Math 547: Mathematical Foundations of Statistical Learning Theory

Lectures 1-3 — Week 9

Instructor: S. Minsker *Scribe: S. Minsker*

1.1 More on Vapnik - Chervonenkis combinatorics

The following result allows one to build new, more complex, VC classes from existing simpler ones. Given a finite collection of sets $C_1, \ldots, C_k$, let $\mathcal{A}(C_1, ..., C_k)$ be the algebra generated by these sets, that is, a collection of all sets that can be obtained from $C_1, \ldots, C_k$ via set-theoretic operations (unions, intersections, complements).

Theorem 1. Assume that $\mathcal{C}$ has finite VC dimension. Define

$$\mathcal{C}^{(k)} = \bigcup \left\{\mathcal{A}(C_1, ..., C_k) : \ C_1, \ldots, C_k \in \mathcal{C}\right\}.$$

Then $\mathcal{C}^{(k)}$ has finite VC dimension.

Proof. First notice that $\mathrm{card}(\mathcal{A}(C_1, ..., C_k)) \leq 2^{2^k}$ (why?). Let F be a finite set with $\mathrm{card}(F) = n$. Observe that

$$\mathcal{C}^{(k)} \cap F = (\mathcal{C} \cap F)^{(k)} = \cup\{\mathcal{A}(C_1, ..., C_k) : C_1 \in \mathcal{C} \cap F, ..., C_k \in \mathcal{C} \cap F\}.$$

Then clearly $\Delta^{\mathcal{C}(k)}(F) = \mathrm{card}((\mathcal{C} \cap F)^{(k)})$. Note that $\mathrm{card}(\mathcal{C} \in F) \leq m^{\mathcal{C}}(n)$ and that there at at most $m^{\mathcal{C}}(n)$ choices of $C_1, ..., C_k \in \mathcal{C} \in F$. For each of these, there are at most 2^{2^k} subsets in $\mathcal{A}(C_1, ..., C_k)$. Then

$$\mathrm{card}((\mathcal{C} \cap F)^{(k)}) \leq (m^{\mathcal{C}}(n))^k 2^{2^k}.$$

Since $\mathcal{C}$ has finite VC dimension V, it implies that $m^{\mathcal{C}}(n) \leq \left(\frac{ne}{V}\right)^V$. It follows that $\left(m^{\mathcal{C}}(n)\right)^k 2^{2^k} \leq \left(\frac{ne}{v}\right)^{vk} 2^{2^k}$ still grows polynomially in n, hence $\mathcal{C}^{(k)}$ must have finite VC dimension. □

Exercise. Assume that $\mathcal{C}_1, \mathcal{C}_2$ have finite VC-dimension, and show that $\mathcal{C}_1 \cap \mathcal{C}_2 = \{C_1 \cap C_2, C_1 \in \mathcal{C}_1, C_2 \in \mathcal{C}_2\}$ is also a VC class.

Example 1. The collection of polygons in $\mathbb{R}^d$ with $k \geq 1$ faces has finite VC-dimension (since they are defined by the intersection of finitely many half-spaces). For instance, triangles in $\mathbb{R}^2$ have VC-dimension equal to 7. What is the upper bound for this class of triangles implied by Theorem 1?

Our next goal is to derive sharper bounds for the uniform deviations of empirical means from the expectations using chaining techniques. To this end, we will need to prove the estimates for the covering numbers associated with VC classes of sets. To this end, let $\mathcal{C}$ be a VC-class of sets and $X_1, \ldots, X_n$ – an i.i.d. sample from distribution Π. We know that

$$\mathbb{E} \sup_{C \in \mathcal{C}} \left| \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}\left(X_j \in C\right) - \Pi(C) \right| \leq K \sqrt{\frac{\log m^{\mathcal{C}}(n)}{n}} \leq K' \sqrt{\frac{V(\mathcal{C}) \log n}{n}}.$$

Let's try to apply the chaining argument to obtain a sharper bound (without the log-factor). By the symmetrization inequality,

$$\mathbb{E}\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\left(X_j\in C\right)-\Pi(C)\right|\le 2\mathbb{E}_X\mathbb{E}_\varepsilon\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\mathbb{I}\left(X_j\in C\right)\right| \tag{1.1}$$
$$=\frac{2}{\sqrt{n}}\mathbb{E}_X\mathbb{E}_\varepsilon\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{j=1}^{n}\frac{\varepsilon_j\mathbb{I}\left(X_j\in C\right)}{\sqrt{n}}\right|=\frac{2}{\sqrt{n}}\mathbb{E}_X\mathbb{E}_\varepsilon\sup_{t\in T}\left|\sum_{j=1}^{n}\frac{\varepsilon_j t_j}{\sqrt{n}}\right|,$$

where $T=\{t=\left(\mathbb{I}\left(x_1\in C\right),...,\mathbb{I}\left(x_n\in C\right)\right),C\in\mathcal{C}\}\subset\{0,1\}^n$. Now we can apply Dudley's entropy integral bound to the Rademacher process $t\mapsto\sum_{j=1}^{n}\frac{\varepsilon_j t_j}{\sqrt{n}}$. Next, we will derive the bounds for the covering number of the set T associated with the VC class $\mathcal{C}$.

1.2 Haussler's Theorem

Note that the following stochastic process (indexed by the class C)

$$C\mapsto\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\varepsilon_j\mathbb{I}\left(X_j\in C\right)$$

has sub-Gaussian increments with sub-Gaussian parameter

$$\frac{1}{n}\sum_{j=1}^{n}\left(\mathbb{I}\left(X_j\in C_1\right)-\mathbb{I}\left(X_j\in C_2\right)\right)^2=\left\|\mathbb{I}_{C_1}-\mathbb{I}_{C_2}\right\|^2_{L_2(P_n)}=P_n\left(\mathbb{I}_{C_1}-\mathbb{I}_{C_2}\right)^2.$$

Let's define the (random!) semi-metric $d_n^2(f,g):=\|f-g\|_{L_2(\Pi_n)}$ where Π_n is the empirical distribution corresponding to the training data. Then, by Dudley's theorem

$$\mathbb{E}\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\left(x_j\in C\right)-P(C)\right|\le\frac{12\sqrt{2}}{\sqrt{n}}\int_0^\infty\sqrt{\log N(d_n,\mathcal{C},\varepsilon)}d\varepsilon. \tag{1.2}$$

The goal here is to find the (non-random!) bound for the random covering number $N(d_n,\mathcal{C},\varepsilon)$ when $\mathcal{C}$ is a VC class.

Definition 1. Given class of functions $\mathcal{F}:=\{f:S\to\mathbb{R}\}$,

$$H(\mathcal{F},\varepsilon):=\sup_{Q\in\mathcal{P}(S)}N(\mathcal{F},L_2(Q),\varepsilon),$$

where $\mathcal{P}(S)$ is the set of all probability measures on S, and

$$\|f-g\|_{L_2(Q)}=\int_S(f(x)-g(x))^2dQ(x).$$

Theorem 2 (Haussler)**.** If $\mathcal{C}$ is a VC class and $\mathcal{F}=\{I_C,C\in\mathcal{C}\}$, then

$$H(\mathcal{F},\varepsilon)\le 5V(\mathcal{C})\log\frac{B}{\varepsilon},$$

for an absolute constant B and for all $\varepsilon\le 1$.

Remark 1. Note that the bound of Haussler's theorem combined with (**??**) yields the inequality

$$\mathbb{E}\,\|P_n - P\|_{\mathcal{C}} \le K\sqrt{\frac{V(\mathcal{C})}{n}},$$

where K is an absolute constant (that can be written as a function of B – try to get an explicit value as an exercise). Hence, we improved the previous bound that had an additional log factor.

Proof. Note that for $C_1, C_2 \in \mathcal{C}$

$$\|I_{C_1} - I_{C_2}\|^2_{L_2(Q)} = \int_S (I_{C_1} - I_{C_2})^2(x)dQ = Q(C_1 \triangle C_2),$$

where $C_1 \triangle C_2 = (C_1 \cap \bar{C}_2) \cup (C_2 \cap \bar{C}_1)$ is the symmetric difference of C_1 and C_2. Let $\tilde{\mathcal{C}} = \{C_1 \triangle C_2 : C_1, C_2 \in \mathcal{C}\}$. Then $\tilde{\mathcal{C}}$ is VC class and $m^{\tilde{\mathcal{C}}} \le 2^{2^2}(m^{\mathcal{C}}(n))^2$ (see the lectures from the previous week). Hence for any probability measure P and a i.i.d. sample $X_1, \dots, X_n$ from P,

$$\mathbb{E}\,\|P_n - P\|_{\tilde{\mathcal{C}}} \le K\sqrt{\frac{\log m^{\tilde{\mathcal{C}}}(n)}{n}} \le K_1\sqrt{\frac{\log m^{\mathcal{C}}(n)}{n}},$$

for any probability measure $\mathbb{P}$, $n \ge 1$ and an absolute constant K_1.
Since expectation is the average, there exists an elementary outcome w such that

$$\|P_n(w) - P\|_{\mathcal{C}} \le K_1\sqrt{\frac{\log m^{\mathcal{C}}(n)}{n}},$$

which holds for some particular realization of the sample $X_1(w), \dots, X_n(w)$. It implies that $\forall C_1, C_2 \in \mathcal{C}$

$$P(C_1 \triangle C_2) \le P_n(C_1 \triangle C_2) + K_1\sqrt{\frac{\log m^{\mathcal{C}}(n)}{n}}.$$

Also, $\exists \mathcal{C}' \subseteq \tilde{\mathcal{C}}, \text{card}(\mathcal{C}') \le m^{\tilde{\mathcal{C}}}(n)$ such that $\mathcal{C}' \cap \{X_1(w), \dots, X_n(w)\} = \tilde{\mathcal{C}} \cap \{X_1(w), \dots, X_n(w)\}$. It means that $\forall C \in \mathcal{C}, \exists C' \in \mathcal{C}'$ such that $P_n(C \triangle C') = 0$. Combining this with the previous bound, we see that for any $C \in \mathcal{C}$, $\exists C' \in \mathcal{C}'$ such that

$$P(C \triangle C') \le K_1\sqrt{\frac{\log m^{\mathcal{C}}(n)}{n}},$$

or $\sqrt{P(C \triangle C')} \le K_2\left(\frac{\log m^{\mathcal{C}}(n)}{n}\right)^{1/4}$. The latter is equivalent to

$$\|\mathbb{I}_C - \mathbb{I}_{C'}\|_{L_2(P)} \le K_2\left(\frac{\log m^{\mathcal{C}}(n)}{n}\right)^{1/4}.$$

Therefore, $\mathcal{C}'$ is ε_n-net for $\mathcal{C}$ with respect to metric $L_2(P)$, where $\varepsilon_n \le K_2\left(\frac{\log m^{\mathcal{C}}(n)}{n}\right)^{1/4}$. Now, assume that $\varepsilon > 0$ is fixed. Take $n = \frac{BV(\mathcal{C})}{\varepsilon^5}$ for B large enough. We claim that $\varepsilon_n \le \varepsilon$. Indeed,

$$\varepsilon_n \le K_2\left(\frac{1}{B}\log\frac{Be}{\varepsilon^5}\varepsilon^5\right)^{1/4} \le \varepsilon,$$

for sufficiently large B. For this choice of ε_n,

$$\mathrm{card}(\mathcal{C}') \le \log m^{\mathcal{C}}(n) \le V(\mathcal{C}) \log \frac{ne}{V(\mathcal{C})},$$

since we've shown previously that $m^{\mathcal{C}}(n) \le \left(\frac{ne}{V(\mathcal{C})}\right)^{V(\mathcal{C})}$. Now if $n = \frac{BV(\mathcal{C})}{\varepsilon^5}$, then

$$\log(\mathrm{card}(\mathcal{C}')) \le 5V(\mathcal{C}) \log\left(\frac{Be}{\varepsilon}\right),$$

Hence, $\mathrm{H}(\mathcal{C}, \varepsilon) \le 5V(\mathcal{C}) \log\left(\frac{Be}{\varepsilon}\right)$. □

1.3 VC-subgraph classes

The material of this section is optional. Most of the details presented below won't be covered in class.

Previously developed theory of VC classes allows to control the "size" of the empirical processes indexed by *sets*, or, equivalently, indicator functions of sets. However, in many examples (such as AdaBoost), one has to control supremum of the processes

$$\mathcal{F} \ni f \mapsto (P_n - P)(f) = \frac{1}{n}\sum_{j=1}^{n} (f(X_j) - \mathbb{E}f(X)), \tag{1.3}$$

where $\mathcal{F} = \{f : S \mapsto \mathbb{R}\}$ is a class of bounded functions. For instance, in the AdaBoost case, $\mathcal{F}$ is a convex hull of a collection of binary classifiers. It turns out that a convenient way two approach this general case is to look at the *subgraphs* of the functions.

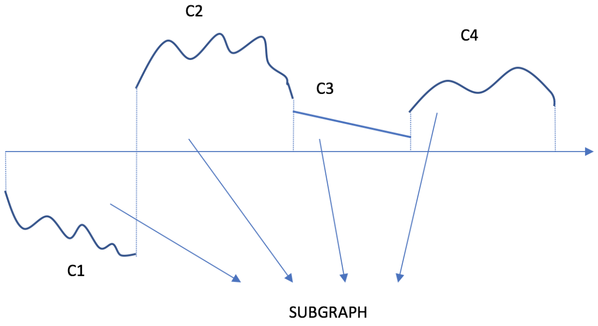

Figure 1.1. Subgraph of a function.

Definition 2. Given $f : S \mapsto \mathbb{R}$, the subgraph $\Gamma(f)$ of f is

$$\Gamma(f) := \{(x,t) \in S \times \mathbb{R} :\ 0 \le t < f(x)\} \cup \{(x,t) \in S \times \mathbb{R} :\ f(x) < t \le 0\}.$$

Definition 3. The class $\mathcal{F} = \{f : S \mapsto \mathbb{R}\}$ is called VC-subgraph if the collection of sets $\{\Gamma_f : \ f \in \mathcal{F}\}$ is a VC class.

Example 2. Let $\mathcal{C}$ be VC class of sets. Then the set of indicator functions $\{I_C, \ C \in \mathcal{C}\}$ is VC-subgraph.

Example 3. Let $P_{k,d}$ be a class of polynomials of d variables of degree at most k, and $\mathcal{C}$ – a VC class. Define

$$\mathcal{F} := \left\{ \sum_{j=1}^{m} p_j(x) I_{C_j} : \, p_j \in P_{k,d}, \, C_j \in \mathcal{C} \right\}$$

to be piecewise polynomial functions. Then $\mathcal{F}$ is VC-subgraph (exercise).

Example 4. Assume that $S = \mathbb{R}^d$, and let φ be a monotone function. Let

$$\mathcal{F} = \{\phi(\| \cdot - \theta\|_2) : \, \theta \in \mathbb{R}^d\}.$$

Then $\mathcal{F}$ is VC-subgraph. Indeed, consider the sets of the form $\{(x,t) : \ 0 \le \varphi(\|x - \theta\|_2) \le t\}$, i.e. the "top parts" of the subgraphs. Then

$$\begin{aligned} 0 \le t \le \varphi(\|x-\theta\|_2) &\Leftrightarrow 0 \le \varphi^{-1}(t) \le \|x-\theta\|_2 \\ &\Leftrightarrow \varphi^{-1}(t)^2 \le \|x-\theta\|_2^2 \\ &\Leftrightarrow \sum_{j=1}^{d} (x_j - \theta_j)^2 - \varphi^{-1}(t)^2 \ge 0 \end{aligned}$$

which is the set of positivity of a function $f(x_1, \ldots, x_d)$ that is a polynomial of degree 2. The dimension of the space of such functions is $d+2$ (to see this, expand the sum of squares).

The latter class is useful in the analysis of kernel density estimators. The kernel, being a function of bounded variation, can be written as a difference of two monotone functions. We will now return to the problem of estimating the supremum of the process defined in (**??**). Recall that we denote

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

and note that by the symmetrization inequality and Dudley's entropy bound,

$$\begin{aligned} \mathbb{E}\|P_n - P\|_{\mathcal{F}} &\le 2\mathbb{E}_X \mathbb{E}_\varepsilon \|R_n\|_{\mathcal{F}} = 2\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum \varepsilon_j f(X_j) \right| \\ &\le \frac{24\sqrt{2}}{\sqrt{n}} \mathbb{E}_X \int_0^\infty \sqrt{\log N(\mathcal{F}, d_n, \varepsilon)} \, d\varepsilon, \end{aligned} \tag{1.4}$$

where $d_n^2(f_1, f_2) = \frac{1}{n} \sum_{j=1}^{n} (f_1(X_j) - f_2(X_j))^2$ is a (random) distance. If we can estimate $N(\mathcal{F}, d_n, \varepsilon)$, then the previous bound will be useful. To this end, we will show that for VC subgraph classes of functions, it is possible to find an upper bound for the *uniform* covering number $\sup_Q N(\mathcal{F}, L_2(Q), \varepsilon)$, where the supremum is taken over all probability measures Q.

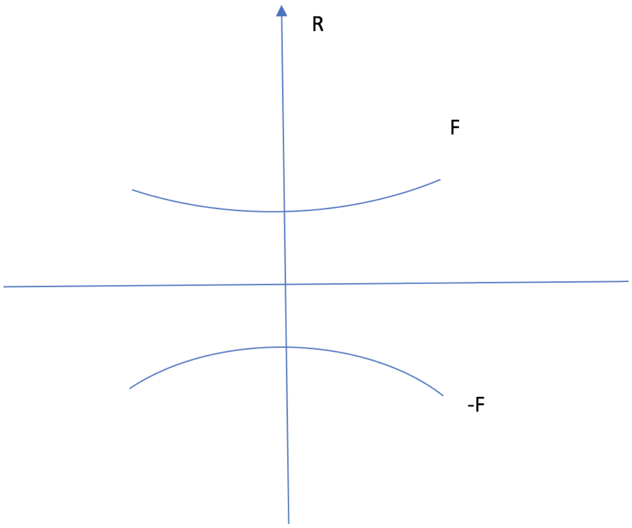

Figure 1.2. Envelope of the class.

We will need some additional notation. Given a class of functions $\mathcal{F}$, define

$$F(x) = \sup_{f \in \mathcal{F}} |f(x)|$$

to be the so-called "envelope" function of the class $\mathcal{F}$. Consider the space $S \times \mathbb{R}$ equipped with measure $Q \times \Lambda$, where Λ is the Lebesgue measure. Note that for any $f, g \in \mathcal{F}$,

$$\begin{aligned} \|f - g\|_{L_1(Q)} &:= \int_S |f - g| \, dQ = \int_S \int_0^{|f-g|} 1 \, dt \, dQ \\ &= \int_S \int_0^{\infty} I\{0 \le t \le |f - g|\} \, dt \, dQ \\ &= \int_0^{\infty} \int_S I\{0 \le t \le |f - g|\} \, dQ \, dt \\ &= (Q \times \Lambda)(\Gamma_f \Delta \Gamma_g), \end{aligned}$$

where we used Fubini-Tonelli theorem to change the order of integration. Here, $A \Delta B$ stands for the symmetric difference of sets A and B; the symmetric difference of two subgraphs corresponds to the area between the curves in figure **??**. Next, it is easy to see that the area corresponding to $\Gamma_f \Delta \Gamma_g$ is between the graphs of F and $-F$, hence, we can make $Q \times \Lambda$ a probability measure by normalizing it by $2 \int F \, dQ = (Q \times \Lambda)(\Gamma_F \Delta \Gamma_{-F})$. In other words, let

$$\mu(\Gamma_f \Delta \Gamma_g) := \frac{(Q \times \Lambda)(\Gamma_f \Delta \Gamma_g)}{2 \int F \, dQ},$$

whence $\|f - g\|_{L_1(Q)} = \mu(\Gamma_f \Delta \Gamma_g) \cdot 2 \int F \, dQ$. We will next use this identity to estimate the covering number of $\mathcal{F}$ with respect to $\| \cdot \|_{L_1(Q)}$ by the covering number of the VC class

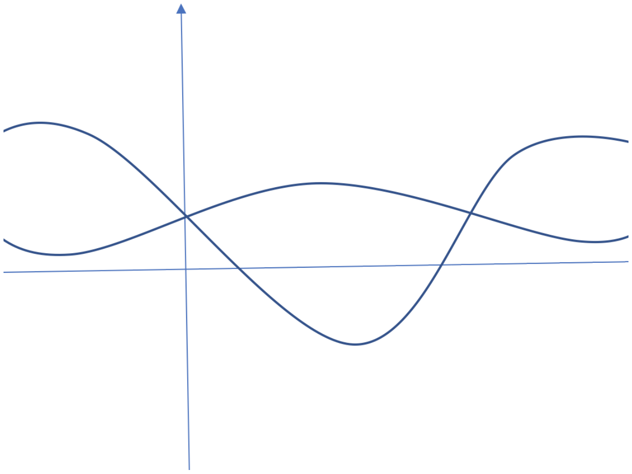

Figure 1.3. The are between the curves corresponds to the symmetric difference of subgraphs.

$\{\Gamma_f,\ f \in \mathcal{F}\}$ with respect to μ. We have just shown that

$$\int_S |f-g|dQ = \underbrace{\frac{(Q\times\Lambda)(\Gamma_f \Delta \Gamma_g)}{2\int F dQ}}_{\mu(\Gamma_f\Delta\Gamma_g)} \cdot 2 \int F dQ.$$

From now on, assume [1] that

$$\sup_{x\in S} |F(x)| \leqslant M.$$

Then

$$\int |f-g|dQ \leqslant \varepsilon \iff \underbrace{\mu(\Gamma_f \Delta \Gamma_g)}_{\|I_{\Gamma_f} - I_{\Gamma_g}\|^2_{L_2(\mu)}} \leqslant \frac{\varepsilon}{2\int F dQ} := \varepsilon'.$$

By our assumptions, $\Gamma_{\mathcal{F}} := \{\Gamma_f,\ f\in\mathcal{F}\}$ is the class of VC dimension V. Using the previously established bound on its covering numbers (Haussler's theorem), we get

$$\begin{aligned} N(\mathcal{F}, L_1(Q), \varepsilon) &\leqslant N(\Gamma_{\mathcal{F}}, L_2(\mu), \sqrt{\varepsilon'}) \\ &\leqslant \frac{5}{2} V(\Gamma_{\mathcal{F}}) \log \frac{2B^2\|F\|_{L_1(Q)}}{\varepsilon} \leqslant \frac{5}{2} V(\Gamma_{\mathcal{F}}) \log \frac{2B^2 M}{\varepsilon}, \end{aligned}$$

where B is an absolute constant. To obtain the bound in $L^2(Q)$ instead of $L_1(Q)$, note that

$$\|f-g\|^2_{L_2(Q)} = \int_S (f-g)^2 dQ \leqslant \int |f-g| \cdot 2F\, dQ.$$

Hence,

$$\|f-g\|_{L_1(Q)} \leqslant \frac{\varepsilon^2}{2M} \implies \int_S |f-g|\cdot 2F\, dQ \leqslant \varepsilon^2 \implies \|f-g\|_{L_2(Q)} \leqslant \varepsilon.$$

[1] This is a simplifying assumption that holds in examples we are interested in, but it is not necessary in general.

We deduce that

$$N(\mathcal{F}, L_2(Q), \varepsilon) \leqslant N\left(F, L_1(Q), \frac{\varepsilon^2}{2M}\right) \leqslant \frac{5}{2}V(\Gamma_{\mathcal{F}})\log\left(\frac{2B^2M\cdot 2M}{\varepsilon^2}\right) = 5V(\Gamma_{\mathcal{F}})\log\left(\frac{2BM}{\varepsilon}\right).$$

As we showed earlier in the course, Dudley's entropy integral bound implies that

$$\begin{aligned}\mathbb{E}\|P_n - P\|_{\mathcal{F}} &\leq \frac{24\sqrt{2}}{\sqrt{n}}\mathbb{E}_X\int_0^\infty \sqrt{\log N(\mathcal{F}, d_n, \varepsilon)}\,d\varepsilon \\ &\leq \frac{24\sqrt{2}}{\sqrt{n}}\int_0^\infty \sqrt{\log\left(\sup_Q N(\mathcal{F}, L_2(Q), \varepsilon)\right)}\,d\varepsilon,\end{aligned}$$

we obtain the bound

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}} \leq C\cdot M\sqrt{\frac{V(\Gamma_{\mathcal{F}})}{n}}$$

where $C > 0$ is another absolute constant.

1.4 Lower bounds

Next, we will establish a lower bound on $\mathbb{E}\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{j=1}^n \mathbb{I}\left(X_j \in C\right) - P(C)\right|$ for a class $\mathcal{C}$ of VC dimension V. To this end, let $\mathcal{A}$ be a class of distributions of (X, Y) such that if $P \in \mathcal{A}$, then there exists $C \in \mathcal{C}$ such that $Y = 1 \Leftrightarrow X \in C$. In other words, Y can be perfectly classified via $Y = \mathbb{I}\left(X \in C\right) - \mathbb{I}\left(X \in \bar{C}\right)$ for some $C \in \mathcal{C}$ (here, $\bar{C}$ is the complement of C). We called this the "realizable learning" framework in the beginning of the semester.

Theorem 3. The following inequality holds:

$$\inf_{g_n: S\to\{\pm 1\}}\sup_{P\in\mathcal{A}}\mathbb{E}_P\Pr\left(Y \neq g_n(X) | (X_j, Y_j)_{j=1}^n\right) \geqslant \frac{V-1}{2en}\left(1 - \frac{1}{n}\right),$$

where $g_n = g_n((X_1, Y_1), \cdots, (X_n, Y_n))$ is a binary classifier constructed by any algorithm.

Remark 2. We know that if $\hat{g}_n$ is the empirical risk minimizer over the binary classifiers that take value $+1$ on the sets $C \in \mathcal{C}$, then (taking into account the fact that the perfect classifier achieves 0 error)

$$\Pr\left(Y \neq \hat{g}_n(X) | (X_j, Y_j)_{j=1}^n\right) \leq 2\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{j=1}^n \mathbb{I}\left(X_j \in C\right) - P(C)\right|.$$

Therefore,

$$\begin{aligned}\sup_{P\in\mathcal{A}}\mathbb{E}_P\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum_{j=1}^n \mathbb{I}\left(X_j \in C\right) - P(C)\right| &\geq \frac{1}{2}\sup_{P\in\mathcal{A}}\mathbb{E}\Pr\left(Y \neq \hat{g}_n(X) | (X_j, Y_j)_{j=1}^n\right) \\ &\geq \frac{1}{2}\inf_{g_n: S\to\{\pm 1\}}\sup_{P\in\mathcal{A}}\mathbb{E}\Pr\left(Y \neq g_n(X) | (X_j, Y_j)_{j=1}^n\right),\end{aligned}$$

hence the bound of the theorem translates to the bound for the expected supremum.

Proof. Since the VC dimension of $\mathcal{C}$ is V, there exists a collection of points $\{x_1, \cdots, x_v\} \subseteq S$ that is shattered by $\mathcal{C}$. Consider the following family $\mathcal{A}' \subset \mathcal{A}$ of distributions:

$$X = \begin{cases} x_1, & \text{with probability } \dfrac{1}{n} \\ \vdots, & \\ x_{V-1}, & \text{with probability } \dfrac{1}{n} \\ x_V, & \text{with probability } 1 - \dfrac{V-1}{n}, \end{cases}$$

and

$$Y = f_b(X) = \begin{cases} b_i, & \text{if } X = x_i, i \leqslant V-1 \\ -1, & X = x_V, \end{cases}$$

where $b = \begin{pmatrix} b_1 \\ \vdots \\ b_{V-1} \end{pmatrix}$ takes values in the binary cube $\{-1,1\}^{V-1}$; let P_b be the corresponding distribution of (X,Y). Observe that

$$\begin{aligned} &\sup_{(X,Y)\sim P \in \mathcal{A}} \Pr\left(Y \neq g_n(X; (X_1,Y_1), \ldots (X_n,Y_n))\right) \\ &\geq \sup_{(X,Y)\sim P_b \in \mathcal{A}'} \Pr\left(Y \neq g_n(X; (X_1,Y_1), \ldots (X_n,Y_n))\right) \\ (\text{sup is} \geq \text{the avg.}) &\geq \sum_{b \in \{-1,1\}^{V-1}} P_b\left(y \neq g_n(x, (X_1,Y_1), \ldots, (X_n,Y_n)\right) \cdot \frac{1}{2^{V-1}} \\ &= \Pr\left(g_n(X, (X_1,Y_1), \ldots (X_n,Y_n)) \neq f_B(X)\right), \end{aligned}$$

where B has uniform distribution on $\{-1,1\}^{V-1}$ and is independent of everything else.

Assume that $Y = f_B(X)$. Note that if X is not among the observed $X_1, \ldots, X_n$ and is not equal to x_V, then it is misclassified with probability $1/2$ by any $\hat{g}_n$, hence

$$\begin{aligned} \Pr\left(g_n(X, (X_1,Y_1), \ldots (X_n,Y_n)) \neq f_B(X)\right) &\geq \frac{1}{2} \Pr\left(X \neq X_1, \ldots, X \neq X_n, X \neq x_V\right) \\ &= \frac{1}{2} \sum_{j=1}^{V-1} \frac{1}{n} \Pr\left(x_j \neq X_i \ , i = 1, \ldots, n\right) \\ &= \frac{1}{2} \sum_{j=1}^{V-1} \frac{1}{n} \left(1 - \frac{1}{n}\right)^n \\ &= \frac{1}{2} \frac{V-1}{n} \left(1 - \frac{1}{n}\right)^{n-1} \left(1 - \frac{1}{n}\right). \end{aligned}$$

Since $\left(1 - \frac{1}{n}\right)^{n-1} \geq \frac{1}{e}$, we obtain the result. □

Remark 3. Obtained rate $\frac{1}{n}$ is sharp for realizable learning in general. However, in the agnostic learning framework that we have mainly focused on, the lower bound can be improved to the order of $\sqrt{\frac{V}{n}}$. The corresponding result, and its proof, can be found in Chapter 28 of the book "Understanding Machine Learning."