## A. Proof of Convergence Results

We first introduce several useful function properties.

**Definition 1.** *A function $f(x) : \mathbb{R}^d \rightarrow \mathbb{R}$ 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 $\rho$-bounded gradients if $\|\nabla f(x)\| \leq \rho, \forall x \in \mathbb{R}^d$.*

**Definition 3.** *A function $f(x)$ has $\mathcal{B}$-bounded Hessian if $\|\nabla^2 f(x)\| \leq \mathcal{B}, \forall x \in \mathbb{R}^d$.*

Then, we prove the main results about convergence.

**Theorem 1.** *(Convergence.) Suppose the supervised loss function is Lipschitz-smooth with constant $L \leq 2$, and the supervised loss and unsupervised loss have $\rho$-bounded gradients, then follow our optimization algorithm, the labeled loss always monotonically decreases with the iteration $t$, i.e.,*

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

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

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

*if and only if*

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

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

$$\begin{aligned}
&\mathcal{L}^{outer}(\theta_{t+1}) - \mathcal{L}^{outer}(\theta_t) \\
&= \mathcal{L}^{outer}(\theta_t - \eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_t, \alpha_t)) - \mathcal{L}^{outer}(\theta_t) \\
&\leq \left\langle \nabla_\theta \mathcal{L}^{outer}(\theta_t), -\eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_t, \alpha_t) \right\rangle + \\
&\frac{L}{2}\| -\eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_t, \alpha_t)\| \\
&\leq (\frac{L}{2} - 1)\eta_\theta \rho^2 \leq 0.
\end{aligned} \tag{2}$$

The first inequality holds since the loss function is Lipschitz-smooth with constant $L$ and the second inequality holds since both the supervised and unsupervised loss function has $\rho$-bounded gradients. The third inequality holds since $L \leq 2$.

Moreover, it is obviously that if and only if $\nabla_\alpha \mathcal{L}^{outer}(\theta_t) = 0$, the optimization will converge and $\mathcal{L}^{outer}(\theta_{t+1}) = \mathcal{L}^{outer}(\theta_t)$. ☐

**Theorem 2.** *(Convergence Rate.) Suppose the aforementioned conditions hold, let the step size $\eta_\theta$ for $\theta$ 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 $\frac{\sqrt{T}}{C} \leq L$. 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,*

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

*where $C$ is some constant independent to the convergence process.*

*Proof.* First, according to the updating rule, we have:

$$\begin{aligned}
&\mathcal{L}^{outer}(\theta_{t+1}) - \mathcal{L}^{outer}(\theta_t) \\
&= \mathcal{L}^{outer}(\theta_t - \eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_t, \alpha_t)) \\
&- \mathcal{L}^{outer}(\theta_{t-1} - \eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_{t-1}, \alpha_{t-1})) \\
&= \{\mathcal{L}^{outer}(\theta_t - \eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_t, \alpha_t)) \\
&- \mathcal{L}^{outer}(\theta_{t-1} - \eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_t, \alpha_t))\} \\
&+ \{\mathcal{L}^{outer}(\theta_{t-1} - \eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_t, \alpha_t)) \\
&- \mathcal{L}^{outer}(\theta_{t-1} - \eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_{t-1}, \alpha_{t-1}))\}
\end{aligned} \tag{3}$$

and

$$\begin{aligned}
&\mathcal{L}^{outer}(\theta_t - \eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_t, \alpha_t)) - \\
&\mathcal{L}^{outer}(\theta_{t-1} - \eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_t, \alpha_t)) \\
&\leq \left\langle \nabla_\theta \mathcal{L}^{outer}[\theta_{t-1} - \eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_t, \alpha_t)], \theta_t - \theta_{t-1} \right\rangle \\
&+ \frac{L}{2}\|\theta_t - \theta_{t-1}\|_2^2 \\
&\leq -\eta_\theta \rho^2 + \frac{L}{2}\eta_\theta \rho^2 = \eta_\theta \rho^2 (\frac{L}{2} - 1)
\end{aligned} \tag{4}$$

For the second term, we can adopt a Lipschitz-continuous function as $w$ to make $\mathcal{L}^{outer}$ smooth w.r.t. $\alpha$. Then we have:

$$\begin{aligned}
&\mathcal{L}^{outer}(\theta_{t-1} - \eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_t, \alpha_t)) \\
&- \mathcal{L}^{outer}(\theta_{t-1} - \eta_\theta \nabla_\theta \mathcal{L}^{inner}(\theta_{t-1}, \alpha_{t-1})) \\
&\leq \left\langle \nabla_\alpha \mathcal{L}^{outer}(\theta_t), \alpha_t - \alpha_{t-1} \right\rangle + \frac{L}{2}\|\alpha_t - \alpha_{t-1}\|_2^2 \\
&= \left\langle \nabla_\alpha \mathcal{L}^{outer}(\theta_t), -\eta_\alpha \nabla_\alpha \mathcal{L}^{outer}(\theta_t) \right\rangle + \frac{L}{2}\eta_\alpha^2 \|\nabla_\alpha \mathcal{L}^{outer}(\theta_t)\|_2^2 \\
&= -(\eta_\alpha - \frac{L}{2}\eta_\alpha^2)\|\nabla_\alpha \mathcal{L}^{outer}(\theta_t)\|_2^2
\end{aligned} \tag{5}$$

Therefore,

$$\begin{aligned}
&\mathcal{L}^{outer}(\theta_{t+1}) - \mathcal{L}^{outer}(\theta_t) \\
&\leq \eta_\theta \rho^2(-1 + \frac{L}{2}) - (\eta_\alpha - \frac{L}{2}\eta_\alpha^2)\|\nabla_\alpha \mathcal{L}^{outer}(\theta_t)\|_2^2
\end{aligned} \tag{6}$$

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

$$\begin{aligned}
&\sum_{t=1}^{T}(\eta_\alpha - \frac{L}{2}\eta_\alpha^2)\|\nabla_\alpha \mathcal{L}^{outer}(\theta_t)\|_2^2 \\
&\leq \mathcal{L}^{outer}(\theta_1) - \mathcal{L}^{outer}(\theta_{T+1}) + \eta_\theta \rho^2(-T + \frac{LT}{2}) \\
&\leq \mathcal{L}^{outer}(\theta_1) + \eta_\theta \rho^2(-T + \frac{LT}{2})
\end{aligned} \tag{7}$$

Further, we can deduce that,

$$\begin{aligned}
&\min_t \mathbb{E}[\|\nabla_\alpha \mathcal{L}^{outer}(\theta_t)\|_2^2] \\
&\leq \frac{1}{2\sum_{t=1}^T(\eta_\alpha - L\eta_\alpha^2)}[2\mathcal{L}^{outer}(\theta_1) + \eta_\theta \rho^2(-2T + LT)] \\
&\leq \frac{1}{\sum_{t=1}^T \eta_\alpha}[2\mathcal{L}^{outer}(\theta_1) + \eta_\theta \rho^2(-2T + LT)] \\
&= \frac{2\mathcal{L}^{outer}(\theta_1)}{T}\frac{1}{\eta_\alpha} + \frac{\eta_\theta\rho^2(-2+L)}{\eta_\alpha} \\
&= \frac{2\mathcal{L}^{outer}(\theta_1)}{T}\max\{L, \frac{\sqrt{T}}{C}\} \\
&+ \min\{1, \frac{k}{T}\}\max\{L, \frac{\sqrt{T}}{C}\}\rho^2(-2+L) \\
&\leq \frac{2\mathcal{L}^{outer}(\theta_1)}{C\sqrt{T}} + \frac{k\rho^2(-2+L)}{C\sqrt{T}} = O(\frac{1}{\sqrt{T}})
\end{aligned} \tag{8}$$

□

## B. Proof of Theoretical Studies

We first introduce several useful definitions.

**Definition 4.** *(Hoeffding's inequality). Let $Z_1, \cdots, Z_n$ be independent bounded random variables with $Z_i \in [0,1]$ for all $i$. Then*

$$P(\frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}(Z_i) \geq t) \leq \exp(-2n\epsilon^2)$$

*and*

$$P(\frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}(Z_i) \leq -t) \leq \exp(-2n\epsilon^2)$$

*for all $t \geq 0$.*

**Definition 5.** *($\epsilon$-cover). A set $\mathcal{A}$ is and $\epsilon$-cover of $\mathcal{B}$, if $\forall \alpha \in \mathcal{B}, \exists \alpha' \in \mathcal{A}$ satisfies $||\alpha - \alpha'|| \leq \epsilon$.*

Then we prove the main results to show the safeness results of our proposal.

**Theorem 3.** *(Safeness.) Let $\theta^{SL}$ be the supervised model, i.e., $\theta^{SL} = \arg\min_{\theta\in\Theta}\sum_{i=1}^n \ell(h(\mathbf{x}_i;\theta), \mathbf{y}_i)$. Define the empirical risk as:*

$$\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^n [\ell(h(\mathbf{x}_i;\theta), \mathbf{y}_i)]$$

*Then we have the empirical risk of $\hat{\theta}$ that returned by* DS$^4$L *is never worse than $\theta^{SL}$ that learned from merely labeled data, i.e., $\hat{R}(\hat{\theta}) \leq \hat{R}(\theta^{SL})$.*

*Proof.* Suppose $\hat{R}(\hat{\theta}) > \hat{R}(\theta^{SL})$, obviously we can always set all weights of unlabeled examples to zero and obtain $\hat{R}(\hat{\theta}) = \hat{R}(\theta^{SL})$. Therefore, $\hat{\theta}$ is never worse than $\theta^{SL}$. □

**Theorem 4.** *(Generalization.) Assume the loss function is $\lambda$-Lipschitz continuous w.r.t. $\alpha$. Let $\alpha \in \mathbb{B}^d$ be the parameter of example weighting function $w$ in a $d$-dimensional unit ball. Let $n$ be the labeled data size. Define the generalization risk as:*

$$R(\theta) = \mathbb{E}_{(X,Y)}[\ell(h(X;\theta), Y)]$$

*Let $\alpha^* = \arg\min_{\alpha\in\mathbb{B}^d} R(\hat{\theta}(\alpha))$ be the optimal parameter in the unit ball, and $\hat{\alpha} = \arg\min_{\alpha\in\mathcal{A}} \hat{R}(\hat{\theta}(\alpha))$ be the empirically optimal among a candidate set $\mathcal{A}$. With probability at least $1-\delta$ we have,*

$$R(\hat{\theta}(\alpha^*)) \leq R(\hat{\theta}(\hat{\alpha})) + \frac{(3\lambda + \sqrt{4d\ln(n) + 8\ln(2/\delta)})}{\sqrt{n}}$$

*Proof.* Let $\epsilon = \frac{3}{\sqrt{n}}$ and $\Delta = \frac{\sqrt{2d\ln(3/\epsilon) + 2\ln(2/\delta)}}{\sqrt{n}}$. For any fixed $\alpha$, according to Hoeffding's inequality, we have,

$$\begin{aligned} P\{|\hat{R}(\hat{\theta}(\alpha)) - R(\hat{\theta}(\alpha))| > \Delta\} &\leq 2\exp(-\frac{N\Delta^2}{2}) \\ &= \frac{\delta}{(3/\epsilon)^d} \end{aligned} \tag{9}$$

Let $\mathcal{A}$ be an $\epsilon$-cover of $\mathbb{B}^d$, then we have

$$|\mathcal{A}| \leq (1 + 2/\epsilon)^d \leq (3/\epsilon)^d.$$

Then, using union bound over all elements of $\mathcal{A}$, with probability no less than $1-\delta$ we have

$$\forall \alpha \in \mathcal{A} : |\hat{R}(\hat{\theta}(\alpha)) - R(\hat{\theta}(\alpha))| \leq \sqrt{\frac{2d\ln(3/\epsilon) + 2\ln(2/\delta)}{n}} \tag{10}$$

Then, $\forall \alpha' \in \mathcal{A}$, we can obtain

$$\begin{aligned}
R(\hat{\theta}(\hat{\alpha})) &\geq \hat{R}(\hat{\theta}(\hat{\alpha})) - \sqrt{\frac{2d\ln(3/\epsilon) + 2\ln(2/\delta)}{n}} && (11)\\
&\geq \hat{R}(\hat{\theta}(\alpha')) - \sqrt{\frac{2d\ln(3/\epsilon) + 2\ln(2/\delta)}{n}} && (12)\\
&\geq R(\hat{\theta}(\alpha')) - 2\sqrt{\frac{2d\ln(3/\epsilon) + 2\ln(2/\delta)}{n}} && (13)
\end{aligned}$$

The first and third inequality holds since Eq.(10) and the second inequality holds since $\hat{\alpha} = \arg\min_{\alpha\in\mathcal{A}} \hat{R}(\hat{\theta}(\alpha))$.

According to the Lipschitz-continuity of $\ell$ w.r.t. to $\alpha$, $\forall \alpha \in \mathbb{B}^d$, we have

$$\begin{aligned}
R(\hat{\theta}(\alpha)) &\leq R(\hat{\theta}(\hat{\alpha})) + \lambda\epsilon + 2\sqrt{\frac{2d\ln(3/\epsilon) + \ln(2/\delta)}{n}} \\
&\leq R(\hat{\theta}(\hat{\alpha})) + \frac{(3\lambda + \sqrt{4d\ln(n) + 8\ln(2/\delta)})}{\sqrt{n}}
\end{aligned}$$

□