Shifted Interpolation for Differential Privacy

Jinho Bok [1] **Weijie J. Su** [1] **Jason Altschuler** [1]

Abstract

Noisy gradient descent and its variants are the predominant algorithms for differentially private machine learning. It is a fundamental question to quantify their privacy leakage, yet tight characterizations remain open even in the foundational setting of convex losses. This paper improves over previous analyses by establishing (and refining) the "privacy amplification by iteration" phenomenon in the unifying framework of f-differential privacy—which tightly captures all aspects of the privacy loss and immediately implies tighter privacy accounting in other notions of differential privacy, e.g., (ε, δ)-DP and Rényi DP. Our key technical insight is the construction of *shifted interpolated processes* that unravel the popular shifted-divergences argument, enabling generalizations beyond divergence-based relaxations of DP. Notably, this leads to the first *exact* privacy analysis in the foundational setting of strongly convex optimization. Our techniques extend to many settings: convex/strongly convex, constrained/unconstrained, full/cyclic/stochastic batches, and all combinations thereof. As an immediate corollary, we recover the f-DP characterization of the exponential mechanism for strongly convex optimization in Gopi et al. (2022), and moreover extend this result to more general settings.

1. Introduction

Private optimization is the primary approach for private machine learning. The goal is to train good models while not leaking sensitive attributes of the training data. Differential privacy (DP) is the gold standard for measuring information leakage (Dwork et al., 2006; Dwork & Roth, 2014), and noisy gradient descent and its variants are the predominant algorithms for private optimization. It is therefore a central question to quantify the differential privacy of these algorithms. However, tight characterizations remain open, even in the seemingly simple setting of convex optimization.

In words, DP measures how distinguishable the output of a (randomized) algorithm is when run on two adjacent datasets, i.e., two datasets that differ in only one individual record. There are several ways to measure distinguishability—leading to many relaxations of DP, e.g., (Bun & Steinke, 2016; Mironov, 2017; Dong et al., 2022). Different DP notions lead to different privacy analyses, and a long line of work has sought to prove sharp privacy bounds for noisy gradient descent and its variants (Bassily et al., 2014; Abadi et al., 2016; Feldman et al., 2018; Chourasia et al., 2021; Ye & Shokri, 2022; Altschuler & Talwar, 2022).

A common approach is to use the composition theorem, which essentially pays a price in privacy for every intermediate iterate along the optimization trajectory, leading to possibly suboptimal privacy bounds. Recent work has significantly improved the privacy analysis in the case of convex and strongly convex losses by leveraging stability properties of (stochastic) gradient descent (Chourasia et al., 2021; Ye & Shokri, 2022; Altschuler & Talwar, 2022) in order to show that the privacy leakage does not increase ad infinitum in the number of iterations t. This is in stark contrast to the composition-based approach, which gives privacy bounds that scale as $\sqrt{t}$.

All these "convergent" privacy bounds were proved in the Rényi DP framework, which is inherently lossy. To achieve the tightest possible privacy bound on private gradient methods, a natural goal is to use the f-DP framework (Dong et al., 2022) for analysis, since it is an information-theoretically lossless definition of DP. This definition measures distinguishability in terms of the Type I vs Type II error tradeoff curve f for the hypothesis testing problem of whether a given user was in the training dataset. The f-DP framework is desirable because: (1) f-DP exactly characterizes all relevant aspects of the hypothesis testing problem defining DP, and thus (optimal) f-DP bounds can be losslessly converted to (optimal) bounds in other notions of privacy such as (ε, δ)-DP or Rényi DP, (2) f-DP is lossless under composition of multiple private mechanisms, which is the most ubiquitous operation in DP since it enables combining

[1] Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA, USA. Correspondence to: Jinho Bok <jinhobok@upenn.edu>.

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

building blocks, and (3) f-DP is easily interpretable in terms of the original hypothesis testing definition of DP.

However, analyzing privacy leakage in the f-DP framework is often challenging since quantifying the entire tradeoff between Type I/II error is substantially more difficult than quantifying (less informative) alternative notions of privacy. Consequently, the analysis toolbox for f-DP is currently limited. These limitations are pronounced for the fundamental problem of analyzing the privacy loss of noisy gradient descent and its variants. To put it into perspective, existing privacy guarantees based on f-DP *diverge* as the number of iterations t increases, whereas the aforementioned recent work has used divergence-based DP definitions to show that for convex problems, noisy gradient descent and its variants can remain private even when run indefinitely (Chourasia et al., 2021; Ye & Shokri, 2022; Altschuler & Talwar, 2022). Convergent privacy bounds complement celebrated results for minimax-optimal privacy-utility tradeoffs (Bassily et al., 2014; 2019) because they enable longer training—which is useful since typical learning problems are not worst-case and benefit from training longer.

Can convergent privacy bounds be achieved directly[1] in the tight framework of f-DP? All current arguments are tailored to Rényi DP—an analytically convenient but inherently lossy relaxation of DP—and do not appear to extend. Answering this question necessitates developing fundamentally different techniques for f-DP, since convergent privacy bounds require only releasing the algorithm's *final* iterate—in sharp contrast to existing f-DP techniques such as the composition theorem which can only argue about the accumulated privacy loss of releasing *all* intermediate iterates. Tight f-DP analyses typically require closed-form expressions for the random variable in question—in order to argue about the tradeoff of Type I/II error—but this is impossible for the final iterate of (stochastic) gradient descent due to the non-linearity intrinsic to each iteration.

1.1. Contribution

Our primary technical contribution is establishing (and refining) the "privacy amplification by iteration" phenomenon in the unifying framework of f-DP. This enables directly analyzing the privacy loss of the final iterate of noisy gradient descent (and its variants), leading to the first direct f-DP analysis that is convergent as the number of iterations $t \to \infty$. §1.2 overviews this new analysis technique.

Notably, this leads to the first *exact* privacy analysis in the foundational setting of strongly convex losses. To our knowledge, there is no other setting where exact convergent privacy analyses are known for any $t > 1$, except for the setting of convex quadratic losses which is analytically trivial because all iterates are explicit Gaussians.[2]

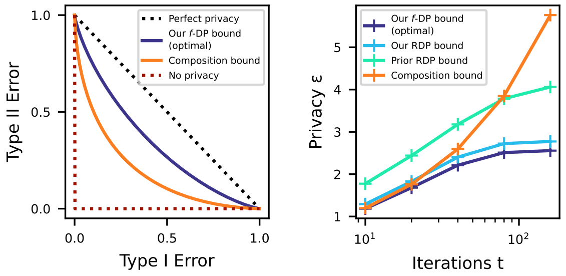

Figure 1. Left: improved f-DP vs the standard composition analysis. Right: improved (ε, δ)-DP by losslessly converting from f-DP. Our privacy bound is optimal in all parameters, here for `NoisyGD` on strongly convex losses; see §D for the parameter choices and other settings. Our f-DP analysis also implies optimal bounds for the Rényi DP framework (previously unknown), but f-DP is strictly better since it captures all aspects of the privacy leakage, whereas Rényi DP is intrinsically lossy.

We emphasize that our techniques are versatile and readily extend to many settings—a well-known challenge for other convergent analyses, even for simpler relaxations of DP like Rényi DP (Chourasia et al., 2021; Ye & Shokri, 2022; Altschuler & Talwar, 2022). In §4, we illustrate how our analysis extends to convex/strongly convex losses, constrained/unconstrained optimization, full/cyclic/stochastic batches, and all combinations thereof.

Since our improved privacy guarantees are for f-DP (Figure 1, left), lossless conversions immediately imply improved guarantees for other notions of privacy like Rényi DP and (ε, δ)-DP (Figure 1, right). For example, for the strongly convex setting, our exact bound improves over previous results by a factor of 2 in Rényi DP, and thus by even more in (ε, δ)-DP due to the intrinsic lossiness of Rényi DP that we overcome by directly analyzing in f-DP. In practice, improving the privacy by a factor of two enables training with half the noise, while satisfying the same privacy budget. Although this paper's focus is the theoretical methodology, preliminary numerics in §4.4 corroborate that our improved privacy guarantees can be helpful in practice.

Since our privacy bounds are convergent in the number of iterations t, we can take the limit $t \to \infty$ to bound the f-DP of the stationary distributions of these optimization algorithms. As an immediate corollary, we recover the recent f-DP characterization of the exponential mechanism for

[1] Convergent RDP bounds can of course be *lossily* converted to convergent f-DP bounds, but that defeats the purpose of using the lossless f-DP framework.

[2] The standard analysis approach based on the composition theorem is nearly tight for small numbers of iterations t, but as mentioned above, yields an arbitrarily loose bound (in fact vacuous) for convex losses as $t \to \infty$.

strongly convex losses in Gopi et al. (2022), and moreover extend this result to more general settings in §5.

1.2. Techniques

The core innovation underlying our results is the construction of certain auxiliary processes, *shifted interpolated processes*, which enable directly analyzing the Type I/II error tradeoff between the final iterates of two stochastic processes—even when their laws are complicated and non-explicit. Informally, this argument enables running *coupling arguments*—traditionally possible only for Wasserstein analysis—to analyze tradeoff functions for the first time. In this paper, the two processes are noisy (stochastic, projected) gradient descent run on two adjacent datasets, but the technique is more general and we believe may be of independent interest. See §3 for a detailed discussion.

Crucially, our argument is *geometrically aware*: it exploits (strong) convexity of losses via (strong) contractivity of gradient descent updates, in order to argue that the sensitive gradient queries have (exponentially) decaying privacy leakage, the longer ago they were performed. This is essential for convergent privacy bounds, and is impossible with the standard composition-based analysis—which only exploits the sensitivity of the losses, and is oblivious to any further geometric phenomena like convexity or contractivity.

A key motivation behind the construction of our auxiliary sequence is that it demystifies the popular privacy amplification by iteration analysis (Feldman et al., 2018), which has been used in many contexts, and in particular was recently shown to give convergent Rényi DP bounds (Altschuler & Talwar, 2022). Those arguments rely on *shifted divergences*, which combine Rényi divergence and Wasserstein distance, and it was an open question whether this ad-hoc potential function could be simplified. Our shifted interpolated process answers this: its iterates coincide with the optimal "shifts" in the shifted divergence argument, which allows us to disentangle the Rényi and Wasserstein components of the shifted divergence argument; details in §B. Crucially, this disentanglement enables generalizations beyond divergence-based relaxations of DP, to f-DP.[3]

1.3. Organization

§2 recalls relevant preliminaries from differential privacy and convex optimization. §3 introduces our core technique of shifted interpolation. §4 uses this technique to establish improved privacy bounds for noisy gradient descent and its variants for the foundational settings of convex and strongly convex losses. §5 describes how, as immediate corollaries of these convergent privacy bounds, taking an appropriate limit recovers and generalizes recent results on the f-DP of the exponential mechanism. §6 discusses future directions motivated by our results. Code reproducing our numerics can be found here: `https://github.com/jinhobok/shifted_interpolation_dp`.

2. Preliminaries

2.1. Differential privacy

DP measures the distinguishability between outputs of a randomized algorithm run on adjacent datasets, i.e., datasets that differ on at most one data point (Dwork et al., 2006). The most popular definition is (ε, δ)-DP.

Definition 2.1 ((ε, δ)-DP)**.** A randomized algorithm $\mathcal{A}$ is (ε, δ)-*DP* if for any adjacent datasets S, S' and event E,

$$\mathbb{P}(\mathcal{A}(S) \in E) \leq e^{\varepsilon}\mathbb{P}(\mathcal{A}(S') \in E) + \delta\,.$$

However, the most precise quantification of DP is based on the (optimal) hypothesis testing formulation (Wasserman & Zhou, 2010; Kairouz et al., 2017). This is formalized as f-DP (Dong et al., 2022), where f denotes a tradeoff function, i.e., a curve of hypothesis testing errors.

Definition 2.2 (f-DP)**.** For distributions P, Q on the same space, the *tradeoff function* $T(P, Q) : [0, 1] \to [0, 1]$ is

$$T(P, Q)(\alpha) = \inf\{1 - \mathbb{E}_Q\phi : \mathbb{E}_P\phi \leq \alpha, 0 \leq \phi \leq 1\}$$

A randomized algorithm $\mathcal{A}$ is f-*DP* if for any adjacent datasets S and S', $T(\mathcal{A}(S), \mathcal{A}(S')) \geq f$.

The following lemma provides a useful characterization of tradeoff functions (Dong et al., 2022, Proposition 1). It follows that the most private tradeoff function is $\mathrm{Id} : [0, 1] \to [0, 1]$, given by $\mathrm{Id}(\alpha) = 1 - \alpha$. See Figure 2.

Lemma 2.3 (Characterization of tradeoff functions)**.** *A function $f : [0, 1] \to [0, 1]$ is a tradeoff function iff f is decreasing, convex and $f(\alpha) \leq 1 - \alpha$ for all $\alpha \in [0, 1]$.*

See §A.2 for further details on tradeoff functions. Gaussian tradeoff functions are a particularly useful family, providing a notion of Gaussian DP (GDP) parametrized by a single scalar. These are central to our analysis due to the Gaussian noise in noisy (stochastic) gradient descent.

Definition 2.4 (GDP)**.** For GDP parameter $\mu \geq 0$, the *Gaussian tradeoff function* $G(\mu)$ is defined as $G(\mu) = T(\mathcal{N}(0, 1), \mathcal{N}(\mu, 1))$. Its value at $\alpha \in [0, 1]$ is given as $G(\mu)(\alpha) = \Phi(\Phi^{-1}(1 - \alpha) - \mu)$, where Φ denotes the CDF of $\mathcal{N}(0, 1)$. A randomized algorithm $\mathcal{A}$ is μ-*GDP* if for any adjacent datasets S and S', $T(\mathcal{A}(S), \mathcal{A}(S')) \geq G(\mu)$.

[3] Naïvely extending the "shifted divergence" argument to "shifted tradeoffs" runs into subtle but fundamental issues since tradeoff functions do not enjoy key properties that divergences do. Details in §B.

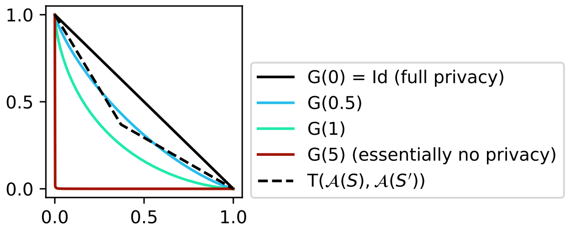

Figure 2. Illustration of f-DP and GDP. Gaussian tradeoff functions $G(\mu)$ are less private as μ increases from 0 (full privacy) to ∞ (no privacy). The closer to Id, the more private. Here $\mathcal{A}$ is 1-GDP but not 0.5-GDP because its tradeoff function is pointwise above $G(1)$ but not pointwise above $G(0.5)$.

We now recall two key properties of tradeoff functions that are central to our analysis. The first states that post-processing two distributions in the same way cannot make them easier to distinguish (Dong et al., 2022, Lemma 1).

Lemma 2.5 (Post-processing)**.** *For any probability distributions P, Q and (random) map Proc, we have $T(Proc(P), Proc(Q)) \geq T(P, Q)$.*

The next lemma enables analyzing the composition of multiple private mechanisms (Dong et al., 2022, Definition 5 & Lemma C.1).

Definition 2.6 (Composition)**.** The *composition* of two tradeoff functions $f = T(P, Q)$ and $g = T(P', Q')$ is defined as $f \otimes g = T(P \times P', Q \times Q')$. The n-fold composition of f with itself is denoted $f^{\otimes n}$.

Lemma 2.7 (Strong composition)**.** *Let K_1, K_1', K_2, K_2' be (random) maps such that for all y, $T(K_1(y), K_1'(y)) \geq T(K_2(y), K_2'(y))$. Then $T((P, K_1(P)), (Q, K_1'(Q))) \geq T((P, K_2(P)), (Q, K_2'(Q)))$. If $g = T(K_2(y), K_2'(y))$ does not depend on y, then $T((P, K_1(P)), (Q, K_1'(Q))) \geq T(P, Q) \otimes g$.*

2.2. Convex optimization

This paper focuses on convex losses because tight privacy guarantees for noisy gradient descent (and variants) are open even in this seemingly simple setting. We make use of the following two basic facts from convex optimization. Below, we say a function is contractive if it is 1-Lipschitz. Recall that a function f is M-smooth if ∇f is M-Lipschitz, and is m-strongly convex if $x \mapsto f(x) - \frac{m}{2}\|x\|^2$ is convex.

Lemma 2.8. *If f is convex and M-smooth, then the gradient descent update $g(x) = x - \eta \nabla f(x)$ is contractive for each $\eta \in [0, \frac{2}{M}]$. If f is additionally m-strongly convex and $\eta \in (0, \frac{2}{M})$, then g is c-Lipschitz where $c = \max\{|1 - \eta m|, |1 - \eta M|\} < 1$.*

Lemma 2.9. *Let $\mathcal{K}$ be a closed and convex set in $\mathbb{R}^d$. Then the projection $\Pi_{\mathcal{K}}(x) = \arg\min_{z \in \mathcal{K}} \|z - x\|$ is well-defined and is a contraction.*

2.3. Algorithms

Throughout, we consider a private optimization setting in which the goal is to minimize the objective function $F(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$, where the i-th loss function f_i is associated with the i-th data point in a dataset S. An adjacent dataset S' corresponds to loss functions $\{f_i'\}_{i \in [n]}$ where $f_i \equiv f_i'$ except for a single index i^*.

Noisy gradient descent and its variants follow the general template of

$$X_{k+1} \leftarrow \Pi_{\mathcal{K}}\left[X_k - \eta\left(\frac{1}{b}\sum_{i \in B_k} \nabla f_i(X_k) + Z_{k+1}\right)\right], \quad k = 0, 1, \ldots, t-1 \qquad (1)$$

where X_0 is the initialization (e.g., zero), η is the learning rate, $Z_{k+1} \sim \mathcal{N}(0, \sigma^2 I_d)$ independently, σ is the noise rate, $\mathcal{K}$ is the constraint set, and t is the number of steps. The batch B_k of size b can be chosen in several ways:

- Full batches (`NoisyGD`): $B_k \equiv [n]$.
- Cyclic batches (`NoisyCGD`): Partition $[n]$ into batches of sizes b and cycle through them.
- Stochastic batches (`NoisySGD`): Choose a batch of size b uniformly at random from $[n]$.

The advantage of the latter two variants is that they avoid computing the gradient of the objective, which can be computationally burdensome when n is large.

A standard assumption in private optimization is the following notion of gradient sensitivity:

Definition 2.10 (Gradient sensitivity)**.** A family of loss functions $\mathcal{F}$ (defined on $\mathcal{X}$) has *gradient sensitivity* L if $\sup_{f,g \in \mathcal{F}, x \in \mathcal{X}} \|\nabla f(x) - \nabla g(x)\| \leq L$.

For example, a family of L-Lipschitz loss functions has gradient sensitivity $2L$. Another example is loss functions of the form $f_i = \ell_i + r$, where ℓ_i are convex, L-Lipschitz losses, and r is a (non-Lipschitz) strongly convex regularization—the point being that this family of loss functions $\{f_i\}$ has finite gradient sensitivity $2L$ despite each f_i not being Lipschitz.

3. Shifted interpolation for f-DP

Here we explain the key conceptual ideas enabling our convergent f-DP bounds (see §1.2 for a high-level overview). To preserve the logical flow of ideas we defer proofs to §C.1. Below, in §3.1 we first recall the standard f-DP analysis based on the composition theorem and why it yields divergent bounds. Then in §3.2 we describe our technique of shifted interpolated processes and how this enables convergent f-DP bounds.

To explain the ideas in their simplest form, we consider here the setting of full-batch gradients and unconstrained optimization. Let $\{f_i\}_{i\in[n]}$ and $\{f'_i\}_{i\in[n]}$ be the losses corresponding to two adjacent datasets, where $f_i \equiv f'_i$ except for one index i^*. Then `NoisyGD` forms the iterates

$$X_{k+1} = \phi(X_k) + Z_{k+1} \tag{2}$$
$$X'_{k+1} = \phi'(X'_k) + Z'_{k+1} \tag{3}$$

where $X_0 = X'_0$, $\phi(x) := x - \frac{\eta}{n}\sum_{i=1}^n \nabla f_i(x)$, $\phi'(x') := x' - \frac{\eta}{n}\sum_{i=1}^n \nabla f'_i(x')$, and $Z_{k+1}, Z'_{k+1} \sim \mathcal{N}(0, \eta^2\sigma^2 I_d)$.

3.1. Previous (divergent) f-DP bounds, via composition

f-DP requires bounding $T(X_t, X'_t)$. The standard approach, based on the composition theorem, argues as follows:

$$\begin{aligned} T(X_t, X'_t) &\geq T(X_{t-1}, X'_{t-1}) \otimes G(c) \\ &\geq T(X_{t-2}, X'_{t-2}) \otimes G(c\sqrt{2}) \\ &\cdots \\ &\geq \underbrace{T(X_0, X'_0)}_{=\text{Id since } X_0 = X'_0} \otimes G(c\sqrt{t})\,. \end{aligned} \tag{4}$$

Here, the composition theorem simultaneously "unrolls" both processes, at some price $G(c)$ in each iteration. (These prices are collected via a basic GDP identity, see Lemma A.2.) This is due to the following simple lemma, which relies on the f-DP of the Gaussian mechanism using different updates ϕ, ϕ' (Dong et al., 2022, Theorem 2).

Lemma 3.1. *Suppose* $\|\phi(x) - \phi'(x)\| \leq s$ *for all* x. *Then* $T(\phi(X) + \mathcal{N}(0,\sigma^2 I_d), \phi'(X') + \mathcal{N}(0,\sigma^2 I_d)) \geq T(X, X') \otimes G(\frac{s}{\sigma})$.

Bounding s via sensitivity enables the argument (4) and gives the appropriate c. See Dong et al. (2022) for details.

However, while this argument (4) is reasonably tight for small t, it is vacuous as $t \to \infty$. Conceptually, this is because this analysis considers releasing all intermediate iterates, hence it bounds $T((X_1,\ldots,X_t),(X'_1,\ldots X'_t)) \geq G(c\sqrt{t})$. Concretely, this is because the above analysis requires *completely* unrolling to iteration 0. Indeed, the identical initialization $X_0 = X'_0$ ensures $T(X_0, X'_0) = \text{Id}$, whereas at any other iteration $k > 0$ it is unclear how to directly bound $T(X_k, X'_k)$ as $X_k \neq X'_k$. This inevitably leads to final privacy bounds which *diverge* in t since a penalty is incurred in each of the t iterations.

3.2. Convergent f-DP bounds, via shifted interpolation

The central idea underlying our analysis is the construction of a certain *auxiliary process* $\{\widetilde{X}_k\}$ that interpolates between the two processes in the sense that $\widetilde{X}_\tau = X'_\tau$ at some intermediate time τ and $\widetilde{X}_t = X_t$ at the final time. See Figure 3. Crucially, this enables running the argument (5) where we unroll only from t to τ, rather than all the way to initialization:

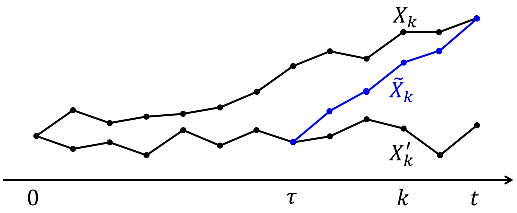

Figure 3. Illustration of shifted interpolated process (6). The interpolated process $\{\widetilde{X}_k\}$ starts from one process ($\widetilde{X}_\tau = X'_\tau$) and ends at the other ($\widetilde{X}_t = X_t$). The intermediate time τ is an analysis parameter that we optimize to get the best final privacy bound.

$$\begin{aligned} T(X_t, X'_t) &= T(\widetilde{X}_t, X'_t) \\ &\geq T(\widetilde{X}_{t-1}, X'_{t-1}) \otimes G(a_t) \\ &\geq T(\widetilde{X}_{t-2}, X'_{t-2}) \otimes G\big(\big(a_t^2 + a_{t-1}^2\big)^{1/2}\big) \\ &\cdots \\ &\geq \underbrace{T(\widetilde{X}_\tau, X'_\tau)}_{=\text{Id since } \widetilde{X}_\tau = X'_\tau} \otimes G\Big(\Big(\sum_{k=\tau+1}^{t} a_k^2\Big)^{1/2}\Big)\,. \end{aligned} \tag{5}$$

Intuitively, this argument replaces the divergent $\sqrt{t}$ dependence of prior f-DP bounds with something scaling in $t - \tau$. Here τ is an analysis parameter that we can optimize based on the following intuitive tradeoff: larger τ enables unrolling less, whereas smaller τ gives the auxiliary process $\widetilde{X}_k$ more time to interpolate between X'_τ and X_t which leads to smaller penalties a_k for unrolling at each iteration.

Formalizing (5) leads to two interconnected questions:

- **Q1.** How to construct the auxiliary process $\{\widetilde{X}_k\}$?
- **Q2.** How to unroll each iteration? I.e., what is the analog of Lemma 3.1?

3.2.1. Shifted interpolated process

For Q1, we initialize $\widetilde{X}_\tau = X'_\tau$ and define

$$\widetilde{X}_{k+1} = \lambda_{k+1}\phi(X_k) + (1 - \lambda_{k+1})\phi'(\widetilde{X}_k) + Z_{k+1} \tag{6}$$

for $k = \tau, \ldots, t-1$. Intuitively, this auxiliary process $\{X_k\}$ uses a convex combination of the updates performed by the two processes, enabling it to gracefully interpolate from its initialization at one process to its termination at the other. Here λ_k controls the speed at which we *shift* from one process to the other. We set $\lambda_t = 1$ so that $\widetilde{X}_t = X_t$ achieves the desired interpolation; the other $\{\lambda_k\}$ are analysis parameters that we optimize to get the best final bound. An important technical remark is that this auxiliary process uses the same noise increments $\{Z_k\}$ as $\{X_k\}$; this coupling enables bounding the distance between X_k and $\widetilde{X}_k$ by a deterministic value (i.e., in the ∞-Wasserstein distance W_∞).

We remark that auxiliary interpolating processes have been used in the context of proving Harnack inequalities (or equivalently, Rényi reverse transport inequalities) for diffusions on manifolds (Arnaudon et al., 2006; Wang, 2013; 2014; Altschuler & Chewi, 2023b;c). Two key challenges posed by the present setting are that f-DP requires tradeoff functions (rather than Rényi divergences), and also tracking stochastic processes that undergo *different* dynamics (rather than the same diffusion). This requires constructing and analyzing the auxiliary process (6).

3.2.2. Geometrically aware composition

For Q2, we develop the following lemma, which generalizes Lemma 3.1 by allowing for an auxiliary process $\widetilde{X}$ and a shift parameter λ (Lemma 3.1 is recovered in the special case $\lambda = 1$ and $\widetilde{X} = X$). A key feature is that unlike Lemma 3.1, this lemma is *geometrically aware* in that it exploits the Lipschitzness of the gradient descent updates ϕ, ϕ'—recall from Lemma 2.8 that ϕ, ϕ' are (strongly) contractive whenever the losses are (strongly) convex. Intuitively, this contractivity ensures that long-ago gradient queries incur (exponentially) less privacy loss, thus making the total privacy loss convergent; c.f., the discussion in §1.2.

Lemma 3.2. *Suppose that $\|\phi(x) - \phi'(x)\| \le s$ for all x and that ϕ, ϕ' are c-Lipschitz. Then for any $\lambda \ge 0$ and any random variable $\widetilde{X}$ satisfying $\|X - \widetilde{X}\| \le z$, $T(\lambda\phi(X) + (1-\lambda)\phi'(\widetilde{X}) + \mathcal{N}(0,\sigma^2 I_d), \phi'(X') + \mathcal{N}(0,\sigma^2 I_d)) \ge T(\widetilde{X}, X') \otimes G(\frac{\lambda(cz+s)}{\sigma})$.*

3.2.3. Convergent f-DP bounds

Combining our answers to Q1 (shifted interpolated process) and Q2 (geometrically aware composition) enables formalizing the argument (5). The remaining proof details are straightforward and deferred to §C.2. For clarity, we state this result as a "meta-theorem" where the shifts λ_k and intermediate time τ are parameters; our final bounds are obtained by optimizing them, see §4.

Theorem 3.3. *Consider the stochastic processes $\{X_k\}, \{X'_k\}, \{\widetilde{X}_k\}$ defined in (2), (3), (6), with $\lambda_t = 1$. Suppose that $\|\phi(x) - \phi'(x)\| \le s$ for all x and that ϕ, ϕ' are c-Lipschitz. For any sequence $\{z_k\}$ such that $\|X_k - \widetilde{X}_k\| \le z_k$,*

$$T(X_t, X'_t) \ge G\left(\frac{1}{\sigma}\sqrt{\sum_{k=\tau+1}^{t} a_k^2}\right)$$

where $a_k = \lambda_k(cz_{k-1} + s)$.

We emphasize that although this technique-overview section focused on the simple case of full-batch gradients and strongly convex losses for clarity, these techniques readily extend to more general settings. Briefly, for constrained optimization, projections are handled by using the post-processing inequality for tradeoff functions; for (non-strongly) convex optimization, the optimal shifts a_k will be of similar size rather than geometrically increasing; for cyclic batches, the update functions ϕ_k, ϕ'_k and corresponding sensitivity s_k are time-varying; and for stochastic batches, the analog of Lemma 3.2 incorporates the celebrated privacy amplification by subsampling phenomenon. Details in §4.

4. Improved privacy for noisy optimization algorithms

Here we apply the shifted interpolation technique developed in §3 to establish improved privacy bounds for noisy gradient descent and its variants. We showcase the versatility of our techniques by investigating gradient descent with full-batch gradients in §4.1, cyclic batches in §4.2, and stochastic batches in §4.3. In all cases, we show convergent f-DP bounds for unconstrained strongly convex and constrained convex settings; the constrained strongly convex setting is similar and omitted for brevity (and the unconstrained convex setting has divergent privacy). The proofs are similar for all these different settings, based on the approach in §3; for brevity the proofs are deferred to §C. See also §D for numerical illustrations of the improvements of our bounds.

Below, recall from §2 that we denote the learning rate by η, the noise rate by σ, the number of data points by n, the batch size by b, the constraint set by $\mathcal{K}$, and its diameter by D. Throughout we denote by $c = \max\{|1-\eta m|, |1-\eta M|\}$ the Lipschitz constant for a step of gradient descent on m-strongly convex and M-smooth losses (c.f., Lemma 2.8).

4.1. Noisy gradient descent

Here we consider full-batch gradient descent. For comparison, we first recall the standard f-DP bound implied by the composition theorem (Dong et al., 2022).

Theorem 4.1. *Consider loss functions with gradient sensitivity L. Then `NoisyGD` is μ-GDP where $\mu = \frac{L}{n\sigma}\sqrt{t}$.*

This (divergent) bound is tight without further assumptions on the losses. Below we show convergent f-DP bounds for `NoisyGD` in the setting of strongly convex losses, and the setting of constrained convex losses.

Theorem 4.2. *Consider m-strongly convex, M-smooth loss functions with gradient sensitivity L. Then for any $\eta \in$*

$(0, 2/M)$, `NoisyGD` is μ-GDP where

$$\mu = \sqrt{\frac{1-c^t}{1+c^t}\frac{1+c}{1-c}}\frac{L}{n\sigma}.$$

For $\eta \in (0, \frac{2}{M+m}]$, this bound is optimal.

Theorem 4.3. *Consider convex, M-smooth loss functions with gradient sensitivity L and constraint set $\mathcal{K}$ of diameter D. Then for any $\eta \in [0, 2/M]$ and $t \geq \frac{Dn}{\eta L}$, `NoisyGD` is μ-GDP where*

$$\mu = \frac{1}{\sigma}\sqrt{\frac{3LD}{\eta n} + \frac{L^2}{n^2}\left\lceil\frac{Dn}{\eta L}\right\rceil}.$$

Theorem 4.2 is exactly tight in all parameters, and improves over the composition-based analysis (Theorem 4.1) for all $t > 1$. Theorem 4.3 is tight up to a constant factor (see Theorem C.16), and for $t > \frac{4Dn}{\eta L}$ it dominates Theorem 4.1 since its convergent nature outweighs the slightly suboptimal constant.

4.2. Noisy cyclic gradient descent

We now turn to cyclic batches. For simplicity, suppose that the number of batches per epoch $l = n/b$ and the number of epochs $E = t/l$ are integers. We state our results with respect to E rather than t. For comparison, we first state the standard (divergent) f-DP bound implied by the composition theorem (Dong et al., 2022).

Theorem 4.4. *Consider loss functions with gradient sensitivity L. Then `NoisyCGD` is μ-GDP where $\mu = \frac{L}{b\sigma}\sqrt{E}$.*

Below we show convergent f-DP bounds for `NoisyCGD` in the setting of strongly convex losses, and the setting of constrained convex losses.

Theorem 4.5. *Consider m-strongly convex, M-smooth loss functions with gradient sensitivity L. Then for any $\eta \in (0, 2/M)$, `NoisyCGD` is μ-GDP where*

$$\mu = \sqrt{1 + c^{2l-2}\frac{1-c^2}{(1-c^l)^2}\frac{1-c^{l(E-1)}}{1+c^{l(E-1)}}}\frac{L}{b\sigma}.$$

Theorem 4.6. *Consider convex, M-smooth loss functions with gradient sensitivity L and constraint set $\mathcal{K}$ of diameter D. Then for any $\eta \in [0, 2/M]$ and $E \geq \frac{Db}{\eta L}$, `NoisyCGD` is μ-GDP where*

$$\mu = \frac{1}{\sigma}\sqrt{\frac{3LD}{\eta bl} + \left(\frac{L}{b}\right)^2 + \frac{L^2}{b^2 l}\left\lceil\frac{Db}{\eta L}\right\rceil}.$$

The convergent nature of these bounds ensures that they dominate Theorem 4.4 when `NoisyCGD` is run long enough. This threshold is roughly $E \approx c^{2l-2}\frac{1-c^2}{(1-c^l)^2}$ for Theorem 4.5 and $E \approx \frac{4Db}{\eta \ell L}$ for Theorem 4.6.

4.3. Noisy stochastic gradient descent

Compared to `NoisyGD`, the privacy leakage in `NoisySGD` only occurs when the index i^* is in the sampled batch. This phenomenon is known as privacy amplification by subsampling (Kasiviswanathan et al., 2011), which is formulated in f-DP as follows (Dong et al., 2022, Definition 6).

Definition 4.7. For tradeoff function f and $p \in [0, 1]$, define $f_p = pf + (1-p)\mathrm{Id}$. The *subsampling operator* C_p (with respect to f) is defined as $C_p(f) = \min\{f_p, (f_p)^{-1}\}^{**}$ where $^{-1}$ denotes the (left-continuous) inverse and * denotes the convex conjugate.

For comparison, we first recall the standard f-DP bound based on composition (Dong et al., 2022, Theorem 9).

Theorem 4.8. *Consider loss functions with gradient sensitivity L. Then `NoisySGD` is f-DP where $f = C_{b/n}(G(\frac{L}{b\sigma}))^{\otimes t}$.*

This (divergent) bound is tight for $t = 1$ without further assumptions on the losses. Below we show convergent f-DP bounds for `NoisySGD` in the setting of strongly convex losses, and the setting of constrained convex losses.

Theorem 4.9. *Consider m-strongly convex, M-smooth loss functions with gradient sensitivity L. Then for any $\eta \in (0, 2/M)$, `NoisySGD` is f-DP for*

$$f = G(\frac{2\sqrt{2}L}{b\sigma}\frac{c^{t-\tau+1} - c^t}{1-c}) \\ \otimes C_{b/n}(G(\frac{2\sqrt{2}L}{b\sigma})) \otimes C_{b/n}(G(\frac{2L}{b\sigma}))^{\otimes(t-\tau)}$$

for any $\tau = 0, 1, \ldots, t-1$.

Theorem 4.10. *Consider convex, M-smooth loss functions with gradient sensitivity L and constraint set $\mathcal{K}$ of diameter D. Then for any $\eta \in [0, 2/M]$, `NoisySGD` is f-DP where*

$$f = (\frac{\sqrt{2}D}{\eta\sigma\sqrt{t-\tau}}) \otimes C_{b/n}(G(\frac{2\sqrt{2}L}{b\sigma}))^{\otimes(t-\tau)}$$

for any $\tau = 0, 1, \ldots, t-1$.

Both theorems give convergent privacy by taking $t - \tau$ constant as $t \to \infty$. In contrast, Theorem 4.8 is convergent in the regime $t = O(\frac{n^2}{b^2})$, but yields a vacuous privacy as $t \to \infty$ for fixed $\frac{b}{n}$ (Dong et al., 2022). We remark that for finite but large t, one can set $t - \tau$ to be sufficiently large and apply CLT (Lemma A.5) to approximate the composition of $C_p(G(\cdot))$; see Lemma A.11 and §C.4.3. We also remark that by choosing $t - \tau = \Theta(\frac{Dn}{\eta L})$, Theorem 4.10 recovers the asymptotically tight Rényi DP bound of (Altschuler & Talwar, 2022).

4.4. Numerical example

As a proof of concept, here we consider regularized logistic regression on MNIST (LeCun et al., 2010). We compare our results with the state-of-the-art Rényi DP bounds, and existing f-DP bounds (based on the composition theorem) which we denote as GDP Composition. For a fair comparison, we use the same algorithm `NoisyCGD`, with all parameters unchanged, and only focus on the privacy accounting. Table 1 demonstrates that for this problem, our privacy guarantees are tighter, enabling longer training for the same privacy budget—which helps both training and testing accuracy (c.f., Table 2). For full details of the experiment, see §D.2.

Table 1. Privacy ε of `NoisyCGD` on regularized logistic regression for $\delta = 10^{-5}$ in (ε, δ)-DP. Our results provide better privacy than both GDP Composition and RDP bounds in all cases.

Epochs	GDP Composition	RDP	Our Bounds
50	30.51	5.82	4.34
100	49.88	7.61	5.60
200	83.83	9.88	7.58

Table 2. Training and test accuracy (%) of `NoisyCGD` for regularized logistic regression, averaged over 10 runs. Both the training and test accuracy improve as the number of epochs increases.

Epochs	Training	Test
50	89.36 ± 0.03	90.12 ± 0.04
100	90.24 ± 0.03	90.94 ± 0.07
200	90.85 ± 0.02	91.37 ± 0.08

5. f-DP of the exponential mechanism

Since we show convergent f-DP bounds for randomized algorithms, we can take the limit $t \to \infty$ to obtain f-DP bounds for their stationary distributions. We focus here on `NoisyGD` because, up to a simple rescaling, it is equivalent to Langevin Monte Carlo (`LMC`), one of the most well-studied sampling algorithms in the statistics literature; see, e.g., (Robert et al., 1999; Liu, 2001; Andrieu et al., 2003). Our results for (strongly) convex losses not only imply new results for (strongly) log-concave sampling for `LMC`, but also imply f-DP bounds for the exponential mechanism (McSherry & Talwar, 2007)—a foundational concept in DP—since it is obtained from `LMC`'s stationary distribution in the limit as the stepsize $\eta \to 0$.

5.1. Strongly log-concave targets

Our optimal f-DP bounds for `NoisyGD` immediately imply optimal[4] f-DP bounds for `LMC`.

Proposition 5.1. *Suppose that F, F' are m-strongly convex, M-smooth and $F - F'$ is L-Lipschitz. Consider the* `LMC` *updates*

$$X_{k+1} = X_k - \eta \nabla F(X_k) + Z_{k+1}$$
$$X'_{k+1} = X'_k - \eta \nabla F'(X'_k) + Z'_{k+1}$$

where $X_0 = X'_0$ and $Z_{k+1}, Z'_{k+1} \sim \mathcal{N}(0, 2\eta I_d)$. Then for any $\eta \in (0, \frac{2}{M+m}]$, $T(X_t, X'_t) \geq G\left(\sqrt{\frac{2-\eta m}{2}} \frac{L}{\sqrt{m}}\right)$.

Proof. `LMC` is a special case of `NoisyGD` with $n = 1$, $f_1 = F, f'_1 = F'$, and $\sigma = \sqrt{2/\eta}$. Apply Theorem 4.2. □

Taking $t \to \infty$ gives f-DP guarantees for the stationary distributions $\pi(\eta)$ and $\pi'(\eta)$ of these `LMC` chains. We also obtain f-DP guarantees between the exponential mechanisms $\pi \propto e^{-F}$ and $\pi' \propto e^{-F'}$ for F and F'.

Corollary 5.2. *In the setting of Proposition 5.1, $T(\pi(\eta), \pi'(\eta)) \geq G\left(\sqrt{\frac{2-\eta m}{2}} \frac{L}{\sqrt{m}}\right)$ and $T(\pi, \pi') \geq G\left(\frac{L}{\sqrt{m}}\right)$.*

Proof. It is well-known that under these assumptions, LMC converges to its stationary distribution in TV as $t \to \infty$, and the stationary distribution converges to the exponential mechanism as $\eta \to 0$, see e.g., (Chewi, 2023). By Lemma A.10, tradeoff functions converge under TV. □

Thus, we recover the recent result (Gopi et al., 2022, Theorem 4) which characterizes the f-DP of the exponential mechanism. The proof in (Gopi et al., 2022) is entirely different, based on the Gaussian isoperimetry inequality (Ledoux, 1999) rather than connecting `LMC` to the exponential mechanism. Our results can be viewed as algorithmic generalizations of theirs in the sense that we also obtain tight f-DP bounds on the iterates of `LMC` and its stationary distribution.

Remark 5.3 (Tightness). As noted in (Gopi et al., 2022), the exponential mechanism bound in Corollary 5.2 is tight by considering $F(x) = \frac{m}{2}\|x\|^2, F'(x) = \frac{m}{2}\|x - \frac{L}{m}v\|^2$ (where v is a unit vector) which yields $\pi = \mathcal{N}(0, \frac{1}{m}I_d), \pi' = \mathcal{N}(\frac{L}{m}v, \frac{1}{m}I_d)$. With the same loss functions, it is straightforward to check that this construction also shows optimality for our results on the f-DP of `LMC` and its stationary distribution.

5.2. Log-concave targets

A similar story holds in the setting of convex losses, although this requires a constrained setting since otherwise stationary distributions may not exist. Hence we consider projected `NoisyGD` (Theorem 4.3), which corresponds to projected `LMC`. As above, this leads to f-DP bounds for the exponential mechanism due to known TV convergence

[4] Although here we bound the optimal constants for simplicity.

results, for projected LMC to its stationary distribution as $t \to \infty$ (Altschuler & Talwar, 2023), and from that distribution to the exponential mechanism as $\eta \to 0$ (Bubeck et al., 2018).

Corollary 5.4. *Let F, F' be convex, M-smooth and L-Lipschitz functions and $\mathcal{K}$ be a convex body with diameter D containing a unit ball. Then for $\pi \propto e^{-F}\mathbf{1}_{\mathcal{K}}$ and $\pi' \propto e^{-F'}\mathbf{1}_{\mathcal{K}}$, $T(\pi, \pi') \geq G(2\sqrt{LD})$.*
Furthermore, for $\eta \in (0, 2/M]$, the respective stationary distributions $\pi(\eta), \pi'(\eta)$ of the projected `LMC` *satisfy $T(\pi(\eta), \pi'(\eta)) \geq G(\sqrt{4LD + 2\eta L^2})$.*

Unlike the strongly convex case (Minami et al., 2016; Gopi et al., 2022), we are unaware of any results in this setting beyond the standard analysis (McSherry & Talwar, 2007) on the exponential mechanism. That yields $(2LD, 0)$-DP, and our result provides nontrivial improvement in privacy when $LD > 0.677$. See §D.4.

6. Discussion

The techniques and results of this paper suggest several directions for future work. One natural direction is whether convergent f-DP bounds can be shown in more general settings, e.g., (structured) non-convex landscapes, heteroscedastic or correlated noise (Choquette-Choo et al., 2023), adaptive first-order algorithms, or second-order algorithms (Ganesh et al., 2023). A technical question is whether one can relax the W_∞ bounds between our shifted interpolated process $\{\widetilde{X}_k\}$ and the target process $\{X_k\}$, and if this can enable tighter analyses of stochastic algorithms. While W_∞ has traditionally been used for privacy amplification by iteration (Feldman et al., 2018), (Altschuler & Chewi, 2023a) recently showed that some of this analysis extends to the Orlicz–Wasserstein distance, which is even necessary in some applications. Another natural direction is more computationally tractable f-DP bounds. Although the f-DP framework provides an information-theoretically lossless quantification of DP, it is often computationally burdensome, e.g., for `NoisySGD` bounds expressed as the composition of many tradeoff functions. Recent work has developed useful tools for approximation (Zheng et al., 2020; Gopi et al., 2021; Zhu et al., 2022), and further developments would help practitioners who need to adhere to given privacy budgets.

Acknowledgements

We thank Sinho Chewi, Kunal Talwar, and Jiayuan Ye for insightful conversations about the literature and the anonymous reviewers for helpful comments. We also thank Jiayuan Ye for sharing helpful code for numerical experiments.

Impact Statement

This paper establishes better privacy guarantees for learning from sensitive user data. Therefore we believe that the broader societal implications can only be positive. Indeed, we are proposing to use the same standard learning algorithms, but now with improved privacy guarantees.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Conference on Computer and Communications Security*, pp. 308–318, 2016.

Altschuler, J. and Talwar, K. Privacy of noisy stochastic gradient descent: More iterations without more privacy loss. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3788–3800, 2022.

Altschuler, J. and Talwar, K. Resolving the mixing time of the Langevin algorithm to its stationary distribution for log-concave sampling. In *Conference on Learning Theory*, volume 195, pp. 2509–2510. PMLR, 2023.

Altschuler, J. M. and Chewi, S. Faster high-accuracy log-concave sampling via algorithmic warm starts. In *Symposium on Foundations of Computer Science (FOCS)*, pp. 2169–2176, 2023a.

Altschuler, J. M. and Chewi, S. Shifted composition I: Harnack and reverse transport inequalities. *arXiv preprint arXiv:2311.14520*, 2023b.

Altschuler, J. M. and Chewi, S. Shifted composition II: Shift Harnack inequalities and curvature upper bounds. *arXiv preprint arXiv:2401.00071*, 2023c.

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43, 2003.

Arnaudon, M., Thalmaier, A., and Wang, F.-Y. Harnack inequality and heat kernel estimates on manifolds with curvature unbounded below. *Bulletin des Sciences Mathématiques*, 130(3):223–233, 2006.

Asoodeh, S., Liao, J., Calmon, F. P., Kosut, O., and Sankar, L. Three variants of differential privacy: Lossless conversion and applications. *IEEE Journal on Selected Areas in Information Theory*, 2(1):208–222, 2021.

Awan, J. and Dong, J. Log-concave and multivariate canonical noise distributions for differential privacy. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34229–34240, 2022.

Balle, B. and Wang, Y.-X. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, volume 80, pp. 394–403. PMLR, 2018.

Balle, B., Barthe, G., Gaboardi, M., Hsu, J., and Sato, T. Hypothesis testing interpretations and Renyi differential privacy. In *International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 2496–2506. PMLR, 2020.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Symposium on Foundations of Computer Science (FOCS)*, pp. 464–473, 2014.

Bassily, R., Feldman, V., Talwar, K., and Guha Thakurta, A. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Bu, Z., Dong, J., Long, Q., and Su, W. Deep learning with Gaussian differential privacy. *Harvard Data Science Review*, 2(3), 2020.

Bubeck, S., Eldan, R., and Lehec, J. Sampling from a log-concave distribution with projected Langevin Monte Carlo. *Discrete & Computational Geometry*, 59(4):757–783, 2018.

Bun, M. and Steinke, T. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *International Conference on Theory of Cryptography*, pp. 635–658, 2016.

Chewi, S. *Log-concave sampling*. Forthcoming, 2023. Draft available at https://chewisinho.github.io/.

Choquette-Choo, C. A., Dvijotham, K., Pillutla, K., Ganesh, A., Steinke, T., and Thakurta, A. Correlated noise provably beats independent noise for differentially private learning. *arXiv preprint arXiv:2310.06771*, 2023.

Chourasia, R., Ye, J., and Shokri, R. Differential privacy dynamics of Langevin diffusion and noisy gradient descent. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14771–14781, 2021.

Dong, J., Roth, A., and Su, W. J. Gaussian differential privacy. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 84(1):3–54, 2022. With discussions and a reply by the authors.

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

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, volume 3876 of *Lecture Notes in Computuer Science*, pp. 265–284. Springer, Berlin, 2006.

Feldman, V., Mironov, I., Talwar, K., and Thakurta, A. Privacy amplification by iteration. In *Symposium on Foundations of Computer Science (FOCS)*, pp. 521–532. 2018.

Ganesh, A., Haghifam, M., Steinke, T., and Thakurta, A. Faster differentially private convex optimization via second-order methods. *arXiv preprint arXiv:2305.13209*, 2023.

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

Gopi, S., Lee, Y. T., and Liu, D. Private convex optimization via exponential mechanism. In *Conference on Learning Theory*, volume 178, pp. 1948–1989. PMLR, 2022.

Kairouz, P., Oh, S., and Viswanath, P. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049, 2017.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

LeCun, Y., Cortes, C., and Burges, C. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Ledoux, M. Concentration of measure and logarithmic Sobolev inequalities. In Azéma, J., Émery, M., Ledoux, M., and Yor, M. (eds.), *Séminaire de Probabilités XXXIII*, pp. 120–216, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

Liu, J. S. *Monte Carlo strategies in scientific computing*, volume 75. Springer, 2001.

McSherry, F. and Talwar, K. Mechanism design via differential privacy. In *Foundations of Computer Science (FOCS)*, pp. 94–103, 2007.

Minami, K., Arai, H., Sato, I., and Nakagawa, H. Differential privacy without sensitivity. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Mironov, I. Rényi differential privacy. In *Computer Security Foundations Symposium*, pp. 263–275, 2017.

Robert, C. P., Casella, G., and Casella, G. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

Wang, C., Su, B., Ye, J., Shokri, R., and Su, W. Unified enhancement of privacy bounds for mixture mechanisms via f-differential privacy. In *Advances in Neural Information Processing Systems*, volume 36, pp. 55051–55063, 2023.

Wang, F.-Y. *Harnack inequalities for stochastic partial differential equations*, volume 1332. Springer, 2013.

Wang, F.-Y. *Analysis for diffusion processes on Riemannian manifolds*, volume 18. World Scientific, 2014.

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

Ye, J. and Shokri, R. Differentially private learning needs hidden state (or much faster convergence). In *Advances in Neural Information Processing Systems*, volume 35, pp. 703–715, 2022.

Zheng, Q., Dong, J., Long, Q., and Su, W. Sharp composition bounds for Gaussian differential privacy via Edgeworth expansion. In *International Conference on Machine Learning*, volume 119, pp. 11420–11435. PMLR, 2020.

Zhu, Y., Dong, J., and Wang, Y.-X. Optimal accounting of differential privacy via characteristic function. In *International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 4782–4817. PMLR, 2022.

A. Rényi DP and tradeoff functions

Here we provide helper lemmas and other relevant background about Rényi DP (§A.1), tradeoff functions (§A.2), and their convergence properties (§A.3).

A.1. Rényi DP

A popular notion of DP that is often analytically tractable is Rényi DP (RDP) (Mironov, 2017).

Definition A.1 (RDP)**.** The *Rényi divergence* of order $\alpha > 1$ between probability distributions P, Q is defined as

$$D_\alpha(P \,\|\, Q) = \frac{1}{\alpha - 1} \log \int \left(\frac{dP}{dQ}(\omega) \right)^\alpha dQ(\omega)\,.$$

A randomized algorithm $\mathcal{A}$ is (α, ε)*-RDP* if for any adjacent datasets S and S',

$$D_\alpha(\mathcal{A}(S) \,\|\, \mathcal{A}(S')) \leq \varepsilon\,.$$

Numerical conversion. Conversion from RDP to (ε, δ)-DP is inherently lossy and there are many proposed formulae for this. Since the RDP bounds mentioned in this text are of the form $(\alpha, \rho\alpha)$-RDP for all $\alpha > 1$, given ρ and a fixed level of δ the corresponding converted value of $\varepsilon = \varepsilon(\alpha, \rho, \delta)$ can be found by optimizing over α. Also, in addition to RDP, results on zero concentrated DP (Bun & Steinke, 2016) can be applied. Throughout, we calculate the minimum ε (aka the best bound) among the following formulae: (Bun & Steinke, 2016, Lemma 3.5), (Mironov, 2017, Proposition 3), (Balle et al., 2020, Theorem 20), and (Asoodeh et al., 2021, Lemma 1).

A.2. Lemmas on tradeoff functions

Here we recall various useful facts about tradeoff functions. The first lemma records basic properties of tradeoff functions that we use repeatedly (Dong et al., 2022, Proposition D.1).

Lemma A.2 (Basic properties)**.** *For tradeoff functions f, g_1, g_2 and $\mu = (\mu_1, \ldots, \mu_d) \in \mathbb{R}^d$,*

(a) $g_1 \geq g_2 \Rightarrow f \otimes g_1 \geq f \otimes g_2$.

(b) $f \otimes \mathit{Id} = \mathit{Id} \otimes f = f$.

(c) $T(\mathcal{N}(0, \sigma^2 I_d), \mathcal{N}(\mu, \sigma^2 I_d)) = G(|\mu_1|/\sigma) \otimes \cdots \otimes G(|\mu_d|/\sigma) = G(\|\mu\|/\sigma)$.

Next, we recall tight conversion formulae from GDP to other standard notions of DP, namely (ε, δ)-DP (Balle & Wang, 2018, Theorem 8) and RDP (Dong et al., 2022, Corollary B.6).

Lemma A.3 (GDP to (ε, δ)-DP)**.** *A μ-GDP mechanism is $(\varepsilon, \delta(\varepsilon))$-DP for all $\varepsilon > 0$ where*

$$\delta(\varepsilon) = \Phi\left(-\frac{\varepsilon}{\mu} + \frac{\mu}{2} \right) - e^\varepsilon \Phi\left(-\frac{\varepsilon}{\mu} - \frac{\mu}{2} \right)\,.$$

Lemma A.4 (GDP to RDP)**.** *A μ-GDP mechanism is $(\alpha, \frac{1}{2}\mu^2\alpha)$-RDP for any $\alpha > 1$.*

An appealing property of tradeoff functions is that they admit a central limit theorem (CLT) that approximates multiple compositions to GDP. In particular, the subsampled GDP can be approximated as follows (Dong et al., 2022, Corollary 4).

Lemma A.5 (CLT)**.** *Let $\mu \geq 0$ and assume that $p\sqrt{t} \to p_0$ as $t \to \infty$. Then*

$$C_p(G(\mu))^{\otimes t} \to G\left(\sqrt{2} p_0 \sqrt{e^{\mu^2}\Phi(1.5\mu) + 3\Phi(-0.5\mu) - 2} \right)\,.$$

A.3. Convergence of tradeoff functions

Here we present results about the convergence of distributions as measured by tradeoff functions. The main results are Lemma A.6 and Lemma A.10, which state that this is equivalent to convergence in TV distance; we also present intermediate results which may be of independent interest. For notation, we use P_n, P, Q_n, Q to denote probability distributions, and α, α' to respectively denote elements in $[0, 1]$ and $(0, 1]$. Also, we use $a \vee b$ and $a \wedge b$ to respectively denote $\max\{a, b\}$ and $\min\{a, b\}$.

Lemma A.6. *The following are equivalent.*

(a) $T(P_n, P) \to$ *Id.*

(b) $T(P, P_n) \to$ *Id.*

(c) $TV(P_n, P) \to 0$.

Proof. On one hand, if $\mathrm{TV}(P, P_n) \to 0$, then $T(P, P_n) \to \mathrm{Id}$ since

$$1 - \mathrm{TV}(P, P_n) \leq \alpha + T(P, P_n)(\alpha) \leq 1\,.$$

On the other hand, if $\mathrm{TV}(P, P_n) \nrightarrow 0$ then by taking a subsequence $\{n'\}$ such that $\mathrm{TV}(P, P_{n'}) \geq \varepsilon > 0$ we know that the first equality holds for some $\alpha = \alpha_{n'}$ and thus

$$T(P, P_{n'})(\alpha_{n'}) \leq 1 - \varepsilon - \alpha_{n'}\,.$$

By taking a further subsequence $\{n''\}$ of $\{n'\}$ such that $\alpha_{n''} \to \alpha$ for some α (note that $\alpha_{n'} \leq 1 - \varepsilon$ for all n' and thus $\alpha \leq 1 - \varepsilon$), there exists $N \in \mathbb{N}$ such that $n'' > N \Rightarrow \alpha_{n''} < \alpha + \varepsilon/2$, from which we have

$$T(P, P_{n''})(\alpha + \frac{\varepsilon}{2}) \leq 1 - \varepsilon - \alpha_{n''}$$

for all $n'' > N$. Thus

$$\liminf_{n''} T(P, P_{n''})(\alpha + \frac{\varepsilon}{2}) \leq 1 - \frac{\varepsilon}{2} - (\alpha + \frac{\varepsilon}{2})\,,$$

implying that $T(P, P_n)$ does not converge to Id. □

Lemma A.7. *If $T(P_n, P) \to$ Id then for any probability distribution Q,*

$$\lim_n T(P_n, Q)(\alpha') = T(P, Q)(\alpha')$$

for every $\alpha' \in (0, 1]$. In particular, if $T(P, Q)(0) = 1$ then $\lim_n T(P_n, Q) = T(P, Q)$.

Proof. From (Dong et al., 2022, Lemma A.5) we have

$$\begin{aligned} T(P, Q)(\alpha') &\geq T(P_n, Q)(1 - T(P, P_n)(\alpha')) \\ T(P_n, Q)(\alpha) &\geq T(P, Q)(1 - T(P_n, P)(\alpha))\,. \end{aligned}$$

By taking $\liminf_n$ in the second line, we have $\liminf_n T(P_n, Q)(\alpha) \geq T(P, Q)(\alpha)$.

On the other hand, for any $\alpha' \in (0, 1]$ and sufficiently small $\varepsilon > 0$ we have $1 - T(P, P_n)(\alpha') \leq (\alpha' + \varepsilon) \wedge 1$ for all sufficiently large n, from which in the first line we have

$$T(P, Q)(\alpha') \geq T(P_n, Q)((\alpha' + \varepsilon) \wedge 1)\,.$$

Taking $\limsup_n$ (it is straightforward to check that a limit supremum of tradeoff function is continuous on $(0, 1)$) and letting $\varepsilon \to 0$, we have $T(P, Q)(\alpha') \geq \limsup_n T(P_n, Q)(\alpha')$. □

Remark A.8 (Necessity of the restriction on α'). The restriction $\alpha' \in (0, 1]$ is necessary. For example, if $P = \delta_0$, $Q = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, and $P_n = (1 - \frac{1}{n})\delta_0 + \frac{1}{n}\delta_1$ (here, δ_x denotes the Dirac measure at x and $pP + (1-p)Q$ denotes the mixture of (P, Q) with mixing rate $(p, 1-p)$), then $\mathrm{TV}(P_n, P) \to 0$ implies $T(P_n, P) \to \mathrm{Id}$ and $T(P, Q)(\alpha) = \frac{1}{2}(1 - \alpha)$, yet

$$T(P_n, Q)(\alpha) = \begin{cases} 1 - \frac{1}{2}n\alpha & \alpha \leq \frac{1}{n} \\ \frac{1}{2}(\frac{1-\alpha}{1-\frac{1}{n}}) & \alpha > \frac{1}{n} \end{cases} \Rightarrow \lim_n T(P_n, Q)(\alpha) = \begin{cases} 1 & \alpha = 0 \\ \frac{1}{2}(1 - \alpha) & \alpha > 0\,. \end{cases}$$

However, if the limit is switched to the second argument, then this restriction on α' simplifies and is unnecessary, as proven in the following lemma.

Lemma A.9. *If $T(Q_n, Q) \to$ Id then for any probability distribution P,*

$$\lim_n T(P, Q_n) = T(P, Q)\,.$$

Proof. Again, from (Dong et al., 2022, Lemma A.5) we have

$$\begin{aligned} T(P, Q_n)(\alpha) &\geq T(Q, Q_n)(1 - T(P, Q)(\alpha)) \\ T(P, Q)(\alpha) &\geq T(Q_n, Q)(1 - T(P, Q_n)(\alpha))\,. \end{aligned}$$

Taking $\liminf_n$ in the first line, we have $\liminf_n T(P, Q_n)(\alpha) \geq T(P, Q)(\alpha)$. On the other hand, we know that the limit $T(Q_n, Q) \to$ Id is uniform over $[0, 1]$—see, for example, (Dong et al., 2022, Lemma A.7)—and thus for any $\varepsilon > 0$ we have $T(Q_n, Q)(\alpha) \geq 1 - \alpha - \varepsilon$ for all $\alpha \in [0, 1]$ when n is sufficiently large, from which we have

$$T(P, Q)(\alpha) \geq T(P, Q_n)(\alpha) - \varepsilon\,.$$

Taking $\limsup_n$ and letting $\varepsilon \to 0$, we have $T(P, Q)(\alpha) \geq \limsup_n T(P, Q_n)(\alpha)$. □

Lemma A.10. *If TV$(P_n, P) \to 0$ and TV$(Q_n, Q) \to 0$ then*

$$\lim_n T(P_n, Q_n)(\alpha') = T(P, Q)(\alpha')$$

for every $\alpha' \in (0, 1]$.[5] *In particular, if $T(P, Q)(0) = 1$ then $\lim_n T(P_n, Q_n) = T(P, Q)$.*

Proof. From $T(P_n, Q_n)(\alpha) \geq T(P, Q_n)(1 - T(P_n, P)(\alpha))$ and $T(P, Q_n) \to T(P, Q)$ uniformly over $[0, 1]$ (by Lemma A.6 and Lemma A.9), taking $\liminf_n$ we have $\liminf_n T(P_n, Q_n)(\alpha) \geq T(P, Q)(\alpha)$.

From $T(P, Q_n)(\alpha) \geq T(P_n, Q_n)(1 - T(P, P_n)(\alpha))$, for any $\alpha' \in (0, 1]$ and sufficiently small $\varepsilon > 0$, for all sufficiently large n we have $1 - T(P, P_n)(\alpha') \leq (\alpha' + \varepsilon) \wedge 1$ and thus

$$T(P, Q_n)(\alpha') \geq T(P_n, Q_n)((\alpha' + \varepsilon) \wedge 1)\,.$$

Taking $\limsup_n$ and letting $\varepsilon \to 0$, we have $T(P, Q)(\alpha') \geq \limsup_n T(P_n, Q_n)(\alpha')$. □

The final lemma shows how composition and limit of tradeoff functions can be combined. This is useful when, for example, we have a lower bound of the form $G(\mu) \otimes g_t$, and g_t converges to $G(\nu)$ as $t \to \infty$ (e.g., by CLT), which can be approximated by the lemma as $G(\mu) \otimes g_t \approx G(\sqrt{\mu^2 + \nu^2})$.

Lemma A.11. *Let f, g, g_n be tradeoff functions such that $g(\alpha) > 0$ for all $\alpha < 1$*[6] *and $g_n \to g$. Then*

$$\liminf_n (f \otimes g_n) \geq f \otimes g\,.$$

Proof. Fix $0 < \delta < 1$, and let h_δ be the tradeoff function defined as

$$h_\delta(\alpha) = \begin{cases} 1 - \delta - \alpha & \alpha \leq 1 - \delta \\ 0 & \alpha > 1 - \delta\,. \end{cases}$$

Then it is known—see (Dong et al., 2022, Equation 12)—that for any tradeoff function f,

$$f \otimes h_\delta = \begin{cases} (1 - \delta) f(\frac{\alpha}{1-\delta}) & \alpha \leq 1 - \delta \\ 0 & \alpha > 1 - \delta\,. \end{cases}$$

[5] In (Awan & Dong, 2022), this result is stated without the restriction on $\alpha' \in (0, 1]$. However, this restriction is needed, as evidenced by the counterexample in Remark A.8.

[6] This condition is technical and is not necessary; the same proof applies by defining $r(\delta)$ as the minimum over $\alpha \in [0, z(1 - \delta)]$ where $z = \inf\{\alpha : g(\alpha) = 0\}$.

Now we approximate g by $g \otimes h_\delta$. Defining $r(\delta) = \min_{0 \le \alpha \le 1-\delta} |g(\alpha) - (g \otimes h_\delta)(\alpha)|$ (the minimum exists as the function is continuous and $[0, 1-\delta]$ is compact), we have $r(\delta) > 0$ because for any $\alpha \in [0, 1-\delta]$

$$g(\alpha) - (g \otimes h_\delta)(\alpha) = g(\alpha) - g(\frac{\alpha}{1-\delta}) + \delta g(\frac{\alpha}{1-\delta}) \ge \delta g(\frac{\alpha}{1-\delta}) \ge 0\,,$$

where the first inequality is from g decreasing; if this value is 0 then we should have $\alpha = 1-\delta$ from the second inequality, but then $g(\alpha) - g(\frac{\alpha}{1-\delta}) = g(1-\delta) - g(1) > 0$, a contradiction.

Since the limit $g_n \to g$ is uniform, for all sufficiently large n we have $g_n \ge (g - r(\delta)) \vee 0 \ge g \otimes h_\delta$, implying

$$\liminf_n (f \otimes g_n) \ge f \otimes (g \otimes h_\delta) = h_\delta \otimes (f \otimes g)\,.$$

Then from $\lim_{\delta \to 0} h_\delta \otimes (f \otimes g) = f \otimes g$, we obtain the result. □

B. Disentangling the shift in shifted divergences

As mentioned in §1.2, a key motivation behind the construction of our shifted interpolated process (6) is that it demystifies the popular privacy amplification by iteration analysis for Rényi DP (Feldman et al., 2018), which has been used in many contexts, and in particular was recently shown to give convergent Rényi DP bounds for `NoisyGD` and variants (Altschuler & Talwar, 2022). Here we explain this connection.

Briefly, privacy amplification by iteration arguments for Rényi DP use as a Lyapunov function the *shifted Rényi divergence* $D_\alpha^{(z)}(P \,\|\, Q) = \inf_{P' : W_\infty(P,P') \le z}(P' \,\|\, Q)$, which combines the Rényi divergence D_α and ∞-Wasserstein distance W_∞. (Feldman et al., 2018) bounds the Rényi DP via an argument of the form

$$\begin{aligned} D_\alpha(X_t \,\|\, X_t') &= D_\alpha^{(z_t)}(X_t \,\|\, X_t') \\ &\le D_\alpha^{(z_{t-1})}(X_{t-1} \,\|\, X_{t-1}') + O(a_t^2) \\ &\le D_\alpha^{(z_{t-2})}(X_{t-2} \,\|\, X_{t-2}') + O(a_t^2 + a_{t-1}^2) \qquad (7) \\ &\cdots \\ &\le \underbrace{D_\alpha^{(z_0)}(X_0 \,\|\, X_0')}_{=0 \text{ since } X_0 = X_0'} + O\Big(\sum_{k=1}^{t} a_k^2\Big) \qquad (8) \end{aligned}$$

where $z_t = 0$ and $z_{k+1} = cz_k + s - a_{k+1}$. (Altschuler & Talwar, 2022) obtained convergent Rényi DP bounds by essentially unrolling this argument only to an intermediate time τ, and then arguing that the shifted Rényi divergence $D_\alpha^{(z_\tau)}(X_\tau, X_\tau') = 0$ if the shift z_τ is made sufficiently large.

Several open questions remained: (1) Can this argument be performed without using shifted divergences, which is an admittedly ad-hoc combination of Rényi divergences and Wasserstein distances? (2) Can this argument be extended beyond divergence-based relaxations of DP, namely to f-DP? Our paper answers both questions.

For (1), our argument makes *explicit* the surrogates implicit in the shifted divergences

$$D_\alpha^{(z_k)}(X_k \,\|\, X_k') = \inf_{\widetilde{X}_k \,:\, W_\infty(\widetilde{X}_k, X_k) \le z_k} D_\alpha(\widetilde{X}_k \,\|\, X_k')$$

in each intermediate iteration of the argument. Indeed, it can be shown that our shifted interpolating process $\{\widetilde{X}_k\}$, defined in (6), gives such a random variable that achieves the value required by this shifted divergence argument. This enables re-writing the argument (8) without any notion of *shifted* divergences, in terms of the auxiliary process $\{\widetilde{X}_k\}$, as we did for f-DP in §3.2. This completely disentangles the Rényi divergence and Wasserstein distance in the shifted divergence argument.

For (2), the disentangling we achieve in (1) appears essential. The naïve approach of directly extending the shifted divergence argument to "shifted tradeoffs" $T^{(z)}(P, Q) = \sup_{P' : W_\infty(P,P') \le z}(P', Q)$ runs into several subtle technical issues. For example, the argument appears to require the existence of an optimal shift P'. For the shifted Rényi argument, it suffices to find a nearly-optimal shift $D_\alpha^{(z)}(P \,\|\, Q) = \inf_{P' : W_\infty(P,P') \le z} D_\alpha(P' \,\|\, Q)$, and moreover have the shift be nearly-optimal for a given Rényi parameter α but perhaps not uniformly so over all α. Due to the more involved calculus of tradeoff functions, these issues become subtle but important problems, and have led others to state the problem of privacy amplification by iteration in f-DP as open, e.g., (Wang et al., 2023). Although the general problem of finding an optimal shift for general tradeoff functions remains open, the answer to (1)—our shifted interpolated process—explicitly constructs an optimal shift for the tradeoff functions specifically needed to analyze two contractive noisy iterations.

C. Deferred details for §4

In this section we provide details for the proofs in §4. See §3 for a high-level overview of the analysis approach. We formalize the technique of shifted interpolated processes in a general context in §C.1, then prove the results of §4.1, §4.2, §4.3 in §C.2, §C.3, §C.4, respectively.

C.1. Shifted interpolation for contractive noisy iterations

We begin by providing definitions that unify the presentation of the different settings. The first definition abstracts the fundamental reason underlying why noisy gradient descent and all its variants enjoy the phenomenon of privacy amplification by iteration for convex optimization—and is why the results stated in this section are for contractive noisy iterations (CNI). This is based on the observation that the variants of noisy gradient descent update by alternately applying contraction maps and noise convolutions (Feldman et al., 2018, Definition 19).

Definition C.1 (CNI)**.** The CNI corresponding to a sequence of contractive functions $\{\phi_k\}_{k\in[t]}$, a sequence of noise distributions $\{\xi_k\}_{k\in[t]}$, and a closed and convex set $\mathcal{K}$, is the stochastic process

$$X_{k+1} = \Pi_{\mathcal{K}}(\phi_{k+1}(X_k) + Z_{k+1}) \tag{9}$$

where $Z_{k+1} \sim \xi_{k+1}$ is independent of $(X_0, \ldots, X_k)$.

Although CNI$(X_0, \{\phi_k\}_{k\in[t]}, \{\xi_k\}_{k\in[t]}, \mathcal{K})$ usually refers to the distribution of the final iterate X_t, we occasionally abuse notation by using this to refer to the entire sequence of iterates $\{X_k\}$.

The second definition abstracts the idea of shifted interpolated processes at the level of generality of CNI. See §3.2 for an informal overview.

Definition C.2 (Shifted interpolated process)**.** Consider processes $\{X_k\}$ and $\{X'_k\}$ corresponding respectively to CNI$(X_0, \{\phi_k\}_{k\in[t]}, \{\xi_k\}_{k\in[t]}, \mathcal{K})$ and CNI$(X'_0, \{\phi'_k\}_{k\in[t]}, \{\xi_k\}_{k\in[t]}, \mathcal{K})$. The *shifted interpolated process* between these two CNI is the auxiliary process $\{\widetilde{X}_k\}$ satisfying $\widetilde{X}_\tau = X'_\tau$ and

$$\widetilde{X}_{k+1} = \Pi_{\mathcal{K}}\left(\lambda_{k+1}\phi_{k+1}(X_k) + (1-\lambda_{k+1})\phi'_{k+1}(\widetilde{X}_k) + Z_{k+1}\right) \tag{10}$$

for all $k = \tau, \ldots, t-1$. Here, the noise $Z_k \sim \xi_k$ is coupled between the processes $\{X_k\}$ and $\{\widetilde{X}_k\}$. The parameters $\tau \in \{0, \ldots, t\}$, and $\lambda_k \in [0,1]$ can be chosen arbitrarily, with the one restriction that $\lambda_t = 1$ so that $\widetilde{X}_t = X_t$.

The upshot of shifted interpolation is the following meta-theorem. See §3.2 for a high-level overview of this result, its proof, and its uses. Here, we state this meta-theorem in the more general framework of CNI.

Theorem C.3 (Meta-theorem for shifted interpolation)**.** *Let X_t and X'_t respectively be the output of CNI$(X_0, \{\phi_k\}_{k\in[t]}, \{\mathcal{N}(0,\sigma^2 I_d)\}_{k\in[t]}, \mathcal{K})$ and CNI$(X_0, \{\phi'_k\}_{k\in[t]}, \{\mathcal{N}(0,\sigma^2 I_d)\}_{k\in[t]}, \mathcal{K})$ such that each ϕ_k, ϕ'_k is c-Lipschitz and $\|\phi_k(x) - \phi'_k(x)\| \leq s_k$ for all x and $k \in [t]$. Then for any intermediate time τ and shift parameters $\lambda_{\tau+1}, \ldots, \lambda_t \in [0,1]$ with $\lambda_t = 1$,*

$$T(X_t, X'_t) \geq G\left(\frac{1}{\sigma}\sqrt{\sum_{k=\tau+1}^{t} a_k^2}\right)$$

where $a_{k+1} = \lambda_{k+1}(cz_k + s_{k+1})$, $z_{k+1} = (1-\lambda_{k+1})(cz_k + s_{k+1})$, and $\|X_\tau - X'_\tau\| \leq z_\tau$.

To prove Theorem C.3, we first prove two helper lemmas. The first lemma characterizes the worst-case tradeoff function between a Gaussian and its convolution with a bounded random variable. The lemma is tight, with equality achieved when the random variable is a constant.

Lemma C.4. *For $s \geq 0$, let $R(s, \sigma) = \inf\{T(W + Z, Z) : \|W\| \leq s, Z \sim \mathcal{N}(0, \sigma^2 I_d), W, Z$ are independent$\}$, where the infimum is taken pointwise.*[7] *Then*

$$R(s, \sigma) = G(\frac{s}{\sigma}) .$$

Proof. For any random variable W with $\|W\| \leq s$, the post-processing inequality (Lemma 2.5) implies

$$T(W + Z, Z) \geq T((W, Z), (W, -W + Z)) .$$

Letting $K_1(y) = Z$ and $K_1'(y) = -y + Z$, we have $T(K_1(y), K_1'(y)) = G(\frac{\|y\|}{\sigma}) \geq G(\frac{s}{\sigma})$ for any fixed y with $\|y\| \leq s$ and thus by strong composition (Lemma 2.7),

$$T((W, Z), (W, -W + Z)) \geq T(W, W) \otimes G(\frac{s}{\sigma}) = G(\frac{s}{\sigma}) .$$

The bound is tight since equality holds with $W = sv$ for any fixed unit vector v. □

The second lemma, Lemma 3.2, is the "one-step" version of the desired result Theorem C.3. It uses the first lemma in its proof.

Proof of Lemma 3.2. For shorthand, let $Z, Z' \sim \mathcal{N}(0, \sigma^2 I_d)$ be independent. Then

$$\begin{aligned} T(\lambda\phi(X) + (1 - \lambda)\phi'(\widetilde{X}) + Z, \phi'(X') + Z') &\geq T((\widetilde{X}, \lambda(\phi(X) - \phi'(\widetilde{X})) + Z), (X', Z')) \\ &\geq T(\widetilde{X}, X') \otimes R(\lambda(cz + s), \sigma) \\ &= T(\widetilde{X}, X') \otimes G(\frac{\lambda(cz + s)}{\sigma}) . \end{aligned}$$

Above, the first step is by the post-processing inequality (Lemma 2.5) for the post-processing function $(x, y) \mapsto \phi'(x) + y$. The second step is by strong composition (Lemma 2.7), which we can apply since $\lambda(\|\phi(X) - \phi'(\widetilde{X})\|) \leq \lambda(\|\phi(X) - \phi(\widetilde{X})\| + \|\phi(\widetilde{X}) - \phi'(\widetilde{X})\|) \leq \lambda(cz + s)$. The final step is by Lemma C.4. □

Proof of Theorem C.3. Let $\{\widetilde{X}_k\}$ be as in (10). By induction, $\|X_k - \widetilde{X}_k\| \leq z_k$ for all $k = \tau, \ldots, t$ from

$$\begin{aligned} \|X_{k+1} - \widetilde{X}_{k+1}\| &\leq (1 - \lambda_{k+1})(\|\phi_{k+1}(X_k) - \phi'_{k+1}(X_k)\| + \|\phi'_{k+1}(X_k) - \phi'_{k+1}(\widetilde{X}_k)\|) \\ &\leq (1 - \lambda_{k+1})(s_{k+1} + cz_k) , \end{aligned}$$

where the first line holds from Lemma 2.9. Letting $Z_{k+1}, Z'_{k+1} \sim \mathcal{N}(0, \sigma^2 I_d)$ be independent noises,

$$\begin{aligned} T(\widetilde{X}_{k+1}, X'_{k+1}) &\geq T(\lambda_{k+1}\phi_{k+1}(X_k) + (1 - \lambda_{k+1})\phi'_{k+1}(\widetilde{X}_k) + Z_{k+1}, \phi'_{k+1}(X'_k) + Z'_{k+1}) \\ &\geq T(\widetilde{X}_k, X'_k) \otimes G(\frac{a_{k+1}}{\sigma}) , \end{aligned}$$

where the first inequality is by the post-processing inequality (Lemma 2.5) with respect to $\Pi_{\mathcal{K}}$, and the second inequality is by Lemma 3.2. Repeating this for $k = t - 1, \ldots, \tau$, and using the fact that the shifted interpolated process satisfy $\widetilde{X}_t = X_t$ (from $\lambda_t = 1$) and $\widetilde{X}_\tau = X'_\tau$, we conclude the desired bound

$$T(X_t, X'_t) = T(\widetilde{X}_t, X'_t) \geq T(\widetilde{X}_\tau, X'_\tau) \otimes G\left(\frac{1}{\sigma}\sqrt{\sum_{k=\tau+1}^{t} a_k^2}\right) = G\left(\frac{1}{\sigma}\sqrt{\sum_{k=\tau+1}^{t} a_k^2}\right) .$$

□

[7] The infimum of tradeoff functions is in general not a tradeoff function; however, we prove a lower bound that is in fact a tradeoff function. That is, we show that $T(W + Z, Z) \geq G(\frac{s}{\sigma})$ for all W, Z satisfying the conditions in the definition of $R(s, \sigma)$. An analogous discussion also applies to Lemma C.12.

C.2. Deferred proofs for §4.1

C.2.1. Proof of Theorem 4.2

First, we consider the following setting where the contractive factor is strictly less than 1, which corresponds to the strongly convex setting for `NoisyGD`.

Theorem C.5. *In the setting of Theorem C.3, additionally assume that $0 < c < 1$ and $s_k \equiv s$. Then*

$$T(X_t, X'_t) \geq G\left(\sqrt{\frac{1-c^t}{1+c^t}\frac{1+c}{1-c}}\frac{s}{\sigma}\right)$$

with equality holding if $X_0 = X'_0 = 0$, $\phi_k(x) = cx$, $\phi'_k(x) = cx + sv$ for any unit vector v and $\mathcal{K} = \mathbb{R}^d$.

Proof. In Theorem C.3, we can take $\tau = 0$ and $z_\tau = 0$. Then the values of $\{\lambda_k\}, \{z_k\}, \{a_k\}$ obtained from the elementary optimization problem (Lemma C.6) yield the desired result. Finally, for the equality case, by direct calculation we have $X_t \sim \mathcal{N}(0, \frac{1-c^{2t}}{1-c^2}\sigma^2 I_d)$ and $X'_t \sim \mathcal{N}(\frac{1-c^t}{1-c}sv, \frac{1-c^{2t}}{1-c^2}\sigma^2 I_d)$, giving

$$T(X_t, X'_t) = G(\sqrt{\frac{1-c^t}{1+c^t}\frac{1+c}{1-c}}\frac{s}{\sigma}).$$

□

Lemma C.6. *Given $s > 0$ and $0 < c < 1$, the optimal value of*

$$\begin{aligned} &\textit{minimize} \sum_{k=1}^{t} a_k^2 \\ &\textit{subject to } z_{k+1} = (1-\lambda_{k+1})(cz_k + s), a_{k+1} = \lambda_{k+1}(cz_k + s) \\ &\qquad\qquad z_k, a_k \geq 0, z_0 = z_t = 0 \\ &\qquad\qquad \lambda_k \in [0, 1] \end{aligned}$$

is $\frac{1-c^t}{1+c^t}\frac{1+c}{1-c}s^2$.

Proof. Since $z_{k+1} = (1-\lambda_{k+1})(cz_k+s)$ and $a_{k+1} = \lambda_{k+1}(cz_k+s)$, we have $z_{k+1}+a_{k+1} = s+cz_k$ for $k = 0, \ldots, t-1$, from which we obtain

$$z_t = c^t z_0 + (1 + c + \cdots + c^{t-1})s - (a_t + ca_{t-1} + \cdots + c^{t-1}a_1).$$

From $z_0 = z_t = 0$, we have

$$a_t + ca_{t-1} + \cdots + c^{t-1}a_1 = \frac{1-c^t}{1-c}s.$$

By the Cauchy-Schwarz inequality,

$$\sum_{k=1}^{t} a_k^2 \geq \frac{(a_t + ca_{t-1} + \cdots + c^{t-1}a_1)^2}{\sum_{k=1}^{t} c^{2(t-k)}} = \frac{1-c^t}{1+c^t}\frac{1+c}{1-c}s^2$$

where equality holds if the corresponding equality criterion of the Cauchy-Schwarz inequality is satisfied. The explicit formulae are $z_k = \frac{(1-c^k)(1-c^{t-k})}{(1+c^t)(1-c)}s$, $a_k = \frac{c^{t-k}(1+c)}{1+c^t}s$, and $\lambda_k = \frac{c^{t-k}(1-c^2)}{1-c^{t-k+2}-c^k+c^t}$. □

Proof of Theorem 4.2. This follows as a direct corollary of Theorem C.5 with $\phi_k(x) \equiv \phi(x) = x - \frac{\eta}{n}\sum_{i=1}^{n}\nabla f_i(x)$ and $\phi'_k(x') \equiv \phi'(x') = x' - \frac{\eta}{n}\sum_{i=1}^{n}\nabla f'_i(x')$. For this application, consider parameters $c = \max\{|1-\eta m|, |1-\eta M|\} < 1$ (by Lemma 2.8), $s \leftarrow \eta L/n$, and $\sigma \leftarrow \eta\sigma$ (rescaling to simplify notation). The equality case is a straightforward calculation in the setting that $\mathcal{K} = \mathbb{R}^d$, $X_0 = 0$, $\nabla f_i(x) = mx$ for all $i \in [n]$, and $\nabla f'_i(x)$ defined as mx for $i \neq i^*$, and otherwise $mx - Lv$ for some unit vector v. □

C.2.2. Proof of Theorem 4.3

We consider here the setting of optimization over a bounded constraint set.

Theorem C.7. *In the setting of Theorem C.3, additionally assume that $\mathcal{K}$ has a finite diameter D and $s_k \equiv s$. Then for any integer $0 \le \tau < t$,*

$$T(X_t, X_t') \ge G\left(\frac{1}{\sigma}(s\sqrt{t-\tau} + \frac{D}{\sqrt{t-\tau}})\right).$$

In particular, if $t \ge D/s$ then

$$T(X_t, X_t') \ge G\left(\frac{1}{\sigma}\sqrt{3sD + s^2\left\lceil\frac{D}{s}\right\rceil}\right).$$

Proof. From $\|X_\tau - \widetilde{X}_\tau\| \le D$ for all τ, in Theorem C.3 we can take $z_\tau = D$ and $c = 1$. The values of $\{\lambda_k\}, \{z_k\}, \{a_k\}$ are obtained by analyzing the following elementary optimization problem (Lemma C.8). □

Lemma C.8. *Given $s > 0$ and $D > 0$, the optimal value of*

$$\begin{aligned} \textit{minimize} \quad & \sum_{k=\tau+1}^{t} a_k^2 \\ \textit{subject to} \quad & z_{k+1} = (1-\lambda_{k+1})(z_k + s), a_{k+1} = \lambda_{k+1}(z_k + s) \\ & z_k, a_k \ge 0, z_\tau = D, z_t = 0 \\ & \lambda_k \in [0,1] \end{aligned}$$

is $\left(s + \frac{D}{t-\tau}\right)^2 (t-\tau)$. As a function of $t-\tau \in (0,\infty)$, this value is minimized when $t - \tau = D/s$.

Proof. By adding the equations $z_{k+1} = (1-\lambda_{k+1})(z_k + s)$ and $a_{k+1} = \lambda_{k+1}(z_k + s)$ for $k = \tau, \ldots, t-1$, we obtain (with $z_\tau = D$ and $z_t = 0$)

$$a_t + a_{t-1} + \cdots + a_{\tau+1} = D + (t-\tau)s.$$

By the Cauchy-Schwarz inequality, the minimum value of $\sum_{k=\tau+1}^{t} a_k^2$ is $(s+R)^2(t-\tau)$ and is obtained when $z_k = R(t-k)$, $a_k \equiv s + R$, and $\lambda_k = \frac{s+R}{s+R(t-k+1)}$, where we use the shorthand $R := \frac{D}{t-\tau}$. The last part is straightforward from the strict convexity of the one-dimensional function $z \mapsto z\left(s + \frac{D}{z}\right)^2 = s^2 z + \frac{D^2}{z} + 2sD, z > 0$. □

Proof of Theorem 4.3. This follows by considering $s \leftarrow \frac{\eta L}{n}$ and $\sigma \leftarrow \eta\sigma$ as in the proof of Theorem 4.2. □

C.3. Deferred proofs for §4.2

As done in the case of `NoisyGD`, we first characterize `NoisyCGD` as a particular instance of CNI and proceed to the proofs of the theorems. The following proposition holds straight from the definition; recall that $l = n/b$ is the number of batches.

Proposition C.9. *For $t = lE$ and $k = 0, 1, \ldots, t-1$, let $B_1, \ldots, B_l$ be a fixed partition of $[n]$ with size b, and define $\phi_{k+1}(x) = x - \frac{\eta}{b}\sum_{i \in B_r} \nabla f_i(x)$ and $\phi_{k+1}'(x') = x' - \frac{\eta}{b}\sum_{i\in B_r}\nabla f_i'(x')$ where $r = k + 1 - l\lfloor\frac{k}{l}\rfloor$. Then the f-DP of `NoisyCGD` is equal to that between $X_t = \mathrm{CNI}(X_0, \{\phi_k\}_{k\in[t]}, \{\mathcal{N}(0, \eta^2\sigma^2 I_d)\}_{k\in[t]}, \mathcal{K})$ and $X_t' = \mathrm{CNI}(X_0, \{\phi_k'\}_{k\in[t]}, \{\mathcal{N}(0,\eta^2\sigma^2 I_d)\}_{k\in[t]}, \mathcal{K})$.*

Proof of Theorem 4.5. Let $j^* \in [l]$ be the index such that $i^* \in B_{j^*}$ and consider the setting in Proposition C.9. We will establish a lower bound on $T(X_{t^*}, X_{t^*}')$ where $t^* = t + j^* - l - 1$; the lower bound on $T(X_t, X_t')$ is then given by

$$T(X_t, X_t') \ge T(X_{t+j^*-l}, X_{t+j^*-l}') \ge T(X_{t+j^*-l-1}, X_{t+j^*-l-1}') \otimes G(\frac{L}{b\sigma})$$

where the first inequality holds from the post-processing inequality with $\phi_k \equiv \phi_k'$ for all $k = t + j^* - l + 1, \ldots, t$, and the second inequality holds from $\|\phi_{t+j^*-l}(x) - \phi_{t+j^*-l}'(x)\| \le \frac{\eta L}{b}$ for all x with Lemma 3.1.

In general, $\phi_{k+1} = \phi'_{k+1}$ when $r = r(k) = k + 1 - l\lfloor \frac{k}{l} \rfloor$ is not equal to j^*; otherwise $\|\phi_{k+1}(x) - \phi'_{k+1}(x)\| \leq \frac{\eta L}{b}$ for all x. Thus, in Theorem C.3 we can take $\tau = 0, z_\tau = 0$ and $s_{k+1} = s\mathbf{1}_{\{r=j^*\}}$ where $s = \frac{\eta L}{b}$. Using $\{\lambda_k\}, \{z_k\}, \{a_k\}$ obtained from the following result (Lemma C.10),

$$T(X_t, X'_t) \geq G\left(\frac{1}{\sigma}\sqrt{\left(\frac{L}{b}\right)^2 + \frac{1}{\eta^2}\sum_{k=1}^{t^*} a_k^2}\right).$$

□

Lemma C.10. *Given $s > 0$ and $0 < c < 1$, let $t^* = t + j^* - l - 1$ and consider a system*

$$\begin{aligned} z_{k+1} &= (1-\lambda_{k+1})(cz_k + s\mathbf{1}_{\{r=j^*\}}), a_{k+1} = \lambda_{k+1}(cz_k + s\mathbf{1}_{\{r=j^*\}}) \\ z_k, a_k &\geq 0, z_0 = z_{t^*} = 0 \\ \lambda_k &\in [0,1]. \end{aligned}$$

Then $\{a_k\}_{1\leq k\leq t^*}, \{z_k\}_{0\leq k\leq t^*}, \{\lambda_k\}_{1\leq k\leq t^*}$ defined as

$$\begin{aligned} a_k &= \begin{cases} \frac{c^{t-k+j^*-2}}{1-c^l}\frac{1-c^2}{1+c^{t-l}}s & k \geq j^* \\ 0 & k < j^* \end{cases} \\ z_{k+1} &= cz_k + s\mathbf{1}_{\{r=j^*\}} - a_{k+1}, z_0 = 0 \\ \lambda_k &= \begin{cases} \frac{a_k}{z_k + a_k} & z_k + a_k > 0 \\ 0 & z_k + a_k = 0 \end{cases} \end{aligned}$$

is a solution, where $r = r(k) = k + 1 - l\lfloor \frac{k}{l} \rfloor$.

Proof. From the stated formulae and $z_{k+1} + a_{k+1} = cz_k + s\mathbf{1}_{\{r=j^*\}}$, every condition except $z_k \geq 0$ and $z_{t^*} = 0$ are straightforward to check.

If $k < j^*$ then $z_k \equiv 0$. For $k \geq j^*$, let q be the integer such that $l(q-1) + j^* \leq k < lq + j^*$ and $r' = k - (l(q-1) + j^*)$. Then

$$\begin{aligned} z_k &= c^{r'}(1 + c^l + \cdots + c^{l(q-1)})s - (a_k + ca_{k-1} + \cdots + c^{k-j^*}a_{j^*}) \\ &= \left(c^{r'}(1 + c^l + \cdots + c^{l(q-1)}) - c^{t-k+j^*-2}\frac{1-c^2}{(1-c^l)(1+c^{t-l})}(1 + c^2 + \cdots + c^{2(k-j^*)})\right)s. \end{aligned}$$

For any fixed q, this is a decreasing function in r' and thus it suffices to consider $r' = l - 1$. Then

$$\begin{aligned} &c^{r'}(1 + c^l + \cdots + c^{l(q-1)}) - c^{t-k+j^*-2}\frac{1-c^2}{(1-c^l)(1+c^{t-l})}(1 + c^2 + \cdots + c^{2(k-j^*)}) \\ &= c^{l-1}\frac{1-c^{lq}}{1-c^l} - c^{l(E-q)-1}\frac{1-c^{2lq}}{(1-c^l)(1+c^{t-l})} = c^{l-1}\frac{1-c^{lq}}{1-c^l}\frac{1-c^{l(E-q-1)}}{1+c^{E(l-1)}}, \end{aligned}$$

which is nonnegative for $q \leq E - 1$. Also, for $q = E - 1$ this is equal to 0, implying $z_{l(E-1)+j^*-1} = z_{t^*} = 0$ and thus $\lambda_{t^*} = 1$. □

Proof of Theorem 4.6. As in the proof of Theorem 4.5, we establish a lower bound on $T(X_{t^*}, X'_{t^*})$ for $t^* = t + j^* - l - 1$. For any τ, letting $\tau^* = j^* + l(\tau - 1)$ we have $\|X_{\tau^*} - \widetilde{X}_{\tau^*}\| \leq D$. Thus in Theorem C.3 we can take $z_{\tau^*} = D, s_{k+1} = s\mathbf{1}_{\{r=j^*\}}(s = \frac{\eta L}{b})$ and $c = 1$. The sequences $\{\lambda_k\}, \{z_k\}, \{a_k\}$ can be chosen as in the following result (Lemma C.11), which yields a bound of

$$T(X_t, X'_t) \geq G\left(\frac{1}{\sigma}\sqrt{\left(\frac{L}{b}\right)^2 + \frac{1}{\eta^2}\sum_{k=\tau^*+1}^{t^*} a_k^2}\right) \geq G\left(\frac{1}{\sigma}\sqrt{\left(\frac{L}{b}\right)^2 + \frac{(D/\eta + L(E-\tau)/b)^2}{l(E-\tau)}}\right)$$

by Proposition C.9. Optimizing over the choice of $E - \tau$ can be done similarly as in Theorem C.7; in particular, one can take $E - \tau = \lceil \frac{Db}{\eta L} \rceil$ when $E \geq \frac{Db}{\eta L}$. □

Lemma C.11. *Given $s > 0$ and $D > 0$, let $t^* = t + j^* - l - 1, \tau^* = j^* + l(\tau - 1)$ and consider a system*

$$\begin{aligned} &z_{k+1} = (1 - \lambda_{k+1})(z_k + s\mathbf{1}_{\{r=j^*\}}), a_{k+1} = \lambda_{k+1}(z_k + s\mathbf{1}_{\{r=j^*\}}) \\ &z_k, a_k \geq 0, z_{\tau^*} = D, z_{t^*} = 0 \\ &\lambda_k \in [0, 1] . \end{aligned}$$

Then $\{a_k\}_{\tau^*+1 \leq k \leq t^*}, \{z_k\}_{\tau^* \leq k \leq t^*}, \{\lambda_k\}_{\tau^*+1 \leq k \leq t^*}$ defined as

$$\begin{aligned} a_k &\equiv \frac{D + s(E - \tau)}{l(E - \tau)} \\ z_{k+1} &= z_k + s\mathbf{1}_{\{r=j^*\}} - a_{k+1}, z_{\tau^*} = D \\ \lambda_k &= \begin{cases} \frac{a_k}{z_k + a_k} & z_k + a_k > 0 \\ 0 & z_k + a_k = 0 \end{cases} \end{aligned}$$

is a solution, where $r = r(k) = k + 1 - l\lfloor \frac{k}{l} \rfloor$.

Proof. As in the proof of Lemma C.10, it suffices to check that $z_k \geq 0$ and $z_{t^*} = 0$. Let $q \geq \tau$ be the integer such that $l(q - 1) + j^* \leq k < lq + j^*$ and $r' = k - (l(q - 1) + j^*)$. Then

$$\begin{aligned} z_k &= D + (q - \tau + 1)s - (l(q - \tau) + r' + 1)\frac{D + s(E - \tau)}{l(E - \tau)} \\ &\geq D + (q - \tau + 1)s - (q - \tau + 1)\frac{D + s(E - \tau)}{(E - \tau)} \\ &= D(1 - \frac{q - \tau + 1}{E - \tau}) \geq 0 , \end{aligned}$$

where the inequality is from that the first line is minimized when $r' = l - 1$ for any fixed q. Also, $z_k = 0$ and $\lambda_k = 1$ when $r' = l - 1$ and $q = E - 1$, i.e., $k = t + j^* - l - 1 = t^*$. □

C.4. Deferred proofs for §4.3

C.4.1. NoisySGD AS STOCHASTIC VERSION OF CNI

We first revisit the composition bound (Theorem 4.8). The key point in this proof relevant to our new results is the following formulation, which can be considered as a stochastic version of CNI (9) with each map $x \mapsto \psi_s(x), \phi_s(x), \phi'_s(x)$ being contractive.

$$\begin{aligned} X_{k+1} &= \Pi_{\mathcal{K}}(\psi_{S_k}(X_k) + V_k(\phi_{S_k} - \psi_{S_k})(X_k) + Z_{k+1}) \\ X'_{k+1} &= \Pi_{\mathcal{K}}(\psi_{S'_k}(X'_k) + V'_k(\phi'_{S'_k} - \psi_{S'_k})(X'_k) + Z'_{k+1}) \end{aligned} \tag{11}$$

Proof of Theorem 4.8. Let X_k and X_{k+1} respectively be the k-th and $(k+1)$-th iterate of NoisySGD with losses $\{f_i\}_{i\in[n]}$, and similarly define X'_k and X'_{k+1} for NoisySGD with losses $\{f'_i\}_{i\in[n]}$. It suffices to show

$$T(X_{k+1}, X'_{k+1}) \geq T(X_k, X'_k) \otimes C_{b/n}(G(\frac{L}{b\sigma})) .$$

For the corresponding random batch B_k, we sample a random pair of set and element $S_k = (R_k, C_k)$ as described below. This S_k will be here and after used as a representation for the random batch B_k.

1. Sample a set A_1 of size b in $[n] \setminus \{i^*\}$ uniformly at random.
2. Sample an element A_2 from A_1 uniformly at random. This element will serve as a candidate to be (potentially) replaced by i^*.
3. Let $R_k = A_1 \setminus \{A_2\}, C_k = A_2$.

Finally, let $V_k \sim \text{Ber}(p)$ be a Bernoulli random variable with success probability $p = b/n$, which serves as an indicator denoting whether $i^* \in B_k$ (i.e., $V_k = 1$) or not (i.e., $V_k = 0$). Then

$$B_k = \begin{cases} R_k \cup \{C_k\} & V_k = 0 \\ R_k \cup \{i^*\} & V_k = 1 \end{cases}$$

is a valid sampling procedure for B_k (i.e., the marginal distribution of B_k is uniform over size b subsets of $[n]$). These can be defined similarly for X'_{k+1} as B'_k, V'_k and S'_k.

The reason for formulating this alternative sampling scheme is to separate the subsampling part—which only depends on whether the index i^* is included in the batch—from the rest of the information on the batch. In particular, in (11), S_k and V_k are independent and V_k is still distributed as $\text{Ber}(p)$ after conditioning on S_k.

Now for a pair of set and element $S = (R, C)$, define

$$\begin{aligned} \phi_S(x) &= x - \frac{\eta}{b}(\nabla f_{i^*} + \sum_{i \in R} \nabla f_i)(x) \\ \phi'_S(x) &= x - \frac{\eta}{b}(\nabla f'_{i^*} + \sum_{i \in R} \nabla f_i)(x) \\ \psi_S(x) &= x - \frac{\eta}{b} \sum_{i \in R \cup \{C\}} \nabla f_i(x) \,. \end{aligned} \tag{12}$$

Then the updates for X_{k+1} and X'_{k+1} can be respectively written as (11), where $Z_{k+1}, Z'_{k+1} \sim \mathcal{N}(0, \eta^2 \sigma^2 I_d)$ are independent of anything else. Now the tradeoff function between X_{k+1} and X'_{k+1} satisfies

$$T(X_{k+1}, X'_{k+1}) \geq T((X_k, S_k, V_k(\phi_{S_k} - \psi_{S_k})(X_k) + Z_{k+1}), (X'_k, S'_k, V'_k(\phi'_{S'_k} - \psi_{S'_k})(X'_k) + Z'_{k+1}))$$

by the post-processing inequality with respect to $(x, s, y) \mapsto \Pi_{\mathcal{K}}(\psi_s(x) + y)$. For any fixed realization $(x, s) = (x, (r, c))$ of the first two arguments, we find a lower bound on

$$T(V_k(\phi_s - \psi_s)(x) + Z_{k+1}, V'_k(\phi'_s - \psi_s)(x) + Z'_{k+1}) \,. \tag{13}$$

In fact, this is tradeoff function of the subsampled Gaussian mechanism as presented in (Dong et al., 2022, Theorem 9). To see this, we construct a new private setting as follows:

- Datasets: $S = \{y_1, y_2, \ldots, y_n\}, S' = \{y'_1, y_2, \ldots, y_n\}$ where $y'_1, y_1, \ldots, y_n$ are distinct alphabets. Note that the "datasets" here are considered only for this part of the proof and are irrelevant with the original datasets in the private optimization setting.
- Mechanisms:
  (a) $\texttt{Sample}_b$: From a set of size n, sample a set of size b uniformly at random.
  (b) $\texttt{M}$: Given a set R of size b, output $\theta(R) + \mathcal{N}(0, \eta^2 \sigma^2 I_d)$ where

$$\theta(R) = \begin{cases} (\phi_s - \psi_s)(x) & y_1 \in R, y'_1 \notin R \\ (\phi'_s - \psi_s)(x) & y_1 \notin R, y'_1 \in R \\ 0 & \text{else} \,. \end{cases}$$

Then a lower bound f on (13) is equivalent to $\texttt{M} \circ \texttt{Sample}_b$ being f-DP (when considered as being applied to S and S'). Note that from

$$\begin{aligned} \phi_s - \psi_s &= \frac{\eta}{b}(\nabla f_c - \nabla f_{i^*}) \\ \phi'_s - \psi_s &= \frac{\eta}{b}(\nabla f_c - \nabla f'_{i^*}) \\ \phi_s - \phi'_s &= \frac{\eta}{b}(\nabla f'_{i^*} - \nabla f_{i^*}) \,, \end{aligned}$$

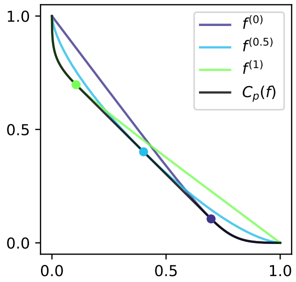

Figure 4. Illustration of $C_p(G(\frac{L}{b\sigma}))$ and $f^{(\lambda)}$, for $\lambda \in \{0, 0.5, 1\}$ with $p = 0.25, L/(b\sigma) = 2.5$.

θ has (l_2-)sensitivity $\frac{\eta L}{b}$ and thus $\mathtt{M}$ is $G(\frac{L}{b\sigma})$-DP by (Dong et al., 2022, Theorem 1). Then by (Dong et al., 2022, Theorem 9), $\mathtt{M} \circ \mathtt{Sample}_b$ is $C_{b/n}(G(\frac{L}{b\sigma}))$-DP. Thus,

$$\begin{aligned}
&T((X_k, S_k, V_k(\phi_{S_k} - \psi_{S_k})(X_k) + Z_{k+1}), (X'_k, S'_k, V'_k(\phi'_{S'_k} - \psi_{S'_k})(X'_k) + Z'_{k+1}))\\
&\geq T((X_k, S_k), (X'_k, S'_k)) \otimes C_{b/n}(G(\frac{L}{b\sigma}))\\
&= T(X_k, X'_k) \otimes C_{b/n}(G(\frac{L}{b\sigma}))
\end{aligned}$$

where the equality is from that $S_k(S'_k)$ is independent of $X_k(X'_k)$, and that S_k and S'_k have the same distribution. □

One step optimality. Now we show the optimality of Theorem 4.8 for $t = 1$, i.e.,

$$T(X_1, X'_1) \geq C_{b/n}(G(\frac{L}{b\sigma}))\,.$$

Let $X_0 = X'_0 = 0, \mathcal{K} = \mathbb{R}^d$ and for $\lambda \in [0, 1]$, consider the gradients

$$\begin{aligned}
\nabla f_i = \nabla f'_i &= 0\\
\nabla f_{i^*} &= (1-\lambda)Lu\\
\nabla f'_{i^*} &= -\lambda Lu
\end{aligned}$$

where u is a unit vector. Then

$$\begin{aligned}
T(X_1, X'_1) &= T(-(1-\lambda)\frac{\eta L}{b}uV_0 + \mathcal{N}(0, \eta^2\sigma^2 I_d), \lambda\frac{\eta L}{b}uV'_0 + \mathcal{N}(0, \eta^2\sigma^2 I_d))\\
&= T(-(1-\lambda)\frac{L}{b\sigma}V_0 + \mathcal{N}(0,1), \lambda\frac{L}{b\sigma}V'_0 + \mathcal{N}(0,1))
\end{aligned}$$

where $V_0, V'_0 \sim \mathrm{Ber}(p)$. Denoting the corresponding tradeoff function as $f^{(\lambda)}$, a valid lower bound for $T(X_1, X'_1)$ is (pointwise) at most $\inf_{\lambda\in[0,1]} f^{(\lambda)}$ and thus it suffices to show that $\inf_{\lambda\in[0,1]} f^{(\lambda)} = C_{b/n}(G(\frac{L}{b\sigma}))$. Now the rest of the proof is a combination of following facts.

(a) $f^{(1)}(\alpha) \geq C_{b/n}(G(\frac{L}{b\sigma}))(\alpha)$ with equality holding for all $\alpha \in [0, \Phi(-\frac{L}{2b\sigma})]$.

(b) $f^{(0)}(\alpha) \geq C_{b/n}(G(\frac{L}{b\sigma}))(\alpha)$ with equality holding for all $\alpha \in [p\Phi(-\frac{L}{2b\sigma}) + (1-p)\Phi(\frac{L}{2b\sigma}), 1]$.

(c) For $\lambda \in (0, 1)$, $f^{(\lambda)}(\alpha) \geq C_{b/n}(G(\frac{L}{b\sigma}))(\alpha)$ with equality holding at $\alpha = p\Phi(-\frac{L}{2b\sigma}) + (1-p)\Phi((\frac{1}{2} - \lambda)\frac{L}{b\sigma})$ (note that as λ varies, this covers the range of α at which $C_{b/n}(G(\frac{L}{b\sigma}))$ is linear with slope -1 and interpolates the boundaries in (a) and (b)).

The first two facts are straightforward from Definition 4.7, with $G(\mu)_p = T(\mathcal{N}(0,1), p\mathcal{N}(\mu,1) + (1-p)\mathcal{N}(0,1))$ and $(f^{(0)})^{-1} = f^{(1)}$.[8] For (c), note that as a mixture of one-dimensional Gaussians the likelihood ratio between the two distributions is monotone and thus for any $z \in \mathbb{R}$, with $\alpha = 1 - (1-p)\Phi(z) - p\Phi(z + (1-\lambda)\frac{L}{b\sigma})$ we have

$$f^{(\lambda)}(\alpha) = (1-p)\Phi(z) + p\Phi(z - \lambda\frac{L}{b\sigma})\,.$$

Thus from (here φ denotes the probability density function of $\mathcal{N}(0,1)$)

$$\begin{aligned}\frac{d\alpha}{dz} &= -(1-p)\varphi(z) - p\varphi(z + (1-\lambda)\frac{L}{b\sigma})\\ \frac{df^{(\lambda)}(\alpha)}{dz} &= (1-p)\varphi(z) + p\varphi(z - \lambda\frac{L}{b\sigma})\,,\end{aligned}$$

at $z = (\lambda - \frac{1}{2})\frac{L}{b\sigma}$ we have $\alpha = p\Phi(-\frac{L}{2b\sigma}) + (1-p)\Phi((\frac{1}{2}-\lambda)\frac{L}{b\sigma})$ where $\alpha + f^{(\lambda)}(\alpha) = (1+p)\Phi(-\frac{L}{2b\sigma}) + (1-p)\Phi(\frac{L}{2b\sigma})$ and $\frac{df^{(\lambda)}}{d\alpha}(\alpha) = \frac{df_\lambda(\alpha)}{dz}/\frac{d\alpha}{dz} = -1$. This implies that $f^{(\lambda)}$ is tangent to $C_{b/n}(G(\frac{L}{b\sigma}))$ at the point, and (c) follows by Lemma 2.3.

C.4.2. Proofs of new results

As in the case of NoisyGD (Lemma C.4), we start by establishing a lower bound for tradeoff function between convolutions of Gaussian random variables with bounded random variables—now including the subsampling.

Lemma C.12. *For $s \geq 0$ and $p = b/n$, let*

$$R(s,\sigma,p) = \inf\{T(VW + Z, VW' + Z) : V \sim \textit{Ber}(p), \|W\|, \|W'\| \leq s, Z \sim \mathcal{N}(0, \sigma^2 I_d)\}$$

where the infimum is taken pointwise and is over independent V, W, W', Z. Then $R(s,\sigma,p) \geq C_p(G(\frac{2s}{\sigma}))$.[9]

Proof. The proof is fairly similar to the subsampling part in the proof of Theorem 4.8. Let V, W, W', Z be as in the definition of $R(s,\sigma,p)$, and consider the following private setting:

- Datasets: $S = \{y_1, y_2, \ldots, y_n\}$, $S' = \{y_1', y_2, \ldots, y_n\}$ where $y_1', y_1, \ldots, y_n$ are distinct alphabets.
- Mechanisms:
  (a) $\texttt{Sample}_b$: From a set of size n, sample a set of size b uniformly at random.
  (b) $\texttt{M}$: Given a set R of size b, output $\theta(R) + Z$ where

$$\theta(R) = \begin{cases} W & y_1 \in R, y_1' \notin R \\ W' & y_1 \notin R, y_1' \in R \\ 0 & \text{else}\,. \end{cases}$$

From $\|W\|, \|W'\|, \|W - W'\| \leq 2s$, θ has sensitivity $2s$ and thus $\texttt{M}$ is a $G(\frac{2s}{\sigma})$-DP mechanism by (Dong et al., 2022, Theorem 1). By (Dong et al., 2022, Theorem 9), $\texttt{M} \circ \texttt{Sample}_b$ is $C_p(G(\frac{2s}{\sigma}))$-DP, which is equivalent to $T(VW + Z, VW' + Z) \geq C_p(G(\frac{2s}{\sigma}))$. □

Now we proceed to the proofs of the new results. The key point here is that we build shifted interpolated processes by not only coupling the noise Z_{k+1} but also the subsampling indicator V_k; see Figure 5. For the strongly convex and smooth setting, we state and prove a general theorem that allows one to choose the sequences of shift and sensitivity.

[8] In fact, since tradeoff functions are convex, (a) and (b) are enough to conclude that $C_p(G(\mu))$ is the best tradeoff function bound; (c) provides an additional explanation on the linear part of $C_p(G(\mu))$. See Figure 4.

[9] We conjecture that a strictly better lower bound holds, which corresponds to the case when W and W' are constant vectors aligned in the opposite direction, i.e., $R(s,\sigma,p) = T(p\mathcal{N}(-\frac{s}{\sigma},1) + (1-p)\mathcal{N}(0,1), p\mathcal{N}(\frac{s}{\sigma},1) + (1-p)\mathcal{N}(0,1))$.

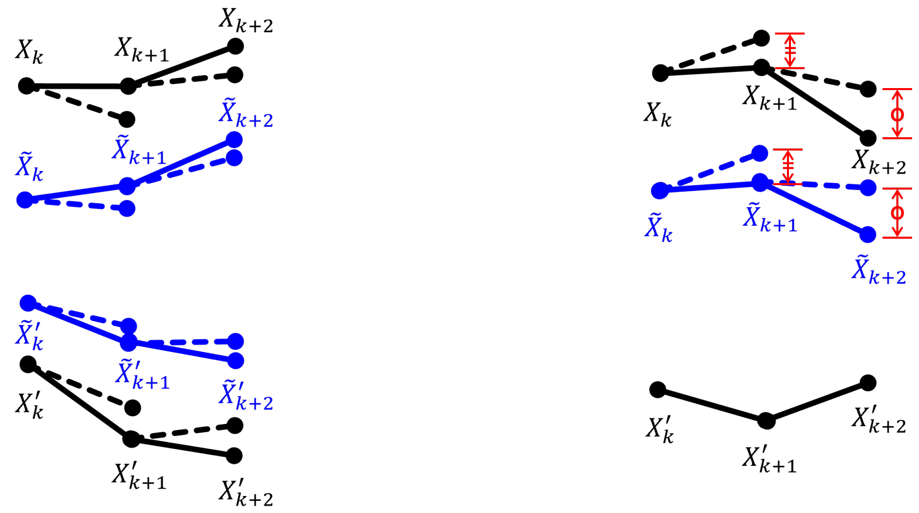

Figure 5. Illustration of shifted interpolated processes in the proofs of Theorem 4.9 (left) and Theorem 4.10 (right). The solid lines denote the updates based on the realized values of $\{V_k\}$, and the dashed lines denote the alternative updates based on their unrealized values; each interpolated process uses the same (coupled) values of $\{V_k\}$ as expressed in the figure. In Theorem 4.9, we build two processes, each of which tracks its corresponding original process. In Theorem 4.10, only one process is built and it inherits the identical deviation based on the realizations of $\{V_k\}$.

Theorem C.13. *Consider m-strongly convex, M-smooth loss functions with gradient sensitivity L. Then for any $\eta \in (0, 2/M)$, $\texttt{NoisySGD}$ is f-DP where*

$$f = G\Big(\frac{2\sqrt{2}cz_{t-1}}{\eta\sigma}\Big) \otimes \bigotimes_{k=0}^{t-1} C_{b/n}\Big(G\Big(\frac{2a_k}{\eta\sigma}\Big)\Big)$$

for any sequence $\{z_k\}_{0\le k\le t-1}, \{a_k\}_{0\le k\le t-1}$ such that $z_0 = 0, a_0 = \frac{\sqrt{2}\eta L}{b}, a_k \le \frac{\eta L}{b}$ for all $k \ge 1$ and $z_{t-1} = \frac{1-c^{t-1}}{1-c}\frac{\eta L}{b} - \sum_{k=1}^{t-1} c^{t-k-1}a_k$ and where $c = \max\{|1-\eta m|, |1-\eta M|\}$.

Proof. As in (11) and (12), the iterates of $\texttt{NoisySGD}$ with respect to $\{f_i\}_{i\in[n]}$ and $\{f'_i\}_{i\in[n]}$ are

$$\begin{aligned} X_{k+1} &= \Pi_{\mathcal{K}}(\psi_{S_k}(X_k) + V_k(\phi_{S_k} - \psi_{S_k})(X_k) + Z_{k+1}) \\ X'_{k+1} &= \Pi_{\mathcal{K}}(\psi_{S'_k}(X'_k) + V'_k(\phi'_{S'_k} - \psi_{S'_k})(X'_k) + Z'_{k+1}), \end{aligned}$$

where $Z_{k+1}, Z'_{k+1} \sim \mathcal{N}(0, \eta^2\sigma^2 I_d)$. Now consider shifted interpolated processes defined as

$$\begin{aligned} \widetilde{X}_{k+1} &= \Pi_{\mathcal{K}}(\psi_{S_k}(\widetilde{X}_k) + \lambda_{k+1}V_k(\phi_{S_k}(X_k) - \psi_{S_k}(\widetilde{X}_k)) + Z_{k+1}) \\ \widetilde{X}'_{k+1} &= \Pi_{\mathcal{K}}(\psi_{S'_k}(\widetilde{X}'_k) + \lambda_{k+1}V'_k(\phi'_{S'_k}(X'_k) - \psi_{S'_k}(\widetilde{X}'_k)) + Z'_{k+1}), \end{aligned}$$

with $\widetilde{X}_0 = \widetilde{X}'_0 = X_0$ and $\lambda_k = \frac{a_k}{z_k + a_k} \cdot \mathbf{1}_{\{z_k + a_k > 0\}}$ for $\{z_k\}_{0\le k\le t-1}$ and $\{a_k\}_{0\le k\le t-1}$ such that $z_0 = 0, a_0 = \frac{\sqrt{2}\eta L}{b}$ and $z_{k+1} = cz_k + \frac{\eta L}{b} - a_{k+1}$ for all $k \ge 0$. Then inductively $\|\widetilde{X}_k - X_k\| \le z_k$ for all k from

$$\|\widetilde{X}_{k+1} - X_{k+1}\| \le \begin{cases} \|\psi_{S_k}(X_k) - \psi_{S_k}(\widetilde{X}_k)\| \le cz_k & V_k = 0 \\ \|(1-\lambda_{k+1})(\phi_{S_k}(X_k) - \psi_{S_k}(\widetilde{X}_k))\| \le cz_k + \frac{\eta L}{b} - a_{k+1} & V_k = 1 \end{cases}$$

and $\|\lambda_{k+1}(\phi_{S_k}(X_k) - \psi_{S_k}(\widetilde{X}_k))\| \le a_{k+1}$; similar results hold for $\{X'_k\}$. Thus as in the proof of Theorem 4.8 (see also Theorem C.3), with Lemma C.12

$$T(\widetilde{X}_{t-1}, \widetilde{X}'_{t-1}) \ge \bigotimes_{k=1}^{t-1} C_{b/n}\Big(G\Big(\frac{2a_k}{\eta\sigma}\Big)\Big).$$

To relate this with $T(X_t, X'_t)$, note that there is no choice of λ_t that yields $\widetilde{X}_t = X_t$. Instead, we can proceed as follows: write down the corresponding update (before taking the projection) as

$$
\begin{aligned}
&\psi_{S_{t-1}}(X_{t-1}) + V_{t-1}(\phi_{S_{t-1}} - \psi_{S_{t-1}})(X_{t-1}) + Z_t \\
&= \psi_{S_{t-1}}(\widetilde{X}_{t-1}) + \psi_{S_{t-1}}(X_{t-1}) - \psi_{S_{t-1}}(\widetilde{X}_{t-1}) + Z_t^{(1)} + V_{t-1}(\phi_{S_{t-1}} - \psi_{S_{t-1}})(X_{t-1}) + Z_t^{(2)}
\end{aligned}
$$

where $Z_t^{(1)}, Z_t^{(2)} \sim \mathcal{N}(0, \frac{\eta^2\sigma^2}{2} I_d)$[10] are independent, $\psi_{S_{t-1}}(X_{t-1}) - \psi_{S_{t-1}}(\widetilde{X}_{t-1})$ is bounded by cz_{t-1} and $(\phi_{S_{t-1}} - \psi_{S_{t-1}})(X_{t-1})$ is bounded by $\frac{\eta L}{b}$. Then

$$
\begin{aligned}
&T(X_t, X'_t) \\
&\geq T((\widetilde{X}_{t-1}, S_{t-1}, \psi_{S_{t-1}}(X_{t-1}) - \psi_{S_{t-1}}(\widetilde{X}_{t-1}) + Z_t^{(1)}), (\widetilde{X}'_{t-1}, S'_{t-1}, \psi_{S'_{t-1}}(X'_{t-1}) - \psi_{S'_{t-1}}(\widetilde{X}'_{t-1}) + Z_t^{(1)'})) \\
&\quad \otimes R(\frac{\eta L}{b}, \frac{\eta\sigma}{\sqrt{2}}, b/n) \\
&\geq T((\widetilde{X}_{t-1}, S_{t-1}), (\widetilde{X}'_{t-1}, S'_{t-1})) \otimes R(cz_{t-1}, \frac{\eta\sigma}{\sqrt{2}}, 1) \otimes R(\frac{\eta L}{b}, \frac{\eta\sigma}{\sqrt{2}}, b/n) \\
&\geq T(\widetilde{X}_{t-1}, \widetilde{X}'_{t-1}) \otimes G(\frac{2\sqrt{2}cz_{t-1}}{\eta\sigma}) \otimes C_{b/n}(G(\frac{2\sqrt{2}L}{b\sigma}))\,.
\end{aligned}
$$

□

In this formulation, optimizing over the sequences $\{z_k\}$ and $\{a_k\}$ is intractable because of the analytically complicated nature of the subsampled operator and composition of tradeoff functions. Heuristically, when b/n is small, each individual $C_{b/n}(G(\cdot))$ is very close to Id and the most substantial factor is the GDP part. In this sense, sequences that make z_{t-1} small can be considered as a reasonable choice.

Proof of Theorem 4.9. Consider $a_{t-1} = \cdots = a_\tau = \frac{\eta L}{b}$ and $a_k = 0$ for all $1 \leq k < \tau$ in Theorem C.13. □

Proof of Theorem 4.10. For the iterates (11) and (12), consider the shifted interpolated process

$$\widetilde{X}_{k+1} = \Pi_{\mathcal{K}}(\psi_{S_k}(\widetilde{X}_k) + \lambda_{k+1}(\psi_{S_k}(X_k) - \psi_{S_k}(\widetilde{X}_k)) + V_k(\phi_{S_k} - \psi_{S_k})(X_k) + Z_{k+1})$$

where $\lambda_{k+1} = \frac{1}{t-k}$ and $\widetilde{X}_\tau = X'_\tau$. Then for any $k \geq \tau$, $\|\widetilde{X}_k - X_k\| \leq z_k$ and $\|\lambda_{k+1}(\psi_{S_k}(X_k) - \psi_{S_k}(\widetilde{X}_k))\| \leq a_{k+1}$ where

$$
\begin{aligned}
z_k &= \frac{D}{t-\tau}(t-k) \\
a_{k+1} &\equiv \frac{D}{t-\tau}\,.
\end{aligned}
$$

The first inequality is inductively from $\|\widetilde{X}_\tau - X_\tau\| = \|X'_\tau - X_\tau\| \leq D$ and

$$\|\widetilde{X}_{k+1} - X_{k+1}\| \leq (1 - \lambda_{k+1})\|\widetilde{X}_k - X_k\| \leq z_{k+1}\,.$$

The second inequality is from $\|\lambda_{k+1}(\psi_{S_k}(X_k) - \psi_{S_k}(\widetilde{X}_k))\| \leq \lambda_{k+1} z_k = a_{k+1}$. Also, note that $\widetilde{X}_t = X_t$. As in the proof of Theorem 4.9, we can write down as

$$
\begin{aligned}
&\psi_{S_k}(\widetilde{X}_k) + \lambda_{k+1}(\psi_{S_k}(X_k) - \psi_{S_k}(\widetilde{X}_k)) + V_k(\phi_{S_k}(X_k) - \psi_{S_k}(X_k)) + Z_{k+1} \\
&= \psi_{S_k}(\widetilde{X}_k) + \lambda_{k+1}(\psi_{S_k}(X_k) - \psi_{S_k}(\widetilde{X}_k)) + Z_{k+1}^{(1)} + V_k(\phi_{S_k} - \psi_{S_k})(X_k) + Z_{k+1}^{(2)}
\end{aligned}
$$

[10] In general, we can split the noise into $Z_t = Z_t^{(1)} + Z_t^{(2)}$ where $Z_t^{(1)} \sim \mathcal{N}(0, \frac{\eta^2\sigma^2}{\alpha^2} I_d)$ and $Z_t^{(2)} \sim \mathcal{N}(0, \frac{\eta^2\sigma^2}{\beta^2} I_d)$ are independent and $1/\alpha^2 + 1/\beta^2 = 1$. Then the part $G(\frac{2\sqrt{2}cz_{t-1}}{\eta\sigma}) \otimes C_{b/n}(G(\frac{2\sqrt{2}L}{b\sigma}))$ in the last line of the proof is replaced with $G(\frac{2\alpha cz_{t-1}}{\eta\sigma}) \otimes C_{b/n}(G(\frac{2\beta L}{b\sigma}))$.

where $Z_{k+1}^{(1)}, Z_{k+1}^{(2)} \sim \mathcal{N}(0, \frac{\eta^2\sigma^2}{2} I_d)$[11] are independent and similarly

$$\psi_{S_k'}(X_k') + V_k'(\phi_{S_k'}' - \psi_{S_k'})(X_k') + Z_{k+1}' = \psi_{S_k'}(X_k') + Z_{k+1}^{(1)'} + V_k'(\phi_{S_k'}' - \psi_{S_k'})(X_k') + Z_{k+1}^{(2)'} .$$

Thus

$$\begin{aligned} T(\widetilde{X}_{k+1}, X_{k+1}') &\geq T((\widetilde{X}_k, S_k, \lambda_{k+1}(\psi_{S_k}(X_k) - \psi_{S_k}(\widetilde{X}_k)) + Z_{k+1}^{(1)}), (X_k', S_k', Z_{k+1}^{(1)'})) \otimes R(\frac{\eta L}{b}, \frac{\eta\sigma}{\sqrt{2}}, b/n) \\ &\geq T((\widetilde{X}_k, S_k), (X_k', S_k')) \otimes R(a_{k+1}, \frac{\eta\sigma}{\sqrt{2}}) \otimes R(\frac{\eta L}{b}, \frac{\eta\sigma}{\sqrt{2}}, b/n) \\ &\geq T(\widetilde{X}_k, X_k') \otimes G(\frac{\sqrt{2}D}{(t-\tau)\eta\sigma}) \otimes C_{b/n}(G(\frac{2\sqrt{2}L}{b\sigma})) . \end{aligned}$$

Repeating this for $k = t-1, \dots, \tau$ yields the result. □

C.4.3. Choice of $t - \tau$ based on approximation

Since Theorem 4.9 and Theorem 4.10 hold for every $t - \tau$, we can calculate the corresponding f-DP bound for each $t - \tau$ and then take the pointwise maximum as a valid privacy guarantee; however, this may be computationally burdensome if t is large. One way to bypass this calculation is to approximate the composition of subsampled Gaussian mechanisms via CLT, (Lemma A.5), where the resulting f-DP bound becomes a GDP bound and thus optimization over $t - \tau$ is analytically tractable.

Proposition C.14. *In the setting of Theorem 4.9, by choosing (modulo floor or ceiling)*

$$t - \tau = -\frac{\log \frac{b^2\sigma(1-c)}{2\sqrt{2}nL\sqrt{\log(1/c)}} \sqrt{e^{4L^2/(b\sigma)^2}\Phi(\frac{3L}{b\sigma}) + 3\Phi(-\frac{L}{b\sigma}) - 2}}{\log(1/c)} - 1$$

`NoisySGD` *is approximately μ-GDP, where*

$$\mu = \sqrt{8\left(\frac{L}{b\sigma}\frac{c^{t-\tau+1}}{1-c}\right)^2 + \frac{2b^2}{n^2}(t-\tau)(e^{4L^2/(b\sigma)^2}\Phi(\frac{3L}{b\sigma}) + 3\Phi(-\frac{L}{b\sigma}) - 2)} . \tag{14}$$

Proof. By Lemma A.5,

$$C_{b/n}(G(\frac{2\sqrt{2}L}{b\sigma})) \otimes C_{b/n}(G(\frac{2L}{b\sigma}))^{\otimes(t-\tau)} \approx G\left(\sqrt{2}\frac{b}{n}\sqrt{(t-\tau)(e^{4L^2/(b\sigma)^2}\Phi(\frac{3L}{b\sigma}) + 3\Phi(-\frac{L}{b\sigma}) - 2)}\right) .$$

Also, by bounding

$$\frac{2\sqrt{2}L}{b\sigma}\frac{c^{t-\tau+1} - c^t}{1-c} \leq \frac{2\sqrt{2}L}{b\sigma}\frac{c^{t-\tau+1}}{1-c}$$

we obtain an approximate lower bound $G(\mu)$ of the form (14). As a function of $t - \tau \in (0, \infty)$ it is convex, and the first-order optimality condition provides the stated formula for $t - \tau$. □

Proposition C.15. *In the setting of Theorem 4.10, by choosing (modulo floor or ceiling)*

$$t - \tau = \frac{Dn}{b\eta\sigma\sqrt{e^{8L^2/(b\sigma)^2}\Phi(\frac{3\sqrt{2}L}{b\sigma}) + 3\Phi(-\frac{\sqrt{2}L}{b\sigma}) - 2}}$$

`NoisySGD` *is approximately μ-GDP, where*

$$\mu = \sqrt{\frac{2D^2}{\eta^2\sigma^2(t-\tau)} + 2\frac{b^2}{n^2}(t-\tau)(e^{8L^2/(b\sigma)^2}\Phi(\frac{3\sqrt{2}L}{b\sigma}) + 3\Phi(-\frac{\sqrt{2}L}{b\sigma}) - 2)} . \tag{15}$$

Proof. As in the proof of Proposition C.14, the CLT approximation of $C_{b/n}(G(\frac{2\sqrt{2}L}{b\sigma}))^{\otimes(t-\tau)}$ provides a lower bound $G(\mu)$ of the form (15), which is a convex function in $t - \tau$; the first-order optimality condition yields the stated result. □

[11] As before, setting $Z_{k+1}^{(1)} \sim \mathcal{N}(0, \frac{\eta^2\sigma^2}{\alpha^2} I_d)$ and $Z_{k+1}^{(2)} \sim \mathcal{N}(0, \frac{\eta^2\sigma^2}{\beta^2} I_d)$ with $1/\alpha^2 + 1/\beta^2 = 1$ replaces $G(\frac{\sqrt{2}D}{(t-\tau)\eta\sigma}) \otimes C_{b/n}(G(\frac{2\sqrt{2}L}{b\sigma}))$ with $G(\frac{\alpha D}{(t-\tau)\eta\sigma}) \otimes C_{b/n}(G(\frac{2\beta L}{b\sigma}))$.

C.5. Lower bounds

Here we elaborate on lower bounds for the amount of privacy preserved (i.e., upper bounds on the f-DP guarantee) that complement our results in §4. Note that an exactly matching bound for `NoisyGD` in the strongly convex setting was obtained in Theorem 4.2, and an asymptotically matching bound for `NoisySGD` in the constrained convex setting was obtained in (Altschuler & Talwar, 2022). Below, we present results for the other related settings using similar techniques. For the strongly convex setting, these lower bounds are built based on convex quadratics which yield iterates with explicit Gaussians; and for the constrained convex setting, these are obtained by comparing symmetric and biased (projected) Gaussians. We refer the readers to (Altschuler & Talwar, 2022) for further discussion about these constructions.

Theorem C.16. *Consider the setting of Theorem 4.3 or Theorem 4.6. There exist universal constants $0 < c_0 < 1/5, c_1 > 0$ such that if $\sigma^2 \le c_0 \frac{LD}{\eta n}$ and $\mu = c_1 \frac{1}{\sigma}\sqrt{\frac{LD}{\eta n}}$, then*

(a) `NoisyGD` is not μ-GDP for all $t \ge \frac{Dn}{\eta L} \ge \frac{1}{2}$.

(b) `NoisyCGD` is not μ-GDP for all $E \ge \frac{Db}{\eta L} \ge \frac{1}{2}$.

Proof. (a) For `NoisyGD`, let $\mu_0 = \frac{\mu}{c_1} = \frac{1}{\sigma}\sqrt{\frac{LD}{\eta n}}$. Consider $d = 1$ and loss functions such that $\nabla f_i(x) = 0$ for all $i \in [n]$, $\nabla f_i'(x) = 0$ for all $i \neq i^*$ and $\nabla f_{i^*}'(x) = -L$.[12] Also, let $X_0 = 0$ and $\mathcal{K} = [-\frac{D}{2}, \frac{D}{2}]$. Note that by Lemma A.3, a μ-GDP algorithm is $(\mu^2, \Phi(-\frac{\mu}{2}))$-DP. We will show that for $E = [-\frac{D}{2}, 0]$,

$$\mathbb{P}(X_t \in E) = \frac{1}{2}$$
$$\mathbb{P}(X_t' \in E) < \exp(-\mu^2)(\frac{1}{2} - \Phi(-\frac{\mu}{2}))$$

which implies that `NoisyGD` is not $(\mu^2, \Phi(-\frac{\mu}{2}))$-DP and thus `NoisyGD` is not μ-GDP. First, recall that

$$X_{k+1} = \Pi_{\mathcal{K}}(X_k + Z_{k+1})$$
$$X_{k+1}' = \Pi_{\mathcal{K}}(X_k' + \frac{\eta L}{n} + Z_{k+1}')$$

where $Z_{k+1}, Z_{k+1}' \sim \mathcal{N}(0, \eta^2\sigma^2)$. Since the distribution of X_k is symmetric for all k, $\mathbb{P}(X_t \in E) = \frac{1}{2}$. On the other hand, for $t_0 = t - \lceil 0.8\frac{Dn}{\eta L}\rceil + 1$ consider a process $\{X_k''\}_{t_0 \le k \le t}$ such that $X_{t_0}'' = -\frac{D}{2}$ and

$$X_{k+1}'' = \min\{X_k'' + \frac{\eta L}{n} + Z_{k+1}', \frac{D}{2}\}\,.$$

Then inductively, $\mathbb{P}(X_k' \le z) \le \mathbb{P}(X_k'' \le z)$ for all z. Letting $E_0 = \{\max_{t_0 \le k \le t} \sum_{j=t_0}^{k} Z_j \le 0.1D\}$, by Doob's submartingale inequality we have

$$\mathbb{P}(E_0^c) \le \exp(-\frac{(0.1D)^2}{2 \times \lceil 0.8\frac{Dn}{\eta L}\rceil \times (\eta\sigma)^2}) \le \exp(-\frac{0.01LD}{5.6n\eta\sigma^2}) = \exp(-\frac{0.01}{5.6}\mu_0^2)\,.$$

[12] For general $d > 1$, a similar argument (with slightly different constants) can be made by considering $\nabla f_{i^*}'(x) = -Le_1$ and $\mathcal{K} = [-\Theta(D), \Theta(D)] \times [-\Theta(D/\sqrt{d-1}), \Theta(D/\sqrt{d-1})]^{d-1}$ (constant factors chosen such that $\mathcal{K}$ has diameter D).

Also, conditioning on E_0, $X''_t = -\frac{D}{2} + \frac{\eta L}{n} \times \lceil 0.8\frac{Dn}{\eta L}\rceil + \sum_{j=t_0}^{t} Z_j \geq 0.3D + \sum_{j=t_0}^{t} Z_j$. Thus

$$\begin{aligned}
\mathbb{P}(X'_t \notin E) &\geq \mathbb{P}(X''_t > 0) \\
&\geq \mathbb{P}(\{X''_t > 0\} \cap E_0) \\
&\geq \mathbb{P}(\{0.3D + \sum_{j=t_0}^{t} Z_j > 0\} \cap E_0) \\
&\geq \mathbb{P}(0.3D + \sum_{j=t_0}^{t} Z_j > 0) - \exp(-\frac{0.01}{5.6}\mu_0^2) \\
&\geq \Phi(\frac{0.3}{\sqrt{2.8}}\mu_0) - \exp(-\frac{0.01}{5.6}\mu_0^2) \\
&\geq 1 - \exp(-\frac{0.9}{5.6}\mu_0^2) - \exp(-\frac{0.01}{5.6}\mu_0^2)
\end{aligned}$$

where the penultimate inequality is from that $\sum_{j=t_0}^{t} Z_j$ is a mean zero Gaussian with variance $\lceil 0.8\frac{Dn}{\eta L}\rceil \eta^2\sigma^2 \leq \frac{2.8Dn\eta\sigma^2}{L} = \frac{2.8D^2}{\mu_0^2}$, and the last inequality is from $\Phi(x) \geq 1 - \exp(-\frac{1}{2}x^2)$ for all $x \geq \frac{1}{\sqrt{2\pi}}$ (with $0.3\mu_0/\sqrt{2.8} \geq 0.3/\sqrt{2.8c_0} \geq 1/\sqrt{2\pi}$). By taking sufficiently small $c_1 < \sqrt{\frac{0.01}{5.6}}$ and $c_0 < c_1^2$ such that

$$\exp(-\frac{0.9}{5.6}\mu_0^2) + \exp(-\frac{0.01}{5.6}\mu_0^2) \leq \exp(-c_1^2\mu_0^2)(\frac{1}{2} - \Phi(-\frac{1}{2}))$$

for all $\mu_0^2 \geq \frac{1}{c_0}$, we have

$$\begin{aligned}
\exp(-\mu^2)(\frac{1}{2} - \Phi(-\frac{\mu}{2})) &= \exp(-c_1^2\mu_0^2)(\frac{1}{2} - \Phi(-\frac{c_1\mu_0}{2})) \\
&\geq \exp(-c_1^2\mu_0^2)(\frac{1}{2} - \Phi(-\frac{c_1}{2\sqrt{c_0}})) \\
&> \exp(-c_1^2\mu_0^2)(\frac{1}{2} - \Phi(-\frac{1}{2})) \\
&\geq \exp(-\frac{0.9}{5.6}\mu_0^2) + \exp(-\frac{0.01}{5.6}\mu_0^2) \\
&\geq \mathbb{P}(X'_t \in E)
\end{aligned}$$

as desired.

(b) The proof for `NoisyCGD` is similar to that for `NoisyGD` (recall that $t = lE$ and $n = lb$); consider the same loss functions, initialization, constraint set with $i^* \in B_l$. Then

$$\begin{aligned}
X_{k+1} &= \Pi_{\mathcal{K}}(X_k + Z_{k+1}) \\
X'_{k+1} &= \Pi_{\mathcal{K}}(X'_k + \frac{\eta L}{b}\mathbf{1}_{\{r(k)=l\}} + Z'_{k+1})
\end{aligned}$$

where $r(k) = k + 1 - l\lfloor\frac{k}{l}\rfloor$. For $t_0 = l(E - \lceil 0.8\frac{Db}{\eta L}\rceil) + 1$, consider a process $\{X''_k\}_{t_0 \leq k \leq t}$ such that $X''_{t_0} = -\frac{D}{2}$ and

$$X''_{k+1} = \min\{X''_k + \frac{\eta L}{b}\mathbf{1}_{\{r(k)=l\}} + Z'_{k+1}, \frac{D}{2}\}\,.$$

Then with the same events E and E_0, $\mathbb{P}(X_t \in E) = 1/2$ and

$$\begin{aligned}
\mathbb{P}(X'_k \leq z) &\leq \mathbb{P}(X''_k \leq z) \text{ for all } z \\
\mathbb{P}(E_0^c) &\leq \exp(-\frac{(0.1D)^2}{2 \times l\lceil 0.8\frac{Db}{\eta L}\rceil \times (\eta\sigma)^2}) \leq \exp(-\frac{0.01}{5.6}\mu_0^2)
\end{aligned}$$

and conditioning on E_0, $X''_t = -\frac{D}{2} + \frac{\eta L}{b} \times \lceil 0.8\frac{Db}{\eta L}\rceil \geq 0.3D + \sum_{j=t_0}^{t} Z_j$; the rest are identical.

□

Theorem C.17. *In the setting of Theorem 4.5, any valid f-DP lower bound for* `NoisyCGD` *satisfies*

$$G(\mu) \geq f$$

where $\mu = \frac{L}{b\sigma}\sqrt{\frac{1-c^{lE}}{1+c^{lE}} \frac{1-c^2}{(1-c^l)^2}}$.

Proof. Consider the loss functions in the proof of Theorem 4.2, with $X_0 = 0, \mathcal{K} = \mathbb{R}^d$ and $i^* \in B_l$. By direct calculation $X_{lE} = \mathcal{N}(0, \frac{1-c^{2lE}}{1-c^2}\eta^2\sigma^2 I_d)$ and $X'_{lE} = \mathcal{N}(\frac{\eta L}{b}\frac{1-c^{lE}}{1-c^l}v, \frac{1-c^{2lE}}{1-c^2}\eta^2\sigma^2 I_d)$, implying $T(X_{lE}, X'_{lE}) = G(\mu)$ with μ as stated. □

D. Numerical details and results

In this section, we provide numerical details of the figures and experiments in the main text and additional numerical results for different algorithms. Code reproducing these numerics can be found here: `https://github.com/jinhobok/shifted_interpolation_dp`.

D.1. Details for Figure 1

In Figure 1, we consider 1-strongly convex and 10-smooth loss functions with learning rate $\eta = 0.05$, effective sensitivity $L/(n\sigma) = 0.1$, and $t \in \{10, 20, 40, 80, 160\}$, with $t = 160$ in the left figure and $\delta = 10^{-5}$ in the right figure. Our f-DP bound is from Theorem 4.2, our RDP bound is from Theorem 4.2 and Lemma A.4, the prior RDP bound is from (Ye & Shokri, 2022, Theorem D.6), and the composition bound is from Theorem 4.1. For conversion from GDP and RDP to (ε, δ)-DP, see §A.1 and §A.2. We emphasize that different choices of parameters lead to qualitatively similar plots; see §D.3 for further numerical comparisons in other settings.

D.2. Details for §4.4

Here we provide further numerical details for the experiment in §4.4. The purpose of this simple numerical example is to corroborate our theoretical findings by comparing them with existing privacy bounds. As such, we simply compare algorithms with the same hyperparameters, and do not attempt to optimize these choices for individual algorithms.

In §4.4, Table 1 shows that our results provide improved privacy bounds. That table considers the privacy leakage of `NoisyCGD` in (ε, δ)-DP with regularization parameter $\lambda = 0.002$. Table 3 and Table 4 provide more details on this numerical comparison by also considering another algorithm (`NoisySGD`), another notion of privacy leakage (GDP), and another parameter ($\lambda = 0.004$). Details on these tables: for the GDP Composition privacy bound on `NoisySGD`, we present the approximate value of the GDP parameter provided by CLT since this is computationally tractable; for (ε, δ)-DP we compute the corresponding ε to an error of 10^{-3} using the numerical procedure in §D.5; and we convert the currently known best RDP bounds provided by (Ye & Shokri, 2022, Theorem 3.3) to (ε, δ)-DP using the numerical procedure in §A.1.

Table 3. More detailed version of Table 1, for GDP. Lists the GDP parameters of private algorithms for the regularized logistic regression problem. Note that GDP Composition yields the same privacy bound regardless of the regularization parameter. Our results provide improved privacy.

Epochs	GDP Composition		Our Bounds	
Algorithms	`NoisyCGD`	`NoisySGD`	`NoisyCGD`	
λ	$\{0.002, 0.004\}$		0.002	0.004
50	4.71	1.03	0.99	0.99
100	6.67	1.45	1.24	1.22
200	9.43	2.05	1.59	1.51

Table 4. More detailed version of Table 1, for (ε, δ)-DP. Lists ε of private algorithms on the regularized logistic regression problem for $\delta = 10^{-5}$. Note that GDP Composition yields the same privacy bound regardless of λ. Our results provide improved privacy over both GDP Composition and RDP.

Epochs	GDP Composition		RDP		Our Bounds	
Algorithms	`NoisyCGD`	`NoisySGD`	`NoisyCGD`		`NoisyCGD`	
λ	$\{0.002, 0.004\}$		0.002	0.004	0.002	0.004
50	30.51	4.44	5.82	5.61	4.34	4.32
100	49.88	6.65	7.61	7.00	5.60	5.51
200	83.83	10.11	9.88	8.38	7.58	7.09

These tables show that compared to our results (Theorem 4.5), the standard GDP Composition bound for `NoisyCGD` (Theorem 4.4) provides essentially no privacy. This is because that standard bound incurs a large privacy loss in each epoch (at the step in which the adjacent datasets use different gradients), and this privacy leakage accumulates indefinitely—whereas our analysis captures the contractivity of the algorithm's updates, which effectively ensures that previous gradient queries leak less privacy the longer ago they were performed. See §3 for a further discussion of this. Combined with the lossless conversion enabled by our f-DP analysis, our results also provide better privacy than the state-of-the-art RDP bounds.

Table 5 and Table 6 (reporting (mean) ± (standard deviation) of accuracies over 10 runs) show that (1) `NoisyCGD` and `NoisySGD` have comparable training and test accuracy for this problem, and (2) both algorithms improve when run longer, thus necessitating better privacy guarantees in order to achieve a target error (for either training or test) given a fixed privacy budget. Note that while `NoisySGD` enjoys better privacy bounds than `NoisyCGD` using the standard GDP Composition argument, our new privacy guarantees for `NoisyCGD` improve over GDP Composition bounds for both algorithms (c.f., Table 3 and Table 4). In particular, observe that while running algorithms longer leads to better accuracy, the privacy leak in `NoisySGD` from GDP Composition grows faster relative to our results (e.g., compare the values of ε when $E = 50$ and $E = 200$). This highlights the convergent dynamics of our privacy bounds and exemplifies how this enables algorithms to be run longer while preserving privacy.

Table 5. More detailed version of Table 2. Lists *training* accuracy (%) of `NoisyCGD` and `NoisySGD` for regularized logistic regression. Note that both algorithms perform similarly and improve when run longer.

Epochs	`NoisyCGD`		`NoisySGD`	
λ	0.002	0.004	0.002	0.004
50	89.36 ± 0.03	89.23 ± 0.02	89.36 ± 0.04	89.22 ± 0.04
100	90.24 ± 0.03	90.00 ± 0.03	90.25 ± 0.02	89.99 ± 0.03
200	90.85 ± 0.02	90.39 ± 0.04	90.84 ± 0.03	90.37 ± 0.02

Table 6. More detailed version of Table 2. Lists *test* accuracy (%) of `NoisyCGD` and `NoisySGD` for regularized logistic regression. Again, note that both algorithms perform similarly and improve when run longer.

Epochs	`NoisyCGD`		`NoisySGD`	
λ	0.002	0.004	0.002	0.004
50	90.12 ± 0.04	90.03 ± 0.07	90.12 ± 0.08	90.00 ± 0.06
100	90.94 ± 0.07	90.70 ± 0.05	90.97 ± 0.04	90.75 ± 0.03
200	91.37 ± 0.08	91.02 ± 0.07	91.40 ± 0.07	91.01 ± 0.04

For the experiment, we closely follow the setting considered in (Ye & Shokri, 2022)—for proofs and details on theoretical guarantees with respect to the setting, see (Ye & Shokri, 2022, Section 5). The MNIST dataset has $n = 60000$ training data points and 10000 test data points; for both `NoisyCGD` and `NoisySGD`, we set the parameters as $C = 8$, $\eta = 0.05$, $b = 1500$, $\sigma = 1/100$, $L = 10$, $E \in \{50, 100, 200\}$ and $\lambda \in \{0.002, 0.004\}$. First, we clip the feature so that it has norm

C. For the loss function $l(\theta, (x, y))$ of the (unregularized) logistic regression, we calculate the gradient for each data point (x, y) as

$$\nabla f(\theta, (x, y)) = \frac{\nabla l(\theta, (x, y))}{\|\nabla l(\theta, (x, y))\|} \cdot \min\{\|\nabla l(\theta, (x, y))\|, \frac{L}{2}\} + \lambda\theta \, .$$

In other words, we first clip the gradient by $L/2$ so that the gradient sensitivity is L, and add a gradient $\lambda\theta$ of the regularization term $(\lambda/2)\|\theta\|^2$ (which does not affect the gradient sensitivity).

D.3. Additional numerics

Here we provide additional numerical results to illustrate our privacy bounds in §4, by comparing our f-DP bounds with the counterparts derived by the standard GDP Composition analysis. We cover the settings and algorithms covered in the main text over a broad range of parameters, emphasizing the convergent dynamics of our privacy bounds. The different settings lead to qualitatively similar comparisons. Recall that the relevant parameters of the algorithms are the learning rate η, noise rate σ, number of data points n, batch size b, gradient sensitivity L, and diameter D of the constraint set $\mathcal{K}$; see §2.3.

D.3.1. NoisyGD

Figure 6 shows our results for f-DP (left) and its conversion into (ε, δ)-DP (right) for NoisyGD in the strongly convex setting (Theorem 4.2), where our bound is exact. In contrast, observe that while the bound from GDP Composition is nearly tight for a small number of iterations t, the guarantee becomes vacuous as t increases. This is also evident from the (ε, δ)-DP plot, where the discrepancy between the two bounds increases in t.

In Table 7, since we obtain GDP bounds, we provide the GDP parameter μ as a function of the number of iterations t and the contractivity $c = \max\{|1 - \eta m|, |1 - \eta M|\}$. All values in Table 7 scale linearly in the effective sensitivity $L/(n\sigma)$; for simplicity we set it to 0.1. Note that the GDP Composition bound is independent of c because it is not "geometrically aware" in the sense described in §3. Our bound is optimal and always improves over GDP composition—substantially so as t increases.

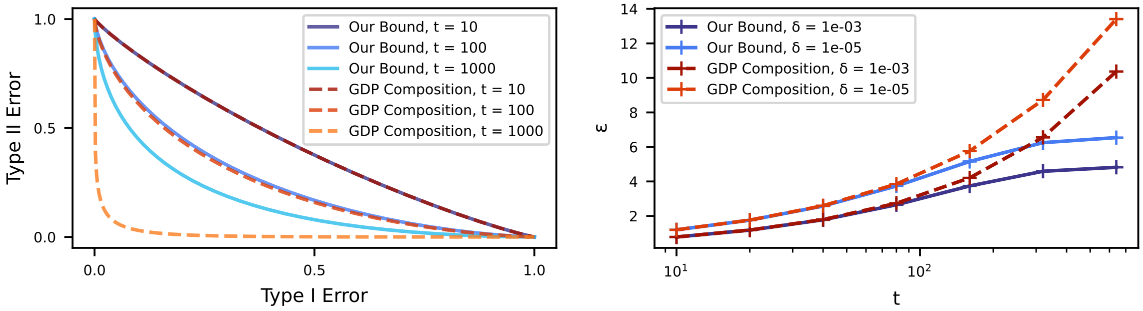

Figure 6. Comparison of our exact privacy characterization (Theorem 4.2) with the standard GDP Composition bound (Theorem 4.1) for NoisyGD, for $c = 0.99$. Shown for f-DP (left) and (ε, δ)-DP (right).

Table 7. GDP parameter μ from our exact privacy characterization (Theorem 4.2), for varying t and c.

Steps	GDP Composition	Our Bounds				
c	$\{0.92, 0.96, 0.98, 0.99, 0.995\}$	0.92	0.96	0.98	0.99	0.995
10	0.316	0.308	0.314	0.316	0.316	0.316
100	1.000	0.490	0.688	0.871	0.961	0.990
1000	3.162	0.490	0.700	0.995	1.411	1.984

Figure 7 and Table 8 turn to the setting of constrained convex optimization in Theorem 4.3. In the (ε, δ)-DP figure, we plot the minimum ε between Theorem 4.3 and GDP Composition. A distinctive feature from both plots is that our privacy bound stays constant after a number of iterations, compared to GDP Composition. In particular, there exists a threshold

$t^* = t^*(L/n, \eta)$ such that the algorithm can run beyond the threshold (and even indefinitely) with a provable guarantee of μ^*-GDP. To highlight this fact, we provide the pairs of (t^*, μ^*) in the table over multiple combinations of parameters. We set the diameter of the constraint set $\mathcal{K}$ to be $D = 1$ and noise parameter to be $\sigma = 8$; note that as in the previous case, the GDP parameters in this setting scale linearly with respect to $1/\sigma$. Other parameter choices lead to qualitatively similar comparisons.

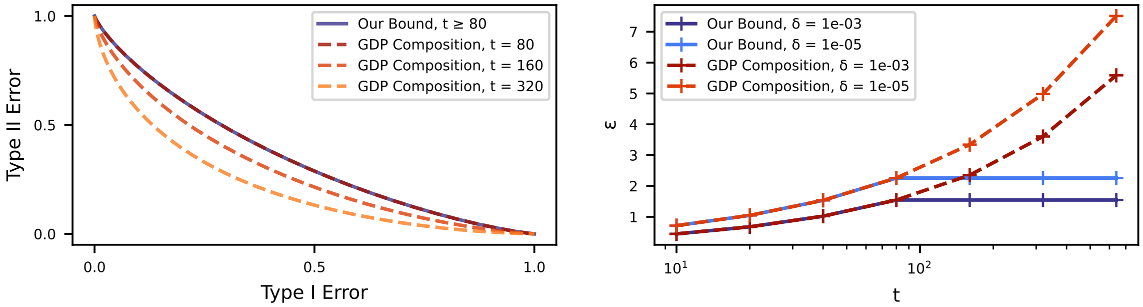

Figure 7. Comparison of our bound (Theorem 4.3) with the standard GDP Composition bound (Theorem 4.1) for `NoisyGD`, for $L/n = 0.5$ and $\eta = 0.1$. Shown for f-DP (left) and (ε, δ)-DP (right).

Table 8. Threshold number of iterations t^* at which point our GDP bound μ^* from Theorem 4.3 no longer increases. Shown for varying parameters L/n and η.

$L/n \setminus \eta$	0.2	0.1	0.05
0.25	(80, 0.280)	(160, 0.395)	(320, 0.559)
0.5	(40, 0.395)	(80, 0.559)	(160, 0.791)
1	(20, 0.559)	(40, 0.791)	(80, 1.118)

D.3.2. NoisyCGD

Figure 8 and Table 9 show the analog of Figure 6 and Table 7, now for `NoisyCGD` rather than `NoisyGD`. Recall that l denotes the number of batches and $c = \max\{|1 - \eta m|, |1 - \eta M|\}$ is the contraction factor. All GDP parameters scale linearly in the effective sensitivity $L/(b\sigma)$; we set it to 0.2 for concreteness. The improvement of our bounds over the standard GDP Composition bound is pronounced: our bounds yield strong privacy in both f-DP (left) and (ε, δ)-DP (right), whereas the GDP Composition bound becomes effectively non-private as the number of epochs E increases. This is also evident from Table 9, where our bound produces better privacy (even with $E = 500$ epochs) than GDP Composition (even with $E = 5$).

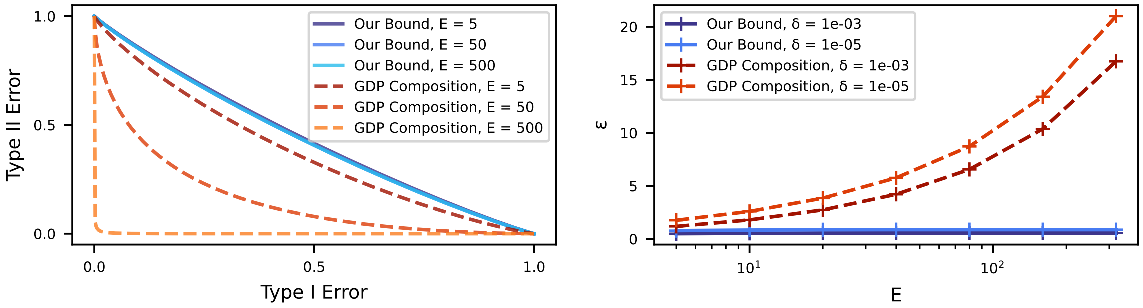

Figure 8. (Left) f-DP, (Right) (ε, δ)-DP comparison of our bound (Theorem 4.5) with the standard GDP Composition bound (Theorem 4.4) for `NoisyCGD`, for $c = 0.99$ and $l = 20$.

Table 9. GDP parameter μ, for varying number of epochs E and contractivity c.

Epochs	GDP Composition	Our Bounds								
l	$\{10, 20, 40\}$		10			20			40	
c	$\{0.98, 0.99, 0.995\}$	0.98	0.99	0.995	0.98	0.99	0.995	0.98	0.99	0.995
5	0.447	0.229	0.233	0.235	0.211	0.215	0.217	0.202	0.205	0.208
50	1.414	0.270	0.334	0.410	0.216	0.237	0.275	0.203	0.208	0.219
500	4.472	0.270	0.336	0.439	0.216	0.237	0.276	0.203	0.208	0.219

Next we turn to the setting of constrained convex losses from Theorem 4.6. Again, our bounds converge in the number of epochs quickly and uniformly improve over the bounds from GDP Composition after only a few number of epochs—for the (ε, δ)-DP plot, we show the minimum ε between Theorem 4.6 and GDP Composition. In particular, from the table one can observe that there are even a few cases in which our bounds are better than those from GDP Composition after less than 10 epochs. We chose $D = 1$ and $\sigma = 3$; the GDP parameters scale linear in $1/\sigma$.

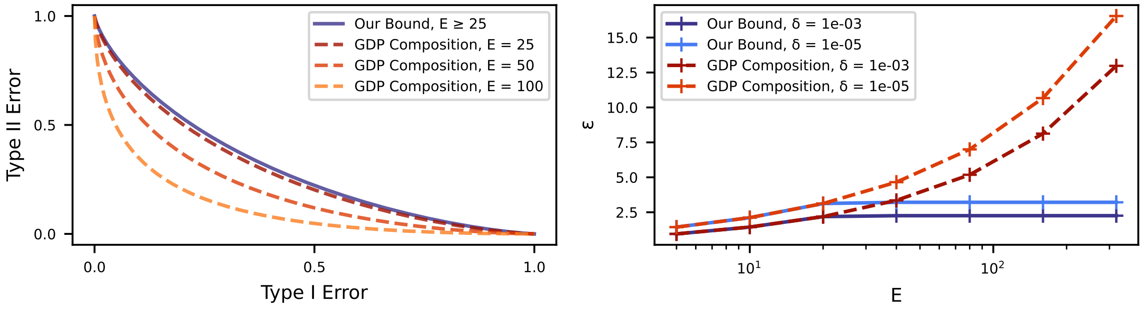

Figure 9. (Left) f-DP, (Right) (ε, δ)-DP comparison of our bound (Theorem 4.6) with the existing bound from GDP Composition (Theorem 4.4) for `NoisyCGD` under constrained set, with $\eta = 0.02$, $L/b = 0.5$ and $l = 20$.

Table 10. (E^*, μ^*) over different values of $(L/b, \eta)$, with $l = 10$.

$L/b \setminus \eta$	0.04	0.02	0.01
0.25	(31, 0.534)	(62, 0.750)	(123, 1.057)
0.5	(17, 0.764)	(33, 1.067)	(65, 1.500)
1	(10, 1.106)	(20, 1.528)	(40, 2.134)

Table 11. (E^*, μ^*) over different values of $(L/b, \eta)$, with $l = 20$.

$L/b \setminus \eta$	0.04	0.02	0.01
0.25	(16, 0.382)	(31, 0.534)	(62, 0.750)
0.5	(9, 0.553)	(17, 0.764)	(33, 1.067)
1	(5, 0.816)	(10, 1.106)	(20, 1.528)

Table 12. (E^*, μ^*) over different values of $(L/b, \eta)$, with $l = 40$.

$L/b \setminus \eta$	0.04	0.02	0.01
0.25	(8, 0.276)	(16, 0.382)	(31, 0.534)
0.5	(5, 0.408)	(9, 0.553)	(17, 0.764)
1	(3, 0.624)	(5, 0.816)	(10, 1.106)

D.3.3. NoisySGD

For brevity, here we consider just the setting of constrained convex losses; similar plots can be obtained for the strongly convex setting. Figure 10 compares our new privacy bound (Theorem 4.10) with the standard GDP Composition bound (Theorem 4.8), by illustrating the f-DP tradeoff curves of both bounds for a broad range of parameters. For most parameter choices, our bounds provide reasonable privacy that is valid even beyond the number of iterations in the plots. On the other hand, the divergence of the GDP Composition bound clearly degrades the privacy as the number of iterations increases.

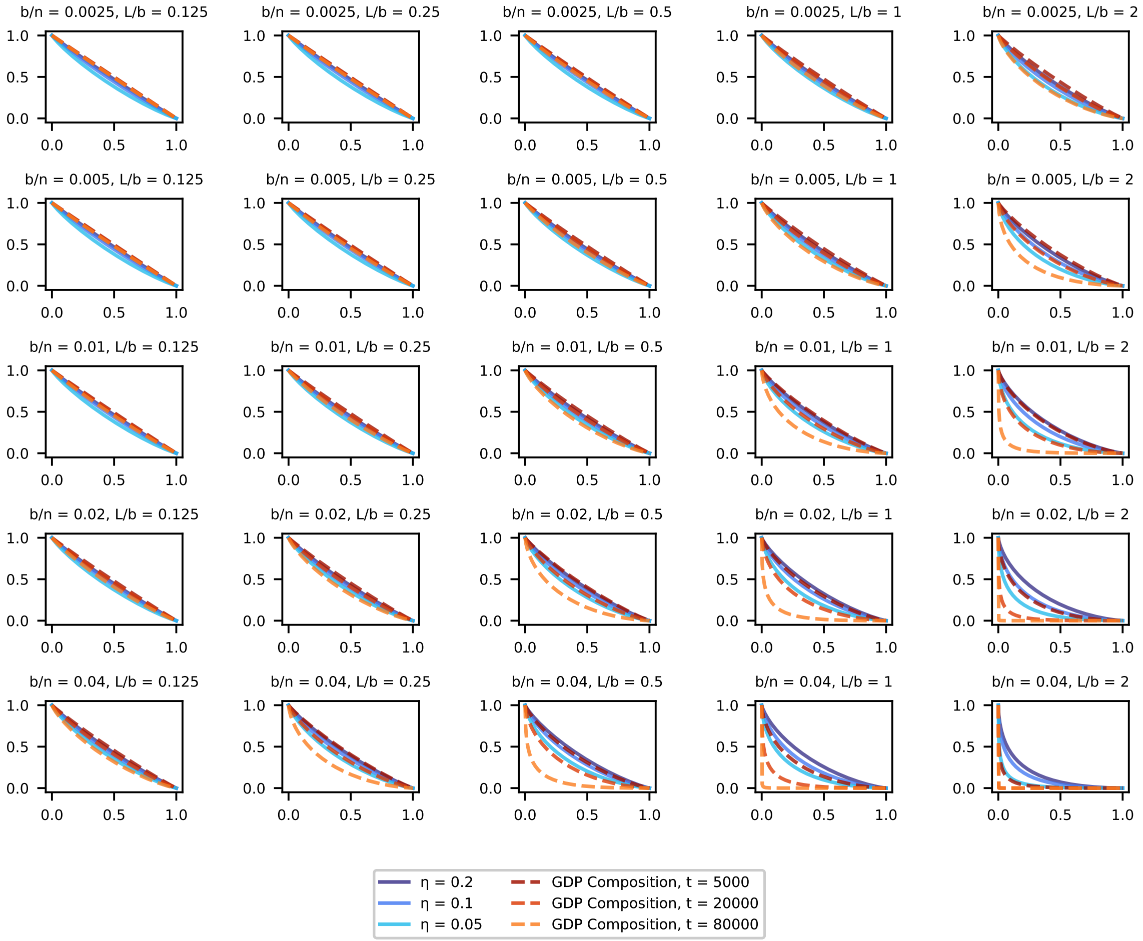

Figure 10. Comparison of our new privacy bound (Theorem 4.10) with the standard GDP Composition bound (Theorem 4.8) for NoisySGD in the setting of constrained convex losses. Each subplot illustrates the f-DP tradeoff curve of these bounds, for a given relative batch size b/n and effective sensitivity L/b.

To make this figure, we approximated the compositions of $C_{b/n}(G(\cdot))$ in Theorem 4.10 by CLT, by choosing the best privacy bound among $t - \tau \in \{100, 200, \dots, 4900\}$ after applying CLT; this is a valid approximation by Lemma A.11. We also note that for improved numerical results, instead of the respective factors of $(\sqrt{2}, \sqrt{2})$ inside $G(\cdot)$ and $C_{b/n}(G(\cdot))$ of the statement of Theorem 4.10 we used $(\sqrt{10}, \sqrt{10}/3)$ (see the proof of Theorem 4.10 for details). For parameters unspecified in the plots we used $D = 1$ and $\sigma = 3$.

D.4. Comparison of privacy bounds for the exponential mechanism

Let f_ε be the tradeoff function corresponding to $(\varepsilon, 0)$-DP. From (Dong et al., 2022, Proposition 3), it is straightforward to check that $f_\varepsilon \geq G(\mu)$ iff

$$e^\varepsilon \leq \frac{1 - \Phi(-\frac{\mu}{2})}{\Phi(-\frac{\mu}{2})} .$$

Plugging in $\varepsilon = 2LD$ and $\mu = 2\sqrt{LD}$ from Corollary 5.4, one can check (using standard nonlinear equation solvers) that this holds iff $LD \leq c^*$ where $c^* \approx 0.676$.

D.5. Numerical composition of subsampled GDP

Here we mention details on the numerical procedure for calculating an (ε, δ)-DP bound from an f-DP bound of the form $f = C_p(G(\mu))^{\otimes t}$; this is used in §4.4. This formula appears in multiple settings of `NoisySGD`, with each $C_p(G(\mu))$ representing the f-DP of subsampled Gaussian mechanism. This conversion process is important for both notions: First, while the composition can be approximated by CLT (Lemma A.5), in practice the approximation is not enough to guarantee whether the algorithm achieves a given privacy budget, typically expressed in (ε, δ)-DP. On the other hand, if one can obtain an accurate collection of different (ε, δ)-DP bounds, it can be converted into an f-DP bound of comparable accuracy due to the duality between the two notions (Dong et al., 2022, Proposition 5 & 6).

We implement the framework of *privacy loss random variables* (PRV)—which is an equivalent notion of f-DP—and the corresponding analytical procedure provided in (Gopi et al., 2021). Given a fixed value of δ and (possibly different) compositions of private mechanisms, the algorithm presented in the paper allows one to numerically calculate ε with user-specified margin of error. We refer the readers to (Gopi et al., 2021) for the background and overview of the PRV framework and only present relevant results for the problem of our interest.[13]

The privacy curve and PRV are characterized as follows (Gopi et al., 2021, Definition 2.1 & 3.1).

Definition D.1 (Privacy curve and PRV). Let $f = T(X, Y)$ be a tradeoff function. Then the *privacy curve* $\delta : \mathbb{R} \to [0, 1]$ with respect to (X, Y) is defined as $\delta(X||Y)(\varepsilon) = \sup_E\{\mathbb{P}(Y \in E) - e^\varepsilon \mathbb{P}(X \in E)\}$ where the supermum is over all events. Conversely, given a privacy curve $\delta : \mathbb{R} \to [0, 1]$, (X, Y) are *privacy loss random variables* if the following holds.

- X, Y are supported on the extended real line $\bar{\mathbb{R}}$.
- $\delta(X||Y) \equiv \delta$.
- Let $X(t), Y(t)$ respectively be the probability density functions of X, Y. Then $Y(t) = e^t X(t)$ and $Y(-\infty) = X(\infty) = 0$.

The probability density functions of PRVs can be calculated from the privacy curve (Gopi et al., 2021, Theorem 3.3).

Lemma D.2 (Conversion). *Given a privacy curve $\delta : \mathbb{R} \to [0, 1]$, the probability density functions of its PRVs (X, Y) are given as $Y(t) = \delta''(t) - \delta'(t)$ and $X(t) = e^t(\delta''(t) - \delta'(t))$.*

Also, symmetric tradeoff functions have the simple form of PRVs (X, Y) with $X = -Y$ (Gopi et al., 2021, Proposition D.9).

Lemma D.3 (Symmetry). *If (X, Y) are PRVs for a privacy curve $\delta(P||Q)$, the PRVs for $\delta(Q||P)$ are $(-Y, -X)$. In particular, if the privacy curve is symmetric (i.e., $\delta(P||Q) = \delta(Q||P)$; equivalently, the corresponding tradeoff function is symmetric) then $X = -Y$.*

By the following result, we can numerically calculate the (ε, δ)-DP converted from f-DP for $f = C_p(G(\mu))^{\otimes t}$.

Proposition D.4. *Let (X, Y) be such that the CDF of Y is given as*

$$F_Y(t) = \begin{cases} p\Phi(\frac{\varepsilon^+}{\mu} - \frac{\mu}{2}) + (1-p)\Phi(\frac{\varepsilon^+}{\mu} + \frac{\mu}{2}) & t > 0 \\ \Phi(-\frac{\varepsilon^-}{\mu} - \frac{\mu}{2}) & t \leq 0 \end{cases}$$

[13] We also note that while a corresponding result for the one-sided version of $G(\mu)_p = T(\mathcal{N}(0, 1), p\mathcal{N}(\mu, 1) + (1-p)\mathcal{N}(0, 1))$ was already presented in (Gopi et al., 2021) and is often interchangeably used, it is quantitatively different from $C_p(G(\mu))$, even in the limiting regime of CLT. Compare, for example, Lemma A.5 and (Bu et al., 2020).

where $\varepsilon^+ = \log((p-1+e^t)/p), \varepsilon^- = \log((p-1+e^{-t})/p)$ *and* $X = -Y$. *Then* (X, Y) *are PRVs for the tradeoff function* $C_p(G(\mu))$.

Proof. Let δ, δ_0 respectively be the privacy curves of $C_p(G(\mu))$ and $G(\mu)_p$. Then it is straightforward to check that

$$\delta(t) = \begin{cases} \delta_0(t) & t > 0 \\ 1 - e^t(1 - \delta_0(-t)) & t \le 0 \end{cases}$$

from the definition of $C_p(G(\mu))$ (as a symmetrized version of $G(\mu)_p$; see Definition 4.7) and the duality between (ε, δ)-DP and f-DP. By taking antiderivative from Lemma D.2, the CDF of Y is given as

$$F_Y(t) = \begin{cases} \delta_0'(t) - \delta_0(t) + C & t > 0 \\ -e^t \delta_0'(-t) - 1 + C & t \le 0 \end{cases}$$

for some constant C. By either obtaining $\delta_0(t)$ directly from (Dong et al., 2022, Lemma 2) or comparing the $t > 0$ part of $F_Y(t)$ with (Gopi et al., 2021, Proposition C.4), one can derive the formula of $F_Y(t)$ as stated with $C = 1$. Also, $X = -Y$ from Lemma D.3. □