f-Differential Privacy Filters: Validity and Approximate Solutions

Long Tran[1] **Antti Koskela**[2] **Ossi Räisä**[3*] **Antti Honkela**[1]

[1]Department of Computer Science, University of Helsinki, Finland
[2]Nokia Bell Labs, Espoo, Finland
[3]CISPA Helmholtz Center for Information Security, Saarbrücken, Germany

{thanh-long.tran,antti.honkela}@helsinki.fi,
antti.h.koskela@nokia-bell-labs.com, ossi.raisa@cispa.de

Abstract

Accounting for privacy loss under fully adaptive composition—where mechanism choice and privacy parameters may depend on the history of prior outputs—is a central challenge in differential privacy (DP). Here, privacy filters are stopping rules ensuring a prescribed global budget is not exceeded. A leading candidate for optimal filter design is f-DP, which characterizes the full extent of adversarial hypothesis testing and recovers (ε, δ)-DP through piece-wise linear trade-off functions, while enabling tight (ε, δ)-DP accounting in standard compositions via tensor products. Yet whether such filters can be correctly defined under f-DP remains unclear. We show that the natural f-DP filter—tracking path-wise accumulating tensor products and stopping when the prescribed curve is crossed—is fundamentally invalid, precluding the direct use of standard efficient numerical Fast-Fourier-Transform accounting in the fully adaptive setting. We characterize this failure, establishing necessary and sufficient conditions for the natural filter's validity. Furthermore, we prove a fully adaptive central limit theorem for f-DP, establishing Gaussian convergence of cumulative privacy losses under full adaptivity. As a demonstration, we construct a closed-form approximate GDP filter for subsampled Gaussian mechanisms that provably outperforms RDP-based accounting in asymptotic regimes ($q \ll 1$ and $q \approx 1$) without tracking the full trade-off function, demonstrating that the slack in RDP is not intrinsic to adaptive composition—though CLT-based approximations are known to be optimistic at realistic subsampling rates, a limitation that remains an open challenge.

1 Introduction

Differential privacy (DP) rigorously quantifies privacy loss from computations on sensitive data (Dwork et al., 2006; Dwork and Roth, 2014). A key property is composition: complex algorithms can be analyzed by composing privacy guarantees of simpler mechanisms. Common composition results require knowing privacy parameters beforehand. However, fully adaptive compositions—where mechanisms and their privacy parameters depend on prior outputs—are essential for iterative algorithms like differentially private stochastic gradient descent and adaptive data analysis. Accounting for privacy loss in this setting is a central challenge.

Privacy filters and odometers formalize fully adaptive composition for (ε, δ)-DP (Rogers et al., 2016), and Rényi DP (RDP) admits simple, tight filters under full adaptivity (Feldman and Zrnic, 2021), making RDP-based accounting the current state of the art. In contrast, much less is understood about

*Work primarily done while at the University of Helsinki.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

fully adaptive composition for f-DP (Dong et al., 2022). Only in the special case of Gaussian DP (GDP) can accurate fully adaptive bounds be obtained (Koskela et al., 2023). Applying the f-DP tensor product gives the tightest composition results for non-adaptive compositions—and since any (ε, δ)-DP guarantee admits an exact equivalent piece-wise linear trade-off function (Dong et al., 2022), this tightness extends to the (ε, δ)-DP framework as well. This raises a natural question:

Does f-DP admit valid filters under fully adaptive composition, in analogy to RDP and GDP?

A natural candidate filter tracks the accumulating tensor product of trade-off functions and halts once it crosses the prescribed f-DP budget, mirroring standard f-DP composition and adaptive RDP/GDP accounting schemes. Our first main result shows that this approach is fundamentally invalid.

Failure of natural f-DP filters. We show that composing trade-off functions and stopping at a threshold does not, in general, guarantee that the resulting stopped interaction satisfies the same f-DP bound. This failure occurs even when each individual mechanism is validly f-DP and even when the composed trade-off function never exceeds the threshold along the realized history of outcomes. We give explicit negative examples and a general construction that explains when and why such failures arise. Moreover, we show that this phenomenon occurs for fully adaptive compositions of subsampled Gaussian mechanisms, the main workhorse of private machine learning. As a consequence, the natural filter rules out the most direct route to efficient numerical Fast-Fourier-Transform (FFT) accounting (Koskela et al., 2020; Gopi et al., 2021; Doroshenko et al., 2022) in the fully adaptive setting, blocking the standard approach for accurate (ε, δ) bounds.

When f-DP filters are valid. While the natural f-DP filter fails in general, we show that it becomes valid under specific structural conditions on the trade-off functions. We provide necessary and sufficient conditions under which stopping based on composed trade-off curves yields a correct f-DP guarantee. These conditions clarify why divergence-based notions such as RDP admit simple filters under full adaptivity, while trade-off-function-based notions do not.

Fully adaptive f-DP central limit theorem. We prove a central limit theorem (CLT) for adaptive privacy loss processes, extending the non-fully adaptive CLT of Dong et al. (2022) to the fully adaptive setting via a modern variant of the martingale Berry–Esseen bound (Fan, 2019). This CLT suggests a general methodology for constructing approximate GDP filters: any mechanism family admitting closed-form PLRV moment approximations yields a closed-form approximate filter, with the CLT error quantifying approximation quality. As a demonstration, we instantiate this methodology for subsampled Gaussian mechanisms. At realistic subsampling rates, the resulting filter underestimates privacy loss—a known limitation of CLT-based Gaussian approximations even in the non-adaptive setting (Zheng et al., 2020; Gopi et al., 2021; Alghamdi et al., 2023). In the extreme low- and high-subsampling regimes, however, it yields asymptotically valid guarantees strictly tighter than the fully adaptive RDP accounting of Feldman and Zrnic (2021), without ever tracking the full trade-off function, demonstrating that the slack in RDP stems from the lossy reduction to Rényi divergences rather than any fundamental barrier in adaptive composition.

Contributions. The main contributions of this paper are as follows:

1. We show that the natural f-DP filter — which composes individual trade-off functions and stops at a prescribed threshold — is not valid under fully adaptive composition, precluding the direct extension of FFT numerical accounting to such setting. We provide explicit counterexamples and a general construction of this failure. We show that this failure occurs even when we restrict to subsampled Gaussian mechanisms. (Corollary 3.3, Figure 1)
2. We establish necessary and sufficient conditions under which the natural f-DP filter becomes valid, characterizing when trade-off-function-based accounting is possible under full adaptivity. (Theorems 3.5 and 3.6)
3. We prove a fully adaptive f-DP central limit theorem, establishing that cumulative privacy losses converge to Gaussian distributions, yielding GDP bounds under full adaptivity. This extends the non-adaptive CLT of Dong et al. (2022) via martingale Berry–Esseen techniques, and suggests a general methodology for constructing approximate GDP filters for any mechanism family admitting closed-form PLRV moment approximations. (Theorem 4.1)
4. As a demonstration of our CLT, we instantiate this methodology for subsampled Gaussian mechanisms, constructing a closed-form approximate GDP filter that yields strictly tighter guarantees than fully adaptive RDP accounting in asymptotic regimes $q \ll 1$ and $q \approx 1$, without tracking the full trade-off function. (Algorithm 2)

Related work. Privacy filters were introduced by Rogers et al. (2016), who formalized fully adaptive composition for (ε, δ)-DP and established the foundations of filter-based accounting. Feldman and Zrnic (2021) showed that Rényi DP admits simple and tight privacy filters under full adaptivity, with the added flexibility of individual privacy accounting. Whitehouse et al. (2023) subsequently constructed approximate Rényi DP filters achieving essentially no loss in tightness relative to advanced standard composition. In the GDP setting, Koskela et al. (2023) developed an accurate fully adaptive filter exploiting the Gaussian structure. Haney et al. (2025) studied the transfer of filters from the noninteractive to the interactive setting under a general framework that includes f-DP, but did not analyze specific filter constructions. Independently and concurrently, Regehr et al. (2026) also prove that the natural f-DP filter is invalid and provide complementary validity conditions. Beyond this shared result, we additionally prove a fully adaptive f-DP central limit theorem that suggests a general methodology for constructing approximate GDP filters via mechanism-specific PLRV moment approximations, of which the subsampled Gaussian filter is one instantiation.

2 Background

Differential Privacy. A mechanism $\mathcal{M}$ is a function that maps a dataset input $S = (X_1, \ldots, X_n)$ to a distribution $\mathcal{M}(S)$ on a space $\mathcal{O}$ of possible outputs. We call (S, S^-) remove neighbours if S^- is obtained by removing one datapoint from S.

Definition 2.1 (Dwork et al., 2006)**.** A mechanism $\mathcal{M}$ is (ε, δ)-DP if for any remove pair S and S^-,

$$P[\mathcal{M}(S) \in A] \leq e^{\varepsilon} P\big[\mathcal{M}\big(S^-\big) \in A\big] + \delta \quad \text{and} \quad P\big[\mathcal{M}\big(S^-\big) \in A\big] \leq e^{\varepsilon} P[\mathcal{M}(S) \in A] + \delta$$

for all set A in the output space $\mathcal{O}$. When $\delta = 0$, we shorten $(\varepsilon, 0)$-DP into ε-DP.

Definition 2.2 (Dong et al., 2022)**.** Given distributions P, Q on a common space $\mathcal{O}$, the trade-off function is defined as

$$T[P, Q](\alpha) = \inf_{\phi : \mathcal{O} \to [0,1]} \{1 - \mathbb{E}_Q[\phi] : \mathbb{E}_P[\phi] \leq \alpha\}.$$

When treating the pair (P, Q) as a binary testing problem of $H_0 : y \sim P$ vs. $H_1 : y \sim Q$, the function ϕ is a randomized rejection rule that maps $y \in \mathcal{O}$ to the probability $\phi(y)$ of accepting H_1. Then, $\mathbb{E}_P[\phi]$ is the false positive rate (FPR), while $1 - \mathbb{E}_Q[\phi]$ is the false negative rate (FNR), and hence $T[P, Q]$ represents the optimal FNR at any FPR $\alpha \in [0, 1]$.

Theorem 2.3 (Dong et al., 2022)**.** *A function $f : [0, 1] \to [0, 1]$ is the trade-off function of some distribution pair if and only if $f(\alpha) \leq 1 - \alpha$, f is decreasing, convex, continuous.*

An everywhere smaller trade-off function means a uniformly easier testing problem, and hence weaker privacy protection.

Definition 2.4. Two trade-off functions f_1, f_2 are Blackwell ordered if $f_1 \geq f_2$ or $f_1 \leq f_2$.

Definition 2.5 (Dong et al., 2022)**.** A mechanism $\mathcal{M}$ is f-DP where f is a trade-off function, if

$$T\big[\mathcal{M}(S), \mathcal{M}\big(S^-\big)\big] \geq f \quad \text{and} \quad T\big[\mathcal{M}\big(S^-\big), \mathcal{M}(S)\big] \geq f$$

point-wise for any remove pair S, S^-. In particular, $\mathcal{M}$ is μ-GDP if it is G_μ-DP:

$$G_\mu = T[\mathcal{N}(0, 1), \mathcal{N}(\mu, 1)] = T[\mathcal{N}(0, 1/\mu^2), \mathcal{N}(1, 1/\mu^2)].$$

Following Zhu et al. (2022), we also consider f-DP under remove adjacency, requiring only the first inequality above. Composing directional guarantees before symmetrizing can yield tighter bounds than symmetrizing each step; see Appendix A.2 and Lebeda et al. (2025).

The subsampled Gaussian trade-off function of a remove pair is $SG^-_{q,\mu} = \big(SG^+_{q,\mu}\big)^{-1}$ where

$$SG^+_{q,\mu}(\alpha) = T[\mathcal{N}(0, 1), (1 - q)\mathcal{N}(0, 1) + q\mathcal{N}(\mu, 1)] = (1 - q)(1 - \alpha) + qG_\mu(\alpha).$$

Proposition 2.6 (Dong et al., 2022)**.** *A mechanism $\mathcal{M}$ is (ε, δ)-DP if and only if it is $f_{\varepsilon,\delta}$-DP for*

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

By deploying (ε, δ)-DP as $f_{\varepsilon,\delta}$-DP, all results for f-DP immediately apply to (ε, δ)-DP as well. In particular, the f-DP standard composition results is expressed via tensor products

$$T[P_1, Q_1] \otimes T[P_2, Q_2] = T[P_1 \times P_2, Q_1 \times Q_2],$$

from which we can derive the tight (ε, δ)-DP of standard compositions.

Theorem 2.7 (Dong et al., 2022)**.** *If $\mathcal{M}_1$ is tightly f_1-DP, and the adaptively chosen $\mathcal{M}_2$ is f_2-DP where f_2 is fixed and tight, then $(\mathcal{M}_1, \mathcal{M}_2)$ is $f_1 \otimes f_2$-DP. In particular, if $\mathcal{M}_1$ is μ_1-GDP and $\mathcal{M}_2$ is μ_2-GDP, then $(\mathcal{M}_1, \mathcal{M}_2)$ is $\sqrt{\mu_1^2 + \mu_2^2}$-GDP. Moreover, $f_1 \otimes f_2$ can be converted into the tight (ε, δ)-DP guarantee at each $\varepsilon > 0$,*

$$\min\{\delta \geq 0 : (\mathcal{M}_1, \mathcal{M}_2) \text{ is } (\varepsilon, \delta)\text{-DP}\} = \min\{\delta \geq 0 : f_1 \otimes f_2 \geq f_{\varepsilon,\delta}\}.$$

Definition 2.8. For distributions (P, Q) on a common space $\mathcal{O}$, denote $L = \ln \frac{P}{Q}$. The PLRVs are defined as $L(y)$, $y \sim P$ and $L(y)$, $y \sim Q$. The PLRV density is called the privacy loss density (PLD).

The cdf of each PLRV characterizes the FPR and FNR of likelihood ratio tests, directly representing the trade-off function via the Neyman-Pearson lemma. Under compositions, PLRVs sum as $\ln \frac{\mathcal{M}_{1:T}(S)}{\mathcal{M}_{1:T}(S^-)} = \sum \ln \frac{\mathcal{M}_t(S)}{\mathcal{M}_t(S^-)}$, indicating CLT-based convergence to Gaussian distributions.

Extended DP Notions. In adaptive composition, the next distributions $P_2(y_1)$ and $Q_2(y_1)$ are chosen based on the realization $y_1 \sim P_1$ and $y_1 \sim Q_1$, respectively. We denote then the adaptive product as

$$T[P_1, Q_1] \otimes_{\text{adapt}} T[P_2(y_1), Q_2(y_1)] = T[P_1 \times P_2(y_1), Q_1 \times Q_2(y_1)]. \tag{1}$$

Let $\mathcal{F}$ be a family of trade-off functions. The greatest lower bound of $\mathcal{F}$ is not the point-wise infimum, but rather the lower convex envelope of that infimum, denoted $f^{\downarrow} = \text{ce}\inf_{f \in \mathcal{F}} f$. For a single trade-off function f, we define its symmetrization as $\text{sym}(f) = \text{ce}\min(f, f^{-1})$.

Not all trade-off functions are Blackwell ordered, so the bound $f^{\downarrow}$ of a family $\mathcal{F}$ can deviate significantly from every $f \in \mathcal{F}$. Kaissis et al. (2024) proposed Δ-divergence to measure the distance between two arbitrary trade-off functions. We then define convergence of trade-off functions via Δ-divergence, and prove its equivalence to other convergence notions (see Theorem C.8).

Definition 2.9. The lower closure $\bar{\mathcal{F}}$ of a family $\mathcal{F}$ of trade-off functions is the set of all limits of any everywhere decreasing sequence in $\mathcal{F}$. We furthermore define $\mathcal{F}^{-\otimes\mathcal{F}} = \left\{f \otimes g : f \in \mathcal{F}, g \in \bar{\mathcal{F}}\right\}$.

In addition, we use Δ-divergence to extend GDP to privacy guarantees that converge to it.

Definition 2.10. A mechanism $\mathcal{M}$ is Δ-approximately μ-GDP if its f-DP guarantee satisfies

$$G_\mu(\alpha + \Delta) - \Delta \leq f(\alpha) \leq G_\mu(\alpha - \Delta) + \Delta \text{ for all } \alpha.$$

Fully Adaptive Composition. Fully adaptive composition extends standard composition by allowing mechanism choice and privacy parameters to depend arbitrarily on the history of prior outputs. This is crucial in practice: in DP-SGD, for instance, committing to a fixed schedule of q_t and σ_t in advance is often suboptimal for utility, yet adapting them based on observed model behaviour or remaining privacy budget requires exactly this framework.

Under full adaptivity, privacy parameters become random variables dependent on the stochastic outcomes of previous mechanisms, requiring path-dependent privacy accounting methods where the composition length itself may be random. To manage this complexity, Rogers et al. (2016) introduced privacy filters and privacy odometers as two complementary approaches. Privacy filters are adaptive stopping rules that halt mechanism releases to ensure pre-specified privacy guarantees, while privacy odometers track cumulative privacy loss without requiring a fixed budget in advance. Since odometers must ensure valid privacy bounds at any analyst-chosen stopping time, they present significant theoretical and practical difficulties. In this

Algorithm 1 Privacy filter in fully adaptive composition

input Adaptively chosen mechanisms $\mathcal{M}_t^{(y_{1:t-1})}(\cdot)$ for $t = 1, \ldots, T$, privacy budget ω_B, dataset S.
for $t = 1$ to T **do**
  compute privacy level ω_t of $\mathcal{M}_t^{(y_{1:t-1})}$
  if $F(\omega_B, \omega_1, \ldots, \omega_t) = \text{HALT}$ **then**
    return $(y_1, \ldots, y_{t-1})$
  compute $y_t \leftarrow \mathcal{M}_t^{(y_{1:t-1})}(S)$
return $(y_1, \ldots, y_T)$

work, we focus on privacy filters (Algorithm 1), which maintain fixed privacy guarantees at the cost of random stopping times. Here, the privacy level is denoted ω_t due to the freedom to choose DP definitions. Central to the composition, the privacy filter F compares the privacy parameters $\omega_{1:t}$ realized so far to the budget ω_B to determine whether the upcoming mechanism release is allowed.

Definition 2.11. A privacy filter F is valid for a family Ω of per-step guarantees ($\omega_t \in \Omega\ \forall t$) if, for every sequence of mechanisms $\mathcal{M}_1, \ldots, \mathcal{M}_T$, each satisfying its selected guarantee $\omega_t \in \Omega$, and every budget ω_B, Algorithm 1 produces a composition satisfying the prescribed budget ω_B.

Among several valid filters developed in previous work (e.g. Feldman and Zrnic, 2021; Whitehouse et al., 2023), the GDP filter by Koskela et al. (2023) directly adapts the standard GDP composition result to the filtering framework, and provides a foundation for our more general f-DP results.

Theorem 2.12 (Koskela et al., 2023)**.** *With GDP guarantees $\omega_t = \mu_t$, the following filter is valid (cf. Theorem 2.7):*

$$F(\mu_B, \mu_1, \ldots, \mu_t) = \begin{cases} \text{CONT} & \text{if } \sum_{i=1}^{t} \mu_i^2 \leq \mu_B^2, \\ \text{HALT} & \text{otherwise.} \end{cases}$$

3 Natural f-DP Filter

We first give the formal definition of the natural f-DP filter construction.

Definition 3.1. The natural f-DP filter is defined as

$$F(f_B, f_1, \ldots, f_t) = \begin{cases} \text{CONT} & \text{if } f_1 \otimes \ldots \otimes f_t \geq f_B, \\ \text{HALT} & \text{otherwise.} \end{cases}$$

This filter keeps track of the selected per-step guarantees f_t and compares their tensor product with the budget f_B. Throughout this section, f-DP guarantees need not be symmetric. For Poisson-subsampled DP-SGD, the filter can be used to track f-DP guarantees under remove adjacency: if it successfully guarantees the budget f_B, then symmetrizing f_B yields the add/remove guarantee.

Such a procedure is motivated by the standard f-DP composition theorem (Theorem 2.7), and it is a direct generalization of the valid GDP filter (Theorem 2.12). It is therefore tempting to expect validity under full adaptivity as well—which would also allow numerical PLD composition methods (Koskela et al., 2020; Gopi et al., 2021; Doroshenko et al., 2022) to extend to this setting. We show that this validity claim is false in general.

Calibration Point and Filter Invalidity. To understand this failure, we reduce to a two-step adaptive composition, since the general case can be viewed through this setting. The first mechanism has a trade-off of f_1, and the second adaptively chosen mechanism has a trade-off of $f_2^{(y_1)}$ which can be selected arbitrarily from a family $\mathcal{F}_2$. The true composition is described by the adaptive product $f_1 \otimes_{\text{adapt}} f_2^{(y_1)}$, whereas the natural f-DP filter only tracks the standard product $f_1 \otimes f_2^{(\hat{y}_1)}$ when $y_1 = \hat{y}_1$ is realized.

Theorem 3.2 (Adaptive Tightness of Theorem 2.7)**.** *Let f_1 be a fixed trade-off function, and let $\mathcal{F}_2$ be a family of trade-off functions. Denote the greatest lower bound $f_2^{\downarrow} = \text{ce}\inf_{f_2 \in \mathcal{F}_2} f_2$. Then, for any selection rule $y_1 \mapsto f_2^{(y_1)} \in \mathcal{F}_2$, the adaptive product of f_1 and $f_2^{(y_1)}$ is lower bounded by $f_1 \otimes f_2^{\downarrow}$. Moreover, this bound is tight:*

$$f_1 \otimes f_2^{\downarrow} = \text{ce}\inf_{\text{selection rule}} \left(f_1 \otimes_{\text{adapt}} f_2^{(y_1)} \right).$$

More precisely, at each FPR $\alpha \in (0, 1)$, there exists a selection rule $y_1 \mapsto f_2^{(y_1)}$ such that the adaptive product $f_1 \otimes_{\text{adapt}} f_2^{(y_1)}$ is arbitrarily close to the lower bound $f_1 \otimes f_2^{\downarrow}$ at α (but not necessarily at other FPRs).

Suppose there exist $f_{21}, f_{22} \in \mathcal{F}_2$ that intersect at an FPR α_0 (a calibration point, see Kaissis et al., 2024). Their lower envelope $f_2^{\downarrow} = \text{ce}\min(f_{21}, f_{22})$ then falls below both f_{21} and f_{22} on some interval containing α_0. Consequently, $f_1 \otimes f_2^{\downarrow}$ can be strictly smaller than both $f_1 \otimes f_{21}$ and $f_1 \otimes f_{22}$: the former captures all extent of the true effect achievable under adaptive selection over $\mathcal{F}_2$, whereas the latter are the per-branch bounds monitored by the natural f-DP filter.

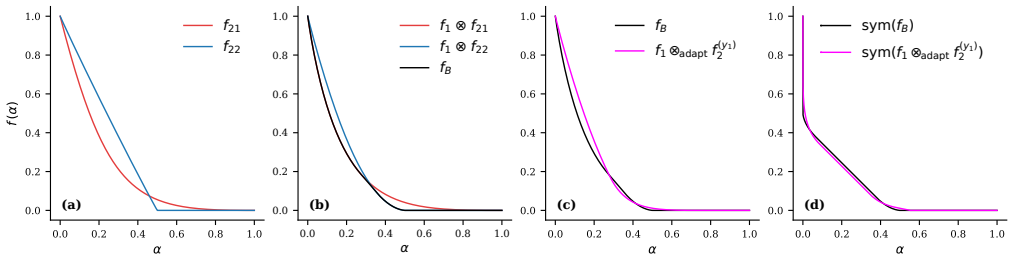


Figure 1: Counterexample showing that the natural f-DP filter can fail in fully adaptive composition of subsampled Gaussian mechanisms (Corollary 3.3). We fix $q = 0.5$, set $f_1 = SG^-_{q,1.3}$, $f_{21} = SG^-_{q,2} \otimes SG^-_{q,2}$, $f_{22} = SG^-_{q,0.1} \otimes SG^-_{q,10}$, and define the budget $f_B = \mathrm{ce}\min(f_1 \otimes f_{21},\ f_1 \otimes f_{22})$. **(a)** The two candidate future trade-off functions cross (not Blackwell-ordered). **(b)** The budget f_B is formed as the convex lower envelope of the two fixed-branch products. **(c)** Output-dependent selection between the continuations produces an adaptive composition below f_B, violating validity. **(d)** The violation persists after symmetrization into an appropriate f-DP guarantee.

Corollary 3.3. *The f-DP filter is not valid for subsampled Gaussian mechanisms (and is thus not valid in general): There exist f_1, f_{21}, f_{22} that are (tensor products of) subsampled Gaussian trade-off functions, and a budget f_B, such that $f_1 \otimes f_{21} \geq f_B$ and $f_1 \otimes f_{22} \geq f_B$, but $(f_1 \otimes \mathrm{ce}\min(f_{21}, f_{22}))(\alpha) < f_B(\alpha)$ at some FPR $\alpha \in (0, 1)$. See Figure 1.*

Such crossings are generic in common mechanism families, including Laplace mechanisms. The absence of crossings for (tensor products of) Gaussian trade-off functions is a notable special case. Additional details and proofs for the material up to this point are deferred to Appendix B.

Blackwell Chains and Filter Characterization. By contrast, this issue does not occur for RDP or GDP filters, since their guarantees are scalar-valued and therefore always comparable. This motivates imposing structural assumptions under which the natural f-DP filter becomes valid.

Definition 3.4 (Blackwell Chain)**.** A family $\mathcal{F}$ of trade-off functions is a Blackwell chain if $\mathcal{F}$ is totally ordered with respect to the Blackwell order: $f_1 \geq f_2$ or $f_1 \leq f_2$ for any $f_1, f_2 \in \mathcal{F}$.

For any Blackwell chain $\mathcal{F}$, the trade-off functions $f \in \mathcal{F}$ can approach the greatest lower bound $f^{\downarrow} = \mathrm{ce}\inf_{f \in \mathcal{F}} f$ arbitrarily close FPR-uniformly. Gaussian trade-off functions form such a chain, with G_μ decreasing in μ and their lower bound approximated uniformly. Thus, the Blackwell chain assumption removes the gap between adaptive products and standard products (as in Corollary 3.3) and ensures validity of the f-DP filter.

Theorem 3.5. *Let $\mathcal{F}$ be the family of trade-off functions used as per-step f-DP guarantees. The natural f-DP filter is valid for $\mathcal{F}$ in fully adaptive composition of up to $T \geq 2$ steps if $\mathcal{F}^{(-\otimes\mathcal{F})(T-2)}$ is a Blackwell chain.*

The form $\mathcal{F}^{(-\otimes\mathcal{F})(T-2)}$ arises from reducing T-step adaptive composition to a two-step setting, where the first and second trade-off functions are drawn from $\mathcal{F}$ and $\mathcal{F}^{(-\otimes\mathcal{F})(T-2)}$, respectively. The proof (Appendix D) shows that the Blackwell chain assumption allows one to conclude that if every $f_2 \in \mathcal{F}^{(-\otimes\mathcal{F})(T-2)}$ satisfies $f_1 \otimes f_2 \geq f_B$, then $f_2^{\downarrow}$ does as well. This deduction is a generalization of the RDP filter proof in Feldman and Zrnic (2021) and the GDP filter proof in Koskela et al. (2023), extended from scalar to function objects. As Gaussian trade-off functions are closed under tensor products, Theorem 3.5 straightforwardly recovers the valid GDP filter as a special case.

The notion of lower closure $\bar{\mathcal{F}}$ within $\mathcal{F}^{-\otimes\mathcal{F}}$ used above relies on convergence of trade-off functions, for which we adopt Δ-divergence (Kaissis et al., 2024) as the underlying metric. As an independent technical contribution, we prove that convergence in Δ-divergence is equivalent to pointwise convergence on $(0, 1]$ and to uniform convergence on compact subsets within $(0, 1]$ (Theorem C.8, Appendix C), providing a flexible toolkit for theoretical analyses of trade-off function approximations.

We conclude this section by sharpening Corollary 3.3 into a general converse statement.

Theorem 3.6. *Let $\mathcal{F}$ be the family of trade-offs used as per-step f-DP guarantees. Suppose that:*

1. *Some $f_1 \in \mathcal{F}$ is differentiable on $(\alpha_1, \alpha_2) \subset (0, 1)$ and has a nonconstant derivative there;*

2. $\mathcal{F}^{(-\otimes\mathcal{F})(T-2)}$ *contains two trade-off functions* f_{21} *and* f_{22} *that cross at some FPR* $\alpha_0 \in (0,1)$ *with different slopes* $f'_{21}(\alpha_0) \neq f'_{22}(\alpha_0)$.

then the natural f*-DP filter is invalid for* $\mathcal{F}$*: there exists a fully adaptive composition of at most* T *steps, defined on a single neighboring input pair whose trade-off function at each step belongs to* $\mathcal{F}$*, that passes every filter check but violates its prescribed budget.*[2]

The first condition in Theorem 3.6 includes Laplace and Poisson subsampled Gaussian mechanisms, but does not generally cover discrete-noise mechanisms such as discrete Gaussian. Theorem 3.5 and Theorem 3.6 together yield a sharp characterisation: *Under regularity assumptions, the natural* f*-DP filter is valid if and only if* $\mathcal{F}^{(-\otimes\mathcal{F})(T-2)}$ *is a Blackwell chain.*

4 Fully Adaptive f-DP Central Limit Theorem

Fully Adaptive Central Limit Theorem. Having established when the natural exact f-DP filter is and is not valid, we move on to proving a central limit theorem for fully adaptive privacy loss processes. In contrast to the previous section, we analyze the composition structure through the stochastic process of PLRVs. At each step t, the output y_t extends the observed history $y_{1:t-1}$, which in turn governs the choice of the next pair of distributions (P_{t+1}, Q_{t+1}) and hence the next PLRV L_{t+1}. If a filter halts the composition early, all subsequent PLRVs are set to zero. The total privacy loss is thus the terminal value of a stochastic process — a running sum of sequentially realized log-likelihood ratios, each drawn from a distribution that may depend on all preceding outputs. Despite this adaptive feedback, the cumulative sum remains approximately Gaussian under some control and regularity conditions.

Theorem 4.1. *Let* $P_1, \ldots, P_T$ *and* $Q_1, \ldots, Q_T$ *be adaptive sequences of distributions, where for each* t*, the choices of* P_t *and* Q_t *may depend on the past realizations* $y_1, \ldots, y_{t-1}$ *drawn under* $P_1, \ldots, P_{t-1}$ *and* $Q_1, \ldots, Q_{t-1}$*, respectively. Define* $L_t = \ln \frac{P_t}{Q_t}$ *as the PLRV at step* t*, and let* $P = \bigotimes_{t=1}^T P_t$*,* $Q = \bigotimes_{t=1}^T Q_t$*, and* $y = (y_1, \ldots, y_T)$*. We assume that the following conditions hold with probability 1:*

1. $\left|\sum_{t=1}^T \mathbb{E}_{y\sim P}[L_t(y_t) \,|\, y_{1:t-1}] - m_1\right| \leq \eta_1$,
2. $\left|\sum_{t=1}^T \mathbb{E}_{y\sim Q}[L_t(y_t) \,|\, y_{1:t-1}] + m_2\right| \leq \eta_2$,
3. $\left|\sum_{t=1}^T \mathrm{Var}_{y\sim P}[L_t(y_t) \,|\, y_{1:t-1}] - v\right| \leq \kappa$,
4. $\left|\sum_{t=1}^T \mathrm{Var}_{y\sim Q}[L_t(y_t) \,|\, y_{1:t-1}] - v\right| \leq \kappa$,
5. $\mathbb{E}_{y\sim P}\left[\left|L_t - \mathbb{E}_{y\sim P}[L_t(y_t) \,|\, y_{1:t-1}]\right|^3 \,\middle|\, y_{1:t-1}\right] \leq \rho \cdot \mathrm{Var}_{y\sim P}[L_t(y_t) \,|\, y_{1:t-1}]$ *for all* t,
6. $\mathbb{E}_{y\sim Q}\left[\left|L_t - \mathbb{E}_{y\sim Q}[L_t(y_t) \,|\, y_{1:t-1}]\right|^3 \,\middle|\, y_{1:t-1}\right] \leq \rho \cdot \mathrm{Var}_{y\sim Q}[L_t(y_t) \,|\, y_{1:t-1}]$ *for all* t.

Here, we require that $4\rho^2 \leq v$ *and* $4\kappa \leq v$*. Under these assumptions, the trade-off function* $T[P,Q]$ *satisfies the following inequalities for all* α*:*

$$G_{\mu+\phi}(\alpha + \Delta) - \Delta \leq T[P,Q](\alpha) \leq G_{\mu-\phi}(\alpha - \Delta) + \Delta,$$

where $\mu = \frac{m_1+m_2}{\sqrt{v}}$*,* $\phi = \frac{\eta_1+\eta_2}{\sqrt{v}}$*, and* $\Delta = C\left(\frac{\rho}{\sqrt{v}}\left|\ln\frac{\rho}{\sqrt{v}}\right| + \sqrt{\frac{\kappa}{v}}\right)$ *for some universal constant* C.

Our fully adaptive CLT generalizes the non-adaptive CLT of Dong et al. (2022). Instead of relying on the classical Berry–Esseen theorem, the proof's main idea (detailed in Appendix E) applies the martingale Berry–Esseen theorem of Fan (2019) to the martingale differences

$$\frac{L_t(y_t) - \mathbb{E}_{y\sim P}[L_t(y_t) \,|\, y_{1:t-1}]}{\sqrt{v}}, y \sim P \quad \text{and} \quad \frac{L_t(y_t) - \mathbb{E}_{y\sim Q}[L_t(y_t) \,|\, y_{1:t-1}]}{\sqrt{v}}, y \sim Q.$$

Conditions 5–6 control higher-order moments relative to variance, ensuring no small number of steps dominates privacy loss and allowing a Gaussian approximation with error Δ of such form—where the constant C, inherited from Fan (2019), is not tracked explicitly as is standard in this literature.

[2] We thank Matthew Regehr for pointing out a gap in an earlier version of Theorem 3.6.

Conditions 1–4 ensure path-realized sums of conditional means and variances remain approximately constant, reflecting moment computation from Dong et al. (2022).

More broadly, Conditions 1–4 suggest a general methodology for constructing approximate GDP filters: any mechanism family for which the cumulative conditional means and variances of PLRVs can be approximated analytically admits a closed-form approximate filter via Theorem 4.1. The resulting filter tracks these moment approximations in place of the full trade-off function, with the CLT error Δ quantifying the approximation quality.

Demonstration: Approximate GDP Filter for Subsampled Gaussian Mechanisms. We derive closed-form PLRV moment approximations for subsampled Gaussian mechanisms in the asymptotic regimes $q \ll 1$ and $q \approx 1$. For an arbitrary remove pair S, S^-, let $P = \mathcal{M}(S)$ and $Q = \mathcal{M}(S^-)$ denote the mechanism outputs, with σ the Gaussian noise and μ the norm of the removed datapoint's contribution. The leading-order Taylor expansions give:

1. For $q \ll 1$, $\mathbb{E}_P\left[\ln \frac{P}{Q}\right] \approx -\mathbb{E}_Q\left[\ln \frac{P}{Q}\right] \approx \frac{1}{2}\mathrm{Var}_P\left[\ln \frac{P}{Q}\right] \approx \frac{1}{2}\mathrm{Var}_Q\left[\ln \frac{P}{Q}\right] \approx \frac{1}{2}q^2\left(\mathrm{e}^{\frac{\mu^2}{\sigma^2}} - 1\right)$.
2. For $q \approx 1$, $\mathbb{E}_P\left[\ln \frac{P}{Q}\right] \approx -\mathbb{E}_Q\left[\ln \frac{P}{Q}\right] \approx \frac{1}{2}\mathrm{Var}_P\left[\ln \frac{P}{Q}\right] \approx \frac{1}{2}\mathrm{Var}_Q\left[\ln \frac{P}{Q}\right] \approx \frac{1}{2}q^2\frac{\mu^2}{\sigma^2}$.

Crucially, these terms depend on the remove pair only through μ, making them pair-invariant and enabling a path-wise filter that certifies privacy uniformly over all neighbouring pairs. We denote these per-step approximations $\mathrm{Budg}(\tilde{q}, q, \sigma, \mu)$, with $\tilde{q} \in \{0, 1\}$ indicating the regime. Approximation errors are numerically negligible for $q < 0.2$ and $q > 0.8$; proofs and error bounds are in Appendix F.

Algorithm 2 instantiates one approximate GDP filter, allowing $q_t \in (0, \bar{q}]$ with $\bar{q} \ll 1$ (regime $\tilde{q} = 0$) or $q_t \in [\bar{q}, 1]$ with $\bar{q} \approx 1$ (regime $\tilde{q} = 1$). The function $\mathrm{invBudg}$ is the algebraic inverse of Budg in μ, returning the clipping level that exactly uses the remaining budget. Since the PLRV mean equals the Rényi divergence of order one, the filter can be interpreted as an approximate order-one RDP-style filter, while providing a final guarantee of approximate GDP.

The following scaling conditions ensure the total privacy spend remains non-degenerate as T grows: $T\bar{q}^2 = \Theta(1)$ as $\bar{q} \to 0$ for regime $\tilde{q} = 0$, and $T\bar{\sigma}^{-2} = \Theta(1)$ with $\hat{q} := 1 - \bar{q} = O(\bar{\sigma}^{-2})$ as $\bar{\sigma} \to \infty$ for regime $\tilde{q} = 1$.

Theorem 4.2. *Under the above scaling assumptions, the composition of subsampled Gaussian mechanisms produced by Algorithm 2 satisfies:*

1. *Near $\tilde{q} = 0$: Δ-approx. $\left(\sqrt{2B} + O(\bar{q})\right)$-GDP, with an error $\Delta = O(\sqrt{\bar{q}})$ as $\bar{q} \to 0$.*
2. *Near $\tilde{q} = 1$: Δ-approx. $\left(\sqrt{2B} + O(\bar{\sigma}^{-2})\right)$-GDP, with an error $\Delta = O\left(\frac{\ln \bar{\sigma}}{\bar{\sigma}}\right)$ as $\bar{\sigma} \to \infty$.*

Algorithm 2: Approximate GDP Filter in Fully Adaptive Composition of Subsampled Gaussian Mechanisms

common input Boolean flag $\tilde{q}$, dataset S. Adaptively chosen sequence of objects for $t = 1, \ldots, T$: functions $g_t(\cdot, y_{1:t-1})$, sampling rates $q_t(y_{1:t-1})$, noise scales $\sigma_t(y_{1:t-1})$ in $[\bar{\sigma}, \infty)$, clips $C_t(y_{1:t-1})$.

additional input privacy budget B.

for $t = 1$ to T **do**
  choose g_t, q_t, σ_t, C_t based on $y_{1:t-1}$
  if $t = 1$ **then** $B_t \leftarrow B$
  else $B_t \leftarrow B_{t-1} - \mathrm{Budg}(\tilde{q}, q_{t-1}, \sigma_{t-1}, 1)$
  $C_{B,t} \leftarrow C_t \cdot \mathrm{invBudg}(\tilde{q}, q_t, \sigma_t, B_t)$
  sample S_t from S at rate q_t
  for $X_j \in S_t$ **do**
    $\grave{y}_{t,j} \leftarrow g_t(X_j) \cdot \min\left[1, \frac{\min(C_t, C_{B,t})}{\|g_t(X_j)\|}\right]$
  $y_t \leftarrow \sum_{X_j \in S_t} \grave{y}_{t,j} + \mathcal{N}\left(0, \sigma_t^2 C_t^2 \mathbf{I}\right)$
  if $C_{B,t} \leq C_t$ **then break**
return $y_1, \ldots, y_t$

Algorithm 3: Individual Privacy Accounting for Subsampled Gaussian Mechanisms via Approximate GDP Filtering

common input Boolean flag $\tilde{q}$, dataset S. Adaptively chosen sequence of objects for $t = 1, \ldots, T$: functions $g_t(\cdot, y_{1:t-1})$, sampling rates $q_t(y_{1:t-1})$, noise scales $\sigma_t(y_{1:t-1})$ in $[\bar{\sigma}, \infty)$, clips $C_t(y_{1:t-1})$.

additional input dataset size n, per-datapoint budgets B_j for $j = 1, \ldots, n$.
for $t = 1$ to T **do**
  choose g_t, q_t, σ_t, C_t based on $y_{1:t-1}$
  for X_j in S **do**
    if $t = 1$ **then** $B_{t,j} \leftarrow B_j$
    else $B_{t,j} \leftarrow B_{t-1,j} - \mathrm{Budg}(\tilde{q}, q_{t-1}, \sigma_{t-1} C_{t-1}, \|\grave{y}_{t-1,j}\|)$
    $C_{B,t,j} \leftarrow \mathrm{invBudg}(\tilde{q}, q_t, \sigma_t C_t, B_{t,j})$
    $\grave{y}_{t,j} \leftarrow g_t(X_j) \cdot \min\left[1, \frac{\min(C_t, C_{B,t,j})}{\|g_t(X_j)\|}\right]$
  sample S_t from S at rate q_t
  $y_t \leftarrow \sum_{X_j \in S_t} \grave{y}_{t,j} + \mathcal{N}\left(0, \sigma_t^2 C_t^2 \mathbf{I}\right)$
return $y_1, \ldots, y_T$

Theorem 4.2 is derived by applying Theorem 4.1 with $m_1 = m_2 = B$ and $v = 2B$ to an arbitrary remove pair (Appendix E). Verification that the CLT conditions hold in each regime is given by Theorems F.1 and F.2. The errors here vanish as $\bar{q} \to 0$ or $\bar{\sigma} \to \infty$, so the guarantee approaches exact $\sqrt{2B}$-GDP in each regime. Since the RDP-to-f-DP conversion is inherently lossy by a nonvanishing factor (Zhu et al., 2022), appropriate parameter selection yields strictly tighter guarantees than fully adaptive RDP accounting.

The CLT methodology extends naturally to individual privacy accounting. Algorithm 3 replaces the single global budget of Algorithm 2 with per-individual budgets $B_{t,j}$ for each datapoint $X_j \in S$, updating privacy costs using $\|\dot{g}_{t-1,j}\|$ rather than the shared worst-case bound C_{t-1}. The budget-aware clipping threshold $C_{B,t,j}$ is chosen so that the next update cannot overspend $B_{t,j}$; once depleted, $C_{B,t,j} = 0$ and X_j contributes no further. In fact, Algorithm 3 generalizes the individual accounting of Feldman and Zrnic (2021) from $q = 1$ to both asymptotic regimes $q \ll 1$ and $q \approx 1$.

Theorem 4.3. *Suppose the initial individual budgets are uniform, i.e. $B_j = B$ for all $j = 1, \ldots, n$. Then Algorithm 3 satisfies identical asymptotic privacy guarantees as stated in Theorem 4.2.*

Numerical Validation via Monte Carlo Simulation. We validate Theorem 4.2 through Monte Carlo simulation of DP-SGD-like fully adaptive compositions with Poisson subsampling. At a fixed subsampling rate q, adaptive noise and clipping selections follow an AdaGrad-inspired rule depending on previous outputs, see Appendix G. We note that we choose and evaluate one sensible adaptive selection policy; searching for the worst-case policy is outside the scope of this paper.

For each fixed q and each add/remove direction, we simulate 10^{10} trajectories and compute empirical (ε, δ) curves using EVR-style pessimistic estimation (Wang et al., 2023), with a simultaneous $1-10^{-9}$ confidence band over the ε-grid used. We compare these curves with the (ε, δ) curves obtained by converting from the theoretical $\sqrt{2B}$-GDP bounds and from the single-order $(1, B)$-RDP filter—the structurally equivalent baseline, as the approximate GDP filter operates by tracking no more than order-one Rényi divergence (see below).

Figure 2 illustrates Gaussian convergence of the empirical PLD and the corresponding privacy bound comparison. At $q = 0.1$, the $\sqrt{2B}$-GDP approximation underestimates privacy loss—a known

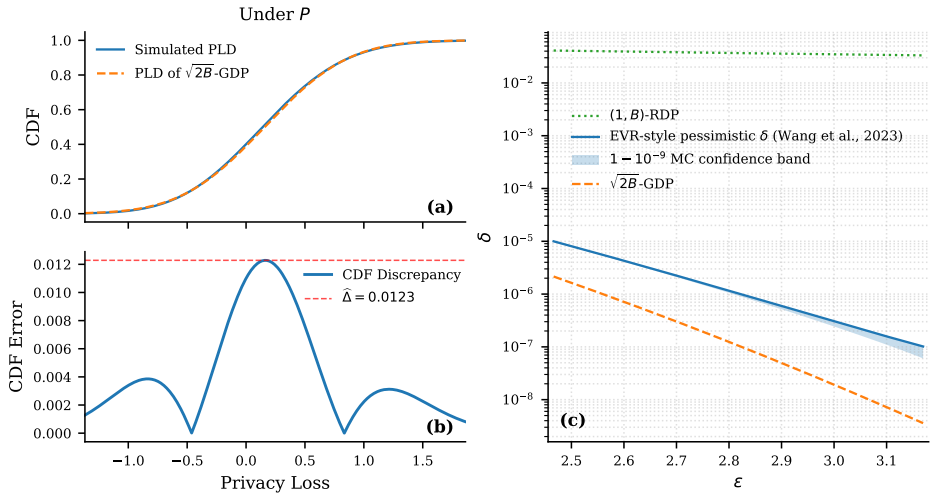


Figure 2: Monte Carlo evaluation of the approximate $\sqrt{2B}$-GDP filter (Theorems 4.2 and 4.1) for DP-SGD-like fully adaptive composition with Poisson subsampling at $q = 0.1$. Adaptive noise and clipping follow an Adagrad-inspired rule depending on previous outputs; we simulate 10^{10} trajectories of privacy losses (Appendix G). **(a)** Empirical CDF of privacy loss vs. its Gaussian approximation, under P. **(b)** Absolute CDF discrepancy, with maximum $\Delta \approx 0.0123$, a large deviation from $\sqrt{2B}$-GDP bound. **(c)** Empirical EVR-style pessimistic upper δ curve (Wang et al., 2023) maximized over add and remove directions, with a simultaneous $1 - 10^{-9}$ confidence band over the ε-grid, compared with the baseline $(1, B)$-RDP guarantee and the exact $\sqrt{2B}$-GDP used when applying Theorem 4.2 in practice with Δ omitted.

limitation of CLT-based approximations even in the non-adaptive case (Zheng et al., 2020; Gopi et al., 2021; Alghamdi et al., 2023)—while the $(1, B)$-RDP bound is notably extremely pessimistic.

The tightness over RDP has a clear interpretation: since the PLRV mean equals the order-one Rényi divergence, the approximate GDP filter effectively operates as a single order-one RDP filter. Compared to more restrictive multi-order RDP filters, this single-order structure can permit more mechanisms to proceed, while yielding tighter guarantees than the $(1, B)$-RDP guarantee given by the single-order RDP filter—though at the cost of validity outside the asymptotic regimes.

Table 1: Empirical error Δ and $\sqrt{2B}$-GDP values across subsampling rates. Δ decreases as $q \to 0$ or $q \to 1$.

q	Δ Error	$\sqrt{2B}$-GDP
0.01	0.003533	0.1775
0.1	0.01450	0.5612
0.199	0.02584	0.7897
0.801	0.01461	1.416
0.95	0.004735	1.454

Table 1 shows smaller Δ toward the low- and high-subsampling regimes, consistent with Theorem 4.2. The theorem further shows that this error can be made arbitrarily small by selecting sufficiently extreme q, making the $\sqrt{2B}$-GDP approximation arbitrarily accurate. Together with the improvement over the matched $(1, B)$-RDP guarantee, this demonstrates that the slack in RDP stems from its lossy Rényi divergence representation rather than any fundamental barrier in adaptive composition. Whether tighter multi-order-equivalent approximate filters—constructed via the same CLT framework by tracking higher-order PLRV moments—can close the remaining gap at moderate subsampling rates is a natural direction for future work.

5 Discussion and Conclusions

This work resolves a central open question in fully adaptive privacy accounting for f-DP. We show that the natural f-DP filter is invalid under full adaptivity and establish necessary and sufficient conditions for validity—a sharp characterization that explains why divergence-based notions such as RDP admit simple filters while trade-off-function-based notions do not. On the constructive side, we prove a fully adaptive f-DP central limit theorem, establishing Gaussian convergence of cumulative privacy losses under full adaptivity, suggesting a general methodology for constructing approximate GDP filters via mechanism-specific PLRV moment approximations. As one instantiation, we construct a closed-form approximate GDP filter for subsampled Gaussian mechanisms that, in asymptotic regimes, yields valid guarantees strictly tighter than fully adaptive RDP accounting without tracking the full trade-off function—demonstrating that the slack in RDP stems from its lossy Rényi divergence representation rather than any fundamental barrier in adaptive composition.

Acknowledgements

We would like to thank Matthew Regehr for identifying a gap in an earlier version of Theorem 3.6, which prompted the revised assumptions and proof. This research was conducted in collaboration with Nokia Bell Labs, as part of Nokia's Veturi program. The program is supported by funding from Business Finland, whose contribution is gratefully acknowledged. This work was also supported by the Research Council of Finland (Flagship programme: Finnish Center for Artificial Intelligence, FCAI, Grant 356499 and Grant 359111), the Strategic Research Council at the Research Council of Finland (Grant 358247), and the European Union (Project 101070617). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them. The authors acknowledge the research environment provided by ELLIS Institute Finland.

Use of LLMs and AI agents. The authors developed the research and wrote the initial manuscript, with LLM assistance limited to language polishing. During subsequent revisions, LLMs and coding agents assisted with implementing and refining experiments, discussing and checking mathematical arguments, literature searches, and manuscript revision. The authors determined the scientific direction and final decisions and take responsibility for the paper.

References

W. Alghamdi, J. F. Gomez, S. Asoodeh, F. Calmon, O. Kosut, and L. Sankar. The saddle-point method in differential privacy. In *Proceedings of the 40th International Conference on Machine Learning*, pages 508–528, 2023.

G. Andrew, O. Thakkar, H. B. McMahan, and S. Ramaswamy. Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems*, 2021.

J. Awan and S. Vadhan. Canonical noise distributions and private hypothesis tests. In *NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021.

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

C. L. Canonne, G. Kamath, and T. Steinke. The discrete Gaussian for differential privacy. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

J. Dong, A. Roth, and W. J. Su. Gaussian Differential Privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 2022.

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

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284, 2006.

X. Fan. Exact rates of convergence in some martingale central limit theorems. *Journal of Mathematical Analysis and Applications*, 469(2):1028–1044, 2019.

X. Fan and Q.-M. Shao. Berry–Esseen bounds for self-normalized martingales. *Communications in Mathematics and Statistics*, 6, 2017.

V. Feldman and T. Zrnic. Individual Privacy Accounting via a Rényi Filter. In *Advances in Neural Information Processing Systems*, pages 28080–28091, 2021.

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

S. Haney, M. Shoemate, G. Tian, S. Vadhan, A. Vyrros, V. Xu, and W. Zhang. Concurrent composition for interactive differential privacy with adaptive privacy-loss parameters. *Journal of Privacy and Confidentiality*, 15, 2025.

C. C. Heyde and B. M. Brown. On the Departure from Normality of a Certain Class of Martingales. *The Annals of Mathematical Statistics*, 41(6):2161–2165, 1970.

G. Kaissis, S. Kolek, B. Balle, J. Hayes, and D. Rueckert. Beyond the Calibration Point: Mechanism Comparison in Differential Privacy. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

A. Koskela, J. Jälkö, and A. Honkela. Computing tight differential privacy guarantees using FFT. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 2560–2569, 2020.

A. Koskela, M. Tobaben, and A. Honkela. Individual Privacy Accounting with Gaussian Differential Privacy. In *The Eleventh International Conference on Learning Representations*, 2023.

C. J. Lebeda, M. Regehr, G. Kamath, and T. Steinke. Avoiding pitfalls for privacy accounting of subsampled mechanisms under composition. In *Proceedings of the IEEE Conference on Secure and Trustworthy Machine Learning*, pages 996–1006, 2025.

A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

M. Regehr, B. Hu, E. Leeman, P. Manurangsi, P. Tholoniat, and M. Lécuyer. Privacy filters are captured by residues: A characterization of free natural filters and the cost of adaptivity. *arXiv preprint arXiv:2602.15815*, 2026.

A. Riess, J. F. Gomez, F. Calmon, J. Schnabel, and G. Kaissis. Optimal conversion from Rényi Differential Privacy to f-Differential Privacy. In *Proceedings of the 43rd International Conference on Machine Learning*, pages 105180–105192, 2026.

R. Rogers, A. Roth, J. Ullman, and S. Vadhan. Privacy odometers and filters: pay-as-you-go composition. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1929–1937, 2016.

J. T. Wang, S. Mahloujifar, T. Wu, R. Jia, and P. Mittal. A randomized approach to tight privacy accounting. In *Advances in Neural Information Processing Systems*, pages 33856–33893, 2023.

J. Whitehouse, A. Ramdas, R. Rogers, and Z. S. Wu. Fully-adaptive composition in differential privacy. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Q. Zheng, J. Dong, Q. Long, and W. Su. Sharp composition bounds for gaussian differential privacy via edgeworth expansion. *Proceedings of machine learning research*, 119:11420–11435, 2020.

Y. Zhu, J. Dong, and Y.-X. Wang. Optimal Accounting of Differential Privacy via Characteristic Function. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 4782–4817, 2022.

Contents

1 Introduction **1**

2 Background **3**
Differential Privacy 3
Extended DP Notions 4
Fully Adaptive Composition 4

3 Natural f-DP Filter **5**
Calibration Point and Filter Invalidity 5
Blackwell Chains and Filter Characterization 6

4 Fully Adaptive f-DP Central Limit Theorem **7**
Fully Adaptive Central Limit Theorem 7
Demonstration: Approximate GDP Filter for Subsampled Gaussian Mechanisms . . . 8
Numerical Validation via Monte Carlo Simulation 9

5 Discussion and Conclusions **10**

Acknowledgements **10**

References **12**

Appendix **14**

A Additional Background: Privacy Profile **14**
A.1 Privacy Profile vs. Trade-off Function 14
A.2 Tensor Product of Trade-off Functions/Privacy Profiles 15

B Counterexamples Against f-DP Filter Validity **18**
B.1 Details of the Counterexample in Corollary 3.3 19
B.2 Proof of Adaptive Tightness of Standard Composition Results 20

C Convergence of Privacy Profiles/Trade-off Functions **22**
C.1 Existence of Convergent Sequence in Blackwell Chains 22
C.2 Unifying Various Notions of Convergence 22

D Proofs for f-DP Filter Validity via Blackwell Chains **25**
D.1 Blackwell Chain Property Implies f-DP Filter Validity 25
D.2 Failure of Blackwell Chain Property Implies f-DP Filter Invalidity 27

E Fully Adaptive Central Limit Theorem **29**
E.1 Proof to the Fully Adaptive f-DP Central Limit Theorem 29
E.2 Proofs for Demonstration of Approximate GDP Filter 31

F Privacy Loss Moments: Approximations **32**
F.1 Numerical Illustration of Moment Approximations 33
F.2 Derivations Using Taylor Expansions 35

G Further Monte Carlo Simulations for the Approximate GDP Filter **42**
G.1 Conversion of RDP Order 1 to (ε, δ)-DP 42
G.2 Details of Monte Carlo Simulations 43

NeurIPS Paper Checklist **54**

A Additional Background: Privacy Profile

A.1 Privacy Profile vs. Trade-off Function

The results here are standard; proofs can be found in Dong et al. (2022), Zhu et al. (2022), Kaissis et al. (2024). The hockey-stick divergence is an integral that directly captures (ε, δ)-DP bounds. Define $[x]_+ = \max(x, 0)$.

Definition A.1. The hockey-stick divergence at order $\gamma \geq 0$ of a pair of distributions P, Q with densities p, q is

$$D_\gamma[P \,||\, Q] = \int_{\mathcal{O}} [p(y) - \gamma q(y)]_+ \, \mathrm{d}y = \max_{A \subset \mathcal{O}} P(A) - \gamma Q(A) \,.$$

A mechanism $\mathcal{M}$ is (ε, δ)-DP if $D_{e^\varepsilon}[\mathcal{M}(S) \,||\, \mathcal{M}(S^-)] \leq \delta$ and $D_{e^\varepsilon}[\mathcal{M}(S^-) \,||\, \mathcal{M}(S)] \leq \delta$ for all remove pair S, S^-.

The privacy profile on a mechanism level defined by Balle et al. (2020) is given as Definition A.5 below. We first give a definition on the distribution level.

Definition A.2. The privacy profile of a pair of distributions (P, Q) is a function $H : [0, \infty) \to [0, 1]$ that represents the hockey-stick divergence $H(\gamma) = D_\gamma[P \,||\, Q]$ of all orders $\gamma \geq 0$.

Theorem A.3. *A function $H : [0, \infty) \to [0, 1]$ is the privacy profile of some pair of distributions if and only if $H(0) = 1$, H is decreasing, convex, and $H(\gamma) \geq [1 - \gamma]_+$.*

We can describe the spectrum of all (ε, δ)-DP bounds using privacy profiles.

Definition A.4. A mechanism $\mathcal{M}$ is dominated by a privacy profile H if for all remove pair S, S^-,

$$D_\gamma\big[\mathcal{M}(S) \,||\, \mathcal{M}\big(S^-\big)\big] \leq H(\gamma) \quad \text{and} \quad D_\gamma\big[\mathcal{M}\big(S^-\big) \,||\, \mathcal{M}(S)\big] \leq H(\gamma)$$

at all $\gamma \geq 0$. $\mathcal{M}$ is dominated by a privacy profile H under remove adjacency if only the first inequality holds. $\mathcal{M}$ is dominated by a pair of distributions (P, Q) (under remove adjacency) if it is dominated by the privacy profile of (P, Q) (under remove adjacency).

For a family of privacy profiles $\mathcal{H}$, the point-wise supremum $H^\uparrow$ of all $H \in \mathcal{H}$ is another privacy profile by Theorem A.3, and therefore $H^\uparrow$ is the smallest upper bound of $\mathcal{H}$.

Definition A.5. The privacy profile of a mechanism $\mathcal{M}$ is the one that tightly dominates $\mathcal{M}$:

$$H_{\mathcal{M}}(\gamma) = \sup_{S, S^- \text{ remove pair}} \max\big(D_\gamma\big[\mathcal{M}(S) \,||\, \mathcal{M}\big(S^-\big)\big], D_\gamma\big[\mathcal{M}\big(S^-\big) \,||\, \mathcal{M}(S)\big]\big).$$

$H_{\mathcal{M}}$ represents tight δ-bounds in (ε, δ)-DP for all $\varepsilon \in \mathbb{R}$. In addition, the privacy profile of $\mathcal{M}$ under remove adjacency is

$$H^-_{\mathcal{M}}(\gamma) = \sup_{S, S^- \text{ remove pair}} D_\gamma\big[\mathcal{M}(S) \,||\, \mathcal{M}\big(S^-\big)\big] \,.$$

In fact, tight f-DP is equivalent to a privacy profile of (ϵ, δ)-DP bounds. This relationship justifies the conventional reason why the privacy profile includes the parameters $\gamma = e^\varepsilon \in (0, 1)$, or $\varepsilon < 0$.

Theorem A.6. *There is a bijection between the trade-off functions and the privacy profiles. Specifically, let f and H be the trade-off function and privacy profile of (P, Q) respectively, then*

1. $H(\gamma) = 1 + \big(f^{-1}\big)^\star(-\gamma)$ *for all* $\gamma \geq 0$,
2. $f^{-1}(\alpha) = 1 + H^\star(-\alpha)$ *for all* $\alpha \in [0, 1]$,

where * denotes the convex conjugate operation and $f^{-1} = T[Q, P]$ is the inverse function of f (see Proposition A.8).

Furthermore, this bijection carries over the Blackwell ordering.

Theorem A.7. *Let f_1, f_2 be trade-off functions, and let H_1, H_2 be the corresponding privacy profiles respectively. Then the following are equivalent.*

1. $f_1(\alpha) \geq f_2(\alpha)$ *for all* $\alpha \in [0,1]$.
2. $H_1(\gamma) \leq H_2(\gamma)$ *for all* $\gamma \geq 0$.

Similar to trade-off functions, we say that the privacy profiles H_1 and H_2 are Blackwell ordered if $H_1 \leq H_2$ or $H_1 \geq H_2$ point-wise. In light of Theorem A.6 and Theorem A.7, Definition 2.5 of f-DP is equivalent to Definition A.4 of the dominating privacy profile.

For a family of trade-off functions $\mathcal{F}$, their point-wise infimum does not maintain convexity in general, and hence it is not necessarily a trade-off function. Instead, the greatest lower bound $f^{\downarrow}$ of all $f \in \mathcal{F}$ can be obtained by first converting $\mathcal{F}$ to the corresponding family of privacy profiles $\mathcal{H}$ by using Theorem A.6, taking the point-wise supremum $H^{\uparrow}$ and converting $H^{\uparrow}$ to $f^{\downarrow}$ using Theorem A.6 again. This is equivalently the lower convex envelope of the point-wise infimum of $\mathcal{F}$.

Proposition A.8. *For a pair of distributions* (P, Q)*,*

1. $T[Q,P]$ *is the reflection of* $T[P,Q]$ *across the identity line.*
2. *For all* $\gamma > 0$*,*
$$D_\gamma[Q\,||\,P] = 1 - \gamma + \gamma D_{\frac{1}{\gamma}}[P\,||\,Q]\,.$$

Corollary A.9. *Let* $H_{\mathcal{M}}$ *be the privacy profile of a mechanism* $\mathcal{M}$*, and let* $f_{\mathcal{M}}$ *be the corresponding tightest trade-off function such that* $\mathcal{M}$ *is* $f_{\mathcal{M}}$*-DP. Then the following holds.*

1. *The graph of* $f_{\mathcal{M}}(\alpha)$ *is symmetric w.r.t. the identity line.*
2. *For all* $\gamma > 0$*,* $H_{\mathcal{M}}(\gamma) = 1 - \gamma + \gamma H_{\mathcal{M}}\left(\frac{1}{\gamma}\right)$.

A privacy profile $H^{-}_{\mathcal{M}}$ of $\mathcal{M}$ under remove adjacency implies that the privacy profile $H_{\mathcal{M}}$ is

$$H_{\mathcal{M}}(\gamma) = \max\left(H^{-}_{\mathcal{M}}(\gamma)\,, 1 - \gamma + \gamma H^{-}_{\mathcal{M}}\left(\frac{1}{\gamma}\right)\right).$$

It is therefore sufficient to analyze privacy profiles under remove adjacency.

Definition A.10. A privacy profile H is symmetric if $H(\gamma) = 1 - \gamma + \gamma H\left(\frac{1}{\gamma}\right)$.

For an arbitrary privacy profile H, denote $\hat{H}(\gamma) = 1 - \gamma + \gamma H\left(\frac{1}{\gamma}\right)$ and $\mathrm{sym}(H) = \max\left(H, \hat{H}\right)$. Note that if H is symmetric, then $H = \hat{H} = \mathrm{sym}(H)$.

A.2 Tensor Product of Trade-off Functions/Privacy Profiles

We give a self-contained introduction to tensor products of trade-off functions and privacy profiles. Building on Dong et al. (2022), we present the theory from the privacy-profile perspective.

Zhu et al. (2022) showed that the dominating pairs of distributions transform nicely under standard compositions.

Theorem A.11. *If* $\mathcal{M}_1$ *is dominated by* (P_1, Q_1)*, and the adaptively chosen* $\mathcal{M}_2$ *is dominated by a fixed pair* (P_2, Q_2)*, then* $(\mathcal{M}_1, \mathcal{M}_2)$ *is dominated by* $(P_1 \times P_2, Q_1 \times Q_2)$.

This motivates the definition of the tensor product of privacy profiles, or equivalently, trade-off functions.

Definition A.12. Let H_1, H_2 be privacy profiles. The tensor product $H_1 \otimes H_2$ is defined as
$$(H_1 \otimes H_2)(\gamma) = D_\gamma[P_1 \times P_2\,||\,Q_1 \times Q_2]\,,$$
where (P_1, Q_1) is any pair of distributions that has a privacy profile of H_1, and (P_2, Q_2) is any pair of distributions that has a privacy profile of H_2. Using Fubini's theorem,

$$\begin{aligned}(H_1 \otimes H_2)(\gamma) &= \int_{\mathcal{O}_1 \times \mathcal{O}_2} \left[p_1(y_1)\, p_2(y_2) - \gamma q_1(y_1)\, q_2(y_2)\right]_+ \mathrm{d}y_1 \mathrm{d}y_2 \\ &= \int_{\mathcal{O}_1} H_2\left(\gamma \frac{q_1(y_1)}{p_1(y_1)}\right) p_1(y_1)\, \mathrm{d}y_1.\end{aligned}$$

Proposition A.13. *The tensor product of privacy profiles satisfies all following properties.*

0. *It is well-defined.*
1. *It is commutative and associative.*
2. *It carries over the Blackwell ordering: If $H_{21} \geq H_{22}$, then $H_1 \otimes H_{21} \geq H_1 \otimes H_{22}$.*
3. *$H \otimes H_{\text{Id}} = H_{\text{Id}} \otimes H = H$, where $H_{\text{Id}}(\gamma) = [1-\gamma]_+$ is the perfectly private privacy profile.*
4. *$\widehat{(H_1 \otimes H_2)} = \hat{H}_1 \otimes \hat{H}_2$.*

Proof. Proof for 0. We need to show that the resulting privacy profile $H_1 \otimes H_2$ is independent of the distribution pair (P_1, Q_1) and (P_2, Q_2) behind H_1 and H_2. It is sufficient to show that property 2 holds: By property 2, $D_\gamma[P_{21}, Q_{21}] = D_\gamma[P_{22}, Q_{22}]$ for all γ implies that $D_\gamma[P_1 \times P_{21}, Q_1 \times Q_{21}] = D_\gamma[P_1 \times P_{22}, Q_1 \times Q_{22}]$ for all γ.

Proof for 2. Using Fubini's theorem,

$$(H_1 \otimes H_{21})(\gamma) = \int_{\mathcal{O}_1} H_{21}\left(\gamma \frac{q_1(y_1)}{p_1(y_1)}\right) p_1(y_1)\,\mathrm{d}y_1 \\ \geq \int_{\mathcal{O}_1} H_{22}\left(\gamma \frac{q_1(y_1)}{p_1(y_1)}\right) p_1(y_1)\,\mathrm{d}y_1 = (H_1 \otimes H_{22})(\gamma)\,.$$

Proof for 1. $H_1 \otimes H_2 = H_2 \otimes H_1$ follows from

$$p_1(y_1)\,p_2(y_2) = p_2(y_2)\,p_1(y_1)\,, \\ q_1(y_1)\,q_2(y_2) = q_2(y_2)\,q_1(y_1)\,.$$

$(H_1 \otimes H_2) \otimes H_3 = H_1 \otimes (H_2 \otimes H_3)$ follows from the Fubini's theorem combined with

$$p_1(y_1)\,p_2(y_2)\,p_3(y_3) = p_1(y_1)[p_2(y_2)\,p_3(y_3)]\,, \\ q_1(y_1)\,q_2(y_2)\,q_3(y_3) = q_1(y_1)[q_2(y_2)\,q_3(y_3)]\,.$$

Proof for 3. H_{Id} is the privacy profile of any pair of identical distributions (P_2, P_2). Therefore

$$(H_1 \otimes H_2)(\gamma) = \int_{\mathcal{O}_1 \times \mathcal{O}_2} [p_1(y_1)\,p_2(y_2) - \gamma q_1(y_1)\,p_2(y_2)]_+ \,\mathrm{d}y_1 \mathrm{d}y_2 \\ = \int_{\mathcal{O}_1 \times \mathcal{O}_2} [p_1(y_1) - \gamma q_1(y_1)]_+\,p_2(y_2)\,\mathrm{d}y_1 \mathrm{d}y_2 = \int_{\mathcal{O}_1} [p_1(y_1) - \gamma q_1(y_1)]_+ \,\mathrm{d}y_1 = H_1(\gamma)\,.$$

Proof for 4. By the definition of tensor product and Proposition A.8,

$$\widehat{(H_1 \otimes H_2)}(\gamma) = D_\gamma[Q_1 \times Q_2 \,||\, P_1 \times P_2] = \left(\hat{H}_1 \otimes \hat{H}_2\right)(\gamma)\,.$$

□

Properties 2 and 3 from Proposition A.13 together imply that $H_1 \otimes H_2 \geq H_1$, i.e. tensor product weakens privacy protections. The following Theorem A.14 is equivalent to Theorem A.11

Theorem A.14. *If $\mathcal{M}_1$ is dominated by H_1, and the adaptively chosen $\mathcal{M}_2$ is dominated by a fixed H_2, then $(\mathcal{M}_1, \mathcal{M}_2)$ is dominated by $H_1 \otimes H_2$.*

Alternatively, we can define the tensor product of trade-off functions. We list results by Dong et al. (2022), which are equivalent to the corresponding versions in terms of privacy profiles presented above.

Definition A.15. Let f_1, f_2 be trade-off functions. The tensor product $f_1 \otimes f_2$ is defined as

$$f_1 \otimes f_2 = T[P_1 \times P_2, Q_1 \times Q_2]\,,$$

where (P_1, Q_1) is any pair of distributions that satisfies $f_1 = T[P_1, Q_1]$, and (P_2, Q_2) is any pair of distributions that satisfies $f_2 = T[P_2, Q_2]$.

Proposition A.16. *The tensor product of trade-off functions satisfies all the following properties.*

0. *It is well-defined.*
1. *It is commutative and associative.*
2. *It carries over the Blackwell ordering: If $f_{21} \leq f_{22}$, then $f_1 \otimes f_{21} \leq f_1 \otimes f_{22}$.*
3. *$f \otimes \mathrm{Id} = \mathrm{Id} \otimes f = f$, where $\mathrm{Id}(\alpha) = [1-\alpha]_+$ is the perfectly private privacy profile.*
4. *$(f_1 \otimes f_2)^{-1} = f_1^{-1} \otimes f_2^{-1}$.*

Proof. The proposition is equivalent to Proposition A.13. □

Theorem A.17. *If $\mathcal{M}_1$ is f_1-DP, and the adaptively chosen $\mathcal{M}_2$ is f_2-DP with a fixed f_2, then $(\mathcal{M}_1, \mathcal{M}_2)$ is $f_1 \otimes f_2$-DP.*

The standard adaptive composition results in GDP follow by using a simple relationship.

Proposition A.18. *Gaussian trade-off functions satisfy $G_{\mu_1} \otimes G_{\mu_2} = G_{\sqrt{\mu_1^2+\mu_2^2}}$.*

Proof. The hockey-stick divergence is an F-divergence, constructed using the convex function $[x-\gamma]_+ = \max(x-\gamma, 0)$. Therefore, it satisfies the data-processing inequality: for any randomized function $\mathcal{T}$, $D_\gamma[\mathcal{T}(P), \mathcal{T}(Q)] \leq D_\gamma[P, Q]$ for all γ. In particular, if $\mathcal{T}$ is a deterministic bijection, then $D_\gamma[\mathcal{T}(P), \mathcal{T}(Q)] = D_\gamma[P, Q]$ for all γ, and equivalently $T[\mathcal{T}(P), \mathcal{T}(Q)] = T[P, Q]$ by Theorem A.6.

For Gaussian trade-off functions G_{μ_1} and G_{μ_2}, we apply the rotation transformation as a bijection,

$$\begin{aligned} G_{\mu_1} \otimes G_{\mu_2} = T[\mathcal{N}(0, \mathbf{I}_2), \mathcal{N}((\mu_1, \mu_2), \mathbf{I}_2)] &= T\left[\mathcal{N}(0, \mathbf{I}_2), \mathcal{N}\left(\left(\sqrt{\mu_1^2+\mu_2^2}, 0\right), \mathbf{I}_2\right)\right] \\ &= T\left[\mathcal{N}(0,1), \mathcal{N}\left(\sqrt{\mu_1^2+\mu_2^2}, 1\right)\right] = G_{\sqrt{\mu_1^2+\mu_2^2}}. \end{aligned}$$

□

Corollary A.19. *If $\mathcal{M}_1$ is μ_1-GDP, and the adaptively chosen $\mathcal{M}_2$ is μ_2-GDP with a fixed μ_2, then $(\mathcal{M}_1, \mathcal{M}_2)$ is $\sqrt{\mu_1^2+\mu_2^2}$-GDP.*

The following Lemma A.20 shows that performing analysis per pair—applying tensor products on the pair's trade-off functions before taking the infimum—can yield tighter bounds than the standard composition Theorem A.17, which directly composes the mechanism-level privacy bounds.

Lemma A.20. *Let $\mathcal{F}_1 = \{f_{1,i}\}_{i \in I}$ and $\mathcal{F}_2 = \{f_{2,i}\}_{i \in I}$ be families of trade-off functions with common indexing by $i \in I$. Then*

$$\operatorname{ce}\inf_{i \in I} f_{1,i} \otimes f_{2,i} \geq \left(\operatorname{ce}\inf_{i \in I} f_{1,i}\right) \otimes \left(\operatorname{ce}\inf_{i \in I} f_{2,i}\right),$$

where ce inf *denotes the convex envelope of the point-wise infimum. There exists an example in which the inequality is strict.*

Proof. For all $i \in I$, $f_{1,i} \geq \operatorname{ce}\inf_{i \in I} f_{1,i}$ and $f_{2,i} \geq \operatorname{ce}\inf_{i \in I} f_{2,i}$. By Property 2 of Proposition A.16, $f_{1,i} \otimes f_{2,i} \geq (\operatorname{ce}\inf_{i \in I} f_{1,i}) \otimes (\operatorname{ce}\inf_{i \in I} f_{2,i})$ for all $i \in I$, and hence we can take the ce inf of the left hand side.

An example in which the inequality is strict: For $\mu \in [0,1]$, let $f_{1,\mu} = G_\mu$ and $f_{2,\mu} = G_{1-\mu}$, then

$$\operatorname{ce}\inf_{i \in I} f_{1,i} \otimes f_{2,i} = G_1 \geq G_{\sqrt{2}} = \left(\operatorname{ce}\inf_{i \in I} f_{1,i}\right) \otimes \left(\operatorname{ce}\inf_{i \in I} f_{2,i}\right).$$

□

Corollary A.21. *For arbitrary trade-off functions f_1, f_2, $\operatorname{sym}(f_1 \otimes f_2) \geq \operatorname{sym}(f_1) \otimes \operatorname{sym}(f_2)$.*

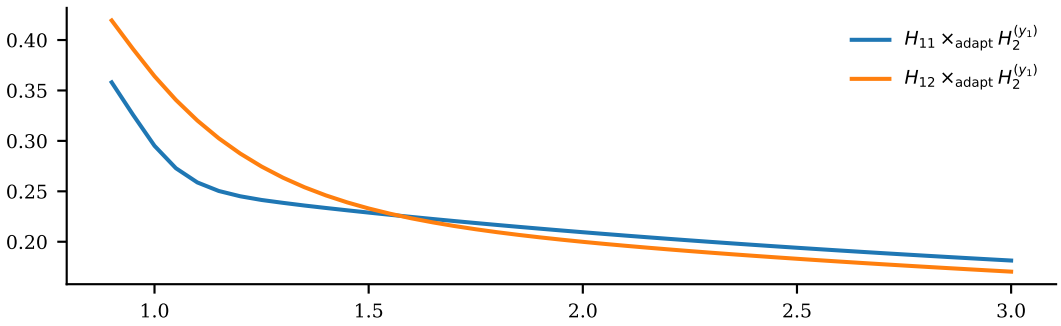


Figure 3: Privacy profiles $H_{11} \otimes_{\text{adapt}} H_2^{(y_1)}$ and $H_{12} \otimes_{\text{adapt}} H_2^{(y_1)}$ for (P_{11}, Q_{11}) being 0.1-GDP and (P_{12}, Q_{12}) being 0.3-GDP, with the same selection rule for step 2 that assigns $H_{G_{0.01}}$ when $y_1 < 0.3$ and H_{G_2} otherwise. The two curves cross near $\gamma \approx 1.5$, demonstrating that the Blackwell ordering $H_{11} \leq H_{12}$ of the first profiles is not preserved under adaptive tensor product.

Adaptive Tensor Product. Translating the adaptive tensor product (1) via Theorem A.6 and Definition A.12, the adaptive tensor product of privacy profiles takes the form

$$\left(H_1 \otimes_{\text{adapt}} H_2^{(y_1)}\right)(\gamma) = \int_{\mathcal{O}_1} H_2^{(y_1)}\left(\gamma \, \frac{q_1(y_1)}{p_1(y_1)}\right) p_1(y_1) \, \mathrm{d}y_1,$$

where the selection rule $H_2^{(y_1)}$ picks a privacy profile for the second mechanism based on the first output y_1. Analogously to Property 2 in Proposition A.13, if $H_{21}^{(y_1)} \leq H_{22}^{(y_1)}$ for every y_1, then

$$H_1 \otimes_{\text{adapt}} H_{21}^{(y_1)} \leq H_1 \otimes_{\text{adapt}} H_{22}^{(y_1)}$$

immediately by Fubini's theorem. However, ordering in the first component behaves very differently.

Remark A.22. The adaptive tensor product is not well-defined as an operation on privacy profiles/trade-off functions alone. Unlike the standard tensor product, the implication

$$H_{11} \leq H_{12} \;\not\Rightarrow\; H_{11} \otimes_{\text{adapt}} H_2^{(y_1)} \leq H_{12} \otimes_{\text{adapt}} H_2^{(y_1)}$$

fails in general. The reason is that the selection rule maps each first-mechanism output y_1 to a second-mechanism profile via the likelihood ratio $q_1(y_1)/p_1(y_1)$, so the result depends on how (P_1, Q_1) is organised over y_1, not merely on the profile H_1 it induces. Two distribution pairs sharing the same privacy profile can therefore yield different adaptive tensor products under the same selection rule. To see this concretely, let (P_{11}, Q_{11}) be 0.1-GDP and (P_{12}, Q_{12}) be 0.3-GDP, so that $H_{11} \leq H_{12}$. Fix the selection rule

$$H_2^{(y_1)} = \begin{cases} H_{G_{0.01}} & \text{if } y_1 < 0.3, \\ H_{G_2} & \text{otherwise,} \end{cases}$$

where H_{G_μ} denotes the privacy profile corresponding to the trade-off function G_μ of μ-GDP via Theorem A.6. Figure 3 plots $H_{11} \otimes_{\text{adapt}} H_2^{(y_1)}$ and $H_{12} \otimes_{\text{adapt}} H_2^{(y_1)}$: the two curves cross, so neither dominates the other pointwise despite the first profiles being Blackwell ordered.

Nevertheless, all results on adaptive composition in this work are valid at the level of distribution pairs, and we abuse the notation $\otimes_{\text{adapt}}$ as a shorthand in the main text to better communicate the structure of these results.

B Counterexamples Against f-DP Filter Validity

We compute tensor products in terms of privacy profiles, as Definition A.12 offers a direct integral form. We can then convert between privacy profiles and trade-off functions using Theorem A.6. By Theorem A.7, the natural f-DP filter can be equivalently expressed using privacy profiles as

$$F(H_B, H_1, \ldots, H_t) = \begin{cases} \text{CONT} & \text{if } H_1 \otimes \ldots \otimes H_t \leq H_B, \\ \text{HALT} & \text{otherwise.} \end{cases} \tag{2}$$

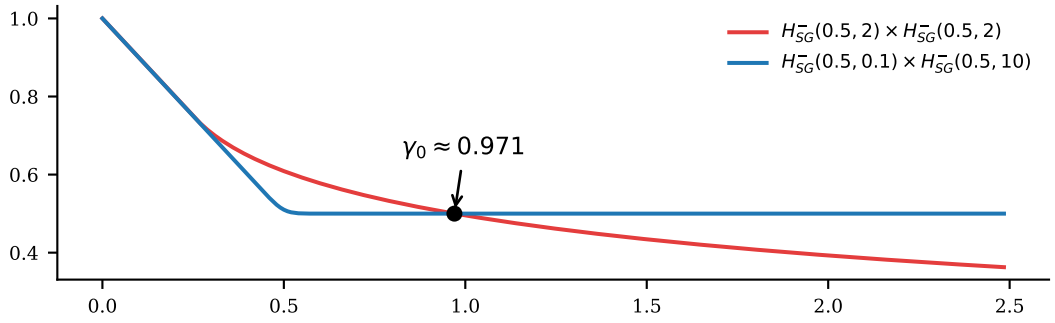


Figure 4: Tensor products of future privacy profiles, crossing at calibration point $\gamma_0 \approx 0.971$. These correspond to the trade-off functions in the left panel of Figure 1.

Motivated by Corollary A.21, we use the f-DP filter with a budget H_B^- for privacy profiles $H_1^-, \ldots, H_t^-$ under remove adjacency, since the validity of the filter would imply that the produced composition has a privacy profile $\mathrm{sym}\big(H_B^-\big)$.

B.1 Details of the Counterexample in Corollary 3.3

We work throughout with remove-adjacency privacy profiles, as motivated in the paragraph above. Let the datapoint space be $\mathcal{O} = \{X\}$, so the only remove pair is $S = (X)$ and $S^- = \emptyset$. Fix composition length $T = 3$, subsampling rate $q = 0.5$, and additive Gaussian noise $\sigma = 1$ for all mechanisms $\mathcal{M}_{1:3}$. Under input S^-, every mechanism outputs mean 0; under input S, the output means are

$$\mu_1 = 1.3, \qquad (\mu_2, \mu_3) = \begin{cases} (0.1,\ 10) & \text{if } y_1 \le 0.65, \\ (2,\ 2) & \text{if } y_1 > 0.65. \end{cases}$$

Note that $\mathcal{M}_3$ depends only on y_1, the output of $\mathcal{M}_1$, and not on y_2.

Budget construction. Denote by $H^-_{SG,q,\mu}$ the remove-adjacency privacy profile of the subsampled Gaussian mechanism with parameters (q, μ), converted from $SG^-_{q,\mu}$ as given by Theorem A.6. Set

$$H_{2\times 3,1} = H^-_{SG,0.5,0.1} \otimes H^-_{SG,0.5,10}, \qquad H_{2\times 3,2} = H^-_{SG,0.5,2} \otimes H^-_{SG,0.5,2},$$

which correspond to the future privacy profiles of $(\mathcal{M}_2, \mathcal{M}_3)$ conditional on $y_1 \le 0.65$ and $y_1 > 0.65$, respectively. These are illustrated in Figure 4; they cross at a calibration point $\gamma_0 \approx 0.971$, confirming there is no Blackwell order between them.

The tightest budget that permits release of y_2 and y_3 regardless of y_1 is, by (2),

$$H_{B,\text{tight}} = \max\Big(H^-_{SG,0.5,1.3} \otimes H_{2\times 3,1},\ H^-_{SG,0.5,1.3} \otimes H_{2\times 3,2}\Big),$$

which passes the filter check for each branch by construction.

True adaptive guarantee. Let p_t and q_t denote the output densities of $\mathcal{M}_t$ under inputs S and S^-, respectively. Using Definition A.12, the tight privacy profile of the fully adaptive composition is

$$\begin{aligned} H_{\text{adapt}}(\gamma) &= \int_{\mathbb{R}^3} \Big[p_1(y_1)\, p_2(y_2 \mid y_1)\, p_3(y_3 \mid y_1) - \gamma\, q_1(y_1)\, q_2(y_2 \mid y_1)\, q_3(y_3 \mid y_1)\Big]_+ \,\mathrm{d}y_1\,\mathrm{d}y_2\,\mathrm{d}y_3 \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}^2} \Big[p_2(y_2 \mid y_1)\, p_3(y_3 \mid y_1) - \gamma \tfrac{q_1(y_1)}{p_1(y_1)}\, q_2(y_2 \mid y_1)\, q_3(y_3 \mid y_1)\Big]_+ \,\mathrm{d}y_2\,\mathrm{d}y_3 \right) p_1(y_1)\,\mathrm{d}y_1 \\ &= \int_{-\infty}^{0.65} H_{2\times 3,1}\Big(\gamma \tfrac{q_1(y_1)}{p_1(y_1)}\Big)\, p_1(y_1)\,\mathrm{d}y_1 + \int_{0.65}^{\infty} H_{2\times 3,2}\Big(\gamma \tfrac{q_1(y_1)}{p_1(y_1)}\Big)\, p_1(y_1)\,\mathrm{d}y_1. \end{aligned}$$

All privacy profiles are computed as integrals numerically; the resulting $H_{B,\text{tight}}$ and H_{adapt} are compared in Figure 5.

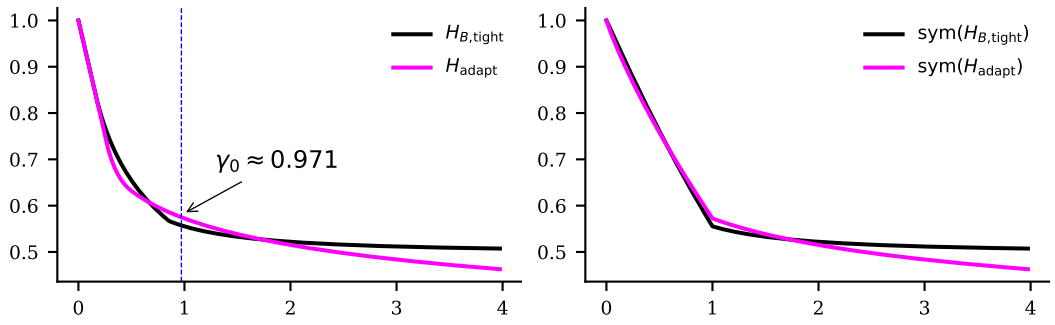


Figure 5: Left: H_{adapt} exceeds $H_{B,\text{tight}}$ in the region surrounding $\gamma_0 \approx 0.971$. Right: the violation persists after symmetrization. These privacy profiles correspond to panels (c) and (d) in Figures 1.

Why the failure occurs. We now show that $H_{\text{adapt}}(\gamma_0) > H_{B,\text{tight}}(\gamma_0)$, which establishes the invalidity claimed in Corollary 3.3.

The two tensor products underlying $H_{B,\text{tight}}$ each average a *fixed* future profile over all outputs y_1:

$$\big(H^{-}_{SG,0.5,1.3} \otimes H_{2\times 3,i}\big)(\gamma) = \int_{\mathbb{R}} H_{2\times 3,i}\Big(\gamma \tfrac{q_1(y_1)}{p_1(y_1)}\Big)\, p_1(y_1)\,\mathrm{d}y_1, \quad i = 1, 2.$$

By contrast, H_{adapt} is free to select a different profile per output. At $\gamma = \gamma_0$, the ratio $q_1(y_1)/p_1(y_1) \geq 1$ when $y_1 \leq 0.65$ and < 1 when $y_1 > 0.65$, so the rescaled argument $\gamma_0 q_1(y_1)/p_1(y_1)$ lies above γ_0 in the first y_1-region and below in the second. Because $H_{2\times 3,1} > H_{2\times 3,2}$ above γ_0 and $H_{2\times 3,2} > H_{2\times 3,1}$ below γ_0 (the calibration point), the adaptive integrand takes the larger profile in each region simultaneously:

$$H_{\text{adapt}}(\gamma_0) = \int_{-\infty}^{0.65} H_{2\times 3,1}\left(\gamma_0 \frac{q_1(y_1)}{p_1(y_1)}\right) p_1(y_1)\,\mathrm{d}y_1 + \int_{0.65}^{\infty} H_{2\times 3,2}\left(\gamma_0 \frac{q_1(y_1)}{p_1(y_1)}\right) p_1(y_1)\,\mathrm{d}y_1,$$

whereas each fixed-branch integral must use the *same* profile across both regions. It follows that

$$H_{\text{adapt}}(\gamma_0) > \max\Big[\big(H^{-}_{SG,0.5,1.3} \otimes H_{2\times 3,1}\big)(\gamma_0),\ \big(H^{-}_{SG,0.5,1.3} \otimes H_{2\times 3,2}\big)(\gamma_0)\Big] = H_{B,\text{tight}}(\gamma_0).$$

By continuity of privacy profiles, the strict inequality holds on a neighbourhood of γ_0. This confirms that the filter is violated, with the calibration point as the precise cause.

B.2 Proof of Adaptive Tightness of Standard Composition Results

We prove the result in terms of privacy profiles, then convert to trade-off functions via Theorem A.6.

Theorem B.1. *Let (P_1, Q_1) be a pair of distributions on $\mathcal{O}_1$ with privacy profile H_1. For each $y_1 \in \mathcal{O}_1$, let $(P_2^{(y_1)}, Q_2^{(y_1)})$ be freely chosen from a family of pairs of distributions on $\mathcal{O}_2$, and let $\mathcal{H}_2$ denote the family of all achievable privacy profiles at step 2. Set $H_2^{\uparrow} = \sup_{H_2 \in \mathcal{H}_2} H_2$ (pointwise). Then the adaptive product $(P_1 \times P_2^{(y_1)},\ Q_1 \times Q_2^{(y_1)})$ for any selection rule is dominated by $H_1 \otimes H_2^{\uparrow}$. Furthermore, this bound is tight: for every $\gamma \geq 0$ and $\eta > 0$, there exists a selection rule $y_1 \mapsto (P_2^{(y_1)}, Q_2^{(y_1)})$ such that*

$$\Big(H_1 \otimes H_2^{\uparrow}\Big)(\gamma) - \eta \ \leq\ D_\gamma\Big[P_1 \times P_2^{(y_1)} \,\Big\|\, Q_1 \times Q_2^{(y_1)}\Big] \ \leq\ \Big(H_1 \otimes H_2^{\uparrow}\Big)(\gamma).$$

Remark B.2. The tightness statement is pointwise in γ: a single selection rule need not bring the full privacy profile uniformly close to $H_1 \otimes H_2^{\uparrow}$.

Proof. The upper bound follows immediately from Property 2 of Proposition A.13.

For the lower tightness bound, fix $\gamma > 0$ and $\eta > 0$. Since $H_2^{\uparrow} = \sup_{H_2 \in \mathcal{H}_2} H_2$, for each $y_1 \in \mathcal{O}_1$ we may choose $H_2^{(y_1)} \in \mathcal{H}_2$ such that

$$H_2^{\uparrow}\left(\gamma \frac{q_1(y_1)}{p_1(y_1)}\right) - \eta \;<\; H_2^{(y_1)}\left(\gamma \frac{q_1(y_1)}{p_1(y_1)}\right) \;\le\; H_2^{\uparrow}\left(\gamma \frac{q_1(y_1)}{p_1(y_1)}\right).$$

Integrating against $p_1(y_1)\,\mathrm{d}y_1$ and applying Definition A.12 gives

$$\begin{aligned}\left(H_1 \otimes H_2^{\uparrow}\right)(\gamma) - \eta \;&\le\; \int_{\mathcal{O}_1} H_2^{(y_1)}\left(\gamma \frac{q_1(y_1)}{p_1(y_1)}\right) p_1(y_1)\,\mathrm{d}y_1 \\ &= D_\gamma\left[P_1 \times P_2^{(y_1)} \,\middle\|\, Q_1 \times Q_2^{(y_1)}\right] \;\le\; \left(H_1 \otimes H_2^{\uparrow}\right)(\gamma).\end{aligned}$$

□

The trade-off function counterpart follows by translating between privacy profiles and trade-off functions.

Corollary B.3. *Under the same setup, let $f_2^{\downarrow} = \mathrm{ce}\inf_{f_2 \in \mathcal{F}_2} f_2$. Then the adaptive product is lower bounded by $f_1 \otimes f_2^{\downarrow}$, and for every $\alpha \in [0,1]$ and $\eta > 0$ there exists a selection rule such that*

$$\left(f_1 \otimes f_2^{\downarrow}\right)(\alpha) \;\le\; T\left[P_1 \times P_2^{(y_1)} \,\middle\|\, Q_1 \times Q_2^{(y_1)}\right](\alpha) \;\le\; \left(f_1 \otimes f_2^{\downarrow}\right)(\alpha) + \eta.$$

Proof. Combine Theorem B.1 with the correspondence in Theorem A.6, using the fact that $H_2^{\uparrow}$ in the privacy profile formulation corresponds to $f_2^{\downarrow}$ in the trade-off function formulation under Theorem A.7. □

Connection to the counterexample. Corollary B.3 implies that the f-DP filter can fail whenever $H_1 \otimes H_2^{\uparrow} > H_B$ at some point, since the adaptive composition can approach $H_1 \otimes H_2^{\uparrow}$ arbitrarily closely there.

Corollary B.4. *In fully adaptive composition of up to 2 steps,*[3] *a counterexample to filter validity exists whenever there is some $\gamma_0 > 0$ such that $\left(H_1 \otimes H_2^{\uparrow}\right)(\gamma_0) > H_B(\gamma_0)$.*

The proof of Theorem B.1 also makes explicit how calibration points lead to filter failure: when $H_{2,1}, H_{2,2} \in \mathcal{H}_2$ cross at γ_0, the pointwise maximum $H_2^{\uparrow} = \max(H_{2,1}, H_{2,2})$ satisfies

$$\left(H_1 \otimes \max(H_{2,1}, H_{2,2})\right)(\gamma_0) \;>\; \max\left(\left(H_1 \otimes H_{2,1}\right)(\gamma_0),\, \left(H_1 \otimes H_{2,2}\right)(\gamma_0)\right),$$

and the extremizing selection rule is simply $y_1 \mapsto \arg\max_{H_2 \in \{H_{2,1}, H_{2,2}\}} H_2\left(\gamma_0 \frac{q_1(y_1)}{p_1(y_1)}\right)$. This is precisely the mechanism used in Appendix B.1, where $H_{2\times 3}^{\uparrow} = \max(H_{2\times 3,1}, H_{2\times 3,2})$. As shown in Figure 6, H_{adapt} touches $H_1 \otimes H_{2\times 3}^{\uparrow}$ at $\gamma_0 \approx 0.971$ but not everywhere, consistent with the pointwise nature of the tightness result.

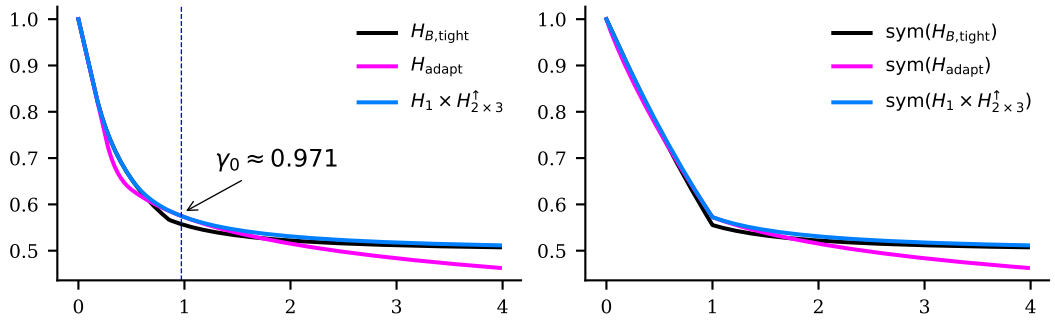


Figure 6: $H_1 \otimes H_{2\times 3}^{\uparrow}$ compared to H_{adapt} and $H_{B,\mathrm{tight}}$, illustrating Corollary B.4. The adaptive composition reaches $H_1 \otimes H_{2\times 3}^{\uparrow}$ at the calibration point γ_0 but not globally.

[3] If y_1 causes the filter to halt, we set $H_2^{(y_1)}(\gamma) = H_{\mathrm{Id}}(\gamma) = [1-\gamma]_+$.

C Convergence of Privacy Profiles/Trade-off Functions

C.1 Existence of Convergent Sequence in Blackwell Chains

We can impose the structural assumption of Blackwell chain on privacy profiles, in a similar way to trade-off functions.

Definition C.1. A family $\mathcal{H}$ of privacy profiles is a Blackwell chain if any pair $H_1, H_2 \in \mathcal{H}$ is Blackwell ordered, i.e. $H_1 \leq H_2$ or $H_1 \geq H_2$.

By Theorem A.7, $\mathcal{H}$ is a Blackwell chain if and only if its corresponding family $\mathcal{F}$ of trade-off functions is a Blackwell chain.

Lemma C.2. *In a Blackwell chain $\mathcal{H}$ of privacy profiles with the point-wise supremum $H^{\uparrow}$, there exists an everywhere increasing sequence $(H_n)_{n\in\mathbb{N}}$ such that for all $n \in \mathbb{N}$,*

$$H^{\uparrow}\left(\frac{i}{2^n}\right) - \frac{1}{n} < H_n\left(\frac{i}{2^n}\right) \leq H^{\uparrow}\left(\frac{i}{2^n}\right) \text{ for } i = 1, 2, \ldots, n2^n. \tag{3}$$

Lemma C.2 enables inequalities at finitely many points covering regions $[0, n]$ of increasing sizes n and increasing resolutions $\frac{1}{2^n}$. As $\cup_{n\in\mathbb{N}}[0, n] = [0, \infty)$ and $\frac{1}{2^n} \searrow 0$, these inequalities serve to connect point-wise properties to set-wise properties.

Proof. Since $H^{\uparrow}(\gamma) = \sup_{H\in\mathcal{H}} H(\gamma)$, for all $n \in \mathbb{N}$, for all $i = 1, 2, \ldots, n2^n$, there exists $H_{ni} \in \mathcal{H}$ such that at $\gamma = \frac{i}{2^n}$,

$$H^{\uparrow}\left(\frac{i}{2^n}\right) - \frac{1}{n} < H_{ni}\left(\frac{i}{2^n}\right) \leq H^{\uparrow}\left(\frac{i}{2^n}\right).$$

Let $\hat{H}_n = \max_{1\leq i\leq n2^n} H_{ni}$, then $\hat{H}_n$ satisfies the required inequality (3), and $\hat{H}_n$ is in $\mathcal{H}$ since $\mathcal{H}$ is a Blackwell chain. Finally, since $\left(\frac{i}{2^n}\right)_{i=1}^{n2^n} \subset \left(\frac{i}{2^{n+1}}\right)_{i=1}^{(n+1)2^{n+1}}$, we choose $H_n = \max_{k\leq n} \hat{H}_k$ so that $(H_n)_{n\in\mathbb{N}}$ is increasing while maintaining inequality (3). □

Lemma C.3. *In a Blackwell chain of privacy profiles $\mathcal{H}$ with a smallest upper bound $H^{\uparrow} = \sup_{H\in\mathcal{H}} H$, there exists an everywhere increasing sequence $(H_n)_{n\in\mathbb{N}}$ such that $H_n \nearrow_{n\to\infty} H^{\uparrow}$ point-wise.*

Proof. Apply Lemma C.2 to find an everywhere increasing sequence H_n such that for all $n \in \mathbb{N}$, for $i = 1, 2, \ldots, n2^n$,

$$H^{\uparrow}\left(\frac{i}{2^n}\right) - \frac{1}{n} < H_n\left(\frac{i}{2^n}\right) \leq H^{\uparrow}\left(\frac{i}{2^n}\right).$$

We show that this sequence converges point-wise to $H^{\uparrow}$. Fix $\gamma > 0$, then $n \geq \gamma$ once n is big enough, say, $n \geq N$. If γ has a finite binary representation, then once n is big enough,

$$H^{\uparrow}(\gamma) \geq H_n(\gamma) > H^{\uparrow}(\gamma) - \frac{1}{n} \nearrow_{n\to\infty} H^{\uparrow}(\gamma).$$

Otherwise if γ has an infinite binary representation, let $(\gamma_n)_{n\geq N}$ be a decreasing sequence that converges to γ, where each γ_n has a finite binary representation and $H_n(\gamma_n) > H^{\uparrow}(\gamma_n) - \frac{1}{n}$. Then

$$H^{\uparrow}(\gamma) \geq H_n(\gamma) \geq H_n(x_n) > H^{\uparrow}(\gamma_n) - \frac{1}{n} \nearrow_{n\to\infty} H^{\uparrow}(\gamma).$$

Here, we used the fact that $H^{\uparrow}(\gamma_n) \nearrow_{n\to\infty} H^{\uparrow}(\gamma)$, which is due to the continuity of $H^{\uparrow}$. □

C.2 Unifying Various Notions of Convergence

In addition to the point-wise convergence shown in Lemma C.3, we discuss other notions of convergence.

1. Uniform convergence $\sup_{\gamma\geq 0}|H_n(\gamma)-H(\gamma)| \xrightarrow{n\to\infty} 0$, much stronger than point-wise convergence.
2. Uniform convergence over any compact set: $\sup_{\gamma\in[0,\gamma_{\max}]}|H_n(\gamma)-H(\gamma)| \xrightarrow{n\to\infty} 0$ for all $\gamma_{\max}>0$.
3. Convergence in the symmetrized Δ-divergence introduced by Kaissis et al. (2024):
$$\sup_{\gamma\geq 0}\frac{|H_n(\gamma)-H(\gamma)|}{1+\gamma} \xrightarrow{n\to\infty} 0$$

Proposition C.4. *Uniform convergence implies convergence in Δ-divergence.*

Proof. The Δ-divergence is smaller or equal to the $\mathcal{L}^\infty$ norm,
$$\sup_{\gamma\geq 0}\frac{|H(\gamma)-H_n(\gamma)|}{1+\gamma} \leq \sup_{\gamma\geq 0}|H(\gamma)-H_n(\gamma)|.$$
□

The following example shows that the implication in Proposition C.4 is strictly one-sided.

Example C.5. For $n\in\mathbb{N}$, let
$$H_n(\gamma)=\begin{cases}1-\frac{\gamma}{n} & \text{if } \gamma\leq n,\\ 0 & \text{otherwise.}\end{cases}$$
Then H_n converges to $H=1$ point-wise but not uniformly. A closer examination reveals that for all n, $\frac{H(\gamma)-H_n(\gamma)}{1+\gamma}$ reaches its maximum at $\gamma=n$. Therefore
$$\max_{\gamma\geq 0}\frac{H(\gamma)-H_n(\gamma)}{1+\gamma}=\frac{1}{1+n} \xrightarrow{n\to\infty} 0,$$
i.e. H_n converges to H in Δ-divergence.

Theorem C.6. *The following notions of convergence are equivalent for a sequence of privacy profiles H_n that converges to a privacy profile H.*

1. *Convergence in Δ-divergence.*
2. *Point-wise convergence.*
3. *Uniform convergence over any compact set.*

Proof. *1 $\implies$ 2*. At any order $\gamma_0>0$,
$$|H_n(\gamma_0)-H(\gamma_0)|=(1+\gamma_0)\frac{|H_n(\gamma_0)-H(\gamma_0)|}{1+\gamma_0}\leq(1+\gamma_0)\sup_{\gamma\geq 0}\frac{|H_n(\gamma)-H(\gamma)|}{1+\gamma} \xrightarrow{n\to\infty} 0.$$

2 $\implies$ 3. Since H_n is decreasing for all n and H is continuous, uniform convergence over any compact set follows from Pólya's theorem, which is proven e.g. as Lemma A.7 by Dong et al. (2022).

3 $\implies$ 1. Fix $\epsilon>0$. We find N such that for all $n>N$,
$$\sup_{\gamma\geq 0}\frac{|H_n(\gamma)-H(\gamma)|}{1+\gamma}<\epsilon.$$
Choose $\gamma_{\max}=\frac{1}{\epsilon}-1$ so that $\frac{1}{1+\gamma_{\max}}=\epsilon$. Note that H_n converges uniformly to H over $\gamma\in[0,\gamma_{\max}]$, which implies that there exists N such that for all $n>N$,
$$|H_n(\gamma)-H(\gamma)|<\epsilon \text{ for all } \gamma\in[0,\gamma_{\max}].$$
Therefore, for all $n>N$,
$$\begin{aligned}\sup_{\gamma\geq 0}\frac{|H_n(\gamma)-H(\gamma)|}{1+\gamma}&=\max\left[\sup_{\gamma\in[0,\gamma_{\max}]}\frac{|H_n(\gamma)-H(\gamma)|}{1+\gamma},\sup_{\gamma>\gamma_{\max}}\frac{|H_n(\gamma)-H(\gamma)|}{1+\gamma}\right]\\&\leq\max\left(\epsilon,\frac{1}{1+\gamma_{\max}}\right)=\epsilon.\end{aligned}$$
□

Corollary C.7. *In a Blackwell chain $\mathcal{H}$ of privacy profiles with a smallest upper bound $H^\uparrow = \sup_{H\in\mathcal{H}} H$, there exists an everywhere increasing sequence $(H_n)_{n\in\mathbb{N}}$ such that $H_n \nearrow^{n\to\infty} H^\uparrow$ under any notion of convergence in Theorem C.6.*

Proof. Combine Lemma C.3 with Theorem C.6. □

Theorem C.8. *The following notions of convergence are equivalent for a sequence of trade-off functions f_n that converges to a trade-off function f.*

1. *Convergence in Δ-divergence Kaissis et al. (2024):*
$$\min\left\{\Delta \geq 0 : \forall\alpha \in [0, 1-\Delta], \begin{array}{l} f_n(\alpha+\Delta)-\Delta \leq f(\alpha) \\ f(\alpha+\Delta)-\Delta \leq f_n(\alpha)\end{array}\right\} \xrightarrow{n\to\infty} 0.$$
2. *Point-wise convergence over $(0, 1]$.*
3. *Uniform convergence over $[\alpha_{\min}, 1]$ for all $\alpha_{\min} \in (0, 1]$.*

Proof. 1 $\implies$ 2. f_n converges to f in Δ-divergence if and only if there exists a decreasing sequence of positive numbers $\Delta_n > 0$ such that $\Delta_n \searrow^{n\to\infty} 0$ and for all n,
$$f(\alpha+\Delta_n)-\Delta_n \leq f_n(\alpha) \text{ for all } \alpha \in [0, 1-\Delta_n],$$
$$f_n(\alpha) \leq f(\alpha-\Delta_n)+\Delta_n \text{ for all } \alpha \in [\Delta_n, 1].$$
Fix $\alpha \in (0, 1)$. Then once n is large enough, say, $n > N$, $\Delta_n < \alpha < 1-\Delta_n$ since $\Delta_n \searrow^{n\to\infty} 0$. Therefore, for all $n > N$,
$$f(\alpha+\Delta_n)-\Delta_n \leq f_n(\alpha) \leq f(\alpha-\Delta_n)+\Delta_n.$$
Since $\lim_{n\to\infty} f(\alpha+\Delta_n)-\Delta_n = \lim_{n\to\infty} f(\alpha-\Delta_n)+\Delta_n = f(\alpha)$ due to the continuity of f, $\lim_{n\to\infty} f_n(\alpha) = f(\alpha)$.

2 $\implies$ 3. Since f_n is decreasing for all n and f is continuous, uniform convergence over $[\alpha_{\min}, 1]$ for all $\alpha_{\min} \in (0, 1]$ follows from Pólya's theorem, which is proven e.g. as Lemma A.7 by Dong et al. (2022).

3 $\implies$ 1. Fix $\Delta > 0$. Since f_n converges to f uniformly over $[\Delta, 1]$, there exists some N such that for all $n > N$,
$$-\Delta \leq f_n(\alpha) - f(\alpha) \leq \Delta \text{ for all } \alpha \in [\Delta, 1].$$
The first and second inequalities imply, respectively,
$$f_n(\alpha) \leq f(\alpha)+\Delta \leq f(\alpha-\Delta)+\Delta,$$
$$f(\alpha) \leq f_n(\alpha)+\Delta \leq f_n(\alpha-\Delta)+\Delta$$
for all $\alpha \in [\Delta, 1]$. Since this holds for all $\Delta > 0$, f_n converges to f in Δ-divergence. □

Corollary C.9. *In a Blackwell chain $\mathcal{F}$ of trade-off functions with a greatest lower bound $f^\downarrow = \text{ce}\inf_{f\in\mathcal{F}} f$, there exists an everywhere decreasing sequence $(f_n)_{n\in\mathbb{N}}$ such that $f_n \searrow^{n\to\infty} f^\downarrow$ under any notion of convergence in Theorem C.8.*

Proof. Denote $\mathcal{H}$ the corresponding family of privacy profiles obtained by converting the trade-off functions in $\mathcal{F}$ using Theorem A.6. Since $\mathcal{F}$ is a Blackwell chain, $\mathcal{H}$ is also a Blackwell chain by Theorem A.7. By Corollary C.7, there exists an everywhere increasing sequence H_n that converges to $H^\uparrow = \sup_{H\in\mathcal{H}} H$ in Δ-divergence.

By Theorem A.7, the sequence of trade-off functions f_n that corresponds to H_n is everywhere decreasing; and furthermore, the smallest upper bound $H^\uparrow$ of privacy profiles in $\mathcal{H}$ corresponds to the greatest lower bound $f^\downarrow$ of trade-off functions in $\mathcal{F}$. Kaissis et al. (2024) showed that Δ-divergence can be converted between privacy profiles and trade-off functions, and hence f_n converges to $f^\downarrow$ in Δ-divergence. By Theorem C.8, f_n converges to $f^\downarrow$ under all three notions of convergence. □

D Proofs for f-DP Filter Validity via Blackwell Chains

D.1 Blackwell Chain Property Implies f-DP Filter Validity

We first restrict ourselves to the simplest case with composition length $T = 2$ as a fundamental result before extending it to the general case of fully adaptive composition of up to $T \geq 3$. The following Proposition D.1 is a direct consequence of Theorem 2.7.

Proposition D.1. *Let $\mathcal{M}_1$ be f_1-DP, and let the adaptively chosen $\mathcal{M}_2$ given output y_1 from $\mathcal{M}_1$ be $f_2^{(y_1)}$-DP. If there exists a trade-off function $f_2^{\downarrow}$ such that*

1. *$f_2^{(y_1)} \geq f_2^{\downarrow}$ for all $y_1 \in \mathcal{O}_1$,*
2. *$f_1 \otimes f_2^{\downarrow} \geq f_B$,*

then the composition $(\mathcal{M}_1, \mathcal{M}_2)$ is f_B-DP.

Among the lower bounds of the family of options $f_2^{(y_1)} \in \mathcal{F}$, i.e. the trade-off functions that satisfy Condition 1 above, it is enough to verify whether the greatest lower bound $f_2^{\downarrow} = \operatorname{ce}\inf_{f_2 \in \mathcal{F}_2} f_2$ satisfies Condition 2, as otherwise no other lower bound does.

Figure 6 shows that calibration points cause Condition 2 to fail. As shown in Appendix B, a diverse family of mechanisms exhibiting a calibration point γ_0 in their privacy profiles may cause $\left(H_1 \otimes H_2^{\uparrow}\right)(\gamma_0) > H_B(\gamma_0)$, or equivalently $\left(f_1 \otimes f_2^{\downarrow}\right)(\alpha_0) < f_B(\alpha_0)$ at the corresponding α_0. This suggests that the Blackwell chain structural assumption is necessary for both conditions in Proposition D.1 to hold.

Theorem D.2. *Let $\mathcal{M}_1$ be f_1-DP, and let the adaptively chosen $\mathcal{M}_2$ given output y_1 from $\mathcal{M}_1$ be $f_2^{(y_1)}$-DP. If*

1. *$f_1 \otimes f_2^{(y_1)} \geq f_B$ for all $y_1 \in \mathcal{O}_1$,*
2. *$\mathcal{F} = \left\{f_2^{(y_1)}\right\}_{y_1 \in \mathcal{O}_1}$ is a Blackwell chain,*

then the greatest lower bound $f_2^{\downarrow} = \operatorname{ce}\inf_{y_1 \in \mathcal{O}_1} f_2^{(y_1)}$ satisfies $f_1 \otimes f_2^{\downarrow} \geq f_B$. As a consequence, the composition $(\mathcal{M}_1, \mathcal{M}_2)$ is f_B-DP by Proposition D.1.

This can be equivalently formulated using privacy profiles. Let the privacy profile of $\mathcal{M}_1$ be H_1, and let the privacy profile of the adaptively chosen $\mathcal{M}_2$ given output y_1 from $\mathcal{M}_1$ be $H_2^{(y_1)}$. If

1. *$H_1 \otimes H_2^{(y_1)} \leq H_B$ for all $y_1 \in \mathcal{O}_1$,*
2. *$\mathcal{H} = \left\{H_2^{(y_1)}\right\}_{y_1 \in \mathcal{O}_1}$ is a Blackwell chain,*

then the smallest upper bound $H_2^{\uparrow} = \sup_{y_1 \in \mathcal{O}_1} H_2^{(y_1)}$ satisfies $H_1 \otimes H_2^{\uparrow} \leq H_B$.

Proof. Let (P_1, Q_1) be any pair of distributions that has a privacy profile H_1. For a fixed $\hat{y}_1$, $H_1 \otimes H_2^{(\hat{y}_1)} \leq H_B$ can be written explicitly as

$$\int_{\mathcal{O}_1} H_2^{(\hat{y}_1)}\left(\gamma \frac{q_1(y_1)}{p_1(y_1)}\right) p_1(y_1)\,\mathrm{d}y_1 \leq H_B(\gamma) \text{ for all } \gamma \geq 0.$$

While $H_1 \otimes H_2^{\uparrow} \leq H_B$ can be written explicitly as

$$\int_{\mathcal{O}_1} H_2^{\uparrow}\left(\gamma \frac{q_1(y_1)}{p_1(y_1)}\right) p_1(y_1)\,\mathrm{d}y_1 \leq H_B(\gamma) \text{ for all } \gamma \geq 0.$$

Lemma C.3 states that there exists an everywhere increasing sequence $\left(H_2^{(\hat{y}_{1,n})}\right)_{n\in\mathbb{N}}$ in $\mathcal{H}_2$ such that $H_2^{(\hat{y}_{1,n})} \nearrow_{n\to\infty} H_2^{\uparrow}$ point-wise. By the monotone convergence theorem, for any $\gamma \geq 0$,

$$\left(H_1 \otimes H_2^{\uparrow}\right)(\gamma) = \lim_{n\to\infty} \left(H_1 \otimes H_2^{\hat{y}_{1,n}}\right)(\gamma) \leq H_B(\gamma)\,.$$

□

In fact, Theorem D.2 has already covered the case $T = 2$ in Theorem 3.5, as we can equivalently choose $f_2^{(y_1)}(\alpha) = \mathrm{Id}(\alpha) = 1 - \alpha$ when y_1 halts the filter early. We now give the remaining proof to Theorem 3.5.

Theorem 3.5. *Let $\mathcal{F}$ be the family of trade-off functions used as per-step f-DP guarantees. The natural f-DP filter is valid for $\mathcal{F}$ in fully adaptive composition of up to $T \geq 2$ steps if $\mathcal{F}^{(-\otimes\mathcal{F})(T-2)}$ is a Blackwell chain.*

Proof. The case $T = 2$ has been proven as Theorem D.2. We present a proof for the case of $T = 3$ for clarity, where $\mathcal{F}^{(-\otimes\mathcal{F})(T-2)} = \mathcal{F}^{-\otimes\mathcal{F}}$. The general case $T > 3$ follows by iteratively applying the same argument.

Without loss of generality, when the previous outputs halt the filter early, we can set the f-DP guarantees for remaining mechanisms as the perfectly private trade-off function $\mathrm{Id}(\alpha) = 1 - \alpha$. Hence, the filter inequality becomes, for all outputs $y_1 \in \mathcal{O}_1$ and $y_2 \in \mathcal{O}_2$,

$$f_1 \otimes f_2^{(y_1)} \otimes f_3^{(y_1,y_2)} \geq f_B,$$

where f_t is the f-DP guarantee of $\mathcal{M}_t$ for $t = 1, 2, 3$.

Fix $y_1 \in \mathcal{O}_1$. Then $f_1 \otimes f_2^{(y_1)}$ is fixed and $f_3^{(y_1,y_2)}$ only depends on $y_2 \in \mathcal{O}_2$. Since $\left\{f_3^{(y_1,y_2)}\right\}_{y_2\in\mathcal{O}_2} \subset \mathcal{F} \subset \mathcal{F}^{-\otimes\mathcal{F}}$, $\left\{f_3^{(y_1,y_2)}\right\}_{y_2\in\mathcal{O}_2}$ is a Blackwell chain, and thus by Theorem D.2,

$$f_1 \otimes f_2^{(y_1)} \otimes f_3^{(y_1,\downarrow)} \geq f_B,$$

where $f_3^{(y_1,\downarrow)} = \mathrm{ce}\inf_{y_2\in\mathcal{O}_2} f_3^{(y_1,y_2)}$ is the greatest lower bound of $\left\{f_3^{(y_1,y_2)}\right\}_{y_2\in\mathcal{O}_2}$.

The deduction so far holds for all $y_1 \in \mathcal{O}_1$. Now, when we let $y_1 \in \mathcal{O}_1$ change as a variable, f_1 is fixed and $f_2^{(y_1)} \otimes f_3^{(y_1,\downarrow)}$ depends on y_1. Since $\left\{f_2^{(y_1)} \otimes f_3^{(y_1,\downarrow)}\right\}_{y_1\in\mathcal{O}_1} \subset \mathcal{F}^{-\otimes\mathcal{F}}$, $\left\{f_2^{(y_1)} \otimes f_3^{(y_1,\downarrow)}\right\}_{y_1\in\mathcal{O}_1}$ is a Blackwell chain, and consequently, by Theorem D.2 again,

$$f_1 \otimes f_{2\times 3}^{\downarrow\downarrow} \geq f_B,$$

where $f_{2\times 3}^{\downarrow\downarrow} = \mathrm{ce}\inf_{y_1\in\mathcal{O}_1} f_2^{(y_1)} \otimes f_3^{(y_1,\downarrow)}$ is the greatest lower bound of $\left\{f_2^{(y_1)} \otimes f_3^{(y_1,\downarrow)}\right\}_{y_1\in\mathcal{O}_1}$.

Conditioned on a fixed $y_1 \in \mathcal{O}_1$ as first output, $\mathcal{M}_2^{(y_1)}$ is $f_2^{(y_1)}$-DP, and for all $y_2 \in \mathcal{O}_2$, $\mathcal{M}_3^{(y_1,y_2)}$ is $f_3^{(y_1,y_2)}$-DP and hence $f_3^{(y_1,\downarrow)}$-DP. Therefore, by Theorem 2.7, the composition $\left(\mathcal{M}_2^{(y_1)}, \mathcal{M}_3^{(y_1,y_2)}\right)$ is $f_2^{(y_1)} \otimes f_3^{(y_1,\downarrow)}$-DP for all y_2.

It now follows that for all $y_1 \in \mathcal{O}_1$ and all $y_2 \in \mathcal{O}_2$, $\left(\mathcal{M}_2^{(y_1)}, \mathcal{M}_3^{(y_1,y_2)}\right)$ is $f_{2\times 3}^{\downarrow\downarrow}$-DP. Since $\mathcal{M}_1$ is f_1-DP, by Theorem 2.7, $(\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3)$ is $f_1 \otimes f_{2\times 3}^{\downarrow\downarrow}$-DP and hence f_B-DP. □

The f-DP central limit theorem Dong et al. (2022) states that a composition of trade-off functions $f_1 \otimes \ldots \otimes f_n$ converges to some Gaussian trade-off function in Δ-divergence as $n \to \infty$. Consequently, as T increases, the trade-off functions in $\mathcal{F}^{(-\otimes\mathcal{F})(T-2)}$ increasingly resemble Gaussian trade-off functions.

Example D.3. Let f be a trade-off function and consider the n-fold self-compositions $\mathcal{F} = (f^{\otimes n})_{n\in\mathbb{N}\cup\{\infty\}}$, in which $f^{\otimes\infty} = 0$ if f is not Id, and $\mathrm{Id}^{\otimes\infty} = \mathrm{Id}$. In this case, $\mathcal{F}^{(-\otimes\mathcal{F})(T-2)} = (f^{\otimes n})_{n\in\mathbb{N}\cup\{\infty\}}$ for all $T \geq 2$, which is a Blackwell chain by Proposition A.16. By Theorem 3.5, the f-DP filter is valid for mechanisms with tight f-DP guarantees matching $f^{\otimes n}$ for some n. A construction of such mechanisms when f is symmetric is given by the canonical noise distribution Awan and Vadhan (2021).

D.2 Failure of Blackwell Chain Property Implies f-DP Filter Invalidity

Definition D.4. For a family $\mathcal{H}$ of privacy profiles, define

$$\bar{\mathcal{H}} = \left\{\bar{H} : \exists H_n \in \mathcal{H} \text{ s.t. } H_n \nearrow \bar{H}\right\}$$

as the family of limits $\bar{H}$ of some everywhere increasing sequence $H_n \in \mathcal{H}$ under all notions of convergence in Theorem C.6. Furthermore, define

$$\mathcal{H}^{-\otimes\mathcal{H}} = \mathcal{H} \otimes \bar{\mathcal{H}} = \left\{H \otimes \bar{H} : H \in \mathcal{H}, \bar{H} \in \bar{\mathcal{H}}\right\}.$$

Lemma D.5. *Let $\mathcal{H}$ be an arbitrary family of privacy profiles, and consider $\mathcal{H}^{(-\otimes\mathcal{H})(T-2)}$ for some $T \geq 2$. For all $H^* \in \mathcal{H}^{(-\otimes\mathcal{H})(T-2)}$, for all $\gamma_{\max} > 0$ and all $\eta > 0$, there exist $T-1$ privacy profiles $H_2, H_3, \ldots, H_T \in \mathcal{H}$ such that*

$$\sup_{\gamma\in[0,\gamma_{\max}]} |(H_2 \otimes H_3 \otimes \ldots \otimes H_T)(\gamma) - H^*(\gamma)| \leq \eta.$$

Proof. For clarity, we present the proof for the case $T = 4$. The general case involves iteratively applying the same kind of argument here.

Now $H^* \in \left(\mathcal{H}^{-\otimes\mathcal{H}}\right)^{-\otimes\mathcal{H}}$. There exists $H_2 \in \mathcal{H}$ and $\overline{H_{3\times4}} \in \overline{\mathcal{H}^{-\otimes\mathcal{H}}}$ such that $H_2 \otimes \overline{H_{3\times4}} = H^*$. Furthermore, there exists an everywhere increasing sequence $H_{3\times4,n} \in \mathcal{H}^{-\otimes\mathcal{H}}$ such that $H_{3\times4,n} \nearrow_{n\to\infty} \overline{H_{3\times4}}$ under all notions of Theorem C.6. By the monotone convergence theorem, $H_2 \otimes H_{3\times4,n} \nearrow_{n\to\infty} H_2 \otimes \overline{H_{3\times4}}$ point-wise, and hence uniformly over any compact set by Theorem C.6. Therefore, there exists n_0 such that

$$\sup_{\gamma\in[0,\gamma_{\max}]} \left|(H_2 \otimes H_{3\times4,n_0})(\gamma) - \left(H_2 \otimes \overline{H_{3\times4}}\right)(\gamma)\right| \leq \frac{\eta}{2}.$$

Now $H_{3\times4,n_0} \in \mathcal{H}^{-\otimes\mathcal{H}}$. There exists $H_3 \in \mathcal{H}$ and $\bar{H}_4 \in \bar{\mathcal{H}}$ such that $H_3 \otimes \bar{H}_4 = H_{3\times4,n_0}$. Furthermore, there exists an everywhere increasing sequence $H_{4,k} \in \mathcal{H}$ such that $H_{4,k} \nearrow_{k\to\infty} \bar{H}_4$ under all notions of Theorem C.6. By the monotone convergence theorem, $H_2 \otimes H_3 \otimes H_{4,k} \nearrow_{k\to\infty} H_2 \otimes H_3 \otimes \bar{H}_4$ point-wise, and hence uniformly over any compact set by Theorem C.6. Therefore, there exists k_0 such that

$$\sup_{\gamma\in[0,\gamma_{\max}]} \left|(H_2 \otimes H_3 \otimes H_{4,k_0})(\gamma) - \left(H_2 \otimes H_3 \otimes \bar{H}_4\right)(\gamma)\right| \leq \frac{\eta}{2}.$$

Since $H_2 \otimes H_3 \otimes \bar{H}_4 = H_2 \otimes H_{3\times4,n_0}$ and $H_2 \otimes \overline{H_{3\times4}} = H^*$, by the triangle inequality,

$$\sup_{\gamma\in[0,\gamma_{\max}]} |(H_2 \otimes H_3 \otimes H_{4,k_0})(\gamma) - H^*(\gamma)| \leq \frac{\eta}{2} + \frac{\eta}{2} = \eta.$$

□

Theorem 3.6. *Let $\mathcal{F}$ be the family of trade-offs used as per-step f-DP guarantees. Suppose that:*

1. *Some $f_1 \in \mathcal{F}$ is differentiable on $(\alpha_1, \alpha_2) \subset (0, 1)$ and has a nonconstant derivative there;*
2. *$\mathcal{F}^{(-\otimes\mathcal{F})(T-2)}$ contains two trade-off functions f_{21} and f_{22} that cross at some FPR $\alpha_0 \in (0, 1)$ with different slopes $f'_{21}(\alpha_0) \neq f'_{22}(\alpha_0)$.*

then the natural f-DP filter is invalid for $\mathcal{F}$: there exists a fully adaptive composition of at most T steps, defined on a single neighboring input pair whose trade-off function at each step belongs to $\mathcal{F}$, that passes every filter check but violates its prescribed budget.[4]

[4]We thank Matthew Regehr for pointing out a gap in an earlier version of Theorem 3.6.

Proof. Let $\mathcal{H}$ be the family of privacy profiles corresponding to $\mathcal{F}$, and let H_1^*, H_2^* be the pair of profiles corresponding to $f_{21}, f_{22} \in \mathcal{F}^{(-\otimes\mathcal{F})(T-2)}$. Since f_{21}, f_{22} cross at α_0 with different slopes $f'_{21}(\alpha_0)$ and $f'_{22}(\alpha_0)$, by Theorem A.6, H_1^*, H_2^* cross and exchange order locally at a calibration point γ_0 lying between $-1/f'_{21}(\alpha_0)$ and $-1/f'_{22}(\alpha_0)$.

Let P_1, Q_1 be any distribution pair with trade-off function f_1. From the assumptions on f_1, we show that $\left(q_1(y_1)/p_1(y_1)\right)_{y_1 \in \mathcal{O}_1}$ contains some interval in $\mathbb{R}$. By the Neyman-Pearson lemma, each pair of error rates $(\alpha, f_1(\alpha))$ is attained by some likelihood ratio test that uses q_1/p_1. The decrease in FNR per unit increase in FPR equals the likelihood-ratio threshold, which is therefore $-f_1'(\alpha)$. Formally, for any $0 < r_1 < r_2$,

$$\begin{aligned} P_1\left(r_1 < \frac{q_1}{p_1} < r_2\right) &= \text{length}\{\alpha \in (0,1) : r_1 < -f_1'(\alpha) < r_2\} \\ &\geq \text{length}\{\alpha \in (\alpha_1, \alpha_2) : r_1 < -f_1'(\alpha) < r_2\}. \end{aligned}$$

Over $\alpha \in (\alpha_1, \alpha_2)$, since f_1 is convex and differentiable, $-f_1'$ is continuous. Together with the assumption that $-f_1'$ is not constant, this implies that $\left(q_1(y_1)/p_1(y_1)\right)_{y_1 \in \mathcal{O}_1}$ contains some interval in $\mathbb{R}$.

As H_1^*, H_2^* cross at γ_0, we can find γ_{scale} such that γ_0 lies in the interior of $\left(\gamma_{\text{scale}} q_1(y_1)/p_1(y_1)\right)_{y_1 \in \mathcal{O}_1}$. Consequently,

$$\left(\gamma_{\text{scale}} \frac{q_1(y_1)}{p_1(y_1)}\right)_{y_1 \in \mathcal{O}_1} \text{ overlaps with both } \{H_1^* > H_2^*\} \text{ and } \{H_1^* < H_2^*\}.$$

We apply Lemma D.5 with small $\eta > 0$ to find $H_{2,1}, H_{3,1}, \ldots, H_{T,1}$ and $H_{2,2}, H_{3,2}, \ldots, H_{T,2}$ that respectively approximate H_1^* and H_2^* on $[0, \gamma_{\max}]$, with $\gamma_{\max}$ large enough. Denote

$$\begin{aligned} \tilde{H}_1^* &= H_{2,1} \otimes H_{3,1} \otimes \ldots \otimes H_{T,1}, \\ \tilde{H}_2^* &= H_{2,2} \otimes H_{3,2} \otimes \ldots \otimes H_{T,2}. \end{aligned}$$

Note that for $\gamma \in [0, \gamma_{\max}]$,

$$\begin{aligned} \left\{\tilde{H}_1^* > \tilde{H}_2^*\right\} &\supset \{H_1^* > H_2^* \text{ and } H_1^* - H_2^* > 2\eta\}, \\ \left\{\tilde{H}_2^* > \tilde{H}_1^*\right\} &\supset \{H_2^* > H_1^* \text{ and } H_2^* - H_1^* > 2\eta\}. \end{aligned}$$

Since $\bigcup_{\eta>0} \{H_1^* - H_2^* > 2\eta\} = \{H_1^* > H_2^*\}$ and $\bigcup_{\eta>0} \{H_2^* - H_1^* > 2\eta\} = \{H_2^* > H_1^*\}$, once $\eta > 0$ is chosen small enough,

$$\left(\gamma_{\text{scale}} \frac{q_1(y_1)}{p_1(y_1)}\right)_{y_1 \in \mathcal{O}_1} \text{ overlaps with both } \left\{\tilde{H}_1^* > \tilde{H}_2^*\right\} \text{ and } \left\{\tilde{H}_1^* < \tilde{H}_2^*\right\}.$$

From this, we can construct a counterexample to the f-DP filter validity. For $t = 2, \ldots, T$, let $(P_{t,1}, Q_{t,1})$ and $(P_{t,2}, Q_{t,2})$ be the pairs of distributions that have privacy profiles $H_{t,1}$ and $H_{t,2}$ respectively. Choose the adaptive selection rule at step t:

$$\begin{aligned} P_t^{(y_1,\ldots,y_{t-1})} \coloneqq P_t^{(y_1)} &= \begin{cases} P_{t,1} & \text{if } \gamma_{\text{scale}} \frac{q_1(y_1)}{p_1(y_1)} \in \left\{\tilde{H}_1^* \geq \tilde{H}_2^*\right\}, \\ P_{t,2} & \text{otherwise,} \end{cases} \\ Q_t^{(y_1,\ldots,y_{t-1})} \coloneqq Q_t^{(y_1)} &= \begin{cases} Q_{t,1} & \text{if } \gamma_{\text{scale}} \frac{q_1(y_1)}{p_1(y_1)} \in \left\{\tilde{H}_1^* \geq \tilde{H}_2^*\right\}, \\ Q_{t,2} & \text{otherwise.} \end{cases} \end{aligned}$$

Writing H_1 for the privacy profile corresponding to f_1, we obtain the following inequality:

$$\begin{aligned} &\int_{y_1 \in \mathcal{O}_1} \max\left[\tilde{H}_1^*\left(\gamma_{\text{scale}} \frac{q_1(y_1)}{p_1(y_1)}\right), \tilde{H}_2^*\left(\gamma_{\text{scale}} \frac{q_1(y_1)}{p_1(y_1)}\right)\right] p_1(y_1)\, \mathrm{d}y_1 \\ &\qquad > \max\left(\left(H_1 \otimes \tilde{H}_1^*\right)(\gamma_{\text{scale}}), \left(H_1 \otimes \tilde{H}_2^*\right)(\gamma_{\text{scale}})\right). \end{aligned}$$

Here, the LHS is the true privacy profile of the composition, while the RHS is chosen as the budget used in the filter. □

E Fully Adaptive Central Limit Theorem

E.1 Proof to the Fully Adaptive f-DP Central Limit Theorem

First, we present the martingale Berry–Esseen theorem from Fan (2019).

Theorem E.1. *Let $(\Omega, \mathcal{G}, P)$ be a probability space and let $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \ldots \subset \mathcal{G}_T \subset \mathcal{G}$ be a filtration of increasing σ-algebras. Let $(\xi_i)_{i=1}^T$ be a sequence of martingale differences, which means that $\sum_{i=1}^t \xi_i$ for $t = 1, \ldots, T$ forms a martingale, or equivalently $\mathbb{E}_P[\xi_i \,|\, \mathcal{G}_{i-1}] = 0$ for all i. Assume that the following holds.*

1. *There exists $\kappa \in \left[0, \frac{1}{4}\right]$ such that $\left|\sum_{i=1}^T \mathbb{E}_P\left[\xi_i^2 \,|\, \mathcal{G}_{i-1}\right] - 1\right| \le \kappa$ almost surely.*
2. *There exists $\rho \in \left(0, \frac{1}{2}\right]$ such that $\mathbb{E}_P\left[|\xi_i|^3 \,|\, \mathcal{G}_{i-1}\right] \le \rho \mathbb{E}_P\left[\xi_i^2 \,|\, \mathcal{G}_{i-1}\right]$ almost surely for all i.*

Then the cdf of $\sum_{i=1}^T \xi_i$ satisfies

$$\sup_{x \in \mathbb{R}} \left| P\left(\sum_{i=1}^T \xi_i \le x\right) - \Phi(x) \right| \le C\left(\rho |\ln \rho| + \sqrt{\kappa}\right),$$

where Φ is the cdf of the standard Gaussian distribution, and C is a universal constant.

Note that as T grows, v increases linearly while ρ remains constant, indicating a convergence rate of $O\left(\frac{\ln T}{\sqrt{T}}\right)$. Fan (2019) showed that for martingale difference sequences satisfying the conditions in Theorem E.1, the order of convergence $\rho|\ln \rho|$ is optimal. To the best of our knowledge, this represents the best convergence rate among existing martingale Berry–Esseen variants, though it is weaker than the classical $O\left(\frac{1}{\sqrt{T}}\right)$ rate for i.i.d. sequences due to the martingale setting. Additional variants of the martingale Berry–Esseen theorem were derived by Heyde and Brown (1970) and Fan and Shao (2017).

Theorem 4.1. *Let $P_1, \ldots, P_T$ and $Q_1, \ldots, Q_T$ be adaptive sequences of distributions, where for each t, the choices of P_t and Q_t may depend on the past realizations $y_1, \ldots, y_{t-1}$ drawn under $P_1, \ldots, P_{t-1}$ and $Q_1, \ldots, Q_{t-1}$, respectively. Define $L_t = \ln \frac{P_t}{Q_t}$ as the PLRV at step t, and let $P = \bigotimes_{t=1}^T P_t$, $Q = \bigotimes_{t=1}^T Q_t$, and $y = (y_1, \ldots, y_T)$. We assume that the following conditions hold with probability 1:*

1. $\left|\sum_{t=1}^T \mathbb{E}_{y\sim P}[L_t(y_t) \,|\, y_{1:t-1}] - m_1\right| \le \eta_1$,
2. $\left|\sum_{t=1}^T \mathbb{E}_{y\sim Q}[L_t(y_t) \,|\, y_{1:t-1}] + m_2\right| \le \eta_2$,
3. $\left|\sum_{t=1}^T \mathrm{Var}_{y\sim P}[L_t(y_t) \,|\, y_{1:t-1}] - v\right| \le \kappa$,
4. $\left|\sum_{t=1}^T \mathrm{Var}_{y\sim Q}[L_t(y_t) \,|\, y_{1:t-1}] - v\right| \le \kappa$,
5. $\mathbb{E}_{y\sim P}\left[\left|L_t - \mathbb{E}_{y\sim P}[L_t(y_t) \,|\, y_{1:t-1}]\right|^3 \,\middle|\, y_{1:t-1}\right] \le \rho \cdot \mathrm{Var}_{y\sim P}[L_t(y_t) \,|\, y_{1:t-1}]$ *for all t,*
6. $\mathbb{E}_{y\sim Q}\left[\left|L_t - \mathbb{E}_{y\sim Q}[L_t(y_t) \,|\, y_{1:t-1}]\right|^3 \,\middle|\, y_{1:t-1}\right] \le \rho \cdot \mathrm{Var}_{y\sim Q}[L_t(y_t) \,|\, y_{1:t-1}]$ *for all t.*

Here, we require that $4\rho^2 \le v$ and $4\kappa \le v$. Under these assumptions, the trade-off function $T[P, Q]$ satisfies the following inequalities for all α:

$$G_{\mu+\phi}(\alpha + \Delta) - \Delta \le T[P, Q](\alpha) \le G_{\mu-\phi}(\alpha - \Delta) + \Delta,$$

where $\mu = \frac{m_1+m_2}{\sqrt{v}}$, $\phi = \frac{\eta_1+\eta_2}{\sqrt{v}}$, and $\Delta = C\left(\frac{\rho}{\sqrt{v}}\left|\ln \frac{\rho}{\sqrt{v}}\right| + \sqrt{\frac{\kappa}{v}}\right)$ for some universal constant C.

Proof. We apply Theorem E.1 to the sequences of martingale differences

$$\frac{L_t(y_t) - \mathbb{E}_{y\sim P}[L_t(y_t) \,|\, y_{1:t-1}]}{\sqrt{v}}, y \sim P \quad \text{and} \quad \frac{L_t(y_t) - \mathbb{E}_{y\sim Q}[L_t(y_t) \,|\, y_{1:t-1}]}{\sqrt{v}}, y \sim Q$$

for $t = 1, \ldots, T$ to arrive at the inequalities

$$\sup_{x\in\mathbb{R}}\left|P\left[\frac{\sum_{t=1}^T L_t(y_t) - \sum_{t=1}^T \mathbb{E}_{y\sim P}[L_t(y_t)\,|\,y_{1:t-1}]}{\sqrt{v}} \le x\right] - \Phi(x)\right| \le C\left(\frac{\rho}{\sqrt{v}}\left|\ln\frac{\rho}{\sqrt{v}}\right| + \sqrt{\frac{\kappa}{v}}\right),$$
$$\sup_{x\in\mathbb{R}}\left|Q\left[\frac{\sum_{t=1}^T L_t(y_t) - \sum_{t=1}^T \mathbb{E}_{y\sim Q}[L_t(y_t)\,|\,y_{1:t-1}]}{\sqrt{v}} \le x\right] - \Phi(x)\right| \le C\left(\frac{\rho}{\sqrt{v}}\left|\ln\frac{\rho}{\sqrt{v}}\right| + \sqrt{\frac{\kappa}{v}}\right).$$

Denote $L(y) = \sum_{t=1}^T L_t(y_t)$ and $\Delta = C\left(\frac{\rho}{\sqrt{v}}\left|\ln\frac{\rho}{\sqrt{v}}\right| + \sqrt{\frac{\kappa}{v}}\right)$. By Conditions 1 and 2, the sums of conditional means are almost surely bounded,

$$m_1 - \eta_1 \le \sum_{t=1}^T \mathbb{E}_{y\sim P}[L_t(y_t)\,|\,y_{1:t-1}] \le m_1 + \eta_1,$$
$$-m_2 - \eta_2 \le \sum_{t=1}^T \mathbb{E}_{y\sim Q}[L_t(y_t)\,|\,y_{1:t-1}] \le -m_2 + \eta_2.$$

Therefore for all $x \in \mathbb{R}$,

$$P\left(\frac{L(y) - m_1 + \eta_1}{\sqrt{v}} \le x\right) \le \Phi(x) + \Delta, \tag{4}$$
$$P\left(\frac{L(y) - m_1 - \eta_1}{\sqrt{v}} \le x\right) \ge \Phi(x) - \Delta, \tag{5}$$
$$Q\left(\frac{L(y) + m_2 + \eta_2}{\sqrt{v}} \le x\right) \le \Phi(x) + \Delta, \tag{6}$$
$$Q\left(\frac{L(y) + m_2 - \eta_2}{\sqrt{v}} \le x\right) \ge \Phi(x) - \Delta. \tag{7}$$

Now, we convert these inequalities to their trade-off function form. Let $\alpha \in (0, 1-\Delta)$ and let $x \in \mathbb{R}$ such that the test

$$\phi(y) = \begin{cases} 1 & \text{if } L(y) \le x, \\ 0 & \text{otherwise} \end{cases}$$

for the hypothesis testing problem $H_0 : y \sim P$ vs. $H_1 : y \sim Q$ has an FPR $P(L(y) \le x) = \alpha$.[5] By (5),

$$\Phi\left(\frac{x - m_1 - \eta_1}{\sqrt{v}}\right) \le P\left(\frac{L(y) - m_1 - \eta_1}{\sqrt{v}} \le \frac{x - m_1 - \eta_1}{\sqrt{v}}\right) + \Delta = \alpha + \Delta.$$

Since Φ^{-1} is an increasing function,

$$\frac{x - m_1 - \eta_1}{\sqrt{v}} \le \Phi^{-1}(\alpha + \Delta),$$

which implies that

$$\frac{x + m_2 + \eta_2}{\sqrt{v}} \le \Phi^{-1}(\alpha + \Delta) + \frac{m_1 + m_2}{\sqrt{v}} + \frac{\eta_1 + \eta_2}{\sqrt{v}}.$$

Combining (6) with the fact that Φ is increasing,

$$\begin{aligned} Q(L(y) \le x) = Q\left(\frac{L(y) + m_2 + \eta_2}{\sqrt{v}} \le \frac{x + m_2 + \eta_2}{\sqrt{v}}\right) &\le \Phi\left(\frac{x + m_2 + \eta_2}{\sqrt{v}}\right) + \Delta \\ &\le \Phi\left(\Phi^{-1}(\alpha + \Delta) + \frac{m_1 + m_2}{\sqrt{v}} + \frac{\eta_1 + \eta_2}{\sqrt{v}}\right) + \Delta. \end{aligned}$$

[5] For those tests with $\phi(y) < 1$ when $L(y) = x$, (8) can be proven once (8) is established for those tests with $\phi(y) = 1$ when $L(y) = x$.

Therefore the FNR $Q(L(y) > x)$ satisfies

$$Q(L(y) > x) \geq 1 - \Phi\left(\Phi^{-1}(\alpha + \Delta) + \frac{m_1 + m_2}{\sqrt{v}} + \frac{\eta_1 + \eta_2}{\sqrt{v}}\right) - \Delta.$$

We modify this into the final expression using the relationship $1 - \Phi(x) = \Phi(-x)$ and $\Phi^{-1}(\alpha) = -\Phi^{-1}(1 - \alpha)$,

$$Q(L(y) > x) \geq \Phi\left(\Phi^{-1}(1 - (\alpha + \Delta)) - \frac{m_1 + m_2}{\sqrt{v}} - \frac{\eta_1 + \eta_2}{\sqrt{v}}\right) - \Delta = G_{\mu+\phi}(\alpha + \Delta) - \Delta,$$

where $\mu = \frac{m_1+m_2}{\sqrt{v}}$ and $\phi = \frac{\eta_1+\eta_2}{\sqrt{v}}$. Finally, $T[P, Q](\alpha) = Q(L(y) > x)$ by the Neyman-Pearson lemma, so we conclude

$$T[P, Q](\alpha) \geq G_{\mu+\phi}(\alpha + \Delta) - \Delta \text{ for all } \alpha \in [0, 1 - \Delta]\,. \tag{8}$$

The bound is extended from $\alpha \in (0, 1 - \Delta)$ to $\alpha \in [0, 1 - \Delta]$ due to the continuity of trade-off functions. The proof for the remaining bound $T[P, Q](\alpha) \leq G_{\mu-\phi}(\alpha - \Delta) + \Delta$ is identical, which utilizes the remaining inequalities (4) and (7). □

E.2 Proofs for Demonstration of Approximate GDP Filter

The limit conditions in Theorem 4.2 are:

1. Near $\tilde{q} = 0$, $T\bar{q}^2 = \Theta(1)$ as $\bar{q} \to 0$.
2. Near $\tilde{q} = 1$, $T\frac{1}{\bar{\sigma}^2} = \Theta(1)$ and $\hat{q} = 1 - \bar{q} = O\big(\frac{1}{\bar{\sigma}^2}\big)$ as $\bar{\sigma} \to \infty$.

Theorem 4.2. *Under the above scaling assumptions, the composition of subsampled Gaussian mechanisms produced by Algorithm 2 satisfies:*

1. *Near $\tilde{q} = 0$: Δ-approx. $\big(\sqrt{2B} + O(\bar{q})\big)$-GDP, with an error $\Delta = O(\sqrt{\bar{q}})$ as $\bar{q} \to 0$.*
2. *Near $\tilde{q} = 1$: Δ-approx. $\big(\sqrt{2B} + O(\bar{\sigma}^{-2})\big)$-GDP, with an error $\Delta = O\big(\frac{\ln \bar{\sigma}}{\bar{\sigma}}\big)$ as $\bar{\sigma} \to \infty$.*

Proof. Fix a remove pair S, S^- that differs in some datapoint X. After T steps, the sum of conditional expectations of PLRVs between S and S^- may remain below B, because

1. The filter takes into account an amount of $\frac{1}{2}q_t^2\left(\mathrm{e}^{\frac{1}{\sigma_t^2}} - 1\right)$ or $\frac{1}{2}q_t^2\frac{1}{\sigma_t^2}$, which upper bounds the actual amount $\frac{1}{2}q_t^2\left(\mathrm{e}^{\frac{\|\dot{g}_t(X)\|^2}{\sigma_t^2 C_t^2}} - 1\right)$ or $\frac{1}{2}q_t^2\frac{\|\dot{g}_t(X)\|^2}{\sigma_t^2 C_t^2}$ spent by the S, S^-.
2. After T steps, the total budget accounted for may not have reached B.

In this case, we can artificially create an extra mechanism $\mathcal{M}_{T+1}$ at step $T + 1$, tailored to the pair S, S^-, that adaptively consumes the remaining budget, and then apply projection as a post-processing to extract the first T outputs from $T + 1$ outputs. Therefore, w.l.o.g. we can assume that the total budget B is almost surely exhausted by S, S^- after T steps.

We apply Theorem 4.1 in combination with Theorems F.1 and F.2 to the pair S, S^-:

1. In the regime near $\tilde{q} = 0$, as $\bar{q} \to 0$, the approximate GDP guarantee is

$$\frac{B + B}{\sqrt{2B}} + T \cdot O\big(\bar{q}^3\big) = \sqrt{2B} + O(\bar{q})\,,$$

with Δ-error

$$O(\bar{q})\,|\ln O(\bar{q})| + \sqrt{\frac{T \cdot O(\bar{q}^3)}{2B}} = O\big(\sqrt{\bar{q}}\big)\,.$$

2. In the regime near $\tilde{q} = 1$, as $\bar{\sigma} \to \infty$ the approximate GDP guarantee is

$$\frac{B+B}{\sqrt{2B}} + T\hat{q} \cdot O\left(\frac{1}{\bar{\sigma}^4}\right) + T \cdot O(\hat{q}^2) = \sqrt{2B} + O\left(\frac{1}{\bar{\sigma}^2}\right),$$

with Δ-error

$$O\left(\frac{1}{\bar{\sigma}}\right)\left|\ln O\left(\frac{1}{\bar{\sigma}}\right)\right| + \sqrt{\frac{T\hat{q} \cdot O\left(\frac{1}{\bar{\sigma}^4}\right) + T \cdot O(\hat{q}^2)}{2B}} = O\left(\frac{\ln \bar{\sigma}}{\bar{\sigma}}\right).$$

Since the Δ-approximate μ-GDP guarantee is symmetric, the privacy guarantees above apply to both trade-off functions $T[\mathcal{M}_{1:t}(S), \mathcal{M}_{1:t}(S^-)]$ and $T[\mathcal{M}_{1:t}(S^-), \mathcal{M}_{1:t}(S)]$. □

Theorem 4.3. *Suppose the initial individual budgets are uniform, i.e. $B_j = B$ for all $j = 1, \ldots, n$. Then Algorithm 3 satisfies identical asymptotic privacy guarantees as stated in Theorem 4.2.*

Proof. The proof mirrors the proof idea of Feldman and Zrnic (2021). Fix a remove pair $S = (X_1, \ldots, X_n, X)$ and $S^- = (X_1, \ldots, X_n)$. At each step t, the subset of datapoints in the input dataset that have positive remaining budget, and thus are still participating as proper inputs, is called the active subset. Denote $\tilde{S}_t$ the active subset of S at step t, and denote $\tilde{S}_t^-$ that of S^-.

At step t, conditioned on a given sequence of previous outputs $y_{1:t-1}$, the total budget spent so far for each individual X_j can be computed, and hence it is deterministic which datapoint remains in $\tilde{S}_t$ and $\tilde{S}_t^-$. Consequently, $\tilde{S}_t$ and $\tilde{S}_t^-$ are either identical, or differ only in the datapoint X.

Therefore, the pair of PLRVs of S, S^- at step t, conditioned on a given sequence of previous outputs $y_{1:t-1}$, is the pair of PLRVs of $\tilde{S}_t, \tilde{S}_t^-$, which implies that Theorems F.1 and F.2 still apply. We then apply Theorem 4.1 and obtain identical results to Theorem 4.2.

This proof can be seen as an extension of the proof to Theorem 4.2 by allowing the pair of active subsets adaptively shrink through composition. □

F Privacy Loss Moments: Approximations

For a subsampled Gaussian mechanism $\mathcal{M}$, let $P = \mathcal{M}(S)$ and $Q = \mathcal{M}(S^-)$ denote the output distributions for an arbitrary remove pair S, S^- differing by a single datapoint X. In the regimes $q \ll 1$ and $q \approx 1$, the trade-off function is nearly symmetric and close to Gaussian, which gives rise to the asymptotic relationships

$$\mathbb{E}_P\left[\ln \tfrac{P}{Q}\right] \approx -\mathbb{E}_Q\left[\ln \tfrac{P}{Q}\right], \qquad \mathrm{Var}_P\left[\ln \tfrac{P}{Q}\right] \approx \mathrm{Var}_Q\left[\ln \tfrac{P}{Q}\right],$$

and, most notably,

$$\frac{\mathbb{E}_P\left[\ln \frac{P}{Q}\right]}{\mathrm{Var}_P\left[\ln \frac{P}{Q}\right]} \approx \frac{-\mathbb{E}_Q\left[\ln \frac{P}{Q}\right]}{\mathrm{Var}_Q\left[\ln \frac{P}{Q}\right]} \approx \frac{1}{2}.$$

To build intuition, consider each regime in turn. When $q \approx 1$, the subsampled Gaussian mechanism is close to the standard Gaussian mechanism, whose PLRV has mean-to-variance ratio exactly $1/2$. When $q \ll 1$, the PLRV mean is small and the variance $\mathrm{Var}[L]$ is well approximated by the second moment $\mathbb{E}[L^2]$. Setting $L = \ln P/Q$ and applying the Taylor expansions $e^y \approx 1 + y + \frac{y^2}{2}$ and $e^{-y} \approx 1 - y + \frac{y^2}{2}$ yields

$$1 = \mathbb{E}_P\left[\tfrac{Q}{P}\right] = \mathbb{E}_P\left[e^{-L}\right] \approx 1 - \mathbb{E}_P[L] + \tfrac{1}{2}\mathbb{E}_P[L^2],$$
$$1 = \mathbb{E}_Q\left[\tfrac{P}{Q}\right] = \mathbb{E}_Q\left[e^{L}\right] \approx 1 + \mathbb{E}_Q[L] + \tfrac{1}{2}\mathbb{E}_Q[L^2].$$

Theorem F.1. *Fix a datapoint X. Let h be a deterministic function on datasets with $h(\emptyset) = 0$, and let $\mu = \|h((X))\|$. Let $\mathcal{M}$ be the subsampled Gaussian mechanism derived from h, with sampling rate q and noise level σ. For an arbitrary remove-neighbor pair $S = (X_1, \ldots, X_n, X)$ and $S^- = (X_1, \ldots, X_n)$, set $P = \mathcal{M}(S)$ and $Q = \mathcal{M}(S^-)$. As $q \to 0$:*

1. $\mathbb{E}_P\left[\ln \frac{P}{Q}\right] = \frac{1}{2}q^2\left(e^{\mu^2/\sigma^2} - 1\right) + O(q^3)$,
2. $\mathbb{E}_Q\left[\ln \frac{P}{Q}\right] = -\frac{1}{2}q^2\left(e^{\mu^2/\sigma^2} - 1\right) + O(q^3)$,
3. $\mathrm{Var}_P\left[\ln \frac{P}{Q}\right] = q^2\left(e^{\mu^2/\sigma^2} - 1\right) + O(q^3)$,
4. $\mathrm{Var}_Q\left[\ln \frac{P}{Q}\right] = q^2\left(e^{\mu^2/\sigma^2} - 1\right) + O(q^3)$,
5. $\mathbb{E}_P\left|\ln \frac{P}{Q} - \mathbb{E}_P\left[\ln \frac{P}{Q}\right]\right|^3 \le q^3\left(e^{3\mu^2/\sigma^2} + 1\right) + O(q^4)$,
6. $\mathbb{E}_Q\left|\ln \frac{P}{Q} - \mathbb{E}_Q\left[\ln \frac{P}{Q}\right]\right|^3 \le q^3\left(e^{3\mu^2/\sigma^2} + 1\right) + O(q^4)$.

Each error term $O(q^3)$ or $O(q^4)$ may depend on q, σ, X, and $X_1, \ldots, X_n$.

Theorem F.2. *Fix a datapoint X. Let h be a summation $h(X_1, \ldots, X_n) = \sum_i g(X_i)$ for some function g, and let $\mu = \|g(X)\|$. Let $\mathcal{M}$ be the subsampled Gaussian mechanism derived from h, with sampling rate q and noise level σ, and write $\hat{q} = 1 - q$. For an arbitrary remove-neighbor pair $S = (X_1, \ldots, X_n, X)$ and $S^- = (X_1, \ldots, X_n)$, set $P = \mathcal{M}(S)$ and $Q = \mathcal{M}(S^-)$. Then:*

1. $\mathbb{E}_P\left[\ln \frac{P}{Q}\right] = \frac{1}{2}q^2\frac{\mu^2}{\sigma^2} + \hat{q}\cdot O\left(\sigma^{-4}\right) + O(\hat{q}^2)$,
2. $\mathbb{E}_Q\left[\ln \frac{P}{Q}\right] = -\frac{1}{2}q^2\frac{\mu^2}{\sigma^2} + \hat{q}\cdot O\left(\sigma^{-4}\right) + O(\hat{q}^2)$,
3. $\mathrm{Var}_P\left[\ln \frac{P}{Q}\right] = q^2\frac{\mu^2}{\sigma^2} + \hat{q}\cdot O\left(\sigma^{-4}\right) + O(\hat{q}^2)$,
4. $\mathrm{Var}_Q\left[\ln \frac{P}{Q}\right] = q^2\frac{\mu^2}{\sigma^2} + \hat{q}\cdot O\left(\sigma^{-4}\right) + O(\hat{q}^2)$,
5. $\mathbb{E}_P\left|\ln \frac{P}{Q} - \mathbb{E}_P\left[\ln \frac{P}{Q}\right]\right|^3 \le (2 - q^3)\sqrt{\frac{8}{\pi}}\frac{\mu^3}{\sigma^3} + \hat{q}\cdot O\left(\sigma^{-4}\right) + O(\hat{q}^2)$,
6. $\mathbb{E}_Q\left|\ln \frac{P}{Q} - \mathbb{E}_Q\left[\ln \frac{P}{Q}\right]\right|^3 \le (2 - q^3)\sqrt{\frac{8}{\pi}}\frac{\mu^3}{\sigma^3} + \hat{q}\cdot O\left(\sigma^{-4}\right) + O(\hat{q}^2)$.

Here $O(\hat{q}^2)$ is in the limit $\hat{q} \to 0$, and $O(\sigma^{-4})$ is in the limit $\sigma \to \infty$. All error terms may depend on q, σ, X, and $X_1, \ldots, X_n$.

F.1 Numerical Illustration of Moment Approximations

We numerically illustrate the asymptotic approximations of Theorems F.1 and F.2 for the means and variances of the PLRVs $L = \ln(P/Q)$ of the Poisson subsampled Gaussian mechanism. The accurate values are computed by direct numerical integration on a fine grid, using the one-dimensional distributions $P = q \cdot \mathcal{N}(1, \sigma^2) + (1 - q) \cdot \mathcal{N}(0, \sigma^2)$ and $Q = \mathcal{N}(0, \sigma^2)$.

Small-q regime. Figures 7 and 8 illustrate the accuracy of approximations 1. and 3. of Theorem F.1, respectively. We plot the relative error between the numerical quantity and its $O(q^2)$ approximation as $q \to 0$. As predicted by the theorem, the error decreases polynomially in q, and the accuracy improves for larger noise levels σ.

Large-q regime. Figures 9 and 10 illustrate the accuracy of approximations 1. and 3. of Theorem F.2. Here q is close to 1, and we again plot relative errors as functions of q for several values of σ. As predicted by the theorem, the relative error decreases as $\hat{q} = 1 - q \to 0$ and σ increases.

The numerical illustrations of Figures 7 to 10 confirm that the PLRV mean and variance satisfy the predicted scalings, and that the resulting mean-to-variance ratios approach $1/2$, as stated by Theorems F.1 and F.2.

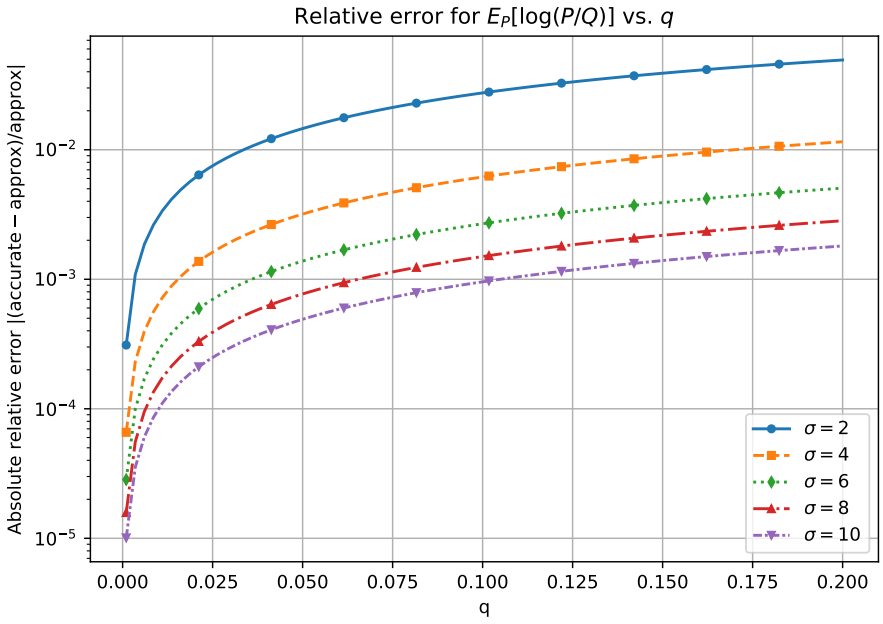


Figure 7: Relative error of $\mathbb{E}_P[\ln(P/Q)]$ versus the approximation $\frac{1}{2}q^2(\exp(\mu^2/\sigma^2) - 1)$ as $q \to 0$, illustrating the approximation of 1. in Theorem F.1.

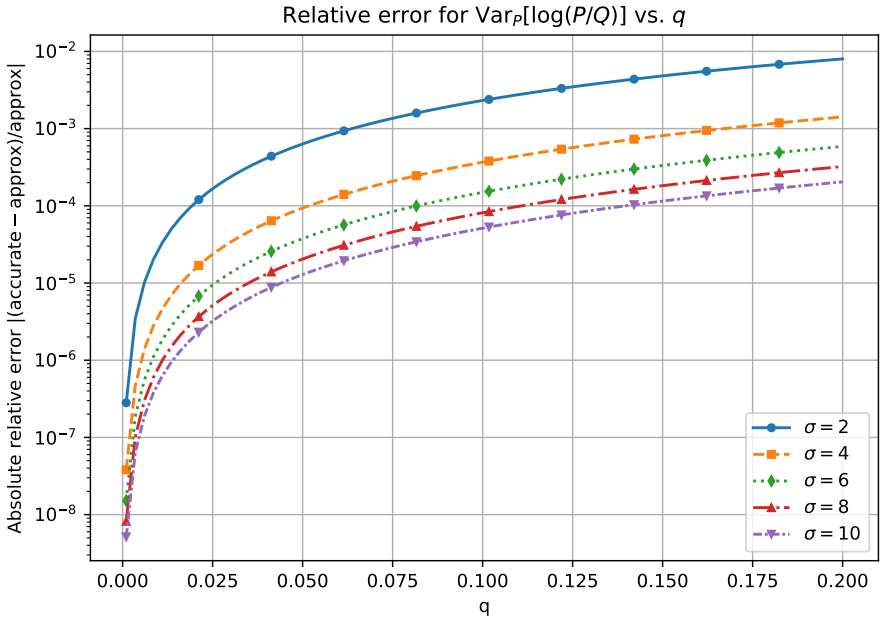


Figure 8: Relative error of $\mathrm{Var}_P[\ln(P/Q)]$ versus the approximation $q^2(\exp(\mu^2/\sigma^2) - 1)$ as $q \to 0$, illustrating the approximation 3. in Theorem F.1.

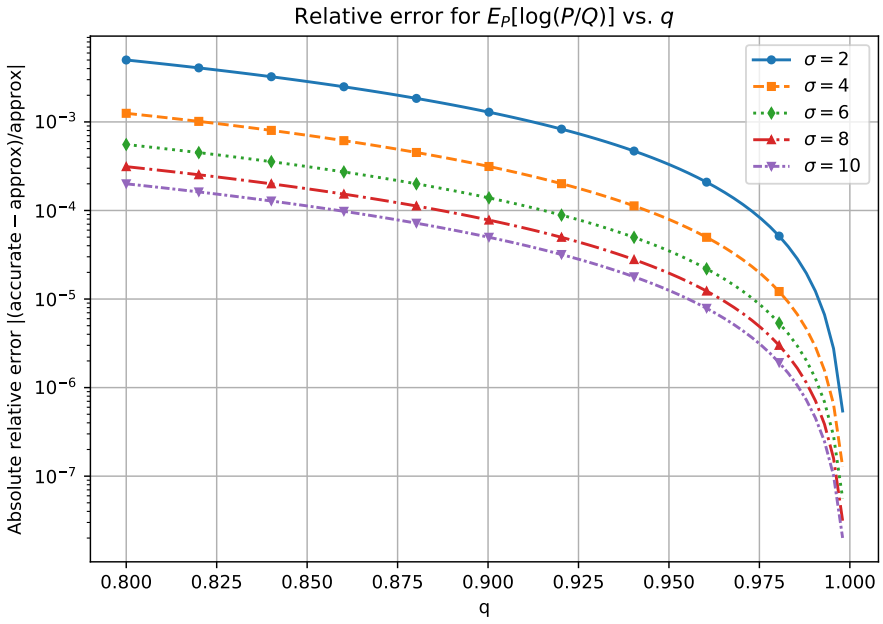


Figure 9: Relative error of $\mathbb{E}_P[\ln(P/Q)]$ versus the approximation $\frac{1}{2}q^2\mu^2/\sigma^2$ for $q \approx 1$, illustrating the approximation 1. in Theorem F.2.

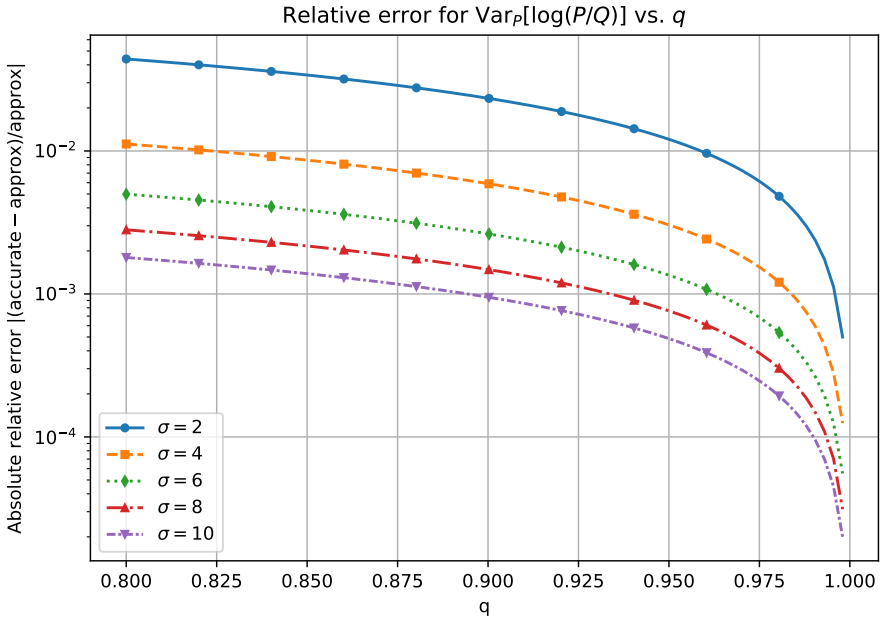


Figure 10: Relative error of $\mathrm{Var}_P[\log(P/Q)]$ versus the approximation $q^2\mu^2/\sigma^2$ for $q \approx 1$, illustrating the approximation 3. in Theorem F.2.

F.2 Derivations Using Taylor Expansions

Proof of Theorem F.1. All big-O terms in this proof are taken in the limit $q \to 0$. Throughout we use the following notation.

- $P_0 = \mathcal{N}(0, \sigma^2\mathbf{I})$.

- $a(y) = \dfrac{\mathcal{N}(h((X)),\, \sigma^2\mathbf{I})(y)}{P_0(y)} - 1.$
- $a_j(y) = \dfrac{\mathcal{N}(h((X_j)),\, \sigma^2\mathbf{I})(y)}{P_0(y)} - 1, \quad j = 1, \ldots, n.$
- $b_k(y) = \dfrac{\mathcal{N}(h((X_{j_1}, X_{j_2})),\, \sigma^2\mathbf{I})(y)}{P_0(y)} - 1$, where X_{j_1}, X_{j_2} is an arbitrary distinct pair from $(X, X_1, \ldots, X_n)$. The index k runs from 1 to $\frac{n(n+1)}{2}$ to cover all such pairs; by convention, $k = 1, \ldots, \frac{n(n-1)}{2}$ covers pairs that exclude X, and the remaining indices cover pairs that include X.

Step 1: Density approximations. The density of $P = \mathcal{M}(S)$ is a Gaussian mixture whose weights depend on q. Applying the Taylor expansion $(1-q)^\alpha = 1 - \alpha q + \frac{\alpha(\alpha-1)}{2} q^2 + q^3 O(1)$ gives two useful approximations:

$$\mathcal{M}(S)(y) = \Big[1 + q\left(\sum_{j=1}^{n} a_j(y) + a(y)\right) + q^2\left(-n\sum_{j=1}^{n} a_j(y) - na(y) + \sum_{k=1}^{\frac{n(n+1)}{2}} b_k(y)\right) + q^3 O(1)\Big] P_0(y), \tag{9}$$

$$\mathcal{M}(S)(y) = \Big[1 + q\left(\sum_{j=1}^{n} a_j(y) + a(y)\right) + q^2 O(1)\Big] P_0(y). \tag{10}$$

The same expansion applied to $Q = \mathcal{M}(S^-)$ yields

$$\mathcal{M}(S^-)(y) = \Big[1 + q\sum_{j=1}^{n} a_j(y) + q^2\left(-(n-1)\sum_{j=1}^{n} a_j(y) + \sum_{k=1}^{\frac{n(n-1)}{2}} b_k(y)\right) + q^3 O(1)\Big] P_0(y), \tag{11}$$

$$\mathcal{M}(S^-)(y) = \Big[1 + q\sum_{j=1}^{n} a_j(y) + q^2 O(1)\Big] P_0(y). \tag{12}$$

Step 2: Approximation of $\mathbb{E}_P[\ln(P/Q)]$. Using (9), (10), and (11),

$$\mathbb{E}_P\left[\ln\frac{P}{Q}\right] = \mathbb{E}_{y\sim P_0}\left[\frac{P(y)}{P_0(y)}\ln\frac{P(y)}{Q(y)}\right] = \mathbb{E}_{y\sim P_0}\left[\left(1 + q\left(\sum_{j=1}^{n} a_j + a\right) + q^2 O(1)\right) \cdot \ln\frac{1 + q\Big(\sum_j a_j + a\Big) + q^2\left(-n\sum_j a_j - na + \sum_{k=1}^{\frac{n(n+1)}{2}} b_k\right) + q^3 O(1)}{1 + q\sum_j a_j + q^2\left(-(n-1)\sum_j a_j + \sum_{k=1}^{\frac{n(n-1)}{2}} b_k\right) + q^3 O(1)}\right].$$

Because $\mathbb{E}_{y\sim P_0}[a(y)] = \mathbb{E}_{y\sim P_0}[a_j(y)] = \mathbb{E}_{y\sim P_0}[b_k(y)] = 0$ for all j, k, the q^2 corrections in the numerator and denominator contribute zero expectation and may be dropped, giving

$$\mathbb{E}_P\left[\ln\frac{P}{Q}\right] = \mathbb{E}_{y\sim P_0}\left[\left(1 + q\left(\sum_j a_j + a\right) + q^2 O(1)\right)\ln\frac{1 + q\Big(\sum_j a_j + a\Big) + q^3 O(1)}{1 + q\sum_j a_j + q^3 O(1)}\right].$$

We simplify the log in two sub-steps.

1. Apply $\frac{1}{1+x} = 1 - x + x^2 + O(x^3)$ to the denominator.
2. Apply $\ln(1+x) = x - \frac{1}{2}x^2 + O(x^3)$ to the resulting expression.

Sub-step 1 yields

$$\begin{aligned}\mathbb{E}_P\left[\ln\frac{P}{Q}\right] &= \mathbb{E}_{y\sim P_0}\Bigg[\left(1 + q\left(\sum_j a_j + a\right) + q^2 O(1)\right) \\ &\qquad \cdot \ln\left(\left(1 + q\left(\sum_j a_j + a\right) + q^3 O(1)\right)\left(1 - q\sum_j a_j + q^2\left(\sum_j a_j\right)^2 + q^3 O(1)\right)\right)\Bigg] \\ &= \mathbb{E}_{y\sim P_0}\left[\left(1 + q\left(\sum_j a_j + a\right) + q^2 O(1)\right)\ln\left(1 + qa(y) - q^2 a(y)\sum_j a_j(y) + q^3 O(1)\right)\right].\end{aligned}$$

Sub-step 2 then gives

$$\begin{aligned}\mathbb{E}_P\left[\ln\frac{P}{Q}\right] &= \mathbb{E}_{y\sim P_0}\Bigg[\left(1 + q\left(\sum_j a_j + a\right) + q^2 O(1)\right) \\ &\qquad \cdot \left(qa(y) - q^2\left(a(y)\sum_j a_j(y) + \tfrac{1}{2}a(y)^2\right) + q^3 O(1)\right)\Bigg] \qquad (13) \\ &= q\,\mathbb{E}_{y\sim P_0}[a(y)] + \tfrac{1}{2}q^2\,\mathbb{E}_{y\sim P_0}[a(y)^2] + O(q^3) \\ &= \tfrac{1}{2}q^2\,\mathbb{E}_{y\sim P_0}[a(y)^2] + O(q^3) \\ &= \tfrac{1}{2}q^2\left(e^{\mu^2/\sigma^2} - 1\right) + O(q^3). \qquad (14)\end{aligned}$$

In the second equality we used that $O(1)$ is integrable, so $\mathbb{E}_{y\sim P_0}[q^3 O(1)] = O(q^3)$. In the third equality we used $\mathbb{E}_{y\sim P_0}[a(y)] = 0$.

The identical argument applied to $\mathbb{E}_Q[\ln(P/Q)]$ yields

$$\begin{aligned}\mathbb{E}_Q\left[\ln\frac{P}{Q}\right] &= \mathbb{E}_{y\sim P_0}\Bigg[\left(1 + q\sum_j a_j + q^2 O(1)\right) \\ &\qquad \cdot \left(qa(y) - q^2\left(a(y)\sum_j a_j(y) + \tfrac{1}{2}a(y)^2\right) + q^3 O(1)\right)\Bigg] \\ &= q\,\mathbb{E}_{y\sim P_0}[a(y)] - \tfrac{1}{2}q^2\,\mathbb{E}_{y\sim P_0}[a(y)^2] + O(q^3) \\ &= -\tfrac{1}{2}q^2\left(e^{\mu^2/\sigma^2} - 1\right) + O(q^3).\end{aligned}$$

Step 3: Variance. Since $\mathbb{E}_P[\ln(P/Q)] = O(q^2)$, we have

$$\mathrm{Var}_P\left[\ln\frac{P}{Q}\right] = \mathbb{E}_P\left[\left(\ln\frac{P}{Q}\right)^2\right] - \left(\mathbb{E}_P\left[\ln\frac{P}{Q}\right]\right)^2 = \mathbb{E}_P\left[\left(\ln\frac{P}{Q}\right)^2\right] + O(q^4),$$

so it suffices to approximate $\mathbb{E}_P[(\ln(P/Q))^2]$. The same two sub-steps give

$$\begin{aligned}
\mathbb{E}_P\left[\left(\ln\frac{P}{Q}\right)^2\right] &= \mathbb{E}_{y\sim P_0}\left[\left(1+q\left(\sum_j a_j + a\right)+q^2O(1)\right)\right. \\
&\qquad \left.\cdot\left(qa(y)-q^2\left(a(y)\sum_j a_j(y)+\tfrac{1}{2}a(y)^2\right)+q^3O(1)\right)^2\right] \\
&= \mathbb{E}_{y\sim P_0}\left[\left(1+q\left(\sum_j a_j + a\right)+q^2O(1)\right)\left(q^2a(y)^2+q^3O(1)\right)\right] \\
&= q^2\,\mathbb{E}_{y\sim P_0}[a(y)^2]+O(q^3) \\
&= q^2\Big(e^{\mu^2/\sigma^2}-1\Big)+O(q^3).
\end{aligned}$$

Hence $\mathrm{Var}_P[\ln(P/Q)] = q^2(e^{\mu^2/\sigma^2}-1)+O(q^3)$, and the case of $\mathrm{Var}_Q[\ln(P/Q)]$ is identical.

Step 4: Third absolute central moment.

$$\begin{aligned}
\mathbb{E}_P\left|\ln\frac{P}{Q}-\mathbb{E}_P\left[\ln\frac{P}{Q}\right]\right|^3 &= \mathbb{E}_{y\sim P_0}\left[(1+qO(1))\cdot\left|\ln\frac{1+q(\sum_j a_j+a)+q^2O(1)}{1+q\sum_j a_j+q^2O(1)}\right.\right. \\
&\qquad \left.\left.-\tfrac{1}{2}q^2(e^{\mu^2/\sigma^2}-1)+q^3O(1)\right|^3\right] \\
&= \mathbb{E}_{y\sim P_0}\Big[(1+qO(1))\,\big|qa(y)+q^2O(1)\big|^3\Big] \\
&= \mathbb{E}_{y\sim P_0}\Big[\big|qa(y)+q^2O(1)\big|^3\Big]+O(q^4).
\end{aligned}$$

Applying the triangle inequality and using $|O(1)| = O(1)$,

$$\begin{aligned}
\mathbb{E}_P\left|\ln\frac{P}{Q}-\mathbb{E}_P\left[\ln\frac{P}{Q}\right]\right|^3 &\le \mathbb{E}_{y\sim P_0}\Big[\big(q|a(y)|+q^2|O(1)|\big)^3\Big]+O(q^4) \\
&= q^3\,\mathbb{E}_{y\sim P_0}|a(y)|^3+O(q^4).
\end{aligned}$$

Applying the inequality $|x-1|^3 \le x^3+1$ for $x\ge 0$ to $a(y) = \exp\Big(\frac{\mu\cdot y}{\sigma^2}-\frac{\mu^2}{2\sigma^2}\Big)-1$,

$$\begin{aligned}
\mathbb{E}_P\left|\ln\frac{P}{Q}-\mathbb{E}_P\left[\ln\frac{P}{Q}\right]\right|^3 &\le q^3\,\mathbb{E}_{y\sim P_0}\left|e^{\frac{\mu\cdot y}{\sigma^2}-\frac{\mu^2}{2\sigma^2}}-1\right|^3+O(q^4) \\
&\le q^3\,\mathbb{E}_{y\sim P_0}\left[e^{\frac{3\mu\cdot y}{\sigma^2}-\frac{3\mu^2}{2\sigma^2}}+1\right]+O(q^4) \\
&= q^3\Big(e^{3\mu^2/\sigma^2}+1\Big)+O(q^4).
\end{aligned}$$

The case of $\mathbb{E}_Q|\ln(P/Q)-\mathbb{E}_Q[\ln(P/Q)]|^3$ is identical. □

Proof of Theorem F.2. For clarity we give the proof in the one-dimensional setting, where g outputs real numbers; the argument extends to the multidimensional case without essential change.

All big-O terms are taken in the limit $\hat{q} = 1-q \to 0$ or $\sigma\to\infty$, unless otherwise stated. We introduce the following notation.

- $\mu = g(X)$, and $\mu_j = g(X_j)$ for $j = 1,\dots,n$.
- $\bar{P} = \mathcal{N}\Big(\sum_j \mu_j + \mu,\, \sigma^2\Big)$ and $\bar{Q} = \mathcal{N}\Big(\sum_j \mu_j,\, \sigma^2\Big)$.

- For $i = 1, \dots, n$,

$$A_i(y) = \frac{\mathcal{N}(\sum_j \mu_j + \mu - \mu_i,\, \sigma^2)(y)}{\bar{P}(y)} - 1 = \exp\!\left(-\frac{\mu_i(y - \sum_j \mu_j - \mu)}{\sigma^2} - \frac{\mu_i^2}{2\sigma^2}\right) - 1. \tag{15}$$

- Similarly,

$$A(y) = \frac{\bar{Q}(y)}{\bar{P}(y)} - 1 = \frac{\mathcal{N}(\sum_j \mu_j,\, \sigma^2)(y)}{\bar{P}(y)} - 1 = \exp\!\left(-\frac{\mu(y - \sum_j \mu_j - \mu)}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) - 1. \tag{16}$$

- For $i = 1, \dots, n$,

$$B_i(y) = \frac{\mathcal{N}(\sum_j \mu_j - \mu_i,\, \sigma^2)(y)}{\bar{Q}(y)} - 1 = \exp\!\left(-\frac{\mu_i(y - \sum_j \mu_j)}{\sigma^2} - \frac{\mu_i^2}{2\sigma^2}\right) - 1.$$

Step 1: Density approximations. Applying the Taylor expansion $q^\alpha = (1-\hat{q})^\alpha = 1 - \alpha\hat{q} + O(\hat{q}^2)$ to the mixture weights gives

$$\mathcal{M}(S)(y) = \Big[1 + \hat{q}\left(\sum_{j=1}^{n} A_j(y) + A(y)\right) + \hat{q}^2 O(1)\Big]\bar{P}(y), \tag{17}$$

$$\mathcal{M}(S^-)(y) = \Big[1 + \hat{q}\sum_{j=1}^{n} B_j(y) + \hat{q}^2 O(1)\Big]\bar{Q}(y). \tag{18}$$

Step 2: Approximation of $\mathbb{E}_P[\ln(P/Q)]$. By (17) and (18),

$$\mathbb{E}_P\left[\ln\frac{P}{Q}\right] = \mathbb{E}_{y\sim\bar{P}}\left[\left(1 + \hat{q}\left(\sum_j A_j + A\right) + \hat{q}^2 O(1)\right) \cdot \left(\ln\frac{\bar{P}(y)}{\bar{Q}(y)} + \ln\frac{1 + \hat{q}\Big(\sum_j A_j + A\Big) + \hat{q}^2 O(1)}{1 + \hat{q}\sum_j B_j + \hat{q}^2 O(1)}\right)\right].$$

We split this into two parts, E_1 and E_2.

Part E_1.

$$E_1 = \mathbb{E}_{y\sim\bar{P}}\left[\left(1 + \hat{q}\left(\sum_j A_j + A\right) + \hat{q}^2 O(1)\right)\ln\frac{\bar{P}(y)}{\bar{Q}(y)}\right]$$
$$= \frac{\mu^2}{2\sigma^2} + \hat{q}\,\mathbb{E}_{y\sim\bar{P}}\left[\left(\sum_j A_j + A\right)\ln\frac{\bar{P}}{\bar{Q}}\right] + O(\hat{q}^2).$$

Expanding A_j, A via (15)–(16) and $e^x - 1 = \sum_{k=1}^{\infty} x^k/k!$, and using $\mathbb{E}_{y\sim\bar{P}} y = \sum_j \mu_j + \mu$ and $\mathrm{Var}_{\bar{P}} y = \sigma^2$, the dominant contribution comes from terms of order $\hat{q}\, y^2/\sigma^4$:

$$M_1 \;=\; \hat{q}\,\mathbb{E}_{y\sim\bar{P}}\left[\left(\sum_j A_j + A\right)\ln\frac{\bar{P}}{\bar{Q}}\right] = \hat{q}\,\mathbb{E}_{y\sim\bar{P}}\left[\frac{\mu(-\sum_j \mu_j - \mu)\, y^2}{\sigma^4}\right] + \hat{q}\cdot O(\sigma^{-4})$$
$$\;=\; -\hat{q}\frac{\mu(\sum_j \mu_j + \mu)}{\sigma^2} + \hat{q}\cdot O(\sigma^{-4}),$$
$$E_1 = \frac{\mu^2}{2\sigma^2} - \hat{q}\frac{\mu(\sum_j \mu_j + \mu)}{\sigma^2} + \hat{q}\cdot O(\sigma^{-4}) + O(\hat{q}^2). \tag{19}$$

Part E_2. Applying $\ln(1+q) = q + O(q^2)$ to the second log,

$$E_2 = \mathbb{E}_{y\sim\bar{P}}\left[(1+\hat{q}O(1))\left(\hat{q}\left(\sum_j A_j + A - \sum_j B_j\right) + \hat{q}^2 O(1)\right)\right]$$
$$= \hat{q}\,\mathbb{E}_{y\sim\bar{P}}\left[\sum_j A_j + A - \sum_j B_j\right] + O(\hat{q}^2).$$

Retaining only the y^2/σ^4 terms from the Taylor expansion of A_j, A, B_j and using the moments of $\bar{P}$,

$$E_2 = \hat{q}\frac{\mu\sum_j \mu_j}{\sigma^2} + \hat{q}\cdot O(\sigma^{-4}) + O(\hat{q}^2). \tag{20}$$

Combining (19) and (20),

$$\mathbb{E}_P\left[\ln\frac{P}{Q}\right] = E_1 + E_2 = (1-2\hat{q})\,\frac{\mu^2}{2\sigma^2} + \hat{q}\cdot O(\sigma^{-4}) + O(\hat{q}^2)$$
$$= (1-\hat{q})^2\,\frac{\mu^2}{2\sigma^2} + \hat{q}\cdot O(\sigma^{-4}) + O(\hat{q}^2).$$

The case of $\mathbb{E}_Q[\ln(P/Q)]$ is analogous.

Step 3: Variance. We approximate $\mathbb{E}_P[(\ln(P/Q))^2]$ by splitting it into two parts, V_1 and V_2.

$$\mathbb{E}_P\left[\left(\ln\frac{P}{Q}\right)^2\right] = \mathbb{E}_{y\sim\bar{P}}\left[\left(1+\hat{q}\left(\sum_j A_j + A\right) + \hat{q}^2 O(1)\right)\right.$$
$$\left.\cdot\left[\left(\ln\frac{\bar{P}}{\bar{Q}}\right)^2 + 2\ln\frac{\bar{P}}{\bar{Q}}\ln\frac{1+\hat{q}\left(\sum_j A_j + A\right) + \hat{q}^2 O(1)}{1+\hat{q}\sum_j B_j + \hat{q}^2 O(1)} + \hat{q}^2 O(1)\right]\right].$$

Part V_1.

$$V_1 = \mathbb{E}_{y\sim\bar{P}}\left[\left(1+\hat{q}\left(\sum_j A_j + A\right) + \hat{q}^2 O(1)\right)\left(\ln\frac{\bar{P}}{\bar{Q}}\right)^2\right] \tag{21}$$
$$= \mathbb{E}_{\bar{P}}\left[\left(\ln\frac{\bar{P}}{\bar{Q}}\right)^2\right] + \hat{q}\cdot O(\sigma^{-4}) + O(\hat{q}^2). \tag{22}$$

Part V_2. Retaining only terms of order $\hat{q}\,y^2/\sigma^4$,

$$V_2 = \mathbb{E}_{y\sim\bar{P}}\left[(1+\hat{q}O(1))\cdot 2\ln\frac{\bar{P}(y)}{\bar{Q}(y)}\ln\frac{1+\hat{q}(\sum_j A_j + A) + \hat{q}^2 O(1)}{1+\hat{q}\sum_j B_j + \hat{q}^2 O(1)}\right]$$
$$= \mathbb{E}_{y\sim\bar{P}}\left[2\hat{q}\left(\ln\frac{\bar{P}(y)}{\bar{Q}(y)}\right)\left(\sum_j A_j + A - \sum_j B_j\right)\right] + O(\hat{q}^2)$$
$$= 2\hat{q}\,\mathbb{E}_{y\sim\bar{P}}\left[\frac{\mu(-\sum_j \mu_j - \mu + \sum_j \mu_j)\,y^2}{\sigma^4}\right] + \hat{q}\cdot O(\sigma^{-4}) + O(\hat{q}^2)$$
$$= -2\hat{q}\frac{\mu^2}{\sigma^2} + \hat{q}\cdot O(\sigma^{-4}) + O(\hat{q}^2). \tag{23}$$

Combining (22) and (23),

$$\begin{aligned}\mathrm{Var}_P\left[\ln\frac{P}{Q}\right] &= V_1 + V_2 - \left(\mathbb{E}_P\left[\ln\frac{P}{Q}\right]\right)^2 \\ &= \mathbb{E}_{\bar{P}}\left[\left(\ln\frac{\bar{P}}{\bar{Q}}\right)^2\right] - 2\hat{q}\frac{\mu^2}{\sigma^2} - (1-4\hat{q})\left(\frac{\mu^2}{2\sigma^2}\right)^2 + \hat{q}\cdot O(\sigma^{-4}) + O(\hat{q}^2) \\ &= \left(\frac{\mu^2}{\sigma^2} + \left(\frac{\mu^2}{2\sigma^2}\right)^2\right) - 2\hat{q}\frac{\mu^2}{\sigma^2} - \left(\frac{\mu^2}{2\sigma^2}\right)^2 + \hat{q}\cdot O(\sigma^{-4}) + O(\hat{q}^2) \\ &= (1-2\hat{q})\,\frac{\mu^2}{\sigma^2} + \hat{q}\cdot O(\sigma^{-4}) + O(\hat{q}^2) = (1-\hat{q})^2\frac{\mu^2}{\sigma^2} + \hat{q}\cdot O(\sigma^{-4}) + O(\hat{q}^2).\end{aligned}$$

The case of $\mathrm{Var}_Q[\ln(P/Q)]$ is identical.

Step 4: Third absolute central moment. We have

$$\begin{aligned}&\mathbb{E}_P\left|\ln\frac{P}{Q} - \mathbb{E}_P\left[\ln\frac{P}{Q}\right]\right|^3 \\ &= \mathbb{E}_{y\sim\bar{P}}\left[\left(1+\hat{q}\left(\sum_j A_j + A\right) + \hat{q}^2 O(1)\right)\left|\ln\frac{\bar{P}(y)}{\bar{Q}(y)} + \ln\frac{1+\hat{q}(\sum_j A_j + A) + \hat{q}^2O(1)}{1+\hat{q}\sum_j B_j + \hat{q}^2O(1)}\right.\right. \\ &\qquad\left.\left. -(1-2\hat{q})\frac{\mu^2}{2\sigma^2} + \hat{q}\cdot O(\sigma^{-4}) + \hat{q}^2O(1)\right|^3\right] \\ &= \mathbb{E}_{y\sim\bar{P}}\left[\left(1+\hat{q}\left(\sum_j A_j + A\right) + \hat{q}^2 O(1)\right)\left|\ln\frac{\bar{P}(y)}{\bar{Q}(y)} - \frac{\mu^2}{2\sigma^2}\right.\right. \\ &\qquad\left.\left. +\hat{q}\left(\sum_j A_j + A - \sum_j B_j + \frac{\mu^2}{\sigma^2} + O(\sigma^{-4})\right) + \hat{q}^2O(1)\right|^3\right].\end{aligned}$$

Applying the triangle inequality, noting $|O(1)| = O(1)$,

$$\begin{aligned}&\mathbb{E}_P\left|\ln\frac{P}{Q} - \mathbb{E}_P\left[\ln\frac{P}{Q}\right]\right|^3 \\ &\le \mathbb{E}_{y\sim\bar{P}}\left[\left(1+\hat{q}\left(\sum_j A_j + A\right) + \hat{q}^2 O(1)\right)\cdot\left(\left|\ln\frac{\bar{P}(y)}{\bar{Q}(y)} - \frac{\mu^2}{2\sigma^2}\right|\right.\right. \\ &\qquad\left.\left. +\hat{q}\left|\sum_j A_j + A - \sum_j B_j + O(\sigma^{-2})\right| + \hat{q}^2O(1)\right)^3\right] \\ &= \mathbb{E}_{y\sim\bar{P}}\left[\left(1+\hat{q}\left(\sum_j A_j + A\right) + \hat{q}^2 O(1)\right)\cdot\left(\left|\ln\frac{\bar{P}(y)}{\bar{Q}(y)} - \frac{\mu^2}{2\sigma^2}\right|^3\right.\right. \\ &\qquad\left.\left. +3\hat{q}\left(\ln\frac{\bar{P}(y)}{\bar{Q}(y)} - \frac{\mu^2}{2\sigma^2}\right)^2\left|\sum_j A_j + A - \sum_j B_j + O(\sigma^{-2})\right| + \hat{q}^2O(1)\right)\right].\end{aligned}$$

We split this into T_1 and T_2.

Part T_1.

$$\begin{aligned}
T_1 &= \mathbb{E}_{y\sim\bar{P}}\left[\left(1+\hat{q}\left(\sum_j A_j + A\right)+\hat{q}^2 O(1)\right)\left|\ln\frac{\bar{P}(y)}{\bar{Q}(y)}-\frac{\mu^2}{2\sigma^2}\right|^3\right] \\
&= \mathbb{E}_{y\sim\bar{P}}\left|\ln\frac{\bar{P}(y)}{\bar{Q}(y)}-\frac{\mu^2}{2\sigma^2}\right|^3 + \hat{q}\cdot O(\sigma^{-5}) + O(\hat{q}^2) \\
&= \sqrt{\tfrac{8}{\pi}}\,\frac{\mu^3}{\sigma^3} + \hat{q}\cdot O(\sigma^{-5}) + O(\hat{q}^2). \qquad (24)
\end{aligned}$$

Part T_2. Applying the triangle inequality and retaining terms of order $\hat{q}/\sigma^3$,

$$\begin{aligned}
T_2 &= \mathbb{E}_{y\sim\bar{P}}\left[(1+\hat{q}O(1))\cdot 3\hat{q}\left(\ln\frac{\bar{P}(y)}{\bar{Q}(y)}-\frac{\mu^2}{2\sigma^2}\right)^2\left|\sum_j A_j + A - \sum_j B_j + O(\sigma^{-2})\right|\right] \\
&= \mathbb{E}_{y\sim\bar{P}}\left[3\hat{q}\left(\ln\frac{\bar{P}(y)}{\bar{Q}(y)}-\frac{\mu^2}{2\sigma^2}\right)^2\left|\sum_j A_j + A - \sum_j B_j + O(\sigma^{-2})\right|\right] + O(\hat{q}^2) \\
&\le \mathbb{E}_{y\sim\bar{P}}\left[3\hat{q}\,\frac{\mu^2(y-\sum_j\mu_j-\mu)^2}{\sigma^4}\left|-\frac{\mu(y-\sum_j\mu_j-\mu)}{\sigma^2}\right|\right] + \hat{q}\cdot O(\sigma^{-4}) + O(\hat{q}^2) \\
&= \frac{3\hat{q}\,\mu^3\,\mathbb{E}_{y\sim\bar{P}}|y-\sum_j\mu_j-\mu|^3}{\sigma^6} + \hat{q}\cdot O(\sigma^{-4}) + O(\hat{q}^2) \\
&= 3\hat{q}\sqrt{\tfrac{8}{\pi}}\,\frac{\mu^3}{\sigma^3} + \hat{q}\cdot O(\sigma^{-4}) + O(\hat{q}^2). \qquad (25)
\end{aligned}$$

Combining (24) and (25),

$$\begin{aligned}
\mathbb{E}_P\left|\ln\frac{P}{Q}-\mathbb{E}_P\left[\ln\frac{P}{Q}\right]\right|^3 \le T_1+T_2 &\le (1+3\hat{q})\sqrt{\tfrac{8}{\pi}}\,\frac{\mu^3}{\sigma^3} + \hat{q}\cdot O(\sigma^{-4}) + O(\hat{q}^2) \\
&= (2-q^3)\sqrt{\tfrac{8}{\pi}}\,\frac{\mu^3}{\sigma^3} + \hat{q}\cdot O(\sigma^{-4}) + O(\hat{q}^2).
\end{aligned}$$

The case of $\mathbb{E}_Q|\ln(P/Q)-\mathbb{E}_Q[\ln(P/Q)]|^3$ is identical. □

G Further Monte Carlo Simulations for the Approximate GDP Filter

G.1 Conversion of RDP Order 1 to (ε,δ)-DP

The standard algebraic conversion from RDP to (ε,δ)-DP (Canonne et al., 2020) is undefined at RDP order 1, because the formula involves a factor that diverges there. This is not merely an algebraic singularity — the underlying proof breaks down at order 1. We adapt that proof to derive a conversion dedicated to RDP order 1.

We note that this algebraic conversion is not tight. Optimal conversion is available through the Bernoulli privacy-region characterization of Riess et al. (2026). The algebraic conversion is nonetheless a standard choice in the literature, and we follow that convention here.

Theorem G.1. *If a mechanism satisfies* $(1,B)$*-RDP, then for all* $\varepsilon>0$ *it satisfies* (ε,δ)*-DP with* $\delta=\min\{1,cB\}$*, where* $c>0$ *is the unique solution to*

$$c\Big(e^{1/c}-1\Big)=e^{\varepsilon}.$$

Proof. We find the smallest c such that

$$\big[1-e^{\varepsilon-l}\big]_+ \le c\big(l-1+e^{-l}\big) \quad \text{for all } l\in\mathbb{R}. \qquad (26)$$

The right-hand side is nonnegative since $e^{-l} \geq 1 - l$. This inequality implies that for every remove pair (P, Q) and PLRV $L(y) = \ln(P(y)/Q(y))$,

$$\mathbb{E}_{y\sim P}\left[\left[1 - e^{\varepsilon - L(y)}\right]_+\right] \leq c \cdot \mathbb{E}_{y\sim P}\left[L(y) - 1 + e^{-L(y)}\right] = c \cdot \mathbb{E}_{y\sim P}[L(y)] \leq c \cdot B.$$

Here $\mathbb{E}_{y\sim P}[e^{-L(y)}] = 1$, since the finite RDP bound in both directions implies that P and Q are mutually absolutely continuous. Taking the supremum over all neighboring pairs, the LHS becomes the smallest δ for which the mechanism is (ε, δ)-DP.

Note that the smallest c such that (26) holds is

$$c = \max_{l>\epsilon} \frac{1 - e^{\varepsilon - l}}{l - 1 + e^{-l}}.$$

The inner ratio tends to zero as $l \searrow \varepsilon$ or $l \nearrow \infty$, and its derivative has the sign of $1 + e^{\varepsilon} l - e^{l}$. It therefore has a unique maximum, characterized by the system

$$1 - e^{\varepsilon - l} = c\big(l - 1 + e^{-l}\big), \qquad e^{\varepsilon - l} = c\big(1 - e^{-l}\big). \tag{27}$$

These equations give $l = 1/c$ and hence $c(e^{1/c} - 1) = e^{\varepsilon}$ with $c > 0$. □

G.2 Details of Monte Carlo Simulations

We simulate Algorithm 2 on a one-dimensional remove pair with independent Poisson subsampling at a fixed rate q. Conditional on the released history, the removed datapoint contributes $\mu_t = \min(C_t, C_{B,t})$, giving

$$P_t = (1 - q)\mathcal{N}(0, \sigma_t^2 C_t^2) + q\mathcal{N}(\mu_t, \sigma_t^2 C_t^2), \qquad Q_t = \mathcal{N}(0, \sigma_t^2 C_t^2),$$

and

$$L_t(y_t) = \ln\left[1 - q + q \exp\left(\frac{\mu_t(2y_t - \mu_t)}{2\sigma_t^2 C_t^2}\right)\right]. \tag{28}$$

The first step uses $(\sigma_1, C_1) = (\sigma_0, C_0)$. For $t \geq 2$, we use the AdaGrad-inspired rule (Duchi et al., 2011)

$$\sigma_t = \operatorname{clamp}\left(\frac{\sigma_0}{\sqrt{\sum_{i<t} y_i^2}}; [\sigma_{\min}, \sigma_0]\right), \qquad C_t = \frac{C_0}{\sqrt{\sum_{i<t} y_i^2}},$$

with noise multipliers $\sigma_0 = 8$ and $\sigma_{\min} = 2$, and $C_0 = 1$. The declining C_t models decreasing update norms during training, motivated by adaptive clipping (Andrew et al., 2021). Although the one-step privacy-loss distribution does not depend on C_t, clipping scales the released outputs and hence affects subsequent parameter choices.

For each of the five configurations in Table 3, we simulate $N = 10^{10}$ independent trajectories under each of P and Q. We choose T and B so that the filter exhausts its budget by T with probability at least 98% in each direction. We use $\tilde{q} = 0$ for $q \in \{0.01, 0.1, 0.199\}$ and $\tilde{q} = 1$ for $q \in \{0.801, 0.95\}$, with the budget charges and final clipping rule of Algorithm 2.

Direct privacy-curve estimation. Let L_i^P and L_i^Q denote the total privacy losses from the i-th independently simulated trajectory under the transcript laws P and Q, respectively. Using the step loss in (28), define $L_{i,t} = L_t(y_{i,t})$, so that

$$L_i = \ln(P/Q) = \sum_{t=1}^{T} L_{i,t} = \sum_{t=1}^{\tau_i} L_{i,t}.$$

Here $y_{i,t}$ is the output of trajectory i at step t, and τ_i is its budget-binding step, or T if the horizon is reached first; we set $L_{i,t} = 0$ for $t > \tau_i$. The directed hockey-stick identities yield the estimators

$$\widehat{\delta}_P(\varepsilon) = \frac{1}{N}\sum_{i=1}^{N}\left[1 - \mathrm{e}^{\varepsilon - L_i^P}\right]_+,$$

$$\widehat{\delta}_Q(\varepsilon) = \frac{1}{N}\sum_{i=1}^{N}\left[1 - \mathrm{e}^{\varepsilon + L_i^Q}\right]_+.$$

The P-expectation and Q-expectation of the respective summands give the remove- and add-direction privacy curves. We estimate the add/remove curve by $\widehat{\delta}(\varepsilon) = \max\{\widehat{\delta}_P(\varepsilon), \widehat{\delta}_Q(\varepsilon)\}$.

We construct a simultaneous confidence band using the empirical-Bernstein bound of Maurer and Pontil (2009, Theorem 4). Let $\widehat{V}_P(\varepsilon)$ and $\widehat{V}_Q(\varepsilon)$ be the sample variances of the summands defining $\widehat{\delta}_P(\varepsilon)$ and $\widehat{\delta}_Q(\varepsilon)$, respectively, each with denominator $N-1$. For total Monte Carlo failure probability $\beta = 10^{-9}$ and at most $M = 5 \times 2 \times 700{,}000$ candidate ε values on the evaluation grid across all five q values and both directions, the radii are

$$r_P(\varepsilon) = \sqrt{\frac{2\widehat{V}_P(\varepsilon)\ln(4M/\beta)}{N}} + \frac{7\ln(4M/\beta)}{3(N-1)},$$
$$r_Q(\varepsilon) = \sqrt{\frac{2\widehat{V}_Q(\varepsilon)\ln(4M/\beta)}{N}} + \frac{7\ln(4M/\beta)}{3(N-1)}.$$

By a union bound, with probability at least $1-\beta$, both bounds

$$\left|\widehat{\delta}_P(\varepsilon) - \delta_P(\varepsilon)\right| \le r_P(\varepsilon), \qquad \left|\widehat{\delta}_Q(\varepsilon) - \delta_Q(\varepsilon)\right| \le r_Q(\varepsilon)$$

hold simultaneously at every candidate ε across all five q values. The shaded confidence band extends from $\max\{0, \widehat{\delta}_P - r_P, \widehat{\delta}_Q - r_Q\}$ to $\min\{1, \max[\widehat{\delta}_P + r_P, \widehat{\delta}_Q + r_Q]\}$. We plot the EVR-style pessimistic estimate (Wang et al., 2023)

$$\min\Big\{1, \max\Big[\widehat{\delta}_P(\varepsilon) + r_P(\varepsilon), \widehat{\delta}_Q(\varepsilon) + r_Q(\varepsilon)\Big] + \beta\Big\}.$$

The candidate ε values are the positive points of a fixed uniform grid dividing $[B - 15\sqrt{2B}, B + 20\sqrt{2B}]$ into 700,000 equal intervals.

Gaussian comparison and results. We compare the empirical CDFs of L under P and Q with $\mathcal{N}(B, 2B)$ and $\mathcal{N}(-B, 2B)$, respectively. Writing $\widehat{F}_P, \widehat{F}_Q$ for the empirical CDFs, the reported discrepancies are

$$\widehat{\Delta}_P = \max_x \left|\widehat{F}_P(x) - \Phi\left(\frac{x-B}{\sqrt{2B}}\right)\right|,$$
$$\widehat{\Delta}_Q = \max_x \left|\widehat{F}_Q(x) - \Phi\left(\frac{x+B}{\sqrt{2B}}\right)\right|,$$
$$\widehat{\Delta} = \max(\widehat{\Delta}_P, \widehat{\Delta}_Q),$$

where x ranges over the evaluated points. The privacy-curve panels compare the simulation with the exact $\sqrt{2B}$-GDP curve used when applying Theorem 4.2 in practice with Δ omitted, and with the matched $(1, B)$-RDP baseline from Appendix G.1.

Despite output-dependent noise and clipping, the empirical PLRV means and variances remain close to the Gaussian predictions $\pm B$ and $2B$ (Table 2), supporting the moment approximations underlying the filter. At $\delta = 10^{-6}$, the empirical ε is approximately 2%–11% higher than the $\sqrt{2B}$-GDP prediction across the five settings; for $q = 0.1$, the values are $\varepsilon \approx 2.81$ and $\varepsilon \approx 2.56$, respectively.

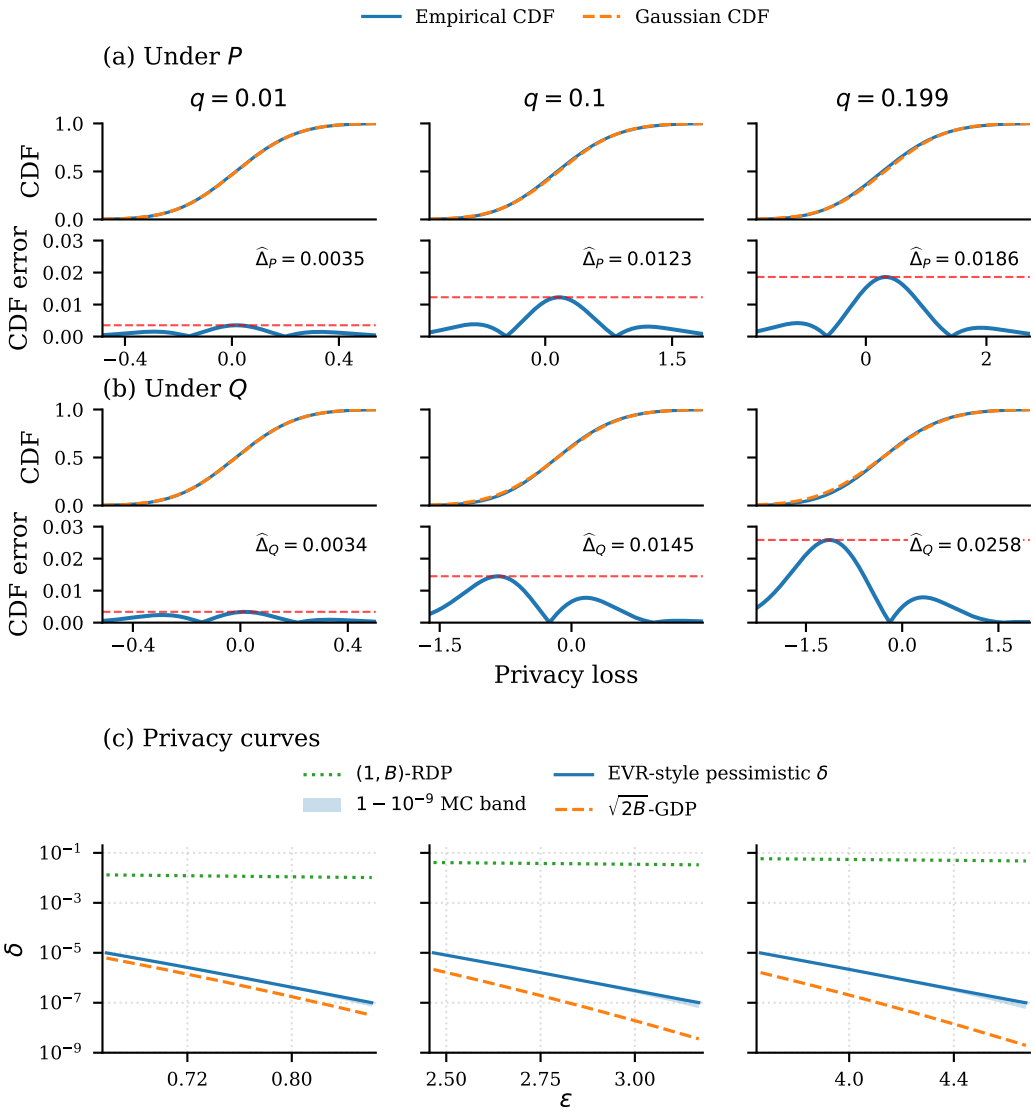


Figure 11: Monte Carlo evaluation in the low-subsampling regime, with 10^{10} trajectories per direction at each q (Table 3). **(a) Under P:** empirical CDF of L versus $\mathcal{N}(B, 2B)$ and absolute CDF discrepancy. **(b) Under Q:** the corresponding comparison with $\mathcal{N}(-B, 2B)$. **(c) Privacy curves:** the EVR-style pessimistic add/remove δ estimate and simultaneous $1 - 10^{-9}$ confidence band compared with $\sqrt{2B}$-GDP with Δ omitted and the matched $(1, B)$-RDP baseline. At $q = 0.01$, halting rates under (P, Q) are $(100.00\%, 100.00\%)$, with conditional mean stopping steps $(1111.2, 1111.2)$ or $(11.11, 11.11)$ epochs. At $q = 0.1$, halting rates under (P, Q) are $(98.87\%, 98.80\%)$, with conditional mean stopping steps $(113.1, 113.1)$ or $(11.31, 11.31)$ epochs. At $q = 0.199$, halting rates under (P, Q) are $(99.24\%, 99.13\%)$, with conditional mean stopping steps $(58.0, 58.0)$ or $(11.53, 11.54)$ epochs.

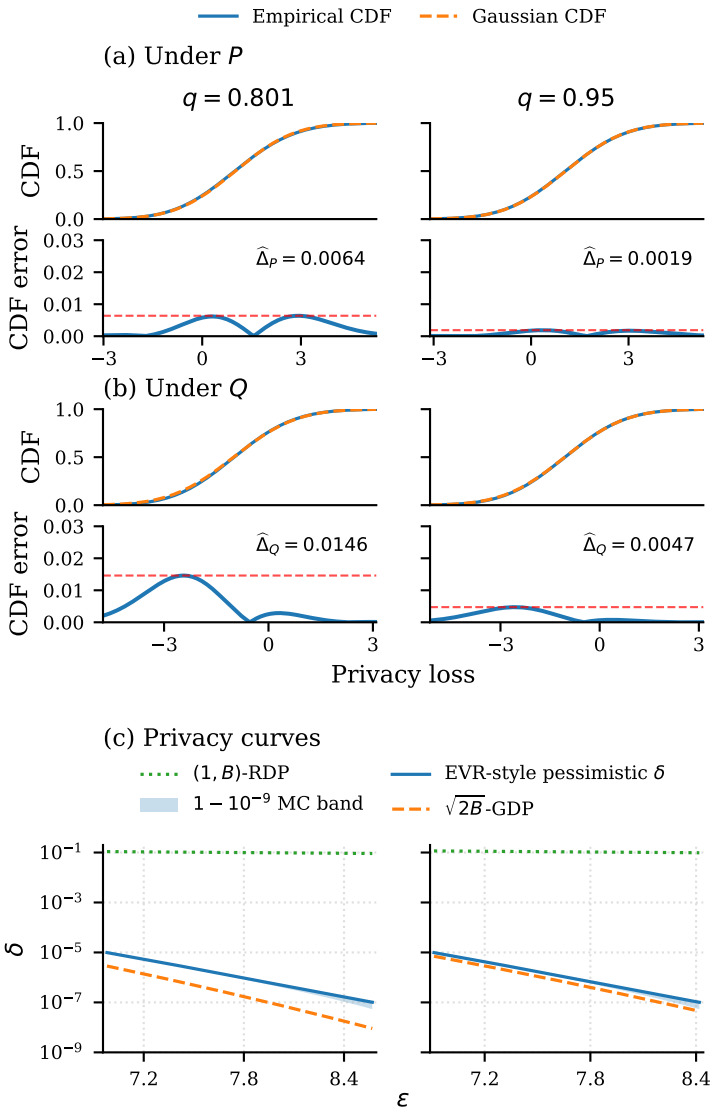


Figure 12: Monte Carlo evaluation in the high-subsampling regime, with 10^{10} trajectories per direction at each q (Table 3). **(a) Under P:** empirical CDF of L versus $\mathcal{N}(B, 2B)$ and absolute CDF discrepancy. **(b) Under Q:** the corresponding comparison with $\mathcal{N}(-B, 2B)$. **(c) Privacy curves:** the EVR-style pessimistic add/remove δ estimate and simultaneous $1 - 10^{-9}$ confidence band compared with $\sqrt{2B}$-GDP with Δ omitted and the matched $(1, B)$-RDP baseline. At $q = 0.801$, halting rates under (P, Q) are $(98.61\%, 98.06\%)$, with conditional mean stopping steps $(14.8, 14.8)$ or $(11.85, 11.88)$ epochs. At $q = 0.95$, halting rates under (P, Q) are $(99.98\%, 99.97\%)$, with conditional mean stopping steps $(11.8, 11.9)$ or $(11.22, 11.29)$ epochs.

Table 2: Empirical PLRV moments and the Gaussian reference. The reference mean is $+B$ under P and $-B$ under Q, with variance $2B$ under both.

		Gaussian reference		Empirical under P		Empirical under Q	
q	B	Mean	Variance	Mean	Variance	Mean	Variance
0.01	0.01575	± 0.01575	0.0315	0.0157	0.0315	-0.01565	0.03121
0.1	0.1575	± 0.1575	0.315	0.1532	0.3142	-0.1494	0.2913
0.199	0.3118	± 0.3118	0.6236	0.2967	0.6187	-0.2846	0.5464
0.801	1.003	± 1.003	2.005	1.006	2.087	-0.9693	1.866
0.95	1.058	± 1.058	2.115	1.058	2.139	-1.046	2.066

Table 3: Simulation configurations and empirical CDF discrepancies. The overall discrepancy is $\widehat{\Delta} = \max(\widehat{\Delta}_P, \widehat{\Delta}_Q)$.

q	T	B	$\widehat{\Delta}_P$	$\widehat{\Delta}_Q$	$\widehat{\Delta}$	$\sqrt{2B}$
0.01	1200	0.01575	0.003533	0.003377	0.003533	0.1775
0.1	120	0.1575	0.01228	0.0145	0.0145	0.5612
0.199	65	0.3118	0.01862	0.02584	0.02584	0.7897
0.801	20	1.003	0.006391	0.01461	0.01461	1.416
0.95	20	1.058	0.001889	0.004735	0.004735	1.454

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction each state the four contributions, which correspond directly to Corollary 3.3, Theorems 3.5 and 3.6, Theorem 4.1, and Algorithm 2 respectively. The suggested general methodology in Contribution 3 is framed as a consequence of the CLT rather than a proven general result, so no overclaiming occurs.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed throughout the paper and summarized in Section 5: the approximate GDP filter underestimates privacy loss at realistic subsampling rates, a known limitation of CLT-based Gaussian approximations acknowledged explicitly in the abstract, Figure 2, and Table 1; and the filter's validity guarantees are restricted to the asymptotic regimes $q \ll 1$ and $q \approx 1$, with the gap to exact accounting quantified by Δ in Theorem 4.2.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theorems state assumptions explicitly in their statement; full proofs appear in the appendices with each theorem cross-referenced to the appendix where the proof lies. Proof sketches and intuition are provided in the main text alongside each result, and intermediate lemmas are stated and proved in the appendix to patch any gaps in the main arguments.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All mechanism parameters and integral formulas for the counterexample are given in Appendix B. Appendix G specifies the Monte Carlo composition rule, five configurations, 10^{10} independent trajectories per direction and configuration, direct privacy-curve estimators, and simultaneous empirical-Bernstein confidence bands.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper does not provide a public code or data link. The counterexample is specified in Appendix B, and the Monte Carlo policy, parameters, estimators and confidence-band construction are specified in Appendix G.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper contains no model training. Appendix B gives the counterexample parameters and privacy-profile integrals. Appendix G gives the adaptive noise and clipping rule, horizons and budgets, subsampling rates, sample sizes, and privacy-curve estimation procedure.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Figure 2 and the privacy-curve panels in Appendix G report simultaneous $1 - 10^{-9}$ confidence bands for Monte Carlo sampling uncertainty across the candidate ε values, all five configurations and both add/remove directions. Appendix G gives the empirical-Bernstein bounds and union-bound allocation. The counterexample privacy profiles are computed by deterministic numerical integration.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper does not report the hardware, memory requirements or execution times for the 10^{10}-trajectory Monte Carlo simulations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper is purely theoretical work on privacy accounting and raises no ethical concerns.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The work improves the theoretical foundations of differential privacy accounting, enabling tighter and more accurate privacy guarantees for sensitive data processing — a direct positive societal impact. No negative societal impacts are foreseen, as the work strengthens rather than weakens privacy protections and has no direct path to misuse.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no models, datasets, or other artifacts that pose misuse risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper uses no existing assets such as code, data, or models.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper releases no new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLM assistance with manuscript preparation, theoretical analysis, and numerical experiments is described in the acknowledgements.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.