Firstly, we give the definition of robustness, and give a series of lemmas including the robustness of MLP, the concentration inequality of polynomial distributions and so on. Then we prove Theorem 1, that is, derive the generalization bounds based on single-clicked historical behavior. Finally, we prove Theorem 2, which is to derive generalization bounds based on the multi-clicked historical behavior sequence.

We first define some symbols. The number of interest implicit state z is N_z. The envelope radius of set $V \subset \mathbb{R}^n$ is $R_0 \in \mathbb{R}^+$. That is, $\exists s \in \mathbb{R}^n$ and $\forall w \in V$, there will be $\|w - s\|_2 < R_0$ defined as $V \in \phi(R_0)$. r-covering set K means $\forall x, y \in K$, the $\|x - y\|_2 \leq r$. r-covering disjoint sets K_1 and K_2 mean that they are both r-covering sets and $K_1 \bigcap K_2 = \emptyset$.

A Preparation

Definition 1 *(Robustness) For a learning algorithm $\mathcal{A}$, and sample $s = \{x, y\}$. $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. $l(\mathcal{A}, s)$ represents the loss function $l(f_{\mathcal{A}}(x), y)$ of model $f_{\mathcal{A}}() : \mathcal{X} \to \mathcal{Y}$ optimized by algorithm $\mathcal{A}$. The domain of the sample s is $\{\mathcal{X}, \mathcal{Y}\}$. It can be partitioned into $L \in \mathbb{N}$ disjoint sets $K_1, K_2, \cdots, K_L$, and $\forall s_1, s_2 \in K_i, \forall i = 1, \cdots, L$, there satisfies $|l(\mathcal{A}, s_1) - l(\mathcal{A}, s_2)| \leq \epsilon$, then the algorithm $\mathcal{A}$ will be defined as (L, ϵ)-robust.*

This definition of robustness is from [3]. The robustness bound ϵ and the number of the disjoint sets L decided the robustness of the algorithm. Intuitively, the large L or large ϵ robust model will cause the fact that output is greatly influenced by input, which will cause the output gap between training set and testing set. Previous literature proves that the generalization bound of the learning algorithm consists of ϵ and L.

Lemma 1 *$\forall\ \Gamma \subset \mathbb{R}^d$ with the radius $R_0 \in \mathbb{R}$. That is, $\Gamma \in \phi(R_0)$. There are $L = (2R_0\sqrt{d}/r)^d$ r-covering disjoint sets of $\phi(R)$, $K_1, \cdots, K_L \subset \phi(R)$ and $\bigcup_{i=1}^{L} K_i = \phi(R_0)$. $\forall x, y \in K_i$, the $\|x - y\|_2 \leq r$.*

The Lemma 1 is a conclusion derived from the Definition 27.1 and Example 27.1 of [1]. It points out the number of subsets that can be divided into r diameter for a region with a radius of R_0.

Lemma 2 *The $s = \{x_1, \cdots, x_n, y^*\}$ is training sample of D layers MLP with ReLU nonlinear function with concatenation of $\{x_1, \cdots, x_n\}$ and $x_i \in \mathcal{X}_i \subset \mathbb{R}^d$ for $\forall i$ as input, $y^* \in \mathcal{Y} = \{0, 1\}$ as the label. The loss function of MLP is $|f(x_1, x_2, \cdots, x_n) - y^*|$. Its forward process is*

$$\begin{aligned} h^0 &= [x_1, \cdots, x_n], h^t = ReLU(W_t h^{t-1} + b_t), \\ y &= sigmoid(W_D h^{D-1} + b_D). \end{aligned} \tag{1}$$

$h^0 \in \mathbb{R}^{nd}$ and $y \in \mathbb{R}$. If $\mathcal{X}_i \in \phi(R_i)$, and the parameter sets $\{W_t, b_t\}$ for $\forall t$ have been trained by the alogrithm $\mathcal{A}$, the alogrithm is $\mathcal{A}$ is $(\|W\|_2^D r\sqrt{n}, 2(2R_{max}\sqrt{d}/r)^{nd})$-robust. Among them, R_{max} is the maximal value in $R_1, \cdots, R_n$, and $\|W\|_2 =$

$\sum_{t=1}^{D} \|W_t\|_2/D$. *In addition, any* $s = \{x, y\}, s' = \{x', y'\}$, *as long as* $y = y'$. $|l(f,s) - l(f,s')| \leq \|W\|_2^D \|x' - x\|_2$ *is satisfied.*

proof A.1 *For each* $\mathcal{X}_i \in \phi(R_i)$, *there are* $L_i = (2R_i\sqrt{d}/r)^d$ r*-covering disjoint sets* $K_1^i, \cdots, K_{L_i}^i \subset \mathcal{X}_i$ *according to Lemma 1.*

If two samples $s_1 = \{x^1, y^1\}$ *and* $s_2 = \{x^2, y^2\}$ *satisfy that* $y^1 = y^2 = y$ *and* $\forall i = 1, \cdots, n$, x_i^1, $x_i^2 \in K_{l_i}^i$, *we define that* s_1, $s_2 \in \mathcal{K}(l_1, \cdots, l_n, y)$. *There exist* $N = 2\prod_{i=1}^n L_i = 2\prod_{i=1}^n (2R_i\sqrt{d}/r)^d = 2(2R_{max}\sqrt{d}/r)^{nd}$ *disjoint sets* $\mathcal{K}()$

$\forall s_1$, $s_2 \in \mathcal{K}(l_1, \cdots, l_n, y)$. $\forall i$, $K_{l_i}^i$ *is* r*-covering set. Therefore, the*

$$\|x_i^1 - x_i^2\|_2 \leq r. \tag{2}$$

The difference between their loss functions is

$$\begin{aligned} &|l(f,s_1) - l(f,s_2)| = |y^1 - f(x^1)| - |y^2 - f(x^2)| \leq |f(x^1) - f(x^2)|| \\ &= |a_D(W_D h_1^D + b_D) - a_D(W_D h_2^D + b_D)| \leq \beta_D \|W_D\|_2 \|h_2^D - h_1^D\| \end{aligned} \tag{3}$$

The $\beta_t = max\frac{|a_t(x) - a_t(x')|}{|x - x'|}$ *for* $\forall x, x' \in domaina_t$, *and* $a_t()$ *mean activation function on* t*-th layer. For the* t*-th layer*

$$\begin{aligned} &\|h_2^t - h_2^t\|_2 \leq \|a_D(W^t h_1^{t-1} + b_t) - a_D(W^t h_2^{t-1} + b_t)\|_2 \\ &\leq \beta_t \|(W^t h_2^{t-1} - W^t h_2^{t-1})\|_2 \leq \beta \|W^t\|_2 \|h_1^{t-1} - h_2^{t-1}\|_2. \end{aligned} \tag{4}$$

The equation(4) is the recurrence relation between the 2-norm of the hidden state difference of layer t *and the 2-norm of the* $t-1$ *level hidden state. According to equation (2), there will be*

$$\|x^1 - x^2\|_2 = \sqrt{\sum_{i=1} \|x_i^1 - x_i^2\|_2^2} \leq r\sqrt{n}. \tag{5}$$

Therefore, (3) combination (4) and (5)the difference of the loss function is

$$\begin{aligned} &|l(f,s_1) - l(f,s_2)| \leq \beta_D \|W_D\|_2 \|h_2^D - h_1^D\|_2 \leq \\ &\prod_i \beta_i \|W_i\|_2 \|h_2^0 - h_1^0\|_2 \leq \prod_i \beta_i \|W_i\|_2 \|h_2^0 - h_1^0\|_2 \leq \\ &\prod_i \|W_i\|_2 \|x_2 - x_1\|_2 \leq r\sqrt{n} \prod_i \beta_i \|W_i\|_2. \end{aligned} \tag{6}$$

If $a_t() = ReLU()$, *then* $\beta_t = 1$ *and if* $a_t() = sigmoid()$, *then* $\beta_t < 1$. *Combining the mean inequality* $(\prod_i \beta_i \|W_i\|_2)^{1/D} \leq \frac{\sum_i \|W_i\|_2}{D}$ *with (6)*

$$|l(f,s_1) - l(f,s_2)| \leq (\sum_i^n \frac{\|W_i\|_2}{D})^D \|x^1 - x^2\|_2 \leq r\sqrt{n}(\sum_i^n \frac{\|W_i\|_2}{D})^D \tag{7}$$

According to the definition of the robustness. The D*-th with* $ReLU$ *activation function is* $(\|W\|_2^D r\sqrt{n}$,$2(2R_{max}\sqrt{d}/r)^{nd})$*-robust. The first inequality in (7) also obtained the additional conclusion of Lemma 2.*

Lemma 3 *(Breteganolle-Huber-Carol inequality)The random variable x belongs to domain$X$ with K disjoint sets $C_1, \cdots, C_K$. $\bigcup C_i$ =domain$X$, let n_i be the number of points that fall into the region C_i, n is the total number. there will be*

$$P(\sum_{i=1}^{K} |\frac{n_i}{n} - P(x \in C_i)| \geq \lambda) \leq 2^K \exp(\frac{-n\lambda^2}{2})$$

This is the concentration inequality of multinomial distribution from Proposition A6.6 of [2]

B Generalization bounds based on single-clicked historical behavior

Assumption 1 *A training sample $s = \{x_h, x_t, y^*\}$ consists of clicked items x_h, target item x_t and label y^*. The interest hidden states of clicked items and target item z_h and z_t are sampled from two distribution $P_o(z)$ and $P_t(z)$ respectively. The clicked items x_h and target item x_t are sampled from the conditional distribution $P(x|z)$. The domain of $P(x|z = i)$ is abbreviated as domainP_i. The label y^* of data is sampled from the set $\{0, 1\}$.*

Theorem 1 *Sampling N training data under Assumption 1, $s = \{x_h, x_t, y^*\}$. For the D layers MLP with ReLU $y = f(x_h, x_t)$ and the loss function $l(f, s) = |f(x_h, x_t) - y^*|$ defined the same as Lemma 2. With the probability $1 - \delta$, there will be*

$$\begin{aligned} &|E_s(l(f,s)) - \frac{1}{N}\sum_{i=1}^{N} l(f, s_i)| \leq \\ &\inf_r \sqrt{2}\|W\|_2^D r + l_M \sqrt{\frac{4N_z^2(R_{max}\sqrt{d}/r)^{2d}\ln 2 + 2ln(1/\delta)}{N}}. \end{aligned} \tag{8}$$

Among them, l_M is the maximum value of $l(f, s)$. d is dimension of each input vector, domain $P_i \in \phi(R_{max})$ for all $z = i$. $\|W\|_2$ is the average of 2-norm of all parameter matrices.

proof B.1 *According to Assumption 1, each sample $s = \{x_h, x_t, y\}$ exists a behavior interest hidden state z_h and target interest hidden state z_t.*

According to the Assumption 1, when $z_h = i$ and $z_t = j$ are determined, the domain of x_h and x_t is identified as domain$P_i \in \phi(R_i)$ and domain$P_j \in \phi(R_i)$. According to Lemma 1, domainP_i and domainP_i exist $L_i = (2R_i\sqrt{d}/r)^d$ and $L_j = (2R_j\sqrt{d}/r)^d$ r-covering disjoint sets $\{K_1^i, \cdots, K_{L_i}^i\}$ and $\{K_1^j, \cdots, K_{L_j}^j\}$.

For one sample $s = \{x_h, x_t, y^*\}$ from N training samples, we define the situation that $z_h = i$, $z_t = j$, $x_h \in K_\alpha^i$, $x_t \in K_\beta^j$, and $y^* = q \in \{0, 1\}$ as $s \in \Psi = \psi(i, j, \alpha, \beta, q)$. When i and j remain unchanged, there are $L_i \times L_j \times 2$ cases in this set. Therefore, the total number of the set $\psi(i, j, \alpha, \beta, q)$ is

$$N_\psi = \sum_{i=1}^{N_z}\sum_{j=1}^{N_z} 2L_iL_j \leq \sum_{i=1}^{N_z}\sum_{j=1}^{N_z} 2(2R_{max}\sqrt{d}/r)^{2d} \leq 2N_z^2(2R_{max}\sqrt{d}/r)^{2d} \tag{9}$$

The difference between expected loss and empirical loss is

$$
\begin{aligned}
&|E_s(l(f,s)) - \frac{1}{N}\sum_{i=1}^{N} l(f,s_i)| \leq \\
&|E_s(l(f,s)) - \frac{1}{N}\sum_{\Psi} n_\Psi E_{s_\Psi}(l(f,s_\Psi)|s_\Psi \in \Psi)| + \\
&|\frac{1}{N}\sum_{\Psi} n_\Psi E_{s_\Psi}(l(f,s_\Psi)|s_\Psi \in \Psi) - \frac{1}{N}\sum_{i=1}^{N} l(f,s_i)|
\end{aligned}
\tag{10}
$$

In equation (10) we add and subtract one item $\frac{1}{N}\sum_{\Psi} n_\Psi E_{s_\Psi}(l(f,s_\Psi)|s_\Psi \in \Psi)$. *Among them,* n_Ψ *is the number of samples belonging to* Φ *in* N *training samples.* $E_{s_\Psi}(l(f,s_\Psi)|s_\Psi \in \Psi)$ *is the expectation of the loss function when the sample belongs to* Ψ.

Based on the definition of expectation, $E_s(l(f,s)) = \sum_{\Psi} E_{s_\Psi}(l(f,s_\Psi)|s_\Psi \in \Psi)P(s_\Psi \in \Psi)$. $\frac{1}{N}\sum_{i=1}^{N} l(f,s_i) = \frac{1}{N}\sum_{\Psi}\sum_{s_i \in \Psi} l(f,s_i)$. *We put them in equation (10).*

$$
\begin{aligned}
&|E_s(l(f,s)) - \frac{1}{N}\sum_{i=1}^{N} l(f,s_i)| \leq \\
&\sum_{\Psi} |E_{s_\Psi}(l(f,s_i)|s_\Psi \in \Psi)||P(s_\Psi \in \Psi) - \frac{n_\Psi}{N}| + \\
&\frac{1}{N}\sum_{\Psi} |n_\Psi E_{s_\Psi}(l(f,s_\Psi)|s_\Psi \in \Psi) - \sum_{s_i \in \Psi} l(f,s_i)| \leq \\
&l_M \sum_{\Psi} |P(s_i \in \Psi) - \frac{n_\Psi}{N}| + \frac{1}{N}\sum_{\Psi}\sum_{s_i \in \Psi} |E_{s_\Psi}(l(f,s_\Psi)|s_\Psi \in \Psi) - l(f,s_i)|
\end{aligned}
\tag{11}
$$

l_M *is the maximal value of* $l(f,s)$. *According to the concentration inequalities for multinomial distribution in Lemma 3, in this situation,* $K = N_\Psi$. *Therefore,* $P(\sum_{\Psi} |P(s_i \in \Psi) - \frac{n_\Psi}{N}| \leq \lambda) \geq 1 - 2^{N_\Psi}\exp(\frac{-n\lambda^2}{2})$. *Let* $\delta = 2^{N_\Psi}\exp(\frac{-n\lambda^2}{2})$, *then* $\lambda = \sqrt{\frac{2N_\Psi ln2 + 2ln(1/\delta)}{2}}$. *It means that*

$$
\sum_{\Psi} |P(s_i \in \Psi) - \frac{n_\Psi}{N}| \leq \sqrt{\frac{2N_\Psi ln2 + 2ln(1/\delta)}{n}} \tag{12}
$$

with at least probability $1 - \delta$.

Lemma 2 and its proof show that if D-layer MLP input are bound in a region of $\|x - x'\|_2$, *then the difference of loss function will not exceed* $\|W\|_2^D \|x - x'\|_2$.

$\forall a, s \in \Psi$, $a = \{x_h^1, x_t^1, y^1\}$,$s = \{x_h^2, x_t^2, y^2\}$. x_h^1 *and* x_h^2 *belong to the same* r*-covering disjoint set.* $\|x_h^1 - x_h^2\| \leq r$,$\|x_t^1 - x_t^2\| \leq r$, *and* $y^1 = y^2$. *There will be* $|l(f,a) - l(f,s)| \leq \|W\|_2^D \sqrt{\|x_h^1 - x_h^2\|^2 + \|x_t^1 - x_t^2\|^2} \leq \sqrt{2}\|W\|_2^D r$,

$$
|E_{s_\Psi}(l(f,s_\Psi)|s_\Psi \in \Psi) - l(f,s_i)| \leq \sqrt{2}\|W\|_2^D r. \tag{13}
$$

We substitute (12) and (13) in (11).

$$|E_s(l(f,s)) - \frac{1}{N}\sum_{i=1}^{N} l(f,s_i)| \leq \sqrt{\frac{2N_\Psi ln2 + 2ln(1/\delta)}{n}} + \sqrt{2}\|W\|_2^D r \tag{14}$$

r-covering disjoint sets is a method of partitioning, so r can be adjusted arbitrarily. Therefore, the generalization error bound should be smaller than infimum under all r. According to (9) and (12), with at least probability $1-\delta$.

$$\begin{aligned} &|E_s(l(f,s)) - \frac{1}{N}\sum_{i=1}^{N} l(f,s_i)| \leq \\ &\inf_r\{\sqrt{\frac{4N_z^2(2R_{max}\sqrt{d}/r)^{2d}ln2 + 2ln(1/\delta)}{N}} + \sqrt{2}\|W\|_2^D r\} \end{aligned} \tag{15}$$

C Generalization bounds based on multi-clicked historical behavior

Assumption 2 *For simplicity, we assume all p periods are T in length. Sample $s = (\{x^1, \cdots, x^{T\times p}\}_h, x_t, y^*)$ consists of clicked items $\{x^1, \cdots, x^{T\times p}\}$, target item x_t and label y^*. The interest hidden states of target item z_t is sampled from $P_t(z)$, and target item x_t are sampled from the conditional distribution $P(x|z)$. Interest hidden state sequence of clicked items $\tilde{z} = \{z_1, \cdots, z_p\}$ is sampled from a set $\tilde{S}_z \subset \bigcup_1^p Z$. Each element of $\tilde{z}$ controls the user's click behavior in a period of length T. That is, items of historical click behaviors subsequence $\{x^{T\times(i-1)+1}, \cdots, x^{T\times i}\}_h$ is sampled from the conditional distribution $P(x|z_i)$. The element number of set S_z is N_S. The label y^* is sampled from $\{0,1\}$.*

Theorem 2 *If the N training samples are sampled from the distribution under Assumption 2. For the D layers MLP with ReLU, $f(x_h, x_t)$ and the loss function $l(f,s) = |f(x_h,x_t) - y^*|$, $s = \{\{x^1, \cdots, x^{T\times p}\}_h, x_t, y^*\}$. With the probability $1-\delta$, the difference between expected loss and empirical loss will be*

$$\begin{aligned} |E_s(l(f,s)) - \frac{1}{N}\sum_{i=1}^{N} l(f,s_i)| &\leq \inf_r\{\sqrt{Tp+1}\|W\|_2^D r \\ &+ l_M\sqrt{\frac{4N_z N_S(2R_{max}\sqrt{d}/r)^{d(Tp+1)}ln2 + 2ln(1/\delta)}{N}}\} \end{aligned} \tag{16}$$

Among them, l_M is the maximum value of $l(f,s)$. d is the dimension of each input vector. $domainP_i \in \phi(R_{max})$. $\|W\|_2$ is the average of 2-norm of all parameter matrices.

proof C.1 *According to Assumption 2, each sample $\{\{x^1, \cdots, x^{T\times p}\}_h, x_t, y\}$ exists a behavior interest hidden state sequence $\tilde{z}_h$ and target interest hidden state z_t.*

According to the Assumption 2, when $\tilde{z}_h = \{z_1, \cdots, z_p\}$ and $z_t = j$ are determined, the domain of elements in each prieod $x_{ph}^i = \{x^{1+T(i-1)}, \cdots, x^{T+T(i-1)}\}$

is $domainP_{z_i} \in \phi(R_{z_i})$*, and domain of* x_t *is* $domainP_j \in \phi(R_j)$*. According to Lemma 1,* $domainP_{z_i}$ *exists* $L_{z_i} = (2R_{z_i}\sqrt{d}/r)^d$ *r-covering disjoint sets* $\{K_1^{z_i}, \cdots, K_{L_{z_i}}^{z_i}\}$*, and* $domainP_j$ *exists* $L_j = (2R_j\sqrt{d}/r)^d$ *r-covering disjoint sets* $\{K_1^j, \cdots, K_{L_j}^j\}$*.*

For one sample $s = \{\{x^1, \cdots, x^{T\times p}\}_h, x_t, y^*\}$ *in training samples, we define the situation that* $\tilde{z}_h = \tilde{z}$*,* $z_t = j$*, each* x^g *in* $\{x^1, \cdots, x^{T\times p}\}_h$ *and* $x^g \in x_{ph}^i$*.* $x^g \in K_\alpha^{z_i}$*,* $x_t \in K_\beta^j$*, and* $y^* = q \in \{0, 1\}$ *as* $s \in \Psi = \psi(\tilde{z}, j, \alpha, \beta, q)$*. When* $\tilde{z}$ *and* j *remain unchanged, there are* $(\prod_{i=1}^p(\prod_1^T L_{z_i})) \times L_j \times 2$ *cases in this set. Therefore, The total number of the set* $\psi(\tilde{z}, j, \alpha, \beta, q)$ *is*

$$N_\psi = \sum_{\tilde{z}\in\tilde{S}_z}\sum_{j=1}^{N_z} 2(\prod_{i=1}^{p}(\prod_{1}^{T} L_{z_i}))L_j \leq 2N_zN_S(2R_{max}\sqrt{d}/r)^{d(Tp+1)} \tag{17}$$

The difference between expected loss and empirical loss is

$$\begin{aligned}
&|E_s(l(f,s)) - \frac{1}{N}\sum_{i=1}^{N} l(f,s_i)| \leq \\
&|E_s(l(f,s)) - \frac{1}{N}\sum_{\Psi} n_\Psi E_{s_\Psi}(l(f,s_\Psi)|s_\Psi \in \Psi)| + \\
&|\frac{1}{N}\sum_{\Psi} n_\Psi E_{s_\Psi}(l(f,s_\Psi)|s_\Psi \in \Psi) - \frac{1}{N}\sum_{i=1}^{N} l(f,s_i)|
\end{aligned} \tag{18}$$

In (18), we add and subtract one item $\frac{1}{N}\sum_\Psi n_\Psi E_{s_\Psi}(l(f,s_\Psi)|s_\Psi \in \Psi)$*. Among them,* n_Ψ *is the number of samples belonging to* Φ *in* N *training samples.* $E_{s_\Psi}(l(f,s_\Psi)|s_\Psi \in \Psi)$ *is the expectation of the loss function when the sample belongs to* Ψ*.*

Based on the definition of expectation, $E_s(l(f,s)) = \sum_\Psi E_{s_\Psi}(l(f,s_\Psi)|s_\Psi \in \Psi)P(s_\Psi \in \Psi)$*, and* $\frac{1}{N}\sum_{i=1}^N l(f,s_i) = \frac{1}{N}\sum_\Psi\sum_{s_i\in\Psi} l(f,s_i)$*. The equation (18) will be*

$$\begin{aligned}
&|E_s(l(f,s)) - \frac{1}{N}\sum_{i=1}^{N} l(f,s_i)| \leq \\
&\sum_{\Psi} |E_{s_\Psi}(l(f,s_i)|s_\Psi \in \Psi)||P(s_\Psi \in \Psi) - \frac{n_\Psi}{N}| + \\
&\frac{1}{N}\sum_{\Psi} |n_\Psi E_{s_\Psi}(l(f,s_\Psi)|s_\Psi \in \Psi) - \sum_{s_i\in\Psi} l(f,s_i)| \leq \\
&l_M\sum_{\Psi} |P(s_i \in \Psi) - \frac{n_\Psi}{N}| + \frac{1}{N}\sum_{\Psi}\sum_{s_i\in\Psi} |E_{s_\Psi}(l(f,s_\Psi)|s_\Psi \in \Psi) - l(f,s_i)|
\end{aligned} \tag{19}$$

l_M *is the maximal value of* $l(f,s)$*. According to the concentration inequalities for multinomial distribution in Lemma 3, in this situation,* $K = N_\Psi$*. Therefore,*

$P(\sum_{\Psi}|P(s_i \in \Psi) - \frac{n_\Psi}{N}| \leq \lambda) \geq 1 - 2^{N_\Psi}\exp(\frac{-n\lambda^2}{2})$. *Let* $\delta = 2^{N_\Psi}\exp(\frac{-n\lambda^2}{2})$, *then* $\lambda = \sqrt{\frac{2N_\Psi ln2 + 2ln(1/\delta)}{2}}$. *It means that*

$$\sum_{\Psi}|P(s_i \in \Psi) - \frac{n_\Psi}{N}| \leq \sqrt{\frac{2N_\Psi ln2 + 2ln(1/\delta)}{n}} \tag{20}$$

with at least probability $1 - \delta$.

$\forall a, s \in \Psi$, $a = \{\{x^1, \cdots, x^{Tp}\}_h^1, x_t^1, y^1\}$,$s = \{\{x^1, \cdots, x^{Tp}\}_h^2, x_t^2, y^2\}$. *elements of* $\{x^1, \cdots, x^{Tp}\}_h^1$ *and* $\{x^1, \cdots, x^{Tp}\}_h^2$ *belong to the same* r*-covering disjoint set. This is the same as the case in Proof B.1, the difference in the individual component of the input is not more than* r. *The* n *in Lemma 2 is* $T \times p + 1$. *Therefore,* $\|x^1 - x^2\|_2 \leq \sqrt{Tp+1}r$. *According to Lemma 2, there will be* $|l(f,a) - l(f,s)| \leq \sqrt{Tp+1}\|W\|_2^D r$,

$$|E_{s_\Psi}(l(f, s_\Psi)|s_\Psi \in \Psi) - l(f, s_i)| \leq \sqrt{Tp+1}\|W\|_2^D r. \tag{21}$$

We substitute it in (19) as

$$\begin{aligned} &|E_s(l(f,s)) - \frac{1}{N}\sum_{i=1}^{N} l(f, s_i)| \leq \\ &\sqrt{\frac{2N_\Psi ln2 + 2ln(1/\delta)}{n}} + \sqrt{Tp+1}\|W\|_2^D r. \end{aligned} \tag{22}$$

In addition, r *is the dividing radius of the input domain, and it can be adjusted arbitrarily. Therefore, the generalization error bound should be smaller than infimum under all* r. *According to (17), (20),with at least probability* $1 - \delta$

$$\begin{aligned} &|E_s(l(f,s)) - \frac{1}{N}\sum_{i=1}^{N} l(f, s_i)| \leq \\ &\inf_r\{\sqrt{\frac{4N_z N_S(2R_{max}\sqrt{d}/r)^{d(Tp+1)}ln2 + 2ln(1/\delta)}{N}} + \sqrt{Tp+1}\|W\|_2^D r\}. \end{aligned} \tag{23}$$

References

[1] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[2] Aad W. Van Der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

[3] Huan Xu. Robustness and generalization. *Machine Learning*, 86(3):391–423, 2012.