Supplementary File for One Gradient Frank-Wolfe for Decentralized Online Convex and Submodular Optimization

Appendix A. Theoretical Analysis for Section 3

In the analysis, we note $\sigma_q(k)$ to be the permutation of k at phase q. We define the average function of the remaining $(K-k)$ functions as

$$\bar{F}_{q,k}(\boldsymbol{x}) = \frac{1}{K-k}\sum_{\ell=k+1}^{K} F_{\sigma_q(\ell)}(\boldsymbol{x}) = \frac{1}{K-k}\sum_{\ell=k+1}^{K}\frac{1}{n}\sum_{i=1}^{n} f^i_{\sigma_q(\ell)}(\boldsymbol{x}) \tag{10}$$

where $F_{\sigma_q(\ell)}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} f^i_{\sigma_q(\ell)}(\boldsymbol{x})$. We also define

$$\hat{f}^i_{q,k} = \frac{1}{K-k}\sum_{\ell=k+1}^{K} f^i_{\sigma_q(\ell)}(\boldsymbol{x}^i_{q,\ell}), \quad \nabla\hat{f}^i_{q,k} = \frac{1}{K-k}\sum_{\ell=k+1}^{K}\nabla f^i_{\sigma_q(\ell)}(\boldsymbol{x}^i_{q,\ell}) \tag{11}$$

as the average of the remaining $(K-k)$ functions and stochastic gradients of $f^i_{\sigma_q(\ell)}(\boldsymbol{x}^i_{q,\ell})$ respectively. Then we note,

$$\hat{F}_{q,k} = \frac{1}{n}\sum_{i=1}^{n}\hat{f}^i_{q,k}, \quad \nabla\hat{F}_{q,k} = \frac{1}{n}\sum_{i=1}^{n}\nabla\hat{f}^i_{q,k}, \tag{12}$$

In the same spririt of $\hat{f}^i_{q,k}$, we define

$$\hat{\boldsymbol{g}}^i_{q,k} = \frac{1}{K-k}\sum_{\ell=k+1}^{K}\boldsymbol{g}^i_{q,\ell}, \quad \hat{\boldsymbol{d}}^i_{q,k} = \frac{1}{K-k}\sum_{\ell=k+1}^{K}\boldsymbol{d}^i_{q,\ell} \tag{13}$$

In the rest of the analysis, we let $\mathcal{F}_{q,1}\subset\cdots\subset\mathcal{F}_{q,k}$ to be the σ-field generated by the permutation up to time k and $\mathcal{H}_{q,1}\subset\cdots\subset\mathcal{H}_{q,k}$ another σ-field generated by the randomness of the stochastic gradient estimate up to time k.

Assumption 5 *Let $\left\{\widetilde{\boldsymbol{d}}_t\right\}_1^T$ be a sequence such that $\mathbb{E}\left[\widetilde{\boldsymbol{d}}_t \mid \mathcal{H}_{t-1}\right] = \boldsymbol{d}_t$ where $\mathcal{H}_{t-1}$ is the filtration of the stochastic estimate up to $t-1$.*

Lemma 8 (Fact 1, Wai et al. (2017)) *Let $\boldsymbol{x}^1,\ldots,\boldsymbol{x}^n\in\mathbb{R}^d$ be a set of vector and and $\boldsymbol{x}_{avg} := \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$ their average. Let $\mathbf{W}$ be non-negative doubly stochastic matrix. The output of one round of the average consensus update $\bar{\boldsymbol{x}}^i = \sum_{j=1}^{n} W_{ij}\boldsymbol{x}^j$ satisfy :*

$$\sqrt{\sum_{i=1}^{n}\left\|\bar{\boldsymbol{x}}^i - \boldsymbol{x}_{avg}\right\|^2} \le |\lambda_2(\mathbf{W})|\cdot\sqrt{\sum_{i=1}^{n}\left\|\boldsymbol{x}^i - \boldsymbol{x}_{avg}\right\|^2}$$

where $\lambda_2(\mathbf{W})$ is the second largest eigenvalue of $\mathbf{W}$.

Lemma 9 *For all $1 \leq q \leq Q$ and $1 \leq k \leq K$, we have*

$$\overline{\boldsymbol{x}}_{q,k+1} = \overline{\boldsymbol{x}}_{q,k} + \eta_k \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{v}_{q,k}^i - \overline{\boldsymbol{x}}_{q,k} \right) \tag{14}$$

for convex case, and

$$\overline{\boldsymbol{x}}_{q,k+1} = \overline{\boldsymbol{x}}_{q,k} + \frac{1}{K} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{v}_{q,k}^i \right) \tag{15}$$

for submodular case.

Proof

$$\begin{aligned}
\overline{\boldsymbol{x}}_{q,k+1} &= \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_{q,k+1}^i && \text{(Definition of } \overline{\boldsymbol{x}}_{q,k+1}\text{)} \\
&= \frac{1}{n} \sum_{i=1}^{n} \big((1-\eta_k) \boldsymbol{y}_{q,k}^i + \eta_k \boldsymbol{v}_{q,k}^i \big) && \text{(Definition of } \boldsymbol{x}_{q,k}^i\text{)} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[(1-\eta_k) \left(\sum_{j=1}^{n} \boldsymbol{W}_{ij} \boldsymbol{x}_{q,k}^j \right) + \eta_k \boldsymbol{v}_{q,k}^i \right] && \text{(Definition of } \boldsymbol{y}_{q,k}^i\text{)} \\
&= (1-\eta_k) \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{j=1}^{n} \boldsymbol{W}_{ij} \boldsymbol{x}_{q,k}^j \right] + \frac{1}{n} \eta_k \sum_{i=1}^{n} \boldsymbol{v}_{q,k}^i \\
&= (1-\eta_k) \frac{1}{n} \sum_{j=1}^{n} \left[\boldsymbol{x}_{q,k}^j \sum_{i=1}^{n} \boldsymbol{W}_{ij} \right] + \frac{1}{n} \eta_k \sum_{i=1}^{n} \boldsymbol{v}_{q,k}^i \\
&= (1-\eta_k) \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{x}_{q,k}^j + \frac{1}{n} \eta_k \sum_{i=1}^{n} \boldsymbol{v}_{q,k}^i && \textstyle (\sum_{i=1}^{n} W_{ij} = 1 \text{ for every } j) \\
&= (1-\eta_k) \overline{\boldsymbol{x}}_{q,k} + \frac{1}{n} \eta_k \sum_{i=1}^{n} \boldsymbol{v}_{q,k}^i \\
&= \overline{\boldsymbol{x}}_{q,k} + \eta_k \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{v}_{q,k}^i - \overline{\boldsymbol{x}}_{q,k} \right)
\end{aligned}$$

where we use a property of W which is $\sum_{i=1}^{n} W_{ij} = 1$ for every j. The proof of the second equation is similar. ■

Lemma 10 *For all $k \in \{1, \ldots, K\}$, it holds that*

$$\nabla \hat{F}_{q,k} = \frac{1}{n} \sum_{i=1}^{n} \nabla \hat{f}_{q,k}^i = \frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{g}}_{q,k}^i = \frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{d}}_{q,k}^i \tag{16}$$

Proof First, we verify that $\forall \ell \in \{1, \dots, K\}$

$$\frac{1}{n}\sum_{i=1}^{n} \nabla f^i_{\sigma_q(\ell)}(\boldsymbol{x}^i_{q,\ell}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}^i_{q,\ell} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{d}^i_{q,\ell} \tag{17}$$

For the base step $\ell = 1$, we have $\boldsymbol{g}^i_{q,\ell} = \nabla f^i_{\sigma_q(1)}(\boldsymbol{x}^i_{q,1})$. Averaging over n yields

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}^i_{q,1} = \frac{1}{n}\sum_{i=1}^{n} \nabla f^i_{\sigma_q(1)}(\boldsymbol{x}^i_{q,1})$$

Since $\mathbf{W}$ is doubly stochastic, we have

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{d}^i_{q,1} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\boldsymbol{g}^j_{q,1} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}^j_{q,1}\sum_{i=1}^{n} W_{ij} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}^j_{q,1} \tag{18}$$

For the recurrence step, recall the definition of $\boldsymbol{g}^i_{q,\ell}$,

$$\boldsymbol{g}^i_{q,\ell} = \nabla f^i_{\sigma_q(\ell)}(\boldsymbol{x}^i_{q,\ell}) - \nabla f^i_{\sigma_q(\ell-1)}(\boldsymbol{x}^i_{q,\ell-1}) + \boldsymbol{d}^i_{q,\ell-1}$$

Averaging over n and using the recurrence hypothesis $\frac{1}{n}\sum_{i=1}^{n} \nabla f^i_{\sigma_q(\ell-1)}(\boldsymbol{x}^i_{q,\ell-1}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{d}^i_{q,\ell-1}$, we deduce that

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}^i_{q,\ell} = \frac{1}{n}\sum_{i=1}^{n} \nabla f^i_{\sigma_q(\ell)}(\boldsymbol{x}^i_{q,\ell}) \tag{19}$$

Also, using the same techniques in equation (18) for $\boldsymbol{d}^i_{q,\ell}$, we complete the verification for equation (17). The proof of Lemma 10 can be deduced from equation (17) by averaging over $\ell \in \{k+1, \dots, K\}$ ∎

Lemma 11 *Suppose that each of $f^i_{\sigma_q(k)}$ is β-smooth. Using the Frank-Wolfe update of $\boldsymbol{x}^i_{q,k}$, the average of the remaining $(K-k)$ gradient approximation $\hat{\boldsymbol{d}}^i_{q,k}$ satisfies*

$$\max_{i\in[1,n]} \mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k} - \nabla \hat{F}_{q,k}\right\|\right] \leq \begin{cases} \dfrac{N}{k} & k \in \left[1, \dfrac{K}{2}\right] \\ \dfrac{N}{K-k+1} & k \in \left[\dfrac{K}{2}+1, K\right] \end{cases}$$

where $N = nGk_0 \max\{\lambda_2\left(1 + \frac{2}{1-\lambda_2}\right), 2\}$.

Proof We will prove the lemma by induction following the idea from Lemma 2 of Wai et al. (2017). Let's define following variables

$$\hat{\boldsymbol{d}}^{cat}_{q,k} = \left[\hat{\boldsymbol{d}}^{1\top}_{q,k}, \dots, \hat{\boldsymbol{d}}^{n\top}_{q,k}\right]^\top, \quad \hat{\boldsymbol{g}}^{cat}_{q,k} = \left[\hat{\boldsymbol{g}}^{1\top}_{q,k}, \dots, \hat{\boldsymbol{g}}^{n\top}_{q,k}\right]^\top, \quad \nabla \hat{F}^{cat}_{q,k} = \left[\nabla \hat{F}^\top_{q,k}, \dots, \nabla \hat{F}^\top_{q,k}\right]^\top \tag{20}$$

and let the slack variables as

$$\delta^i_{q,k} := \nabla \hat{f}^i_{q,k} - \nabla \hat{f}^i_{q,k-1}, \quad \bar{\delta}_{q,k} := \frac{1}{n}\sum_{i=1}^{n}\left(\nabla \hat{f}^i_{q,k} - \nabla \hat{f}^i_{q,k-1}\right) = \nabla \hat{F}_{q,k} - \nabla \hat{F}_{q,k-1} \tag{21}$$

then, following the definition in 20, we note

$$\delta^{cat}_{q,k} = \left[\delta^{1\top}_{q,k}, \ldots, \delta^{n\top}_{q,k}\right]^\top, \quad \bar{\delta}^{cat}_{q,k} = \left[\bar{\delta}^\top_{q,k}, \ldots, \bar{\delta}^\top_{q,k}\right]^\top$$

By Lemma 8, we have

$$\begin{aligned}
\left\|\hat{\boldsymbol{d}}^{cat}_{q,k} - \nabla \hat{F}^{cat}_{q,k}\right\|^2 &= \sum_{i=1}^{n}\left\|\hat{\boldsymbol{d}}^i_{q,k} - \nabla \hat{F}_{q,k}\right\|^2 \\
&\le \lambda_2^2 \sum_{i=1}^{n}\left\|\hat{\boldsymbol{g}}^i_{q,k} - \nabla \hat{F}_{q,k}\right\|^2 \\
&= \lambda_2^2\left\|\hat{\boldsymbol{g}}^{cat}_{q,k} - \nabla \hat{F}^{cat}_{q,k}\right\|^2
\end{aligned} \tag{22}$$

We can deduce that

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^{cat}_{q,k} - \nabla \hat{F}^{cat}_{q,k}\right\|\right] &\le \lambda_2 \mathbb{E}\left[\left\|\hat{\boldsymbol{g}}^{cat}_{q,k} - \nabla \hat{F}^{cat}_{q,k}\right\|\right] \\
&= \lambda_2 \mathbb{E}\left[\left\|\delta^{cat}_{q,k} + \hat{\boldsymbol{d}}^{cat}_{q,k-1} - \nabla \hat{F}^{cat}_{q,k} + \nabla \hat{F}^{cat}_{q,k-1} - \nabla \hat{F}^{cat}_{q,k-1}\right\|\right] \\
&\le \lambda_2\left(\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^{cat}_{q,k-1} - \nabla \hat{F}^{cat}_{q,k-1}\right\|\right] + \mathbb{E}\left[\left\|\delta^{cat}_{q,k} - \bar{\delta}^{cat}_{q,k}\right\|\right]\right) \\
&\le \lambda_2\left(\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^{cat}_{q,k-1} - \nabla \hat{F}^{cat}_{q,k-1}\right\|\right] + \mathbb{E}\left[\left\|\delta^{cat}_{q,k}\right\|\right]\right)
\end{aligned} \tag{23}$$

since

$$\left\|\delta^{cat}_{q,k} - \bar{\delta}^{cat}_{q,k}\right\|^2 = \sum_{i=1}^{n}\left\|\delta^i_{q,k} - \bar{\delta}_{q,k}\right\|^2 \le \sum_{i=1}^{n}\left\|\delta^i_{q,k}\right\|^2 - n\left\|\bar{\delta}_{q,k}\right\|^2 \le \sum_{i=1}^{n}\left\|\delta^i_{q,k}\right\|^2 = \left\|\delta^{cat}_{q,k}\right\|^2 \tag{24}$$

Notice that we can bound the expected value of δ^{cat} by

$$\begin{aligned}
\mathbb{E}\left[\left\|\delta^{cat}_{q,k}\right\|^2\right] &= \mathbb{E}\left[\sum_{i=1}^{n}\left\|\delta^i_{q,k}\right\|^2\right] = \sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla \hat{f}^i_{q,k} - \nabla \hat{f}^i_{q,k-1}\right\|^2\right] \\
&= \sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\left\|\nabla \hat{f}^i_{q,k} - \nabla \hat{f}^i_{q,k-1}\right\|^2 \mid \mathcal{F}_{q,k-1}\right]\right] \\
&= \sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\left\|\frac{\sum_{\ell=k+1}^{K}\nabla f^i_{\sigma_q(\ell)}(\boldsymbol{x}^i_{q,\ell})}{K-k} - \frac{\sum_{\ell=k}^{K}\nabla f^i_{\sigma_q(\ell)}(\boldsymbol{x}^i_{q,\ell})}{K-k+1}\right\|^2 \mid \mathcal{F}_{q,k-1}\right]\right] \\
&= \sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\left\|\frac{\sum_{\ell=k+1}^{K}\nabla f^i_{\sigma_q(\ell)}(\boldsymbol{x}^i_{q,\ell})}{(K-k)(K-k+1)} - \frac{\nabla f^i_{\sigma_q(k)}(\boldsymbol{x}^i_{q,k})}{K-k+1}\right\|^2 \mid \mathcal{F}_{q,k-1}\right]\right] \\
&\le n\left(\frac{2G}{K-k+1}\right)^2
\end{aligned} \tag{25}$$

using Jensen's inequality, we can deduce that

$$\mathbb{E}\left[\left\|\delta_{q,k}^{cat}\right\|\right] \leq \sqrt{\mathbb{E}\left[\left\|\delta_{q,k}^{cat}\right\|^2\right]} \leq \frac{2\sqrt{n}G}{K-k+1} \tag{26}$$

We are now proving the lemma by induction, when $k = 1$, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,1}^{cat} - \nabla \hat{F}_{q,1}^{cat}\right\|^2\right] &= \mathbb{E}\left[\sum_{i=1}^{n}\left\|\hat{\boldsymbol{d}}_{q,1}^{i} - \nabla \hat{F}_{q,1}\right\|^2\right] \leq \lambda_2^2\mathbb{E}\left[\sum_{i=1}^{n}\left\|\hat{\boldsymbol{g}}_{q,1}^{i} - \nabla \hat{F}_{q,1}\right\|^2\right] \\
&\leq \lambda_2^2\mathbb{E}\left[\sum_{i=1}^{n}\left\|\nabla \hat{f}_{q,1}^{i} - \nabla \hat{F}_{q,1}\right\|^2\right] \leq \lambda_2^2\mathbb{E}\left[\sum_{i=1}^{n}\left\|\nabla \hat{f}_{q,1}^{i}\right\|^2\right] \leq n\lambda_2^2 G^2
\end{aligned}$$

where we have used Lipschitzness of f in the last inequality. We now suppose that $1 \leq k \leq k_0$, from equations (23) and (26)

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k}^{cat} - \nabla \hat{F}_{q,k}^{cat}\right\|\right] &\leq \lambda_2\left(\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{cat} - \nabla \hat{F}_{q,k-1}^{cat}\right\|\right] + \mathbb{E}\left[\left\|\delta_{q,k}^{cat}\right\|\right]\right) \\
&\leq \lambda_2^{k-1}\sqrt{n}G + 2\sum_{\tau=1}^{k}\lambda_2^{\tau}\sqrt{n}G \\
&\leq \lambda_2\sqrt{n}G + 2\frac{\lambda_2}{1-\lambda_2}\sqrt{n}G \\
&= \lambda_2\sqrt{n}G\left(1 + \frac{2}{1-\lambda_2}\right)
\end{aligned} \tag{27}$$

We set $N_0 = k_0\sqrt{n}G\max\{\lambda_2\left(1+\frac{2}{1-\lambda_2}\right), 2\}$, we claim that $\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k}^{cat} - \nabla \hat{F}_{q,k}^{cat}\right\|\right] \leq \frac{N_0}{k}$ for $k \in \left[k_0, \frac{K}{2}+1\right]$. Recall that $K - k + 1 \geq k - 1$, by equations (23) and (26) and induction hypothesis, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k}^{cat} - \nabla \hat{F}_{q,k}^{cat}\right\|\right] &\leq \lambda_2\left(\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{cat} - \nabla \hat{F}_{q,k-1}^{cat}\right\|\right] + \mathbb{E}\left[\left\|\delta_{q,k}^{cat}\right\|\right]\right) \\
&\leq \lambda_2\left(\frac{N_0}{k-1} + \frac{2\sqrt{n}G}{K-k+1}\right) \\
&\leq \lambda_2\left(\frac{N_0}{k-1} + \frac{2\sqrt{n}G}{k-1}\right) \\
&\leq \lambda_2\left(\frac{N_0 + 2\sqrt{n}G}{k-1}\right) \\
&\leq \lambda_2\left(N_0\frac{k_0+1}{k_0(k-1)}\right) \\
&\leq \frac{N_0}{k}
\end{aligned} \tag{28}$$

where we have used the fact that $\lambda_2(\mathbf{W})\dfrac{k_0+1}{k_0(k-1)} \leq \dfrac{1}{k}$ in the last inequality. When $k \in \left[\frac{K}{2}+1, K\right]$, we claim that $\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k}^{cat} - \nabla \hat{F}_{q,k}^{cat}\right\|\right] \leq \dfrac{N_0}{K-k+1}$. The base case $k = \frac{K}{2}+1$

is verified by equation (28),

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k}^{cat} - \nabla \hat{F}_{q,k}^{cat}\right\|\right] \leq \frac{N_0}{\frac{K}{2}+1} \leq \frac{N_0}{\frac{K}{2}} \leq \frac{N_0}{K-(\frac{K}{2}+1)+1} \tag{29}$$

For $k \geq \frac{K}{2}+2$, using equations (23) and (26) and the induction hypothesis, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k}^{cat} - \nabla \hat{F}_{q,k}^{cat}\right\|\right] &\leq \lambda_2 \left(\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{cat} - \nabla \hat{F}_{q,k-1}^{cat}\right\|\right] + \mathbb{E}\left[\left\|\delta_{q,k}^{cat}\right\|\right]\right) \\
&\leq \lambda_2 \left(\frac{N_0}{K-k+2} + \frac{2\sqrt{n}G}{K-k+1}\right) \\
&\leq \lambda_2 \left(\frac{N_0+2G}{K-k+1}\right) \\
&\leq \lambda_2 \left(N_0 \frac{k_0+1}{k_0(K-k+1)}\right) \\
&\leq \frac{N_0}{K-k+1}
\end{aligned} \tag{30}$$

Recall that

$$\frac{1}{\sqrt{n}}\mathbb{E}\left[\sum_{i=1}^{n}\left\|\hat{\boldsymbol{d}}_{q,k}^{i} - \nabla \hat{F}_{q,k}\right\|\right] \leq \mathbb{E}\left[\left(\sum_{i=1}^{n}\left\|\hat{\boldsymbol{d}}_{q,k}^{i} - \nabla \hat{F}_{q,k}\right\|^2\right)^{1/2}\right] = \mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k}^{cat} - \nabla \hat{F}_{q,k}^{cat}\right\|\right] \tag{31}$$

The desired result followed from equations (28), (30) and (31) where $N = \sqrt{n}N_0$

$$\max_{i\in[1,n]} \mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k}^{i} - \nabla \hat{F}_{q,k}\right\|\right] \leq \begin{cases} \frac{N}{k} & k \in \left[1, \frac{K}{2}\right] \\ \frac{N}{K-k+1} & k \in \left[\frac{K}{2}+1, K\right] \end{cases} \tag{32}$$

■

Lemma 1 *For $i \in [n], k \in [K]$. Let $V_d = 2nG\left(\frac{\lambda_2}{1-\lambda_2}+1\right)$, the local gradient is upper-bounded, i.e $\left\|\boldsymbol{d}_{q,k}^{i}\right\| \leq V_d$*

Proof We use the same notation introduced in equation (20). Let's define

$$\boldsymbol{d}_{q,k}^{cat} = \left[\boldsymbol{d}_{q,k}^{1\top}, \ldots, \boldsymbol{d}_{q,k}^{n\top}\right]^\top \in \mathbb{R}^{nd}, \quad \nabla f_{\sigma_q(k)}^{cat} = \left[\nabla f_{\sigma_q(k)}^{1}(\boldsymbol{x}_{q,k}^{1})^\top, \ldots, \nabla f_{\sigma_q(k)}^{n}(\boldsymbol{x}_{q,k}^{n})^\top\right]^\top \in \mathbb{R}^{nd} \tag{33}$$

and

$$\nabla F_{\sigma_q(k)}^{cat} = \left[\nabla F_{\sigma_q(k)}^\top, \ldots, \nabla F_{\sigma_q(k)}^\top\right]^\top = \left[\frac{1}{n}\sum_{i=1}^{n}\nabla f_{\sigma_q(k)}^{i}(\boldsymbol{x}_{q,k}^{i})^\top, \ldots, \frac{1}{n}\sum_{i=1}^{n}\nabla f_{\sigma_q(k)}^{i}(\boldsymbol{x}_{q,k}^{i})^\top\right]^\top \tag{34}$$

Using the local gradient update, we have

$$\begin{aligned}
\boldsymbol{d}_{q,k}^{cat} &= (\mathbf{W} \otimes I_d)\left(\nabla f_{\sigma_q(k)}^{cat} - \nabla f_{\sigma_q(k-1)}^{cat} + \boldsymbol{d}_{q,k-1}^{cat}\right) \\
&= (\mathbf{W} \otimes I_d)\left(\nabla f_{\sigma_q(k)}^{cat} - \nabla f_{\sigma_q(k-1)}^{cat}\right) + (\mathbf{W} \otimes I_d)^2\left(\nabla f_{\sigma_q(k-1)}^{cat} - \nabla f_{\sigma_q(k-2)}^{cat} + \boldsymbol{d}_{q,k-2}^{cat}\right) \\
&= \sum_{\tau=1}^{k-1} (\mathbf{W} \otimes I_d)^{k-\tau}\left(\nabla f_{\sigma_q(\tau+1)}^{cat} - \nabla f_{\sigma_q(\tau)}^{cat}\right) + (\mathbf{W} \otimes I_d)^k \nabla f_{\sigma_q(1)}^{cat} \\
&= \sum_{\tau=1}^{k-1} (\mathbf{W} \otimes I_d)^{k-\tau}\left(\nabla f_{\sigma_q(\tau+1)}^{cat} - \nabla f_{\sigma_q(\tau)}^{cat}\right) + (\mathbf{W} \otimes I_d)^k \nabla f_{\sigma_q(1)}^{cat} \\
&\quad - \sum_{\tau=1}^{k-1}\left(\nabla F_{\sigma_q(\tau+1)}^{cat} - \nabla F_{\sigma_q(\tau)}^{cat}\right) - \nabla F_{\sigma_q(1)}^{cat} + \nabla F_{\sigma_q(k)}^{cat} \\
&= \sum_{\tau=1}^{k-1}\left[\left(\mathbf{W}^{k-\tau} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right) \otimes I_d\right]\left(\nabla f_{\sigma_q(\tau+1)}^{cat} - \nabla f_{\sigma_q(\tau)}^{cat}\right) \\
&\quad + \left[\left(\mathbf{W}^{k} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right) \otimes I_d\right]\nabla f_{\sigma_q(1)}^{cat} + \left(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T \otimes I_d\right)\nabla f_{\sigma_q(k)}^{cat}
\end{aligned} \tag{35}$$

where the fourth equality holds since $\nabla F_{\sigma_q(k)}^{cat} - \sum_{\tau=1}^{k-1}\left(\nabla F_{\sigma_q(\tau+1)}^{cat} - \nabla F_{\sigma_q(\tau)}^{cat}\right) - \nabla F_{\sigma_q(1)}^{cat} = 0$. The fifth equality can be deduced using $\nabla F_{\sigma_q(k)}^{cat} = \left(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T \otimes I_d\right)\nabla f_{\sigma_q(k)}^{cat}$ and $(\mathbf{W} \otimes I_d)^k = \left(\mathbf{W}^k \otimes I_d\right)$. Recall that $\|\mathbf{W} \otimes I_d\| = \|\mathbf{W}\|$. Taking the norm on equation (35), we have

$$\left\|\boldsymbol{d}_{q,k}^{cat}\right\| \leq 2\sqrt{n}G\sum_{\tau=1}^{k-1}\lambda_2^{k-\tau} + \sqrt{n}G\left(\lambda_2^k + 1\right) \leq 2\sqrt{n}G\left(\frac{\lambda_2}{1-\lambda_2} + 1\right) \tag{36}$$

where we have used $\left\|\nabla f_{\sigma_q(\tau+1)}^{cat} - \nabla f_{\sigma_q(\tau)}^{cat}\right\| \leq 2\sqrt{n}G$, $\left\|\mathbf{W}^k - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right\| \leq \lambda_2^k$ and $\left\|\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right\| \leq 1$ in the first inequality. We have $\forall i \in [n]$

$$\left\|\boldsymbol{d}_{q,k}^{i}\right\| \leq \sum_{i=1}^{n}\left\|\boldsymbol{d}_{q,k}^{i}\right\| \leq \sqrt{n}\left(\sum_{i=1}^{n}\left\|\boldsymbol{d}_{q,k}^{i}\right\|^2\right)^{1/2} = \sqrt{n}\left\|\boldsymbol{d}_{q,k}^{cat}\right\| \tag{37}$$

one can obtain the desired result. ■

Lemma 2 *Under Assumption 2 and let* $\sigma_1^2 = 4n\left[\left(\frac{G+G_0}{\frac{1}{\lambda_2}-1}\right)^2 + 2\sigma_0^2\right]$. *For* $i \in [n], k \in [K]$, *the variance of the local stochastic gradient is uniformly bounded.*

$$\mathbb{E}\left[\left\|\boldsymbol{d}_{q,k}^{i} - \widetilde{\boldsymbol{d}}_{q,k}^{i}\right\|^2\right] \leq \sigma_1^2$$

Proof We denote $\widetilde{\boldsymbol{d}}^{cat}$ the stochastique version of $\boldsymbol{d}^{cat}$, following equation (35), we have

$$\begin{aligned}\widetilde{\boldsymbol{d}}_{q,k}^{cat} &= \sum_{\tau=1}^{k-1}\left[\left(\mathbf{W}^{k-\tau}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\otimes I_d\right]\left(\widetilde{\nabla} f_{\sigma_q(\tau+1)}^{cat}-\widetilde{\nabla} f_{\sigma_q(\tau)}^{cat}\right)\\&+\left[\left(\mathbf{W}^{k}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\otimes I_d\right]\widetilde{\nabla} f_{\sigma_q(1)}^{cat}+\left(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\otimes I_d\right)\widetilde{\nabla} f_{\sigma_q(k)}^{cat}\end{aligned}\tag{38}$$

Then, we have

$$\begin{aligned}\boldsymbol{d}_{q,k}^{cat}-\widetilde{\boldsymbol{d}}_{q,k}^{cat} &= \sum_{\tau=1}^{k-1}\left[\left(\mathbf{W}^{k-\tau}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\otimes I_d\right]\left(\nabla f_{\sigma_q(\tau+1)}^{cat}-\widetilde{\nabla} f_{\sigma_q(\tau+1)}^{cat}+\widetilde{\nabla} f_{\sigma_q(\tau)}^{cat}-\nabla f_{\sigma_q(\tau)}^{cat}\right)\\&+\left[\left(\mathbf{W}^{k}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\otimes I_d\right]\left(\nabla f_{\sigma_q(1)}^{cat}-\widetilde{\nabla} f_{\sigma_q(1)}^{cat}\right)\\&+\left(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\otimes I_d\right)\left(\nabla f_{\sigma_q(k)}^{cat}-\widetilde{\nabla} f_{\sigma_q(k)}^{cat}\right)\end{aligned}\tag{39}$$

By Assumption 2 and Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla f_{\sigma_q(\tau)}^{cat}-\widetilde{\nabla} f_{\sigma_q(\tau)}^{cat}\right\|^2\right] &= \mathbb{E}\left[\sum_{i=1}^{n}\left\|\nabla f_{\sigma_q(\tau)}^{i}\left(\boldsymbol{x}_{q,\tau}^i\right)-\widetilde{\nabla} f_{\sigma_q(\tau)}^{i}\left(\boldsymbol{x}_{q,\tau}^i\right)\right\|^2\right]\\&\leq \sqrt{\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_{\sigma_q(\tau)}^{i}\left(\boldsymbol{x}_{q,\tau}^i\right)-\widetilde{\nabla} f_{\sigma_q(\tau)}^{i}\left(\boldsymbol{x}_{q,\tau}^i\right)\right\|^2\right]}\leq\sqrt{n}\sigma_0\end{aligned}\tag{40}$$

The second moment of equation (39) is written as

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{d}_{q,k}^{cat}-\widetilde{\boldsymbol{d}}_{q,k}^{cat}\right\|^2\right]\\
=&\mathbb{E}\sum_{\tau=1}^{k-1}\left[\left(\mathbf{W}^{k-\tau}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\otimes I_d\right]\left(\nabla f_{\sigma_q(\tau+1)}^{cat}-\widetilde{\nabla} f_{\sigma_q(\tau+1)}^{cat}+\widetilde{\nabla} f_{\sigma_q(\tau)}^{cat}-\nabla f_{\sigma_q(\tau)}^{cat}\right)\\
&\quad+\left[\left(\mathbf{W}^{k}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\otimes I_d\right]\left(\nabla f_{\sigma_q(1)}^{cat}-\widetilde{\nabla} f_{\sigma_q(1)}^{cat}\right)+\left(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\otimes I_d\right)\left(\nabla f_{\sigma_q(k)}^{cat}-\widetilde{\nabla} f_{\sigma_q(k)}^{cat}\right)^2\\
\le&\mathbb{E}\left[\left(\sum_{\tau=1}^{k-1}\left\|\mathbf{W}^{k-\tau}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right\|\left\|\nabla f_{\sigma_q(\tau+1)}^{cat}-\widetilde{\nabla} f_{\sigma_q(\tau+1)}^{cat}+\widetilde{\nabla} f_{\sigma_q(\tau)}^{cat}-\nabla f_{\sigma_q(\tau)}^{cat}\right\|\right)^2\right]\\
&\quad+\mathbb{E}\left[\left\|\left(\mathbf{W}^{k}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\otimes I_d\right)\left(\nabla f_{\sigma_q(1)}^{cat}-\widetilde{\nabla} f_{\sigma_q(1)}^{cat}\right)+\left(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\otimes I_d\right)\left(\nabla f_{\sigma_q(k)}^{cat}-\widetilde{\nabla} f_{\sigma_q(k)}^{cat}\right)\right\|^2\right]\\
\le&\mathbb{E}\left[\left(\sum_{\tau=1}^{k-1}\left\|\mathbf{W}^{k-\tau}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right\|\left\|\nabla f_{\sigma_q(\tau+1)}^{cat}-\widetilde{\nabla} f_{\sigma_q(\tau+1)}^{cat}+\widetilde{\nabla} f_{\sigma_q(\tau)}^{cat}-\nabla f_{\sigma_q(\tau)}^{cat}\right\|\right)^2\right]\\
&\quad+4\left(\mathbb{E}\left[\left\|\mathbf{W}^{k}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right\|^2\left\|\nabla f_{\sigma_q(1)}^{cat}-\widetilde{\nabla} f_{\sigma_q(1)}^{cat}\right\|^2\right]+\mathbb{E}\left[\left\|\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right\|^2\left\|\nabla f_{\sigma_q(k)}^{cat}-\widetilde{\nabla} f_{\sigma_q(k)}^{cat}\right\|^2\right]\right)\\
\le&4n\left(G+G_0\right)^2\left(\sum_{\tau=1}^{k-1}\lambda_2^{k-\tau}\right)^2+4n\sigma_0^2\left(\lambda_2^{2k}+1\right)\\
\le&4n\left(G+G_0\right)^2\left(\frac{\lambda_2}{1-\lambda_2}\right)^2+4n\sigma_0^2(\lambda_2+1)\le 4n\left[\left(\frac{G+G_0}{\frac{1}{\lambda_2}-1}\right)^2+2\sigma_0^2\right]
\end{aligned}
\tag{41}
$$

where the first inequality holds since $\mathbb{E}\left[\nabla f_{\sigma_q(\tau+1)}^{cat}-\widetilde{\nabla} f_{\sigma_q(\tau+1)}^{cat}+\widetilde{\nabla} f_{\sigma_q(\tau)}^{cat}-\nabla f_{\sigma_q(\tau)}^{cat}\right]=0$. The second inequality follows the fact that $\|a+b\|^2\le 4\left(\|a\|^2+\|b\|^2\right)$. The third inequality comes from Assumption 2 and the analysis in Lemma 1. Finally, one can obtain the desired result by noticing $\mathbb{E}\left[\left\|\boldsymbol{d}_{q,k}^{i}-\widetilde{\boldsymbol{d}}_{q,k}^{i}\right\|^2\right]\le\sum_{i=1}^{n}\mathbb{E}\left[\left\|\boldsymbol{d}_{q,k}^{i}-\widetilde{\boldsymbol{d}}_{q,k}^{i}\right\|^2\right]=\mathbb{E}\left[\left\|\boldsymbol{d}_{q,k}^{cat}-\widetilde{\boldsymbol{d}}_{q,k}^{cat}\right\|^2\right]$
■

Lemma 12 (Lemma 6, Zhang et al. (2019)) *Under Assumption 5, Lemma 1, Lemma 2 and setting* $\rho_k=\frac{2}{(k+3)^{2/3}}$ *and* $\rho_k=\frac{1.5}{(K-k+2)^{2/3}}$ *for* $k\in\left[\frac{K}{2}\right]$ *and* $k\in\left[\frac{K}{2}+1,K\right]$ *respectively, we have*

$$
\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{i}-\widetilde{\boldsymbol{a}}_{q,k}^{i}\right\|\right]\le\begin{cases}\dfrac{\sqrt{M}}{(k+4)^{1/3}} & k\in\left[\dfrac{K}{2}\right]\\[2ex] \dfrac{\sqrt{M}}{(K-k+1)^{1/3}} & i\in\left[\dfrac{K}{2}+1,K\right]\end{cases}
\tag{42}
$$

where $M = \max\{M_1, M_2\}$ *where* $M_1 = \max\{5^{2/3}\left(V_{\boldsymbol{d}} + L_0\right)^2, M_0\}$, $M_0 = 4\left(V_{\boldsymbol{d}}^2 + \sigma^2\right) + 32\sqrt{2}V_{\boldsymbol{d}}$ *and* $M_2 = 2.55\left(V_{\boldsymbol{d}}^2 + \sigma^2\right) + \dfrac{7\sqrt{2}V_{\boldsymbol{d}}}{3}$ *and* $L_0 = \frac{2}{4^{2/3}}\left\|\widetilde{\boldsymbol{d}}^i_{q,1}\right\|$

Proof In order to prove the lemma, we only need to bound $\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{a}}^i_{q,k}\right\|^2\right]$, following the decomposition in Zhang et al. (2019), we have

$$\begin{aligned}
\mathbb{E}&\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{a}}^i_{q,k}\right\|^2\right] = \mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - (1-\rho_k)\widetilde{\boldsymbol{a}}^i_{q,k-1} - \rho_k\widetilde{\boldsymbol{d}}^i_{q,k}\right\|^2\right] \\
&= \rho_k^2\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{d}}^i_{q,k})\right\|^2\right] + (1-\rho_k)^2\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \hat{\boldsymbol{d}}^i_{q,k-2}\right\|^2\right] \qquad (43) \\
&\quad + (1-\rho_k)^2\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\|^2\right] \\
&\quad + 2\rho_k(1-\rho_k)\mathbb{E}\left[\left\langle\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{d}}^i_{q,k}, \hat{\boldsymbol{d}}^i_{q,k-1} - \hat{\boldsymbol{d}}^i_{q,k-2}\right\rangle\right] \\
&\quad + 2\rho_k(1-\rho_k)\mathbb{E}\left[\left\langle\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{d}}^i_{q,k}, \hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\rangle\right] \\
&\quad + 2(1-\rho_k)^2\mathbb{E}\left[\left\langle\hat{\boldsymbol{d}}^i_{q,k-1} - \hat{\boldsymbol{d}}^i_{q,k-2}, \hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\rangle\right] \qquad (44)
\end{aligned}$$

The first part of the above equation is written as

$$\begin{aligned}
\mathbb{E}&\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{d}}^i_{q,k})\right\|^2\right] = \mathbb{E}\left[\mathbb{E}_\sigma\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{d}}^i_{q,k})\right\|^2 \mid \mathcal{F}_{q,k-1}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}_\sigma\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \boldsymbol{d}^i_{q,k} + \boldsymbol{d}^i_{q,k} - \widetilde{\boldsymbol{d}}^i_{q,k})\right\|^2 \mid \mathcal{F}_{q,k-1}\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{E}_\sigma\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \boldsymbol{d}^i_{q,k}\right\|^2 + \left\|\boldsymbol{d}^i_{q,k} - \widetilde{\boldsymbol{d}}^i_{q,k})\right\|^2 + 2\langle\hat{\boldsymbol{d}}^i_{q,k-1} - \boldsymbol{d}^i_{q,k}, \boldsymbol{d}^i_{q,k} - \widetilde{\boldsymbol{d}}^i_{q,k}\rangle \mid \mathcal{F}_{q,k-1}\right]\right] \qquad (45)
\end{aligned}$$

Using the definition of $\hat{\boldsymbol{d}}^i_{q,k-1}$, Lemma 1 and Lemma 2 and law of total expectation, we have

$$\mathbb{E}\left[\mathbb{E}_\sigma\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \boldsymbol{d}^i_{q,k}\right\|^2 \mid \mathcal{F}_{q,k-1}\right]\right] = \mathbb{E}\left[Var_\sigma\left(\boldsymbol{d}^i_{q,k} \mid \mathcal{F}_{q,k-1}\right)\right] \leq \mathbb{E}\left[\mathbb{E}_\sigma\left[\left\|\boldsymbol{d}^i_{q,k}\right\|^2 \mid \mathcal{F}_{q,k-1}\right]\right] \leq V_{\boldsymbol{d}}^2 \qquad (46)$$

$$\mathbb{E}\left[\mathbb{E}_\sigma\left[\left\|\boldsymbol{d}^i_{q,k} - \widetilde{\boldsymbol{d}}^i_{q,k})\right\|^2\right] |\mathcal{F}_{q,k-1}\right] \leq \sigma_1^2 \qquad (47)$$

Recall that $\mathcal{H}_{q,k}$ is the filtration related to the randomness of $\widetilde{\boldsymbol{d}}^i_{q,k}$ and $\hat{\boldsymbol{d}}^i_{q,k-1}$ and $\boldsymbol{d}^i_{q,k}$ is $\mathcal{F}_{q,k}$-measurable, then one can write

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}_\sigma\left[\left\langle \hat{\boldsymbol{d}}^i_{q,k-1} - \boldsymbol{d}^i_{q,k}, \boldsymbol{d}^i_{q,k} - \widetilde{\boldsymbol{d}}^i_{q,k}\right\rangle \mid \mathcal{F}_{q,k-1}\right]\right] \\
=&\mathbb{E}\left[\left\langle \hat{\boldsymbol{d}}^i_{q,k-1} - \boldsymbol{d}^i_{q,k}, \boldsymbol{d}^i_{q,k} - \widetilde{\boldsymbol{d}}^i_{q,k}\right\rangle\right] \\
=&\mathbb{E}\left[\mathbb{E}_\sigma\left[\left\langle \hat{\boldsymbol{d}}^i_{q,k-1} - \boldsymbol{d}^i_{q,k}, \boldsymbol{d}^i_{q,k} - \widetilde{\boldsymbol{d}}^i_{q,k}\right\rangle \mid \mathcal{F}_{q,k}\right]\right] \\
=&\mathbb{E}\left[\left\langle \hat{\boldsymbol{d}}^i_{q,k-1} - \boldsymbol{d}^i_{q,k}, \mathbb{E}_\sigma\left[\boldsymbol{d}^i_{q,k} - \widetilde{\boldsymbol{d}}^i_{q,k} \mid \mathcal{F}_{q,k}\right]\right\rangle\right] && \text{(by } \mathcal{F}_{q,k}\text{-measurability)} \\
=&\mathbb{E}\left[\mathbb{E}_{\widetilde{\boldsymbol{d}}}\left[\left\langle \hat{\boldsymbol{d}}^i_{q,k-1} - \boldsymbol{d}^i_{q,k}, \mathbb{E}_\sigma\left[\boldsymbol{d}^i_{q,k} - \widetilde{\boldsymbol{d}}^i_{q,k} \mid \mathcal{F}_{q,k}\right]\right\rangle\right] \mid \mathcal{H}_{q,k-1}\right] \\
=&\mathbb{E}\left[\left\langle \hat{\boldsymbol{d}}^i_{q,k-1} - \boldsymbol{d}^i_{q,k}, \mathbb{E}_{\widetilde{\boldsymbol{d}}}\left[\mathbb{E}_\sigma\left[\boldsymbol{d}^i_{q,k} - \widetilde{\boldsymbol{d}}^i_{q,k} \mid \mathcal{F}_{q,k}\right] \mid \mathcal{H}_{q,k-1}\right]\right\rangle\right] \\
=&\mathbb{E}\left[\left\langle \hat{\boldsymbol{d}}^i_{q,k-1} - \boldsymbol{d}^i_{q,k}, \mathbb{E}_\sigma\left[\mathbb{E}_{\widetilde{\boldsymbol{d}}}\left[\boldsymbol{d}^i_{q,k} - \widetilde{\boldsymbol{d}}^i_{q,k} \mid \mathcal{H}_{q,k-1}\right] \mid \mathcal{F}_{q,k}\right]\right\rangle\right] && \text{(by Fubini's theorem)} \\
=&0 && (48)
\end{aligned}$$

where the last equation holds since $\mathbb{E}_{\widetilde{\boldsymbol{d}}}\left[\widetilde{\boldsymbol{d}}^i_{q,k} \mid \mathcal{H}_{q,k-1}\right] = \boldsymbol{d}^i_{q,k}$. Combining equations (46) to (48), equation (45) is upper bounded by

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{d}}^i_{q,k})\right\|^2\right] \leq V_{\boldsymbol{d}}^2 + \sigma_1^2 \triangleq V \tag{49}$$

We are now bounding $\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \hat{\boldsymbol{d}}^i_{q,k-2}\right\|^2\right]$, using the definition of $\hat{\boldsymbol{d}}^i_{q,k}$ and Lemma 1. We have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \hat{\boldsymbol{d}}^i_{q,k-2}\right\|^2\right] \\
=&\mathbb{E}\left[\mathbb{E}_\sigma\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \hat{\boldsymbol{d}}^i_{q,k-2}\right\|^2 \mid \mathcal{F}_{q,k-2}\right]\right] \\
=&\mathbb{E}\left[\mathbb{E}_\sigma\left[\left\|\frac{\sum_{\ell=k}^K \boldsymbol{d}^i_{q,\ell}}{K-k+1} - \frac{\sum_{\ell=k-1}^K \boldsymbol{d}^i_{q,\ell}}{K-k+2}\right\|^2 \mid \mathcal{F}_{q,k-2}\right]\right] \\
=&\mathbb{E}\left[\mathbb{E}_\sigma\left[\left\|\frac{\sum_{\ell=k}^K \boldsymbol{d}^i_{q,\ell}}{K-k+1} - \frac{\sum_{\ell=k}^K \boldsymbol{d}^i_{q,\ell}}{K-k+2} - \frac{\boldsymbol{d}^i_{q,k-1}}{K-k+2}\right\|^2 \mid \mathcal{F}_{q,k-2}\right]\right] \\
=&\mathbb{E}\left[\mathbb{E}_\sigma\left[\left\|\frac{\sum_{\ell=k}^K \boldsymbol{d}^i_{q,\ell}}{(K-k+1)(K-k+2)} - \frac{\boldsymbol{d}^i_{q,k-1}}{K-k+2}\right\|^2 \mid \mathcal{F}_{q,k-2}\right]\right] \\
\leq&\mathbb{E}\left[\mathbb{E}_\sigma\left[\left(\frac{\sum_{\ell=k}^K \left\|\boldsymbol{d}^i_{q,\ell}\right\|}{(K-k+1)(K-k+2)} + \frac{\left\|\boldsymbol{d}^i_{q,k-1}\right\|}{K-k+2}\right)^2 \mid \mathcal{F}_{q,k-2}\right]\right] \\
\leq&\frac{4V_{\boldsymbol{d}}^2}{(K-k+2)^2} \triangleq \frac{L}{(K-k+2)^2} && (50)
\end{aligned}$$

More over, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{d}}^i_{q,k}, \hat{\boldsymbol{d}}^i_{q,k-1} - \hat{\boldsymbol{d}}^i_{q,k-2} \right\rangle\right] \\
=&\mathbb{E}\left[\mathbb{E}_{\sigma,\widetilde{\boldsymbol{d}}}\left[\left\langle \hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{d}}^i_{q,k}, \hat{\boldsymbol{d}}^i_{q,k-1} - \hat{\boldsymbol{d}}^i_{q,k-2} \right\rangle \mid \mathcal{F}_{q,k-1}, \mathcal{H}_{q,k-1}\right]\right] \\
=&\mathbb{E}\left[\left\langle \mathbb{E}_{\sigma,\widetilde{\boldsymbol{d}}}\left[\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{d}}^i_{q,k} \mid \mathcal{F}_{q,k-1}, \mathcal{H}_{q,k-1}\right], \hat{\boldsymbol{d}}^i_{q,k-1} - \hat{\boldsymbol{d}}^i_{q,k-2} \right\rangle\right] \\
=&0
\end{aligned} \tag{51}$$

since $\mathbb{E}_{\widetilde{\boldsymbol{d}}}\left[\widetilde{\boldsymbol{d}}^i_{q,k} \mid \mathcal{H}_{q,k-1}\right] = \boldsymbol{d}^i_{q,k}$ and $\mathbb{E}_{\sigma}\left[\boldsymbol{d}^i_{q,k} \mid \mathcal{F}_{q,k-1}\right] = \hat{\boldsymbol{d}}^i_{q,k}$. Using the same argument, we can deduce

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{d}}^i_{q,k}, \hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1} \right\rangle\right] \\
=&\mathbb{E}\left[\mathbb{E}_{\sigma,\widetilde{\boldsymbol{d}}}\left[\left\langle \hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{d}}^i_{q,k}, \hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1} \right\rangle \mid \mathcal{F}_{q,k-1}, \mathcal{H}_{q,k-1},\right]\right] \\
=&\mathbb{E}\left[\left\langle \mathbb{E}_{\sigma,\widetilde{\boldsymbol{d}}}\left[\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{d}}^i_{q,k} \mid \mathcal{F}_{q,k-1}, \mathcal{H}_{q,k-1}\right], \hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1} \right\rangle\right] \\
=&0
\end{aligned} \tag{52}$$

where we have use law of total expectation and conditional unbiasness of $\widetilde{\boldsymbol{d}}^i_{q,k}$. Using Young's inequality and equation (50), one can write

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \hat{\boldsymbol{d}}^i_{q,k-1} - \hat{\boldsymbol{d}}^i_{q,k-2}, \hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1} \right\rangle\right] \\
&\quad\le \mathbb{E}\left[\frac{1}{2\alpha_k}\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \hat{\boldsymbol{d}}^i_{q,k-2}\right\|^2 + \frac{\alpha_k}{2}\left\|\hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\|^2\right] \\
&\quad\le \frac{L}{2\alpha_k(K-k+2)^2} + \frac{\alpha_k}{2}\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\|^2\right]
\end{aligned} \tag{53}$$

With the above analysis, we can deduce that

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{a}}^i_{q,k}\right\|^2\right] \le \rho_k^2 V + (1-\rho_k)^2\frac{L}{(K-k+2)^2} + (1-\rho_k)^2\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\|^2\right] \\
+ (1-\rho_k)^2\left(\frac{L}{\alpha_k(K-k+2)^2} + \alpha_k\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\|^2\right]\right)
\end{aligned} \tag{54}$$

Setting $\alpha_k = \frac{\rho_k}{2}$, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{a}}^i_{q,k}\right\|^2\right] &\le \rho_k^2 V + (1-\rho_k)^2\left(1+\frac{2}{\rho_k}\right)\frac{L}{(K-k+2)^2} \\
&\quad + (1-\rho_k)^2\left(1+\frac{\rho_k}{2}\right)\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\|^2\right] \\
&\le \rho_k^2 V + \left(1+\frac{2}{\rho_k}\right)\frac{L}{(K-k+2)^2} + (1-\rho_k)\,\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\|^2\right]
\end{aligned} \tag{55}$$

For $k \leq \frac{K}{2} + 1$, we set $\rho_k = \frac{2}{(k+3)^{2/3}}$ and recall that $K - k + 2 \geq k$, equation (55) is written as:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{a}}^i_{q,k}\right\|^2\right] \\
&\leq \frac{4}{(k+3)^{4/3}}V + \left(1 + (k+3)^{2/3}\right)\frac{L}{k^2} + \left(1 - \frac{2}{(k+3)^{2/3}}\right)\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\|^2\right] \\
&\leq \frac{4}{(k+3)^{4/3}}V + \left(1 + (k+3)^{2/3}\right)\frac{16L}{(k+3)^2} + \left(1 - \frac{2}{(k+3)^{2/3}}\right)\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\|^2\right] \\
&\leq \frac{4}{(k+3)^{4/3}}V + \frac{16L}{(k+3)^{4/3}} + \frac{16L}{(k+3)^{4/3}} + \left(1 - \frac{2}{(k+3)^{2/3}}\right)\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\|^2\right] \\
&\leq \frac{4V + 32L}{(k+3)^{4/3}} + \left(1 - \frac{2}{(k+3)^{2/3}}\right)\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\|^2\right] \\
&\triangleq \frac{M_0}{(k+3)^{4/3}} + \left(1 - \frac{2}{(k+3)^{2/3}}\right)\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\|^2\right] \qquad (56)
\end{aligned}
$$

We consider the base step where $k = 1$,

$$
\begin{aligned}
\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,0} - \widetilde{\boldsymbol{a}}^i_{q,1}\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{K}\sum_{\ell=1}^{K}\boldsymbol{d}^i_{q,\ell} - \frac{2}{4^{2/3}}\widetilde{\boldsymbol{d}}^i_{q,1}\right\|^2\right] \\
&\leq \left(V_{\boldsymbol{d}} + \frac{2}{4^{2/3}}\left\|\widetilde{\boldsymbol{d}}^i_{q,1}\right\|\right)^2 \\
&\leq \left(V_{\boldsymbol{d}} + \frac{2}{4^{2/3}}G_0\right)^2 \\
&\triangleq \left(V_{\boldsymbol{d}} + L_0\right)^2 \qquad (57)
\end{aligned}
$$

Set $M_1 = \max\left\{5^{2/3}\left(V_{\boldsymbol{d}} + L_0\right)^2, M_0\right\}$. For $k \in \left[\frac{K}{2} + 1\right]$, we claim that $\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{a}}^i_{q,k}\right\|^2\right] \leq \frac{M_1}{(k+4)^{2/3}}$. Suppose the claim holds for $k-1$, we have

$$
\begin{aligned}
\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{a}}^i_{q,k}\right\|^2\right] &\leq \frac{M_1}{(k+3)^{4/3}} + \mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\|^2\right]\left(1 - \frac{2}{(k+3)^{2/3}}\right) \\
&\leq \frac{M_1}{(k+3)^{4/3}} + \frac{M_1}{(k+3)^{2/3}} \cdot \frac{(k+3)^{2/3} - 2}{(k+3)^{2/3}} \\
&\leq \frac{M_1\left((k+3)^{2/3} - 1\right)}{(k+3)^{4/3}} \\
&\leq \frac{M_1}{(k+4)^{2/3}} \qquad (58)
\end{aligned}
$$

since $\frac{(k+3)^{2/3}-1}{k+3)^{4/3}} \leq \frac{1}{(k+4)^{2/3}}$. For $k \in \left[\frac{K}{2}+1, K\right]$, we set $\rho_k = \frac{1.5}{(K-k+2)^{2/3}}$, thus

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{a}}^i_{q,k}\right\|^2\right] &\leq \frac{2.55V}{(K-k+2)^{4/3}} + \left(1 + \frac{4}{3}(K-k+2)^{2/3}\right)\frac{L}{(K-k+2)^2} \\
&\quad + \mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\|^2\right]\left(1 - \frac{1.5}{(K-k+2)^{2/3}}\right) \\
&\leq \frac{2.55V}{(K-k+2)^{4/3}} + \frac{L}{(K-k+2)^{4/3}} + \frac{4}{3}\frac{L}{(K-k+2)^{4/3}} \\
&\quad + \mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\|^2\right]\left(1 - \frac{1.5}{(K-k+2)^{2/3}}\right) \\
&\leq \frac{2.55V + 7L/3}{(K-k+2)^{4/3}} + \mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\|^2\right]\left(1 - \frac{1.5}{(K-k+2)^{2/3}}\right) \\
&\triangleq \frac{M_2}{(K-k+2)^{4/3}} + \mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-2} - \widetilde{\boldsymbol{a}}^i_{q,k-1}\right\|^2\right]\left(1 - \frac{1.5}{(K-k+2)^{2/3}}\right)
\end{aligned} \tag{59}$$

Let $M = \max\{M_1, M_2\}$ and $k \in \left[\frac{K}{2}+1, K\right]$, we claim that $\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{a}}^i_{q,k}\right\|^2\right] \leq \frac{M}{(K-k+1)^{2/3}}$. The base step is verified by equation (58). We now suppose the claim holds for $k-1$, let's prove for k.

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{a}}^i_{q,k}\right\|^2\right] &\leq \frac{M}{(K-k+2)^{4/3}} + \frac{M}{(K-k+2)^{2/3}} \cdot \frac{(K-k+2)^{2/3} - 1.5}{(K-k+2)^{2/3}} \\
&= \frac{M\left((K-k+2)^{2/3} - 0.5\right)}{(K-k+2)^{4/3}} \\
&\leq \frac{M}{(K-k+1)^{2/3}}
\end{aligned} \tag{60}$$

Thus, from equation (58) and equation (60), we have

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{a}}^i_{q,k}\right\|^2\right] \leq \begin{cases} \dfrac{M}{(k+4)^{2/3}} & k \in \left[1, \dfrac{K}{2}\right] \\ \dfrac{M}{(K-k+1)^{2/3}} & k \in \left[\dfrac{K}{2}+1, K\right] \end{cases} \tag{61}$$

Thus, using Jensen inequality, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{a}}^i_{q,k}\right\|\right] &\leq \sqrt{\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{a}}^i_{q,k}\right\|^2\right]} \\
&= \sqrt{\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^i_{q,k-1} - \widetilde{\boldsymbol{a}}^i_{q,k}\right\|^2\right]} \\
&\leq \begin{cases} \dfrac{\sqrt{M}}{(k+4)^{1/3}} & k \in \left[1, \dfrac{K}{2}\right] \\ \dfrac{\sqrt{M}}{(K-k+1)^{1/3}} & k \in \left[\dfrac{K}{2}+1, K\right] \end{cases}
\end{aligned} \tag{62}$$

■

Claim 1

$$\mathbb{E}\left[\left\|\nabla \overline{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) - \nabla \hat{F}_{q,k-1}\right\|\right] \leq \beta D \tag{63}$$

Proof of claim. Recall the definition of $\overline{F}_{q,k-1}$ and $\hat{F}_{q,k-1}$,

$$\begin{aligned}
\overline{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) &= \frac{1}{K-k+1}\sum_{\ell=k}^{K}\frac{1}{n}\sum_{i=1}^{n} f^i_{\sigma_q(\ell)}(\overline{\boldsymbol{x}}_{q,k}) \\
\hat{F}_{q,k-1} &= \frac{1}{K-k+1}\sum_{\ell=k}^{K}\frac{1}{n}\sum_{i=1}^{n} f^i_{\sigma_q(\ell)}(\boldsymbol{x}^i_{q,\ell})
\end{aligned}$$

we have,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla \overline{F}_{q,k-1}(\overline{\boldsymbol{x}}^k_q) - \nabla \hat{F}_{q,k-1}\right\|\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{K-k+1}\cdot\frac{1}{n}\sum_{\ell=k}^{K}\sum_{i=1}^{n}\left(\nabla f^i_{\sigma_q(\ell)}\left(\overline{\boldsymbol{x}}_{q,k}\right) - \nabla f^i_{\sigma_q(\ell)}\left(\boldsymbol{x}^i_{q,\ell}\right)\right)\right\|\right] \\
&\leq \mathbb{E}\left[\frac{1}{K-k+1}\cdot\frac{1}{n}\sum_{\ell=k}^{K}\sum_{i=1}^{n}\left\|\nabla f^i_{\sigma_q(\ell)}\left(\overline{\boldsymbol{x}}_{q,k}\right) - \nabla f^i_{\sigma_q(\ell)}\left(\boldsymbol{x}^i_{q,\ell}\right)\right\|\right] \\
&\leq \mathbb{E}\left[\frac{1}{K-k+1}\cdot\frac{1}{n}\sum_{\ell=k}^{K}\sum_{i=1}^{n}\beta\left\|\overline{\boldsymbol{x}}_{q,k} - \boldsymbol{x}^i_{q,\ell}\right\|\right] \qquad \text{(by } \beta\text{-smoothness)} \\
&\leq \beta D
\end{aligned} \tag{64}$$

Claim 2

$$\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla \overline{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) - \widetilde{\boldsymbol{a}}^i_{q,k}\right\|\right] \leq \beta D + \left(N + \sqrt{M}\right)3K^{2/3} \tag{65}$$

Proof of claim.

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla\overline{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) - \tilde{\boldsymbol{a}}^{i}_{q,k}\right\|\right] \\
&\qquad \leq \sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla\overline{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) - \nabla\hat{F}_{q,k-1}\right\|\right] + \sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla\hat{F}_{q,k-1} - \widetilde{\boldsymbol{a}}^{i}_{q,k}\right\|\right] \\
&\qquad \leq \beta D + \sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla\hat{F}_{q,k-1} - \hat{\boldsymbol{d}}^{i}_{q,k-1}\right\|\right] + \sum_{k=1}^{K}\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^{i}_{q,k-1} - \widetilde{\boldsymbol{a}}^{i}_{q,k}\right\|\right] \qquad (66)
\end{aligned}
$$

where we have used Claim 1 and triangle inequality in the last inequality. Using Lemma 11, we have

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla\hat{F}_{q,k-1} - \hat{\boldsymbol{d}}^{i}_{q,k-1}\right\|\right] \\
&\qquad = \sum_{k=1}^{K/2}\mathbb{E}\left[\left\|\nabla\hat{F}_{q,k-1} - \hat{\boldsymbol{d}}^{i}_{q,k-1}\right\|\right] + \sum_{k=K/2+1}^{K}\mathbb{E}\left[\left\|\nabla\hat{F}_{q,k-1} - \hat{\boldsymbol{d}}^{i}_{q,k-1}\right\|\right] \\
&\qquad \leq \sum_{k=1}^{K/2}\frac{N}{k} + \sum_{k=K/2+1}^{K}\frac{N}{K-k+1} \qquad (67)
\end{aligned}
$$

By Lemma 12, we also have

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^{i}_{q,k-1} - \widetilde{\boldsymbol{a}}^{i}_{q,k}\right\|\right] \\
&\qquad = \sum_{k=1}^{K/2}\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^{i}_{q,k-1} - \widetilde{\boldsymbol{a}}^{i}_{q,k}\right\|\right] + \sum_{k=K/2+1}^{K}\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^{i}_{q,k-1} - \widetilde{\boldsymbol{a}}^{i}_{q,k}\right\|\right] \\
&\qquad \leq \sum_{k=1}^{K/2}\frac{\sqrt{M}}{(k+4)^{1/3}} + \sum_{k=K/2+1}^{K}\frac{\sqrt{M}}{(K-k+1)^{1/3}} \qquad (68)
\end{aligned}
$$

Combining equation (67) and equation (68), equation (66) is written as

$$\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla \overline{F}_{q,k-1}(\overline{\boldsymbol{x}}_q^k)-\tilde{\boldsymbol{a}}_{q,k}^i\right\|\right] \\
&\quad\leq \beta D+\sum_{k=1}^{K/2}\left(\frac{N}{k}+\frac{\sqrt{M}}{(k+4)^{1/3}}\right)+\sum_{k=K/2+1}^{K}\left(\frac{N}{K-k+1}+\frac{\sqrt{M}}{(K-k+1)^{1/3}}\right) \\
&\quad\leq \beta D+\left(N+\sqrt{M}\right)\sum_{k=1}^{K/2}\frac{1}{(k+4)^{1/3}}+\left(N+\sqrt{M}\right)\sum_{k=K/2+1}^{K}\frac{1}{(K-k+1)^{1/3}} \\
&\quad\leq \beta D+\left(N+\sqrt{M}\right)\sum_{k=1}^{K/2}\frac{1}{k^{1/3}}+\left(N+\sqrt{M}\right)\sum_{l=1}^{K/2}\frac{1}{l^{1/3}} \\
&\quad\leq \beta D+2\left(N+\sqrt{M}\right)\int_0^{K/2}\frac{1}{s^{1/3}}ds \\
&\quad\leq \beta D+2\left(N+\sqrt{M}\right)\frac{3}{2}\left(\frac{K}{2}\right)^{2/3} \\
&\quad\leq \beta D+\left(N+\sqrt{M}\right)3K^{2/3}
\end{aligned} \tag{69}$$

A.1. Proof of Theorem 3

Theorem 3 *Given a convex set $\mathcal{K}$ with diameters D. Assume that function F_t are convex, β-smooth and $\|\nabla F_t\|\leq G$ for $t\in[T]$. Setting $Q=T^{2/5}, K=T^{3/5}, T=QK$ and step-size $\eta_k=\frac{1}{k}$. Let $\rho_k=\frac{2}{(k+3)^{2/3}}$ and $\rho_k=\frac{1.5}{(K-k+2)^{2/3}}$ when $k\in\left[1,\frac{K}{2}\right]$ and $k\in\left[\frac{K}{2}+1,K\right]$ respectively. Then, the expected regret of Algorithm 1 is at most*

$$\mathbb{E}\left[\mathcal{R}_T\right]\leq\left(GD+2\beta D^2\right)T^{2/5}+\left(C+6D\left(N+\sqrt{M}\right)\right)T^{4/5}+\frac{3}{5}\beta D^2T^{2/5}\log(T) \tag{70}$$

where $N=k_0\cdot nG\max\{\lambda_2\left(1+\frac{2}{1-\lambda_2}\right),2\}$ and $M=\max\{M_1,M_2\}$ where $M_0=4\left(V_{\boldsymbol{d}}^2+\sigma_1^2\right)+128V_{\boldsymbol{d}}^2$, $M_1=\max\left\{5^{2/3}\left(V_{\boldsymbol{d}}+\frac{2}{4^{2/3}}G_0\right)^2,M_0\right\}$ and $M_2=2.55\left(V_{\boldsymbol{d}}^2+\sigma_1^2\right)+\frac{28V_{\boldsymbol{d}}^2}{3}$. All the constant are defined in Lemma 1, Lemma 2, Lemma 11 and Lemma 12.

Proof

$$\begin{aligned}
&\mathbb{E}\left[\bar{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k+1})-\bar{F}_{q,k-1}(\boldsymbol{x}^*)\right] \\
&\quad\leq(1-\eta_k)\mathbb{E}\left[\bar{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k})-\bar{F}_{q,k-1}(\boldsymbol{x}^*)\right]+\frac{\eta_k}{n}\sum_{i=1}^{n}\mathbb{E}\left[\langle\tilde{\boldsymbol{a}}_{q,k}^i,\boldsymbol{v}_{q,k}^i-\boldsymbol{x}^*\rangle\right] \\
&\quad\quad+\frac{\eta_k}{n}D\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla\bar{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k})-\tilde{\boldsymbol{a}}_{q,k}^i\right\|\right]+\frac{\beta}{2}\eta_k^2D^2
\end{aligned} \tag{71}$$

As $\mathbb{E}\left[\bar{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) - \bar{F}_{q,k-1}(\boldsymbol{x}^*)\right] = \mathbb{E}\left[\bar{F}_{q,k-2}(\overline{\boldsymbol{x}}_{q,k}) - \bar{F}_{q,k-2}(\boldsymbol{x}^*)\right]$, we can apply equation (71) recursively for $k \in \{1, \ldots, K\}$, thus

$$\begin{aligned}
&\mathbb{E}\left[\bar{F}_{q,0}(\overline{\boldsymbol{x}}_q) - \bar{F}_{q,0}(\boldsymbol{x}^*)\right] \\
&\leq \prod_{k=1}^{K}(1-\eta_k)\mathbb{E}\left[\bar{F}_{q,0}(\overline{\boldsymbol{x}}_{q,1}) - \bar{F}_{q,0}(\boldsymbol{x}^*)\right] + \sum_{k=1}^{K}\prod_{k'=k+1}^{K}(1-\eta_{k'})\frac{\eta_k}{n}\sum_{i=1}^{n}\mathbb{E}\left[\langle \widetilde{\boldsymbol{a}}_{q,k}^i, \boldsymbol{v}_{q,k}^i - \boldsymbol{x}^*\rangle\right] \\
&\quad + \sum_{k=1}^{K}\prod_{k'=k+1}^{K}(1-\eta_{k'})\frac{\eta_k}{n}D\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla \bar{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) - \widetilde{\boldsymbol{a}}_{q,k}^i\right\|\right] + \frac{\beta}{2}D^2\sum_{k=1}^{K}\prod_{k'=k+1}^{K}(1-\eta_{k'})\eta_k^2
\end{aligned} \tag{72}$$

Choosing $\eta_k = \frac{1}{k}$, we have

$$\prod_{k=r}^{K}(1-\eta_k) \leq \exp\left(-\sum_{k=r}^{K}\frac{1}{k}\right) \leq \frac{r}{K}$$

We have then,

$$\begin{aligned}
&\mathbb{E}\left[\bar{F}_{q,0}(\overline{\boldsymbol{x}}_q) - \bar{F}_{q,0}(\boldsymbol{x}^*)\right] \\
&\leq \frac{1}{K}\mathbb{E}\left[\bar{F}_{q,0}(\overline{\boldsymbol{x}}_{q,1}) - \bar{F}_{q,0}(\boldsymbol{x}^*)\right] + \sum_{k=1}^{K}\frac{k+1}{K}\cdot\frac{1}{k}\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\langle \tilde{\boldsymbol{a}}_{q,k}^i, \boldsymbol{v}_{q,k}^i - \boldsymbol{x}^*\rangle\right] \\
&\quad + \sum_{k=1}^{K}\frac{k+1}{K}\cdot\frac{1}{k}\cdot\frac{1}{n}D\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla \bar{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) - \tilde{\boldsymbol{a}}_{q,k}^i\right\|\right] + \frac{\beta}{2}D^2\sum_{k=1}^{K}\frac{k+1}{K}\cdot\frac{1}{k^2}
\end{aligned} \tag{73}$$

Which maybe simplified by using $\frac{k+1}{K}\cdot\frac{1}{k} \leq \frac{2}{K}$.

$$\begin{aligned}
&\mathbb{E}\left[\bar{F}_{q,0}(\overline{\boldsymbol{x}}_q) - \bar{F}_{q,0}(\boldsymbol{x}^*)\right] \\
&\leq \frac{1}{K}\mathbb{E}\left[\bar{F}_{q,0}(\overline{\boldsymbol{x}}_{q,1}) - \bar{F}_{q,0}(\boldsymbol{x}^*)\right] + \frac{2}{K}\cdot\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbb{E}\left[\langle \widetilde{\boldsymbol{a}}_{q,k}^i, \boldsymbol{v}_{q,k}^i - \boldsymbol{x}^*\rangle\right] \\
&\quad + \frac{2}{K}\cdot\frac{1}{n}D\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla \bar{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) - \widetilde{\boldsymbol{a}}_{q,k}^i\right\|\right] + \frac{\beta D^2}{2}\frac{2}{K}\sum_{k=1}^{K}\frac{1}{k} \\
&\leq \frac{GD}{K} + \frac{2}{K}\cdot\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbb{E}\left[\langle \widetilde{\boldsymbol{a}}_{q,k}^i, \boldsymbol{v}_{q,k}^i - \boldsymbol{x}^*\rangle\right] \\
&\quad + \frac{2}{K}\cdot D\left(\beta D + \left(N + \sqrt{M}\right)3K^{2/3}\right) + \frac{\beta D^2}{K}\log K
\end{aligned} \tag{74}$$

where we have used Claim 2, G-Lipschitz property of $\bar{F}_{q,0}$ and boundedness of $\mathcal{K}$. Since $T = QK$ and assume that the oracle at round k has a regret of order $\mathcal{O}\left(\sqrt{Q}\right)$, i.e

$$\mathbb{E}\left[\sum_{q=1}^{Q} \langle \widetilde{\boldsymbol{a}}_{q,k}^{i}, \boldsymbol{v}_{q,k}^{i} - \boldsymbol{x}^{*} \rangle\right] \leq C\sqrt{Q}$$

then, the expected regret of the algorithm upper bounded by

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}_T\right] &= \mathbb{E}\left[\sum_{q=1}^{Q} K\left(\bar{F}_{q,0}(\overline{\boldsymbol{x}}_q) - \bar{F}_{q,0}(\boldsymbol{x}^*)\right)\right] \\
&\leq QGD + CKQ^{1/2} + 2QD\left(\beta D + \left(N + \sqrt{M}\right)3K^{2/3}\right) + Q\beta D^2 \log K \\
&\leq QGD + CKQ^{1/2} + 2Q\beta D^2 + 6D\left(N + \sqrt{M}\right)QK^{2/3} + Q\beta D^2 \log K \\
&\leq \left(GD + 2\beta D^2\right)Q + CKQ^{1/2} + 6D\left(N + \sqrt{M}\right)QK^{2/3} + Q\beta D^2 \log K \qquad (75)
\end{aligned}$$

Setting $Q = T^{2/5}$ and $K = T^{3/5}$, we have

$$\mathbb{E}\left[\mathcal{R}_T\right] \leq \left(GD + 2\beta D^2\right)T^{2/5} + \left(C + 6D\left(N + \sqrt{M}\right)\right)T^{4/5} + \frac{3}{5}\beta D^2 T^{2/5}\log(T) \qquad (76)$$

■

A.2. Proof of Theorem 4

Lemma 13 *If F_t is monotone continous DR-submodular and β-smoothness, $\boldsymbol{x}_{t,k+1} = \boldsymbol{x}_{t,k} + \frac{1}{K}\boldsymbol{v}_{t,k}$ for $k \in [1, \cdots, K]$, then*

$$\begin{aligned}
F_t(\boldsymbol{x}^*) - F_t(\boldsymbol{x}_{t,k+1}) &\leq (1 - 1/K)\left[F_t(\boldsymbol{x}^*) - F_t(\boldsymbol{x}_{t,k})\right] \qquad (77) \\
&\quad - \frac{1}{K}\left[-\left\|\nabla F_t(\boldsymbol{x}_{t,k}) - \boldsymbol{d}_{t,k}\right\| D + \langle \boldsymbol{d}_{t,k}, \boldsymbol{v}_{t,k} - \boldsymbol{x}^*\rangle\right] + \frac{\beta D^2}{2K^2}
\end{aligned}$$

Proof The proof is essentially based on the analysis of Chen et al. (2018). By β-smoothness of F_t,

$$\begin{aligned}
F_t(\boldsymbol{x}_{t,k+1}) &\geq F_t(\boldsymbol{x}_{t,k}) + \langle F_t(\boldsymbol{x}_{t,k}), \boldsymbol{x}_{t,k+1} - \boldsymbol{x}_{t,k}\rangle - \frac{\beta}{2}\left\|\boldsymbol{x}_{t,k+1} - \boldsymbol{x}_{t,k}\right\|^2 \\
&\geq F_t(\boldsymbol{x}_{t,k}) + \frac{1}{K}\left\langle F_t(\boldsymbol{x}_{t,k}), \boldsymbol{v}_{t,k}^{i}\right\rangle - \frac{\beta}{2}\frac{D^2}{K^2} \qquad \text{(since } \|\boldsymbol{v}_{t,k}\| \leq D\text{)} \\
&\geq F_t(\boldsymbol{x}_{t,k}) + \frac{1}{K}\left[\langle \nabla F_t(\boldsymbol{x}_{t,k}) - \boldsymbol{d}_{t,k}, \boldsymbol{v}_{t,k} - \boldsymbol{x}^*\rangle + \langle \nabla F_t(\boldsymbol{x}_{t,k}), \boldsymbol{x}^*\rangle + \langle \boldsymbol{d}_{t,k}, \boldsymbol{v}_{t,k} - \boldsymbol{x}^*\rangle\right] - \frac{\beta}{2}\frac{D^2}{K^2} \qquad (78)
\end{aligned}$$

By Cauchy-Schwarz's inequality, note that,

$$\langle \nabla F_t(\boldsymbol{x}_{t,k}) - \boldsymbol{d}_{t,k}, \boldsymbol{v}_{t,k} - \boldsymbol{x}^*\rangle \geq -\left\|\nabla F_t(\boldsymbol{x}_{t,k}) - \boldsymbol{t}_{t,k}\right\| D$$

Using concavity along non-negative direction and monotonicity of F_t, we have,

$$\begin{aligned} F_t(\boldsymbol{x}^*) - F_t(\boldsymbol{x}_{t,k}) &\leq F_t(\boldsymbol{x}^* \vee \boldsymbol{x}_{t,k}) - F_t(\boldsymbol{x}_{t,k}) \\ &\leq \langle \nabla F_t(\boldsymbol{x}_{t,k}), (\boldsymbol{x}^* \vee \boldsymbol{x}_{t,k}) - \boldsymbol{x}_{t,k} \rangle \\ &= \langle \nabla F_t(\boldsymbol{x}_{t,k}), (\boldsymbol{x}^* - \boldsymbol{x}_{t,k}) \vee 0 \rangle \\ &\leq \langle \nabla F_t(\boldsymbol{x}_{t,k}), \boldsymbol{x}^* \rangle \end{aligned} \tag{79}$$

then, equation (78) becomes

$$\begin{aligned} F_t(\boldsymbol{x}_{t,k+1}) &\geq F_t(\boldsymbol{x}_{t,k}) + \langle F_t(\boldsymbol{x}_{t,k}), \boldsymbol{x}_{t,k+1} - \boldsymbol{x}_{t,k} \rangle - \frac{\beta}{2} \left\| \boldsymbol{x}_{t,k+1} - \boldsymbol{x}_{t,k} \right\|^2 \\ &\geq F_t(\boldsymbol{x}_{t,k}) + \frac{1}{K} \left[- \left\| \nabla F_t(\boldsymbol{x}_{t,k}) - \boldsymbol{t}_{t,k} \right\| D + F_t(\boldsymbol{x}^*) - F_t(\boldsymbol{x}_{t,k}) + \langle \boldsymbol{d}_{t,k}, \boldsymbol{v}_{t,k} - \boldsymbol{x}^* \rangle \right] - \frac{\beta}{2} \frac{D^2}{K^2} \end{aligned} \tag{80}$$

Adding and substracting $F_t(\boldsymbol{x}^*)$ and multiply both side by -1 yields lemma 13. ■

Theorem 4 *Given a convex set $\mathcal{K}$ with diameters D. Assume that functions F_t are monotone continuous DR-Submodular, β-smooth and G-Lipschitz. Setting $Q = T^{2/5}, K = T^{3/5}, T = QK$ and step-size $\eta_k = \frac{1}{K}$. Let $\rho_k = \frac{2}{(k+3)^{2/3}}$ and $\rho_k = \frac{1.5}{(K-k+2)^{2/3}}$ when $1 \leq k \leq \frac{K}{2} + 1$ and $\frac{K}{2} + 1 \leq k \leq K$ respectively. Then, the expected $\left(1 - \frac{1}{e}\right)$-regret is at most*

$$\mathbb{E}\left[\mathcal{R}_T\right] \leq \frac{3}{2} \beta D^2 T^{2/5} + \left(C + 3D(N + \sqrt{M} \right) T^{4/5} \tag{81}$$

where the constant are defined in Theorem 3

Proof

We apply Lemma 13 with $F_t = \bar{F}_{q,k-1}$, $\boldsymbol{x}_{t,k} = \overline{\boldsymbol{x}}_{q,k}$ and $\boldsymbol{d}_{t,k} = \frac{1}{n} \sum_{i=1}^n \widetilde{\boldsymbol{a}}^i_{q,k}$, we have

$$\begin{aligned} \bar{F}_{q,k-1}(\boldsymbol{x}^*) - \bar{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k+1}) &\leq \left(1 - \frac{1}{K}\right) \left[\bar{F}_{q,k-1}(\boldsymbol{x}^*) - \bar{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) \right] \\ &+ \frac{1}{K} \cdot \frac{1}{n} \sum_{i=1}^n \left[\left\| \nabla \bar{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) - \widetilde{\boldsymbol{a}}^i_{q,k} \right\| D + \left\langle \widetilde{\boldsymbol{a}}^i_{q,k}, \boldsymbol{x}^* - \boldsymbol{v}^i_{q,k} \right\rangle \right] + \frac{\beta}{2} \frac{D^2}{K^2} \end{aligned} \tag{82}$$

As $\mathbb{E}\left[\bar{F}_{q,k-1}(\boldsymbol{x}^*) - \bar{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k})\right] = \mathbb{E}\left[\bar{F}_{q,k-2}(\boldsymbol{x}^*) - \bar{F}_{q,k-2}(\overline{\boldsymbol{x}}_{q,k})\right]$, we can apply equation (82) recursively for $k \in \{1, \ldots, K\}$, thus

$$\begin{aligned} \mathbb{E}\left[\bar{F}_{q,0}(\boldsymbol{x}^*) - \bar{F}_{q,0}(\overline{\boldsymbol{x}}_q)\right] &\leq \left(1 - \frac{1}{K}\right)^K \mathbb{E}\left[\bar{F}_{q,0}(\boldsymbol{x}^*) - \bar{F}_{q,0}(\overline{\boldsymbol{x}}_{q,1})\right] \\ &+ \frac{1}{K} \cdot \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \mathbb{E}\left[\left\| \nabla \bar{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) - \widetilde{\boldsymbol{a}}^i_{q,k} \right\| D \right] + \frac{1}{K} \cdot \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \mathbb{E}\left[\left\langle \widetilde{\boldsymbol{a}}^i_{q,k}, \boldsymbol{x}^* - \boldsymbol{v}^i_{q,k} \right\rangle \right] + \frac{\beta}{2} \frac{D^2}{K} \end{aligned} \tag{83}$$

Note that $\left(1-\frac{1}{K}\right)^K \le \frac{1}{e}$ and $\bar{F}_{q,0}(\overline{\boldsymbol{x}}_{q,1}) \ge 0$, we have

$$\mathbb{E}\left[\left(1-\frac{1}{e}\right)\bar{F}_{q,0}(\boldsymbol{x}^*) - \bar{F}_{q,0}(\overline{\boldsymbol{x}}_q)\right] \le \frac{1}{K}\cdot\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla \bar{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) - \widetilde{\boldsymbol{a}}^i_{q,k}\right\| D\right] + \frac{1}{K}\cdot\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}\left[\langle \widetilde{\boldsymbol{a}}^i_{q,k}, \boldsymbol{x}^* - \boldsymbol{v}^i_{q,k}\rangle\right] + \frac{\beta}{2}\frac{D^2}{K} \tag{84}$$

Let $T = QK$, using Claim 2 and note that the oracle has a regret $\mathcal{R}_{\mathcal{Q}} \le C\sqrt{Q}$. We have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}_T\right] &= \mathbb{E}\left[\sum_{q=1}^{Q} K\left[\left(1-\frac{1}{e}\right)\bar{F}_{q,0}(\boldsymbol{x}^*) - \bar{F}_{q,0}(\overline{\boldsymbol{x}}_q)\right]\right] \\
&\le \frac{D}{n}\sum_{q=1}^{Q}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla \bar{F}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) - \widetilde{\boldsymbol{a}}^i_{q,k}\right\|\right] + \frac{1}{n}\sum_{q=1}^{Q}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}\left[\langle \widetilde{\boldsymbol{a}}^i_{q,k}, \boldsymbol{x}^* - \boldsymbol{v}^i_{q,k}\rangle\right] + \frac{\beta}{2}QD^2 \\
&\le QD\left(\beta D + \left(N+\sqrt{M}\right)3K^{2/3}\right) + KC\sqrt{Q} + \frac{\beta Q D^2}{2}
\end{aligned} \tag{85}$$

Setting $Q = T^{2/5}$ and $K = T^{3/5}$, the expected regret of the algorithm is upper bounded by

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}_T\right] &\le T^{2/5}\left(\beta D^2 + \left(N+\sqrt{M}\right)3T^{2/5}\right) + CT^{4/5} + \frac{\beta D^2 T^{2/5}}{2} \\
&\le \frac{3}{2}\beta D^2 T^{2/5} + \left(C + 3D(N+\sqrt{M}\right)T^{4/5}
\end{aligned} \tag{86}$$

■

Appendix B. Theoretical analysis for Section 4

Let $f^\delta_t(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{v}\in\mathbb{B}^d}\left[f_t\left(\boldsymbol{x}+\delta\boldsymbol{v}\right)\right]$ and recall its gradient $\nabla f^\delta_t(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\left[\frac{d}{\delta}f_t\left(\boldsymbol{x}+\delta\boldsymbol{u}\right)\boldsymbol{u}\right]$. We define the average function

$$\bar{F}^\delta_{q,k}(\boldsymbol{x}) = \frac{1}{L-k}\sum_{\ell=k+1}^{L}F^\delta_{\sigma_q(\ell)}(\boldsymbol{x}) = \frac{1}{L-k}\sum_{\ell=k+1}^{L}\frac{1}{n}\sum_{i=1}^{n}f^{i,\delta}_{\sigma_q(\ell)}(\boldsymbol{x}) \tag{87}$$

and the average of the remaining $(L-k)$ functions of $f^{i,\delta}_{\sigma_q(\ell)}(\boldsymbol{x}^i_{q,\ell})$ over n agents as

$$\hat{F}^\delta_{q,k} = \frac{1}{n}\sum_{i=1}^{n}\hat{f}^{i,\delta}_{q,k} = \frac{1}{L-k}\sum_{\ell=k+1}^{L}\frac{1}{n}\sum_{i=1}^{n}f^{i,\delta}_{\sigma_q(\ell)}(\boldsymbol{x}^i_{q,\ell}) \tag{88}$$

where $F^\delta_{\sigma_q(\ell)}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}f^{i,\delta}_{\sigma_q(\ell)}(\boldsymbol{x})$ and $\hat{f}^{i,\delta}_{q,k} = \frac{1}{L-k}\sum_{\ell=k+1}^{L}f^{i,\delta}_{\sigma_q(\ell)}(\boldsymbol{x}^i_{q,\ell})$. Then, the one-point gradient $\nabla\bar{F}^\delta_{q,k}$ and $\nabla\hat{F}^\delta_{q,k}$ come naturally with the above definitions. Let $\mathcal{H}_{q,1} \subset \cdots \subset \mathcal{H}_{q,k}$

be the σ-fields generated by the randomness of the stochastic gradient estimate up to time k.

$$\boldsymbol{g}_{q,k}^{i,\delta} = \mathbb{E}\left[\widetilde{\boldsymbol{g}}_{q,k}^{i}|\mathcal{H}_{q,k-1}\right], \quad \boldsymbol{d}_{q,k}^{i,\delta} = \mathbb{E}\left[\widetilde{\boldsymbol{d}}_{q,k}^{i}|\mathcal{H}_{q,k-1}\right], \quad \nabla f_{\sigma_q(k)}^{i,\delta}(\boldsymbol{x}_{q,k}^{i}) = \mathbb{E}\left[\widetilde{\boldsymbol{h}}_{q,k}^{i}\right] \tag{89}$$

and

$$\hat{\boldsymbol{g}}_{q,k}^{i,\delta} = \frac{1}{L-k}\sum_{\ell=k+1}^{L}\boldsymbol{g}_{q,\ell}^{i,\delta}, \quad \hat{\boldsymbol{d}}_{q,k}^{i,\delta} = \frac{1}{L-k}\sum_{\ell=k+1}^{L}\boldsymbol{d}_{q,\ell}^{i,\delta}, \tag{90}$$

Lemma 14 *For $i \in [n], k \in [K]$. Let $V_d^\delta = 2n\frac{d}{\delta}B\left(\frac{\lambda_2}{1-\lambda_2}+1\right)$, the local gradient is upper-bounded, i.e $\left\|\boldsymbol{d}_{q,k}^{i,\delta}\right\| \leq V_d^\delta$*

Lemma 15 *Under Assumption 3, the variance of the local gradient estimate is uniformly bounded, i.e*

$$\mathbb{E}\left[\left\|\boldsymbol{d}_{q,k}^{i,\delta} - \widetilde{\boldsymbol{d}}_{q,k}^{i,\delta}\right\|^2\right] \leq 4n\left(\frac{d}{\delta}B\right)^2\left[\frac{1}{\left(\frac{1}{\lambda_2}-1\right)^2}+2\right] \tag{91}$$

Proof By Assumption 3, we have

$$\mathbb{E}\left[\left\|\nabla f_{\sigma_q(\tau)}^{cat} - \widetilde{\boldsymbol{h}}_{q,\tau}^{cat}\right\|^2\right] = \mathbb{E}\left[\sum_{i=1}^{n}\left\|\nabla f_{\sigma_q(\tau)}^{i,\delta}\left(\boldsymbol{x}_{q,\tau}^{i}\right) - \widetilde{\boldsymbol{h}}_{q,\tau}^{i}\right\|^2\right] \leq n\left(\frac{d}{\delta}B\right)^2 \tag{92}$$

Following the same analysis in equation (41), we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{d}_{q,k}^{cat} - \widetilde{\boldsymbol{d}}_{q,k}^{cat}\right\|^2\right] \\
\leq &\mathbb{E}\left[\left(\sum_{\tau=1}^{k-1}\left\|\mathbf{W}^{k-\tau} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right\|\left\|\nabla f_{\sigma_q(\tau+1)}^{cat} - \widetilde{\boldsymbol{h}}_{q,\tau+1}^{cat} + \widetilde{\boldsymbol{h}}_{q,\tau}^{cat} - \nabla f_{\sigma_q(\tau)}^{cat}\right\|\right)^2\right] \\
&+ 4\left(\mathbb{E}\left[\left\|\mathbf{W}^{k} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right\|^2\left\|\nabla f_{\sigma_q(1)}^{cat} - \widetilde{\boldsymbol{h}}_{q,1}^{cat}\right\|^2\right] + \mathbb{E}\left[\left\|\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right\|^2\left\|\nabla f_{\sigma_q(k)}^{cat} - \widetilde{\boldsymbol{h}}_{q,k}^{cat}\right\|^2\right]\right) \\
\leq &4n\left(\frac{d}{\delta}B\right)^2\left(\sum_{\tau=1}^{k-1}\lambda_2^{k-\tau}\right)^2 + 4n\left(\frac{d}{\delta}B\right)^2\left(\lambda_2^{2k}+1\right) \\
\leq &4n\left(\frac{d}{\delta}B\right)^2\left(\frac{\lambda_2}{1-\lambda_2}\right)^2 + 4n\left(\frac{d}{\delta}B\right)^2(\lambda_2+1) \leq 4n\left(\frac{d}{\delta}B\right)^2\left[\frac{1}{\left(\frac{1}{\lambda_2}-1\right)^2}+2\right]
\end{aligned} \tag{93}$$

The lemma follows by remarking that $\mathbb{E}\left[\left\|\boldsymbol{d}_{q,k}^{i,\delta} - \widetilde{\boldsymbol{d}}_{q,k}^{i}\right\|^2\right] \leq \mathbb{E}\left[\left\|\boldsymbol{d}_{q,k}^{cat} - \widetilde{\boldsymbol{d}}_{q,k}^{cat}\right\|^2\right]$ ■

Let $\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}\in\mathcal{K}}\sum_{t=1}^{T} f_t(\boldsymbol{x})$, $\boldsymbol{x}_\delta^* = \arg\max_{\boldsymbol{x}\in\mathcal{K}'}\sum_{t=1}^{T} f_t(\boldsymbol{x})$ Let $\boldsymbol{z}_{q,k}^{i} = \boldsymbol{x}_{q,k}^{i} + \delta\boldsymbol{u}_{q,k}^{i}$, we define $\overline{\boldsymbol{z}}_{q,k} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_{q,k}^{i}$ for $1 \leq k \leq K$.

Lemma 16 *Under Assumption 1 and Assumption 3. Let* $N = k_0 \cdot nB\frac{d}{\delta}\max\left\{\lambda_2\left(1+\frac{2}{1-\lambda_2}\right), 2\right\}$. *Then, for* $k \in [K]$,*we have*

$$\max_{i\in[1,n]} \mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k}^{i,\delta} - \nabla \hat{F}_{q,k}^{\delta}\right\|\right] \leq \frac{N}{k} \tag{94}$$

Proof The proof is essentially based on the one of Lemma 11. Note that we keep the same notation with a superscipt δ to indicate the smooth version of f and related variables. By definition of the one-point gradient estimator and Assumption 3, equation (25) becomes

$$\begin{aligned}
\mathbb{E}\left[\left\|\delta_{q,k}^{cat,\delta}\right\|^2\right] &= \mathbb{E}\left[\sum_{i=1}^{n}\left\|\delta_{q,k}^{i,\delta}\right\|^2\right] = \sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla \hat{f}_{q,k}^{i,\delta} - \nabla \hat{f}_{q,k-1}^{i,\delta}\right\|^2\right] \\
&= \sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\left\|\nabla \hat{f}_{q,k}^{i,\delta} - \nabla \hat{f}_{q,k-1}^{i,\delta}\right\|^2 \mid \mathcal{F}_{q,k-1}\right]\right] \\
&= \sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\left\|\frac{\sum_{\ell=k+1}^{L}\nabla f_{\sigma_q(\ell)}^{i,\delta}(\boldsymbol{x}_{q,\ell}^{i})}{L-k} - \frac{\sum_{\ell=k}^{L}\nabla f_{\sigma_q(\ell)}^{i,\delta}(\boldsymbol{x}_{q,\ell}^{i})}{L-k+1}\right\|^2 \mid \mathcal{F}_{q,k-1}\right]\right] \\
&= \sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\left\|\frac{\sum_{\ell=k+1}^{L}\nabla f_{\sigma_q(\ell)}^{i,\delta}(\boldsymbol{x}_{q,\ell}^{i})}{(L-k)(L-k+1)} - \frac{\nabla f_{\sigma_q(k)}^{i,\delta}(\boldsymbol{x}_{q,k}^{i})}{L-k+1}\right\|^2 \mid \mathcal{F}_{q,k-1}\right]\right] \\
&\leq n\left(\frac{2B\frac{d}{\delta}}{L-k+1}\right)^2
\end{aligned} \tag{95}$$

By Jensen's inequality, we deduce that

$$\mathbb{E}\left[\left\|\delta_{q,k}^{cat,\delta}\right\|\right] \leq \sqrt{\mathbb{E}\left[\left\|\delta_{q,k}^{cat,\delta}\right\|^2\right]} \leq \frac{2\sqrt{n}B\frac{d}{\delta}}{L-k+1} \tag{96}$$

When $k=1$, following the same derivation in equation (27), we have

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,1}^{cat,\delta} - \nabla \hat{F}_{q,1}^{cat,\delta}\right\|^2\right] \leq \lambda_2^2\mathbb{E}\left[\sum_{i=1}^{n}\left\|\hat{\boldsymbol{g}}_{q,1}^{i,\delta} - \nabla \hat{F}_{q,1}^{\delta}\right\|^2\right] \leq n\lambda_2^2\frac{d^2}{\delta^2}B^2$$

Let $k \in [2, k_0]$, from equation (23) and equation (96)

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k}^{cat,\delta} - \nabla \hat{F}_{q,k}^{cat,\delta}\right\|\right] &\leq \lambda_2\left(\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{cat,\delta} - \nabla \hat{F}_{q,k-1}^{cat,\delta}\right\|\right] + \mathbb{E}\left[\left\|\delta_{q,k}^{cat,\delta}\right\|\right]\right) \\
&\leq \lambda_2^{k-1}\sqrt{n}\frac{d}{\delta}B + 2\sum_{\tau=1}^{k}\lambda_2^{\tau}\sqrt{n}\frac{d}{\delta}B \\
&\leq \lambda_2\sqrt{n}\frac{d}{\delta}B + 2\frac{\lambda_2}{1-\lambda_2}\sqrt{n}\frac{d}{\delta}B \\
&= \lambda_2\sqrt{n}\frac{d}{\delta}B\left(1+\frac{2}{1-\lambda_2}\right)
\end{aligned} \tag{97}$$

Let $N_0 = k_0 \cdot \sqrt{n} \max\left\{\lambda_2 B \frac{d}{\delta}\left(1 + \frac{2}{1-\lambda_2}\right), 2B\frac{d}{\delta}\right\}$. We claim that $\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k}^{cat,\delta} - \nabla \hat{F}_{q,k}^{cat,\delta}\right\|\right] \leq \frac{N_0}{k}$ when $k \in [k_0, K]$. Let $L \geq 2K$, we have then $\frac{1}{L-k+1} \leq \frac{1}{2K-k+1} \leq \frac{1}{K+1} \leq \frac{1}{k+1}$. Thus, using the induction hypothesis, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k}^{cat,\delta} - \nabla \hat{F}_{q,k}^{cat,\delta}\right\|\right] &\leq \lambda_2 \left(\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{cat,\delta} - \nabla \hat{F}_{q,k-1}^{cat,\delta}\right\|\right] + \mathbb{E}\left[\left\|\delta_{q,k}^{cat,\delta}\right\|\right]\right) \\
&\leq \lambda_2 \left(\frac{N_0}{k-1} + \frac{2\sqrt{n}B\frac{d}{\delta}}{L-k+1}\right) \\
&\leq \lambda_2 \left(\frac{N_0}{k-1} + \frac{2\sqrt{n}B\frac{d}{\delta}}{k+1}\right) \\
&\leq \lambda_2 \left(N_0 \frac{k_0+1}{k_0(k-1)}\right) \\
&\leq \frac{N_0}{k} \qquad (98)
\end{aligned}$$

Using the inequality in equation (31) and the above result, the lemma is then proven. ∎

Lemma 17 (Lemma 10, Lemma 11 Zhang et al. (2019)) *Under Lemma 14 and lemma 15 and setting* $\rho_k = \frac{2}{(k+3)^{2/3}}$, *we have*

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta} - \widetilde{\boldsymbol{a}}_{q,k}^{i}\right\|\right] \leq \frac{\sqrt{M_0}}{(k+3)^{1/3}}, \qquad k \in [K] \qquad (99)$$

where $M_0 = 4^{2/3}\frac{d^2}{\delta^2}B^2\left[24n^2\left(\frac{1}{\frac{1}{\lambda_2}-1} + 1\right)^2 + 8n\left(\frac{1}{\left(\frac{1}{\lambda_2}-1\right)^2} + 2\right)\right]$

Proof The proof follows the same idea in Lemma 10 and Lemma 11 of Zhang et al. (2019) with some changes in the constant values. We will evoques in details in the following section. Following the same decomposition in the proof of Lemma 12, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta} - \widetilde{\boldsymbol{a}}_{q,k}^{i}\right\|^2\right] &= \mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta} - (1-\rho_k)\widetilde{\boldsymbol{a}}_{q,k-1}^{i} - \rho_k \widetilde{\boldsymbol{d}}_{q,k}^{i,\delta}\right\|^2\right] \\
&= \rho_k^2 \mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta} - \widetilde{\boldsymbol{d}}_{q,k}^{i,\delta})\right\|^2\right] + (1-\rho_k)^2 \mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta} - \hat{\boldsymbol{d}}_{q,k-2}^{i,\delta}\right\|^2\right] \\
&\quad + (1-\rho_k)^2 \mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-2}^{i,\delta} - \widetilde{\boldsymbol{a}}_{q,k-1}^{i}\right\|^2\right] \\
&\quad + 2\rho_k(1-\rho_k)\mathbb{E}\left[\left\langle \hat{\boldsymbol{d}}_{q,k-1}^{i,\delta} - \widetilde{\boldsymbol{d}}_{q,k}^{i,\delta}, \hat{\boldsymbol{d}}_{q,k-1}^{i,\delta} - \hat{\boldsymbol{d}}_{q,k-2}^{i,\delta}\right\rangle\right] \\
&\quad + 2\rho_k(1-\rho_k)\mathbb{E}\left[\left\langle \hat{\boldsymbol{d}}_{q,k-1}^{i,\delta} - \widetilde{\boldsymbol{d}}_{q,k}^{i,\delta}, \hat{\boldsymbol{d}}_{q,k-2}^{i,\delta} - \widetilde{\boldsymbol{a}}_{q,k-1}^{i}\right\rangle\right] \\
&\quad + 2(1-\rho_k)^2 \mathbb{E}\left[\left\langle \hat{\boldsymbol{d}}_{q,k-1}^{i,\delta} - \hat{\boldsymbol{d}}_{q,k-2}^{i,\delta}, \hat{\boldsymbol{d}}_{q,k-2}^{i,\delta} - \widetilde{\boldsymbol{a}}_{q,k-1}^{i}\right\rangle\right]
\end{aligned}$$

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta}-\widetilde{\boldsymbol{d}}_{q,k}^{i,\delta})\right\|^2\right]=\mathbb{E}\left[\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta}-\widetilde{\boldsymbol{d}}_{q,k}^{i,\delta})\right\|^2\mid\mathcal{F}_{q,k-1}\right]\right]$$
$$\leq\mathbb{E}\left[\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta}-\boldsymbol{d}_{q,k}^{i,\delta}\right\|^2+\left\|\boldsymbol{d}_{q,k}^{i,\delta}-\widetilde{\boldsymbol{d}}_{q,k}^{i,\delta})\right\|^2+2\langle\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta}-\boldsymbol{d}_{q,k}^{i,\delta},\boldsymbol{d}_{q,k}^{i,\delta}-\widetilde{\boldsymbol{d}}_{q,k}^{i,\delta}\rangle\mid\mathcal{F}_{q,k-1}\right]\right] \tag{100}$$

By the definition in equation (90), we have $\mathbb{E}\left[\boldsymbol{d}_{q,k}^{i,\delta}|\mathcal{F}_{q,k-1}\right]=\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta}$, using Lemma 14, we have

$$\mathbb{E}\left[\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta}-\boldsymbol{d}_{q,k}^{i,\delta}\right\|^2\mid\mathcal{F}_{q,k-1}\right]\right]\leq(V_{\boldsymbol{d}}^{\delta})^2 \tag{101}$$

Invoking Lemma 15, we have

$$\mathbb{E}\left[\left\|\boldsymbol{d}_{q,k}^{i,\delta}-\widetilde{\boldsymbol{d}}_{q,k}^{i,\delta})\right\|^2\right]\leq\sigma_2^2 \tag{102}$$

and

$$\mathbb{E}\left[\mathbb{E}\left[\left\langle\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta}-\boldsymbol{d}_{q,k}^{i,\delta},\boldsymbol{d}_{q,k}^{i,\delta}-\widetilde{\boldsymbol{d}}_{q,k}^{i,\delta}\right\rangle\mid\mathcal{F}_{q,k-1}\right]\right]=0 \tag{103}$$

by following the same analysis in equation (48). We now claim that equation (100) is bounded above by

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta}-\widetilde{\boldsymbol{d}}_{q,k}^{i,\delta})\right\|^2\right]\leq(V_{\boldsymbol{d}}^{\delta})^2+\sigma_2^2\triangleq V^{\delta} \tag{104}$$

More over, taking the idea from equations (50) to (53), we have

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta}-\hat{\boldsymbol{d}}_{q,k-2}^{i,\delta}\right\|^2\right]\leq\frac{4(V_{\boldsymbol{d}}^{\delta})^2}{(L-k+2)^2}\triangleq\frac{L^{\delta}}{(L-k+2)^2} \tag{105}$$

$$\mathbb{E}\left[\left\langle\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta}-\widetilde{\boldsymbol{d}}_{q,k}^{i,\delta},\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta}-\hat{\boldsymbol{d}}_{q,k-2}^{i,\delta}\right\rangle\right]=0 \tag{106}$$

$$\mathbb{E}\left[\left\langle\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta}-\widetilde{\boldsymbol{d}}_{q,k}^{i,\delta},\hat{\boldsymbol{d}}_{q,k-2}^{i,\delta}-\widetilde{\boldsymbol{a}}_{q,k-1}^{i}\right\rangle\right]=0 \tag{107}$$

and

$$\mathbb{E}\left[\left\langle\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta}-\hat{\boldsymbol{d}}_{q,k-2}^{i,\delta},\hat{\boldsymbol{d}}_{q,k-2}^{i,\delta}-\widetilde{\boldsymbol{a}}_{q,k-1}^{i}\right\rangle\right]\leq\frac{L^{\delta}}{2\alpha_k(L-k+2)^2}+\frac{\alpha_k}{2}\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-2}^{i,\delta}-\widetilde{\boldsymbol{a}}_{q,k-1}^{i}\right\|^2\right] \tag{108}$$

by using Young's inequality. Setting $\alpha_k=\frac{\rho_k}{2}$ similarly to Lemma 12, we have

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta}-\widetilde{\boldsymbol{a}}_{q,k}^{i}\right\|^2\right]\leq\rho_k^2V^{\delta}+\left(1+\frac{2}{\rho_k}\right)\frac{L^{\delta}}{(L-k+2)^2}+(1-\rho_k)\,\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-2}^{i,\delta}-\widetilde{\boldsymbol{a}}_{q,k-1}^{i}\right\|^2\right] \tag{109}$$

Setting $L \geq 2K$ and $\rho_k = \frac{2}{(k+2)^{2/3}}$, we have then $\frac{1}{L-k+2} \leq \frac{1}{2K-k+2} \leq \frac{1}{K+2} \leq \frac{1}{k+2}$. Following the derivation from Lemma 11 of Zhang et al. (2019).Equation (109) can be bounded above by

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta} - \widetilde{\boldsymbol{a}}_{q,k}^{i}\right\|^2\right] &\leq \rho_k^2 V^\delta + \left(1 + \frac{2}{\rho_k}\right)\frac{L^\delta}{(k+2)^2} + (1-\rho_k)\,\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-2}^{i,\delta} - \widetilde{\boldsymbol{a}}_{q,k-1}^{i}\right\|^2\right] \\ &\leq \frac{4^{2/3}\left(2V^\delta + L^\delta\right)}{(k+2)^{4/3}} + \left(1 - \frac{2}{(k+2)^{2/3}}\right)\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-2}^{i,\delta} - \widetilde{\boldsymbol{a}}_{q,k-1}^{i}\right\|^2\right] \\ &\triangleq \frac{M_0}{(k+2)^{4/3}} + \left(1 - \frac{2}{(k+2)^{2/3}}\right)\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-2}^{i,\delta} - \widetilde{\boldsymbol{a}}_{q,k-1}^{i}\right\|^2\right] \qquad (110)\end{aligned}$$

Assume that $\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta} - \widetilde{\boldsymbol{a}}_{q,k}^{i}\right\|^2\right] \leq \frac{M_0}{(k+3)^{2/3}}$ for $k \in [K]$. When $k = 1$, by definition of $\widetilde{\boldsymbol{a}}_{q,1}^{i}$ and $\hat{\boldsymbol{d}}_{q,0}^{i,\delta}$, we have

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,0}^{i,\delta} - \widetilde{\boldsymbol{a}}_{q,1}^{i}\right\|^2\right] \leq \left(V_{\boldsymbol{d}}^\delta + \frac{2}{3^{2/3}}\frac{d}{\delta}B\right)^2 \qquad (111)$$

Thus, since $\sigma_2 \geq \frac{2}{3^{2/3}}\frac{d}{\delta}B$, one can observe that

$$\frac{M_0}{(1+2)^{2/3}} = 2V^\delta + L^\delta \geq 2V^\delta = 2\left((V_{\boldsymbol{d}}^\delta)^2 + \sigma_2^2\right) \geq \left(V_{\boldsymbol{d}}^\delta) + \sigma_2\right)^2 \geq \mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,0}^{i,\delta} - \widetilde{\boldsymbol{a}}_{q,1}^{i}\right\|^2\right] \qquad (112)$$

Suppose that the induction hypothesis holds for $k-1$, one can easily verify for k since

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-1}^{i,\delta} - \widetilde{\boldsymbol{a}}_{q,k}^{i}\right\|^2\right] &\leq \frac{M_0}{(k+2)^{4/3}} + \left(1 - \frac{2}{(k+2)^{2/3}}\right)\mathbb{E}\left[\left\|\hat{\boldsymbol{d}}_{q,k-2}^{i,\delta} - \widetilde{\boldsymbol{a}}_{q,k-1}^{i}\right\|^2\right] \\ &\leq \frac{M_0}{(k+2)^{4/3}} + \left(1 - \frac{2}{(k+2)^{2/3}}\right)\frac{M_0}{(k+2)^{2/3}} \\ &\leq M_0\frac{(k+2)^{2/3} - 1}{(k+3)^{4/3}} \\ &\leq \frac{M_0}{(k+3)^{2/3}} \qquad (113)\end{aligned}$$

■

Claim 3 *Under Claim 1, Lemma 16 and Lemma 17.*

$$\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla\bar{F}_{q,k-1}^{\delta}(\overline{\boldsymbol{x}}_{q,k}) - \widetilde{\boldsymbol{a}}_{q,k}^{i}\right\|\right] \leq \beta D + \frac{3}{2}\left(N + \sqrt{M_0}\right)K^{2/3} \qquad (114)$$

Proof of claim.

$$\begin{aligned}
\sum_{k=1}^{K} \mathbb{E}\left[\left\|\nabla \bar{F}^{\delta}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) - \tilde{\boldsymbol{a}}^{i}_{q,k}\right\|\right] &\leq \sum_{k=1}^{K} \mathbb{E}\left[\left\|\nabla \bar{F}^{\delta}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) - \nabla \hat{F}^{\delta}_{q,k-1}\right\|\right] + \sum_{k=1}^{K} \mathbb{E}\left[\left\|\nabla \hat{F}^{\delta}_{q,k-1} - \tilde{\boldsymbol{a}}^{i}_{q,k}\right\|\right] \\
&\leq \beta D + \sum_{k=1}^{K} \mathbb{E}\left[\left\|\nabla \hat{F}^{\delta}_{q,k-1} - \hat{\boldsymbol{d}}^{i,\delta}_{q,k-1}\right\|\right] + \sum_{k=1}^{K} \mathbb{E}\left[\left\|\hat{\boldsymbol{d}}^{i,\delta}_{q,k-1} - \tilde{\boldsymbol{a}}^{i}_{q,k}\right\|\right] \\
&\leq \beta D + \sum_{k=1}^{K} \frac{N}{k} + \sum_{k=1}^{K} \frac{\sqrt{M_0}}{(k+3)^{1/3}} \\
&\leq \beta D + \left(N + \sqrt{M_0}\right) \sum_{k=1}^{K} \frac{1}{(k+3)^{1/3}} \\
&\leq \beta D + \frac{3}{2}\left(N + \sqrt{M_0}\right) K^{2/3}
\end{aligned} \tag{115}$$

where Claim 1 is still verified in the second inequality since $f^{i,\delta}_{\sigma_q(\ell)}$ is β-smooth and the third inequality is the result of Lemma 16 and Lemma 17

Claim 4

$$\mathbb{E}\left[\sum_{q=1}^{Q} \sum_{\ell=1}^{L} \left(1 - \frac{1}{e}\right) F^{\delta}_{\sigma_q(\ell)}(\boldsymbol{x}^*_{\delta}) - F^{\delta}_{\sigma_q(\ell)}(\overline{\boldsymbol{x}}_q)\right] \leq \frac{L\beta D^2}{K} + \frac{3LD\left(N + \sqrt{M_0}\right)}{2K^{1/3}} + LC\sqrt{Q} + \frac{\beta Q L D^2}{2K} \tag{116}$$

Proof of claim. Using Lemma 13 with $F_t = \bar{F}^{\delta}_{q,k-1}$, $\boldsymbol{x}_{t,k} = \overline{\boldsymbol{x}}_{q,k}$ and $\boldsymbol{d}_{t,k} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\boldsymbol{a}}^{i}_{q,k}$, we have

$$\begin{aligned}
\bar{F}^{\delta}_{q,k-1}(\boldsymbol{x}^*_{\delta}) - \bar{F}^{\delta}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k+1}) &\leq \left(1 - \frac{1}{K}\right)\left[\bar{F}^{\delta}_{q,k-1}(\boldsymbol{x}^*_{\delta}) - \bar{F}^{\delta}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k})\right] \\
&\quad + \frac{1}{K} \cdot \frac{1}{n} \sum_{i=1}^{n} \left[\left\|\nabla \bar{F}^{\delta}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) - \tilde{\boldsymbol{a}}^{i}_{q,k}\right\| D + \langle \tilde{\boldsymbol{a}}^{i}_{q,k}, \boldsymbol{x}^*_{\delta} - \boldsymbol{v}^{i}_{q,k}\rangle\right] + \frac{\beta}{2}\frac{D^2}{K^2}
\end{aligned} \tag{117}$$

Similarly to the proof of Theorem 4 and using Claim 3, we note

$$\begin{aligned}
&\mathbb{E}\left[\left(1 - \frac{1}{e}\right)\bar{F}^{\delta}_{q,0}(\boldsymbol{x}^*_{\delta}) - \bar{F}^{\delta}_{q,0}(\overline{\boldsymbol{x}}_q)\right] \\
&\leq \frac{1}{K} \cdot \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}\left[\left\|\nabla \bar{F}^{\delta}_{q,k-1}(\overline{\boldsymbol{x}}_{q,k}) - \tilde{\boldsymbol{a}}^{i}_{q,k}\right\| D\right] + \frac{1}{K} \cdot \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}\left[\langle \tilde{\boldsymbol{a}}^{i}_{q,k}, \boldsymbol{x}^*_{\delta} - \boldsymbol{v}^{i}_{q,k}\rangle\right] + \frac{\beta}{2}\frac{D^2}{K} \\
&\leq \frac{D}{K}\left(\beta D + \frac{3}{2}\left(N + \sqrt{M_0}\right) K^{2/3}\right) + \frac{1}{K} \cdot \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}\left[\langle \tilde{\boldsymbol{a}}^{i}_{q,k}, \boldsymbol{x}^*_{\delta} - \boldsymbol{v}^{i}_{q,k}\rangle\right] + \frac{\beta}{2}\frac{D^2}{K}
\end{aligned} \tag{118}$$

Thus, we can write

$$\begin{aligned}
&\mathbb{E}\left[\sum_{q=1}^{Q}\sum_{\ell=1}^{L}\left(1-\frac{1}{e}\right)F^{\delta}_{\sigma_q(\ell)}(\boldsymbol{x}^*_\delta)-F^{\delta}_{\sigma_q(\ell)}(\overline{\boldsymbol{x}}_q)\right]=\mathbb{E}\left[\sum_{q=1}^{Q}L\left(1-\frac{1}{e}\right)\bar{F}^{\delta}_{q,0}(\boldsymbol{x}^*_\delta)-\bar{F}^{\delta}_{q,0}(\overline{\boldsymbol{x}}_q)\right]\\
&\le\frac{LD}{K}\left(\beta D+\frac{3}{2}\left(N+\sqrt{M_0}\right)K^{2/3}\right)+LC\sqrt{Q}+\frac{\beta}{2}\frac{QLD^2}{K}\\
&\le\frac{L\beta D^2}{K}+\frac{3LD\left(N+\sqrt{M_0}\right)}{2K^{1/3}}+LC\sqrt{Q}+\frac{\beta QLD^2}{2K} \qquad (119)
\end{aligned}$$

Theorem 7 *Let $\mathcal{K}$ be a down-closed convex and compact set. We suppose the δ-interior $\mathcal{K}'$ following Lemma 6. Let $Q=T^{2/9}, L=T^{7/9}, K=T^{2/3}$, $\delta=\frac{r}{\sqrt{d}+2}T^{-1/9}$ and $\rho_k=\frac{2}{(k+2)^{2/3}}$, $\eta_k=\frac{1}{K}$. Then the expected $\left(1-\frac{1}{e}\right)$-regret is upper bounded*

$$\mathbb{E}\left[\mathcal{R}_T\right]\le ZT^{8/9}+\frac{\beta D^2}{2}T^{1/9}+\frac{3}{2}D\frac{d\left(\sqrt{d}+2\right)}{r}P_{n,\lambda_2}T^{2/9}+\beta D^2T^{3/9} \qquad (120)$$

where we note $Z=\left(1-\frac{1}{e}\right)\left(\sqrt{d}\left(\frac{R}{e}+1\right)+\frac{R}{r}\right)G\frac{r}{\sqrt{d}+2}+\left(2-\frac{1}{e}\right)G\frac{r}{\sqrt{d}+2}+2\beta+C$ and $P_{n,\lambda_2}=k_0\cdot nB\max\left\{\lambda_2\left(1+\frac{2}{1-\lambda_2}\right),2\right\}+4^{1/3}\left(24n^2\left(\frac{1}{\frac{1}{\lambda_2}-1}+1\right)^2+8n\left(\frac{1}{\left(\frac{1}{\lambda_2}-1\right)^2}+2\right)\right)^{1/2}$

Proof Recall the values of N and M_0 from Lemma 16 and Lemma 17, we have

$$N=k_0\cdot nB\frac{d}{\delta}\max\left\{\lambda_2\left(1+\frac{2}{1-\lambda_2}\right),2\right\}$$

$$M_0=4^{2/3}\frac{d^2}{\delta^2}B^2\left[24n^2\left(\frac{1}{\frac{1}{\lambda_2}-1}+1\right)^2+8n\left(\frac{1}{\left(\frac{1}{\lambda_2}-1\right)^2}+2\right)\right]$$

Let $P_{n,\lambda_2}=k_0{\cdot}nB\max\left\{\lambda_2\left(1+\frac{2}{1-\lambda_2}\right),2\right\}+4^{1/3}\left(24n^2\left(\frac{1}{\frac{1}{\lambda_2}-1}+1\right)^2+8n\left(\frac{1}{\left(\frac{1}{\lambda_2}-1\right)^2}+2\right)\right)^{1/2}$. Then, one can easily see that $N+\sqrt{M_0}=\frac{d}{\delta}BP_{n,\lambda_2}$ where P_{n,λ_2} is a constant depending on n and λ_2. For the next step, we set $\delta=\frac{r}{\sqrt{d}+2}T^{-1/9}$, then $\frac{d}{\delta}=\frac{d\left(\sqrt{d}+2\right)}{r}T^{1/9}$, $Q=T^{2/9}, L=T^{7/9}$ and $K=T^{2/3}$. From the analysis in Theorem 4 of Zhang et al. (2019), Lemma 6 and Claim 4, we have

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{R}_{\mathcal{T}}\right] \leq& \left(1-\frac{1}{e}\right)\left(\sqrt{d}\left(\frac{R}{e}+1\right)+\frac{R}{r}\right)GT\delta^{\gamma}+\left(2-\frac{1}{e}\right)GT\delta+2BQK \\
&+\sum_{q=1}^{Q}\sum_{\ell=1}^{L}\left(1-\frac{1}{e}\right)F^{\delta}_{\sigma_q(\ell)}(\boldsymbol{x}^*_{\delta})-F^{\delta}_{\sigma_q(\ell)}(\overline{\boldsymbol{x}}_q) \\
\leq& \left(1-\frac{1}{e}\right)\left(\sqrt{d}\left(\frac{R}{e}+1\right)+\frac{R}{r}\right)GT\delta^{\gamma}+\left(2-\frac{1}{e}\right)GT\delta+2BQK \\
&+\frac{L\beta D^2}{K}+\frac{3LD\left(N+\sqrt{M_0}\right)}{2K^{1/3}}+LC\sqrt{Q}+\frac{\beta QLD^2}{2K} \\
\leq& \left(1-\frac{1}{e}\right)\left(\sqrt{d}\left(\frac{R}{e}+1\right)+\frac{R}{r}\right)GT\delta^{\gamma}+\left(2-\frac{1}{e}\right)GT\delta+2BQK \\
&+\frac{L\beta D^2}{K}+\frac{3LD\frac{d}{\delta}P_{n,\lambda_2}}{2K^{1/3}}+LC\sqrt{Q}+\frac{\beta QLD^2}{2K} \\
\leq& \left(1-\frac{1}{e}\right)\left(\sqrt{d}\left(\frac{R}{e}+1\right)+\frac{R}{r}\right)GT\frac{r}{\sqrt{d}+2}T^{-1/9}+\left(2-\frac{1}{e}\right)GT\frac{r}{\sqrt{d}+2}T^{-1/9} \\
&+2\beta T^{2/9}T^{2/3}+T^{7/9}\beta D^2T^{-2/3}+\frac{3}{2}T^{7/9}D\frac{d\left(\sqrt{d}+2\right)}{r}T^{1/9}P_{n,\lambda_2}T^{-2/3} \\
&+T^{7/9}CT^{1/9}+\frac{\beta}{2}T^{2/9}T^{7/9}D^2T^{-2/3} \\
\leq& \left[\left(1-\frac{1}{e}\right)\left(\sqrt{d}\left(\frac{R}{e}+1\right)+\frac{R}{r}\right)G\frac{r}{\sqrt{d}+2}+\left(2-\frac{1}{e}\right)G\frac{r}{\sqrt{d}+2}+C\right]T^{8/9} \\
&+\frac{\beta D^2}{2}T^{6/9}+\left[2\beta+\frac{3}{2}D\frac{d\left(\sqrt{d}+2\right)}{r}P_{n,\lambda_2}\right]T^{5/9}+\beta D^2T^{4/9} \\
\leq& \left[\left(1-\frac{1}{e}\right)\left(\sqrt{d}\left(\frac{R}{e}+1\right)+\frac{R}{r}\right)G\frac{r}{\sqrt{d}+2}+\left(2-\frac{1}{e}\right)G\frac{r}{\sqrt{d}+2}+2\beta+C\right]T^{8/9} \\
&+\frac{\beta D^2}{2}T^{1/9}+\frac{3}{2}D\frac{d\left(\sqrt{d}+2\right)}{r}P_{n,\lambda_2}T^{2/9}+\beta D^2T^{3/9}
\end{aligned}
\tag{121}
$$

■