Uniform Law of Large Numbers

Tianwei Gao

January 21, 2025

1 Uniform Law of Large Numbers

1.1 Main theorem

Definition 1.1. *Fix a probability space $(\mathcal{X}, \Omega, \mathbb{P})$, and $\mathcal{F}$ a class of integrable functions, $\{X_i\}$ are i.i.d. random variables following distribution F.*

1. *We define $\|F_n - F\|_{\mathcal{F}}$ to be the random variable*

$$\|F_n - F\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} f(X_i) - f(X)|.$$

2. *Denote by $\mathcal{F}(x_1^n)$ the subset*

$$\mathcal{F}(x_1^n) := \{(f(x_1), f(x_2), \dots, f(x_n)) \in \mathbb{R}^n | f \in \mathcal{F}\} \subset \mathbb{R}^n,$$

we define the empirical Radamacher complexity to be

$$\mathcal{R}(\mathcal{F}(x_1^n)/n)$$

where $\mathcal{R}$ is the usual Rademacher complexity.

3. *The Rademacher complexity of function class $\mathcal{F}$ is defined by*

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_X[\mathcal{R}(\mathcal{F}(x_1^n)/n)].$$

4. *We say that $\mathcal{F}$ is a Glivenko–Cantelli class for the distribution F if*

$$\|F_n - F\|_{\mathcal{F}} \to 0 \text{ as } n \to \infty,$$

in probability.

Remark 1.1. *When the $\mathcal{F}$ has one single function, then this is just a common convergence law. When proving some technical lemma, we can always take this as a starting point.*

Lemma 1.1. *The following inequalities holds,*

1.

$$\sup_{f \in \mathcal{F}} \mathbb{E}_X[f(X)] \leq \mathbb{E}[\sup_f f(X)].$$

2. *For any non-decrease and convex function Φ, we have*

$$\sup_{f \in \mathcal{F}} \Phi(\mathbb{E}[f(X)]) \leq \mathbb{E}[\Phi(\sup_f f(X))]$$

Proof. We prove 2 here. For any $f \in \mathcal{F}$,

$$\begin{aligned}\Phi(\mathbb{E}[f(X)]) &\leq \mathbb{E}[\Phi(f(X))] \\ &\leq \mathbb{E}[\Phi(\sup_f f(X))].\end{aligned}$$

Taking sup on both sides leads to the inequality. □

To bound $\|F_n - F\|_{\mathcal{F}}$, we first bound the expectation

$$\mathbb{E}_X[\|F_n - F\|_{\mathcal{F}}]$$

as follows:

Lemma 1.2.

$$\mathbb{E}_X[\|F_n - F\|_{\mathcal{F}}] \leq 2\mathcal{R}_n(\mathcal{F})$$

Proof. Consider $\{Y_i\}_{i=1}^n$ are i.i.d also follows F and independent of X_1^n, then

$$\begin{aligned}
\mathbb{E}_{X_1^n}\left[\sup_f |\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X)|\right] &= \mathbb{E}_{X_1^n}\left[\sup_f |\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_{Y_1^n}(\frac{1}{n}\sum_{i=1}^n f(Y_i))|\right] \\
&\leq \mathbb{E}_{X_i,Y_i}\left[\frac{1}{n}\sup_f |\sum_{i=1}^n (f(X_i) - f(Y_i))|\right] \\
&= \mathbb{E}_{X_i,Y_i,\varepsilon_i}\left[\frac{1}{n}\sup_f |\sum_{i=1}^n \varepsilon_i(f(X_i) - f(Y_i))|\right] \\
&= 2\mathbb{E}_{X_i,\varepsilon_i}\left[\frac{1}{n}\sup_f |\sum_{i=1}^n \varepsilon_i(f(X_i)|\right] \\
&= 2\mathcal{R}_n(\mathcal{F})
\end{aligned}$$

□

Now we come to the proof of the following main theorem:

Theorem 1.3. *Assume that $\mathcal{F}$ is uniformly bounded by $b > 0$, for any $\delta > 0$,*

$$\mathbb{P}(\|F_n - F\|_{\mathcal{F}} \geq 2\mathcal{R}_n(\mathcal{F}) + \delta) \leq \exp(-\frac{n\delta^2}{2b^2}).$$

Proof. With the above lemma, we only need to prove the concentration part.

We set $G(x_i^n) : \mathbb{R}^d \to \mathbb{R}$ to be

$$G(x_1^n) = \sup_f |\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}f(x)|,$$

we can see that, for $x_1^n, y_1^n \in \mathbb{R}^n$, we have

$$\begin{aligned}
|\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}f(x)| - G(y_1^n) &\leq |\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}f(x)| - |\frac{1}{n}\sum_{i=1}^n f(y_i) - \mathbb{E}f(X)| \\
&\leq \frac{1}{n}\sum_{i=1}^n |f(x_i) - f(y_i)| \\
&\leq 2b,
\end{aligned}$$

thus

$$G(x_1^n) - G(y_1^n) \leq 2b,$$

by symmetrization,

$$|G(x_1^n) - G(y_1^n)| \leq 2b,$$

thus we have the concentration formula for the bounded random variable $G(X_1^n)$

$$\mathbb{P}(G(X_1^n) - \mathbb{E}G(X_1^n) \geq \delta) \leq \exp(-\frac{n\delta^2}{2b^2}).$$

Now with $\mathbb{E}G(X_1^n) \leq 2\mathcal{R}_n(\mathcal{F})$, we finished the proof. □

1.2 Rademacher Complexity and VC dimension

To make the bound obtained in main theorem meaningful, we need to bound $\mathcal{R}_n(\mathcal{F})$.

Intuitively speaking, for a random variable X and its i.i.d samples X_1^n, the larger the set $\mathcal{F}(X_1^n)$ is, the larger the complexity is.

Here is a first thought based on this intuition:

Proposition 1.4. *Assume that we have some $p > 0$, such that for any $x_1^n \in \mathbb{R}^n$,*

$$\mathcal{F}(x_1^n) \subset B(0, D),$$
$$\#(\mathcal{F}(x_1^n)) \leq (n+1)^p.$$

Then

$$\mathcal{R}(\mathcal{F}(x_1^n)/n) \leq \frac{C_1 D}{n} \sqrt{p \log(n+1)}$$

Proof. After rescaling, we can see that this amounts to say that a set $T \subset B(0,1)$ with $\#(T) \leq (n+1)^p$,

$$\mathcal{R}(T) \leq C_1 \sqrt{p \log(n+1)}.$$

Note that

$$\begin{aligned} \mathcal{R}(T) &= \mathbb{E}_\varepsilon \sup_{\theta \in T} \langle \epsilon, \theta \rangle \\ &= \mathbb{E}_\varepsilon \max_{\theta \in T} Y_\theta, \end{aligned}$$

where $Y_\theta := \langle \varepsilon, \theta \rangle$ is a random variable defined on $\mathbb{H}^n$. Note that by the following lemma, Y_θ are zero-mean, bound, 1-sub-gaussian variables.

Lemma 1.5. $\mathbb{Y}_\theta$ *is* 1*-sub-gaussian.*

Proof.

$$\begin{aligned} \mathbb{E}[\exp(\lambda Y_\theta)] &= \prod_{i=1}^n \mathbb{E}[\exp(\lambda \theta_i \varepsilon_i)] \\ &\leq \prod_{i=1}^n \exp(\frac{\lambda^2 \theta^2}{2}) \\ &= \exp(\frac{\lambda^2}{2}). \end{aligned}$$

□

Now the result follows from a concentration argument. □

For function class satisfies the condition

$$\#\mathcal{F}(x_1^n) \leq (n+1)^p,$$

the function class $\mathcal{F}$ is said to have polynomial discrimination of order p.

Definition 1.2.

1. *Given a binary-valued function class $\mathcal{F}$, we say that $(x_1, \ldots, x_n)$ is shattered by $\mathcal{F}$ if*

$$\#(\mathcal{F}(x_1^n)) = 2^n.$$

The VC-dimension of $\mathcal{F}$ is to be the largest k such that there is a set of cardinal k, that is shattered by $\mathcal{F}$, we denote this number by $\nu(\mathcal{F})$.

2. *For a general function class $\mathcal{F}$, we define its VC dimension to be the VC dimension of $\mathcal{F}_{sub}$, where*

$$\mathcal{F}_{sub} = \{\mathbf{1}_{f \leq 0},\ f \in \mathcal{F}\}.$$

Remark 1.2. *Indeed, the definition can be generalized to $x_i \in \mathcal{X}$ for general space $\mathcal{X}$.*

Now we state the Vapnik-Chervonenkis theorem, it states that $\mathcal{F}$ has finite VC-dimension would imply polynomial discrimination or order $\nu(\mathcal{F})$.

Theorem 1.6. *Assume that $\nu(\mathcal{F}) = k < \infty$, then*

$$\#\mathcal{F}(x_1^n) \leq \sum_{j=0}^{k} C_n^j.$$

Corollary 1.7.

$$\#(\mathcal{F}(x_1^n)) \leq (n+1)^{\nu(\mathcal{F})}$$

Proof. This follows from the main theorem, and the following identity:

$$\sum_{j=0}^{k} C_n^j \leq \sum_{j=0}^{k} n^j \leq \sum_{j=0}^{k} n^j C_k^j = (n+1)^k.$$

$\square$

With the above discussion, we are able to prove one version of uniform law of large numbers.

Corollary 1.8 (Classical Glivenko–Cantelli theorem)**.** *Let $\mathcal{F} := \{\mathbf{1}_{(-\infty,t]:t\in\mathbb{R}}\}$, then its VC-dimension is 1, hence By the proposition and theorem, we have*

$$\mathcal{R}(\mathcal{F}(x_1^n)/n) \leq \frac{C_1}{n}\sqrt{\log(n+1)},$$

and

$$\mathcal{R}_n(\mathcal{F}) \leq \frac{C_1}{n}\sqrt{\log(n+1)}.$$

Now by the concentration theorem of $\|F_n - F\|_{\mathcal{F}}$, we have

$$\mathbb{P}\left(\|F_n - F\|_{\mathcal{F}} \geq 2\frac{C_1}{n}\sqrt{\log(n+1)}\right) \leq \exp(-\frac{n\delta^2}{2b^2}).$$

Note that $\|F_n - F\|_{\mathcal{F}}$ can be written as

$$\begin{aligned}\|F_n - F\|_{\mathcal{F}} &= \sup_t |\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i \leq t) - P(X \leq t)| \\ &= \|\hat{F}_n - F\|_\infty,\end{aligned}$$

where $\hat{F}_n$ is the empirical density funciton and $\|\cdot\|_\infty$ is the supreme norm. This shows that

$$\|\hat{F}_n - F\|_\infty \to 0, a.s.$$

Now we present a criterion on the finiteness of VC-dimension.

Proposition 1.9. *Let $\mathcal{G}$ be a vector space of functions $g : \mathbb{R}^d \mapsto \mathbb{R}$, then the VC-dimension of $\mathcal{G}$ is at most* $\dim(\mathcal{G})$.

Proof. If not, assume $\dim(\mathcal{G}) = k$, $P = \{x_1, \ldots, x_n\}$ is a set of n points that can be shattered by $\mathcal{G}$ while $n > k$.

In the following, for simplicity we write $i \in S$ for $x_i \in S$.

Assume $g_1, \ldots, g_k$ form a basis of $\mathcal{G}$, denote this vector by v_g and denote by

$$v_i := (g_j(x_i))_{j=1}^k,$$

WLOG, we assume that $v_n = \sum_{i=1}^{n-1} c_i v_i$. Now, we construct a set S as follows:

1. For each $i \in [1, n-1]$, $i \in S$ if $c_i > 0$,

2. For $i = n$, $i \notin S$.

Assume that $g = \sum_{j=1}^{k} b_j g_j = \langle v_g, b\rangle$ satisfies:

$$g(x_i \leq 0) \Leftrightarrow i \in S,$$

then $c_i\langle v_i, b\rangle \leq 0$ for $i \in [0, n-1]$, which means

$$g(x_n) = \sum_{i=1}^{n-1} c_i\langle v_i, b\rangle \leq 0 \Rightarrow n \in S,$$

this gives contradiction. $\square$

Example 1.1. *From the above proposition, we can see the VC-dimension of family of d-dimensional spheres is* 1.