Appendix: Convergence Proofs

The inequality

$$\|\sum_{i=1}^{n} a_i\|^2 \leq n \sum_{i=1}^{n} \|a_i\|^2 \tag{1}$$

is used frequently in our proofs.

Proof of Theorem 1

In QRP, the updating rule for global mode ω_t can be written as

$$\omega_{t+1} = \omega_t - \frac{\eta}{P} \sum_{i=1}^{P} Q(g(\omega_t^i; \xi_t^i)). \tag{2}$$

Then we have

$$\begin{aligned} & F(\omega_{t+1}) - F(\omega_t) \\ \leq & \nabla F(\omega_t)(\omega_{t+1} - \omega_t)^T + \frac{L}{2}\|\omega_{t+1} - \omega_t\|^2 \\ = & \nabla F(\omega_t)(-\frac{\eta}{P}\sum_{i=1}^{P} Q(g(\omega_t^i; \xi_t^i)))^T + \frac{L}{2}\|\frac{\eta}{P}\sum_{i=1}^{P} Q(g(\omega_t^i; \xi_t^i))\|^2 \end{aligned} \tag{3}$$

Taking the expectation for both side with respect to ξ_t, the batch in iteration t, we have

$$\begin{aligned} & \mathbb{E}_{\xi_t}[F(\omega_{t+1})] - F(\omega_t) \\ = & \mathbb{E}_{\xi_t}[\nabla F(\omega_t)(-\frac{\eta}{P}\sum_{i=1}^{P} Q(g(\omega_t^i; \xi_t^i)))^T] + \frac{L}{2}\mathbb{E}_{\xi_t}[\|\frac{\eta}{P}\sum_{i=1}^{P} Q(g(\omega_t^i; \xi_t^i))\|^2] \\ \stackrel{(a)}{=} & \nabla F(\omega_t)(-\frac{\eta}{P}\sum_{i=1}^{P} \nabla F(\omega_t^i; \xi_t^i))^T + \frac{L\eta^2}{2P^2}\mathbb{E}_{\xi_t}[\|\sum_{i=1}^{P} Q(g(\omega_t^i; \xi_t^i))\|^2] \\ = & -\eta\nabla F(\omega_t)(-\frac{1}{P}\sum_{i=1}^{P} \nabla F(\omega_t^i; \xi_t^i))^T + \frac{L\eta^2}{2P^2}\mathbb{E}_{\xi_t}[\|\sum_{i=1}^{P} Q(g(\omega_t^i; \xi_t^i))\|^2] \\ \stackrel{(b)}{\leq} & -\frac{\eta}{2}\|\nabla F(\omega_t)\|^2 - \frac{\eta}{2P^2}\|\sum_{i=1}^{P}\nabla F(\omega_t^i)\|^2 + \frac{\eta}{2}\|\nabla F(\omega_t) - \frac{1}{P}\sum_{i=1}^{P}\nabla F(\omega_t^i)\|^2 + \frac{L\eta^2}{2P^2}\mathbb{E}_{\xi_t}\|\sum_{i=1}^{P} Q(g(\omega_t^i; \xi_t^i))\|^2 \\ = & -\frac{\eta}{2}\|\nabla F(\omega_t)\|^2 - \frac{\eta}{2P^2}\|\sum_{i=1}^{P}\nabla F(\omega_t^i)\|^2 + \frac{\eta}{2P^2}\|\sum_{i=1}^{P}[\nabla F(\omega_t) - \nabla F(\omega_t^i)]\|^2 + \frac{L\eta^2}{2P^2}\mathbb{E}_{\xi_t}\|\sum_{i=1}^{P} Q(g(\omega_t^i; \xi_t^i))\|^2 \\ \stackrel{(c)}{\leq} & -\frac{\eta}{2}\|\nabla F(\omega_t)\|^2 - \frac{\eta}{2P^2}\|\sum_{i=1}^{P}\nabla F(\omega_t^i)\|^2 + \frac{\eta}{2P}\sum_{i=1}^{P}\|\nabla F(\omega_t) - \nabla F(\omega_t^i)\|^2 + \frac{L\eta^2}{2P^2}\mathbb{E}_{\xi_t}\|\sum_{i=1}^{P} Q(g(\omega_t^i; \xi_t^i))\|^2 \end{aligned} \tag{4}$$

where (a) follows according to the unbiasedness of quantization and stochastic gradient, (b) comes from $<\mathbf{a}, \mathbf{b}> = \frac{1}{2}\|\mathbf{a}\|^2 + \frac{1}{2}\|\mathbf{b}\|^2 - \frac{1}{2}\|\mathbf{a} - \mathbf{b}\|^2$, and (c) follows after (1).

First we derive the bound of $\|\nabla F(\omega_t) - \nabla F(\omega_t^i)\|^2$.

For any worker i and iteration t, if worker i last updated the global parameter in iteration $t-k$, where $k = 0, 1, \ldots, t-1$, i.e., $\omega_{t-k}^i = \omega_{t-k}$ then we have the formulation for local parameter and global parameter:

$$\omega_t^i = \omega_{t-k} - \sum_{j=1}^{k} \eta Q(g(\omega_{t-j}^i; \xi_{t-j}^i)) \tag{5}$$

$$\omega_t = \omega_{t-k} - \frac{1}{P}\sum_{j=1}^{k}\sum_{i=1}^{P} \eta Q(g(\omega_{t-j}^i; \xi_{t-j}^i)) \tag{6}$$

Under the Assumption 1, we have that

$$
\begin{aligned}
&\|\nabla F(\omega_t)-\nabla F(\omega_t^i)\|^2\\
\le & 2L^2\left[\|\omega_t-\omega_t^i\|^2\right]\\
= & 2L^2\|\sum_{j=1}^{k}\eta Q(g(\omega_{t-j}^i;\xi_{t-j}^i))-\frac{1}{P}\sum_{j=1}^{k}\sum_{i=1}^{P}\eta Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2\\
\le & 4L^2\|\sum_{j=1}^{k}\eta Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2+\frac{4L^2}{P^2}\|\sum_{j=1}^{k}\sum_{i=1}^{P}\eta Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2\\
= & 4\eta^2L^2\|\sum_{j=1}^{k}Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2+\frac{4\eta^2L^2}{P^2}\|\sum_{j=1}^{k}\sum_{i=1}^{P}Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2
\end{aligned}
\tag{7}
$$

Based on the pulling probability r, worker i last updated the global parameter in iteration $t-k$ with the probability $r(1-r)^k$, for any $k=0,1,\dots,t-1$. Then we can get expectation of $\|\nabla F(\omega_t)-\nabla F(\omega_t^i)\|^2$ with respect to k.

$$
\begin{aligned}
&\mathbb{E}_k\|\nabla F(\omega_t)-\nabla F(\omega_t^i)\|^2\\
\le & 4\eta^2L^2\mathbb{E}_k\|\sum_{j=1}^{k}Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2+\frac{4\eta^2L^2}{P^2}\mathbb{E}_k\|\sum_{j=1}^{k}\sum_{i=1}^{P}Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2\\
= & 4\eta^2L^2\sum_{\ell=0}^{t-1}[\mathbb{P}[k=\ell]\|\sum_{j=1}^{\ell}Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2]+\frac{4\eta^2L^2}{P^2}\sum_{\ell=0}^{t-1}[\mathbb{P}[k=\ell]\|\sum_{j=1}^{\ell}\sum_{i=1}^{P}Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2]\\
= & 4\eta^2L^2\sum_{\ell=0}^{t-1}[r(1-r)^\ell\|\sum_{j=1}^{\ell}Q(g(\omega_{t-j}^i;\xi_{t-j}^i)\|^2)]+\frac{4\eta^2L^2}{P^2}\sum_{\ell=0}^{t-1}[r(1-r)^\ell\|\sum_{j=1}^{\ell}\sum_{i=1}^{P}Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2]\\
\le & 4\eta^2L^2r\sum_{\ell=0}^{t-1}[(1-r)^\ell\ell\sum_{j=1}^{\ell}\|Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2]+\frac{4\eta^2L^2}{P^2}\sum_{\ell=0}^{t-1}[r(1-r)^\ell\ell\sum_{j=1}^{\ell}\|\sum_{i=1}^{P}Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2]\\
\le & 4\eta^2L^2r\sum_{\ell=0}^{t-1}[(1-r)^\ell\ell\sum_{j=1}^{\ell}\|Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2]+\frac{4\eta^2L^2}{P}\sum_{\ell=0}^{t-1}[r(1-r)^\ell\ell\sum_{j=1}^{\ell}\sum_{i=1}^{P}\|Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2]
\end{aligned}
\tag{8}
$$

With respect to quantization, the bound of the expect squared magnitude of quantized gradient can be formulated by non-compression gradient:

$$
\begin{aligned}
&\mathbb{E}\|Q(g(\omega;\xi))\|^2\\
= & \mathbb{E}\|Q(g(\omega;\xi))-g(\omega;\xi)+g(\omega;\xi)\|^2\\
\le & 2\mathbb{E}\|Q(g(\omega;\xi))-g(\omega;\xi)\|^2+2\|g(\omega;\xi)\|^2\\
\le & 2\epsilon^2\|g(\omega;\xi)\|^2+2\|g(\omega;\xi)\|^2\\
\le & 2(1+\epsilon^2)\|g(\omega;\xi)\|^2
\end{aligned}
\tag{9}
$$

Replacing the $\|Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2$ in (8) by the result of (9), we have

$$
\begin{aligned}
&\mathbb{E}_k\|\nabla F(\omega_t)-\nabla F(\omega_t^i)\|^2\\
\le & 8\eta^2L^2r(1+\epsilon^2)\sum_{\ell=0}^{t-1}[(1-r)^\ell\ell\sum_{j=1}^{\ell}\|g(\omega_{t-j}^i;\xi_{t-j}^i)\|^2]+\frac{8\eta^2L^2r(1+\epsilon^2)}{P}\sum_{\ell=0}^{t-1}[(1-r)^\ell\ell\sum_{j=1}^{\ell}\sum_{i=1}^{P}\|g(\omega_{t-j}^i;\xi_{t-j}^i)\|^2]\\
\overset{(a)}{\le} & 16\eta^2L^2rG^2(1+\epsilon^2)\sum_{\ell=0}^{t-1}(1-r)^\ell\ell^2\\
\overset{(b)}{\le} & \frac{16\eta^2L^2G^2(1-r)(2-r)(1+\epsilon^2)}{r^2}
\end{aligned}
\tag{10}
$$

where (a) comes after bounded gradient assumption, and (b) follows according to that $\lim_{t\to\infty}\sum_{\ell=0}^{t-1}(1-r)^\ell\ell^2=\frac{(1-r)(2-r)}{r^3}$.

Then, we derive the bound of $\mathbb{E}_{\xi_t}\|\sum_{i=1}^{P} Q(g(\omega_t^i;\xi_t^i))\|^2$.

$$
\begin{aligned}
&\mathbb{E}_{\xi_t}[\|\sum_{i=1}^{P} Q(g(\omega_t^i;\xi_t^i))\|^2] \\
&= \mathbb{E}_{\xi_t}[\|\sum_{i=1}^{P} Q(g(\omega_t^i;\xi_t^i)) - \sum_{i=1}^{P} g(\omega_t^i;\xi_t^i) + \sum_{i=1}^{P} g(\omega_t^i;\xi_t^i)\|^2] \\
&= \mathbb{E}_{\xi_t}[\|\sum_{i=1}^{P} Q(g(\omega_t^i;\xi_t^i)) - \sum_{i=1}^{P} g(\omega_t^i;\xi_t^i) + \sum_{i=1}^{P} g(\omega_t^i;\xi_t^i) - \sum_{i=1}^{P} \nabla F(\omega_t^i) + \sum_{i=1}^{P} \nabla F(\omega_t^i)\|^2] \\
&\leq 3\mathbb{E}_{\xi_t}[\|\sum_{i=1}^{P} Q(g(\omega_t^i;\xi_t^i)) - \sum_{i=1}^{P} g(\omega_t^i;\xi_t^i)\|^2] + 3\mathbb{E}_{\xi}\|\sum_{i=1}^{P}[g(\omega_t^i;\xi_t^i) - \nabla F(\omega_t^i)]\|^2 + 3\mathbb{E}_{\xi}\|\sum_{i=1}^{P} \nabla F(\omega_t^i)\|^2 \\
&\overset{(a)}{\leq} 3\sum_{i=1}^{P} \mathbb{E}_{\xi_t}[\|Q(g(\omega_t^i;\xi_t^i)) - g(\omega_t^i;\xi_t^i)\|^2] + 3\sum_{i=1}^{P} \mathbb{E}_{\xi}[\|g(\omega_t^i;\xi_t^i) - \nabla F(\omega_t^i)\|^2] + 3\mathbb{E}_{\xi}\|\sum_{i=1}^{P} \nabla F(\omega_t^i)\|^2 \\
&\overset{(b)}{\leq} 3P\epsilon^2\|g(\omega_t^i;\xi_t^i)\|^2 + 3P\sigma^2 + 3\mathbb{E}_{\xi}\|\sum_{i=1}^{P} \nabla F(\omega_t^i)\|^2 \\
&\leq 3PG^2\epsilon^2 + 3P\sigma^2 + 3\mathbb{E}_{\xi}\|\sum_{i=1}^{P} \nabla F(\omega_t^i)\|^2
\end{aligned}
\tag{11}
$$

Based on (10) and (11), and replacing $\mathbb{E}_k\|\nabla F(\omega_t) - \nabla F(\omega_t^i)\|^2$ and $\mathbb{E}_{\xi_t}[\|\sum_{i=1}^{P} Q(g(\omega_t^i;\xi_t^i))\|^2]$ in (4), we have

$$
\begin{aligned}
&\mathbb{E}_{\xi_t}[F(\omega_{t+1})] - F(\omega_t) \\
\leq& -\frac{\eta}{2}\|\nabla F(\omega_t)\|^2 - \frac{\eta}{2P^2}\|\sum_{i=1}^{P} \nabla F(\omega_t^i)\|^2 + \frac{\eta}{2P}\sum_{i=1}^{P}\|\nabla F(\omega_t) - \nabla F(\omega_t^i)\|^2 + \frac{L\eta^2}{2P^2}\mathbb{E}_{\xi_t}\|\sum_{i=1}^{P} Q(g(\omega_t^i;\xi_t^i))\|^2 \\
=& -\frac{\eta}{2}\|\nabla F(\omega_t)\|^2 - \frac{\eta}{2P^2}\|\sum_{i=1}^{P} \nabla F(\omega_t^i)\|^2 + \frac{\eta}{2P}\sum_{i=1}^{P}[\frac{16\eta^2L^2G^2(1-r)(2-r)(1+\epsilon^2)}{r^2}] + \frac{L\eta^2}{2P^2}[3PG^2\epsilon^2 + 3P\sigma^2 + 3\mathbb{E}_{\xi}\|\sum_{i=1}^{P} \nabla F(\omega_t^i)\|^2] \\
=& -\frac{\eta}{2}\|\nabla F(\omega_t)\|^2 + \frac{3L\eta^2 - \eta}{2P^2}\|\sum_{i=1}^{P} \nabla F(\omega_t^i)\|^2 + \frac{8\eta^3L^2G^2(1-r)(2-r)(1+\epsilon^2)}{r^2} + \frac{3L\eta^2(G^2\epsilon^2 + \sigma^2)}{2P}
\end{aligned}
\tag{12}
$$

Let $\frac{3L\eta^2 - \eta}{2P^2} \leq 0$, i.e., $\eta \leq \frac{1}{3L}$, we get

$$
\mathbb{E}_{\xi_t}[F(\omega_{t+1})] \leq F(\omega_t) - \frac{\eta}{2}\|\nabla F(\omega_t)\|^2 + \frac{8\eta^3L^2G^2(1-r)(2-r)(1+\epsilon^2)}{r^2} + \frac{3L\eta^2(G^2\epsilon^2 + \sigma^2)}{2P}
\tag{13}
$$

Since the objective $F(\cdot)$ is μ-strongly convex, we can bound the optimality gap at any given point in terms of the squared L_2 norm of the gradient as follows.

$$
\|\nabla F(\omega_t)\|^2 \geq 2\mu[F(\omega_t) - F(\omega^*)]
\tag{14}
$$

where ω^* is the optimal solution for $F(\cdot)$.

Therefore,

$$
\mathbb{E}_{\xi_t}[F(\omega_{t+1})] \leq F(\omega_t) - \eta\mu[F(\omega_t) - F(\omega^*)] + \frac{8\eta^3L^2G^2(1-r)(2-r)(1+\epsilon^2)}{r^2} + \frac{3L\eta^2(G^2\epsilon^2 + \sigma^2)}{2P}
\tag{15}
$$

Subtracting $F(\omega^*)$ from both sides and takeing total expectation for both sides, this yields that:

$$
\mathbb{E}[F(\omega_{t+1}) - F(\omega^*)] \leq (1-\eta\mu)\mathbb{E}[F(\omega_t) - F(\omega^*)] + \frac{8\eta^3L^2G^2(1-r)(2-r)(1+\epsilon^2)}{r^2} + \frac{3L\eta^2(G^2\epsilon^2 + \sigma^2)}{2P}
\tag{16}
$$

Subtracting $\frac{8\eta^2L^2G^2(1-r)(2-r)(1+\epsilon^2)}{\mu r^2} + \frac{3L\eta(G^2\epsilon^2 + \sigma^2)}{2P\mu}$ from both sides and rearranging yield that

$$
\begin{aligned}
&\mathbb{E}[F(\omega_{t+1}) - F(\omega^*) - \frac{3L\eta(G^2\epsilon^2 + \sigma^2)}{2P\mu} - \frac{8\eta^2L^2G^2(1-r)(2-r)(1+\epsilon^2)}{\mu r^2}] \\
\leq& (1-\eta\mu)\mathbb{E}[F(\omega_t) - F(\omega^*) - \frac{3L\eta(G^2\epsilon^2 + \sigma^2)}{2P\mu} - \frac{8\eta^2L^2G^2(1-r)(2-r)(1+\epsilon^2)}{\mu r^2}]
\end{aligned}
\tag{17}
$$

Applying (17) repeatedly for iteration 1 to $t-1$, we have

$$\begin{aligned}&\mathbb{E}[F(\omega_{t+1}) - F(\omega^*) - \frac{3L\eta(G^2\epsilon^2+\sigma^2)}{2P\mu} - \frac{8\eta^2L^2G^2(1-r)(2-r)(1+\epsilon^2)}{\mu r^2}] \\ \leq & (1-\eta\mu)^t[F(\omega_1) - F(\omega^*) - \frac{3L\eta(G^2\epsilon^2+\sigma^2)}{2P\mu} - \frac{8\eta^2L^2G^2(1-r)(2-r)(1+\epsilon^2)}{\mu r^2}]\end{aligned} \tag{18}$$

which is

$$\begin{aligned}&\mathbb{E}F(\omega_{t+1}) - F(\omega^*) \\ \leq & (1-\eta\mu)^t[F(\omega_1) - F(\omega^*) - \frac{3L\eta(G^2\epsilon^2+\sigma^2)}{2P\mu} - \frac{8\eta^2L^2G^2(1-r)(2-r)(1+\epsilon^2)}{\mu r^2}] + \frac{3L\eta(G^2\epsilon^2+\sigma^2)}{2P\mu} \\ & + \frac{8\eta^2L^2G^2(1-r)(2-r)(1+\epsilon^2)}{\mu r^2}]\end{aligned} \tag{19}$$

The proof is completed.

Proof of Theorem 2

In non-convex case, according to (4), we have

$$\begin{aligned}&\mathbb{E}_\xi[F(\omega_{t+1}) - F(\omega_t)] \\ \leq & -\frac{\eta}{2}\mathbb{E}_\xi\|\nabla F(\omega_t)\|^2 - \frac{\eta}{2P^2}\mathbb{E}_\xi\|\sum_{i=1}^P \nabla F(\omega_t^i)\|^2 + \frac{\eta}{2P}\sum_{i=1}^P \mathbb{E}_\xi\|\nabla F(\omega_t) - \nabla F(\omega_t^i)\|^2 + \frac{L\eta^2}{2P^2}\mathbb{E}_\xi\|\sum_{i=1}^P Q(g(\omega_t^i;\xi_t^i))\|^2\end{aligned} \tag{20}$$

The main difference is that we have to taking the expectation for both side with respect to ξ, the whole stochastic batch space in all iteration $1, 2, \ldots, t$.

Summing this inequality for iteration $1, 2, \ldots, T$ for both sides, we have

$$\begin{aligned}&\mathbb{E}_\xi[F(\omega_{T+1}) - F(\omega_1)] \\ \leq & -\frac{\eta}{2}\sum_{t=1}^T \mathbb{E}_\xi\|\nabla F(\omega_t)\|^2 - \frac{\eta}{2P^2}\sum_{t=1}^T \mathbb{E}_\xi\|\sum_{i=1}^P \nabla F(\omega_t^i)\|^2 + \frac{\eta}{2P}\sum_{i=1}^P\sum_{t=1}^T \mathbb{E}_\xi\|\nabla F(\omega_t) - \nabla F(\omega_t^i)\|^2 + \frac{L\eta^2}{2P^2}\sum_{t=1}^T \mathbb{E}_\xi\|\sum_{i=1}^P Q(g(\omega_t^i;\xi_t^i))\|^2 \\ \leq & -\frac{\eta}{2}\sum_{t=1}^T \mathbb{E}_\xi\|\nabla F(\omega_t)\|^2 - \frac{\eta - 3L\eta^2}{2P^2}\sum_{t=1}^T \mathbb{E}_\xi\|\sum_{i=1}^P \nabla F(\omega_t^i)\|^2 + \frac{\eta}{2P}\sum_{i=1}^P\sum_{t=1}^T \mathbb{E}_\xi\|\nabla F(\omega_t) - \nabla F(\omega_t^i)\|^2 + \frac{3TL\eta^2(G^2\epsilon^2+\sigma^2)}{2P}\end{aligned} \tag{21}$$

where the last inequality follows according to (11) which also holds by taking the expectation with respect to ξ.

Different from strongly convex case, $\mathbb{E}_\xi\|\nabla F(\omega_t) - \nabla F(\omega_t^i)\|^2$ are summarized for all iteration and such that we can derive a tighter bound. According to (8), we have

$$
\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}_\xi\|\nabla F(\omega_t)-\nabla F(\omega_t^i)\|^2\\
\le\, & 4\eta^2L^2r\sum_{t=1}^{T}\sum_{\ell=0}^{t-1}[(1-r)^\ell\ell\sum_{j=1}^{\ell}\|Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2]+\frac{4\eta^2L^2}{P^2}\sum_{t=1}^{T}\sum_{\ell=0}^{t-1}[r(1-r)^\ell\ell\sum_{j=1}^{\ell}\|\sum_{i=1}^{P}Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2]\\
\le\, & \frac{8T\eta^2L^2G^2(1-r)(2-r)(1+\epsilon^2)}{r^2}+\frac{4\eta^2L^2}{P^2}\sum_{t=1}^{T}\sum_{\ell=0}^{t-1}[r(1-r)^\ell\ell\sum_{j=1}^{\ell}\|\sum_{i=1}^{P}Q(g(\omega_{t-j}^i;\xi_{t-j}^i))\|^2]\\
\le\, & \frac{8T\eta^2L^2G^2(1-r)(2-r)(1+\epsilon^2)}{r^2}+\frac{4\eta^2L^2}{P^2}\sum_{t=1}^{T}\sum_{\ell=0}^{t-1}\left[r(1-r)^\ell\ell\sum_{j=1}^{\ell}[3PG^2\epsilon^2+3P\sigma^2+3\mathbb{E}_\xi\|\sum_{i=1}^{P}\nabla F(\omega_{t-j}^i)\|^2]\right]\\
=\, & \frac{8T\eta^2L^2G^2(1-r)(2-r)(1+\epsilon^2)}{r^2}+\frac{4\eta^2L^2}{P^2}\sum_{t=1}^{T}\sum_{\ell=0}^{t-1}\left[r(1-r)^\ell\ell^2(3PG^2\epsilon^2+3P\sigma^2)\right]\\
&+\frac{12\eta^2L^2}{P^2}\mathbb{E}_\xi\sum_{t=1}^{T}\sum_{\ell=0}^{t-1}r(1-r)^\ell\ell\sum_{j=1}^{\ell}\|\sum_{i=1}^{P}\nabla F(\omega_{t-j}^i)\|^2]]\\
=\, & \frac{8T\eta^2L^2G^2(1-r)(2-r)(1+\epsilon^2)}{r^2}+\frac{12T\eta^2L^2(1-r)(2-r)(G^2\epsilon^2+\sigma^2)}{Pr^2}+\frac{12\eta^2L^2}{P^2}\mathbb{E}_\xi\sum_{t=1}^{T}\sum_{\ell=0}^{t-1}r(1-r)^\ell\ell\sum_{j=1}^{\ell}\|\sum_{i=1}^{P}\nabla F(\omega_{t-j}^i)\|^2\\
\overset{(a)}{\le}\, & \frac{8T\eta^2L^2G^2(1-r)(2-r)(1+\epsilon^2)}{r^2}+\frac{12T\eta^2L^2(1-r)(2-r)(G^2\epsilon^2+\sigma^2)}{Pr^2}\\
&+\frac{12\eta^2L^2}{P^2}\sum_{t=1}^{T-1}\sum_{m=1}^{t-1}[(1-r)^{t-m}(t-m)^2\mathbb{E}_\xi\|\sum_{i=1}^{P}\nabla F(\omega_t^i;\xi_t^i)\|^2]\\
\le\, & \frac{8T\eta^2L^2G^2(1-r)(2-r)(1+\epsilon^2)}{r^2}+\frac{12T\eta^2L^2(1-r)(2-r)(G^2\epsilon^2+\sigma^2)}{Pr^2}+\frac{12\eta^2L^2(1-r)(2-r)}{P^2r^2}\mathbb{E}_\xi\|\sum_{i=1}^{P}\nabla F(\omega_t^i;\xi_t^i)\|^2]
\end{aligned}
\tag{22}
$$

where (a) follows according to that for each ℓ, the part $(1-r)^{t-m}(t-m)\mathbb{E}_\xi\|\sum_{i=1}^{P}g(\omega_\ell^i;\xi_\ell^i)\|^2$ occurs at most $t-m$ times, considering the constraints of $m\le\ell$ and $\ell<t$.

Therefore, we have

$$
\begin{aligned}
&\mathbb{E}_\xi[F(\omega_{T+1})-F(\omega_1)]\\
\le & -\frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}_\xi\|\nabla F(\omega_t)\|^2-\frac{\eta-3L^2\eta^2}{2P^2}\sum_{t=1}^{T}\mathbb{E}_\xi\|\sum_{i=1}^{P}\nabla F(\omega_t^i)\|^2+\frac{\eta}{2P}\sum_{i=1}^{P}\sum_{t=1}^{T}\mathbb{E}_\xi\|\nabla F(\omega_t)-\nabla F(\omega_t^i)\|^2+\frac{3TL\eta^2(G^2\epsilon^2+\sigma^2)}{2P}\\
\le & -\frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}_\xi\|\nabla F(\omega_t)\|^2-\frac{r^2\eta-3r^2L^2\eta^2-12\eta^3L^2(1-r)(2-r)}{2P^2}\sum_{t=1}^{T}\mathbb{E}_\xi\|\sum_{i=1}^{P}\nabla F(\omega_t^i)\|^2+\frac{3TL\eta^2(G^2\epsilon^2+\sigma^2)}{2P}\\
&+\frac{4T\eta^3L^2G^2(1-r)(2-r)(1+\epsilon^2)}{r^2}+\frac{6T\eta^3L^2(1-r)(2-r)(G^2\epsilon^2+\sigma^2)}{Pr^2}
\end{aligned}
\tag{23}
$$

Let

$$
r^2\eta-3r^2L^2\eta^2-12\eta^3L^2(1-r)(2-r)\ge 0, \tag{24}
$$

we have

$$
\begin{aligned}
&\mathbb{E}_\xi[F(\omega_{T+1})-F(\omega_1)]\\
\le & -\frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}_\xi\|\nabla F(\omega_t)\|^2+\frac{3TL\eta^2(G^2\epsilon^2+\sigma^2)}{2P}+\frac{4T\eta^3L^2G^2(1-r)(2-r)(1+\epsilon^2)}{r^2}+\frac{6T\eta^3L^2(1-r)(2-r)(G^2\epsilon^2+\sigma^2)}{Pr^2}
\end{aligned}
\tag{25}
$$

This immediately yields

$$\begin{aligned}&\frac{1}{T}\mathbb{E}_{\xi}\|\nabla F(\omega_t)\|^2\\ \le&\frac{2|F(\omega_1-\omega_*)|}{\eta T}+\frac{3L\eta(G^2\epsilon^2+\sigma^2)}{P}+\frac{8\eta^2L^2G^2(1-r)(2-r)(1+\epsilon^2)}{r^2}+\frac{12\eta^2L^2(1-r)(2-r)(G^2\epsilon^2+\sigma^2)}{Pr^2}\end{aligned} \tag{26}$$

which completes the proof.

Proof of Theorem 3

By setting the learning rate as

$$\eta=\sqrt{\frac{2|F(\boldsymbol{\omega}_1)-F(\boldsymbol{\omega}^*)|P}{3L(\epsilon^2G^2+\sigma^2)T}} \tag{27}$$

we have

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla F(\boldsymbol{\omega}_t)\right\|^2\\ \le&2\sqrt{\frac{6[F(\boldsymbol{\omega}_1)-F(\boldsymbol{\omega}^*)]L(\epsilon^2G^2+\sigma^2)}{P}}*\frac{1}{\sqrt{T}}\\ &+\frac{16[F(\boldsymbol{\omega}_1)-F(\boldsymbol{\omega}^*)]PLG^2(1-r)(2-r)(1+\epsilon^2)}{3(\epsilon^2G^2+\sigma^2)r^2}*\frac{1}{T}\\ &+\frac{24[F(\boldsymbol{\omega}_1)-F(\boldsymbol{\omega}^*)]L(1-r)(2-r)}{3r^2}*\frac{1}{T}\end{aligned} \tag{28}$$

Combining with the constraint of stepsize in (24), we can derive the condition of T

$$T\ge[R^+(h(\eta)=0)]^{-2}\frac{2|F(\boldsymbol{\omega}_1)-F(\boldsymbol{\omega}^*)|P}{3L(\epsilon^2G^2+\sigma^2)T} \tag{29}$$

where $R^+(h(\eta)=0)$ denotes the positive root of $h(\eta)=0$ and $h(\eta)=[r^2\eta-3r^2L^2\eta^2-12\eta^3L^2(1-r)(2-r)]$.