This is the corrigendum for the paper *"Multirate Training of Neural Networks"* which appeared in proceedings: https://proceedings.mlr.press/v162/vlaar22b.html. The authors wish to thank Katerina Karoni for providing valuable comments that inspired this corrigendum. To make the document self-contained we provide the full proof provided in Appendix B below. The updated paper can be found on the arXiv: https://arxiv.org/abs/2106.10771. Please do not hesitate to contact the authors for any further questions regarding this file or the paper itself.

Recall our main assumptions:

Assumption B.1. We assume function $f : \mathbb{R}^n \to \mathbb{R}$ to be L-smooth, i.e., f is continuously differentiable and its gradient is Lipschitz continuous with Lipschitz constant $L > 0$

$$\|\nabla f(\varphi) - \nabla f(\theta)\|_2 \leq L\|\varphi - \theta\|_2,\ \forall \theta, \varphi \in \mathbb{R}^n. \tag{1}$$

Assumption B.2. We assume that the second moment of the stochastic gradient is bounded above, i.e., there exists a constant M for any sample x_i such that

$$\|\nabla f_{x_i}(\theta)\|_2^2 \leq M,\ \ \forall \theta \in \mathbb{R}^n. \tag{2}$$

Lemma B.3. *If $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth then $\forall \theta, \varphi \in \mathbb{R}^n$*

$$|f(\varphi) - (f(\theta) + \nabla f(\theta)^T(\varphi - \theta))| \leq \frac{L}{2}\|\varphi - \theta\|_2^2. \tag{3}$$

As a starting point for our layer-wise multirate approach we partition the parameters as $\theta = \{\theta_F, \theta_S\}$, with $\theta_F \in \mathbb{R}^{n_F}, \theta_S \in \mathbb{R}^{n_S}$, $n = n_F + n_S$. The multirate method update for base algorithm SGD is

$$\theta_\ell^{t+1} = \theta_\ell^t - h\nabla f_{\ell,x_i}(\theta^t), \tag{4}$$

where $\ell \in \{F, S\}$, θ_ℓ^t are the parameter groups at iteration t, h is the stepsize, and $\nabla f_{\ell,x_i}$ denotes the gradient of the loss of the ith training example for parameters θ_ℓ^t, where $\nabla f_{F,x_i}(\theta^t) = \nabla f_{F,x_i}(\theta^t)$ and with linear drift: for any $t \in [\tau, \tau + k - 1]$, where τ is divisible by k, $\nabla f_{S,x_i}(\theta^t) = \nabla f_{S,x_i}(\theta^\tau)$. The total number of iterations T is always set to be a multiple of k. In the following we denote $\nabla f_{x_i}(\theta^t) = \{\nabla f_{F,x_i}(\theta^t), \nabla f_{S,x_i}(\theta^t)\}$ and $g_{x_i}(\theta^t) = \{\nabla f_{F,x_i}(\theta^t), \nabla f_{S,x_i}(\theta^\tau)\}$, such that the parameter update rule becomes

$$\theta^{t+1} = \theta^t - hg_{x_i}(\theta^t). \tag{5}$$

Theorem B.4. *Assume that Assumptions B.1 and B.2 hold. Then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\theta^t)\|_2^2\right] \leq \frac{2(f(\theta^0) - f(\theta^*))}{hT} + hLM\ell\left(\frac{1}{3}hLk^2 + 1\right), \tag{6}$$

where θ^* is the optimal solution to $f(\theta)$.

Proof of Theorem B.4. Because f is L-smooth, from Lemma B.3 it follows that

$$\begin{aligned} f(\theta^{t+1}) &\leq f(\theta^t) + \nabla f(\theta^t) \cdot (\theta^{t+1} - \theta^t) + \frac{L}{2}\|\theta^{t+1} - \theta^t\|_2^2 \\ &\leq f(\theta^t) - h\nabla f(\theta^t) \cdot g_{x_i}(\theta^t) + \frac{h^2L}{2}\left\|g_{x_i}(\theta^t)\right\|_2^2 \end{aligned} \tag{7}$$

Taking the double expectation gives (because of unbiased gradient $\mathbb{E}_{x_i \sim p(X)}[g_{x_i}(\theta^t)] = g(\theta^t)$ and Assumption B.2):

$$\mathbb{E}[f(\theta^{t+1}) - f(\theta^t)] \leq -h\mathbb{E}\left[\nabla f(\theta^t) \cdot g(\theta^t)\right] + h^2LM\ell/2$$

for number of parameter groups ℓ and where $\mathbb{E}[..]$ is the expectation with respect to the parameters. So in T iterations we have θ^T such that (using a telescoping sum):

$$f(\theta^*) - f(\theta^0) \leq \mathbb{E}[f(\theta^T)] - f(\theta^0) \leq -h\underbrace{\sum_{t=0}^{T-1}\mathbb{E}\left[\nabla f(\theta^t) \cdot g(\theta^t)\right]}_{\mathcal{A}} + \frac{h^2LM\ell}{2}T. \tag{8}$$

For term $\mathcal{A}$ we get: $$\mathcal{A} = \sum_{t=0}^{T-1} a_t = \sum_{t=0}^{k-1} a_t + \sum_{t=k}^{2k-1} a_t + \cdots + \sum_{t=\tau}^{\tau+k-1} a_t + \cdots + \sum_{t=T-k}^{T-1} a_t, \tag{9}$$

where $\sum_{t=\tau}^{\tau+k-1} a_t$ is given by

$$\begin{aligned}
\sum_{t=\tau}^{\tau+k-1} \mathbb{E}\left[\nabla f(\theta^t)\cdot g(\theta^t)\right] &= \sum_{t=\tau}^{\tau+k-1} \mathbb{E}\left[\{\nabla f_F(\theta^t), \nabla f_S(\theta^t)\}\cdot\{\nabla f_F(\theta^t), \nabla f_S(\theta^\tau)\}\right] \\
&= \sum_{t=\tau}^{\tau+k-1} \mathbb{E}\left[\|\nabla f_F(\theta^t)\|_2^2\right] + \sum_{t=\tau}^{\tau+k-1} \mathbb{E}[\nabla f_S(\theta^t)\cdot(\nabla f_S(\theta^\tau) - \nabla f_S(\theta^t) + \nabla f_S(\theta^t))] \\
&= \sum_{t=\tau}^{\tau+k-1} \mathbb{E}\left[\|\nabla f(\theta^t)\|_2^2\right] + \underbrace{\sum_{t=\tau}^{\tau+k-1} \mathbb{E}[\nabla f_S(\theta^t)\cdot(\nabla f_S(\theta^\tau) - \nabla f_S(\theta^t))]}_{\mathcal{B}}.
\end{aligned}$$

Because $xy \leq \frac{1}{2}\|x\|_2^2 + \frac{1}{2}\|y\|_2^2$ (combination of Cauchy-Schwarz and Young's inequality) (gives 1st inequality) and Assumption B.1 (gives 2nd inequality) we get for term $\mathcal{B}$

$$\begin{aligned}
\mathcal{B} &\leq \frac{1}{2}\sum_{t=\tau}^{\tau+k-1} \mathbb{E}\left[\|\nabla f_S(\theta^t)\|_2^2\right] + \frac{1}{2}\sum_{t=\tau}^{\tau+k-1} \mathbb{E}\left[\|\nabla f_S(\theta^\tau) - \nabla f_S(\theta^t)\|_2^2\right] \\
&\leq \frac{1}{2}\sum_{t=\tau}^{\tau+k-1} \mathbb{E}\left[\|\nabla f_S(\theta^t)\|_2^2\right] + \frac{L^2}{2}\mathbb{E}\left[\underbrace{\sum_{t=\tau+1}^{\tau+k-1}\|\theta^\tau - \theta^t\|_2^2}_{\mathcal{C}}\right].
\end{aligned}$$

We get for term $\mathcal{C}$ from Eq. (4) (gives 2nd equality), $\|a_1 + \cdots + a_m\|_2^2 \leq m(\|a_1\|_2^2 + \cdots + \|a_m\|_2^2)$ (gives 1st inequality), Assumption B.2 (gives 2nd inequality), and $k > 1$ (final inequality):

$$\begin{aligned}
\mathcal{C} &= \|\theta^\tau - \theta^{\tau+1}\|_2^2 + \|\theta^\tau - \theta^{\tau+2}\|_2^2 + \cdots + \|\theta^\tau - \theta^{\tau+k-1}\|_2^2 \\
&= h^2\left(\|g_{x_i}(\theta^\tau)\|_2^2 + \left\|g_{x_i}(\theta^\tau) + g_{x_i}(\theta^{\tau+1})\right\|_2^2 + \cdots + \left\|g_{x_i}(\theta^\tau) + \cdots + g_{x_i}(\theta^{\tau+k-2})\right\|_2^2\right) \\
&\leq h^2\left(\sum_{m=1}^{k-1} m\,\|g_{x_i}(\theta^\tau)\|_2^2 + \sum_{m=2}^{k-1} m\left\|\,g_{x_i}(\theta^{\tau+1})\right\|_2^2 + \cdots + (k-1)\left\|g_{x_i}(\theta^{\tau+k-2})\right\|_2^2\right) \\
&\leq h^2 M\ell\left((k-1)^2 + (k-2)^2 + \cdots + 1\right) = h^2 M\ell\sum_{m=1}^{k-1} m^2 = h^2 M\ell\left(k/6 - k^2/2 + k^3/3\right) \leq h^2 M\ell k^3/3.
\end{aligned}$$

So overall for term $-h\mathcal{A}$ we get

$$\begin{aligned}
-h\sum_{t=0}^{T-1}\mathbb{E}[\nabla f(\theta^t)\cdot g(\theta^t)] &\leq -h\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\theta^t)\|_2^2\right] + h\left|\sum_\tau \mathcal{B}\right| \\
&\leq -\frac{h}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\theta^t)\|_2^2\right] + \frac{1}{6}h^3L^2M\ell k^2 T. \qquad (10)
\end{aligned}$$

Substituting this into Eq. (8) gives

$$\begin{aligned}
f(\theta^*) - f(\theta^0) &\leq \mathbb{E}[f(\theta^T)] - f(\theta^0) \\
&\leq -\frac{h}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\theta^t)\|_2^2\right] + \frac{1}{6}h^3L^2M\ell k^2 T + \frac{h^2 LM\ell}{2}T \\
&= -\frac{h}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\theta^t)\|_2^2\right] + \frac{1}{2}h^2LM\ell T\left(\frac{1}{3}hLk^2 + 1\right). \qquad (11)
\end{aligned}$$

This gives Theorem B.4

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\theta^t)\|_2^2\right] \leq \frac{2(f(\theta^0) - f(\theta^*))}{hT} + hLM\ell\left(\frac{1}{3}hLk^2 + 1\right).$$

□