PROOF OF LEMMA 8

The vector $\overline{\mathbf{w}}^{t+1}$ can be decomposed as in the following.

$$\begin{aligned}
\overline{\mathbf{w}}^{t+1} &= \sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}\mathbf{w}^{L,m,u,t}, \\
&= \sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}(\mathbf{w}^{L-1,m,u,t}-\eta_t\widetilde{\nabla}f_{m,u}(\mathbf{w}^{L-1,m,u,t})), \qquad (38a)\\
&= \mathbf{w}^t-\eta_t\Big[\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}\sum_{l=0}^{L-1}\widetilde{\nabla}f_{m,u}(\mathbf{w}^{l,m,u,t})\Big], \qquad (38b)\\
&= \mathbf{w}^t-\eta_t\Big[\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}\sum_{l=0}^{L-1}(\widetilde{\nabla}f_{m,u}(\mathbf{w}^{l,m,u,t})-\nabla f(\mathbf{b}^{l,t})+\nabla f(\mathbf{b}^{l,t}))\Big],\\
&= \mathbf{w}^t-\eta_t\Big[\mathbf{e}^{0,t}+\nabla f(\mathbf{b}^{0,t})+\cdots+\mathbf{e}^{L-1,t}+\nabla f(\mathbf{b}^{L-1,t})\Big],\\
&= \mathbf{w}^t-\eta_t(\mathbf{g}^t+\mathbf{e}^t). \qquad (38c)
\end{aligned}$$

where (38a) comes from Step 8 of Algorithm 1 and (38b) is due to the fact that $\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}=1$.

From (38c), the following holds

$$\begin{aligned}
\mathbb{E}[\|\overline{\mathbf{w}}^{t+1}-\mathbf{w}^\star\|_2^2] &= \mathbb{E}[\|\mathbf{w}^t-\eta_t\mathbf{g}^t-\mathbf{w}^\star-\eta_t\mathbf{e}^t\|_2^2],\\
&= \mathbb{E}[\|\mathbf{w}^t-\eta_t\mathbf{g}^t-\mathbf{w}^\star\|_2^2]-2\eta_t\langle\mathbf{w}^t-\eta_t\mathbf{g}^t-\mathbf{w}^\star,\mathbf{e}^t\rangle+\eta_t^2\mathbb{E}[\|\mathbf{e}^t\|_2^2],\\
&\le \mathbb{E}[\|\mathbf{w}^t-\eta_t\mathbf{g}^t-\mathbf{w}^\star\|_2^2]+K\eta_t^2\mathbb{E}[\|\mathbf{w}^t-\eta_t\mathbf{g}^t-\mathbf{w}^\star\|_2^2]+\frac{1}{K}\mathbb{E}[\|\mathbf{e}^t\|_2^2]+\eta_t^2\mathbb{E}[\|\mathbf{e}^t\|_2^2], \qquad (39a)\\
&= (1+K\eta_t^2)\mathbb{E}[\|\mathbf{w}^t-\eta_t\mathbf{g}^t-\mathbf{w}^\star\|_2^2]+\Big(\frac{1}{K}+\eta_t^2\Big)\mathbb{E}[\|\mathbf{e}^t\|_2^2],\\
&= (1+K\eta_t^2)\mathbb{E}[\|\mathbf{b}^{L,t}-\mathbf{w}^\star\|_2^2]+\Big(\frac{1}{K}+\eta_t^2\Big)\mathbb{E}[\|\mathbf{e}^t\|_2^2], \qquad (39b)
\end{aligned}$$

where (39a) is due to the inequality $-2\langle\mathbf{x},\mathbf{y}\rangle\le\alpha\|\mathbf{x}\|_2^2+\frac{1}{\alpha}\|\mathbf{y}\|_2^2$ for $\alpha>0$ and (39b) follows from the definition of $\{\mathbf{b}^{l,t}\}_{l=0}^{L}$ in (18).

Next, applying the Cauchy-Schwarz inequality yields

$$\mathbb{E}[\|\mathbf{e}^t\|_2^2] \le L\sum_{l=0}^{L-1}\mathbb{E}[\|\mathbf{e}^{l,t}\|_2^2] = L\mathbb{E}[\|\mathbf{e}^{0,t}\|_2^2] + L\sum_{l=1}^{L-1}\mathbb{E}[\|\mathbf{e}^{l,t}\|_2^2]. \tag{40}$$

We have

$$\begin{aligned}
\mathbb{E}[\|\mathbf{e}^{0,t}\|_2^2] &= \mathbb{E}\left[\left\|\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}[\widetilde{\nabla} f_{m,u}(\mathbf{w}^{0,m,u,t}) - \nabla f(\mathbf{b}^{0,t})]\right\|_2^2\right], \\
&= \mathbb{E}\left[\left\|\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}[\widetilde{\nabla} f_{m,u}(\mathbf{w}^{t}) - \nabla f(\mathbf{w}^{t})]\right\|_2^2\right], \\
&= \mathbb{E}\left[\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}^2\|\widetilde{\nabla} f_{m,u}(\mathbf{w}^{t}) - \nabla f(\mathbf{w}^{t})\|_2^2\right], &\text{(41a)} \\
&\le \zeta\mathbb{E}_{\{\mathcal{C}_m\}}\left[\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}^2\right], &\text{(41b)} \\
&\le \zeta, &\text{(41c)}
\end{aligned}$$

where (41a) is due to the unbiasedness in Assumption 3 and the fact that $\{\widetilde{\nabla} f_{m,u}(\mathbf{w}^t)\}$ are independent, (41b) follows from the boundedness Assumption 3, and (41c) is due to the fact that $\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}^2 \le \sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u} = 1$. Now, the term $\mathbb{E}[\|\mathbf{e}^{l,t}\|_2^2]$ in (40) is rewritten by

$$\begin{aligned}
\mathbb{E}[\|\mathbf{e}^{l,t}\|_2^2] &= \mathbb{E}\left[\left\|\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}\left[\widetilde{\nabla} f_{m,u}(\mathbf{w}^{l,m,u,t}) - \nabla f(\mathbf{b}^{l,t})\right]\right\|_2^2\right], \\
&= \mathbb{E}\left[\left\|\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}\left[\widetilde{\nabla} f_{m,u}(\mathbf{w}^{l,m,u,t}) - \nabla f(\mathbf{w}^{l,m,u,t}) + \nabla f(\mathbf{w}^{l,m,u,t}) - \nabla f(\mathbf{b}^{l,t})\right]\right\|_2^2\right], \\
&= \mathbb{E}\left[\left\|\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}\left[\widetilde{\nabla} f_{m,u}(\mathbf{w}^{l,m,u,t}) - \nabla f(\mathbf{w}^{l,m,u,t})\right]\right\|_2^2\right] \\
&\qquad + \mathbb{E}\left[\left\|\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}\left[\nabla f(\mathbf{w}^{l,m,u,t}) - \nabla f(\mathbf{b}^{l,t})\right]\right\|_2^2\right], \qquad \text{(42a)}
\end{aligned}$$

where (42a) follows from Assumption 3 that $\mathbb{E}[\widetilde{\nabla} f_{m,u}(\mathbf{w}^{l,m,u,t}) - \nabla f(\mathbf{w}^{l,m,u,t})] = \mathbf{0}$. Thus, the following holds

$$\mathbb{E}[\|\mathbf{e}^{l,t}\|_2^2] \leq \zeta\mathbb{E}\left[\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}^2\right] + \mathbb{E}\left[\Big(\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}^2\Big)\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\left\|\nabla f(\mathbf{w}^{l,m,u,t})-\nabla f(\mathbf{b}^{l,t})\right\|_2^2\right], \tag{43a}$$

$$\leq \zeta\mathbb{E}\left[\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}^2\right] + \mathbb{E}\left[\Big(\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m.u}^2\Big)\sum_{m=1}^{M}\sum_{d_{m.u}\in\mathcal{C}_m}\gamma^2\left\|\mathbf{w}^{l,m,u,t}-\mathbf{b}^{l,t}\right\|_2^2\right], \tag{43b}$$

$$\leq \zeta + \gamma^2\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\mathbb{E}\left[\left\|\mathbf{w}^{l,m,u,t}-\mathbf{b}^{l,t}\right\|_2^2\right], \tag{43c}$$

$$= \zeta + K\gamma^2 a^{l,t}, \tag{43d}$$

where (43a) follows from $\mathbb{E}[\|\widetilde{\nabla} f_{m,u}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2] \leq \zeta$, (43b) follows from Assumption 1, (43c) is due to the fact that $\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u} = 1$, and (43d) is due to the definition of $a^{l,t}$ in (19).

From (20b), plugging (41c) and (43d) into (40) leads to

$$\begin{aligned}
\mathbb{E}[\|\mathbf{e}^t\|_2^2] &\leq L\zeta + \sum_{l=1}^{L-1} L\Big(\zeta + K\gamma^2 a^{l,t}\Big),\\
&= L^2\zeta + LK\gamma^2\sum_{l=1}^{L-1}a^{l,t},\\
&\leq L^2\zeta + LK\gamma^2(L-1)\frac{N\eta_t^2L^2\zeta}{K}(1+L\eta_t^2\gamma^2)^L,\\
&\leq L^2\zeta + NL^4\gamma^2\zeta\eta_t^2e^{L^2\eta_t^2\gamma^2}, && (44a)\\
&\leq L^2\zeta + NL^4\gamma^2\zeta\eta_t^2e, && (44b)
\end{aligned}$$

where (44a) is due to the fact that $(1+x)\leq e^x$, for $x\geq 0$, and (44b) follows from $\eta_t \leq \frac{1}{L\gamma}$.

Plugging (44b) into (39b) and using (20a) give

$$\mathbb{E}[\|\overline{\mathbf{w}}^{t+1}-\mathbf{w}^\star\|_2^2] \leq (1+K\eta_t^2)(1-\mu\eta_t)^L\mathbb{E}[\|\mathbf{w}^t-\mathbf{w}^\star\|_2^2] + \left(\frac{1}{K}+\eta_t^2\right)\Big(L^2\zeta + NL^4\gamma^2\zeta\eta_t^2e\Big),$$

which completes the proof.