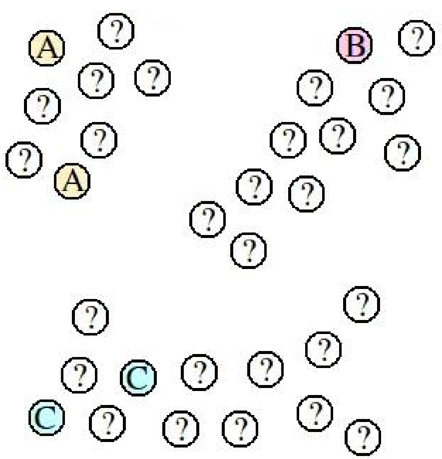
inductive learning / transductive learning

the number of categories. We focus on the transductive learning setting in this work, *i.e.*, all features together with the randomly sampled labels are constructed as training set. Let $S = \{\mathbf{x}_i, y_i\}_{i=1}^{m+u}$ be the set of instances where $m + u = n$. Without loss of generality (w.l.o.g.), let $\{y_i\}_{i=1}^{m}$ be the selected labels, our task is to predict the labels of samples $\{\mathbf{x}_i\}_{i=m+1}^{m+u}$ by a learner (model) trained on $\{\mathbf{x}_i\}_{i=1}^{m+u} \bigcup \{y_i\}_{i=1}^{m}$. This setting is widely adopted in node classification task (Yang et al., 2016; Kipf & Welling, 2017) where the training and test nodes are determined by a random partition.



Assumptions

1: 2-norm of the node feature x is bounded $\|\mathrm{x}\|_2 \leq c_X$

2: the parameters during learning process is bounded $\|W_h\| \leq c_W$

3: activation function is α-Holder Smooth

Assumption 3.4. Assume that the activation function $\sigma(\cdot)$ is $\tilde{\alpha}$-Hölder smooth. To be specific, let $P > 0$ and $\tilde{\alpha} \in (0, 1]$, for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$,

$$\|\sigma'(\mathbf{u}) - \sigma'(\mathbf{v})\|_2 \leq P\|\mathbf{u} - \mathbf{v}\|_2^{\tilde{\alpha}}.$$

Theorem 4.3

$$\text{transductive generalization gap}: \varepsilon_{\text{gen}} = \left| R_m(w) - R_{\text{u}}(w) \right|$$

$$R_m(w) = \frac{1}{m} \sum_{i=1}^{m} l(w; z_i)$$

$$R_u(w) = \frac{1}{u} \sum_{i=m+1}^{m+u} l(w; z_i)$$

Theorem 4.3 shows that the transductive generalization gap depends on
the training/test data size m/u,
network architecture related Lipschitz continuity constant L_F
and the number of iterations T

(a). If $\alpha \in (0, \frac{1}{2})$, *we have*

$$R_u(\mathbf{w}_1^{(T+1)}) - R_m(\mathbf{w}^{(T+1)})$$
$$= \mathcal{O}\left(L_{\mathcal{F}} \frac{(m+u)^{\frac{3}{2}}}{mu} \log^{\frac{1}{2}}(T) T^{\frac{1-2\alpha}{2}} \log\left(\frac{1}{\delta}\right)\right).$$

(b). If $\alpha = \frac{1}{2}$, *we have*

$$R_u(\mathbf{w}^{(T+1)}) - R_m(\mathbf{w}^{(T+1)})$$
$$= \mathcal{O}\left(L_{\mathcal{F}} \frac{(m+u)^{\frac{3}{2}}}{mu} \log(T) \log\left(\frac{1}{\delta}\right)\right).$$

(c). If $\alpha \in (\frac{1}{2}, 1]$, *we have*

$$R_u(\mathbf{w}^{(T+1)}) - R_m(\mathbf{w}^{(T+1)})$$
$$= \mathcal{O}\left(L_{\mathcal{F}} \frac{(m+u)^{\frac{3}{2}}}{mu} \log^{\frac{1}{2}}(T) \log\left(\frac{1}{\delta}\right)\right).$$

Proof Abstract

Step 1:From (El-Yaniv & Pechyony, 2007)

$$R_u(\mathbf{w}^{(T+1)}) \leq R_m(\mathbf{w}^{(T+1)}) + \mathcal{R}_{m+u}(\mathbf{w}) + c_0 Q\sqrt{\min(m,u)} + \sqrt{\frac{SQ}{2}\log\frac{2}{\delta}}$$

Step 2:Bound the Transductive Rademacher Complexity

$$\begin{aligned}\mathcal{R}_{m+u}(\mathbf{w}) \leq & 12\frac{(m+u)^{\frac{3}{2}}}{mu}\sqrt{d}\int_0^{L_{\mathcal{F}}R}\sqrt{\log\left(3L_{\mathcal{F}}R/r\right)}\,\mathrm{d}r \\ \leq & 12\frac{(m+u)^{\frac{3}{2}}}{mu}\sqrt{d}\left(\sqrt{\log 3}+\frac{3}{2}\sqrt{\pi}\right)L_{\mathcal{F}}R.\end{aligned}$$

Step 3:Combine with (Li & Liu, 2021)

$$\|\mathbf{w}_{t+1}\| = \begin{cases} \mathcal{O}\left(\log^{\frac{1}{2}}(T)T^{(1-2\alpha)/2}\log\left(\frac{1}{\delta}\right)\right) & \text{if } \alpha\in(0,\frac{1}{2}), \\ \mathcal{O}\left(\log(T)\log\left(\frac{1}{\delta}\right)\right) & \text{if } \alpha = 1/2, \\ \mathcal{O}\left(\log^{\frac{1}{2}}(T)\log\left(\frac{1}{\delta}\right)\right) & \text{if } \alpha\in(\frac{1}{2},1]. \end{cases}$$

Definition of Transductive Rademacher Complexity

$$\hat{\mathcal{R}}_S(\mathcal{G}) = E_\sigma\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\Sigma_{i=1}^{m}\sigma_i g(z_i)\right] \tag{2.1}$$

其中随机变量 $\sigma = (\sigma_1, ..., \sigma_m) \in U\{-1,1\}^m$

2.2 Transductive Rademacher Complexity 2007

We adapt the inductive Rademacher complexity to our transductive setting but generalize it a bit to also include "neutral" Rademacher values.

Definition 1 (Transductive Rademacher complexity) *Let $\mathcal{V} \subseteq \mathbb{R}^{m+u}$ and $p \in [0, 1/2]$. Let $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{m+u})^T$ be a vector of i.i.d. random variables such that*

$$\sigma_i \triangleq \begin{cases} 1 & \text{with probability} \quad p; \\ -1 & \text{with probability} \quad p; \\ 0 & \text{with probability} \quad 1-2p. \end{cases} \tag{1}$$

The transductive Rademacher complexity *with parameter p is*

$$R_{m+u}(\mathcal{V}, p) \triangleq \left(\frac{1}{m} + \frac{1}{u}\right) \cdot \mathbf{E}_{\boldsymbol{\sigma}}\left\{\sup_{\mathbf{v}\in\mathcal{V}} \boldsymbol{\sigma}^T \cdot \mathbf{v}\right\} .$$

Proof Abstract

Inequality 1 $$\mathcal{R}_{m+u}(\mathbf{w}) \le \left(\frac{1}{m}+\frac{1}{u}\right)\mathbb{E}_{\epsilon}\left[\sup_{\mathbf{w}\in B_R}\sum_{i=1}^{m+u}\epsilon_i\ell(\mathbf{w};z_i)\right]$$ From (El-Yaniv & Pechyony, 2007)

Inequality 2

$$\mathbb{E}_{\epsilon}\left[\sup_{\mathbf{w}\in B_R}\sum_{i=1}^{m+u}\epsilon_i\ell(\mathbf{w};z_i)\right] \le \mathbb{E}_{\epsilon}\left[\sup_{\mathbf{w}\in B_R}\left(\sum_{i=1}^{m+u}\epsilon_i(\ell(\mathbf{w};z_i)-\ell(\mathbf{w}^N;z_i)[\mathbf{w}])\right)\right]$$

$$+\sum_{j=1}^{N}\mathbb{E}_{\epsilon}\left[\sup_{\mathbf{w}\in B_R}\left(\sum_{i=1}^{m+u}\epsilon_i(\ell(\mathbf{w}^j;z_i)[\mathbf{w}]-\ell(\mathbf{w}^{j-1};z_i)[\mathbf{w}])\right)\right]$$

$$+\mathbb{E}_{\epsilon}\left[\sum_{i=1}^{m+u}\epsilon_i\ell(\mathbf{w}^{(1)};z_i)\right]$$

Inequality 3 (first item in inequality 2)

$$\mathbb{E}_{\epsilon}\left[\sup_{\mathbf{w}\in B_R}\left(\sum_{i=1}^{m+u}\epsilon_i(\ell(\mathbf{w};z_i)-\ell(\mathbf{w}^N;z_i)[\mathbf{w}])\right)\right] \le \left(\mathbb{E}_{\epsilon}\left[\sum_{i=1}^{m+u}\epsilon_i^2\right]\right)^{\frac{1}{2}}\left(\sup_{\mathbf{w}\in B_R}\sum_{i=1}^{m+u}(\ell(\mathbf{w};z_i)-\ell(\mathbf{w}^N;z_i)[\mathbf{w}])^2\right)^{\frac{1}{2}} \le (m+u)\alpha_N$$

Inequality 4 (second item in inequality 2 / single)

$$\mathbb{E}_{\epsilon}\left[\sup_{\mathbf{w}\in B_R}\left(\sum_{i=1}^{m+u}\epsilon_i(\ell(\mathbf{w}^j;z_i)[\mathbf{w}]-\ell(\mathbf{w}^{j-1};z_i)[\mathbf{w}])\right)\right] \le \sqrt{m+u}\sup_{\mathbf{w}\in B_R} d_{\mathcal{H}_S}(\mathbf{w}^j,\mathbf{w}^{j-1})\sqrt{2\log|T_j||T_{j-1}|}.$$

inequality 5 (last item in inequality 2)

$$\mathbb{E}_{\epsilon}\left[\sum_{i=1}^{m+u}\epsilon_i\ell(\mathbf{w}^{(1)};z_i)\right] \le \left(\sum_{i=1}^{m+u}\ell^2(\mathbf{w}^{(1)};z_i)\right)^{\frac{1}{2}} \le b_\ell\sqrt{m+u}.$$

Proof Abstract

inequality 6 (item in inequality 4)

$$\sup_{\mathbf{w}\in B_R} d_{\mathcal{H}_S}(\mathbf{w}^j, \mathbf{w}^{j-1}) \leq 3\alpha_j$$

inequality 7 (sum of inequality 4)

$$\sum_{j=1}^{N} \mathbb{E}_\epsilon \left[\sup_{\mathbf{w}\in B_R} \left(\sum_{i=1}^{m+u} \epsilon_i (\ell(\mathbf{w}^j; z_i)[\mathbf{w}] - \ell(\mathbf{w}^{j-1}; z_i)[\mathbf{w}]) \right) \right] \leq 12\sqrt{m+u} \int_{\alpha_{N+1}}^{\infty} \sqrt{\log \mathcal{N}(\alpha, \mathcal{H}_R, d_{\mathcal{H}_S})}\, \mathrm{d}\alpha$$

Inequality 8

$$d_{\mathcal{H}_S} \leq \left(\frac{1}{m+u} \sum_{i=1}^{m+u} \left[\max_{\mathbf{w},\tilde{\mathbf{w}}\in B_R, z\in\mathcal{Z}} \ell(\mathbf{w}; z_i) - \ell(\tilde{\mathbf{w}}; z_i) \right]^2 \right)^{\frac{1}{2}} \leq d_{\mathcal{H}_R}$$

Inequality 9

$$\log \mathcal{N}(r, \mathcal{H}_R, d_{\mathcal{H}_S}) \leq d \log \left(\frac{3L_{\mathcal{F}}R}{r} \right)$$

Combine all above

$$\begin{aligned} \mathcal{R}_{m+u}(\mathbf{w}) &\leq 12 \frac{(m+u)^{\frac{3}{2}}}{mu} \sqrt{d} \int_0^{L_{\mathcal{F}}R} \sqrt{\log (3L_{\mathcal{F}}R/r)}\, \mathrm{d}r \\ &\leq 12 \frac{(m+u)^{\frac{3}{2}}}{mu} \sqrt{d} \left(\sqrt{\log 3} + \frac{3}{2}\sqrt{\pi} \right) L_{\mathcal{F}}R. \end{aligned}$$

Inquality 2

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\mathbf{w}\in B_R}\sum_{i=1}^{m+u}\epsilon_i\ell(\mathbf{w};z_i)\right]\\
=&\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\mathbf{w}\in B_R}\left(\sum_{i=1}^{m+u}\left(\epsilon_i(\ell(\mathbf{w};z_i)-\ell(\mathbf{w}^N;z_i))[\mathbf{w}]+\sum_{j=1}^{N}\epsilon_i(\ell(\mathbf{w}^j;z_i)[\mathbf{w}]-\ell(\mathbf{w}^{j-1};z_i)[\mathbf{w}])+\epsilon_i\ell(\mathbf{w}^{(1)};z_i)\right)\right)\right]\\
\leq&\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\mathbf{w}\in B_R}\left(\sum_{i=1}^{m+u}\epsilon_i(\ell(\mathbf{w};z_i)-\ell(\mathbf{w}^N;z_i)[\mathbf{w}])\right)\right]+\sum_{j=1}^{N}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\mathbf{w}\in B_R}\left(\sum_{i=1}^{m+u}\epsilon_i(\ell(\mathbf{w}^j;z_i)[\mathbf{w}]-\ell(\mathbf{w}^{j-1};z_i)[\mathbf{w}])\right)\right]\\
&+\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sum_{i=1}^{m+u}\epsilon_i\ell(\mathbf{w}^{(1)};z_i)\right].
\end{aligned}
$$

item1: max loss between Final parameter and arbitrary w

item2: max loss during the learning process(Stability)

item3: model performance under the initialiazation parameter

Inequality 3

$$\mathbb{E}_{\epsilon}\left[\sup_{\mathbf{w}\in B_R}\left(\sum_{i=1}^{m+u}\epsilon_i(\ell(\mathbf{w};z_i)-\ell(\mathbf{w}^N;z_i)[\mathbf{w}])\right)\right]$$

$$\leq\left(\mathbb{E}_{\epsilon}\left[\sum_{i=1}^{m+u}\epsilon_i^2\right]\right)^{\frac{1}{2}}\left(\sup_{\mathbf{w}\in B_R}\sum_{i=1}^{m+u}(\ell(\mathbf{w};z_i)-\ell(\mathbf{w}^N;z_i)[\mathbf{w}])^2\right)^{\frac{1}{2}}\leq(m+u)\alpha_N.$$

Step1:Cauchy-Schwarz inequality $\sqrt{\left(\sum_{i=1}^{n}a_ib_i\right)^2}\leq\sqrt{\sum_{i=1}^{n}a_i^2}\cdot\sqrt{\sum_{i=1}^{n}\mathrm{b}_i^2}$

Step2: $$\sum_{i=1}^{m+u}\left(l(w,z_i)-l(w^N,z_i)[w]\right)^2=(m+u)d_{H_S}{}^2$$

$$\varepsilon_i{}^2\equiv1\qquad item1=\sqrt{m+u}$$

$$E_{\varepsilon}\leq(m+n)d_{H_S}$$

Inequality 4

$$\mathbb{E}_{\epsilon}\left[\sup_{\mathbf{w}\in B_R}\left(\sum_{i=1}^{m+u}\epsilon_i(\ell(\mathbf{w}^j;z_i)[\mathbf{w}]-\ell(\mathbf{w}^{j-1};z_i)[\mathbf{w}])\right)\right]\le\sqrt{m+u}\sup_{\mathbf{w}\in B_R} d_{\mathcal{H}_S}(\mathbf{w}^j,\mathbf{w}^{j-1})\sqrt{2\log|T_j||T_{j-1}|}.$$

Theorem 3.7 (Massart's lemma) *Let $\mathcal{A}\subseteq\mathbb{R}^m$ be a finite set, with $r=\max_{\mathbf{x}\in\mathcal{A}}\|\mathbf{x}\|_2$, then the following holds:*

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sup_{\mathbf{x}\in\mathcal{A}}\sum_{i=1}^{m}\sigma_i x_i\right]\le\frac{r\sqrt{2\log|\mathcal{A}|}}{m}, \tag{3.20}$$

where σ_is are independent uniform random variables taking values in $\{-1,+1\}$ and $x_1,\ldots,x_m$ are the components of vector $\mathbf{x}$.

x in Theorem 3.7 equals to the difference of l in ineaulity 4

r means max[the measure of x],Here we also have denoted the measure of diff l

$$\sum_{i=1}^{m+u}\left(l(w,z_i)-l(w^N,z_i)[w]\right)^2=(m+u)d_{H_S}{}^2$$

$|T_j||T_{j-1}|$ provides a upper bound of the number to cover the set of w(BR)

$\boxed{\sup_{\mathbf{w}\in B_R} d_{\mathcal{H}_S}(\mathbf{w}^j, \mathbf{w}^{j-1})}$ By the Minkowski inequality

$$
\begin{aligned}
&= \sup_{\mathbf{w}\in B_R} \left(\frac{1}{m+u} \sum_{i=1}^{m+u} \left[\ell(\mathbf{w}^j; z_i)[\mathbf{w}] - \ell(\mathbf{w}; z) + \ell(\mathbf{w}; z) - \ell(\mathbf{w}^{j-1}; z_i)[\mathbf{w}] \right]^2 \right)^{\frac{1}{2}} \\
&\leq \sup_{\mathbf{w}\in B_R} \left(\frac{1}{m+u} \sum_{i=1}^{m+u} \left[\ell(\mathbf{w}^j; z_i)[\mathbf{w}] - \ell(\mathbf{w}; z) \right]^2 \right)^{\frac{1}{2}} + \sup_{\mathbf{w}\in B_R} \left(\frac{1}{m+u} \sum_{i=1}^{m+u} \left[\ell(\mathbf{w}; z) - \ell(\mathbf{w}^{j-1}; z_i)[\mathbf{w}] \right]^2 \right)^{\frac{1}{2}} \\
&\boxed{= \sup_{\mathbf{w}\in B_R} d_{\mathcal{H}_S}(\mathbf{w}^j, \mathbf{w}) + \sup_{\mathbf{w}\in B_R} d_{\mathcal{H}_S}(\mathbf{w}, \mathbf{w}^{j-1})} \leq \alpha_j + \alpha_{j-1} = 3\alpha_j.
\end{aligned}
\tag{20}
$$

where ϵ_i is the standard Rademacher random variable. Now we give an upper bound of the Transductive Rademacher Complexity by Dudley's integral technique. Denote by $d_{\mathcal{H}_S}(\mathbf{w}, \widetilde{\mathbf{w}}) = \left(\frac{1}{m+u} \sum_{i=1}^{m+u} [\ell(\mathbf{w}; z_i) - \ell(\widetilde{\mathbf{w}}; z_i)]^2 \right)^{\frac{1}{2}}$. For $j \in \mathbb{N}$, let $\alpha_j = 2^{-j} M$ with $M = \sup_{\mathbf{w}\in B_R} d_{\mathcal{H}_S}(\mathbf{w}, \mathbf{w}^{(1)})$. Denote by T_j the minimal α_j-cover of B_R and $\ell(\mathbf{w}^j; z)[\mathbf{w}]$ the element in T_j that covers $\ell(\mathbf{w}; z)$. Specifically, since $\{\ell(\mathbf{w}^{(1)}; z)\}$ is a M-cover of B_R, we set $\ell(\mathbf{w}^0; z)[\mathbf{w}] = \ell(\mathbf{w}^{(1)}; z)$

$$
\sum_{j=1}^{N} \sqrt{m+u} \cdot 3a_j \cdot \sqrt{2\log|T_j||T_{j+1}|} \quad \leq \quad 6\sqrt{m+u} \sum_{j=1}^{N} a_j \cdot \sqrt{\log|T_j|}
$$

$$
\alpha_j = 2(\alpha_j - \alpha_{j+1}) \qquad = 12\sqrt{m+u} \sum_{j=1}^{N} (\alpha_j - \alpha_{j+1}) \sqrt{\log|T_j|}
$$

$$
= 12\sqrt{m+u} \sum_{j=1}^{N} (\alpha_j - \alpha_{j+1}) \sqrt{\log \mathcal{N}(\alpha_j, \mathcal{H}_R, d_{\mathcal{H}_S})}
$$

$$
\leq 12\sqrt{m+u} \int_{\alpha_{N+1}}^{\alpha_0} \sqrt{\log \mathcal{N}(\alpha, \mathcal{H}_R, d_{\mathcal{H}_S})}\, d\alpha \leq 12\sqrt{m+u} \int_{\alpha_{N+1}}^{\infty} \sqrt{\log \mathcal{N}(\alpha, \mathcal{H}_R, d_{\mathcal{H}_S})}\, d\alpha
$$

Inequality 5

$$\mathbb{E}_{\epsilon}\left[\sum_{i=1}^{m+u}\epsilon_i\ell(\mathbf{w}^{(1)};z_i)\right] \leq \left(\sum_{i=1}^{m+u}\ell^2(\mathbf{w}^{(1)};z_i)\right)^{\frac{1}{2}} \leq b_\ell\sqrt{m+u}.$$

Khintchine-Kahane inequality When p = 2

Let $\{\varepsilon_n\}_{n=1}^N$ be i.i.d. random variables with $P(\varepsilon_n = \pm 1) = \frac{1}{2}$ for $n = 1, \ldots, N$, i.e., a sequence with Rademacher distribution. Let $0 < p < \infty$ and let $x_1, \ldots, x_N \in \mathbb{C}$. Then

$$A_p\left(\sum_{n=1}^N |x_n|^2\right)^{1/2} \leq \left(\mathrm{E}\left|\sum_{n=1}^N \varepsilon_n x_n\right|^p\right)^{1/p} \leq B_p\left(\sum_{n=1}^N |x_n|^2\right)^{1/2}$$

Combine inqeuality 1/2/3/4/5/6/7:

$$\mathcal{R}_{m+u}(\mathbf{w}) \leq b_\ell\frac{(m+u)^{\frac{3}{2}}}{mu} + 12\frac{(m+u)^{\frac{3}{2}}}{mu}\int_0^\infty \sqrt{\log\mathcal{N}(r,\mathcal{H}_R,d_{\mathcal{H}_S})}\,\mathrm{d}r$$

Inequality 8

Propose a new measure of space H_R

$$d_{H_R}(\ell(\mathbf{w};\cdot),\ell(\widetilde{\mathbf{w}};\cdot)) = \max_{z\in\mathcal{Z}} |\ell(\mathbf{w};z) - \ell(\widetilde{\mathbf{w}};z)|$$

$$d_{\mathcal{H}_S}(\mathbf{w},\widetilde{\mathbf{w}}) = \left(\frac{1}{m+u}\sum_{i=1}^{m+u}\left[\ell(\mathbf{w};z_i) - \ell(\widetilde{\mathbf{w}};z_i)\right]^2\right)^{\frac{1}{2}}$$

$$d_{\mathcal{H}_S} \leq \left(\frac{1}{m+u}\sum_{i=1}^{m+u}\left[\max_{\mathbf{w},\widetilde{\mathbf{w}}\in B_R, z\in\mathcal{Z}} \ell(\mathbf{w};z_i) - \ell(\widetilde{\mathbf{w}};z_i)\right]^2\right)^{\frac{1}{2}} \leq d_{\mathcal{H}_R}$$

By the definition of covering number, we have $\mathcal{N}(r,\mathcal{H}_R,d_{\mathcal{H}_S}) \leq \mathcal{N}(r,\mathcal{H}_R,d_{\mathcal{H}_R})$

如果一个较小的度量空间可以被较大的度量空间覆盖，
那么较小空间的覆盖数不会超过较大空间的覆盖数

Inequality 9

Lipschitz contiguity of loss function

$$d_{\mathcal{H}_R} = \max_{z \in \mathcal{Z}} |\ell(\mathbf{w}; z) - \ell(\widetilde{\mathbf{w}}; z)| \le L_{\mathcal{F}} \|\mathbf{w} - \widetilde{\mathbf{w}}\|_2.$$

measure base on 2-norm

$$\mathcal{N}(r, \mathcal{H}_R, d_{\mathcal{H}_R}) \;\le\; \mathcal{N}\left(\tfrac{r}{L_{\mathcal{F}}}, B_R, d_2\right).$$

$$\log \mathcal{N}(r, B_R, d_2) \le d \log(3R/r)$$ From (Pisier, 1989)

$$\log N\left(r, H_R, d_{H_S}\right) \le \log N\left(r, H_R, d_{H_R}\right) \le \log N\left(\frac{r}{L_F}, B_R, d_2\right) \le d \log\left(\frac{3 L_F R}{r}\right)$$

$$\int_0^\infty \sqrt{\log \mathcal{N}(r, \mathcal{H}_R, d_{\mathcal{H}_S})}\mathrm{d}r = \int_0^{L_{\mathcal{F}} R} \sqrt{\log \mathcal{N}(r, \mathcal{H}_R, d_{\mathcal{H}_S})}\,\mathrm{d}r.$$

Calculation

$$R_u(\mathbf{w}^{(T+1)}) \leq R_m(\mathbf{w}^{(T+1)}) + \mathcal{R}_{m+u}(\mathbf{w}) + c_0 Q\sqrt{\min(m,u)} + \sqrt{\frac{SQ}{2}\log\frac{2}{\delta}}$$

$$\mathcal{R}_{m+u}(\mathbf{w}) \leq b_\ell \frac{(m+u)^{\frac{3}{2}}}{mu} + 12\frac{(m+u)^{\frac{3}{2}}}{mu}\int_0^\infty \sqrt{\log \mathcal{N}(r, \mathcal{H}_R, d_{\mathcal{H}_S})}\,\mathrm{d}r$$

Combining Eq. (23), Eq. (24), and Eq. (25) yields

$$\begin{aligned}\mathcal{R}_{m+u}(\mathbf{w}) &\leq 12\frac{(m+u)^{\frac{3}{2}}}{mu}\sqrt{d}\int_0^{L_{\mathcal{F}}R} \sqrt{\log\left(3L_{\mathcal{F}}R/r\right)}\,\mathrm{d}r \\ &\leq 12\frac{(m+u)^{\frac{3}{2}}}{mu}\sqrt{d}\left(\sqrt{\log 3} + \frac{3}{2}\sqrt{\pi}\right)L_{\mathcal{F}}R.\end{aligned} \tag{26}$$

$$R_{m+u}(w) \leq b_l \frac{(m+u)^{\frac{3}{2}}}{mu} + 12\frac{(m+u)^{\frac{3}{2}}}{mu}\cdot\sqrt{d}\cdot\frac{3\sqrt{\pi}}{2}\cdot L_F R$$

Result

Lemma 43 in (Li & Liu, 2021),Page 32

$$\|\mathbf{w}_{t+1}\| = \begin{cases} \mathcal{O}\left(\log^{\frac{1}{2}}(T)T^{(1-2\alpha)/2}\log\left(\frac{1}{\delta}\right)\right) & \text{if } \alpha \in (0, \frac{1}{2}), \\ \mathcal{O}\left(\log(T)\log\left(\frac{1}{\delta}\right)\right) & \text{if } \alpha = 1/2, \\ \mathcal{O}\left(\log^{\frac{1}{2}}(T)\log\left(\frac{1}{\delta}\right)\right) & \text{if } \alpha \in (\frac{1}{2}, 1]. \end{cases}$$

(a). If $\alpha \in (0, \frac{1}{2})$, we have

$$\begin{aligned} & R_u(\mathbf{w}_1^{(T+1)}) - R_m(\mathbf{w}^{(T+1)}) \\ =& \mathcal{O}\left(L_{\mathcal{F}}\frac{(m+u)^{\frac{3}{2}}}{mu}\log^{\frac{1}{2}}(T)T^{\frac{1-2\alpha}{2}}\log\left(\frac{1}{\delta}\right)\right) \end{aligned}$$

Conclusion:

Generalization gap depends on

Data size

Lipschitz constant(Network Architecture)

Iteritions

Remainings

Q1:From (El-Yaniv & Pechyony, 2007)

$$R_u(\mathbf{w}^{(T+1)}) \le R_m(\mathbf{w}^{(T+1)}) + \mathcal{R}_{m+u}(\mathbf{w}) + c_0 Q\sqrt{\min(m,u)} + \sqrt{\frac{SQ}{2}\log\frac{2}{\delta}}$$

$$\mathcal{R}_{m+u}(\mathbf{w}) \le \left(\frac{1}{m}+\frac{1}{u}\right)\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\mathbf{w}\in B_R}\sum_{i=1}^{m+u}\epsilon_i \ell(\mathbf{w}; z_i)\right]$$

Q2:From (Pisier, 1989)

$$\log \mathcal{N}(r, B_R, d_2) \le d\log(3R/r)$$

Q3:From (Li & Liu, 2021)

$$\|\mathbf{w}_{t+1}\| = \begin{cases} \mathcal{O}\left(\log^{\frac{1}{2}}(T)T^{(1-2\alpha)/2}\log\left(\frac{1}{\delta}\right)\right) & \text{if } \alpha \in (0, \frac{1}{2}), \\ \mathcal{O}\left(\log(T)\log\left(\frac{1}{\delta}\right)\right) & \text{if } \alpha = 1/2, \\ \mathcal{O}\left(\log^{\frac{1}{2}}(T)\log\left(\frac{1}{\delta}\right)\right) & \text{if } \alpha \in (\frac{1}{2}, 1]. \end{cases}$$

Theorem 4.5

Our second main result is high probability bounds of the gradients on training and test data.

Theorem 4.5. *Suppose Assumptions 3.1, 3.2, 3.4, 3.6, and 3.8 hold. Suppose that the learning rate $\{\eta_t\}$ satisfies $\eta_t = \frac{1}{t+t_0}$ such that $t_0 \ge \max\{(2P)^{1/\alpha}, 1\}$. For any $\delta \in (0,1)$, with probability $1-\delta$,*

(a). If $\alpha \in (0, \frac{1}{2})$, we have

$$\left\|\nabla R_m(\mathbf{w}^{(T+1)}) - \nabla R_u(\mathbf{w}^{(T+1)})\right\|_2 = \mathcal{O}\left(\frac{(m+u)^{\frac{3}{2}}}{mu}\log^{\frac{1}{2}}(T)T^{\frac{1-2\alpha}{2}}\log\left(\frac{1}{\delta}\right)\right).$$