Math 547: Mathematical Foundations of Statistical Learning Theory

Lectures — Week 7

Instructor: Stas Minsker *Scribe: S. Minsker*

1.1 Metric Entropy and Dudley's Theorem

Recall that $\{X(t), t \in T\}$ is Gaussian $\Leftrightarrow \forall k \geq 1, t_1, \ldots, t_k \in T$, $(X(t_1), \ldots X(t_n))$ has multivariate normal distribution. Assume that $\mathbb{E}X(t) = 0 \ \forall t \in T$. Then $\{X(t), t \in T\}$ has sub-Gaussian increments with respect to

$$d(s,t) = \sqrt{\mathrm{Var}(X(s) - X(t))}.$$

It is a natural metric associated to the Gaussian process.

Definition 1. Let (T, d) be a relative compact metric space. The ε-covering number of (T, d), denoted $N(T, d, \epsilon)$, is defined as

$$N(T, d, \epsilon) = \min(n \geq 1 \text{ s.t. } \exists t_1, \ldots, t_n \ : T \subseteq \bigcup_{j=1}^{n} B(t_j, \varepsilon)),$$

where $B(t, \varepsilon) = (s \in T : d(s,t) \leq \varepsilon)$.

Note that $t_1, \ldots, t_n$ do not necessarily have to belong to T (for instance, this becomes important if we want to estimate the covering number of a subset of $S \subset T$).

Definition 2. The metric entropy of (T, d) is defined as

$$H(\varepsilon) = \log(N(\varepsilon)).$$

Example 1. Let $T = [0,1]^d, d \geq 1$

(a) If $d(t,s) = \|t - s\|_\infty$, then $N(\varepsilon) = \left(\frac{1}{\varepsilon}\right)^d, H(\varepsilon) = d \cdot \log(1/2)$.

(b) If $d(t,s) = \|t - s\|_2 = \sqrt{\sum_{j=1}^{d}(t_j - s_j)^2}$, then

$N(\varepsilon) \leq \left(\frac{\sqrt{d}}{\varepsilon}\right)^d$ and $N(\varepsilon) \geq \frac{1}{\mathrm{Vol}(B_2(s))} = C\left(\frac{\sqrt{d}}{\epsilon}\right)^d$.

Example 2 (Kolmogorov, Tikhomirov)**.** Let T be a set of smooth functions

$$f : [0,1]^d \longrightarrow, f \in \mathbb{C}^k$$

with distance $d(f,g) = \sup_{x \in [0,1]^d} |f(x) - g(x)|$. Then

$$H(T, d, \varepsilon) \sim \varepsilon^{-d/k}.$$

Theorem 1 (R. M. Dudley)**.** Let $\{X(t), t \in T\}$ have sub-Gaussian increments with respect to $d(\cdot, \cdot)$. Then $\forall t_0 \in T$

$$\mathbb{E} \sup_{t \in T} |X(t) - X(t_0)| \leq 12\sqrt{2} \int_0^\infty \sqrt{H(\varepsilon)} d\varepsilon$$

Let

$$D(T) = \text{diameter}(T) = \sup_{t,s \in T} d(t, s).$$

It is easy to see that $H(\varepsilon) = 0$ whenever $\varepsilon \geq D(T, d)$, so the upper limit of integration can be replaced by $D(T)$.

Proof. One proof readily follows from the (stronger) generic chaining bound that we proved in class earlier (this is a problem in your homework assignment). Here, we will give a direct "historical" argument, again using the classical chaining ideas.
Let $T_0 = \{t_0\}$ and T_j be the smallest $D(T)2^{-j}$ - net for T, meaning that

$$T_n = \{t_1, ..., t_{N(T,d,D2^j)}\}$$

such that $T \subseteq \bigcup_{i=1}^{N(T,d,D2^j)} B(t_i, D2^{-n})$. Define $\pi_j : T \to T_j$ via

$$\pi_j(t) = \arg\min_{s \in T_j} d(t, s),$$

the closest point to t in T_j. In particular, $d(t, \pi_j t) = d(t, T_j)$. Recall that $T_0 = \{t_0\}$ and note that

$$X(t) - X(t_0) = \sum_{j=0}^{\infty} (X(\pi_{j+1} t) - X(\pi_j t)).$$

Remark 1. T_j is not necessarily a subset of T (but may be some $\mathcal{A} \supset T$).

Next, it is easy to see that

$$\sup_{t \in T} |X(t) - X(t_0)| \leq \sum_{j \geq 0} \sup_{t \in T} |X(\pi_{j+1} t) - X(\pi_j t)|,$$

hence $\mathbb{E} \sup_{t \in T} |X(t) - X(t_0)| \leq \sum_{j \geq 0} \mathbb{E} \sup_{t \in T} |X(\pi_{j+1} t) - X(\pi_j t)|$. Note that

$$d(\pi_j t, \pi_{j+1} t) \leq d(t, \pi_j t) + d(t, \pi_{j+1} t) = D(T)2^{-j} + D(T)2^{-(j+1)},$$

hence

$$\mathbb{E} \sup_{t \in T} |X(\pi_{j+1} t) - X(\pi_j t)| \leq \mathbb{E} \max_{t_1 \in T_j,\ t_2 \in T_{j+1}} |X(t_1) - X(t_2)|$$

where the last maximum is taken over all t_1, t_2 such that $d(t_1, t_2) \leq 3D(T)2^{-(j+1)}$. The total number of such pairs is at most $\text{card}(T_j)\,\text{card}(T_{j+1}) = N(D(T)2^{-j}) \cdot N(D(T)2^{-(j+1)})$. Using the bound for the expected maximum of sub-Gaussian random variables, we deduce that

$$\mathbb{E} \sup_{t \in T} |X(\pi_{j+1} t) - X(\pi_j t)| \leq 3D(T)2^{-(j+1)}\, 2\sqrt{2H\left(D(T)2^{-(j+1)}\right)},$$

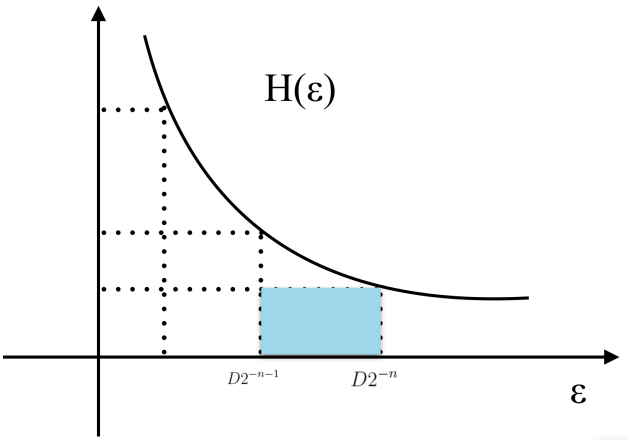

Figure 1.1. The area of the shadow is upper bounded by the area (integral) under the curve.

hence

$$\mathbb{E}\sup_{t\in T}|X(\pi_{j+1}t)-X(\pi_j t)|\le 12\sqrt{2}\sum_{j\ge 0}D(T)2^{-(j+2)}\sqrt{H\left(D(T)2^{-(j+1)}\right)}.$$

The latter can be viewed as a (lower) Riemann sum for the integral (see figure 1.1)

$$\int_0^\infty \sqrt{H(\varepsilon)}d\varepsilon,$$

hence the result follows. □

Is the obtained bound optimal? The (partial) answer is provided by the so-call "Sudakov's minoration" bound. It can be seen as a weak converse to Dudley's Inequality.

Theorem 2 (Sudakov minoration)**.** Let $\{X(t), t\in T\}$ be a Gaussian process and $d(t,s)=\sqrt{\mathrm{Var}(X(t)-X(s))}$. Then for any $t_0\in T$,

$$\mathbb{E}\sup_{t\in T}|X(t)-X(t_0)|\ge \mathrm{C}\cdot\sup_{\varepsilon>0}\varepsilon\sqrt{H(T,d,\varepsilon)},$$

where C is a numerical constant.

Proof. See section 2.4.2 in the book by Nickl and Giné. □

A sharp result was proven much later by M. Talagrand. It states that the "correct" way to measure the expected supremum of a Gaussian process is via the generic chaining complexity.

Theorem 3 (Talagrand)**.** Let $\{X(t), t\in T\}$ be a Gaussian process and $d(t,s)=\sqrt{\mathrm{Var}(X(t)-X(s))}$. Then there exists a numerical constant $C>0$ such that

$$\frac{1}{C}\gamma_2(T,d)\le \sup_{t\in T}(X(t)-X(t_0))\le C\gamma_2(T,d),$$

where $\gamma_2(T,d)=\inf_{\{T_n \text{ admissible }\}}\sup_{t\in T}\sum_{n\ge 0}2^{n/2}d(t,T_n)$ is the generic chaining complexity (see previous week's lectures).

We proved the upper bound, but the surprising (and more difficult) fact is that the lower bound in terms of the generic chaining complexity holds.

1.2 Empirical and Rademacher Processes.

Recall that our goal is to understand the ways to estimate the random variable

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^{n} I\left\{Y_j \neq f(X_j)\right\} - P(Y\neq f(X)\right|.$$

Given a sequence $X_1,\dots,X_n \in S$, recall that P_n is the empirical measure concentrated on $X_1,...,X_n$, $P_n = \frac{1}{n}\sum_j^n \delta_{X_j}$. In particular, $P_n f = \int f dP_n = \frac{1}{n}\sum_j^n f(X_j)$. If $X_1 \sim P$, then $Pf = \mathbb{E}P_n f = \int f dP = \mathbb{E}f(X_1)$. For example, if $X_1,\dots,X_n \in \mathbb{R}$ and $\mathcal{F} = \{I\left\{(-\infty,t]\right\},\ t\in\mathbb{R}\}$, then $P_n I\left\{(-\infty,t]\right\} = \frac{1}{n}\sum_{j=1}^n I\left\{X_j\leq t\right\} := F_n(t)$ is the empirical distribution function. We define the *empirical process* indexed by the class $\mathcal{F}$ as

$$Z_n(f) = P_n f - Pf,\ f\in\mathcal{F}.$$

Given a class $\mathcal{F}$ and a functional $G:\mathcal{F}\mapsto\mathbb{R}$, define $\|G\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}}|G(f)|$ (it is easy to check that $\|\cdot\|_{\mathcal{F}}$ defines a pseudo-norm). Our question can be reformulated as follows: what is the size of

$$\sup_{f\in\mathcal{F}}|Z_n(f)| = \sup_{f\in\mathcal{F}}|P_n f - Pf| = \|P_n - P\|_{\mathcal{F}}?$$

A useful result that will help us to answer this question is the so-called *symmetrization inequality.*

1.3 Symmetrization inequality.

We start with the definition of a Rademacher process: let $\varepsilon_1,...,\varepsilon_n$ be i.i.d. Rademacher random variables. Assume that $\{\varepsilon_j\}_{j=1}^n$ and $\{X_j\}_{j=1}^n$ are jointly independent. Then the *Rademacher process* indexed by a class $\mathcal{F}$ is defined as

$$R_n(f) = \frac{1}{n}\sum_{j=1}^n \varepsilon_j f(X_j) = \frac{1}{n}\left\langle \begin{pmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_n\end{pmatrix}, \begin{pmatrix} f(x_1)\\ \vdots\\ f(x_n)\end{pmatrix}\right\rangle,\ f\in\mathcal{F}.$$

In other words, Rademacher process is akin to the "empirical correlation" of the function f with the random noise. Intuitively, if the class $\mathcal{F}$ is very large, then the maximum correlation should also be large. Precise meaning to this intuition is given by the symmetrization inequality.

Theorem 4 (Symmetrization inequality, due to E. Gine and J. Zinn)**.** Under the previously stated assumptions,

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_j^n f(X_j) - \mathbb{E}f(X_j)\right| \leq 2\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^n \varepsilon_j f(X_j)\right| := 2\mathbb{E}\|R_n\|_{\mathcal{F}}.$$

Proof. Let $Y_1, ..., Y_n$ be i.i.d. copies of $X_1, ..., X_n$, and independent of $X_1, \dots, X_n$. Note that $Pf = \mathbb{E}f(X_j) = \mathbb{E}(Y_j) = \frac{1}{n}\sum_{j=1}^n \mathbb{E}f(Y_j)$, hence

$$\mathbb{E}\sup_f |P_n f - Pf| = \mathbb{E}\sup_f |P_n f - \widetilde{\mathbb{E}}\widetilde{P}_n f|,$$

where $\widetilde{\mathbb{E}}$ is the expectation with respect to $Y_1, ..., Y_n$, and $\widetilde{P}_n f = \frac{1}{n}\sum_{j=1}^n f(Y_j)$.
Next, by Jensen's inequality (and since $\|\cdot\|_{\mathcal{F}}$, being a pseudo-norm, is a convex function),

$$\left|P_n f - \widetilde{\mathbb{E}}\widetilde{P}_n f\right| \le \widetilde{\mathbb{E}}\sup_f \left|P_n f - \widetilde{P}_n f\right|,$$

and

$$\mathbb{E}\sup_f \left|P_n f - \widetilde{\mathbb{E}}\widetilde{P}_n f\right| \le \mathbb{E}\sup_f \left|P_n f - \widetilde{P}_n f\right| = \mathbb{E}\sup_f \left|\sum_{j=1}^n \frac{1}{n}(f(X_j) - f(Y_j))\right|.$$

Note that $f(X_j) - f(Y_j) \stackrel{d}{=} (-1)(f(X_j) - f(Y_j))$ for any $1 \le j \le n$ (here, $\stackrel{d}{=}$ means equality of distributions). Hence, for any fixed sequence $\sigma_j \in \{\pm 1\}$, $j = 1, ..., n$,

$$\mathbb{E}\sup_f \left|\sum_{j=1}^n \frac{1}{n}(f(X_j) - f(Y_j))\right| = \mathbb{E}\sup_f \left|\sum_{j=1}^n \frac{1}{n}\sigma_j(f(X_j) - f(Y_j))\right|.$$

Taking the average over all possible choices of $\{\sigma_j\}_{j=1}^n$, we get

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}} \left|\sum_{j=1}^n \frac{1}{n}(f(X_j) - f(Y_j))\right| &= \frac{1}{2^n}\sum_{\{\sigma_j\}\in\{-1,1\}^n} \mathbb{E}\sup_{f\in\mathcal{F}} \left|\sum_{j=1}^n \frac{1}{n}\sigma_j(f(X_j) - f(Y_j))\right| \\
&= \mathbb{E}_\varepsilon \mathbb{E}_{X,Y} \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{j=1}^n \varepsilon_j(f(X_j) - f(Y_j))\right| \\
&\le \mathbb{E}_{\varepsilon,X} \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{j=1}^n \varepsilon_j f(X_j)\right| + \mathbb{E}_{\varepsilon,Y} \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{j=1}^n \varepsilon_j f(Y_j)\right| = 2\mathbb{E}\|R_n\|_{\mathcal{F}}.
\end{aligned}$$

□

Theorem 5 (de-Symmetrization inequality)**.**

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}} \ge \frac{1}{2}\mathbb{E}\|R_n\|_{\mathcal{F}_c} = \frac{1}{2}\mathbb{E}\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{j=1}^n \varepsilon_j(f(X_j) - \mathbb{E}f(X_j))\right|,$$

where $F_c = \{f - Pf, f \in F\}$.

Remark 2. Note that $\mathbb{E}\|R_n\|_{\mathcal{F}_c} \ge \mathbb{E}\|R_n\|_{\mathcal{F}} - \frac{1}{\sqrt{n}}\sup_{f\in\mathcal{F}}|Pf|$. Indeed,

$$\left|\frac{1}{n}\sum_{j=1}^n \varepsilon_j(f(X_j) - Pf)\right| \ge \left|\frac{1}{n}\sum_{j=1}^n \varepsilon_j f(X_j)\right| - \left|\frac{1}{n}Pf\sum_{j=1}^n \varepsilon_j\right|.$$

Moreover,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\cdot Pf\right| \le \sup_{f\in\mathcal{F}}|Pf|\cdot\mathbb{E}\left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\right|$$
$$\le \sup_{f\in\mathcal{F}}|Pf|\,\mathbb{E}^{1/2}\left(\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\right)^2 = \sup_{f\in\mathcal{F}}|Pf|\cdot\frac{1}{\sqrt{n}}.$$

Next, we will see how to apply this result to bound the excess risk. Let $\mathcal{C}$ be a collection of measurable subsets of S. We will identify $\mathcal{C}$ with a class of indicator functions $\{I_C:\ C\in\mathcal{C}\}$ and will write $\|P_n-P\|_{\mathcal{C}}$ for $\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{j=1}^{n}I\{X_j\in C\}-\Pr(X\in C)\right|$.
Recall that, given a finite collection of random variables $\{X_t,\ t\in T\}$ such that $X_t\sim \mathrm{SG}(\sigma_t^2)$,

$$\mathbb{E}\sup_{t\in T}|X_t|\le\sqrt{2}\max_{t\in T}\sigma_t\sqrt{\log\left(2\,\mathrm{card}(T)\right)}.$$

For example, let $X_t=\sum_{j=1}^{n}t_j\varepsilon_j$, where $t=(t_1,...,t_n)\in T\subset\mathbb{R}^n$, and $\varepsilon_1,...,\varepsilon_n$ are i.i.d. Rademacher random variables. By properties of sub-Gaussian random variables, $X_t\in\mathrm{SG}\left(\|t\|_2^2\right)$. If we set $R_n(T):=\sup_{t\in T}\left|\sum_{j=1}^{n}\varepsilon_j t_j\right|$, it follows that

$$\mathbb{E}R_n(T)\le\sqrt{2}\max_{t\in T}\|t\|_2\sqrt{\log\left(2\,\mathrm{card}(T)\right)}.$$

1.4 First applications of the symmetrization inequality

Let F $\subseteq S$ be a finite set (a collection of points in S), and let $\mathcal{C}$ be a collection of subsets of S. The shattering number is defined as

$$\Delta^{\mathcal{C}}(F)=\mathrm{card}\{\{C\cap F\},C\subseteq\mathcal{C}\},$$

where $\mathrm{card}(\mathcal{A})$ stands for the cardinality of a set $\mathcal{A}$.

Example 3. Let

$$F=\{x_1,...,x_n\}\subset\mathbb{R},$$
$$\mathcal{C}=\{(-\infty,t],t\in\mathbb{R}\}$$

Then $\Delta^{\mathcal{C}}(F)=n+1$.

If $\Delta^{\mathcal{C}}(F)=2^{\mathrm{card}(F)}$, we say that F is *shattered* by $\mathcal{C}$. Given an i.i.d. sample $X_1,X_2,\ldots,X_n$ and a set $C\in\mathcal{C}$, consider the stochastic process

$$\frac{1}{\sqrt{n}}Z_n(I_C):=\frac{1}{\sqrt{n}}\left(P_n-P\right)(C)=\frac{1}{n}\sum_{i=1}^{n}I\{X_i\in C\}-\Pr(X\in C).$$

Note that by the symmetrization inequality,

$$\frac{1}{\sqrt{n}}\mathbb{E}\sup_{C\in\mathcal{C}}|Z_n(I_C)| \le 2\mathbb{E}_X\mathbb{E}_\varepsilon\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{j=1}^n \varepsilon_j \underbrace{I(X_j\in C)}_{=t_j}\right| = 2\mathbb{E}_X\mathbb{E}_\varepsilon\sup_{t\in T}\left|\frac{1}{n}\sum_{j=1}^n\varepsilon_j t_j\right|.$$

Let $T=\{(I(X_1\in C)),\cdots,I(X_n\in C)\}$, and note that $\mathrm{card}(T)=\Delta^{\mathcal{C}}(F)$. Since the random variables $R_n(t)=\frac{1}{n}\sum_{j=1}^n\varepsilon_j t_j$ are sub-Gaussian $\mathrm{SG}(\|t\|_2^2/n^2)$ and

$$\sup_{t\in T}\frac{\|t\|_2^2}{n}=\sup_{C\in\mathcal{C}}P_n(C)\le 1, \tag{1.1}$$

the bound on the expected maximum of sub-Gaussian random variables gives

$$2\mathbb{E}_X\mathbb{E}_\varepsilon\sup_{t\in T}\left|\frac{1}{n}\sum_{j=1}^n\varepsilon_j t_j\right|\le 2\sqrt{2}\mathbb{E}_X\sqrt{\frac{1}{n}}\sqrt{\log 2\Delta^{\mathcal{C}}(\{X_1,\cdots,X_n\})}.$$

We have just proven the following result:

Theorem 6. Let $X_1,X_2,...,X_n$ be i.i.d. from P, and let $\mathcal{C}$ be a collection of subset. Then

$$\mathbb{E}\sup_{\mathcal{C}\in C}|P_n(C)-P(C)|\le\frac{2\sqrt{2}}{\sqrt{n}}\mathbb{E}\sqrt{\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)}.$$

Corollary 1. Assume that $X_1,...,X_n\in\mathbb{R}, X_1\sim F$ are i.i.d., and $\mathcal{C}=\{(-\infty,t],t\in\mathbb{R}\}$.

$$\mathbb{E}\sup_{t\in R}|F_n(t)-F(t)|\le\frac{2\sqrt{2}}{n}\sqrt{\log(2n+2)}.$$

Remark 3.

1. the logarithmic factor in the previous bound can be removed (we will later demonstrate the way to achieve this).
2. Since the mapping $x\mapsto\sqrt{x}$ is concave, Jensen's inequality implies that $\mathbb{E}\sqrt{Z}\le\sqrt{\mathbb{E}Z}$, hence
$$\mathbb{E}\sup_{\mathcal{C}\in C}|P_n(C)-P(C)|\le\frac{2\sqrt{2}}{n}\sqrt{\mathbb{E}\log 2\Delta^{\mathcal{C}}(\{X_1,\ldots,X_n\})}.$$

Note that the previous bound does not depend on the "size" of the subsets in $\mathcal{C}$. We will now proved an improved bound that takes $\sup_{C\in\mathcal{C}}P(C)$ into account: specifically, note that the estimate (1.1) is quite crude if $\sup_{C\in\mathcal{C}}P(C)$ is much smaller than 1. In what follows, we will write $a\vee b$ for $\max(a,b)$. Note: this result has not been proven during the lectures.

Theorem 7. Let $X_1,X_2,...,X_n$ be i.i.d. from P, and let $\mathcal{C}$ be a collection of subsets of S. Then there exists a numerical constant $K>0$ such that

$$\mathbb{E}\|P_n-P\|_{\mathcal{C}}\le K\left(\sqrt{\sup_{C\in\mathcal{C}}P(C)}\mathbb{E}\sqrt{\frac{\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)}{n}}\vee\frac{\mathbb{E}\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)}{n}\right).$$

Note: compare this result with Theorem 6 and convince yourself that it is stronger.

Proof. Proceeding as in the proof of Theorem 6, we deduce that

$$\begin{aligned}
\mathbb{E}\|P_n - P\|_{\mathcal{C}} &\le 2\mathbb{E}\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_1^n \varepsilon_j I(X_j \in C)\right| \le \frac{2\sqrt{2}}{n}\mathbb{E}\left(\sup_{C\in\mathcal{C}}\sqrt{P_n(C)}\sqrt{\log 2\Delta^{\mathcal{C}}(x_1,..x_n)}\right)\\
&= \frac{2\sqrt{2}}{n}\mathbb{E}\left(\sup_{C\in\mathcal{C}}\sqrt{(P_n-P)(C)+P(C)}\sqrt{\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)}\right)\\
&\le \frac{2\sqrt{2}}{n}\mathbb{E}\left(\sup_{C\in\mathcal{C}}\sqrt{(P_n-P)(C)}+\sup_{C\in\mathcal{C}}\sqrt{P(C)}\right)\sqrt{\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)})\\
\le \frac{2\sqrt{2}}{\sqrt{n}}\sqrt{\sup_{C\in\mathcal{C}}P(C)}\mathbb{E}\sqrt{\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)}&+\frac{2\sqrt{2}}{\sqrt{n}}\mathbb{E}\sqrt{\sup_{C\in\mathcal{C}}|P_n-P|(C)}\mathbb{E}\sqrt{\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)}\\
\le \frac{2\sqrt{2}}{\sqrt{n}}\sqrt{\sup_{C\in\mathcal{C}}P(C)}\mathbb{E}\sqrt{\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)}&+\frac{2\sqrt{2}}{\sqrt{n}}\mathbb{E}^{1/2}\|P_n-P\|_{\mathcal{C}}\mathbb{E}^{1/2}\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n),
\end{aligned}$$

where we used the bound $\sqrt{a+b} \le \sqrt{a}+\sqrt{b},\ a,b>0$, as well as the Cauchy-Schwartz inequality in the last step.
Note that for any positive α_1,α_2, $\alpha_1+\sqrt{w}\alpha_2 \le 2\max(\alpha_1,\sqrt{w}\alpha_2)$. Moreover,

$$w \le \alpha_1+\sqrt{w}\alpha_2 \implies w \le 2\alpha_1 \vee 4\alpha_2^2.$$

Take

$$\begin{aligned}
w &= \mathbb{E}\|P_n-P\|_{\mathcal{C}},\\
\alpha_1 &= \frac{2\sqrt{2}}{\sqrt{n}}\sqrt{\sup_{C\in\mathcal{C}}P(C)}\mathbb{E}\sqrt{\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)},\\
\alpha_2 &= \frac{2\sqrt{2}}{\sqrt{n}}\mathbb{E}^{1/2}\|P_n-P\|_{\mathcal{C}}\mathbb{E}^{1/2}\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)
\end{aligned}$$

to complete the proof. □

1.5 Vapnik - Chervonenkis combinatorics

Let $\mathcal{C}$ be a collection of subsets of S, and let $F=\{x_1,...,x_n\}$ be a finite set. Recall the definition of the shattering number:

$$\Delta^{\mathcal{C}}(F) = \text{card}\{C\cap F, C\in\mathcal{C}\}.$$

The growth function of the class $\mathcal{C}$ is defined via

$$m^{\mathcal{C}}(n) = \sup_{F\subseteq S,\text{card}(F)=n}\Delta^{\mathcal{C}}(F) = \sup\{\Delta^{\mathcal{C}}\{x_1,...,x_n\}, x_1,\ldots,x_n\in S\}.$$

Note that there are 2 possibilities:

1. $\forall n \geq 1, m^{\mathcal{C}}(n) = 2^n$,
2. for some $n \in \mathbb{N}$, $m^{\mathcal{C}}(n) < 2^n$.

The Vapnik-Chervonenkis (VC) dimension of $\mathcal{C}$ is defined as

$$V(\mathcal{C}) = \sup\{n \geq 1 : m^{\mathcal{C}}(n) = 2^n\}.$$

If $m^{\mathcal{C}}(n) = 2^n \ \forall n$, then $V(\mathcal{C}) = \infty$. If $V(\mathcal{C}) < \infty$, we say that $\mathcal{C}$ is the VC class of sets.

Example 4. Let $\mathcal{C} = \{(-\infty, t], t \in R\}$. For $n = 1$, all subsets are shattered. For $n = 2$, there exists a subset of 2 points that is not shattered, hence $V(\mathcal{C}) = 1$.

Example 5. Let $\mathcal{C} = \{$all rectangles in $\mathbb{R}^2$ with axis parrallel to the coordinates axes$\}$. We claim that $V(\mathcal{C}) = 4$. Do show it, we need to find a subsets of 4 points that are shattered, and show that one can not shatter any subset of 5 distinct points. Clearly, points with coordinates $(0, 1),\ (1, 0),\ (0, -1),\ (-1, 0)$ are shattered. Next, given $\{x_1, x_2, ..., x_5\} \subset \mathbb{R}^2$, let $x^{(1)}$ be the one with smallest x coordinate; $x^{(2)}$ be the one with largest x coordinates; $x^{(3)}$ be the one with smallest y coordinate; $x^{(4)}$ be the one with largest y coordinate. A rectangle that contains $x^{(1)}, ..., x^{(4)}$ will necessarily include the last point as well, hence the collection of points $\{x_1, x_2, ..., x_5\}$ is not shattered.
This example can be generalized to show that for the collection $\mathcal{C}$ of "boxes" in $\mathbb{R}^d$ with faces parallel to the coordinate planes, $V(\mathcal{C}) = 2d$.

Example 6. Let $\mathcal{C} = \{$all convex sets of $\mathbb{R}^2\}$. Then $V(\mathcal{C}) = \infty$. To demonstrate it, pick a collection of points on a circle and show that it is shattered by convex polygones.

Exercise (Hard!)**.** What is the VC dimension of the set of triangles in $\mathbb{R}^2$?