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

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

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

Therefore, we can deduce the following proof.

$$\begin{aligned}
&\mathbb{E}f(\boldsymbol{x}_{t+1}) - \mathbb{E}f(\boldsymbol{x}_t)\\
&\le \mathbb{E}\left\langle \boldsymbol{x}_{t+1} - \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t)\right\rangle + \frac{L}{2}\mathbb{E}||\boldsymbol{x}_{t+1} - \boldsymbol{x}_t||^2\\
&= \mathbb{E}\left\langle -\gamma\nabla g(\boldsymbol{x}_t), \nabla g(\boldsymbol{x}_t) + p\xi_t\right\rangle + \frac{L\gamma^2}{2}\mathbb{E}||\nabla g(\boldsymbol{x}_t) + p\xi_t||^2 \quad (\nabla g(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) - \xi \ \mathit{and}\ \boldsymbol{x_{t+1}} = \boldsymbol{x_t} - \gamma\nabla g(\boldsymbol{x_t}))\\
&\le -\gamma\mathbb{E}\left\langle \nabla g(\boldsymbol{x}_t), \nabla g(\boldsymbol{x}_t) + p\xi_t\right\rangle + \frac{L\gamma^2}{2}\mathbb{E}||\nabla g(\boldsymbol{x}_t) + p\xi_t||^2\\
&\le -\gamma\mathbb{E}||\nabla g(\boldsymbol{x}_t)||^2 + \frac{L\gamma^2}{2}p^2\mathbb{E}||\nabla g(\boldsymbol{x}_t)||^2 + \frac{L\gamma^2}{2}p^2\mathbb{E}\xi_t^2 \quad (\mathbb{E}[\xi] = 0)\\
&\le (-\gamma + \frac{L\gamma^2}{2})\mathbb{E}||\nabla g(\boldsymbol{x}_t)||^2 + \frac{L\gamma^2}{2}p^2\sigma^2 \quad (\mathbb{E}[||\xi||^2] \le \sigma^2)
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}f(\boldsymbol{x}_t) - \mathbb{E}f(\boldsymbol{x}_0) \le (\frac{L\gamma^2}{2} - \gamma)\sum_{t=0}^{t-1}\mathbb{E}||\nabla g(\boldsymbol{x}_t)||^2 + \frac{L\gamma^2}{2}p^2\sigma^2 T\\
&(\gamma - \frac{L\gamma^2}{2})\sum_{t=0}^{t-1}\mathbb{E}||\nabla g(\boldsymbol{x}_t)||^2 \le \mathbb{E}f(\boldsymbol{x}_0) - \mathbb{E}f(\boldsymbol{x}_t) + \frac{L\gamma^2}{2}p^2\sigma^2 T\\
&(\gamma - \frac{L\gamma^2}{2})\frac{1}{T}\sum_{t=0}^{t-1}\mathbb{E}||\nabla g(\boldsymbol{x}_t)||^2 \le \frac{\mathbb{E}f(\boldsymbol{x}_0) - \mathbb{E}f(\boldsymbol{x}_t)}{T} + \frac{L\gamma^2}{2}p^2\sigma^2\\
&\frac{1}{T}\sum_{t=0}^{t-1}\mathbb{E}||\nabla g(\boldsymbol{x}_t)||^2 \le \frac{2L(\mathbb{E}f(\boldsymbol{x}_0) - \mathbb{E}f(\boldsymbol{x}_t))}{T} + p^2\sigma^2
\end{aligned}$$

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.

$$\frac{1}{T}\sum_{t=0}^{t-1}\mathbb{E}||\nabla g(\boldsymbol{x}_t)||^2 \le \frac{2L(\mathbb{E}f(\boldsymbol{x}_0) - \mathbb{E}f(\boldsymbol{x}_t)) + \sigma^2}{T} \tag{2}$$

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

$$\frac{1}{T}\sum_{t=0}^{t-1}\mathbb{E}||\nabla g(\boldsymbol{x}_t)||^2 \le \frac{2L(\mathbb{E}f(\boldsymbol{x}_0) - \mathbb{E}f(\boldsymbol{x}_t))}{T} + \frac{\sigma^2}{\sqrt{T}} \tag{3}$$