# Appendix

## Table of Contents

# A Table of Notations

Table 2: Symbol reference

| Symbol | Meaning |
|---|---|
| $d$ | Dimension of model parameters. |
| $\mathbb{R}^d$ | Unconstrained model parameter space of dimension $d$. |
| $\mathcal{C}$ | A closed convex Model parameter set $\mathcal{C} \subseteq \mathbb{R}^d$ for convex optimization. |
| $\Pi_{\mathcal{C}}(\theta)$ | Projection of $\theta$ to $\mathcal{C}$. |
| $\mathcal{X}, \mathcal{X}^n$ | Data universe and Domain of all datasets of size $n$. |
| $n$ | Dataset size. |
| $D, D'$ | Neighbouring Dataset of size $n$. |
| $\mathbf{x}_i$ | $i$-th data point in dataset $D$. |
| $\ell(\theta; \mathbf{x})$ | Risk of parameter $\theta$ w.r.t data point $\mathbf{x}$. |
| $\mathcal{L}_D(\theta)$ | Emprical risk optimization objective. |
| $U_1(\theta)$ | Average between gradients on neighbouring datasets $D$ and $D'$. |
| $U_2(\theta)$ | Half of difference between gradients on neighbouring datasets $D$ and $D'$. |
| $g(\theta; D)$ | Sum of risk gradients at $\theta$ for all data points in $D$. |
| $V_t, V_t'$ | Time-variable vector fields on $\mathbb{R}^d$. |
| $S_g$ | $\ell_2$-sensitivity of total loss gradient $g(\theta; D)$ |
| $S_v$ | maximum $\ell_2$ distance between $V_t$ and $V_t'$ for all $t > 0$. |
| $\theta^*$ | Parameter minimizing the empirical risk $\mathcal{L}_{\mathcal{D}}(\theta)$. |
| $\mathcal{L}$ | Potential function for Langevin diffusion. |
| $\mathbf{W}_t$ | Standard Brownian motion aka. Wiener process. |
| $\alpha$ | Rényi differential privacy order. |
| $\delta$ | Probability of uncontrolled breach in standard DP. |
| $\varepsilon$ | Rényi or standard DP privacy parameter. |
| $\mathcal{A}$ | Randomized algorithm. |
| $\nu$ , $\nu'$ | Two probability measures. |
| $p$ , $p'$ | Two Probability densities over parameter space $\mathbb{R}^d$. |
| $\Theta, \Theta'$ | Two random variables distributed as $p, p'$ respectively. |
| $\sigma^2$ | Noise variance in noisy GD and Langevin diffusion. |
| $\mathbb{I}_d$ | $d$-dimensional identity matrix. |
| $\mathcal{N}(0, \mathbb{I}_d)$ | Standard gaussion distribution with dimension $d$. |
| $Z, Z_1, Z_2, \cdots$ | Random variables taken from $\mathcal{N}(0, \mathbb{I}_d)$. |
| $\eta$ | Step size of updates in noisy GD. |
| $\lambda$ | Strong convexity parameter of risk function. |
| $\beta$ | Smoothness parameter of risk function. |
| $B$ | Bound on range of risk function. |
| $L$ | Lipschitzness parameter of risk function. |
| $K, k$ | Number of update steps and intermediate step index in noisy GD. |
| $\theta_k, \theta_k'$ | Parameter at step $k$ of noisy GD on $\mathcal{D}, \mathcal{D}'$. |
| $T, t$ | Termination time and intermediate time stamp for diffusion. |
| $\Theta_t, \Theta_t'$ | Model parameter random variable at time $t$ of diffusion on $\mathcal{D}, \mathcal{D}'$. |
| $p_t, p_t'$ | Probability densities or random variables $\Theta_t, \Theta_t'$ |
| $p_{t_1,t_2}$ | Joint density between diffusion random variables $(\Theta_{t_1}, \Theta_{t_2})$. |
| $p'_{t_1,t_2}$ | Joint density between diffusion random variables $(\Theta'_{t_1}, \Theta'_{t_2})$. |
| $p_{t_1\|t_2}(\theta\|\theta_{t_2})$ | Conditional density for $\Theta_{t_1}$ given $\Theta_{t_2} = \theta_{t_2}$. |
| $p'_{t_1\|t_2}(\theta\|\theta_{t_2})$ | Conditional density for $\Theta'_{t_1}$ given $\Theta'_{t_2} = \theta_{t_2}$. |
| $R_\alpha(\Theta_t\|\|\Theta_t')$ | Rényi divergence of distribution of $\Theta_t$ w.r.t $\Theta_t'$. |
| $E_\alpha(\Theta_t\|\|\Theta_t')$ | $\alpha^{\text{th}}$ moment of likelihood ratio r.v. between $\Theta_t, \Theta_t'$. |
| $I_\alpha(\Theta_t\|\|\Theta_t')$ | Rényi Information of distribution of $\Theta_t$ w.r.t $\Theta_t'$. |
| $c$ | Constant in Log sobolev inequality. |
| $\ll$ | Absolute continuity with respect to measure. |

## B Preliminaries

### B.1 Divergence measures

A measure $\nu$ is said to be absolutely continuous with respect to another measure $\nu'$ on same space (denoted as $\nu \ll \nu'$) if for all measurable set $S$, $\nu(S) = 0$ whenever $\nu'(S) = 0$.

**Definition B.1** ($\alpha$-Rényi Divergence)**.** *For $\alpha > 1$, and any two measures $\nu, \nu'$ with $\nu \ll \nu'$, the* $\alpha$-Rényi Divergence $R_\alpha(.\|.)$ *of $\nu$ with respect to $\nu'$ is defined as*

$$R_\alpha(\nu\|\nu') = \frac{1}{\alpha - 1} \log E_\alpha(\nu\|\nu'), \tag{24}$$

*where $E_\alpha(\nu\|\nu')$ is defined as:*

$$E_\alpha(\nu\|\nu') = \int \left(\frac{d\nu}{d\nu'}\right)^\alpha d\nu', \tag{25}$$

*Additionally, if $\nu$ and $\nu'$ are absolutely continuous with Lebesgue measures on $\mathbb{R}^d$ (i.e. they are continuous distributions on $\mathbb{R}^d$) with densities $p$ and $p'$ respectively, $E_\alpha(\nu\|\nu')$ is same as*

$$E_\alpha(\nu\|\nu') = \mathop{\mathbb{E}}_{\theta\sim p'}\left[\frac{p(\theta)^\alpha}{p'(\theta)^\alpha}\right]. \tag{26}$$

As an example, the $\alpha$-Rényi divergence between two Gaussian distributions centered at $\mu, \mu' \in \mathbb{R}^d$, with covariance matrix $\sigma^2\mathbb{I}_d$ is $\frac{\alpha\|\mu-\mu'\|_2^2}{2\sigma^2}$ [17, Proposition 7].

**Definition B.2** (Rényi information [22])**.** *Let $1 < \alpha < \infty$. For any two measures $\nu, \nu'$ with $\nu \ll \nu'$, if the Radon-Nikodym derivative $\frac{d\nu}{d\nu'}$ is differentiable, the* $\alpha$-Rényi Information $I_\alpha(\cdot\|\cdot)$ *of $\nu$ with respect to $\nu'$ is*

$$I_\alpha(\nu\|\nu') = \int \left(\frac{d\nu}{d\nu'}\right)^\alpha \left\|\nabla \log \frac{d\nu}{d\nu'}\right\|_2^2 d\nu'. \tag{27}$$

*Additionally, if $\nu$ and $\nu'$ are absolutely continuous with Lebesgue measures (i.e. they are continuous distributions on $\mathbb{R}^d$) with densities $p$ and $p'$ respectively, $I_\alpha(\nu\|\nu')$ is same as*

$$I_\alpha(\nu\|\nu') = \frac{4}{\alpha^2}\mathop{\mathbb{E}}_{\theta\sim p'}\left[\left\|\nabla \frac{p(\theta)^{\frac{\alpha}{2}}}{p'(\theta)^{\frac{\alpha}{2}}}\right\|_2^2\right] = \mathop{\mathbb{E}}_{\theta\sim p'}\left[\frac{p(\theta)^{\alpha-2}}{p'(\theta)^{\alpha-2}}\left\|\nabla \frac{p(\theta)}{p'(\theta)}\right\|_2^2\right]. \tag{28}$$

### B.2 Differential privacy

Let $\mathcal{X}$ be a data universe. Let a dataset be a vector of $n$ records from $\mathcal{X}$: $D = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n) \in \mathcal{X}^n$.

**Definition B.3** (Neighboring datasets)**.** *Two datasets $D$ and $D'$ are* neighboring*, denoted by $D \sim D'$, if $|D| = |D'|$, and they differ in exactly one data record, i.e., $|D \oplus D'| = 2$.*

**Definition B.4** (Differential drivacy [7])**.** *A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathbb{R}^d$ satisfies* $(\varepsilon, \delta)$-differential privacy (DP) *if for any two neighboring datasets $D, D' \in \mathcal{X}^n$, and for all sets $S \in \mathbb{R}^d$,*

$$Pr[\mathcal{A}(D) \in S] \le e^\varepsilon Pr[\mathcal{A}(D') \in S] + \delta. \tag{29}$$

**Definition B.5** (Rényi differential privacy [17])**.** *Let $\alpha > 1$. A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathbb{R}^d$ satisfies* $(\alpha, \varepsilon)$-Rényi Differential Privacy(RDP)*, if for any two neighboring datasets $D, D' \in \mathcal{X}^n$:*

$$R_\alpha(\mathcal{A}(D)\|\mathcal{A}(D')) \le \varepsilon. \tag{30}$$

In this paper, we mainly use Rényi differential privacy notion to analyze the privacy loss of algorithms. We refer to $R_\alpha(\mathcal{A}(D)\|\mathcal{A}(D'))$ as the *Rényi privacy loss* of algorithm $\mathcal{A}$ on datasets $D, D'$.

**Theorem 6** (RDP composition theorem [17, Proposition 1])**.** *Let $\mathcal{A}_1 : \mathcal{X}^n \to \mathbb{R}^d$ and $\mathcal{A}_2 : \mathbb{R}^d \times \mathcal{X}^n \to \mathbb{R}^d$ be two randomized algorithms that satisfy $(\alpha, \varepsilon_1)$ and $(\alpha, \varepsilon_2)$-RDP, respectively. The composed algorithm defined as $\mathcal{A}(D) = (\mathcal{A}_1(D), \mathcal{A}_2(D))$ satisfies $(\alpha, \varepsilon_1 + \varepsilon_2)$-Rényi DP.*

An RDP guarantee can be converted to a DP guarantee as per the following theorem.

**Theorem 7** (DP Conversion [17, Proposition 3])**.** *If a randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathbb{R}^d$ satisfies $(\alpha, \varepsilon)$-RDP, then it also satisfies the standard $(\varepsilon + \frac{\log 1/\delta}{\alpha - 1}, \delta)$-DP guarantee for any $0 < \delta < 1$.*

### B.3 Langevin diffusion

We focus on the Langevin diffusion process in $\mathbb{R}^d$ with noise variance $\sigma^2$, described by the following stochastic differential equation (SDE).

$$d\Theta_t = -\nabla\mathcal{L}(\Theta_t)dt + \sqrt{2\sigma^2}d\mathbf{W}_t, \tag{31}$$

where $d\mathbf{W}_t = \mathbf{W}_{t+dt} - \mathbf{W}_t \sim \sqrt{dt}\mathcal{N}(0, \mathbb{I}_d)$ characterizes the $d$-dimensional Wiener process.

The joint effect of this drag force (i.e. $-\nabla\mathcal{L}$) and Brownian fluctuations on the probability density $p_t$ of position random variable $\Theta_t$ is characterized through the Fokker-Planck equation [10],

$$\frac{\partial p_t(\theta)}{\partial t} = \nabla \cdot (p_t(\theta)\nabla\mathcal{L}(\theta)) + \sigma^2\Delta p_t(\theta), \tag{32}$$

which describes the rate of change in probability density at any position $\theta \in \mathbb{R}^d$. It's important to point out that Fokker-Planck equation isn't a property of Langevin diffusion, but rather a general equation quantifying the distributional change under *any* drag force in presence of Brownian fluctuations.

Under mild regularity conditions on the potential $\mathcal{L}(\theta)$, this diffusion process has a stationary distribution $\nu$, given by the solution to $\frac{\partial p_t(\theta)}{\partial t} = 0$, which is the following Gibbs distribution.

$$\nu(\theta) = \frac{1}{V}e^{-\mathcal{L}(\theta)/\sigma^2}, \text{ where } V = \int_{\mathbb{R}^d} e^{-\mathcal{L}(\theta)/\sigma^2}d\theta. \tag{33}$$

### B.4 Loss function properties

For any data record $\mathbf{x} \in \mathcal{X}$, a loss function $\ell(\theta; \mathbf{x}) : \mathcal{C} \to \mathbb{R}$ on a closed convex set $\mathcal{C}$ maps parameter $\theta \in \mathcal{C} \subseteq \mathbb{R}^d$ to a real value. Let $\nabla\ell(\theta; \mathbf{x})$ be its loss gradient vector with respect to $\theta$.

**Definition B.6** (Lipschitz continuity)**.** *Function $\ell(\theta; \mathbf{x})$ is* $L$-Lipschitz continuous *if for all $\theta, \theta' \in \mathcal{C}$ and $\mathbf{x} \in \mathcal{X}$,*

$$|\ell(\theta; \mathbf{x}) - \ell(\mathbf{x}; \theta')| \leq L \left\|\theta - \theta'\right\|_2. \tag{34}$$

**Definition B.7** (Smoothness)**.** *Differentiable function $\ell(\theta; \mathbf{x})$ is* $\beta$-smooth *over $\mathcal{C}$ if for all $\theta, \theta' \in \mathcal{C}$ and $\mathbf{x} \in \mathcal{X}$,*

$$\left\|\nabla\ell(\theta; \mathbf{x}) - \nabla\ell(\theta'; \mathbf{x})\right\|_2 \leq \beta \left\|\theta - \theta'\right\|_2. \tag{35}$$

**Definition B.8** (Strong convexity)**.** *Differentiable function $\ell(\theta; \mathbf{x})$ is* $\lambda$-strongly convex *if for all $\theta, \theta' \in \mathbb{R}^d$ and $\mathbf{x} \in \mathcal{X}$,*

$$\ell(\theta'; \mathbf{x}) \geq \ell(\theta; \mathbf{x}) + \nabla\ell(\theta; \mathbf{x})^T(\theta' - \theta) + \frac{\lambda}{2}\left\|\theta' - \theta\right\|_2^2. \tag{36}$$

**Definition B.9** (Vector field sensitivity)**.** *For two vector fields $V, V'$ on $\mathbb{R}^d$, we define $S_v$ to be the $l_2$-sensitivity between them:*

$$S_v = \max_{\theta\in\mathbb{R}^d} \left\|V(\theta) - V'(\theta)\right\|_2. \tag{37}$$

**Definition B.10** (Sensitivity of total gradient)**.** *For a differentiable function $\ell(\theta; \mathbf{x})$, we define $S_g$ to be the $l_2$-sensitivity of its total gradient $g(\theta; D) = \sum_{\mathbf{x}\in D}\nabla\ell(\theta; \mathbf{x})$ on neighboring datasets $D, D' \in \mathcal{X}^n$:*

$$S_g = \max_{D\sim D'} \max_{\theta\in\mathbb{R}^d} \left\|g(\theta; D) - g(\theta; D')\right\|_2. \tag{38}$$

In Appendix C, we briefly present the basic vector calculus that we require in this paper.

## C Calculus Refresher

Given a smooth function $\mathcal{L} : \Theta \to \mathbb{R}$, where $\Theta \subset \mathbb{R}^d$, its gradient $\nabla\mathcal{L} : \mathcal{X}^n \to \mathbb{R}^d$ is the vector of partial derivatives

$$\nabla\mathcal{L}(\theta) = \left(\frac{\partial\mathcal{L}(\theta)}{\partial\theta_1}, \cdots, \frac{\partial\mathcal{L}(\theta)}{\partial\theta_2}\right). \tag{39}$$

Its Hessian $\nabla^2 \mathcal{L} : \Theta \to \mathbb{R}^{d \times d}$ is the matrix of second partial derivatives

$$\nabla^2 \mathcal{L}(\theta) = \left( \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta_i \theta_j} \right)_{1 \le i,j \le d}. \tag{40}$$

Its Laplacian $\Delta \mathcal{L} : \Theta \to \mathbb{R}$ is the trace of its Hessian $\nabla^2 \mathcal{L}$, i.e.,

$$\Delta \mathcal{L}(\theta) = \mathsf{Tr}\left( \nabla^2 \mathcal{L}(\theta) \right). \tag{41}$$

Given a smooth vector field $\mathbf{v} = (\mathbf{v}_1, \cdots, \mathbf{v}_d) : \Theta \to \mathbb{R}^d$, its divergence $\nabla \cdot \mathbf{v} : \Theta \to \mathbb{R}$ is

$$(\nabla \cdot \mathbf{v})(\theta) = \sum_{i=1}^{d} \frac{\partial \mathbf{v}_i(\theta)}{\partial \theta_i}. \tag{42}$$

Some identities that we would rely on:

1. Divergence of gradient is the Laplacian, i.e.,

$$\nabla \cdot \nabla \mathcal{L}(\theta) = \sum_{i=1}^{d} \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta_i^2} = \Delta \mathcal{L}(\theta). \tag{43}$$

2. For any function $f : \Theta \to \mathbb{R}$ and a vector field $\mathbf{v} : \Theta \to \mathbb{R}^d$ with sufficiently fast decay to a constant at the border of $\Theta$,

$$\int_{\Theta} \langle \mathbf{v}(\theta), \nabla f(\theta) d\theta \rangle = -\int_{\Theta} f(\theta)(\nabla \cdot \mathbf{v})(\theta) d\theta. \tag{44}$$

3. For any two twice continuously differentiable functions $f, g : \Theta \to \mathbb{R}$, out of which at least for one the gradient decays sufficiently fast at infinity, the following also holds.

$$\int_{\Theta} f(\theta) \Delta g(\theta) d\theta = -\int_{\Theta} \langle \nabla f(\theta), \nabla g(\theta) \rangle \, d\theta = \int_{\Theta} g(\theta) \Delta f(\theta) d\theta. \tag{45}$$

   This identity comes from the Green's first identity, which is the higher dimensional equivalent of integration by part.

4. Based on Young's inequality, for two vector fields $\mathbf{v}_1, \mathbf{v}_2 : \Theta \to \mathbb{R}^d$, and any $a, b \in \mathbb{R}$ such that $ab = 1$, the following inequality holds.

$$\langle \mathbf{v}_1, \mathbf{v}_2 \rangle (\theta) \le \frac{1}{2a} \left\| \mathbf{v}_1(\theta) \right\|_2^2 + \frac{1}{2b} \left\| \mathbf{v}_2(\theta) \right\|_2^2. \tag{46}$$

Wherever it is clear, we would drop $(\theta)$ for brevity. For example, we would represent $(\nabla \cdot \mathbf{v})(\theta)$ as only $\nabla \cdot \mathbf{v}$.

# D Proofs for Section 3: Privacy analysis of noisy gradient descent

## D.1 Proofs for Section 3.1: Tracing diffusion for Noisy GD

**Lemma 1.** *For coupled diffusion processes* (5) *in time $\eta k < t < \eta(k+1)$, the equivalent Fokker-Planck equations are*

$$\begin{cases} \frac{\partial p_t(\theta)}{\partial t} = \nabla \cdot (p_t(\theta) V_t(\theta)) + \sigma^2 \Delta p_t(\theta) \\ \frac{\partial p'_t(\theta)}{\partial t} = \nabla \cdot (p'_t(\theta) V'_t(\theta)) + \sigma^2 \Delta p'_t(\theta), \end{cases} \tag{47}$$

*where $V_t(\theta) = -V'_t(\theta) = \mathop{\mathbb{E}}_{\theta_k \sim p_{\eta k | t}} [U_2(\theta_k) | \theta]$ are time-dependent vector fields on $\mathbb{R}^d$, and $U_2(\theta) = \frac{\nabla \mathcal{L}_D(\theta) - \nabla \mathcal{L}_{D'}(\theta)}{2}$ is the difference between gradients on neighboring datasets $D$ and $D'$.*

*Proof.* We only prove $\frac{\partial p_t(\theta)}{\partial t} = \nabla \cdot (p_t(\theta) V_t(\theta)) + \sigma^2 \Delta p_t(\theta)$. The proof for the other Fokker-Planck equation is similar.

Recall that conditionals of joint distribution $p_{\eta k,t}$ is

$$p_{\eta k,t}(\theta_k, \theta) = p_{\eta k}(\theta_k) p_{t|\eta k}(\theta|\theta_k) = p_t(\theta) p_{\eta k|t}(\theta_k|\theta). \tag{48}$$

By marginalizing away $\theta_k$ in (48), and taking partial derivative w.r.t. $t$ on both sides, we obtain the following:

$$\begin{aligned}
\frac{\partial p_t(\theta)}{\partial t} &= \int_{\mathbb{R}^d} \frac{\partial p_{t|\eta k}(\theta|\theta_k)}{\partial t} p_{\eta k}(\theta_k) d\theta_k \\
&= \int_{\mathbb{R}^d} \left( \nabla \cdot (p_{\eta k,t}(\theta_k, \theta) U_2(\theta_k)) + \sigma^2 \Delta p_{\eta k,t}(\theta_k, \theta) \right) d\theta_k && \text{(By (8))} \\
&= \nabla \cdot \left( p_t(\theta) \int_{\mathbb{R}^d} p_{\eta k|t}(\theta_k|\theta) U_2(\theta_k) d\theta_k \right) + \sigma^2 \Delta p_t(\theta) \\
&= \nabla \cdot \left( p_t(\theta) \mathop{\mathbb{E}}_{\theta_k \sim p_{\eta k|t}} [U_2(\theta_k)|\theta] \right) + \sigma^2 \Delta p_t(\theta) \\
&= \nabla \cdot (p_t(\theta) \cdot V_t(\theta)) + \sigma^2 \Delta p_t(\theta) && \text{(where } V_t(\theta) = \mathop{\mathbb{E}}_{\theta_k \sim p_{\eta k|t}} [U_2(\theta_k)|\theta])
\end{aligned}$$

□

### D.2 Proofs for Section 3.2: Privacy erosion in tracing diffusion

**Lemma 7** (Leibniz integral rule)**.** *Suppose $f_t : \mathbb{R}^d \to \mathbb{R}$ is Lebesgue-integrable for each $t \geq 0$. If for almost all $\theta \in \mathbb{R}^d$, the derivative $\frac{df_t}{dt}$ exists and there exists an integrable function $g : \mathbb{R}^d \to \mathbb{R}$ such that $\left| \frac{df_t}{dt}(\theta) \right| \leq g(\theta)$ for all $t \geq 0$ and almost every $\theta \in \mathbb{R}^d$, then*

$$\frac{d}{dt} \int_{\mathbb{R}^d} f_t(\theta) d\theta = \int_{\mathbb{R}^d} \frac{df_t}{dt}(\theta) d\theta, \quad \textit{for all } t \geq 0. \tag{49}$$

**Lemma 2** ( Rate of Rényi divergence )**.** *Let $V_t$ and $V_t'$ be two vector fields on $\mathbb{R}^d$ with $\max_{\theta \in \mathbb{R}^d} \|V_t(\theta) - V_t'(\theta)\|_2 \leq S_v$ for all $t \geq 0$. Then, for corresponding coupled diffusions $\{\Theta_t\}_{t \geq 0}$ and $\{\Theta_t'\}_{t \geq 0}$ under $V_t$ and $V_t'$ with noise variance $\sigma^2$, the rate of Rényi divergence at any $t \geq 0$ is upper bounded by*

$$\frac{\partial R_\alpha \left( \Theta_t \| \Theta_t' \right)}{\partial t} \leq \frac{1}{\gamma} \frac{\alpha S_v^2}{4\sigma^2} - (1 - \gamma) \sigma^2 \alpha \frac{I_\alpha \left( \Theta_t \| \Theta_t' \right)}{E_\alpha \left( \Theta_t \| \Theta_t' \right)}. \tag{50}$$

*where $\gamma > 0$ is a tuning parameter that we can fix arbitrarily according to our need.*

*Proof.* For brevity, let the functions $R(\alpha, t) = R_\alpha \left( p_t \| p_t' \right)$, $E(\alpha, t) = E_\alpha \left( p_t \| p_t' \right)$, and $I(\alpha, t) = I_\alpha \left( p_t \| p_t' \right)$. Under the stated assumptions $\frac{\partial E(\alpha, t)}{\partial t}$ is bounded as follows.

$$\frac{\partial E(\alpha, t)}{\partial t} = \frac{\partial}{\partial t} \int_{\mathbb{R}^d} \frac{p_t^\alpha}{p_t'^{\alpha - 1}} d\theta \tag{51}$$

By Leibniz integral rule (Lemma 7), we exchange order of derivative and integration in (51). The necessary conditions are satisfied because of the following properties about $p_t$ and $p_t'$:

1. $p_t$ and $p_t'$ have the same support, and their Renyi divergence is well-defined.
2. The distributions of coupled tracing diffusions $\{\theta_t\}_{\eta k < t < \eta(k+1)}$ and $\{\theta_t'\}_{\eta k < t < \eta(k+1)}$ have full support and smooth densities $p_t$ and $p_t'$ (due to convolution with Gaussian noise).
3. The evolutions of probability densities $p_t$ and $p_t'$ with regard to time $t$ satisfy the Fokker-Planck equations (8).

Therefore, we obtain:

$$\begin{aligned}
\frac{\partial E(\alpha,t)}{\partial t} &= \alpha\int_{\mathbb{R}^d}\frac{\partial p_t}{\partial t}\left(\frac{p_t}{p_t'}\right)^{\alpha-1}d\theta-(\alpha-1)\int_{\mathbb{R}^d}\frac{\partial p_t'}{\partial t}\left(\frac{p_t}{p_t'}\right)^{\alpha}d\theta && \text{(By Lemma 7)}\\
&= \alpha\int_{\mathbb{R}^d}\left(\sigma^2\Delta p_t+\nabla\cdot(p_tV_t)\right)\left(\frac{p_t}{p_t'}\right)^{\alpha-1}d\theta\\
&\quad-(\alpha-1)\int_{\mathbb{R}^d}\left(\sigma^2\Delta p_t'+\nabla\cdot(p_t'V_t')\right)\left(\frac{p_t}{p_t'}\right)^{\alpha}d\theta && \text{(From (9))}\\
&= \underbrace{\sigma^2\alpha\int_{\mathbb{R}^d}\left(\frac{p_t}{p_t'}\right)^{\alpha-1}\Delta p_t d\theta\ -\sigma^2(\alpha-1)\int_{\mathbb{R}^d}\left(\frac{p_t}{p_t'}\right)^{\alpha}\Delta p_t' d\theta}_{\stackrel{\text{def}}{=}F_1}\\
&\quad+\underbrace{\alpha\int_{\mathbb{R}^d}\left(\frac{p_t}{p_t'}\right)^{\alpha-1}\nabla\cdot(p_tV_t)d\theta\ -(\alpha-1)\int_{\mathbb{R}^d}\left(\frac{p_t}{p_t'}\right)^{\alpha}\nabla\cdot(p_t'V_t')d\theta}_{\stackrel{\text{def}}{=}F_2}
\end{aligned}$$

We simplify $F_1$ as following:

$$\begin{aligned}
F_1 &= \sigma^2(\alpha-1)\int_{\mathbb{R}^d}\left\langle\nabla\left(\frac{p_t}{p_t'}\right)^{\alpha},\nabla p_t'\right\rangle d\theta-\sigma^2\alpha\int_{\mathbb{R}^d}\left\langle\nabla\left(\frac{p_t}{p_t'}\right)^{\alpha-1},\nabla p_t\right\rangle d\theta && \text{(From (45))}\\
&= \sigma^2\alpha(\alpha-1)\int_{\mathbb{R}^d}\left(\frac{p_t}{p_t'}\right)^{\alpha-2}\left\langle\nabla\frac{p_t}{p_t'},\frac{p_t}{p_t'^2}\nabla p_t'-\frac{\nabla p_t}{p_t'}\right\rangle p_t' d\theta\\
&= -\sigma^2\alpha(\alpha-1)\mathop{\mathbb{E}}_{p_t'}\left[\left(\frac{p_t}{p_t'}\right)^{\alpha-2}\left\|\nabla\frac{p_t}{p_t'}\right\|_2^2\right] && (\because\nabla\tfrac{p_t}{p_t'}=\tfrac{\nabla p_t}{p_t'}-\tfrac{p_t}{p_t'^2}\nabla p_t')\\
&= -\sigma^2\alpha(\alpha-1)I(\alpha,t) && \text{(From (28))}
\end{aligned}$$

We upper bound $F_2$ as following:

$$\begin{aligned}
F_2 &= -\alpha\int_{\mathbb{R}^d}\left\langle\nabla\left(\frac{p_t}{p_t'}\right)^{\alpha-1},p_tV_t\right\rangle d\theta\ +(\alpha-1)\int_{\mathbb{R}^d}\left\langle\nabla\left(\frac{p_t}{p_t'}\right)^{\alpha},p_t'V_t'\right\rangle d\theta && \text{(From (44))}\\
&= \alpha(\alpha-1)\int_{\mathbb{R}^d}\left(\frac{p_t}{p_t'}\right)^{\alpha-2}\left\langle\nabla\frac{p_t}{p_t'},\frac{p_t}{p_t'}(V_t'-V_t)\right\rangle p_t' d\theta\\
&\le \gamma\alpha(\alpha-1)\sigma^2\int_{\mathbb{R}^d}\left(\frac{p_t}{p_t'}\right)^{\alpha-2}\left\|\nabla\frac{p_t}{p_t'}\right\|_2^2 p_t' d\theta && \text{(From (46) with } b=2\gamma\sigma^2)\\
&\quad+\frac{\alpha(\alpha-1)S_v^2}{4\gamma\sigma^2}\int_{\mathbb{R}^d}\left(\frac{p_t}{p_t'}\right)^{\alpha-2}\times\left(\frac{p_t}{p_t'}\right)^2 p_t' d\theta && (\because\max_{\theta\in\mathbb{R}^d}\|V_t(\theta)-V_t'(\theta)\|_2\le S_v)\\
&= \gamma\sigma^2\alpha(\alpha-1)I(\alpha,t)+\frac{1}{\gamma}\frac{\alpha(\alpha-1)S_v^2}{4\sigma^2}E(\alpha,t) && \text{(From (26) \& (28))}
\end{aligned}$$

Therefore, we get the following bound on the rate of Renyi divergence:

$$\begin{aligned}
\frac{\partial R(\alpha,t)}{\partial t} &= \frac{1}{\alpha-1}\times\frac{1}{E(\alpha,t)}\times\frac{\partial E(\alpha,t)}{\partial t}\\
&\le -(1-\gamma)\sigma^2\alpha\frac{I(\alpha,t)}{E(\alpha,t)}+\frac{1}{\gamma}\frac{\alpha S_v^2}{4\sigma^2}
\end{aligned}$$

□

**Discussions about the terms in Lemma 2** Lemma 2 bounds the rate of privacy loss with various terms. Generally speaking, the term $\frac{\alpha S_v^2}{4\sigma^2}$ bounds the worst-case privacy loss growth due to noisy gradient update when $S_v=\frac{S_g}{n}$, while the term $\frac{I_\alpha(\Theta_t\|\Theta_t')}{E_\alpha(\Theta_t\|\Theta_t')}$ amplifies our bound for the rate of privacy loss, as the Rényi privacy loss accumulates during the process. We offer more explanations as the following.

1. $\frac{\alpha S_g^2}{4\sigma^2 n^2}$: This is the first term in the right hand side of (15). It quantifies the worst-case privacy loss due of one noisy gradient update in noisy GD Algorithm 1. The term $\frac{S_g}{n}$ is the sensitivity of average loss gradient $\mathcal{L}_D(\theta)$ over two neighboring datasets $D, D'$. The larger $S_g$ is, the further apart the parameters $\theta$ and $\theta'$ after the gradient descent updates on two neighboring dataset $D, D'$ could be, where $\theta = \theta_0 - \eta \nabla \mathcal{L}_D(\theta_0)$ and $\theta' = \theta_0 - \eta \mathcal{L}_{D'}(\theta_0)$. The term $\sigma^2$ is the variance of Gaussian noise. Because additive noise shrink the expected trajectory difference between $\theta$ and $\theta'$ in noisy GD updates, the larger $\sigma^2$ is, the more indistinguishable the distributions of sum of $\theta, \theta'$ and Gaussian noise will be, therefore the smaller the privacy loss (Rényi divergence between end distributions) will be.

2. $\frac{I_\alpha(\Theta_t \| \Theta_t')}{E_\alpha(\Theta_t \| \Theta_t')}$: This term is the second term in the right hand side of (10), which originates from the derivative of $p_t, p_t'$ with regard to time $t$. To obtain the expression $I_\alpha / E_\alpha$, we are using the Fokker Planck equation to replace the terms related to $\frac{\partial p_t}{\partial t}, \frac{\partial p_t'}{\partial t}$ with terms determined by the gradient and Laplacian of $p_t, p_t'$ over $\theta$.
   The term $I_\alpha(\Theta_t \| \Theta_t')$ is the **Rényi information** defined in Definition 2.2., which equals $\mathbb{E}_{\theta \sim p_t'} \left[ \left\| \nabla \log \frac{p_t(\theta)}{p_t'(\theta)} \right\|_2^2 \left( \frac{p_t(\theta)}{p_t'(\theta)} \right)^\alpha \right]$. The term $E_\alpha(\Theta_t \| \Theta_t')$ is the **moment of likelihood ratio** defined in Definition 2.1., which equals $\mathbb{E}_{\theta \sim p_t'} \left[ \left( \frac{p_t(\theta)}{p_t'(\theta)} \right)^\alpha \right]$. These two terms differ by a **multiplicative ratio** $\left\| \nabla \log \frac{p_t(\theta)}{p_t'(\theta)} \right\|_2^2$ for their quantities inside expectation. This ratio characterizes the variation of log likelihood ratio function across $\theta$, where $\theta$ is taken from distribution $p_t'$. This is intuitive in the one dimensional case, because $\int_{\theta_1}^{\theta_2} \nabla \log \frac{p_t(\theta)}{p_t'(\theta)} d\theta = \log \frac{p_t(\theta_2)}{p_t'(\theta_2)} - \log \frac{p_t(\theta_1)}{p_t'(\theta_2)}$. Meanwhile since $p_t(\theta), p_t(\theta)'$ are continuous and $\int p_t(\theta) d\theta = \int p_t'(\theta) d\theta = 1$, by mean value theorem, there exists $\tilde{\theta} \in \mathbb{R}^d$ such that the log likelihood ratio $\log \frac{p_t(\tilde{\theta})}{p_t'(\tilde{\theta})}$ is zero. Therefore the variation of log likelihood ratio across $\theta$ implicitly increases the largest log likelihood ratio $\max_{\theta \in \mathbb{R}^d} \left[ \log(\frac{p_t(\theta)}{p_t(\theta)}) - \log(\frac{p_t(\tilde{\theta})}{p_t(\tilde{\theta})}) \right] = \max_{\theta \in \mathbb{R}^d} \left[ \log(\frac{p_t(\theta)}{p_t(\theta)}) \right]$ across $\theta$ , which reflects the Rényi privacy loss $R_\alpha$.
   As a result, intuitively, under some conditions, the larger the Rényi privacy loss $R_\alpha$ is, the larger the variation of log likelihood ratio across $\theta$ will be, and therefore the larger the term $\frac{I_\alpha(\Theta_t \| \Theta_t')}{E_\alpha(\Theta_t \| \Theta_t')}$ will be. Therefore when the Rényi privacy loss $R_\alpha$ is large, the bound for the rate of privacy loss in (10) Lemma 2 will also be smaller (under $(1 - \gamma) > 0$).

3. $\gamma$ is a tuning constant to balance the privacy growth rate estimated using the above two terms, thus helping us tune the privacy loss accumulation. See the tightness results in Appendix E for more details.

**Theorem 1** (Linear Rényi divergence bound)**.** *Let $V_t$ and $V_t'$ be two vector fields on $\mathbb{R}^d$, with $\max_{\theta \in \mathbb{R}^d} \|V_t(\theta) - V_t'(\theta)\|_2 \leq S_v$ for all $t \geq 0$. Then, the diffusion under vector fields $V_t$ and $V_t'$ with noise variance $\sigma^2$ for time $T$ has $\alpha$-Rényi divergence of output distributions bounded by $\varepsilon = \frac{\alpha S_v^2 T}{4\sigma^2}$.*

*Proof.* Setting $\gamma = 1$ in Lemma 2 gives constant privacy loss rate. Integrating over $t$ suffices. □

**Controlling Rényi privacy loss rate under isoperimetry**

**Lemma 3** ( [22] $c$-LSI in terms of Rényi Divergence )**.** *Suppose $\Theta_t, \Theta_t' \in \mathbb{R}^d$ are random variables such that probability density ratio between $\Theta_t$ and $\Theta_t'$ lies in $\mathcal{F}_{\Theta_t'}$. Then for any $\alpha \geq 1$,*

$$R_\alpha(\Theta_t \| \Theta_t') + \alpha(\alpha - 1) \frac{\partial R_\alpha(\Theta_t \| \Theta_t')}{\partial \alpha} \leq \frac{\alpha^2}{2c} \frac{I_\alpha(\Theta_t \| \Theta_t')}{E_\alpha(\Theta_t \| \Theta_t')}, \tag{52}$$

*if and only if $\Theta'$ satisfies $c$-LSI.*

*Proof.* Let $p$ and $p'$ denote the probability density functions of $\Theta_t$ and $\Theta_t'$ respectively. For brevity, let the functions $R(\alpha) = R_\alpha(\Theta_t \| \Theta_t')$, $E(\alpha) = E_\alpha(\Theta_t \| \Theta_t')$, and $I(\alpha) = I_\alpha(\Theta_t \| \Theta_t')$. Let function

$g^2(\theta) = \left(\frac{p(\theta)}{p'(\theta)}\right)^\alpha$. Then,

$$\mathop{\mathbb{E}}_{p'}\left[g^2\right] = \mathop{\mathbb{E}}_{p'}\left[\left(\frac{p}{p'}\right)^\alpha\right] = E_\alpha\left(p\|p'\right), \qquad \text{(From (26))}$$

and,

$$\begin{aligned}
\mathop{\mathbb{E}}_{p'}\left[g^2 \log g^2\right] &= \mathop{\mathbb{E}}_{p'}\left[\left(\frac{p}{p'}\right)^\alpha \log\left(\frac{p}{p'}\right)^\alpha\right] \\
&= \alpha\frac{\partial}{\partial\alpha}\mathop{\mathbb{E}}_{p'}\left[\int_\alpha \left(\frac{p}{p'}\right)^\alpha \log\left(\frac{p}{p'}\right) d\alpha\right] && \text{(Lebniz's rule)} \\
&= \alpha\frac{\partial}{\partial\alpha}\mathop{\mathbb{E}}_{p'}\left[\left(\frac{p}{p'}\right)^\alpha\right] = \alpha\frac{\partial E(\alpha)}{\partial\alpha}. && \text{(From (26))}
\end{aligned}$$

Moreover, from (28),

$$\mathop{\mathbb{E}}_{p'}\left[\|\nabla g\|_2^2\right] = \mathop{\mathbb{E}}_{p'}\left[\left\|\nabla\left(\frac{p}{p'}\right)^{\frac{\alpha}{2}}\right\|_2^2\right] = \frac{\alpha^2}{4}I(\alpha). \tag{53}$$

On substituting the above equalities in (11), we get:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{p'}\left[g^2\log g^2\right] - \mathop{\mathbb{E}}_{p'}\left[g^2\right]\log\mathop{\mathbb{E}}_{p'}\left[g^2\right] \le \frac{2}{c}\mathop{\mathbb{E}}_{p'}\left[\|\nabla g\|_2^2\right] \\
\iff &\alpha\frac{\partial E(\alpha)}{\partial\alpha} - E(\alpha)\log E(\alpha) \le \frac{\alpha^2}{2c}I(\alpha) \\
\iff &\alpha\frac{\partial \log E(\alpha)}{\partial\alpha} - \log E(\alpha) \le \frac{\alpha^2}{2c}\frac{I(\alpha)}{E(\alpha)} \\
\iff &\alpha\frac{\partial}{\partial\alpha}\left((\alpha-1)R(\alpha)\right) - (\alpha-1)R(\alpha) \le \frac{\alpha^2}{2c}\frac{I(\alpha)}{E(\alpha)} && \text{(From (24))} \\
\iff &R(\alpha) + \alpha(\alpha-1)\frac{\partial R(\alpha)}{\partial\alpha} \le \frac{\alpha^2}{2c}\frac{I(\alpha)}{E(\alpha)}
\end{aligned}$$

□

### D.3 Proofs for Section 3.3: Privacy guarantee for Noisy GD

**Lemma 4.** *Let $\ell(\theta;\mathbf{x})$ be a loss function on closed convex set $\mathcal{C}$, with a finite total gradient sensitivity $S_g$. Let $\{\Theta_t\}_{t\ge0}$ and $\{\Theta'_t\}_{t\ge0}$ be the coupled tracing diffusions for noisy GD on neighboring datasets $D, D' \in \mathcal{X}^n$, under loss $\ell(\theta;\mathbf{x})$ and noise variance $\sigma^2$. Then the difference between underlying vector fields $V_t$ and $V'_t$ for coupled tracing diffusions is bounded by*

$$\max_{\theta\in\mathbb{R}^d}\|V_t(\theta) - V'_t(\theta)\|_2 \le \frac{S_g}{n}, \tag{54}$$

*where $V_t(\theta)$ and $V'_t(\theta)$ are time-dependent vector fields on $\mathbb{R}^d$, defined in Lemma 1.*

*Proof.* By triangle inequality, for any $\theta \in \mathbb{R}^d$,

$$\begin{aligned}
\|V_t(\theta) - V'_t(\theta)\|_2 &\le \|V_t(\theta)\|_2 + \|V'_t(\theta)\|_2 \\
&\le \frac{1}{2}\mathop{\mathbb{E}}_{\theta_k\sim p_{\eta k|t}}\left[\|\nabla\mathcal{L}_D(\theta_k) - \nabla\mathcal{L}_{D'}(\theta_k)\|_2|\theta\right] && (55) \\
&\quad + \frac{1}{2}\mathop{\mathbb{E}}_{\theta'_k\sim p'_{\eta k|t}}\left[\|\nabla\mathcal{L}_{D'}(\theta'_k) - \nabla\mathcal{L}_D(\theta'_k)\|_2|\theta\right]. && \text{(From Jensen's inequality)}
\end{aligned}$$

By definition of total gradient sensitivity, for any $\theta_k$ and $\theta'_k$, we have

$$\|\nabla\mathcal{L}_D(\theta_k) - \nabla\mathcal{L}_{D'}(\theta_k)\|_2 \le \frac{S_g}{n}, \quad \|\nabla\mathcal{L}_{D'}(\theta'_k) - \nabla\mathcal{L}_D(\theta'_k)\|_2 \le \frac{S_g}{n}.$$

Therefore, by applying this inequality in equation (55) we obtain (54). □

**Theorem 2.** *Let $\{\Theta_t\}_{t\geq 0}$ and $\{\Theta'_t\}_{t\geq 0}$ be the tracing diffusion for $\mathcal{A}_{\text{Noisy-GD}}$ on neighboring datasets $D$ and $D'$, under noise variance $\sigma^2$ and loss function $\ell(\theta;\mathbf{x})$. Let $\ell(\theta;\mathbf{x})$ be a loss function on closed convex set $\mathcal{C}$, with a finite total gradient sensitivity $S_g$. If for any neighboring datasets $D$ and $D'$, the corresponding coupled tracing diffusions $\Theta_t$ and $\Theta'_t$ satisfy c-LSI throughout $0 \leq t \leq \eta K$, then $\mathcal{A}_{\text{Noisy-GD}}$ satisfies $(\alpha, \varepsilon)$ Rényi Differential Privacy for*

$$\varepsilon = \frac{\alpha S_g^2}{2c\sigma^4 n^2}(1 - e^{-\sigma^2 c\eta K}). \tag{56}$$

*Proof.* The RDP evolution equation (15) holds for projected noisy GD during the tracing diffusion in every time piece $\eta k < t < \eta(k+1)$. Therefore, for $\eta k < t < \eta(k+1)$, the following differential inequality holds:

$$\frac{\partial R(\alpha,t)}{\partial t} \leq \frac{1}{\gamma}\frac{\alpha S_g^2}{4\sigma^2 n^2} - 2(1-\gamma)\sigma^2 c\left[\frac{R(\alpha,t)}{\alpha} + (\alpha-1)\frac{\partial R(\alpha,t)}{\partial \alpha}\right] \tag{57}$$

Let $a_1 = 2(1-\gamma)\sigma^2 c$, $a_2 = \frac{1}{\gamma}\frac{\alpha S_g^2}{4\sigma^2 n^2}$, and $y = log(\alpha - 1)$.

We define $u(t,y) = \begin{cases} \frac{R(\alpha, \lim_{t\to\eta k^+} t)}{\alpha} - \frac{a_2}{a_1} & \text{if } t = \eta k \\ \frac{R(\alpha,t)}{\alpha} - \frac{a_2}{a_1} & \text{if } \eta k < t < \eta(k+1) \end{cases}$, where we denote the limit privacy at start of a step with $R(\alpha, \lim_{t\to\eta k^+} t) = R_\alpha(\lim_{t\to\eta k^+}\Theta_t \| \lim_{t\to\eta k^+}\Theta'_t)$. Then we can include starting time $t = \eta k$ in the time piece for evolution of $u(t,y)$ and re-write (57) as the following:

$$\frac{\partial u}{\partial t} + a_1 u + a_1 \frac{\partial u}{\partial y} \leq 0, \quad \text{when } \eta k \leq t < \eta(k+1), \tag{58}$$

with initial condition $u(\eta k, y) = \frac{R(\alpha, \lim_{t\to\eta k^+} t)}{\alpha} - \frac{a_2}{a_1}$.

We introduce auxiliary variables $\tau = t$, and $z = t - \frac{1}{a_1}y$. By defining $v(\tau, z) = u(t,y)$, we get $\frac{\partial v}{\partial \tau} + a_1 v \leq 0$ from (58), with initial condition $v(\eta k, z) = u(\eta k, -a_1(z - \eta k))$. This PDI implies that for every $z$, the rate of decay of $v$ is proportional to its present value. The solution for this PDI is $v(\tau, z) \leq v(\eta k, z)e^{-a_1(\tau - \eta k)}$. By bringing back the original variables, we have

$$u(t,y) \leq u(\eta k, y - a_1(t - \eta k))e^{-a_1(t-\eta k)}, \quad \text{when } \eta k \leq t < \eta(k+1). \tag{59}$$

On undoing the substitution $u(t,y)$, we have

$$R(\alpha,t) - \frac{a_2}{a_1}\alpha \leq (R(\alpha, \lim_{t\to\eta k^+} t) - \frac{a_2}{a_1}\alpha)e^{-a_1(t-\eta k)}, \quad \text{when } \eta k \leq t < \eta(k+1). \tag{60}$$

On taking the limit $t \to \eta(k+1)^-$, we have

$$R(\alpha, \lim_{t\to\eta(k+1)^-} t) - \frac{a_2}{a_1}\alpha \leq (R(\alpha, \lim_{t\to\eta k^+} t) - \frac{a_2}{a_1}\alpha)e^{-a_1\eta}. \tag{61}$$

Meanwhile, the tracing diffusion expression (5) gives us

$$\lim_{t\to\eta k^+}\Theta_t = \phi(\Pi_{\mathcal{C}}(\lim_{t\to\eta k^-}\Theta_t)), \quad \text{and} \quad \lim_{t\to\eta k^+}\Theta'_t = \phi(\Pi_{\mathcal{C}}(\lim_{t\to\eta k^-}\Theta'_t)), \tag{62}$$

where $\phi(\theta) = \theta - \eta\cdot\frac{1}{2}(\mathcal{L}_D(\theta) + \mathcal{L}_{D'}(\theta))$ is a mapping on parameter set $\mathcal{C} \subseteq \mathbb{R}^d$. This mapping is the same for neighboring dataset $D$ and $D'$, because its definition only uses the average gradient between neighboring datasets $D$ and $D'$. Therefore by post-processing property of Rényi divergence, we have

$$R(\alpha, \lim_{t\to\eta k^+} t) \leq R(\alpha, \lim_{t\to\eta k^-} t). \tag{63}$$

Combining the above two inequalities, we immediately have the following recursive equation:

$$R(\alpha, \lim_{t\to\eta(k+1)^-} t) - \frac{a_2}{a_1}\alpha \leq (R(\alpha, \lim_{t\to\eta k^-} t) - \frac{a_2}{a_1}\alpha)e^{-a_1\eta} \tag{64}$$

Repeating this step for $k = 0, \cdots, K-1$, we have

$$R(\alpha, \lim_{t \to \eta K^-} t) - \frac{a_2}{a_1}\alpha \leq (R(\alpha, \lim_{t \to 0^-} t) - \frac{a_2}{a_1}\alpha)e^{-a_1 \eta K}. \tag{65}$$

Because coupled tracing diffusion have the same start parameter, we have $R(\alpha, \lim_{t\to 0^-} t) = 0$. Moreover, since projection is post-processing mapping, we have $R(\alpha, \eta K) \leq R(\alpha, \lim_{t \to \eta K^-} t)$. Therefore, taking the value $a_1 = 2(1-\gamma)\sigma^2 c$, $a_2 = \frac{1}{\gamma}\frac{S_g^2}{4\sigma^2 n^2}$ in (65), we have

$$R(\alpha, \eta K) \leq \frac{\alpha S_g^2}{8\gamma(1-\gamma)c\sigma^4 n^2}(1 - e^{-2(1-\gamma)\sigma^2 c \eta K}). \tag{66}$$

Setting $\gamma = \frac{1}{2}$ suffices to prove the Rényi privacy loss bound in the theorem. □

**Isoperimetry constants for noisy GD** To prove that LSI holds for the tracing diffusion for noisy GD, we first note that the diffusion process (5) can be written as composition of Lipschitz mapping and Gaussian noise for any $\eta k < t < \eta(k+1)$. Meanwhile, the projection at the end of a step is 1-Lipstchitz mapping. Then, we rely on the following two lemmas that show Lipschitz transformation and Gaussian perturbation of a probability distribution preserve its LSI property.

**Lemma 8** (LSI under Lipschitz transformation [15])**.** *Suppose a probability distribution $p$ on $\mathbb{R}^d$ satisfies LSI with constant $c > 0$. Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a differentiable and $L$-Lipschitz transformation. The push-forward distribution $T_{\#p}$, representing $T(\Theta)$ when $\Theta \sim p$, satisfies LSI with constant $\frac{c}{L^2}$.*

**Lemma 9** (LSI under Gaussian convolution [15])**.** *Suppose a probability distribution $p$ on $\mathbb{R}^d$ satisfies LSI with constant $c > 0$. For $t > 0$, the probability distribution $p * \mathcal{N}(0, 2t\mathbb{I}_d)$ satisfies LSI with constant $\left(\frac{1}{c} + 2t\right)^{-1}$. A special case of this is that $\mathcal{N}(0, 2t\mathbb{I}_d)$ satisfies LSI with constant $\frac{1}{2t}$.*

**Lemma 5.** *If loss function $\ell(\theta; \mathbf{x})$ is $\lambda$-strongly convex and $\beta$-smooth over a closed convex set $\mathcal{C}$, the step-size is $\eta < \frac{1}{\beta}$, and initial distribution is $\Theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, \frac{2\sigma^2}{\lambda}\mathbb{I}_d))$, then the coupled tracing diffusion processes $\{\Theta_t\}_{t\geq 0}$ and $\{\Theta'_t\}_{t\geq 0}$ for noisy GD on any neighboring datasets $D$ and $D'$ satisfy $c$-LSI for any $t \geq 0$ with $c = \frac{\lambda}{2\sigma^2}$.*

*Proof.* We only prove $c$-LSI for the tracing diffusion process $\{\Theta_t\}_{t\geq 0}$ on dataset $D$. The proof for $\{\Theta'_t\}_{t\geq 0}$ is similar.

For any $D \in \mathcal{X}^n$, and any $0 < s < \eta$, recall that the update step in tracing diffusion (5) equals the following random mapping:

$$\Theta_{\eta k + s} = \begin{cases} T_s(\Theta_{\eta k}) + \sqrt{2s\sigma^2}\mathbf{Z}, & \text{if } 0 \leq s < \eta \\ \Pi_{\mathcal{C}}(T_s(\Theta_{\eta k}) + \sqrt{2s\sigma^2}\mathbf{Z}), & \text{if } s = \eta \end{cases} \tag{67}$$

where the mapping $T_s(\theta) = \theta - \eta \cdot \frac{1}{2}\left(\nabla \mathcal{L}_D(\theta) + \nabla \mathcal{L}_{D'}(\theta)\right) - s \cdot \frac{1}{2}\left(\nabla \mathcal{L}_D(\theta) - \nabla \mathcal{L}_{D'}(\theta)\right)$. We first show that $T_s(\theta)$ is $(1-\eta\lambda)$-Lipschitz. For any $w, v \in \mathcal{C}$, we have

$$\begin{aligned} T_s(w) - T_s(v) &= w - v - \frac{\eta + s}{2}[\nabla \mathcal{L}_D(w) - \nabla \mathcal{L}_D(v)] - \frac{\eta - s}{2}[\nabla \mathcal{L}_{D'}(w) - \nabla \mathcal{L}_{D'}(v)] \\ &= w - v - \left[\frac{\eta + s}{2}\nabla^2 \mathcal{L}_D(z) + \frac{\eta - s}{2}\nabla^2 \mathcal{L}_{D'}(z')\right](w - v) \\ &\qquad \text{(for some } z, z' \in \mathcal{C} \text{ by the mid-value theorems)} \\ &= \left(I - \left[\frac{\eta + s}{2}\nabla^2 \mathcal{L}_D(z) + \frac{\eta - s}{2}\nabla^2 \mathcal{L}_{D'}(z')\right]\right)(w - v) \end{aligned}$$

By $\lambda$-strong convexity and $\beta$-smoothness of loss function $\ell(\theta; \mathbf{x})$ on $\mathcal{C}$, we prove that $\nabla^2 \mathcal{L}_D(z)$ and $\nabla^2 \mathcal{L}_{D'}(z')$ both have eigenvalues in the range $[\lambda, \beta]$. Since $s < \eta < \frac{1}{\beta}$, all eigenvalues of $I - \left[\frac{\eta+s}{2}\nabla^2 \mathcal{L}_D(z) + \frac{\eta-s}{2}\nabla^2 \mathcal{L}_{D'}(z')\right]$ is in $(0, 1-\eta\lambda]$. So, $T_s$ is $(1-\eta\lambda)$-Lipshitz.

Now, using induction we prove $p_t$ statisfies $c$-LSI for $c = \frac{\lambda}{2\sigma^2}$ for any $t \geq 0$.

**Base step:** Being a projection of Gaussian with variance $\frac{\lambda}{2\sigma^2}$ in every dimension, $\Theta_0$ satisfies $c$-LSI with the given constant by Lemma 8 (because projection is 1-Lipschitz) and Lemma 9.

**Induction step:** Suppose $\Theta_{\eta k}$ satisfies $c$-LSI with the above constant for some $k \in \mathbb{N}$. Distribution $\Theta_t$ for $\eta k < t < \eta(k+1)$ is same as $T_s$ pushover distribution plus gaussian noise distribution, i.e. $\Theta_t = \Theta_{\eta k \#T_s} * \mathcal{N}(0, 2s\sigma^2 \mathbb{I}_d)$ for $s = t - \eta k$. By using Lemma 8 and 9, we get $\left(\frac{c}{(1-\eta\lambda)^2 + 2s\sigma^2 c}\right)$-LSI for $\Theta_t$. Since $s < \eta < \frac{1}{\lambda}$, we have

$$(1 - s\lambda)^2 + 2s\sigma^2 c < 1 - s\lambda + 2s\sigma^2 c = 1.$$

Hence, for $\eta k < t < \eta(k+1)$, $\Theta_t$ satisfies $c'$-LSI with constant $c' > c$, which means it also satisfies $c$-LSI by definition.

By (67), $\Theta_{\eta(k+1)}$ undergoes an additional projection $\Pi_{\mathcal{C}}(\cdot)$. Since projection is a 1-Lipschitz map, by Lemma 8, it preserves $c$-LSI. So distribution $\Theta_{\eta(k+1)}$ also satisfies $c$-LSI. □

# E Proofs and discussions for Section 4: Tightness analysis

**Theorem 3.** *There exist two neighboring datasets $D, D' \in \mathcal{X}^n$, a start distribution $p_0$, and a smooth loss function $\ell(\theta; \mathbf{x})$ whose total gradient $g(\theta; D)$ has finite sensitivity $S_g$ on unconstrained convex set $\mathcal{C} = \mathbb{R}^d$, such that for any step-size $\eta < 1$, noise variance $\sigma^2 > 0$, and $K \in \mathbb{N}$, the Rényi privacy loss of $\mathcal{A}_{\text{Noisy-GD}}$ on $D, D'$ is lower-bounded by*

$$R_\alpha\left(\Theta_{\eta K} \middle\| \Theta'_{\eta K}\right) \geq \frac{\alpha S_g^2}{4\sigma^2 n^2}\left(1 - e^{-\eta K}\right). \tag{68}$$

*Proof.* We give lower bounds for the Rényi DP guarantee of noisy gradient descent algorithm for minimizing any smooth loss function $\ell(\theta; \mathbf{x})$ with finite total sensitivity $S_g$. We consider the following $L_2$-norm squared loss function with bounded data universe.

$$\ell(\theta; \mathbf{x}) = \frac{1}{2}\left\|\theta - \mathbf{x}\right\|_2^2, \text{where } \theta \in \mathbb{R}^d, \mathbf{x} \in \mathbb{R}^d \text{ and } \left\|\mathbf{x}\right\|_2 \leq \frac{S_g}{2}. \tag{69}$$

For any dataset $D = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ of size $n$, and any $\theta \in \mathbb{R}^d$, the loss is

$$\mathcal{L}_D(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\left\|\theta - \mathbf{x}_i\right\|_2^2.$$

It is easy to verify that $\mathcal{L}_D(\theta)$ is 1-smooth. The total gradient of $D$ is

$$g(\theta; D) = \sum_{\mathbf{x} \in D} \nabla \ell(\theta; \mathbf{x}) = n\theta - \sum_{\mathbf{x} \in D} \mathbf{x},$$

with a finite sensitivity $S_g$.

We construct the two neighboring datasets $D, D' \in \mathcal{X}^n$ such that $D = (\mathbf{x}_1, 0^d, \cdots, 0^d)$ and $D' = (\mathbf{x}'_1, 0^d, \cdots, 0^d)$, where $\mathbf{x}_1, \mathbf{x}'_1 \in \mathcal{X}$ are two records that are $S_g$ distance apart (i.e. $\left\|\mathbf{x}_1 - \mathbf{x}'_1\right\|_2 = S_g$).

Under dataset $D$, we can express the random variable $\Theta_{\eta K}$ at the $K$'th iteration of noisy GD using the following recursion with starting parameter $\Theta_0 = 0^d$.

$$\begin{aligned}
\Theta_{\eta K} =& (1-\eta)\Theta_{\eta(K-1)} + \eta\frac{\mathbf{x}_1}{n} + \sqrt{2\eta\sigma^2} \cdot \mathbf{Z}_{K-1} \\
=& (1-\eta)^K \Theta_0 + \eta \sum_{i=0}^{K-1} (1-\eta)^i \frac{\mathbf{x}_1}{n} + \sqrt{2\eta\sigma^2} \sum_{i=0}^{K-1} (1-\eta)^{K-1-i} \mathbf{Z}_i \\
=& \frac{\eta \mathbf{x}_1}{n} \sum_{i=0}^{K-1} (1-\eta)^i + \sqrt{2\eta\sigma^2 \sum_{i=0}^{K-1} (1-\eta)^{2i}} \cdot \mathbf{Z} \qquad \text{(where } \mathbf{Z}_i, \mathbf{Z} \sim \mathcal{N}(0, \mathbb{I}_d)\text{)}
\end{aligned}$$

A similar recursion can be used for $\Theta'_K$ in Noisy GD under dataset $D'$. Both $\Theta_K$ and $\Theta'_K$ are Gaussian random variables with variance $2\eta\sigma^2\sum_{i=0}^{K-1}(1-\eta)^{2i}$ in each dimension. Thus, we can calculate their exact divergence.

$$\begin{aligned}
R_\alpha\left(\Theta_{\eta K}\middle\|\Theta'_{\eta K}\right) &= \frac{\alpha\cdot\left\|\eta(\mathbf{x}_1-\mathbf{x}'_1)\sum_{i=0}^{K-1}(1-\eta)^i\right\|_2^2}{2\cdot 2\eta\sigma^2 n^2\sum_{i=0}^{K-1}(1-\eta)^{2i}} \\
&= \frac{\alpha\eta^2 S_g^2}{4\eta\sigma^2 n^2}\cdot\frac{\left(1-(1-\eta)^K\right)^2/\eta^2}{(1-(1-\eta)^{2K})/\left(\eta(2-\eta)\right)} \\
&= \frac{\alpha S_g^2}{4\sigma^2 n^2}\cdot\frac{2-\eta}{1+(1-\eta)^K}\left(1-(1-\eta)^K\right) \\
&\geq \frac{\alpha S_g^2}{4\sigma^2 n^2}(1-e^{-\eta K})
\end{aligned}$$

This inequality concludes the proof. □

**Corollary 2.** *Given $\ell_2$-norm squared loss function $\ell(\theta;\mathbf{x})=\frac{1}{2}\|\theta-\mathbf{x}\|_2^2$ on unconstrained convex set $\mathcal{C}=\mathbb{R}^d$ and bounded data domain with range $S_g$, and initial parameter $\theta_0=0^d$, for any two neighboring datasets $D, D'\in\mathcal{X}^n$, step-size $\eta$, noise variance $\sigma^2$, and $K\in\mathbb{N}$, the Rényi privacy loss of $\mathcal{A}_{\text{Noisy-GD}}$ on $D, D'$ is upper-bounded by*

$$R_\alpha\left(\Theta_{\eta K}\middle\|\Theta'_{\eta K}\right)\leq\frac{\alpha S_g^2}{(2-\eta)\sigma^2 n^2}(1-e^{-\frac{2-\eta}{2}\eta K}). \tag{70}$$

*Proof.* To use Theorem 2, we still need to verify $c$-LSI for the tracing diffusion on $\ell_2$-norm squared loss.

We use the explicit expression for tracing diffusion proved in Theorem 3 to prove $c$-LSI. We utilize the fact that $\Theta_{\eta K}$, the tracing diffusion for $L_2$-norm squared loss at discrete update time $\eta K$, is Gaussian with bounded variance $2\eta\sigma^2\sum_{i=0}^{K-1}(1-\eta)^{2i}\leq\frac{2\sigma^2}{2-\eta}$ in each dimension. Therefore, based on Lemma 9, which shows the LSI properties of Gaussian distributions, $\Theta_{K\eta}$ satisfies $c$-LSI with $c=\frac{2-\eta}{2\sigma^2}$. Similarly, by computing the explicit expression for tracing diffusion at time $\eta k<t<\eta(k+1)$, one can verify $\Theta_t$ satisfies $c$-LSI.

Now, we can directly use Theorem 2 to derive an upper-bound for RDP for Noisy GD under $L_2$-squared norm loss. □

**Discussion about tightness results** Figure 2 shows the gap between this lower bound and our RDP guarantee derived by Corollary 2, under small step-size $\eta=0.02$. The upper bound is roughly two times larger than the lower bound, which shows tightness of our privacy guarantee up to a rough constant of two. As comparison, we compute and plot the composition-based bound, which grows as fast as the lower bound in early iterations, but linearly grows above the lower bound, and our RDP guarantee, as $K$ increases to $\Omega(\frac{1}{\eta})\approx 100\ll n=5000$. Moreover, the larger the RDP order $\alpha$ is, the smaller the required number of iterations $K$ is for our RDP guarantee to be superior to the composition-based privacy bound.

**Gap between our upper bound and lower bound** There is a gap between the exponent and constant of our privacy upper bound Corollary 2 and the lower bound Theorem 3. We analyze the gap as follows.

1. **The gap in exponent:** There is a $\frac{2-\eta}{2}$ multiplicative gap between the exponent of our privacy upper bound and the lower bound. In hindsight, this is because discretized noisy GD converges to a biased stationary distribution. Therefore, our LSI constant bound $c=\frac{2-\eta}{2\sigma^2}$ depends on the discretization bias caused by step-size $\eta$, thus causing the exponent gap in our privacy bound.
2. **The gap in constant:** Our upper bound is larger than the lower bound by roughly a multiplicative constant of two. This is due to the **balancing ratio** $\gamma>0$ in Lemma 2 for bounding the rate of privacy loss.

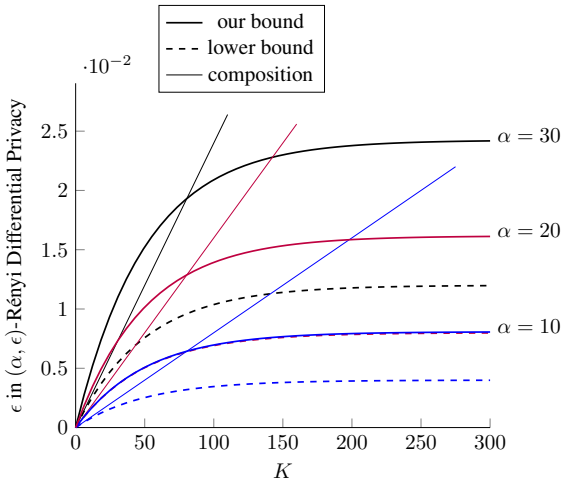


Figure 2: Tightness analysis of our RDP guarantee for the noisy GD algorithm. We show the changes of $\alpha$-RDP guarantee computed using Corollary 2, over $K$ iterations (number of full passes over the dataset) versus the lower-bounds (dashed lines) which are computed using Theorem 3. The loss function is the $\ell_2$-norm squared function (69), noise standard deviation is $\sigma = 0.02$, the step size is $\eta = 0.02$, the size of the dataset is $n = 5000$, and the finite $\ell_2$-sensitivity for total gradient is $S_g = 4$. The expression for computing the privacy loss in Baseline composition-based analysis (derived by moment accountant [1] with details in Appendix E) is: $\varepsilon = \frac{\alpha S_g^2}{4n^2\sigma^2} \cdot \eta K$

(a) **At the start of Noisy GD:** setting $\gamma = 1$ in (10) results in a smaller privacy loss rate bound. This is because, at the start of noisy GD, the accumulated privacy loss $R_\alpha(\Theta_t \| \Theta_t')$ is small, thus leading to a small second term $I_\alpha / E_\alpha$ in (10), by Lemma 3. Setting $\gamma = 1$ reduces the coefficient $\frac{1}{\gamma}$ for the dominating first term of (10), at a small cost of increasing the coefficient for the smaller second term $I_\alpha / E_\alpha$. This facilitates a smaller privacy loss rate bound, and is reflected in the similar growth of composition bound (equivalent to setting $\gamma = 1$) and our lower bound in Figure 2.

(b) **As Noisy GD converges:** setting $\gamma \to 0$ in (10) results in a smaller privacy loss rate bound. This is because, at convergence, the accumulated privacy loss $R_\alpha(\Theta_t \| \Theta_t')$ is larger, thus leading to more significant second term $I/E$ in (10). Setting $\gamma \to 0$ in (10) reduces the coefficient $-(1-\gamma)$ for the dominating second term $I/E$, thus facilitate a smaller bound for the privacy loss rate.

3. In our proof for Theorem 2, we set $\gamma = \frac{1}{2}$ to **balance** privacy loss rate estimates at the start and convergence of noisy GD, thus obtaining the smallest privacy bound at convergence, as shown in the proof.

**Derivation for Baseline composition-based privacy bound** Abadi et al. [1] introduce the moments accountant $\alpha(\lambda)$ for noisy SGD in Eq (2) of their paper, which effectively tracks the scaled Renyi divergence between processes. Therefore in Figure 1, we plot moment accountant bound in Abadi et al. [1] as baseline composition privacy analysis.

1. We first use **moments bound on the Gaussian mechanism** (following Lemma 3 in Abadi et al. [1] ) to bound the log moment $\alpha_{\mathcal{M}}(\lambda)$ of data-sensitive computation one update $M : \mathcal{M}(D) = \frac{\eta}{n} \sum_{x_i \in D} \nabla \ell(\theta; x_i) + \mathcal{N}(0, 2\eta\sigma^2 \mathbb{I}_d)$ in our Algorithm 1.

   By Eq (2) in Abadi et al. [1], and that $M(D), M(D')$ are Gaussian distributions (with variance $2\eta\sigma^2$ in every dimension and means at most $\frac{\eta}{n} S_g$ apart in $\ell_2$ norm), we bound $\alpha_{\mathcal{M}}(\lambda) \leq \frac{\lambda(\lambda+1)\eta S_g^2}{4n^2\sigma^2}$.

2. We then **compose log moment bound for $K$ iterations** by Theorem 2 [Composibility] of log moement bound in Abadi et al. [1]., and we obtain $\alpha(\lambda) \leq K \cdot \alpha_{\mathcal{M}}(\lambda) = \frac{K\lambda(\lambda+1)\eta S_g^2}{4n^2\sigma^2}$.

3. Finally by definition of log moment (Eq (2) of Abadi et al. [1]) and Renyi divergence ((1) in our paper), we take $\lambda \leftarrow \alpha - 1$ and $R_\alpha(\Theta_K \| \Theta_K) \leftarrow \frac{\alpha(\lambda)}{\lambda}$, and obtain the **baseline composition privacy bound** $\epsilon = \frac{\alpha S_g^2}{4n^2\sigma^2} \cdot \eta K$ from the log moment bound. We use this expression in Figure 1 and 2.

# F Proofs for Section 5: Utility analysis

**Theorem 4.** *For Lipschitz smooth strongly convex loss function $\ell(\theta; \mathbf{x})$ on a bounded closed convex set $\mathcal{C} \subseteq \mathbb{R}^d$, and dataset $D \in \mathcal{X}^n$ of size $n$, if the step-size $\eta = \frac{\lambda}{2\beta^2}$ and the initial parameter $\theta_0 \sim \Pi_{\mathcal{C}}(\mathcal{N}(0, \frac{2\sigma^2}{\lambda}\mathbb{I}_d))$, then the noisy GD Algorithm 1 is $(\alpha, \varepsilon')$-Rényi differentially private, where $\alpha > 1$ and $\varepsilon' > 0$, and satisfies*

$$\mathbb{E}[\mathcal{L}_D(\theta_{K^*}) - \mathcal{L}_D(\theta^*)] = O(\frac{\alpha\beta d L^2}{\varepsilon' \lambda^2 n^2}), \tag{71}$$

*by setting noise variance $\sigma^2 = \frac{4\alpha L^2}{\lambda \varepsilon' n^2}$, and number of updates $K^* = \frac{2\beta^2}{\lambda^2}\log(\frac{n^2\varepsilon'}{\alpha d})$.*

*Equivalently, for $\varepsilon \leq 2\log(1/\delta)$ and $\delta > 0$, Algorithm 1 is $(\varepsilon, \delta)$-differentially private, and satisfies*

$$\mathbb{E}[\mathcal{L}_D(\theta_{K^*}) - \mathcal{L}_D(\theta^*)] = O(\frac{\beta d L^2 \log(1/\delta)}{\epsilon^2 \lambda^2 n^2}), \tag{72}$$

*by setting noise variance $\sigma^2 = \frac{8L^2(\varepsilon + 2\log(1/\delta))}{\lambda \varepsilon^2 n^2}$, and number of updates $K^* = \frac{2\beta^2}{\lambda^2}\log(\frac{n^2\varepsilon^2}{4\log(1/\delta)d})$.*

*Proof.* From Lemma 6, we have

$$\mathbb{E}[\mathcal{L}_D(\theta_K) - \mathcal{L}_D(\theta^*)] \leq \frac{2\beta L^2}{\lambda^2} e^{-\lambda\eta K} + \frac{2\beta d\sigma^2}{\lambda}. \tag{73}$$

Since $\eta = \frac{\lambda}{2\beta^2} \leq \frac{1}{\beta}$, by Corollary 1, the noisy GD with $K$ iterations will be $(\alpha, \varepsilon')$-RDP as long as $\sigma^2 \geq \frac{4\alpha L^2}{\lambda\varepsilon' n^2}(1 - e^{-\lambda\eta K/2})$. Therefore, if we set $\sigma^2 = \frac{4\alpha L^2}{\lambda\varepsilon' n^2}$, noisy GD is $(\alpha, \varepsilon')$-RDP for any $K$. On substituting this noise variance in (73), we get

$$\mathbb{E}[\mathcal{L}_D(\theta_K) - \mathcal{L}_D(\theta^*)] \leq \frac{2\beta L^2}{\lambda^2} e^{-\lambda\eta K} + \frac{8\alpha L^2 \beta d}{\lambda^2 \varepsilon' n^2}. \tag{74}$$

By setting $K^* = \frac{1}{\lambda\eta}\log(\frac{\varepsilon' n^2}{\alpha d}) = \frac{2\beta^2}{\lambda^2}\log(\frac{\varepsilon' n^2}{\alpha d})$, we can control the empirical risk to be

$$\mathbb{E}[\mathcal{L}_D(\theta_{K^*}) - \mathcal{L}_D(\theta^*)] \leq \frac{10\alpha L^2 \beta d}{\lambda^2 \varepsilon' n^2}. \tag{75}$$

Now, we convert the optimal excess risk guarantee under an $(\alpha, \varepsilon')$ RDP constraint to an optimal excess risk guarantee under $(\varepsilon, \delta)$ DP constraint. Let $\varepsilon > 0$ and $0 < \delta < 1$ be two constants such that $\varepsilon \leq 2\log(1/\delta)$. As per DP transition Theorem 7, $(\alpha, \varepsilon')$-RDP implies $(\varepsilon, \delta)$-DP for $\alpha = 1 + \frac{2}{\varepsilon}\log(1/\delta)$ and $\varepsilon' = \frac{\varepsilon}{2}$. By using this conversion, we bound (75) in terms of DP parameters as

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_D(\theta_{K^*}) - \mathcal{L}_D(\theta^*)] &\leq \frac{10L^2\beta d}{\lambda^2 n^2}\frac{\alpha}{\varepsilon'} \\
&= \frac{10L^2\beta d}{\lambda^2 n^2}\frac{1 + \frac{2}{\varepsilon}\log(1/\delta)}{\frac{\varepsilon}{2}} \\
\because \varepsilon \leq 2\log(1/\delta) \quad &\leq \frac{10L^2\beta d}{\lambda^2 n^2}\frac{8\log(1/\delta)}{\varepsilon^2}.
\end{aligned}$$

The amount of noise needed in terms of DP parameters is

$$\begin{aligned}
\sigma^2 &= \frac{4L^2}{\lambda n^2}\frac{\alpha}{\varepsilon'} \\
&= \frac{4L^2}{\lambda n^2} \cdot \frac{1 + \frac{2}{\varepsilon}\log(1/\delta)}{\frac{\varepsilon}{2}}
\end{aligned}$$

The optimal number of updates $K^*$ in terms of DP parameters is bounded as

$$\begin{aligned} K^* &= \frac{2\beta^2}{\lambda^2}\log(\frac{n^2}{d}\cdot\frac{\varepsilon'}{\alpha}) \\ &= \frac{2\beta^2}{\lambda^2}\log(\frac{n^2}{d}\cdot\frac{\frac{\varepsilon}{2}}{1+\frac{2}{\varepsilon}\log(1/\delta)}) \\ &\leq \frac{2\beta^2}{\lambda^2}\log(\frac{n^2}{d}\cdot\frac{\varepsilon^2}{4\log(1/\delta)}). \end{aligned}$$

□

**Lemma 6.** *For L-Lipschitz, $\lambda$-strongly convex and $\beta$-smooth loss function $\ell(\theta;\mathbf{x})$ over a closed convex set $\mathcal{C}\subseteq\mathbb{R}^d$, step-size $\eta\leq\frac{\lambda}{2\beta^2}$, and start parameter $\theta_0\sim\Pi_{\mathcal{C}}(\mathcal{N}(0,\frac{2\sigma^2}{\lambda}\mathbb{I}_d))$, the excess empirical risk of Algorithm 1 is bounded by*

$$\mathbb{E}[\mathcal{L}_D(\theta_K)-\mathcal{L}_D(\theta^*)]\leq\frac{2\beta L^2}{\lambda^2}e^{-\lambda\eta K}+\frac{2\beta d\sigma^2}{\lambda}, \tag{76}$$

*where $\theta^*$ is the minimizer of $\mathcal{L}_D(\theta)$ in the relative interior of convex set $\mathcal{C}$, and $\mathrm{d}$ is the dimension of parameter.*

*Proof.* By the noisy GD update equation we have

$$\theta_{k+1}=\Pi_{\mathcal{C}}(\theta_k-\eta\nabla\mathcal{L}_D(\theta_k)+\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbb{I}_d)). \tag{77}$$

From the definition of projection $\Pi_{\mathcal{C}}(\cdot)$, we have:

$$\begin{aligned} \Pi_{\mathcal{C}}(\theta^*-\eta\nabla\mathcal{L}_D(\theta^*)) &= \arg\min_{\theta\in\mathcal{C}}\|\theta-\theta^*+\eta\nabla\mathcal{L}_D(\theta^*)\|_2^2 \\ &= \arg\min_{\theta\in\mathcal{C}}\|\theta-\theta^*\|_2^2+2\eta\langle\theta-\theta^*,\nabla\mathcal{L}_D(\theta^*)\rangle+\eta^2\|\nabla\mathcal{L}_D(\theta^*)\|_2^2 \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(by optimality of } \theta^* \text{ in } \mathcal{C}) \\ &= \arg\min_{\theta\in\mathcal{C}}\|\theta-\theta^*\|_2^2+\eta^2\|\nabla\mathcal{L}_D(\theta^*)\|_2^2 \\ &= \theta^* \end{aligned}$$

Therefore, by combining the above two, and from contractivity of projection $\Pi_{\mathcal{C}}(\cdot)$ [8, Proposition 17] we have

$$\begin{aligned} \|\theta_{k+1}-\theta^*\|_2^2 \leq& \|\theta_k-\eta\nabla\mathcal{L}_D(\theta_k)+\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbb{I}_d)-(\theta^*-\eta\nabla\mathcal{L}_D(\theta^*))\|_2^2 \\ =& \|\theta_k-\theta^*\|_2^2+\eta^2\|\nabla\mathcal{L}_D(\theta_k)-\nabla\mathcal{L}_D(\theta^*)\|_2^2+2\eta\sigma^2\|\mathcal{N}(0,\mathbb{I}_d)\|_2^2 \\ &+2\langle\theta_k-\theta^*,\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbb{I}_d)\rangle-2\eta\langle\nabla\mathcal{L}_D(\theta_k)-\nabla\mathcal{L}_D(\theta^*),\sqrt{2\eta\sigma^2}\mathcal{N}(0,\mathbb{I}_d)\rangle \\ &-2\eta\langle\theta_k-\theta^*,\nabla\mathcal{L}_D(\theta_k)-\nabla\mathcal{L}_D(\theta^*)\rangle. \end{aligned}$$

By $\beta$-smoothness of $\mathcal{L}_D$ and $\eta=\frac{\lambda}{2\beta^2}$, we have

$$\eta^2\|\nabla\mathcal{L}_D(\theta_k)-\nabla\mathcal{L}_D(\theta^*)\|_2^2\leq\eta\lambda\|\theta_k-\theta^*\|_2^2. \tag{78}$$

By strong convexity of $\mathcal{L}_D$, we have

$$\begin{aligned} \mathbb{E}[\langle\nabla\mathcal{L}_D(\theta_k),\theta_k-\theta^*\rangle] &\geq \mathbb{E}[\mathcal{L}_D(\theta_k)-\mathcal{L}_D(\theta^*)]+\frac{\lambda}{2}\mathbb{E}[\|\theta_k-\theta^*\|^2] \\ &\geq \frac{\lambda}{2}\mathbb{E}[\|\theta_k-\theta^*\|^2]+\frac{\lambda}{2}\mathbb{E}[\|\theta_k-\theta^*\|^2] \\ &\geq \lambda\mathbb{E}[\|\theta_k-\theta^*\|^2]. \end{aligned}$$

By taking expectations on the controlling inequality, and plugging the above results, we get

$$\mathbb{E}[\|\theta_{k+1}-\theta^*\|_2^2]\leq(1-\lambda\eta)\mathbb{E}[\|\theta_k-\theta^*\|_2^2]+2\eta\sigma^2 d. \tag{79}$$

By $\beta$-smoothness,

$$\mathcal{L}_D(\theta_k) - \mathcal{L}_D(\theta^*) \leq \langle \nabla \mathcal{L}_D(\theta^*), \theta_k - \theta^* \rangle + \frac{\beta}{2} \|\theta_k - \theta^*\|_2^2.$$

By the optimality of $\theta^*$ in the relative interior of convex set $\mathcal{C}$ and the fact that $\theta_K \in \mathcal{C}$, we prove

$$\langle \nabla \mathcal{L}_D(\theta^*), \theta_K - \theta^* \rangle = 0.$$

Therefore, $\mathcal{L}_D(\theta_K) - \mathcal{L}_D(\theta^*) \leq \frac{\beta}{2} \|\theta_K - \theta^*\|_2^2$. On taking expectation over $\theta_K$, we have

$$\mathbb{E}[\mathcal{L}_D(\theta_K) - \mathcal{L}_D(\theta^*)] \leq \frac{\beta}{2} \mathbb{E}[\|\theta_K - \theta^*\|^2].$$

On unrolling the recursion in (79), we have

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_D(\theta_K) - \mathcal{L}_D(\theta^*)] \leq & \frac{\beta}{2}(1 - \eta\lambda)^K \mathbb{E}[\|\theta_0 - \theta^*\|_2^2] + 2\beta d\sigma^2 \sum_{k=0}^{K-1} (1 - \eta\lambda)^k \\
\leq & \frac{\beta}{2} e^{-\lambda\eta K} \mathbb{E}[\|\theta_0 - \theta^*\|_2^2] + \frac{2\beta d\sigma^2}{\lambda}.
\end{aligned}$$

Since we always have $\|\mathcal{C}\|_2 \leq 2L/\lambda$, we can bound $\mathbb{E}[\|\theta_0 - \theta^*\|_2^2] \leq \frac{4L^2}{\lambda^2}$ as both $\theta_0, \theta^* \in \mathcal{C}$. Therefore, we have

$$\mathbb{E}[\mathcal{L}_D(\theta_K) - \mathcal{L}_D(\theta^*)] \leq \frac{2\beta L^2}{\lambda^2} e^{-\lambda\eta K} + \frac{2\beta d\sigma^2}{\lambda}.$$

□