Mathematics Theory of Neural Network Models **Summer 2019**

Lecture 3: Concentration Inequalities Rademacher Complexity

Lecturer: Chao Ma *Scribe: Zehao Wang, Haoran Wang*

1 Some backgrounds in Concentration Inequalities

3.1 Markov Inequality

Theorem 3.1 (Markov Inequality) *Let X be a random variable that is non-negative with expectation $\mathrm{E}(X)$. Then, for every constant $a > 0$,*

$$\Pr(X \geq a) \leq \frac{\mathrm{E}(X)}{a}.$$

This inequality gives a tight upper bound of the tail probability of X when we only know the first order moment of X.

Proof:

$$\Pr(X \geq a) = \mathrm{E}(\mathbf{1}_{[X \geq a]}), \mathbf{1}_{[X \geq a]} \leq \frac{X}{a} \Rightarrow \Pr(X \geq a) \leq \frac{\mathrm{E}(X)}{a}$$

■

3.2 Chebyshev Inequality

When we know the first order moment and the second order moment of X, we can give a more specific bound of tail probability using Chebyshev Inequality.

Theorem 3.2 (Chebyshev Inequality) *For every constant $a > 0$,*

$$\Pr(|X - \mathrm{E}(X)| \geq a) \leq \frac{\mathrm{var}(X)}{a^2}.$$

This inequality can be derived from the Markov inequality easily.
Go a step further, think about the case when $\mathrm{E}(X), \mathrm{E}(X^2), ..., \mathrm{E}(X^r)$ is known, the straight forward upper bound will become:

$$\Pr(X \geq a) \leq \min_{k \in \{1,...,r\}} \frac{\mathrm{E}(X^k)}{a^k}.$$

3.3 Chernoff Inequality

The generic Chernoff bound requires all the moments of X, or the Moment Generative Function defined as:

$$M_X(t) = \mathrm{E}(\mathrm{e}^{tX}).$$

In fact these two conditions are equivalent, if we expand the function $M_x(t)$ we can get:

$$M_X(t) = \sum_{k=0}^{\infty} \frac{\mathrm{E}(X^k)}{k!} t^k.$$

Which means by expanding the Moment Generative Function we can get all the moments of X as the parameters in the series.[1]

Theorem 3.3 (Chernoff Inequality) *Based on Markov's inequality, for every $t > 0$:*

$$\Pr(X \geq a) \leq \frac{\mathrm{E}(\mathrm{e}^{tX})}{\mathrm{e}^{ta}}.$$

Proof: $\forall t > 0$

$$\Pr(X \geq a) = \Pr(\mathrm{e}^{tX} \geq \mathrm{e}^{ta}) \leq \frac{\mathrm{E}(\mathrm{e}^{tX})}{\mathrm{e}^{ta}}$$

The last step is exactly Markov's inequality. ■

3.4 Chernoff Bound

Theorem 3.4 *Let $X_1, \ldots, X_n$ be a set of n i.i.d. Bernoulli random variables, $EX = p$,then for all $\epsilon > 0$, the following inequality holds:*

$$P(\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq \epsilon) \leq e^{-nD_e^{(B)}(p+\epsilon||p)}$$

Theorem 3.5 *Let $X_1, \ldots, X_n$ be a set of n random variables satisfying $X_i \in [0,1]$ and $EX_i = p$ for $i = 1, \ldots, n$,then for all $\epsilon > 0$, the following inequality holds:*

$$P(\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq \epsilon) \leq e^{-nD_e^{(B)}(p+\epsilon||p)}$$

Proof: Exponent function is convex and use Jensen's inequality, for all t and $x \in (0,1)$ we can write:

$$Ee^{tx} \leq E(xe^t) + E((1-x)e^0) = pe^t + 1 - p$$

Using this inequality, we can prove the theorem like Chernoff Bound. ■

Theorem 3.6 *Let $X_1, \ldots, X_n$ be a set of n random variables satisfying $X_i \in [0,1]$ and $EX_i = p_i$ for $i = 1, \ldots, n$,then for all $\epsilon > 0$, the following inequality holds for $p = \frac{1}{n}\sum_{i=1}^{n} p_i$:*

$$P(\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq \epsilon) \leq e^{-nD_e^{(B)}(p+\epsilon||p)}$$

[1] We should notice that the moment-generating function of a real-valued distribution does not always exist, while the characteristic function does. And most distributions' moment-generating function is just to replace the *it* in the characteristic function with t. For example we consider $X \sim \mathrm{U}(a, b)$,it's characteristic function is $\frac{\mathrm{e}^{itb} - \mathrm{e}^{ita}}{it(b-a)}$ while the moment-generating function is $\frac{\mathrm{e}^{tb} - \mathrm{e}^{ta}}{t(b-a)}$

Proof: Logarithmic function is concave and use Jensen's inequality, for all t we can write:

$$\frac{\sum_{i=1}^{n} ln(1-p_i+p_ie^t)}{n} \leq ln(1-p+pe^t)$$

then

$$\prod_{i=1}^{n}(1-p_i+p_ie^t) \leq (1-p+pe^t)^n$$

Using this inequality, we can prove the theorem like Chernoff Bound. ∎

3.5 Hoeffding Inequality

Lemma 3.7 ***(Hoeffding's Lemma)****Let $X_1, ..., X_m$ be independent random variables with $E[X] = 0$ and $a \leq X \leq b$.Then for any $t > 0$,the following inequality holds:*

$$E[e^{tX}] \leq e^{\frac{t^2(b-a)^2}{8}}$$

Proof: Since $f(x) = e^{tx}$ is a convex function of x,the following holds:

$$e^{tx} \leq \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb}$$

Then,using $E[X] = 0$,

$$E[e^{tX}] \leq E[\frac{b-X}{b-a}e^{ta} + \frac{X-a}{b-a}e^{tb}] = \frac{b}{b-a}e^{ta} + \frac{-a}{b-a}e^{tb} = e^{\phi(t)}$$

where,

$$\phi(t) = \ln(\frac{b}{b-a}e^{ta} + \frac{-a}{b-a}e^{tb})$$

Taking derivative of $\phi(t)$,note that$\phi(0) = \phi'(0) = 0$,and that $\phi''(t) \leq \frac{(b-a)^2}{4}$.Thus by the second order expansion of function ϕ,there exists $\theta \in [0, t]$,such that:

$$\phi(t) = \phi(0) + t\phi'(0) + \frac{t^2}{2}\phi''(\theta) \leq t^2\frac{(b-a)^2}{8},$$

which completes the proof. ∎

Theorem 3.8 ***(Hoeffding's inequality)*** *Let $X_1, X_2, ...X_n$ be independent random variables where $X_i \in [a_i, b_i]$, and Let $\mu = \frac{\sum_{i=1}^{n} E[X_i]}{n}$,the following inequality holds:*

$$P(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \epsilon) \leq \exp(\frac{-2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2})$$

Proof: Let $S_n = \sum_{i=1}^{n} X_i$,Then for any $t \geq 0$,

$$\begin{aligned} P(S_n - E[S_n] \geq n\epsilon) &\leq e^{(-tn\epsilon)} E[e^{t(S_n - E[S_n])}] \\ &= \prod_{i=1}^{n} e^{-t\epsilon} E[e^{t(X_i - E[X_i])}] \\ &\leq \prod_{i=1}^{n} e^{-t\epsilon} e^{\frac{t^2 (b_i - a_i)^2}{8}} \qquad (Lemma 2.6) \\ &= e^{-tn\epsilon} e^{t^2 \sum_{i=1}^{n} \frac{(b_i - a_i)^2}{8}} \\ &\leq e^{(\frac{-2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2})} \end{aligned}$$

Where we chose $t = 4n\epsilon / \sum_{i=1}^{n}(b_i - a_i)^2$ to minimize the upper bound.And so,

$$P(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \epsilon) \leq \exp(\frac{-2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2})$$

■

3.6 McDiarmid Lemma

Theorem 3.9 *Assume* $\forall i, \forall x_1, x_2, ..., x_n, x_i^{'}, |f(x_1, ..., x_i, ..., x_n) - f(x_1, ..., x_i^{'}, ..., x_n) \mid \leq c_i$ *if* $x_1, x_2, ..., x_n$ *are independent random variables,then*

$$P(\mid f(x_1, ..., x_n) - E[f(x_1, ..., x_n]) \mid \geq \epsilon) \leq \exp(\frac{-2\epsilon^2}{\sum_{i=1}^{n} c_i^2})$$

Proof: Let $f(S)$ denote $f(x_1, ..., x_n)$
Define a sequence of random variables $V_k, k \in [1, m]$, as follows: $V = f(S) - E[f(S)]$, $V1 = E[V|x_1] - E[V]$, and for $k < 1$,

$$V_k = E[V|x_1, ..., x_k] - E[V|x_1, ..., x_{k-1}].$$

Note that $V = \sum_{k=1}^{m} V_k$. Furthermore, the random variable $E[V|x_1, ..., x_k]$ is a function of $x_1, ...x_k$. Conditioning on $x_1, ..., x_k$ and taking its expectation is therefore:

$$E[E[V|x_1, ..., x_k]|x_1, ..., x_{k-1}] = E[V|x_1, ..., x_{k-1}],$$

which implies $E[V|x_1, ..., x_k] = 0$. Thus, the sequence $(V_k)_{k\in[1,m]}$ is a martingale difference sequence. Next, observe that, since $E[f(S)]$ is a scalar, V_k can be expressed as follows:

$$V_k = E[f(S)|x_1, ..., x_k] - E[f(S)|x_1, ..., x_{k-1}]$$

Thus, we can define an upper bound W_k and lower bound U_k for V_k by:

$$W_k = sup_x E[f(S)|x_1, ..., x_{k-1}, x] - E[f(S)|x_1, ..., x_{k-1}]$$

$$U_k = inf_x E[f(S)|x_1, ..., x_{k-1}, x] - E[f(S)|x_1, ..., x_{k-1}]$$

Now, $\forall k \in [1, m]$, the following holds:

$$W_k - U_k = sup_{x,x'} E[f(S)|x_1, ..., x_{k-1}, x] - E[f(S)|x_1, ..., x_{k-1}, x^{'}] \leq c_k,$$

thus, $U_k \leq V_k \leq U_k + c_k$. In the view of these inequalities, we can apply Azuma's inequality to

$$V = \sum_{k=1}^{m} V_k,$$

which yields the desired inequality. ■

2 Rademacher complexity and its estimations

3.1 Definition

Definition 3.10 (Rademacher Complexity) *Let $\mathcal{F}$ be a collection of functions on X,$S = \{x_i\}_{i=1}^n$ be a sample of distribution D on X.Then we write*

$$Rad_n(\mathcal{F}) = E_\tau \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \tau_i f(x_i)$$

where τ_i takes the value ± 1 with probability $\frac{1}{2}$ for each.

Theorem 3.11 *For each $f \in \mathcal{F}$ we have $f \in [0,1]$,then, w.p. $1-\delta$ over the choice of S,we have*

$$\sup_{f \in \mathcal{F}} \left[E_D f(x) - \hat{E}_S f(x) \right] \leq 2Rad_n(\mathcal{F}) + 3\sqrt{\frac{\log \frac{1}{\delta}}{n}}$$

Proof: Let

$$\varphi(x_1, \cdots, x_n) = \sup_{f \in \mathcal{F}} \left[E_D f(x) - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right]$$

Then,

$$\begin{aligned}
& |\varphi(x_1, \cdots, x_n) - \varphi(x_1', \cdots, x_n)| \\
= & \left| \sup_{f \in \mathcal{F}} \left[E_D f(x) - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right] - \sup_{f \in \mathcal{F}} \left[E_D f(x) - \frac{1}{n} f(x_1') - \frac{1}{n} \sum_{i=2}^{n} f(x_i) \right] \right| \\
\leq & \left| \sup_{f \in \mathcal{F}} \frac{1}{n} \left(f(x_1) - f(x_1') \right) \right| \\
\leq & \frac{1}{n}
\end{aligned}$$

That is ,φ satisfies the condition of McDiarmid lemma with $c_i = \frac{1}{n}$,then we have

$$P(\varphi(x_1, \cdots, x_n) - E\varphi(x_1, \cdots, x_n) \geq t) \leq \exp(-2nt^2)$$

that is, w.p.$\geq 1-\delta$,

$$\varphi(x_1, \cdots, x_n) \leq E\varphi(x_1, \cdots, x_n) + \sqrt{\frac{\log \frac{1}{\delta}}{n}}$$

Then we estimate $E\varphi(x_1, \cdots, x_n)$.By definition we have

$$E\varphi(x_1, \cdots, x_n) = E_S \sup_{f\in\mathcal{F}} \left[E_D f(x) - \frac{1}{n}\sum_{i=1}^{n} f(x_i)\right]$$

Let $S' = \{x_i'\}_{i=1}^n$ be a sample i.i.d of S,then

$$\begin{aligned} E\varphi(x_1, \cdots, x_n) &= E_S \sup_{f\in\mathcal{F}} {}_{S'}\left[\hat{E}_{S'} f(x) - \hat{E}_S f(x)\right] \\ &\leq E_{S,S'} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} (f(x_i) - f(x_i')) \\ &= E_{\tau,S,S'} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \tau_i (f(x_i') - f(x_i)) \\ &\leq E_{\tau,S,S'} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \tau_i f(x_i) + E_{\tau,S,S'} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \tau_i f(x_i') \\ &= 2E_S Rad_n(\mathcal{F}) \end{aligned}$$

Similarly,let

$$\psi(x_1, \cdots, x_n) = E_\tau \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \tau_i f(x_i)$$

and we can verify that

$$|\psi(x_1, \cdots, x_n) - \psi(x_1', \cdots, x_n)| \leq \frac{1}{n}$$

Then,according to Mcdiarmid lemma we have,w.p$\geq 1 - \delta$

$$E_S Rad_n(\mathcal{F}) \leq Rad_n(\mathcal{F}) + \sqrt{\frac{\log\frac{1}{\delta}}{n}}$$

Then we finish the proof. ∎

As an example,we estimate the Rademacher complexity of the set

$$\mathcal{F} = \{w^T x : \|w\|_2 \leq W\}, \|x_i\|_2 \leq X$$

$$\begin{aligned} Rad_n(\mathcal{F}) &= \frac{1}{n} E_\tau \sup_{\|w\|_2\leq W} \sum_{i=1}^{n} \tau_i w^T x_i \\ &= \frac{1}{n} E_\tau \sup_{\|w\|_2\leq W} w^T \sum_{i=1}^{n} \tau_i x_i \\ &= \frac{1}{n} W E_\tau \left\|\sum_{i=1}^{n} \tau_i x_i\right\|_2 \\ &\leq \frac{W}{n}\sqrt{E_\tau \left\|\sum_{i=1}^{n} \tau_i x_i\right\|_2^2} \\ &= \frac{W}{n}\sqrt{\sum_{i=1}^{n} \|x_i\|_2^2} \leq \frac{WX}{\sqrt{n}} \end{aligned}$$

3.2 Properties of Rademacher Complexity

The following two properties are trivial

$$Rad(\mathcal{F} + f_0) = Rad(\mathcal{F})$$

$$Rad(\lambda\mathcal{F}) = \lambda Rad(\mathcal{F})$$

Theorem 3.12 *Let φ be a Lipschitz-continuous function with Lipschitz-constant L,and*

$$\varphi \circ \mathcal{F} = \{\varphi \circ f : f \in \mathcal{F}\}$$

then,

$$Rad(\varphi \circ \mathcal{F}) \leq LRad(\mathcal{F})$$

Proof:

$$\begin{aligned}
Rad(\varphi \circ \mathcal{F}) &= E \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \tau_i \varphi \circ f(x_i) \\
&= \frac{1}{n} E \left[\sup_{f\in\mathcal{F}} \left[\varphi \circ f(x_1) + \sum_{i=2}^{n} \tau_i \varphi \circ f(x_i) \right] + \sup_{f\in\mathcal{F}} \left[-\varphi \circ f(x_1) + \sum_{i=2}^{n} \tau_i \varphi \circ f(x_i) \right] \right] \\
&= \frac{1}{n} E \sup_{f,f'\in\mathcal{F}} \left[\varphi \circ f(x_1) + \sum_{i=2}^{n} \tau_i \varphi \circ f(x_i) - \varphi \circ f'(x_1) + \sum_{i=2}^{n} \tau_i \varphi \circ f'(x_i) \right] \\
&\leq \frac{1}{n} E \sup_{f,f'\in\mathcal{F}} \left[L|f(x_1) - f'(x_1)| + \sum_{i=2}^{n} \tau_i \varphi \circ f(x_i) + \sum_{i=2}^{n} \tau_i \varphi \circ f'(x_i) \right] \\
&= \frac{1}{n} E \sup_{f,f'\in\mathcal{F}} \left[L(f(x_1) - f'(x_1)) + \sum_{i=2}^{n} \tau_i \varphi \circ f(x_i) + \sum_{i=2}^{n} \tau_i \varphi \circ f'(x_i) \right] \\
&= \frac{1}{n} E \left[\sup_{f\in\mathcal{F}} \left[Lf(x_1) + \sum_{i=2}^{n} \tau_i \varphi \circ f(x_i) \right] + \sup_{f\in\mathcal{F}} \left[-Lf(x_1) + \sum_{i=2}^{n} \tau_i \varphi \circ f(x_i) \right] \right] \\
&= \frac{1}{n} E \sup_{f\in\mathcal{F}} \left[\tau_1 Lf(x_1) + \sum_{i=2}^{n} \tau_i \varphi \circ f(x_i) \right]
\end{aligned}$$

Repeat this process for index $i = 2, \cdots, n$,and we have

$$Rad(\varphi \circ \mathcal{F}) \leq E \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \tau_i Lf(x_i) = LRad(\mathcal{F})$$

■

3.3 Generalization to Subset of $\mathbb{R}^n$

For a subset A of $\mathbb{R}^n$,we write

$$Rad_n(A) = \frac{1}{n} E \sup_{a\in A} \tau^T a$$

Definition 3.13 (Covering number) *Let (S,ρ) be a metric space,$T \subset S, \alpha > 0.T' \subset T$ is an $\alpha-$cover of T,if $\forall x \in T, \exists x' \in T'$,s.t. $\rho(x,x') \leq \alpha$.We let the covering number be*

$$N(\alpha, T, \rho) = \min |T'|$$

where the minimum is taken over all α-covering T'.

Lemma 3.14 (Massart) *Assume that $|A| < \infty$,$r = max_{a \in A} \|a\|_2$,then $Rad(A) \leq \frac{r\sqrt{2\log|A|}}{n}$.*

Proof:

$$\begin{aligned}
\exp\left(\lambda E \max_{a\in A} \tau^T a\right) &\leq E\exp\left(\lambda \max_{a\in A} \tau^T a\right) \\
&\leq E\sum_{a\in A} \exp\left(\lambda \tau^T a\right) \\
&= \sum_{a\in A} E\exp\left(\lambda \sum_{i=1}^n \tau_i a_i\right) \\
&= \sum_{a\in A} \prod_{i=1}^n E\exp\left(\lambda \tau_i a_i\right) \\
&\leq \sum_{a\in A} \prod_{i=1}^n \exp\frac{(2\lambda a_i)^2}{8} \\
&\leq |A| \exp\frac{r^2\lambda^2}{2}
\end{aligned}$$

In the 5th line we have adopted Hoeffding inequality.Take the logarithm of both sides,we have

$$\max_{a\in A} \tau^T a \leq \frac{r^2\lambda}{2} + \frac{1}{\lambda}\log|A|$$

As $\lambda > 0$ is chosen arbitrarily, we can choose proper λ to minimize the right side.Then we have

$$Rad(A) \leq \frac{r\sqrt{2\log|A|}}{n}$$

■

Theorem 3.15

$$Rad(A) \leq \inf_{\alpha>0}\left\{\max_{a\in A}\|a\|_2 \frac{\sqrt{2\log N(\sqrt{n}\alpha, A, l_2)}}{n} + \alpha\right\}$$

Proof: For $\alpha > 0$,let A' be a $\sqrt{n}\alpha-$cover of A, $|A'| = N(\sqrt{n}\alpha, A, l_2)$.

$$\begin{aligned}
Rad(A) &= \frac{1}{n}E\sup_{a\in A}\tau^T a \\
&\leq \frac{1}{n}E\sup_{a'\in A'}\tau^T a' + \frac{1}{n}E\sup_{a\in A}\tau^T(a-a') \\
&\leq \max_{a\in A}\|a\|_2 \frac{\sqrt{2\log|A'|}}{n} + \alpha
\end{aligned}$$

■

Theorem 3.16 *Let A be a bounded subset of $\mathbb{R}^n$, then*

$$Rad(A) \le 4 \int_0^{+\infty} \frac{\sqrt{2 \log N(\alpha, A, l_2)}}{n} d\alpha$$

Proof: Let $r = \max_{a \in A} \|a\|_2$, $\hat{A}^j$ be a $2^{-j}r$-cover of A which has the least elements, and for fixed $a \in A$, let $\hat{a}^j$ be an element in $\hat{A}^j$ s.t. $\|a - \hat{a}^j\| \le 2^{-j}r$.We can choose $\hat{A}^0$ to be $\{0\}$.
For any sufficiently big integer N, we have

$$\begin{aligned}
Rad(A) &= \frac{1}{n} E \sup_{a \in A} \tau^T a \\
&\le \frac{1}{n} E \sup_{a \in \hat{A}^N} \tau^T a + \frac{1}{n} E \sup_{a \in A} \tau^T (a - \hat{a}^N) \\
&\le \frac{1}{n} E \sup_{a \in \hat{A}^{N-1}} \tau^T a + \frac{1}{n} E \sup_{a \in \hat{A}^N} \tau^T (a - \hat{a}^{N-1}) + \frac{1}{n} E \sup_{a \in A} \tau^T (a - \hat{a}^N) \\
&\le \cdots \le \sum_{j=1}^{N} \frac{1}{n} E \sup_{a \in \hat{A}^j} \tau^T (a - \hat{a}^{j-1}) + \frac{1}{n} E \sup_{a \in A} \tau^T (a - \hat{a}^N) \\
&\le \sum_{j=1}^{N} \frac{2^{-j+1} r}{n} \sqrt{2 \log N(2^{-j}r, A, l_2)} + \frac{2^{-N} r}{\sqrt{n}} \\
&\le 4 \int_0^{+\infty} \frac{\sqrt{2 \log N(\alpha, A, l_2)}}{n} d\alpha + \frac{2^{-N} r}{\sqrt{n}}
\end{aligned}$$

Let $N \to \infty$,we get the inequaliy to be proved. ■

References

[1] MOHRI.M, ROSTAMIZADEH.A and TALWALKAR.A, Foundations of Machine Learning, MIT Press (2012)
[2] https://en.wikipedia.org/wiki/Azuma%27s_inequality
[3] Hang Li, Statistical Learning Method, Tsinghua University Press(2012)
[4] Ian Goodfellow and Yoshua Bengio and Aaron Courville, Deep Learning, MIT Press(2016)