A APPENDIX

A.1   Proof of Lemma 1.

In order to demonstrate the convergence of our update scheme, we make the following assumptions:

- Lipschitzian gradient: $f(x)$ is L-Lipschitz smooth, i.e., $||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$, $\forall x, y \in \mathbb{R}^d$.
- Bounded variance: the variance of the stochastic gradient, denoted by ξ, is bounded such that $\mathbb{E}[||\xi||^2] \leq \sigma^2$.

Due to the L-Lipschitz condition satisfied by f(x), the derivative of the formula $||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$ with respect to x yields $\nabla^2 f(x) \leq L$. Therefore, we can obtain the second-order Taylor expansion of f(x) around $\nabla^2 f(x) \leq L$.

$$f(x) \leq f(y) + \nabla f(y)^T(y - x) + \frac{L}{2}||y - x||^2 \tag{1}$$

Therefore, we can deduce the following proof.

$$\mathbb{E}f(x_{t+1}) - \mathbb{E}f(x_t)$$

$$\leq \mathbb{E}\langle x_{t+1} - x_t, \nabla f(x_t)\rangle + \frac{L}{2}\mathbb{E}||x_{t+1} - x_t||^2$$

$$= \mathbb{E}\langle -\gamma\nabla g(x_t), \nabla g(x_t) + p\xi_t\rangle + \frac{L\gamma^2}{2}\mathbb{E}||\nabla g(x_t) + p\xi_t||^2 \quad (\nabla g(x) = \nabla f(x) - \xi \; and \; x_{t+1} = x_t - \gamma\nabla g(x_t))$$

$$\leq -\gamma\mathbb{E}\langle\nabla g(x_t), \nabla g(x_t) + p\xi_t\rangle + \frac{L\gamma^2}{2}\mathbb{E}||\nabla g(x_t) + p\xi_t||^2$$

$$\leq -\gamma\mathbb{E}||\nabla g(x_t)||^2 + \frac{L\gamma^2}{2}p^2\mathbb{E}||\nabla g(x_t)||^2 + \frac{L\gamma^2}{2}p^2\mathbb{E}\xi_t^2 \quad (\mathbb{E}[\xi] = 0)$$

$$\leq (-\gamma + \frac{L\gamma^2}{2})\mathbb{E}||\nabla g(x_t)||^2 + \frac{L\gamma^2}{2}p^2\sigma^2 \quad (\mathbb{E}[||\xi||^2] \leq \sigma^2)$$

Using the method of telescoping sum, we can obtain the following equations.

$$\mathbb{E}f(x_t) - \mathbb{E}f(x_0) \leq (\frac{L\gamma^2}{2} - \gamma)\sum_{t=0}^{t-1}\mathbb{E}||\nabla g(x_t)||^2 + \frac{L\gamma^2}{2}p^2\sigma^2 T$$

$$(\gamma - \frac{L\gamma^2}{2})\sum_{t=0}^{t-1}\mathbb{E}||\nabla g(x_t)||^2 \leq \mathbb{E}f(x_0) - \mathbb{E}f(x_t) + \frac{L\gamma^2}{2}p^2\sigma^2 T$$

$$(\gamma - \frac{L\gamma^2}{2})\frac{1}{T}\sum_{t=0}^{t-1}\mathbb{E}||\nabla g(x_t)||^2 \leq \frac{\mathbb{E}f(x_0) - \mathbb{E}f(x_t)}{T} + \frac{L\gamma^2}{2}p^2\sigma^2$$

$$\frac{1}{T}\sum_{t=0}^{t-1}\mathbb{E}||\nabla g(x_t)||^2 \leq \frac{2L(\mathbb{E}f(x_0) - \mathbb{E}f(x_t))}{T} + p^2\sigma^2$$

Lemma 1 suggests that by setting the probability of selecting hybrid batches for training to $p = 1/\sqrt{T}$, we achieve a convergence rate of $O(1/T)$, equivalent to that of full-vertex batch training.

$$\frac{1}{T}\sum_{t=0}^{t-1}\mathbb{E}||\nabla g(x_t)||^2 \leq \frac{2L(\mathbb{E}f(x_0) - \mathbb{E}f(x_t)) + \sigma^2}{T} \tag{2}$$

Setting $p = 1/\sqrt[4]{T}$, however, leads to a convergence rate of $O(1/\sqrt{T})$, akin to mini-batch training.

$$\frac{1}{T}\sum_{t=0}^{t-1}\mathbb{E}||\nabla g(x_t)||^2 \leq \frac{2L(\mathbb{E}f(x_0) - \mathbb{E}f(x_t))}{T} + \frac{\sigma^2}{\sqrt{T}} \tag{3}$$