8. Proof for Lemma 2

Proof. By re-organizing the update rule of accumulated quantization error, we have:

$$\begin{aligned} \mathbf{h}_p^{(t)} &= \beta \mathbf{h}_p^{(t-1)} + (\mathbf{g}_p^{(t-1)} - \tilde{\mathbf{g}}_p^{(t-1)}) \\ &= \beta \mathbf{h}_p^{(t-1)} + (-\alpha \mathbf{h}_p^{(t-1)} - \boldsymbol{\varepsilon}_p^{(t-1)}) \\ &= (\beta - \alpha) \cdot \mathbf{h}_p^{(t-1)} - \boldsymbol{\varepsilon}_p^{(t-1)} \\ &= -\sum_{t'=0}^{t-1} (\beta - \alpha)^{t-t'-1} \cdot \boldsymbol{\varepsilon}_p^{(t')} \end{aligned} \tag{27}$$

which indicates that $\mathbf{h}_p^{(t)}$ is the linear combination of all the previous quantization errors.

Taking the expectation of squared l_2-norm of both sides of the second to the last equality in (27), we have:

$$\begin{aligned} \mathbb{E}\|\mathbf{h}_p^{(t)}\|_2^2 &= \mathbb{E}\|(\beta - \alpha) \cdot \mathbf{h}_p^{(t-1)} - \boldsymbol{\varepsilon}_p^{(t-1)}\|_2^2 \\ &= (\beta - \alpha)^2 \cdot \mathbb{E}\|\mathbf{h}_p^{(t-1)}\|_2^2 + \mathbb{E}\|\boldsymbol{\varepsilon}_p^{(t-1)}\|_2^2 \end{aligned} \tag{28}$$

and the last equality holds due to the independence between $\mathbf{h}_p^{(t-1)}$ and $\boldsymbol{\varepsilon}_p^{(t-1)}$ (recall that all quantization errors are *i.i.d.* random noises).

Since the quantization error $\boldsymbol{\varepsilon}_p^{(t-1)}$ have the following variance bound (from Theorem 1):

$$\begin{aligned} \mathbb{E}\|\boldsymbol{\varepsilon}_p^{(t-1)}\|_2^2 &\leq \gamma \cdot \mathbb{E}\|\mathbf{g}_p^{(t-1)} + \alpha \mathbf{h}_p^{(t-1)}\|_2^2 \\ &= \gamma \cdot \mathbb{E}\|\mathbf{g}_p^{(t-1)}\|_2^2 + \alpha^2 \gamma \cdot \mathbb{E}\|\mathbf{h}_p^{(t-1)}\|_2^2 \\ &\leq \gamma B + \alpha^2 \gamma \cdot \mathbb{E}\|\mathbf{h}_p^{(t-1)}\|_2^2 \end{aligned} \tag{29}$$

where the second equality is also derived from the independence between $\mathbf{g}_p^{(t-1)}$ and $\mathbf{h}_p^{(t-1)}$.

By substituting (29) into (28), we have:

$$\begin{aligned} \mathbb{E}\|\mathbf{h}_p^{(t)}\|_2^2 &\leq [\alpha^2 \gamma + (\beta - \alpha)^2] \cdot \mathbb{E}\|\mathbf{h}_p^{(t-1)}\|_2^2 + \gamma B \\ &\leq \sum_{t'=0}^{t-1} [\alpha^2 \gamma + (\beta - \alpha)^2]^{t-t'-1} \cdot \gamma B \\ &= \frac{1 - \lambda^t}{1 - \lambda} \cdot \gamma B \end{aligned} \tag{30}$$

where $\lambda = \alpha^2 \gamma + (\beta - \alpha)^2$.

By substituting (30) into the variance bound of quantization error at the t-th iteration, we have:

$$\begin{aligned} \mathbb{E}\|\boldsymbol{\varepsilon}_p^{(t)}\|_2^2 &\leq \gamma B + \alpha^2 \gamma \cdot \mathbb{E}\|\mathbf{h}_p^{(t)}\|_2^2 \\ &\leq \left[1 + \alpha^2 \gamma \cdot \frac{1 - \lambda^t}{1 - \lambda}\right] \cdot \gamma B \end{aligned} \tag{31}$$

which completes the proof. □

9. Proof for Theorem 1

Proof. Recall the update rule in ECQ-SGD:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta(\mathbf{A}\mathbf{w}^{(t)} + \mathbf{b} + \boldsymbol{\xi}^{(t)} + \alpha \mathbf{h}^{(t)} + \boldsymbol{\varepsilon}^{(t)}) \tag{32}$$

By applying the optimality of $\mathbf{w}^*$, and subtracting $\mathbf{w}^*$ from both sides of the above equality, we arrive at (note that we introduce $\mathbf{H} = \mathbf{I} - \eta \mathbf{A}$ for simplicity):

$$\begin{aligned} \mathbf{w}^{(t+1)} - \mathbf{w}^* &= \mathbf{H}(\mathbf{w}^{(t)} - \mathbf{w}^*) - \eta(\boldsymbol{\xi}^{(t)} + \alpha \mathbf{h}^{(t)} + \boldsymbol{\varepsilon}^{(t)}) \\ &= \mathbf{H}^{t+1}(\mathbf{w}^{(0)} - \mathbf{w}^*) - \boldsymbol{\Psi}^{(t)} - \boldsymbol{\Phi}^{(t)} \end{aligned} \tag{33}$$

where:

$$\begin{aligned} \mathbf{\Psi}^{(t)} &= \eta \sum_{t'=0}^{t} \mathbf{H}^{t-t'} \boldsymbol{\xi}^{(t')} \\ \mathbf{\Phi}^{(t)} &= \eta \sum_{t'=0}^{t} \mathbf{H}^{t-t'} (\alpha \mathbf{h}^{(t')} + \boldsymbol{\varepsilon}^{(t')}) \end{aligned} \tag{34}$$

Since each accumulated quantization error $\mathbf{h}^{(t')}$ is the linear combination of all previous quantization errors:

$$\mathbf{h}^{(t')} = \sum_{t''=0}^{t'-1} (\beta - \alpha)^{t'-t''-1} \cdot \boldsymbol{\varepsilon}^{(t'')} \tag{35}$$

we can further simplify $\mathbf{\Phi}^{(t)}$ as:

$$\mathbf{\Phi}^{(t)} = \eta \sum_{t'=0}^{t} \mathbf{\Theta}^{(t')} \boldsymbol{\varepsilon}^{(t')} \tag{36}$$

where:

$$\mathbf{\Theta}^{(t')} = \mathbf{H}^{t-t'} - \sum_{t''=t'+1}^{t} \alpha (\beta - \alpha)^{t''-t'-1} \mathbf{H}^{t-t''} \tag{37}$$

Due to the independence between all the random noises ($\{\boldsymbol{\xi}^{(t')}\}$ and $\{\boldsymbol{\varepsilon}^{(t')}\}$), the expectation of squared Euclidean distance between $\mathbf{w}^{(t+1)}$ and $\mathbf{w}^*$ is bounded by:

$$\begin{aligned} \mathbb{E}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 &= \mathbb{E}\|\mathbf{H}^{t+1}(\mathbf{w}^{(0)} - \mathbf{w}^*)\|_2^2 + \eta^2 \sum_{t'=0}^{t} \left[\mathbb{E}\|\mathbf{H}^{t-t'} \boldsymbol{\xi}^{(t')}\|_2^2 + \mathbb{E}\|\mathbf{\Theta}^{(t')} \boldsymbol{\varepsilon}^{(t')}\|_2^2 \right] \\ &\leq R^2 \|\mathbf{H}^{t+1}\|_2^2 + \eta^2 \sigma^2 \sum_{t'=0}^{t} \|\mathbf{H}^{t'}\|_2^2 + \eta^2 \mathbb{E}\|\boldsymbol{\varepsilon}^{(t)}\|_2^2 + \eta^2 \sum_{t'=0}^{t-1} \|\mathbf{\Theta}^{(t')}\|_2^2 \cdot \mathbb{E}\|\boldsymbol{\varepsilon}^{(t')}\|_2^2 \end{aligned} \tag{38}$$

which completes the proof. □

10. Proof for Lemma 3

Proof. Since $\mathbf{A} \succeq a_1 \mathbf{I}$, and the learning rate satisfies $\eta a_1 < 1$, we have $\mathbf{I} - \eta \mathbf{A} \preceq (1 - \eta a_1)\mathbf{I}$, which implies that $(\mathbf{I} - \eta \mathbf{A})^{t''} \preceq (1 - \eta a_1)^{t''} \mathbf{I}$ holds for any positive integer t''. Therefore, we can derive the following inequality:

$$\begin{aligned} (\mathbf{I} - \eta \mathbf{A})^{t} &= (\mathbf{I} - \eta \mathbf{A})^{t-t'} (\mathbf{I} - \eta \mathbf{A})^{t'} \\ &\preceq (1 - \eta a_1)^{t-t'} (\mathbf{I} - \eta \mathbf{A})^{t'} \end{aligned} \tag{39}$$

By substituting the above inequality into the definition of $\mathbf{\Theta}^{(t')}$ ($t' < t$), we arrive at:

$$\begin{aligned} \mathbf{\Theta}^{(t')} &\preceq \left[1 - \sum_{t''=t'+1}^{t} \frac{\alpha (\beta - \alpha)^{t''-t'-1}}{(1 - \eta a_1)^{t''-t'}} \right] \cdot (\mathbf{I} - \eta \mathbf{A})^{t-t'} \\ &= \left[1 - \frac{\alpha}{\beta - \alpha} \sum_{t''=1}^{t-t'} \left(\frac{\beta - \alpha}{1 - \eta a_1} \right)^{t''} \right] \cdot (\mathbf{I} - \eta \mathbf{A})^{t-t'} \\ &= \left[1 - \frac{\alpha}{1 - \eta a_1} \frac{1 - \nu^{t-t'}}{1 - \nu} \right] \cdot (\mathbf{I} - \eta \mathbf{A})^{t-t'} \end{aligned} \tag{40}$$

where $\nu = (\beta - \alpha)/(1 - \eta a_1)$. □

11. Proof for Lemma 4

Proof. Here we use $\Delta t = t - t'$ to denote the time gap. With $\beta = 1 - \eta a_1$ and $0 < \alpha < \beta$, we have $\nu = \frac{\beta - \alpha}{1 - \eta a_1} \in (0, 1)$, which leads to:

$$\lim_{\Delta t \to \infty} \nu^{\Delta t} = 0 \tag{41}$$

Recall that the upper bound of reduction ratio is given by:

$$\frac{\tau^{(t-\Delta t)}}{\tau_{QSGD}^{(t-\Delta t)}} < \left(1 - \frac{\alpha}{1 - \eta a_1} \cdot \frac{1 - \nu^{\Delta t}}{1 - \nu}\right)^2 \cdot \left(1 + \frac{\alpha^2 \gamma}{1 - \lambda}\right) \tag{42}$$

and substituting $\beta = 1 - \eta a_1$ into it, we arrive at:

$$\lim_{\Delta t \to \infty} \frac{\tau^{(t-\Delta t)}}{\tau_{QSGD}^{(t-\Delta t)}} = \left(1 - \frac{\alpha}{1 - \eta a_1} \cdot \frac{1}{1 - \frac{1 - \eta a_1 - \alpha}{1 - \eta a_1}}\right)^2 \cdot \left(1 + \frac{\alpha^2 \gamma}{1 - \lambda}\right) = 0 \tag{43}$$

which completes the proof. □