9. Supplementary Material

9.1. Proof of Proposition 1

Define $\mathbf{x}_{con} = [\mathbf{x}_1; \dots; \mathbf{x}_n] \in \mathbb{R}^{np}$ and $\mathbf{v}_{con} = [\mathbf{v}_1; \dots; \mathbf{v}_n] \in \mathbb{R}^{np}$ as the concatenation of the local variables and descent directions, respectively. Using these definitions and the update in (8) we can write

$$\mathbf{x}_{con}^{t+1} = (\mathbf{W} \otimes \mathbf{I})\mathbf{x}_{con}^t + \frac{1}{T}\mathbf{v}_{con}^t, \tag{26}$$

where $\mathbf{W} \otimes \mathbf{I} \in \mathbb{R}^{np \times np}$ is the Kronecker product of the matrices $\mathbf{W} \in \mathbb{R}^{n \times n}$ and $\mathbf{I} \in \mathbb{R}^{p \times p}$. If we set $\mathbf{x}_i^0 = \mathbf{0}_p$ for all nodes i, it follows that $\mathbf{x}_{con}^0 = \mathbf{0}_{np}$. Hence, by applying the update in (26) recursively we obtain that the iterate $\mathbf{x}_{con}^t$ is equal to

$$\mathbf{x}_{con}^t = \frac{1}{T}\sum_{s=0}^{t-1}(\mathbf{W} \otimes \mathbf{I})^{t-1-s}\mathbf{v}_{con}^s. \tag{27}$$

We proceed by showing that if the local blocks of a vector $\mathbf{v}_{con} \in \mathbb{R}^{np}$ belong to the feasible set $\mathcal{C}$, i.e., $\mathbf{v}_i \in \mathcal{C}$ for $i = 1, \dots, n$, then the local vectors of $\mathbf{y}_{con} = (\mathbf{W} \otimes \mathbf{I})\mathbf{v}_{con} \in \mathbb{R}^{np}$ also in the set $\mathcal{C}$. Note that if the condition $\mathbf{y}_{con} = (\mathbf{W} \otimes \mathbf{I})\mathbf{v}_{con}$ holds, then the i-th block of $\mathbf{y}_{con} = [\mathbf{y}_1; \dots; \mathbf{y}_n]$ can be written as

$$\mathbf{y}_i = \sum_{j=1}^{n} w_{ij}\mathbf{v}_j. \tag{28}$$

Since we assume that all $\{\mathbf{v}_j\}_{j=1}^n$ belong to the set $\mathcal{C}$ and the set $\mathcal{C}$ is convex, the weighted average of these vectors also is in the set $\mathcal{C}$, i.e., $\mathbf{y}_i \in \mathcal{C}$. This argument indeed holds for all blocks $\mathbf{y}_i$ and therefore $\mathbf{y}_i \in \mathcal{C}$ for $i = 1, \dots, n$. This argument verifies that if we apply any power of the matrix $\mathbf{W} \otimes \mathbf{I}$ to a vector $\mathbf{v}_{con} \in \mathbb{R}^{np}$ whose blocks belong to the set $\mathcal{C}$, then the local components of the output vector also belong to the set $\mathcal{C}$. Therefore, the local components of each of the terms $(\mathbf{W} \otimes \mathbf{I})^{t-1-s}\mathbf{v}_{con}^s$ in (27) belong to the set $\mathcal{C}$. The fact that $\mathbf{x}_i$ which is the i-th block of the vector $\mathbf{x}_{con}^t$, is the average of T terms that are in the set $\mathcal{C}$ ($\mathbf{x}_{con}^t$ is the average of the vectors $(\mathbf{W} \otimes \mathbf{I})^{t-1}\mathbf{v}_{con}^0, \dots, (\mathbf{W} \otimes \mathbf{I})^0\mathbf{v}_{con}^{t-1}$ with weights $1/T$ and the vector $\mathbf{0}_{np}$ with weight $(T-t)/T$), implies that $\mathbf{x}_i^t \in \mathcal{C}$. This result holds for all $i \in \{1, \dots, n\}$ and the proof is complete.

9.2. Proof of Lemma 1

By averaging both sides of the update in (8) over the nodes in the network and using the fact $w_{ij} = 0$ if i and j are not neighbors we can write

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^{t+1} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j \in \mathcal{N}_i \cup \{i\}} w_{ij}\mathbf{x}_j^t + \frac{1}{T}\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i^t \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\mathbf{x}_j^t + \frac{1}{T}\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i^t \\
&= \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j^t\sum_{i=1}^{n} w_{ij} + \frac{1}{T}\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i^t \\
&= \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j^t + \frac{1}{T}\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i^t,
\end{aligned} \tag{29}$$

where the last equality holds since $\mathbf{W}^T\mathbf{1}_n = \mathbf{1}_n$ (i.e. $\mathbf{W}$ is a doubly stochastic matrix). By using the definition of the average iterate vector $\bar{\mathbf{x}}^t$ and the result in (29) it follows that

$$\bar{\mathbf{x}}^{t+1} = \bar{\mathbf{x}}^t + \frac{1}{T}\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i^t \tag{30}$$

Since $\mathbf{v}_i^t$ belongs to the convex set $\mathcal{C}$ its Euclidean norm is bounded by $\|\mathbf{v}_i^t\| \leq D$ according to Assumption 2. This inequality and the expression in (30) yield

$$\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\| \leq \frac{D}{T}, \tag{31}$$

and the claim in (17) follows.

9.3. Proof of Lemma 2

Recall the definitions $\mathbf{x}_{con} = [\mathbf{x}_1; \ldots; \mathbf{x}_n] \in \mathbb{R}^{np}$ and $\mathbf{v}_{con} = [\mathbf{v}_1; \ldots; \mathbf{v}_n] \in \mathbb{R}^{np}$ for the concatenation of the local variables and descent directions, respectively. These definitions along with the update in (8) lead to the expression

$$\mathbf{x}_{con}^t = \frac{1}{T} \sum_{s=0}^{t-1} (\mathbf{W} \otimes \mathbf{I})^{t-1-s} \mathbf{v}_{con}^s. \tag{32}$$

If we premultiply both sides of (32) by the matrix $(\frac{\mathbf{1}_n \mathbf{1}_n^\dagger}{n} \otimes \mathbf{I})$ which is the Kronecker product of the matrices $(1/n)(\mathbf{1}_n \mathbf{1}_n^\dagger) \in \mathbb{R}^{n \times n}$ and $\mathbf{I} \in \mathbb{R}^{p \times p}$ we obtain

$$\left(\frac{\mathbf{1}_n \mathbf{1}_n^\dagger}{n} \otimes \mathbf{I}\right) \mathbf{x}_{con}^t = \frac{1}{T} \sum_{s=0}^{t-1} \left(\left(\frac{\mathbf{1}_n \mathbf{1}_n^\dagger}{n} \mathbf{W}^{t-1-s} \right) \otimes \mathbf{I} \right) \mathbf{v}_{con}^s. \tag{33}$$

The left hand side of (33) can be simplified to

$$\left(\frac{\mathbf{1}_n \mathbf{1}_n^\dagger}{n} \otimes \mathbf{I}\right) \mathbf{x}_{con}^t = \bar{\mathbf{x}}_{con}^t, \tag{34}$$

where $\bar{\mathbf{x}}_{con}^t = [\bar{\mathbf{x}}^t; \ldots; \bar{\mathbf{x}}^t]$ is the concatenation of n copies of the average vector $\bar{\mathbf{x}}^t$. Using the equality in (34) and the simplification $\mathbf{1}_n \mathbf{1}_n^\dagger \mathbf{W} = \mathbf{1}_n \mathbf{1}_n^\dagger$, we can rewrite (33) as

$$\bar{\mathbf{x}}_{con}^t = \frac{1}{T} \sum_{s=0}^{t-1} \left(\frac{\mathbf{1}_n \mathbf{1}_n^\dagger}{n} \otimes \mathbf{I}\right) \mathbf{v}_{con}^s. \tag{35}$$

Using the expressions in (32) and (35) we can derive an upper bound on the difference $\|\mathbf{x}_{con}^t - \bar{\mathbf{x}}_{con}^t\|$ as

$$\begin{aligned}
\|\mathbf{x}_{con}^t - \bar{\mathbf{x}}_{con}^t\| &= \frac{1}{T} \left\| \sum_{s=0}^{t-1} \left(\left[\mathbf{W}^{t-1-s} - \frac{\mathbf{1}_n \mathbf{1}_n^\dagger}{n} \right] \otimes \mathbf{I} \right) \mathbf{v}_{con}^s \right\| \\
&\leq \frac{1}{T} \sum_{s=0}^{t-1} \left\| \mathbf{W}^{t-1-s} - \frac{\mathbf{1}_n \mathbf{1}_n^\dagger}{n} \right\| \|\mathbf{v}_{con}^s\| \\
&\leq \frac{\sqrt{n} D}{T} \sum_{s=0}^{t-1} \left\| \mathbf{W}^{t-1-s} - \frac{\mathbf{1}_n \mathbf{1}_n^\dagger}{n} \right\|,
\end{aligned} \tag{36}$$

where the first inequality follows from the Cauchy-Schwarz inequality and the fact that the norm of a matrix does not change if we Kronecker it by the identity matrix, the second inequality holds since $\|\mathbf{v}_i^t\| \leq D$ and therefore $\|\mathbf{v}_{con}^t\| \leq \sqrt{n} D$. Note that the eigenvectors of the matrices $\mathbf{W}$ and $\mathbf{W}^{t-s-1}$ are the same for all $s = 0, \ldots, t-1$. Therefore, the largest eigenvalue of $\mathbf{W}^{t-s-1}$ is 1 with eigenvector $\mathbf{1}_n$ and its second largest magnitude of the eigenvalues is β^{t-1-s}, where β is the second largest magnitude of the eigenvalues of $\mathbf{W}$. Also, note that since $\mathbf{W}^{t-1-s}$ has $\mathbf{1}_n$ as one of its eigenvectors, then all the other eigenvectors of $\mathbf{W}$ are orthogonal to $\mathbf{1}_n$. Hence, we can bound the norm $\|\mathbf{W}^{t-1-s} - (\mathbf{1}_n \mathbf{1}_n^\dagger)/(n)\|$ by β^{t-1-s}. Applying this substitution into the right hand side of (36) yields

$$\|\mathbf{x}_{con}^t - \bar{\mathbf{x}}_{con}^t\| \leq \frac{\sqrt{n} D}{T} \sum_{s=0}^{t-1} \beta^{t-1-s} \leq \frac{\sqrt{n} D}{T(1-\beta)}. \tag{37}$$

Since $\|\mathbf{x}_{con}^t - \bar{\mathbf{x}}_{con}^t\|^2 = \sum_{i=1}^n \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2$, the claim in (19) follows.

9.4. Proof of Lemma 3

Recall the definition of the vector $\mathbf{x}_{con} = [\mathbf{x}_1; \dots; \mathbf{x}_n] \in \mathbb{R}^{np}$ as the concatenation of the local variables, and define $\mathbf{d}_{con} = [\mathbf{d}_1; \dots; \mathbf{d}_n] \in \mathbb{R}^{np}$ as the concatenation of the local approximate gradients. Further, consider the function $F_{con} : \mathcal{X}^n \to \mathbb{R}$ which is defined as $F_{con}(\mathbf{x}_{con}) = F_{con}(\mathbf{x}_1, \dots, \mathbf{x}_n) := \sum_{i=1}^n F_i(\mathbf{x}_i)$. According to these definitions and the update in (6), we can show that

$$\mathbf{d}_{con}^t = (1-\alpha)(\mathbf{W} \otimes \mathbf{I})\mathbf{d}_{con}^{t-1} + \alpha \nabla F_{con}(\mathbf{x}_{con}^t), \tag{38}$$

where $\mathbf{W} \otimes \mathbf{I} \in \mathbb{R}^{np \times np}$ is the Kronecker product of the matrices $\mathbf{W} \in \mathbb{R}^{n \times n}$ and $\mathbf{I} \in \mathbb{R}^{p \times p}$. Considering the initialization $\mathbf{d}_{con}^0 = \mathbf{0}_p$, applying the update in (38) recursively from step 1 to t leads to

$$\mathbf{d}_{con}^t = \alpha \sum_{s=1}^{t} (((1-\alpha)\mathbf{W})^{t-s} \otimes \mathbf{I}) \nabla F_{con}(\mathbf{x}_{con}^s) \tag{39}$$

If we multiply both sides of (39) from left by the matrix $(\frac{\mathbf{1}_n \mathbf{1}_n^\dagger}{n} \otimes \mathbf{I}) \in \mathbb{R}^{np \times np}$ and use the properties of the weight matrix $\mathbf{W}$, i.e., $\mathbf{1}_n^\dagger \mathbf{W}^{t-s} = \mathbf{1}_n^\dagger$, we obtain that

$$\bar{\mathbf{d}}_{con}^t = \alpha \sum_{s=1}^{t} (1-\alpha)^{t-s} \left(\frac{\mathbf{1}_n \mathbf{1}_n^\dagger}{n} \otimes \mathbf{I} \right) \nabla F_{con}(\mathbf{x}_{con}^s), \tag{40}$$

where $\bar{\mathbf{d}}_{con}^t = [\bar{\mathbf{d}}^t; \dots; \bar{\mathbf{d}}^t]$ is the concatenation of n copies of the average vector $\bar{\mathbf{d}}^t$. Hence, the difference $\|\mathbf{d}_{con}^t - \bar{\mathbf{d}}_{con}^t\|$ can be upper bounded by

$$\begin{aligned}
\|\mathbf{d}_{con}^t - \bar{\mathbf{d}}_{con}^t\| &= \left\| \alpha \sum_{s=1}^{t} (1-\alpha)^{t-s} (\mathbf{W}^{t-s} \otimes \mathbf{I}) \nabla F_{con}(\mathbf{x}_{con}^s) - \alpha \sum_{s=1}^{t} (1-\alpha)^{t-s} \left(\frac{\mathbf{1}_n \mathbf{1}_n^\dagger}{n} \otimes \mathbf{I} \right) \nabla F_{con}(\mathbf{x}_{con}^s) \right\| \\
&= \left\| \alpha \sum_{s=1}^{t} (1-\alpha)^{t-s} \left[(\mathbf{W}^{t-s} - \frac{\mathbf{1}_n \mathbf{1}_n^\dagger}{n}) \otimes \mathbf{I} \right] \nabla F_{con}(\mathbf{x}_{con}^s) \right\| \\
&\le \alpha \sqrt{n} G \sum_{s=1}^{t} (1-\alpha)^{t-s} \beta^{t-s} \\
&\le \frac{\alpha \sqrt{n} G}{1 - \beta(1-\alpha)},
\end{aligned} \tag{41}$$

where the first equality is implied by replacing $\mathbf{d}_{con}^t$ and $\bar{\mathbf{d}}_{con}^t$ with the expressions in (39) and (40), respectively, the second equality is achieved by regrouping the terms, the first inequality holds since $\|\nabla F_i(x_i^s)\| \le G$ and $\|\mathbf{W}^{t-s-1} - (\mathbf{1}_n \mathbf{1}_n^\dagger)/n\| \le \beta^{t-s-1}$, and finally the last inequality is valid since $\sum_{s=1}^t ((1-\alpha)\beta)^{t-s} \le \frac{1}{1-(\beta(1-\alpha))}$. Now considering the result in (41) and the expression $\|\mathbf{d}_{con}^t - \bar{\mathbf{d}}_{con}^t\|^2 = \sum_{i=1}^n \|\mathbf{d}_i^t - \bar{\mathbf{d}}^t\|^2$, the claim in (20) follows.

9.5. Proof of Lemma 4

Considering the update in (6), we can write the sum of local ascent directions $\mathbf{d}_i^t$ at step t as

$$\begin{aligned}
\sum_{i=1}^{n} \mathbf{d}_i^t &= (1-\alpha) \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \mathbf{d}_j^{t-1} + \alpha \sum_{i=1}^{n} \nabla F_i(\mathbf{x}_i^t) \\
&= (1-\alpha) \sum_{j=1}^{n} \mathbf{d}_j^{t-1} \sum_{i=1}^{n} w_{ij} + \alpha \sum_{i=1}^{n} \nabla F_i(\mathbf{x}_i^t) \\
&= (1-\alpha) \sum_{j=1}^{n} \mathbf{d}_j^{t-1} + \alpha \sum_{i=1}^{n} \nabla F_i(\mathbf{x}_i^t),
\end{aligned} \tag{42}$$

where the last equality holds since $\sum_{i=1}^n w_{ij} = 1$ which is the consequence of $\mathbf{W}^\dagger \mathbf{1}_n = \mathbf{1}_n$. Now, we use the expression in (42) to bound the difference $\| \sum_{i=1}^n \mathbf{d}_i^t - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t) \|$. Hence,

$$\begin{aligned}
&\left\| \sum_{i=1}^n \mathbf{d}_i^t - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t) \right\| \\
&= \left\| (1-\alpha) \sum_{j=1}^n \mathbf{d}_j^{t-1} + \alpha \sum_{i=1}^n \nabla F_i(\mathbf{x}_i^t) - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t) \right\| \\
&= \left\| (1-\alpha) \sum_{j=1}^n \mathbf{d}_j^{t-1} - (1-\alpha) \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^{t-1}) + (1-\alpha) \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^{t-1}) + \alpha \sum_{i=1}^n \nabla F_i(\mathbf{x}_i^t) - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t) \right\| \\
&= \left\| (1-\alpha) \left[\sum_{j=1}^n \mathbf{d}_j^{t-1} - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^{t-1}) \right] + (1-\alpha) \left[\sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^{t-1}) - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t) \right] + \alpha \left[\sum_{i=1}^n \nabla F_i(\mathbf{x}_i^t) - \nabla F_i(\bar{\mathbf{x}}^t) \right] \right\| \\
&\leq (1-\alpha) \left\| \sum_{j=1}^n \mathbf{d}_j^{t-1} - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^{t-1}) \right\| + (1-\alpha) \left\| \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^{t-1}) - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t) \right\| + \alpha \left\| \sum_{i=1}^n \nabla F_i(\mathbf{x}_i^t) - \nabla F_i(\bar{\mathbf{x}}^t) \right\|.
\end{aligned} \tag{43}$$

The first equality is the outcome of replacing $\sum_{i=1}^n \mathbf{d}_i^t$ by the expression in (42), the second equality is obtained by adding and subtracting $(1-\alpha)\sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^{t-1})$, in the third equality we regroup the terms, and the inequality follows from applying the triangle inequality twice. Applying the Cauchy–Schwarz inequality to the second and third summands in (43) and using the Lipschitz continuity of the gradients lead to

$$\left\| \sum_{i=1}^n \mathbf{d}_i^t - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t) \right\| \leq (1-\alpha) \left\| \sum_{j=1}^n \mathbf{d}_j^{t-1} - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^{t-1}) \right\| + (1-\alpha) L \sum_{i=1}^n \left\| \bar{\mathbf{x}}^{t-1} - \bar{\mathbf{x}}^t \right\| + \alpha L \sum_{i=1}^n \| \mathbf{x}_i^t - \bar{\mathbf{x}}^t \| \tag{44}$$

According to the result in Lemma 1, we can bound the $\sum_{i=1}^n \|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\|$ by nD/T. Further, the result in Lemma 2 shows that $(\sum_{i=1}^n \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2)^{1/2} \leq \frac{\sqrt{n}D}{T(1-\beta)}$. Since by the Cauchy–Swartz inequality it holds that $(\sum_{i=1}^n \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2)^{1/2} \geq \frac{1}{\sqrt{n}} \sum_{i=1}^n \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|$, it follows that $\sum_{i=1}^n \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\| \leq (nD)/(T(1-\beta))$. Applying these substitutions into (44) yields

$$\left\| \sum_{i=1}^n \mathbf{d}_i^t - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t) \right\| \leq (1-\alpha) \left\| \sum_{j=1}^n \mathbf{d}_j^{t-1} - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^{t-1}) \right\| + \frac{(1-\alpha)LnD}{T} + \frac{\alpha LnD}{T(1-\beta)} \tag{45}$$

By multiplying both of sides of (45) by $1/n$ and applying the resulted inequality recessively for t steps we obtain

$$\begin{aligned}
\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{d}_i^t - \frac{1}{n} \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t) \right\| &\leq (1-\alpha)^t \left\| \frac{1}{n} \sum_{j=1}^n \mathbf{d}_j^0 - \frac{1}{n} \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^0) \right\| + \left(\frac{(1-\alpha)LD}{T} + \frac{\alpha LD}{T(1-\beta)} \right) \sum_{s=0}^{t-1} (1-\alpha)^s \\
&\leq (1-\alpha)^t \frac{1}{n} \sum_{i=1}^n \left\| \nabla F_i(\bar{\mathbf{x}}^0) \right\| + \frac{(1-\alpha)LD}{\alpha T} + \frac{LD}{T(1-\beta)} \\
&\leq (1-\alpha)^t G + \frac{(1-\alpha)LD}{\alpha T} + \frac{LD}{T(1-\beta)},
\end{aligned} \tag{46}$$

where the second inequality holds since $\sum_{j=1}^n \mathbf{d}_j^0 = \mathbf{0}_p$ and $\sum_{s=0}^{t-1} (1-\alpha)^s \leq 1/\alpha$, and the last inequality follows from Assumption 4.

9.6. Proof of Theorem 1

Recall the definition of $\bar{\mathbf{x}}^t = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^t$ as the average of local variables at step t. Since the gradients of the global objective function are L-Lipschitz we can write

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^{t+1}) - \frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^t) &\geq \frac{1}{n}\langle \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t), \bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\rangle - \frac{L}{2}\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\|^2 \\ &= \frac{1}{T}\langle \frac{1}{n}\sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t), \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^t\rangle - \frac{L}{2T^2}\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^t\right\|^2, \end{aligned} \tag{47}$$

where the equality holds due to the expression in (30). Note that the term $\|(1/n)\sum_{i=1}^n \mathbf{v}_i^t\|^2$ can be upper bounded by D^2 according to Assumption 2, since $(1/n)\sum_{i=1}^n \mathbf{v}_i^t \in \mathcal{C}$. Apply this substition into (47) and add and subtract $(1/nT)\sum_{i=1}^n \mathbf{d}_i^t$ to obtain

$$\frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^{t+1}) - \frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^t) \geq \frac{1}{T}\langle \frac{1}{n}\sum_{i=1}^n \mathbf{d}_i^t, \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^t\rangle + \frac{1}{T}\langle \frac{1}{n}\sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t) - \frac{1}{n}\sum_{i=1}^n \mathbf{d}_i^t, \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^t\rangle - \frac{LD^2}{2T^2}. \tag{48}$$

Now by rewriting the inner product $\langle \sum_{i=1}^n \mathbf{d}_i^t, \sum_{i=1}^n \mathbf{v}_i^t\rangle$ as $\sum_{i=1}^n\sum_{j=1}^n \langle \mathbf{d}_i^t, \mathbf{v}_j^t\rangle = \sum_{j=1}^n \langle \sum_{i=1}^n \mathbf{d}_i^t, \mathbf{v}_j^t\rangle$, we can rewrite the right hand side of (48) as

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^{t+1}) - \frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^t) \\ &\geq \frac{1}{n^2T}\sum_{j=1}^n \langle \sum_{i=1}^n \mathbf{d}_i^t, \mathbf{v}_j^t\rangle + \frac{1}{T}\langle \frac{1}{n}\sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t) - \frac{1}{n}\sum_{i=1}^n \mathbf{d}_i^t, \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^t\rangle - \frac{LD^2}{2T^2} \\ &= \frac{1}{nT}\sum_{j=1}^n \langle \mathbf{d}_j^t, \mathbf{v}_j^t\rangle + \frac{1}{nT}\sum_{j=1}^n \langle (\frac{1}{n}\sum_{i=1}^n \mathbf{d}_i^t - \mathbf{d}_j^t), \mathbf{v}_j^t\rangle + \frac{1}{T}\langle \sum_{i=1}^n \frac{1}{n}\nabla F_i(\bar{\mathbf{x}}^t) - \frac{1}{n}\sum_{i=1}^n \mathbf{d}_i^t, \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^t\rangle - \frac{LD^2}{2T^2}. \end{aligned} \tag{49}$$

Note that in the last step we added and and subtracted $(1/nT)\sum_{j=1}^n \langle \mathbf{d}_j^t, \mathbf{v}_j^t\rangle$. Now according to the update in (7) we can write, $\langle \mathbf{d}_j^t, \mathbf{v}_j^t\rangle = \max_{\mathbf{v}\in\mathcal{C}} \langle \mathbf{d}_j^t, \mathbf{v}\rangle \geq \langle \mathbf{d}_j^t, \mathbf{x}^*\rangle$. Hence, we can replace $\langle \mathbf{d}_j^t, \mathbf{v}_j^t\rangle$ by its lower bound $\langle \mathbf{d}_j^t, \mathbf{x}^*\rangle$ to obtain

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^{t+1}) - \frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^t) \\ &\geq \frac{1}{nT}\sum_{j=1}^n \langle \mathbf{d}_j^t, \mathbf{x}^*\rangle + \frac{1}{nT}\sum_{j=1}^n \langle (\frac{1}{n}\sum_{i=1}^n \mathbf{d}_i^t - \mathbf{d}_j^t), \mathbf{v}_j^t\rangle + \frac{1}{T}\langle \frac{1}{n}\sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t) - \frac{1}{n}\sum_{i=1}^n \mathbf{d}_i^t, \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^t\rangle - \frac{LD^2}{2T^2}. \end{aligned} \tag{50}$$

Adding and subtracting $\frac{1}{n^2T}\sum_{j=1}^n \langle \sum_{i=1}^n \mathbf{d}_i^t, \mathbf{x}^*\rangle$ and regrouping the terms lead to

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^{t+1}) - \frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^t) &\geq \frac{1}{n^2T}\sum_{j=1}^n \langle \sum_{i=1}^n \mathbf{d}_i^t, \mathbf{x}^*\rangle + \frac{1}{nT}\sum_{j=1}^n \langle \mathbf{d}_j^t - \frac{1}{n}\sum_{i=1}^n \mathbf{d}_i^t, \mathbf{x}^*\rangle + \frac{1}{nT}\sum_{j=1}^n \langle \frac{1}{n}\sum_{i=1}^n \mathbf{d}_i^t - \mathbf{d}_j^t, \mathbf{v}_j^t\rangle \\ &+ \frac{1}{T}\langle \frac{1}{n}\sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t) - \frac{1}{n}\sum_{i=1}^n \mathbf{d}_i^t, \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^t\rangle - \frac{LD^2}{2T^2} \end{aligned} \tag{51}$$

Further add and subtract the expression $\frac{1}{n^2T}\sum_{j=1}^n\langle\sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t),\mathbf{x}^*\rangle$ and combine the terms to obtain

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^{t+1})-\frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^t)\\
&\geq \frac{1}{n^2T}\sum_{j=1}^n\langle\sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t),\mathbf{x}^*\rangle+\frac{1}{nT}\sum_{j=1}^n\langle(\frac{1}{n}\sum_{i=1}^n\mathbf{d}_i^t-\frac{1}{n}\sum_{i=1}^n\nabla F_i(\bar{\mathbf{x}}^t),\mathbf{x}^*\rangle+\frac{1}{nT}\sum_{j=1}^n\langle(\mathbf{d}_j^t-\frac{1}{n}\sum_{i=1}^n\mathbf{d}_i^t,\mathbf{x}^*\rangle\\
&\quad+\frac{1}{nT}\sum_{j=1}^n\langle(\frac{1}{n}\sum_{i=1}^n\mathbf{d}_i^t-\mathbf{d}_j^t),\mathbf{v}_j^t\rangle+\frac{1}{T}\langle\sum_{i=1}^n\frac{1}{n}\nabla F_i(\bar{\mathbf{x}}^t)-\frac{1}{n}\sum_{i=1}^n\mathbf{d}_i^t,\frac{1}{n}\sum_{i=1}^n\mathbf{v}_i^t\rangle-\frac{LD^2}{2T^2}\\
&=\frac{1}{nT}\langle\sum_{i=1}^n\nabla F_i(\bar{\mathbf{x}}^t),\mathbf{x}^*\rangle+\frac{1}{nT}\langle\sum_{i=1}^n\mathbf{d}_i^t-\sum_{i=1}^n\nabla F_i(\bar{\mathbf{x}}^t),\mathbf{x}^*-\frac{1}{n}\sum_{i=1}^n\mathbf{v}_i^t\rangle+\frac{1}{nT}\sum_{j=1}^n\langle(\frac{1}{n}\sum_{i=1}^n\mathbf{d}_i^t-\mathbf{d}_j^t),\mathbf{v}_j^t-\mathbf{x}^*\rangle-\frac{LD^2}{2T^2}.
\end{aligned}\tag{52}$$

The monotonicity of the average function $(1/n)\sum_{i=1}^n F_i(\mathbf{x})$ and its concavity along positive directions imply that $\langle(1/n)\sum_{i=1}^n\nabla F_i(\bar{\mathbf{x}}^t),\mathbf{x}^*\rangle\geq(1/n)\sum_{i=1}^n F_i(\mathbf{x}^*)-(1/n)\sum_{i=1}^n F_i(\bar{\mathbf{x}}^t)$. By applying this substitution into (52) and using the Cauchy-Schwarz inequality we obtain

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^{t+1})-\frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^t)\geq\frac{1}{nT}\left[\sum_{i=1}^n F_i(\mathbf{x}^*)-\sum_{i=1}^n F_i(\bar{\mathbf{x}}^t)\right]-\frac{1}{nT}\left\|\sum_{i=1}^n\mathbf{d}_i^t-\sum_{i=1}^n\nabla F_i(\bar{\mathbf{x}}^t)\right\|\left\|\mathbf{x}^*-\frac{1}{n}\sum_{i=1}^n\mathbf{v}_i^t\right\|\\
-\frac{1}{nT}\sum_{j=1}^n\left\|\frac{1}{n}\sum_{i=1}^n\mathbf{d}_i^t-\mathbf{d}_j^t\right\|\|\mathbf{v}_j^t-\mathbf{x}^*\|-\frac{LD^2}{2T^2}.
\end{aligned}\tag{53}$$

Now we proceed to derive lower bounds for the negative terms on the right hand side of (53). Note that all $\mathbf{v}_i^t$ for $i=1,\dots,n$ belong to the convex set $\mathcal{C}$ and therefore the average vector $\frac{1}{n}\sum_{i=1}^n\mathbf{v}_i^t$ is also in the set. Hence, we can bound the difference $\|\mathbf{x}^*-\frac{1}{n}\sum_{i=1}^n\mathbf{v}_i^t\|$ by D according to Assumption 2. Indeed, the norm $\|\mathbf{v}_j^t-\mathbf{x}^*\|$ is also upper bounded by D and hence we can write

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^{t+1})-\frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^t)\\
&\geq\frac{1}{nT}\left[\sum_{i=1}^n F_i(\mathbf{x}^*)-\sum_{i=1}^n F_i(\bar{\mathbf{x}}^t)\right]-\frac{D}{nT}\left\|\sum_{i=1}^n\mathbf{d}_i^t-\sum_{i=1}^n\nabla F_i(\bar{\mathbf{x}}^t)\right\|-\frac{D}{nT}\sum_{j=1}^n\left\|\frac{1}{n}\sum_{i=1}^n\mathbf{d}_i^t-\mathbf{d}_j^t\right\|-\frac{LD^2}{2T^2}.
\end{aligned}\tag{54}$$

The result in Lemma 3 implies that $(\sum_{i=1}^n\|\mathbf{d}_i^t-\bar{\mathbf{d}}^t\|^2)^{1/2}\leq\frac{\alpha\sqrt{n}G}{1-\beta(1-\alpha)}$. Note that based on the Cauchy–Swartz inequality it holds that $(\sum_{i=1}^n\|\mathbf{d}_i^t-\bar{\mathbf{d}}^t\|^2)^{1/2}\geq\frac{1}{\sqrt{n}}\sum_{i=1}^n\|\mathbf{d}_i^t-\bar{\mathbf{d}}^t\|$, and hence, $\sum_{i=1}^n\|\mathbf{d}_i^t-\bar{\mathbf{d}}^t\|\leq\frac{\alpha nG}{1-\beta(1-\alpha)}$. Using this result and recalling the definition $\bar{\mathbf{d}}^t:=(1/n)\sum_{i=1}^n\mathbf{d}_i^t$, we obtain that

$$\frac{1}{n}\sum_{j=1}^n\left\|\frac{1}{n}\sum_{i=1}^n\mathbf{d}_i^t-\mathbf{d}_j^t\right\|\leq\frac{\alpha G}{1-\beta(1-\alpha)}.\tag{55}$$

Replace the term $\frac{1}{n}\sum_{j=1}^n\left\|\frac{1}{n}\sum_{i=1}^n\mathbf{d}_i^t-\mathbf{d}_j^t\right\|$ in (54) by its upper bound in (55) and use the result in Lemma 4 to replace $\frac{1}{n}\|\sum_{i=1}^n\mathbf{d}_i^t-\sum_{i=1}^n\nabla F_i(\bar{\mathbf{x}}^t)\|$ by its upper bound in (21). Applying these substitutions yields

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^{t+1})-\frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^t)\\
&\geq\frac{1}{T}\left[\frac{1}{n}\sum_{i=1}^n F_i(\mathbf{x}^*)-\frac{1}{n}\sum_{i=1}^n F_i(\bar{\mathbf{x}}^t)\right]-\frac{(1-\alpha)^tGD}{T}-\frac{(1-\alpha)LD^2}{\alpha T^2}-\frac{LD^2}{(1-\beta)T^2}-\frac{\alpha GD}{(1-\beta(1-\alpha))T}-\frac{LD^2}{2T^2}
\end{aligned}\tag{56}$$

Set $\alpha = 1/\sqrt{T}$ and regroup the terms to obtain

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} F_i(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^{n} F_i(\bar{\mathbf{x}}^{t+1}) \\
&\leq \left(1 - \frac{1}{T}\right)\left[\frac{1}{n}\sum_{i=1}^{n} F_i(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^{n} F_i(\bar{\mathbf{x}}^{t})\right] + \frac{(1-(1/\sqrt{T}))^t GD}{T} + \frac{LD^2}{T^{3/2}} + \frac{LD^2}{(1-\beta)T^2} + \frac{GD}{(1-\beta)T^{3/2}} + \frac{LD^2}{2T^2}
\end{aligned} \tag{57}$$

By applying the inequality in (57) recursively for $t = 0, \ldots, T-1$ we obtain

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} F_i(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^{n} F_i(\bar{\mathbf{x}}^{T}) \leq \left(1 - \frac{1}{T}\right)^T \left[\frac{1}{n}\sum_{i=1}^{n} F_i(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^{n} F_i(\bar{\mathbf{x}}^{0})\right] + \sum_{t=0}^{T-1}\frac{\left(1-1/\sqrt{T}\right)^t GD}{T} + \sum_{t=0}^{T-1}\frac{LD^2}{T^{3/2}} \\
+ \sum_{t=0}^{T-1}\frac{LD^2}{(1-\beta)T^2} + \sum_{t=0}^{T-1}\frac{GD}{(1-\beta)T^{3/2}} + \sum_{t=0}^{T-1}\frac{LD^2}{2T^2}.
\end{aligned} \tag{58}$$

By using the inequality $\sum_{t=0}^{T-1}(1-1/\sqrt{T})^t \leq \sqrt{T}$ and simplifying the terms on the right hand side (58) we obtain that to the expression

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} F_i(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^{n} F_i(\bar{\mathbf{x}}^{T}) \\
&\leq \frac{1}{e}\left[\frac{1}{n}\sum_{i=1}^{n} F_i(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^{n} F_i(\bar{\mathbf{x}}^{0})\right] + \frac{GD}{T^{1/2}} + \frac{LD^2}{T^{1/2}} + \frac{LD^2}{(1-\beta)T} + \frac{GD}{(1-\beta)T^{1/2}} + \frac{LD^2}{2T} \\
&= \frac{1}{e}\left[\frac{1}{n}\sum_{i=1}^{n} F_i(\mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^{n} F_i(\bar{\mathbf{x}}^{0})\right] + \frac{LD^2 + GD(1+(1-\beta)^{-1})}{T^{1/2}} + \frac{LD^2(0.5+(1-\beta)^{-1})}{T},
\end{aligned} \tag{59}$$

where to derive the first inequality we used $(1-1/T)^T \leq 1/e$. Note that we set $\mathbf{x}_i^0 = \mathbf{0}_p$ for all $i \in \mathcal{N}$ and therefore $\bar{\mathbf{x}}^0 = \mathbf{0}_p$. Since we assume that $F_i(\mathbf{0}_p) \geq 0$ for all $i \in \mathcal{N}$, it implies that $\frac{1}{n}\sum_{i=1}^{n} F_i(\bar{\mathbf{x}}^0) = \frac{1}{n}\sum_{i=1}^{n} F_i(\mathbf{0}_p) \geq 0$ and the expression in (59) can be simplified to

$$\frac{1}{n}\sum_{i=1}^{n} F_i(\bar{\mathbf{x}}^{T}) \geq (1-e^{-1})\frac{1}{n}\sum_{i=1}^{n} F_i(\mathbf{x}^*) - \frac{LD^2 + GD(1+(1-\beta)^{-1})}{T^{1/2}} - \frac{LD^2(0.5+(1-\beta)^{-1})}{T}. \tag{60}$$

Also, since the norm of local gradients is uniformly bounded by G, the local functions F_i are G-Lipschitz. This observation implies that

$$\left|\frac{1}{n}\sum_{i=1}^{n} F_i(\bar{\mathbf{x}}^{T}) - \frac{1}{n}\sum_{i=1}^{n} F_i(\mathbf{x}_j^{T})\right| \leq \frac{G}{n}\sum_{i=1}^{n}\|\bar{\mathbf{x}}^T - \mathbf{x}_j^T\| \leq \frac{GD}{T(1-\beta)}, \tag{61}$$

where the second inequality holds by using the result in Lemma 2 and the Cauchy-Schwartz inequality. Therefore, by combining the results in (60) and (61) we obtain that for all $j = \mathcal{N}$

$$\frac{1}{n}\sum_{i=1}^{n} F_i(\mathbf{x}_j^{T}) \geq (1-e^{-1})\frac{1}{n}\sum_{i=1}^{n} F_i(\mathbf{x}^*) - \frac{LD^2 + GD(1+(1-\beta)^{-1})}{T^{1/2}} - \frac{GD(1-\beta)^{-1} + LD^2(0.5+(1-\beta)^{-1})}{T}, \tag{62}$$

and the claim in (22) follows.

9.7. How to Construct an Unbiased Estimator of the Gradient in Multilinear Extensions

In this section, we provide an unbiased estimator for the gradient of a multilinear extension. We thus consider an arbitrary submodular set function $h : 2^V \to \mathbb{R}$ with multilinear H. Our goal is to provide an unbiased estimator for $\nabla H(\mathbf{x})$. We have $H(\mathbf{x}) = \sum_{S\subseteq V} \prod_{i\in S} \mathbf{x}_i \prod_{j\notin S}(1-x_j)h(S)$. Now, it can easily be shown that (see ())

$$\frac{\partial H}{\partial x_i} = H(\mathbf{x}; \mathbf{x}_i \leftarrow 1) - H(\mathbf{x}; \mathbf{x}_i \leftarrow 0).$$

where for example by $(\mathbf{x}; \mathbf{x}_i \leftarrow 1)$ we mean a vector which has value 1 on its i-th coordinate and is equal to $\mathbf{x}$ elsewhere. To create an unbiased estimator for $\frac{\partial H}{\partial x_i}$ at a point $\mathbf{x}$ we can simply sample a set S by including each element in it independently with probability x_i and use $h(S \cup \{i\}) - h(S \setminus \{i\})$ as an unbiased estimator for the i-th partial derivative. We can sample one single set S and use the above trick for all the coordinates. This involves n function computations for h. Having a mini-batch size B we can repeat this procedure B times and then average.

Note that since every element of the unbiased estimator is of the form $h(S \cup \{i\}) - h(S \setminus \{i\})$ for some chosen set S, then due to submodularity of the function h every element of the unbiased estimator is bounded above by the maximum marginal value of h (i.e. $\max_{i\in V}$ h({i})). As a result, the norm of the unbiased estimator (of the gradient of H) is bounded above by $\sqrt{|V|}\max_{i\in V} h(\{i\})$.

9.8. Proof of Theorem 2

The steps of the proof are similar to the one for Theorem 1. In particular, for the Discrete DCG method we can also show that the expressions in (47)-(54) hold and we can write

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} F_i(\bar{\mathbf{x}}^{t+1}) - \frac{1}{n}\sum_{i=1}^{n} F_i(\bar{\mathbf{x}}^{t}) \\ &\geq \frac{1}{nT}\left[\sum_{i=1}^{n} F_i(\mathbf{x}^*) - \sum_{i=1}^{n} F_i(\bar{\mathbf{x}}^{t})\right] - \frac{D}{nT}\left\|\sum_{i=1}^{n}\mathbf{d}_i^t - \sum_{i=1}^{n}\nabla F_i(\bar{\mathbf{x}}^t)\right\| - \frac{D}{nT}\sum_{j=1}^{n}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{d}_i^t - \mathbf{d}_j^t\right\| - \frac{LD^2}{2T^2}. \qquad (63)\end{aligned}$$

Now we proceed to derive upper bounds for the norms on the right hand side of (63). To derive these bounds we use the results in Lemmata 1 and 2 which also hold for the Discrete DCG algorithm.

We first derive an upper bound for the sum $\sum_{j=1}^{n}\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{d}_i^t - \mathbf{d}_j^t\|$ in (63). To achieve this goal the following lemma is needed.

Lemma 5 *Consider the proposed Discrete DCG method defined in Algorithm 2. If Assumptions 4 and 5 hold, then for all $i \in \mathcal{N}$ and $t \geq 0$ the expected squared norm $\mathbb{E}\left[\|\mathbf{g}_i^t\|^2\right]$ is bounded above by*

$$\mathbb{E}\left[\|\mathbf{g}_i^t\|^2\right] \leq K^2, \qquad (64)$$

where $K^2 = \sigma^2 + G^2$.

Proof: Considering the condition in Assumption 5 on the variance of stochastic gradients, we can define $K^2 := \sigma^2 + G^2$ as an upper bound on the expected norm of stochastic gradients, i.e., for all $\mathbf{x} \in \mathcal{C}$ and $i \in \mathcal{N}$

$$\mathbb{E}\left[\|\nabla \tilde{F}_i(\mathbf{x}_i^t)\|^2\right] \leq K^2. \qquad (65)$$

Now we use an induction argument to show that the expected norm $\mathbb{E}\left[\|\mathbf{g}_i^t\|^2\right] \leq K^2$. Since the iterates are initialized at $\mathbf{g}_i^0 = \mathbf{0}$, the update in (12) implies that $\mathbb{E}\left[\|\mathbf{g}_i^1\|^2 \mid \mathbf{x}_i^1\right] = \phi^2\mathbb{E}\left[\|\nabla \tilde{F}_i(\mathbf{x}_i^1)\|^2 \mid \mathbf{x}_i^1\right] \leq \phi^2K^2 \leq K^2$. Since $\mathbb{E}\left[\mathbb{E}\left[\|\mathbf{g}_i^1\|^2 \mid \mathbf{x}_i^1\right]\right] = \mathbb{E}\left[\|\mathbf{g}_i^1\|^2\right]$ it follows that $\mathbb{E}\left[\|\mathbf{g}_i^1\|^2\right] \leq K^2$. Now we proceed to show that if $\mathbb{E}\left[\|\mathbf{g}_i^{t-1}\|^2\right] \leq K^2$ then $\mathbb{E}\left[\|\mathbf{g}_i^t\|^2\right] \leq K^2$.

Recall the update of $\mathbf{g}_i^t$ in (12). By computing the squared norm of both sides and using the Cauchy-Schwartz inequality we obtain that

$$\|\mathbf{g}_i^t\|^2 \leq (1-\phi)^2\|\mathbf{g}_i^{t-1}\|^2 + \phi^2\|\nabla \tilde{F}_i(\mathbf{x}_i^t)\|^2 + 2\phi(1-\phi)\|\mathbf{g}_i^{t-1}\|\|\nabla \tilde{F}_i(\mathbf{x}_i^t)\|. \qquad (66)$$

Compute the expectation with respect to the random variable corresponding to the stochastic gradient $\nabla \tilde{F}_i(\mathbf{x}_i^t)$ to obtain

$$\mathbb{E}\left[\|\mathbf{g}_i^t\|^2 \mid \mathbf{x}_i^t\right] \leq (1-\phi)^2\|\mathbf{g}_i^{t-1}\|^2 + \phi^2\mathbb{E}\left[\|\nabla \tilde{F}_i(\mathbf{x}_i^t)\|^2 \mid \mathbf{x}_i^t\right] + 2\phi(1-\phi)\|\mathbf{g}_i^{t-1}\|\mathbb{E}\left[\|\nabla \tilde{F}_i(\mathbf{x}_i^t)\| \mid \mathbf{x}_i^t\right]. \qquad (67)$$

Note that according to Jensen's inequality $\mathbb{E}\left[\|\nabla \tilde{F}_i(\mathbf{x}_i^t)\|^2\right] \leq K^2$ implies that $\mathbb{E}\left[\|\nabla \tilde{F}_i(\mathbf{x}_i^t)\|\right] \leq K$. Replacing these bounds into (67) yields

$$\mathbb{E}\left[\|\mathbf{g}_i^t\|^2 \mid \mathbf{x}_i^t\right] \leq (1-\phi)^2\|\mathbf{g}_i^{t-1}\|^2 + \phi^2K^2 + 2K\phi(1-\phi)\|\mathbf{g}_i^{t-1}\|. \tag{68}$$

Now by computing the expectation of both sides with respect to all sources of randomness from $t = 0$ and using the simplification $\mathbb{E}\left[\mathbb{E}\left[\|\mathbf{g}_i^t\|^2 \mid \mathbf{x}_i^t\right]\right] = \mathbb{E}\left[\|\mathbf{g}_i^t\|^2\right]$ we can write

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{g}_i^t\|^2 \mid\right] &\leq (1-\phi)^2\mathbb{E}\left[\|\mathbf{g}_i^{t-1}\|^2\right] + \phi^2K^2 + 2K\phi(1-\phi)\mathbb{E}\left[\|\mathbf{g}_i^{t-1}\|\right] \\ &\leq (1-\phi)^2K^2 + \phi^2K^2 + 2K\phi(1-\phi)K \\ &= K^2, \end{aligned}\tag{69}$$

and the claim in (64) follows by induction. ■

We use the result in Lemma 5 to find an upper bound for the sum $(1/n)\sum_{j=1}^n \left\|\bar{\mathbf{d}}^t - \mathbf{d}_j^t\right\|$ on the right hand side of (63).

Lemma 6 *Consider the proposed Discrete DCG method defined in Algorithm 2. If Assumptions 1, 4 and 5 hold, then for all $i \in \mathcal{N}$ and $t \geq 0$ we have*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|\mathbf{d}_i^t - \bar{\mathbf{d}}^t\|\right] \leq \frac{\alpha K}{1-\beta(1-\alpha)}, \tag{70}$$

where $K = (\sigma^2 + G^2)^{1/2}$.

Proof: Define the vector $\mathbf{g}_{con}^t = [\mathbf{g}_1^t; \ldots; \mathbf{g}_n^t]$ as the concatenation of the local vectors $\mathbf{g}_i^t$ at time t. Further, recall the definitions of the vectors $\mathbf{x}_{con} = [\mathbf{x}_1; \ldots; \mathbf{x}_n] \in \mathbb{R}^{np}$ and $\mathbf{d}_{con} = [\mathbf{d}_1; \ldots; \mathbf{d}_n] \in \mathbb{R}^{np}$ as the concatenation of the local variables and local approximate gradients, respectively, and the definition of $\bar{\mathbf{d}}_{con}^t = [\bar{\mathbf{d}}^t; \ldots; \bar{\mathbf{d}}^t]$ as the concatenation of n copies of the average vector $\bar{\mathbf{d}}^t$. By following the steps of the proof for Lemma 3, it can be shown that

$$\begin{aligned}\|\mathbf{d}_{con}^t - \bar{\mathbf{d}}_{con}^t\| &= \left\|\alpha\sum_{s=1}^t (1-\alpha)^{t-s}\left[(\mathbf{W}^{t-s} - \frac{\mathbf{1}_n\mathbf{1}_n^\dagger}{n}) \otimes \mathbf{I}\right]\mathbf{g}_{con}^t\right\| \\ &\leq \alpha\sum_{s=1}^t (1-\alpha)^{t-s}\left\|(\mathbf{W}^{t-s} - \frac{\mathbf{1}_n\mathbf{1}_n^\dagger}{n}) \otimes \mathbf{I}\right\|\left\|\mathbf{g}_{con}^t\right\| \\ &\leq \alpha\sum_{s=1}^t (1-\alpha)^{t-s}\beta^{t-s}\left\|\mathbf{g}_{con}^t\right\|. \end{aligned}\tag{71}$$

By computing the expected value of both sides and using the result in (64) we obtain that

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{d}_{con}^t - \bar{\mathbf{d}}_{con}^t\|\right] &\leq \alpha\sqrt{n}K\sum_{s=1}^t (1-\alpha)^{t-s}\beta^{t-s} \\ &\leq \frac{\alpha\sqrt{n}K}{1-\beta(1-\alpha)}, \end{aligned}\tag{72}$$

where in the first inequality we use the fact that $\mathbb{E}\left[\|\mathbf{g}_{con}^t\|\right] \leq (\mathbb{E}\left[\|\mathbf{g}_{con}^t\|^2\right])^{1/2} = (\mathbb{E}\left[(\sum_{i=1}^n \|\mathbf{g}_i^t\|^2)\right])^{1/2} = (\sum_{i=1}^n \mathbb{E}\left[\|\mathbf{g}_i^t\|^2\right])^{1/2} \leq \sqrt{n}K$. By combining the result in (72) with the inequality

$$\frac{1}{n}\sum_{i=1}^n \|\mathbf{d}_i^t - \bar{\mathbf{d}}^t\| \leq \frac{1}{\sqrt{n}}\left[\sum_{i=1}^n \|\mathbf{d}_i^t - \bar{\mathbf{d}}^t\|^2\right]^{1/2} = \frac{1}{\sqrt{n}}\|\mathbf{d}_{con}^t - \bar{\mathbf{d}}_{con}^t\|, \tag{73}$$

the claim in (70) follows. ■

The result in Lemma 6 shows that the sum $\frac{1}{n}\sum_{i=1}^n \|\mathbf{d}_i^t - \bar{\mathbf{d}}^t\|$ is bounded above by $(\alpha K)/(1-\beta(1-\alpha))$ in expectation. To bound the second sum in (63), which is $\|\sum_{i=1}^n \mathbf{d}_i^t - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t)\|$, we first introduce the following lemma, which was presented in (Mokhtari et al., 2018a) in a slightly different form.

Lemma 7 *Consider the proposed Discrete DCG method defined in Algorithm 2. If Assumptions 1-5 hold and we set $\phi = 1/T^{2/3}$, then for all $i \in \mathcal{N}$ and $t \geq 0$ we have*

$$\mathbb{E}\left[\sum_{i=1}^{n} \|\nabla F_i(\mathbf{x}_i^t) - \mathbf{g}_i^t\|^2\right] \leq \left(1 - \frac{1}{2T^{2/3}}\right)^t nG^2 + \frac{6nL^2D^2C}{T^{4/3}} + \frac{2n\sigma^2 + 12nL^2D^2C}{T^{2/3}}, \tag{74}$$

where $C := 1 + (2/(1-\beta)^2)$.

Proof: Use the update $\mathbf{g}_i^t := (1-\phi)\mathbf{g}_i^{t-1} + \phi\nabla\tilde{F}(\mathbf{x}_i^t)$ to write the squared norm $\|\nabla F_i(\mathbf{x}_i^t) - \mathbf{g}_i^t\|^2$ as

$$\|\nabla F_i(\mathbf{x}_i^t) - \mathbf{g}_i^t\|^2 = \|\nabla F_i(\mathbf{x}_i^t) - (1-\phi)\mathbf{d}_{t-1} - \phi\nabla\tilde{F}_i(\mathbf{x}_i^t)\|^2. \tag{75}$$

Add and subtract the term $(1-\phi)\nabla F_i(\mathbf{x}_i^{t-1})$ to the right hand side of (75) and regroup the terms to obtain

$$\|\nabla F_i(\mathbf{x}_i^t) - \mathbf{g}_i^t\|^2 = \|\phi(\nabla F_i(\mathbf{x}_i^t) - \nabla\tilde{F}_i(\mathbf{x}_i^t)) + (1-\phi)(\nabla F_i(\mathbf{x}_i^t) - \nabla F_i(\mathbf{x}_i^{t-1})) + (1-\phi)(\nabla F_i(\mathbf{x}_i^{t-1}) - \mathbf{g}_i^{t-1})\|^2. \tag{76}$$

Define $\mathcal{F}^t$ as a sigma algebra that measures the history of the system up until time t. Expanding the square and computing the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}^t]$ of the resulted expression yield

$$\begin{aligned}\mathbb{E}\left[\|\nabla F_i(\mathbf{x}_i^t) - \mathbf{g}_i^t\|^2 \mid \mathcal{F}^t\right] &= \phi^2\mathbb{E}\left[\|\nabla F_i(\mathbf{x}_i^t) - \nabla\tilde{F}_i(\mathbf{x}_i^t)\|^2 \mid \mathcal{F}^t\right] + (1-\phi)^2\|\nabla F_i(\mathbf{x}_i^{t-1}) - \mathbf{g}_i^{t-1}\|^2 \\ &+ (1-\phi)^2\|\nabla F_i(\mathbf{x}_i^t) - \nabla F_i(\mathbf{x}_i^{t-1})\|^2 + 2(1-\phi)^2\langle\nabla F_i(\mathbf{x}_i^t) - \nabla F_i(\mathbf{x}_i^{t-1}), \nabla F_i(\mathbf{x}_i^{t-1}) - \mathbf{g}_i^{t-1}\rangle,\end{aligned} \tag{77}$$

where we have used the fact $\mathbb{E}\left[\nabla\tilde{F}_i(\mathbf{x}_i^t) \mid \mathcal{F}^t\right] = \nabla F_i(\mathbf{x}_i^t)$. The term $\mathbb{E}\left[\|\nabla F_i(\mathbf{x}_i^t) - \nabla\tilde{F}_i(\mathbf{x}_i^t)\|^2 \mid \mathcal{F}^t\right]$ can be bounded above by σ^2 according to Assumption 5. Based on Assumption 3, we can also show that the squared norm $\|\nabla F_i(\mathbf{x}_i^t) - \nabla F_i(\mathbf{x}_i^{t-1})\|^2$ is upper bounded by $L^2\|\mathbf{x}_i^t - \mathbf{x}_i^{t-1}\|^2$. Moreover, the inner product $2\langle\nabla F_i(\mathbf{x}_i^t) - \nabla F_i(\mathbf{x}_i^{t-1}), \nabla F_i(\mathbf{x}_i^{t-1}) - \mathbf{d}_{t-1}\rangle$ can be upper bounded by $\zeta\|\nabla F_i(\mathbf{x}_i^{t-1}) - \mathbf{d}_{t-1}\|^2 + (1/\zeta)L^2\|\mathbf{x}_i^t - \mathbf{x}_i^{t-1}\|^2$ using Young's inequality (i.e., $2\langle\mathbf{a}, \mathbf{b}\rangle \leq \zeta\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2/\beta$ for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ and $\zeta > 0$) and the condition in Assumption 3, where $\zeta > 0$ is a free scalar. Applying these substitutions into (77) leads to

$$\mathbb{E}\left[\|\nabla F_i(\mathbf{x}_i^t) - \mathbf{g}_i^t\|^2 \mid \mathcal{F}^t\right] \leq \phi^2\sigma^2 + (1-\phi)^2(1+\zeta^{-1})L^2\|\mathbf{x}_i^t - \mathbf{x}_i^{t-1}\|^2 + (1-\phi)^2(1+\zeta)\|\nabla F_i(\mathbf{x}_i^{t-1}) - \mathbf{g}_i^{t-1}\|^2. \tag{78}$$

By setting $\zeta = \phi/2$ we can replace $(1-\phi)^2(1+\zeta^{-1})$ and $(1-\phi)^2(1+\zeta)$ by their upper bounds $(1+2\phi^{-1})$ and $(1-\phi/2)$, respectively. Applying theses substitutions and summing up both sides of the resulted inequality for $i = 1, \ldots, n$ lead to

$$\mathbb{E}\left[\sum_{i=1}^{n}\|\nabla F_i(\mathbf{x}_i^t) - \mathbf{g}_i^t\|^2 \mid \mathcal{F}^t\right] \leq n\phi^2\sigma^2 + L^2(1+2\phi^{-1})\sum_{i=1}^{n}\|\mathbf{x}_i^t - \mathbf{x}_i^{t-1}\|^2 + \left(1 - \frac{\phi}{2}\right)\sum_{i=1}^{n}\|\nabla F_i(\mathbf{x}_i^{t-1}) - \mathbf{g}_i^{t-1}\|^2. \tag{79}$$

Now we proceed to derive an upper bound for the sum $\sum_{i=1}^n \|\mathbf{x}_i^t - \mathbf{x}_i^{t-1}\|^2$. Note that using the Cauchy-Schwartz inequality and the results in Lemmata 1 and 2 we can show that

$$\begin{aligned}\sum_{i=1}^{n}\|\mathbf{x}_i^t - \mathbf{x}_i^{t-1}\|^2 &\leq \sum_{i=1}^{n}\left(3\left\|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\right\|^2 + 3\left\|\bar{\mathbf{x}}^t - \bar{\mathbf{x}}^{t-1}\right\|^2 + 3\left\|\bar{\mathbf{x}}^{t-1} - \mathbf{x}_i^{t-1}\right\|^2\right) \\ &\leq \frac{3nD^2}{T^2(1-\beta)^2} + \frac{3nD^2}{T^2} + \frac{3nD^2}{T^2(1-\beta)^2} \\ &= \frac{3nD^2}{T^2}\left(1 + \frac{2}{(1-\beta)^2}\right)\end{aligned} \tag{80}$$

Replace the sum $\sum_{i=1}^n \|\mathbf{x}_i^t - \mathbf{x}_i^{t-1}\|^2$ in (79) by its upper bound in (80) and compute the expectation with respect to $\mathcal{F}_0$ to obtain

$$\mathbb{E}\left[\sum_{i=1}^{n}\|\nabla F_i(\mathbf{x}_i^t) - \mathbf{g}_i^t\|^2\right] \leq \left(1 - \frac{\phi}{2}\right)\mathbb{E}\left[\sum_{i=1}^{n}\|\nabla F_i(\mathbf{x}_i^{t-1}) - \mathbf{g}_i^{t-1}\|^2\right] + n\phi^2\sigma^2 + (1+2\phi^{-1})\frac{3nL^2D^2}{T^2}\left(1 + \frac{2}{(1-\beta)^2}\right) \tag{81}$$

Set $\phi = T^{-2/3}$ to obtain

$$\mathbb{E}\left[\sum_{i=1}^{n} \|\nabla F_i(\mathbf{x}_i^t) - \mathbf{g}_i^t\|^2\right] \leq \left(1 - \frac{1}{2T^{2/3}}\right) \mathbb{E}\left[\sum_{i=1}^{n} \|\nabla F_i(\mathbf{x}_i^{t-1}) - \mathbf{g}_i^{t-1}\|^2\right] + \frac{n\sigma^2}{T^{4/3}} + \frac{3nL^2D^2C}{T^2} + \frac{6nL^2D^2C}{T^{4/3}} \tag{82}$$

where $C := \left(1 + \frac{2}{(1-\beta)^2}\right)$. Applying the expression in (82) recursively leads to

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{n} \|\nabla F_i(\mathbf{x}_i^t) - \mathbf{g}_i^t\|^2\right] \\
&\leq \left(1 - \frac{1}{2T^{2/3}}\right)^t \sum_{i=1}^{n} \|\nabla F_i(\mathbf{x}_i^0) - \mathbf{d}_0\|^2 + \left(\frac{n\sigma^2}{T^{4/3}} + \frac{3nL^2D^2C}{T^2} + \frac{6nL^2D^2C}{T^{4/3}}\right) \sum_{s=0}^{t-1} \left(1 - \frac{1}{2T^{2/3}}\right)^s \\
&\leq \left(1 - \frac{1}{2T^{2/3}}\right)^t \sum_{i=1}^{n} \|\nabla F_i(\mathbf{x}_i^0) - \mathbf{d}_0\|^2 + \frac{2n\sigma^2}{T^{2/3}} + \frac{6nL^2D^2C}{T^{4/3}} + \frac{12nL^2D^2C}{T^{2/3}} \\
&\leq \left(1 - \frac{1}{2T^{2/3}}\right)^t nG^2 + \frac{2n\sigma^2}{T^{2/3}} + \frac{6nL^2D^2C}{T^{4/3}} + \frac{12nL^2D^2C}{T^{2/3}},
\end{aligned} \tag{83}$$

and the claim in (74) follows. ■

We use the result in Lemma 7 to derive an upper bound for $\left\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{d}_i^t - \frac{1}{n}\sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^t)\right\|$ in expectation.

Lemma 8 *Consider the proposed Discrete DCG method defined in Algorithm 2. If Assumptions 1-5 hold and we set $\alpha = 1/\sqrt{T}$ and $\phi = 1/T^{2/3}$, then for all $i \in \mathcal{N}$ and $t \geq 0$ we have*

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{d}_i^t - \frac{1}{n}\sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^t)\right\|\right] &\leq G\left(1 - \frac{1}{T^{1/2}}\right)^t + G\left(1 - \frac{1}{2T^{2/3}}\right)^{t/2} + \frac{LD}{T^{1/2}} \\
&\quad + \frac{LD}{T(1-\beta)} + \frac{\sqrt{6}LDC^{1/2}}{T^{2/3}} + \frac{\sqrt{2}\sigma + \sqrt{12}LDC^{1/2}}{T^{1/3}},
\end{aligned} \tag{84}$$

where $C := 1 + (2/(1-\beta)^2)$.

Proof: The steps of this proof are similar to the ones in the proof of Lemma 4. It can be shown that

$$\begin{aligned}
&\left\|\sum_{i=1}^{n} \mathbf{d}_i^t - \sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^t)\right\| \\
&= \left\|(1-\alpha)\sum_{j=1}^{n} \mathbf{d}_j^{t-1} + \alpha\sum_{i=1}^{n} \mathbf{g}_i^t - \sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^t)\right\| \\
&= \left\|(1-\alpha)\sum_{j=1}^{n} \mathbf{d}_j^{t-1} - (1-\alpha)\sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^{t-1}) + (1-\alpha)\sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^{t-1}) + \alpha\sum_{i=1}^{n} \mathbf{g}_i^t - \sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^t)\right\| \\
&= \left\|(1-\alpha)\left[\sum_{j=1}^{n} \mathbf{d}_j^{t-1} - \sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^{t-1})\right] + (1-\alpha)\left[\sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^{t-1}) - \sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^t)\right] + \alpha\left[\sum_{i=1}^{n} \mathbf{g}_i^t - \sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^t)\right]\right\| \\
&\leq (1-\alpha)\left\|\sum_{j=1}^{n} \mathbf{d}_j^{t-1} - \sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^{t-1})\right\| + (1-\alpha)\left\|\sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^{t-1}) - \sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^t)\right\| + \alpha\left\|\sum_{i=1}^{n} \mathbf{g}_i^t - \sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^t)\right\| \\
&\leq (1-\alpha)\left\|\sum_{j=1}^{n} \mathbf{d}_j^{t-1} - \sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^{t-1})\right\| + (1-\alpha)\left\|\sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^{t-1}) - \sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^t)\right\| + \alpha\left\|\sum_{i=1}^{n} \mathbf{g}_i^t - \sum_{i=1}^{n} \nabla F_i(\mathbf{x}_i^t)\right\| \\
&\quad + \alpha\left\|\sum_{i=1}^{n} \nabla F_i(\mathbf{x}_i^t) - \sum_{i=1}^{n} \nabla F_i(\bar{\mathbf{x}}^t)\right\|
\end{aligned} \tag{85}$$

The first equality is the outcome of replacing $\sum_{i=1}^n \mathbf{d}_i^t$ by the expression in (42), the second equality is obtained by adding and subtracting $(1-\alpha)\sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^{t-1})$, in the third equality we regroup the terms, and the inequality follows from applying the triangle inequality twice. Applying the Cauchy–Schwarz inequality to the second and third summands in (43) and using the Lipschitz continuity of the gradients lead to

$$\begin{aligned}
&\left\|\sum_{i=1}^n \mathbf{d}_i^t - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t)\right\| \\
&\leq (1-\alpha)\left\|\sum_{j=1}^n \mathbf{d}_j^{t-1} - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^{t-1})\right\| + (1-\alpha)L\sum_{i=1}^n \left\|\bar{\mathbf{x}}^{t-1}-\bar{\mathbf{x}}^t\right\| + \alpha L \sum_{i=1}^n \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\| + \alpha\left\|\sum_{i=1}^n \mathbf{g}_i^t - \sum_{i=1}^n \nabla F_i(\mathbf{x}_i^t)\right\| \\
&\leq (1-\alpha)\left\|\sum_{j=1}^n \mathbf{d}_j^{t-1} - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^{t-1})\right\| + \frac{(1-\alpha)LnD}{T} + \frac{\alpha LnD}{T(1-\beta)} + \alpha\sum_{i=1}^n \left\|\mathbf{g}_i^t - \nabla F_i(\mathbf{x}_i^t)\right\|, \qquad (86)
\end{aligned}$$

where the last inequality follows from Lemmata 1 and 2. Using the inequality

$$\frac{1}{\sqrt{n}}\mathbb{E}\left[\sum_{i=1}^n \left\|\mathbf{g}_i^t - \nabla F_i(\mathbf{x}_i^t)\right\|\right] \leq \mathbb{E}\left[\left(\sum_{i=1}^n \left\|\mathbf{g}_i^t - \nabla F_i(\mathbf{x}_i^t)\right\|^2\right)^{1/2}\right] \leq \left(\mathbb{E}\left[\sum_{i=1}^n \left\|\mathbf{g}_i^t - \nabla F_i(\mathbf{x}_i^t)\right\|^2\right]\right)^{1/2}, \qquad (87)$$

and the result in Lemma 8 we obtain that

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^n \left\|\mathbf{g}_i^t - \nabla F_i(\mathbf{x}_i^t)\right\|\right] &\leq \sqrt{n}\left[\left(1-\frac{1}{2T^{2/3}}\right)^t nG^2 + \frac{2n\sigma^2}{T^{2/3}} + \frac{6nL^2D^2C}{T^{4/3}} + \frac{12nL^2D^2C}{T^{2/3}}\right]^{1/2} \\
&\leq nG\left(1-\frac{1}{2T^{2/3}}\right)^{t/2} + \frac{\sqrt{2}n\sigma}{T^{1/3}} + \frac{\sqrt{6}nLDC^{1/2}}{T^{2/3}} + \frac{\sqrt{12}nLDC^{1/2}}{T^{1/3}}, \qquad (88)
\end{aligned}$$

where the second inequality holds since $\sum_i a_i^2 \leq (\sum_i a_i)^2$ for $a_i \geq 0$. Compute the expected value of both sides of (86) and replace $\mathbb{E}\left[\sum_{i=1}^n \|\mathbf{g}_i^t - \nabla F_i(\mathbf{x}_i^t)\|\right]$ by its upper bound in (88) to obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{i=1}^n \mathbf{d}_i^t - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t)\right\|\right] &\leq (1-\alpha)\mathbb{E}\left[\left\|\sum_{j=1}^n \mathbf{d}_j^{t-1} - \sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^{t-1})\right\|\right] + \frac{(1-\alpha)LnD}{T} + \frac{\alpha LnD}{T(1-\beta)} \\
&\quad + \alpha nG\left(1-\frac{1}{2T^{2/3}}\right)^{t/2} + \frac{\sqrt{6}\alpha nLDC^{1/2}}{T^{2/3}} + \frac{\sqrt{2}n\alpha\sigma + \sqrt{12}\alpha nLDC^{1/2}}{T^{1/3}} \qquad (89)
\end{aligned}$$

By multiplying both of sides of (45) by $1/n$ and applying the resulted inequality recessively for t steps we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{d}_i^t - \frac{1}{n}\sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^t)\right\|\right] \\
&\leq (1-\alpha)^t\left\|\frac{1}{n}\sum_{j=1}^n \mathbf{d}_j^0 - \frac{1}{n}\sum_{i=1}^n \nabla F_i(\bar{\mathbf{x}}^0)\right\| \\
&\quad + \left(\frac{(1-\alpha)LD}{T} + \frac{\alpha LD}{T(1-\beta)} + \alpha G\left(1-\frac{1}{2T^{2/3}}\right)^{t/2} + \frac{\sqrt{6}\alpha LDC^{1/2}}{T^{2/3}} + \frac{\sqrt{2}\alpha\sigma + \sqrt{12}\alpha LDC^{1/2}}{T^{1/3}}\right)\sum_{s=0}^{t-1}(1-\alpha)^s \\
&\leq (1-\alpha)^t\frac{1}{n}\sum_{i=1}^n \left\|\nabla F_i(\bar{\mathbf{x}}^0)\right\| + \frac{(1-\alpha)LD}{\alpha T} + \frac{LD}{T(1-\beta)} + G\left(1-\frac{1}{2T^{2/3}}\right)^{t/2} + \frac{\sqrt{6}LDC^{1/2}}{T^{2/3}} + \frac{\sqrt{2}\sigma + \sqrt{12}LDC^{1/2}}{T^{1/3}} \\
&\leq \left(1-\frac{1}{T^{1/2}}\right)^t G + \frac{(1-\alpha)LD}{\alpha T} + \frac{LD}{T(1-\beta)} + G\left(1-\frac{1}{2T^{2/3}}\right)^{t/2} + \frac{\sqrt{6}LDC^{1/2}}{T^{2/3}} + \frac{\sqrt{2}\sigma + \sqrt{12}LDC^{1/2}}{T^{1/3}}, \qquad (90)
\end{aligned}$$

which follows the claim in (84). ■

Now we can complete the proof of Theorem 2 using the results in Lemmata 6 and 8 as well as the expression in (63). Replace the terms on the right hand side of (63) by their upper bounds in Lemmata 6 and 8 to obtain

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}F_i(\bar{\mathbf{x}}^{t+1})-\frac{1}{n}\sum_{i=1}^{n}F_i(\bar{\mathbf{x}}^{t})\right]\\
&\geq \mathbb{E}\left[\frac{1}{nT}\left[\sum_{i=1}^{n}F_i(\mathbf{x}^{*})-\sum_{i=1}^{n}F_i(\bar{\mathbf{x}}^{t})\right]\right]-\left(1-\frac{1}{T^{1/2}}\right)^{t}\frac{DG}{T}-\frac{(1-\alpha)LD^2}{\alpha T^2}-\frac{LD^2}{T^2(1-\beta)}+\frac{DG}{T}\left(1-\frac{1}{2T^{2/3}}\right)^{t/2}\\
&\quad-\frac{\sqrt{6}LD^2C^{1/2}}{T^{5/3}}-\frac{\sqrt{2}\sigma+\sqrt{12}LD^2C^{1/2}}{T^{4/3}}-\frac{D(\sigma^2+G^2)^{1/2}}{T^{3/2}(1-\beta(1-\alpha))}-\frac{LD^2}{2T^2}
\end{aligned}\tag{91}$$

Regrouping the terms implies that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}F_i(\mathbf{x}^{*})-\frac{1}{n}\sum_{i=1}^{n}F_i(\bar{\mathbf{x}}^{t+1})\right]\\
&\leq \left(1-\frac{1}{T}\right)\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}F_i(\mathbf{x}^{*})-\frac{1}{n}\sum_{i=1}^{n}F_i(\bar{\mathbf{x}}^{t})\right]+\left(1-\frac{1}{T^{1/2}}\right)^{t}\frac{DG}{T}+\frac{(1-\alpha)LD^2}{\alpha T^2}+\frac{LD^2}{T^2(1-\beta)}\\
&\quad+\frac{DG}{T}\left(1-\frac{1}{2T^{2/3}}\right)^{t/2}+\frac{\sqrt{6}LD^2C^{1/2}}{T^{5/3}}+\frac{\sqrt{2}\sigma+\sqrt{12}LD^2C^{1/2}}{T^{4/3}}+\frac{D(\sigma^2+G^2)^{1/2}}{T^{3/2}(1-\beta(1-\alpha))}+\frac{LD^2}{2T^2}
\end{aligned}\tag{92}$$

Now apply the expression in (92) for $t=0,\dots,T-1$ to obtain

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}F_i(\mathbf{x}^{*})-\frac{1}{n}\sum_{i=1}^{n}F_i(\bar{\mathbf{x}}^{T})\right]\\
&\leq \left(1-\frac{1}{T}\right)^{T}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}F_i(\mathbf{x}^{*})-\frac{1}{n}\sum_{i=1}^{n}F_i(\bar{\mathbf{x}}^{0})\right]+\frac{(1-\alpha)LD^2}{\alpha T}\\
&\quad+\frac{LD^2}{T(1-\beta)}+\frac{\sqrt{6}LD^2C^{1/2}}{T^{2/3}}+\frac{\sqrt{2}\sigma+\sqrt{12}LD^2C^{1/2}}{T^{1/3}}\\
&\quad+\frac{D(\sigma^2+G^2)^{1/2}}{T^{1/2}(1-\beta(1-\alpha))}+\frac{LD^2}{2T}+\sum_{t=0}^{T}\left(1-\frac{1}{T^{1/2}}\right)^{t}\frac{DG}{T}+\sum_{t=0}^{T}\frac{DG}{T}\left(1-\frac{1}{2T^{2/3}}\right)^{t/2}\\
&\leq \left(1-\frac{1}{T}\right)^{T}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}F_i(\mathbf{x}^{*})-\frac{1}{n}\sum_{i=1}^{n}F_i(\bar{\mathbf{x}}^{0})\right]+\frac{(1-\alpha)LD^2}{\alpha T}\\
&\quad+\frac{LD^2}{T(1-\beta)}+\frac{\sqrt{6}LD^2C^{1/2}}{T^{2/3}}+\frac{\sqrt{2}\sigma+\sqrt{12}LD^2C^{1/2}}{T^{1/3}}\\
&\quad+\frac{D(\sigma^2+G^2)^{1/2}}{T^{1/2}(1-\beta(1-\alpha))}+\frac{LD^2}{2T}+\frac{DG}{T^{1/2}}+\frac{4DG}{T^{1/3}},
\end{aligned}\tag{93}$$

where in the last inequality we use the inequalities $\sum_{t=0}^{T}\left(1-\frac{1}{2T^{2/3}}\right)^{t/2}\leq\frac{1}{1-(1-\frac{1}{2T^{2/3}})^{1/2}}\leq 4T^{2/3}$ and $\sum_{t=0}^{T}\left(1-\frac{1}{T^{1/2}}\right)^{t}\leq T^{1/2}$. Regrouping the terms and using the inequality $(1-1/T)^T\leq 1/e$ lead to

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}F_i(\bar{\mathbf{x}}^{T})\right]&\geq(1-e^{-1})\frac{1}{n}\sum_{i=1}^{n}F_i(\mathbf{x}^{*})-\frac{LD^2}{T^{1/2}}-\frac{LD^2}{T(1-\beta)}-\frac{\sqrt{6}LD^2C^{1/2}}{T^{2/3}}-\frac{\sqrt{2}\sigma+\sqrt{12}LD^2C^{1/2}}{T^{1/3}}\\
&\quad-\frac{D(\sigma^2+G^2)^{1/2}}{T^{1/2}(1-\beta)}-\frac{LD^2}{2T}-\frac{DG}{T^{1/2}}-\frac{4DG}{T^{1/3}}
\end{aligned}\tag{94}$$

Now using the argument in (61), we can show that the result in (94) implies that for all $j = \mathcal{N}$ it holds

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} F_i({\mathbf{x}_j}^T)\right] \geq (1-e^{-1})\frac{1}{n}\sum_{i=1}^{n} F_i(\mathbf{x}^*) - \frac{LD^2}{T^{1/2}} - \frac{GD+LD^2}{T(1-\beta)} - \frac{\sqrt{6}LD^2C^{1/2}}{T^{2/3}} - \frac{\sqrt{2}\sigma + \sqrt{12}LD^2C^{1/2}}{T^{1/3}} - \frac{D(\sigma^2+G^2)^{1/2}}{T^{1/2}(1-\beta)} - \frac{LD^2}{2T} - \frac{DG}{T^{1/2}} - \frac{4DG}{T^{1/3}}. \tag{95}$$

Since $C := 1 + \frac{2}{(1-\beta)^2}$ it can be shown that $C^{1/2} = (1+\frac{2}{(1-\beta)^2})^{1/2} \leq 1 + \frac{\sqrt{2}}{1-\beta}$. Applying this upper bound into (95) yields the claim in (25).