Differentially Private Stochastic Convex Optimization in (Non)-Euclidean Space Revisited (Supplementary material)

Jinyan Su[1] **Changhong Zhao**[2] **Di Wang**[3,4,5]

[1]Mohamed bin Zayed University of Artificial Intelligence
[2]Department of Information Engineering,, The Chinese University of Hong Kong
[3]Provable Responsible AI and Data Analytics Lab
[4]Computational Bioscience Research Center
[5]Division of CEMSE, King Abdullah University of Science and Technology

7 NOTATION SUMMARY

$\mathcal{C}$: constraint set	$G_{\mathcal{C}}$: Gaussian width of set $\mathcal{C}$
d: dimension	n: sample size
ϵ, δ: privacy parameters	ℓ: convex loss funtion
L: Lipschitz constant	β: smoothness constant
λ: regularization parameter	α: optimization accurancy
ℓ_p^d: Normed space corresponds to $\|\cdot\|_p$, where $\|x\|_p = (\sum_{j=1}^d \|x_j\|^p)^{1/p}$	$\mathcal{L}(\theta)$: population risk
$\hat{\mathcal{L}}(\theta, D)$: empirical risk	κ: κ-regular space
$\|\|\cdot\|\|_{\mathcal{C}}$: Minkowski norm, $\|\cdot\|_{\mathcal{C}} = \min\{r \in \mathbb{R}^+ : v \in r\mathcal{C}\}$	$\|\cdot\|_{\mathcal{C}^*}$: dual norm of $\|\cdot\|_{\mathcal{C}}$
σ: the variance of Gaussian noise	$\|\cdot\|_+$: the smooth norm for $(\mathbf{E}, \|\cdot\|_*)$

Table 3: Notation summary of the paper.

8 OMITTED PROOFS IN SECTION 4

8.1 PROOF OF THEOREM 1

Algorithm 5 $\mathcal{A}_{\text{ObjP}}$: Objective perturbation

1: **Input:** Dateset D, loss function ℓ, regularization parameter λ.
2: Sample $\mathbf{G} \sim \mathcal{N}(0, \sigma_1^2 \mathbb{I}_d)$ where $\sigma_1^2 = \frac{32L^2 \log(1/\delta)}{\epsilon^2}$. Set $\lambda \geq \frac{r\beta}{2\epsilon n}$, where $r = \min\{d, 2 \cdot \text{rank}(\nabla^2 \ell(\theta, x))\}$ with $\text{rank}(\nabla^2 \ell(\theta, x))$ being the maximal rank of the Hessian of ℓ for all $\theta \in \mathcal{C}$ and $x \sim \mathcal{P}$.
3: Let $\mathcal{J}(\theta, D) = \hat{\mathcal{L}} + \frac{\langle \mathbf{G}, \theta \rangle}{n} + \lambda ||\theta||_2^2$.
4: **return** $\theta_1 = \arg\min_{\theta \in \mathcal{C}} \mathcal{J}(\theta, D)$.

Proof. Let $\theta_1 = \arg\min_{\theta \in \mathcal{C}} \mathcal{J}(\theta, D)$, where $\mathcal{J}(\theta, D) = \hat{\mathcal{L}}(\theta, D) + \frac{\langle \mathbf{G}, w \rangle}{n} + \lambda ||\theta||_2^2$. Let $\theta_2 = \mathcal{O}(\mathcal{J}, \alpha)$ where $\mathcal{O}$ is the optimizer defined in the algorithm. Notice that one can compute $\hat{\theta}$ from tuple $(\theta_1, \theta_2 - \theta_1 + \mathbf{H})$ by simple post-processing. Furthermore, the algorithm that outputs θ_1 is (ϵ, δ)-DP by the following theorem.

Lemma 5 (Theorem 1 in [Iyengar et al., 2019])**.** Suppose Assumption 1 holds and that the smoothness parameter satisfy $\beta \leq \frac{\epsilon n \lambda}{r}$, the algorithm $\mathcal{A}_{\text{ObjP}}$ (Algorithm 5) that outputs $\theta_1 = \arg\min_{\theta \in \mathcal{C}} \mathcal{J}(\theta, D)$ is (ϵ, δ)-DP.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Next, we will bound the term $||\theta_2 - \theta_1||$ to make $(\theta_2 - \theta_1 + \mathbf{H})$ differentially private, conditioned on θ_1. As $\mathcal{J}(\theta, D)$ is λ-strongly convex, we have $\mathcal{J}(\theta_2, D) \geq \mathcal{J}(\theta_1, D) + \frac{\lambda}{2}||\theta_2 - \theta_1||_2^2$, which implies that

$$||\theta_2 - \theta_1||_2 \leq \sqrt{\frac{2}{\lambda}(\mathcal{J}(\theta_2, D) - \mathcal{J}(\theta_1, D))} \leq \sqrt{\frac{2\alpha}{\lambda}}. \tag{2}$$

Thus, conditioned on θ_1, $\theta_2 - \theta_1$ has the l_2 sensitivity of $\sqrt{\frac{8\alpha}{\lambda}}$. Therefore, $(\theta_2 - \theta_1) + \mathbf{H}$ is $(\epsilon/2, \delta.2)$-DP. By the standard composition in Dwork et al. [2014], the tuple $(\theta_1, \theta_2 - \theta_1 + \mathbf{H})$ satisfies (ϵ, δ)-DP and hence $\hat{\theta}$ satisfies (ϵ, δ)-DP. □

8.2 PROOF OF THEOREM 2

Proof. Let θ_1 be the exact minimizer of $\mathcal{J}(\theta, D)$. We split the objective $\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta^*)$ into two parts and bound them separately.

$$\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta^*) = \mathbb{E}[\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta_1)] + \mathbb{E}[\mathcal{L}(\theta_1)] - \mathcal{L}(\theta^*). \tag{3}$$

In the following, we bound the term $\mathbb{E}[\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta_1)]$ and the term $\mathbb{E}[\mathcal{L}(\theta_1)] - \mathcal{L}(\theta^*)$ separately. To bound the term $\mathbb{E}[\mathcal{L}(\theta_1)] - \mathcal{L}(\theta^*)$, we need the following two lemmas. The first lemma states the excess empirical risk of θ_1 while the second lemma states the stability property of regularized empirical risk minimization.

Lemma 6. (Excess empirical loss of θ_1 in $\mathcal{A}_{\text{ObjP}}$). Let $D \sim \mathcal{P}^n$, under Assumption 1, the excess empirical loss of θ_1 satisfies

$$\mathbb{E}[\hat{\mathcal{L}}(\theta_1, D)] - \min_{\theta \in \mathcal{C}} \hat{\mathcal{L}}(\theta, D) \leq O\left(\frac{LG_{\mathcal{C}}\sqrt{\log(1/\delta)}}{\epsilon n} + \lambda||\mathcal{C}||_2^2\right), \tag{4}$$

where the expectation is taken over the randomness induced by Gaussian noise.

Lemma 7. [[Shalev-Shwartz and Ben-David, 2014]] Let $f : \mathcal{C} \times D \to \mathbb{R}$ be a convex, ρ-Lipschitz loss function where $D = \{x_1, \cdots, x_n\} \sim \mathcal{P}^n$. Let $\mathcal{A}$ be an algorithm that outputs $\tilde{\theta} = \arg\min_{\theta \in \mathcal{C}}\{\hat{F}(\theta, D) + \lambda||\theta||^2\}$ with $\lambda > 0$ where $\hat{F}(\theta, D) = \frac{1}{n}\sum_{i=1}^n f(\theta, x_i)$, then $\mathcal{A}$ is $\frac{2\rho^2}{\lambda n}$-uniformly stable, i.e., for all neighboring datasets $D \sim D'$ we have

$$\sup_z |\mathbb{E}[f(\mathcal{A}(D), z) - f(\mathcal{A}(D'), z)]| \leq \frac{2\rho^2}{\lambda n}.$$

The property of uniform stability is described by the following lemma.

Lemma 8 (Bousquet and Elisseeff [2002])**.** Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{C}$ be an α-uniformly stable algorithm w.r.t. loss $\ell : \mathcal{C} \times \mathcal{X} \to \mathbb{R}$. Let $D \sim \mathcal{P}^n$ where $\mathcal{P}$ is the distribution over $\mathcal{X}$. Then,

$$\mathop{\mathbb{E}}_{D \sim \mathcal{P}^n, \mathcal{A}}[\mathcal{L}(\mathcal{A}(D)) - \hat{\mathcal{L}}(\mathcal{A}(D), D)] \leq \alpha.$$

Now we begin to bound the term $\mathcal{L}(\theta_1) - \mathcal{L}(\theta^*)$ using the above three lemmas. Fix any realization of the noise vector $\mathbf{G}$, we define $f_{\mathbf{G}}(\theta, x) = \ell(\theta, x) + \frac{\langle \mathbf{G}, \theta \rangle}{n}$, then $f_{\mathbf{G}}$ is $\left(L + \frac{||\mathbf{G}||_2}{n}\right)$-Lipschitz.

Define $\hat{F}_{\mathbf{G}}(\theta, D) = \frac{1}{n}\sum_{i=1}^n f_{\mathbf{G}}(\theta, x_i)$, and we have $\theta_1 = \arg\min_{\theta \in \mathcal{C}} \hat{F}_{\mathbf{G}}(\theta, D) + \lambda||\theta||_2^2$, so from Lemma 7, the algorithm that outputs θ_1 is $\frac{2\left(L + \frac{||\mathbf{G}||_2}{n}\right)^2}{\lambda n}$-uniformly stable. Denote $F_{\mathbf{G}}(\theta) = \mathop{\mathbb{E}}_{x \sim \mathcal{P}}[f_{\mathbf{G}}(\theta, x)]$, according to Lemma 8, we have

$$\mathop{\mathbb{E}}_{D \sim \mathcal{P}^n}[\mathcal{L}(\theta) - \hat{\mathcal{L}}(\theta, D)] = \mathop{\mathbb{E}}_{D \sim \mathcal{P}^n}[F_{\mathbf{G}}(\theta) - \hat{F}_{\mathbf{G}}(\theta, D)] \leq \frac{2\left(L + \frac{||\mathbf{G}||_2}{n}\right)^2}{\lambda n}.$$

Take the expectation w.r.t. $\mathbf{G} \sim \mathcal{N}(0, \frac{32L^2\log(1/\delta)}{\epsilon^2}\mathbb{I}_d)$ as well, we get

$$\mathbb{E}[\mathcal{L}(\theta) - \hat{\mathcal{L}}(\theta, D)] \leq O\left(\frac{L^2 \cdot \left(1 + \frac{\sqrt{d\log(1/\delta)}}{\epsilon n}\right)^2}{\lambda n}\right) \leq O\left(\frac{L^2}{\lambda n}\right), \tag{5}$$

where we assume $n \geq O(\frac{\sqrt{d\log(1/\delta)}}{\epsilon})$.

Thus

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\theta_1)] - \mathcal{L}(\theta^*) &= \mathbb{E}[\mathcal{L}(\theta_1)] - \min_{\theta\in\mathcal{C}}\mathcal{L}(\theta) \\ &\leq \mathbb{E}[\hat{\mathcal{L}}(\theta_1, D) - \min_{\theta\in\mathcal{C}}\hat{\mathcal{L}}(\theta, D)] + \mathbb{E}[\mathcal{L}(\theta_1) - \hat{\mathcal{L}}(\theta_1, D)] \\ &\leq O\left(\frac{L\cdot G_{\mathcal{C}}\cdot\sqrt{\log(1/\delta)}}{\epsilon n} + \lambda||\mathcal{C}||_2^2 + \frac{L^2}{\lambda n}\right),\end{aligned} \tag{6}$$

where we use the fact that $\mathop{\mathbb{E}}_{D\sim\mathcal{P}^n}[\min_{\theta\in\mathcal{C}}\hat{\mathcal{L}}(\theta, D)] \leq \min_{\theta\in\mathcal{C}}\mathop{\mathbb{E}}_{D\sim\mathcal{P}^n}[\hat{\mathcal{L}}(\theta, D)] = \min_{\theta\in\mathcal{C}}\mathcal{L}(\theta)$ and the last bound is directly from Eq.(4) and Eq.(5).

Now we bound the term $\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta_1)$. Recall that $\theta_2 = \mathcal{O}(\mathcal{J}, \alpha)$ and

$$\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta_1) = \mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta_2) + \mathcal{L}(\theta_2) - \mathcal{L}(\theta_1).$$

Note the term $\mathcal{L}(\theta_2) - \mathcal{L}(\theta_1) \leq L\cdot||\theta_1 - \theta_2||_2 \leq L\cdot\sqrt{\frac{2\alpha}{\lambda}}$ (From Eq.(2)), and the term $\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta_2) \leq L\cdot\mathbb{E}[||\hat{\theta} - \theta_2||_2]$.

Also note that $\hat{\theta} = \text{Proj}_{\mathcal{C}}(\theta_2 + \mathbf{H})$. Let q be the line through θ_2 and $\hat{\theta}$, and let p be the projection of $\theta_3 = \theta_2 + \mathbf{H}$ onto q. The key observation is that p lies on the ray from $\hat{\theta}$ to infinity otherwise p will be a point in $\mathcal{C}$ that is closer to θ_3 than $\hat{\theta}$. Thus we have

$$\begin{aligned}\mathbb{E}[||\hat{\theta} - \theta_2||_2^2] &= \mathbb{E}[\langle\hat{\theta} - \theta_2, \hat{\theta} - \theta_2\rangle] \\ &\leq \mathbb{E}[\langle\hat{\theta} - \theta_2, \theta_3 - \theta_2] \\ &= \mathbb{E}[\langle\mathbf{H}, \hat{\theta} - \theta_2\rangle] \\ &\leq 2\cdot\max_{\theta\in\mathcal{C}}\mathbb{E}[\langle\mathbf{H}, \theta\rangle] \\ &\leq O(\mathbb{E}[\max|\langle\mathbf{H}, \theta\rangle|]) \\ &= O\left(\sqrt{\frac{\alpha\log(1/\delta)}{\lambda}}\cdot\frac{G_{\mathcal{C}}}{\epsilon}\right),\end{aligned}$$

where the last equation is from the definition of Gaussian width.

So we have

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta_1) &\leq L\cdot\sqrt{\frac{2\alpha}{\lambda}} + L\cdot\mathbb{E}[||\hat{\theta} - \theta_2||_2] \\ &\leq O\left(L\cdot\sqrt[4]{\frac{\alpha\log(1/\delta)}{\lambda}}\cdot\sqrt{\frac{G_{\mathcal{C}}}{\epsilon}} + L\sqrt{\frac{\alpha}{\lambda}}\right).\end{aligned} \tag{7}$$

In total, combining Eq.(6) and Eq.(7), we can bound Eq. (3) by

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta^*) &= \mathbb{E}[\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta_1)] + \mathbb{E}[\mathcal{L}(\theta_1)] - \mathcal{L}(\theta^*) \\ &\leq O\left(L\cdot\sqrt[4]{\frac{\alpha\log(1/\delta)}{\lambda}}\cdot\sqrt{\frac{G_{\mathcal{C}}}{\epsilon}} + L\sqrt{\frac{\alpha}{\lambda}} + \frac{L\cdot G_{\mathcal{C}}\cdot\sqrt{\log(1/\delta)}}{\epsilon n} + \lambda||\mathcal{C}||_2^2 + \frac{L^2}{\lambda n}\right).\end{aligned}$$

Since $\alpha \leq \min\left\{\frac{L||\mathcal{C}||_2}{n^{\frac{3}{2}}}, \frac{\epsilon^2 L||\mathcal{C}||_2^3}{G_{\mathcal{C}}^2\log(1/\delta)n^{\frac{5}{2}}}\right\}$, we have $\sqrt{L\cdot||\mathcal{C}||_2\sqrt{n}\alpha} \leq \frac{L\cdot||\mathcal{C}||_2}{\sqrt{n}}$ and $L\cdot\sqrt[4]{\frac{\alpha\log(1/\delta)}{\lambda}}\cdot\sqrt{\frac{G_{\mathcal{C}}}{\epsilon}} \leq \frac{L\cdot||\mathcal{C}||_2}{\sqrt{n}}$. Let $\lambda = \frac{L}{\sqrt{n}||\mathcal{C}||_2}$, then

$$\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta^*) \leq O\left(\frac{L\cdot G_{\mathcal{C}}\cdot\sqrt{\log(1/\delta)}}{\epsilon n} + \frac{L||\mathcal{C}||_2}{\sqrt{n}}\right).$$

Note that we need $\lambda = \frac{L}{\sqrt{n}||\mathcal{C}||_2} \geq \frac{r\beta}{\epsilon n}$, namely, $n \geq \frac{r^2\beta^2||\mathcal{C}||_2^2}{\epsilon^2 L^2}$. □

Proof of Lemma 6. Let $\bar{\mathcal{L}}(\theta, D) = \hat{\mathcal{L}}(\theta, D) + \lambda||\theta||_2^2$ and $\bar{\theta} = \arg\min_{\theta \in \mathcal{C}} \bar{\mathcal{L}}(\theta, D)$. So $\mathcal{J}(\theta, D) = \bar{\mathcal{L}}(\theta, D) + \frac{\langle \mathbf{G}, \theta \rangle}{n}$. Since θ_1 minimizes $\mathcal{J}(\theta, D)$, we have $\mathcal{J}(\bar{\theta}, D) \geq \mathcal{J}(\theta_1, D)$, namely,

$$\bar{\mathcal{L}}(\bar{\theta}, D) + \frac{\langle \mathbf{G}, \bar{\theta} \rangle}{n} \geq \bar{\mathcal{L}}(\theta_1, D) + \frac{\langle \mathbf{G}, \theta_1 \rangle}{n}.$$

Recall that $\mathbf{G} \sim \mathcal{N}(0, \frac{128L^2 \log(1/\delta)}{\epsilon^2} \mathbb{I}_d)$, rearrange the inequality and take the expectation at both sides and we get

$$\begin{aligned}
\mathbb{E}[\bar{\mathcal{L}}(\theta_1, D) - \bar{\mathcal{L}}(\bar{\theta}, D)] &\leq \mathbb{E}[\frac{\langle \mathbf{G}, \bar{\theta} - \theta_1 \rangle}{n}] \\
&\leq 2 \cdot \max_{\theta \in \mathcal{C}} \mathbb{E}\left[\frac{\langle \mathbf{G}, \theta \rangle}{n}\right] \\
&\leq 2 \cdot \mathbb{E}\left[\max_{\theta \in \mathcal{C}} \left|\frac{\langle \mathbf{G}, \theta \rangle}{n}\right|\right] \\
&= O\left(\frac{L \cdot G_{\mathcal{C}} \sqrt{\log(1/\delta)}}{\epsilon n}\right),
\end{aligned}$$

where the last bound is from the definition of Gaussian width.

Thus

$$\begin{aligned}
\mathbb{E}[\hat{\mathcal{L}}(\theta_1, D) - \hat{\mathcal{L}}(\theta^*, D)] &= \mathbb{E}[\bar{\mathcal{L}}(\theta_1, D) - \bar{\mathcal{L}}(\theta^*, D) + \lambda||\theta^*||_2^2 - \lambda||\theta_1||_2^2] \\
&\leq \mathbb{E}[\bar{\mathcal{L}}(\theta_1, D) - \bar{\mathcal{L}}(\theta^*, D) + \lambda||\theta^*||_2^2] \\
&\leq \mathbb{E}[\bar{\mathcal{L}}(\theta_1, D) - \bar{\mathcal{L}}(\bar{\theta}, D) + \lambda||\theta^*||_2^2] \\
&\leq O\left(\frac{L \cdot G_{\mathcal{C}} \sqrt{\log(1/\delta)}}{\epsilon n} + \lambda||\mathcal{C}||_2^2\right).
\end{aligned}$$

□

8.3 PROOF OF THEOREM 3

Proof. The proof is similar to the convex case. Note that $\mathcal{J}(\theta, D)$ is a $\frac{r\beta}{\epsilon n}$-strongly convex function. □

8.4 PROOF OF THEOREM 4

Proof. By the assumptions we made about n, we have $\Delta \geq \frac{L \cdot ||\mathcal{C}||_2}{\sqrt{n}}$ and $\frac{L}{\sqrt{n}||\mathcal{C}||_2} \geq \frac{r\beta}{\epsilon n}$.

Since the loss function is Δ-strongly convex with respect to $||\cdot||_{\mathcal{C}}$, which implies that the loss function is $\frac{\Delta}{||\mathcal{C}||_2^2}$-strongly convex w.r.t. $||\cdot||_2$ and thus $\frac{L}{\sqrt{n}||\mathcal{C}||_2}$-strongly convex w.r.t. $||\cdot||_2$, where we use the fact that $\Delta \geq \frac{L \cdot ||\mathcal{C}||_2}{\sqrt{n}}$ and $||v||_{\mathcal{C}} \geq \frac{||v||_2}{||\mathcal{C}||_2}$ for any vector $v \in \mathcal{C}$.

Since $\Delta \geq \frac{L}{\sqrt{n}||\mathcal{C}||_2} \geq \frac{r\beta}{\epsilon n}$, we have $\lambda = \max\left\{\frac{r\beta}{\epsilon n} - \Delta, 0\right\} = 0$.

The population loss can be disassembled as the following two parts, and we bound them separately.

$$\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta^*) = \mathbb{E}[\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta_1)] + \mathbb{E}[\mathcal{L}(\theta_1)] - \mathcal{L}(\theta^*).$$

We first bound $\mathbb{E}[\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta_1)]$. Note that

$$\mathbb{E}[\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta_1)] = \mathbb{E}[\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta_2)] + \mathbb{E}[\mathcal{L}(\theta_2) - \mathcal{L}(\theta_1)].$$

For term $\mathbb{E}[\mathcal{L}(\theta_2) - \mathcal{L}(\theta_1)]$, since $\mathcal{L}$ is Δ-strongly convex w.r.t. $||\cdot||_{\mathcal{C}}$ and thus $\frac{\Delta}{||\mathcal{C}||_2^2}$-strongly convex w.r.t. $||\cdot||_2$. So by the definition of strong convexity of $\mathcal{L}$, we have

$$\alpha \geq \mathcal{L}(\theta_2) - \mathcal{L}(\theta_1) \geq \frac{\Delta}{2||\mathcal{C}||_2^2}||\theta_2 - \theta_1||_2^2,$$

where α is the optimization accuracy.

Thus,

$$||\theta_2 - \theta_1||_2 \leq \sqrt{\frac{2\alpha||\mathcal{C}||_2^2}{\Delta}}.$$

So using the definition of L-Lipschitz,

$$\mathbb{E}[\mathcal{L}(\theta_2) - \mathcal{L}(\theta_1)] \leq L \cdot \mathbb{E}[||\theta_2 - \theta_1||_2] \leq L \cdot \sqrt{\frac{2\alpha||\mathcal{C}||_2^2}{\Delta}}.$$

For term $\mathbb{E}[\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta_2)]$, it is similar to the convex case, and we have

$$\mathbb{E}[\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta_2)] \leq O\left(L \cdot \sqrt[4]{\frac{\alpha \log(1/\delta)||\mathcal{C}||_2^2}{\Delta}} \cdot \sqrt{\frac{G_{\mathcal{C}}}{\epsilon}}\right).$$

Thus,

$$\mathbb{E}[\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta_1)] \leq O\left(L \cdot \sqrt[4]{\frac{\alpha \log(1/\delta)||\mathcal{C}||_2^2}{\Delta}} \cdot \sqrt{\frac{G_{\mathcal{C}}}{\epsilon}} + L \cdot \sqrt{\frac{2\alpha||\mathcal{C}||_2^2}{\Delta}}\right).$$

Next we bound $\mathbb{E}[\mathcal{L}(\theta_1)] - \mathcal{L}(\theta^*)$. Note that

$$\mathbb{E}[\mathcal{L}(\theta_1)] - \mathcal{L}(\theta^*) \leq \mathbb{E}[\hat{\mathcal{L}}(\theta_1, D) - \min_{\theta \in \mathcal{C}} \hat{\mathcal{L}}(\theta, D)] + \mathbb{E}[\mathcal{L}(\theta_1) - \hat{\mathcal{L}}(\theta_1, D)],$$

where we used the fact that $\mathbb{E}[\min_{\theta \in \mathcal{C}} \hat{\mathcal{L}}(\theta, D)] \leq \min_{\theta \in \mathcal{C}} \mathbb{E}[\hat{\mathcal{L}}(\theta, D)] = \mathcal{L}(\theta^*)$.

For term $\mathbb{E}[\mathcal{L}(\theta_1) - \hat{\mathcal{L}}(\theta_1, D)]$, note that with $\lambda = 0$, $f_{\mathbf{G}}(\theta, x) = \ell(\theta, x) + \frac{\langle \mathbf{G}, \theta \rangle}{n}$ would be $\frac{\Delta}{||\mathcal{C}||_2^2}$ strongly convex w.r.t. $||\cdot||_2$. Using the same notation as in the convex case, where $\hat{F}_{\mathbf{G}}(\theta, D) = \frac{1}{n}\sum_{i=1}^n f_{\mathbf{G}}(\theta, x_i)$ and $F_{\mathbf{G}}(\theta) = \mathop{\mathbb{E}}_{x \sim \mathcal{P}}[f_{\mathbf{G}}(\theta, x)]$, we have

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\theta_1) - \hat{\mathcal{L}}(\theta_1, D)] &= \mathbb{E}[F_{\mathbf{G}}(\theta_1) - \hat{F}_{\mathbf{G}}(\theta_1, D)] \\
&\leq \frac{\left(L + \frac{||\mathbf{G}||_2}{n}\right)^2 ||\mathcal{C}||_2^2}{n\Delta} \text{ (According to Lemma 7)} \\
&\leq O\left(\frac{L^2||\mathcal{C}||_2^2}{n\Delta}\right) \text{ (since } n \geq O\left(\frac{\sqrt{d\log(1/\delta)}}{\epsilon}\right)\text{)}.
\end{aligned}$$

Let $\theta' = \arg\min_{\theta \in \mathcal{C}} \hat{\mathcal{L}}(\theta, D)$. In the following, we bound the term $\mathbb{E}[\hat{\mathcal{L}}(\theta_1, D) - \min_{\theta \in \mathcal{C}} \hat{\mathcal{L}}(\theta, D)] = \mathbb{E}[\hat{\mathcal{L}}(\theta_1, D) - \hat{\mathcal{L}}(\theta', D)]$.

By the definition of strong convexity,

$$\begin{aligned}
&\hat{\mathcal{L}}(\theta_1, D) \geq \hat{\mathcal{L}}(\theta', D) + \frac{\Delta}{2}||\theta_1 - \theta'||_{\mathcal{C}}^2, \\
\Leftrightarrow \hat{\mathcal{L}}(\theta_1, D) + \frac{\langle \mathbf{G}, \theta_1 \rangle}{n} - \frac{\langle \mathbf{G}, \theta_1 \rangle}{n} &\geq \hat{\mathcal{L}}(\theta', D) + \frac{\langle \mathbf{G}, \theta' \rangle}{n} - \frac{\langle \mathbf{G}, \theta' \rangle}{n} + \frac{\Delta}{2}||\theta_1 - \theta'||_{\mathcal{C}}^2, \\
\Leftrightarrow \mathcal{J}(\theta_1, D) - \frac{\langle \mathbf{G}, \theta_1 \rangle}{n} &\geq \mathcal{J}(\theta', D) - \frac{\langle \mathbf{G}, \theta' \rangle}{n} + \frac{\Delta}{2}||\theta_1 - \theta'||_{\mathcal{C}}^2.
\end{aligned}$$

So,

$$\mathcal{J}(\theta_1, D) - \mathcal{J}(\theta', D) + \frac{\langle \mathbf{G}, \theta' - \theta_1 \rangle}{n} \geq \frac{\Delta}{2}||\theta_1 - \theta'||_{\mathcal{C}}^2.$$

Since $\mathcal{J}(\theta_1, D) - \mathcal{J}(\theta', D) \leq 0$ (due to the optimality condition), we get

$$\begin{aligned}
\frac{\langle \mathbf{G}, \theta' - \theta_1 \rangle}{n} &\geq \frac{\Delta}{2}||\theta_1 - \theta'||_{\mathcal{C}}^2, \\
\Rightarrow ||\theta_1 - \theta'||_{\mathcal{C}} &\leq \frac{2 \cdot \langle \mathbf{G}, \frac{\theta' - \theta_1}{||\theta' - \theta_1||_{\mathcal{C}}} \rangle}{n\Delta}, \\
\Rightarrow ||\theta_1 - \theta'||_{\mathcal{C}} &\leq 2 \cdot \max_{\theta \in \mathcal{C}} \frac{\langle \mathbf{G}, \theta \rangle}{n\Delta} = \frac{2||\mathbf{G}||_{\mathcal{C}^*}}{n\Delta}.
\end{aligned} \tag{8}$$

Using $\mathcal{J}(\theta_1, D) - \mathcal{J}(\theta^{'}, D) \leq 0$ again, and take the expectation at both sizes,

$$\mathcal{L}(\theta^{'}) + \mathbb{E}[\frac{\langle \mathbf{G}, \theta^{'} \rangle}{n}] \geq \mathcal{L}(\theta_1) + \mathbb{E}[\frac{\langle \mathbf{G}, \theta_1 \rangle}{n}].$$

Thus

$$\begin{aligned}
\mathcal{L}(\theta_1) - \mathcal{L}(\theta^{'}) &\leq \mathbb{E}[\frac{\langle \mathbf{G}, \theta^{'} - \theta_1 \rangle}{n}] \\
&\leq \mathbb{E}\left[\frac{||\mathbf{G}||_{\mathcal{C}^*}}{n} \cdot ||\theta_1 - \theta^{'}||_{\mathcal{C}}\right] \text{ (Holder's inequality)} \\
&\leq \mathbb{E}\left[\frac{2||\mathbf{G}||^2_{\mathcal{C}^*}}{n^2\Delta}\right] \text{ (according to Eq.(8))} \\
&\leq O\left(\frac{G^2_{\mathcal{C}} L^2 \log(1/\delta)}{\Delta n^2 \epsilon^2}\right).
\end{aligned}$$

Thus $\mathbb{E}[\hat{\mathcal{L}}(\theta_1, D) - \min_{\theta \in \mathcal{C}} \hat{\mathcal{L}}(\theta, D)] \leq O\left(\frac{L^2||\mathcal{C}||_2^2}{n\Delta} + \frac{G^2_{\mathcal{C}} L^2 \log(1/\delta)}{\Delta n^2 \epsilon^2}\right)$. So

$$\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta^*) \leq O\left(\frac{L^2||\mathcal{C}||_2^2}{n\Delta} + \frac{G^2_{\mathcal{C}} L^2 \log(1/\delta)}{\Delta n^2 \epsilon^2} + L \cdot \sqrt[4]{\frac{\alpha \log(1/\delta)||\mathcal{C}||_2^2}{\Delta}} \cdot \sqrt{\frac{G_{\mathcal{C}}}{\epsilon}} + L \cdot \sqrt{\frac{2\alpha||\mathcal{C}||_2^2}{\Delta}}\right).$$

When $\alpha \leq O\left(\min\left\{\frac{L^2||\mathcal{C}||_2^2}{\Delta n^2}, \frac{L^4 \cdot ||\mathcal{C}||_2^6 \epsilon^2}{\Delta^3 n^4 G^2_{\mathcal{C}} \log(1/\delta)}\right\}\right)$, we have $L \cdot \sqrt{\frac{2\alpha||\mathcal{C}||_2^2}{\Delta}} \leq \frac{L^2||\mathcal{C}||_2^2}{n\Delta}$ and $L \cdot \sqrt[4]{\frac{\alpha \log(1/\delta)||\mathcal{C}||_2^2}{\Delta}} \cdot \sqrt{\frac{G_{\mathcal{C}}}{\epsilon}} \leq \frac{L^2||\mathcal{C}||_2^2}{n\Delta}$.

Thus,

$$\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathcal{L}(\theta^*) \leq O\left(\frac{L^2||\mathcal{C}||_2^2}{n\Delta} + \frac{G^2_{\mathcal{C}} L^2 \log(1/\delta)}{\Delta n^2 \epsilon^2}\right).$$

□

8.5 PROOF OF THEOREM 5

Proof. To show the proof, we first prove the following theorem on the lower bound of excess empirical risk and then use reduction from Private ERM to Private SCO to get the lower bound for excess population risk.

Theorem 11. Let $\mathcal{C}$ be a symmetric body contained in the unit Euclidean ball $\mathcal{B}_2^d$ in $\mathbb{R}^d$ and satisfies $\|\mathcal{C}\|_2 = 1$. For any $n = O(\frac{\sqrt{d \log(1/\delta)}}{\epsilon})$, $\epsilon = O(1)$ and $2^{-\Omega(n)} \leq \delta \leq 1/n^{1+\Omega(1)}$, there exists a loss ℓ which is 1-Lipschitz w.r.t. $\|\cdot\|_2$ and $\mathcal{C}^2_{\min}$-strongly convex w.r.t. $\|\cdot\|_{\mathcal{C}}$, and a dataset $D = \{x_1, \cdots, x_n\} \subseteq \mathcal{C}$ such as for any (ϵ, δ)-differentially private algorithm $\mathcal{A}$, its output satisfies

$$\mathbb{E}[\hat{\mathcal{L}}(\mathcal{A}, D)] - \min_{\theta \in \mathcal{C}} \hat{\mathcal{L}}(\theta, D) = \Omega\left(\frac{G^2_{\mathcal{C}} \log(1/\delta)}{(\log(2d))^4 \epsilon^2 n^2}\right),$$

where the expectation is taken over the internal randomness of the algorithm $\mathcal{A}$.

Theorem 12 (Reduction from private ERM to private SCO [Bassily et al., 2019])**.** For any $\gamma > 0$, suppose there is a $\left(\frac{\epsilon}{4\log(1/\delta)}, \frac{e^{-\epsilon}\delta}{8\log(2/\delta)}\right)$-DP algorithm $\mathcal{A}$ such that for any distribution on domain $\mathcal{X}$, $\mathcal{A}$ yields expected population loss $\mathbb{E}_{\mathcal{A}}[\mathcal{L}(\mathcal{A})] - \min_w \mathcal{L}(w) < \gamma$. Then, there is a (ϵ, δ)-DP algorithm $\mathcal{B}$ that given any dataset $D \in \mathcal{X}^n$, it yields expected excess empirical loss $\mathbb{E}_{\mathcal{B}}[\hat{\mathcal{L}}(\mathcal{B}, D)] - \min_w \hat{\mathcal{L}}(w, D) < \gamma$.

From Theorem 12, for any dataset D and any 1-Lipschitz, $\mathcal{C}^2_{\min}$- strongly convex loss ℓ, if there exists an algorithm with excess population loss

$$\mathbb{E}[\mathcal{L}(\theta^{priv})] - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta) = o\left(\frac{G^2_{\mathcal{C}} \log(1/\delta)}{(\log(2d))^4 \epsilon^2 n^2}\right),$$

then there exists an algorithm $\mathcal{B}$ such that the excess empirical loss $\mathbb{E}[\hat{\mathcal{L}}(\mathcal{B}, D)] - \min_{\theta \in \mathcal{C}} \hat{\mathcal{L}}(\theta, D) = o\left(\frac{G_{\mathcal{C}}^2 \log(1/\delta)}{(\log(2d))^4 \epsilon^2 n^2}\right)$, which contradicts Theorem 11.

Thus, $\forall n = O(\frac{\sqrt{d \log(1/\delta)}}{\epsilon})$, there exists a dataset $D = \{x_1, \cdots, x_n\} \subseteq \mathcal{C}$ and a strongly convex loss function ℓ such that for any output θ^{priv}, the excess population loss $\mathbb{E}[\mathcal{L}(\theta^{priv})] - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta) = \Omega\left(\frac{G_{\mathcal{C}}^2 \log(1/\delta)}{(\log(2d))^4 \epsilon^2 n^2}\right)$.

As a result, we have

$$\mathbb{E}[\mathcal{L}(\theta^{priv})] - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta) = \Omega\left(\max\left\{\frac{G_{\mathcal{C}}^2 \log(1/\delta)}{(\log(2d))^4 \epsilon^2 n^2}, \frac{1}{n}\right\}\right),$$

where the first term is the lower bound on excess empirical loss and the second term is the lower bound on excess population loss in the non-private setting.

□

Proof of Theorem 11. Before starting our proof, we give some background on the mean point problem.

Let $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ be the mean of the database D, where $D = \{x_1, \cdots, x_n\}$ is a multiset of points in $\mathcal{C}$. The sample complexity of the mean point problem to achieve an error α with respect to an algorithm $\mathcal{A}$ is defined as

$$SC_{mp}(\mathcal{C}, \mathcal{A}, \alpha) = \min\{n : \sup_{D} (\mathbb{E}||\mathcal{A}(D) - \bar{x}||_2^2)^{1/2} \leq \alpha\},$$

where the supremum is taken over the database D consisting of at most n points from $\mathcal{C}$ and the expectation is taken over the randomness of the algorithm $\mathcal{A}$.

The sample complexity of solving the mean point problem with error α under (ϵ, δ)-differential privacy over convex set $\mathcal{C}$ is defined as the minimum number of samples among all the differentially private algorithm $\mathcal{A}$.

$$SC_{mp}(\mathcal{C}, \alpha) = \min\{SC_{mp}(\mathcal{C}, \mathcal{A}, \alpha) : \mathcal{A} \text{ is } (\epsilon, \delta)\text{-differentially private}\}.$$

Previous work Kattis and Nikolov [2016] shows that we can characterize sample complexity $SC_{mp}(\mathcal{C}, \alpha)$ as a natural property of convex set $\mathcal{C}$.

Lemma 9. Kattis and Nikolov [2016] Let $\mathcal{C}$ be a symmetric convex body contained in the unit Euclidean ball $\mathcal{B}_2^d$ in $\mathbb{R}^d$. Let c be an absolute constant, then for any $\epsilon = O(1), 2^{-\Omega(n)} \leq \delta \leq 1/n^{1+\Omega(1)}$ and any $\alpha \leq \frac{G_{\mathcal{C}}}{c\sqrt{d}(\log 2d)^2}$,

$$SC_{mp}(\mathcal{C}, \alpha) = \Omega\left(\frac{G_{\mathcal{C}} \sqrt{\log(1/\delta)}}{(\log 2d)^2 \alpha \epsilon}\right), \tag{9}$$

$$SC_{mp}(\mathcal{C}, \alpha) = O\left(\min\left\{\frac{G_{\mathcal{C}} \sqrt{\log(1/\delta)}}{\alpha^2 \epsilon}, \frac{\sqrt{d \log(1/\delta)}}{\alpha \epsilon}\right\}\right).$$

When $G_{\mathcal{C}} = \Omega(\sqrt{d})$, then $SC_{mp}(\mathcal{C}, \alpha) = \Theta\left(\frac{\sigma(\epsilon, \delta)\sqrt{d}}{\alpha}\right)$ for any $\alpha \leq 1/c$.

Now we start our proof with the help of the above lemma.

Let $\ell(\theta; x) = \frac{1}{2}||\theta - x||_2^2$ be half of the squared ℓ_2-distance between $\theta \in \mathcal{C} \subseteq \mathcal{B}_2^d$ and $x_i \in \mathcal{C}$, which is 1-Lipschitz and 1-strongly convex w.r.t to $\|\cdot\|_2$. Actually, based on the following lemma we can easily show it is $\mathcal{C}_{\min}^2$-strongly convex w.r.t $\|\cdot\|_{\mathcal{C}}$.

Lemma 10. For any x, we have $\|x\|_2 \geq \|x\|_{\mathcal{C}} \cdot \mathcal{C}_{\min}$.

Proof. By the definition of $\|x\|_{\mathcal{C}}$ we can see it is sufficient to show that $x \in \frac{\|x\|_2}{\mathcal{C}_{\min}} \mathcal{C}$. Note that as $\mathcal{C}$ is symmetric and $\mathcal{C}_{\min}$ is the minimal distance from the original point to the boundary of $\mathcal{C}$, thus, $\frac{\mathcal{C}}{\mathcal{C}_{\min}}$ contains the unit ℓ_2-norm ball, indicating that $x \in \frac{\|x\|_2}{\mathcal{C}_{\min}} \mathcal{C}$. □

The strongly convex decomposable loss function is defined as $\hat{\mathcal{L}}(\theta; D) = \frac{1}{2n}\sum_{i=1}^{n}\ell(\theta; x_i) = \frac{1}{2n}\sum_{i=1}^{n}||\theta - x_i||_2^2$. Notice that the minimizer of $\hat{\mathcal{L}}(\cdot; D)$ over $\mathcal{B}_2^d$ is $\theta^* = \frac{1}{n}\sum_{i=1}^{n} x_i \in \mathcal{C}$, and the excess empirical risk can be written as:

$$\mathbb{E}[\hat{\mathcal{L}}(\theta^{priv}; D)] - \hat{\mathcal{L}}(\theta^*; D) = \frac{1}{2}\mathbb{E}||\theta^{priv} - \theta^*||_2^2 = \frac{1}{2}\mathbb{E}||\theta^{priv} - \frac{1}{n}\sum_{i=1}^{n} x_i||_2^2.$$

We prove the theorem by contradiction. Assume Theorem 11 is false, then for any dataset D, there exists a (ϵ, δ)-differentially private algorithm $\mathcal{A}$, for some $n = O(\frac{\sqrt{d\log(1/\delta)}}{\epsilon})$, it outputs θ^{priv} such that $\mathbb{E}[\hat{\mathcal{L}}(\theta^{priv}; D)] - \hat{\mathcal{L}}(\theta^*; D) = \frac{1}{2}\mathbb{E}||\theta^{priv} - \frac{1}{n}\sum_{i=1}^{n} x_i||_2^2 = o\left(\frac{G_{\mathcal{C}}^2\log(1/\delta)}{(\log(2d))^4\epsilon^2 n^2}\right)$.

In Lemma 9,

$$\begin{aligned} SC_{mp} =& \min\{n : \sup_{D}(\mathbb{E}||\theta^{priv} - \bar{x}||_2^2) \le \alpha^2\} \\ =& \Omega\left(\frac{G_{\mathcal{C}}\sqrt{\log(1/\delta)}}{(\log 2d)^2\alpha\epsilon}\right) \text{(Using Eq.(9))} \\ =& o(n) \quad \text{(By letting } \alpha = o\left(\frac{G_{\mathcal{C}}\sqrt{\log(1/\delta)}}{(\log(2d))^2\epsilon n}\right)\text{),} \end{aligned}$$

which leads to a contradiction. □

9 OMITTED PROOFS IN SECTION 5

9.1 PROOF OF THEOREM 6

Proof. Note that for any neighboring dataset D and $D^{'}$, we have $||\nabla\hat{\mathcal{L}}(w_t, D) - \nabla\hat{\mathcal{L}}(w_t, D^{'})||_* \le \frac{2L}{n}$ by the Lipschitz assumption. Since for ℓ_p^d-space, $||\cdot||_* = ||\cdot||_{\frac{p}{p-1}}$, the space $(\mathbf{E}, ||\cdot||_*)$ is κ-regular with $\kappa = \min\{\frac{p}{p-1} - 1, 2\ln d\} = \min\{\frac{1}{p-1}, 2\ln d\}$, so using the privacy guarantee provided by generalized Gaussian mechanism and the advanced composition theorem, the algorithm is (ϵ, δ)-DP. □

9.2 PROOF OF THEOREM 7

Proof. Observe that $\Phi(x) = \frac{\kappa}{2}||x||_{\kappa_+}^2$ where $\kappa = \min\{\frac{1}{p-1}, 2\ln d\}$ and $\kappa_+ = \frac{\kappa}{\kappa-1}$ is 1-strongly convex w.r.t. $||\cdot||$ by the definition of $||\cdot||_{\kappa^+}$ and the duality between strongly convexity and smoothness. We recall the following lemma showing that adding regularization may impair smoothness, but it also induces good properties such as relatively smooth and strongly convex.

Lemma 11. (Lemma 14 in Attia and Koren [2022]) Let $f(x)$ be a convex and β-smooth function w.r.t. $||\cdot||$ and $\Phi(x)$ be 1-strongly convex w.r.t. $||\cdot||$, then $f^\alpha(x) = f(x) + \alpha\cdot\Phi(x)$ for $\alpha > 0$ is $(\alpha+\beta)$-smooth relative to $\Phi(x)$ as well as α-strongly convex relative to $\Phi(x)$.

Let $w_\alpha^* = \arg\min_{w\in\mathbf{E}}\hat{\mathcal{L}}(w, D) + \alpha\Phi(w)$, $w^* = \arg\min_{w\in\mathbf{E}}\mathcal{L}(w)$ and $\tilde{w}^* = \tilde{w}^*(D) = \arg\min_{w\in\mathbf{E}}\hat{\mathcal{L}}(w, D)$, and $C_D = \Phi^{\frac{1}{2}}(\tilde{w}^*)$. Based on the optimality of w_α^* for the regularized objective function $\hat{\mathcal{L}}(w, D) + \alpha\Phi(w)$, along with the optimality of $\tilde{w}^*$ for the objective $\hat{\mathcal{L}}(w, D)$, we have

$$\begin{aligned} &\hat{\mathcal{L}}(w_\alpha^*, D) + \alpha\Phi(w_\alpha^*) \le \hat{\mathcal{L}}(\tilde{w}^*, D) + \alpha\Phi(\tilde{w}^*), \\ \implies &\Phi(\tilde{w}^*) - \Phi(w_\alpha^*) \ge \frac{\hat{\mathcal{L}}(w_\alpha^*, D) - \hat{\mathcal{L}}(\tilde{w}^*, D)}{\alpha} > 0, \\ \implies &\Phi(\tilde{w}^*) > \Phi(w_\alpha^*). \end{aligned} \tag{10}$$

Since $w_1 = 0 = \arg\min_{w \in \mathbf{E}} \Phi(w)$, from the first-order optimality of w_1, we have $\langle \nabla \Phi(w_1), w_1 - w_\alpha^* \rangle \leq 0$ and thus

$$\begin{aligned} D_\Phi(w_\alpha^*, w_1) &= \Phi(w_\alpha^*) - \Phi(w_1) - \langle \nabla \Phi(w_1), w_\alpha^* - w_1 \rangle \\ &\leq \Phi(w_\alpha^*) - \Phi(w_1) \\ &\leq \Phi(\tilde{w}^*) - \Phi(w_1) (\text{ From Eq.(10)}) \\ &\leq C_D^2 \ (\text{Let } C_D^2 = \Phi(\tilde{w}^*)). \end{aligned}$$

Now we rewrite our objectives in Algorithm 3:

$$\begin{aligned} &\langle \nabla \hat{\mathcal{L}}(w_t, D) + g_t, w - w_t \rangle + \beta \cdot D_\Phi(w, w_t) + \alpha \Phi(w) \\ =&\langle \nabla \hat{\mathcal{L}}(w_t, D) + g_t, w - w_t \rangle + (\beta + \alpha) \cdot D_\Phi(w, w_t) + \alpha \Phi(x) - \alpha \cdot D_\Phi(w, w_t) \\ =&\langle \nabla \hat{\mathcal{L}}(w_t, D) + g_t, w - w_t \rangle + (\alpha + \beta) \cdot D_\Phi(w, w_t) + \alpha \Phi(w) - \alpha \cdot (\Phi(w) - \Phi(w_t) - \langle \nabla \Phi(w_t), w - w_t \rangle) \\ =&\langle \nabla \hat{\mathcal{L}}(w_t, D) + \alpha \nabla \Phi(w_t) + g_t, w - w_t \rangle + (\alpha + \beta) \cdot D_\Phi(w, w_t) + \alpha \Phi(w_t) \\ =&\langle \nabla \hat{\mathcal{L}}^{(\alpha)}(w_t, D) + g_t, w - w_t \rangle + (\alpha + \beta) \cdot D_\Phi(w, w_t) + \alpha \Phi(w_t). \end{aligned}$$

where $\hat{\mathcal{L}}^{(\alpha)}(w, D) \triangleq \hat{\mathcal{L}}(w, D) + \alpha \cdot \Phi(w)$ and note that $\hat{\mathcal{L}}^{(\alpha)}(w, D)$ is $(\alpha+\beta)$-smooth relative to $\Phi(x)$ as well as α-strongly convex relative to $\Phi(w)$ according to Lemma 11. Next, we recall the following "three-point property":

Lemma 12. (**Three point property**) Tseng [2008]. Let $\phi(x)$ be a convex function and $D_\Phi(\cdot, \cdot)$ be the Bregman divergence for $\Phi(\cdot)$. For given z, let $z^* = \arg\min_{x \in \mathbf{E}} \{\phi(x) + D_\Phi(x, z)\}$, then for all $x \in \mathbf{E}$ we have

$$\phi(x) + D_\Phi(x, z) \geq \phi(z^*) + D_\Phi(z^*, z) + D_\Phi(x, z^*).$$

Let $\phi(w) = \frac{1}{\alpha+\beta} \cdot \langle \nabla f(w_t) + g_t, w - w_t \rangle$ where $f(w) = \hat{\mathcal{L}}(w, D) + \alpha \cdot \Phi(w)$, set $z = w_t$ in Lemma 12, we get

$$\frac{1}{\alpha + \beta} \cdot \langle \nabla f(w_t) + g_t, w - w_t \rangle + D_\Phi(w, w_t) \geq \frac{1}{\alpha + \beta} \cdot \langle \nabla f(w_t) + g_t, w_{t+1} - w_t \rangle + D_\Phi(w_{t+1}, w_t) + D_\Phi(w, w_{t+1}),$$

which implies

$$(\alpha + \beta) \cdot D_\Phi(w_{t+1}, w_t) \leq \langle \nabla f(w_t) + g_t, w - w_{t+1} \rangle + (\alpha + \beta) \cdot (D_\Phi(w, w_t) - D_\Phi(w, w_{t+1})).$$

Since $f(w)$ is $(\alpha + \beta)$-smooth relative to $\Phi(w)$, we have

$$\begin{aligned} f(w_{t+1}) \leq& f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + (\alpha + \beta) \cdot D_\Phi(w_{t+1}, w_t) \\ \leq& f(w_t) + \langle \nabla f(w_t), w - w_t \rangle + (\alpha + \beta) \cdot (D_\Phi(w, w_t) - D_\Phi(w, w_{t+1})) + \langle g_t, w - w_{t+1} \rangle. \end{aligned} \tag{11}$$

Since $f(w)$ is α-strongly convex relative to $\Phi(w)$, from the definition, we have

$$f(w_t) + \langle \nabla f(w_t), w - w_t \rangle \leq f(w) - \alpha \cdot D_\Phi(w, w_t).$$

So inequality (11) becomes

$$\begin{aligned} f(w_{t+1}) &\leq f(w) - \alpha \cdot D_\Phi(w, w_t) + (\alpha + \beta) \cdot (D_\Phi(w, w_t) - D_\Phi(w, w_{t+1})) + \langle g_t, w - w_{t+1} \rangle \\ &\leq f(w) + \beta \cdot D_\Phi(w, w_t) - (\alpha + \beta) \cdot D_\Phi(w, w_{t+1}) + \langle g_t, w - w_{t+1} \rangle. \end{aligned} \tag{12}$$

Note that for any constant $a > 0$

$$\begin{aligned} \langle g_t, w - w_{t+1} \rangle \leq& a \cdot ||g_t||_*^2 + \frac{1}{2a} \cdot ||w - w_{t+1}||^2 \\ \leq& a \cdot ||g_t||_*^2 + \frac{1}{2a} \cdot D_\Phi(w, w_{t+1}), \end{aligned}$$

where the last inequality is due to Φ being 1-strongly convex w.r.t. $\|\cdot\|$. Now inequality (12) can be written as

$$f(w_{t+1}) \leq f(w) + \beta \cdot D_\Phi(w, w_t) - (\alpha + \beta - \frac{1}{2a}) \cdot D_\Phi(w, w_{t+1}) + a \cdot ||g_t||_*^2. \tag{13}$$

Let w in Eq. (13) to be $w_\alpha^* = \arg\min f(w)$, let $a = \frac{1}{\alpha}$, we have

$$\begin{aligned} D_\Phi(w_\alpha^*, w_{t+1}) \leq & \frac{\beta}{\alpha + \beta - \frac{1}{2a}} \cdot D_\Phi(w_\alpha^*, w_t) + O\left(\frac{a}{\alpha + \beta - \frac{1}{2a}} \cdot \|g_t\|_*^2\right) \\ \leq & \frac{1}{1 + \frac{\alpha}{2\beta}} \cdot D_\Phi(w_\alpha^*, w_t) + O\left(\frac{1}{\alpha\beta} \cdot ||g_t||_*^2\right). \end{aligned}$$

Letting $t = 1, 2, \cdots, T$, add these inequalities together, we have

$$\begin{aligned} \mathbb{E}[D_\Phi(w_\alpha^*, w_{T+1})] \leq & \left(\frac{1}{1 + \frac{\alpha}{2\beta}}\right)^T \cdot D_\Phi(w_\alpha^*, w_1) + O\left(\frac{1}{\alpha^2} \cdot g^2\right) \\ = & \left(1 + \frac{\alpha}{2\beta}\right)^{-T} \cdot D_\Phi(w_\alpha^*, w_1) + O\left(\frac{1}{\alpha^2} \cdot g^2\right) \\ \leq & 2^{-\frac{\alpha T}{2\beta}} \cdot D_\Phi(w_\alpha^*, w_1) + O\left(\frac{1}{\alpha^2} \cdot g^2\right) \\ \leq & 2^{-\frac{\alpha T}{2\beta}} \cdot C_D^2 + O\left(\frac{1}{\alpha^2} \cdot g^2\right), \end{aligned}$$

where the expectation is taken over all $g_1, \cdots, g_T$ and $g^2 = \mathbb{E}[||g_t||_*^2]$. The last inequality utilizes the fact that $(1 + \frac{1}{x})^x \geq 2$ for all $x \geq 1$ and note that $\frac{2\beta}{\alpha} \geq 1$. Since Φ is strongly convex, we also have

$$\frac{1}{2}\mathbb{E}[||w_\alpha^* - w_{T+1}||^2] \leq \mathbb{E}[D_\Phi(w_\alpha^*, w_{T+1})] \leq 2^{-\frac{\alpha T}{2\beta}} \cdot C_D^2 + O\left(\frac{1}{\alpha^2} \cdot g^2\right).$$

Thus, we have

$$\mathbb{E}[||w_\alpha^* - w_{T+1}||] \leq O\left(2^{-\frac{\alpha T}{4\beta}} \cdot C_D + \frac{1}{\alpha} \cdot g\right).$$

Now we consider a neighboring data D' of D where they differ by the i-th entry. Denote $w_\alpha^{*'} = \hat{\mathcal{L}}(w, D') + \alpha \cdot \Phi(w)$ and $w_{T+1}^{'}$ as the parameters of the algorithm on D'. Then, similar to the previous case we can get

$$\mathbb{E}[||w_\alpha^{*'} - w_{T+1}^{'}||] \leq O\left(2^{-\frac{\alpha T}{4\beta}} \cdot C_D + \frac{1}{\alpha} \cdot g\right).$$

Next, we will bound the term $||w_\alpha^* - w_\alpha^{*'}||$ by the following lemma.

Lemma 13. Let $f_1, f_2 : \mathbf{E} \to \mathbb{R}$ be convex and α-strongly convex (relatively). Let $x_1 = \arg\min_{x \in \mathbf{E}} f_1(x)$ and $x_2 = \arg\min_{x \in \mathbf{E}} f_2(x)$, then

$$||x_2 - x_1|| \leq \frac{2}{\alpha}||\nabla(f_2 - f_1)(x_1)||_*.$$

From the above lemma, let $f_1(w) = \hat{\mathcal{L}}(w, D) + \alpha \cdot \Phi(w)$ and $f_2(w) = \hat{\mathcal{L}}(w, D^{'}) + \alpha \cdot \Phi(w)$, we can get

$$||w_\alpha^* - w_\alpha^{*'}|| \leq \frac{2||\nabla\ell(w_\alpha^*; x_i) - \nabla\ell(w_\alpha^*; x_i^{'})||_*}{n\alpha} \leq \frac{4L}{n\alpha}.$$

In total

$$\begin{aligned} \mathbb{E}[||w_{T+1}^{'} - w_{T+1}||] \leq & O\left(2^{-\frac{\alpha T}{4\beta}} \cdot C_D + \frac{L}{n\alpha} + \frac{g}{\alpha}\right) \\ = & O\left(2^{-\frac{\alpha T}{4\beta}} \cdot C_D + \frac{L}{n\alpha} + \frac{L\sqrt{\log(1/\delta)d\kappa T}}{\alpha n\epsilon}\right). \end{aligned}$$

Similarly, we can also show that for any t we have

$$\begin{aligned}\mathbb{E}[||w_{t+1}^{'} - w_{t+1}||] \leq & O\left(2^{-\frac{\alpha t}{4\beta}} \cdot C_D + \frac{L}{n\alpha} + \frac{g}{\alpha}\right) \\ = & O\left(2^{-\frac{\alpha t}{4\beta}} \cdot C_D + \frac{L}{n\alpha} + \frac{L\sqrt{\log(1/\delta)d\kappa T}}{\alpha n\epsilon}\right).\end{aligned}$$

Now we go back to Eq. (13),

$$\begin{aligned}f(w_{t+1}) - f(w_\alpha^*) \leq & \beta \cdot D_\Phi(w_\alpha^*, w_t) - (\alpha + \beta - \frac{1}{2a}) \cdot D_\Phi(w_\alpha^*, w_{t+1}) + a \cdot ||g_t||_*^2 \\ \leq & \beta \cdot D_\Phi(w_\alpha^*, w_t) - (\beta + \frac{\alpha}{2}) \cdot D_\Phi(w_\alpha^*, w_{t+1}) + O\left(\frac{1}{\alpha} \cdot ||g_t||_*^2\right).\end{aligned}$$

Since

$$\begin{aligned}& \sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t \cdot \mathbb{E}[f(w_{t+1}) - f(w_\alpha^*)] \\ \leq & \beta\left[\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t \cdot D_\Phi(w_\alpha^*, w_t) - \sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^{t+1} \cdot D_\Phi(w_\alpha^*, w_{t+1})\right] + O\left(\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t \cdot \frac{1}{\alpha} g^2\right) \\ = & \beta\left[\frac{2\beta+\alpha}{2\beta} \cdot D_\Phi(w_\alpha^*, w_1) - \left(\frac{2\beta+\alpha}{2\beta}\right)^{T+1} \cdot D_\Phi(w_\alpha^*, w_{T+1})\right] + O\left(\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t \cdot \frac{1}{\alpha} g^2\right) \\ \leq & \frac{2\beta+\alpha}{2} \cdot D_\Phi(w_\alpha^*, w_1) + O\left(\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t \cdot \frac{1}{\alpha} g^2\right).\end{aligned}$$

Let

$$\hat{w} = \frac{\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t \cdot w_{t+1}}{\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t}.$$

And we have

$$
\begin{aligned}
\mathbb{E}[f(\hat{w}) - f(w_\alpha^*)] &= \mathbb{E}\left[f\left(\frac{\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t \cdot w_{t+1}}{\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t} \right) - f(w_\alpha^*) \right] \\
&\leq \mathbb{E}\left[\frac{\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t \cdot f(w_{t+1})}{\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t} - f(w_\alpha^*) \right] \\
&= \frac{\mathbb{E}\left[\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t \cdot (f(w_{t+1}) - f(w_\alpha^*))\right]}{\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t} \\
&= \frac{\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t \cdot \mathbb{E}[f(w_{t+1}) - f(w_\alpha^*)]}{\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t} \\
&\leq \frac{(2\beta+\alpha) \cdot D_\Phi(w_\alpha^*, w_1)}{2 \cdot \sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t} + O\left(\frac{1}{\alpha} g^2\right) \\
&= \frac{\alpha \cdot D_\Phi(w_\alpha^*, w_1)}{2\left[\left(\frac{2\beta+\alpha}{2\beta}\right)^T - 1\right]} + O\left(\frac{1}{\alpha} g^2\right) \\
&\leq \frac{\alpha}{2} \cdot D_\Phi(w_\alpha^*, w_1) + O\left(\frac{1}{\alpha} g^2\right) \qquad (14) \\
&\leq O\left(\alpha \cdot D_\Phi(w_\alpha^*, w_1) + \frac{1}{\alpha} g^2\right),
\end{aligned}
$$

where we used the fact that when $T \geq \frac{2\beta}{\alpha}$,

$$
\left(\frac{2\beta+\alpha}{2\beta}\right)^T = (1 + \frac{\alpha}{2\beta})^T \geq 2
$$

in inequality (14).

Denote $\tilde{w}^* = \arg\min_{w \in \mathbf{E}} \hat{\mathcal{L}}(w, D)$, we have

$$
\begin{aligned}
\mathbb{E}[\hat{\mathcal{L}}(\hat{w}, D) - \hat{\mathcal{L}}(\tilde{w}^*, D)] &= \mathbb{E}[\hat{\mathcal{L}}^{(\alpha)}(\hat{w}, D) - \hat{\mathcal{L}}^{(\alpha)}(\tilde{w}^*, D)] + \alpha \cdot \Phi(\tilde{w}^*) - \alpha \cdot \Phi(\hat{w}) \\
&\leq \mathbb{E}[\hat{\mathcal{L}}^{(\alpha)}(\hat{w}, D) - \hat{\mathcal{L}}^{(\alpha)}(w_\alpha^*, D)] + \alpha \cdot \Phi(\tilde{w}^*) - \alpha \cdot \Phi(\hat{w}) \\
&\leq O\left(\alpha \cdot D_\Phi(w_\alpha^*, w_1)\right) + O\left(\frac{1}{\alpha} g^2\right) + \alpha \cdot \Phi(\tilde{w}^*) - \alpha \cdot \Phi(\hat{w}) \\
&\leq O\left(\alpha \cdot D_\Phi(\tilde{w}^*, w_1)\right) + O\left(\frac{1}{\alpha} g^2\right) + \alpha \cdot C_D^2 \\
&\leq O(\alpha \cdot C_D^2 + \frac{1}{\alpha} g^2).
\end{aligned}
$$

Now we bound the sensitivity of $\hat{w}$:

$$
\begin{aligned}
\mathbb{E}[||\hat{w} - \hat{w}^{'}||] &\leq \frac{\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t \mathbb{E}[||w_{t+1} - w_{t+1}^{'}||]}{\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t} \\
&\leq O\left(\frac{\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t 2^{-\frac{\alpha t}{4\beta}} \cdot C_D}{\sum_{t=1}^{T} \left(\frac{2\beta+\alpha}{2\beta}\right)^t} + \frac{L}{n\alpha} + \frac{L\sqrt{\log(1/\delta) d \kappa T}}{\alpha n \epsilon} \right). \qquad (15)
\end{aligned}
$$

We bound the first term above:

$$\begin{aligned}\frac{\sum_{t=1}^{T}\left(\frac{2\beta+\alpha}{2\beta}\right)^t 2^{-\frac{\alpha t}{4\beta}}\cdot C_D}{\sum_{t=1}^{T}\left(\frac{2\beta+\alpha}{2\beta}\right)^t} =& \frac{C_D\cdot\sum_{t=1}^{T}\left[\frac{2\beta+\alpha}{2\beta}\cdot\left(\frac{1}{2}\right)^{\frac{\alpha}{4\beta}}\right]^t}{\sum_{t=1}^{T}\left(\frac{2\beta+\alpha}{2\beta}\right)^t}\\ =& C_D\cdot\frac{1-\frac{2\beta+\alpha}{2\beta}}{\frac{2\beta+\alpha}{2\beta}\cdot\left[1-\left(\frac{2\beta+\alpha}{2\beta}\right)^T\right]}\cdot\frac{\frac{2\beta+\alpha}{2\beta}\cdot\left(\frac{1}{2}\right)^{\frac{\alpha}{4\beta}}\cdot\left(1-\left[\frac{2\beta+\alpha}{2\beta}\cdot\left(\frac{1}{2}\right)^{\frac{\alpha}{4\beta}}\right]^T\right)}{1-\frac{2\beta+\alpha}{2\beta}\cdot\left(\frac{1}{2}\right)^{\frac{\alpha}{4\beta}}}\\ =& C_D\cdot\left(\frac{1}{2}\right)^{\frac{\alpha}{4\beta}}\cdot\frac{\alpha}{(2\beta+\alpha)\cdot\left(\frac{1}{2}\right)^{\frac{\alpha}{4\beta}}-2\beta}\cdot\frac{\left[\frac{2\beta+\alpha}{2\beta}\cdot\left(\frac{1}{2}\right)^{\frac{\alpha}{4\beta}}\right]^T-1}{\left(\frac{2\beta+\alpha}{2\beta}\right)^T-1}.\end{aligned} \tag{16}$$

Consider function $f(x)=(1+x)\cdot a^x$. Its derivative $f'(x)=\ln a\cdot a^x+a^x+\ln a\cdot x\cdot a^x=a^x(\ln a+1+\ln a\cdot x)$, let $a=\frac{1}{\sqrt{2}}$, then $f'(x)>0$ for $x\in[0,1]$. Thus we have $(1+x)\cdot(\frac{1}{\sqrt{2}})^x>1$. Let $x=\frac{\alpha}{2\beta}$, we have $(1+\frac{\alpha}{2\beta})\cdot(\frac{1}{2})^{\frac{\alpha}{4\beta}}>1$, namely $(2\beta+\alpha)\cdot(\frac{1}{2})^{\frac{\alpha}{4\beta}}-2\beta>0$.

In the following, we bound the term $\frac{\alpha}{(2\beta+\alpha)\cdot\left(\frac{1}{2}\right)^{\frac{\alpha}{4\beta}}-2\beta}$.

$$\begin{aligned}\frac{\alpha}{(2\beta+\alpha)\cdot\left(\frac{1}{2}\right)^{\frac{\alpha}{4\beta}}-2\beta} &= \frac{\alpha}{(2\beta+\alpha)\cdot\left(\left(\frac{1}{2}\right)^{\frac{\alpha}{4\beta}}-1\right)+\alpha}\\ &\le \frac{\alpha}{(2\beta+\alpha)\cdot(-\frac{\alpha}{4\beta})+\alpha}\\ &= \frac{1}{\frac{1}{2}-\frac{\alpha}{4\beta}}\le 4\ (\text{Assume }\frac{\alpha}{\beta}\le 1),\end{aligned}$$

where we use the fact that $(\frac{1}{2})^{\frac{\alpha}{4\beta}}-1\ge-\frac{\alpha}{4\beta}$. (To prove this is to prove that $2^{\frac{\alpha}{4\beta}}(1-\frac{\alpha}{4\beta})\le 1$. Let $f(x)=a^x(1-x)$. The derivative $f'(x)=\ln a\cdot a^x-\ln a\cdot x\cdot a^x-a^x=a^x\cdot(\ln a-x\cdot\ln a-1)<0$ when $a<e$. So $f(x)$ decreases in $[0,1]$, and thus $f(x)\le 1,\forall x\in[0,1]$. Let $a=2$ and $x=\frac{\alpha}{4\beta}$, and we will get $2^{\frac{\alpha}{4\beta}}\cdot(1-\frac{\alpha}{4\beta})\le 1$.)

Now we bound the term $\frac{\left[\frac{2\beta+\alpha}{2\beta}\cdot\left(\frac{1}{2}\right)^{\frac{\alpha}{4\beta}}\right]^T-1}{\left(\frac{2\beta+\alpha}{2\beta}\right)^T-1}$.

$$\begin{aligned}\frac{\left[\frac{2\beta+\alpha}{2\beta}\cdot\left(\frac{1}{2}\right)^{\frac{\alpha}{4\beta}}\right]^T-1}{\left(\frac{2\beta+\alpha}{2\beta}\right)^T-1} &= \frac{\left(\frac{2\beta+\alpha}{2\beta}\right)^T\cdot\left(\frac{1}{2}\right)^{\frac{\alpha T}{4\beta}}-\left(\frac{1}{2}\right)^{\frac{\alpha T}{4\beta}}+\left(\frac{1}{2}\right)^{\frac{\alpha T}{4\beta}}-1}{\left(\frac{2\beta+\alpha}{2\beta}\right)^T-1}\\ &= \left(\frac{1}{2}\right)^{\frac{\alpha T}{4\beta}}+\frac{\left(\frac{1}{2}\right)^{\frac{\alpha T}{4\beta}}-1}{\left(\frac{2\beta+\alpha}{2\beta}\right)^T-1}\\ &< \left(\frac{1}{2}\right)^{\frac{\alpha T}{4\beta}}.\end{aligned}$$

Thus, Eq. (16) becomes

$$\frac{\sum_{t=1}^{T}\left(\frac{2\beta+\alpha}{2\beta}\right)^t 2^{-\frac{\alpha t}{4\beta}}\cdot C_D}{\sum_{t=1}^{T}\left(\frac{2\beta+\alpha}{2\beta}\right)^t}=O\left(C_D\cdot\left(\frac{1}{2}\right)^{\frac{\alpha(T+1)}{4\beta}}\right).$$

Bring this back to Eq.(15) and we can get

$$\mathbb{E}[||\hat{w}-\hat{w}'||]\le O\left(C_D\cdot 2^{\frac{-\alpha(T+1)}{4\beta}}+\frac{L}{n\alpha}+\frac{L\sqrt{\log(1/\delta)d\kappa T}}{\alpha n\epsilon}\right).$$

Since the loss is L-Lipschitz w.r.t $\|\cdot\|$, we can see the generalization error $\mathbb{E}[\mathcal{L}(\hat{w}) - \hat{\mathcal{L}}(\hat{w}, D)] \leq L \cdot O\left(C_D \cdot 2^{\frac{-\alpha(T+1)}{4\beta}} + \frac{L}{n\alpha} + \frac{L\sqrt{\log(1/\delta)d\kappa T}}{\alpha n\epsilon}\right)$.

Take $\alpha = \frac{4\beta}{T+1}\log_2\frac{n}{T}$,

$$
\begin{aligned}
\mathbb{E}[\mathcal{L}(\hat{w})] - \mathcal{L}(w^*) &= \mathbb{E}[\mathcal{L}(\hat{w}) - \hat{\mathcal{L}}(\hat{w}, D)] + \mathbb{E}[\hat{\mathcal{L}}(\hat{w}, D) - \hat{\mathcal{L}}(w^*, D)] \\
&\leq L \cdot \mathbb{E}[||\hat{w} - \hat{w}'||] + \mathbb{E}[\hat{\mathcal{L}}(\hat{w}, D) - \hat{\mathcal{L}}(\tilde{w}^*, D)] \\
&= O\left(L \cdot 2^{\frac{-\alpha(T+1)}{4\beta}} \cdot \mathbb{E}[C_D] + \frac{L^2}{n\alpha} + \frac{L^2\sqrt{\log(1/\delta)d\kappa T}}{\alpha n\epsilon} + \alpha \cdot \mathbb{E}[C_D^2] + \frac{1}{\alpha} \cdot \frac{L^2\log(1/\delta)d\kappa T}{n^2\epsilon^2}\right) \\
&= \tilde{O}\left(\frac{T\sqrt{\kappa}}{n} + \frac{T^{\frac{3}{2}}\sqrt{d\log(1/\delta)\kappa}}{n\epsilon} + \frac{T^2 d\log(1/\delta)\kappa}{n^2\epsilon^2} + \frac{\kappa}{T}\right) \text{ (By substituting } \alpha = \frac{4\beta}{T+1}\log_2\frac{n}{T}) \\
&= \tilde{O}\left(\frac{T\sqrt{\kappa}}{n} + \frac{T^{\frac{3}{2}}\sqrt{d\log(1/\delta)\kappa}}{n\epsilon} + \frac{\kappa}{T}\right) \\
&\leq \tilde{O}\left(\frac{T^{\frac{3}{2}}\sqrt{d\log(1/\delta)\kappa}}{n\epsilon} + \frac{\kappa}{T}\right) \text{ (Since } T = O\left(\sqrt{n\sqrt{\kappa}}\right)) \\
&= \tilde{O}\left(\kappa^{\frac{4}{5}}\left(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)^{\frac{2}{5}}\right) \text{ (By letting } T = \Theta\left(\left(\frac{n\epsilon\sqrt{k}}{\sqrt{d\log(1/\delta)}}\right)^{\frac{2}{5}}\right)),
\end{aligned}
$$

where $\tilde{O}$ hides a factor of $\mathbb{E}[\tilde{C}_D^2]$ with $\tilde{C}_D^2 = \|\tilde{w}^*\|_{\kappa_+}^2$ and $\tilde{w}^* = \arg\min_{w\in\mathbf{E}}\hat{\mathcal{L}}(w, D)$.

(Note that since we assume $n = O\left(\frac{\epsilon^4}{(d\log(1/\delta))^2\kappa^{1/2}}\right)$, the constraint $T = O\left(\sqrt{n\sqrt{\kappa}}\right)$ comes for free when letting $T = \Theta\left(\left(\frac{n\epsilon\sqrt{k}}{\sqrt{d\log(1/\delta)}}\right)^{\frac{2}{5}}\right)$).

9.3 PROOF OF THEOREM 8

To be self-contained, we first review the Phased DP-SGD algorithm in Feldman et al. [2020]. Since we are concerned about the unconstrained case, we slightly modify the original Phased DP-SGD algorithm by eliminating the projection step.

Algorithm 6 Phased-DP-SGD algorithm Feldman et al. [2020]

Input: Dataset $S = \{x_1, \cdots, x_n\}$, convex loss ℓ, step size η (will be specified later), privacy parameter ϵ and (or) δ.
Set $k = \lceil \log_2 n \rceil$. Partite the whole dataset S into k subsets $\{S_1, \cdots, S_k\}$. Denote n_i as the number of samples in S_i, *i.e.,* $|S_i| = n_i$, where $n_i = \lfloor 2^{-i}n \rfloor$. Moreover, set $w_0 = 0$.
for $i = 1, \cdots, k$ **do**
    Let $\eta_i = 4^{-i}\eta$, $w_i^1 = w_{i-1}$.
    for $t = 1, \cdots, n_i$ **do**
        Update $w_i^{t+1} = w_i^t - \eta_i \nabla\ell(w_i^t, x_i^t)$, where x_i^t is the t-th sample of the set S_i.
    end for
    Set $\overline{w}_i = \frac{1}{n_i+1}\sum_{t=1}^{n_i+1} w_i^t$.
    For (ϵ, δ)-DP, $w_i = \overline{w}_i + \xi_i$, where $\xi_i \sim \mathcal{N}(0, \sigma_i^2\mathbb{I}_d)$ with $\sigma_i = \frac{4L\eta_i\sqrt{\log(1/\delta)}}{\epsilon}$.
end for
return w_k

Lemma 14. (Modification of Theorem 4.4 in Feldman et al. [2020]) Let $\ell(\cdot, x)$ be β-smooth, convex and L-Lipschitz function over $\mathbb{R}^d$ for each x. If we set $\eta = \frac{1}{L}\min\{\frac{4}{\sqrt{n}}, \frac{\epsilon}{2\sqrt{d\log(1/\delta)}}\}$ and if $\eta \leq \frac{1}{\beta}$ (*i.e.,* n is sufficiently large), then

Algorithm 6 will be (ϵ, δ)-DP for all $\epsilon \leq 2\log(1/\delta)$. The output satisfies

$$\mathbb{E}[\mathcal{L}(w_k)] - \mathcal{L}(\theta^*) \leq O\left(L\|\theta^*\|_2^2\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon n}\right)\right).$$

Proof. First, we have the following result, which can be found in the standard convergence bounds for SGD

Lemma 15. Consider the Gradient Descent method with initial parameter w_0, fixed stepsize η and iteration number T, assume in the t-the iteration we have w_t, then for any w we have

$$\mathcal{L}(\bar{w}_T, D) - \mathcal{L}(w, D) \leq O(\frac{\|w_0 - w\|_2^2}{\eta T} + \eta L^2), \tag{17}$$

where $\bar{w}_T = \frac{w_0 + w_1 + w_2 + \cdots + w_T}{T+1}$.

Now we focus on the i-th epoch, by Lemma 15 we have for any w

$$\mathbb{E}[\mathcal{L}(\bar{w}_i)] - \mathcal{L}(w) \leq O(\frac{\mathbb{E}[\|w_{i-1} - w\|_2^2]}{\eta T} + \eta L^2). \tag{18}$$

Now let's be back to our proof. We have (denote $\theta^* = \arg\min_{w \in \mathbb{R}^d} \mathcal{L}(w)$)

$$\mathcal{L}(w_k) - \mathcal{L}(\theta^*) = \underbrace{\mathcal{L}(w_k) - \mathcal{L}(\bar{w}_k)}_{A} + \underbrace{\sum_{i=2}^{k}(\mathcal{L}(\bar{w}_i) - \mathcal{L}(\bar{w}_{i-1}))}_{B} + \underbrace{\mathcal{L}(\bar{w}_1) - \mathcal{L}(\theta^*)}_{C}$$

For term A, by the Lipschitz property we have

$$\mathbb{E}[\mathcal{L}(w_k)] - \mathcal{L}(\bar{w}_k) \leq L\mathbb{E}[\|w_k - \bar{w}_k\|_2] \leq L\mathbb{E}\|\zeta_k\|_2.$$

For each term of B by (18) and take $w = \bar{w}_{i-1}$ we have

$$\mathbb{E}[\mathcal{L}(\bar{w}_i)] - \mathcal{L}(\bar{w}_{i-1}) \leq O(\frac{\mathbb{E}[\|w_{i-1} - \bar{w}_{i-1}\|_2^2]}{\eta_i n_i} + \eta_i L^2) = O(\frac{\mathbb{E}[\|\zeta_i\|_2^2]}{\eta_i n_i} + \eta_i L^2) \tag{19}$$

For term C, by (18) and take $w = \theta^*$ we have

$$\mathbb{E}[\mathcal{L}(\bar{w}_1)] - L(\theta^*) \leq O(\frac{\|\theta^*\|_2^2}{\eta_1 n_1} + \eta_1 L^2). \tag{20}$$

Thus, combing (18), (19) and (20), we have

$$\mathbb{E}[\mathcal{L}(w_k)] - \mathcal{L}(\theta^*) \leq O(L\mathbb{E}[\|\zeta_k\|_2] + \frac{\|\theta^*\|_2^2}{\eta_1 n_1} + \eta_1 L^2 + \sum_{i=2}^{k}(\frac{\mathbb{E}[\|\zeta_i\|_2^2]}{\eta_i n_i} + \eta_i L^2) \tag{21}$$

Now, we analyze the case of (ϵ, δ)-DP, it is almost the same for ϵ-DP. Specifically, we have $\mathbb{E}[\|\zeta_i\|_2^2] = O(\frac{dL^2\eta_i^2\log(1/\delta)}{\epsilon^2})$. Thus,

$$\begin{aligned} L\mathbb{E}[\|\zeta_k\|_2] \leq L\sqrt{\mathbb{E}\|\zeta_k\|_2^2} &= L^2 \cdot \frac{\sqrt{d\log(1/\delta)}\eta_k}{\epsilon} \\ &= O(\frac{\sqrt{d\log(1/\delta)}\eta L^2}{n^2\epsilon}) \\ &= O(L(\frac{\sqrt{d\log(1/\delta)}}{n^{2.5}\epsilon} + \frac{1}{n^2})). \end{aligned}$$

where the second inequality is due to $\eta = \frac{1}{L}\min\{\frac{1}{\sqrt{n}}, \frac{\epsilon}{\sqrt{d\log(1/\delta)}}\}$. And

$$\begin{aligned}\frac{\|\theta^*\|_2^2}{\eta_1 n_1} + \eta_1 L^2 &= O(\frac{\|\theta^*\|_2^2}{\eta n} + \eta L^2) \\ &= O(\|\theta^*\|_2^2 L(\frac{1}{n}\max\{\sqrt{n}, \frac{\sqrt{d\log(1/\delta)}}{\epsilon}\} + \frac{1}{\sqrt{n}})) \\ &\leq O(\|\theta^*\|_2^2 L(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon})),\end{aligned}$$

where the second inequality is due to $\eta = \frac{1}{L}\min\{\frac{1}{\sqrt{n}}, \frac{\epsilon}{\sqrt{d\log(1/\delta)}}\}$.

$$\begin{aligned}\sum_{i=2}^{k}(\frac{\mathbb{E}\|\zeta_i\|_2^2}{\eta_i n_i} + \eta_i L^2) &= O(\sum_{i=2}^{k}(\frac{dL^2\eta_i^2\log(1/\delta)}{\eta_i n_i \epsilon^2} + \eta_i L^2) \\ &= O(\sum_{i=2}^{k}\frac{2^{-i}}{n\eta} + 4^{-i}\frac{L}{\sqrt{n}}) \\ &= O(\sum_{i=2}^{k}(2^{-i}(\frac{1}{n\eta} + \frac{L}{\sqrt{n}})) \\ &\leq O(\sum_{i=2}^{\infty}(2^{-i}L(\frac{1}{n}\max\{\sqrt{n}, \frac{\sqrt{d\log(1/\delta)}}{\epsilon}\} + \frac{1}{\sqrt{n}})) \\ &\leq O(L(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon})).\end{aligned}$$

Thus, combining with the previous three bounds into (21), we have our result.

□

Next, we will prove Theorem 8 via Lemma 14. Specifically, we have the following result.

Theorem 13. For the ℓ_p^d space with $1 < p < 2$ and suppose Assumption 3 holds. Then Algorithm 6 will be (ϵ, δ)-DP for all $\epsilon \leq 2\log(1/\delta)$. If we set $\eta = \frac{1}{L}\min\{\frac{4}{\sqrt{n}}, \frac{\epsilon}{2\sqrt{d\log(1/\delta)}}\}$, the output satisfies

$$\mathbb{E}[\mathcal{L}(\hat{w})] - \mathcal{L}(\theta^*) \leq O\left(Ld^{1-\frac{2}{p}}\|\theta^*\|^2\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon n}\right)\right). \tag{22}$$

Proof. We bound the $\|\cdot\|_2$-diameter and Lipschitz constant for the ℓ_p^d-setting. First we have that $\|\theta^*\|_2 \leq d^{\frac{1}{2}-\frac{1}{p}}\|\theta^*\|$. Moreover, since ℓ is Lipschitz w.r.t. $\|\cdot\|$, we can see it is L-Lipschitz w.r.t $\|\cdot\|_2$ as $\|\nabla\ell(w,x)\|_2 \leq \|\nabla\ell(w,x)\|_* \leq L$. Moreover since ℓ is β-smooth w.r.t $\|\cdot\|$, we have $\|\nabla\ell(w,x) - \nabla\ell(w',x)\|_2 \leq \|\nabla\ell(w,x) - \nabla\ell(w',x)\|_2\|_* \leq \beta\|w - w'\| \leq \beta\|w - w'\|_2$, indicating that it is β-smooth w.r.t. $\|\cdot\|_2$. Thus, we have

$$\mathbb{E}[\mathcal{L}(\hat{w})] - \mathcal{L}(\theta^*) \leq O\left(Ld^{1-\frac{2}{p}}\|\theta^*\|^2\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon n}\right)\right). \tag{23}$$

□

9.4 PROOF OF THEOREM 9

Proof. We first recall the following lemma:

Lemma 16. [Feldman et al., 2022] For a domain $\mathcal{D}$, let $\mathcal{R}^{(i)} : f \times \mathcal{D} \to \mathcal{S}^{(i)}$ for $i \in [n]$ be a sequence of algorithms such that $\mathcal{R}^{(i)}(z_{1:i-1}, \cdot)$ is a (ϵ_0, δ_0)-DP local randomizer for all values of auxiliary inputs $z_{1:i-1} \in \mathcal{S}^{(1)} \times \cdots \times \mathcal{S}^{(i-1)}$. Let $\mathcal{A}_{\mathcal{S}} : \mathcal{D}^n \to \mathcal{S}^{(1)} \times \cdots \times \mathcal{S}^{(n)}$ be the algorithm that given a dataset $x_{1:n \in \mathcal{D}^n}$, sample a uniformly random permutation π, then sequentially computes $z_i = \mathcal{R}^{(i)}(z_{1:i-1}, x_{\pi(i)})$ for $i \in [n]$, and the outputs $z_{1:n}$. Then for any $\delta \in [0, 1]$ such that $\epsilon_0 \leq \log\left(\frac{n}{16\log(2/\delta)}\right)$, $\mathcal{A}_{\mathcal{S}}$ is $(\epsilon, \delta + O(e^{\epsilon}\delta_0 n))$-DP where $\epsilon = O\left((1 - e^{-\epsilon_0}) \cdot (\frac{\sqrt{e^{\epsilon_0}\log(1/\delta)}}{\sqrt{n}} + \frac{e^{\epsilon_0}}{n})\right)$.

Now let's get back to the proof. Note that by the Generalized Gaussian mechanism, we can see $\mathcal{R}(x) = g_x + \mathcal{GG}_{||\cdot||_+}(\sigma^2)$ with $\sigma^2 = O\left(\frac{\kappa(\beta M+\lambda)^2\log(1/\delta_0)}{\epsilon_0^2}\right)$ will be a (ϵ_0, δ_0)-DP local minimizer. The output could be considered as the postprocessing of the shuffled output $\mathcal{R}(x)$. Thus, the algorithm will be $(\hat{\epsilon}, \hat{\delta} + O(e^{\hat{\epsilon}}\delta_0 n))$-DP where $\hat{\epsilon} = O\left((1 - e^{-\epsilon_0}) \cdot (\frac{\sqrt{e^{\epsilon_0}\log(1/\hat{\delta})}}{\sqrt{n}} + \frac{e^{\epsilon_0}}{n})\right)$.

Now, assume that $\epsilon_0 \leq \frac{1}{2}$, then $\exists c_1 > 0$, s.t.,

$$\begin{aligned}
\hat{\epsilon} &\leq c_1(1 - e^{-\epsilon_0}) \cdot \left(\frac{\sqrt{e^{\epsilon_0}\log(1/\hat{\delta})}}{\sqrt{n}} + \frac{e^{\epsilon_0}}{n}\right) \\
&\leq c_1 \cdot \left((e^{\epsilon_0/2} - e^{-\epsilon_0/2}) \cdot \sqrt{\frac{\log(1/\hat{\delta})}{n}} + \frac{e^{\epsilon_0} - 1}{n}\right) \\
&\leq c_1 \cdot \left(\left((1 + \epsilon_0) - (1 - \frac{\epsilon_0}{2})\right) \cdot \sqrt{\frac{\log(1/\hat{\delta})}{n}} + \frac{(1 + 2\epsilon_0) - 1}{n}\right) \\
&= c_1 \cdot \epsilon_0 \cdot \left(\frac{3}{2}\sqrt{\frac{\log(1/\hat{\delta})}{n}} + \frac{2}{n}\right).
\end{aligned}$$

Set $\hat{\delta} = \frac{\delta}{2}$, $\delta_0 = c_2 \cdot \frac{\delta}{e^{\hat{\epsilon}}n}$ for some constant $c_2 > 0$ and replace $\epsilon_0 = \frac{c_3 \cdot \kappa(\beta M+\lambda) \cdot \sqrt{\log(1/\delta_0)}}{\sigma_1}$:

$$\begin{aligned}
\hat{\epsilon} &\leq c_1 \cdot c_3 \cdot \frac{\kappa(\beta M + \lambda) \cdot \sqrt{\log(1/\delta_0)}}{\sigma_1} \cdot \left(\frac{3}{2}\sqrt{\frac{\log(1/\hat{\delta})}{n}} + \frac{2}{n}\right) \\
&\leq O\left(\frac{\kappa(\beta M + \lambda) \cdot \sqrt{\log(1/\delta_0)\log(1/\hat{\delta})}}{\sigma_1\sqrt{n}}\right) \\
&\leq O\left(\frac{\kappa(\beta M + \lambda) \cdot \sqrt{\log(1/\delta)\log(e^{\hat{\epsilon}}n/\delta)}}{\sigma_1\sqrt{n}}\right).
\end{aligned}$$

For any $\epsilon \leq 1$, if we set $\sigma = O\left(\frac{\kappa(\beta M+\lambda)\sqrt{\log(1/\delta)\log(n/\delta)}}{\epsilon\sqrt{n}}\right)$, then we have $\hat{\epsilon} \leq \epsilon$. Furthermore, we need $\epsilon_0 = O\left(\frac{\kappa(\beta M+\lambda)\sqrt{\log(1/\delta_0)}}{\sigma}\right) \leq \frac{1}{2}$, which would be ensured if we set $\epsilon = O\left(\sqrt{\frac{\log(n/\delta)}{n}}\right)$. This implies that for $\sigma = O\left(\frac{\kappa(\beta M+\lambda)\cdot\log(n/\delta)}{\epsilon\sqrt{n}}\right)$, algorithm 4 satisfies (ϵ, δ)-DP as long as $\epsilon = O\left(\sqrt{\frac{\log(n/\delta)}{n}}\right)$. □

9.5 PROOF OF THEOREM 10

Proof. Denote $y_t = \frac{1}{|B_t|}\sum_{x \in B_t} g_x$, $z_t = \frac{1}{|B_t|}\sum_{x \in B_t} Z_x^t$ and $\tilde{y}_t = y_t + z_t$. The optimality condition for $w_t = \arg\min_{w \in \mathcal{C}}\left\{\langle \frac{\sum_{x \in B_t} g_x + Z_x^t}{|B_t|}, w\rangle + \gamma_t \cdot D_\Phi(w, w_{t-1})\right\}$ has the form:

$$\langle \tilde{y}_t + \gamma_t(\nabla\Phi(w_t) - \nabla\Phi(w_{t-1})), z - w_t\rangle \geq 0, \forall z \in \mathcal{C}.$$

Equivalently, we have

$$\begin{aligned}\langle \tilde{y}_t, w_t - z\rangle &\leq \gamma_t \langle \nabla\Phi(w_t) - \nabla\Phi(w_{t-1}), z - w_t\rangle \\ &= \gamma_t (D_\Phi(z, w_{t-1}) - D_\Phi(z, w_t) - D_\Phi(w_t, w_{t-1})),\ \forall z \in \mathcal{C}.\end{aligned}$$

Let $\xi_t = y_t - \nabla\mathcal{L}(w_{t-1}) + z_t = \tilde{y}_t - \nabla\mathcal{L}(w_{t-1})$, then we have

$$\langle \nabla\mathcal{L}(w_{t-1}), w_t - z\rangle \leq \gamma_t (D_\Phi(z, w_{t-1}) - D_\Phi(z, w_t) - D_\Phi(w_t, w_{t-1})) - \langle \xi_t, w_t - z\rangle.$$

On the other hand, we know that

$$\begin{aligned}\mathcal{L}(w_t) - \mathcal{L}(z) &= (\mathcal{L}(w_t) - \mathcal{L}(w_{t-1})) + (\mathcal{L}(w_{t-1}) - \mathcal{L}(z)) \\ &= \langle \nabla\mathcal{L}(w_{t-1}), w_t - w_{t-1}\rangle + \beta \cdot D_\Phi(w_t, w_{t-1}) + \langle \nabla\mathcal{L}(w_{t-1}), w_{t-1} - z\rangle \quad (24) \\ &\leq \langle \nabla\mathcal{L}(w_{t-1}), w_t - z\rangle + \frac{\gamma_t}{2} D_\Phi(w_t, w_{t-1}) \quad (25) \\ &\leq \gamma_t (D_\Phi(z, w_{t-1}) - D_\Phi(z, w_t) - \frac{1}{2} D_\Phi(w_t, w_{t-1})) - \langle \xi_t, w_t - z\rangle,\end{aligned}$$

where Eq. (24) uses the fact that $D_\Phi(w_t, w_{t-1}) \geq \frac{1}{2}||w_t - w_{t-1}||^2$ and $\mathcal{L}$ is smooth as well as the convexity of $\mathcal{L}$ while Eq. (25) is because $\gamma_t \geq 2\beta$.

Due to the strong convexity of $D_\Phi(\cdot, w_{t-1})$, we have

$$\begin{aligned}&\langle \xi_t, w_{t-1} - w_t\rangle \leq \frac{\gamma_t \|w_{t-1} - w_t\|_2^2}{4} + \frac{||\xi_t||_*^2}{\gamma_t} \\ \Longrightarrow\ &\langle \xi_t, w_{t-1} - w_t\rangle \leq \frac{\gamma_t}{2} D_\Phi(w_t, w_{t-1}) + \frac{||\xi_t||_*^2}{\gamma_t} \\ \Longrightarrow\ &\langle \xi_t, z - w_t\rangle - \frac{\gamma_t}{2} D_\Phi(w_t, w_{t-1}) \leq \langle \xi_t, z - w_{t-1}\rangle + \frac{||\xi_t||_*^2}{\gamma_t}.\end{aligned}$$

Thus,

$$\begin{aligned}&\mathcal{L}(w_t) - \mathcal{L}(z) \leq \gamma_t (D_\Phi(z, w_{t-1}) - D_\Phi(z, w_t)) - \langle \xi_t, w_{t-1} - z\rangle + \frac{||\xi_t||_*^2}{\gamma_t} \\ \Rightarrow &\frac{1}{\gamma_t}(\mathcal{L}(w_t) - \mathcal{L}(z)) \leq D_\Phi(z, w_{t-1}) - D_\Phi(z, w_t) - \frac{\langle \xi_t, w_{t-1} - z\rangle}{\gamma_t} + \frac{||\xi_t||_*^2}{\gamma_t^2}.\end{aligned}$$

Thus, summing over $t = 1, \cdots, T$,

$$\begin{aligned}&\sum_{t=1}^{T} (\gamma_t^{-1}) \cdot (\mathcal{L}(w_t) - \mathcal{L}(z)) \leq D_\Phi(z, w_0) - D_\Phi(z, w_T) + \sum_{t=1}^{T} \left(\frac{\langle \xi_t, z - w_{t-1}\rangle}{\gamma_t} + \frac{||\xi_t||_*^2}{\gamma_t^2} \right) \\ \Rightarrow &(\sum_{t=1}^{T} \gamma_t^{-1}) \cdot (\mathcal{L}(\frac{\sum_{t=1}^{T} \gamma_t^{-1} w_t}{\sum_{t=1}^{T} \gamma_t^{-1}}) - \mathcal{L}(z)) \leq D_\Phi(z, w_0) - D_\Phi(z, w_T) + \sum_{t=1}^{T} \left(\frac{\langle \xi_t, z - w_{t-1}\rangle}{\gamma_t} + \frac{||\xi_t||_*^2}{\gamma_t^2} \right) \\ \Rightarrow &(\sum_{t=1}^{T} \gamma_t^{-1}) \cdot (\mathcal{L}(\hat{w}) - \mathcal{L}(z)) \leq D_\Phi(z, w_0) + \sum_{t=1}^{T} \left(\frac{\langle \xi_t, z - w_{t-1}\rangle}{\gamma_t} + \frac{||\xi_t||_*^2}{\gamma_t^2} \right).\end{aligned}$$

Take the expectation over the randomness of the noise, we get

$$(\sum_{t=1}^{T} \gamma_t^{-1}) \cdot (\mathbb{E}[\mathcal{L}(\hat{w})] - \mathcal{L}(z)) \leq D_\Phi(z, w_0) + \sum_{t=1}^{T} \frac{\mathbb{E}[\langle \xi_t, z - w_{t-1}\rangle]}{\gamma_t} + \sum_{t=1}^{T} \frac{\mathbb{E}[||\xi_t||_*^2]}{\gamma_t^2}.$$

To bound the term $\sum_{t=1}^{T} \frac{\mathbb{E}[\langle \xi_t, z - w_{t-1}\rangle]}{\gamma_t}$, let $x_t = y_t - \nabla\mathcal{L}(w_{t-1})$ and notice that

$$\begin{aligned}\sum_{t=1}^{T} \frac{\mathbb{E}[\langle \xi_t, z - w_{t-1}\rangle]}{\gamma_t} &= \sum_{t=1}^{T} \frac{\mathbb{E}[\langle y_t - \nabla\mathcal{L}(w_{t-1}), z - w_{t-1}\rangle]}{\gamma_t} \\ &= \sum_{t=1}^{T} \frac{[\langle x_t, z - w_{t-1}\rangle]}{\gamma_t}.\end{aligned}$$

We will bound $\sum_{t=1}^{T} \langle x_t, z - w_{t-1}\rangle = \sum_{t=1}^{T} \psi_t$. First, we recall the following lemma proposed by Nazin et al. [2019].

Lemma 17. When $\beta M \leq \lambda$, we have

$$||x_t||_* \leq 2\beta M + \lambda \leq 3\lambda \Rightarrow |\langle x_t, z - w_{t-1}\rangle| \leq 3\lambda M,$$

$$||\mathbb{E}[x_t]||_* \leq \beta \cdot M \cdot \left(\frac{\sigma}{\lambda}\right)^2 + \frac{\sigma^2}{\lambda} \leq \frac{2\sigma^2}{\lambda} \Rightarrow |\mathbb{E}[\langle x_t, z - w_{t-1}\rangle]| \leq \frac{2\sigma^2 M}{\lambda},$$

$$\left(\mathbb{E}[||x_t||_*^2]\right)^{1/2} \leq \sigma + \beta M \cdot \frac{\sigma}{\lambda} \leq 2\sigma \Rightarrow \left(\mathbb{E}[(\langle x_t, z - w_{t-1}\rangle)^2]\right)^{1/2} \leq 2\sigma M.$$

Next, we recall Bernstein's inequality for martingales Freedman [1975],

Lemma 18. Suppose $X_1, \cdots, X_n$ are a sequence of random variables such that $0 \leq X_i \leq 1$. Define the martingale difference sequence $\{Y_n = \mathbb{E}[X_n|X_1, \cdots, X_{n-1}] - X_n\}$ and denote K_n the sum of the conditional variances

$$K_n = \sum_{t=1}^{n} \text{Var}(X_n|X_1, \cdots, X_{n-1}).$$

Let $S_n = \sum_{i=1}^n X_i$, then for all $\epsilon, k \geq 0$ we have

$$\Pr[\sum_{i=1}^{n} \mathbb{E}[X_n|X_1, \cdots, X_{n-1}] - S_n \geq \epsilon, K_n \leq k] \leq \exp(-\frac{\epsilon^2}{2k + 2\epsilon/3}). \tag{26}$$

we have

$$\begin{aligned}
\Pr\left\{\sum_{t=1}^{T} \psi_t \geq \frac{2TM\sigma^2}{\lambda} + 3 \cdot (2\sigma M)\sqrt{\tau T}\right\} &\leq \exp\left\{-\frac{9 \cdot \tau}{2 + \frac{2}{3} \cdot \frac{3\sqrt{\tau} \cdot (3\lambda M)}{2\sigma M \sqrt{T}}}\right\} \\
&\leq \exp\left\{-\frac{9\tau}{2 + \frac{3\lambda\sqrt{\tau}}{\sigma\sqrt{T}}}\right\} \\
&\leq e^{-\tau}
\end{aligned}$$

for all $\tau = O\left(\frac{\sigma^2 T}{\lambda^2}\right)$.

Thus, for all $\tau = O\left(\frac{\sigma^2 T}{\lambda^2}\right)$ w. p. $1 - e^{-\tau}$,

$$\sum_{t=1}^{T} \psi_t \leq O\left(\frac{TM\sigma^2}{\lambda} + \sigma M\sqrt{T\tau}\right).$$

Next we bound the term of $\sum_{t=1}^{T} \mathbb{E}[||\xi_t||_*^2]$. It is notable that

$$\mathbb{E}[||\xi_t||_*^2] = \mathbb{E}[\|x_t + z_t\|_*^2] \leq 2\|x_t\|_*^2 + 2\mathbb{E}[\|z_t\|_*^2] = 2\|x_t\|_*^2 + 2g^2,$$

with

$$g^2 = O(\frac{1}{|B_t|} \frac{\log(\frac{n}{\delta}) \cdot d\kappa(\beta M + \lambda)^2 \cdot \log(1/\delta)}{n\epsilon^2}) = O(\frac{\log(\frac{n}{\delta}) \cdot dT\kappa(\beta M + \lambda)^2 \cdot \log(1/\delta)}{n^2\epsilon^2}).$$

Thus, it is sufficient for us to bound $\sum_{i=1}^{T} \|x_t\|_*^2 = \sum_{i=1}^{T} \phi_i$. Similar to Lemma 17 we have the following result

Lemma 19. [Nazin et al., 2019] When $M \leq \lambda$, we have

$$\begin{aligned}
&\mathbb{E}[\phi_i] \leq (\sigma + \frac{M\sigma}{\lambda})^2 \leq 4\sigma^2, \\
&\phi_i \leq (2M + \lambda)^2 \leq 9\lambda^2, \\
&[\mathbb{E}(\phi_i^2)]^{\frac{1}{2}} \leq (\sigma + \frac{M\sigma}{\lambda})(2M + \lambda) \leq 6\lambda\sigma.
\end{aligned}$$

Thus, by Berstern's inequality, we have if $\tau = O\left(\frac{\sigma^2 T}{\lambda^2}\right)$

$$\Pr[\sum_{t=1}^{T} ||x_t||_*^2 \geq 4\sigma^2 T + 18\lambda\sigma\sqrt{T\tau}] \leq \exp(-\frac{9\tau}{2 + \frac{3\sqrt{\tau}\lambda}{\sigma\sqrt{T}}}) \leq \exp(-\tau).$$

In total, let $\gamma_t = \bar{\gamma}$, we have with probability at least $1 - 2\exp(-\tau)$

$$\mathbb{E}[\mathcal{L}(\hat{w})] - L(\theta^*) \leq O\left(\frac{D_\Phi(\theta^*, w_0)\cdot\bar{\gamma}}{T} + \frac{M\sigma^2}{\lambda} + \frac{\sigma M\sqrt{\tau}}{\sqrt{T}} + \frac{\sigma^2}{\bar{\gamma}} + \frac{M\sigma\sqrt{\tau}}{\sqrt{T}\bar{\gamma}} + \frac{\log(\frac{n}{\delta})\cdot dT\kappa(\beta M+\lambda)^2\cdot\log(1/\delta)}{n^2\epsilon^2\bar{\gamma}}\right). \tag{27}$$

Let $\frac{\bar{\gamma}}{T} = O(\frac{(\beta M+\lambda)\sqrt{d\log(1/\delta)}}{nM\epsilon})$, and since $D_\Phi(\theta^*, w_0) = \Phi(\theta^*) \leq \frac{\kappa M^2}{2}$ we have

$$\mathbb{E}[\mathcal{L}(\hat{w})] - L(\theta^*) \leq \tilde{O}\left(\frac{M\sigma^2}{\lambda} + \frac{\sigma M\sqrt{\tau}}{\sqrt{T}} + \frac{M\sigma^2}{\bar{\gamma}} + \frac{(\beta M+\lambda)M\kappa\sqrt{d\log(1/\delta)}}{n\epsilon}\right).$$

Let $\lambda = \frac{\sigma\sqrt{n\epsilon}}{\sqrt[4]{\kappa^2 d\log(1/\delta)}} \geq \max\{\beta, 1\}M$, we have

$$\mathbb{E}[\mathcal{L}(\hat{w})] - \mathcal{L}(\theta^*) \leq O\left(\frac{M\sigma\kappa\sqrt[4]{d\log(1/\delta)}}{\sqrt{n\epsilon}} + \frac{\sigma M\sqrt{\tau}}{\sqrt{T}} + \frac{M\sigma^2}{\bar{\gamma}}\right).$$

Let $\bar{\gamma} = \sqrt{T}$, then $\sqrt{T} = O(\frac{Mn\epsilon}{(\beta M+\lambda)\sqrt{d\log(1/\delta)}})$, and it holds that

$$\mathbb{E}[\mathcal{L}(\hat{w})] - \mathcal{L}(\theta^*) \leq O\left(\frac{M\max\{\sigma^2, \sigma\}\sqrt[4]{\kappa^2 d\log(1/\delta)}\sqrt{\log(1/\delta')}}{\sqrt{n\epsilon}}\right)$$

w.p. at least $1 - \delta'$. □

10 ADDITIONAL THEOREMS AND PROOFS

Theorem 14. For the ℓ_p^d space with $1 < p < 2$, suppose Assumption 4 holds and assume n is large enough such that $O((\frac{\sqrt{n\epsilon}M}{\kappa\sqrt[4]{d\log(1/\delta)}})^{\frac{2}{3}}) \geq \max\{\beta, 1\}M$. For any $0 < \epsilon, \delta < 1$, Algorithm 7 is (ϵ, δ)-DP. Moreover, if we set $\{\gamma_t\} = \gamma = \sqrt{T}$, $T = \frac{n\epsilon}{M\lambda\sqrt{d\log(1/\delta)}}$ and $\lambda = O((\frac{\sqrt{n\epsilon}M}{\kappa\sqrt[4]{d\log(1/\delta)}})^{\frac{2}{3}})$. Then for any failure probability δ', the output $\hat{w}$ satisfies the following with probability at least $1 - \delta'$

$$\mathbb{E}[\mathcal{L}(\hat{w})] - L(\theta^*) \leq O\left(M^{\frac{4}{3}}\frac{\kappa^{\frac{2}{3}}(d\log(1/\delta))^{\frac{1}{6}}\sqrt{\log(1/\delta')}}{(n\epsilon)^{\frac{1}{3}}}\right),$$

where the expectation is taken over the randomness of noise, and the probability is w.r.t. the dataset D.

10.1 PROOF OF THEOREM 14

We propose our method in Algorithm 7. Note that there are two key differences compared to Algorithm 4. First, since we do not need the privacy amplification via shuffling, there is no shuffling step. Secondly, instead of adding noise to each truncated gradient g_x, here we add a generalized Gaussian noise to the averages of the gradients for each batch. In the following we will prove our theoretical results in Theorem 14.

Proof. The proof of DP is just by the Generalizer Gaussian mechanism. For utility, the proof is almost the same as in the proof for Theorem 10, while the only difference is the noise. Similar to (27) we have the following result with probability at least $1 - 2\exp(-\tau)$

$$\mathcal{L}(\hat{w}) - \mathcal{L}(\theta^*) \leq O\left(\frac{\kappa M^2\bar{\gamma}}{T} + \frac{M\sigma^2}{\lambda} + \frac{\sigma M\sqrt{\tau}}{\sqrt{T}} + \frac{\sigma^2}{\bar{\gamma}} + \frac{M\sigma\sqrt{\tau}}{\sqrt{T}\bar{\gamma}} + \frac{dT^2\kappa(\beta M+\lambda)^2\cdot\log(1/\delta)}{n^2\epsilon^2\bar{\gamma}}\right). \tag{28}$$

Algorithm 7 Truncated DP Batched Mirror Descent

Input: Dataset D, loss function ℓ, initial point $w_0 = 0$, smooth parameter β and and λ.
Divide the permuted data into T batches $\{B_i\}_{i=1}^T$ where $|B_i| = \frac{n}{T}$ for all $i = 1, \cdots, T$
for $t = 1, \cdots, T$ **do**
  for each $x \in B_t$ **do**
    $g_x = \begin{cases} \nabla \ell(w_{t-1}, x) & \text{if } ||\nabla \ell(w_{t-1}, x)||_* \leq \beta M + \lambda \\ 0 & \text{otherwise} \end{cases}$
  end for
  Let
  $w_t = \arg\min_{w \in \mathcal{C}} \left\{ \langle \frac{\sum_{x \in B_t} g_x}{|B_t|} + Z^t, w \rangle + \gamma_t \cdot D_\Phi(w, w_{t-1}) \right\}$, where $Z^t \sim \mathcal{GG}_{||\cdot||_+}(\sigma_1^2)$ with $\sigma_1^2 = O\left(\frac{\kappa(\beta M + \lambda)^2 \cdot \log(1/\delta)}{|B_t|^2 \epsilon^2}\right)$, $||\cdot||_+$ is the smooth norm for $(\mathbf{E}, ||\cdot||_*)$. $\kappa = \min\{\frac{1}{p-1}, \log d\}$ and $\Phi(x) = \frac{\kappa}{2}||x||_{\kappa_+}^2$ with $\kappa_+ = \frac{\kappa}{\kappa - 1}$.
end for
return $\hat{w} = (\sum_{t=1}^T \gamma_t^{-1})^{-1} \cdot \sum_{t=1}^T \gamma_t^{-1} w_t$

Take $\bar{\gamma} = \sqrt{T}$ then we have

$$\mathcal{L}(\hat{w}) - \mathcal{L}(\theta^*) \leq O\left(\frac{\kappa M^2 \sqrt{\tau}}{\sqrt{T}} + \frac{M^2}{\lambda} + \frac{dT^{3/2} \kappa \lambda^2 \cdot \log(1/\delta)}{n^2 \epsilon^2}\right).$$

Take $T = \frac{n\epsilon}{M\lambda\sqrt{d\log(1/\delta)}}$ we have

$$\mathcal{L}(\hat{w}) - \mathcal{L}(\theta^*) \leq O\left(\frac{\kappa M \sqrt{\lambda} \sqrt[4]{d\log(1/\delta)} \sqrt{\tau}}{\sqrt{n\epsilon}} + \frac{M^2}{\lambda}\right).$$

Take $\lambda = O((\frac{\sqrt{n\epsilon} M}{\kappa \sqrt[4]{d\log(1/\delta)}})^{\frac{2}{3}}) \geq \max\{\beta, 1\} M$ we have w.p at least $1 - \delta'$

$$\mathcal{L}(\hat{w}) - \mathcal{L}(\theta^*) \leq O\left(M^{\frac{4}{3}} \frac{\kappa^{\frac{2}{3}} (d\log(1/\delta))^{\frac{1}{6}} \sqrt{\log(1/\delta')}}{(n\epsilon)^{\frac{1}{3}}}\right).$$

□

Theorem 15. For the ℓ_p^d space with $2 \leq p \leq \infty$, suppose Assumption 4 holds. Then the Algorithm 1 in Kamath et al. [2022] is (ϵ, δ)-DP for all $0 < \epsilon, \delta < 1$. Moreover, suppose the loss function is non-negative, there exists $R = O(1)$ such that $\|\nabla \mathcal{L}(w)\|_* \leq R$ for all $w \in \mathcal{C}$ and 3) in Assumption 5 holds. then the output satisfies

$$\mathbb{E}[\mathcal{L}(w)] - \mathcal{L}(\theta^*) \leq O\left(\frac{d^{\frac{3}{2} - \frac{1}{p}}}{\sqrt{n}} + \frac{d^{\frac{3}{2} - \frac{1}{2p}}}{\sqrt{n\epsilon}}\right). \tag{29}$$

10.2 PROOF OF THEOREM 15

Kamath et al. [2022] study DP-SCO with heavy-tailed data in Euclidean space and propose an (ϵ, δ)-DP algorithm for any $0 < \epsilon, \delta < 1$ that achieves an expected excess population risk of $O(M\frac{d}{\sqrt{n}} + \frac{\sqrt{M} d^{\frac{5}{4}}}{\sqrt{n\epsilon}})$, where M is the ℓ_2-norm diameter of the constraint set $\mathcal{C}$, under the following assumptions

Assumption 5. 1) The loss function $\ell(w, x)$ is non-negative, differentiable and convex for all $w \in \mathcal{C}$. 2) The loss function is β-smooth. 3) The gradient of $\mathcal{L}(w)$ at the optimum is zero. 4) There is a constant σ such that for all $j \in [d]$ and $w \in \mathcal{C}$ we have $\mathbb{E}[\langle \nabla \ell(w, x) - \nabla \mathcal{L}(w), e_j \rangle^2] \leq \sigma^2$, where e_j is the j-th standard basis vector. 5) For any $w \in \mathcal{C}$, the distribution of the gradient has bounded mean, i.e., $\|\nabla \mathcal{L}(w)\|_2 \leq R$.

For ℓ_p^d space, we know that L-Lipschitz w.r.t $\|\cdot\|$ implies L-Lipschitz w.r.t $\|\cdot\|_2$. Moreover, $\mathbb{E}[||\nabla \ell(w, x) - \nabla \mathcal{L}(w)||_*^2] \leq \sigma^2$ implies $\mathbb{E}[||\nabla \ell(w, x) - \nabla \mathcal{L}(w)||_2^2] \leq \sigma^2$ which indicates condition 4) in Assumption 5. For the diameter, it has the diameter of $d^{\frac{1}{2} - \frac{1}{p}} M$ w.r.t $\|\cdot\|_2$. Thus we have the following result.

□