac{\mathrm{e}^{\frac{\varepsilon}{2}}}{\mathrm{e}^{\frac{\varepsilon}{2}}+\mathrm{e}^{-\frac{\varepsilon}{2}}} = \frac{\cosh\frac{\varepsilon}{2}+\sinh\frac{\varepsilon}{2}}{2\cosh\frac{\varepsilon}{2}} = \frac{1}{2}(1+t)\\
q &= \frac{1}{\mathrm{e}^\varepsilon+1} = \frac{\mathrm{e}^{-\frac{\varepsilon}{2}}}{\mathrm{e}^{\frac{\varepsilon}{2}}+\mathrm{e}^{-\frac{\varepsilon}{2}}} = \frac{\cosh\frac{\varepsilon}{2}-\sinh\frac{\varepsilon}{2}}{2\cosh\frac{\varepsilon}{2}} = \frac{1}{2}(1-t).
\end{aligned}$$

The log likelihood ratio is $-\varepsilon$ with probability p and ε with probability q, so

$$\begin{aligned}
\mathrm{kl}(f_{\varepsilon,0}) &= -[(-\varepsilon)\cdot p + \varepsilon\cdot q] = \varepsilon(p-q) = \varepsilon t,\\
\kappa_2(f_{\varepsilon,0}) &= \varepsilon^2(p+q) = \varepsilon^2,\\
\kappa_3(f_{\varepsilon,0}) &= \varepsilon^3(p+q) = \varepsilon^3
\end{aligned}$$

and

$$\begin{aligned}
\bar{\kappa}_3(f_{\varepsilon,0}) &= p|-\varepsilon+\varepsilon(p-q)|^3 + q|\varepsilon+\varepsilon(p-q)|^3\\
&= 8\varepsilon^3\cdot pq(p^2+q^2)\\
&= 2\varepsilon^3\cdot(1-t^2)\cdot\frac{1}{2}(1+t^2)\\
&= \varepsilon^3(1-t^4).
\end{aligned}$$

□

Proof of Corollary D.6 . Follows directly from Theorem 3.4 and Lemma D.8. □

Proof of Corollary D.7. All we need to do is to simplify the expression of μ and γ assuming all $\varepsilon_i = \varepsilon$ and hence $t_i = t = \tanh\frac{\varepsilon}{2}$.

$$\begin{aligned}
\mu &= \frac{2n\varepsilon t}{\sqrt{n\varepsilon^2(1-t^2)}}\\
&= 2\sqrt{n}\cdot\frac{t}{1-t^2} = 2\sqrt{n}\sinh\frac{\varepsilon}{2}\\
\gamma &= 0.56\cdot\frac{n\varepsilon^3(1-t^4)}{\big(n\varepsilon^2(1-t^2)\big)^{3/2}}\\
&= \frac{0.56}{\sqrt{n}}\cdot\frac{1+t^2}{\sqrt{1-t^2}}\\
&= \frac{0.56}{\sqrt{n}}\cdot\frac{\cosh\varepsilon}{\cosh\frac{\varepsilon}{2}}.
\end{aligned}$$

The proof is complete. □

E Proof of Theorem 3.7

This section is devoted to the proof of Theorem 3.7. Since we always assume $\delta = 0$, it is dropped from the subscript and we use f_ε to denote $f_{\varepsilon,0}$. As in the proof of Theorem 3.4, the first step is to express $f_\varepsilon^{\otimes n}$ in the form

$$1 - f_\varepsilon^{\otimes n}(\alpha) = F_n\big[x_n - F_n^{-1}(1-\alpha)\big]$$

with $F_n \to \Phi$ and $x_n \to 1$. Then we show both convergences have rate $1/n$.

Proof of Theorem 3.7. Let's find F_n first. Fix ε and let $p = \frac{1}{1+e^\varepsilon}, q = 1 - p = p \cdot e^\varepsilon$. Recall that $f_\varepsilon^{\otimes n} = T\big(B(n,p), B(n,q)\big)$ and we know that it is the linear interpolation of points given by binomial tails. The main goal here is to avoid the linear interpolation.

For the simple hypothesis testing problem $B(n,p)$ vs $B(n,q)$, we know via Neyman-Pearson that every optimal rejection rule ϕ must have the following form:

$$\phi(x) = \begin{cases} 1, & \text{if } x > k, \\ 0, & \text{if } x < k, \\ 1-c, & \text{if } x = k. \end{cases}$$

It rejects (i.e. decides that the sample comes from $B(n,q)$) if it sees something greater than k, accepts if it sees something smaller than k, and reject with probability $1-c$ if it sees k. Such tests are parametrized by (k,c) where $k \in \{0, 1, \dots, n\}$ and $c \in [0,1)$.

The corresponding type I and type II errors are denoted by $\alpha_{(k,c)}$ and $\beta_{(k,c)}$. Let $X \sim B(n,p)$ and $Y \sim U[0,1]$ be independent random variables. We have

$$\begin{aligned}
\alpha_{(k,c)} &= \mathbb{E}_{x\sim B(n,p)}[\phi(x)] = \mathbb{E}[\phi(X)] \\
&= \mathbb{P}[X > k] + (1-c)\mathbb{P}[X = k] \\
&= \mathbb{P}[X > k] + \mathbb{P}[Y > c] \cdot \mathbb{P}[X = k] \\
&= \mathbb{P}[X + Y > k + c] \\
\beta_{(k,c)} &= \mathbb{E}_{x\sim B(n,q)}[1 - \phi(x)] = \mathbb{E}[1 - \phi(n - X)] \\
&= \mathbb{P}[n - X < k] + c \cdot \mathbb{P}[n - X = k] \\
&= \mathbb{P}[X > n - k] + \mathbb{P}[Y > 1 - c] \cdot \mathbb{P}[X = n - k] \\
&= \mathbb{P}[X + Y > n + 1 - k - c]
\end{aligned}$$

$X+Y$ supports on $[0, n+1]$ and has a piecewise constant density. As a consequence, the cdf F_{X+Y} is a bijection between $[0, n+1]$ and $[0,1]$. So for a fixed type I error $\alpha \in [0,1]$, the optimal testing rule (k,c) is uniquely determined by the formula

$$k + c = F_{X+Y}^{-1}(1-\alpha).$$

And we have for the trade-off function:

$$1 - f_\varepsilon^{\otimes n}(\alpha) = F_{X+Y}\big(n + 1 - F_{X+Y}^{-1}(1-\alpha)\big).$$

Now we proceed to write F_{X+Y} in a form that reveals its central limit behavior. First notice $\mathbb{E}[X+Y] = np + \frac{1}{2}, \mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] = npq + \frac{1}{12}$. For simplicity denote this variance by σ^2. Let F_n be the normalized cdf of $X+Y$, i.e.

$$F_n(x) = P\Big[\tfrac{X+Y-\mathbb{E}[X+Y]}{\sqrt{\mathrm{Var}[X+Y]}} \leqslant x\Big] = F_{X+Y}\Big[np + \tfrac{1}{2} + x\sigma\Big].$$

Simple algebra yields

$$1 - f_\varepsilon^{\otimes n}(\alpha) = F_n\Big[\tfrac{n(q-p)}{\sigma} - F_n^{-1}(1-\alpha)\Big]. \tag{33}$$

It's easy to show that $\frac{n(q-p)}{\sigma} \to 1$ and $F_n \to \Phi$ pointwise. However, we need to show that the convergence rates are both $1/n$, which is technically involved, especially for the convergence of F_n. In view of this, we pack the conclusions into the following lemmas, and provide the proofs later:

Lemma E.1. *With $\varepsilon = 1/\sqrt{n}$ and p, q, σ defined as above,*

$$\frac{n(q-p)}{\sigma} = 1 - \frac{1}{8n} + o(n^{-1}).$$

As a consequence, there exists $C > 0$ such that

$$\big|\tfrac{n(q-p)}{\sigma} - 1\big| \leqslant \tfrac{C}{n}. \tag{34}$$

Lemma E.2. *There is a positive number C such that $|F_n(x) - \Phi(x)| \leqslant \frac{C}{n}$ holds for $n \geqslant 2$.*

Since $\Phi(x) \geqslant F_n(x) - \frac{C}{n}$, setting $x = F_n^{-1}(1-\alpha)$ yields

$$\Phi\big(F_n^{-1}(1-\alpha)\big) \geqslant F_n\big(F_n^{-1}(1-\alpha)\big) - \tfrac{C}{n} = 1 - \alpha - \tfrac{C}{n}.$$

Hence

$$F_n^{-1}(1-\alpha) \geqslant \Phi^{-1}\big(1 - \alpha - \tfrac{C}{n}\big). \tag{35}$$

With (33–35) and Lemma E.2 we have

$$\begin{aligned}
1 - f_\varepsilon^{\otimes n}(\alpha) &= F_n\Big[\tfrac{n(q-p)}{\sigma} - F_n^{-1}(1-\alpha)\Big] \\
&\leqslant \Phi\Big[\tfrac{n(q-p)}{\sigma} - F_n^{-1}(1-\alpha)\Big] + \tfrac{C}{n} \\
&\leqslant \Phi\Big[1 + \tfrac{C}{n} - F_n^{-1}(1-\alpha)\Big] + \tfrac{C}{n} \\
&\leqslant \Phi\Big[1 + \tfrac{C}{n} - \Phi^{-1}(1 - \alpha - \tfrac{C}{n})\Big] + \tfrac{C}{n}.
\end{aligned}$$

The function Φ is $\frac{1}{\sqrt{2\pi}}$-Lipschitz, so

$$1 - f_\varepsilon^{\otimes n}(\alpha) \leqslant \Phi\Big[1 - \Phi^{-1}(1 - \alpha - \tfrac{C}{n})\Big] + \tfrac{1}{\sqrt{2\pi}} \cdot \tfrac{C}{n} + \tfrac{C}{n}.$$

By blowing up the current C and using the symmetry of standard normal, we have

$$\begin{aligned}
f_\varepsilon^{\otimes n}(\alpha) &\geqslant 1 - \Phi\Big[1 - \Phi^{-1}(1 - \alpha - \tfrac{C}{n})\Big] - \tfrac{C}{n} \\
&= \Phi\Big[\Phi^{-1}(1 - \alpha - \tfrac{C}{n}) - 1\Big] - \tfrac{C}{n} \\
&= G_1(\alpha + \tfrac{C}{n}) - \tfrac{C}{n}.
\end{aligned}$$

Similarly, we can show the upper bound

$$f_\varepsilon^{\otimes n}(\alpha) \leqslant G_1(\alpha - \tfrac{C}{n}) + \tfrac{C}{n}.$$

The proof is now complete. □

Next we show Lemma E.1 and Lemma E.2.

Proof of Lemma E.1. The proof is basically careful Taylor expansion. We will frequently use the assumption that $\varepsilon = 1/\sqrt{n}$. First we factor the objective as

$$\frac{n(q-p)}{\sigma} = 2\sqrt{n}(q-p) \cdot \frac{\sqrt{n}}{2\sigma} = \frac{2(q-p)}{\varepsilon} \cdot \frac{\sqrt{n}}{2\sigma}$$

and consider Taylor expansions of the two factors separately. For the first factor, recall that

$$q - p = \frac{e^{\varepsilon} - 1}{e^{\varepsilon} + 1} = \frac{e^{\frac{\varepsilon}{2}} - e^{-\frac{\varepsilon}{2}}}{e^{\frac{\varepsilon}{2}} + e^{-\frac{\varepsilon}{2}}} = \tanh \tfrac{\varepsilon}{2}.$$

Using the Taylor expansion $\tanh x = x - x^3/3 + o(x^4)$, we have

$$\tfrac{2(q-p)}{\varepsilon} = \tanh \tfrac{\varepsilon}{2} \,/\, \tfrac{\varepsilon}{2} = 1 - \tfrac{1}{3}(\tfrac{\varepsilon}{2})^2 + o(\varepsilon^3) = 1 - \tfrac{1}{12n} + o(n^{-3/2}). \tag{36}$$

For the second one, since $p + q = 1$, we have $4pq = (p+q)^2 - (p-q)^2 = 1 - (q-p)^2$. A shorter expansion shows $q - p = \tanh \frac{\varepsilon}{2} = \frac{\varepsilon}{2} + o(\varepsilon^2)$, and hence

$$4pq = 1 - \big(\tfrac{\varepsilon}{2} + o(\varepsilon^2)\big)^2 = 1 - \tfrac{\varepsilon^2}{4} + o(\varepsilon^3) = 1 - \tfrac{1}{4n} + o(n^{-3/2}).$$

Recall that σ is defined to be $\sqrt{npq + \frac{1}{12}}$. Using the above expansion of $4pq$, we have

$$\frac{\sqrt{n}}{2\sigma} = \sqrt{\frac{n}{4\sigma^2}} = \sqrt{\frac{n}{4npq + \frac{1}{3}}} = \big(4pq + \tfrac{1}{3n}\big)^{-1/2} = \big(1 + \tfrac{1}{12n} + o(n^{-3/2})\big)^{-1/2}$$

Since $(1+x)^{-1/2} = 1 - \frac{1}{2}x + o(x)$, we have

$$\frac{\sqrt{n}}{2\sigma} = 1 - \tfrac{1}{2}\big(\tfrac{1}{12n} + o(n^{-3/2})\big) + o(n^{-1}) = 1 - \tfrac{1}{24n} + o(n^{-1}). \tag{37}$$

Combining the expansions (36) and (37),

$$\begin{aligned}\frac{n(q-p)}{\sigma} &= \frac{2(q-p)}{\varepsilon} \cdot \frac{\sqrt{n}}{2\sigma} \\ &= \Big(1 - \frac{1}{12n} + o(n^{-3/2})\Big) \cdot \Big(1 - \frac{1}{24n} + o(n^{-1})\Big) \\ &= 1 - \frac{1}{8n} + o(n^{-1}).\end{aligned}$$

The proof is complete. □

Then we move on to the more challenging Lemma E.2.

Proof of Lemma E.2. The proof is inspired by Problem 6 on page 305 of [Usp37]. Though involved, the idea is not hard: reduce the bound on cdfs to a bound on characteristic functions (ch.f. for short) by an appropriate Fourier inversion, then control the ch.f. by careful Taylor expansion.

Recall that $\varepsilon, p, q, \sigma$ depend on n via

$$\varepsilon = \frac{1}{\sqrt{n}}, \quad p = \frac{1}{1+e^{\varepsilon}}, \quad q = \frac{e^{\varepsilon}}{1+e^{\varepsilon}}, \quad \sigma = \sqrt{npq + \tfrac{1}{12}}.$$

Random variables $X \sim B(n,p), Y \sim U[0,1]$. F_n is the normalized cdf of $X+Y$. More precisely, since

$$\mathbb{E}[X+Y] = np + \frac{1}{2}, \quad \operatorname{Var}[X+Y] = \operatorname{Var} X + \operatorname{Var} Y = npq + \tfrac{1}{12} = \sigma^2,$$

F_n is the cdf of $\sigma^{-1}(X+Y-\frac{1}{2}-np)$. Our goal is to show that $\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| = O(\frac{1}{n})$.

First let's compute the characteristic function (ch.f. for short) φ_n of the distribution F_n.

$$\begin{aligned}\varphi_n(t) &= \mathbb{E}[e^{it\sigma^{-1}(X+Y-\frac{1}{2}-np)}] \\ &= e^{-inpt/\sigma} \cdot \mathbb{E}[e^{it/\sigma(X+Y-\frac{1}{2})}] \\ &= e^{-inpt/\sigma} \cdot \varphi_X(t/\sigma) \cdot \varphi_{Y-\frac{1}{2}}(t/\sigma).\end{aligned}$$

Easy calculation shows that the ch.f. of X is $(pe^{it}+q)^n$ and that of $Y - \frac{1}{2}$ is $\frac{\sin t/2}{t/2}$. So

$$\begin{aligned}\varphi_n(t) &= e^{-inpt/\sigma} \cdot (pe^{it/\sigma}+q)^n \cdot \tfrac{\sin t/2\sigma}{t/2\sigma} \\ &= (pe^{iqt/\sigma} + qe^{-ipt/\sigma})^n \cdot \tfrac{\sin t/2\sigma}{t/2\sigma}.\end{aligned}$$

The base $pe^{iqt/\sigma} + qe^{-ipt/\sigma}$ is a convex combination of two complex numbers on the unit circle, so we have $|\varphi_n(t)| \leqslant \frac{|\sin t/2\sigma|}{|t/2\sigma|} \leqslant \min\{\frac{2\sigma}{|t|}, 1\}$.

Now let's connect back to cdf. We need some form of Fourier inversion formula. Let $\varphi(t) = e^{-t^2/2}$ be the ch.f. of the standard normal.

Lemma E.3. *We have the following inversion formula*

$$F_n(x) - \Phi(x) = -\frac{1}{2\pi i}\int_{-\infty}^{+\infty} e^{-itx} \cdot \frac{\varphi_n(t) - \varphi(t)}{t} \,\mathrm{d}t.$$

The integrand is integrable over $\mathbb{R}$ because: (1) At infinity $|\varphi_n(t)| = O(\frac{1}{t})$, so the integrand has modulus $O(\frac{1}{t^2})$; (2) When $t \to 0$,

$$\tfrac{\varphi_n(t)-\varphi(t)}{t} = \tfrac{\varphi_n(t)-1-\varphi(t)+1}{t} = \tfrac{\varphi_n(t)-\varphi_n(0)}{t} - \tfrac{\varphi(t)-\varphi(0)}{t} \to \varphi_n'(0) - \varphi'(0) = \mathbb{E}_{Z\sim F_n}[Z]$$

is a finite number. So the integrand is continuous at 0.

Lemma E.3 makes it possible to control $F_n(x) - \Phi(x)$ by controling $\varphi_n(t) - \varphi(t)$.

$$\begin{aligned}2\pi|F_n(x) - \Phi(x)| \leqslant\; & \int_{-\infty}^{+\infty} \frac{|\varphi_n(t) - \varphi(t)|}{|t|}\,\mathrm{d}t \\ \leqslant\; & \int_{|t|\leqslant r\sigma} \frac{|\varphi_n(t) - \varphi(t)|}{|t|}\,\mathrm{d}t && (I_1) \\ & + \int_{|t|>r\sigma} \frac{|\varphi_n(t)|}{|t|}\,\mathrm{d}t && (I_2) \\ & + \int_{|t|>r\sigma} \frac{|\varphi(t)|}{|t|}\,\mathrm{d}t && (I_3)\end{aligned}$$

It suffices to find some constant r such that all three integrals are $O(\frac{1}{n})$. This is done via the following three lemmas.

Lemma E.4. *There exist universal constants $r > 0, C > 0$ such that when $|t| \leqslant r\sigma$,*

$$|\varphi_n(t) - \varphi(t)| \leqslant Ce^{-\frac{t^2}{8}} \cdot \big(\tfrac{t^2}{n} + \tfrac{|t|^3}{n} + \tfrac{t^4}{n}\big).$$

Consequently,

$$\begin{aligned} I_1 &= \int_{|t|\leqslant r\sigma} \frac{|\varphi_n(t) - \varphi(t)|}{|t|} \,\mathrm{d}t \\ &\leqslant \int_{\mathbb{R}} Ce^{-\frac{t^2}{8}} \cdot \big(\tfrac{|t|}{n} + \tfrac{t^2}{n} + \tfrac{|t|^3}{n}\big) \,\mathrm{d}t = O(\tfrac{1}{n}). \end{aligned}$$

Lemma E.5. *For $r < \pi$, we have $I_2 \leqslant (2 + \frac{48}{r^2}) \cdot \frac{1}{n}$.*

Lemma E.6. *For $n \geqslant 2$, $I_3 \leqslant \frac{10}{r^2} \cdot \frac{1}{n} \cdot e^{-0.1r^2 n}$ holds for any positive r.*

So we can select a small enough r such that all three estimates hold, which implies $I_1 = O(\frac{1}{n}), I_2 = O(\frac{1}{n})$ and $I_3 \ll \frac{1}{n}$. In summary,

$$|F_n(x) - \Phi(x)| \leqslant \frac{1}{2\pi}(I_1 + I_2 + I_3) = O(\frac{1}{n}).$$

Assuming correctness of Lemmas E.3 to E.6, the proof of Theorem 3.7 is complete. □

The rest is to prove Lemmas E.3 to E.6. We deal with the three integrals first, and then come back to inversion formula.

Proof of Lemma E.4. Let $w = pe^{itq/\sigma} + qe^{-itp/\sigma}$. Then $\varphi_n(t) = w^n \cdot \frac{\sin t/2\sigma}{t/2\sigma}$. We have

$$\begin{aligned} |\varphi_n(t) - \varphi(t)| &= |w^n \cdot \frac{\sin t/2\sigma}{t/2\sigma} - e^{-\frac{1}{2}t^2}| \\ &\leqslant |w^n - e^{-\frac{1}{2}t^2}| + |w|^n \cdot \Big|1 - \frac{\sin t/2\sigma}{t/2\sigma}\Big| \end{aligned} \tag{38}$$

All we need is a positive r such that when $|t| \leqslant r\sigma$, both of the above terms are small. We are going to shrink r as we need from time to time.

First, on the disk $|z| \leqslant r$ we have Taylor expansion $e^z = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + O(|z|^4)$. So for $|t| \leqslant r\sigma$ we have

$$\begin{aligned} pe^{itq/\sigma} &= p\big(1 + \tfrac{itq}{\sigma} + \tfrac{1}{2}\big(\tfrac{itq}{\sigma}\big)^2 + \tfrac{1}{6}\big(\tfrac{itq}{\sigma}\big)^3 + O(\tfrac{t^4}{\sigma^4})\big) \\ qe^{-itp/\sigma} &= q\big(1 - \tfrac{itp}{\sigma} + \tfrac{1}{2}\big(\tfrac{itp}{\sigma}\big)^2 - \tfrac{1}{6}\big(\tfrac{itp}{\sigma}\big)^3 + O(\tfrac{t^4}{\sigma^4})\big) \\ w &= 1 - \tfrac{t^2}{2\sigma^2} \cdot (pq^2 + qp^2) + \tfrac{t^3}{6\sigma^3}\big(- ipq^3 - qp^3(-i)\big) + O(\tfrac{t^4}{\sigma^4}) \\ &= 1 - \tfrac{pq}{\sigma^2} \cdot \tfrac{t^2}{2} + \tfrac{t^3}{6\sigma^3} \cdot ipq(p-q) + O(\tfrac{t^4}{\sigma^4}). \end{aligned} \tag{39}$$

Obviously this implies $w = 1 - \frac{pq}{\sigma^2} \cdot \frac{t^2}{2} + o(\frac{t^2}{\sigma^2})$ (we will return to the more delicate (39) soon). Since $npq \geqslant pq \geqslant \frac{1}{4} - \frac{1}{16n} \geqslant \frac{3}{16}$, we have

$$\tfrac{pq}{\sigma^2} = \tfrac{pq}{npq+\frac{1}{12}} = \tfrac{npq}{npq+\frac{1}{12}} \cdot \tfrac{1}{n} \geqslant \tfrac{1}{n} \cdot \tfrac{3}{16}/(\tfrac{3}{16} + \tfrac{1}{12}) = \tfrac{9}{13n} > \tfrac{2}{3n}.$$

That is, the quadratic term is more than $\frac{t^2}{3n}$. We can tune r so that the little o remainder is even smaller, i.e. $|w - 1 + \frac{pq}{\sigma^2} \cdot \frac{t^2}{2}| < \frac{t^2}{12n}$. This implies

$$|w| < 1 - \tfrac{t^2}{3n} + \tfrac{t^2}{12n} = 1 - \tfrac{t^2}{4n} \leqslant e^{-\frac{t^2}{4n}}.$$

One consequence is we can bound the second term in (38). By Taylor expansion again, $\frac{\sin x}{x} = 1 + O(x^2)$, so

$$|w|^n \cdot \left|1 - \tfrac{\sin t/2\sigma}{t/2\sigma}\right| \leqslant e^{-\frac{t^2}{4}} \cdot O(\tfrac{t^2}{\sigma^2}) = e^{-\frac{t^2}{4}} \cdot O(\tfrac{t^2}{n}). \tag{40}$$

The first term in (38) requires a more careful analysis. Let $z = e^{-\frac{t^2}{2n}}$ and $\gamma = e^{-\frac{t^2}{4n}}$. Our goal is $|w^n - z^n|$. We have proved $|w| < \gamma$, while $|z| = \gamma^2 < \gamma$ is obviously true. We have

$$|w^n - z^n| \leqslant |w^n - w^{n-1}z| + \cdots + |wz^{n-1} - z^n| \leqslant n|w - z| \cdot \gamma^{n-1}.$$

Without loss of generality assume $n \geqslant 2$, then $\gamma^{n-1} = e^{-\frac{t^2}{4} \cdot \frac{n-1}{n}} \leqslant e^{-\frac{t^2}{8}}$. That is,

$$|w^n - e^{-\frac{1}{2}t^2}| \leqslant n|w - e^{-\frac{t^2}{2n}}| \cdot e^{-\frac{1}{8}t^2}. \tag{41}$$

For $n|w - e^{-\frac{t^2}{2n}}|$ we need (39) again. First decompose it as

$$n|w - e^{-\frac{t^2}{2n}}| \leqslant n\left|w - 1 + \tfrac{t^2}{2n}\right| + n\left|e^{-\frac{t^2}{2n}} - 1 + \tfrac{t^2}{2n}\right|. \tag{42}$$

Since $|p - q| = \frac{e^\varepsilon - 1}{e^\varepsilon + 1} \leqslant e^\varepsilon - 1 = O(\varepsilon) = O(\frac{1}{\sqrt{n}})$ and $\sigma^{-1} = O(\frac{1}{\sqrt{n}})$, we have

$$w = 1 - \tfrac{pq}{\sigma^2} \cdot \tfrac{t^2}{2} + \tfrac{t^3}{6\sigma^3} \cdot ipq(p - q) + O(\tfrac{t^4}{\sigma^4}) = 1 - \tfrac{pq}{\sigma^2} \cdot \tfrac{t^2}{2} + O(\tfrac{t^3}{n^2}) + O(\tfrac{t^4}{n^2}).$$

Note that neither of two "big O" dominate each other, because t can be as small as 0 and as large as $r\sigma = O(\sqrt{n})$. Using the more delicate expansion of w and that $\sigma^2 = npq + \frac{1}{12}$, we have

$$\begin{aligned} n\left|w - 1 + \tfrac{t^2}{2n}\right| &= n\left|\tfrac{t^2}{2n} - \tfrac{pq}{\sigma^2} \cdot \tfrac{t^2}{2} + O(\tfrac{t^3}{n^2}) + O(\tfrac{t^4}{n^2})\right| \\ &= \left|\tfrac{t^2}{2} - \tfrac{npq}{\sigma^2} \cdot \tfrac{t^2}{2} + O(\tfrac{t^3}{n}) + O(\tfrac{t^4}{n})\right| \\ &= \left|\tfrac{t^2}{2} \cdot \tfrac{1}{12\sigma^2} + O(\tfrac{t^3}{n^2}) + O(\tfrac{t^4}{n^2})\right| \\ &= O(\tfrac{t^2}{n^2}) + O(\tfrac{t^3}{n^2}) + O(\tfrac{t^4}{n^2}). \end{aligned} \tag{43}$$

By Taylor expansion again, we can tune r so that when $|t| \leqslant r\sigma$, we have

$$n\left|e^{-\frac{t^2}{2n}} - 1 + \tfrac{t^2}{2n}\right| = n \cdot O(\tfrac{t^4}{n^2}) = O(\tfrac{t^4}{n}). \tag{44}$$

Now plug (43) and (44) back into (42), and then into (41) to get

$$|w^n - e^{-\frac{1}{2}t^2}| \leqslant e^{-\frac{1}{8}t^2} \cdot \left(O(\tfrac{t^2}{n^2}) + O(\tfrac{t^3}{n^2}) + O(\tfrac{t^4}{n^2})\right).$$

This ends the analysis of the first term of (38). Combining with the estimate of the first term, we have

$$|\varphi_n(t) - \varphi(t)| \leqslant e^{-\frac{1}{8}t^2} \cdot \big(O(\tfrac{t^2}{n^2}) + O(\tfrac{t^3}{n^2}) + O(\tfrac{t^4}{n^2})\big).$$

□

Proof of Lemma E.5. For this integral we only care about the modulus of $|\varphi_n(t)|$. Let's simplify it first.

$$|\varphi_n(t)| = |pe^{iqt/\sigma} + qe^{-ipt/\sigma}|^n \cdot \left|\tfrac{\sin t/2\sigma}{t/2\sigma}\right|$$

Let $\theta = t/\sigma$. Using $|z|^2 = z\bar{z}$, we have

$$\begin{aligned}
|pe^{iq\theta} + qe^{-ip\theta}|^2 &= \big(pe^{iq\theta} + qe^{-ip\theta}\big) \cdot \big(pe^{-iq\theta} + qe^{ip\theta}\big) \\
&= p^2 + q^2 + pq(e^{i\theta} + e^{-i\theta}) \\
&= 1 - 2pq + 2pq\cos\theta \\
&= 1 - 4pq\sin^2\tfrac{\theta}{2}.
\end{aligned}$$

So

$$|\varphi_n(t)| = \big(1 - 4pq\sin^2\tfrac{t}{2\sigma}\big)^{n/2} \cdot \left|\tfrac{\sin t/2\sigma}{t/2\sigma}\right|.$$

We see from this expression that the integrand of I_2 is an even function. Therefore,

$$\begin{aligned}
\frac{1}{2}I_2 &= \int_{r\sigma}^{+\infty} \tfrac{|\varphi_n(t)|}{t}\,\mathrm{d}t \\
&= \int_{r\sigma}^{+\infty} \tfrac{1}{t}\big(1 - 4pq\sin^2\tfrac{t}{2\sigma}\big)^{n/2} \cdot \left|\tfrac{\sin t/2\sigma}{t/2\sigma}\right| \mathrm{d}t \\
&= \int_{r/2}^{+\infty} \tfrac{1}{t^2}\big(1 - 4pq\sin^2 t\big)^{n/2} \cdot |\sin t|\,\mathrm{d}t
\end{aligned}$$

In the last step we do a change of variable $s = t/2\sigma$ and rename s to t. Next, we break down the integral at $k\pi$, and upper bound the $\frac{1}{t^2}$ factor by its value at the left end of the interval, so that the rest of the integrand is periodic.

$$\begin{aligned}
\frac{1}{2}I_2 \leqslant\ & \int_{r/2}^{\pi} \tfrac{1}{t^2}\big(1 - 4pq\sin^2 t\big)^{n/2} \cdot |\sin t|\,\mathrm{d}t \\
&+ \sum_{k=1}^{+\infty} \int_{k\pi}^{(k+1)\pi} \tfrac{1}{t^2}\big(1 - 4pq\sin^2 t\big)^{n/2} \cdot |\sin t|\,\mathrm{d}t \\
\leqslant\ & \Big(\tfrac{4}{r^2} + \sum_{k=1}^{+\infty} \tfrac{1}{k^2\pi^2}\Big) \underbrace{\int_0^{\pi} \big(1 - 4pq\sin^2 t\big)^{n/2} \cdot \sin t\,\mathrm{d}t}_{J}
\end{aligned}$$

The integral J can be estimated as follows:

$$\begin{aligned} J &= \int_0^\pi \left(1 - 4pq\sin^2 t\right)^{n/2} \cdot \sin t \,\mathrm{d}t \\ &= -\int_0^\pi \left(1 - 4pq(1-\cos^2 t)\right)^{n/2} \mathrm{d}\cos t \\ &= \int_{-1}^1 \left(1 - 4pq(1-x^2)\right)^{n/2} \mathrm{d}x \\ &= 2\int_0^1 (1 - 4pq + 4pqx^2)^{n/2} \,\mathrm{d}x. \end{aligned}$$

We have seen that $1 - 4pq = (p-q)^2 = \tanh^2 \frac{\varepsilon}{2}$. It is easy to show that $\tanh x \leqslant x$ for $x \geqslant 0$. So

$$pq = \tfrac{1}{4}(1 - \tanh^2 \tfrac{\varepsilon}{2}) \geqslant \tfrac{1}{4}(1 - \tfrac{\varepsilon^2}{4}) = \tfrac{1}{4} - \tfrac{1}{16n}.$$

Since $0 \leqslant x \leqslant 1$, we have

$$1 - 4pq + 4pqx^2 \leqslant \tfrac{1}{4n} + (1 - \tfrac{1}{4n})x^2.$$

Hence

$$J \leqslant 2\int_0^1 \left(\tfrac{1}{4n} + (1 - \tfrac{1}{4n})x^2\right)^{n/2} \mathrm{d}x.$$

It's easy to check that $\frac{1}{4n-1}$ and 1 are the two roots of the quadratic equation $\frac{1}{4n} + (1 - \frac{1}{4n})x^2 = x$. So we have $\frac{1}{4n} + (1 - \frac{1}{4n})x^2 \leqslant x$ between the two roots, i.e. for $x \in [\frac{1}{4n-1}, 1]$. For the rest of the interval, we upper bound the integrand by 1. That is,

$$\int_0^1 \left(\tfrac{1}{4n} + (1 - \tfrac{1}{4n})x^2\right)^{n/2} \mathrm{d}x \leqslant \int_0^{\frac{1}{4n-1}} 1 \,\mathrm{d}x + \int_{\frac{1}{4n-1}}^1 x^{n/2} \,\mathrm{d}x \leqslant \tfrac{1}{4n-1} + \tfrac{1}{n/2+1} \leqslant \tfrac{3}{n}.$$

So we have $J \leqslant \frac{6}{n}$. Returning to I_2, with the well-known identity $\sum_{k=1}^{+\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$, we have

$$\begin{aligned} I_2 &\leqslant 2\Big(\tfrac{4}{r^2} + \sum_{k=1}^{+\infty} \tfrac{1}{k^2\pi^2}\Big) \cdot J \\ &= \left(\tfrac{8}{r^2} + \tfrac{\pi^2}{6} \cdot \tfrac{2}{\pi^2}\right) \cdot \tfrac{6}{n} \\ &= \left(2 + \tfrac{48}{r^2}\right) \cdot \tfrac{1}{n} \end{aligned}$$

The estimate of I_2 is complete. □

Proof of Lemma E.6. First notice the following simple facts:

1. When $t > r\sigma$, we have $\frac{1}{t} \leqslant t \cdot \frac{1}{r^2\sigma^2}$.

2. $\sigma^2 > 0.2n$ for any n.

The second follows from a bound we derive in the proof of Lemma E.5: $pq \geqslant \frac{1}{4} - \frac{1}{16n}$. In fact,

$$\sigma^2 = npq + \tfrac{1}{12} \geqslant \tfrac{n}{4} - \tfrac{1}{16} + \tfrac{1}{12} = \tfrac{n}{4} - \tfrac{1}{48} > \tfrac{n}{5}.$$

Using these two facts, we can bound I_3 as follows:

$$\begin{aligned} I_3 &= \int_{|t|>r\sigma} \frac{|\varphi(t)|}{|t|} \,\mathrm{d}t \\ &= 2\int_{r\sigma}^{+\infty} \frac{1}{t} e^{-\frac{t^2}{2}} \,\mathrm{d}t \\ &\leqslant \frac{2}{r^2\sigma^2} \cdot \int_{r\sigma}^{+\infty} te^{-\frac{t^2}{2}} \,\mathrm{d}t \\ &= \frac{2}{r^2\sigma^2} \cdot e^{-\frac{t^2}{2}}\Big|_{+\infty}^{r\sigma} \\ &= \frac{2}{r^2\sigma^2} \cdot e^{-\frac{r^2\sigma^2}{2}} \\ &\leqslant \frac{10}{r^2} \cdot \frac{1}{n} \cdot e^{-0.1r^2 n} \end{aligned}$$

The estimate of I_3 is complete. □

We are done with the three integrals. Before we dive into the proof of the inversion formula E.3, we make a few observations.

First, one cannot hope to obtain this lemma by showing

$$F_n(x) = \tfrac{1}{2\pi} \int_{-\infty}^{+\infty} -e^{-itx} \cdot \tfrac{\varphi_n(t)}{it} \,\mathrm{d}t$$

and a similar expression for $\Phi(x)$ separately because this alternative integrand is not even integrable. To see this, notice $\varphi_n(0) = 1$, so the integrand $\approx \frac{1}{t}$ around 0.

Inversion formula E.3 has the same form as Lemma 3.4.19 of [Dur19]. However, the ch.f.s are assumed to be (absolutely) integrable there, while φ_n is not. To see this, recall that Fourier inversion tells us that if the ch.f. is absolutely integrable, then the probability distribution has continuous density (see e.g. [Dur19], Theorem 3.3.14). This is not true for $X + Y$ because its density is piecewise constant. So φ_n cannot be in $L^1(\mathbb{R})$. There seems to be no shortcut, so let's work out our own proof.

Proof of Lemma E.3. Applying the general inversion formula (see e.g. [Dur19] Theorem 3.3.11) to F_n, we have

$$F_n(x) - F_n(a) = \frac{1}{2\pi} \lim_{T\to+\infty} \int_{-T}^{T} \frac{e^{-ita} - e^{-itx}}{it} \cdot \varphi_n(t) \,\mathrm{d}t$$

φ_n is continuous and decays in the rate $\frac{1}{|t|}$, so the integrand is dominated by $O(t^{-2} \wedge 1)$ and hence the limit on T is equal to the Lebesgue integral. That is,

$$F_n(x) - F_n(a) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{e^{-ita} - e^{-itx}}{it} \cdot \varphi_n(t) \,\mathrm{d}t. \tag{45}$$

Similarly,

$$\Phi(x) - \Phi(a) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{e^{-ita} - e^{-itx}}{it} \cdot \varphi(t)\,\mathrm{d}t.$$

Note that in (45), we cannot let $a \to -\infty$ and use Riemann-Lebesgue lemma because $\frac{\varphi_n(t)}{t}$ is not integrable, as discussed before the proof. However, subtracting the two formula yields

$$\big(F_n(x) - \Phi(x)\big) - \big(F_n(a) - \Phi(a)\big) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{e^{-ita} - e^{-itx}}{it} \cdot \big(\varphi_n(t) - \varphi(t)\big)\,\mathrm{d}t \tag{46}$$

Consider the part involving a

$$\int_{-\infty}^{+\infty} e^{-ita} \cdot \frac{\varphi_n(t) - \varphi(t)}{it}\,\mathrm{d}t.$$

We argued right after introducing Lemma E.3 that $\frac{\varphi_n(t)-\varphi(t)}{it} \in L^1(\mathbb{R})$, so by Riemann-Lebesgue lemma we have the limit

$$\lim_{a\to-\infty}\int_{-\infty}^{+\infty} e^{-ita} \cdot \frac{\varphi_n(t) - \varphi(t)}{it}\,\mathrm{d}t = 0.$$

Take the limit $a \to -\infty$ on both sides of (46) and we have

$$\begin{aligned} F_n(x) - \Phi(x) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{-e^{-itx}}{it} \cdot \big(\varphi_n(t) - \varphi(t)\big)\,\mathrm{d}t \\ &= -\frac{1}{2\pi i}\int_{-\infty}^{+\infty} e^{-itx} \cdot \frac{\varphi_n(t) - \varphi(t)}{t}\,\mathrm{d}t. \end{aligned}$$

The proof is now complete. □

F Omitted Details in Section 4

We begin this appendix with a small example showing our subsampling theorem is generically unimprovable.

Tightness Consider the mechanism $\widetilde{M}$ that randomly releases one individual's private information in the dataset. The privacy analysis is easy: without loss of generality we can assume two neighboring datasets differ in the first individual. Effectively we are trying to distinguish uniform distributions over $\{1, 2, \ldots, n\}$ and $\{1', 2, \ldots, n\}$. It's not hard to see that the trade-off function of these two uniform distributions is $f_{0,1/n}$, i.e. (ε, δ)-DP with $\varepsilon = 0, \delta = 1/n$. This is exact — the adversary has tests that achieve every point on the curve.

Our theorem 4.2 yields the same result, showing its tightness. To see this, let M be the identity map that takes in one individual and outputs his/her entire private information. Then $\widetilde{M} = M \circ \mathtt{Sample}_{\frac{1}{n}}$. Privacy of M is described by $f \equiv 0$. By Theorem 4.2, $\widetilde{M}$ is $C_{1/n}(f)$-DP. Figure 6 shows that $C_{1/n}(f) = f_{0,1/n}$.

Next we show the following two equations:

$$\begin{aligned} \varepsilon' &= \log(1 - p + pe^{\varepsilon}), \\ \delta' &= p\big(1 + f^*(-e^{\varepsilon})\big) \end{aligned}$$

can be re-parametrized into

$$\delta' = 1 + f_p^*(-e^{\varepsilon'}) \tag{15}$$

where $f_p = pf + (1-p)\mathrm{Id}$.

Proof of Equation (15). Since $\varepsilon \mapsto \log(1-p+pe^{\varepsilon})$ maps $[0,+\infty)$ to $[0,+\infty)$ monotonically, we can solve ε from ε' and plug into δ'. We have

$$\frac{1}{p}(1-e^{\varepsilon'}) = 1-e^{\varepsilon} \quad \text{and} \quad \delta' = p\big(1+f^*(-e^{\varepsilon})\big) = p\big(1+f^*(\tfrac{1}{p}(1-e^{\varepsilon'})-1)\big).$$

Let $y = -e^{\varepsilon'}$ and it suffices to show for any $y \leqslant -1$,

$$1 + f_p^*(y) = p\big(1+f^*(\tfrac{1}{p}(1+y)-1)\big). \tag{47}$$

To see this, expand f_p^* as follows

$$\begin{aligned} f_p^*(y) &= \sup_x yx - f_p(x) \\ &= \sup_x yx - pf(x) - (1-p)(1-x) \\ &= p - 1 + \sup_x (y+1-p)x - pf(x) \\ &= p - 1 + p\cdot\sup_x (\tfrac{1}{p}(1+y)-1)x - f(x) \\ &= p - 1 + pf^*(\tfrac{1}{p}(1+y)-1) \end{aligned}$$

(47) follows directly. □

Next we provide the general tool mentioned in Section 4.2 that convert collections of (ε,δ)-DP guarantee in the form of (15) to some f-DP.

The symmetrization operator $\mathrm{Symm}: \mathscr{F} \to \mathscr{F}^S$ maps a general trade-off function to a symmetric trade-off function. It's defined as follows:

Definition F.1. *For $f \in \mathscr{F}$, let $\bar{x} = \inf\{x \in [0,1] : -1 \in \partial f(x)\}$. The symmetrization operator* $\mathrm{Symm}: \mathscr{F} \to \mathscr{F}^S$ *is defined as*

$$\mathrm{Symm}(f) := \begin{cases} \min\{f, f^{-1}\}^{**}, & \text{if } \bar{x} \leqslant f(\bar{x}), \\ \max\{f, f^{-1}\}, & \text{if } \bar{x} > f(\bar{x}). \end{cases}$$

Proposition F.2. *Let $f \in \mathscr{F}$, not necessarily symmetric. Suppose a mechanism is $(\varepsilon, 1+f^*(-e^{\varepsilon}))$-DP for all $\varepsilon \geqslant 0$, then it is* $\mathrm{Symm}(f)$*-DP.*

Recall from basic convex analysis that double convex conjugate f^{**} is the greatest convex lower bound of f. If f itself is convex then $f^{**} = f$. For f symmetric , $f = f^{-1}$. By convexity of f, we have $\mathrm{Symm}(f) = f$ in both cases. So Proposition 2.12 is a special case of Proposition F.2. The first half of Proposition F.2 is Proposition 4.5, the part we used in the proof of our subsampling theorem.

From Figure 13 it's not hard to see that

$$\min\{f, f^{-1}\}^{**}(x) = \begin{cases} f(x), & x \in [0,\bar{x}], \\ \bar{x} + f(\bar{x}) - x, & x \in [\bar{x}, f(\bar{x})], \\ f^{-1}, & x \in [f(\bar{x}), 1]. \end{cases}$$

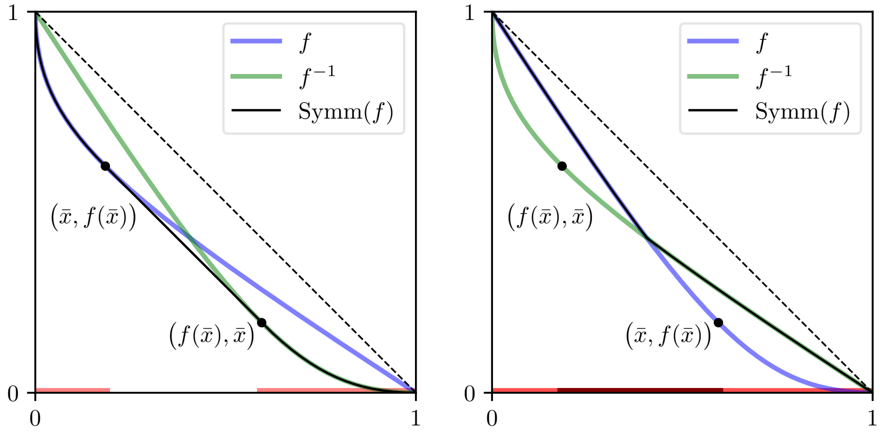

Figure 13: Action of Symm. Left panel: $\bar{x} \leqslant f(\bar{x})$. Right panel: $\bar{x} > f(\bar{x})$. For both panels the effective parts (red bars on x-axes) are $[0, \bar{x}]$ of f and $[f(\bar{x}), 1]$ of f^{-1}. No overlap in the left panel since $\bar{x} < f(\bar{x})$, so interpolate with straight line; overlap in the right panel so the max is taken.

Proof of Proposition F.2. M being $\big(\varepsilon, \delta(\varepsilon)\big)$-DP means that for any neighboring datasets S and S',

$$T\big(M(S), M(S')\big)(x) \geqslant -e^{\varepsilon} x + 1 - \delta(\varepsilon).$$

Fix $x \in [0,1]$. Since the DP condition holds for all $\varepsilon \geqslant 0$, the lower bound still holds when we take the supremum over $\varepsilon \geqslant 0$. In other words, M is f_{env}-DP with

$$f_{\mathrm{env}}(x) = \max\{0, \sup_{\varepsilon \geqslant 0} 1 - \delta(\varepsilon) - e^{\varepsilon} x\}.$$

By Proposition 2.4 M is also $\max\{f_{\mathrm{env}}, f_{\mathrm{env}}^{-1}\}$-DP. The proof will be complete if we can show $\max\{f_{\mathrm{env}}, f_{\mathrm{env}}^{-1}\} = \mathrm{Symm}(f)$.

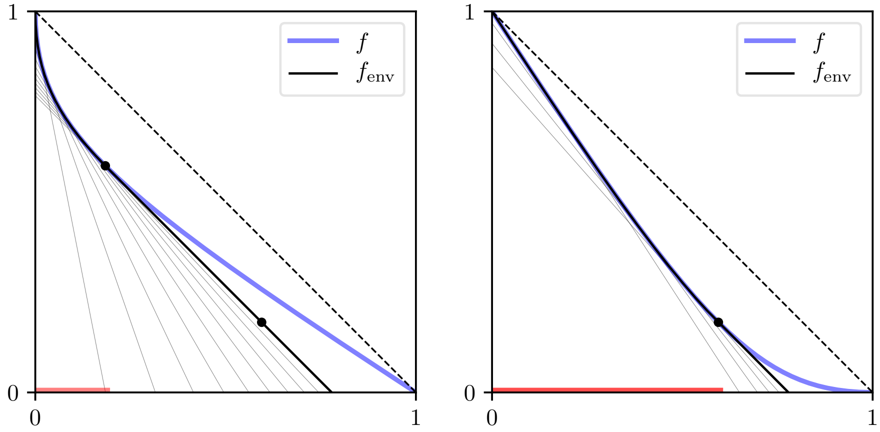

Figure 14: Symm explained. Left panel: $\bar{x} \leqslant f(\bar{x})$. Right panel: $\bar{x} > f(\bar{x})$.

We achieve this by first showing:

$$f_{\mathrm{env}}(x) = \begin{cases} f(x), & x \in [0, \bar{x}], \\ \bar{x} + f(\bar{x}) - x, & x \in [\bar{x}, \bar{x} + f(\bar{x})], \\ 0, & x \in [\bar{x} + f(\bar{x}), 1]. \end{cases}$$

From Figure 14 it is almost obvious. We still provide the argument below.

Plug in $\delta(\varepsilon) = 1 + f^*(-e^\varepsilon)$ and change the variable $y = -e^\varepsilon$:

$$\begin{aligned}\sup_{\varepsilon \geqslant 0}[-e^\varepsilon x + 1 - \delta(\varepsilon)] &= \sup_{\varepsilon \geqslant 0}[-e^\varepsilon x - f^*(-e^\varepsilon)] \\ &= \sup_{y \leqslant -1}[yx - f^*(y)]\end{aligned}$$

From convex analysis we know if $y \in \partial f(x)$ then $yx = f(x) + f^*(y)$. By definition of $\bar{x}$, if $x \leqslant \bar{x}$, then at least one subgradient $y \in \partial f(x)$ is no greater than -1. So this specific $yx - f^*(y) = f(x)$ is involved in the supremum, i.e. $\sup_{y \leqslant -1}[yx - f^*(y)] = f(x)$. This justifies the expression for the first segment.

When $x > \bar{x}$, the supremum is always attained at $y = -1$. In fact, if we let $l_y(x) = yx - f^*(y)$, then

Lemma F.3. *$l_y(x) \leqslant l_{-1}(x)$ when $y \leqslant -1$ and $x > \bar{x}$.*

Proof of Lemma F.3. l_y is the supporting linear function of f with slope y. It suffices to show that $l_y(x)$ is monotone increasing in y. To see this, change the variable from the slope y to the supporting location u. As f is convex, $y = f'(u)$ is increasing in u. In terms of u, $l_y(x) = f(u) + f'(u)(x - u)$. Taking derivative with respect to u:

$$\frac{\partial}{\partial u} l_y(x) = f'(u) + f''(u)(x-u) + f'(u)\cdot(-1) = f''(u)(x-u).$$

$y \leqslant -1$ corresponds to location $u \leqslant \bar{x}$ and hence $u < x$. So we see that $\frac{\partial}{\partial u} l_y(x) = f''(u)(x-u) \geqslant 0$. $l_y(x)$ is increasing in u, and hence increasing in y, completing the proof of the lemma. □

So the supremum is attained at $y = -1$. The value is $\sup_{y \leqslant -1}\big[yx - f^*(y)\big] = l_{-1}(x)$. Support function with slope -1 must support f at $\bar{x}$. The location yields expression $l_{-1}(x) = f(\bar{x}) - (x - \bar{x})$. This justifies the expression for the second segment. The third one is simply the result of thresholding at 0.

It's straightforward to verify that

$$f_{\text{env}}^{-1}(x) = \begin{cases} \bar{x} + f(\bar{x}) - x, & x \in [0, f(\bar{x})], \\ f^{-1}(x), & x \in [f(\bar{x}), 1]. \end{cases}$$

Obviously $\max\{f_{\text{env}}, f_{\text{env}}^{-1}\} = \mathrm{Symm}(f)$. When the intervals $[0, \bar{x}]$ and $[f(\bar{x}), 1]$ are disjoint, f and f^{-1} are effective separately, and the linear interpolation fills the blank. On the other hand when they intersect, max is taken and the linear function is never effective. □

We conclude this appendix with the proof of the classical privacy amplifaction by subsampling theorem. It primarily follows [Ull17], but is written so that potential generalization and improvement are in reach.

Lemma 4.4 ([Ull17])**.** *If M is (ε, δ)-DP, then $M \circ \texttt{Sample}_m$ is (ε', δ')-DP with ε' and δ' defined in Corollary 4.3.*

Proof of Lemma 4.4. Let S and S' be neighboring datasets, each with n individuals. Without loss of generality, assume S and S' differ in the first individual. We are ultimately interested in $\widetilde{M}(S)$ and $\widetilde{M}(S')$. They are generated as follows.

Let $I \subseteq [n]$ be any size m subset of the index set $[n] = \{1, 2, \ldots, n\}$. S_I and S'_I denote the m individuals indexed by I in corresponding datasets, both of which can be the input of M. When I is uniformly sampled from the $\binom{n}{m}$ subsets of $[n]$ of cardinality m, $M(S_I)$ is $\widetilde{M}(S)$.

Let ϕ be an arbitrary rejection rule. For a fixed $I \subseteq [n]$, when ϕ is used for the problem $M(S_I)$ vs $M(S'_I)$, the corresponding type I and type II errors are

$$\alpha_I := \mathbb{E}[\phi(M(S_I))] \quad \text{and} \quad \beta_I := 1 - \mathbb{E}[\phi(M(S'_I))] \tag{48}$$

respectively. The expectations are over the randomness of M.

When ϕ is used for $\widetilde{M}(S)$ vs $\widetilde{M}(S')$, the type I error α and type II error β satisfy

$$\alpha = \mathbb{E}[\phi(\widetilde{M}(S))] = \mathbb{E}_I \, \mathbb{E}[\phi(M(S_I))] = \mathbb{E}_I[\alpha_I]$$

and

$$\beta = 1 - \mathbb{E}[\phi(\widetilde{M}(S'))] = \mathbb{E}_I\big[1 - \mathbb{E}[\phi(M(S'_I))]\big] = \mathbb{E}_I[\beta_I].$$

Ultimately we are going to show that β has a lower bound in terms of α. This is possible because for each I, β_I has a lower bounded in terms of α_I, whose form depends on whether the "difference" individual 1 is sampled.

When $1 \notin I$, $S_I = S'_I$. By definition (48), $\alpha_I + \beta_I = 1$. When $1 \in I$, S_I and S'_I are neighbors in X^m, so (ε, δ) privacy of M becomes effective here. Let $k = -e^\varepsilon, b = 1 - \delta$. From Proposition 2.5 we know $\beta_I \geqslant k\alpha_I + b$.

So we should separate these cases $1 \in I$ and $1 \notin I$ and average them respectively. Define

$$A = \mathbb{E}[\alpha_I | I \ni 1], \quad B = \mathbb{E}[\beta_I | I \ni 1]$$

and

$$\bar{A} = \mathbb{E}[\alpha_I | I \not\ni 1], \quad \bar{B} = \mathbb{E}[\beta_I | I \not\ni 1].$$

Let $p = m/n$. Then $P[I \ni 1] = p, P[I \not\ni 1] = 1 - p$. Further averages of these quantities give us α and β:

$$\alpha = pA + (1-p)\bar{A}, \quad \beta = pB + (1-p)\bar{B}.$$

Linearity passes through expectations, so we have $\bar{A} + \bar{B} = 1$ and

$$B = \mathbb{E}[\beta_I | I \ni 1] \geqslant \mathbb{E}[k\alpha_I + b | I \ni 1] = k\mathbb{E}[\alpha_I | I \ni 1] + b = kA + b.$$

This inequality $B \geqslant kA + b$ comes from the difference between S and S'. The smart observation is, there are a lot more neighboring datasets we can exploit. For example S_I and $S'_{\tilde{I}}$ when $I = \{1, 2, \ldots, m\}, \tilde{I} = \{2, 3, \ldots, m+1\}$. They share individuals $2, 3, \ldots, m$ and differ in the one left. This yields another inequality $B \geqslant k\bar{A} + b$. We fill in the details Jon Ullman omits in his lecture note.

For $I \ni 1$, we can replace 1 with any of the $n-m$ indices not in I and obtain $n-m$ different subsets $I^{(1)}, \dots, I^{(n-m)}$. Note that none of these contains 1. In another word, we have the following correspondence:

$$\begin{array}{ccc} \ni 1 & & \not\ni 1 \\ I & \longleftrightarrow & I^{(1)} \\ & \vdots & \\ I & \longleftrightarrow & I^{(n-m)} \end{array}$$

Each $S_{I(j)}$ is a neighbor of S'_I, so by the privacy of M, we have $\beta_I \geqslant k\alpha_{I(j)} + b$. Sum these up for each $I \ni 1$ and the $n-m$ replacements of I, we have

$$(n-m) \cdot \sum_{I \ni 1} \beta_I \geqslant m \cdot \sum_{I \not\ni 1} k\alpha_I + b.$$

The right hand side factor m comes from a different counting: if $I \not\ni 1$ then each of the m items in I could have been 1 before the replacement. So each $I \not\ni 1$ appears m times in the summation.

Multiply by $\frac{n}{m(n-m)} \cdot \binom{n}{m}^{-1}$,

$$\frac{n}{m} \cdot \left(\sum_{I \ni 1} \beta_I \right) \cdot \binom{n}{m}^{-1} \geqslant b + k \cdot \frac{n}{n-m} \cdot \left(\sum_{I \not\ni 1} \alpha_I \right) \cdot \binom{n}{m}^{-1}.$$

By the simple Bayes' rule, this is $B \geqslant k\bar{A} + b$.

Remember we want a lower bound of β. The best possible lower bound we can ge t via these relations is the minimum of the following linear program:

$$\begin{aligned} \min_{A,B,\bar{A},\bar{B}} \quad & \beta \\ \text{s.t.} \quad & pB + (1-p)\bar{B} = \beta \\ & pA + (1-p)\bar{A} = \alpha \\ & \bar{A} + \bar{B} = 1 \\ & B \geqslant kA + b \\ & B \geqslant k\bar{A} + b \\ & A, B, \bar{A}, \bar{B} \in [0,1] \end{aligned}$$

Lemma F.4. *The minimum of the above linear program is no less than* $p(k\alpha + b) + (1-p)(1-\alpha)$.

Proof of Lemma F.4. We are going to remove the $A, B, \bar{A}, \bar{B} \in [0,1]$ constraint and find the exact minimum of the relaxation, which is a lower bound of the original minimum.

We have $\beta + \alpha = p(A+B) + (1-p)(\bar{A}+\bar{B}) = p(A+B) + (1-p)$. So equivalently we can try to solve

$$\begin{aligned} \min_{A,B,\bar{A}} \quad & A + B \\ \text{s.t.} \quad & pA + (1-p)\bar{A} = \alpha \\ & B \geqslant kA + b \\ & B \geqslant k\bar{A} + b \end{aligned}$$

When $A = \bar{A} = \alpha$, the lower bound they impose on B is $k\alpha + b$. Notice that $k = -\mathrm{e}^\varepsilon \leqslant -1$. The two consequences are:

1. $A > \alpha$ is worse than $A = \alpha$, because the convex combination equality requires $\bar{A} < \alpha$. B has to increase anyway. Both A and B increase, so the objective $A + B$ also increases.
2. If A is decreased from α by some amount, B has to increase by e^ε times that amount, which is not worth it.

So the minimum of the relaxed linear program is achieved at $A = \bar{A} = \alpha, B = k\alpha + b, \bar{B} = 1 - \alpha$, thereby inducing the claimed lower bound. □

So $\beta \geqslant p(k\alpha + b) + (1-p)(1-\alpha)$. Changing back to ε, δ, we have

$$\beta \geqslant p(-\mathrm{e}^\varepsilon \alpha + 1 - \delta) + (1-p)(1-\alpha) = -[p\mathrm{e}^\varepsilon + 1 - p]\alpha + 1 - p\delta = -\mathrm{e}^{\varepsilon'}\alpha + 1 - \delta'.$$

By Proposition 2.5, $\widetilde{M}$ is (ε', δ')-DP. □

G Omitted Proofs in Section 5

Our first goal is to prove

Theorem 5.2. *Suppose f is a symmetric trade-off function such that $f(0) = 1$ and $\int_0^1 (f'(x) + 1)^4 \,\mathrm{d}x < +\infty$. Furthermore, $p\sqrt{T} \to p_0$ as $T \to \infty$. Then we have the uniform limit*

$$C_p(f)^{\otimes T} \to G_{p_0\sqrt{2\chi^2_+(f)}}$$

where

$$\chi^2_+(f) = \int_0^1 \big(|f'(x)| - 1\big)_+^2 \,\mathrm{d}x.$$

First we point out that the functional χ^2_+ is computing a variant of χ^2-divergence. Recall that χ^2-divergence is an F-divergence with $F(t) = (t-1)^2$. We define χ^2_+-divergence to be the F-divergence with $F(t) = (t-1)_+^2 = \begin{cases} 0, & t \leqslant 1, \\ (t-1)^2, & t > 1. \end{cases}$ As in Appendix B, let $z_f = \inf\{x \in [0,1] : f(x) = 0\}$ be the first zero of f.

Proposition G.1. *For a pair of distributions P and Q such that $T(P,Q) = f$ is a symmetric trade-off function with $f(0) = 1$,*

$$\chi^2_+(P\|Q) = \chi^2_+(f).$$

Proof of Proposition G.1. By Proposition B.4, when $f = T(P,Q)$, $\chi^2_+(P\|Q)$ can be computed via the following expression:

$$\chi^2_+(P\|Q) = \int_0^{z_f} \big(|f'(x)|^{-1} - 1\big)_+^2 \cdot |f'(x)| \,\mathrm{d}x + F(0) \cdot (1 - f(0)) + \tau_F \cdot (1 - z_f)$$

where $F(0) = \lim_{p \to 0^+} F(t) = 0, \tau_F = \lim_{p \to +\infty} \frac{F(t)}{t} = +\infty$. Since we assume f is symmetric, $z_f = f(0) = 1$. This also implies that f^{-1} is the ordinary function inverse, i.e. $f(f(x)) = x$. Let

$y = f^{-1}(x) = f(x)$. Then $\mathrm{d}y = f'(x)\,\mathrm{d}x$. On the other hand, $x = f(y)$, $\mathrm{d}x = f'(y)\,\mathrm{d}y$. $x = 1$ corresponds to $y = 0$ and $x = 0$ corresponds to $y = 1$, so

$$\begin{aligned}
\chi^2_+(P\|Q) &= \int_0^1 \big(|f'(x)|^{-1} - 1\big)_+^2 \cdot |f'(x)|\,\mathrm{d}x \\
&= \int_0^1 \Big(\Big|\frac{\mathrm{d}y}{\mathrm{d}x}\Big|^{-1} - 1\Big)_+^2 \cdot |f'(x)|\,\mathrm{d}x \\
&= -\int_1^0 \big(|f'(y)| - 1\big)_+^2\,\mathrm{d}y \\
&= \int_0^1 \big(|f'(y)| - 1\big)_+^2\,\mathrm{d}y \\
&= \chi^2_+(f).
\end{aligned}$$

□

We need some more calculation tools to prove Theorem 5.2.

Lemma G.2. *Let $f \in \mathscr{F}^S$ with $f(0) = 1$ and x^* be its unique fixed point. Then*

$$\begin{aligned}
\chi^2_+(f) &= \int_0^{x^*} (f'(x) + 1)^2\,\mathrm{d}x \\
\mathrm{kl}(f) &= \int_0^{x^*} \big(|f'(x)| - 1\big)\log|f'(x)|\,\mathrm{d}x \\
\kappa_2(f) &= \int_0^{x^*} \big(|f'(x)| + 1\big)\big(\log|f'(x)|\big)^2\,\mathrm{d}x \\
\bar{\kappa}_3(f) &= \int_0^{x^*} \big|\log|f'(x)| + \mathrm{kl}(f)\big|^3 + |f'(x)| \cdot \big|\log|f'(x)| - \mathrm{kl}(f)\big|^3\,\mathrm{d}x \\
\kappa_3(f) &= \int_0^{x^*} \big(|f'(x)| + 1\big)\big(\log|f'(x)|\big)^3\,\mathrm{d}x.
\end{aligned}$$

Proof of Lemma G.2. First we observe that $f'(x) \leqslant -1$ for $x \leqslant x^*$ and $f'(x) \geqslant -1$ for $x \geqslant x^*$. This means the integrand involved in χ^2_+ is 0 in $[x^*, 1]$ and hence proves the first identity.

The rest of the proof is entirely based on a trick we used above. Let $y = f^{-1}(x) = f(x)$. Then $x = f(y)$, $\mathrm{d}x = f'(y)\,\mathrm{d}y$. Since x^* is the fixed point of f, $x = x^*$ corresponds to $y = x^*$. $x = 1$ corresponds to $y = 0$ and $x = 0$ corresponds to $y = 1$.

$$\begin{aligned}
-\int_{x^*}^1 \log|f'(x)|\,\mathrm{d}x &= \int_{x^*}^1 \log|f'(x)|^{-1}\,\mathrm{d}x \\
&= \int_{x^*}^0 \log\Big|\frac{\mathrm{d}x}{\mathrm{d}y}\Big| \cdot f'(y)\,\mathrm{d}y \\
&= \int_0^{x^*} \log|f'(y)| \cdot |f'(y)|\,\mathrm{d}y.
\end{aligned}$$

So

$$\begin{aligned}
\mathrm{kl}(f) &= -\int_0^1 \log|f'(x)|\,\mathrm{d}x \\
&= -\int_0^{x^*} \log|f'(x)|\,\mathrm{d}x - \int_{x^*}^1 \log|f'(x)|\,\mathrm{d}x \\
&= -\int_0^{x^*} \log|f'(x)|\,\mathrm{d}x + \int_0^{x^*} \log|f'(x)| \cdot |f'(x)|\,\mathrm{d}x \\
&= \int_0^{x^*} \big(|f'(x)| - 1\big) \log|f'(x)|\,\mathrm{d}x.
\end{aligned}$$

The rest of identities can be proved in exactly the same way. □

Lemma G.3. *Suppose $f \in \mathscr{F}^S$ and $f(0) = 1$. x^* is its unique fixed point. Let $g(x) = -f'(x) - 1 = |f'(x)| - 1$. Then*

$$\begin{aligned}
\mathrm{kl}(C_p(f)) &= p\int_0^{x^*} g(x) \log\big(1 + pg(x)\big)\,\mathrm{d}x \\
\kappa_2(C_p(f)) &= \int_0^{x^*} \big(2 + pg(x)\big)\big[\log\big(1 + pg(x)\big)\big]^2\,\mathrm{d}x \\
\kappa_3(C_p(f)) &= \int_0^{x^*} \big(2 + pg(x)\big)\big[\log\big(1 + pg(x)\big)\big]^3\,\mathrm{d}x.
\end{aligned}$$

Proof of Lemma G.3. We prove for kl and the rest are similar. Let x_p^* be the fixed point of $C_p(f)$. By Lemma G.2,

$$\mathrm{kl}(C_p(f)) = \int_0^{x_p^*} \big(|C_p(f)'(x)| - 1\big) \log|C_p(f)'(x)|\,\mathrm{d}x.$$

From the expression of $C_p(f)$ (13) we know $\log|C_p(f)'(x)| = 0$ in the interval $[x^*, x_p^*]$, and $C_p(f) = f_p = pf + (1-p)\mathrm{Id}$ in the interval $[0, x^*]$. So

$$\mathrm{kl}(C_p(f)) = \int_0^{x^*} \big(|f_p'(x)| - 1\big) \log|f_p'(x)|\,\mathrm{d}x.$$

In the interval $[0, x^*]$, $g(x) = |f'(x)| - 1 \geqslant 0$. $f_p'(x) = pf'(x) + (1-p)(-1) = p(f'(x)+1) - 1 = -pg(x) - 1$, so $|f_p'(x)| = pg(x) + 1$. When plugged in to the expression above, we have

$$\mathrm{kl}(C_p(f)) = p\int_0^{x^*} g(x) \log\big(1 + pg(x)\big)\,\mathrm{d}x.$$

□

Proof of Theorem 5.2. It suffices to compute the limits in Theorem 3.5, namely

$$T \cdot \mathrm{kl}(C_p(f)),\ T \cdot \kappa_2(C_p(f)) \text{ and } T \cdot \kappa_3(C_p(f)).$$

Since $T \sim p^{-2}$, we can consider $p^{-2}\mathrm{kl}(C_p(f))$ and so on.

As in Lemma G.3, let x^* be the unique fixed point of f and $g(x) = -f'(x) - 1 = |f'(x)| - 1$. Note that $g(x) \geqslant 0$ for $x \in [0, x^*]$. The assumption expressed in terms of g is simply

$$\int_0^1 g(x)^4 \,\mathrm{d}x < +\infty.$$

In particular, it implies $g(x)^k$ are integrable in $[0, x^*]$ for $k = 2, 3, 4$. In addition, $\chi^2_+(f) = \int_0^{x^*} g(x)^2 \,\mathrm{d}x$ by Lemma G.2.

For the functional kl, by Lemma G.3,

$$\begin{aligned} \lim_{p \to 0^+} \frac{1}{p^2} \,\mathrm{kl}(C_p(f)) &= \lim_{p \to 0^+} \int_0^{x^*} g(x) \cdot \frac{1}{p} \log\big(1 + pg(x)\big) \,\mathrm{d}x && (*) \\ &= \int_0^{x^*} g(x) \cdot \lim_{p \to 0^+} \frac{1}{p} \log\big(1 + pg(x)\big) \,\mathrm{d}x \\ &= \int_0^{x^*} g(x)^2 \,\mathrm{d}x = \chi^2_+(f) \end{aligned}$$

Changing the order of the limit and the integral in $(*)$ is approved by dominated converegence theorem. To see this, notice that $\log(1 + x) \leqslant x$. The integrand in $(*)$ satisfies

$$0 \leqslant g(x) \cdot \frac{1}{p} \log\big(1 + pg(x)\big) \leqslant g(x)^2.$$

We already argued that $g(x)^2$ is integrable, so it works as a dominating function and the limit is justified. When $p\sqrt{T} \to p_0$, we have

$$T \cdot \mathrm{kl}(C_p(f)) \to p_0^2 \cdot \chi^2_+(f).$$

So the constant K in Theorem 3.5 is $p_0^2 \cdot \chi^2_+(f)$.

For the functional κ_2 we have

$$\frac{1}{p^2} \kappa_2(C_p(f)) = \int_0^{x^*} \big(2 + pg(x)\big) \Big[\frac{1}{p} \log\big(1 + pg(x)\big)\Big]^2 \,\mathrm{d}x.$$

By a similar dominating function argument,

$$\lim_{p \to 0^+} \frac{1}{p^2} \,\kappa_2(C_p(f)) = 2 \int_0^{x^*} g(x)^2 \,\mathrm{d}x = 2\chi^2_+(f).$$

Adding in the limit $p\sqrt{T} \to p_0$, we know s^2 in Theorem 3.5 is $2p_0^2 \cdot \chi^2_+(f)$. Once again, we have $s^2 = 2K$.

The same argument involving $g(x)^4$ applies to the functional κ_3 and yields

$$\lim_{p \to 0^+} \frac{1}{p^3} \,\kappa_3(C_p(f)) = 2 \int_0^{x^*} g(x)^3 \,\mathrm{d}x.$$

Note the different power in p in the denominator. It means $\kappa_3(C_p(f)) = o(p^2)$ and hence $T \cdot \kappa_3(C_p(f)) \to 0$ when $p\sqrt{T} \to p_0$.

Hence all the limits in Theorem 3.5 check and we have a G_μ limit where

$$\mu = 2K/s = s = \sqrt{2p_0^2 \cdot \chi^2_+(f)} = p_0 \cdot \sqrt{2\chi^2_+(f)}.$$

This completes the proof. □

Lemma 5.3.

$$\chi_+^2(G_\mu) = \mathrm{e}^{\mu^2} \cdot \Phi(3\mu/2) + 3\Phi(-\mu/2) - 2.$$

Proof of Lemma 5.3. We use Proposition G.1 as the tool. Obviously $P = \mathcal{N}(\mu, 1)$ and $Q = \mathcal{N}(0,1)$ satisfy the conditions there. So it suffices to compute $\chi_+^2(\mathcal{N}(\mu,1)\|\mathcal{N}(0,1))$. Recall that χ_+^2 is the F-divergence with $F(t) = (t-1)_+^2$, so $\chi_+^2(P\|Q) = \mathbb{E}_Q\big[(\frac{P}{Q}-1)_+^2\big]$. Let φ and Φ be the density function and cdf of the standard normal. We have

$$\begin{aligned}
\chi_+^2(G_\mu) &= \chi_+^2(\mathcal{N}(\mu,1)\|\mathcal{N}(0,1))\\
&= \mathbb{E}_{x\sim\mathcal{N}(0,1)}\Big[\Big(\frac{\varphi(x-\mu)}{\varphi(x)} - 1\Big)_+^2\Big]\\
&= \int_{\mu/2}^{+\infty} \Big(\frac{\varphi(x-\mu)}{\varphi(x)} - 1\Big)^2 \cdot \varphi(x)\,\mathrm{d}x\\
&= \int_{\mu/2}^{+\infty} \Big(\frac{\varphi(x-\mu)}{\varphi(x)}\Big)^2 \cdot \varphi(x)\,\mathrm{d}x - 2\int_{\mu/2}^{+\infty} \varphi(x-\mu)\,\mathrm{d}x + \int_{\mu/2}^{+\infty} \varphi(x)\,\mathrm{d}x\\
&= \underbrace{\int_{\mu/2}^{+\infty} e^{2\mu x-\mu^2} \cdot \varphi(x)\,\mathrm{d}x}_{I} -2(1-\Phi(\mu/2)) + \Phi(-\mu/2)\\
&= I + 3\Phi(-\mu/2) - 2.
\end{aligned}$$

For the integral I,

$$\begin{aligned}
I &= \int_{\mu/2}^{+\infty} e^{2\mu x-\mu^2} \cdot \varphi(x)\,\mathrm{d}x\\
&= \int_{\mu/2}^{+\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{2\mu x-\mu^2-x^2/2}\,\mathrm{d}x\\
&= \int_{\mu/2}^{+\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-(x-2\mu)^2/2} \cdot e^{\mu^2}\,\mathrm{d}x\\
&= e^{\mu^2} \cdot P[\mathcal{N}(2\mu,1) \geqslant \mu/2]\\
&= e^{\mu^2} \cdot \Phi(3\mu/2)
\end{aligned}$$

This completes the proof. □

Theorem 5.4. *If* $\frac{m}{n} \cdot \sqrt{T} \to c$, *then* NoisySGD *is* μ*-GDP with*

$$\mu = \sqrt{2}c \cdot \sqrt{\mathrm{e}^{\sigma^{-2}} \cdot \Phi(1.5\sigma^{-1}) + 3\Phi(-0.5\sigma^{-1}) - 2}.$$

Proof of Theorem 5.4. Combining Theorems 5.1 and 5.2 and Lemma 5.3, it suffices to check $\int_0^1 (f'(x)+1)^4\,\mathrm{d}x < +\infty$ when $f(x) = G_a(x) = \Phi(\Phi^{-1}(1-x)-a)$. Let $y = \Phi^{-1}(1-x)$. We have $\varphi(y)\,\mathrm{d}y = -\,\mathrm{d}x$. Hence

$$G_a'(x) = \varphi(y-a) \cdot \frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{\varphi(y-a)}{\varphi(y)} = -\mathrm{e}^{ay-\frac{a^2}{2}}.$$

The integral is

$$\int_0^1 (G_a'(x)+1)^4\,\mathrm{d}x = \int_{-\infty}^{+\infty} (-\mathrm{e}^{ay-\frac{a^2}{2}}+1)^4\varphi(y)\,\mathrm{d}y,$$

which is just a linear combination of moment generating functions of the standard normal and hence finite. □

Lemma 5.5. *Let* $Z(x) = \log(p\cdot e^{\mu x-\mu^2/2}+1-p)$*. Then*

$$\begin{aligned}
\mathrm{kl}\big(C_p(G_\mu)\big) &= p\int_{\mu/2}^{+\infty} Z(x)\cdot\big(\varphi(x-\mu)-\varphi(x)\big)\,\mathrm{d}x\\
\kappa_2\big(C_p(G_\mu)\big) &= \int_{\mu/2}^{+\infty} Z^2(x)\cdot\big(p\varphi(x-\mu)+(2-p)\varphi(x)\big)\,\mathrm{d}x\\
\bar\kappa_3\big(C_p(G_\mu)\big) &= \int_{\mu/2}^{+\infty} \big|Z(x)-\mathrm{kl}\big(C_p(G_\mu)\big)\big|^3\cdot(p\varphi(x-\mu)+(1-p)\varphi(x))\,\mathrm{d}x\\
&\quad+\int_{\mu/2}^{+\infty} \big|Z(x)+\mathrm{kl}\big(C_p(G_\mu)\big)\big|^3\cdot\varphi(x)\,\mathrm{d}x.
\end{aligned}$$

Proof of Lemma 5.5. We will use Lemma G.3. It's easy to show the fixed point of G_μ is $x^*=\Phi(-\mu/2)$. So

$$\mathrm{kl}\big(C_p(G_\mu)\big) = p\int_0^{\Phi(-\mu/2)} \big(-G_\mu'(x)-1\big)\log\big(1+p(-G_\mu'(x)-1)\big)\,\mathrm{d}x$$

Using the same change of variable $y=\Phi^{-1}(1-x)=-\Phi^{-1}(x)$, we have

$$\begin{aligned}
\mathrm{kl}\big(C_p(G_\mu)\big) &= p\int_{\mu/2}^{+\infty}\Big(\frac{\varphi(y-\mu)}{\varphi(y)}-1\Big)\log\Big(1+p\Big(\frac{\varphi(y-\mu)}{\varphi(y)}-1\Big)\Big)\varphi(y)\,\mathrm{d}y\\
&= p\int_{\mu/2}^{+\infty} Z(y)\cdot\big(\varphi(y-\mu)-\varphi(y)\big)\,\mathrm{d}y.
\end{aligned}$$

The rest can be proved similarly. □

Theorem 5.6. *Let* $p=m/n, \mu=\sigma^{-1}$ *and*

$$\begin{aligned}
z &= \frac{2\sqrt{T}\cdot\mathrm{kl}\big(C_p(G_\mu)\big)}{\sqrt{\kappa_2\big(C_p(G_\mu)\big)-\mathrm{kl}^2\big(C_p(G_\mu)\big)}},\\
\gamma &= \frac{0.56}{\sqrt{T}}\cdot\frac{\bar\kappa_3\big(C_p(G_\mu)\big)}{\big(\kappa_2\big(C_p(G_\mu)\big)-\mathrm{kl}^2\big(C_p(G_\mu)\big)\big)^{\frac32}}.
\end{aligned}$$

NoisySGD *is f-DP with*

$$f(\alpha) = \max\{G_z(\alpha+\gamma)-\gamma, 0\}.$$

Proof of Theorem 5.6. Follows from plugging in the expressions above into Theorem 3.4. □