CS281B/Stat241B. Statistical Learning Theory. Lecture 6.

Peter Bartlett

1. Concentration inequalities
   (a) Martingale methods.
2. Uniform laws of large numbers
   (a) Motivation.
   (b) Glivenko-Cantelli theorem.

Martingale Difference Sequences: the Doob construction

$$\text{Define} \qquad X = (X_1, \ldots, X_n),$$
$$X_1^i = (X_1, \ldots, X_i),$$
$$Y_0 = \mathbf{E} f(X),$$
$$Y_i = \mathbf{E}[f(X) | X_1^i].$$

$$\text{Then} \qquad f(X) - \mathbf{E} f(X) = Y_n - Y_0 = \sum_{i=1}^{n} D_i,$$

where $D_i = Y_i - Y_{i-1}$. Also, Y_i is a **martingale** w.r.t. X_i, and hence D_i is a **martingale difference sequence**. [Why?]

Concentration Bounds for Martingale Difference Sequences

Theorem: Consider a martingale difference sequence D_n (adapted to a filtration $\mathcal{F}_n$) that satisfies

$$\text{for } |\lambda| \le 1/b_n \text{ a.s., } \mathbf{E}\left[\exp(\lambda D_n) \middle| \mathcal{F}_{n-1}\right] \le \exp(\lambda^2 \sigma_n^2/2).$$

Then $\sum_{i=1}^n D_i$ is sub-exponential, with $(\sigma^2, b) = (\sum_{i=1}^n \sigma_i^2, \max_i b_i)$.

$$P\left(\left|\sum_i D_i\right| \ge t\right) \le \begin{cases} 2\exp(-t^2/(2\sigma^2)) & \text{if } 0 \le t \le \sigma^2/b \\ 2\exp(-t/(2b)) & \text{if } t > \sigma^2/b. \end{cases}$$

Concentration Bounds for Martingale Difference Sequences

Proof:

Concentration Bounds for Martingale Difference Sequences

Theorem: Consider a martingale difference sequence D_i that a.s. falls in an interval of length B_i. Then

$$P\left(\left|\sum_i D_i\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_i B_i^2}\right).$$

Proof:

Bounded Differences Inequality

Theorem: Suppose $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the following **bounded differences inequality**:
for all $x_1, \ldots, x_n, x_i' \in \mathcal{X}$,

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq B_i.$$

Then

$$P\left(|f(X) - \mathbf{E}f(X)| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_i B_i^2}\right).$$

Bounded Differences Inequality

Proof: Use the Doob construction.

$$
\begin{aligned}
Y_i &= \mathbf{E}[f(X)|X_1^i], \\
D_i &= Y_i - Y_{i-1}, \\
f(X) - \mathbf{E}f(X) &= \sum_{i=1}^{n} D_i.
\end{aligned}
$$

Then ...

Examples: Rademacher Averages

For a set $A \subset \mathbb{R}^n$, consider

$$Z = \sup_{a \in A} \langle \epsilon, a \rangle,$$

where $\epsilon = (\epsilon_1, \ldots \epsilon_n)$ is a sequence of i.i.d. uniform $\{\pm 1\}$ random variables. Define the **Rademacher complexity** of A as $R(A) = \mathbf{E}Z/n$. [This is a measure of the size of A.] The bounded differences approach implies that Z is concentrated around $R(A)$:

Theorem: Z is sub-Gaussian with parameter $4 \sum_i \sup_{a \in A} a_i^2$.

Proof:

?

Examples: Empirical Processes

For a class F of functions $f : \mathcal{X} \to [0,1]$, suppose that $X_1, \ldots, X_n, X$ are i.i.d. on $\mathcal{X}$, and consider

$$Z = \sup_{f \in F} \left| \mathbf{E} f(X) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right| =: \left\| \underbrace{P - P_n}_{\text{emp proc}} \right\|_F .$$

If Z converges to 0, this is called a *uniform law of large numbers*. Here, we show that Z is concentrated about $\mathbf{E}Z$:

Theorem: Z is sub-Gaussian with parameter $1/n$.

Proof:

?

Uniform laws of large numbers: Motivation

We are interested in the performance of empirical risk minimization: Choose $f_n \in F$ to minimize $\hat{R}(f)$.

How does $R(f_n)$ behave?

Define $f^* = \arg\min_{f \in F} R(f)$.

How does the excess risk, $R(f_n) - R(f^*)$ behave?

We can write

$$R(f_n) - R(f^*) = \left[R(f_n) - \hat{R}(f_n)\right] + \left[\hat{R}(f_n) - \hat{R}(f^*)\right] + \left[\hat{R}(f^*) - R(f^*)\right]$$

Uniform laws of large numbers: Motivation

One of these terms is a difference between a sample average and an expectation for the fixed function $(x, y) \mapsto \ell(f^*(x), y)$:

$$\hat{R}(f^*) - R(f^*) = \frac{1}{n}\sum_{i=1}^{n} \ell(f^*(X), Y) - P\ell(f^*(X), Y).$$

The law of large numbers shows that this term converges to zero; and with information about the tails of $\ell(f^*(X), Y)$ (such as boundedness), we can get bounds on its value.

Uniform laws of large numbers: Motivation

$\hat{R}(f_n) - \hat{R}(f^*)$ is non-positive, because f_n is chosen to minimize $\hat{R}$.

The other difference, $R(f_n) - \hat{R}(f_n)$, is more interesting. For any fixed f, this difference goes to zero. But f_n is random, since it is chosen using the data. An easy upper bound is

$$R(f_n) - \hat{R}(f_n) \leq \sup_{f \in F} \left| R(f) - \hat{R}(f) \right|,$$

and this motivates the study of uniform laws of large numbers.

Glivenko-Cantelli Theorem

First example of a uniform law of large numbers.

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

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 $\|F - G\|_\infty = \sup_t |F(t) - G(t)|$.

Glivenko-Cantelli Theorem

Why *uniform* law of large numbers?

$$\begin{aligned}\|F_n - F\|_\infty &= \sup_x |F_n(x) - F(x)| \\ &= \sup_x |P_n(X \geq x) - P[X \geq x]| \\ &\stackrel{as}{\to} 0,\end{aligned}$$

where P_n is the empirical distribution that assigns mass $1/n$ to each X_i.

The law of large numbers says that, for all x, $P_n(X \geq x) \stackrel{as}{\to} P(X \geq x)$. The GC Theorem says that this happens uniformly over x.

Glivenko-Cantelli Classes

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

$$\sup_{f \in F} |P_n f - Pf| =: \|P_n - P\|_F \stackrel{P}{\to} 0.$$

Here, P is a distribution on $\mathcal{X}$, $X_1, \ldots, X_n$ are drawn i.i.d. from P, P_n is the empirical distribution (which assigns mass $1/n$ to each of $X_1, \ldots, X_n$), F is a set of measurable real-valued functions on $\mathcal{X}$ with finite expectation under P, $P_n - P$ is an **empirical process**, that is, a stochastic process indexed by a class of functions F, and $\|P_n - P\|_F := \sup_{f \in F} |P_n f - Pf|$.

The GC Theorem is a special case, with $F = \{1[x \geq t] : t \in \mathbb{R}\}$ (and with the stronger conclusion that convergence is almost sure—we say that such an F is a 'strong GC class').

Glivenko-Cantelli Classes

Not all F are Glivenko-Cantelli classes. For instance, recall

$$F = \{1[x \in S] : S \subset \mathbb{R}, \, |S| < \infty\} \, .$$

Then for a continuous distribution P, $Pf = 0$ for any $f \in F$, but $\sup_{f \in F} P_n f = 1$ for all n. So although $P_n f \stackrel{as}{\to} Pf$ for all $f \in F$, this convergence is not uniform over F. F is too large.