High Dimensional Probability and Statistics **2nd Semester, 2023-2024**

Lecture 7: Uniform LLN, VC Dimension and Applications

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2024/04/28)*

Motivation: We are interested in bounding the random variable

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^{n} f(X_k)-\mathbb{E}\left[f(X)\right]\right|, \tag{7.1}$$

where $\mathcal{F}$ is a class of functions. That is, we want to estimate the deviation between $\frac{1}{n}\sum_{k=1}^{n} f(X_k)$ and $\mathbb{E}\left[f(X)\right]$ uniformly over the class $\mathcal{F}$ – hence the name of uniform laws of large numbers (ULLN). Here, we ignore the measurability issue after taking the supremum. This problem arises in a wide range of applications. We first give several typical examples.

Wasserstein law of large numbers Let $X_1,\cdots,X_n$ be i.i.d samples from a population measure $\mathbb{P}$, where $\mathbb{P}$ is a probability measure on $[0,1]$. Define the following empirical measure

$$\mathbb{P}_n=\frac{1}{n}\sum_{k=1}^{n}\delta_{X_k}.$$

Then a natural question is how well $\mathbb{P}_n$ stands for $\mathbb{P}$. For a realization, the Wasserstein distance between $\mathbb{P}_n$ and $\mathbb{P}$ is given by

$$W_1(\mathbb{P}_n,\mathbb{P})=\sup_{f\in\mathcal{F}}\left|\mathbb{E}_{\mathbb{P}_n}[f(Z)]-\mathbb{E}_{\mathbb{P}}[f(Z)]\right|$$

where $\mathcal{F}=\{f\in\mathrm{Lip}\left([0,1],|\cdot|\right):\ 0\leq f\leq 1\}$ is a set of 1-Lipschitz functions. It is evident that $W_1(\mathbb{P}_n,P)$ is in the form of (7.1). In this case, the related result is also known as Wasserstein law of large numbers.

Classical Glivenko–Cantelli theorem Letting $X\sim\mathbb{P}$, the cumulative distribution function (CDF) $F(a)$ is given by $F(a)=\mathbb{P}\left[X\leq a\right]$. Given a set of i.i.d samples $\{X_k\}_{k=1}^{n}$, we can estimate F by the empirical CDF,

$$\widehat{F}_n(a)=\frac{1}{n}\sum_{k=1}^{n}1_{(-\infty,a]}(X_k),$$

i.e., the empirical frequency over $(-\infty,a]$. Then it is natural to ask whether

$$\left|\widehat{F}_n(a)-F(a)\right| \text{ is small uniformly for all } a\in\mathbb{R}?$$

The classical Glivenko–Cantelli theorem answers this question in an affirmative way. Letting $\mathcal{F}=\{1_{(-\infty,a]}(x):\ a\in\mathbb{R}\}$, since $\mathbb{E}\left[1_{(-\infty,a]}(X)\right]=F(a)$, we actually need to bound (7.1), where $\mathcal{F}$ is given by

$$\mathcal{F}=\{1_{(-\infty,a]},\ a\in\mathbb{R}\}. \tag{7.2}$$

Generalization analysis in statistical learning Given a pair of random variables (X, Y), a central task in statical learning is to find the relationship between X and Y. This is typically formed as the problem of finding a function (hypothesis) h in a function class $\mathcal{H}$ such that the population risk

$$R(h) = \mathbb{E}\left[\mathcal{L}(h(X), Y)\right]$$

is minimized. Here $\mathcal{L}(\cdot, \cdot)$ represents certain loss function. However, since we do not know the distribution of by only have access to a set of i.i.d samples $X_1, \cdots, X_n$, a computationally tractable alternative is to minimize the empirical risk,

$$\widehat{R}_n(h) = \frac{1}{n}\sum_{k=1}^{n} \mathcal{L}(h(X_k), Y_k).$$

Letting h^* be the minimizer of $R(h)$ and $\hat{h}_n^*$ be the minimizer of $\widehat{R}_n(h)$, in order for $\hat{h}_n^*$ to generalize well for the entire distribution, we wish $R(\hat{h}_n^*)$ should be close to $R(h^*)$. This can be achieved if $\widehat{R}_n(h)$ is close to $R(h)$ for all $h \in \mathcal{H}$ since then they will have their minimizers close to each other. More precisely, the **excess risk** defined by $R(\hat{h}_n^*) - R(h^*)$ satisfies

$$\begin{aligned} R(\hat{h}_n^*) - R(h^*) &= \left(R(\hat{h}_n^*) - \widehat{R}_n(\hat{h}_n^*)\right) + \left(\widehat{R}_n(\hat{h}_n^*) - \widehat{R}_n(h^*)\right) + \left(\widehat{R}_n(h^*) - R(h^*)\right) \\ &\le \left|R(\hat{h}_n^*) - \widehat{R}_n(\hat{h}_n^*)\right| + \left|\widehat{R}_n(h^*) - R(h^*)\right| \\ &\le 2\sup_{h\in\mathcal{H}} \left|\widehat{R}_n(h) - R(h)\right|. \end{aligned} \tag{7.3}$$

Thus, in order to bound the generalization error $R(\hat{h}_n^*) - R(h^*)$, it suffices to bound

$$\sup_{h\in\mathcal{H}} \left|\widehat{R}_n(h) - R(h)\right| = \sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum_{k=1}^{n} \mathcal{L}(h(X_k), Y_k) - \mathbb{E}\left[\mathcal{L}(h(X, Y))\right]\right| \tag{7.4}$$

If we define

$$f(Z_1, \cdots, Z_n) := \sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum_{k=1}^{n} \mathcal{L}(h(X_k), Y_k) - \mathbb{E}\left[\mathcal{L}(h(X), Y)\right]\right|, \quad \text{where } Z_k = (X_k, Y_k),$$

it is easy to see that (7.4) is a special case of (7.1).

Under some proper conditions (e.g., $\|f\|_\infty \le b$ for $f \in \mathcal{F}$), it is easy to show that the quantity in (7.1) concentrates around its mean, for example by bounded difference inequality. Thus, we will focus on its expectation

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{k=1}^{n} \left(f(X_k) - \mathbb{E}\left[f(X_k)\right]\right)\right|\right]. \tag{7.5}$$

Note that when there is single function in $\mathcal{F}$, we have

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{k=1}^{n} \left(f(X_k) - \mathbb{E}\left[f(X_k)\right]\right)\right|\right] \lesssim \frac{1}{\sqrt{n}}$$

under some mild moment assumptions. It is intriguing to see whether this is also true when $\mathcal{F}$ has many functions, which is also the desirable goal to pursue.

Agenda:

- Wasserstein Law of Large Numbers
- Symmetrization, VC Dimension
- Classical Glivenko-Cantelli Theorem
- Statistical Learning

7.1 Wasserstein Law of Large Numbers

In this section, we consider (7.5) for the case

$$\mathcal{F} = \{f \in \mathrm{Lip}\,([0,1], |\cdot|) : \; 0 \le f \le 1\}.$$

The following theorem establishes the covering number of $\mathcal{F}$ under the infinity norm.

Lemma 7.1 *There is a numerical constant $c > 0$ such that*

$$N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le e^{c/\varepsilon} \text{ for } \varepsilon < \frac{1}{2} \quad \text{and} \quad N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) = 1 \text{ for } \varepsilon \ge \frac{1}{2}.$$

Proof: The claim $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) = 1$ for $\varepsilon \ge \frac{1}{2}$ is trivial since $\left\|f - \frac{1}{2}\right\|_\infty \le \frac{1}{2}$ for each $f \in \mathcal{F}$. The proof of the first claim is basically based on approximating f with piecewise constant functions, see [2] for details. ■

7.1.1 First Effort via Finite Approximation

For ease of notation, let $Z_f = \frac{1}{n}\sum_{k=1}^n f(X_k) - \mathbb{E}[f(X)]$. First note that Z_f is $1/4n$-sub-Gaussian since $f \in [0,1]$ (**check this!**). Letting N be the ε-net of $(\mathcal{F}, \|\cdot\|_\infty)$, by Lemma 7.1, we have $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le e^{c/\varepsilon}$, for $\varepsilon < 1/2$. Thus,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |Z_f|\right] \le \inf_{0<\varepsilon<1/2} \left\{ \mathbb{E}\left[\sup_{f\in\mathcal{F}} |Z_f - Z_{\pi(f)}|\right] + \sqrt{\frac{c}{2n\varepsilon}} \right\}$$

Moreover, we have

$$\begin{aligned} |Z_f - Z_{\pi(f)}| &= \left| \left(\frac{1}{n}\sum_{k=1}^n (f(X_k) - \pi(f)(X_k)) \right) + \mathbb{E}[\pi(f)(X) - f(X)] \right| \\ &\le \frac{2}{n}\sum_{k=1}^n \|f - \pi(f)\|_\infty \\ &\le 2\varepsilon. \end{aligned}$$

It follows that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |Z_f|\right] \le \inf_{0<\varepsilon<1/2} \left\{ 2\varepsilon + \sqrt{\frac{c}{2n\varepsilon}} \right\} \asymp n^{-1/3},$$

which is sub-optimal.

7.1.2 Second Effort via Dudley Integral

With the same definition of Z_f as in the last subsection, we have that

$$Z_f - Z_g = \frac{1}{n}\sum_{k=1}^{n}\left(f(X_k) - g(X_k) - (\mathbb{E}\left[f(X_k)\right] - \mathbb{E}\left[g(X_k)\right]\right)$$

is $\frac{1}{n}\|f-g\|_\infty^2$-sub-Gaussian. Thus, if we define

$$d(f,g) = n^{-1/2}\|f-g\|_\infty,$$

then $Z_f - Z_g$ is $d(f,g)^2$-sub-Gaussian. In addition, it is easily seen that (**check this!**)

$$N(\mathcal{F}, n^{-1/2}\|\cdot\|_\infty, \varepsilon) = N(\mathcal{F}, \|\cdot\|_\infty, n^{1/2}\varepsilon).$$

Thus, the application of the Dudley integral (note that $0 \in \mathcal{F}$) yields

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^{n} f(X_k) - \mathbb{E}\left[f(X)\right]\right|\right] &\lesssim \int_0^\infty \sqrt{\log N(\mathcal{F}, \|\cdot\|_\infty, n^{1/2}\varepsilon)}d\varepsilon \\ &= \frac{1}{\sqrt{n}}\int_0^\infty \sqrt{\log N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)}d\varepsilon \\ &= \frac{1}{\sqrt{n}}\int_0^{1/2} \sqrt{\frac{c}{\varepsilon}}d\varepsilon \\ &\asymp \frac{1}{\sqrt{n}}.\end{aligned}$$

7.2 Symmetrization, VC Dimension

For the problem in the last section, the infinite norm can be used to bring out the sub-Gaussian nature of the process, and thus the tight $1/\sqrt{n}$ bound can be established via Dudley integral. However, for many cases, it is not efficient to bound the increments using the infinite norm and weaker metrics should be considered. Consider $\mathcal{F}$ given in (7.2). Since

$$\|1_{(-\infty,a]} - 1_{(-\infty,a']}\|_\infty = 1 \quad \text{whenever} \quad a \neq a',$$

we have $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) = \infty$ for $\varepsilon < 1$. Thus, substituting this into Dudley integral is not quite meaningful. The symmetrization argument provides a way to overcome this pitfall, which allows us to use the Dudley integral based on covering under potentially a smaller distance through separating the sign (or "Gaussian part") out from its magnitude . To motivate the symmetrization argument, consider the random variable $\sum_{k=1}^n X_k$ where X_k are independent mean zero random variables. When the magnitude of each X_k is of the order $O(1)$, a naive bound for $|\sum_{k=1}^n X_k|$ would be $O(n)$. However, by the central limit theorem, a more desirable bound would be $O(\sqrt{n})$. This is due to that the terms in the sum are independent and centered, so they are likely to have opposite signs, yielding the cancellation effect. Therefore, the random sign $\sum_{k=1}^n \text{sign}(X_k)$ plays an essential role in the Gaussian tail while the magnitudes of X_k only determine the variance.

7.2.1 Symmetrization

As already mentioned, the symmetrization technique separates the sign (or "Gaussian part") of the process out from its magnitude and analyze each part sequentially. This allows us to provide bounds for (7.1) more efficiently.

Lemma 7.2 (Upper bound by symmetrization) *Let $\{X_k\}_{k=1}^n$ be i.i.d random variables. Then,*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n (f(X_k)-\mathbb{E}\left[f(X_k)\right])\right|\right] \le 2\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n \varepsilon_k f(X_k)\right|\right],$$

where $\{\varepsilon_k\}_{k=1}^n$ is a collection of i.i.d Rademacher random variables.

Proof: Let $\{Y_k\}_{k=1}^n$ be i.i.d copies of $\{X_k\}_{k=1}^n$. We have

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n (f(X_k)-\mathbb{E}\left[f(X)\right])\right|\right] &= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n (f(X_k)-\mathbb{E}_Y\left[f(Y_k)\right])\right|\right]\\ &= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}_Y\left[\sum_{k=1}^n (f(X_k)-f(Y_k))\right]\right|\right]\\ &\le \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\mathbb{E}_Y\left[\left|\sum_{k=1}^n (f(X_k)-f(Y_k))\right|\right]\right]\\ &\le \mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n (f(X_k)-f(Y_k))\right|\right].\end{aligned}$$

where the third line follows from Jensen inequality. Noting that $f(X_k)-f(Y_k)$ is symmetric and thus has the same distribution with $\varepsilon_k(f(X_k)-f(Y_k))$, it follows that

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n (f(X_k)-\mathbb{E}\left[f(X)\right])\right|\right] &\le \mathbb{E}_{X,Y,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n (\varepsilon_k(f(X_k)-f(Y_k)))\right|\right]\\ &\le \mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n \varepsilon_k f(X_k)\right|\right] + \mathbb{E}_{Y,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n \varepsilon_k f(Y_k)\right|\right],\end{aligned}$$

which completes the proof since $\{Y_k\}_{k=1}^n$ are i.i.d copies of $\{X_k\}_{k=1}^n$. ■

Lemma 7.3 (Lower bound by symmetrization) *Let $\{X_k\}_{k=1}^n$ be i.i.d random variables. Then,*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n (f(X_k)-\mathbb{E}\left[f(X_k)\right])\right|\right] \ge \frac{1}{2}\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n \varepsilon_k\left(f(X_k)-\mathbb{E}\left[f(X_k)\right]\right)\right|\right],$$

where $\{\varepsilon_k\}_{k=1}^n$ is a collection of i.i.d Rademacher random variables.

Proof: We have

$$\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n \varepsilon_k\left(f(X_k)-\mathbb{E}_X\left[f(X_k)\right]\right)\right|\right]$$

$$
\begin{aligned}
&= \mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\varepsilon_k\left(f(X_k)-\mathbb{E}_Y\left[f(Y_k)\right]\right)\right|\right] \\
&\le \mathbb{E}_{X,Y,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\varepsilon_k\left(f(X_k)-f(Y_k)\right)\right|\right] \\
&= \mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\left(f(X_k)-f(Y_k)\right)\right|\right] \\
&\le \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\left(f(X_k)-\mathbb{E}\left[f(X_k)\right]\right)\right|\right]+\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\left(f(Y_k)-\mathbb{E}\left[f(Y_k)\right]\right)\right|\right],
\end{aligned}
$$

which completes the proof since $\{Y_k\}_{k=1}^n$ are i.i.d copies of $\{X_k\}_{k=1}^n$. ■

Remark 7.4 *Note the right hand side in Lemma 7.3 cannot be replaced by* $\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\varepsilon_k f(X_k)\right|\right]$ *since a counter example can be easily constructed for the* $n=1$ *case.*

To upper bound (7.5), by Lemma 7.2, it suffices to bound

$$
\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\varepsilon_k f(X_k)\right|\right]. \tag{7.6}
$$

For this, we can first condition on $X=(x_1,\cdots,x_n)$ and bound

$$
\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\varepsilon_k f(x_k)\right|\right] \tag{7.7}
$$

and then take expectation with respect to X. It follows that

$$
\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\left(f(X_k)-\mathbb{E}\left[f(X_k)\right]\right)\right|\right] \lesssim \sqrt{\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^{n}f^2(X_k)\right]}\sqrt{\log\Pi_{\mathcal{F}}(n)}, \tag{7.8}
$$

where

$$
\Pi_{\mathcal{F}}(n):=\max_{\{x_1,\cdots,x_n\}\subset\mathcal{X}}\left|\left\{\left(f(x_1),\cdots,f(x_n)\right):\ f\in\mathcal{F}\right\}\right|. \tag{7.9}
$$

Note that when $|\mathcal{F}|=\infty$ in which case a direct bound uniform bound for (7.5) fails. In contrast, it is possible that $\Pi_{\mathcal{F}}(n)$ is finite (e.g., when $\mathcal{F}$ a class of binary value functions for classification problems). Assuming $\|f\|_\infty\le b$ for all $f\in\mathcal{F}$, we can still work out an upper bound for (7.5) through (7.7) and obtain

$$
\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\left(f(X_k)-\mathbb{E}\left[f(X_k)\right]\right)\right|\right] \lesssim \sqrt{\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^{n}f^2(X_k)\right]}\sqrt{\log\Pi_{\mathcal{F}}(n)} \le \sqrt{n}b\sqrt{\log\Pi_{\mathcal{F}}(n)}. \tag{7.10}
$$

Next we will focus on the case when $\mathcal{F}$ a class of binary value functions (and hence $\Pi_{\mathcal{F}}(n)$ is finite, at most 2^n). It can be shown that the growth of $\Pi_{\mathcal{F}}(n)$ is determined by a notion called VC dimension. In other words, VC dimension provides a different way to quantify the complexity of the function class $\mathcal{F}$. Though we only discuss the VC dimension for the families of binary value functions, it can be extended to general classes of functions, see for example Chapter 7.3 of [2].

7.2.2 VC Dimension

Definition 7.5 (Shattering and VC dimension) *Let $\mathcal{F}$ be a class of binary value functions. We say a set $(x_1, \cdots, x_n) \subset \mathcal{X}$ is shattered by $\mathcal{F}$ if*

$$|\{(f(x_1), \cdots, f(x_n)) : \ f \in \mathcal{F}\}| = 2^n.$$

*The VC dimension of $\mathcal{F}$, denoted $v(\mathcal{F})$ or simply v for short, is defined as the largest integer n for which **there exists** a collection of points $(x_1, \cdots, x_n)$ that is shattered by $\mathcal{F}$.*

Remark 7.6 *By the definition, when $n > v$, then for any collection of points $(x_1, \cdots, x_n)$,*

$$|\{(f(x_1), \cdots, f(x_n)) : \ f \in \mathcal{F}\}|$$

must be exactly smaller than 2^n. In terms of the growth function in (7.9), the VC dimension is the largest integer n such that $\Pi_{\mathcal{F}}(n) = 2^n$.

Exercise 7.7 *If there exists n points that can be shattered, why for any $m < n$ there exists m points hat can also be shattered? If there does not exist n points that can be shattered, why for any $m > n$ there does not exist m points that can be shattered?*

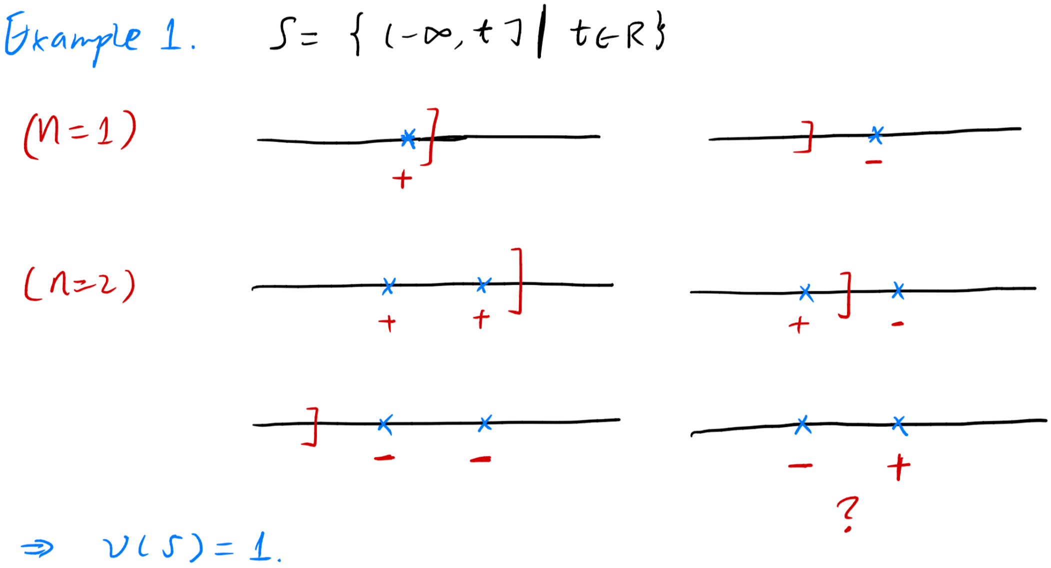


Figure 7.1: Example I

Example 7.8 *Figures 7.1, 7.2 and 7.3 give three examples with finite VC dimension, where*

$$\mathcal{F} = \{1_S(x), \ S \in \mathcal{S}\}.$$

There also exists set $\mathcal{S}$ such that the VC dimension of $\mathcal{F}$ is infinite, see [3].

For the function class having a finite VC dimension, it turns out its growth function is of the polynomial order in n.

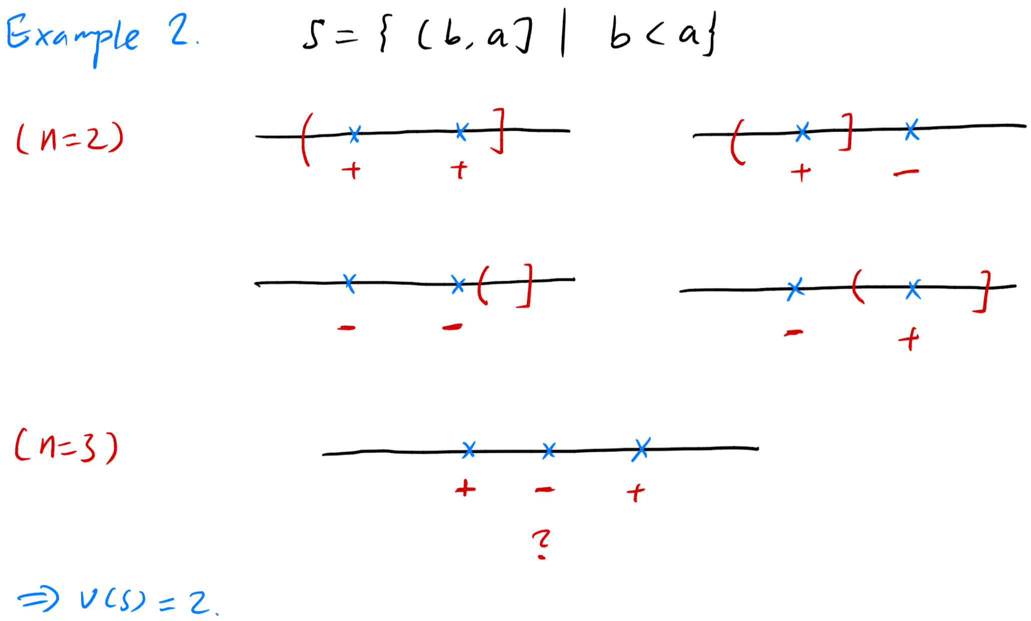


Figure 7.2: Example II

Lemma 7.9 (Sauer-Shelah) *For all $n \geq v$ and $(x_1, \cdots, x_n) \subset \mathcal{X}$, there holds*

$$\Pi_{\mathcal{F}}(n) := \max_{\{x_1,\cdots,x_n\}\subset\mathcal{X}} |\{(f(x_1), \cdots, f(x_k)) : \ f \in \mathcal{F}\}| \leq \sum_{k=0}^{v} \binom{n}{k} \leq \left(\frac{en}{v}\right)^v.$$

Proof: The second inequality follows directly from the combinatorial argument

$$\begin{aligned}
\sum_{k=0}^{v} \binom{n}{k} &\leq \sum_{k=0}^{v} \binom{n}{k} \left(\frac{n}{v}\right)^{v-k} \\
&\leq \sum_{k=0}^{n} \binom{n}{k} \left(\frac{n}{v}\right)^{v-k} \\
&= \left(\frac{n}{v}\right)^{v} \sum_{k=0}^{n} \binom{n}{k} \left(\frac{v}{n}\right)^{k} \\
&= \left(\frac{n}{v}\right)^{v} (1 + v/n)^n \\
&\leq \left(\frac{en}{v}\right)^{v}
\end{aligned}$$

The first inequality follows from an inductive argument and the details will be omitted. Interested readers may find them in [1] and [3]. ∎

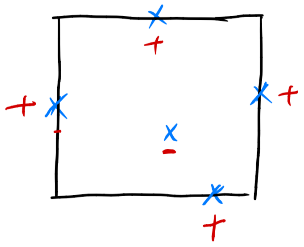
Note that Lemma 7.9 is a truly deep result. For $n > v(\mathcal{F})$, though the definition of VC dimension implies that $|\{(f(x_1), \cdots, f(x_n)) : \ f \in \mathcal{F}\}| < 2^n$ for any $(x_1, \cdots, x_n)$, this does not exclude the possibility that there exists a $(x_1, \cdots, x_n)$ such that $|\{(f(x_1), \cdots, f(x_n)) : \ f \in \mathcal{F}\}| = 2^n - 1$. However, the Sauer-Shelah lemma says that this cannot be true.

Example 3. $S = \{ [a,b] \times [c,d] \mid a \le b, c \le d \}$

(n=4) there exist four points on $\mathbb{R}^2$ that can be shattered.

[Try this]

(n=5) For any given five points in $\mathbb{R}^2$, first find the smallest rectangle that contains all the points. Set the point inside the rectangle to be '−', and the others to be '+',



This configuration cannot be realized by S.

$\Rightarrow$ $v(S) = 4$.

Figure 7.3: Example III

Exercise 7.10 *For the three examples in Example 7.8, show that* $|\{(f(x_1), \cdots, f(x_n)) : f \in \mathcal{F}\}| \lesssim n^v$ *directly rather that using the Sauer-Shelah lemma.*

7.3 Classical Glivenko-Cantelli Theorem

In this section we return back to the problem of estimating $\mathbb{E}\left[\|\widehat{F}_n - F\|_\infty\right]$, where F and $\widehat{F}_n$ are CDF and empirical CDF, respectively. It corresponds to estimating (7.5) for $\mathcal{F} = \{1_{(-\infty,a]}(x) : a \in \mathbb{R}\}$. By Example 7.8, we first know that $v(\mathcal{F}) = 1$. It follows from the Sauer-Shelah lemma that $\Pi_n(\mathcal{F}) \lesssim n$. Together with (7.10), we have

$$\mathbb{E}\left[\|\widehat{F}_n - F\|_\infty\right] \lesssim \sqrt{\frac{\log n}{n}}, \tag{7.11}$$

Remark 7.11 *By certain central limit theorem (Kolmogorov theorem), one can directly show that the optimal rate for* $\|\widehat{F}_n - F\|_\infty$ *is* $1/\sqrt{n}$. *Next, we will remove the log-factor in* (7.11) *by more advanced technique.*

Let $\mathcal{F} = \{1_C,\ C \subset \mathcal{X}\}$ be the set of binary value functions defined on a probability space $(\mathcal{X}, \mathbb{P})$. For any $f, g \in \mathcal{F}$, we define

$$\|f - g\|_{L^2(\mathbb{P})} = \left(\int_{\mathcal{X}} (f(x) - g(x))^2 d\mathbb{P}(x)\right)^{1/2}.$$

Lemma 7.12 *There is a numerical constant $c > 0$ such that*

$$N(\mathcal{F}, \|\cdot\|_{L^2(\mathbb{P})}, \varepsilon) \le \left(\frac{c}{\varepsilon}\right)^{cv} \quad \textit{for } \varepsilon < 1.$$

where v is the VC dimension of $\mathcal{F}$.

The proof of Lemma 7.12 relies on the following lemma.

Lemma 7.13 *Let $f_1, \cdots, f_n$ be functions on $(\mathcal{X}, \mathbb{P})$. If*

$$\|f_i\|_\infty \le 1, \quad \|f_i - f_j\|_{L^2(\mathbb{P})} > \varepsilon \quad \textit{for all } i \ne j,$$

then there exists $m \asymp \varepsilon^{-4} \log n$ points $x_1, \cdots, x_m$ such that

$$\frac{1}{m}\sum_{k=1}^m |f_i(x_k) - f_j(x_k)|^2 > \varepsilon^2/4 \quad \textit{for all } i \ne j. \tag{7.12}$$

Proof: The proof of this lemma uses a very interesting probabilistic argument: we first choose m points randomly and then show (7.12) holds with high probability. Then there must exist such m deterministic points. More precisely, let $X_1, \cdots, X_m \sim \mathbb{P}$ be i.i.d samples. The application of Hoeffding inequality implies that

$$\mathbb{P}\left[\frac{1}{m}\sum_{k=1}^m \left(|f_i(X_k) - f_j(X_k)|^2 - \mathbb{E}\left[|f_i(X_k) - f_j(X_k)|^2\right]\right) \le -t\right] \le \exp\left(-\frac{mt^2}{2}\right).$$

Noting that

$$\mathbb{E}\left[\frac{1}{m}\sum_{k=1}^m |f_i(X_k) - f_j(X_k)|^2\right] = \mathbb{E}\left[|f_i(X_k) - f_j(X_k)|^2\right] = \|f_i - f_j\|^2_{L^2(\mathbb{P})} > \varepsilon^2,$$

we have

$$\mathbb{P}\left[\frac{1}{m}\sum_{k=1}^m |f_i(X_k) - f_j(X_k)|^2 \le \frac{\varepsilon^2}{4}\right] \le \exp\left(-\frac{m\varepsilon^4}{4}\right).$$

Now a union bound gives

$$\mathbb{P}\left[\frac{1}{m}\sum_{k=1}^m |f_i(X_k) - f_j(X_k)|^2 \ge \frac{\varepsilon^2}{4} \text{ for all } i \ne j\right] \ge 1 - n^2 \text{exp}\left(-\frac{m\varepsilon^4}{4}\right) > 0$$

provided $m \asymp \varepsilon^{-4} \log n$. ■

Proof: [of Lemma 7.12] Let $f_1, \cdots, f_n$ be an maximal ε-packing of $(\mathcal{F}, \|\cdot\|_{L^2(\mathbb{P})})$. By Lemma 7.13, there exist $m \asymp \varepsilon^{-4} \log n$ points $x_1, \cdots, x_m$ such that

$$\frac{1}{m} \sum_{k=1}^{m} |f_i(x_k) - f_j(x_k)|^2 > \varepsilon^2/4 \quad \text{for all } i \neq j.$$

Thus, letting $\mathcal{F}_n = \{f_1, \cdots, f_n\}$,

$$n = |\{(f_i(x_1), \cdots, f_i(x_m)) : \ f_i \in \mathcal{F}_n\}| .$$

Note that the VC dimension of $\mathcal{F}_n$ is less or equal than the VC dimension of $\mathcal{F}$. By the Sauer-Shelah lemma we have

$$n \leq \left(\frac{em}{v}\right)^v \leq \left(\frac{c\varepsilon^{-4} \log n}{v}\right)^v,$$

and the claim follows after some simple calculus. ■

Theorem 7.14 (Glivenko-Cantelli) *We have* $\mathbb{E}\left[\|\widehat{F}_n - F\|_\infty\right] \lesssim \frac{1}{\sqrt{n}}$.

Proof: For fixed $(x_1, \cdots, x_n)$, let

$$Z_f = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \varepsilon_k f(x_k).$$

Noting that $Z_f - Z_g = \frac{1}{\sqrt{n}} \sum_{k=1}^n \varepsilon_k (f(x_k) - g(x_k))$ is $\frac{1}{n} \sum_{k=1}^n (f(x_k) - g(x_k))^2$-sub-Gaussian (see Lecture 1). Thus, if we define the metric

$$d(f,g) = \sqrt{\frac{1}{n} \sum_{k=1}^{n} (f(x_k) - g(x_k))^2},$$

then $Z_f - Z_g$ is $d(f,g)^2$-sub-Gaussian. Let $\widetilde{\mathcal{F}} = \{\mathcal{F}, 0\}$, namely we add a 0 function to $\mathcal{F}$. Note that we still have $v(\widetilde{\mathcal{F}}) = 1$ (**check this!**). Thus, Lemma 7.12 implies that

$$N(\widetilde{\mathcal{F}}, d, \varepsilon) \leq \left(\frac{c}{\varepsilon}\right)^c, \quad \text{for } \varepsilon < 1,$$

where $c > 0$ is a universal constant. Moreover, it is easy to see that $d(f,g) \leq 1$ for any $f, g \in \widetilde{F}$, and thus $\operatorname{diam}(\mathcal{F}) \leq 1$. By the Dudley integral (also noting Remark 6.4) we have

$$\begin{aligned} \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \varepsilon_k f(x_k) \right| \right] &= \mathbb{E}_\varepsilon \left[\sup_{f \in \widetilde{F}} \left| \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \varepsilon_k f(x_k) - 0 \right| \right] \\ &\lesssim \int_0^1 \sqrt{\log N(\widetilde{\mathcal{F}}, d, \varepsilon)} d\varepsilon \\ &= O(1), \end{aligned}$$

where $O(1)$ means a constant. Thus,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{k=1}^{n} f(X_k) - \mathbb{E}[f(X)] \right| \right] \lesssim \mathbb{E}_{X,\varepsilon} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{k=1}^{n} \varepsilon_k f(X_k) \right| \right] \lesssim \frac{1}{\sqrt{n}}.$$

The proof is now complete. ■

7.4 Statistical Learning

In this section, we study the generalization error analysis problem in statistical learning, as introduced at the beginning of this lecture. For simplicity, consider the classification problem where $Y = T(X)$ and T is a fixed Boolean function on $\mathcal{X}$. Moreover, we consider the case where $\mathcal{L}$ is squared loss,

$$\mathcal{L}(h(X_k), Y_k) = |h(X_k) - Y_k|^2 = |h(X_k) - T(X_k)|^2.$$

Therefore, by (7.3), we have

$$\begin{aligned} R(\hat{h}_n^*) - R(h^*) &\leq 2 \sup_{h \in \mathcal{H}} \left| \widehat{R}_n(h) - R(h) \right| \\ &= 2 \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{k=1}^{n} \left(|h(X_k) - T(X_k)|^2 - \mathbb{E}\left[|h(X_k) - T(X_k)|^2 \right] \right) \right| \\ &= 2 \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{k=1}^{n} f(X_k) - \mathbb{E}\left[f(X) \right] \right|, \end{aligned}$$

where $f = |h - T|^2$, $\mathcal{F} = \{|h - T|^2,\ h \in \mathcal{H}\}$. Assume $\mathcal{H}$ is a set of Boolean functions: $\{1_C,\ C \subset \mathcal{X}\}$. We have the following result.

Lemma 7.15 *For any set of points $(x_1, \cdots, x_n)$, define $d(f, g) = \sqrt{\frac{1}{n} \sum_{k=1}^{n} (f(x_k) - g(x_k))^2}$. Then*

$$N(\mathcal{F}, d, \varepsilon) \leq N(\mathcal{H}, d, \varepsilon).$$

Proof: Since both h and T are Boolean functions, so does $h - T$. Thus,

$$|h - T|^2 = |h - T|.$$

It follows that

$$||h_1 - T|^2 - |h_2 - T|^2| = ||h_1 - T| - |h_2 - T|| \leq |h_1 - h_2| = |h_1 - h_2|^2.$$

for any $h_1, h_2 \in \mathcal{H}$. Therefore,

$$\begin{aligned} d(h_1 - T, h_2 - T) &= \sqrt{\frac{1}{n} \sum_{k=1}^{n} |h_1(x_k) - T(x_k)|^2 - |h_2(x_k) - T(x_k)|^2} \\ &\leq \sqrt{\frac{1}{n} \sum_{k=1}^{n} |h_1(x_k) - h_2(x_k)|^2} \\ &= d(h_1, h_2), \end{aligned}$$

from which $N(\mathcal{F}, d, \varepsilon) \leq N(\mathcal{H}, d, \varepsilon)$ can be easily established. ■

Theorem 7.16 *Under the previous assumptions, we have*

$$\mathbb{E}\left[R(\hat{h}_n^*) - R(h^*) \right] \lesssim \sqrt{\frac{v(\mathcal{H})}{n}},$$

where $v(\mathcal{H})$ denotes the VC dimension of $\mathcal{H}$.

Proof: We have

$$\begin{aligned}
\mathbb{E}\left[R(\hat{h}_n^*) - R(h^*)\right] &\lesssim \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^{n} f(X_k) - \mathbb{E}\left[f(X)\right]\right|\right] \\
&\lesssim \mathbb{E}_X\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^{n}\varepsilon_k f(X_k)\right|\right] \\
&\lesssim \frac{1}{\sqrt{n}}\cdot\mathbb{E}_X\left[\int_0^1 \sqrt{\log N(\mathcal{F}, d, \varepsilon)}d\varepsilon\right] \\
&\leq \frac{1}{\sqrt{n}}\cdot\mathbb{E}_X\left[\int_0^1 \sqrt{\log N(\mathcal{H}, d, \varepsilon)}d\varepsilon\right] \\
&\lesssim \sqrt{\frac{v(\mathcal{H})}{n}},
\end{aligned}$$

where the last line follows from Lemma 7.12. ■

Reading Materials

[1] Martin Wainwright, *High-dimensional statistics – A non-asymptotic viewpoint*, Chapter 4.

[2] Ramon van Handel, *Probability in High Dimension*, Chapters 5.2, 7.1, 7.2.

[3] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, *Foundations of Machine Learning*, Chapter 3.