Efficient Privacy-Preserving Stochastic Nonconvex Optimization (Supplementary material)

Lingxiao Wang[1] **Bargav Jayaraman**[2] **David Evans**[2] **Quanquan Gu**[3]

[1]Toyota Technological Institute at Chicago
[2]Department of Computer Science, University of Virginia
[3]Department of Computer Science, University of California, Los Angeles

1 ADDITIONAL EXPERIMENTS

In this section, we present additional experiment results on nonconvex logistic regression and convolutional neural networks.

1.1 RESULTS ON *IJCNN1* DATASET

In this subsection, we present the additional experiment of our method on *ijcnn1* dataset. In this dataset, we follow the same settings as before: we set the clipping thresholds $C_1 = 1, C_2 = 0.01$, and set the momentum parameter $\gamma = C_2$. Figures 1 illustrates the objective function value and the gradient norm of different algorithms under various privacy budgets $\epsilon \in \{0.2, 0.5\}$. We can see that our proposed algorithm (DP-SRM) outperforms the other three baseline algorithms (RRPSGD, DP-GD, and DP-AGD) in terms of the objective loss, gradient norm, and convergence rate by a large margin. Table 1 shows the test error of different algorithms as well as the CPU time (in seconds) of the training process on *ijcnn1* dataset. It demonstrates that our algorithm convergences faster and can achieve a better test error on the test set than other baselines.

Table 1: Comparison of different algorithms on *ijcnn1* dataset under different privacy budgets $\epsilon \in \{0.2, 0.5\}$ and $\delta = 10^{-5}$. Note that the non-private baseline denotes the test error of the non-private STORM algorithm [Cutkosky and Orabona, 2019].

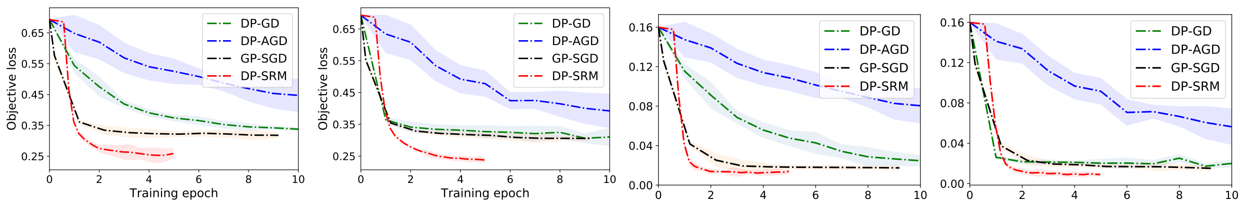
Privacy Budget	Non-private Baseline	Method	Test Error	Data Passes	CPU time	Gradient Norm
$\epsilon = 0.2$	0.2096 (0.002)	DP-GD	0.3160 (0.0120)	20	0.5180	0.0184 (0.0024)
		DP-AGD	0.2645 (0.0044)	346	90.05	0.0133 (0.0018)
		RRPSGD	0.3110 (0.0106)	8	47.64	0.0175 (0.0023)
		DP-SRM	**0.2503 (0.0090)**	**4**	**0.4748**	**0.0117 (0.0008)**
$\epsilon = 0.5$	0.2096 (0.002)	DP-GD	0.2717 (0.0081)	20	0.4990	0.0171 (0.0024)
		DP-AGD	0.2416 (0.0029)	365	94.28	0.0397 (0.0025)
		RRPSGD	0.3033 (0.0110)	10	59.06	0.0160 (0.0018)
		DP-SRM	**0.2341 (0.0042)**	**5**	**0.4368**	**0.0082 (0.0005)**

1.2 ADDITIONAL EXPERIMENTS ON CONVOLUTIONAL NEURAL NETWORKS

In this subsection, we present additional experiment results on training convolutional neural networks. Figures 2 shows the average test error (over 30 trials) and the corresponding 95% confidence interval of different methods versus the number of iterations as well as the training time under different privacy budgets on MNIST and CIFAR-10 datasets.

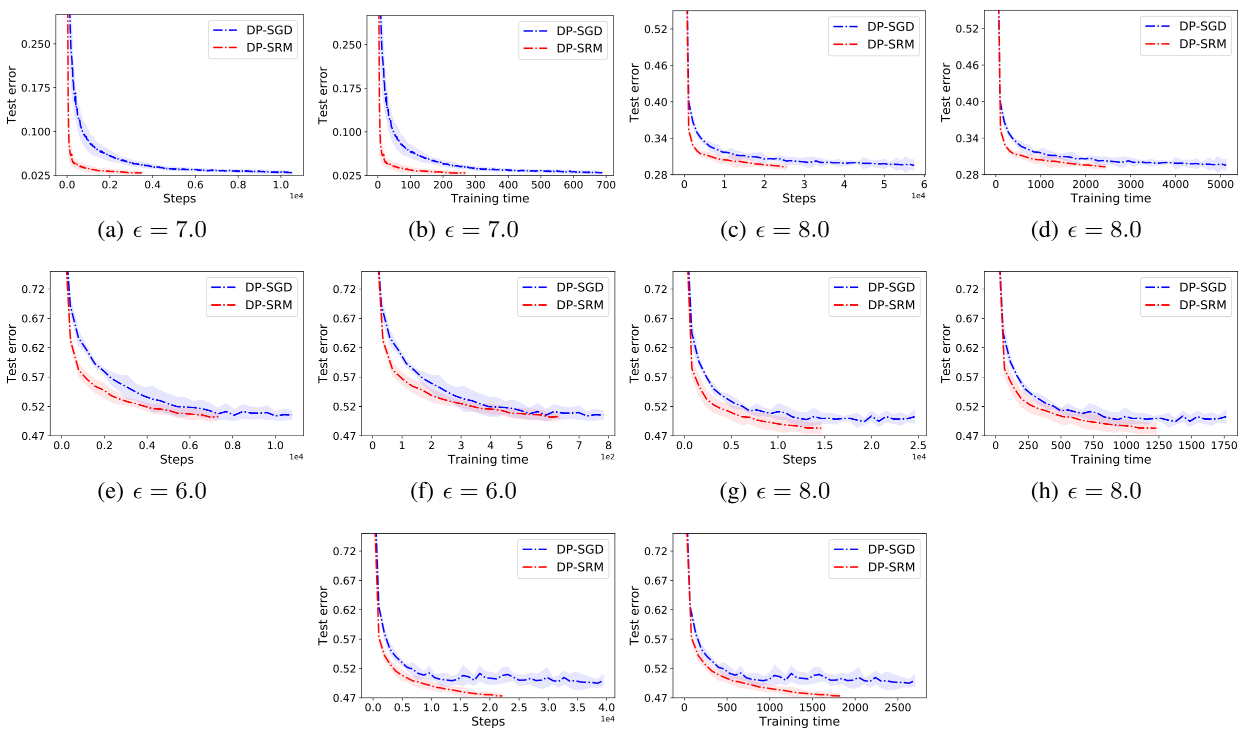
Results on MINST dataset. We can see from Figure 2(a) and Figure 2(b) that our proposed method can achieve 2.91%

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



(a) $\epsilon = 0.2$ (b) $\epsilon = 0.5$ (c) $\epsilon = 0.2$ (d) $\epsilon = 0.5$

Figure 1: Results for nonconvex logistic regression on *ijcnn1* dataset. (a), (b) show the objective loss versus the number of epochs. (c), (d) illustrate the gradient norm versus the number of epochs.



(a) $\epsilon = 7.0$ (b) $\epsilon = 7.0$ (c) $\epsilon = 8.0$ (d) $\epsilon = 8.0$

(e) $\epsilon = 6.0$ (f) $\epsilon = 6.0$ (g) $\epsilon = 8.0$ (h) $\epsilon = 8.0$

(i) $\epsilon = 10.0$ (j) $\epsilon = 10.0$

Figure 2: Results for CNN on MNIST and CIFAR-10 datasets. (a), (b) illustrate the results on MNIST dataset. (c), (d) demonstrate the results for CNN6 on CIFAR-10 dataset. (e)-(j) show the results for CNN5 on CIFAR-10 dataset.

test error when $\epsilon = 7.0$, which is comparable to the 2.93% test errors achieved by DP-SGD. Furthermore, the results show that our method is more efficient than DP-SGD in terms of iteration numbers and the training time. More specifically, our method is more than $2\times$ faster than DP-SGD to achieve the desired test error.

Parameters for CNN5. We choose three different privacy budgets $\epsilon \in \{6.0, 8.0, 10.0\}$, and set $\delta = 10^{-5}$. We set the clipping parameter $C_1 = 2$ for the term $\|\nabla f_i(\theta^t)\|_2$. For the term $\|\nabla f_i(\theta^t) - \nabla f_i(\theta^{t-1})\|_2$, we choose the clipping parameter C_2 by searching the grid $\{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 0.99\}$. For DP-SGD, we tune the batch size by searching the grid $\{32, 64, 128\}$ and the step size by $\{0.01, 0.02, 0.05, 0.1, 0.2\}$. For DP-SRM, we tune the batch size b by searching the grid $\{32, 64, 128\}$, step size by $\{0.01, 0.02, 0.05, 0.1, 0.2\}$, and b_0 by $\{b, 2b, 4b\}$. In addition, we set the momentum parameter $\gamma = C_2$.

Results for CNN5 on CIFAR-10 dataset. Figures 2(e)-2(j) present the average test error of different methods versus the number of iterations as well as the training time under different privacy budgets for CNN5 on CIFAR-10 dataset. The CNN5 trained by the non-private SGD will have 39.5% test error after 100 epochs. The results show that that our proposed method has 50.3%, 48.2% and 47.1% test errors when $\epsilon = 6.0$, $\epsilon = 8.0$ and $\epsilon = 10.0$. Nevertheless, DP-SGD has 51.0%,

50.2% and 49.3% test errors under the privacy budgets $\epsilon = 6.0$, $\epsilon = 8.0$ and $\epsilon = 10.0$, which are worse than our method. Furthermore, we can see from the plots that compared with DP-SGD, our method can reduce both the iteration numbers and the training time.

Results for CNN6 on CIFAR-10 dataset. Figure 2(c) and Figure 2(d) illustrate the average test error of different methods versus the number of iterations and the training time for CNN6 on CIFAR-10 dataset. We can see from the results that that our proposed method can achieve 29.3% test errors given the privacy budget $\epsilon = 8.0$, which are comparable to the results of DP-SGD with 29.4% under the same privacy budget. However, we can see from the plots that our method can significantly reduce the iteration numbers and the training time. When $\epsilon = 8$, DP-SGD takes 5.8×10^4 iterations and 5176 seconds to achiever 29.4% test error. In sharp contrast, our method only takes 2.6×10^4 iterations and 2589 seconds to achieve 29.3% test error.

2 PROOF OF MAIN RESULTS

In this section, we present the proofs of our main results.

2.1 PROOF OF THEOREM 5.1

We will provide the privacy guarantee of Algorithm 1 in this subsection. To this end, we need the following composition rule for RDP.

Lemma 2.1 (Mironov [2017).] If k randomized mechanisms $\mathcal{M}_i : \mathcal{S}^n \to \mathcal{R}$ for $i \in [k]$, satisfy (α, ρ_i)-RDP, then their composition $\big(\mathcal{M}_1(S), \ldots, \mathcal{M}_k(S)\big)$ satisfies $(\alpha, \sum_{i=1}^k \rho_i)$-RDP. Moreover, the input of the i-th mechanism can base on the outputs of previous $(i-1)$ mechanisms.

We will first show that our proposed algorithm satisfies RDP using Lemma 3.7 and Lemma 2.1. Then we will transform it into (ϵ, δ)-DP based on Lemma 3.9. For the given dataset S, we use S' to denote its neighboring dataset with one different example indexed by i' in the following discussion. According to Algorithm 1, we use the following $\mathcal{M}_t$ to denote the mechanism at t-th iteration

$$\mathcal{M}_t = \begin{cases} \nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^t) + (1-\gamma)\big(\mathbf{v}_p^{t-1} - \nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^{t-1})\big) + \mathbf{u}^t, & t > 0, \\ \mathbf{v}^0 + \mathbf{u}^0, & t = 0. \end{cases} \tag{2.1}$$

Therefore, our goal is to show the privacy guarantees of $\mathcal{M}_t$ for $t = 0, 1, \ldots, T$.

Case 1: If $t = 0$, we have $\mathbf{v}^0 = \nabla F_{\mathcal{B}_0}(\boldsymbol{\theta}^0)$ and $\mathcal{M}_0$ is equivalent to the following Gaussian mechanism

$$\mathcal{G}_0 = \nabla F_{\mathcal{B}_0}(\boldsymbol{\theta}^0) + \mathbf{u}^0,$$

where $\mathbf{u}^0 \sim N(0, \sigma_0^2 \mathbf{I}_d)$. Note that the mechanism $\mathcal{G}_0$ is based on the subsampling, thus we will use the results of privacy-amplification by subsampling, i.e., Lemma 3.7, to show that $\mathcal{G}_0$ satisfies RDP given appropriate $\mathbf{u}^0$. To this end, we first consider the following Gaussian mechanism without subsampling

$$\widetilde{\mathcal{G}}_0 = \frac{1}{b_0} \sum_{i=1}^n \nabla f_i(\boldsymbol{\theta}^0) + \mathbf{u}^0.$$

Sensitivity. Consider the query on the dataset S as follows $\widetilde{\mathbf{q}}_0(S) = \sum_{i=1}^n \nabla f_i(\boldsymbol{\theta}^0)/b_0$, where $\widetilde{\mathbf{q}}_0(S)$ denotes that the query is based on the dataset S. Thus, we have

$$\widetilde{\mathbf{q}}_0(S) - \widetilde{\mathbf{q}}_0(S') = \frac{1}{b_0}\big(\nabla f_i(\boldsymbol{\theta}^0) - \nabla f_{i'}(\boldsymbol{\theta}^0)\big).$$

Since each component function is G-Lipschitz, we can obtain the ℓ_2-sensitivity of this query as follows

$$\widetilde{\Delta}_0 = \frac{1}{b_0}\|\nabla f_i(\boldsymbol{\theta}^0) - \nabla f_{i'}(\boldsymbol{\theta}^0)\|_2 \le \frac{2G}{b_0}. \tag{2.2}$$

Privacy guarantee of $\mathcal{G}_0$. By Lemma 3.7, if the Gaussian noise $\mathbf{u}^0$ in $\widetilde{\mathcal{G}}_0$ has the following variance

$$\sigma_0^2 = \frac{14T\alpha G^2}{\beta n^2 \epsilon}, \tag{2.3}$$

the mechanism $\widetilde{\mathcal{G}}_0$ satisfies $\big(\alpha, \beta\epsilon n^2/\big(7b_0^2 T\big)\big)$-RDP. Therefore, according to the privacy-amplification by subsampling result in Lemma 3.7, we have that the mechanism $\mathcal{G}_0$ satisfies (α, ρ_0)-RDP, where $\rho_0 = \beta\epsilon/T$. Furthermore, the variance σ_0^2 should satisfy the following condition

$$\frac{\sigma_0^2}{\widetilde{\Delta}_0^2} = \frac{\sigma_0^2 b_0^2}{4G^2} = \frac{7b_0^2 T\alpha}{\beta n^2 \epsilon} \geq 0.7.$$

And the parameter α should satisfy $\alpha \leq 1 + 2(\sigma_0/\widetilde{\Delta}_0)^2 \log\big(1/\tau\alpha(1 + (\sigma_0/\widetilde{\Delta}_0)^2)\big)/3$.

Case 2: If $t > 0$, according to the definition of $\mathcal{M}_t$ in (2.1), we consider the following Gaussian mechanism

$$\mathcal{G}_t = \nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^t) - (1-\gamma)\nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^{t-1}) + \mathbf{u}^t.$$

Now, we are going to show that $\mathcal{G}_t$ satisfies RDP given appropriate $\mathbf{u}^t$. Since the mechanism $\mathcal{G}_t$ is based on the subsampling, we will use the similar proof procedure as in **Case 1** to show that $\mathcal{G}_t$ satisfies RDP. Thus we consider the following Gaussian mechanism without subsampling

$$\widetilde{\mathcal{G}}_t = \frac{1}{b}\sum_{i=1}^n \nabla f_i(\boldsymbol{\theta}^t) - (1-\gamma)\frac{1}{b}\sum_{i=1}^n \nabla f_i(\boldsymbol{\theta}^{t-1}) + \mathbf{u}^t.$$

Sensitivity. We consider the following query without subsampling

$$\widetilde{\mathbf{q}}_t(S) = \frac{1}{b}\sum_{i=1}^n \nabla f_i(\boldsymbol{\theta}^t) - (1-\gamma)\frac{1}{b}\sum_{i=1}^n \nabla f_i(\boldsymbol{\theta}^{t-1}).$$

Thus we have

$$\widetilde{\mathbf{q}}_t(S) - \widetilde{\mathbf{q}}_t(S') = \frac{1}{b}\big(\nabla f_i(\boldsymbol{\theta}^t) - (1-\gamma)\nabla f_i(\boldsymbol{\theta}^{t-1}) - \nabla f_{i'}(\boldsymbol{\theta}^t) + (1-\gamma)\nabla f_{i'}(\boldsymbol{\theta}^{t-1})\big).$$

As a result, we can obtain the ℓ_2-sensitivity of the query $\widetilde{\mathbf{q}}_t$ as follows

$$\begin{aligned}\widetilde{\Delta}_t &= \frac{1}{b}\big\|(1-\gamma)\big(\nabla f_i(\boldsymbol{\theta}^t) - \nabla f_i(\boldsymbol{\theta}^{t-1}) - \nabla f_{i'}(\boldsymbol{\theta}^t) + \nabla f_{i'}(\boldsymbol{\theta}^{t-1})\big) \\ &\quad + \gamma\big(\nabla f_i(\boldsymbol{\theta}^t) - \nabla f_{i'}(\boldsymbol{\theta}^t)\big)\big\|_2 \\ &\leq \frac{2L(1-\gamma)}{b}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\|_2 + \frac{2\gamma G}{b},\end{aligned}$$

where the inequality is due to L-Lipschitz continuous gradient and G-Lipschitz of each component function. Furthermore, according to the update rule of Algorithm 1 and the definition of η_{t-1}, we have

$$\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\|_2 \leq \eta_{t-1}\|\mathbf{v}_p^{t-1}\|_2 \leq \min\left\{\frac{\zeta}{n_0 L\|\mathbf{v}_p^{t-1}\|_2}, \frac{1}{2n_0 L}\right\} \cdot \|\mathbf{v}_p^{t-1}\|_2 \leq \frac{\zeta}{n_0 L},$$

which implies that

$$\widetilde{\Delta}_t \leq \frac{2L(1-\gamma)}{b}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\|_2 + \frac{2\gamma G}{b} \leq \frac{2\big((1-\gamma)\zeta/n_0 + \gamma G\big)}{b}. \tag{2.4}$$

Privacy guarantee of $\mathcal{G}_t$. By Lemma 3.7, if we the Gaussian noise $\mathbf{u}^t$ in $\widetilde{\mathcal{G}}_t$ has the variance as follows

$$\sigma_t^2 = \frac{14T\alpha\big((1-\gamma)\zeta/n_0 + \gamma G\big)^2}{\beta n^2 \epsilon}, \tag{2.5}$$

the mechanism $\widetilde{\mathcal{G}}_t$ satisfies $\big(\alpha, \beta\epsilon n^2/\big(7b^2T\big)\big)$-RDP. Thus based on the privacy-amplification by subsampling result (Lemma 3.7), we can get that the mechanism $\mathcal{G}_t$ satisfies (α, ρ)-RDP, where $\rho = \beta\epsilon/T$. In addition, the variance σ_t^2 should satisfy the following condition

$$\frac{\sigma_t^2}{\widetilde{\Delta}_t^2} = \frac{\sigma_t^2 b^2}{4\big((1-\gamma)\zeta/n_0 + \gamma G\big)^2} = \frac{7b^2 T\alpha}{\beta n^2 \epsilon} \geq 0.7.$$

And the parameter α should satisfy $\alpha \leq 1 + 2(\sigma_t/\widetilde{\Delta}_t)^2 \log\big(1/\tau\alpha(1+(\sigma_t/\widetilde{\Delta}_t)^2)\big)/3$. As a result, we show that $\mathcal{G}_t$ satisfies (α, ρ)-RDP.

Privacy guarantee of $\mathcal{M}_t$. By the definition of the mechanism $\mathcal{M}_t$ in (2.1), $\mathcal{M}_t$ is a composition of $\mathcal{G}_0, \ldots, \mathcal{G}_t$, i.e., $\mathcal{M}_t = (\mathcal{G}_0, \ldots, \mathcal{G}_t)$. According to the composition property of RDP, i.e., Lemma 2.1, we have $\mathcal{M}_t$ satisfies $(\alpha, \rho_0 + (t-1)\rho)$-RDP. Since $\rho_0 = \rho = \beta\epsilon/T$, we have that after T' iterations of Algorithm 1, it satisfies $(\alpha, \beta T'\epsilon/T)$-RDP. According to Lemma 3.9 and $\alpha = \log(1/\delta)/\big((1-\beta)\epsilon\big) + 1$, we have that after T' iterations, Algorithm 1 satisfies $(T'\epsilon/T, \delta)$-DP. As a result, we have that for each $\boldsymbol{\theta}^t$, where $t = 1, \ldots, T$, it satisfies (ϵ, δ)-DP. Finally, by the definition of $\widetilde{\boldsymbol{\theta}}$, we have $\widetilde{\boldsymbol{\theta}}$ satisfies (ϵ, δ)-DP.

2.2 PROOF OF COROLLARY 5.3

In this subsection, we show that by choosing a larger mini-batch size, we can get rid of the constraints in Theorem 5.1. More specifically, let $b_0^2 = b^2 = n^2\epsilon/T$ and $\beta = 1/2$, we have $\sigma'^2 = 7T\alpha b^2/(\beta n^2\epsilon) = 14\alpha$. Furthermore, we have

$$\tau\alpha\big(1+\sigma'^2\big) \overset{\text{(i)}}{\leq} 15\tau\alpha^2 \overset{\text{(ii)}}{=} 15\big(2\log(1/\delta)/\epsilon + 1\big)^2\sqrt{\epsilon/T},$$

where (i) uses $\sigma'^2 = 14\alpha$, (ii) uses $\tau = b/n = \sqrt{\epsilon/T}$ and $\epsilon = 2\log(1/\delta)/(\alpha-1)$. If $\epsilon \leq 2\log(1/\delta)$, we can obtain $\tau\alpha\big(1+\sigma'^2\big) \leq 1/3$ if T is larger than $O\big(\log^4(1/\delta)/\epsilon^3\big)$. If $\epsilon > 2\log(1/\delta)$, we can obtain $\tau\alpha\big(1+\sigma'^2\big) \leq 1/3$ if T is larger than $O(\epsilon)$. Therefore, we can get $\log(1/\tau\alpha\big(1+\sigma'^2)\big) \geq 1$. As a result, we have $2\big(\sigma'^2\log(1/\tau\alpha\big(1+\sigma'^2)\big)\big)/3 \geq 28\alpha/3 > \alpha - 1$.

2.3 PROOF OF THEOREM 5.4

In this subsection. we provide the utility guarantee of our method. According to the assumption that each component function has L-Lipschitz continuous gradient, we can obtain that

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\|_2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2,$$

which implies that $F(\mathbf{x})$ has L-Lipschitz continuous gradient. Thus we have

$$\begin{aligned}
F(\boldsymbol{\theta}^{t+1}) &\leq F(\boldsymbol{\theta}^t) + \langle \nabla F(\boldsymbol{\theta}^t), \boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\rangle + \frac{L}{2}\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\|_2^2 \\
&= F(\boldsymbol{\theta}^t) - \eta_t\langle \nabla F(\boldsymbol{\theta}^t), \mathbf{v}_p^t\rangle + \frac{\eta_t^2 L}{2}\big\|\mathbf{v}_p^t\big\|_2^2 \\
&= F(\boldsymbol{\theta}^t) + \frac{\eta_t}{2}\big\|\nabla F(\boldsymbol{\theta}^t) - \mathbf{v}_p^t\big\|_2^2 - \frac{\eta_t}{2}\big\|\nabla F(\boldsymbol{\theta}^t)\big\|_2^2 - \eta_t\left(\frac{1}{2} - \frac{\eta_t L}{2}\right)\big\|\mathbf{v}_p^t\big\|_2^2,
\end{aligned}$$

where the last equality is due to the fact that $2\langle \nabla F(\boldsymbol{\theta}^t), \mathbf{v}_p^t\rangle = \big\|\nabla F(\boldsymbol{\theta}^t)\big\|_2^2 + \big\|\mathbf{v}_p^t\big\|_2^2 - \big\|\nabla F(\boldsymbol{\theta}^t) - \mathbf{v}_p^t\big\|_2^2$. Since $\eta_t \leq 1/(2n_0L)$, we can obtain that

$$F(\boldsymbol{\theta}^{t+1}) \leq F(\boldsymbol{\theta}^t) + \frac{1}{4n_0L}\big\|\nabla F(\boldsymbol{\theta}^t) - \mathbf{v}_p^t\big\|_2^2 - \frac{\eta_t}{4}\big\|\mathbf{v}_p^t\big\|_2^2.$$

In addition, we have

$$\frac{\eta_t}{4}\big\|\mathbf{v}_p^t\big\|_2^2 = \frac{\zeta^2}{8n_0L}\min\big\{2\big\|\mathbf{v}_p^t/\zeta\big\|_2, \big\|\mathbf{v}_p^t/\zeta\big\|_2^2\big\} \geq \frac{\zeta\big\|\mathbf{v}_p^t\big\|_2 - 2\zeta^2}{4n_0L}.$$

Thus we have

$$F(\boldsymbol{\theta}^{t+1}) \le F(\boldsymbol{\theta}^t) + \frac{1}{4n_0L}\big\|\nabla F(\boldsymbol{\theta}^t) - \mathbf{v}_p^t\big\|_2^2 - \frac{\zeta\big\|\mathbf{v}_p^t\big\|_2}{4n_0L} + \frac{\zeta^2}{2n_0L}. \tag{2.6}$$

Summing over $t = 0, \ldots, T-1$ and taking expectation in (2.6), we can get

$$\begin{aligned}
\frac{\zeta}{4n_0L}\sum_{t=0}^{T-1}\mathbb{E}\big\|\mathbf{v}_p^t\big\|_2 &\le F(\boldsymbol{\theta}^0) - \mathbb{E}F(\boldsymbol{\theta}^T) + \frac{1}{4n_0L}\sum_{t=0}^{T-1}\mathbb{E}\big\|\nabla F(\boldsymbol{\theta}^t) - \mathbf{v}_p^t\big\|_2^2 + \frac{T\zeta^2}{2n_0L}\\
&\le F(\boldsymbol{\theta}^0) - F(\boldsymbol{\theta}^*) + \frac{1}{4n_0L}\sum_{t=0}^{T-1}\mathbb{E}\big\|\nabla F(\boldsymbol{\theta}^t) - \mathbf{v}_p^t\big\|_2^2 + \frac{T\zeta^2}{2n_0L}.
\end{aligned} \tag{2.7}$$

For the term $\mathbb{E}\big\|\nabla F(\boldsymbol{\theta}^t) - \mathbf{v}_p^t\big\|_2^2$, we can bound it as follows: we first consider the conditional expectation

$$\begin{aligned}
\mathbb{E}_t\big\|\mathbf{v}_p^t - \nabla F(\boldsymbol{\theta}^t)\big\|_2^2 &= \mathbb{E}_t\big\|(1-\gamma)\big(\mathbf{v}_p^{t-1} - \nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^{t-1})\big) + \nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^t) - \nabla F(\boldsymbol{\theta}^t) + \mathbf{u}^t\big\|_2^2\\
&= \mathbb{E}_t\big\|(1-\gamma)\big(\mathbf{v}_p^{t-1} - \nabla F(\boldsymbol{\theta}^{t-1})\big) + (1-\gamma)\nabla F(\boldsymbol{\theta}^{t-1})\\
&\qquad - (1-\gamma)\nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^{t-1}) + \nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^t) - \nabla F(\boldsymbol{\theta}^t)\big\|_2^2 + \mathbb{E}_t\|\mathbf{u}^t\|_2^2\\
&= \mathbb{E}_t\big\|(1-\gamma)\big(\mathbf{v}_p^{t-1} - \nabla F(\boldsymbol{\theta}^{t-1})\big) + (1-\gamma)\big(\nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^t) - \nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^{t-1})\\
&\qquad + \nabla F(\boldsymbol{\theta}^{t-1}) - \nabla F(\boldsymbol{\theta}^t)\big) + \gamma\big(\nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^t) - \nabla F(\boldsymbol{\theta}^t)\big)\big\|_2^2 + \mathbb{E}_t\|\mathbf{u}^t\|_2^2,
\end{aligned} \tag{2.8}$$

where $\mathbb{E}_t$ is taken over the randomness at the t-th iteration given the observations after $(t-1)$-th iteration, the first equation comes from the definition of $\mathbf{v}_p^t$, the second one is due to the independence of the random variables. Therefore, we can obtain that

$$\begin{aligned}
\mathbb{E}_t\big\|\mathbf{v}_p^t - \nabla F(\boldsymbol{\theta}^t)\big\|_2^2 &= (1-\gamma)^2\mathbb{E}_t\big\|\mathbf{v}_p^{t-1} - \nabla F(\boldsymbol{\theta}^{t-1})\big\|_2^2\\
&\quad + 2\gamma^2\mathbb{E}_t\big\|\nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^t) - \nabla F(\boldsymbol{\theta}^t)\big\|_2^2 + \mathbb{E}_t\|\mathbf{u}^t\|_2^2\\
&\quad + 2(1-\gamma)^2\mathbb{E}_t\big\|\nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^t) - \nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^{t-1}) + \nabla F(\boldsymbol{\theta}^{t-1}) - \nabla F(\boldsymbol{\theta}^t)\big\|_2^2,
\end{aligned} \tag{2.9}$$

where the equality is due to the expansion of (2.8) and Cauchy-Schwartz inequality. In addition, we have

$$\begin{aligned}
&\mathbb{E}_t\big\|\nabla F(\boldsymbol{\theta}^t) - \nabla F(\boldsymbol{\theta}^{t-1}) - \nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^t) + \nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^{t-1})\big\|_2^2\\
&\le \frac{1}{b}\cdot\frac{1}{n}\sum_{i=1}^{n}\big\|\nabla F(\boldsymbol{\theta}^t) - \nabla F(\boldsymbol{\theta}^{t-1}) - \nabla f_i(\boldsymbol{\theta}^t) + \nabla f_i(\boldsymbol{\theta}^{t-1})\big\|_2^2\\
&\le \frac{1}{b}\cdot\frac{1}{n}\sum_{i=1}^{n}\big\|\nabla f_i(\boldsymbol{\theta}^t) - \nabla f_i(\boldsymbol{\theta}^{t-1})\big\|_2^2\\
&\le \frac{L^2}{b}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\|_2^2,
\end{aligned}$$

where the first inequality is due to Lemma 4.1, the second one comes from the fact that $\mathbb{E}\|\boldsymbol{X} - \mathbb{E}\boldsymbol{X}\|_2^2 \le \mathbb{E}\|\boldsymbol{X}\|_2^2$ for any random variable $\boldsymbol{X}$, and the last one is due to the gradient Lipschitz property of each component function. According to the update rule, we have

$$\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\|_2 \le \eta_{t-1}\big\|\mathbf{v}_p^{t-1}\big\|_2 \le \min\left\{\frac{\zeta}{n_0L\big\|\mathbf{v}_p^{t-1}\big\|_2}, \frac{1}{2n_0L}\right\}\cdot\big\|\mathbf{v}_p^{t-1}\big\|_2 \le \frac{\zeta}{n_0L},$$

which implies

$$\mathbb{E}_t\big\|\nabla F(\boldsymbol{\theta}^t) - \nabla F(\boldsymbol{\theta}^{t-1}) - \nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^t) + \nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^{t-1})\big\|_2^2 \le \frac{\zeta^2}{n_0^2b}. \tag{2.10}$$

Thus plugging (2.10) into (2.9), we can obtain that

$$
\begin{aligned}
\mathbb{E}_t\big\|\mathbf{v}_p^t - \nabla F(\boldsymbol{\theta}^t)\big\|_2^2 &\le (1-\gamma)^2\big\|\mathbf{v}_p^{t-1} - \nabla F(\boldsymbol{\theta}^{t-1})\big\|_2^2 + \frac{2(1-\gamma)^2L^2}{b}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\|_2^2 \\
&\quad + 2\gamma^2\mathbb{E}_t\big\|\nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^t) - \nabla F(\boldsymbol{\theta}^t)\big\|_2^2 + \mathbb{E}_t\|\mathbf{u}^t\|_2^2 \\
&\le (1-\gamma)^2\big\|\mathbf{v}_p^{t-1} - \nabla F(\boldsymbol{\theta}^{t-1})\big\|_2^2 + \frac{2(1-\gamma)^2\zeta^2}{n_0^2 b} + \frac{2\gamma^2G^2}{b} + \mathbb{E}_t\|\mathbf{u}^t\|_2^2, \qquad (2.11)
\end{aligned}
$$

where the second inequality follows the following inequality (using Lemma 4.1, $\mathbb{E}\|\boldsymbol{X} - \mathbb{E}\boldsymbol{X}\|_2^2 \le \mathbb{E}\|\boldsymbol{X}\|_2^2$, and the G-Lipschitz of each component function)

$$
\mathbb{E}_t\big\|\nabla F_{\mathcal{B}_t}(\boldsymbol{\theta}^t) - \nabla F(\boldsymbol{\theta}^t)\big\|_2^2 \le \frac{1}{b}\cdot\frac{1}{n}\sum_{i=1}^{n}\big\|\nabla f_i(\boldsymbol{\theta}^t)\big\|_2^2 \le \frac{G^2}{b}. \qquad (2.12)
$$

Therefore, taking expectations over all iterations in (2.11), we can get

$$
\mathbb{E}\big\|\mathbf{v}_p^t - \nabla F(\boldsymbol{\theta}^t)\big\|_2^2 \le (1-\gamma)^2\mathbb{E}\big\|\mathbf{v}_p^{t-1} - \nabla F(\boldsymbol{\theta}^{t-1})\big\|_2^2 + \frac{2(1-\gamma)^2\zeta^2}{n_0^2 b} + \frac{2\gamma^2G^2}{b} + d\sigma^2. \qquad (2.13)
$$

Following the proof of Lemma 9 in Yuan et al. [2020], we have

$$
\begin{aligned}
\gamma\sum_{t=0}^{T-1}\mathbb{E}\big\|\mathbf{v}_p^t - \nabla F(\boldsymbol{\theta}^t)\big\|_2^2 &\le \frac{2T(1-\gamma)^2\zeta^2}{n_0^2 b} + \frac{2T\gamma^2G^2}{b} + Td\sigma^2 + \mathbb{E}\big\|\mathbf{v}_p^0 - \nabla F(\boldsymbol{\theta}^0)\big\|_2^2 \\
&\le \frac{2T(1-\gamma)^2\zeta^2}{n_0^2 b} + \frac{2T\gamma^2G^2}{b} + Td\sigma^2 + \frac{G^2}{b_0} + d\sigma_0^2,
\end{aligned}
$$

where the last line comes from the definition of $\mathbf{v}_p^0 = \nabla F_{\mathcal{B}_0}(\boldsymbol{\theta}^0) + \mathbf{u}^0$ and the inequality $\mathbb{E}\big\|\nabla F_{\mathcal{B}_0}(\boldsymbol{\theta}^0) - \nabla F(\boldsymbol{\theta}^0)\big\|_2^2 \le G^2/b_0$ (see equation (2.12)). Therefore, we can obtain that

$$
\sum_{t=0}^{T-1}\mathbb{E}\big\|\mathbf{v}_p^t - \nabla F(\boldsymbol{\theta}^t)\big\|_2^2 \le \frac{2T(1-\gamma)^2\zeta^2}{n_0^2\gamma b} + \frac{2T\gamma G^2}{b} + \frac{Td\sigma^2 + d\sigma_0^2}{\gamma} + \frac{G^2}{\gamma b_0}. \qquad (2.14)
$$

Combining (2.7) and (2.14), we can get

$$
\begin{aligned}
\frac{\zeta}{4n_0L}\sum_{t=0}^{T-1}\mathbb{E}\big\|\mathbf{v}_p^t\big\|_2 &\le F(\boldsymbol{\theta}^0) - F(\boldsymbol{\theta}^*) + \frac{1}{4n_0L}\sum_{t=0}^{T-1}\mathbb{E}\big\|\nabla F(\boldsymbol{\theta}^t) - \mathbf{v}_p^t\big\|_2^2 + \frac{T\zeta^2}{2n_0L} \\
&\le F(\boldsymbol{\theta}^0) - F(\boldsymbol{\theta}^*) + \frac{T(1-\gamma)^2\zeta^2}{2n_0^3L\gamma b} + \frac{T\gamma G^2}{4Ln_0b} \\
&\quad + \frac{Td\sigma^2 + d\sigma_0^2}{4n_0L\gamma} + \frac{G^2}{4L\gamma n_0 b_0} + \frac{T\zeta^2}{2n_0L}.
\end{aligned}
$$

Hence we have

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big\|\mathbf{v}_p^t\big\|_2 &\le \frac{4n_0L}{T\zeta}\big(F(\boldsymbol{\theta}^0) - F(\boldsymbol{\theta}^*)\big) + \frac{2\zeta}{n_0^2\gamma b} + \frac{\gamma G^2}{\zeta b} + \frac{d\sigma^2 + d\sigma_0^2/T}{\zeta\gamma} + \frac{G^2}{T\zeta\gamma b_0} + 2\zeta \\
&\le 6\zeta + \frac{2\zeta}{n_0^2\gamma b} + \frac{\gamma G^2}{\zeta b} + \frac{d\sigma^2 + d\sigma_0^2/T}{\zeta\gamma} + \frac{G^2}{T\zeta\gamma b_0}, \qquad (2.15)
\end{aligned}
$$

where the first inequality is due to $T = \lfloor 4n_0L\big(F(\boldsymbol{\theta}^0) - F(\boldsymbol{\theta}^*)\big)/\zeta^2\rfloor + 1$. In addition, according to (2.14) and Jensen's inequality, we have

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big\|\nabla F(\boldsymbol{\theta}^t) - \mathbf{v}_p^t\big\|_2 \le \frac{\sqrt{2}\zeta}{n_0\sqrt{\gamma b}} + \frac{\sqrt{2\gamma}G}{\sqrt{b}} + \frac{\sqrt{d}\sigma + \sqrt{d}\sigma_0/\sqrt{T}}{\sqrt{\gamma}} + \frac{G}{\sqrt{T\gamma b_0}}. \qquad (2.16)
$$

Thus by the definition of $\widetilde{\boldsymbol{\theta}}$, we have

$$
\begin{aligned}
\mathbb{E}\|\nabla F(\widetilde{\boldsymbol{\theta}})\|_2 &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\boldsymbol{\theta}^t)\|_2 \\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big\|\mathbf{v}_p^t\big\|_2 + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big\|\nabla F(\boldsymbol{\theta}^t) - \mathbf{v}_p^t\big\|_2 \\
&\leq 6\zeta + \frac{2\zeta}{n_0^2\gamma b} + \frac{\gamma G^2}{\zeta b} + \frac{d\sigma^2}{\zeta\gamma} + \frac{d\sigma_0^2}{T\zeta\gamma} + \frac{G^2}{T\zeta\gamma b_0} + \frac{\sqrt{2}\zeta}{n_0\sqrt{\gamma b}} + \frac{\sqrt{2\gamma}G}{\sqrt{b}} \\
&\quad + \frac{\sqrt{d}\sigma}{\sqrt{\gamma}} + \frac{\sqrt{d}\sigma_0}{\sqrt{T\gamma}} + \frac{G}{\sqrt{T\gamma b_0}},
\end{aligned}
\tag{2.17}
$$

where the second inequality comes from (2.15) and (2.16). Let $\gamma^2 = 2\zeta^2/(n_0^2G^2)$, $b = G/(n_0\zeta)$, $b_0 = G^3/(\zeta LD_F)$, where $D_F = F(\boldsymbol{\theta}^0) - F(\boldsymbol{\theta}^*)$ and $F(\boldsymbol{\theta}^*)$ is a global minimum of F, by the definition of T, we can get

$$
\mathbb{E}\|\nabla F(\widetilde{\boldsymbol{\theta}})\|_2 \leq 15\zeta + \frac{d\sigma^2}{\zeta\gamma} + \frac{\sqrt{d}\sigma}{\sqrt{\gamma}} + \frac{d\sigma_0^2}{T\zeta\gamma} + \frac{\sqrt{d}\sigma_0}{\sqrt{T\gamma}}. \tag{2.18}
$$

Furthermore, we have

$$
\sigma^2 = \frac{14T\big((1-\gamma)\zeta/n_0 + \gamma G\big)^2\log(1/\delta)}{n^2\epsilon^2}, \quad \sigma_0^2 = \frac{14TG^2\log(1/\delta)}{n^2\epsilon^2}. \tag{2.19}
$$

Plugging (2.19) into (2.18), we can obtain

$$
\begin{aligned}
\mathbb{E}\|\nabla F(\widetilde{\boldsymbol{\theta}})\|_2 &\leq 15\zeta + \frac{C_1TdG\log(1/\delta)}{n_0n^2\epsilon^2} + \frac{\sqrt{C_1T\zeta dG\log(1/\delta)}}{n\epsilon\sqrt{n_0}} + \frac{C_2dn_0G^3\log(1/\delta)}{n^2\epsilon^2\zeta^2} \\
&\quad + \frac{\sqrt{C_2n_0dG^3\log(1/\delta)}}{n\epsilon\sqrt{\zeta}} \\
&\leq 15\zeta + \frac{C_3LD_FGd\log(1/\delta)}{n^2\epsilon^2\zeta^2} + \frac{\sqrt{C_4GLD_Fd\log(1/\delta)}}{n\epsilon\sqrt{\zeta}} \\
&\quad + \frac{C_5n_0dG^3\log(1/\delta)}{n^2\epsilon^2\zeta^2} + \frac{\sqrt{C_6n_0dG^3\log(1/\delta)}}{n\epsilon\sqrt{\zeta}},
\end{aligned}
\tag{2.20}
$$

where the second inequality is due to the fact that $T = \lfloor 4n_0LD_F/\zeta^2\rfloor + 1$. Without loss of generality, we can assume $G \geq 1$ and $\zeta \leq 1$. Therefore, let $n_0 = LD_F/G^2 \cdot (G/\zeta)^\kappa$ with $\kappa \in [0,1]$, and plugging n_0 into (2.20), we can obtain

$$
\mathbb{E}\|\nabla F(\widetilde{\boldsymbol{\theta}})\|_2 \leq 15\zeta + \frac{C_7LD_FGd\log(1/\delta)G^\kappa}{n^2\epsilon^2\zeta^{2+\kappa}} + \frac{C_8\sqrt{GLD_Fd\log(1/\delta)}G^{\frac{\kappa}{2}}}{n\epsilon\zeta^{\frac{1+\kappa}{2}}}. \tag{2.21}
$$

Thus, choosing

$$
\zeta = C_9\bigg(\frac{G^{\frac{\kappa}{2}}\sqrt{GLD_Fd\log(1/\delta)}}{n\epsilon}\bigg)^{\frac{2}{3+\kappa}}, \tag{2.22}
$$

we can get

$$
\mathbb{E}\|\nabla F(\widetilde{\boldsymbol{\theta}})\|_2 \leq C_{10}\bigg(\frac{G^{\frac{\kappa}{2}}\sqrt{GLD_Fd\log(1/\delta)}}{n\epsilon}\bigg)^{\frac{2}{3+\kappa}}. \tag{2.23}
$$

Note that we require $\gamma \leq 1$, which gives us $n\epsilon \geq O\big(G^2(d\log(1/\delta))^{1/2}/(LD_F)\big)$.

Furthermore, according to Theorem 5.1 to achieve the desired privacy guarantee, we require $\sigma'^2 = \min\{b^2\sigma^2/\big(4((1-\gamma)\zeta/n_0 + \gamma G)^2\big), b_0^2\sigma_0^2/(4G^2)\} \geq 0.7$. Note that $b = G/(n_0\zeta)$, $b_0 = G^3/(\zeta LD_F)$, $n_0 = LD_F/G^2 \cdot (G/\zeta)^\kappa$, we have

$b = b_0 \cdot (\zeta/G)^\kappa$. Thus, the aforementioned requirement reduces to

$$\begin{aligned}\frac{14b_0^2 T\log(1/\delta)}{4n^2\epsilon^2}\cdot\frac{\zeta^{2\kappa}}{G^{2\kappa}} &= \frac{14b_0^2 n_0 LD_F\log(1/\delta)}{\zeta^4 n^2\epsilon^2}\cdot\frac{\zeta^{2\kappa}}{G^{2\kappa}}\\ &\geq \frac{14b_0 n_0 LD_F\log(1/\delta)}{\zeta^4 n^2\epsilon^2}\cdot\frac{\zeta^{2\kappa}}{G^{2\kappa}}\\ &= \frac{14GLD_F\log(1/\delta)}{\zeta^3 n^2\epsilon^2}\cdot\frac{\zeta^{\kappa}}{G^{\kappa}}\\ &\geq 0.7,\end{aligned}$$

where the first equality comes from the definition of T and the first inequality is due to $b_0 \geq 1$. Therefore, we need

$$\zeta \leq \left(4\frac{G^{-\frac{\kappa}{2}}\sqrt{GLD_F d\log(1/\delta)}}{n\epsilon}\right)^{\frac{2}{3-\kappa}}. \tag{2.24}$$

Combining (2.22) and (2.24), we need to choose $\kappa = 0$ in n_0, which gives us

$$\mathbb{E}\|\nabla F(\widetilde{\boldsymbol{\theta}})\|_2 \leq C_{10}\left(\frac{\sqrt{GLD_F d\log(1/\delta)}}{n\epsilon}\right)^{\frac{2}{3}}, \tag{2.25}$$

where $\{C_i\}_{i=1}^{10}$ are absolute constants. Furthermore, the requirement $\alpha - 1 = \log(1/\delta)/\big((1-\beta)\epsilon\big) \leq 2\sigma'^2\log\big(1/\big(\tau\alpha(1+\sigma'^2)\big)\big)/3$ in Theorem 5.1 can be satisfied under our choice of parameters given large enough n. Since we have $\sigma'^2 \geq 0.7$, we have $2\sigma'^2\log\big(1/\big(\tau\alpha(1+\sigma'^2)\big)\big)/3 \geq 0.4\log\big(1/\big(\tau\alpha(1+\sigma'^2)\big)\big) \geq 0.4\log\big(1/\big(3\tau\alpha\sigma'^2\big)\big)$. Furthermore, we have

$$\begin{aligned}\tau\alpha\sigma'^2 &= \frac{G^3}{n\zeta LD_F}\cdot\frac{\log(1/\delta)+(1-\beta)\epsilon}{(1-\beta)\epsilon}\cdot\frac{14GLD_F\log(1/\delta)}{\zeta^3 n^2\epsilon^2}\\ &\leq \frac{28G^4\log^2(1/\delta)}{(1-\beta)n^3\epsilon^3\zeta^4}\\ &\leq C_{11}\frac{G^4\log^2(1/\delta)}{(n\epsilon)^3}\cdot\frac{(n\epsilon)^{8/3}}{(GLD_F d\log(1/\delta))^{4/3}}\\ &= C_{11}\frac{G^{8/3}\log^{2/3}(1/\delta)}{(n\epsilon)^{1/3}(LD_F d)^{4/3}},\end{aligned}$$

where the first inequality comes from assumining $\epsilon \leq \log(1/\delta)$ without loss of generality, and the second inequality is due to the definition of ζ. Thus we have

$$\log\big(1/\big(3\tau\alpha\sigma'^2\big)\big) \geq \log\left(3C_{11}\frac{(n\epsilon)^{1/3}(LD_F d)^{4/3}}{G^{8/3}\log^{2/3}(1/\delta)}\right).$$

As a result, the requirement reduces to

$$0.4\log\left(3C_{11}\frac{(n\epsilon)^{1/3}(LD_F d)^{4/3}}{G^{8/3}\log^{2/3}(1/\delta)}\right) \geq \frac{\log(1/\delta)}{(1-\beta)\epsilon},$$

which can be satisfied if we have

$$n \geq C_{12}\frac{G^8\log^2(1/\delta)}{(LD_F d)^4\epsilon},$$

where C_{11}, C_{12} are some large constants.

Gradient Complexity. Since we have $b = b_0 = G^3/(\zeta LD_F)$, the total gradient complexity is

$$2(T-1)b + b_0 \leq \frac{8LD_F n_0}{\zeta^2}\cdot\frac{G^3}{LD_F\zeta} + \frac{G^3}{LD_F\zeta}.$$

According to the definition of ζ and n_0, we have the total gradient complexity is $O\big(n^2\epsilon^2/(d\log(1/\delta))\big)$.

3 PROOF OF LEMMA 3.7

Without loss of generality, we assume $\Delta(q) = 1$. According to Theorem 9 in Wang et al. [2019b], we have

$$\rho'(\alpha) \leq \frac{1}{\alpha - 1} \log \left(1 + \tau^2 \binom{\alpha}{2} \min \left\{4(e^{\rho(2)} - 1), 2e^{\rho(2)}\right\} + \sum_{j=3}^{\alpha} \tau^j \binom{\alpha}{j} 2e^{(j-1)\rho(j)}\right), \tag{3.1}$$

where τ is the subsample rate, $\rho(j) = j/(2\sigma^2)$. Next, we will show that the summation term in the right hand side of the above inequality is dominated by the second term under certain conditions. First of all, when σ^2 is large, i.e., $\sigma^2 \geq 0.7$, we have

$$\min \left\{4(e^{\rho(2)} - 1), 2e^{\rho(2)}\right\} \leq 6/\sigma^2,$$

which implies that

$$\tau^2 \binom{\alpha}{2} \min \left\{4(e^{\rho(2)} - 1), 2e^{\rho(2)}\right\} \leq \tau^2 \binom{\alpha}{2} 6/\sigma^2.$$

Next, we consider the summation term in (3.1), and we have

$$\begin{aligned} \sum_{j=3}^{\alpha} \tau^j \binom{\alpha}{j} 2e^{(j-1)\rho(j)} &\leq \tau^2 \binom{\alpha}{2} \left(\sum_{j=3}^{\alpha} \tau^{j-2} \alpha^{j-2} e^{\frac{(\alpha-1)j}{2\sigma^2}}\right) \\ &\leq \tau^2 \binom{\alpha}{2} \frac{\tau\alpha e^{\frac{3(\alpha-1)}{2\sigma^2}}}{1 - \tau\alpha e^{\frac{\alpha-1}{2\sigma^2}}}, \end{aligned}$$

where the first inequality is due to the fact that

$$e^{(j-1)\rho(j)} = e^{\frac{(j-1)j}{2\sigma^2}} \leq e^{\frac{(\alpha-1)j}{2\sigma^2}} \quad \text{and} \quad \binom{\alpha}{j} = \frac{\alpha!}{j!(\alpha - j)!} \leq \frac{\alpha^2 \alpha^{j-2}}{3!}.$$

In addition, the last inequality comes from the condition that $\tau\alpha \exp\big((\alpha - 1)/(2\sigma^2)\big) < 1$ and the sum of the geometric sequence. Therefore, as long as

$$\alpha - 1 \leq \frac{2}{3} \sigma^2 \log \frac{1}{\tau\alpha(1 + \sigma^2)}, \tag{3.2}$$

we have

$$\sum_{j=3}^{\alpha} \tau^j \binom{\alpha}{j} 2e^{(j-1)\rho(j)} \leq \tau^2 \binom{\alpha}{2} \frac{1}{\sigma^2}.$$

In addition, we require that $\tau\alpha \exp\big((\alpha - 1)/(2\sigma^2)\big) < 1$. By plugging the condition of α into the above requirement, we can obtain that this condition can hold if $\tau < 1$.

As a result, under the conditions that $\sigma^2 \geq 0.7$, $\alpha \leq \log(1/\tau(1 + \sigma^2))$, we can obtain that

$$\rho'(\alpha) \leq \frac{1}{\alpha - 1} \log \left(1 + \tau^2 \binom{\alpha}{2} \frac{10}{\sigma^2}\right) \leq \frac{1}{\alpha - 1} \tau^2 \binom{\alpha}{2} \frac{7}{\sigma^2} \leq 3.5\alpha\tau^2/\sigma^2.$$

4 AUXILIARY LEMMAS

Lemma 4.1 (Lei et al., 2017)**.** Consider vectors $\mathbf{a}_i$ satisfying $\sum_{i=1}^{n} \mathbf{a}_i = 0$. Let $\mathcal{B}$ be a uniform random subset of $\{1, 2, \ldots, n\}$ with size m, we have

$$\mathbb{E}\left\| \frac{1}{m} \sum_{i \in \mathcal{B}} \mathbf{a}_i \right\|_2^2 \leq \frac{\mathbb{1}\{|\mathcal{B}| < n\}}{mn} \sum_{i=1}^{n} \|\mathbf{a}_i\|_2^2.$$