A Proofs

In this section, we provide proofs for lemmas and theorems in Section 5. Throughout this section, we use $\boldsymbol{x}_{cat} \in \mathbb{R}^{nd}$ to denote the concatenation of n vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, i.e., $\boldsymbol{x}_{cat} \coloneqq [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top]^\top$, $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$ to denote the average of $\boldsymbol{x}_i$, and $\bar{\boldsymbol{x}}_{cat} \coloneqq [\bar{\boldsymbol{x}}^\top, \ldots, \bar{\boldsymbol{x}}^\top]^\top \in \mathbb{R}^{nd}$ the concatenation of n copies of $\bar{\boldsymbol{x}}$. Note that $\bar{\boldsymbol{x}}_{cat} = (\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes \boldsymbol{I}_d)\boldsymbol{x}_{cat}$, where $\boldsymbol{I}_d \in \mathbb{R}^{d\times d}$ is an identity matrix. Besides, we denote $\nabla F_{cat}(\boldsymbol{x}_{cat}) \in \mathbb{R}^{nd}$ as the vector $[\nabla F_i(\boldsymbol{x}_1)^\top, \ldots, \nabla F_i(\boldsymbol{x}_n)^\top]^\top$. We note that the proofs of Lemma 1, 2, and 3 are borrowed from (Mokhtari et al., 2018b) and we state them here for completeness.

A.1 Proof of Lemma 1

Proof. First, we observe that

$$\boldsymbol{x}_{cat}^{(t+1)} = (\boldsymbol{W} \otimes \boldsymbol{I}_d)\boldsymbol{x}_{cat}^{(t)} + \frac{1}{T}\boldsymbol{v}_{cat}^{(t)} = (\boldsymbol{W} \otimes \boldsymbol{I}_d)^t\boldsymbol{x}_{cat}^{(1)} + \frac{1}{T}\sum_{\tau=1}^{t}(\boldsymbol{W} \otimes \boldsymbol{I}_d)^{t-\tau}\boldsymbol{v}_{cat}^{(\tau)} = \frac{1}{T}\sum_{\tau=1}^{t}(\boldsymbol{W} \otimes \boldsymbol{I}_d)^{t-\tau}\boldsymbol{v}_{cat}^{(\tau)}, \tag{23}$$

where the first equality follows from the update rule of $\boldsymbol{x}_i^{(t)}$ and the second equlity holds because $\boldsymbol{x}_i^{(1)} = \mathbf{0}_d$. Let $\boldsymbol{z}_i^{(\tau)} \in \mathbb{R}^d$, $i \in [n]$, denote the i-th block of $(\boldsymbol{W} \otimes \boldsymbol{I}_d)\boldsymbol{v}_{cat}^{(\tau)}$ and $\boldsymbol{y}_i^{(\tau,t)}$ denote the i-th block of $(\boldsymbol{W} \otimes \boldsymbol{I}_d)^{t-\tau}\boldsymbol{v}_{cat}^{(\tau)}$. Note that each $\boldsymbol{v}_i^{(\tau)}$ belongs to the set $\mathcal{C}$, therefore $\boldsymbol{z}_i^{(\tau)} = \sum_{j\in\mathcal{N}(i)\cup\{i\}} \boldsymbol{W}_{ij}\boldsymbol{v}_j^{(\tau)}$ also belongs to $\mathcal{C}$, since $\boldsymbol{z}_i^{(\tau)}$ is a convex combination of $\boldsymbol{v}_j^{(\tau)}$. Hence, $\boldsymbol{y}_i^{(\tau,t)} \in \mathcal{C}$, and thus $\frac{1}{t}\sum_{\tau=1}^{t}\boldsymbol{y}_i^{(\tau,t)} \in \mathcal{C}$. From (23) we can see that, for any $t \in [T]$,

$$\mathbf{0}_d \le \boldsymbol{x}_i^{(t+1)} = \frac{1}{T}\sum_{\tau=1}^{t}\boldsymbol{y}_i^{(\tau,t)} \le \frac{1}{t}\sum_{\tau=1}^{t}\boldsymbol{y}_i^{(\tau,t)}, \tag{24}$$

which implies that $\boldsymbol{x}_i^{(t+1)} \in \mathcal{X}$ since $\mathcal{C} \subset \mathcal{X}$ and $\mathcal{X} = \{\boldsymbol{x} | \mathbf{0}_d \le \boldsymbol{x} \le \boldsymbol{u}\}$. Besides, since $\mathcal{C}$ has a radius R,

$$\|\boldsymbol{x}_i^{(t+1)}\| \le \|\frac{1}{t}\sum_{\tau=1}^{t}\boldsymbol{y}_i^{(\tau,t)}\| \le R. \tag{25}$$

Moreover, when $t = T$, we have $\boldsymbol{x}_i^{(T+1)} = \frac{1}{T}\sum_{\tau=1}^{T}\boldsymbol{y}_i^{(\tau,t)}$, which implies that $\boldsymbol{x}_i^{(T+1)} \in \mathcal{C}$.

□

A.2 Proof of Lemma 2

Proof. For $t = 1$, $\bar{\boldsymbol{x}}^{(1)} = \boldsymbol{x}_i^{(1)} = \mathbf{0}_d$, the claim (11) immediately holds. Notice that $\sqrt{\sum_{i=1}^n \|\boldsymbol{x}_i^{(t+1)} - \bar{\boldsymbol{x}}^{(t+1)}\|^2} = \|\boldsymbol{x}_{cat}^{(t+1)} - \bar{\boldsymbol{x}}_{cat}^{(t+1)}\|$, it suffices to prove that $\|\boldsymbol{x}_{cat}^{(t+1)} - \bar{\boldsymbol{x}}_{cat}^{(t+1)}\| \le \frac{\sqrt{n}R}{T(1-\beta)}$ holds for $t \in [T]$. Observe that,

$$\bar{\boldsymbol{x}}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\boldsymbol{v}_i^{(t)}}{T} + \sum_{j=1}^{n}\boldsymbol{W}_{ij}\boldsymbol{x}_j^{(t)}\Big) = \frac{\bar{\boldsymbol{v}}^{(t)}}{T} + \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n}\boldsymbol{W}_{ij}\boldsymbol{x}_j^{(t)} = \frac{\bar{\boldsymbol{v}}^{(t)}}{T} + \bar{\boldsymbol{x}}^{(t)} = \frac{1}{T}\sum_{\tau=1}^{t}\bar{\boldsymbol{v}}^{(\tau)}, \tag{26}$$

where $\bar{\boldsymbol{v}}^{(t)} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{v}_i^{(t)}$ and the third equality holds because $\boldsymbol{W}$ is doubly stochastic. Therefore,

$$\bar{\boldsymbol{x}}_{cat}^{(t+1)} = \frac{1}{T}\sum_{\tau=1}^{t}\bar{\boldsymbol{v}}_{cat}^{(\tau)} = \frac{1}{T}\sum_{\tau=1}^{t}(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes \boldsymbol{I}_d)\boldsymbol{v}_{cat}^{(\tau)}. \tag{27}$$

Combining (23) and (27), we have

$$\begin{aligned}\|\boldsymbol{x}_{cat}^{(t+1)} - \bar{\boldsymbol{x}}_{cat}^{(t+1)}\| &= \frac{1}{T}\|\sum_{\tau=1}^{t}\big((\boldsymbol{W}^{t-\tau} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top) \otimes \boldsymbol{I}_d\big)\boldsymbol{v}_{cat}^{(\tau)}\| \le \frac{1}{T}\sum_{\tau=1}^{t}\|\boldsymbol{W}^{t-\tau} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\|\|\boldsymbol{v}_{cat}^{(\tau)}\| \\ &\le \frac{\sqrt{n}R}{T}\sum_{\tau=1}^{t}\|\boldsymbol{W}^{t-\tau} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\| \le \frac{\sqrt{n}R}{T}\sum_{\tau=1}^{t}\beta^{t-\tau} \le \frac{\sqrt{n}R}{T(1-\beta)},\end{aligned} \tag{28}$$

where the first inequality follows from the fact that $\|\boldsymbol{A} \otimes \boldsymbol{I}_d\| = \|\boldsymbol{A}\|$ for any matrix $\boldsymbol{A}$; the second inequality holds because $\boldsymbol{v}_i^{(t)} \in \mathcal{C}$. To see the third inequality holds, we observe that $\boldsymbol{W}$ has an eigenvalue 1 with an eigenvector $\mathbf{1}_n$ and $(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top)$ is a rank-one matrix that has an eigenvalue 1 with eigenvector $\mathbf{1}_n$, too. Therefore, the largest eigenvalue of $\boldsymbol{W}^{t-\tau} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ is $\beta^{t-\tau}$, where β is defined in Assumption 1.

We proceed to prove the second part of Lemma 2 as follows.

$$\begin{aligned}|F(\boldsymbol{x}_i^{(t)}) - F(\bar{\boldsymbol{x}}^{(t)})| &= \Big|\frac{1}{n}\sum_{j=1}^n F_j(\boldsymbol{x}_i^{(t)}) - \frac{1}{n}\sum_{i=1}^n F_j(\bar{\boldsymbol{x}}^{(t)})\Big| \\ &\leq G\|\boldsymbol{x}_i^{(t)} - \bar{\boldsymbol{x}}^{(t)}\| \leq G\sqrt{\sum_{j=1}^n \|\boldsymbol{x}_j^{(t)} - \bar{\boldsymbol{x}}^{(t)}\|^2} \leq \frac{\sqrt{n}GR}{T(1-\beta)},\end{aligned} \tag{29}$$

where the first inequality holds because each F_i is G-Lipschitz over $\mathcal{X}$. the last inequality follows from (28). □

A.3 Proof of Lemma 3

Proof. Since each F_i is L-smooth, the global objective F is also L-smooth. Besides, Lemma 1 implies that $\boldsymbol{x}_i^{(t)} \in \mathcal{X}$ and thus $\bar{\boldsymbol{x}}^{(t)}$ also lies in $\mathcal{X}$ for any $t \in [T+1]$. Thus, for $t \in [T]$,

$$\begin{aligned}F(\bar{\boldsymbol{x}}^{(t+1)}) - F(\bar{\boldsymbol{x}}^{(t)}) &\geq \langle \nabla F(\bar{\boldsymbol{x}}^{(t)}), \bar{\boldsymbol{x}}^{(t+1)} - \bar{\boldsymbol{x}}^{(t)}\rangle - \frac{L}{2}\|\bar{\boldsymbol{x}}^{(t+1)} - \bar{\boldsymbol{x}}^{(t)}\|^2 \\ &= \frac{1}{T}\langle \nabla F(\bar{\boldsymbol{x}}^{(t)}), \bar{\boldsymbol{v}}^{(t)}\rangle - \frac{L}{2}\|\frac{1}{T}\bar{\boldsymbol{v}}^{(t)}\|^2 \\ &\geq \frac{1}{T}\langle \nabla F(\bar{\boldsymbol{x}}^{(t)}), \bar{\boldsymbol{v}}^{(t)}\rangle - \frac{LR^2}{2T^2}.\end{aligned} \tag{30}$$

where the first equality follows from (26). Next, we derive a lower bound of $\langle \nabla F(\bar{\boldsymbol{x}}^{(t)}), \bar{\boldsymbol{v}}^{(t)}\rangle$.

$$\begin{aligned}\langle \nabla F(\bar{\boldsymbol{x}}^{(t)}), \bar{\boldsymbol{v}}^{(t)}\rangle &= \langle \bar{\boldsymbol{d}}^{(t)}, \bar{\boldsymbol{v}}^{(t)}\rangle + \langle \nabla F(\bar{\boldsymbol{x}}^{(t)}) - \bar{\boldsymbol{d}}^{(t)}, \bar{\boldsymbol{v}}^{(t)}\rangle \\ &= \frac{1}{n}\sum_{i=1}^n \langle \bar{\boldsymbol{d}}^{(t)} - \boldsymbol{d}_i^{(t)}, \boldsymbol{v}_i^{(t)}\rangle + \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{d}_i^{(t)}, \boldsymbol{v}_i^{(t)}\rangle + \langle \nabla F(\bar{\boldsymbol{x}}^{(t)}) - \bar{\boldsymbol{d}}^{(t)}, \bar{\boldsymbol{v}}^{(t)}\rangle \\ &\geq \frac{1}{n}\sum_{i=1}^n \langle \bar{\boldsymbol{d}}^{(t)} - \boldsymbol{d}_i^{(t)}, \boldsymbol{v}_i^{(t)}\rangle + \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{d}_i^{(t)}, \boldsymbol{x}^*\rangle + \langle \nabla F(\bar{\boldsymbol{x}}^{(t)}) - \bar{\boldsymbol{d}}^{(t)}, \bar{\boldsymbol{v}}^{(t)}\rangle,\end{aligned} \tag{31}$$

where the inequality holds since $\boldsymbol{v}_i^{(t)} = \operatorname{argmax}_{\boldsymbol{v} \in \mathcal{C}} \langle \boldsymbol{d}_i^{(t)}, \boldsymbol{v}_i^{(t)}\rangle$. Add and subtract $\langle \bar{\boldsymbol{d}}^{(t)}, \boldsymbol{x}^*\rangle$, we have

$$\begin{aligned}\langle \nabla F(\bar{\boldsymbol{x}}^{(t)}), \bar{\boldsymbol{v}}^{(t)}\rangle &\geq \frac{1}{n}\sum_{i=1}^n \langle \bar{\boldsymbol{d}}^{(t)} - \boldsymbol{d}_i^{(t)}, \boldsymbol{v}_i^{(t)} - \boldsymbol{x}^*\rangle + \langle \bar{\boldsymbol{d}}^{(t)}, \boldsymbol{x}^*\rangle + \langle \nabla F(\bar{\boldsymbol{x}}^{(t)}) - \bar{\boldsymbol{d}}^{(t)}, \bar{\boldsymbol{v}}^{(t)}\rangle \\ &= \frac{1}{n}\sum_{i=1}^n \langle \bar{\boldsymbol{d}}^{(t)} - \boldsymbol{d}_i^{(t)}, \boldsymbol{v}_i^{(t)} - \boldsymbol{x}^*\rangle + \langle \nabla F(\bar{\boldsymbol{x}}^{(t)}) - \bar{\boldsymbol{d}}^{(t)}, \bar{\boldsymbol{v}}^{(t)} - \boldsymbol{x}^*\rangle + \langle \nabla F(\bar{\boldsymbol{x}}^{(t)}), \boldsymbol{x}^*\rangle \\ &\geq -\frac{D}{n}\sum_{i=1}^n \|\bar{\boldsymbol{d}}^{(t)} - \boldsymbol{d}_i^{(t)}\| - D\|\nabla F(\bar{\boldsymbol{x}}^{(t)}) - \bar{\boldsymbol{d}}^{(t)}\| + \langle \nabla F(\bar{\boldsymbol{x}}^{(t)}), \boldsymbol{x}^*\rangle,\end{aligned} \tag{32}$$

where we add and subtract $\langle \nabla F(\bar{\boldsymbol{x}}^{(t)}), \boldsymbol{x}^*\rangle$ in the first equality, and we use Cauchy-Schwarz in the last inequality. Since F is monotone and continuous DR-submodular, one can show that $\langle \nabla F(\bar{\boldsymbol{x}}^{(t)}), \boldsymbol{x}^*\rangle \geq F(\boldsymbol{x}^*) - F(\bar{\boldsymbol{x}}^{(t)})$ as follows. For any $\boldsymbol{x} \in \mathcal{X}$, define $\boldsymbol{y} = (\boldsymbol{x}^* - \boldsymbol{x}) \vee \mathbf{0}_d$, then

$$\begin{aligned}\langle \nabla F(\boldsymbol{x}), \boldsymbol{x}^*\rangle &\geq \langle \nabla F(\boldsymbol{x}), \boldsymbol{y}\rangle \geq F(\boldsymbol{x} + \boldsymbol{y}) - F(\boldsymbol{x}) \\ &= F(\boldsymbol{x} \vee \boldsymbol{x}^*) - F(\boldsymbol{x}) \geq F(\boldsymbol{x}^*) - F(\boldsymbol{x}),\end{aligned} \tag{33}$$

where the first inequality follows from the monotonicity ($\nabla F(\boldsymbol{x}) \geq 0$) and the fact that $\boldsymbol{y} \leq \boldsymbol{x}^*$; the second inequality follows from the concavity of F along any non-negative direction (see, e.g., (Bian et al., 2017, Propositon

4)); the last inequality follows from the monotonicity of F. Combining (30), (32), and (33) yields

$$F(\bar{\boldsymbol{x}}^{(t+1)}) - F(\bar{\boldsymbol{x}}^{(t)}) \geq \frac{1}{T}\big(F(\boldsymbol{x}^*) - F(\bar{\boldsymbol{x}}^{(t)})\big) - \frac{D}{nT}\sum_{i=1}^{n} \|\bar{\boldsymbol{d}}^{(t)} - \boldsymbol{d}_i^{(t)}\| - \frac{D}{T}\|\nabla F(\bar{\boldsymbol{x}}^{(t)}) - \bar{\boldsymbol{d}}^{(t)}\| - \frac{LR^2}{2T^2}. \tag{34}$$

After rearranging terms of the above inequality, we arrive at (13). □

A.4 Proof of Lemma 4

Proof. First, we show that $\bar{\boldsymbol{d}}^{(t)} = \bar{\boldsymbol{g}}^{(t)} = \frac{1}{n}\sum_{i=1}^{n} \nabla F_i(\boldsymbol{x}_i^{(t)})$. Since $\boldsymbol{g}_i^{(1)} = \nabla F_i(\boldsymbol{x}_i^{(1)})$, $\bar{\boldsymbol{d}}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \nabla F_i(\boldsymbol{x}_i^{(1)})$ immediately holds. For $t \geq 2$,

$$\begin{aligned}
\bar{\boldsymbol{d}}^{(t)} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{d}_i^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{W}_{ij}\boldsymbol{g}_j^{(t)} = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n} \boldsymbol{W}_{ij}\boldsymbol{g}_j^{(t)} \overset{(a)}{=} \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_j^{(t)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \big(\boldsymbol{d}_i^{(t-1)} + \nabla F_i(\boldsymbol{x}_i^{(t)}) - \nabla F_i(\boldsymbol{x}_i^{(t-1)})\big) \\
&= \bar{\boldsymbol{d}}^{(t-1)} + \frac{1}{n}\sum_{i=1}^{n} \big(\nabla F_i(\boldsymbol{x}_i^{(t)}) - \nabla F_i(\boldsymbol{x}_i^{(t-1)})\big) \\
&= \bar{\boldsymbol{d}}^{(1)} + \frac{1}{n}\sum_{i=1}^{n}\sum_{\tau=2}^{t} \big(\nabla F_i(\boldsymbol{x}_i^{\tau}) - \nabla F_i(\boldsymbol{x}_i^{\tau-1})\big) = \frac{1}{n}\sum_{i=1}^{n} \nabla F_i(\boldsymbol{x}_i^{(t)}),
\end{aligned} \tag{35}$$

where (a) follows from Assumption 1.

We proceed to prove the second part of Lemma 4. Notice that $(\boldsymbol{W} \otimes \boldsymbol{I}_d)\boldsymbol{g}_{cat}^{(t)} = \big[(\sum_{j=1}^{n} \boldsymbol{W}_{1j}\boldsymbol{g}_j^{(t)})^\top, \ldots, (\sum_{j=1}^{n} \boldsymbol{W}_{nj}\boldsymbol{g}_j)^\top\big]^\top = \boldsymbol{d}_{cat}^{(t)}$ and $\bar{\boldsymbol{d}}_{cat}^{(t)} = \bar{\boldsymbol{g}}_{cat}^{(t)} = (\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes \boldsymbol{I}_d)\boldsymbol{g}_{cat}^{(t)}$, then it is easy to verify that

$$\begin{aligned}
\sum_{i=1}^{n} \|\boldsymbol{d}_i^{(t)} - \bar{\boldsymbol{d}}^{(t)}\|^2 = \|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\|^2 &= \|\big((\boldsymbol{W} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top) \otimes \boldsymbol{I}_d\big)\boldsymbol{g}_{cat}^{(t)}\|^2 \\
&= \|\big((\boldsymbol{W} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top) \otimes \boldsymbol{I}_d\big)\big(\boldsymbol{g}_{cat}^{(t)} - \bar{\boldsymbol{g}}_{cat}^{(t)}\big)\|^2 \\
&\leq \|\boldsymbol{W} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\|^2 \|\boldsymbol{g}_{cat}^{(t)} - \bar{\boldsymbol{g}}_{cat}^{(t)}\|^2 \\
&\leq \beta^2\|\boldsymbol{g}_{cat}^{(t)} - \bar{\boldsymbol{g}}_{cat}^{(t)}\|^2 = \beta^2\sum_{i=1}^{n} \|\boldsymbol{g}_i^{(t)} - \bar{\boldsymbol{g}}^{(t)}\|^2,
\end{aligned} \tag{36}$$

where the second equality holds because $(\boldsymbol{W} \otimes \boldsymbol{I}_d)\bar{\boldsymbol{g}}_{cat}^{(t)} = (\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes \boldsymbol{I}_d)\bar{\boldsymbol{g}}_{cat}^{(t)} = \bar{\boldsymbol{g}}_{cat}^{(t)}$. □

A.5 Proof of Lemma 5

Proof. First, we notice that

$$\big(\sum_{i=1}^{n} \|\boldsymbol{d}_i^{(t)} - \bar{\boldsymbol{d}}^{(t)}\|\big)^2 \leq n\big(\sum_{i=1}^{n} \|\boldsymbol{d}_i^{(t)} - \bar{\boldsymbol{d}}^{(t)}\|^2\big) = n\|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\|^2, \tag{37}$$

where the first inequality follows from the Cauchy-Schwarz inequality. It suffices to show that $\|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\| \leq \beta^t\sqrt{n}G + \frac{3\sqrt{n}\beta LR}{(1-\beta)^2 T}$ for $t \leq T$.

Applying Lemma 4, we obtain

$$\|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\| \leq \beta\|\boldsymbol{g}_{cat}^{(t)} - \bar{\boldsymbol{g}}_{cat}^{(t)}\| = \beta\|\boldsymbol{g}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\|. \tag{38}$$

For $t = 1$, since $\boldsymbol{g}_i^{(1)} = \nabla F_i(\boldsymbol{x}_i^{(1)})$ for all $i \in [n]$, $\|\boldsymbol{d}_{cat}^{(1)} - \bar{\boldsymbol{d}}_{cat}^{(1)}\|$ can be simply bounded as follows.

$$\begin{aligned}\|\boldsymbol{d}_{cat}^{(1)} - \bar{\boldsymbol{d}}_{cat}^{(1)}\|^2 &\le \beta^2 \sum_{i=1}^{n} \|\boldsymbol{g}_i^{(1)} - \bar{\boldsymbol{g}}^{(1)}\|^2 = \beta^2 \sum_{i=1}^{n} \|\nabla F_i(\boldsymbol{x}_i^{(1)}) - \frac{1}{n}\sum_{j=1}^{n} \nabla F_j(\boldsymbol{x}_j^{(1)})\|^2 \\ &\le \beta^2 \sum_{i=1}^{n} \|\nabla F_i(\boldsymbol{x}_i^{(1)})\|^2 \le n\beta^2 G^2, \end{aligned} \tag{39}$$

where the last inequality holds because of Assumption 3. Therefore, $\|\boldsymbol{d}_{cat}^{(1)} - \bar{\boldsymbol{d}}_{cat}^{(1)}\| \le \sqrt{n}\beta G$, which implies that the claim (14) holds for $t = 1$. Next, we show that the claim holds for $1 < t \le T$. We define $\boldsymbol{\delta}_i^{(t)} := \nabla F_i(\boldsymbol{x}_i^{(t)}) - \nabla F_i(\boldsymbol{x}_i^{(t-1)})$, $\bar{\boldsymbol{\delta}}^{(t)} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\delta}_i^{(t)}$. Then we have $\boldsymbol{g}_{cat}^{(t)} = \boldsymbol{d}_{cat}^{(t-1)} + \boldsymbol{\delta}_{cat}^{(t)}$ and $\bar{\boldsymbol{d}}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t-1)} = \bar{\boldsymbol{\delta}}_{cat}^{(t)}$. Applying (38), we have

$$\begin{aligned}\|\boldsymbol{d}_{cat}^{(t+1)} - \bar{\boldsymbol{d}}_{cat}^{(t+1)}\| &\le \beta\|\boldsymbol{g}_{cat}^{(t+1)} - \bar{\boldsymbol{d}}_{cat}^{(t+1)}\| = \beta\|(\boldsymbol{d}_{cat}^{(t)} + \boldsymbol{\delta}_{cat}^{(t+1)}) - (\bar{\boldsymbol{d}}_{cat}^{(t)} + \bar{\boldsymbol{\delta}}_{cat}^{(t+1)})\| \\ &\le \beta\big(\|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\| + \|\boldsymbol{\delta}_{cat}^{(t+1)} - \bar{\boldsymbol{\delta}}_{cat}^{(t+1)}\|\big) \le \beta\big(\|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\| + \|\boldsymbol{\delta}_{cat}^{(t+1)}\|\big), \end{aligned} \tag{40}$$

where the last inequality holds because $\|\boldsymbol{\delta}_{cat}^{(t+1)} - \bar{\boldsymbol{\delta}}_{cat}^{(t+1)}\|^2 = \sum_{i=1}^{n} \|\boldsymbol{\delta}_i^{(t+1)} - \bar{\boldsymbol{\delta}}^{(t)}\|^2 = \sum_{i=1}^{n} \|\boldsymbol{\delta}_i^{(t+1)}\|^2 - n\|\bar{\boldsymbol{\delta}}^{(t)}\|^2 \le \sum_{i=1}^{n} \|\boldsymbol{\delta}_i^{(t+1)}\|^2 = \|\boldsymbol{\delta}_{cat}^{(t+1)}\|^2$.

Notice that for any $t \le T$, $\|\boldsymbol{\delta}_{cat}^{(t+1)}\|$ can be bounded as follows.

$$\begin{aligned}\|\boldsymbol{\delta}_{cat}^{(t+1)}\| &= \|\nabla F_{cat}(\boldsymbol{x}_{cat}^{(t+1)}) - \nabla F_{cat}(\boldsymbol{x}_{cat}^{(t)})\| \overset{(a)}{\le} L\|\boldsymbol{x}_{cat}^{(t+1)} - \boldsymbol{x}_{cat}^{(t)}\| \\ &= L\|(\boldsymbol{W} \otimes \boldsymbol{I}_d)\boldsymbol{x}_{cat}^{(t)} + \frac{1}{T}\boldsymbol{v}_{cat}^{(t)} - \boldsymbol{x}_{cat}^{(t)}\| \\ &= L\|(\boldsymbol{W} \otimes \boldsymbol{I}_d)(\boldsymbol{x}_{cat}^{(t)} - \bar{\boldsymbol{x}}_{cat}^{(t)}) + \frac{1}{T}\boldsymbol{v}_{cat}^{(t)} - (\boldsymbol{x}_{cat}^{(t)} - \bar{\boldsymbol{x}}_{cat}^{(t)})\| \\ &\le L(\|\boldsymbol{W}\| + 1)\|\boldsymbol{x}_{cat}^{(t)} - \bar{\boldsymbol{x}}_{cat}^{(t)}\| + \frac{L}{T}\|\boldsymbol{v}_{cat}^{(t)}\| \\ &\overset{(b)}{\le} L(2\|\boldsymbol{x}_{cat}^{(t)} - \bar{\boldsymbol{x}}_{cat}^{(t)}\| + \frac{\sqrt{n}R}{T}) \\ &\overset{(c)}{\le} L(\frac{2\sqrt{n}R}{T(1-\beta)} + \frac{\sqrt{n}R}{T}) \le \frac{3\sqrt{n}LR}{T(1-\beta)} \end{aligned} \tag{41}$$

where (a) follows from the smoothness of F_i; (b) holds because $\|\boldsymbol{W}\| = 1$ and the fact that $\boldsymbol{v}_i^{(t)} \in \mathcal{C}$; (c) follows from Lemma 2.

Let $\zeta_t = \|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\|$ and $\Gamma = \frac{3\sqrt{n}LR}{T(1-\beta)}$. Combining (40) and (41), we have

$$\zeta_{t+1} \le \beta(\zeta_t + \Gamma),\ t \in \{1, \dots, T\} \tag{42}$$

Applying (42) recursively, we get

$$\zeta_{t+1} \le \beta^t \zeta_1 + \frac{\beta}{1-\beta}\Gamma,\ t \in \{1, \dots, T\} \tag{43}$$

where $\zeta_1 = \|\boldsymbol{d}_{cat}^{(1)} - \bar{\boldsymbol{d}}_{cat}^{(1)}\| \le \sqrt{n}\beta G$. Thus, we have $\|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\| \le \beta^t\sqrt{n}G + \frac{3\sqrt{n}\beta LR}{(1-\beta)^2 T}$ for any $t \in \{1, \dots, T\}$, which is the desired result. □

A.6 Proof of Theorem 1

Proof. Since Lemma 3 holds under Theorem 1's conditions, we proceed to bound $\big(\frac{D}{T}\|\bar{\boldsymbol{d}}^{(t)} - \nabla F(\bar{\boldsymbol{x}}^{(t)})\| + \frac{D}{nT}\sum_{i=1}^{n} \|\boldsymbol{d}_i^{(t)} - \bar{\boldsymbol{d}}^{(t)}\|\big)$ using Lemma 2 and Lemma 5. Recall that $\bar{\boldsymbol{d}}^{(t)} = \frac{1}{n}\sum_{i=1}^{n} \nabla F_i(\boldsymbol{x}_i^{(t)})$, then

$$\|\bar{\boldsymbol{d}}^{(t)} - \nabla F(\bar{\boldsymbol{x}}^{(t)})\| \le \frac{1}{n}\sum_{i=1}^{n} \|\nabla F_i(\boldsymbol{x}_i^{(t)}) - \nabla F_i(\bar{\boldsymbol{x}}^{(t)})\| \le \frac{1}{n}\sum_{i=1}^{n} L\|\boldsymbol{x}_i^{(t)} - \bar{\boldsymbol{x}}^{(t)}\| \le \frac{LR}{T(1-\beta)}, \tag{44}$$

where the second inequality holds because of Lemma 1 and the fact that $\bar{\boldsymbol{x}}^{(t)}$ is a convex combination of $\boldsymbol{x}_i^t$; the last inequality follows from Lemma 2 and the Cauchy-Schwarz inequality. Combining (44), Lemma 3, and Lemma 5, we have

$$F(\boldsymbol{x}^*) - F(\bar{\boldsymbol{x}}^{(t+1)}) \le (1 - \frac{1}{T})(F(\boldsymbol{x}^*) - F(\bar{\boldsymbol{x}}^{(t)})) + \frac{LRD}{T^2(1-\beta)} + \frac{\beta^t GD}{T} + \frac{3\beta LRD}{(1-\beta)^2 T^2} + \frac{LR^2}{2T^2} \tag{45}$$

With the above recursion, we obtain

$$\begin{aligned} F(\boldsymbol{x}^*) - F(\bar{\boldsymbol{x}}^{(T+1)}) &\le (1 - \frac{1}{T})^T (F(\boldsymbol{x}^*) - F(\bar{\boldsymbol{x}}^{(1)})) + \sum_{t=1}^{T} \Big(\frac{LRD}{T^2(1-\beta)} + \frac{\beta^t GD}{T} + \frac{3\beta LRD}{(1-\beta)^2 T^2} + \frac{LR^2}{2T^2} \Big) \\ &\le \frac{1}{e}(F(\boldsymbol{x}^*) - F(\bar{\boldsymbol{x}}^{(1)})) + \frac{LRD + \beta GD}{T(1-\beta)} + \frac{3\beta LRD}{T(1-\beta)^2} + \frac{LR^2}{2T} \end{aligned} \tag{46}$$

Rearrange terms and recall that $\boldsymbol{x}_i^{(1)} = \mathbf{0}_d$ for $i \in [n]$ and $F(\mathbf{0}_d) \ge 0$, we arrive at

$$\begin{aligned} F(\bar{\boldsymbol{x}}^{(T+1)}) &\ge (1 - \frac{1}{e})F(\boldsymbol{x}^*) + \frac{1}{e}F(\mathbf{0}_d) - \frac{LRD + \beta GD}{T(1-\beta)} - \frac{3\beta LRD}{T(1-\beta)^2} - \frac{LR^2}{2T} \\ &\ge (1 - \frac{1}{e})F(\boldsymbol{x}^*) - \frac{LRD + \beta GD}{T(1-\beta)} - \frac{3\beta LRD}{T(1-\beta)^2} - \frac{LR^2}{2T}. \end{aligned} \tag{47}$$

Combining (47) and (12), we obtain

$$F(\boldsymbol{x}_i^{(T+1)}) \ge (1 - \frac{1}{e})F(\boldsymbol{x}^*) - \frac{LRD + \beta GD}{T(1-\beta)} - \frac{3\beta LRD}{T(1-\beta)^2} - \frac{LR^2}{2T} - \frac{\sqrt{n}GR}{T(1-\beta)}. \tag{48}$$

To ensure that $F(\bar{\boldsymbol{x}}^{(T+1)})$ or each $F(\boldsymbol{x}_i^{(T+1)})$ is greater than $(1 - 1/e)F(\boldsymbol{x}^*) - \epsilon$, the number of iterations T should be $T = \mathcal{O}(\frac{1}{\epsilon})$. Since at each iteration, DeGTFW requires one communication round and one full local gradient evaluation at each node, we conclude that both the communication and gradient evaluation complexities of DeGTFW are $\mathcal{O}(\frac{1}{\epsilon})$. □

A.7 Proof of Lemma 6

Proof. Using the same argument as (35), one can show that $\bar{\boldsymbol{d}}^{(t)} = \bar{\boldsymbol{g}}^{(t)} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\nabla}_i^{(t)}$. On the other hand, since the gradient estimate $\tilde{\nabla}_i^{(t)}$ is the average of t^2 samples of $\tilde{\nabla} F_i(\boldsymbol{x}_i^{(t)})$, then

$$\mathbb{E}\Big[\big\|\tilde{\nabla}_i^{(t)} - \nabla F_i(\boldsymbol{x}_i^{(t)})\big\|^2 \Big| \boldsymbol{x}_i^{(t)}\Big] \le \frac{\sigma^2}{t^2} \tag{49}$$

Therefore, we have

$$\begin{aligned} \|\bar{\boldsymbol{d}}^{(t)} - \nabla F(\bar{\boldsymbol{x}}^{(t)})\| &= \|\frac{1}{n}\sum_{i=1}^{n} \big(\tilde{\nabla}_i^{(t)} - \nabla F_i(\bar{\boldsymbol{x}}^{(t)})\big)\| \le \frac{1}{n}\sum_{i=1}^{n} \|\tilde{\nabla}_i^{(t)} - \nabla F_i(\bar{\boldsymbol{x}}^{(t)})\| \\ &\le \frac{1}{n}\sum_{i=1}^{n} \big(\|\tilde{\nabla}_i^{(t)} - \nabla F_i(\boldsymbol{x}_i^{(t)})\| + \|\nabla F_i(\boldsymbol{x}_i^{(t)}) - \nabla F_i(\bar{\boldsymbol{x}}^{(t)})\|\big) \\ &\le \frac{1}{n}\sum_{i=1}^{n} \|\tilde{\nabla}_i^{(t)} - \nabla F_i(\boldsymbol{x}_i^{(t)})\| + \frac{LR}{T(1-\beta)} \end{aligned} \tag{50}$$

where the last inequality follows from (44). Taking expectation on both sides yields

$$\mathbb{E}\big[\|\bar{\boldsymbol{d}}^{(t)} - \nabla F(\bar{\boldsymbol{x}}^{(t)})\|\big] \le \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\Big[\mathbb{E}\Big[\big\|\tilde{\nabla}_i^{(t)} - \nabla F_i(\boldsymbol{x}_i^{(t)})\big\| \;\Big|\; \boldsymbol{x}_i^{(t)}\Big]\Big] + \frac{LR}{T(1-\beta)} \quad \le \frac{\sigma}{t} + \frac{LR}{T(1-\beta)}, \tag{51}$$

where the last inequality follows from (49) and Jensen's inequality. □

A.8 Proof of Lemma 7

Proof. We prove the lemma by induction, which uses ideas in (Wai et al., 2017, Section D). Recall from (37) that $\sum_{i=1}^n \|\boldsymbol{d}_i^{(t)} - \bar{\boldsymbol{d}}^{(t)}\| \leq \sqrt{n}\|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\|$. It suffices to prove the claim $\mathbb{E}[\|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\|] \leq \sqrt{n}\tilde{M}/t$ for $t \leq T$.

To begin with, using essentially the same argument as (38), one can show that

$$\|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\|^2 \leq \beta^2 \|\boldsymbol{g}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\|^2. \tag{52}$$

Then for $t = 1$, we have

$$\begin{aligned}\|\boldsymbol{d}_{cat}^{(1)} - \bar{\boldsymbol{d}}_{cat}^{(1)}\|^2 &\leq \beta^2 \sum_{i=1}^n \|\boldsymbol{g}_i^{(1)} - \bar{\boldsymbol{g}}^{(1)}\|^2 = \beta^2 \sum_{i=1}^n \|\tilde{\nabla}_i^{(1)} - \frac{1}{n}\sum_{i=1}^n \nabla F_i(\boldsymbol{x}_i^{(1)})\|^2 \\ &= \beta^2 \sum_{i=1}^n \Big(\|\tilde{\nabla}_i^{(1)} - \nabla F_i(\boldsymbol{x}_i^{(1)})\|^2 + \|\nabla F_i(\boldsymbol{x}_i^{(1)}) - \frac{1}{n}\sum_{j=1}^n \nabla F_j(\boldsymbol{x}_j^{(1)})\|^2 \\ &\quad - 2\langle \tilde{\nabla}_i^{(1)} - \nabla F_i(\boldsymbol{x}_i^{(1)}), \nabla F_i(\boldsymbol{x}_i^{(1)}) - \frac{1}{n}\sum_{i=1}^n \nabla F_i(\boldsymbol{x}_i^{(1)})\rangle\Big).\end{aligned} \tag{53}$$

Taking expectation on both sides, we get

$$\mathbb{E}[\|\boldsymbol{d}_{cat}^{(1)} - \bar{\boldsymbol{d}}_{cat}^{(1)}\|^2] \leq \beta^2 \Big(\sum_{i=1}^n \mathbb{E}\big[\|\tilde{\nabla}_i^{(1)} - \nabla F_i(\boldsymbol{x}_i^{(1)})\|^2\big] + nG^2\Big) \leq n\beta^2(\sigma^2 + G^2), \tag{54}$$

where the first inequality follows from the same argument as (39) and the unbiasedness of $\tilde{\nabla}_i^{(t)}$. Thus, $\mathbb{E}[\|\boldsymbol{d}_{cat}^{(1)} - \bar{\boldsymbol{d}}_{cat}^{(1)}\|] \leq (\mathbb{E}[\|\boldsymbol{d}_{cat}^{(1)} - \bar{\boldsymbol{d}}_{cat}^{(1)}\|^2])^{1/2} \leq \beta\sqrt{n(\sigma^2 + G^2)} \leq \sqrt{n}\tilde{M}$, which implies that (18) holds for $t = 1$.

In the remaining of the paper, we define $\boldsymbol{\delta}_i^{(t)} := \tilde{\nabla}_i^{(t)} - \tilde{\nabla}_i^{(t-1)}$ and $\bar{\boldsymbol{\delta}}^{(t)} := \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_i^{(t)}$. Recall that $\bar{\boldsymbol{d}}^{(t)} = \frac{1}{n}\sum_{i=1}^n \tilde{\nabla}_i^{(t)}$, then we have $\bar{\boldsymbol{d}}_{cat}^{(t)} = (\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes \boldsymbol{I}_d)\tilde{\nabla}_{cat}^{(t)}$ and $\bar{\boldsymbol{\delta}}_{cat}^{(t+1)} = \bar{\boldsymbol{d}}_{cat}^{(t+1)} - \bar{\boldsymbol{d}}_{cat}^{(t)}$. Note that $(1/\sqrt{n})\sum_{i=1}^n \|\boldsymbol{d}_i^{(t)} - \bar{\boldsymbol{d}}^{(t)}\| \leq (\sum_{i=1}^n \|\boldsymbol{d}_i^{(t)} - \bar{\boldsymbol{d}}^{(t)}\|^2)^{1/2} = \|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\|$. It suffices to prove that $\mathbb{E}[\|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\|] \leq \sqrt{n}\tilde{M}/t$ for $t \leq T$.

First, we show that the claim (18) holds for $t \leq t_0$, where t_0 is defined in Lemma 7. Using the same argument as (40), one can easily show that

$$\|\boldsymbol{d}_{cat}^{(t+1)} - \bar{\boldsymbol{d}}_{cat}^{(t+1)}\| \leq \beta(\|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\| + \|\boldsymbol{\delta}_{cat}^{(t+1)}\|). \tag{55}$$

Notice that $\|\nabla F_{cat}(\boldsymbol{x}_{cat}^{(t+1)}) - \nabla F_{cat}(\boldsymbol{x}_{cat}^{(t)})\|$ can be simply bounded as follows.

$$\begin{aligned}\|\nabla F_{cat}(\boldsymbol{x}_{cat}^{(t+1)}) - \nabla F_{cat}(\boldsymbol{x}_{cat}^{(t)})\| &= \sqrt{\sum_{i=1}^n \|\nabla F_i(\boldsymbol{x}_i^{(t+1)}) - \nabla F_i(\boldsymbol{x}_i^{(t)})\|^2} \leq \sqrt{\sum_{i=1}^n L^2\|\boldsymbol{x}_i^{(t+1)} - \boldsymbol{x}_i^{(t)}\|^2} \\ &\leq \sqrt{\sum_{i=1}^n 2L^2(\|\boldsymbol{x}_i^{(t+1)}\|^2 + \|\boldsymbol{x}_i^{(t)}\|^2)} \leq 2\sqrt{n}LR,\end{aligned} \tag{56}$$

where the last inequality follows from Lemma 1 and Assumption 2. Then the expectation of $\|\boldsymbol{\delta}_{cat}^{(t+1)}\|$ can be bounded as follows:

$$\begin{aligned}&\mathbb{E}\big[\|\boldsymbol{\delta}_{cat}^{(t+1)}\|\big] = \mathbb{E}\Big[\|\tilde{\nabla}_{cat}^{(t+1)} - \tilde{\nabla}_{cat}^{(t)}\|\Big] \\ &\leq \mathbb{E}\Big[\|\tilde{\nabla}_{cat}^{(t+1)} - \nabla F_{cat}(\boldsymbol{x}_{cat}^{(t+1)})\| + \|\tilde{\nabla}_{cat}^{(t)} - \nabla F_{cat}(\boldsymbol{x}_{cat}^{(t)})\| + \|\nabla F_{cat}(\boldsymbol{x}_{cat}^{(t+1)}) - \nabla F_{cat}(\boldsymbol{x}_{cat}^{(t)})\|\Big] \\ &\leq \Big(\mathbb{E}[\|\tilde{\nabla}_{cat}^{(t+1)} - \nabla F_{cat}(\boldsymbol{x}_{cat}^{(t+1)})\|^2]\Big)^{\frac{1}{2}} + \Big(\mathbb{E}[\|\tilde{\nabla}_{cat}^{(t)} - \nabla F_{cat}(\boldsymbol{x}_{cat}^{(t)})\|^2]\Big)^{\frac{1}{2}} + 2\sqrt{n}LR \\ &\leq \sqrt{n}(\frac{\sigma}{t+1} + \frac{\sigma}{t} + 2LR) \leq 2\sqrt{n}(\sigma + LR).\end{aligned} \tag{57}$$

where the second inequality follows from Jensen's inequality and (56); the third inequality holds because

$$\mathbb{E}\Big[\Big\|\tilde{\nabla}_{cat}^{(t)} - \nabla F_{cat}(\boldsymbol{x}_{cat}^{(t)})\Big\|^2\Big] = \sum_{i=1}^{n} \mathbb{E}\Big[\mathbb{E}\Big[\Big\|\tilde{\nabla}_i^{(t)} - \nabla F_i(\boldsymbol{x}_i^{(t)})\Big\|^2 \Big| \boldsymbol{x}_i^{(t)}\Big]\Big] = \frac{n\sigma^2}{t^2}. \tag{58}$$

For $t_0 = 1$, (54) already holds for $t = 1$. Now suppose that $t + 1 \le t_0$, then taking expectation on both sides of (55) and applying (54) and (57) yields

$$\begin{aligned}\mathbb{E}\big[\|\boldsymbol{d}_{cat}^{(t+1)} - \bar{\boldsymbol{d}}_{cat}^{(t+1)}\|\big] &\le \beta\mathbb{E}\big[\|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\| + \|\boldsymbol{\delta}_{cat}^{(t+1)}\|\big] \\ &\le \beta^t\mathbb{E}\big[\|\boldsymbol{d}_{cat}^{(1)} - \bar{\boldsymbol{d}}_{cat}^{(1)}\|\big] + \sum_{\tau=1}^{t} 2\beta^\tau\sqrt{n}(\sigma + LR) \\ &\le \beta\sqrt{n(\sigma^2 + G^2)} + \frac{2\beta\sqrt{n}(\sigma + LR)}{1 - \beta} \le \sqrt{n}\tilde{M}/t_0 \le \sqrt{n}\tilde{M}/(t+1)\end{aligned} \tag{59}$$

where the third inequality holds because of Jensen's inequality ($\|\mathbb{E}[\boldsymbol{x}]\|^2 \le \mathbb{E}[\|\boldsymbol{x}\|^2]$). Therefore, $\mathbb{E}[\|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\|] \le \sqrt{n}\tilde{M}/t$ holds for $t \le t_0$.

If $t_0 \ge T$, then Lemma 7 immediately holds. Suppose that $t_0 < T$, and that $\mathbb{E}[\|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\|] \le \sqrt{n}\tilde{M}/t$ holds for some $t \in [t_0, T]$, we will show that the same claim holds for $t+1$. Note that $\|\boldsymbol{\delta}_{cat}^{(t+1)}\|$ can be bounded as follows:

$$\begin{aligned}\|\boldsymbol{\delta}_{cat}^{(t+1)}\| &= \|\tilde{\nabla}_{cat}^{(t+1)} - \tilde{\nabla}_{cat}^{(t)}\| \\ &= \|\tilde{\nabla}_{cat}^{(t+1)} - \nabla F_{cat}(\boldsymbol{x}_{cat}^{(t+1)}) - (\tilde{\nabla}_{cat}^{(t)} - \nabla F_{cat}(\boldsymbol{x}_{cat}^{(t)})) + \nabla F_{cat}(\boldsymbol{x}_{cat}^{(t+1)}) - \nabla F_{cat}(\boldsymbol{x}_{cat}^{(t)})\| \\ &\le \|\tilde{\nabla}_{cat}^{(t+1)} - \nabla F_{cat}(\boldsymbol{x}_{cat}^{(t+1)})\| + \|\tilde{\nabla}_{cat}^{(t)} - \nabla F_{cat}(\boldsymbol{x}_{cat}^{(t)})\| + \frac{3\sqrt{n}LR}{T(1-\beta)},\end{aligned} \tag{60}$$

where the inequality follows from the same argument as (41). Taking expectation on both sides, we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\delta}_{cat}^{(t+1)}\|] &\le \mathbb{E}[\|\tilde{\nabla}_{cat}^{(t+1)} - \nabla F_{cat}(\boldsymbol{x}_{cat}^{(t+1)})\|] + \mathbb{E}[\|\tilde{\nabla}_{cat}^{(t)} - \nabla F_{cat}(\boldsymbol{x}_{cat}^{(t)})\|] + \frac{3\sqrt{n}LR}{T(1-\beta)} \\ &\le \sqrt{n}(\frac{\sigma}{t+1} + \frac{\sigma}{t} + \frac{3LR}{T(1-\beta)}) \le \frac{\sqrt{n}}{t}(2\sigma + \frac{3LR}{1-\beta}),\end{aligned} \tag{61}$$

where the second inequality follows from Jensen's inequality and (58).

Taking expectation on both sides of (55) and applying (61) yields

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{d}_{cat}^{(t+1)} - \bar{\boldsymbol{d}}_{cat}^{(t+1)}\|] &\le \beta\big(\mathbb{E}[\|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\|] + \frac{\sqrt{n}}{t}(2\sigma + \frac{3LR}{1-\beta})\big) \\ &\le \frac{\beta}{t}\big(\sqrt{n}\tilde{M} + \sqrt{n}(2\sigma + \frac{3LR}{1-\beta})\big) \le \frac{\sqrt{n}\tilde{M}\beta(1 + 1/t_0)}{t} \\ &\le \sqrt{n}\tilde{M}\big(\frac{1 + 1/t_0}{t} \cdot \frac{1}{(1 + 1/t_0)^2}\big) \\ &\le \sqrt{n}\tilde{M}\big(\frac{1}{t} \cdot \frac{1}{1 + 1/t}\big) = \frac{\sqrt{n}\tilde{M}}{t+1}.\end{aligned} \tag{62}$$

Therefore, we have proved that $\mathbb{E}[\|\boldsymbol{d}_{cat}^{(t)} - \bar{\boldsymbol{d}}_{cat}^{(t)}\|] \le \sqrt{n}\tilde{M}/t$ for $t \in \{1, \dots, T+1\}$, which is the desired result. □

A.9 Proof of Theorem 2

Proof. Combining Lemma 3, Lemma 6 and Lemma 7 and taking expectation leads to

$$\mathbb{E}[F(\boldsymbol{x}^*) - F(\bar{\boldsymbol{x}}^{(t+1)})] \le (1 - \frac{1}{T})\mathbb{E}[F(\boldsymbol{x}^*) - F(\bar{\boldsymbol{x}}^{(t)})] + \frac{LR^2}{2T^2} + \frac{LRD}{T^2(1-\beta)} + \frac{\sigma D}{Tt} + \frac{\tilde{M}D}{Tt} \tag{63}$$

Applying the aboe inequality recursively yields

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{x}^*) - F(\bar{\boldsymbol{x}}^{(T+1)})] &\leq (1-\frac{1}{T})^T\mathbb{E}[F(\boldsymbol{x}^*) - F(\bar{\boldsymbol{x}}^{(1)})] + \sum_{t=1}^{T}\Big(\frac{LR^2}{2T^2} + \frac{LRD}{T^2(1-\beta)} + \frac{\sigma D + \tilde{M}D}{Tt}\Big) \\ &\leq \frac{1}{e}(F(\boldsymbol{x}^*) - F(\bar{\boldsymbol{x}}^{(1)})) + \frac{LR^2}{2T} + \frac{LRD}{T(1-\beta)} + \frac{(\sigma D + \tilde{M}D)(\log T + 1)}{T}. \end{aligned} \tag{64}$$

Finally, recall that $\boldsymbol{x}_i^{(1)} = \mathbf{0}_d$ for $i \in [n]$ and $F(\mathbf{0}_d) \geq 0$, we arrive at

$$\begin{aligned}\mathbb{E}[F(\bar{\boldsymbol{x}}^{(T+1)})] &\geq (1-\frac{1}{e})F(\boldsymbol{x}^*) + \frac{1}{e}F(\mathbf{0}_d) - \frac{LR^2}{2T} - \frac{LRD}{T(1-\beta)} - \frac{(\sigma D + \tilde{M}D)(\log T + 1)}{T} \\ &\geq (1-\frac{1}{e})F(\boldsymbol{x}^*) - \frac{LR^2}{2T} - \frac{LRD}{T(1-\beta)} - \frac{(\sigma D + \tilde{M}D)(\log T + 1)}{T}. \end{aligned} \tag{65}$$

Combining (65) and (12), we obtain

$$\mathbb{E}[F(\boldsymbol{x}_i^{(T+1)})] \geq (1-\frac{1}{e})F(\boldsymbol{x}^*) - \frac{LR^2}{2T} - \frac{LRD}{T(1-\beta)} - \frac{(\sigma D + \tilde{M}D)(\log T + 1)}{T} - \frac{\sqrt{n}GR}{T(1-\beta)}. \tag{66}$$

To ensure that $\mathbb{E}[F(\bar{\boldsymbol{x}}^{(T+1)})]$ or each $\mathbb{E}[F(\boldsymbol{x}_i^{(T+1)})]$ is greater than $(1-1/e)F(\boldsymbol{x}^*) - \epsilon$, the number of iterations T should be $T = \tilde{\mathcal{O}}(\frac{1}{\epsilon})$. Since at each iteration t, DeSGTFW requires one communication round and t^2 stochastic gradient evaluations, we conclude that the communication and stochastic gradient evaluation complexities of DeSGTFW are $\tilde{\mathcal{O}}(\frac{1}{\epsilon})$ and $\tilde{\mathcal{O}}(\frac{1}{\epsilon^3})$, respectively. □