

A CONVERGENCE PROOF FOR ADAGRAD-WINDOW

In this section, we give the proof of convergence for the two versions of AdaGrad-window. To avoid the inversion of singular matrices, the second-moment matrix $\mathbf{V}_{t,i}$ is often added by $\delta \mathbf{I}_d$ where $\delta > 0$ is a small fixed constant and $\mathbf{I}_d \in \mathbb{R}^{d \times d}$ is the identity matrix. In the following, to avoid discussions of whether $\mathbf{V}_{t,i}$ is singular, we let $\mathbf{V}_{t,i} = \delta \mathbf{I}_d + \sum_{j=m-i+1}^m \mathbf{g}_{t-1,j}^2 + \sum_{j=1}^i \mathbf{g}_{t,j}^2$ where the small constant δ is less than $1/T$ where T is the total number of epochs.

The algorithm can be summarized as follows.

Algorithm 1: Full and Diagonal AdaGrad-window

Input: The step size $\eta \sim m^{-5/4}$, the iteration number in one epoch m .
1 **Initialization:** $\mathbf{x}_{1,1}$; random shuffle all samples to get a partition $\mathbb{B}_1, \dots, \mathbb{B}_m$; matrix $\mathbf{G} = 0^{d \times m}$ to save the gradients;
2 **for** $t \leftarrow 1$ **to** T **do**
3 $\eta_t = \eta / \sqrt{t}$;
4 **for** $i \leftarrow 1$ **to** m **do**
5 Calculate the mini-batch stochastic gradient $\mathbf{g}_{t,i} = \nabla f_{\mathbb{B}_i}(\mathbf{x}_{t,i})$;
6 $\mathbf{G}_{i,:} = \mathbf{g}_{t,i}$, $\mathbf{V}_{t,i} = \delta \mathbf{I}_d + \mathbf{G} \mathbf{G}^\top$;
7 Full: $\mathbf{x}_{t,i+1} = \mathbf{x}_{t,i} - \eta_t \mathbf{V}_{t,i}^{-\frac{1}{2}} \mathbf{g}_{t,i}$;
8 Diagonal: $\mathbf{x}_{t,i+1} = \mathbf{x}_{t,i} - \eta_t \text{diag}(\mathbf{V}_{t,i})^{-\frac{1}{2}} \mathbf{g}_{t,i}$;
9 **end**
10 Let $\mathbf{x}_{t+1,1} = \mathbf{x}_{t,m+1}$;
11 **end**
12 **return** $\mathbf{x} = \arg \min_{\mathbf{x} \in \{\mathbf{x}_{1,1}, \dots, \mathbf{x}_{T+1,1}\}} \|\nabla f(\mathbf{x})\|$

A.1 FULL ADAGRAD-WINDOW

A.1.1 USEFUL LEMMAS

Lemma 8. For any $t > 0$ and $1 \leq i \leq m$, we have

$$\|\mathbf{V}_{t,i}^{-1/2} \mathbf{g}_{t,i}\| \leq 1, \quad \|\mathbf{V}_{t,m}^{-1/2} \mathbf{g}_{t,i}\| \leq 1, \quad \|\mathbf{V}_{t,m}^{-1/2} \sum_{i=1}^m \mathbf{g}_{t,i}\| \leq \sqrt{m}.$$

Proof. Since $\mathbf{V}_{t,i} = \mathbf{H} + \mathbf{g}_{t,i}^2$ for some positive definite matrix $\mathbf{H} \succ 0$, by the Sherman-Morrison formula, we have

$$\mathbf{g}_{t,i}^\top \mathbf{V}_{t,i}^{-1} \mathbf{g}_{t,i} = \frac{\mathbf{g}_{t,i}^\top \mathbf{H}^{-1} \mathbf{g}_{t,i}}{1 + \mathbf{g}_{t,i}^\top \mathbf{H}^{-1} \mathbf{g}_{t,i}} \leq 1.$$

Similarly, we can prove $\|\mathbf{V}_{t,m}^{-1/2} \mathbf{g}_{t,i}\| \leq 1$. For the last inequality, since $\mathbf{V}_{t,m} = \delta \mathbf{I}_d + \mathbf{G}_t \mathbf{G}_t^\top$, we have:

$$\begin{aligned} \|\mathbf{V}_{t,m}^{-1/2} \sum_{i=1}^m \mathbf{g}_{t,i}\|^2 &= \mathbf{e}^\top \mathbf{G}_t^\top (\delta \mathbf{I}_d + \mathbf{G}_t \mathbf{G}_t^\top)^{-1} \mathbf{G}_t \mathbf{e} \\ &= \mathbf{e}^\top (\mathbf{I}_m - (\mathbf{I}_m + \delta^{-1} \mathbf{G}_t^\top \mathbf{G}_t)^{-1}) \mathbf{e} \\ &\leq \mathbf{e}^\top \mathbf{e} = m. \end{aligned}$$

□

Lemma 9. For any $t > 1$ and $1 \leq i \leq m$,

$$\|\mathbf{x}_{t-1,i} - \mathbf{x}_{t,i}\| \leq (m - i + 1)\eta_{t-1} + (i - 1)\eta_t.$$

Proof. First, we have:

$$\|\mathbf{x}_{t-1,i} - \mathbf{x}_{t,i}\| \leq \|\mathbf{x}_{t-1,i} - \mathbf{x}_{t,1}\| + \|\mathbf{x}_{t,1} - \mathbf{x}_{t,i}\|.$$

Next, there is:

$$\begin{aligned} \|\mathbf{x}_{t-1,i} - \mathbf{x}_{t,1}\| &= \left\| \sum_{j=i}^m (\mathbf{x}_{t-1,j+1} - \mathbf{x}_{t-1,j}) \right\| \\ &\leq \eta_{t-1} \sum_{j=i}^m \|(\mathbf{V}_{t-1,j})^{-1/2} \mathbf{g}_{t-1,j}\| \\ &\leq (m-i+1)\eta_{t-1}, \end{aligned}$$

where we use Lemma 8 in the last inequality. Similarly, we may have $\|\mathbf{x}_{t,1} - \mathbf{x}_{t,i}\| \leq (i-1)\eta_t$. Thus $\|\mathbf{x}_{t-1,i} - \mathbf{x}_{t,i}\| \leq (m-i+1)\eta_{t-1} + (i-1)\eta_t$. \square

A.1.2 PROOF OF LEMMA 2

Proof. Use the L-smooth assumption, we have

$$\begin{aligned} \|m\nabla f(\mathbf{x}_{t,1}) - \sum_{i=1}^m \mathbf{g}_{t,i}\| &= \left\| \sum_{i=1}^m (\nabla f_{\mathbb{B}_i}(\mathbf{x}_{t,1}) - \nabla f_{\mathbb{B}_i}(\mathbf{x}_{t,i})) \right\| \\ &\leq \sum_{i=1}^m \|\nabla f_{\mathbb{B}_i}(\mathbf{x}_{t,1}) - \nabla f_{\mathbb{B}_i}(\mathbf{x}_{t,i})\| \\ &\leq \sum_{i=1}^m L\|\mathbf{x}_{t,1} - \mathbf{x}_{t,i}\|, \end{aligned}$$

where $\nabla f_{\mathbb{B}_i}$ corresponds to the mini-batch gradient calculated using the mini batch of $\mathbf{g}_{t,i}$. From Lemma 9, we know $\|\mathbf{x}_{t,1} - \mathbf{x}_{t,i}\| \leq (i-1)\eta_t$. Thus, there is

$$\|\nabla f(\mathbf{x}_{t,1}) - \frac{1}{m} \sum_{i=1}^m \mathbf{g}_{t,i}\| \leq \frac{1}{m} \sum_{i=1}^m L\|\mathbf{x}_{t,i} - \mathbf{x}_{t,1}\| \leq \frac{(m-1)}{2} L\eta_t = \frac{1}{\sqrt{t}} \cdot \frac{\eta(m-1)L}{2}.$$

\square

A.1.3 PROOF OF LEMMA 3

In fact, we will prove a slight modification of Lemma 3 due to the presence of δ .

Lemma 3*. For any $t > 0$, we have either:

$$\left(\sum_{i=1}^m \mathbf{g}_{t,i} \right)^\top \mathbf{V}_{t,m}^{-1/2} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right) \geq \frac{1}{\sqrt{11/3 + 8 \cdot r^2}} \left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\|, \quad (12)$$

or $\|\nabla f(\mathbf{x}_{t,1})\| \leq \frac{1}{\sqrt{t}} \cdot (\eta m L)$ or $\left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\| \leq \sqrt{\delta}$. In addition, there is always $\left(\sum_{i=1}^m \mathbf{g}_{t,i} \right)^\top \mathbf{V}_{t,m}^{-1/2} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right) \geq 0$.

Proof. First, we will show that either:

$$\sum_{i=1}^m \|\mathbf{G}_t \mathbf{e}_i\|^2 \leq (8/3 + 8 \cdot r^2) \|\mathbf{G}_t \mathbf{e}\|^2,$$

or $\|\nabla f(\mathbf{x}_{t,1})\| \leq \frac{1}{\sqrt{t}} \cdot (\eta m L)$.

In the following, we suppose that $\|\nabla f(\mathbf{x}_{t,1})\| > \frac{1}{\sqrt{t}} \cdot (\eta m L)$, otherwise the proof is done. Using Lemma 2, we know that $\left\| \sum_{i=1}^m \mathbf{g}_{t,i} - m\nabla f(\mathbf{x}_{t,1}) \right\| \leq \frac{1}{\sqrt{t}} \cdot (\eta m^2 L/2) \leq \frac{m}{2} \|\nabla f(\mathbf{x}_{t,1})\|$. Therefore,

we have:

$$\begin{aligned}
\|\mathbf{G}_t \mathbf{e}\|^2 &= \|\mathbf{G}_t \mathbf{e} - m \nabla f(\mathbf{x}_{t,1}) + m \nabla f(\mathbf{x}_{t,1})\|^2 \\
&\geq [m \|\nabla f(\mathbf{x}_{t,1})\| - \|\mathbf{G}_t \mathbf{e} - m \nabla f(\mathbf{x}_{t,1})\|]^2 \\
&= [m \|\nabla f(\mathbf{x}_{t,1})\| - \|\sum_{i=1}^m \mathbf{g}_{t,i} - m \nabla f(\mathbf{x}_{t,1})\|]^2 \\
&\geq \frac{m}{4} \|\nabla f(\mathbf{x}_{t,1})\|^2.
\end{aligned}$$

On the other hand, there is:

$$\begin{aligned}
\sum_{i=1}^m \|\mathbf{G}_t \mathbf{e}_i\|^2 &= \sum_{i=1}^m \|\mathbf{g}_{t,i}\|^2 \\
&\leq \sum_{i=1}^m (\|\mathbf{g}_{t,i} - \nabla f_i(\mathbf{x}_{t,1})\| + \|\nabla f_i(\mathbf{x}_{t,1})\|)^2 \\
&\leq \sum_{i=1}^m 2 \cdot (\|\mathbf{g}_{t,i} - \nabla f_i(\mathbf{x}_{t,1})\|^2 + \|\nabla f_i(\mathbf{x}_{t,1})\|^2) \\
&\leq 2 \sum_{i=1}^m (L^2 \|\mathbf{x}_{t,i} - \mathbf{x}_{t,1}\|^2 + \|\nabla f_i(\mathbf{x}_{t,1})\|^2) \\
&\leq 2 \sum_{i=1}^m (L^2 (i-1)^2 \eta^2 / t + \|\nabla f_i(\mathbf{x}_{t,1})\|^2) \\
&\leq 2L^2 m^3 \eta^2 / (3t) + 2 \sum_{i=1}^m \|\nabla f_i(\mathbf{x}_{t,1})\|^2 \\
&\leq (2m/3) \|\nabla f(\mathbf{x}_{t,1})\|^2 + 2 \sum_{i=1}^m \|\nabla f_i(\mathbf{x}_{t,1})\|^2
\end{aligned}$$

Using the strong growth condition, we have:

$$\begin{aligned}
\sum_{i=1}^m \|\mathbf{G}_t \mathbf{e}_i\|^2 &\leq (2m/3) \|\nabla f(\mathbf{x}_{t,1})\|^2 + 2 \sum_{i=1}^m \|\nabla f_i(\mathbf{x}_{t,1})\|^2 \\
&\leq (2m/3 + 2mr^2) \|\nabla f(\mathbf{x}_{t,1})\|^2 \\
&\leq (8/3 + 8r^2) \|\mathbf{G}_t \mathbf{e}\|^2.
\end{aligned}$$

Since $\mathbf{g}_{t,i}^2 \preceq \|\mathbf{g}_{t,i}\|^2 \mathbf{I}_d = \|\mathbf{G}_t \mathbf{e}_i\|^2 \mathbf{I}_d$, we have $\mathbf{V}_{t,m} \preceq (\delta + \sum_{i=1}^m \|\mathbf{G}_t \mathbf{e}_i\|^2) \mathbf{I}_d$. Thus

$$\left(\sum_{i=1}^m \mathbf{g}_{t,i}\right)^\top \mathbf{V}_{t,m}^{-1/2} \left(\sum_{i=1}^m \mathbf{g}_{t,i}\right) \geq \frac{\|\mathbf{G}_t \mathbf{e}\|^2}{\sqrt{\delta + \sum_{i=1}^m \|\mathbf{G}_t \mathbf{e}_i\|^2}}.$$

From equation (12), we have:

$$\delta + \sum_{i=1}^m \|\mathbf{G}_t \mathbf{e}_i\|^2 \leq \delta + (8/3 + 8r^2) \|\mathbf{G}_t \mathbf{e}\|^2.$$

Thus either $\|\mathbf{G}_t \mathbf{e}\| \leq \sqrt{\delta}$ or:

$$\left(\sum_{i=1}^m \mathbf{g}_{t,i}\right)^\top \mathbf{V}_{t,m}^{-1/2} \left(\sum_{i=1}^m \mathbf{g}_{t,i}\right) \geq \frac{\|\mathbf{G}_t \mathbf{e}\|}{\sqrt{11/3 + 8r^2}}.$$

□

A.1.4 PROOF OF LEMMA 4

In order to prove Lemma 4 we need the following lemma, which is a modification of Lemma 3.1 in (Schmitt, 1992).

Lemma 10. *Given any matrices $\mathbf{C} \in \mathbb{R}^{d \times m}$ such that $\mathbf{C} \succeq 0$ and vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{d \times 1}$, we have*

$$\|(\delta \mathbf{I}_d + \mathbf{C} + \mathbf{a}\mathbf{a}^\top)^{1/2} - (\delta \mathbf{I}_d + \mathbf{C} + \mathbf{b}\mathbf{b}^\top)^{1/2}\| \leq \frac{8}{\pi} \|\mathbf{a} - \mathbf{b}\|.$$

Proof. From section v 3.11 in (Kato, 1976), we have the integration formulation:

$$\begin{aligned} & (\delta \mathbf{I}_d + \mathbf{C} + \mathbf{a}\mathbf{a}^\top)^{1/2} - (\delta \mathbf{I}_d + \mathbf{C} + \mathbf{b}\mathbf{b}^\top)^{1/2} \\ &= \frac{1}{\pi} \int_0^\infty \sqrt{\lambda} (\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{a}\mathbf{a}^\top)^{-1} (\mathbf{a}\mathbf{f}^\top + \mathbf{f}\mathbf{b}^\top) (\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{b}\mathbf{b}^\top)^{-1} d\lambda \\ &= \frac{1}{\pi} \int_0^\infty \sqrt{\lambda} (\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{a}\mathbf{a}^\top)^{-1} \mathbf{a}\mathbf{f}^\top (\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{b}\mathbf{b}^\top)^{-1} d\lambda \\ &\quad + \frac{1}{\pi} \int_0^\infty \sqrt{\lambda} (\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{a}\mathbf{a}^\top)^{-1} \mathbf{f}\mathbf{b}^\top (\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{b}\mathbf{b}^\top)^{-1} d\lambda \end{aligned}$$

where $\mathbf{a}\mathbf{f}^\top + \mathbf{f}\mathbf{b}^\top = \mathbf{a}\mathbf{a}^\top - \mathbf{b}\mathbf{b}^\top$ with $\mathbf{f} := \mathbf{a} - \mathbf{b}$. We split the first integral into $[0, s] \cup [s, +\infty]$ with

$$s = \frac{\|\mathbf{a}\|}{\|(\delta \mathbf{I}_d + \mathbf{a}\mathbf{a}^\top)^{-1} \mathbf{a}\|}.$$

To bound the two parts, we first notice that

$$\|(\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{b}\mathbf{b}^\top)^{-1}\| \leq \|(\lambda \mathbf{I}_d + \delta \mathbf{I}_d)^{-1}\| = \frac{1}{\lambda + \delta},$$

and

$$\|\mathbf{b}^\top (\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{b}\mathbf{b}^\top)^{-1}\| \leq \|\mathbf{b}^\top (\delta \mathbf{I}_d + \mathbf{b}\mathbf{b}^\top)^{-1}\|,$$

since $\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{b}\mathbf{b}^\top \succeq \delta \mathbf{I}_d + \mathbf{b}\mathbf{b}^\top \succ 0$ leads to $(\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{b}\mathbf{b}^\top)^{-1} \preceq (\delta \mathbf{I}_d + \mathbf{b}\mathbf{b}^\top)^{-1}$. Similarly, there are

$$\|(\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{a}\mathbf{a}^\top)^{-1}\| \leq \frac{1}{\lambda + \delta},$$

and

$$\|(\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{a}\mathbf{a}^\top)^{-1} \mathbf{a}\| \leq \|(\delta \mathbf{I}_d + \mathbf{a}\mathbf{a}^\top)^{-1} \mathbf{a}\|.$$

For the first part, we have

$$\begin{aligned} & \left\| \int_0^s \sqrt{\lambda} (\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{a}\mathbf{a}^\top)^{-1} \mathbf{a}\mathbf{f}^\top (\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{b}\mathbf{b}^\top)^{-1} d\lambda \right\| \\ & \leq \|\mathbf{f}\| \cdot \|(\delta \mathbf{I}_d + \mathbf{a}\mathbf{a}^\top)^{-1} \mathbf{a}\| \cdot \int_0^s \frac{\sqrt{\lambda}}{\lambda + \delta} d\lambda \\ & \leq \|\mathbf{f}\|_2 \cdot \|(\delta \mathbf{I}_d + \mathbf{a}\mathbf{a}^\top)^{-1} \mathbf{a}\|_2 \cdot 2\sqrt{s}. \end{aligned}$$

For the second part, we have

$$\begin{aligned} & \left\| \int_s^{+\infty} \sqrt{\lambda} (\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{a}\mathbf{a}^\top)^{-1} \mathbf{a}\mathbf{f}^\top (\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{b}\mathbf{b}^\top)^{-1} d\lambda \right\| \\ & \leq \|\mathbf{f}\| \cdot \|\mathbf{a}\| \cdot \int_s^\infty \frac{\sqrt{t}}{(t + \delta)^2} d\lambda \\ & \leq \|\mathbf{f}\| \cdot \|\mathbf{a}\| \cdot \frac{2}{\sqrt{s}}. \end{aligned}$$

Thus, we have

$$\begin{aligned} & \frac{1}{\pi} \int_0^\infty \sqrt{\lambda} (\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{a}\mathbf{a}^\top)^{-1} \mathbf{a}\mathbf{f}^\top (\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{b}\mathbf{b}^\top)^{-1} d\lambda \\ & \leq 4\|\mathbf{a}\|^{1/2} \cdot \|(\delta \mathbf{I}_d + \mathbf{a}\mathbf{a}^\top)^{-1} \mathbf{a}\|^{1/2} \cdot \|\mathbf{f}\|. \end{aligned}$$

Similarly, there is

$$\begin{aligned} & \frac{1}{\pi} \int_0^\infty \sqrt{\lambda} (\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{a}\mathbf{a}^\top)^{-1} \mathbf{f} \mathbf{b}^\top (\lambda \mathbf{I}_d + \delta \mathbf{I}_d + \mathbf{C} + \mathbf{b}\mathbf{b}^\top)^{-1} d\lambda \\ & \leq 4 \|\mathbf{b}\|^{1/2} \cdot \|\mathbf{b}^\top (\delta \mathbf{I}_d + \mathbf{b}\mathbf{b}^\top)^{-1}\|^{1/2} \cdot \|\mathbf{f}\|. \end{aligned}$$

Therefore, we may deduce

$$\begin{aligned} & \|(\delta \mathbf{I}_d + \mathbf{C} + \mathbf{a}\mathbf{a}^\top)^{1/2} - (\delta \mathbf{I}_d + \mathbf{C} + \mathbf{b}\mathbf{b}^\top)^{1/2}\| \\ & \leq \frac{4}{\pi} \|\mathbf{a}\|^{1/2} \cdot \|(\delta \mathbf{I}_d + \mathbf{a}\mathbf{a}^\top)^{-1} \mathbf{a}\|^{1/2} \cdot \|\mathbf{f}\| + \frac{4}{\pi} \|\mathbf{b}\|^{1/2} \cdot \|\mathbf{b}^\top (\delta \mathbf{I}_d + \mathbf{b}\mathbf{b}^\top)^{-1}\|^{1/2} \cdot \|\mathbf{f}\| \\ & \leq \frac{8}{\pi} \|\mathbf{a} - \mathbf{b}\|, \end{aligned}$$

where we use $\|\mathbf{a}\| \cdot \|(\delta \mathbf{I}_d + \mathbf{a}\mathbf{a}^\top)^{-1} \mathbf{a}\| \leq 1$ and $\|\mathbf{b}\| \cdot \|(\delta \mathbf{I}_d + \mathbf{b}\mathbf{b}^\top)^{-1} \mathbf{b}\| \leq 1$ in the last inequality. These two inequalities can be deduced using Sherman-Morrison formula: $\|\mathbf{a}\| \cdot \|(\delta \mathbf{I} + \mathbf{a}\mathbf{a}^\top)^{-1} \mathbf{a}\| = \delta^{-1} \mathbf{a}^\top \mathbf{a} / (1 + \delta^{-1} \mathbf{a}^\top \mathbf{a}) \leq 1$. \square

Proof of Lemma 4 By lemma 10, we have

$$\begin{aligned} & \|\mathbf{V}_{t,i}^{1/2} - \mathbf{V}_{t,m}^{1/2}\| \\ & = \|(\delta \mathbf{I}_d + \sum_{j=1}^i \mathbf{g}_{t,j}^2 + \sum_{j=0}^{m-i-1} \mathbf{g}_{t-1,m-j}^2)^{1/2} - (\delta \mathbf{I}_d + \sum_{j=1}^m \mathbf{g}_{t,j}^2)^{1/2}\| \\ & \leq \sum_{k=0}^{m-i-1} \|(\delta \mathbf{I}_d + \sum_{j=1}^{i+k} \mathbf{g}_{t,j}^2 + \sum_{j=0}^{m-i-1-k} \mathbf{g}_{t-1,m-j}^2)^{1/2} - (\delta \mathbf{I}_d + \sum_{j=1}^{i+k+1} \mathbf{g}_{t,j}^2 + \sum_{j=0}^{m-i-2-k} \mathbf{g}_{t-1,m-j}^2)^{1/2}\| \\ & \leq \sum_{k=0}^{m-i-1} \frac{8}{\pi} \|\mathbf{g}_{t-1,i+k+1} - \mathbf{g}_{t,i+k+1}\| \\ & = \frac{8}{\pi} \sum_{j=0}^{m-i-1} \|\mathbf{g}_{t-1,m-j} - \mathbf{g}_{t,m-j}\|, \end{aligned}$$

where we use lemma 10 in the second inequality $m-i$ times by setting $\mathbf{C} = \sum_{j=1}^{i+k} \mathbf{g}_{t,j}^2 + \sum_{j=0}^{m-i-2-k} \mathbf{g}_{t-1,m-j}^2$ and $\mathbf{a} = \mathbf{g}_{t-1,i+k+1}$, $\mathbf{b} = \mathbf{g}_{t,i+k+1}$.

By the L-smooth assumption and Lemma 2, we have

$$\begin{aligned} \sum_{j=0}^{m-i-1} \|\mathbf{g}_{t-1,m-j} - \mathbf{g}_{t,m-j}\| & \leq \sum_{j=0}^{m-i-1} L \|\mathbf{x}_{t-1,m-j} - \mathbf{x}_{t,m-j}\| \\ & \leq L \sum_{j=0}^{m-i-1} (j\eta_{t-1} + (m-j)\eta_t) \\ & = \frac{(m-i-1)(m-i)}{2} L\eta_{t-1} + \frac{(m-i)(m+i+1)}{2} L\eta_t. \end{aligned}$$

Thus, we deduce

$$\begin{aligned} \|\mathbf{V}_{t,i}^{1/2} - \mathbf{V}_{t,m}^{1/2}\| & = \frac{8}{\pi} \sum_{j=0}^{m-i-1} \|\mathbf{g}_{t-1,m-j} - \mathbf{g}_{t,m-j}\| \\ & \leq \frac{1}{\sqrt{t-1}} \cdot \frac{4\eta(m-i-1)(m-i)L}{\pi} + \frac{1}{\sqrt{t}} \cdot \frac{4\eta(m-i)(m+i+1)L}{\pi} \\ & \leq \frac{1}{\sqrt{t}} \cdot \left(\frac{6\eta(m-i-1)(m-i)L}{\pi} + \frac{4\eta(m-i)(m+i+1)L}{\pi} \right). \end{aligned}$$

\square

A.1.5 PROOF OF LEMMA 5

Proof. We divide it to two parts:

$$\begin{aligned} & \nabla f(\mathbf{x}_{t,1})^\top \left[\mathbf{V}_{t,m}^{-1/2} \sum_{i=1}^m \mathbf{g}_{t,i} - \sum_{i=1}^m \mathbf{V}_{t,i}^{-1/2} \mathbf{g}_{t,i} \right] \\ &= \left[f(\mathbf{x}_{t,1}) - \frac{1}{m} \sum_{i=1}^m \mathbf{g}_{t,i}^\top \right]^\top \left[\mathbf{V}_{t,m}^{-1/2} \sum_{i=1}^m \mathbf{g}_{t,i} - \sum_{i=1}^m \mathbf{V}_{t,i}^{-1/2} \mathbf{g}_{t,i} \right] + \frac{1}{m} \left(\sum_{i=1}^m \mathbf{g}_{t,i}^\top \right) \sum_{i=1}^m [\mathbf{V}_{t,m}^{-1/2} - \mathbf{V}_{t,i}^{-1/2}] \mathbf{g}_{t,i}. \end{aligned} \quad (13)$$

Using Lemma 2 and Lemma 8, the first part can be upper bounded by:

$$\begin{aligned} & \left[f(\mathbf{x}_{t,1}) - \frac{1}{m} \sum_{i=1}^m \mathbf{g}_{t,i}^\top \right]^\top \left[\mathbf{V}_{t,m}^{-1/2} \sum_{i=1}^m \mathbf{g}_{t,i} - \sum_{i=1}^m \mathbf{V}_{t,i}^{-1/2} \mathbf{g}_{t,i} \right] \\ & \leq \left\| f(\mathbf{x}_{t,1}) - \frac{1}{m} \sum_{i=1}^m \mathbf{g}_{t,i} \right\| \cdot \left(\left\| \mathbf{V}_{t,m}^{-1/2} \sum_{i=1}^m \mathbf{g}_{t,i} \right\| + \sum_{i=1}^m \left\| \mathbf{V}_{t,i}^{-1/2} \mathbf{g}_{t,i} \right\| \right) \\ & \leq \frac{1}{\sqrt{t}} \cdot \eta m^2 L. \end{aligned} \quad (14)$$

Using Lemma 4 and Lemma 8, the second part can be upper bounded by:

$$\begin{aligned} & \frac{1}{m} \left(\sum_{i=1}^m \mathbf{g}_{t,i}^\top \right) \sum_{i=1}^m [\mathbf{V}_{t,m}^{-1/2} - \mathbf{V}_{t,i}^{-1/2}] \mathbf{g}_{t,i} \\ &= \frac{1}{m} \sum_{j=1}^m (\mathbf{V}_{t,m}^{-1/2} \sum_{i=1}^m \mathbf{g}_{t,i}^\top) [\mathbf{V}_{t,j}^{1/2} - \mathbf{V}_{t,m}^{1/2}] \mathbf{V}_{t,j}^{-1/2} \mathbf{g}_{t,j} \\ & \leq \frac{1}{m} \sum_{j=1}^m \left\| \mathbf{V}_{t,m}^{-1/2} \sum_{i=1}^m \mathbf{g}_{t,i}^\top \right\| \cdot \left\| \mathbf{V}_{t,j}^{1/2} - \mathbf{V}_{t,m}^{1/2} \right\| \cdot \left\| \mathbf{V}_{t,j}^{-1/2} \mathbf{g}_{t,j} \right\| \\ & \leq \frac{1}{\sqrt{t}} \cdot \frac{1}{\sqrt{m}} \sum_{j=1}^m \left(\frac{6\eta(m-i-1)(m-i)L}{\pi} + \frac{4\eta(m-i)(m+i+1)L}{\pi} \right) \\ & \leq \frac{1}{\sqrt{t}} \cdot \frac{5\eta m^{5/2} L}{\pi} \end{aligned} \quad (15)$$

Plugging equation (14) and equation (15) into equation (13), we have:

$$\nabla f(\mathbf{x}_{t,1})^\top \left[\mathbf{V}_{t,m}^{-1/2} \sum_{i=1}^m \mathbf{g}_{t,i} - \sum_{i=1}^m \mathbf{V}_{t,i}^{-1/2} \mathbf{g}_{t,i} \right] \leq \frac{1}{\sqrt{t}} \cdot \left(\eta m^2 L + \frac{5\eta m^{5/2} L}{\pi} \right).$$

□

A.1.6 PROOF OF LEMMA 1

Again, we will prove a slight modification of Lemma 1 due to the presence of δ .

Lemma 1*. For any $t > 1$, denote $c_1 = \eta / \sqrt{11/3 + 8r^2}$, $c_2 = 5\eta^2 L m^2 / 2 + 5\eta^2 L m^{5/2} / \pi$ as constants independent of t . We have:

$$\frac{1}{\sqrt{t}} \cdot c_1 \|\nabla f(\mathbf{x}_{t,1})\| \leq f(\mathbf{x}_{t,1}) - f(\mathbf{x}_{t,m+1}) + \frac{1}{t} \cdot c_2.$$

or $\|\nabla f(\mathbf{x}_{t,1})\| \leq \frac{1}{\sqrt{t}} \cdot (\eta m L)$ or $\left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\| \leq \sqrt{\delta}$. In addition, there is always $0 \leq f(\mathbf{x}_{t,1}) - f(\mathbf{x}_{t,m+1}) + \frac{1}{t} \cdot c_2$.

Proof. As shown in section 4, there is:

$$\begin{aligned} \underbrace{\frac{\eta}{\sqrt{t}} \cdot \nabla f^\top(\mathbf{x}_{t,1}) \mathbf{V}_{t,m}^{-1/2} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right)}_{S1} &\leq f(\mathbf{x}_{t,1}) - f(\mathbf{x}_{t,m+1}) + \frac{L}{2} \|\mathbf{x}_{t,m+1} - \mathbf{x}_{t,1}\|^2 \\ &\quad + \underbrace{\frac{\eta}{\sqrt{t}} \cdot \nabla f(\mathbf{x}_{t,1})^\top \left[\mathbf{V}_{t,m}^{-1/2} \sum_{i=1}^m \mathbf{g}_{t,i} - \sum_{i=1}^m \mathbf{V}_{t,i}^{-1/2} \mathbf{g}_{t,i} \right]}_{S2}. \end{aligned} \quad (16)$$

From Lemma 9, we have:

$$\frac{L}{2} \|\mathbf{x}_{t,m+1} - \mathbf{x}_{t,1}\|^2 = \frac{L}{2} \|\mathbf{x}_{t+1,1} - \mathbf{x}_{t,1}\|^2 \leq \frac{1}{t} \cdot \frac{\eta^2 m^2 L}{2}. \quad (17)$$

From Lemma 5, we have:

$$\frac{\eta}{\sqrt{t}} \cdot S2 \leq \frac{1}{t} \cdot (\eta^2 m^2 L + \frac{5\eta^2 m^{5/2} L}{\pi}). \quad (18)$$

From Lemma 3*, Lemma 2 and Lemma 9, we have:

$$\begin{aligned} \frac{\eta}{\sqrt{t}} \cdot S1 &= \frac{\eta}{\sqrt{t}} \cdot \left[(\nabla f^\top(\mathbf{x}_{t,1}) - \frac{1}{m} \sum_{i=1}^m \mathbf{g}_{t,i}^\top) \mathbf{V}_{t,m}^{-1/2} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right) + \frac{1}{m} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right)^\top \mathbf{V}_{t,m}^{-1/2} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right) \right] \\ &\geq -\frac{\eta}{\sqrt{t}} \|\nabla f(\mathbf{x}_{t,1}) - \frac{1}{m} \sum_{i=1}^m \mathbf{g}_{t,i}\| \cdot \|\mathbf{V}_{t,m}^{-1/2} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right)\| + \frac{\eta}{m\sqrt{t}} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right)^\top \mathbf{V}_{t,m}^{-1/2} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right) \\ &\geq -\frac{\eta^2 m^{3/2} L}{2t} + \frac{\eta}{m\sqrt{11/3 + 8r^2} \cdot \sqrt{t}} \left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\| \\ &\geq -\frac{\eta^2 m^{3/2} L}{2t} + \frac{\eta}{m\sqrt{11/3 + 8r^2} \cdot \sqrt{t}} \left\| \sum_{i=1}^m \mathbf{g}_{t,i} - m\nabla f(\mathbf{x}_{t,1}) \right\| + \frac{\eta}{\sqrt{11/3 + 8r^2} \cdot \sqrt{t}} \|\nabla f(\mathbf{x}_{t,1})\| \\ &\geq -\frac{\eta^2 m^{3/2} L}{t} + \frac{\eta}{\sqrt{11/3 + 8r^2} \cdot \sqrt{t}} \|\nabla f(\mathbf{x}_{t,1})\| \\ &\geq -\frac{\eta^2 m^2 L}{t} + \frac{\eta}{\sqrt{11/3 + 8r^2} \cdot \sqrt{t}} \|\nabla f(\mathbf{x}_{t,1})\|, \end{aligned} \quad (19)$$

or $\|\nabla f(\mathbf{x}_{t,1})\| \leq \frac{1}{\sqrt{t}} \cdot (\eta m L)$ or $\|\sum_{i=1}^m \mathbf{g}_{t,i}\| \leq \sqrt{\delta}$.

Plugging equation (17), equation (18) and equation (19) into equation (16), we have:

$$\frac{1}{\sqrt{t}} \cdot \frac{\eta}{\sqrt{11/3 + 8r^2}} \|\nabla f(\mathbf{x}_{t,1})\| \leq f(\mathbf{x}_{t,1}) - f(\mathbf{x}_{t,m+1}) + \frac{1}{t} \cdot \left(\frac{5\eta^2 m^2 L}{2} + \frac{5\eta^2 m^{5/2} L}{\pi} \right),$$

or $\|\nabla f(\mathbf{x}_{t,1})\| \leq \frac{1}{\sqrt{t}} \cdot (\eta m L)$ or $\|\sum_{i=1}^m \mathbf{g}_{t,i}\| \leq \sqrt{\delta}$. In addition, from the first inequality in equation (19), we can easily see that there is always:

$$\frac{\eta}{\sqrt{t}} \cdot S1 \geq -\frac{\eta^2 m^{3/2} L}{2t} \geq -\frac{\eta^2 m^2 L}{t}.$$

Therefore there is always $0 \leq f(\mathbf{x}_{t,1}) - f(\mathbf{x}_{t,m+1}) + \frac{1}{t} \cdot c_2$. \square

A.1.7 PROOF OF MAIN THEOREM [1](#)

Proof. If there is a t where $\frac{T}{2} \leq t \leq T$ such that $\|\sum_{i=1}^m \mathbf{g}_{t,i}\| \leq \sqrt{\delta}$, using Lemma [2](#) and the fact that $\delta \leq 1/T$, we have

$$\begin{aligned} \|\nabla f(\mathbf{x}_{t,1})\| &\leq \|\nabla f(\mathbf{x}_{t,1}) - \frac{1}{m} \sum_{i=1}^m \mathbf{g}_{t,i}\| + \frac{1}{m} \|\sum_{i=1}^m \mathbf{g}_{t,i}\| \\ &\leq \frac{1}{\sqrt{t}} \cdot \frac{\eta(m-1)L}{2} + \frac{1}{m} \|\sum_{i=1}^m \mathbf{g}_{t,i}\| \\ &\leq \frac{(m-1)L\eta}{\sqrt{2T}} + \sqrt{\frac{\delta}{m^2}} \\ &\leq \frac{m^{5/4} \sqrt{11/3 + 8r^2}}{\sqrt{2T}} (f(\mathbf{x}_{1,1}) - f^* + G + 5L \ln T), \end{aligned}$$

which completes the proof.

If there is a t where $\frac{T}{2} \leq t \leq T$ such that $\|\nabla f(\mathbf{x}_{t,1})\| \leq \frac{1}{\sqrt{t}} \cdot (\eta mL)$, then it directly leads to:

$$\begin{aligned} \|\nabla f(\mathbf{x}_{t,1})\| &\leq \frac{1}{\sqrt{t}} \cdot (\eta mL) \\ &\leq \frac{m^{5/4} \sqrt{11/3 + 8r^2}}{\sqrt{2T}} (f(\mathbf{x}_{1,1}) - f^* + G + 5L \ln T), \end{aligned}$$

If for all t such that $\frac{T}{2} \leq t \leq T$ and $\|\nabla f(\mathbf{x}_{t,1})\| > \frac{1}{\sqrt{t}} \cdot (\eta mL)$, we have $\|\sum_{i=1}^m \mathbf{g}_{t,i}\|_2 > \sqrt{\delta}$, then from Lemma [1*](#) we have for all $\frac{T}{2} \leq t \leq T$,

$$\frac{1}{\sqrt{t}} \cdot \frac{\eta}{\sqrt{11/3 + 8r^2}} \|\nabla f(\mathbf{x}_{t,1})\| \leq f(\mathbf{x}_{t,1}) - f(\mathbf{x}_{t,m+1}) + \frac{1}{t} \cdot \left(\frac{5\eta^2 m^2 L}{2} + \frac{5\eta^2 m^{5/2} L}{\pi} \right),$$

In addition, we have for all $1 < t < \frac{T}{2}$, there is

$$0 \leq f(\mathbf{x}_{t,1}) - f(\mathbf{x}_{t,m+1}) + \frac{1}{t} \cdot \left(\frac{5\eta^2 m^2 L}{2} + \frac{5\eta^2 m^{5/2} L}{\pi} \right),$$

Summing up the above two inequalities for $t = 2, \dots, T$, we have

$$\begin{aligned} &\sqrt{2T} \cdot \frac{\eta}{\sqrt{11/3 + 8r^2}} \min_{\frac{T}{2} \leq t \leq T} \|\nabla f(\mathbf{x}_{t,1})\| \\ &\leq \frac{\eta}{\sqrt{11/3 + 8r^2}} \sum_{t=\lceil T/2 \rceil}^T \frac{1}{\sqrt{t}} \|\nabla f(\mathbf{x}_{t,i})\| \\ &\leq f(\mathbf{x}_{2,1}) - f(\mathbf{x}_{T,m+1}) + \sum_{t=2}^T \frac{1}{t} \cdot \left(\frac{5\eta^2 m^2 L}{2} + \frac{5\eta^2 m^{5/2} L}{\pi} \right) \\ &\leq f(\mathbf{x}_{2,1}) - f^* + \ln T \cdot \left(\frac{5\eta^2 m^2 L}{2} + \frac{5\eta^2 m^{5/2} L}{\pi} \right). \end{aligned}$$

Plugging $\eta = m^{-5/4}$, we have

$$\min_{\frac{T}{2} \leq t \leq T} \|\nabla f(\mathbf{x}_{t,1})\| \leq \frac{m^{5/4} \sqrt{11/3 + 8r^2}}{\sqrt{2T}} (f(\mathbf{x}_{2,1}) - f^* + 5L \ln T).$$

Since $\min_{1 \leq t \leq T} \|\nabla f(\mathbf{x}_{t,1})\| \leq \min_{\frac{T}{2} \leq t \leq T} \|\nabla f(\mathbf{x}_{t,1})\|$ and $f(\mathbf{x}_{2,1}) - f(\mathbf{x}_{1,1}) \leq G \|\mathbf{x}_{2,1} - \mathbf{x}_{1,1}\| \leq Gm\eta \leq G$ from the gradient bounded assumption and lemma [9](#) the proof is completed. \square

A.2 DIAGONAL ADAGRAD-WINDOW

Most parts in the analysis of the diagonal version is similar to the full version and we omit some proofs that are very easy to derive using same arguments as the full version.

A.2.1 LEMMAS

Lemma 11. For any $t > 0$ and $1 \leq i \leq m$, we have

$$\|\mathbf{V}_{t,i}^{-1/2} \mathbf{g}_{t,i}\| \leq \sqrt{d}, \quad \|\mathbf{V}_{t,m}^{-1/2} \mathbf{g}_{t,i}\| \leq \sqrt{d}, \quad \|\mathbf{V}_{t,m}^{-1/2} \sum_{i=1}^m \mathbf{g}_{t,i}\| \leq \sqrt{md}.$$

Proof. Since $\mathbf{V}_{t,i} = \mathbf{H} + \text{diag}(\mathbf{g}_{t,i}^2)$ for some positive definite diagonal matrix $\mathbf{H} \succ 0$, if $\mathbf{e}_j^\top \mathbf{V}_{t,i} \mathbf{e}_j = 0$, there must be $\mathbf{e}_j^\top \mathbf{g}_{t,i} = 0$. Thus, we have

$$\mathbf{g}_{t,i}^\top \mathbf{V}_{t,i}^{-1} \mathbf{g}_{t,i} \leq \sum_{j=1}^d 1 = d.$$

Similarly, we can prove $\|\mathbf{V}_{t,m}^{-1/2} \mathbf{g}_{t,i}\| \leq \sqrt{d}$. For the last inequality, we have:

$$\begin{aligned} \left(\sum_{i=1}^m \mathbf{g}_{t,i}\right)^\top \mathbf{V}_{t,m}^{-1} \left(\sum_{i=1}^m \mathbf{g}_{t,i}\right) &= \sum_{j=1}^d \frac{(\mathbf{e}_j^\top \mathbf{g}_{t,1} + \dots + \mathbf{e}_j^\top \mathbf{g}_{t,m})^2}{\delta + \|\mathbf{e}_j^\top \mathbf{g}_{t,1}\|^2 + \dots + \|\mathbf{e}_j^\top \mathbf{g}_{t,m}\|^2} \\ &\leq \sum_{j=1}^d \frac{m(\|\mathbf{e}_j^\top \mathbf{g}_{t,1}\|^2 + \dots + \|\mathbf{e}_j^\top \mathbf{g}_{t,m}\|^2)}{\delta + \|\mathbf{e}_j^\top \mathbf{g}_{t,1}\|^2 + \dots + \|\mathbf{e}_j^\top \mathbf{g}_{t,m}\|^2} \leq md \end{aligned}$$

□

Lemma 12. For any $t > 1$ and $1 \leq i \leq m$, we have:

$$\|\mathbf{x}_{t-1,i} - \mathbf{x}_{t,i}\| \leq (m-i+1)\sqrt{d}\eta_{t-1} + (i-1)\sqrt{d}\eta_t.$$

Proof. The proof is similar to Lemma 9 where we use Lemma 11 instead. □

Lemma 13. For any $t > 0$, we have:

$$\|\nabla f(\mathbf{x}_{t,1}) - \frac{1}{m} \sum_{i=1}^m \mathbf{g}_{t,i}\| \leq \frac{1}{\sqrt{t}} \cdot \frac{\eta(m-1)L\sqrt{d}}{2}. \quad (20)$$

Proof. The proof is similar to Lemma 2 where we use Lemma 12 instead. □

Lemma 14. For any $t > 0$, we have either:

$$\left(\sum_{i=1}^m \mathbf{g}_{t,i}\right)^\top \mathbf{V}_{t,m}^{-1/2} \left(\sum_{i=1}^m \mathbf{g}_{t,i}\right) \geq \frac{1}{\sqrt{11/3 + 8r^2}} \left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\|,$$

or $\|\nabla f(\mathbf{x}_{t,1})\| \leq \frac{1}{\sqrt{t}} \cdot (\eta m L)$ or $\left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\| \leq \sqrt{\delta/m}$.

Proof. Since $\mathbf{V}_{t,m} \preceq (\delta + \sum_{i=1}^m \|\mathbf{g}_{t,i}\|^2) \mathbf{I}_d$, we have:

$$\left(\sum_{i=1}^m \mathbf{g}_{t,i}\right)^\top \mathbf{V}_{t,m}^{-1/2} \left(\sum_{i=1}^m \mathbf{g}_{t,i}\right) \geq \frac{\left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\|^2}{\sqrt{\delta + \sum_{i=1}^m \|\mathbf{g}_{t,i}\|^2}}.$$

The rest of the proof is the same as Lemma 3*. □

Lemma 15. For any $t > 1$ and $1 \leq i \leq m$, we have:

$$\|\mathbf{V}_{t,m}^{1/2} - \mathbf{V}_{t,i}^{1/2}\| \leq \frac{1}{\sqrt{t}} \cdot \left(\frac{6\eta(m-i-1)(m-i)L\sqrt{d}}{\pi} + \frac{4\eta(m-i)(m+i+1)L\sqrt{d}}{\pi} \right). \quad (21)$$

Proof. Denote the diagonal matrix $\mathbf{D} = \delta \mathbf{I}_d + \text{diag}(\sum_{j=1}^i \mathbf{g}_{t,j}^2)$ and $\mathbf{A}_i = \mathbf{A}_{i,i}$ for any diagonal matrix \mathbf{A} , we have:

$$\begin{aligned}
& \|\mathbf{V}_{t,i}^{1/2} - \mathbf{V}_{t,m}^{1/2}\| \\
&= \|(\mathbf{D} + \text{diag}(\sum_{j=i+1}^m \mathbf{g}_{t-1,j}^2))^{1/2} - (\mathbf{D} + \text{diag}(\sum_{j=i+1}^m \mathbf{g}_{t,j}^2))^{1/2}\|_2 \\
&= \max_{1 \leq k \leq d} |(\mathbf{D}_k + \sum_{j=i+1}^m (\mathbf{e}_k^\top \mathbf{g}_{t-1,j})^2)^{1/2} - (\mathbf{D}_k + \sum_{j=i+1}^m (\mathbf{e}_k^\top \mathbf{g}_{t,j})^2)^{1/2}| \\
&= \max_{1 \leq k \leq d} \left| \frac{\sum_{j=i+1}^m (\mathbf{e}_k^\top \mathbf{g}_{t-1,j})^2 - (\mathbf{e}_k^\top \mathbf{g}_{t,j})^2}{(\mathbf{D}_k + \sum_{j=i+1}^m (\mathbf{e}_k^\top \mathbf{g}_{t-1,j})^2)^{1/2} + (\mathbf{D}_k + \sum_{j=i+1}^m (\mathbf{e}_k^\top \mathbf{g}_{t,j})^2)^{1/2}} \right| \\
&\leq \max_{1 \leq k \leq d} \sum_{j=i+1}^m \frac{|(\mathbf{e}_k^\top \mathbf{g}_{t-1,j})^2 - (\mathbf{e}_k^\top \mathbf{g}_{t,j})^2|}{|\mathbf{e}_k^\top \mathbf{g}_{t-1,j}| + |\mathbf{e}_k^\top \mathbf{g}_{t,j}|} \\
&\leq \sum_{j=i+1}^m \max_{1 \leq k \leq d} |\mathbf{e}_k^\top \mathbf{g}_{t-1,j} - \mathbf{e}_k^\top \mathbf{g}_{t,j}| \\
&\leq \sum_{j=i+1}^m \|\mathbf{g}_{t-1,j} - \mathbf{g}_{t,j}\|.
\end{aligned}$$

The rest can be deduced similar to lemma 4 where we use lemma 12 instead. \square

Lemma 16. For any $t > 1$, we have:

$$\nabla f(\mathbf{x}_{t,1})^\top \left[m \mathbf{V}_{t,m}^{-1/2} \nabla f(\mathbf{x}_{t,1}) - \sum_{i=1}^m \mathbf{V}_{t,i}^{-1/2} \mathbf{g}_{t,i} \right] \leq \frac{1}{\sqrt{t}} \cdot (5 \frac{\eta m^{5/2} L d^{3/2}}{\pi} + \eta m^2 L d). \quad (22)$$

Proof. The proof is similar to Lemma 5 where we use Lemma 11, Lemma 13 and Lemma 15 instead. \square

Lemma 17. For any $t > 1$, denote $c'_1 = \eta / \sqrt{11/3 + 8r^2}$, $c'_2 = 5\eta^2 m^2 L d^{3/2} / 2 + 5\eta^2 m^{5/2} L d / \pi$ as constants independent of t . We have either:

$$\frac{1}{\sqrt{t}} \cdot c'_1 \|\nabla f(\mathbf{x}_{t,1})\| \leq f(\mathbf{x}_{t,1}) - f(\mathbf{x}_{t,m+1}) + \frac{1}{t} \cdot c'_2.$$

or $\|\nabla f(\mathbf{x}_{t,1})\| \leq \frac{1}{\sqrt{t}} \cdot (\eta m L)$ or $\|\sum_{i=1}^m \mathbf{g}_{t,i}\| \leq \sqrt{\delta}$. In addition, there is always $0 \leq f(\mathbf{x}_{t,1}) - f(\mathbf{x}_{t,m+1}) + \frac{1}{t} \cdot c'_2$.

Proof. The proof is similar to Lemma 1* where we use lemmas corresponding to the diagonal version instead. \square

A.2.2 PROOF OF MAIN THEOREM 2

Proof. The proof is similar to Theorem 1 where we use the diagonal lemmas in the last section. \square

B MISSING PROOFS ON THE STRONG GROWTH CONDITION

B.1 PROOF OF LEMMA 6

In order to prove Lemma 6 we need the following results, which is a special case of the main result in (Milne, 2019).

Lemma (Milne, 2019). *There exists closed sets $B_i, i = 1, \dots, L$, such that:*

$$U = \bigcup_{i=1}^L B_i \cap U,$$

and smooth functions $\phi_i : V_i \rightarrow \mathbb{R}$ with $V_i \supset B_i \cap U$ open, satisfying:

$$f|_{B_i \cap U} = \phi_i|_{B_i \cap U},$$

and strong convexity $\nabla^2 \phi_i(\mathbf{x}) \succeq \frac{\lambda}{2} \mathbf{I}, \forall \mathbf{x} \in V_i$.

Proof of Lemma 6 For any set D , denote D° to be the set containing all the inner points of D . We will prove that all points in $U' = \bigcup_{i=1}^L (B_i \cap U)^\circ$ satisfy the strong growth condition with constant $\frac{2L}{\lambda}$.

By Lemma (Milne, 2019), we know that f is strongly convex with constant $\frac{\lambda}{2}$ in $(B_i \cap U)^\circ$. Therefore $\forall \mathbf{x}, \mathbf{y} \in (B_i \cap U)^\circ$, there is $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\lambda}{4} \|\mathbf{x} - \mathbf{y}\|^2$. We have:

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{x}^*) &\leq f(\mathbf{x}) - f(\mathbf{y}) \\ &\leq \nabla f(\mathbf{x})^\top (\mathbf{x} - \mathbf{y}) - \frac{\lambda}{4} \|\mathbf{x} - \mathbf{y}\|^2 \\ &\leq \frac{1}{\lambda} \|\nabla f(\mathbf{x})\|^2 + \frac{\lambda}{4} \|\mathbf{x} - \mathbf{y}\|^2 - \frac{\lambda}{4} \|\mathbf{x} - \mathbf{y}\|^2 \\ &= \frac{1}{\lambda} \|\nabla f(\mathbf{x})\|^2. \end{aligned}$$

On the other hand, from the L -smoothness, we have:

$$f_i(\mathbf{x}) - f_i(\mathbf{x} - \frac{1}{L} \nabla f_i(\mathbf{x})) + \frac{L}{2} \|\frac{1}{L} \nabla f_i(\mathbf{x})\|^2 \geq \nabla f_i(\mathbf{x})^\top (\frac{1}{L} \nabla f_i(\mathbf{x})).$$

Since \mathbf{x}^* minimize f_i as well, there is:

$$\begin{aligned} f_i(\mathbf{x}^*) &\leq f_i(\mathbf{x} - \frac{1}{L} \nabla f_i(\mathbf{x})) \\ &\leq f_i(\mathbf{x}) - \frac{1}{2L} \|\nabla f_i(\mathbf{x})\|^2. \end{aligned}$$

Thus,

$$\begin{aligned} \frac{1}{\lambda} \|\nabla f(\mathbf{x})\|^2 &\geq f(\mathbf{x}) - f(\mathbf{x}^*) \\ &= \frac{1}{n} \sum_{i=1}^n (f_i(\mathbf{x}) - f(\mathbf{x}^*)) \\ &\geq \frac{1}{n} \sum_{i=1}^n \frac{1}{2L} \|\nabla f_i(\mathbf{x})\|^2. \end{aligned}$$

Therefore \mathbf{x} satisfies the strong growth condition with constant $\frac{2L}{\lambda}$. The proof is complete. \square

B.2 PROOF OF LEMMA 7

Proof. Denote the data to be $\{(\mathbf{a}_i, o_i)\}_{i=1}^n$, where $o_i \in \{0, 1\}$. From the condition, we know that $\|\mathbf{a}_i\| \leq c$ and $\exists \mathbf{x}^*$ with $\|\mathbf{x}^*\| = 1$ such that $\mathbf{a}_i^\top \mathbf{x}^* \leq -\tau$ if $o_i = 0$ and $\mathbf{a}_i^\top \mathbf{x}^* \geq \tau$ if $o_i = 1$. Now

$\forall \mathbf{x}$, we have for the cross entropy loss:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(\mathbf{x})\|^2 &= \frac{1}{n} \sum_{i=1}^n \left(o_i - \frac{1}{e^{\mathbf{a}_i^\top \mathbf{x}} + 1}\right)^2 \|\mathbf{a}_i\|^2 \\
&\leq \frac{c^2}{n} \left[\sum_{o_i=1} \left(\frac{e^{\mathbf{a}_i^\top \mathbf{x}}}{e^{\mathbf{a}_i^\top \mathbf{x}} + 1}\right)^2 + \sum_{o_i=0} \left(\frac{1}{e^{\mathbf{a}_i^\top \mathbf{x}} + 1}\right)^2 \right] \\
&\leq \frac{c^2}{n} \left[\sum_{o_i=1} \left(\frac{e^{\mathbf{a}_i^\top \mathbf{x}}}{e^{\mathbf{a}_i^\top \mathbf{x}} + 1}\right) + \sum_{o_i=0} \left(\frac{1}{e^{\mathbf{a}_i^\top \mathbf{x}} + 1}\right) \right]^2 \\
&\leq \frac{c^2}{n\tau^2} \left[\sum_{o_i=1} \left(\frac{e^{\mathbf{a}_i^\top \mathbf{x}}}{e^{\mathbf{a}_i^\top \mathbf{x}} + 1}\right) \mathbf{a}_i^\top \mathbf{x}^* + \sum_{o_i=0} \left(\frac{1}{e^{\mathbf{a}_i^\top \mathbf{x}} + 1}\right) \mathbf{a}_i^\top \mathbf{x}^* \right]^2 \\
&= \frac{c^2}{n\tau^2} \left[\left(\sum_{o_i=1} \left(\frac{e^{\mathbf{a}_i^\top \mathbf{x}}}{e^{\mathbf{a}_i^\top \mathbf{x}} + 1}\right) \mathbf{a}_i + \sum_{o_i=0} \left(\frac{1}{e^{\mathbf{a}_i^\top \mathbf{x}} + 1}\right) \mathbf{a}_i \right)^\top \mathbf{x}^* \right]^2 \\
&\leq \frac{c^2}{n\tau^2} \left\| \sum_{o_i=1} \left(\frac{e^{\mathbf{a}_i^\top \mathbf{x}}}{e^{\mathbf{a}_i^\top \mathbf{x}} + 1}\right) \mathbf{a}_i + \sum_{o_i=0} \left(\frac{1}{e^{\mathbf{a}_i^\top \mathbf{x}} + 1}\right) \mathbf{a}_i \right\|^2 \cdot \|\mathbf{x}^*\|^2 \\
&= \frac{c^2}{\tau^2} \|\nabla f(\mathbf{x})\|^2.
\end{aligned}$$

The proof is complete. \square

C DETAIL OF EXPERIMENTS ON MNIST AND CIFAR-10

In both experiments, the loss function is taken to be the cross entropy. The step size η_t is set to η/\sqrt{t} for SGD, AdaGrad, AdaGrad-window and $\eta/t^{1/3}$ for SGD-shuffle, consistent with the theory. The constant η is chosen from $\{0.01, 0.1, 1.0\}$ separately based on test accuracy for each method. The small constant δ is set to default as 10^{-9} .

In the MNIST dataset, the logistic regression takes a 784-dimensional image vector as input and predicts a 10-dimensional class vector as output. Every epoch contains 50 iterations; each iteration uses a mini batch of size 1200. In the CIFAR-10 dataset, we use ResNet-18 (He et al., 2015) as our model where we do not include data augmentation and weight decay schemes in data preprocessing. very epoch contains 50 iterations; each iteration uses a mini batch of size 1000.

D CONVERGENCE PROOF FOR ADAGRAD-TRUNCATION

D.1 THE ANALYSIS OF ADAGRAD-TRUNCATION

In this section, we introduce an AdaGrad variants, Adagrad-truncation, and provide its theoretical analysis.

D.1.1 ALGORITHMS

In this section, we show Adagrad-truncation in Algorithm 2 for minimizing finite sum objective functions formulated as eq. 1. In the following sections, we provide an $\tilde{O}(1/\sqrt{T})$ convergence rate for both Adagrad-truncation and its diagonal version.

D.1.2 TECHNICAL LEMMAS

In this section, we show our technical lemmas for providing the convergence rate of Adagrad-truncation.

Lemma 18 (Sherman-Morrison formula). *Suppose $\mathbf{M} \in \mathbb{R}^{d \times d}$ is an invertible square matrix and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ are column vectors. Then $\mathbf{M} + \mathbf{u}\mathbf{v}^\top$ is invertible if and only if $1 + \mathbf{v}^\top \mathbf{M} \mathbf{u} \neq 0$. In this*

Algorithm 2: Adagrad-truncation

Input: The step size $\eta > 0$, the iteration number in one epoch m , the number of instances n

- 1 **Initialization:** $\mathbf{x}_{1,1}; \mathbf{g}_0 = 1/\sqrt{m}$; Random shuffle all samples;
- 2 **for** $t \leftarrow 1$ **to** T **do**
- 3 Initialize $\mathbf{x}_{t,1} = \mathbf{x}_{t-1,m+1}, \mathbf{g}_t = 0^{d \times 1}$;
- 4 **for** $i \leftarrow 1$ **to** m **do**
- 5 Calculate the mini-batch stochastic gradient $\mathbf{g}_{t,i} = \nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,i})$;
- 6 $\mathbf{g}_t = \mathbf{g}_t + \mathbf{g}_{t,i} \quad \mathbf{V}_{t,i} = m \|\mathbf{g}_{t-1}\|^2 \cdot \mathbf{I}$;
- 7 $\mathbf{x}_{t,i+1} = \mathbf{x}_{t,i} - \eta_t \cdot \mathbf{V}_{t,i}^{-\frac{1}{2}} \mathbf{g}_{t,i}$;
- 8 **end**
- 9 **end**
- 10 **return** $\mathbf{x} = \arg \min_{\mathbf{x} \in \mathbf{x}_{1,1}, \dots, \mathbf{x}_{T+1,1}} \|\nabla f(\mathbf{x})\|$

case,

$$(\mathbf{M} + \mathbf{u}\mathbf{v}^\top)^{-1} = \mathbf{M}^{-1} - \frac{\mathbf{M}^{-1}\mathbf{u}\mathbf{v}^\top\mathbf{M}^{-1}}{1 + \mathbf{v}^\top\mathbf{M}^{-1}\mathbf{u}} \quad (23)$$

Lemma 19 (Lemma 13 in [Duchi et al. \(2011\)](#)). Let $\mathbf{N} \succeq \mathbf{M} \succeq 0$ be symmetric $d \times d$ matrices. Then $\mathbf{N}^{1/2} \succeq \mathbf{M}^{1/2}$.

Proof. This lemma had been proved in [Duchi et al. \(2011\)](#), we include a proof for the convenience of readers. Let λ be a eigenvalue of $\mathbf{N}^{1/2} - \mathbf{M}^{1/2}$, corresponding to some eigenvector \mathbf{x} . Hence, we have $(\mathbf{N}^{1/2} - \lambda\mathbf{I})\mathbf{x} = \mathbf{M}^{1/2}\mathbf{x}$. Taking the inner product of both size with $\mathbf{x}^\top \mathbf{N}^{1/2}$, we have

$$\begin{aligned} \underbrace{\mathbf{x}^\top \mathbf{N} \mathbf{x} - \lambda \mathbf{x}^\top \mathbf{N}^{1/2} \mathbf{x}}_{S_1} &= \mathbf{x}^\top \mathbf{N}^{1/2} (\mathbf{N}^{1/2} - \lambda\mathbf{I}) \mathbf{x} \\ &= \mathbf{x}^\top \mathbf{N}^{1/2} \mathbf{M}^{1/2} \mathbf{x} \leq \|\mathbf{N}^{1/2} \mathbf{x}\| \|\mathbf{M}^{1/2} \mathbf{x}\| = \sqrt{\mathbf{x}^\top \mathbf{N} \mathbf{x} \cdot \mathbf{x}^\top \mathbf{M} \mathbf{x}} \leq \underbrace{\mathbf{x}^\top \mathbf{N} \mathbf{x}}_{S_2}. \end{aligned} \quad (24)$$

Thus, with $S_1 \leq S_2$ and $\mathbf{x}^\top \mathbf{N}^{1/2} \mathbf{x} \geq 0$, we obtain $\lambda \geq 0$ to complete the proof. \square

Lemma 20 (Conjugate Rule in [Halko et al. \(2011\)](#)). Suppose that $\mathbf{M} \succeq 0$. For every \mathbf{A} , the matrix $\mathbf{A}^* \mathbf{M} \mathbf{A} \succeq 0$ where \mathbf{A}^* means the conjugate transpose matrix of \mathbf{A} . In particular,

$$\mathbf{M} \preceq \mathbf{N} \implies \mathbf{A}^* \mathbf{M} \mathbf{A} \preceq \mathbf{A}^* \mathbf{N} \mathbf{A}. \quad (25)$$

Lemma 21. Let $\mathbf{N} \succeq \mathbf{M} \succ 0$ be symmetric $d \times d$ matrices. Then $\mathbf{N}^{-1} \preceq \mathbf{M}^{-1}$.

Proof. Since $\mathbf{N} \succeq \mathbf{M}$, we have $\mathbf{M}^{-\frac{1}{2}} \mathbf{N} \mathbf{M}^{-\frac{1}{2}} = (\mathbf{M}^{-\frac{1}{2}} \mathbf{N}^{\frac{1}{2}}) (\mathbf{N}^{\frac{1}{2}} \mathbf{M}^{-\frac{1}{2}}) \succeq \mathbf{I}$ because of Lemma [20](#). Commuting the product of two matrix does not change the eigenvalues, hence all eigenvalues of $\mathbf{N}^{\frac{1}{2}} \mathbf{M}^{-1} \mathbf{N}^{\frac{1}{2}}$ are larger than 1. Utilizing Lemma [20](#) again, then we obtain $\mathbf{M}^{-1} \succeq \mathbf{N}^{-1}$ to complete the proof. \square

D.1.3 CONVERGENCE RATE FOR THE ADAGRAD-TRUNCATION

In this section, we provide convergence rate analysis for Adagrad-truncation which have an $\tilde{O}(1/\sqrt{T})$ convergence rate for achieving FSPs on non-convex and random shuffling settings. Since the Cauchy-Schwarz inequality and the triangle inequality are used frequently, we will not provide them additional explanation in our proof.

In fact, we will prove a slight modification of Lemma [Lemma 21*](#) due to the presence of strong growth condition.

Lemma 21*. In Adagrad-Truncation, if the step size satisfies

$$\eta_t \leq \left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\| / (4mLr),$$

Then, we have

$$\left(\sum_{i=1}^m \|\mathbf{g}_{t,i}\| \right)^2 \leq m \cdot \sum_{i=1}^m \|\mathbf{g}_{t,i}\|^2 \leq 8m^2 r^2 \left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\|^2.$$

Proof. According to the Assumption (A3) for any \mathbf{x} , we have

$$\frac{1}{m} \sum_{i=1}^m \|\nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,1})\|^2 \leq r^2 \|\nabla f(\mathbf{x}_{t,1})\|^2. \quad (26)$$

According to the Cauchy-Schwarz inequality, we have

$$\|\nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,1})\|^2 \geq \frac{1}{2} \|\mathbf{g}_{t,i}\|^2 - \|\nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,1}) - \mathbf{g}_{t,i}\|^2. \quad (27)$$

Combining eq. 26 and eq. 27, we obtain

$$\begin{aligned} \frac{1}{2m} \sum_{i=1}^m \|\mathbf{g}_{t,i}\|^2 &\leq r^2 \|\nabla f(\mathbf{x}_{t,1})\|^2 + \frac{1}{m} \sum_{i=1}^m \|\nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,1}) - \mathbf{g}_{t,i}\|^2 \\ &\leq 2r^2 \left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\|^2 + 2r^2 \left\| \nabla f(\mathbf{x}_{t,1}) - \sum_{i=1}^m \mathbf{g}_{t,i} \right\|^2 \\ &\quad + \frac{1}{m} \sum_{i=1}^m \|\nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,1}) - \mathbf{g}_{t,i}\|^2 \\ &\leq 2r^2 \left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\|^2 + \underbrace{\left(2mr^2 + \frac{1}{m} \right) \sum_{i=1}^m \|\nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,1}) - \mathbf{g}_{t,i}\|^2}_{S1}. \end{aligned} \quad (28)$$

For each i in S1, we have

$$\begin{aligned} \|\nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,1}) - \mathbf{g}_{t,i}\| &\leq L \|\mathbf{x}_{t,i} - \mathbf{x}_{t,1}\| \leq L \sum_{j=1}^{i-1} \|\mathbf{x}_{t,j+1} - \mathbf{x}_{t,j}\| \\ &\stackrel{\textcircled{1}}{\leq} L m^{-0.5} \eta_t \frac{\sum_{j=1}^m \|\mathbf{g}_{t,j}\|}{\left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|}, \end{aligned} \quad (29)$$

where $\textcircled{1}$ follows from Lemma 22. Hence, we have

$$\begin{aligned} S1 &\leq \left(2mr^2 + \frac{1}{m} \right) \cdot m \cdot \left(L m^{-0.5} \eta_t \frac{\sum_{j=1}^m \|\mathbf{g}_{t,j}\|}{\left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|} \right)^2 \\ &\leq (2m^2 r^2 + 1) L^2 \eta_t^2 \cdot \frac{\sum_{j=1}^m \|\mathbf{g}_{t,j}\|^2}{\left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|^2}. \end{aligned} \quad (30)$$

Hence, if we require $\eta_t \leq \left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\| / (4m^{3/2} L r)$, we have $S1 \leq 1/4m \cdot \sum_{j=1}^m \|\mathbf{g}_{t,i}\|^2$. Plugging the result into eq. 28, we obtain

$$\sum_{i=1}^m \|\mathbf{g}_{t,i}\|^2 \leq 8mr^2 \left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\|^2.$$

Hence, the proof is completed. \square

Lemma 22. In Adagrad-truncation, for any $1 \leq t \leq T$ and $1 \leq i \leq m$, we have

$$\|\mathbf{x}_{t,i+1} - \mathbf{x}_{t,i}\| \leq \eta_t \cdot \frac{\|\mathbf{g}_{t,i}\|}{\sqrt{m} \left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|}$$

Proof. According to the update rule of Adagrad-truncation, we have

$$\|\mathbf{x}_{t,i+1} - \mathbf{x}_{t,i}\| = \left\| \eta_t \mathbf{V}_{t,i}^{-1/2} \mathbf{g}_{t,i} \right\| = \eta_t \sqrt{\mathbf{g}_{t,i}^\top \mathbf{V}_{t,i}^{-1} \mathbf{g}_{t,i}}. \quad (31)$$

If we set $\mathbf{V}_{t,i} = \hat{\mathbf{V}}_{t,i} + \mathbf{g}_{t,i} \mathbf{g}_{t,i}^\top$, then we have

$$\begin{aligned} \mathbf{g}_{t,i}^\top \mathbf{V}_{t,i}^{-1} \mathbf{g}_{t,i} &= \mathbf{g}_{t,i}^\top \left(\hat{\mathbf{V}}_{t,i} + \mathbf{g}_{t,i} \mathbf{g}_{t,i}^\top \right)^{-1} \mathbf{g}_{t,i} \\ &\stackrel{\textcircled{1}}{=} \mathbf{g}_{t,i}^\top \left(\hat{\mathbf{V}}_{t,i}^{-1} - \frac{\hat{\mathbf{V}}_{t,i}^{-1} \mathbf{g}_{t,i} \mathbf{g}_{t,i}^\top \hat{\mathbf{V}}_{t,i}^{-1}}{1 + \mathbf{g}_{t,i}^\top \hat{\mathbf{V}}_{t,i}^{-1} \mathbf{g}_{t,i}} \right) \mathbf{g}_{t,i} \xrightarrow{a := \mathbf{g}_{t,i}^\top \hat{\mathbf{V}}_{t,i}^{-1} \mathbf{g}_{t,i}} a - \frac{a^2}{1+a} \\ &\leq a = \mathbf{g}_{t,i}^\top \hat{\mathbf{V}}_{t,i}^{-1} \mathbf{g}_{t,i} \stackrel{\textcircled{2}}{\leq} \frac{\|\mathbf{g}_{t,i}\|^2}{m \left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|^2}. \end{aligned} \quad (32)$$

Hence, the proof is completed. \square

Lemma 23. In Adagrad-truncation, for any $t > 1$, if step sizes satisfy

$$\eta_t \leq \frac{\left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|}{4m^{3/2} Lr},$$

we have

$$\sum_{i=1}^m \|\mathbf{G}_{t,m} \mathbf{e}_i\| \leq 3mr \left\| \sum_{j=1}^m \mathbf{g}_{t,j} \right\| \leq 12mr \left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|.$$

Proof. According to the definition of \mathbf{G}_t and \mathbf{e} , we have

$$\|\mathbf{G}_t \mathbf{e}\| = \left\| \sum_{i=1}^m (\mathbf{g}_{t,i} - \mathbf{g}_{t-1,i} + \mathbf{g}_{t-1,i}) \right\| \leq \underbrace{\sum_{i=1}^m \|\mathbf{g}_{t,i} - \mathbf{g}_{t-1,i}\|}_{S_1} + \left\| \sum_{i=1}^m \mathbf{g}_{t-1,i} \right\|. \quad (33)$$

For each i in S_1 of eq. 33 if we set

$$\eta_t \leq \frac{\left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|}{4m^{3/2} Lr},$$

for any $t > 1$, we have

$$\begin{aligned} \|\mathbf{g}_{t,i} - \mathbf{g}_{t-1,i}\| &\stackrel{\textcircled{1}}{\leq} L \|\mathbf{x}_{t,i} - \mathbf{x}_{t-1,i}\| \leq L \left(\sum_{j=1}^{i-1} \|\mathbf{x}_{t,j+1} - \mathbf{x}_{t,j}\| + \sum_{j=i}^m \|\mathbf{x}_{t-1,j+1} - \mathbf{x}_{t-1,j}\| \right) \\ &\stackrel{\textcircled{2}}{\leq} Lm^{-1/2} \left(\frac{\eta_t \sum_{j=1}^{i-1} \|\mathbf{g}_{t,j}\|}{\left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|} + \frac{\eta_{t-1} \sum_{j=i}^m \|\mathbf{g}_{t-1,j}\|}{\left\| \sum_{j=1}^m \mathbf{g}_{t-2,j} \right\|} \right) \stackrel{\textcircled{3}}{\leq} \frac{1}{4rm^2} \cdot \sum_{j=1}^{i-1} \|\mathbf{g}_{t,j}\| + \frac{1}{4rm^2} \cdot \sum_{j=i}^m \|\mathbf{g}_{t-1,j}\| \\ &\leq \frac{1}{4rm^2} \cdot \sum_{j=1}^m \|\mathbf{g}_{t,j}\| + \frac{1}{4rm^2} \cdot \sum_{j=1}^m \|\mathbf{g}_{t-1,j}\| \stackrel{\textcircled{4}}{\leq} \frac{1}{4rm^2} \cdot \sum_{j=1}^m \|\mathbf{g}_{t,j}\| + \frac{1}{\sqrt{2}m} \cdot \left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|, \end{aligned} \quad (34)$$

where $\textcircled{1}$ follows from L-smoothness, $\textcircled{2}$ follows from Lemma 22, $\textcircled{3}$ follows from the selection of step size η_t , $\textcircled{4}$ follows from Lemma 21*. Hence, plugging eq 34 into eq 33, S_1 in eq 33 satisfies

$$S_1 = \sum_{i=1}^m \|\mathbf{g}_{t,i} - \mathbf{g}_{t-1,i}\| \leq \frac{1}{4rm} \sum_{j=1}^m \|\mathbf{g}_{t,j}\| + \frac{1}{\sqrt{2}} \left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|. \quad (35)$$

Hence, we obtain

$$\begin{aligned} \sum_{i=1}^m \|\mathbf{G}_t \mathbf{e}_i\| &\leq 3mr \|\mathbf{G}_t \mathbf{e}\| \stackrel{\textcircled{1}}{\leq} \frac{3}{4} \sum_{j=1}^m \|\mathbf{g}_{t,j}\| + 3mr \left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\| \\ &\stackrel{\textcircled{2}}{\leq} \frac{9mr}{4} \|\mathbf{G}_t \mathbf{e}\| + 3mr \left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|, \end{aligned} \quad (36)$$

where $\textcircled{1}$ follows from eq. 33 and eq. 35 $\textcircled{2}$ follows from Lemma D.1.3. Then, the proof is completed. \square

Corollary 3. In Adagrad-truncation, for any $t > 1$, if step sizes satisfy

$$\eta_t \leq \frac{\left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|}{4m^{3/2} Lr},$$

we have

$$\left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\| \leq 4 \left\| \sum_{j=1}^m \mathbf{g}_{t,j} \right\|.$$

Proof. This corollary can be easily extended from Lemma 23 \square

Lemma 24. In Adagrad-truncation, suppose that the Assumption (A3) holds. For any $t > 1$ and $1 \leq i \leq m$, if step sizes satisfy

$$\eta_t \leq \frac{\left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|}{4m^{3/2} Lr},$$

we have

$$\|\mathbf{x}_{t,i} - \mathbf{x}_{t,1}\|^2 \leq 128(i-1)\eta_t^2 r^2.$$

Proof. This lemma can be easily checked when $i = 1$, we only need to prove $i \geq 2$. We have

$$\begin{aligned} \|\mathbf{x}_{t,i} - \mathbf{x}_{t,1}\|^2 &= \left\| \sum_{j=1}^{i-1} (\mathbf{x}_{t,j+1} - \mathbf{x}_{t,j}) \right\|^2 \leq (i-1) \cdot \left(\sum_{j=1}^{i-1} \|\mathbf{x}_{t,j+1} - \mathbf{x}_{t,j}\|^2 \right) \\ &\stackrel{\textcircled{1}}{\leq} (i-1)\eta_t^2 \cdot \frac{\sum_{j=1}^{i-1} \|\mathbf{g}_{t,j}\|^2}{m \left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|^2} \leq (i-1)\eta_t^2 \cdot \frac{\sum_{j=1}^{i-1} \|\mathbf{g}_{t,j}\|^2}{m \left\| \sum_{j=1}^m \mathbf{g}_{t,j} \right\|^2} \cdot \frac{\left\| \sum_{j=1}^m \mathbf{g}_{t,j} \right\|^2}{\left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|^2} \\ &\stackrel{\textcircled{2}}{\leq} 128(i-1)\eta_t^2 r^2, \end{aligned} \quad (37)$$

where $\textcircled{1}$ follows from Lemma 22 and $\textcircled{2}$ follows from Lemma 20* and Lemma 23. \square

Lemma 25. In Adagrad-truncation, if the step sizes satisfy

$$\eta_t \leq \frac{\left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|}{2m} \cdot \min \left\{ \left(2m^{1/2} Lr \right)^{-1}, 1 \right\}.$$

Then, for any $t > 1$, we have

$$\frac{3\eta_t}{4m} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right)^\top \mathbf{V}_{t,m}^{-\frac{1}{2}} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right) \leq f(\mathbf{x}_{t,1}) - f(\mathbf{x}_{t,m+1}) + 32L^2\eta_t^2 m^{1/2} r^2 + 64L\eta_t^2 m r^2.$$

for the t -th epoch.

Proof. With the L -smoothness, we have

$$\begin{aligned}
f(\mathbf{x}_{t,m+1}) - f(\mathbf{x}_{t,1}) &\leq \nabla f^\top(\mathbf{x}_{t,1}) (\mathbf{x}_{t,m+1} - \mathbf{x}_{t,1}) + \frac{L}{2} \|\mathbf{x}_{t,m+1} - \mathbf{x}_{t,1}\|^2 \\
&= \nabla f^\top(\mathbf{x}_{t,1}) \left(\sum_{i=1}^m -\eta_t \mathbf{V}_{t,i}^{-1/2} \mathbf{g}_{t,i} \right) + \frac{L}{2} \|\mathbf{x}_{t,m+1} - \mathbf{x}_{t,1}\|^2 \\
&= \frac{L}{2} \|\mathbf{x}_{t,m+1} - \mathbf{x}_{t,1}\|^2 - \sum_{i=1}^m \eta_t \nabla f^\top(\mathbf{x}_{t,1}) \mathbf{V}_{t,m}^{-1/2} \mathbf{g}_{t,i} \\
&= \underbrace{\frac{L}{2} \|\mathbf{x}_{t,m+1} - \mathbf{x}_{t,1}\|^2}_{S_2} + \underbrace{\frac{\eta_t}{m} \left(\sum_{i=1}^m (\mathbf{g}_{t,i} - \nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,1})) \right)^\top \mathbf{V}_{t,m}^{-1/2} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right)}_{S_1} \\
&\quad - \frac{\eta_t}{m} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right)^\top \mathbf{V}_{t,m}^{-1/2} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right)
\end{aligned} \tag{38}$$

We next bound S_1 and S_2 separately. Before providing the upper bound of S_1 in eq. [38](#), if the step size in t -th epoch satisfies

$$\eta_t \leq \frac{\left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|}{2m},$$

we obtain

$$\eta_t \sqrt{m} \cdot \mathbf{I} \preceq \frac{\sqrt{m} \left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|}{2m} \mathbf{I} \stackrel{\textcircled{1}}{\preceq} \frac{1}{2m} \mathbf{V}_{t,m}^{1/2} \stackrel{\textcircled{2}}{\Longleftrightarrow} \eta_t^2 \sqrt{m} \mathbf{V}_{t,m}^{-1} \preceq \frac{\eta_{:,t}}{2m} \mathbf{V}_{t,m}^{-1/2}, \tag{39}$$

where $\textcircled{1}$ follows from Lemma [19](#) and $\textcircled{2}$ follows from Lemma [20](#). With such condition, we have S_1 in eq. [38](#) satisfies

$$\begin{aligned}
S_1 &\leq \frac{1}{2m^{5/2}} \left\| \sum_{i=1}^m (\mathbf{g}_{t,i} - \nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,1})) \right\|^2 + \frac{\sqrt{m} \eta_t^2}{2} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right)^\top \mathbf{V}_{t,m}^{-1} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right) \\
&\stackrel{\textcircled{1}}{\leq} 32L^2 \eta_t^2 m^{1/2} r^2 + \frac{\sqrt{m} \eta_t^2}{2} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right)^\top \mathbf{V}_{t,m}^{-1} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right) \\
&\stackrel{\textcircled{2}}{\leq} 32L^2 \eta_t^2 m^{1/2} r^2 + \frac{\eta_t}{4m} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right)^\top \mathbf{V}_{t,m}^{-1/2} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right),
\end{aligned} \tag{40}$$

where $\textcircled{1}$ follows from L -smoothness and Lemma [24](#), $\textcircled{2}$ follows from eq. [39](#).

For S_2 in eq. [38](#), we have

$$\eta_t \leq \frac{\left\| \sum_{j=1}^m \mathbf{g}_{t,j} \right\|}{4m^{3/2} Lr} \Rightarrow S_2 \stackrel{\textcircled{1}}{\leq} 64L\eta_t^2 m r^2, \tag{41}$$

where $\textcircled{1}$ follows from Lemma [24](#).

As a result, we plug eq. [40](#) and eq. [41](#) into eq. [38](#) and obtain

$$\begin{aligned}
\Delta_t &\leq -\frac{3\eta_t}{4m} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right)^\top \mathbf{V}_{t,m}^{-\frac{1}{2}} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right) + 32L^2 \eta_t^2 m^{1/2} r^2 \\
&\quad + 64L\eta_t^2 m r^2.
\end{aligned} \tag{42}$$

Hence, the proof is completed. \square

Lemma 26. In Adagrad-truncation, if the step sizes satisfy

$$\eta_t \leq \frac{\left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|}{2m} \cdot \min \left\{ (2m^{1/2} Lr)^{-1}, 1 \right\}.$$

Then, for any $t > 1$, we have

$$\left(\sum_{i=1}^m \mathbf{g}_{t,i} \right)^\top \mathbf{V}_{t,m}^{-1/2} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right) \geq \frac{1}{4} m^{-1/2} \cdot \left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\|$$

Proof. We have

$$\begin{aligned} \mathbf{V}_{t,m} &\preceq \left(m \left\| \sum_{i=1}^m \mathbf{g}_{t-1,i} \right\|^2 \right) \cdot \mathbf{I} \\ &\preceq m \left\| \sum_{i=1}^m \mathbf{g}_{t-1,i} \right\|^2 \cdot \mathbf{I} \stackrel{\textcircled{1}}{\preceq} m \left\| \sum_{i=1}^m \mathbf{g}_{t-1,i} \right\|^2 \cdot \mathbf{I} \stackrel{\textcircled{2}}{\preceq} 16m \left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\|^2, \end{aligned}$$

where $\textcircled{1}$ follows from Assumption [\(A3\)](#) and $\textcircled{2}$ follows from Corollary [3](#). With Lemma [19](#) and Lemma [21](#), we have

$$\mathbf{V}_{t,m}^{-1/2} \succeq \frac{1}{4} \cdot m^{-1/2} \left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\|^{-1} \cdot \mathbf{I}. \quad (43)$$

That means

$$\left(\sum_{i=1}^m \mathbf{g}_{t,i} \right)^\top \mathbf{V}_{t,m}^{-1/2} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right) \geq \frac{1}{4} m^{-1/2} \cdot \left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\| \quad (44)$$

□

Proof of Theorem [3](#) For epoch t and $t > 2$, we have

$$\begin{aligned} \left\| \sum_{i=1}^m \nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,1}) \right\| &\leq \left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\| + \left\| \sum_{i=1}^m (\mathbf{g}_{t,i} - \nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,1})) \right\| \\ &\leq \left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\| + \sum_{i=1}^m \left\| \mathbf{g}_{t,i} - \nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,1}) \right\| \stackrel{\textcircled{1}}{\leq} \left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\| + L \sum_{i=1}^m \|\mathbf{x}_{t,i} - \mathbf{x}_{t,1}\| \\ &\stackrel{\textcircled{2}}{\leq} \left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\| + L \sum_{i=1}^m \sqrt{128(i-1)\eta_t^2 r^2} \leq \left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\| + 8L\eta_t m^{3/2} r, \end{aligned} \quad (45)$$

where $\textcircled{1}$ follows from L -smoothness, and $\textcircled{2}$ follows from Lemma [24](#). Then, we obtain

$$\begin{aligned} &\eta_t \left\| \sum_{i=1}^m \nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,1}) \right\| \stackrel{\textcircled{1}}{\leq} \eta_t \left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\| + 8L\eta_t^2 m^{3/2} r \\ &= \eta_t \cdot 4m^{1/2} \cdot \frac{\left\| \sum_{i=1}^m \mathbf{g}_{t,i} \right\|}{4m^{1/2}} + 8L\eta_t^2 m^{3/2} r \\ &\stackrel{\textcircled{2}}{\leq} \eta_t \cdot 4m^{1/2} \cdot \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right)^\top \mathbf{V}_{t,m}^{-\frac{1}{2}} \left(\sum_{i=1}^m \mathbf{g}_{t,i} \right) + 8L\eta_t^2 m^{3/2} r \\ &\stackrel{\textcircled{3}}{\leq} \eta_t \cdot m^{1/2} \cdot \frac{6m}{\eta_t} \cdot \left[f(\mathbf{x}_{t,1}) - f(\mathbf{x}_{t,m+1}) + 32L^2 \eta_t^2 m^{1/2} r^2 \right. \\ &\quad \left. + 64L\eta_t^2 m r^2 \right] + 8L\eta_t^2 m^{3/2} r \\ &= 6m^{3/2} \cdot (f(\mathbf{x}_{t,1}) - f(\mathbf{x}_{t,m+1})) + 200r^2 m^2 L^2 \cdot \eta_t^2 \\ &\quad + 400r^2 m^{5/2} L \cdot \eta_t^2, \end{aligned} \quad (46)$$

where $\textcircled{1}$ follows from eq. [45](#), $\textcircled{2}$ follows from Lemma [26](#) and $\textcircled{3}$ follows from Lemma [25](#). Without loss of generality, we assume $2m^{1/2} Lr \geq 1$, and set the step size in epoch t as

$$\eta_t = \frac{\epsilon'}{4m^{3/2} Lr} \stackrel{\textcircled{1}}{\leq} \frac{\left\| \sum_{j=1}^m \mathbf{g}_{t-1,j} \right\|}{2m} \cdot \min \left\{ (2m^{1/2} Lr)^{-1}, 1 \right\},$$

Then, there exists

$$\begin{aligned}
& \frac{(T-2)\epsilon'}{4m^{3/2}Lr} \min_t \left\{ \left\| \frac{1}{m} \sum_{i=1}^m \nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,1}) \right\| \right\} \leq \sum_{t=3}^T \frac{\eta_t}{m} \left\| \sum_{i=1}^m \nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,1}) \right\| \\
& \stackrel{\textcircled{1}}{\leq} 6m^{1/2} (f(\mathbf{x}_{3,1}) - f^*) + \sum_{t=3}^T 200r^2 m L^2 \cdot \eta_t^2 + \sum_{t=3}^T 400r^2 m^{3/2} L \cdot \eta_t^2 \\
& \leq 6m^{1/2} (f(\mathbf{x}_{3,1}) - f^*) + (200r^2 L^2 m^{-2} + 400r^2 m^{-3/2} L) T (\epsilon')^2 / (16L^2 r^2). \quad (47)
\end{aligned}$$

For $t = 1$, with gradient bounded,

$$f(\mathbf{x}_{t,1}) - f(\mathbf{x}_{t+1,1}) \leq G \|\mathbf{x}_{t,1} - \mathbf{x}_{t+1,1}\| \leq G^2. \quad (48)$$

where $\textcircled{1}$ follows from Lemma 22, $\textcircled{2}$ follows from the selection of η_t and the setting $\epsilon \leq 1$ without loss of generality. With the fact $m, r \geq 1$, we have

$$\begin{aligned}
& \frac{(T-2)\epsilon'}{4m^{3/2}Lr} \min_t \left\{ \left\| \frac{1}{m} \sum_{i=1}^m \nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,1}) \right\| \right\} \\
& \leq 6m^{1/2} (f(\mathbf{x}_{1,1}) - f^* + G^2) + (200r^2 L^2 m^{-2} + 400r^2 m^{-3/2} L) T (\epsilon')^2 / (16L^2 r^2), \quad (49)
\end{aligned}$$

through combining eq. 47 and eq. 48. If we set $T \geq 4$ and $L \geq 1$, it can be easily verified that $T - 2 \geq T/2$ and $2Lr \geq Lr + 1$. Hence, rearrange the terms in eq. 47, we obtain

$$\begin{aligned}
\min_t \left\{ \left\| \frac{1}{m} \sum_{i=1}^m \nabla f_{\mathbb{B}_{t,i}}(\mathbf{x}_{t,1}) \right\| \right\} & \leq 48rm^2 (f(\mathbf{x}_{1,1}) - f^* + G^2) LT^{-1} (\epsilon')^{-1} \\
& + (100rLm^{-1/2} + 200r)(\epsilon'). \quad (50)
\end{aligned}$$

For achieving FSPs, i.e., $\min_t \{\|\nabla f(\mathbf{x}_{t,1})\|\} = O(\epsilon)$, we require

$$\begin{cases} (100rLm^{-1/2} + 200r)(\epsilon') \leq \epsilon \\ 48rm^2 (f(\mathbf{x}_{1,1}) - f^* + G^2) LT^{-1} (\epsilon')^{-1} \leq \epsilon, \end{cases} \quad (51)$$

Plugging eq. 51 to eq. 50, and setting

$$\eta_t = \frac{\sqrt{3}}{10m^{1/2}Lr} \cdot \sqrt{\frac{f(\mathbf{x}_{1,1}) - f^* + G^2}{L+2}} \cdot \frac{1}{\sqrt{T}} \quad (52)$$

we have

$$\min_t \{\|\nabla f(\mathbf{x}_{t,1})\|\} \leq 80m(Lr+2) \sqrt{f(\mathbf{x}_{1,1}) - f^* + G^2} \cdot \frac{1}{\sqrt{T}} \quad (4)$$

Hence, the proof is completed. \square