High Dimensional Probability and Statistics **2nd Semester, 2021-2022**

Lecture 6: Uniform Law of Large Numbers

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2022/04/23)*

Motivation: We are interested in bounding the random variable

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^{n}f(X_k)-\mathbb{E}\left[f(X)\right]\right|, \tag{6.1}$$

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. Here, we ignore the measurability issue after taking the supremum. It is worth noting that we can view (6.1) as the (random) distance between the empirical probability measure $\mathbb{P}_n := \frac{1}{n}\sum_{k=1}^{n}\delta_{X_k}$ and the probability measure $\mathbb{P}$ where the distance is defined in the linear functional/operator sense (cf. Wasserstein distance where $\mathcal{F}$ is a particular class of functitons).

Recall that the quantity in (6.1) plays an important role in the generalization analysis of empirical risk minimization methods. Apart from this, it also has a close connection with the classical Glivenko-Cantelli theorem. Letting $X\sim\mathbb{P}$, the cumulative distribution function (CDF) $F(a)$ is given by $F(a)=\mathbb{P}\left[X\le a\right]$. 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

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

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 (6.1), where $\mathcal{F}$ is given by

$$\mathcal{F}=\{1_{(-\infty,a]},\ a\in\mathbb{R}\}. \tag{6.2}$$

Under some proper conditions (e.g., $\|f\|_\infty\le b$ for $f\in\mathcal{F}$), it is easy to show that the quantity in (6.1) concentrates around its mean, for example by bounded difference property. Thus, we only pay attention to its expectation

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\left(f(X_k)-\mathbb{E}\left[f(X_k)\right]\right)\right|\right]. \tag{6.3}$$

In particularly, we focus on the case when $\mathcal{F}$ is a set of binary value functions in (6.1), with the Glivenko-Cantelli theorem as a special example. Define

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

By the discussion in Lecture 5.2.2, we have

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^{n}f(X_k)-\mathbb{E}\left[f(X)\right]\right|\right]\lesssim\int_0^\infty\sqrt{\log N(\mathcal{F},\|\cdot\|_\infty,n^{1/2}\varepsilon)}d\varepsilon$$
$$=\frac{1}{\sqrt{n}}\int_0^\infty\sqrt{\log N(\mathcal{F},\|\cdot\|_\infty,\varepsilon)}d\varepsilon. \tag{6.4}$$

However, for $\mathcal{F}$ given in (6.2), 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, the bound in (6.4) 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 $\sum_{k=1}^n \operatorname{sign}(X_k)$ plays an essential role in the Gaussian tail while the magnitudes of X_k only determine the variance.

Agenda:

- Symmetrization
- VC Dimension and Sauer-Shelah Lemma
- Classical Glivenko-Cantelli Theorem

6.1 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 (6.1) more efficiently.

Lemma 6.1 (Upper bound by symmetrization) *Let $\{X_k\}_{k=1}^n$ be i.i.d random variables. Then,*

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

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

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

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

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

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

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

Lemma 6.2 (Lower bound by symmetrization) *Let $\{X_k\}_{k=1}^n$ be i.i.d random variables. Then,*

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

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

Proof: We have

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

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

Remark 6.3 *Note the right hand side in Lemma 6.2 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 (6.3), by Lemma 6.1, it suffices to bound

$$\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\varepsilon_k f(X_k)\right|\right]. \tag{6.5}$$

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

$$\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\varepsilon_k f(x_k)\right|\right] \tag{6.6}$$

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

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\left(f(X_k)-\mathbb{E}\left[f(X_k)\right]\right)\right|\right]\lesssim\sqrt{\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^{n}f^2(X_k)\right]}\sqrt{\log\Pi_{\mathcal{F}}(n)}, \tag{6.7}$$

where

$$\Pi_{\mathcal{F}}(n):=\max_{\{x_1,\cdots,x_n\}\subset\mathcal{X}}\left|\left\{(f(x_1),\cdots,f(x_n)):\ f\in\mathcal{F}\right\}.\right| \tag{6.8}$$

Note that when $|\mathcal{F}|=\infty$ in which case a direct bound uniform bound for (6.3) fails. In contrast, it is possible that $\Pi_{\mathcal{F}}(n)$ is finite (e.g., when $\mathcal{F}$ a class of binary value functions or for classification problems). In this case we can still work out an upper bound for (6.3) through (6.6) and obtain

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{n}\left(f(X_k)-\mathbb{E}\left[f(X_k)\right]\right)\right|\right]\lesssim\sqrt{\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^{n}f^2(X_k)\right]}\sqrt{\log\Pi_{\mathcal{F}}(n)}\leq\sqrt{n}b\sqrt{\log\Pi_{\mathcal{F}}(n)}. \tag{6.9}$$

Next we will focus on the case when $\mathcal{F}$ a class of binary value functions (and hence $\Pi_{\mathcal{F}}(n)$ is finite) 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].

6.2 VC Dimension and Sauer-Shelah Lemma

Definition 6.4 (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*

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

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

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

must be exactly smaller than 2^n. In terms of the growth function in (6.8)*, the VC dimension is the largest integer n such that $\Pi_{\mathcal{F}}(n)=2^n$.*

Exercise 6.6 *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?*

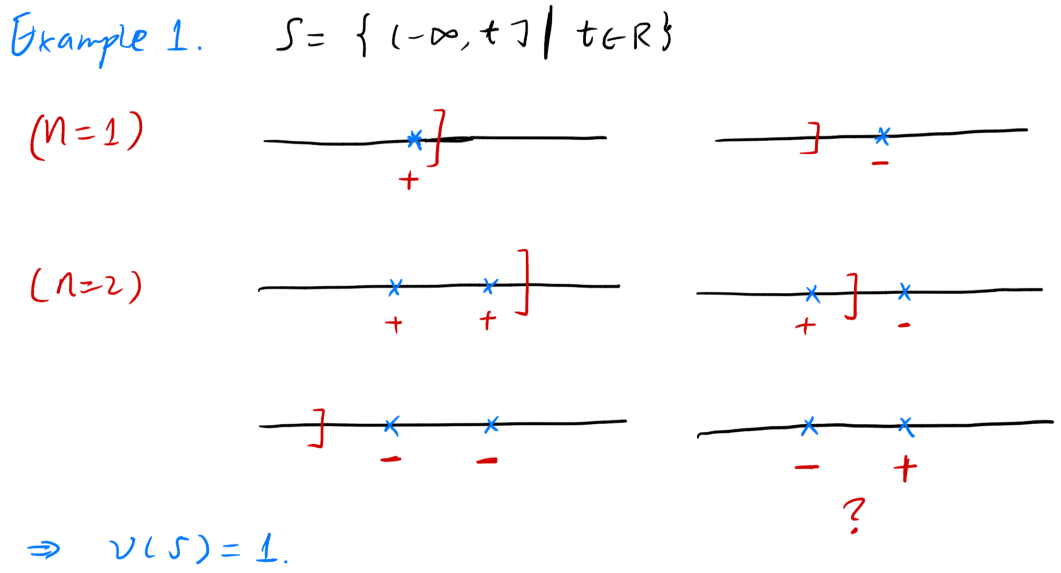


Figure 6.1: Example I

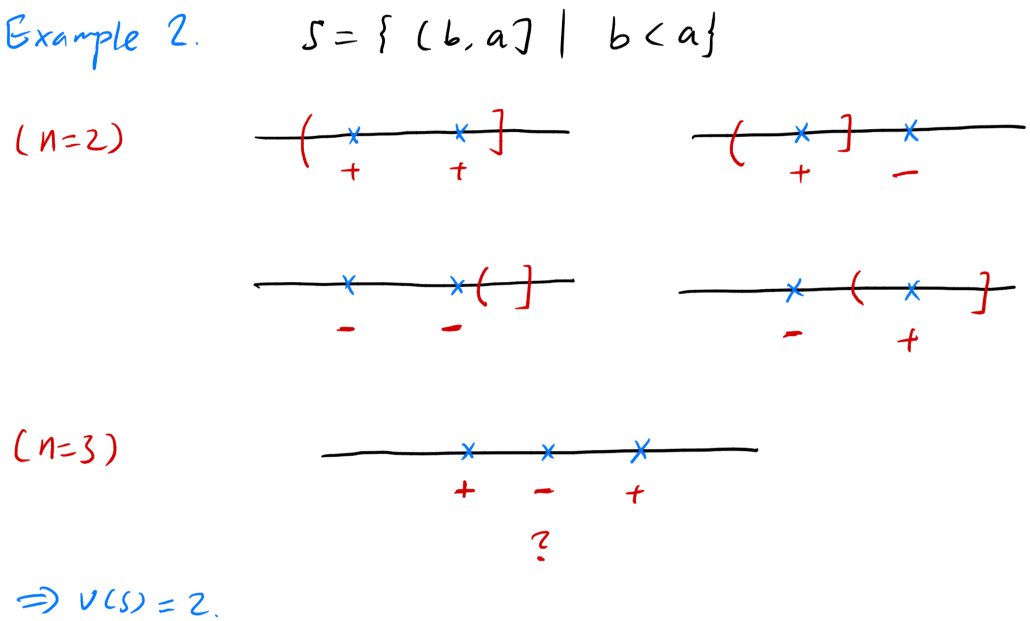


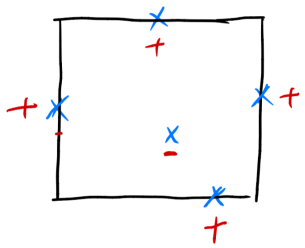
Figure 6.2: Example II

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.

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

Figure 6.3: Example III

Example 6.7 *Figures 6.1, 6.2 and 6.3 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.

Lemma 6.8 (Sauer-Shelah) *For all $n \ge v$ and $(x_1, \cdots, x_n) \subset \mathcal{X}$, there holds*

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

Proof: The second inequality follows directly from the combinatorial argument

$$\sum_{k=0}^{v} \binom{n}{k} \le \sum_{k=0}^{v} \binom{n}{k} \left(\frac{n}{v}\right)^{v-k}$$

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

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 6.8 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.

Exercise 6.9 *For the three examples in Example 6.7, show that* $|\{(f(x_1), \cdots, f(x_n)) : \ f \in \mathcal{F}\}| \lesssim n^v$ *directly rather that using the Sauer-Shelah lemma.*

6.3 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 (6.3) for $\mathcal{F} = \{1_{(-\infty,a]}(x) : \ a \in \mathbb{R}\}$. By Example 6.7, we first know that $v(\mathcal{F}) = 1$. It follows from the Sauer-Shelah lemma that $\Pi_n(\mathcal{F}) \lesssim n$. Together with (6.9), we have

$$
\mathbb{E}\left[\|\widehat{F}_n - F\|_\infty\right] \lesssim \sqrt{\frac{\log n}{n}}, \tag{6.10}
$$

Remark 6.10 *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* (6.10) *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 6.11 *There is a numerical constant* $c > 0$ *such that*

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

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

The proof of Lemma 6.11 relies on the following lemma.

Lemma 6.12 *Let $f_1, \cdots, f_n$ be functions on $(\mathcal{X}, \mathbb{P})$. If*

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

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

$$\frac{1}{m} \sum_{k=1}^{m} |f_i(x_k) - f_j(x_k)|^2 > \varepsilon^2/4 \quad \textit{for all } i \neq j. \tag{6.11}$$

Proof: The proof of this lemma uses a very interesting probabilistic argument: we first choose m points randomly and then show (6.11) 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

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

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

Proof: [of Lemma 6.11] Let $f_1, \cdots, f_n$ be an maximal ε-packing of $(\mathcal{F}, \|\cdot\|_{L^2(\mathbb{P})})$. By Lemma 6.12, 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 $\mathcal{F}_n = \{f_1, \cdots, f_n\}$,

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

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

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

and the claim follows after some simple calculus. ∎

Theorem 6.13 (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 6.11 implies that

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

where $c > 0$ is a universal constant. Moreover, it is easy to see that $d(f,g) \le 1$ for any $f, g \in \widetilde{F}$, and thus $\mathrm{diam}(\mathcal{F}) \le 1$. 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,

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

The proof is now complete. ■

Reading Materials

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

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

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