Errata of "Safe Deep Semi-Supervised Learning for Unseen-Class Unlabeled Data"

Abstract

We fix an error in Theorem 1 and some typos in Theorem 2. To be more clear, we rewrite the proof of the Theorem.

A. Proof of Convergence Results

We first introduce several useful function properties.

Definition 1. *A function $f(x) : \mathbb{R}^d \to \mathbb{R}$ is said to be Lipshitz-smooth with constant L if*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \forall x, y \in \mathbb{R}^d$$

Definition 2. *A function $f(x)$ has ρ-bounded gradients if $\|\nabla f(x)\| \leq \rho, \forall x \in \mathbb{R}^d$.*

Then, we prove the main results about convergence.

Theorem 1. *(Convergence.) Suppose the loss function is L-Lipschitz smooth and*

$$(\nabla_\theta \mathcal{L}^{outer}(\theta))^\top \nabla_\theta \mathcal{L}^{inner}(\theta, \alpha) \geq G\|\nabla_\theta \mathcal{L}^{inner}(\theta, \alpha)\|_2^2$$

where $G \geq 0$[1]. If the step size $\eta_\theta \leq \frac{2G}{L}$, then follow our optimization algorithm, the labeled loss always monotonically decreases with the iteration t, i.e.,

$$\mathcal{L}^{outer}(\theta_{t+1}) \leq \mathcal{L}^{outer}(\theta_t)$$

Furthermore, the equality in Eq.(1) holds when the gradient of the outer objective respect to α becomes 0 at some iteration t, i.e.,

$$\mathcal{L}^{outer}(\theta_{t+1}) = \mathcal{L}^{outer}(\theta_t)$$

if

$$\nabla_\alpha \mathcal{L}^{outer}(\theta_t) = 0$$

Proof. The change of outer-level objective from iteration t to $t+1$ is:

$$\begin{aligned}
&\mathcal{L}^{outer}(\theta_{t+1}) - \mathcal{L}^{outer}(\theta_t) \\
&= \mathcal{L}^{outer}(\theta_t - \eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_t, \alpha_t)) - \mathcal{L}^{outer}(\theta_t) \\
&\leq -\eta_\theta (\nabla_\theta \mathcal{L}^{outer}(\theta_t))^\top \nabla_\theta \mathcal{L}^{inner}(\theta_t, \alpha_t) + \\
&\frac{L}{2}\| - \eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_t, \alpha_t)\|_2^2 \\
&\leq (\frac{L\eta_\theta^2}{2} - \eta_\theta G)\|\nabla_\theta \mathcal{L}^{inner}(\theta_t, \alpha_t)\|_2^2 \\
&\leq 0
\end{aligned}$$

The first inequality holds since the loss function is Lipschitz-smooth with constant L and the second inequality holds since the gradient of supervised loss is dominant in the optimization process of semi-supervised learning. The third inequality holds since $0 \leq \eta_\theta \leq \frac{2G}{L}$. □

Theorem 2. *(Convergence Rate.) Suppose the loss function are L-Lipschitz smooth and have ρ-bounded gradients, let the step size η_θ for θ satisfies $\eta_\theta = \min\{1, \frac{k}{T}\}$ for some constant $k > 0$, such that $\frac{k}{T} < 1$ and $\eta_\alpha = \min\{\frac{1}{L}, \frac{c}{\sqrt{T}}\}$ for some constant $c > 0$, such that $\frac{\sqrt{T}}{c} \geq L$. Then, the approximation algorithm can achieve $\mathbb{E}[\|\nabla_\alpha \mathcal{L}^{outer}(\theta_t)\|_2^2] \leq \epsilon$ in $\mathcal{O}(1/\epsilon^2)$. And more specifically,*

$$\min_{0 \leq t \leq T} \mathbb{E}[\|\nabla_\alpha \mathcal{L}^{outer}(\theta_t)\|_2^2] \leq \mathcal{O}(\frac{1}{\sqrt{T}})$$

Proof. For the writing simplicity, we write $g(\theta_t, \alpha_t) = \theta_t - \eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_t, \alpha_t)$. According to the updating rule, we have

$$\begin{aligned}
&\mathcal{L}^{outer}(\theta_{t+1}) - \mathcal{L}^{outer}(\theta_t) \\
&= \mathcal{L}^{outer}(g(\theta_t, \alpha_t)) - \mathcal{L}^{outer}(g(\theta_{t-1}, \alpha_{t-1})) \\
&= \{\mathcal{L}^{outer}(g(\theta_t, \alpha_t)) - \mathcal{L}^{outer}(g(\theta_{t-1}, \alpha_t))\} + \\
&\{\mathcal{L}^{outer}(g(\theta_{t-1}, \alpha_t)) - \mathcal{L}^{outer}(g(\theta_{t-1}, \alpha_{t-1}))\}
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
&\mathcal{L}^{outer}(g(\theta_t, \alpha_t)) - \mathcal{L}^{outer}(g(\theta_{t-1}, \alpha_t)) \\
&\leq \left\langle \nabla_\theta \mathcal{L}^{outer}(g(\theta_{t-1}, \alpha_t)), g(\theta_t, \alpha_t) - g(\theta_{t-1}, \alpha_t) \right\rangle \\
&+ \frac{L}{2}\|g(\theta_t, \alpha_t) - g(\theta_{t-1}, \alpha_t)\|_2^2 \\
&\leq \eta_\theta \rho^2 + \frac{L}{2}\eta_\theta^2 \rho^2 = \eta_\theta \rho^2 (\frac{\eta_\theta L}{2} + 1)
\end{aligned}$$

[1] The assumption holds naturally since in practice the gradient of supervised loss is dominant in the optimization process of semi-supervised learning, for example, G is often larger than $1/4$.

The first inequality holds since the loss function is Lipschitz-smooth and the second inequality holds since the gradients of loss functions are bounded by ρ.

For the second term, we can adopt a weight learning function w to make $\mathcal{L}^{outer}$ smooth w.r.t. α, i.e.,

$$\|\nabla_\alpha \mathcal{L}^{outer}(g(\theta,\alpha_t)) - \nabla_\alpha \mathcal{L}^{outer}(g(\theta,\alpha_{t+1}))\| \leq L\|\alpha_t - \alpha_{t+1}\|, \forall t$$

Then we have:

$$\begin{aligned}
&\mathcal{L}^{outer}(g(\theta_{t-1},\alpha_t)) - \mathcal{L}^{outer}(g(\theta_{t-1},\alpha_{t-1})) \\
&\leq \left\langle \nabla_\alpha \mathcal{L}^{outer}(g(\theta_{t-1},\alpha_{t-1})), \alpha_t - \alpha_{t-1} \right\rangle \\
&+ \frac{L}{2}\|\alpha_t - \alpha_{t-1}\|_2^2 \\
&= -(\eta_\alpha - \frac{L}{2}\eta_\alpha^2)\|\nabla_\alpha \mathcal{L}^{outer}(\theta_t)\|_2^2
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathcal{L}^{outer}(\theta_{t+1}) - \mathcal{L}^{outer}(\theta_t) \\
&\leq \eta_\theta \rho^2 (\frac{\eta_\theta L}{2} + 1) - (\eta_\alpha - \frac{L}{2}\eta_\alpha^2)\|\nabla_\alpha \mathcal{L}^{outer}(\theta_t)\|_2^2
\end{aligned}$$

Summing up the above inequalities and rearranging the terms, we can obtain

$$\begin{aligned}
&\sum_{t=1}^{T} (\eta_\alpha - \frac{L}{2}\eta_\alpha^2)\|\nabla_\alpha \mathcal{L}^{outer}(\theta_t)\|_2^2 \\
&\leq \mathcal{L}^{outer}(\theta_1) - \mathcal{L}^{outer}(\theta_{T+1}) + \eta_\theta \rho^2 (\frac{\eta_\theta LT}{2} + T) \\
&\leq \mathcal{L}^{outer}(\theta_1) + \eta_\theta \rho^2 (\frac{\eta_\theta LT}{2} + T)
\end{aligned}$$

Further, we can deduce that,

$$\begin{aligned}
&\min_t \mathbb{E}[\|\nabla_\alpha \mathcal{L}^{outer}(\theta_t)\|_2^2] \\
&\leq \frac{\sum_{t=1}^{T} (\eta_\alpha - \frac{L}{2}\eta_\alpha^2)\|\nabla_\alpha \mathcal{L}^{outer}(\theta_t)\|_2^2}{\sum_{t=1}^{T} (\eta_\alpha - \frac{L}{2}\eta_\alpha^2)} \\
&\leq \frac{1}{T(2\eta_\alpha - L\eta_\alpha^2)} [2\mathcal{L}^{outer}(\theta_1) + \eta_\theta \rho^2 (2T + \eta_\theta LT)] \\
&\leq \frac{1}{T\eta_\alpha} [2\mathcal{L}^{outer}(\theta_1) + \eta_\theta \rho^2 (2T + \eta_\theta LT)] \\
&\leq \frac{2\mathcal{L}^{outer}(\theta_1)}{T} \frac{1}{\eta_\alpha} + \frac{\eta_\theta \rho^2 (2+L)}{\eta_\alpha} \\
&= \frac{2\mathcal{L}^{outer}(\theta_1)}{T} \max\{L, \frac{\sqrt{T}}{C}\} \\
&+ \min\{1, \frac{k}{T}\} \max\{L, \frac{\sqrt{T}}{C}\} \rho^2 (2+L) \\
&\leq \frac{2\mathcal{L}^{outer}(\theta_1)}{C\sqrt{T}} + \frac{k\rho^2(2+L)}{C\sqrt{T}} = O(\frac{1}{\sqrt{T}})
\end{aligned}$$

The third inequality holds since $\eta_\alpha \leq \frac{1}{L}$ and the fourth inequality holds since $\eta_\theta \leq 1$. □