STAT 583: Advanced Theory of Statistical Inference **Spring 2018**

Lecture 2: Uniform Entropy

Lecturer: Fang Han *April 16*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Lecturer.*

"God help us – for art is long, and life so short."

— Faust, Part I

The ultimate goal of this chapter is to determine, for a given possibly very general function class $\mathcal{F}$, the values of

$$\text{(Glivenko} - \text{Cantalli property)} \quad E\sup_{f\in\mathcal{F}}|(\mathbb{P}_n - P)f|$$

and

$$\text{(Donsker property)} \quad E\Big(\sup_{f,g\in\mathcal{F}:d_P(f,g)\leq\delta}|\sqrt{n}(\mathbb{P}_n - P)(f-g)|\Big).$$

It is clear that they both reduce to proving a maximal inequality, which is literally the core of empirical processes techniques.

2.1 VC classes of sets

2.1.1 Basic properties

Consider a class of sets $\mathcal{C} := \{C \in \mathcal{C}, C \subset \mathcal{X}\}$ and any sample $x_1^n = \{x_1, \ldots, x_n\} \subset \mathcal{X}$ of size n. We define $\mathcal{C}$'s growth function as follows.

Definition 1. *The growth function $\Pi_{\mathcal{C}}(n)$ is defined as*

$$\Pi_{\mathcal{C}}(n) := \max_{x_1^n\subset\mathcal{X}}|x_1^n \cap \mathcal{C}|.$$

Definition 2 (shattering)**.** *$\mathcal{C}$ is said to shatter a class $T \subset \mathcal{X}$ if $|T \cap \mathcal{C}| = 2^{|T|}$.*

Definition 3 (VC dimension)**.** *The VC dimension (or called VC index) of $\mathcal{C}$, written as $\nu(\mathcal{C})$, is the largest n such that there exists a set $T \subset \mathcal{X}$, $|T| = n$, and $\mathcal{C}$ shatters it.*

When the quantity $\nu(\mathcal{C})$ is finite, the class of sets $\mathcal{C}$ is said to be a *VC-class*.

Example 4. *Consider the class $\mathcal{C}_{\text{left}} := \{(-\infty, a]; a \in \mathbb{R}\}$. We have $\nu(\mathcal{C}_{\text{left}}) = 1$. On the other hand, it is easy to derive that $\Pi_{\mathcal{C}_{\text{left}}}(n) \leq n + 1 = (n + 1)^{\nu(\mathcal{C}_{\text{left}})}$.*

Example 5. *Consider the class $\mathcal{C}_{\text{two}} := \{(b, a]; a, b \in \mathbb{R}\}$. We have $\nu(\mathcal{C}_{\text{two}}) = 2$. On the other hand, it is easy to derive that $\Pi_{\mathcal{C}_{\text{left}}}(n) \leq (n + 1)^2 = (n + 1)^{\nu(\mathcal{C}_{\text{two}})}$.*

The following result shows that, for any VC class, the cardinality of $x_1^n \cap \mathcal{C}$ can grow at most polynomially in n. This is named the Sauer's Lemma.

Lemma 6 (Vapnik-Chervonenkis, Sauer, and Shelah)**.** *Consider a set class $\mathcal{C}$ with $\nu(\mathcal{C}) < \infty$. Then, for any collection of points $x_1^n = (x_1, \ldots, x_n)$, we have*

$$|x_1^n \cap \mathcal{C}| \le \sum_{i=0}^{\nu(\mathcal{C})} \binom{n}{i} \le \min\Big\{(n+1)^{\nu(\mathcal{C})}, \Big(\frac{en}{\nu(\mathcal{C})}\Big)^{\nu(\mathcal{C})}\Big\}.$$

Proof. The first inequality could be established through the following more general inequality. The rest two are simple algebra and are left to the readers.

Lemma 7. *Let A be a finite set and let $\mathcal{U}$ be a class of subsets of A. Then*

$$|\mathcal{U}| \le \Big|\Big\{B \subset A \mid B \text{ is shattered by } \mathcal{U}\Big\}\Big|.$$

To see how this lemma immediately proves Sauer's lemma, note that $B \subset A$ is shattered by $\mathcal{C}$ meaning that $|B| \le \nu(\mathcal{C})$. Consequently, if we let $A = x_1^n$ and set $\mathcal{U} = \mathcal{C} \cap A$, then Lemma 7 yields

$$|x_1^n \cap \mathcal{C}| = |\mathcal{C} \cap A| \le \Big|\Big\{B \subset A \mid B \le \nu(\mathcal{C})\Big\}\Big| \le \sum_{i=0}^{\nu(\mathcal{C})} \binom{n}{i}.$$

It remains to prove Lemma 7. For a given $x \in A$, let's define an operator on sets $U \in \mathcal{U}$ via

$$T_x(U) = \begin{cases} U \setminus \{x\} & \text{if } x \in U \text{ and } U \setminus \{x\} \notin \mathcal{U} \\ U & \text{otherwise.} \end{cases}$$

We let $T_x(\mathcal{U})$ be the new class of sets defined by applying T_x to each member of $\mathcal{U}$, namely, $T_x(\mathcal{U}) := \Big\{T_x(U) \mid U \in \mathcal{U}\Big\}$.

(1) We first show that T_x is a one-to-one map between $\mathcal{U}$ and $T_x(\mathcal{U})$, and hence $|\mathcal{U}| = |T_x(\mathcal{U})|$. This is equivalent to proving that, for any sets $U, U' \in \mathcal{U}$ such that $T_x(U) = T_x(U')$, we must have $U = U'$ (the reverse is simple). This is by the following case-by-case investigation:

- Case 1: $x \notin U$ and $x \notin U'$. We then have $U = T_x(U) = T_x(U') = U'$.
- Case 2: $x \notin U$ and $x \in U'$. In this case, we have $U = T_x(U) = T_x(U')$, so that $x \in U'$ but $x \notin T_x(U')$. But this means that $T_x(U') = U' \setminus \{x\} \notin \mathcal{U}$, which contradicts the fact that $T_x(U') = U \in \mathcal{U}$. By symmetry, the case $x \in U$ and $x \notin U'$ is identical.
- Case 3: $x \in U \cap U'$. If both $U \setminus \{x\}$ and $U' \setminus \{x\}$ belong to $\mathcal{U}$, then $U = T_x(U) = T_x(U') = U'$. If neither $U \setminus \{x\}$ nor $U \setminus \{x\}$ belongs to $\mathcal{U}$, then we also have $U \setminus \{x\} = U' \setminus \{x\}$, yielding $U = U'$. Lastly, if $U \setminus \{x\} \notin \mathcal{U}$ but $U' \setminus \{x\} \in \mathcal{U}$, then $T_x(U) = U \setminus \{x\} \notin \mathcal{U}$ but $T_x(U') = U' \in \mathcal{U}$, which is a contradiction.

(2) We secondly show that if $T_x(\mathcal{U})$ shatters a set B, then so does $\mathcal{U}$. If $x \notin B$, then both $\mathcal{U}$ and $T_x(\mathcal{U})$ pick out the same set of subsets of B, and the claim must be true. Otherwise, if $x \in B$, since $T_x(\mathcal{U})$ shatters B, for any subset $B' \subset B \setminus \{x\}$, there is a subset $T \in T_x(\mathcal{U})$ such that $T \cap B = B' \cup \{x\}$. Since $T = T_x(\mathcal{U})$ for some subset $U \in \mathcal{U}$ and $x \in T$, we conclude that both U and $U \setminus \{x\}$ must belong to $\mathcal{U}$, so that $\mathcal{U}$ also shatters B.

(3) We now conclude the lemma. Define the weight function $\omega(\mathcal{U}) = \sum_{U \in \mathcal{U}} |U|$. Note that applying a transformation T_x can only reduce this weight function: $\omega(T_x(\mathcal{U})) \le \omega(\mathcal{U})$. Consequently, by applying the transformations $\{T_x\}$ to $\mathcal{U}$ repeatedly, we can obtain a new class of sets $\mathcal{U}'$ such that $|\mathcal{U}| = |\mathcal{U}'|$ and the weight $\omega(\mathcal{U}')$ is minimal. Then, for any $U \in \mathcal{U}'$ and any $x \in U$, we have $U \setminus \{x\} \in \mathcal{U}'$ (otherwise, we have $\omega(T_x(\mathcal{U}')) < \omega(\mathcal{U}')$, contradicting minimality). Therefore, the set class $\mathcal{U}'$ shatters any one of its elements. Noting that $\mathcal{U}$ shatters at least as many subsets as $\mathcal{U}'$, and $|\mathcal{U}| = |\mathcal{U}'|$, the proof is complete. □

2.1.2 VC stability

The property of having finite VC-dimension is preserved under a number of basic operations, as summarized in the following (refer to, for example, Lemma 9.7 in K2008, Proposition 3.6.7 in GN2015, and Theorem 13.5 in "A Probabilistic Theory of Pattern Recognition" by Luc Devroye, Lszl Gyorfi, and Gabor Lugosi). They are also known as stability results in David Pollard's sense.

Theorem 8 (Stability)**.** *Let $\mathcal{C}$ and $\mathcal{D}$ be VC-classes on $\mathcal{X}$ with growth functions $\Pi_{\mathcal{C}}(n)$ and $\Pi_{\mathcal{D}}(n)$ and VC dimensions $V_{\mathcal{C}}$ and $V_{\mathcal{D}}$. Let $\mathcal{E}$ be VC-class on $\mathcal{W}$ with growth function $\Pi_{\mathcal{E}}(n)$ and VC dimension $V_{\mathcal{E}}$. We then have*

(1) $\mathcal{C}^C$ has VC-dimension $V_{\mathcal{C}}$ and growth function $\Pi_{\mathcal{C}}(n)$;

(2) $\mathcal{C} \cap \mathcal{D} = \{C \cap D; C \in \mathcal{C}, D \in \mathcal{D}\}$ has growth function $\leq \Pi_{\mathcal{C}}(n)\Pi_{\mathcal{D}}(n)$;

(3) $\mathcal{C} \cup \mathcal{D} = \{C \cup D; C \in \mathcal{C}, D \in \mathcal{D}\}$ has growth function $\leq \Pi_{\mathcal{C}}(n)\Pi_{\mathcal{D}}(n)$;

(4) $\mathcal{D} \times \mathcal{E}$ has growth function $\leq \Pi_{\mathcal{C}}(n)\Pi_{\mathcal{D}}(n)$;

(5) $\phi(\mathcal{C})$ has VC-dimension $V_{\mathcal{C}}$ if ϕ is one-to-one;

(6) $\psi^{-1}(\mathcal{C})$ has VC-dimension $\leq V_{\mathcal{C}}$.

Remark 9. *When you have an upper bound on the growth function of a given class of sets, by the definition of VC dimension, you also obtain an upper bound on that class by noticing that $\nu(\mathcal{C})$ is the largest n such that $2^n = \Pi_{\mathcal{C}}(n)$, and for any $n \in \mathbb{N}$, $\Pi_{\mathcal{C}}(n) \leq (n+1)^{\nu(\mathcal{C})}$.*

Theorem 8 is a nice result, but we still need something to begin with. Regarding any given real-valued function $g : \mathcal{X} \to \mathbb{R}$, it defines a "classification" function by the set $S_g := \{x \in \mathcal{X} \mid g(x) \leq 0\}$. In this way, we can associate the function class $\mathcal{G}$ with the collection of subsets $\mathcal{S}(\mathcal{G}) := \{S_g; g \in \mathcal{G}\}$.

In case the function class $\mathcal{G}$ is a vector space, the following result upper bounds the VC-dimension of the associated "classification" class $\mathcal{S}(\mathcal{G})$.

Proposition 10. *Let $\mathcal{G}$ be a vector space of functions $g : \mathbb{R}^d \to \mathbb{R}$ with dimension $\dim(\mathcal{G}) < \infty$. Then the class $\mathcal{S}(\mathcal{G})$ has VC-dimension at most $\dim(\mathcal{G})$.*

Proof. By definition of VC-dimension, we need to show that no collection of $n = \dim(\mathcal{G}) + 1$ points in $\mathbb{R}^d$ can be shattered by $\mathcal{S}(\mathcal{G})$. To this end, fix a collection x_1^n of n points in $\mathbb{R}^d$, and consider the following sets:

$$\Big\{(g(x_1), \ldots, g(x_n))^T, g \in \mathcal{G}\Big\}.$$

We then have, the range of the above sets is a linear subspace of $\mathbb{R}^n$ with dimension at most $\dim(\mathcal{G}) = n - 1 < n$. Therefore, there must exist a non-zero vector $a \in \mathbb{R}^n$ such that $\langle a, (g(x_1), \ldots, g(x_n))^T \rangle = 0$ for all $g \in \mathcal{G}$. We may assume, W.L.O.G., that at least one entry a_i of a is positive, and then write

$$\sum_{\{i|a_i>0\}} a_i g(x_i) = \sum_{\{i;a_i<0\}} (-a_i) g(x_i) \quad \text{for all } g \in \mathcal{G}.$$

Now suppose that there exists some $g \in \mathcal{G}$ such that the associate classification class $S_g = \{x \in \mathbb{R}^d; g(x) \leq 0\}$ includes only the subset $\{x_i : a_i \leq 0\}$. For such a function g, the LHS of the above equation would be strictly positive, while the RHS would be non-positive, which is a contradiction. We thus proved that $\mathcal{S}(\mathcal{G})$ cannot shatter x_1^n, and finish the proof. □

Example 11 (Linear functions in $\mathbb{R}^d$)**.** *For a pair $(a,b) \in \mathbb{R}^d \times \mathbb{R}$, consider the function class $f_{a,b}(x) = a^T x + b$, and consider the family $\mathcal{L}^d = \{f_{a,b} \mid (a,b) \in \mathbb{R}^d \times \mathbb{R}\}$. The associated classification is the collection of all half-spaces of the form $H_{a,b} := \{x \in \mathbb{R}^d \mid a^T x + b \le 0\}$. Since the family $\mathcal{L}^d$ forms a vector space of dimension $d+1$, we have $\mathcal{S}(\mathcal{L}^d)$ has VC-dimension at most $d+1$.*

Example 12 (Sphere in $\mathbb{R}^d$)**.** *Consider the sphere $S_{a,b} := \{x \in \mathbb{R}^d; \|x-a\|_2 \le b\}$ where $(a,b) \in \mathbb{R}^d \times \mathbb{R}^+$. Let $\mathbb{S}^d$ denote the collection of all such spheres. If we define the function*

$$f_{a,b}(x) := \|x\|_2^2 - 2\sum_{j=1}^{d} a_j x_j + \|a\|_2^2 - b^2$$

Then we have $\mathcal{S}_{a,b} = \{x \in \mathbb{R}^d; f_{a,b}(x) \le 0\}$, so that the sphere is a classification set of the function $f_{a,b}$. In order to leverage Proposition 10, we define a feature map $\phi : \mathbb{R}^d \to \mathbb{R}^{d+1}$ via

$$\phi(x) = (x_1, \ldots, x_d, 1),$$

and then consider the functions of the form

$$g_c(x) := c^T \phi(x) + \|x\|_2^2, \quad \text{where } x \in \mathbb{R}^{d+1}.$$

The family of functions $\{g_c; c \in \mathbb{R}^{d+1}\}$ is a vector space of dimension $d+2$, and it contains the functions $f_{a,b}$. We thus conclude $\nu(\mathbb{S}^d) \le d+2$.

Remark 13. *The VC-dimension should never be confused with the degree of freedom (or simply the number of parameters) in statistics. in fact, if you have a nonlinear classification function class, it is very possible that you will have a much higher VC-dimension than the number of parameters in your function. As an extreme case, the function class $\{\mathbb{1}(\sin ax > 0); a \in \mathbb{R}\}$ can have infinite VC-dimension.*

2.2 VC subgraph

2.2.1 Subgaussian processes and the chaining

Below is a quick glance of the chaining technique.

Definition 14 (Definition 2.3.5 in GN2015)**.** *A centered stochastic process $\{X(t), t \in T\}$ is subgaussian with respect to a pseudo-distance d on T if its increments satisfy the subgaussian inequality:*

$$E \exp(\lambda(X(t) - X(s))) \le \exp(\lambda^2 d^2(s,t)/2) \quad \text{for } \lambda \in \mathbb{R}, s, t \in T.$$

Remark 15. *As has been highlighted in Lecture 0, the process $\{\sqrt{n}(P_n - P)f; f \in \mathcal{F}\}$ is strongly related to the following associated process (called randomized empirical process):*

$$\Big\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i f(X_i); f \in \mathcal{F} \Big\},$$

where $\{\epsilon_i\}$ is the Rademacher sequence. Conditionally on the data $\{X_i\}$, it is a subgaussian process with respect to $d(f,g) = \|f - g\|_{L_2(\mathbb{P}_n)}$.

Given a pseudo-metric space (T,d), for any $\epsilon > 0$, its *covering number* $N(T,d,\epsilon)$ stands for the smallest number of closed ball with radius ϵ that could cover T. Analogously, its packing $D(T,d,\epsilon)$ stands for the largest number of closes balls with radius $\epsilon/2$ such that they could be packed into T. It is hence clear that

$$N(T,d,\epsilon) \le D(T,d,\epsilon) \le N(T,d,\epsilon/2).$$

Its *metric entropy* is defined as $\log N(T,d,\epsilon)$.

Theorem 16 (Theorems 2.3.6 and 2.3.7 in GN2015)**.** *Let (T, d) be a pseudo-metric space, and let $\{X(t), t \in T\}$ be a subgaussian process w.r.t the pseudo-metric d. Then,*

(1) For all finite subset $S \subset T$ and points $t_0 \in T$,

$$E \max_{t \in S} |X(t)| \leq E|X(t_0)| + 4\sqrt{2} \int_0^{D/2} \sqrt{\log 2N(T, d, \epsilon)} d\epsilon,$$

where D is the diameter of (T, d).

(2) Assume that

$$\int_0^{\infty} \sqrt{\log N(T, d, \epsilon)} d\epsilon < \infty.$$

Then

$$E \sup_{t \in T} |X(t)| \leq E|X(t_0)| + 4\sqrt{2} \int_0^{D/2} \sqrt{\log 2N(T, d, \epsilon)} d\epsilon.$$

2.2.2 VC subgraph classes of functions

Viewing Remark 15 and Theorem 16, for tackling a given maximal inequality $P \sup_{f \in \mathcal{F}} |\mathbb{G}_n f|$, the remaining part is to derive an upper bound for the metric entropy $\log N(\mathcal{F}, \|\cdot\|_{L_2(\mathbb{P}_n)}, \epsilon)$. This is doable via the Dudley-Pollard-Koltchinskii Universality Theorem by noticing that

$$\log N(\mathcal{F}, \|\cdot\|_{L_2(\mathbb{P}_n)}, \epsilon) \leq \sup_Q \log N(\mathcal{F}, \|\cdot\|_{L_2(Q)}, \epsilon),$$

where Q is any finitely discrete probability measure such that $\|F\|_{L_2(Q)} > 0$.

Definition 17 (Definition 3.6.8 in GN2015)**.** *The subgraph of a real function f on $\mathcal{X}$ is the set*

$$G_f := \Big\{(x, t) : x \in \mathcal{X}, t \in \mathbb{R}, t \leq f(x)\Big\}.$$

A class of functions $\mathcal{F}$ is VC subgraph of index (VC dimension) ν if the class of sets $\mathcal{C} := \{G_f; f \in \mathcal{F}\}$ is VC of index ν.

Example 18. *Suppose $\mathcal{C}$ is a VC class of index $\nu(\mathcal{C})$, then by definition the class of functions $\mathcal{F} := \{\mathbb{1}_C; C \in \mathcal{C}\}$ is VC subgraph of index $\nu(\mathcal{C})$.*

Example 19 (Lemma 2.6.15 in VW1996)**.** *Any finite-dimensional vector space $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ is VC-subgraph of index $\leq \dim(\mathcal{F}) + 1$.*

Proof. The proof resembles that of Proposition 10. Take any collection of $n = \dim(\mathcal{F})+2$ points $(x_1, t_1), \ldots, (x_n, t_n)$ in $\mathcal{X} \times \mathbb{R}$. Since $\mathcal{F}$ is a vector space, we have

$$\{(f(x_1) - t_1, \ldots, f(x_n) - t_n)^T, f \in \mathcal{F}\}$$

are contained in a $(\dim \mathcal{F} + 1) = (n - 1)$-dimensional subspace of $\mathbb{R}^n$. Hence, there exists a nondegenerate vector $a \neq 0$ such that

$$\sum_{a_i > 0} a_i \big(f(x_i) - t_i\big) = \sum_{a_i < 0} (-a_i)(f(x_i) - t_i), \quad \text{for every } f \in \mathcal{F},$$

where by default the sum over an empty set is set to be 0. WLOG, we pick out an a such that there exists at least one positive entry. For this vector, the set $\{(x_i, t_i) : a_i > 0\}$ cannot be of the form $\{(x_i, t_i) : t_i < f(x_i)\}$, since if then the LHS of the equation would be positive, and the RHS will be nonpositive. This concludes that the $\mathcal{F}$ is VC-subgraph of index $\leq \dim \mathcal{F} + 1$. □

We are now ready to state the main theorem in this chapter.

Theorem 20 (Dudley-Pollard-Koltchinskii-vanderVaart-Wellner Universality Theorem, Theorem 3.6.9 in GN2015)**.** *Let $\mathcal{F}$ be a non-empty VC subgraph class of index ν, and have an envelop $F \in L^p(\Omega, \mathcal{A}, Q)$ for some $1 \le p < \infty$. Set*

$$m_{v,w} := \max\Big\{m \in \mathbb{N} : \log m \ge m^{1/\nu - 1/w}\Big\}$$

for some $w > \nu$. We then have

$$D(\mathcal{F}, L^p(Q), \epsilon\|F\|_{p,Q}) \le m_{v,w} \vee \Big[2^{w/\nu}\Big(\frac{2^{p+1}}{\epsilon^p}\Big)^w\Big].$$

Proof. The proof uses probabilistic method tracing back to Paul Erdos and many other mathematicians who worked on number theory via probabilistic construction techniques. We omit Q in the norm when no confusion is made.

Let $f_1, \dots, f_m$ be a maximal collection of functions in $\mathcal{F}$ satisfying

$$Q|f_i - f_j|^p > \epsilon^p Q F^p, \quad \text{for } i \ne j,$$

so that $m = D(\mathcal{F}, L^p(Q), \epsilon\|F\|_p)$. For some k to be specified later, let $\{(x_i, t_i); i \in [k]\}$ be i.i.d. random vectors with law

$$\Pr\Big\{(x,t) \in A \times [a,b]\Big\} = \frac{\int_A \lambda[(-F(x)) \vee a, F(x) \wedge b] F^{p-1}(x) dQ(x)}{2QF^p}$$

for $A \subset \mathcal{X}$, real numbers $a < b$, and Lebesgue measure λ. In other words, x_i is chosen according to the law $P_F(A) = Q(\mathbb{1}_A F^p)/QF^p$, and given x_i, t_i is chosen uniformly on $[-F(x_i), F(x_i)]$.

The probability that at least two graphs have the same intersection with the sample $\{(x_i, t_i), i \in [k]\}$ is at most

$$\begin{aligned}
&\binom{m}{2} \max_{i \ne j} \Pr(C_i \text{ and } C_j \text{ have the same intersection with the sample}) \\
=&\binom{m}{2} \max_{i \ne j} \prod_{r=1}^k \Pr\{(x_r, t_r) \notin C_i \Delta C_j\} \\
=&\binom{m}{2} \max_{i \ne j} \prod_{r=1}^k \Big[1 - \Pr\Big\{(x_r, t_r) \in C_i \Delta C_j\Big\}\Big] \\
=&\binom{m}{2} \max_{i \ne j} \prod_{r=1}^k \Big[1 - \Pr\Big\{(x_r, t_r) : t_r \text{ is between } f_i(s_r), f_j(s_r)\Big\}\Big] \\
=&\binom{m}{2} \max_{i \ne j} \Big[1 - \frac{1}{\|F\|_p^p} \int \frac{|f_i - f_j|}{2F} F^p dQ\Big]^k \\
\le&\binom{m}{2} \max_{i \ne j} \Big[1 - \frac{1}{\|F\|_p^p} \int \frac{|f_i - f_j|^p}{(2F)^p} F^p dQ\Big]^k \\
\le&\binom{m}{2} \Big[1 - \frac{\epsilon^p}{2^p}\Big]^k \\
\le&\binom{m}{2} \exp(-\epsilon^p k / 2^p),
\end{aligned}$$

where in the last equation we use $1 - x \le \exp(-x)$.

Let k be such that this probability is less than 1. Then there exists a set of k elements such that graphs $C_i \in \mathcal{C},\ 1 \le i \le m$, intersect different subsets of this set, which implies that $\prod_{\mathcal{C}}(k) \ge m$. On the other hand, the smallest k such that $\binom{m}{2}\exp(-\epsilon^p k/2^p) < 1$ satisfies $k \le (2^{p+1}/\epsilon^p)\log m$. Then, by Sauer's Lemma, we have

$$m \le 2k^\nu \le 2\Big(\frac{2^{p+1}}{\epsilon^p}\log m\Big)^\nu.$$

Some algebra then gives the desired bound. □

The Universality Theorem, combined with the chaining theorem, renders the following corollary.

Corollary 21 (Theorems 3.5.1 and 3.5.4 in GN2015)**.** *Assuming $0 \in \mathcal{F}$, we have*

$$\begin{aligned}
E\sqrt{n}\|\mathbb{P}_n - P\|_{\mathcal{F}} &\le 8\sqrt{2}\cdot E\Big[\int_0^{\sqrt{\mathbb{P}_n F^2}} \sqrt{\log 2D(\mathcal{F}, L_2(P_n), \tau)}d\tau\Big] \\
&= 8\sqrt{2}\cdot E\Big[\|F\|_{\mathbb{P}_n,2}\int_0^1 \sqrt{\log 2D(\mathcal{F}, L_2(P_n), \epsilon\|F\|_{\mathbb{P}_n,2}}d\epsilon\Big] \\
&\le 8\sqrt{2}\cdot E\Big[\|F\|_{\mathbb{P}_n,2}\cdot\int_0^1 \sup_Q \sqrt{\log 2D(\mathcal{F}, L_2(Q), \epsilon\|F\|_{Q,2})}d\epsilon\Big] \\
&\le 8\sqrt{2}\|F\|_{P,2}\int_0^1 \sup_Q \sqrt{\log 2D(\mathcal{F}, L_2(Q), \epsilon\|F\|_{Q,2})}d\epsilon \\
&\lesssim \|F\|_{P,2}\int_0^1 \sqrt{v\log(A/\epsilon)}d\epsilon.
\end{aligned}$$

Example 22. *Using the above corollary, it is immediate to prove that*

$$E\sqrt{n}\|\mathbb{P}_n - P\|_{\mathcal{G}} = O(1)$$

by noticing that $\mathcal{G}$ is a VC-subgraph of index 1 and $\int_0^1 \sqrt{\log(A/\epsilon)}d\epsilon < \infty$.

We close this section with the VC-subgraph stability result, which is left for the students to verify.

Lemma 23 (VC-subgraph stability, Lemma 2.6.18 in VW1996)**.** *Let $\mathcal{F}$ and $\mathcal{G}$ be VC-subgraph classes of functions on a set $\mathcal{X}$ and $g : \mathcal{X} \to \mathbb{R}$, $\phi : \mathbb{R} \to \mathbb{R}$, and $\psi : \mathcal{Z} \to \mathcal{X}$ fixed functions. Then*

(i) $\mathcal{F} \wedge \mathcal{G} := \{f \wedge g : f \in \mathcal{F}, g \in \mathcal{G}\}$ *is VC-subgraph;*

(ii) $\mathcal{F} \vee \mathcal{G}$ *is VC-subgraph;*

(iii) $\{\mathcal{F} > 0\} := \{\{f > 0\} : f \in \mathcal{F}\}$ *is VC;*

(iv) $-\mathcal{F}$ *is VC-subgraph;*

(v) $\mathcal{F} + g := \{f + g : f \in \mathcal{F}\}$ *is VC-subgraph;*

(vi) $\mathcal{F} \cdot g := \{fg : f \in \mathcal{F}\}$ *is VC-subgraph;*

(vii) $\mathcal{F} \circ \psi := \{f(\psi) : f \in \mathcal{F}\}$ *is VC-subgraph;*

(viii) $\phi \circ \mathcal{F}$ *is VC-subgraph for monotone ϕ.*

2.2.3 VC-hull and VC-major

This section briefly introduces the VC-hull and VC-major classes, without touching too much detail due to the time limit. VC-hull and VC-major classes generalize the VC-subgraph (sometimes just referred to as VC) classes of functions.

Definition 24 (convex hull, Definition 3.6.13 in GN2015)**.** *Given a class of functions $\mathcal{F}$, $\mathrm{co}(\mathcal{F})$ is defined as the convex hull of $\mathcal{F}$, that is*

$$\mathrm{co}(\mathcal{F}) = \Big\{ \sum_{f\in\mathcal{F}} \lambda_f f : f \in \mathcal{F}, \sum_f \lambda_f = 1, \lambda_f > 0, \lambda_f \neq 0 \text{ only for finitely many } f \Big\},$$

and $\overline{\mathrm{co}}(\mathcal{F})$ is defined as the pointwise sequential closure of $\mathrm{co}(\mathcal{F})$, that is, $f \in \overline{\mathrm{co}}(\mathcal{F})$ if there exist $f_n \in \mathrm{co}(\mathcal{F})$ such that $f_n(x) \to f(x)$ for all $x \in \mathcal{X}$ as $n \to \infty$.

Definition 25 (VC-hull, Definition 3.6.13 in GN2015)**.** *If the class $\mathcal{F}$ is VC-subgraph, then we say that $\overline{\mathrm{co}}(\mathcal{F})$ is a VC-hull class of functions.*

Example 26 (Example 3.6.14 in GN2015)**.** *Let $\mathcal{F}$ be the class of all monotone nondecreasing functions $f : \mathbb{R} \to [0,1]$. Then $\mathcal{F} \in \overline{\mathrm{co}}(\mathcal{G})$, where $\mathcal{G} := \{\mathbb{1}_{(x,\infty)}, \mathbb{1}_{[x,\infty)} : x \in \mathbb{R}\}$.*

Proof. For any $f : \mathbb{R} \to [0,1]$, we could define

$$f_n = \frac{1}{n}\sum_{i=1}^{n-1} \mathbb{1}_{\{f>i/n\}} = \sum_{j=0}^{n-1} \frac{j}{n} \mathbb{1}_{\{j/n<f\le(j+1)/n\}}.$$

It is immediate that

$$\sup_{x\in\mathbb{R}} |f_n(x) - f(x)| \le 1/n.$$

On the other hand, since f is monotone nondecreasing (with possible jumps), we have the sets $\{f > i/n\}$ are all half lines, rendering that $\mathbb{1}_{\{f>i/n\}} \in \mathcal{G}$. □

Definition 27 (VC-major)**.** *$\mathcal{F}$ is a VC-major class if the collection of set $\{x : f(x) \ge t\}_{t\in\mathbb{R}, f\in\mathcal{F}}$ is a VC-class.*

Lemma 28 (Lemma 2.6.13 in VW1996)**.** *A bounded VC-major class is a scalar multiple of a VC-hull class.*

Proof. A given function $f : \mathcal{X} \to [0,1]$ is the uniform limit of the sequence

$$f_m = \sum_{i=1}^{m} \frac{1}{m} \mathbb{1}(f > i/m).$$

Thus, a given class of functions $f : \mathcal{X} \to [0,1]$ is contained in the pointwise sequential closure of the convex hull of $\{\mathbb{1}(f > t) : f \in \mathcal{F}, t \in \mathbb{R}\}$, which is VC-subgraph using Example 18 and the definition of VC-major class. This then finishes the proof. □

The last result in this chapter, which we shall not prove, is the Universality Theorem on VC-hull, and hence also on bounded VC-major classes.

Theorem 29 (Universality Theorem on VC-hull, Theorem 3.6.17 in GN2015)**.** *Let Q be a probability measure on $(\mathcal{X}, \sigma(\mathcal{X}))$, and let $\mathcal{F}$ be a collection of measurable functions with envelop $F \in L_2(Q)$ such that*

$$N(\mathcal{F}, L_2(Q), \epsilon\|F\|_{L_2(Q)}) \le C\epsilon^{-w}, \quad \text{for } 0 < \epsilon \le 1.$$

Then there exists a constant K depending only on C and w such that

$$\log N(\overline{\mathrm{co}}(\mathcal{F}), L_2(Q), \epsilon\|F\|_{L_2(Q)}) \le K\epsilon^{-2w/(w+1)}, \quad \text{for } 0 < \epsilon \le 1.$$

2.3 Donsker preservation

Following Chapter 1, in this section we will provide sufficient conditions for a class of measurable functions $\mathcal{F}$ to be Donsker. As a consequence of the Universality Theorem, let's consider the following uniform entropy integral (Equation (3.169) in GN2015, named as Kolchinskii-Pollard entropy) of a class of measurable functions $\mathcal{F}$ with envelope F:

$$J(\mathcal{F}, F, \delta) := \int_0^\delta \sup_Q \sqrt{\log 2N(\mathcal{F}, L_2(Q), \epsilon \|F\|_{L_2(Q)})} d\epsilon.$$

Obviously, there is no need to consider $\delta > 1$.

2.3.1 Donsker Theorem

The following theorem states a sufficient condition for $\mathcal{F}$ to be Donsker (asymptotic equicontinuity plus totally bounded).

Theorem 30 (Theorem 2.5.2 in VW1996)**.** *Let $\mathcal{F}$ be a class of measurable functions that satisfies*

(1) $J(\mathcal{F}, F, 1) < \infty$*;*

(2) the function classes

$$\mathcal{F}_\delta := \{f - g : f, g \in \mathcal{F}, \|f - g\|_{P,2} < \delta\} \quad \text{and} \quad \mathcal{F}_\infty^2 := \{(f - g)^2 : f, g \in \mathcal{F}\}$$

are P-measruable for every $\delta > 0$;

(3) $PF^2 < \infty$*.*

Then $\mathcal{F}$ is P-Donsker.

Proof. (i) We first prove asymptotic equicontinuity. Proving it is equivalent to proving

$$P(\|\mathbb{G}_n\|_{\mathcal{F}_{\delta_n}} > x) \to 0$$

for every fixed x and sequence $\delta_n \to 0$. Using Markov inequality, symmetrization, and chaining, we have

$$P(\|\mathbb{G}_n\|_{\mathcal{F}_{\delta_n}} > x) \le \frac{2}{x} E \Big\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(X_i) \Big\|_{\mathcal{F}_{\delta_n}} \le \frac{C}{x} \cdot E \int_0^{\theta_n} \sqrt{\log N(\mathcal{F}_{\delta_n}, L_2(P_n), \epsilon)} d\epsilon,$$

where

$$\theta_n^2 := \sup_{f \in \mathcal{F}_{\delta_n}} \|f\|_{\mathbb{P}_n,2}^2 := \sup_{f \in \mathcal{F}_{\delta_n}} \|f\|_n^2.$$

Noticing that, for any probability measure Q,

$$N(\mathcal{F}_{\delta_n}, L_2(Q), \epsilon) \le N(\mathcal{F}_\infty, L_2(Q), \epsilon) \le N^2(\mathcal{F}, L_2(Q), \epsilon/2),$$

we have

$$\int_0^{\theta_n} \sqrt{\log N(\mathcal{F}_{\delta_n}, L_2(P_n), \epsilon)} d\epsilon \le C \|F\|_n \int_0^{\theta_n / 2\|F\|_n} \sup_Q \sqrt{\log N(\mathcal{F}, L_2(Q), \epsilon \|F\|_{Q,2})} d\epsilon.$$

By Cauchy-Swartz and dominated convergence theorem, via Condition (1), it now suffices to proving $\theta_n \xrightarrow{P} 0$ as $\delta_n \to 0$.

We have, by Condition (2), $\sup\{Pf^2, f \in \mathcal{F}_\delta\} \to 0$ and $\mathcal{F}_{\delta_n} \subset \mathcal{F}_\infty$. Thusly, proving $\theta_n \xrightarrow{P} 0$ can be reduced to proving

$$\|\mathbb{P}_n f^2 - Pf^2\|_{\mathcal{F}_\infty} \xrightarrow{P} 0.$$

This is to prove G-C property for $\mathcal{F}_\infty^2$. By Condition (3) it has integrable envelope $(2F)^2$ and is measurable. In addition, for any pair $f, g \in \mathcal{F}_\infty$,

$$\mathbb{P}_n|f^2 - g^2| \le \mathbb{P}_n|f-g|4F \le \|f-g\|_n \|4F\|_n,$$

entailing

$$N(\mathcal{F}_\infty^2, L_1(\mathbb{P}_n), \epsilon\|2F\|_n^2) \le N(\mathcal{F}_\infty, L_2(\mathbb{P}_n), \epsilon\|F\|_n) \le N^2(\mathcal{F}, L_2(\mathbb{P}_n), \epsilon\|F\|_n/2).$$

Theorem 2.4.3 in VW1996 then kicks in to conclude.

(ii) Total boundedness is a direct consequence of the previous arguments. In detail, previously we proved there exists a sequence of discrete (non-random) measures P_n such that $\|(P_n - P)f^2\|_{\mathcal{F}_\infty} \to 0$, we could take n sufficiently large such that $\|(P_n - P)f^2\|_{\mathcal{F}_\infty} \le \epsilon^2$. Triangle inequality then yields that the ϵ-net for $\mathcal{F}$ in $L_2(P_n)$ is a $\sqrt{2}\epsilon$-net in $L_2(P)$. By assumption, $N(\mathcal{F}, L_2(P_n), \epsilon) < \infty$, which then renders $N(\mathcal{F}, L_2(P), \epsilon) < \infty$ and completes the proof. □

Theorem 31 (Theorem 2.4.3 in VW1996)**.** *Let $\mathcal{F}$ be a measurable class of measurable functions with envelop F such that $PF < \infty$. If $\log N(\mathcal{F}, L_1(\mathbb{P}_n), \epsilon) = o_P(n)$ for every $\epsilon > 0$, then $\|\mathbb{P}_n - P\|_{\mathcal{F}} \xrightarrow{a.s.} 0$.*

2.3.2 BUEI classes

A key part in Donsker preservation is the preservation of the boundedness of uniform entropy integrals. In the sequel, following K2008, we call $\mathcal{F}$ to have bounded uniform entropy integral (BUEI) with envelope F if $J(\mathcal{F}, F, 1) < \infty$.

Lemma 32 (Lemma 9.14 in K2008)**.** *Let $\mathcal{F}$ and $\mathcal{G}$ be BUEI with envelopes F and G, and let $\phi : \mathbb{R} \to \mathbb{R}$ be a Lipschitz continuous function with Lipschitz constant $c < \infty$. Then*

(i) $\mathcal{F} \wedge \mathcal{G}$ is BUEI with envelope $F + G$;

(ii) $\mathcal{F} \vee \mathcal{G}$ is BUEI with envelope $F + G$;

(iii) $\mathcal{F} + \mathcal{G}$ is BUEI with envelope $F + G$;

(iv) $\phi(\mathcal{F})$ is BUEI with envelope $|\phi(f_0)| + c(|f_0| + F)$, provided $f_0 \in \mathcal{F}$.

Its proof is a HW problem.

An important property of BUEI is that it is closed w.r.t. multiplication. This, combined with Theorem 30, will give a set of sufficient conditions for $\mathcal{F} \cdot \mathcal{G}$ to be Donsker (which is a very useful result).

Theorem 33 (Theorem 9.15 in K2008)**.** *Let $\mathcal{F}$ and $\mathcal{G}$ be BUEI classes with envelopes F and G. Then $\mathcal{F} \cdot \mathcal{G} := \{fg : f \in \mathcal{F}, g \in \mathcal{G}\}$ is BUEI with envelope FG.*

Its proof is pretty cute, yet straightforward. It is left for the readers to verify.

2.3.3 Donsker preservation

Donsker class is preserved under several function operators.

For a class $\mathcal{F}$ of real-valued, measurable functions on the sample space $\mathcal{X}$, let $\overline{\mathcal{F}}$ be the set of all functions $f : \mathcal{X} \to \mathbb{R}$ for which there exists a sequence of functions $f_m \in \mathcal{F}$ such that $f_m \to f$ both pointwise and in $L_2(P)$. Let $\text{sconv}\mathcal{F}$ denote the set of convex combinations $\sum_{i=1}^{\infty} \lambda_i f_i$ with $f_i \in \mathcal{F}$ and $\sum |\lambda_i| \leq 1$ and the series converges both pointwise and in $L_2(P)$.

Theorem 34 (Theorem 2.10.1-3 in VW1996, Theorem 9.30 in K2008)**.** *Let $\mathcal{F}$ be a P-Donsker class. Then*

(i) For any $\mathcal{G} \subset \mathcal{F}$, $\mathcal{G}$ is P-Donsker.

(ii) $\overline{\mathcal{F}}$ is P-Donsker.

(iii) $\text{sconv}\mathcal{F}$ *is P-Donsker.*

Theorem 35 (Corollary 9.32 in K2008)**.** *Let $\mathcal{F}$ and $\mathcal{G}$ be Donsker classes. Then*

(i) $\mathcal{F} \cup \mathcal{G}$ and $\mathcal{F} + \mathcal{G}$ are Donsker.

(ii) If $\|P\|_{\mathcal{F} \cup \mathcal{G}} < \infty$, then $\mathcal{F} \vee \mathcal{G}$ and $\mathcal{F} \wedge \mathcal{G}$ are both Donsker.

(iii) If $\mathcal{F}$ and $\mathcal{G}$ are both uniformly bounded, then $\mathcal{F} \cdot \mathcal{G}$ is Donsker (a stronger version exists via Theorems 33 and 30).

(iv) If $\psi : \overline{R} \to \mathbb{R}$ is Lipschitz continuous, where R is the range of functions in $\mathcal{F}$, and $\|\psi(f)\|_{P,2} < \infty$ for at least one $f \in \mathcal{F}$, then $\psi(\mathcal{F})$ is Donsker.

(v) If $\|P\|_{\mathcal{F}} < \infty$ and g is a uniformly bounded and measurable function, then $\mathcal{F} \cdot g$ is Donsker.