A Proof of Theorem 1

By applying the Theorem 11 in (Dwork et al. [2015c]) to gradient computation, we get the following Lemma.

Statement of Lemma 1 *Let $\mathcal{M}$ be an ϵ-differentially private gradient descent algorithm and $S_t \sim \mathcal{P}^n$ be the training set. Let $w_s = \mathcal{M}(S_t)$ be the corresponding output for $s \in 1, ..., T$ and $\hat{g}(w_s)$ be the empirical gradient on S_t. For any $\sigma > 0$, $i \in 1, ..., d$ and $s \in 1, ..., T$, setting $\epsilon \leqslant \frac{\sigma}{2G}$ ensures*

$$P\{|\hat{g}_i(w_s) - g_i(w_s)| \geqslant \sigma\} \leqslant 6\sqrt{2}\exp\left(\frac{-n\sigma^2}{4G^2}\right). \quad (15)$$

Lemma 1 illustrates that differential privacy enables the reused training set to maintain the statistical guarantee as a fresh set under the condition that the privacy parameter ϵ is bounded by the estimation error σ. Next, we analyze the privacy parameter ϵ of StGD.

Statement of Lemma 2 *StGD satisfies $\frac{2TG}{n\sigma}$-differentially private.*

Proof. At each iteration s, the algorithm is composed of two sequential parts: differential private estimate $\tilde{g}(w_s)$ of the gradient on training set S_t and gradient descent based on the estimated gradient $\tilde{g}(w_s)$. We mark the differential private estimate as part $\mathcal{A}$ and the gradient descent as part $\mathcal{B}$. We first show $\mathcal{A}$ preserves $\frac{2G}{n\sigma}$-differential privacy. Then according to the *post-processing property* of differential privacy (Proposition 2.1 in Dwork and Roth [2014]) we have $\mathcal{B} \circ \mathcal{A}$ is also $\frac{2G}{n\sigma}$-differentially private.

The part $\mathcal{A}$ is an instantiation of basic tools from differential privacy, the "Sparse Vector Algorithm" (Algorithm 2 in Dwork and Roth [2014]) and the "Laplace Mechanism" (Definition 3.3 in Dwork and Roth [2014]). The sparse vector algorithm takes as input a sequence of c sensitivity $1/n$ queries (here $c = T$, the iteration time), and for each query, attempts to determine whether the value of the query, evaluated on the private dataset, is above a fixed threshold T or below it. In our instantiation, the training set S_t is the private data set, and each function corresponds to the gradient computation function $g^t(w_s)$ which is of sensitivity G/n. StGD is equivalent to the following procedure: we run the sparse vector algorithm with $c = T$, queries f for each gradient computation function $g^t(w_s)$, and noise rate σ. By the privacy guarantee of the sparse vector algorithm, the sparse vector portion of StGD satisfies $G/n\sigma$-differential privacy. The Laplace mechanism portion of StGD satisfies $G/n\sigma$-differential privacy by (Theorem 3.6 in Dwork and Roth [2014]). Finally, the composition of two mechanisms satisfies $\frac{2G}{n\sigma}$-differential privacy. After all the iterations, by the advanced composition theorem (Theorem 3.20 in Dwork and Roth [2014]), T applications of a $\frac{2G}{n\sigma}$-differentially private algorithm is $\frac{2TG}{n\sigma}$-differentially private. So StGD preserves $\frac{2TG}{n\sigma}$ differential privacy. □

In order to achieve the gradient concentration bound described in Lemma 1 by considering the guarantee of Lemma 2 (i.e. to guarantee that for every w_s, we have $P\{|\hat{g}_i(w_s) - g_i(w_s)| \geqslant \sigma\} \leqslant 6\sqrt{2}\exp(\frac{-n\sigma^2}{4G^2})$), we need to set $\frac{2TG}{n\sigma} \leqslant \frac{\sigma}{2G}$ so that we achieve ϵ-differential privacy for $\epsilon \leqslant \frac{\sigma}{2G}$. As a result, we get the upper bound of iteration time T in StGD as $T = \frac{\sigma^2 n}{4G^2}$.

Statement of Theorem1: *Given parameter $\sigma > 0$, let $w_1, w_2, ..., w_T$ be the adaptively updated points by StGD and $\tilde{g}(w_1), ..., \tilde{g}(w_T)$ be the corresponding output gradient. If we set $T = \frac{\sigma^2 n}{4G^2}$, then for all $s \in 1, ..., T$ and for all $t > 0$, we have*

$$\begin{aligned} P\{ \quad & \|\tilde{g}(w_s) - g(w_s)\|^2 \geqslant d(6t+1)^2\sigma^2\} \\ & \leqslant 2d\exp(-t) + 6\sqrt{2}d\exp\left(\tfrac{-n\sigma^2}{4G^2}\right). \end{aligned} \quad (16)$$

Proof. We first prove the concentration bound of each coordinate:

$$\begin{aligned} P\{ \quad & |\tilde{g}_i(w_s) - g_i(w_s)| \geqslant (6t+1)\sigma\} \\ & \leqslant 2\exp(-t) + 6\sqrt{2}\exp\left(\tfrac{-n\sigma^2}{4G^2}\right). \end{aligned} \quad (17)$$

The above equation can be decomposed the error into two part:

$$\begin{aligned} & P\{|\tilde{g}_i(w_s) - g_i(w_s)| \geqslant (6t+1)\cdot\sigma\} \\ & \quad \leqslant P\{|\tilde{g}_i(w_s) - g_i^t(w_s))| \geqslant \sigma \cdot 6t\} \\ & \qquad + P\{|g_i^t(w_s) - g_i(w_s)| \geqslant \sigma\}. \end{aligned} \quad (18)$$

There are two types of error we need to control. The first type results from the first term in Equation 18: the deviation between the differentially private estimate gradient $\tilde{g}_i(w_s)$ and the empirical gradient $g_i^t(w_s)$. The second type is the second term in Equation 18: the deviation between empirical $g_i^t(w_s)$ and the population gradient $g_i(w_s)$. Lemma 4.2 has already give the bound of the second type.

We now bound the first term of Equation 18 by considering two cases, depending on whether StGD answer the estimated gradient $\tilde{g}_i(w_s)$ by returning $\tilde{g}_i(w_s) = g_i^t(w_s) + \xi$ or by returning $\tilde{g}_i(w_s) = g_i^h(w_s)$. In the first case, we have

$$|\tilde{g}_i(w_s) - g_i^t(w_s)| = |\xi|.$$

In the second case we have

$$|\tilde{g}_i(w_s) - g_i^t(w_s)| \leqslant \gamma + \delta \leqslant |\gamma| + |\delta|.$$

Combining these two cases implies that

$$\begin{aligned} &P\{|\tilde{g}_i(w_s) - g_i^t(w_s)| \geqslant 6t\sigma\} \\ &\quad \leqslant \max\{P\{|\xi| \geqslant 6t\sigma\}, P\{|\gamma| + |\delta| \geqslant 6t\sigma\}\}. \end{aligned} \tag{19}$$

We note that the noise variables are chosen from Laplace distribution. By properties of the Laplace distribution, we have

$$P\{|\xi| \geqslant 6t\sigma\} \leqslant \exp(-6t), \tag{20}$$

and

$$\begin{aligned} &P\{|\gamma| + |\delta| \geqslant 6t\sigma\} \\ &\quad \leqslant P\{|\gamma| \geqslant 2t\sigma\} + P\{|\delta| \geqslant 4t\sigma\} \leqslant 2exp(-t). \end{aligned} \tag{21}$$

Combining Equation 20 and 21 with Equation 19, we have:

$$P\{|\tilde{g}_i(w_s) - g_i^t(w_s)| \geqslant 6t \cdot \sigma\} \leqslant 2\exp(-t). \tag{22}$$

Bring the Equation 22 and the result in Lemma 4.21 into Equation 18, we get Equation 17. Then, applying the union bound over all the coordinate $i = 1, ..., d$, we complete the proof of Theorem 1. □

B Proof of Theorem 2 and Theorem 3

We first prove that the convergence rate of a gradient-based iterative algorithm is related to the gradient concentration error ε and its iteration time T. Besides, the probability of the convergence rate is related to the probability δ of the gradient concentration bound. The details are given in the following theorem.

Theorem 6. *If there exists a gradient descent algorithm $\mathcal{A}$ with with T iterations and initial point w_0: For $s = 0, .., T$, $\mathcal{A}$ queries the training data at w_s to get a estimated gradient $\tilde{g}(w_s)$ such that $P\{\|\tilde{g}(w_s) - g(w_s)\|^2 \geq \varepsilon\} \leq \delta$ and updates $w_{s+1} = w_s + \eta_s \tilde{\nabla} F(w_s)$. For $\mathcal{A}$ we have the following.*

*1. F is L-**Lipschitz and** α-**strongly convex**: If we set the step size $\eta_s = \frac{2}{\alpha(s+1)}$, then we have the following excess risk bound:*

$$F(\bar{w}_T) - F(w^\star) \leqslant \frac{4L^2 \ln(T+1)}{4\alpha T} + \frac{\varepsilon}{\alpha} \tag{23}$$

with probability at least $1 - T\delta$.

*2. F is β **smooth and** α-**strongly convex**: If we set the step size $\eta = \frac{1}{\alpha+\beta}$, then we have the following excess risk bound:*

$$\begin{aligned} F(w_T) - F(w^\star) &\leqslant \tfrac{\beta}{2} exp\left(-\tfrac{\alpha\beta T}{(\alpha+\beta)^2}\right) \|w_1 - w^\star\|^2 \\ &\quad + \tfrac{2\alpha\beta + (\alpha+\beta)^2}{4\alpha^2\beta} \cdot \varepsilon \end{aligned} \tag{24}$$

with probability at least $1 - T\delta$.

Proof. In the case of L-Lipschitz and α-strongly convex function, with updating rule $w_{s+1} = w_s - \eta_s \cdot \tilde{g}(w_s)$ we have the following:

$$\begin{aligned} &F(w_s) - F(w^\star) \\ &\leqslant g(w_s)^T(w_s - w^\star) - \tfrac{\alpha}{2}\|w_s - w^\star\|^2 \\ &= (\tilde{g}(w_s) + g(w_s) - \tilde{g}(w_s))^T(w_s - w^\star) \\ &\quad - \tfrac{\alpha}{2}\|w_s - w^\star\|^2 \\ &= \tfrac{1}{\eta_s}(w_s - w_{s+1})^T(w_s - w^\star) \\ &\quad + (g(w_s) - \tilde{g}(w_s))^T(w_s - w^\star) \\ &\quad - \tfrac{\alpha}{2}\|w_s - w^\star\|^2 \\ &\leqslant \tfrac{1}{2\eta_s}(\|w_s - w^\star\|^2 + \eta_s^2\|\tilde{g}(w_s)\|^2 \\ &\quad - \|w_{s+1} - w^\star\|^2) \\ &\quad + \tfrac{\alpha}{4}(\tfrac{4}{\alpha^2}\|g(w_s) - \tilde{g}(w_s)\|^2 + \|w_s - w^\star\|^2) \\ &\quad - \tfrac{\alpha}{2}\|w_s - w^\star\|^2 \end{aligned}$$

$$
\begin{aligned}
&= (\tfrac{1}{2\eta_s} - \tfrac{\alpha}{4})\|w_s - w^\star\|^2 - \tfrac{1}{2\eta_s}\|w_{s+1} - w^\star\|^2 \\
&\quad + \tfrac{\eta_s}{2}\|\tilde{g}(w_s)\|^2 + \tfrac{1}{\alpha}\|g(w_s) - \tilde{g}(w_s)\|^2 \\
&\leqslant \tfrac{\alpha s}{4}\|w_s - w^\star\|^2 - \tfrac{\alpha(s+1)}{4}\|w_{s+1} - w^\star\|^2 \\
&\quad + \tfrac{L^2}{\alpha(s+1)} + \tfrac{1}{\alpha}\|g(w_s) - \tilde{g}(w_s)\|^2.
\end{aligned} \tag{25}
$$

Sum the above inequality over $s = 0$ to $s = T$, and apply Jensen's inequality:

$$
\begin{aligned}
&F(\hat{w}_n) - F(w^\star) \\
&\leqslant \tfrac{4L^2 \ln(T+1)}{4\alpha T} + \tfrac{\sum_{s=0}^{s=T} \|g(w_s) - \tilde{g}(w_s)\|^2}{(T+1)\alpha}.
\end{aligned} \tag{26}
$$

Bringing the condition that

$$
P\{\|\tilde{g}(w_s) - g(w_s)\|^2 \geq \varepsilon\} \leq \delta,
$$

into Equation 26, then applying the union bound over all $s = 0, ..., T$, we can complete the proof in the case of L-Lipschitz and α-strongly convex function.

Now we prove the Equation 24 in the case of the β smooth and α-strongly convex function.

First note that by β-smoothness, for all w_s, $w^\star \in \mathcal{W}$, one has

$$
F(w_s) - F(w^\star) \leqslant \frac{\beta}{2}\|w_s - w^\star\|. \tag{27}
$$

Now using Lemma 3.11 in (Bubeck [2015]): For β-smooth and α-strongly convex function F, $w, w' \in \mathcal{W}$, one has

$$
\begin{aligned}
(g(w) - g(w'))(w - w') \geqslant \tfrac{\alpha\beta}{\alpha+\beta}\|w - w'\|^2 \\
+ \tfrac{1}{\alpha+\beta}\|g(w) - g(w')\|^2.
\end{aligned} \tag{28}
$$

By this property, we obtain

$$
\begin{aligned}
&\|w_{s+1} - w^\star\| \\
&= \|w_s - \eta\tilde{g}(w_s) - w^\star\| \\
&= \|w_s - w^\star\|^2 - 2\eta\tilde{g}(w_s)^T(w_s - w^\star) \\
&\quad + \eta^2\|\tilde{g}(w_s)\|^2 \\
&= \|w_s - w^\star\|^2 - 2\eta(g(w_s) + \tilde{g}(w_s) \\
&\quad - g(w_s))^T(w_s - w^\star) \\
&\quad + \eta^2\|g(w_s) + \tilde{g}(w_s) - g(w_s)\|^2 \\
&\leqslant \|w_s - w^\star\|^2 - 2\eta(g(w_s))^T(w_s - w^\star) \\
&-2\eta\sqrt{\tfrac{\alpha+\beta}{\alpha\beta}}(\tilde{g}(w_s) - g(w_s))^T\sqrt{\tfrac{\alpha\beta}{\alpha+\beta}}(w_s - w^\star) \\
&\quad + 2\eta^2\|g(w_s)\|^2 + 2\eta^2\|\tilde{g}(w_s) - g(w_s)\|^2 \\
&\leqslant \|w_s - w^\star\|^2 - \tfrac{2\eta\alpha\beta}{\alpha+\beta}\|w_s - w^\star\|^2 - \tfrac{2\eta}{\alpha+\beta}\|g(w_s)\|^2 \\
&\quad + \eta\tfrac{\alpha+\beta}{\alpha\beta}\|\tilde{g}(w_s) - g(w_s)\|^2 + \eta\tfrac{\alpha\beta}{\alpha+\beta}\|w_s - w^\star\|^2 \\
&\quad + 2\eta^2\|g(w_s)\|^2 + 2\eta^2\|\tilde{g}(w_s) - g(w_s)\|^2 \\
&= (1 - \tfrac{\eta\alpha\beta}{\alpha+\beta})\|w_s - w^\star\|^2 + (2\eta^2 - \tfrac{2\eta}{\alpha+\beta})\|g(w_s)\|^2 \\
&\quad + (2\eta^2 + \eta\tfrac{\alpha+\beta}{\alpha\beta})\|\tilde{g}(w_s) - g(w_s)\|^2 \\
&= (1 - \tfrac{\alpha\beta}{(\alpha+\beta)^2})\|w_s - w^\star\|^2 \\
&\quad + \tfrac{2\alpha\beta + (\alpha+\beta)^2}{\alpha\beta(\alpha+\beta)^2}\|\tilde{g}(w_s) - g(w_s)\|^2.
\end{aligned} \tag{29}
$$

Applying the condition in Theorem 6 that

$$
P\{\|\tilde{g}(w_s) - g(w_s)\|^2 \geq \varepsilon\} \leq \delta,
$$

into the above equation as well as the union bound over all iterations we can complete the proof. □

Statement of Theorem 2: *For L-Lipschitz and α-strongly convex function F, given $2n$ available samples, set noise parameter $\sigma^2 = 4G^2\rho_{n,d}/\sqrt{n}$, step size $\eta_s = \frac{2}{\alpha(s+1)}$ and iteration time $T = \rho_{n,d}\sqrt{n}$ for StGD. Let $\hat{w}_n = \sum_{s=0}^{T} w_s/(T+1)$, StGD achieves:*

$$
F(\hat{w}_n) - F(w^\star) \leqslant O\left(\tfrac{\ln(\sqrt{n}\rho_{n,d})}{\sqrt{n}\rho_{n,d}}\right) + O\left(\tfrac{d\rho_{n,d}^3}{\sqrt{n}}\right), \tag{30}
$$

with probability at least $1 - O(\frac{\rho_{n,d}}{\sqrt{n}})$

Proof. First consider the gradient concentration bound of achieve by StGD (Theorem 1). Then, bringing the settings that the noise parameter $\sigma^2 = 4G^2(\ln n + \ln d)/\sqrt{n}$, the step size $\eta_s = \frac{2}{\alpha(s+1)}$ and the iteration time $T = (\ln n + \ln d)\sqrt{n}$ into the result in Theorem 1, Combining the results with Equation 23 in Theorem 6, we can complete the proof of this Theorem. □

Statement of Theorem 3: *For β-smooth and α-strongly convex function F, given $2n$ available samples, set noise parameter $\sigma^2 = \frac{\rho_{n,d}(4G^2\alpha+\beta)^2}{n\alpha\beta}$, step size $\eta = \frac{1}{\alpha+\beta}$ and iteration time $T = (\kappa + \frac{1}{\kappa} + 2)\rho_{n,d}$ where $\kappa = \beta/\alpha$. Let $\hat{w}_n = w_T$ be the output of StGD, we have the following excess risk bound:*

$$
F(\hat{w}_n) - F(w^\star) \leqslant O\left(\frac{\|w_1 - w^\star\|^2}{n}\right) + O\left(\frac{d\rho_{n,d}^3}{n}\right) \tag{31}
$$

with probability at least $1 - O\left(\frac{\rho_{n,d}}{n^4 d^3}\right)$.

Proof. First consider the gradient concentration bound of achieve by StGD (Theorem 1). Then, bringing the settings that the noise parameter $\sigma^2 = \frac{(4G^2\alpha+\beta)^2}{n\alpha\beta}(\ln n + \ln d)$, step size $\eta = \frac{1}{\alpha+\beta}$ and iteration time $T = (\kappa + \frac{1}{\kappa} + 2)(\ln n + \ln d)$ into the result in Theorem 1. C combining the results with Equation 24 in Theorem 6, we can complete the proof of this Theorem. □

C Proof of Theorem 4 and Theorem 5

Statement of Theorem 4: *Given* $2n$ *available samples, mini-batch SGD can achieve the following:*

1. F is L**-Lipschitz and** α**-strongly convex**: If we set the step size $\eta_s = \frac{2}{\alpha(s+1)}$, batch size $m = \sqrt{n}$ and iteration time $T = 2n/m$, output $\hat{w}_n = \sum_{s=1}^{T} w_s/T$ of mini-batch SGD satisfies:

$$F(\hat{w}_n) - F(w^\star) \leqslant O\left(\frac{\ln(\sqrt{n}+1)}{\sqrt{n}}\right) + O\left(\frac{d\ln\sqrt{n}}{\sqrt{n}}\right) \tag{32}$$

with probability at least $1 - d/\sqrt{n}$.

2. F is β **smooth and** α**-strongly convex**: If we set the step size $\eta = \frac{1}{\alpha+\beta}$, $m = \frac{\alpha\beta n}{(\alpha+\beta)^2 \ln n}$, $T = 2n/m$, output $\hat{w}_n = w_T$ of StGD satisfies:

$$F(\hat{w}_n) - F(w^\star) \leqslant O\left(\frac{\|w_1 - w^\star\|^2}{n}\right) + O\left(\frac{d\ln^2 n}{n}\right) \tag{33}$$

with probability $1 - O\left(\frac{\ln n}{n}\right)$.

Proof. By Hoeffding bound and union bound for every coordinate, at each iteration s, for batch size m and any $\mu > 0$, the gradient $\hat{g}(w_s)$ computed by mini-batch SGD enjoys

$$P\{\|\hat{g}(w_s) - g(w_s)\|^2 \leq d\mu^2\} \geq 1 - 2d\exp(\frac{-2m\mu^2}{4G^2}). \tag{34}$$

For L-Lipschitz and α-strongly convex function, bringing the settings that batch size $m = \sqrt{n}$ and $\mu = 2G(\ln\sqrt{n}/\sqrt{n})^{1/2}$ into Equation 34, then combining the result and settings that $\eta_s = \frac{2}{\alpha(s+1)}$ and iteration time $T = 2n/m$ with Equation 23 in Theorem 6, we can obtain Equation 32.

For β-smooth and α-strongly convex function bringing the settings that $m = \frac{\alpha\beta n}{(\alpha+\beta)^2 \ln n}$ into Equation 34, then combining the result and settings that $\eta = \frac{1}{\alpha+\beta}$, and iteration time $T = 2n/m$ with Equation 24 in Theorem 6, we can obtain Equation 33.

□

Statement of Theorem 5: *Given* $2n$ *available samples, mini-batch StGD can achieve the following:*

1. F is L**-Lipschitz and** α**-strongly convex**: If we set the step size $\eta_s = \frac{2}{\alpha(s+1)}$, batch size $m = \sqrt{n}$, $T = 2n/m$, noise parameter $\sigma^2 = 8G^2\ln n/\sqrt{n}$ and $T_1 = \ln n$, output $\hat{w}_n = \sum_{s=1}^{T} w_s/T$ of mini-batch StGD satisfies:

$$F(\hat{w}_n) - F(w^\star) \leqslant O\left(\frac{\ln(\sqrt{n}+1)}{\sqrt{n}\ln n}\right) + O\left(\frac{\ln^3 n}{\sqrt{n}}\right) \tag{35}$$

with probability at least $1 - d/\sqrt{n}$.

2. F is β **smooth and** α**-strongly convex**: If we set the step size $\eta = \frac{1}{\alpha+\beta}$, $m = \frac{\alpha\beta n}{(\alpha+\beta)^2 \ln n}$, $T = 2n/m$, $T_1 = \ln n$, noise parameter $\sigma^2 = \frac{4G^2(\alpha+\beta)^2(\ln n)^2}{\alpha\beta n}$, output $\hat{w}_n = w_T$ of mini-batch StGD satisfies:

$$F(\hat{w}_n) - F(w^\star) \leqslant O\left(\frac{\|w_1 - w^\star\|^2}{n^{\ln n}}\right) + O\left(\frac{d\ln^4 n}{n}\right) \tag{36}$$

with probability at least $1 - O\left(\frac{\ln^2 n}{n}\right)$.

Proof. When running StGD on each batch S_s, where $s \in 0, ..., T-1$, there are T_1 gradient computations to update w_{s+1}: With $\tilde{w}_0 = w_s$ being the initial point, sub-algorithm StGD updates $\tilde{w}_{k+1} = \tilde{w}_k + \eta\tilde{g}(\tilde{w}_k)$ for $k = 0, ..., T_1 - 1$ and $w_{s+1} = \tilde{w}_{T_1}$. For every batch of size m, $\tilde{g}(\tilde{w}_k)$ produced by StGD satisfy the gradient concentration bound in Theorem 1

$$\begin{aligned} P\{ \quad & \|\tilde{g}(\tilde{w}_k) - g(\tilde{w}_k)\|^2 \leqslant (6t+1)^2\sigma^2\} \\ & \leqslant 1 - 2d\exp(-t) - 6\sqrt{2}d\exp\left(\tfrac{-m\sigma^2}{4G^2}\right). \end{aligned} \tag{37}$$

Mini-batch StGD repeats StGD T times to go through T batches, then, the total gradient

computations is $T \cdot T_1$. For L-Lipschitz and α-strongly convex function, bringing the settings that batch size $m = \sqrt{n}$ and $\sigma^2 = 8G^2 \ln n / \sqrt{n}$ into Equation 37, then combining the result, the settings that $\eta_s = \frac{2}{\alpha(s+1)}$ and total gradient computations $T \cdot T_1 = 2\sqrt{n} \ln n$ with Equation 23 in Theorem 6, we can obtain Equation 35.

For β-smooth and α-strongly convex function bringing the settings that $m = \frac{\alpha\beta n}{(\alpha+\beta)^2 \ln n}$ and $\sigma^2 = \frac{4G^2(\alpha+\beta)^2(\ln n)^2}{\alpha\beta n}$ into Equation 37, then combining the result and settings that $\eta = \frac{1}{\alpha+\beta}$, and total gradient computations $T \cdot T_1 = \frac{2(\alpha+\beta)^2(\ln n)^2}{\alpha\beta}$ with Equation 24 in Theorem 6, we can obtain Equation 36. □