Correction to: Training Deep Neural Networks Using Conjugate Gradient-like Methods (Electronics 2020, 9, 1809; doi:10.3390/electronics9111809)

Hideaki Iiduka and Yu Kobayashi

Let $\alpha_n := 1/n^\eta$, $\beta_n := \beta^n$, and $\gamma_n := \gamma^n$ or $1/n^\kappa$, where $\eta \in [1/2, 1), \kappa > 1 - \eta$, and $\beta, \gamma \in (0,1)$. The rate of convergence of Algorithm 1 in Theorem 2,

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[\langle \boldsymbol{x} - \boldsymbol{x}_k, \nabla f(\boldsymbol{x}_k)\rangle\right] \geq \begin{cases} -\mathcal{O}\left(\sqrt{\dfrac{1+\ln n}{n}}\right) & \text{if } \eta = \frac{1}{2}, \\ -\mathcal{O}\left(\dfrac{1}{n^{1-\eta}}\right) & \text{if } \eta \in \left(\frac{1}{2}, 1\right), \end{cases} \tag{1}$$

can be replaced with

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[\langle \boldsymbol{x} - \boldsymbol{x}_k, \nabla f(\boldsymbol{x}_k)\rangle\right] \geq -\mathcal{O}\left(\frac{1}{n^{1-\eta}}\right). \tag{2}$$

The correction implies that Algorithm 1 achieves a better convergence rate (2) than (1). In particular, Algorithm 1 with $\eta = 1/2$ achieves an $\mathcal{O}(1/\sqrt{n})$ convergence rate.

Proof of (2). We have that

$$\frac{1}{n\alpha_n} = \frac{1}{n^{1-\eta}}$$

and

$$\frac{1}{n}\sum_{k=1}^{n} \alpha_k \leq \frac{1}{n}\left(1 + \int_1^n \frac{\mathrm{d}t}{t^\eta}\right) \leq \frac{1}{1-\eta}\frac{1}{n^\eta} \leq \frac{1}{1-\eta}\frac{1}{n^{1-\eta}}. \tag{3}$$

We also have that

$$\frac{1}{n}\sum_{k=1}^{n} \beta_k \leq \frac{1}{n}\sum_{k=1}^{+\infty} \beta^k = \frac{\beta}{(1-\beta)n}.$$

First, let us consider the case where $\gamma_n := \gamma^n$. Then,

$$\frac{1}{n}\sum_{k=1}^{n}\gamma_k \leq \frac{1}{n}\sum_{k=1}^{+\infty}\gamma^k = \frac{\gamma}{(1-\gamma)n}.$$

Therefore, Theorem A1 implies (2).

Next, consider the case where $\gamma_n := 1/n^\kappa$, where $\kappa > 1-\eta$. An argument similar to the one for obtaining (3) implies that

$$\frac{1}{n}\sum_{k=1}^{n}\gamma_k \leq \frac{1}{1-\kappa}\frac{1}{n^\kappa} \leq \frac{1}{1-\kappa}\frac{1}{n^{1-\eta}}.$$

Therefore, Theorem A1 implies (2). This completes the proof. □

Under the convex setting (Proposition 2), we can replace

$$\mathbb{E}\left[f(\tilde{\boldsymbol{x}}_n) - f^\star\right] = \begin{cases} \mathcal{O}\left(\sqrt{\dfrac{1+\ln n}{n}}\right) & \text{if } \eta = \frac{1}{2}, \\ \mathcal{O}\left(\dfrac{1}{n^{1-\eta}}\right) & \text{if } \eta \in \left(\frac{1}{2}, 1\right) \end{cases}$$

with

$$\mathbb{E}\left[f(\tilde{\boldsymbol{x}}_n) - f^\star\right] = \mathcal{O}\left(\frac{1}{n^{1-\eta}}\right).$$