PROOF OF LEMMA 7

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

$$\begin{aligned}
\overline{\mathbf{w}}^{t+1} &= \sum_{m=1}^{M} \sum_{d_{m,u}\in\mathcal{C}_m} \omega_{m,u} \mathbf{w}^{L,m,u,t}, \\
&= \sum_{m=1}^{M} \sum_{d_{m,u}\in\mathcal{C}_m} \omega_{m,u} (\mathbf{w}^{L-1,m,u,t} - \eta_t \widetilde{\nabla} f_{m,u}(\mathbf{w}^{L-1,m,u,t})), && \text{(39a)} \\
&= \mathbf{w}^t - \eta_t \Big[\sum_{m=1}^{M} \sum_{d_{m,u}\in\mathcal{C}_m} \omega_{m,u} \sum_{l=0}^{L-1} \widetilde{\nabla} f_{m,u}(\mathbf{w}^{l,m,u,t}) \Big], && \text{(39b)} \\
&= \mathbf{w}^t - \eta_t \Big[\sum_{m=1}^{M} \sum_{d_{m,u}\in\mathcal{C}_m} \omega_{m,u} \sum_{l=0}^{L-1} (\widetilde{\nabla} f_{m,u}(\mathbf{w}^{l,m,u,t}) - \nabla f(\mathbf{b}^{l,t}) + \nabla f(\mathbf{b}^{l,t})) \Big], \\
&= \mathbf{w}^t - \eta_t \Big[\mathbf{e}^{0,t} + \nabla f(\mathbf{b}^{0,t}) + \cdots + \mathbf{e}^{L-1,t} + \nabla f(\mathbf{b}^{L-1,t}) \Big], \\
&= \mathbf{w}^t - \eta_t (\mathbf{g}^t + \mathbf{e}^t). && \text{(39c)}
\end{aligned}$$

where (39a) comes from Step 8 of Algorithm 1 and (39b) is due to the fact that $\sum_{m=1}^{M} \sum_{d_{m,u}\in\mathcal{C}_m} \omega_{m,u} = 1$.

From (39c), the following holds

$$\begin{aligned}
\mathbb{E}[\|\overline{\mathbf{w}}^{t+1} - \mathbf{w}^\star\|_2^2] &= \mathbb{E}[\|\mathbf{w}^t - \eta_t \mathbf{g}^t - \mathbf{w}^\star - \eta_t \mathbf{e}^t\|_2^2], \\
&= \mathbb{E}[\|\mathbf{w}^t - \eta_t \mathbf{g}^t - \mathbf{w}^\star\|_2^2] - 2\eta_t \mathbb{E}\left[\langle \mathbf{w}^t - \eta_t \mathbf{g}^t - \mathbf{w}^\star, \mathbf{e}^t \rangle\right] + \eta_t^2 \mathbb{E}[\|\mathbf{e}^t\|_2^2], \\
&\le \mathbb{E}[\|\mathbf{w}^t - \eta_t \mathbf{g}^t - \mathbf{w}^\star\|_2^2] + K\eta_t^2 \mathbb{E}[\|\mathbf{w}^t - \eta_t \mathbf{g}^t - \mathbf{w}^\star\|_2^2] + \frac{1}{K}\mathbb{E}[\|\mathbf{e}^t\|_2^2] + \eta_t^2 \mathbb{E}[\|\mathbf{e}^t\|_2^2], && \text{(40a)} \\
&= (1 + K\eta_t^2) \mathbb{E}[\|\mathbf{w}^t - \eta_t \mathbf{g}^t - \mathbf{w}^\star\|_2^2] + \Big(\frac{1}{K} + \eta_t^2\Big) \mathbb{E}[\|\mathbf{e}^t\|_2^2], \\
&= (1 + K\eta_t^2) \mathbb{E}[\|\mathbf{b}^{L,t} - \mathbf{w}^\star\|_2^2] + \Big(\frac{1}{K} + \eta_t^2\Big) \mathbb{E}[\|\mathbf{e}^t\|_2^2], && \text{(40b)}
\end{aligned}$$

where (40a) 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 (40b) follows from the definition of $\{\mathbf{b}^{l,t}\}_{l=0}^{L}$ in (18).

Next, applying Jensen's inequality to ℓ_2-norm yields

$$\mathbb{E}[\|\mathbf{e}^t\|_2^2] \leq L\sum_{l=0}^{L-1}\mathbb{E}[\|\mathbf{e}^{l,t}\|_2^2] = L\mathbb{E}[\|\mathbf{e}^{0,t}\|_2^2] + L\sum_{l=1}^{L-1}\mathbb{E}[\|\mathbf{e}^{l,t}\|_2^2]. \tag{41}$$

Regarding the first term on the r.h.s. of (41), we have

$$\begin{aligned}
\mathbb{E}[\|\mathbf{e}^{0,t}\|_2^2] &= \mathbb{E}\left[\left\|\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}[\widetilde{\nabla}f_{m,u}(\mathbf{w}^{0,m,u,t}) - \nabla f(\mathbf{b}^{0,t})]\right\|_2^2\right],\\
&= \mathbb{E}\left[\left\|\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}[\widetilde{\nabla}f_{m,u}(\mathbf{w}^{t}) - \nabla f(\mathbf{w}^{t})]\right\|_2^2\right],\\
&= \mathbb{E}\left[\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}^2\|\widetilde{\nabla}f_{m,u}(\mathbf{w}^{t}) - \nabla f(\mathbf{w}^{t})\|_2^2\right], && \text{(42a)}\\
&\leq \zeta\mathbb{E}_{\{\mathcal{C}_m\}_{m=1}^M}\left[\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}^2\right], && \text{(42b)}\\
&\leq \zeta, && \text{(42c)}
\end{aligned}$$

where (42a) is due to the unbiasedness in Assumption 3 and the fact that $\{\widetilde{\nabla}f_{m,u}(\mathbf{w}^t)\}$ are independent, (42b) follows from the boundedness Assumption 3, and (42c) is due to the fact that $\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}^2 \leq 1$. Now, the term $\mathbb{E}[\|\mathbf{e}^{l,t}\|_2^2]$ in (41) is rewritten by

$$\begin{aligned}
\mathbb{E}[\|\mathbf{e}^{l,t}\|_2^2] &= \mathbb{E}\left[\left\|\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}\left[\widetilde{\nabla}f_{m,u}(\mathbf{w}^{l,m,u,t}) - \nabla f(\mathbf{b}^{l,t})\right]\right\|_2^2\right],\\
&= \mathbb{E}\left[\left\|\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}\left[\widetilde{\nabla}f_{m,u}(\mathbf{w}^{l,m,u,t}) - \nabla f(\mathbf{w}^{l,m,u,t}) + \nabla f(\mathbf{w}^{l,m,u,t}) - \nabla f(\mathbf{b}^{l,t})\right]\right\|_2^2\right],\\
&= \mathbb{E}\left[\left\|\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}\left[\widetilde{\nabla}f_{m,u}(\mathbf{w}^{l,m,u,t}) - \nabla f(\mathbf{w}^{l,m,u,t})\right]\right\|_2^2\right]\\
&\quad + \mathbb{E}\left[\left\|\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}\left[\nabla f(\mathbf{w}^{l,m,u,t}) - \nabla f(\mathbf{b}^{l,t})\right]\right\|_2^2\right], && \text{(43a)}
\end{aligned}$$

where (43a) 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

$$\begin{aligned}
\mathbb{E}[\|\mathbf{e}^{l,t}\|_2^2] &\leq \zeta+\mathbb{E}\left[\Big(\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}^2\Big)\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\left\|\nabla f(\mathbf{w}^{l,m,u,t})-\nabla f(\mathbf{b}^{l,t})\right\|_2^2\right], && \text{(44a)}\\
&\leq \zeta+\mathbb{E}\left[\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\left\|\nabla f(\mathbf{w}^{l,m,u,t})-\nabla f(\mathbf{b}^{l,t})\right\|_2^2\right], && \text{(44b)}\\
&\leq \zeta+\mathbb{E}_{\{\mathcal{C}_m\}_{m=1}^{M}}\left[\mathbb{E}\left[\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\gamma^2\left\|\mathbf{w}^{l,m,u,t}-\mathbf{b}^{l,t}\right\|_2^2\Big|\{\mathcal{C}_m\}_{m=1}^{M}\right]\right], && \text{(44c)}\\
&= \zeta+\mathbb{E}_{\{\mathcal{C}_m\}_{m=1}^{M}}\left[\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{G}_m}\mathbb{1}_{\mathcal{C}_m}(d_{m,u})\gamma^2\mathbb{E}\left[\left\|\mathbf{w}^{l,m,u,t}-\mathbf{b}^{l,t}\right\|_2^2\right]\right],\\
&= \zeta+\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{G}_m}\frac{c_m}{g_m}\gamma^2\mathbb{E}\left[\left\|\mathbf{w}^{l,m,u,t}-\mathbf{b}^{l,t}\right\|_2^2\right], && \text{(44d)}\\
&\leq \zeta+\gamma^2\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{G}_m}\mathbb{E}\left[\left\|\mathbf{w}^{l,m,u,t}-\mathbf{b}^{l,t}\right\|_2^2\right], && \text{(44e)}\\
&= \zeta+K\gamma^2a^{l,t}, && \text{(44f)}
\end{aligned}$$

where (44a) follows from $\mathbb{E}[\|\widetilde{\nabla} f_{m,u}(\mathbf{x})-\nabla f(\mathbf{x})\|_2^2]\leq\zeta$ and the result in (11), (44b) is due to the fact that $\sum_{m=1}^{M}\sum_{d_{m,u}\in\mathcal{C}_m}\omega_{m,u}=1$, (44c) follows from Assumption 1, (44d) is due to the fact that $\Pr[d_{m,u}\in\mathcal{C}_m]=\frac{1}{g_m}$, (44e) follows from $\frac{c_m}{g_m}\leq 1$, and (44f) is due to the definition of $a^{l,t}$ in (19).

From (20b), plugging (42c) and (44f) into (41) leads to

$$\begin{aligned}
\mathbb{E}[\|\mathbf{e}^{t}\|_2^2] &\leq L\zeta+\sum_{l=1}^{L-1}L\Big(\zeta+K\gamma^2a^{l,t}\Big),\\
&= L^2\zeta+LK\gamma^2\sum_{l=1}^{L-1}a^{l,t},\\
&\leq L^2\zeta+LK\gamma^2(L-1)\eta_t^2L\zeta(1+L\eta_t^2\gamma^2)^L,\\
&\leq L^2\zeta+KL^3\gamma^2\zeta\eta_t^2e^{L^2\eta_t^2\gamma^2}, && \text{(45a)}\\
&\leq L^2\zeta+KL^3\gamma^2\zeta\eta_t^2e, && \text{(45b)}
\end{aligned}$$

where (45a) is due to the fact that $(1+x)\leq e^x$ for $x\geq 0$, and (45b) follows from $\eta_t\leq\frac{1}{L\gamma}$.

Plugging (45b) into (40b) and using (20a) give

$$\mathbb{E}[\|\overline{\mathbf{w}}^{t+1}-\mathbf{w}^\star\|_2^2]\leq(1+K\eta_t^2)(1-\mu\eta_t)^L\mathbb{E}[\|\mathbf{w}^t-\mathbf{w}^\star\|_2^2]+\Big(\frac{1}{K}+\eta_t^2\Big)\Big(L^2\zeta+KL^3\gamma^2\zeta\eta_t^2e\Big),$$

which completes the proof.