On the Provable Importance of Gradients for Autonomous Language-Assisted Image Clustering

Supplementary Material

1. Notations and Datasets

Here we summarize the important notations in Table 1 and the details of datasets in Table 2.

2. Derivation of Eq. (6) in Main Content

$$\begin{aligned}
\left\|\frac{\partial \ell\big(h(\tilde{\mathbf{r}}_i;\mathbf{W}^\star),\tilde{y}_i\big)}{\partial \mathbf{W}^\star}\right\|_F^2 &= \sum_{k=1}^{C}\left\|\frac{\partial \ell\big(h(\tilde{\mathbf{r}}_i;\mathbf{W}^\star),\tilde{y}_i\big)}{\partial \mathbf{w}_k^\star}\right\|_2^2 \\
&= \sum_{k=1}^{C}\|\tau\cdot[\tilde{\pi}_{ik}-\mathbb{I}(k=\tilde{y}_i)]\tilde{\mathbf{r}}_i\|_2^2 \\
&= \tau^2\cdot\sum_{k=1}^{C}\|(\tilde{\pi}_{ik}-\mathbb{I}(k=\tilde{y}_i))\|^2 \\
&= \tau^2\sum_{k\neq y_i}^{C}\tilde{\pi}_{ij}^2+\tau^2(\max_{j\in[C]}\tilde{\pi}_{ij}-1)^2 \\
&= \tau^2\cdot\left(\sum_{k\in[C]}\tilde{\pi}_{ik}^2+1-2\max_{j\in[C]}\tilde{\pi}_{ij}\right),
\end{aligned}$$

where the last two step holds due to the fact that $\tilde{y}_i = \arg\min_{j\in[C]}\ell\big(h(\tilde{\mathbf{r}}_i;\mathbf{W}^\star),j\big) = \arg\max_{k\in[C]}\tilde{\pi}_{ij}$.

3. Assumptions, Propositions and Lemmas

Assumption 1 (γ-smoothness)**.** *The loss function $\ell(\cdot,\cdot)$ (defined over $\mathcal{Z}\times\mathcal{Y}$) is γ-smooth such that, for any $\mathbf{z}\in\mathcal{Z}$, $y\in[C]$, and $\mathbf{W},\mathbf{W}'\in\mathcal{W}$,*

$$\big|\ell\big(h(\mathbf{z};\mathbf{W}),y\big)-\ell\big(h(\mathbf{z};\mathbf{W}'),y\big)\big|\le\gamma\,\|\mathbf{W}-\mathbf{W}'\|_F\,.$$

Assumption 2 ((ρ,ϵ,δ)-Boundness)**.** *The parameter space $\mathcal{W}\subset\big\{\mathbf{W}\in\mathbb{R}^{d\times C}:\|\mathbf{W}-\mathbf{W}_0\|_F\le\rho\big\}$ is within a Frobenius ball of radius ρ around the given point $\mathbf{W}_0$ that should satisfy the following properties:*

1. $\sup_{(\mathbf{z},y)\sim\mathbb{P}_{\mathcal{ZY}}}\ell\big(h(\mathbf{z};\mathbf{W}_0),y\big)=\epsilon$;
2. $\sup_{(\mathbf{z},y)\sim\mathbb{P}_{\mathcal{ZY}}}\big\|\partial\ell\big(h(\mathbf{z};\mathbf{W}_0),y\big)/\partial\mathbf{W}_0\big\|_F=\delta$.

Remark 1. *It can be easily checked that, for the classifier $h(\cdot;\mathbf{W})$ with softmax output function, the Frobenius norm of the Hessian matrix of the cross-entropy function with regard to the weight matrix $\mathbf{W}$ is bounded given a bounded parameter space. As a results, it is always true that the cross-entropy function is γ-smooth, therefore justifying the above assumptions.*

Proposition 1. *if Assumptions 1 and 2 holds, we have:*

$$\sup_{\mathbf{W}\in\mathcal{W}}\sup_{(\mathbf{z},y)\sim\mathbb{P}_{ZY}}\ell\big(h(\mathbf{z};\mathbf{W}),y\big)\le A,$$

where $A=\gamma\rho^2+\delta\rho+\epsilon$.

Proof. One can prove this by Mean Value Theorem of Integrals easily. □

Proposition 2 (Self-bounding Property)**.** *if Assumptions 1 and 2 holds, for any $\mathbf{W}\in\mathcal{W}$, we have:*

$$\big\|\partial\ell\big(h(\mathbf{z};\mathbf{W}),y\big)/\partial\mathbf{W}\big\|_F^2\le2\gamma\cdot\ell\big(h(\mathbf{z};\mathbf{W}),y\big).\quad(1)$$

Proof. The detailed proof of Proposition 2 can be found in Appendix B of Lei and Ying [1]. □

Proposition 3. *If Assumptions 1 and 2, for any empirical dataset $\mathcal{D}\sim\mathbb{P}_{ZY}^{|\mathcal{D}|}$, we have:*

$$\left\|\mathbb{E}_{(\mathbf{z},y)\in\mathcal{D}}\frac{\partial\ell\big(h(\mathbf{z};\mathbf{W}),y\big)}{\partial\mathbf{W}}\right\|_F^2\le2\gamma\mathbb{E}_{(\mathbf{z},y)\in\mathcal{D}}\ell\big(h(\mathbf{z};\mathbf{W}),y\big),$$

$$\left\|\mathbb{E}_{(\mathbf{z},y)\sim\mathbb{P}}\frac{\partial\ell\big(h(\mathbf{z};\mathbf{W}),y\big)}{\partial\mathbf{W}}\right\|_F^2\le2\gamma\mathbb{E}_{(\mathbf{z},y)\sim\mathbb{P}}\ell\big(h(\mathbf{z};\mathbf{W}),y\big),$$

where we use $\mathbb{P}$ as the abbreviation of $\mathbb{P}_{ZY}$ for brevity.

Proof. Given that the squared Frobenius norm $\|\cdot\|_F^2$ is a convex function, Jensen's inequality and Proposition 2 imply that

$$\begin{aligned}
\left\|\mathbb{E}_{(\mathbf{z},y)\in\mathcal{D}}\frac{\partial\ell\big(h(\mathbf{z};\mathbf{W}),y\big)}{\partial\mathbf{W}}\right\|_F^2 &\le\mathbb{E}_{(\mathbf{z},y)\in\mathcal{D}}\left\|\frac{\partial\ell\big(h(\mathbf{z};\mathbf{W}),y\big)}{\partial\mathbf{W}}\right\|_F^2 \\
&\le\mathbb{E}_{(\mathbf{z},y)\in\mathcal{D}}2\gamma\cdot\ell\big(h(\mathbf{z};\mathbf{W}),y\big) \\
&=2\gamma\cdot\mathbb{E}_{(\mathbf{z},y)\in\mathcal{D}}\ell\big(h(\mathbf{z};\mathbf{W}),y\big)
\end{aligned}$$

$$\begin{aligned}
\left\|\mathbb{E}_{(\mathbf{z},y)\sim\mathbb{P}}\frac{\partial\ell\big(h(\mathbf{z};\mathbf{W}),y\big)}{\partial\mathbf{W}}\right\|_F^2 &\le\mathbb{E}_{(\mathbf{z},y)\sim\mathbb{P}}\left\|\frac{\partial\ell\big(h(\mathbf{z};\mathbf{W}),y\big)}{\partial\mathbf{W}}\right\|_F^2 \\
&\le\mathbb{E}_{(\mathbf{z},y)\sim\mathbb{P}}2\gamma\cdot\ell\big(h(\mathbf{z};\mathbf{W}),y\big) \\
&=2\gamma\cdot\mathbb{E}_{(\mathbf{z},y)\sim\mathbb{P}}\ell\big(h(\mathbf{z};\mathbf{W}),y\big).
\end{aligned}$$

□

Lemma 1. *For any empirical dataset $\mathcal{D}\sim\mathbb{P}^N$ and $\mathbf{W}\in\mathcal{W}$, with the probability at least $1-\zeta>0$, we have:*

$$\begin{aligned}
&\mathbb{E}_{(\mathbf{z},y)\in\mathcal{D}}\ell\big(h(\mathbf{z};\mathbf{W}),y\big) \\
\le&\mathbb{E}_{(\mathbf{z},y)\sim\mathbb{P}}\ell\big(h(\mathbf{z};\mathbf{W}),y\big)+A\sqrt{\frac{\log(1/\zeta)}{2N}}.
\end{aligned}$$

Table 1. Main notations and their descriptions.

Notation	Description
Δ	Prompt template
$f_{\mathcal{X}}$	CLIP image encoder
$f_{\mathcal{T}}$	CLIP text encoder
$\mathcal{Z}, \mathcal{Y}, \mathcal{W}$	CLIP feature space, Pseudo-label space, Parameter space
$h, \mathbf{W}$	Classifier, Parameters of h
$\mathcal{D}_{\mathcal{X}}, N$	Unlabeled image dataset, The size of $\mathcal{D}_{\mathcal{X}}$
$\mathcal{D}_{\mathcal{T}}, M$	Unlabeled wild textual dataset, The size of $\mathcal{D}_{\mathcal{T}}$
$\mathcal{P}_{\mathcal{T}}(k), M_k$	the ground-truth set of positive semantics whose predicted pseudo-label is k, The size of $\mathcal{P}_{\mathcal{T}}(k)$
$\mathbf{x}$	Unlabeled image
$\mathbf{e}$	CLIP feature of unlabeled image
y	Image pseudo-label produced by k-means
$\tilde{\mathbf{t}}$	wild textual data
$\tilde{\mathbf{r}}$	CLIP feature of wild textual data
$\tilde{y}$	The predicted pseudo-label of wild textual data from h
T_k	The filtering threshold for wild text data whose predicted pseudo-label is k
$\|\cdot\|_F, \|\cdot\|_2$	Frobenius norm, L_2 norm

Table 2. A summary of datasets used for evaluation.

Dataset	Training Split	Test Split	# of Training	# of Test	# of Classes
STL-10	Train	Test	5000	8000	10
CIFAR-10	Train	Test	50000	10000	10
CIFAR-20	Train	Test	50000	10000	20
ImageNet-10	Train	Test	13000	500	10
ImageNet-Dogs	Train	Test	19500	750	15
DTD	Train+Val	Test	3760	1880	47
UCF-101	Train	Test	9537	3783	101
ImageNet-1K	Train	Test	1281167	50000	1000

Proof. Without loss of generality, let

$$\Omega(\mathbf{W}, \mathcal{D}) = \mathbb{E}_{(\mathbf{z},y)\in\mathcal{D}}\ell\big(h(\mathbf{z};\mathbf{W}), y\big),$$

$$\Omega(\mathbf{W}, \mathbb{P}) = \mathbb{E}_{(\mathbf{z},y)\sim\mathbb{P}}\ell\big(h(\mathbf{z};\mathbf{W}), y\big).$$

Given that

$$\mathbb{E}_{\mathcal{D}\sim\mathbb{P}^N}\left[\Omega(\mathbf{W}, \mathcal{D})\right] = \Omega(\mathbf{W}, \mathbb{P}),$$

Hoeffding's inequality implies that, with the probability at least $1-\zeta > 0$, we have:

$$\Omega(\mathbf{W}^\star, \mathcal{D}) - \Omega(\mathbf{W}^\dagger, \mathbb{P}) \le \Omega(\mathbf{W}^\dagger, \mathcal{D}) - \Omega(\mathbf{W}^\dagger, \mathbb{P}) \le A\sqrt{\frac{\log(1/\zeta)}{2N}}.$$

□

Lemma 2. *If Assumptions 1 and 2 holds, for any empirical dataset $\mathcal{D}\sim\mathbb{P}^N$ and $\mathbf{W}\in\mathcal{W}$, with the probability at least $1-\zeta>0$, we have:*

$$\begin{aligned} d_{\mathbf{W}}(\mathcal{D},\mathbb{P}) =& \Omega(\mathbf{W}, \mathcal{D}) - \Omega(\mathbf{W}, \mathbb{P}) \\ \le& A\sqrt{\frac{\log(1/\zeta)}{2N}} + U\sqrt{\frac{A(A-\epsilon)D}{N}}, \end{aligned}$$

$$\begin{aligned} -d_{\mathbf{W}}(\mathcal{D},\mathbb{P}) =& \Omega(\mathbf{W}, \mathbb{P}) - \Omega(\mathbf{W}, \mathcal{D}) \\ \le& A\sqrt{\frac{\log(1/\zeta)}{2N}} + U\sqrt{\frac{A(A-\epsilon)D}{N}}, \end{aligned}$$

where D is the dimension of the parameter space $\mathcal{W}$, U is a uniform constant, and

$$\Omega(\mathbf{W}, \mathcal{D}) = \mathbb{E}_{(\mathbf{z},y)\in\mathcal{D}}\ell\big(h(\mathbf{z};\mathbf{W}), y\big),$$

$$\Omega(\mathbf{W}, \mathbb{P}) = \mathbb{E}_{(\mathbf{z},y)\sim\mathbb{P}}\ell\big(h(\mathbf{z};\mathbf{W}), y\big).$$

Proof. Since it can be easily checked that

$$\mathbb{E}_{D\sim\mathbb{P}^N}\left[d_{\mathbf{W}}(\mathcal{D},\mathbb{P})\right]=0,$$

For any $\mathbf{W}\in\mathcal{W}$ and $\mathbf{W}'\in\mathcal{W}$, Proposition 2.6.1 and Lemma 2.6.8 in Vershynin [2] imply that

$$\begin{aligned}&\|d_{\mathbf{W}}(\mathcal{D},\mathbb{P})-d_{\mathbf{W}'}(\mathcal{D},\mathbb{P})\|_{\Phi}\\ \leq&\frac{u_0}{\sqrt{N}}\left\|\ell\big(h(\mathbf{z};\mathbf{W}),y\big)-\ell\big(h(\mathbf{z};\mathbf{W}'),y\big)\right\|_{L^\infty(\mathcal{Z}\times\mathcal{Y})},\end{aligned}$$

where $||\cdot||_{\Phi}$ is the sub-gaussian norm and u_0 is a uniform constant. Therefore, the Dudley's entropy integral [2] implies that

$$\begin{aligned}&\mathbb{E}_{D\sim\mathbb{P}^N}\sup_{\mathbf{W}\in\mathcal{W}}d_{\mathbf{W}}(\mathcal{D},\mathbb{P})\\ \leq&\frac{u_1}{\sqrt{N}}\int_0^{+\infty}\sqrt{\log\Upsilon(\mathcal{F},o,L^\infty)}\mathrm{d}o,\end{aligned}$$

where $\mathcal{F}=\left\{\ell\big(h(\mathbf{z}|\mathbf{W}),y\big):\mathbf{W}\in\mathcal{W}\right\}$, u_1 is anther uniform constant, and $\Upsilon(\mathcal{F},o,||\cdot||_{\max})$ is the covering number under the L^∞ norm. Due to the fact that

$$\begin{aligned}&\mathbb{E}_{D\sim\mathbb{P}^N}\sup_{\mathbf{W}\in\mathcal{W}}d_{\mathbf{W}}(\mathcal{D},\mathbb{P})\\ \leq&\frac{u_1}{\sqrt{N}}\int_0^{+\infty}\sqrt{\log\Upsilon(\mathcal{F},o,L^\infty)}\mathrm{d}o\\ &\frac{u_1}{\sqrt{N}}\int_0^{A}\sqrt{\log\Upsilon(\mathcal{F},o,L^\infty)}\mathrm{d}o\\ =&\frac{u_1}{\sqrt{N}}A\int_0^{1}\sqrt{\log\Upsilon(\mathcal{F},A\cdot o,L^\infty)}\mathrm{d}o,\end{aligned}$$

according to the McDiarmid's Inequality, for any $\mathbf{W}\in\mathcal{W}$, with the probability at least $1-\zeta>0$, we have either

$$\begin{aligned}&d_{\mathbf{W}}(\mathcal{D},\mathbb{P})\\ \leq&\frac{u_1}{\sqrt{N}}A\int_0^{1}\sqrt{\log\Upsilon(\mathcal{F},A\cdot o,L^\infty)}\mathrm{d}o+A\sqrt{\frac{\log(1/\zeta)}{2N}}\end{aligned}$$

or

$$\begin{aligned}&-d_{\mathbf{W}}(\mathcal{D},\mathbb{P})\\ \leq&\frac{u_1}{\sqrt{N}}A\int_0^{1}\sqrt{\log\Upsilon(\mathcal{F},A\cdot o,L^\infty)}\mathrm{d}o+A\sqrt{\frac{\log(1/\zeta)}{2N}}.\end{aligned}$$

Note that $\ell\big(h(\mathbf{z};\mathbf{W}),y\big)$ is $(\gamma\rho+\delta)$-Lipschitz with regard to $\mathbf{W}$ under $||\cdot||_F$. Then

$$\begin{aligned}&\Upsilon(\mathcal{F},A\cdot o,L^\infty)\\ \leq&\Upsilon(\mathcal{W},A\cdot o/(\gamma\rho+\delta),||\cdot||_{\mathrm{F}})\\ \leq&(1+\frac{2\rho(\gamma\rho+\delta)}{A\cdot o})^D\\ \leq&(1+\frac{2(A-\epsilon)}{A\cdot o})^D,\end{aligned}$$

such that

$$\begin{aligned}&\frac{u_1}{\sqrt{N}}A\int_0^{1}\sqrt{\log\Upsilon(\mathcal{F},A\cdot o,L^\infty)}\mathrm{d}o\\ =&\frac{u_1}{\sqrt{N}}A\int_0^{1}\sqrt{\log(1+\frac{2(A-\epsilon)}{A\cdot o})^D}\mathrm{d}o\\ =&\frac{u_1}{\sqrt{N}}A\int_0^{1}\sqrt{D\log(1+\frac{2(A-\epsilon)}{A\cdot o})}\mathrm{d}o\\ \leq&\frac{u_1}{\sqrt{N}}A\sqrt{D}\int_0^{1}\sqrt{\frac{2(A-\epsilon)}{A\cdot o}}\mathrm{d}o\\ =&2\frac{u_1}{\sqrt{N}}A\sqrt{D}\sqrt{\frac{2(A-\epsilon)}{A}}\\ =&U\sqrt{\frac{A(A-\epsilon)D}{N}},\end{aligned}$$

where $U=2\sqrt{2}u_1$. □

Lemma 3. *If Assumptions 1 and 2 hold, for any empirical dataset $\mathcal{D}\sim\mathbb{P}^N$ and $\mathcal{D}'\sim\mathbb{P}^{N'}$, with the probability at least $(1-\zeta)^3>0$, we have*

$$\begin{aligned}&\Omega(\mathbf{W}^\star,\mathcal{D}')\\ \leq&\Omega(\mathbf{W}^\dagger,\mathbb{P})+A\sqrt{\frac{\log(1/\zeta)}{2N'}}+U\sqrt{\frac{A(A-\epsilon)D}{N'}}\\ &+2A\sqrt{\frac{\log(1/\zeta)}{2N}}+U\sqrt{\frac{A(A-\epsilon)D}{N}},\end{aligned}$$

where D is the dimension of the parameter space $\mathcal{W}$, U is a uniform constant, and

$$\begin{aligned}\mathbf{W}^\star&=\arg\min_{\mathbf{W}\in\mathcal{W}}\mathbb{E}_{(\mathbf{z},y)\in\mathcal{D}}\ell\big(h(\mathbf{z};\mathbf{W}),y\big)\\ &=\arg\min_{\mathbf{W}\in\mathcal{W}}\Omega(\mathbf{W},\mathcal{D}),\\ \mathbf{W}^\dagger&=\arg\min_{\mathbf{W}\in\mathcal{W}}\mathbb{E}_{(\mathbf{z},y)\in\mathbb{P}}\ell\big(h(\mathbf{z};\mathbf{W}),y\big)\\ &=\arg\min_{\mathbf{W}\in\mathcal{W}}\Omega(\mathbf{W},\mathbb{P}),\\ \Omega(\mathbf{W}^\star,\mathcal{D}')&=\mathbb{E}_{(\mathbf{z},y)\in\mathcal{D}'}\ell\big(h(\mathbf{z};\mathbf{W}^\star),y\big).\end{aligned}$$

Proof. Given that

$$\begin{aligned}&\Omega(\mathbf{W}^\star,\mathcal{D}')-\Omega(\mathbf{W}^\dagger,\mathbb{P})\\ =&\Omega(\mathbf{W}^\star,\mathcal{D}')-\Omega(\mathbf{W}^\star,\mathbb{P})+\Omega(\mathbf{W}^\star,\mathbb{P})-\Omega(\mathbf{W}^\star,\mathcal{D})\\ &+\Omega(\mathbf{W}^\star,\mathcal{D})-\Omega(\mathbf{W}^\dagger,\mathbb{P})\\ \leq&\Omega(\mathbf{W}^\star,\mathcal{D}')-\Omega(\mathbf{W}^\star,\mathbb{P})+\Omega(\mathbf{W}^\star,\mathbb{P})-\Omega(\mathbf{W}^\star,\mathcal{D})\\ &+\Omega(\mathbf{W}^\dagger,\mathcal{D})-\Omega(\mathbf{W}^\dagger,\mathbb{P})\\ =&d_{\mathbf{W}^\star}(\mathcal{D}',\mathbb{P})-d_{\mathbf{W}^\star}(\mathbb{P},\mathcal{D})+\Omega(\mathbf{W}^\dagger,\mathcal{D})-\Omega(\mathbf{W}^\dagger,\mathbb{P}),\end{aligned}$$

Lemmas 1 and 2 imply that, with the probability at least $(1-\zeta)^3>0$, we have all of the following:

$$d_{\mathbf{W}}(\mathcal{D}',\mathbb{P})\leq A\sqrt{\frac{\log(1/\zeta)}{2N'}}+U\sqrt{\frac{A(A-\epsilon)D}{N'}},$$

$$-d_{\mathbf{W}^\star}(\mathbb{P},\mathcal{D}) \le A\sqrt{\frac{\log(1/\zeta)}{2N}} + U\sqrt{\frac{A(A-\epsilon)D}{N}}.$$

$$\Omega(\mathbf{W}^\dagger,\mathcal{D}) - \Omega(\mathbf{W}^\dagger,\mathbb{P}) \le A\sqrt{\frac{\log(1/\zeta)}{2N}}.$$

□

Lemma 4. *If Assumptions 1 and 2 hold, for any empirical dataset $\mathcal{D} \sim \mathbb{P}^N$ and $\mathcal{D}' \sim \mathbb{P}^{N'}$, with the probability at least $(1-\zeta)^3 > 0$, we have*

$$\begin{aligned}&\mathbb{E}_{(\mathbf{z},y)\in\mathcal{D}'}\left\|\partial\ell\big(h(\mathbf{z};\mathbf{W}^\star),\hat{y}\big)/\partial\mathbf{W}^\star\right\|_F^2\\ \le& 2\gamma\Omega(\mathbf{W}^\dagger,\mathbb{P}) + 2\gamma(A\sqrt{\frac{\log(1/\zeta)}{2N'}} + U\sqrt{\frac{A(A-\epsilon)D}{N'}}\\ &+2A\sqrt{\frac{\log(1/\zeta)}{2N}} + U\sqrt{\frac{A(A-\epsilon)D}{N}}),\end{aligned}$$

where D is the dimension of the parameter space $\mathcal{W}$, U is a uniform constant, $\hat{y} = \arg\min_{k\in[C]}\ell\big(h(\mathbf{z};\mathbf{W}^\star),k\big)$, and

$$\begin{aligned}\mathbf{W}^\star &= \arg\min_{\mathbf{W}\in\mathcal{W}}\mathbb{E}_{(\mathbf{z},y)\in\mathcal{D}}\ell\big(h(\mathbf{z};\mathbf{W}),y\big)\\ &= \arg\min_{\mathbf{W}\in\mathcal{W}}\Omega(\mathbf{W},\mathcal{D}).\end{aligned}$$

Proof. By Proposition 2 and Lemma 3, with the probability at least $(1-\zeta)^3 > 0$, we have

$$\begin{aligned}&\mathbb{E}_{(\mathbf{z},y)\in\mathcal{D}'}\left\|\partial\ell\big(h(\mathbf{z};\mathbf{W}),\hat{y}\big)/\partial\mathbf{W}\right\|_F^2\\ \le&\mathbb{E}_{(\mathbf{z},y)\in\mathcal{D}'}2\gamma\cdot\ell\big(h(\mathbf{z};\mathbf{W}),\hat{y}\big)\\ \le&\mathbb{E}_{(\mathbf{z},y)\in\mathcal{D}'}2\gamma\cdot\ell\big(h(\mathbf{z};\mathbf{W}),y\big)\\ =&2\gamma\Omega(\mathbf{W},\mathcal{D})\\ \le&2\gamma\Omega(\mathbf{W}^\dagger,\mathbb{P}) + 2\gamma(A\sqrt{\frac{\log(1/\zeta)}{2N'}} + U\sqrt{\frac{A(A-\epsilon)D}{N'}}\\ &+2A\sqrt{\frac{\log(1/\zeta)}{2N}} + U\sqrt{\frac{A(A-\epsilon)D}{N}}).\end{aligned}$$

□

Lemma 5. *Let us define the ground-truth set of positive semantics from the wild data as*

$$\mathcal{P}_\mathcal{T}(k) = \left\{\tilde{\mathbf{t}}_i \in \mathcal{D}_\mathcal{T} : \tilde{\mathbf{t}}_i \sim \mathbb{P}_{pos} \text{ and } k = \arg\max_{j\in[L]}\pi_{ij}\right\}$$

and $|\mathcal{P}_\mathcal{T}(k)| = B_k$. If Assumptions 1 and 2 hold, with the probability at least $(1-\zeta)^3 > 0$, we have the following:

$$\begin{aligned}&\mathbb{E}_{\tilde{\mathbf{t}}_i\in\mathcal{P}_\mathcal{T}(k)}\left\|\partial\ell\big(h(\tilde{\mathbf{r}}_i;\mathbf{W}^\star),\tilde{y}\big)/\partial\mathbf{W}^\star\right\|_F^2\\ \le&2\gamma\Omega(\mathbf{W}^\dagger,\mathbb{P}) + 2\gamma(A\sqrt{\frac{\log(1/\zeta)}{2B_k}} + U\sqrt{\frac{A(A-\epsilon)D}{B_k}}\\ &+2A\sqrt{\frac{\log(1/\zeta)}{2N}} + U\sqrt{\frac{A(A-\epsilon)D}{N}}),\end{aligned}$$

where D is the dimension of the parameter space $\mathcal{W}$, U is a uniform constant, $\tilde{y}_i = \arg\min_{k\in[C]}\ell\big(h(\tilde{\mathbf{t}}_i;\mathbf{W}^\star),k\big)$, and

$$\mathbf{W}^\star = \arg\min_{\mathbf{W}\in\mathcal{W}}\frac{1}{N}\sum_{i=1}^{N}\ell\big(h(\mathbf{e}_i;\mathbf{W}),y_i\big).$$

Proof. Lemma 4 directly implies this result. □

4. Proof of Theorem 1 in Main Content

Theorem 1. *Let us define the ground-truth set of positive semantics from the wild data as*

$$\mathcal{P}_\mathcal{T}(k) = \left\{\tilde{\mathbf{t}}_i \in \mathcal{D}_\mathcal{T} : \tilde{\mathbf{t}}_i \sim \mathbb{P}_{pos} \text{ and } k = \arg\max_{j\in[L]}\pi_{ij}\right\}$$

and $|\mathcal{P}_\mathcal{T}(k)| = B_k$. If Assumptions 1 and 2 hold, with the probability at least 0.97, we have the following:

$$\begin{aligned}ERR_{pos}(k) &\triangleq \frac{\left|\left\{\tilde{\mathbf{t}}_i\in\mathcal{P}_\mathcal{T}(k) : S(\tilde{\mathbf{t}}_i) > T_k\right\}\right|}{O_k}\\ &\le \frac{2\gamma}{T_k}\left[\min_{\mathbf{W}\in\mathcal{W}}\Omega(\mathbf{W}) + O(\sqrt{\frac{1}{B_k}}) + O(\sqrt{\frac{1}{N}})\right],\end{aligned}$$

where $O(1/N, 1/B_k) \ge 0$ is a uniform constant that is positively correlated to $1/N$ and $1/O_k$, and $\Omega(\mathbf{W}) = \mathbb{E}_{(\mathbf{z},y)\in\mathbb{P}_{ZY}}\ell\big(h(\mathbf{z};\mathbf{W}),y\big)$ denotes the expected risk.

Proof. Let S_k be the uniform random variable with $\mathcal{P}_\mathcal{T}(k)$ as the support and $S_k(\tilde{\mathbf{t}}_i) = \Phi(\tilde{\mathbf{t}}_i)$ for any $\tilde{\mathbf{t}}_i \in \mathcal{P}_\mathcal{T}(k)$, then by the Markov inequality, we have

$$\begin{aligned}\mathrm{ERR}_{\mathrm{pos}}(k) &\triangleq \frac{\left|\left\{\tilde{\mathbf{t}}_i\in\mathcal{P}_\mathcal{T}(k) : S(\tilde{\mathbf{t}}_i) > T_k\right\}\right|}{O_k}\\ &\le \frac{1}{T_k}\mathbb{E}_{\tilde{\mathbf{t}}_i\in\mathcal{P}_\mathcal{T}(k)}\left[S_k(\tilde{\mathbf{t}}_i)\right].\end{aligned}$$

As implied by Lemma 5, with the probability at least $(1-\zeta)^3 > 0$, we have the following:

$$\begin{aligned}&\mathbb{E}_{\tilde{\mathbf{t}}_i\in\mathcal{P}_\mathcal{T}(k)}\left[S_k(\tilde{\mathbf{t}}_i)\right]\\ =&\mathbb{E}_{\tilde{\mathbf{t}}_i\in\mathcal{P}_\mathcal{T}(k)}\left\|\partial\ell\big(h(\tilde{\mathbf{r}}_i;\mathbf{W}^\star),\tilde{y}\big)/\partial\mathbf{W}^\star\right\|_F^2\\ \le&2\gamma\Omega(\mathbf{W}^\dagger,\mathbb{P}) + 2\gamma(A\sqrt{\frac{\log(1/\zeta)}{2B_k}} + U\sqrt{\frac{A(A-\epsilon)D}{B_k}}\\ &+2A\sqrt{\frac{\log(1/\zeta)}{2N}} + U\sqrt{\frac{A(A-\epsilon)D}{N}}).\end{aligned}$$

If we set $\zeta = 0.01$, with the probability at least $(1-$

$0.01)^3 = 0.97$, we have:

$$\frac{\left|\{\tilde{\mathbf{t}}_i \in \mathcal{P}_{\mathcal{T}}(k) : S(\tilde{\mathbf{t}}_i) > T_k\}\right|}{B_k}$$

$$\leq \frac{2\gamma}{T_k}\Omega(\mathbf{W}^\dagger, \mathbb{P}) + \frac{2\gamma}{T_k}\underbrace{(A\sqrt{\frac{\log 10}{B_k}} + U\sqrt{\frac{A(A-\epsilon)D}{B_k}})}_{O(\sqrt{1/B_k})}$$

$$+\frac{2\gamma}{T_k}\underbrace{(2A\sqrt{\frac{\log 10}{N}} + U\sqrt{\frac{A(A-\epsilon)D}{N}})}_{O(\sqrt{1/N})}.$$

□

References

[1] Yunwen Lei and Yiming Ying. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations*, 2021. 1

[2] Roman Vershynin. High-dimensional probability, 2009. 3