Robust Semi-supervised Learning by Wisely Leveraging Open-set Data

Yang Yang, *Member, IEEE*, Nan Jiang, Yi Xu, and De-Chuan Zhan

Index Terms—Semi-supervised Learning, OOD Detection, Open-set Data.

References

Appendix A
Proof of the Theorem 1

Proof. (a) Following the standard analysis, we have the following inequality in expectation

$$\begin{aligned}
&\mathrm{E}[\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t)] \\
\overset{\text{(i)}}{\leq} &\mathrm{E}[\langle \theta_{t+1} - \theta_t, \nabla \mathcal{L}(\theta_t)\rangle] + \frac{L}{2}\mathrm{E}[\|\theta_{t+1} - \theta_t\|^2] \\
\overset{\text{(ii)}}{=} &-\eta \mathrm{E}[\langle \widehat{g}(\theta_t), \nabla \mathcal{L}(\theta_t)\rangle] + \frac{\eta^2 L}{2}\mathrm{E}[\|\widehat{g}(\theta_t)\|^2] \\
\overset{\text{(iii)}}{=} &-\eta(1 - \frac{\eta L}{2})\|\nabla \mathcal{L}(\theta_t)\|^2 + \frac{\eta^2 L}{2}\mathrm{E}[\|\nabla \mathcal{L}(\theta_t) - \widehat{g}(\theta_t)\|^2] \\
\overset{\text{(iv)}}{\leq} &-\frac{\eta}{2}\|\nabla \mathcal{L}(\theta_t)\|^2 + \frac{\eta^2 L}{2}\mathrm{E}[\|\nabla \mathcal{L}(\theta_t) - \widehat{g}(\theta_t)\|^2], \quad (1)
\end{aligned}$$

where (i) is due to the smoothness of loss function $\mathcal{L}$ in Assumption **??**; (ii) is due to the update of SGD; (iii) is due to $\mathrm{E}[\widehat{g}(\theta)] = \nabla \mathcal{L}(\theta)$; (iv) is due to $\eta \leq 1/L$. By the definition of $\widehat{g}(\theta_t)$ and the convexity of norm $\|\cdot\|^2$, we know

$$\begin{aligned}
&\mathrm{E}[\|\nabla \mathcal{L}(\theta_t) - \widehat{g}(\theta_t)\|^2] \\
\leq &\lambda \mathrm{E}[\|\nabla \mathcal{L}(\theta_t) - g_{\text{id}}(\theta_t)\|^2] \\
&+ (1-\lambda)(\tau \mathrm{E}[\|\nabla \mathcal{L}(\theta_t) - g_{\text{fr}}(\theta_t)\|^2 \\
&+ (1-\tau)\mathrm{E}[\|\nabla \mathcal{L}(\theta_t) - g_{\text{uf}}(\theta_t)\|^2]) \\
\leq &\lambda\sigma^2 + (1-\lambda)[\tau(\frac{\epsilon}{2}\|\nabla \mathcal{L}(\theta_t)\|^2 + \sigma^2) \\
&+ (1-\tau)(\frac{\nu}{2}\|\nabla \mathcal{L}(\theta_t)\|^2 + \sigma^2)] \\
= &\sigma^2 + (1-\lambda)\frac{\tau\epsilon + (1-\tau)\nu}{2}\|\nabla \mathcal{L}(\theta_t)\|^2, \quad (2)
\end{aligned}$$

where the second inequality is due to Assumptions **??** and **??**. Therefore, by (1) and (2) we have

$$\begin{aligned}
&\mathrm{E}[\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t)] \\
\leq &-\frac{\eta}{2}\left(1 - \eta L(1-\lambda)\frac{\tau\epsilon + (1-\tau)\nu}{2}\right)\mathrm{E}[\|\nabla \mathcal{L}(\theta_t)\|^2] + \frac{\eta^2 L\sigma^2}{2} \\
\leq &-\frac{\mu\eta}{2}\mathrm{E}[\mathcal{L}(\theta_t) - \mathcal{L}(\theta_*)] + \frac{\eta^2 L\sigma^2}{2}, \quad (3)
\end{aligned}$$

where the last inequality is due to Assumption **??** and $\eta = \frac{1}{(1-\lambda)(\tau\epsilon + (1-\tau)\nu)L}$. Therefore, we have

$$\mathrm{E}[\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_*)] \leq \left(1 - \frac{\mu\eta}{2}\right)\mathrm{E}[\mathcal{L}(\theta_t) - \mathcal{L}(\theta_*)] + \frac{\eta^2 L\sigma^2}{2}, \quad (4)$$

which implies

$$\begin{aligned}
&\mathrm{E}[\mathcal{L}(\theta_{n+m+m'+1}) - \mathcal{L}(\theta_*)] \\
\leq &\exp\left(-\mu\eta(n+m+m')/2\right)\left(\mathcal{L}(\theta_0) - \mathcal{L}(\theta_*)\right) + \frac{\eta L\sigma^2}{\mu}, \quad (5)
\end{aligned}$$

Since ν is large enough, then we know that $\eta' := \frac{1}{(1-\lambda)(\tau\epsilon + (1-\tau)\nu)L}$ is small enough such that $\exp\left(-\mu\eta'(n+m+m')/2\right) \gg \frac{L\sigma^2}{(n+m+m')\mu^2(\mathcal{L}(\theta_0) - \mathcal{L}(\theta_*))}$. Thus, we have

$$\begin{aligned}
&\mathrm{E}[\mathcal{L}(\theta_{n+m+m'+1}) - \mathcal{L}(\theta_*)] \\
\leq &\exp\left(-\mu\eta'(n+m+m')/2\right)\left(\mathcal{L}(\theta_0) - \mathcal{L}(\theta_*)\right) + \frac{\eta' L\sigma^2}{\mu} \quad (6)
\end{aligned}$$

Let's consider a special case of $n + m + m' = \frac{2(1-\lambda)(\tau\epsilon + (1-\tau)\nu)L}{\mu}$, then $\eta' = \frac{2}{(n+m+m')\mu}$, implying

$$\mathrm{E}[\mathcal{L}(\theta_{n+m+m'+1}) - \mathcal{L}(\theta_*)] \leq O\left(\mathcal{L}(\theta_0) - \mathcal{L}(\theta_*)\right) \quad (7)$$

(b) Next, let's consider the case without either far OOD data or near OOD data, i.e., $\lambda = 1$. Then, following the similar analysis in (a), we have

$$\begin{aligned}
&\mathrm{E}[\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t)] \\
\leq &-\frac{\eta}{2}\mathrm{E}[\|\nabla \mathcal{L}(\theta_t)\|^2] + \frac{\eta^2 L\sigma^2}{2} \\
\leq &-\frac{\mu\eta}{2}\mathrm{E}[\mathcal{L}(\theta_t) - \mathcal{L}(\theta_*)] + \frac{\eta^2 L\sigma^2}{2}, \quad (8)
\end{aligned}$$

- *Yang Yang is with the school of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China. E-mail: yyang@njust.edu.cn*
- *Nan Jiang and De-Chuan Zhan are with the National Key Laboratory for Novel Software Technology, Nanjing University, and also with the School of Artificial Intelligence, Nanjing University, Nanjing 210023, China. E-mail: jiangn@lamda.nju.edu.cn, zhandc@nju.edu.cn*
- *Yi Xu is with the School of Control Science and Engineering, Dalian University of Technology, Dalian 116081, China. E-mail: yxu@dlut.edu.cn*

Corresponding authors: Yi Xu and De-Chuan Zhan.

where the last inequality is due to Assumption **??**. Therefore, we have

$$\mathrm{E}[\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_*)] \leq \left(1 - \frac{\mu\eta}{2}\right) \mathrm{E}[\mathcal{L}(\theta_t) - \mathcal{L}(\theta_*)] + \frac{\eta^2 L\sigma^2}{2}, \tag{9}$$

which implies

$$\mathrm{E}[\mathcal{L}(\theta_{n+1}) - \mathcal{L}(\theta_*)] \leq \exp\left(-\mu\eta n/2\right) \left(\mathcal{L}(\theta_0) - \mathcal{L}(\theta_*)\right) + \frac{\eta L\sigma^2}{\mu}, \tag{10}$$

By setting $\eta = \frac{2}{n\mu} \log\left(\frac{n\mu^2(\mathcal{L}(\theta_0) - \mathcal{L}(\theta_*))}{\sigma^2 L}\right)$,

$$\begin{aligned}
&\mathrm{E}[\mathcal{L}(\theta_{n+1}) - \mathcal{L}(\theta_*)] \\
\leq& \frac{L\sigma^2}{n\mu^2} + \frac{2L\sigma^2}{n\mu^2} \log\left(\frac{n\mu^2(\mathcal{L}(\theta_0) - \mathcal{L}(\theta_*))}{\sigma^2 L}\right) \\
\leq& O\left(\frac{\log(n)}{n}\right)
\end{aligned} \tag{11}$$

(c) Finally, let's consider the case without far OOD data (but with near OOD data), i.e., $\tau = 1$. Then, following the similar analysis in (a), we have

$$\begin{aligned}
&\mathrm{E}[\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t)] \\
\leq& -\frac{\eta}{2}\left(1 - \frac{(1-\lambda)\epsilon\eta L}{2}\right) \mathrm{E}[\|\nabla\mathcal{L}(\theta_t)\|^2] + \frac{\eta^2 L\sigma^2}{2} \\
\leq& -\frac{\mu\eta}{2}\mathrm{E}[\mathcal{L}(\theta_t) - \mathcal{L}(\theta_*)] + \frac{\eta^2 L\sigma^2}{2},
\end{aligned} \tag{12}$$

where the last inequality is due to Assumption **??** and $\eta \leq \frac{1}{(1-\lambda)\epsilon\nu L}$. Therefore, we have

$$\mathrm{E}[\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_*)] \leq \left(1 - \frac{\mu\eta}{2}\right) \mathrm{E}[\mathcal{L}(\theta_t) - \mathcal{L}(\theta_*)] + \frac{\eta^2 L\sigma^2}{2}, \tag{13}$$

which implies

$$\begin{aligned}
&\mathrm{E}[\mathcal{L}(\theta_{n+m+1}) - \mathcal{L}(\theta_*)] \\
&\leq \exp\left(-\mu\eta(n+m)/2\right) \left(\mathcal{L}(\theta_0) - \mathcal{L}(\theta_*)\right) + \frac{\eta L\sigma^2}{\mu},
\end{aligned} \tag{14}$$

By setting $\eta = \frac{2}{(n+m)\mu} \log\left(\frac{(n+m)\mu^2(\mathcal{L}(\theta_0) - \mathcal{L}(\theta_*))}{\sigma^2 L}\right)$,

$$\begin{aligned}
&\mathrm{E}[\mathcal{L}(\theta_{n+m+1}) - \mathcal{L}(\theta_*)] \\
\leq& \frac{L\sigma^2}{(n+m)\mu^2} + \frac{2L\sigma^2}{(n+m)\mu^2} \log\left(\frac{(n+m)\mu^2(\mathcal{L}(\theta_0) - \mathcal{L}(\theta_*))}{\sigma^2 L}\right) \\
\leq& O\left(\frac{\log(n+m)}{(n+m)}\right)
\end{aligned} \tag{15}$$

□