The convergence of the privacy loss in strongly convex optimization

Aayushman
IIT Madras
aayushman@dsai.iitm.ac.in

Kaushik Doddamani
IIT Madras
kaushik.doddamani@dsai.iitm.ac.in

Pranav
IIT Madras
cs22b015@smail.iitm.ac.in

1 Motivation

Differential privacy provides a rigorous mathematical framework for quantifying how much the output of an algorithm can change when a single training example is modified. It was introduced by Dwork et al. Dwork et al. [2006] and has become a standard definition for privacy-preserving data analysis and machine learning Dwork and Roth [2014]. Formally, a differentially private algorithm is required to produce statistically similar outputs on adjacent datasets, where adjacency means that the two datasets differ in one individual record.

In private machine learning, one of the most widely used training procedures is noisy stochastic gradient descent, also known as DP-SGD or Noisy-SGD. The basic idea is to modify ordinary stochastic gradient descent by clipping per sample gradients and adding Gaussian noise to the minibatch gradient at every iteration. This algorithm was popularized in deep learning by Abadi et al. Abadi et al. [2016], and related implementations are available in libraries such as TensorFlow Privacy, Opacus, and JAX-based systems TensorFlow Authors [2019], Yousefpour et al. [2021], Bradbury et al. [2018]. Noisy-SGD has become a standard algorithmic primitive for private learning, because of its simplicity and scalability.

The question studied in this report is not the optimization error of Noisy-SGD, but its privacy loss as a function of the number of iterations. Classical analyses usually treat each noisy gradient step as a separate randomized mechanism and then apply privacy composition over all T iterations. In the Rényi differential privacy framework introduced by Mironov Mironov [2017], this gives bounds of the form

$$\varepsilon_{\mathrm{RDP}}(T) \lesssim \frac{\alpha L^2}{n^2\sigma^2} T.$$

Such bounds grow linearly with the number of iterations. This type of behavior appears in standard analyses based on privacy amplification by sampling and composition Abadi et al. [2016], Wang et al. [2019]. Consequently, these bounds suggest that every additional training iteration consumes additional privacy budget, even if the optimization procedure has already nearly converged.

The work of Altschuler and Talwar [2023] shows that this linear growth is not the correct qualitative behavior in the smooth convex bounded setting. They consider Noisy-SGD for convex, L-Lipschitz, and M-smooth losses over a bounded convex domain $K \subset \mathbb{R}^d$ with diameter D. For step size $\eta \leq 2/M$, their result gives the informal RDP scaling

$$\varepsilon_{\mathrm{RDP}}(T) \lesssim \frac{\alpha L^2}{n^2\sigma^2} \min\left\{T, \frac{Dn}{L\eta}\right\}.$$

Thus the privacy loss grows initially, but stops increasing after the time scale

$$\bar{T} \asymp \frac{Dn}{L\eta}.$$

This phenomenon is called privacy saturation: after the burn-in period $\bar{T}$, running Noisy-SGD for more iterations does not increase the privacy loss beyond constant factors Altschuler and Talwar [2023].

The reason for this saturation is that Noisy-SGD should not be viewed only as a sequence of independent privacy-releasing mechanisms. It is also a noisy iterative process. In the smooth convex setting, a gradient step with step size $\eta \leq 2/M$ is non-expansive, and projection onto a convex set is also non-expansive. This contractive structure is the basis of privacy amplification by iteration Dwork and Feldman [2018], Altschuler and Talwar [2023]. Gaussian noise then smooths the distributions of the iterates, making it possible for the dynamics to partially forget discrepancies from earlier iterations.

The proof combines two privacy mechanisms. The first is privacy amplification by sampling, at each iteration, the changed datapoint appears in the minibatch only with probability b/n. Hence the privacy cost of a single noisy minibatch update is smaller than the cost of always using the changed datapoint. This mechanism is standard in the analysis of DP-SGD and subsampled RDP mechanisms Abadi et al. [2016], Wang et al. [2019]. Applied to a block of R iterations, it gives a contribution of order

$$\frac{\alpha L^2}{n^2\sigma^2}R.$$

The second mechanism is privacy amplification by iteration. If two runs of the algorithm start the final R iterations from different points in K, their distance is at most D, because K has diameter D. Since the subsequent updates are contractive and noisy, the distinguishability due to this initial displacement decreases with R. In the analysis of Altschuler and Talwar [2023], this gives a contribution of order

$$\frac{\alpha D^2}{\eta^2\sigma^2 R}.$$

Balancing the increasing sampling term and the decreasing iteration term gives

$$R \asymp \frac{Dn}{L\eta},$$

which is exactly the saturation time.

The boundedness of the domain is essential in this argument. If both Noisy-SGD trajectories remain in a set K of diameter D, then at any intermediate time τ, their distance is at most D. Therefore the proof can restart the privacy analysis from time $T - R$ and analyze only the final R iterations. The earlier iterations are not released, and their effect is summarized only through the fact that both trajectories lie inside the same bounded set. This is the key step that allows the privacy bound to stop growing with T Altschuler and Talwar [2023].

This report focuses on explaining the upper-bound argument for privacy saturation. The main proof is based on the interaction between sampling, contraction, Gaussian smoothing, and the bounded geometry of the feasible set. The optimal-transport viewpoint enters through the use of Wasserstein-type control of the distance between intermediate distributions, which is closely related to the privacy amplification by iteration argument Dwork and Feldman [2018], Peyré and Cuturi [2020], Altschuler and Talwar [2023]. The lemmas used in the proof are stated in the preliminaries, while their detailed proofs are deferred to the appendix.

2 Problem Setup

Let

$$X = \{x_1, \ldots, x_n\} \qquad \text{and} \qquad X' = \{x'_1, \ldots, x'_n\}$$

be adjacent datasets, meaning that they differ in exactly one data point. Without loss of generality, assume that they differ at index $i^\star$.

For each data point x_i, let $f_i : K \to \mathbb{R}$ be the corresponding loss function. Throughout the report, we assume that $K \subset \mathbb{R}^d$ is a closed convex set with diameter

$$\operatorname{diam}(K) = \sup_{u,v \in K} \|u - v\| = D.$$

We also assume that each loss f_i is convex, L-Lipschitz, and M-smooth on K.

Noisy-SGD starts from an initialization

$$W_0 = w_0 \in K.$$

At iteration t, a minibatch $B_t \subset [n]$ of size b is sampled, and the update is

$$W_{t+1} = \Pi_K \left[W_t - \eta \left(G_t(W_t) + Z_t \right) \right], \tag{1}$$

where

$$G_t(W_t) = \frac{1}{b} \sum_{i \in B_t} \nabla f_i(W_t), \qquad Z_t \sim \mathcal{N}(0, \sigma^2 I_d). \tag{2}$$

Here Π_K denotes Euclidean projection onto K. Since the noise is multiplied by the step size η, the Gaussian noise added to the iterate has covariance

$$\eta^2 \sigma^2 I_d.$$

Let W_T and W_T' denote the final iterates obtained by running Noisy-SGD on X and X', respectively. The objective is to bound the Rényi divergence

$$\mathcal{D}_\alpha(P_{W_T} \| P_{W_T'}),$$

where P_{W_T} and $P_{W_T'}$ denote the distributions of the final outputs.

The main theorem shows that this privacy loss satisfies

$$\mathcal{D}_\alpha(P_{W_T} \| P_{W_T'}) \lesssim \frac{\alpha L^2}{n^2 \sigma^2} \min \left\{ T, \frac{Dn}{L\eta} \right\}.$$

Thus, the privacy loss grows linearly only up to the time scale $Dn/(L\eta)$, and then saturates.

3 Preliminaries

Definition 1 (Asymptotic notation)**.** *Asymptotic notation suppresses only universal numerical constants. More precisely, for two nonnegative quantities A and B, we write*

$$A \lesssim B$$

if there exists a universal constant $C > 0$, independent of the problem parameters, such that

$$A \leq CB.$$

We write

$$A \asymp B$$

if both $A \lesssim B$ and $B \lesssim A$ hold. Equivalently, there exist universal constants $c, C > 0$ such that

$$cB \leq A \leq CB.$$

Definition 2 ((ε, δ)-differential privacy)**.** *A randomized algorithm $\mathcal{A}$ satisfies (ε, δ)-differential privacy if for all adjacent datasets X, X' and all events S,*

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

In this report, we mainly use Rényi differential privacy because it composes cleanly across iterations.

Definition 3 (Rényi divergence)**.** *Let P and Q be two distributions with densities p and q. For $\alpha > 1$, the order-α Rényi divergence is*

$$\mathcal{D}_\alpha(P \| Q) = \frac{1}{\alpha - 1} \log \int p(x)^\alpha q(x)^{1-\alpha} \, dx.$$

Definition 4 ((α, ε)-RDP)**.** *A randomized algorithm $\mathcal{A}$ satisfies (α, ε)-Rényi differential privacy if for every pair of adjacent datasets X, X',*

$$\mathcal{D}_\alpha(P_{\mathcal{A}(X)} \| P_{\mathcal{A}(X')}) \leq \varepsilon.$$

An RDP guarantee can be converted into an ordinary DP guarantee. If $\mathcal{A}$ satisfies $(\alpha, \varepsilon_\alpha)$-RDP, then $\mathcal{A}$ satisfies $(\varepsilon_\delta, \delta)$-DP with

$$\varepsilon_\delta = \varepsilon_\alpha + \frac{\log(1/\delta)}{\alpha - 1}.$$

Lemma 1 (Gradient-step contraction)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and M-smooth. If $\eta \le 2/M$, then for all $u, v \in \mathbb{R}^d$,*

$$\Big\| \big(u - \eta \nabla f(u)\big) - \big(v - \eta \nabla f(v)\big) \Big\| \le \|u - v\|.$$

Thus a gradient step with step size $\eta \le 2/M$ is non-expansive. Since projection onto a closed convex set is also non-expansive,

$$\|\Pi_K(u) - \Pi_K(v)\| \le \|u - v\|,$$

the projected gradient step is non-expansive as well.

Lemma 2 (Post-processing)**.** *Let μ and ν be two probability distributions. For any function h,*

$$D_\alpha(h_\# \mu \,\|\, h_\# \nu) \le D_\alpha(\mu \,\|\, \nu)$$

This allows us to upper-bound the privacy loss of W_T by analyzing a larger object, such as $(W_T, Z_{\tau:T-1})$.

Lemma 3 (Conditional composition)**.** *Let*

$$X_{1:k} = (X_1, \ldots, X_k), \qquad Y_{1:k} = (Y_1, \ldots, Y_k)$$

be two sequence of random variables. Then

$$\mathcal{D}_\alpha(P_{X_{1:k}} \| P_{Y_{1:k}}) \le \sum_{i=1}^{k} \sup_{x_{<i}} \mathcal{D}_\alpha \left(P_{X_i | X_{<i} = x_{<i}} \middle\| P_{Y_i | Y_{<i} = x_{<i}} \right),$$

where $x_{<i} = (x_1, \ldots, x_{i-1})$.

Definition 5 (Rényi Divergence of the Sampled Gaussian Mechanism)**.** *For Rényi parameter $\alpha \ge 1$, mixing probability $q \in (0, 1)$, and noise parameter $\sigma > 0$, define*

$$S_\alpha(q, \sigma) := \mathcal{D}_\alpha\Big(\mathcal{N}(0, \sigma^2) \;\Big\|\; (1-q)\mathcal{N}(0, \sigma^2) + q\mathcal{N}(1, \sigma^2) \Big).$$

Lemma 4 (Extrema for Rényi Divergence of the Sampled Gaussian Mechanism)**.** *For any Rényi parameter $\alpha \ge 1$, mixing probability $q \in (0, 1)$, noise level $\sigma > 0$, dimension $d \in \mathbb{N}$, and radius $R > 0$,*

$$\sup_{\mu \in \mathcal{P}(\mathbb{B}_d)} \mathcal{D}_\alpha\Big(\mathcal{N}(0, \sigma^2 I_d) \;\Big\|\; (1-q)\mathcal{N}(0, \sigma^2 I_d) + q\big(\mathcal{N}(0, \sigma^2 I_d) * \mu\big) \Big) = S_\alpha(q, \sigma/R),$$

where $\mathcal{P}(\mathbb{B}_d)$ denotes the set of Borel probability distributions supported on the ball of radius R in $\mathbb{R}^d$.

Lemma 5 (Bound on Rényi Divergence of the Sampled Gaussian Mechanism)**.** *Consider Rényi parameter $\alpha > 1$, mixing probability $q \in (0, 1/5)$, and noise level $\sigma \ge 4$. If $\alpha \le \alpha^*(q, \sigma)$, then*

$$S_\alpha(q, \sigma) \le \frac{2\alpha q^2}{\sigma^2}.$$

Proposition 1 (New PABI bound)**.** *Let X_T, X'_T, and s_t be as in 2. Consider any $\tau \in \{0, \ldots, T-1\}$ and any sequence $a_{\tau+1}, \ldots, a_T$ such that*

$$z_t := D + \sum_{i=\tau+1}^{t} (s_i - a_i)$$

is nonnegative for all t and satisfies $z_T = 0$. If K has diameter D, then

$$\mathcal{D}_\alpha\Big(\mathbb{P}_{X_T} \;\Big\|\; \mathbb{P}_{X'_T} \Big) \le \frac{\alpha}{2} \sum_{t=\tau+1}^{T} \frac{a_t^2}{\sigma_t^2}.$$

4 Computational Optimal Transport and Wasserstein Distance

Optimal transport gives a geometric way to compare probability distributions. Instead of only comparing their densities pointwise, it asks how much mass must be moved, and how far, to transform one distribution into another Peyré and Cuturi [2020].

Let μ and ν be probability distributions on $\mathbb{R}^d$. A *coupling* of μ and ν is a joint distribution π over pairs (X, Y) such that $X \sim \mu$ and $Y \sim \nu$. The set of all such couplings is denoted by $\Pi(\mu, \nu)$ Peyré and Cuturi [2020].

Definition 6 (p-Wasserstein distance)**.** *For $p \geq 1$, the p-Wasserstein distance is*

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^p \, d\pi(x, y) \right)^{1/p}.$$

The version most relevant for this proof is W_∞, which represents maximum cost rather than average cost from W_p Peyré and Cuturi [2020].

Definition 7 (W_∞ distance)**.** *The W_∞ distance between μ and ν is*

$$W_\infty(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} {}_{(X,Y)\sim\pi} \|X - Y\|.$$

Thus $W_\infty(\mu, \nu) \leq z$ means that μ and ν can be coupled so that the two random variables are always within distance z.

Computational optimal transport studies how such distances and couplings can be computed or approximated. For discrete distributions, the optimal transport problem becomes a linear program over a transport matrix Peyré and Cuturi [2020]. If

$$\mu = \sum_{i=1}^{m} a_i \delta_{x_i}, \qquad \nu = \sum_{j=1}^{r} b_j \delta_{y_j},$$

then a coupling is a nonnegative matrix $P \in \mathbb{R}_+^{m \times r}$ with row sums a_i and column sums b_j. The transport cost is

$$\sum_{i,j} P_{ij} c(x_i, y_j).$$

Modern computational optimal transport often uses entropic regularization, replacing the linear program by

$$\min_P \sum_{i,j} P_{ij} c(x_i, y_j) + \varepsilon \sum_{i,j} P_{ij} (\log P_{ij} - 1),$$

which can be solved efficiently by Sinkhorn iterations Peyré and Cuturi [2020].

In the privacy proof, optimal transport enters through the idea of allowing a controlled geometric shift before measuring Rényi divergence. This is captured by shifted Rényi divergence, which is used in privacy amplification by iteration arguments for noisy contractive processes Altschuler and Talwar [2023].

Definition 8 (Shifted Rényi divergence)**.** *For $z \geq 0$, define*

$$D_\alpha^{(z)}(\mu \| \nu) = \inf_{\tilde{\mu}:\, W_\infty(\tilde{\mu}, \mu) \leq z} D_\alpha(\tilde{\mu} \| \nu).$$

The interpretation is direct: before comparing μ and ν in Rényi divergence, we are allowed to move μ by at most z in W_∞. This is exactly the right object for noisy contractive processes. Contraction reduces distances between states, and Gaussian noise converts a remaining geometric displacement into a Rényi-divergence cost Mironov [2017], van Erven and Harremoes [2014], Altschuler and Talwar [2023]. The following lemma states the same idea in a mathematically precise way.

Lemma 6 (Shift-reduction lemma)**.** *Let μ and ν be probability distributions on $\mathbb{R}^d$. For any $a \geq 0$,*

$$\mathcal{D}_\alpha^{(z)} \left(\mu * \mathcal{N}(0, \sigma^2 I_d) \,\middle\|\, \nu * \mathcal{N}(0, \sigma^2 I_d) \right) \leq \mathcal{D}_\alpha^{(z+a)}(\mu \| \nu) + \frac{\alpha a^2}{2\sigma^2}.$$

Lemma 7 (Contraction-reduction lemma for random contractions)**.** *Suppose ϕ, ϕ' are random functions from $\mathbb{R}^d$ to $\mathbb{R}^d$ such that: (i) each is a contraction almost surely; (ii) there exists a coupling of (ϕ, ϕ') under which $\sup_x \|\phi(x) - \phi'(x)\| \le s$ almost surely.*

Then for any probability distributions μ and μ' on $\mathbb{R}^d$,

$$\mathcal{D}_\alpha^{(z+s)}\left(\phi_\# \mu \,\middle\|\, \phi'_\# \mu'\right) \le \mathcal{D}_\alpha^{(z)}(\mu \| \mu').$$

5 Main Theorem and Proof

We now prove the privacy saturation bound for Noisy-SGD. Throughout this section, we use the setup and notation from Section 2.

Theorem 1 (Privacy saturation of Noisy-SGD)**.** *Under the assumptions in Section 2, suppose that $\eta \le 2/M$. Then, under the standard parameter regime for the sampled Gaussian mechanism, Noisy-SGD satisfies*

$$\mathcal{D}_\alpha(P_{W_T} \| P_{W'_T}) \lesssim \frac{\alpha L^2}{n^2 \sigma^2} \min\left\{T, \frac{Dn}{L\eta}\right\}.$$

Equivalently, by Definition 4, Noisy-SGD satisfies (α, ε)-RDP with

$$\varepsilon \lesssim \frac{\alpha L^2}{n^2 \sigma^2} \min\left\{T, \frac{Dn}{L\eta}\right\}.$$

Proof. Fix adjacent datasets X, X' differing at index $i^\star$. Let $\{W_t\}_{t=0}^T$ and $\{W'_t\}_{t=0}^T$ be the corresponding Noisy-SGD iterates, started from the same initialization $w_0 \in K$. We couple the two runs using the same minibatches $B_0, \dots, B_{T-1}$.

Split the Gaussian noise into two independent parts:

$$\sigma^2 = \sigma_1^2 + \sigma_2^2,$$

and write the corresponding noise added to the iterate as

$$Y_t \sim \mathcal{N}(0, \eta^2 \sigma_1^2 I_d), \qquad Z_t \sim \mathcal{N}(0, \eta^2 \sigma_2^2 I_d).$$

The Z_t-noise will account for the sampled Gaussian privacy cost, while the Y_t-noise will be used for the iteration-forgetting argument.

With this decomposition, the two updates can be written as

$$W_{t+1} = \Pi_K\left[W_t - \frac{\eta}{b}\sum_{i \in B_t} \nabla f_i(W_t) + Y_t + Z_t\right], \tag{3}$$

$$W'_{t+1} = \Pi_K\left[W'_t - \frac{\eta}{b}\sum_{i \in B_t} \nabla f_i(W'_t) + Y_t + Z'_t\right], \tag{4}$$

where $\widetilde{Z}_t \sim \mathcal{N}(0, \eta^2 \sigma_2^2 I_d)$, and

$$Z'_t = \widetilde{Z}_t + \frac{\eta}{b}\big(\nabla f'_{i^\star}(W'_t) - \nabla f_{i^\star}(W'_t)\big)\mathbf{1}_{\{i^\star \in B_t\}}.$$

Thus Z'_t differs from a centered Gaussian only when the changed example appears in the minibatch.

Fix $\tau \in \{0, \dots, T-1\}$, and let

$$R := T - \tau$$

be the number of final iterations to be analyzed.

By the post-processing property, Lemma 2,

$$\mathcal{D}_\alpha(P_{W_T} \| P_{W'_T}) \le \mathcal{D}_\alpha(P_{W_T, Z_{\tau:T-1}} \| P_{W'_T, Z'_{\tau:T-1}}). \tag{5}$$

Applying the conditional composition lemma, Lemma 3, to the pair

$$(Z_{\tau:T-1}, W_T) \qquad \text{and} \qquad (Z'_{\tau:T-1}, W'_T)$$

gives

$$\mathcal{D}_\alpha(P_{W_T}\|P_{W'_T}) \le \underbrace{\mathcal{D}_\alpha(P_{Z_{\tau:T-1}}\|P_{Z'_{\tau:T-1}})}_{\text{sampling term}} + \underbrace{\sup_z \mathcal{D}_\alpha\Big(P_{W_T \mid Z_{\tau:T-1}=z}\Big\|P_{W'_T \mid Z'_{\tau:T-1}=z}\Big)}_{\text{iteration term}}. \tag{6}$$

We bound the two terms separately.

Sampling term. By another application of Lemma 3,

$$\mathcal{D}_\alpha(P_{Z_{\tau:T-1}}\|P_{Z'_{\tau:T-1}}) \le \sum_{t=\tau}^{T-1} \sup_{z_{\tau:t-1}} \mathcal{D}_\alpha\Big(P_{Z_t \mid Z_{\tau:t-1}=z_{\tau:t-1}}\Big\|P_{Z'_t \mid Z'_{\tau:t-1}=z_{\tau:t-1}}\Big). \tag{7}$$

For each t, $Z_t \sim \mathcal{N}(0, \eta^2\sigma_2^2 I_d)$, while Z'_t is a sampled shifted Gaussian. The shift is

$$m_t = \frac{\eta}{b}\big(\nabla f'_{i^\star}(W'_t) - \nabla f_{i^\star}(W'_t)\big),$$

and by L-Lipschitz property,

$$\|m_t\| \le \frac{\eta}{b}\left(\|\nabla f'_{i^\star}(W'_t)\| + \|\nabla f_{i^\star}(W'_t)\|\right) \le \frac{2\eta L}{b}.$$

The changed example is selected with probability b/n. Hence, by the sampled Gaussian bound, Lemma 5,

$$\mathcal{D}_\alpha\Big(P_{Z_t \mid Z_{\tau:t-1}=z_{\tau:t-1}}\Big\|P_{Z'_t \mid Z'_{\tau:t-1}=z_{\tau:t-1}}\Big) \le S_\alpha\left(\frac{b}{n}, \frac{b\sigma_2}{2L}\right) \lesssim \frac{\alpha L^2}{n^2\sigma_2^2}.$$

Therefore,

$$\mathcal{D}_\alpha(P_{Z_{\tau:T-1}}\|P_{Z'_{\tau:T-1}}) \lesssim \frac{\alpha L^2}{n^2\sigma_2^2} R. \tag{8}$$

Iteration term. Fix $z = (z_\tau, \dots, z_{T-1})$ and condition on

$$Z_{\tau:T-1} = Z'_{\tau:T-1} = z.$$

Define

$$\phi_t(\omega) = \omega - \frac{\eta}{b}\sum_{i\in B_t} \nabla f_i(\omega) + z_t.$$

Then, conditionally on z, the two chains evolve as

$$W_{t+1} = \Pi_K[\phi_t(W_t) + Y_t], \qquad W'_{t+1} = \Pi_K[\phi_t(W'_t) + Y_t].$$

By Lemma 1, the gradient step is contractive because $\eta \le 2/M$. Adding z_t does not affect distances, and projection onto K is non-expansive. Hence the conditional dynamics is a projected noisy contractive process.

Since both W_τ and W'_τ lie in K, and $\operatorname{diam}(K) = D$, the initial discrepancy at time τ is at most D. Applying the diameter-aware PABI bound, Lemma 1, with

$$a_t = \frac{D}{R}, \qquad t = \tau+1, \dots, T,$$

gives

$$\mathcal{D}_\alpha\Big(P_{W_T \mid Z_{\tau:T-1}=z}\Big\|P_{W'_T \mid Z'_{\tau:T-1}=z}\Big) \le \frac{\alpha}{2}\sum_{t=\tau+1}^{T} \frac{a_t^2}{\eta^2\sigma_1^2} = \frac{\alpha D^2}{2\eta^2\sigma_1^2 R}. \tag{9}$$

The bound is uniform in z, so it also bounds the supremum in (6).

Combining (6), (8), and (9),

$$\mathcal{D}_\alpha(P_{W_T}\|P_{W'_T}) \lesssim \frac{\alpha L^2}{n^2\sigma_2^2}R + \frac{\alpha D^2}{\eta^2\sigma_1^2 R}. \tag{10}$$

It remains to choose R. Write the right-hand side as

$$F(R) = AR + \frac{B}{R},$$

where

$$A = \frac{\alpha L^2}{n^2\sigma_2^2}, \qquad B = \frac{\alpha D^2}{\eta^2\sigma_1^2}.$$

Treating R as continuous,

$$F'(R) = A - \frac{B}{R^2}.$$

Setting $F'(R) = 0$ gives

$$R^2 = \frac{B}{A} = \frac{D^2 n^2 \sigma_2^2}{\eta^2 L^2 \sigma_1^2}.$$

Thus

$$R^\star = \frac{Dn}{L\eta} \cdot \frac{\sigma_2}{\sigma_1}.$$

By setting $\sigma_1^2 = \sigma_2^2 = \frac{\sigma^2}{2}$ gives $R^\star = \frac{Dn}{L\eta}$. Define

$$\bar{T} := \left\lceil \frac{Dn}{L\eta} \right\rceil.$$

If $T \le \bar{T}$, we use ordinary sampled Gaussian composition over all T iterations:

$$\mathcal{D}_\alpha(P_{W_T}\|P_{W'_T}) \lesssim \frac{\alpha L^2}{n^2\sigma^2}T.$$

If $T > \bar{T}$, choose $\tau = T - \bar{T}$, so that $R = \bar{T}$. Then (10) gives

$$\mathcal{D}_\alpha(P_{W_T}\|P_{W'_T}) \lesssim \frac{\alpha L^2}{n^2\sigma^2}\bar{T} + \frac{\alpha D^2}{\eta^2\sigma^2\bar{T}}.$$

Since

$$\bar{T} \asymp \frac{Dn}{L\eta},$$

the two terms are of the same order, and hence

$$\mathcal{D}_\alpha(P_{W_T}\|P_{W'_T}) \lesssim \frac{\alpha L^2}{n^2\sigma^2}\bar{T}.$$

Combining the two regimes,

$$\mathcal{D}_\alpha(P_{W_T}\|P_{W'_T}) \lesssim \frac{\alpha L^2}{n^2\sigma^2}\min\{T, \bar{T}\}.$$

Substituting the definition of $\bar{T}$ proves

$$\mathcal{D}_\alpha(P_{W_T}\|P_{W'_T}) \lesssim \frac{\alpha L^2}{n^2\sigma^2}\min\left\{T, \left\lceil \frac{Dn}{L\eta} \right\rceil\right\}.$$

□

6 Conclusion

The privacy behavior of Noisy-SGD is not captured correctly by simply composing all iterations. In the smooth convex bounded setting, the final iterate remembers less about early data-dependent perturbations because the algorithm repeatedly applies contractive maps and adds Gaussian noise. The proof formalizes this using a combination of privacy amplification by sampling and privacy amplification by iteration.

The main technical insight is to restart the privacy analysis from the last R iterations. Sampling makes the cost grow like R, while iteration-based forgetting makes the cost decay like $1/R$. Balancing these terms gives

$$R \asymp \frac{Dn}{L\eta},$$

which is exactly the burn-in time after which privacy loss stops increasing. Wasserstein distance and shifted Rényi divergence provide the right language for this argument because they separate geometric displacement from distributional distinguishability. This is why optimal transport is not merely background material here, it is part of the proof mechanism.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16, page 308–318. ACM, October 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

Jason M. Altschuler and Kunal Talwar. Privacy of noisy stochastic gradient descent: More iterations without more privacy loss. *arXiv preprint arXiv:2205.13710*, 2023.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. Jax: composable transformations of python+numpy programs. https://github.com/jax-ml/jax, 2018.

Cynthia Dwork and Vitaly Feldman. Privacy-preserving prediction. *CoRR*, abs/1803.10266, 2018. URL http://arxiv.org/abs/1803.10266.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284, 2006. URL https://journalprivacyconfidentiality.org/index.php/jpc/article/download/405/388/.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, page 263–275. IEEE, August 2017. doi: 10.1109/csf.2017.11. URL http://dx.doi.org/10.1109/CSF.2017.11.

Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020. URL https://arxiv.org/abs/1803.00567.

TensorFlow Authors. Tensorflow privacy. https://github.com/tensorflow/privacy, 2019.

Tim van Erven and Peter Harremoes. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, July 2014. ISSN 1557-9654. doi: 10.1109/tit.2014.2320500. URL http://dx.doi.org/10.1109/TIT.2014.2320500.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled renyi differential privacy and analytical moments accountant. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1226–1235. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/wang19b.html.

Ashkan Yousefpour et al. Opacus: User-friendly differential privacy library in pytorch. https://opacus.ai/, 2021.

A Disclosure of LLM Usage

We used an LLM only for permitted support tasks such as understanding background material, locating related work, and clarifying technical concepts. All factual claims, references, equations, and proof steps used in the report were independently verified from the original papers and standard scholarly sources.

B Introduction to Optimal Transport

Given a cost function $\mathbf{C}_{m\times n}$, where $m = n$, the optimal assignment problem seeks for a bijection σ in the set $Perm(n)$ of permutations of n elements solving

$$min\frac{1}{n}\sum_{i=1}^{n}\mathbf{C}_{i,\sigma(i)} \tag{11}$$

$$\sigma \in Perm(n) \tag{12}$$

B.1 Monge Formulation

We define the Monge problem between discrete measures. Consider $\alpha = \sum_{i=1}^{n}\mathbf{a}_i\delta_{x_i}$ and $\beta = \sum_{j=1}^{m}\mathbf{b}_j\delta_{y_j}$, where $\mathbf{a}$ denotes the source mass distribution and $\mathbf{b}$ represents the destination mass distribution. δ_x is the Dirac delta at x.

The Monge problem seeks a map that associates each point x_j a single point y_j to which the mass of α should be pushed towards the mass of β. In other words, we need to find a map $T : \{x_1, \cdots, x_n\} \to \{y_1, \cdots y_m\}$ such that

$$\forall j \in \{1, \cdots m\}, \mathbf{b}_j = \sum_{i:T(x_i)=y_j}\mathbf{a}_i \tag{13}$$

Which can be written in compact form as $T_{\#}\alpha = \beta$. Because all elements of $\mathbf{b}$ are positive, that map is necessarily surjective. The objective is to minimize a transportation cost parametrized by a cost function $c(x, y)$ for points $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

$$min_T\{\sum_i c(x_i, T(x_i)) : T_{\#}\alpha = \beta\} \tag{14}$$

B.2 Kantarovich Formulation

The key idea of Kantarovich is to relax the deterministic nature of transportation, namely that a source point x_i could only be moved to a single point $y_{\sigma(i)}$. Instead, the mass at x_i could potentially be dispatched across multiple locations. This leads to a formulation with probabilistic transport, encoded with a coupling matrix $\mathbf{P} \in \mathbb{R}_+^{n\times m}$, where $\mathbf{P}_{i,j}$ describes the amount of mass flowing from bin i to bin j. The characterization for a coupling to be admissible is as follows -

$$\mathbf{U}(\mathbf{a}, \mathbf{b}) = \{\mathbf{P} \in \mathbb{R}_+^{n\times m} : \mathbf{P}\mathbb{1}_m = \mathbf{a} \text{ and } \mathbf{P}^T\mathbb{1}_n = \mathbf{b}\} \tag{15}$$

where $\mathbf{P}\mathbb{1}_m = (\sum_j \mathbf{P}_{i,j})_i \in \mathbb{R}^n$.

B.3 Relation to the Wasserstein Metric

For continuous mass distribution, the conditions written in $\mathbf{U}(\mathbf{a}, \mathbf{b})$ can be re-written in using a coupling function $\gamma(x, y)$ satisfying the following properties :

1. $\int \gamma(x, y)dy = \mathbf{a}(x)$

2. $\int \gamma(x,y)dx = \mathbf{b}(y)$

This is equivalent to the requirement that γ be a joing PDF with marginals $\mathbf{a}$ and $\mathbf{b}$. The cost of the optimal plan, therefore, is

$$\int\int c(x,y)\gamma(x,y)dxdy = \int c(x,y)d\gamma(x,y) \tag{16}$$

The optimal cost is

$$C = inf_{\gamma\in\Gamma(a,b)} \int c(x,y)d\gamma(x,y) \tag{17}$$

For a distance metric $\mathbf{D}$ which satisfies the following three properties -

- $\mathbf{D} \in \mathbb{R}_+^{n\times n}$ is symmetric
- $\mathbf{D}_{i,j} = 0$ if and only if $i = j$.
- $\forall (i,j,k) \in \{1,\cdots n\}^3, \mathbf{D}_{i,k} \leq \mathbf{D}_{i,j} + \mathbf{D}_{j,k}$

we denote the $p-$ Wasserstein distance on Σ_n as

$$W_p(\mathbf{a},\mathbf{b}) = L_{\mathbf{D}^p}(\mathbf{a},\mathbf{b})^{\frac{1}{p}} \tag{18}$$

C Proofs for Intermediate Lemmas

proof of lemma 1 contraction. Let

$$T(u) := u - \eta\nabla f(u).$$

We want to show that

$$\|T(u) - T(v)\| \leq \|u - v\|.$$

Expanding the squared distance gives

$$\begin{aligned}\|T(u) - T(v)\|^2 &= \left\|u - v - \eta(\nabla f(u) - \nabla f(v))\right\|^2 \\ &= \|u-v\|^2 - 2\eta\langle u - v, \nabla f(u) - \nabla f(v)\rangle + \eta^2\|\nabla f(u) - \nabla f(v)\|^2. \end{aligned} \tag{19}$$

Since f is convex and M-smooth, its gradient is $1/M$-cocoercive:

$$\langle u - v, \nabla f(u) - \nabla f(v)\rangle \geq \frac{1}{M}\|\nabla f(u) - \nabla f(v)\|^2.$$

Substituting this into (19), we get

$$\begin{aligned}\|T(u) - T(v)\|^2 &\leq \|u-v\|^2 - \frac{2\eta}{M}\|\nabla f(u) - \nabla f(v)\|^2 + \eta^2\|\nabla f(u) - \nabla f(v)\|^2 \\ &= \|u-v\|^2 + \left(\eta^2 - \frac{2\eta}{M}\right)\|\nabla f(u) - \nabla f(v)\|^2. \end{aligned} \tag{20}$$

Since $\eta \leq 2/M$, we have

$$\eta^2 - \frac{2\eta}{M} = \eta\left(\eta - \frac{2}{M}\right) \leq 0.$$

Therefore,

$$\|T(u) - T(v)\|^2 \leq \|u - v\|^2.$$

Taking square roots gives

$$\|T(u) - T(v)\| \leq \|u - v\|.$$

□

proof of lemma 2 Post-processing for RDP. Let $Y = h(X)$, and let μ_Y, ν_Y be the distributions of Y when $X \sim \mu, \nu$, respectively. We prove that

$$\mathcal{D}_\alpha(\mu_Y \| \nu_Y) \le \mathcal{D}_\alpha(\mu \| \nu).$$

For $\alpha > 1$,

$$e^{(\alpha-1)D_\alpha(\mu_Y\|\nu_Y)} = \sum_y \nu_Y(y) \left(\frac{\mu_Y(y)}{\nu_Y(y)}\right)^\alpha = \sum_y \frac{\mu_Y(y)^\alpha}{\nu_Y(y)^{\alpha-1}}.$$

Since

$$\mu_Y(y) = \sum_{x:\, h(x)=y} \mu(x), \qquad \nu_Y(y) = \sum_{x:\, h(x)=y} \nu(x),$$

we get

$$e^{(\alpha-1)D_\alpha(\mu_Y\|\nu_Y)} = \sum_y \frac{\left(\sum_{x:\, h(x)=y} \mu(x)\right)^\alpha}{\left(\sum_{x:\, h(x)=y} \nu(x)\right)^{\alpha-1}}.$$

Now for fix y. we can write $A_y := \sum_{x:\, h(x)=y} \nu(x)$. Then

$$\frac{\left(\sum_{x:\, h(x)=y} \mu(x)\right)^\alpha}{\left(\sum_{x:\, h(x)=y} \nu(x)\right)^{\alpha-1}} = A_y \left(\sum_{x:\, h(x)=y} \frac{\mu(x)}{\nu(x)} \cdot \frac{\nu(x)}{A_y}\right)^\alpha.$$

Since $t \mapsto t^\alpha$ is convex, Jensen's inequality gives

$$\begin{aligned} A_y \left(\sum_{x:\, h(x)=y} \frac{\mu(x)}{\nu(x)} \cdot \frac{\nu(x)}{A_y}\right)^\alpha &\le A_y \sum_{x:\, h(x)=y} \left(\frac{\mu(x)}{\nu(x)}\right)^\alpha \frac{\nu(x)}{A_y} \\ &= \sum_{x:\, h(x)=y} \frac{\mu(x)^\alpha}{\nu(x)^{\alpha-1}}. \end{aligned}$$

Therefore, for every y,

$$\frac{\left(\sum_{x:\, h(x)=y} \mu(x)\right)^\alpha}{\left(\sum_{x:\, h(x)=y} \nu(x)\right)^{\alpha-1}} \le \sum_{x:\, h(x)=y} \frac{\mu(x)^\alpha}{\nu(x)^{\alpha-1}}.$$

Summing over all y,

$$\begin{aligned} e^{(\alpha-1)D_\alpha(\mu_Y\|\nu_Y)} &\le \sum_y \sum_{x:\, h(x)=y} \frac{\mu(x)^\alpha}{\nu(x)^{\alpha-1}} \\ &= \sum_x \frac{\mu(x)^\alpha}{\nu(x)^{\alpha-1}} \\ &= e^{(\alpha-1)D_\alpha(\mu\|\nu)}. \end{aligned}$$

Taking logarithms gives

$$\mathcal{D}_\alpha(\mu_Y \| \nu_Y) \le \mathcal{D}_\alpha(\mu \| \nu).$$

□

Lemma 8 (Joint quasi-convexity of Rényi divergence)**.** *For any Rényi parameter $\alpha \ge 1$, any mixing probability $\lambda \in [0, 1]$, and any two pairs of probability distributions μ, ν and μ', ν',*

$$\mathcal{D}_\alpha\big(\lambda\mu + (1-\lambda)\mu' \,\big\|\, \lambda\nu + (1-\lambda)\nu'\big) \le \max\left\{\mathcal{D}_\alpha(\mu\|\nu), \mathcal{D}_\alpha(\mu'\|\nu')\right\}.$$

proof of lemma 8 Quasi Convexity.

$$M := \max\{\mathcal{D}_\alpha(\mu\|\nu),\, \mathcal{D}_\alpha(\mu'\|\nu')\}.$$

$$\mu_\lambda = \lambda\mu + (1-\lambda)\mu', \qquad \nu_\lambda = \lambda\nu + (1-\lambda)\nu'.$$

Then, for each x,

$$\frac{\mu_\lambda(x)}{\nu_\lambda(x)} = \frac{\lambda\mu(x) + (1-\lambda)\mu'(x)}{\lambda\nu(x) + (1-\lambda)\nu'(x)} = \frac{\lambda\nu(x)\frac{\mu(x)}{\nu(x)} + (1-\lambda)\nu'(x)\frac{\mu'(x)}{\nu'(x)}}{\lambda\nu(x) + (1-\lambda)\nu'(x)}.$$

Since $t \mapsto t^\alpha$ is convex for $\alpha \geq 1$, Jensen's inequality yields

$$\left(\frac{\mu_\lambda(x)}{\nu_\lambda(x)}\right)^\alpha \leq \frac{\lambda\nu(x)\left(\frac{\mu(x)}{\nu(x)}\right)^\alpha + (1-\lambda)\nu'(x)\left(\frac{\mu'(x)}{\nu'(x)}\right)^\alpha}{\lambda\nu(x) + (1-\lambda)\nu'(x)}.$$

Multiplying by $\nu_\lambda(x) = \lambda\nu(x) + (1-\lambda)\nu'(x)$ and integrating, we obtain

$$\begin{aligned}
\mathbb{E}_{X\sim\nu_\lambda}\left[\left(\frac{\mu_\lambda(X)}{\nu_\lambda(X)}\right)^\alpha\right] &\leq \lambda\,\mathbb{E}_{X\sim\nu}\left[\left(\frac{\mu(X)}{\nu(X)}\right)^\alpha\right] + (1-\lambda)\,\mathbb{E}_{X\sim\nu'}\left[\left(\frac{\mu'(X)}{\nu'(X)}\right)^\alpha\right] \\
&\leq \lambda e^{(\alpha-1)M} + (1-\lambda)e^{(\alpha-1)M} \\
&= e^{(\alpha-1)M}.
\end{aligned}$$

Therefore,

$$\mathcal{D}_\alpha(\mu_\lambda\|\nu_\lambda) \leq M,$$

□

proof of Lemma 3 strong composition. Let

$$P := P_{X_{1:k}}, \qquad Q := P_{Y_{1:k}}.$$

Write the joint densities in conditional form:

$$p(x_{1:k}) = \prod_{i=1}^k p_i(x_i \mid x_{<i}), \qquad q(x_{1:k}) = \prod_{i=1}^k q_i(x_i \mid x_{<i}),$$

where $x_{<i} = (x_1, \ldots, x_{i-1})$.

Starting from the definition of Rényi divergence,

$$\begin{aligned}
e^{(\alpha-1)D_\alpha(P\|Q)} &= \int \left(\frac{p(x_{1:k})}{q(x_{1:k})}\right)^\alpha q(x_{1:k})\,dx_{1:k} \\
&= \int \prod_{i=1}^k \left(\frac{p_i(x_i \mid x_{<i})}{q_i(x_i \mid x_{<i})}\right)^\alpha q_i(x_i \mid x_{<i})\,dx_{1:k}. \qquad (21)
\end{aligned}$$

Now define, for each i,

$$M_i := \sup_{x_{<i}} \int \left(\frac{p_i(x_i \mid x_{<i})}{q_i(x_i \mid x_{<i})}\right)^\alpha q_i(x_i \mid x_{<i})\,dx_i.$$

The key point is that M_i is exactly the exponential form of the worst-case conditional Rényi divergence:

$$M_i = \exp\left((\alpha-1)\sup_{x_{<i}} \mathcal{D}_\alpha\big(P_{X_i|X_{<i}=x_{<i}} \,\|\, P_{Y_i|Y_{<i}=x_{<i}}\big)\right).$$

From (21), we now integrate one variable at a time. At each step, the inner integral is bounded by the corresponding M_i. Therefore

$$e^{(\alpha-1)\mathcal{D}_\alpha(P\|Q)} \leq \prod_{i=1}^k M_i. \qquad (22)$$

Substituting the expression for M_i into (22), we get

$$e^{(\alpha-1)D_\alpha(P\|Q)} \le \prod_{i=1}^{k} \exp\left((\alpha-1)\sup_{x_{<i}} D_\alpha\big(P_{X_i|X_{<i}=x_{<i}} \parallel P_{Y_i|Y_{<i}=x_{<i}}\big)\right)$$
$$= \exp\left((\alpha-1)\sum_{i=1}^{k}\sup_{x_{<i}} \mathcal{D}_\alpha\big(P_{X_i|X_{<i}=x_{<i}} \parallel P_{Y_i|Y_{<i}=x_{<i}}\big)\right).$$

Taking logarithms and dividing by $\alpha-1$ proves

$$\mathcal{D}_\alpha(P_{X_{1:k}}\|P_{Y_{1:k}}) \le \sum_{i=1}^{k}\sup_{x_{<i}} \mathcal{D}_\alpha\big(P_{X_i|X_{<i}=x_{<i}} \parallel P_{Y_i|Y_{<i}=x_{<i}}\big).$$

□

proof of lemma 4.

$$\mathbb{B}_d(R) := \{x \in \mathbb{R}^d : \|x\| \le R\},$$

and $\mathcal{P}(\mathbb{B}_d(R))$ denotes the set of probability distributions supported on $\mathbb{B}_d(R)$, i.e.,

$$\mu \in \mathcal{P}(\mathbb{B}_d(R)) \quad \iff \quad \mu(\mathbb{B}_d(R)) = 1.$$

Let $Q_\mu := (1-q)\mathcal{N}(0,\sigma^2 I_d) + q\big(\mathcal{N}(0,\sigma^2 I_d) * \mu\big)$. If $Z \sim \mu$, then

$$\mathcal{N}(0,\sigma^2 I_d) * \mu$$

is simply the distribution of

$$G + Z, \qquad G \sim \mathcal{N}(0,\sigma^2 I_d).$$

Equivalently, Q_μ can be viewed as the following mixture: with probability $1-q$, sample from $\mathcal{N}(0,\sigma^2 I_d)$, and with probability q, first sample $Z \sim \mu$ and then sample from $\mathcal{N}(Z,\sigma^2 I_d)$. Hence

$$Q_\mu = \mathbb{E}_{Z\sim\mu}\Big[(1-q)\mathcal{N}(0,\sigma^2 I_d) + q\mathcal{N}(Z,\sigma^2 I_d)\Big].$$

Therefore,

$$\mathcal{D}_\alpha\Big(\mathcal{N}(0,\sigma^2 I_d)\,\Big\|\,Q_\mu\Big) = \mathcal{D}_\alpha\left(\mathcal{N}(0,\sigma^2 I_d)\,\middle\|\,\mathbb{E}_{Z\sim\mu}\Big[(1-q)\mathcal{N}(0,\sigma^2 I_d) + q\mathcal{N}(Z,\sigma^2 I_d)\Big]\right)$$
$$\le \sup_{\|z\|\le R} \mathcal{D}_\alpha\Big(\mathcal{N}(0,\sigma^2 I_d)\,\Big\|\,(1-q)\mathcal{N}(0,\sigma^2 I_d) + q\mathcal{N}(z,\sigma^2 I_d)\Big), \qquad \text{(using lemma 8)}$$

Now the Gaussian distribution is rotation invariant, so the last expression depends only on the norm $r := \|z\|$. Hence

$$\sup_{\|z\|\le R} \mathcal{D}_\alpha\Big(\mathcal{N}(0,\sigma^2 I_d)\,\Big\|\,(1-q)\mathcal{N}(0,\sigma^2 I_d) + q\mathcal{N}(z,\sigma^2 I_d)\Big)$$
$$= \sup_{0\le r\le R} \mathcal{D}_\alpha\Big(\mathcal{N}(0,\sigma^2)\,\Big\|\,(1-q)\mathcal{N}(0,\sigma^2) + q\mathcal{N}(r,\sigma^2)\Big).$$

Since the divergence increases with the shift size r, the supremum is attained at $r = R$. Thus

$$\sup_{0\le r\le R} \mathcal{D}_\alpha\Big(\mathcal{N}(0,\sigma^2)\,\Big\|\,(1-q)\mathcal{N}(0,\sigma^2) + q\mathcal{N}(r,\sigma^2)\Big)$$
$$= \mathcal{D}_\alpha\Big(\mathcal{N}(0,\sigma^2)\,\Big\|\,(1-q)\mathcal{N}(0,\sigma^2) + q\mathcal{N}(R,\sigma^2)\Big).$$

Finally, rescale by $x \mapsto x/R$. Then

$$\mathcal{N}(0,\sigma^2) \mapsto \mathcal{N}\left(0,\left(\frac{\sigma}{R}\right)^2\right), \qquad \mathcal{N}(R,\sigma^2) \mapsto \mathcal{N}\left(1,\left(\frac{\sigma}{R}\right)^2\right).$$

Therefore,

$$\begin{aligned}
&\mathcal{D}_\alpha\Big(\mathcal{N}(0,\sigma^2)\,\Big\|\,(1-q)\mathcal{N}(0,\sigma^2)+q\mathcal{N}(R,\sigma^2)\Big)\\
&=\mathcal{D}_\alpha\left(\mathcal{N}\left(0,\left(\frac{\sigma}{R}\right)^2\right)\,\middle\|\,(1-q)\mathcal{N}\left(0,\left(\frac{\sigma}{R}\right)^2\right)+q\mathcal{N}\left(1,\left(\frac{\sigma}{R}\right)^2\right)\right)\\
&=S_\alpha(q,\sigma/R).
\end{aligned}$$

Hence

$$\sup_{\mu\in\mathcal{P}(RB_d)}\mathcal{D}_\alpha\Big(\mathcal{N}(0,\sigma^2I_d)\,\Big\|\,(1-q)\mathcal{N}(0,\sigma^2I_d)+q\big(\mathcal{N}(0,\sigma^2I_d)*\mu\big)\Big)=S_\alpha(q,\sigma/R).$$

□

proof of lemma 5. By definition, $S_\alpha(q,\sigma)=D_\alpha\big(\mathcal{N}(0,\sigma^2)\,\|\,(1-q)\mathcal{N}(0,\sigma^2)+q\mathcal{N}(1,\sigma^2)\big)$. Let

$$P=\mathcal{N}(0,\sigma^2),\qquad Q=(1-q)\mathcal{N}(0,\sigma^2)+q\mathcal{N}(1,\sigma^2),$$

and let p, q_σ denote their corresponding densities. Then

$$q_\sigma(x)=(1-q)p(x)+q\,p_1(x),$$

where p_1 is the density of $\mathcal{N}(1,\sigma^2)$. Hence

$$q_\sigma(x)=p(x)\left[(1-q)+q\frac{p_1(x)}{p(x)}\right].$$

For Gaussian densities, $\frac{p_1(x)}{p(x)}=\exp\big(\frac{x}{\sigma^2}-\frac{1}{2\sigma^2}\big)$. Therefore

$$q_\sigma(x)=p(x)\left[(1-q)+q\exp\left(\frac{x}{\sigma^2}-\frac{1}{2\sigma^2}\right)\right].$$

Now use the definition of Rényi divergence:

$$\begin{aligned}
S_\alpha(q,\sigma)&=\frac{1}{\alpha-1}\log\int\left(\frac{p(x)}{q_\sigma(x)}\right)^{\alpha-1}p(x)\,dx\\
&=\frac{1}{\alpha-1}\log\mathbb{E}_{X\sim\mathcal{N}(0,\sigma^2)}\left[\left((1-q)+q\exp\left(\frac{X}{\sigma^2}-\frac{1}{2\sigma^2}\right)\right)^{1-\alpha}\right].
\end{aligned}$$

Define $L(X):=\exp\Big(\frac{X}{\sigma^2}-\frac{1}{2\sigma^2}\Big)-1$. Then and so

$$S_\alpha(q,\sigma)=\frac{1}{\alpha-1}\log\mathbb{E}\big[(1+qL(X))^{1-\alpha}\big].$$

Next,

$$\mathbb{E}[L(X)]=\mathbb{E}\left[\exp\left(\frac{X}{\sigma^2}-\frac{1}{2\sigma^2}\right)\right]-1=1-1=0.$$

Thus the linear term in q vanishes. Expanding $(1+qL)^{1-\alpha}$ to second order gives

$$(1+qL)^{1-\alpha}=1-(\alpha-1)qL+\frac{\alpha(\alpha-1)}{2}q^2L^2+O(q^3),$$

so taking expectation and using $\mathbb{E}[L]=0$,

$$\mathbb{E}[(1+qL)^{1-\alpha}]=1+\frac{\alpha(\alpha-1)}{2}q^2\mathbb{E}[L^2]+O(q^3).$$

Therefore,

$$S_\alpha(q,\sigma)\approx\frac{\alpha}{2}q^2\mathbb{E}[L^2].$$

Now compute

$$\begin{aligned}\mathbb{E}[L^2] &= \mathbb{E}\left[\left(\exp\left(\frac{X}{\sigma^2} - \frac{1}{2\sigma^2}\right) - 1\right)^2\right] \\ &= \mathbb{E}\left[\exp\left(\frac{2X}{\sigma^2} - \frac{1}{\sigma^2}\right)\right] - 2\mathbb{E}\left[\exp\left(\frac{X}{\sigma^2} - \frac{1}{2\sigma^2}\right)\right] + 1 \\ &= e^{1/\sigma^2} - 1.\end{aligned}$$

Hence

$$S_\alpha(q,\sigma) \approx \frac{\alpha}{2} q^2 \big(e^{1/\sigma^2} - 1\big).$$

Since $\sigma \geq 4$, we have $e^{1/\sigma^2} - 1 \lesssim \sigma^{-2}$, and thus

$$S_\alpha(q,\sigma) = O\left(\frac{\alpha q^2}{\sigma^2}\right).$$

The rigorous bound

$$S_\alpha(q,\sigma) \leq \frac{2\alpha q^2}{\sigma^2}$$

under the conditions $q \in (0, 1/5)$, $\sigma \geq 4$, and $\alpha \leq \alpha^*(q,\sigma)$ follows from the sharp upper bound for the sampled Gaussian mechanism, as stated in Wang et al. [2019]. □

proof of rdp to dp conversion formula in def 4. Let P, Q be the outputs on adjacent datasets, and suppose

$$D_\alpha(P\|Q) \leq \varepsilon_\alpha = \mathbb{E}_{z\sim P}\left[\left(\frac{P(z)}{Q(z)}\right)^{\alpha-1}\right] \leq e^{(\alpha-1)\varepsilon_\alpha}.$$

Fix any event S, and split it as

$$S = S_1 \cup S_2, \qquad \text{where} \qquad S_1 = \left\{z \in S : \frac{P(z)}{Q(z)} \leq e^\varepsilon\right\}, \qquad S_2 = \left\{z \in S : \frac{P(z)}{Q(z)} > e^\varepsilon\right\}.$$

Then

$$P(S) = P(S_1) + P(S_2).$$

Also,

$$P(S_1) \leq e^\varepsilon Q(S).$$

For the second term, Markov's inequality gives

$$\begin{aligned}P(S_2) &\leq \frac{\mathbb{E}_{z\sim P}\left[\left(\frac{P(z)}{Q(z)}\right)^{\alpha-1}\right]}{e^{(\alpha-1)\varepsilon}} \\ &\leq e^{(\alpha-1)(\varepsilon_\alpha - \varepsilon)}.\end{aligned}$$

Choosing

$$\varepsilon = \varepsilon_\alpha + \frac{\log(1/\delta)}{\alpha - 1}$$

makes the last bound equal to δ. □

proof of Lemma 6 shift reduction. Let

$$G := \mathcal{N}(0, \sigma^2 I_d).$$

Fix $\varepsilon > 0$. By the definition of shifted Rényi divergence, choose a distribution $\widetilde{\mu}$ such that

$$W_\infty(\mu, \widetilde{\mu}) \leq z + a$$

and

$$\mathcal{D}_\alpha(\widetilde{\mu}\|\nu) \le \mathcal{D}_\alpha^{(z+a)}(\mu\|\nu) + \varepsilon.$$

Since $W_\infty(\mu, \widetilde{\mu}) \le z + a$, we can couple

$$X \sim \mu, \qquad \widetilde{X} \sim \widetilde{\mu}$$

so that

$$\|X - \widetilde{X}\| \le z + a$$

almost surely.

We split this displacement into a z-part and an a-part. Define

$$\bar{X} = X + \frac{z}{z+a}(\widetilde{X} - X),$$

when $z + a > 0$. If $z + a = 0$, set $\bar{X} = X = \widetilde{X}$. Let

$$\rho := P_{\bar{X}}.$$

Then

$$\|X - \bar{X}\| \le z, \qquad \|\bar{X} - \widetilde{X}\| \le a.$$

Hence

$$W_\infty(\mu, \rho) \le z, \qquad W_\infty(\rho, \widetilde{\mu}) \le a.$$

Since $W_\infty(\mu, \rho) \le z$, convolving both distributions with the same Gaussian noise gives

$$W_\infty(\mu * G, \rho * G) \le z.$$

Therefore, by the definition of shifted Rényi divergence,

$$\mathcal{D}_\alpha^{(z)}\left(\mu * G \parallel \nu * G\right) \le \mathcal{D}_\alpha\left(\rho * G \parallel \nu * G\right). \tag{23}$$

It remains to bound the ordinary Rényi divergence on the right-hand side. We claim that

$$\mathcal{D}_\alpha\left(\rho * G \parallel \nu * G\right) \le \mathcal{D}_\alpha(\widetilde{\mu}\|\nu) + \frac{\alpha a^2}{2\sigma^2}. \tag{24}$$

To prove the claim, compare the two joint random variables

$$(\widetilde{X}, \bar{X} + G) \qquad \text{and} \qquad (Y, Y + G),$$

where

$$Y \sim \nu, \qquad G \sim \mathcal{N}(0, \sigma^2 I_d).$$

The second coordinates have distributions

$$\rho * G \qquad \text{and} \qquad \nu * G.$$

Thus, by post-processing,

$$\mathcal{D}_\alpha(\rho * G\|\nu * G) \le \mathcal{D}_\alpha\left(P_{\widetilde{X}, \bar{X}+G} \,\middle\|\, P_{Y, Y+G}\right).$$

We now bound the joint divergence using conditional composition. The first coordinate contributes

$$\mathcal{D}_\alpha(P_{\widetilde{X}}\|P_Y) = \mathcal{D}_\alpha(\widetilde{\mu}\|\nu).$$

For the second coordinate, condition on the first coordinate being x. Under the second joint distribution,

$$Y + G \mid Y = x \sim \mathcal{N}(x, \sigma^2 I_d).$$

Under the first joint distribution, since

$$\|\bar{X} - \widetilde{X}\| \le a,$$

the conditional distribution of $\bar{X} + G$ given $\widetilde{X} = x$ is a mixture of Gaussians

$$\mathcal{N}(u, \sigma^2 I_d)$$

with centers satisfying

$$\|u - x\| \le a.$$

By joint quasi-convexity of Rényi divergence, this mixture is bounded by the worst such Gaussian shift. Hence

$$\mathcal{D}_\alpha\left(P_{\bar{X}+G|\widetilde{X}=x} \,\middle\|\, \mathcal{N}(x, \sigma^2 I_d)\right) \le \sup_{\|u-x\|\le a} \mathcal{D}_\alpha\left(\mathcal{N}(u, \sigma^2 I_d) \,\middle\|\, \mathcal{N}(x, \sigma^2 I_d)\right).$$

For two Gaussians with the same covariance,

$$\mathcal{D}_\alpha\left(\mathcal{N}(u, \sigma^2 I_d) \,\middle\|\, \mathcal{N}(x, \sigma^2 I_d)\right) = \frac{\alpha\|u - x\|^2}{2\sigma^2}.$$

Therefore,

$$\mathcal{D}_\alpha\left(P_{\bar{X}+G|\widetilde{X}=x} \,\middle\|\, \mathcal{N}(x, \sigma^2 I_d)\right) \le \frac{\alpha a^2}{2\sigma^2}.$$

Conditional composition now gives

$$\mathcal{D}_\alpha\left(P_{\widetilde{X},\bar{X}+G} \,\middle\|\, P_{Y,Y+G}\right) \le \mathcal{D}_\alpha(\widetilde{\mu}\|\nu) + \frac{\alpha a^2}{2\sigma^2}.$$

This proves (24).

Combining (23) and (24), we obtain

$$\mathcal{D}_\alpha^{(z)}\left(\mu * G \parallel \nu * G\right) \le \mathcal{D}_\alpha(\widetilde{\mu}\|\nu) + \frac{\alpha a^2}{2\sigma^2}.$$

Using the choice of $\widetilde{\mu}$,

$$\mathcal{D}_\alpha^{(z)}\left(\mu * \gamma_\sigma \parallel \nu * \gamma_\sigma\right) \le \mathcal{D}_\alpha^{(z+a)}(\mu\|\nu) + \frac{\alpha a^2}{2\sigma^2} + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, letting $\varepsilon \downarrow 0$ gives

$$\mathcal{D}_\alpha^{(z)}\left(\mu * \mathcal{N}(0, \sigma^2 I_d) \,\middle\|\, \nu * \mathcal{N}(0, \sigma^2 I_d)\right) \le \mathcal{D}_\alpha^{(z+a)}(\mu\|\nu) + \frac{\alpha a^2}{2\sigma^2}.$$

□

proof of Lemma 7 *contraction reduction.* Fix $\varepsilon > 0$. By the definition of shifted Rényi divergence, choose a distribution $\widetilde{\mu}$ such that

$$W_\infty(\mu, \widetilde{\mu}) \le z$$

and

$$\mathcal{D}_\alpha(\widetilde{\mu}\|\mu') \le \mathcal{D}_\alpha^{(z)}(\mu\|\mu') + \varepsilon.$$

We will show that $\phi'_\#\widetilde{\mu}$ is a valid candidate in the definition of

$$\mathcal{D}_\alpha^{(z+s)}(\phi_\#\mu\|\phi'_\#\mu').$$

Since $W_\infty(\mu, \widetilde{\mu}) \le z$, there exists a coupling $(X, \widetilde{X})$ such that

$$X \sim \mu, \qquad \widetilde{X} \sim \widetilde{\mu}, \qquad \|X - \widetilde{X}\| \le z$$

almost surely. Also, by assumption, there exists a coupling of the random maps (ϕ, ϕ') such that

$$\sup_x \|\phi(x) - \phi'(x)\| \le s$$

almost surely.

Using these two couplings together, we have

$$\begin{aligned}\|\phi(X) - \phi'(\widetilde{X})\| &\le \|\phi(X) - \phi(\widetilde{X})\| + \|\phi(\widetilde{X}) - \phi'(\widetilde{X})\| \\ &\le \|X - \widetilde{X}\| + s \\ &\le z + s.\end{aligned}$$

The second inequality uses that ϕ is a contraction. Therefore,

$$W_\infty(\phi_\# \mu, \phi'_\# \widetilde{\mu}) \le z + s.$$

Hence, by the definition of shifted Rényi divergence,

$$\mathcal{D}_\alpha^{(z+s)}(\phi_\# \mu \| \phi'_\# \mu') \le \mathcal{D}_\alpha(\phi'_\# \widetilde{\mu} \| \phi'_\# \mu').$$

It remains to bound the right-hand side. The same random map ϕ' is applied to both $\widetilde{\mu}$ and μ'. Applying a deterministic map cannot increase Rényi divergence, and the same remains true for a random map by conditioning on the map and using joint quasi-convexity of Rényi divergence. Therefore,

$$\mathcal{D}_\alpha(\phi'_\# \widetilde{\mu} \| \phi'_\# \mu') \le \mathcal{D}_\alpha(\widetilde{\mu} \| \mu').$$

Using the choice of $\widetilde{\mu}$, we obtain

$$\mathcal{D}_\alpha^{(z+s)}(\phi_\# \mu \| \phi'_\# \mu') \le \mathcal{D}_\alpha^{(z)}(\mu \| \mu') + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, letting $\varepsilon \downarrow 0$ gives

$$\mathcal{D}_\alpha^{(z+s)}(\phi_\# \mu \| \phi'_\# \mu') \le \mathcal{D}_\alpha^{(z)}(\mu \| \mu').$$

□

Proposition 2 (Original PABI bound)**.** *Consider two contractive noisy iteration processes*

$$X_t = \Pi_K\big[\phi_t(X_{t-1}) + \xi_t\big], \qquad X'_t = \Pi_K\big[\phi'_t(X'_{t-1}) + \xi_t\big], \qquad t = 1, \dots, T,$$

where $K \subset \mathbb{R}^d$ is convex, each ϕ_t and ϕ'_t is contractive, and $\xi_t \sim \mathcal{N}(0, \sigma_t^2 I_d)$. For each t, define

$$s_t := \sup_{x \in \mathbb{R}^d} \|\phi_t(x) - \phi'_t(x)\|.$$

Let $a_1, \dots, a_T$ be any nonnegative sequence such that

$$z_t := \sum_{i=1}^{t} (s_i - a_i)$$

is nonnegative for every $t = 1, \dots, T$, and satisfies $z_T = 0$. Then

$$\mathcal{D}_\alpha(P_{X_T} \| P_{X'_T}) \le \frac{\alpha}{2} \sum_{t=1}^{T} \frac{a_t^2}{\sigma_t^2}.$$

proof lemma 2 original PABI. For $t = 0, 1, \dots, T$, define

$$\Delta_t := D_\alpha^{(z_t)}\big(P_{X_t} \,\|\, P_{X'_t}\big), \qquad z_t := \sum_{i=1}^{t} (s_i - a_i).$$

By assumption, $z_t \ge 0$ for all t, and $z_T = 0$. Hence

$$\begin{aligned} D_\alpha(P_{X_T} \| P_{X'_T}) &= D_\alpha^{(z_T)}(P_{X_T} \| P_{X'_T}) \\ &= \Delta_T. \qquad \big[\text{since } z_T = 0\big] \end{aligned} \tag{25}$$

We now derive a one-step recursion for Δ_t. Using the update rule of the projected CNI,

$$X_t = \Pi_K\big[\phi_t(X_{t-1}) + Z_t\big], \qquad X'_t = \Pi_K\big[\phi'_t(X'_{t-1}) + Z'_t\big].$$

Therefore

$$\begin{aligned} \Delta_t &= D_\alpha^{(z_t)}\Big(P_{\Pi_K[\phi_t(X_{t-1}) + Z_t]} \,\Big\|\, P_{\Pi_K[\phi'_t(X'_{t-1}) + Z'_t]}\Big) \\ &\le D_\alpha^{(z_t)}\Big(P_{\phi_t(X_{t-1}) + Z_t} \,\Big\|\, P_{\phi'_t(X'_{t-1}) + Z'_t}\Big) \qquad \big[\text{post-processing for } \Pi_K\big]. \end{aligned} \tag{26}$$

Now use the identity

$$z_t = z_{t-1} + s_t - a_t.$$

Applying the shift-reduction lemma with shift increment a_t, we obtain

$$\Delta_t \le D_\alpha^{(z_{t-1}+s_t)}\Big(P_{\phi_t(X_{t-1})} \,\Big\|\, P_{\phi'_t(X'_{t-1})}\Big) + \frac{\alpha a_t^2}{2\sigma_t^2} \qquad \big[\text{Lemma 6}\big]. \tag{27}$$

Next, by definition,

$$s_t = \sup_x \|\phi_t(x) - \phi'_t(x)\|,$$

and ϕ_t, ϕ'_t are contractions. Hence the contraction-reduction lemma gives

$$\begin{aligned} D_\alpha^{(z_{t-1}+s_t)}\Big(P_{\phi_t(X_{t-1})} \,\Big\|\, P_{\phi'_t(X'_{t-1})}\Big) &\le D_\alpha^{(z_{t-1})}(P_{X_{t-1}} \| P_{X'_{t-1}}) \\ &= \Delta_{t-1}. \qquad \big[\text{Lemma 7}\big] \end{aligned} \tag{28}$$

Combining (27) and (28), we get

$$\Delta_t \le \Delta_{t-1} + \frac{\alpha a_t^2}{2\sigma_t^2}. \tag{29}$$

Now iterate (29) from $t = T$ down to $t = 1$:

$$\Delta_T \le \Delta_0 + \frac{\alpha}{2}\sum_{t=1}^{T} \frac{a_t^2}{\sigma_t^2}. \tag{30}$$

Finally,

$$\begin{aligned} \Delta_0 &= D_\alpha^{(z_0)}(P_{X_0} \| P_{X'_0}) \\ &= D_\alpha^{(0)}(P_{X_0} \| P_{X'_0}) \\ &= D_\alpha(P_{X_0} \| P_{X'_0}) \\ &= 0. \qquad \big[\text{same initialization } X_0 = X'_0\big] \end{aligned} \tag{31}$$

Substituting (31) into (30), and then using (25), we conclude that

$$D_\alpha(P_{X_T} \| P_{X'_T}) \le \frac{\alpha}{2}\sum_{t=1}^{T} \frac{a_t^2}{\sigma_t^2}.$$

□

proof of proposition 1 new pabi bound. The proof is the same as proposition 2, except that we stop the recursion at time τ instead of unrolling it all the way to 0.

Define

$$\Delta_t := D_\alpha^{(z_t)}\big(P_{X_t} \,\|\, P_{X'_t}\big), \qquad z_t := D + \sum_{i=\tau+1}^{t} (s_i - a_i),$$

so that $z_t \ge 0$ for all t and $z_T = 0$. Hence

$$\begin{aligned} D_\alpha(P_{X_T} \| P_{X'_T}) &= D_\alpha^{(z_T)}(P_{X_T} \| P_{X'_T}) \\ &= \Delta_T. \qquad (\text{since } z_T = 0) \end{aligned} \tag{32}$$

Now repeat exactly the same one-step argument as in Proposition 2.17: using post-processing for the projection, then Lemma 2.15, and then Lemma 2.16, we obtain

$$\Delta_t \le \Delta_{t-1} + \frac{\alpha a_t^2}{2\sigma_t^2}, \qquad t = \tau+1, \dots, T. \tag{33}$$

Unrolling only from $t = T$ down to $t = \tau + 1$, this gives

$$\Delta_T \leq \Delta_\tau + \frac{\alpha}{2} \sum_{t=\tau+1}^{T} \frac{a_t^2}{\sigma_t^2}. \tag{34}$$

It remains to bound Δ_τ. By definition,

$$z_\tau = D.$$

Since both X_τ and X'_τ lie in K, and K has diameter D, we have

$$\|X_\tau - X'_\tau\| \leq D \qquad \text{almost surely,}$$

and therefore

$$W_\infty(P_{X_\tau}, P_{X'_\tau}) \leq D.$$

Thus, in the definition of shifted Rényi divergence, we may choose $\widetilde{\mu} = P_{X'_\tau}$, which gives

$$\begin{aligned} \Delta_\tau &= D_\alpha^{(D)}(P_{X_\tau} \| P_{X'_\tau}) \\ &\leq D_\alpha(P_{X'_\tau} \| P_{X'_\tau}) \\ &= 0. \qquad \text{(since } W_\infty(P_{X_\tau}, P_{X'_\tau}) \leq D\text{)} \end{aligned} \tag{35}$$

Hence $\Delta_\tau = 0$. Substituting into (34) and using (32), we obtain

$$D_\alpha(P_{X_T} \| P_{X'_T}) \leq \frac{\alpha}{2} \sum_{t=\tau+1}^{T} \frac{a_t^2}{\sigma_t^2}.$$

□