CS281B/Stat241B. Statistical Learning Theory. Lecture 7.

Peter Bartlett

1. Uniform laws of large numbers
   (a) Glivenko-Cantelli theorem proof:
   Concentration. Symmetrization. Restrictions.
   (b) Symmetrization: Rademacher complexity.
   (c) Restrictions: growth function, VC dimension, ...

Glivenko-Cantelli Theorem

First example of a uniform law of large numbers.

Theorem: $\|F_n - F\|_\infty \overset{as}{\to} 0.$

Here, F is a cumulative distribution function, F_n is the empirical cumulative distribution function,

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1[X_i \geq x],$$

where $X_1, \ldots, X_n$ are i.i.d. with distribution F, and $\|F - G\|_\infty = \sup_t |F(t) - G(t)|$.

Proof of Glivenko-Cantelli Theorem

Theorem: $\|F_n - F\|_\infty \overset{as}{\to} 0$.
That is, $\|P - P_n\|_G \overset{as}{\to} 0$, where $G = \{x \mapsto 1[x \geq t] : t \in \mathbb{R}\}$.

We'll look at a proof that we'll then extend to a more general sufficient condition for a class to be Glivenko-Cantelli.

The proof involves three steps:

1. Concentration: with probability at least $1 - \exp(-2\epsilon^2 n)$,
$$\|P - P_n\|_G \leq \mathbf{E}\|P - P_n\|_G + \epsilon.$$
2. Symmetrization: $\mathbf{E}\|P - P_n\|_G \leq 2\mathbf{E}\|R_n\|_G$, where we've defined the **Rademacher process** $R_n(g) = (1/n)\sum_{i=1}^n \epsilon_i g(X_i)$ (and this leads us to consider restrictions of step functions $g \in G$ to the data),
3. Simple restrictions.

Proof of Glivenko-Cantelli Theorem: Concentration

First, since $g(X_i) \in \{0, 1\}$, we have that the following function of the random variables $X_1, \ldots, X_n$ satisfies the bounded differences property with bound $1/n$:

$$\sup_{g \in G} |Pg - P_n g|$$

The bounded differences inequality implies that, with probability at least $1 - \exp(-2\epsilon^2 n)$,

$$\|P - P_n\|_G \leq \mathbf{E}\|P - P_n\|_G + \epsilon.$$

Proof of Glivenko-Cantelli Theorem: Symmetrization

Second, we symmetrize by replacing Pg by $P_n'g = \frac{1}{n}\sum_{i=1}^n g(X_i')$. In particular, we have

$$\mathbf{E}\|P - P_n\|_G \leq \mathbf{E}\|P_n' - P_n\|_G.$$

[Why?]

Proof of Glivenko-Cantelli Theorem: Symmetrization

Now we symmetrize again: for any $\epsilon_i \in \{\pm 1\}$,

$$\mathbf{E} \sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} (g(X_i') - g(X_i)) \right| = \mathbf{E} \sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (g(X_i') - g(X_i)) \right|,$$

This follows from the fact that X_i and X_i' are i.i.d., and so the distribution of the supremum is unchanged when we swap them. And so in particular the expectation of the supremum is unchanged. And since this is true for any ϵ_i, we can take the expectation over any random choice of the ϵ_i. We'll pick them independently and uniformly.

Proof of Glivenko-Cantelli Theorem: Symmetrization

$$\begin{aligned}
&\mathbf{E}\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(g(X_i')-g(X_i))\right| \\
&\leq \mathbf{E}\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i')\right| + \sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i)\right| \\
&\leq 2\,\underbrace{\mathbf{E}\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i)\right|}_{\text{Rademacher complexity}} = 2\mathbf{E}\|R_n\|_G,
\end{aligned}$$

where we've defined the **Rademacher process** $R_n(g) = (1/n)\sum_{i=1}^{n}\epsilon_i g(X_i)$.

Proof of Glivenko-Cantelli Theorem: Restrictions

We consider the set of restrictions $G(X_1^n) = \{(g(X_1), \ldots, g(X_n)) : g \in G\}$:

$$2\mathbf{E}\|R_n\|_G = 2\mathbf{E} \sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(X_i) \right| = 2\mathbf{E}\mathbf{E} \left[\sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(X_i) \right| \middle| X_1^n \right].$$

But notice that the cardinality of $G(X_1^n)$ does not change if we order the data. That is,

$$\begin{aligned} |G((X_1, \ldots, X_n))| &= \left| G((X_{(1)}, \ldots, X_{(n)})) \right| \\ &= \left| \left\{ (1[X_{(1)} \geq t], \ldots, 1[X_{(n)} \geq t]) : t \in \mathbb{R} \right\} \right| \leq n + 1, \end{aligned}$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ is the data in sorted order (and so $X_{(i)} \geq t$ implies $X_{(i+1)} \geq t$).

Proof of Glivenko-Cantelli Theorem: Rademacher Averages

Finally, we use the following result.

Lemma: [Finite Classes] For $A \subseteq \mathbb{R}^n$ with $R^2 = \dfrac{\max_{a \in A} \|a\|_2^2}{n}$,

$$\mathbf{E} \sup_{a \in A} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i a_i \leq \sqrt{\frac{2R^2 \log |A|}{n}}.$$

Hence

$$\mathbf{E} \sup_{a \in A} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i a_i \right| = \mathbf{E} \sup_{a \in A \cup -A} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i a_i \leq \sqrt{\frac{2R^2 \log(2|A|)}{n}}.$$

Proof of Rademacher Averages Result

Proof of Glivenko-Cantelli Theorem

For the class G of step functions, $R \leq 1/\sqrt{n}$ and $|A| \leq n+1$. Thus, with probability at least $1 - \exp(-2\epsilon^2 n)$,

$$\|P - P_n\|_G \leq \sqrt{\frac{8\log(2(n+1))}{n}} + \epsilon.$$

By Borel-Cantelli, $\|P - P_n\|_G \stackrel{as}{\to} 0$.

Recall: Glivenko-Cantelli Classes

Definition: F is a **Glivenko-Cantelli class** for P if

$$\|P_n - P\|_F \overset{P}{\to} 0.$$

GC Theorem:

$$\|P_n - P\|_G \overset{as}{\to} 0,$$

for $G = \{x \mapsto 1[x \leq \theta] : \theta \in \mathbb{R}\}$.

Uniform laws and Rademacher complexity

The proof of the Glivenko-Cantelli Theorem involved three steps:

1. Concentration of $\|P - P_n\|_F$ about its expectation.
2. Symmetrization, which bounds $\mathbf{E}\|P - P_n\|_F$ in terms of the Rademacher complexity of F, $\mathbf{E}\|R_n\|_F$.
3. A combinatorial argument showing that the set of restrictions of F to X_1^n is small, and a bound on the **Rademacher complexity** using this fact.

We'll follow a similar path to prove a more general uniform law of large numbers.

Uniform laws and Rademacher complexity

Definition: The **Rademacher complexity** of F is $\mathbf{E}\|R_n\|_F$, where the empirical process R_n is defined as

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i),$$

where the $\epsilon_1, \ldots, \epsilon_n$ are Rademacher random variables: i.i.d. uniform on $\{\pm 1\}$.

Note that this is the expected supremum of the alignment between the random $\{\pm 1\}$-vector ϵ and $F(X_1^n)$, the set of n-vectors obtained by restricting F to the sample $X_1, \ldots, X_n$.

Uniform laws and Rademacher complexity

Theorem: For any F, $\mathbf{E}\|P - P_n\|_F \leq 2\mathbf{E}\|R_n\|_F$.
If $F \subset [0,1]^{\mathcal{X}}$,

$$\frac{1}{2}\mathbf{E}\|R_n\|_F - \sqrt{\frac{\log 2}{2n}} \leq \mathbf{E}\|P - P_n\|_F \leq 2\mathbf{E}\|R_n\|_F,$$

and, with probability at least $1 - 2\exp(-2\epsilon^2 n)$,

$$\mathbf{E}\|P - P_n\|_F - \epsilon \leq \|P - P_n\|_F \leq \mathbf{E}\|P - P_n\|_F + \epsilon.$$

Thus, $\mathbf{E}\|R_n\|_F \to 0$ iff $\|P - P_n\|_F \stackrel{as}{\to} 0$.

That is, the sup of the empirical process $P - P_n$ is concentrated about its expectation, and its expectation is about the same as the expected sup of the Rademacher process R_n.

Uniform laws and Rademacher complexity

The first result is the symmetrization that we saw earlier:

$$\begin{aligned}
\mathbf{E}\|P - P_n\|_F &\le \mathbf{E}\|P'_n - P_n\|_F \\
&= \mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(f(X'_i) - f(X_i))\right\|_F \\
&\le 2\mathbf{E}\|R_n\|_F.
\end{aligned}$$

where R_n is the Rademacher process $R_n(f) = (1/n)\sum_{i=1}^{n}\epsilon_i f(X_i)$.

Uniform laws and Rademacher complexity

The second inequality (*desymmetrization*) follows from:

$$
\begin{aligned}
\mathbf{E}\|R_n\|_F &\le \mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(f(X_i)-\mathbf{E}f(X_i)\right)\right\|_F + \mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbf{E}f(X_i)\right\|_F \\
&\le \mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(f(X_i)-f(X_i')\right)\right\|_F + \|P\|_F\,\mathbf{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right| \\
&= \mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(f(X_i)-\mathbf{E}f(X_i)+\mathbf{E}f(X_i')-f(X_i')\right)\right\|_F \\
&\qquad + \|P\|_F\,\mathbf{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right| \\
&\le 2\mathbf{E}\,\|P_n-P\|_F + \sqrt{\frac{2\log 2}{n}}.
\end{aligned}
$$

Uniform laws and Rademacher complexity

And this shows that $\|P - P_n\|_F \overset{as}{\to} 0$ implies $\mathbf{E}\|R_n\|_F \to 0$.

The last inequality follows from the triangle inequality and the Finite Classes Lemma.

And Borel-Cantelli implies that $\mathbf{E}\|R_n\|_F \to 0$ implies $\|P - P_n\|_F \overset{as}{\to} 0$.

Controlling Rademacher complexity

So how do we control $\mathbf{E}\|R_n\|_F$? We'll look at several approaches:

1. $|F(X_1^n)|$ small. ($\max |F(x_1^n)|$ is the **growth function**)
2. For binary-valued functions: Vapnik-Chervonenkis dimension. Bounds rate of growth function. Can be bounded for parameterized families.
3. Structural results on Rademacher complexity: Obtaining bounds for function classes constructed from other function classes.
4. Covering numbers. Dudley entropy integral, Sudakov lower bound.
5. For real-valued functions: scale-sensitive dimensions.

Controlling Rademacher complexity: Growth function

For the class of distribution functions, $G = \{x \mapsto 1[x \leq \alpha] : \alpha \in \mathbb{R}\}$, we saw that the set of restrictions,

$$G(x_1^n) = \{(g(x_1), \ldots, g(x_n)) : g \in G\}$$

is always small: $|G(x_1^n)| \leq \Pi_G(n) = n + 1$.

Definition: For a class $F \subseteq \{0, 1\}^{\mathcal{X}}$, the **growth function** is

$$\Pi_F(n) = \max\{|F(x_1^n)| : x_1, \ldots, x_n \in \mathcal{X}\}.$$

Controlling Rademacher complexity: Growth function

Lemma: [Finite Class Lemma] For $f \in F$ satisfying $|f(x)| \leq 1$,

$$\begin{aligned} \mathbf{E}\|R_n\|_F &\leq \mathbf{E}\sqrt{\frac{2\log(|F(X_1^n) \cup -F(X_1^n)|)}{n}} \\ &\leq \sqrt{\frac{2\log(2\mathbf{E}|F(X_1^n)|)}{n}}. \end{aligned}$$

[where R_n is the Rademacher process:

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i).$$

and $F(X_1^n)$ is the set of restrictions of functions in F to $X_1, \ldots, X_n$.]

Controlling Rademacher complexity: Growth function

Proof: For $A \subseteq \mathbb{R}^n$ with $R^2 = \frac{\max_{a\in A} \|a\|_2^2}{n}$, we saw that

$$\mathbf{E} \sup_{a\in A} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i a_i \right| \le \sqrt{\frac{2R^2 \log(|A \cup -A|)}{n}}.$$

Here, we have $A = F(X_1^n)$, so $R \le 1$, and we get

$$\begin{aligned}
\mathbf{E}\|R_n\|_F &= \mathbf{E}\mathbf{E}\left[\|R_n\|_{F(X_1^n)} | X_1, \ldots, X_n\right] \\
&\le \mathbf{E}\sqrt{\frac{2\log(2|F(X_1^n)|)}{n}} \\
&\le \sqrt{\frac{2\mathbf{E}\log(2|F(X_1^n)|)}{n}} \\
&\le \sqrt{\frac{2\log(2\mathbf{E}|F(X_1^n)|)}{n}}.
\end{aligned}$$

Controlling Rademacher complexity: Growth function

e.g. For the class of distribution functions, $G = \{x \mapsto 1[x \geq \alpha] : \alpha \in \mathbb{R}\}$, we saw that $|G(x_1^n)| \leq n+1$. So $\mathbf{E}\|R_n\|_F \leq \sqrt{\frac{2 \log 2(n+1)}{n}}$.

e.g. F parameterized by k bits:

If g maps to $[0,1]$,

$$\begin{aligned} F &= \left\{x \mapsto g(x, \theta) : \theta \in \{0,1\}^k\right\}, \\ |F(x_1^n)| &\leq 2^k, \\ \mathbf{E}\|R_n\|_F &\leq \sqrt{\frac{2(k+1)\log 2}{n}}. \end{aligned}$$

Notice that $\mathbf{E}\|R_n\|_F \to 0$.

Growth function

Definition: For a class $F \subseteq \{0,1\}^{\mathcal{X}}$, the **growth function** is

$$\Pi_F(n) = \max\{|F(x_1^n)| : x_1, \ldots, x_n \in \mathcal{X}\}.$$

- $\mathbf{E}\|R_n\|_F \leq \sqrt{\frac{2\log(2\Pi_F(n))}{n}}$.
- $\Pi_F(n) \leq |F|$, $\lim_{n\to\infty} \Pi_F(n) = |F|$.
- $\Pi_F(n) \leq 2^n$. (But then this gives no useful bound on $\mathbf{E}\|R_n\|_F$.)
- Notice that $\log \Pi_F(n) = o(n)$ implies $\mathbf{E}\|R_n\|_F \to 0$.

Vapnik-Chervonenkis dimension

Definition: A class $F \subseteq \{0,1\}^{\mathcal{X}}$ **shatters** $\{x_1, \ldots, x_d\} \subseteq \mathcal{X}$ means that $|F(x_1^d)| = 2^d$.

The Vapnik-Chervonenkis dimension of F is

$$\begin{aligned} d_{VC}(F) &= \max\left\{d : \text{some } x_1, \ldots, x_d \in \mathcal{X} \text{ is shattered by } F\right\} \\ &= \max\left\{d : \Pi_F(d) = 2^d\right\}. \end{aligned}$$

Vapnik-Chervonenkis dimension: “Sauer’s Lemma”

Theorem: [Vapnik-Chervonenkis] $d_{VC}(F) \leq d$ implies

$$\Pi_F(n) \leq \sum_{i=0}^{d} \binom{n}{i}.$$

If $n \geq d$, the latter sum is no more than $\left(\frac{en}{d}\right)^d$.

So the VC-dimension is a single integer summary of the growth function: either it is finite, and $\Pi_F(n) = O(n^d)$, or $\Pi_F(n) = 2^n$. No other growth is possible.

$$\Pi_F(n) \begin{cases} = 2^n & \text{if } n \leq d, \\ \leq (e/d)^d \, n^d & \text{if } n > d. \end{cases}$$

Vapnik-Chervonenkis dimension: “Sauer’s Lemma”

Thus, for $d_{VC}(F) \leq d$ and $n \geq d$, we have

$$\mathbf{E}\|R_n\|_F \leq \sqrt{\frac{2\log(2\Pi_F(n))}{n}} \leq \sqrt{\frac{2\log 2 + 2d\log(en/d)}{n}}.$$