Math 547: Mathematical Foundations of Statistical Learning Theory Fall 2017

Lecture 16 — September 27

Instructor: Stas Minsker *Scribe: Mose Wintner*

16.1 Bernstein's inequality

Theorem 1. Let $X_1, \dots, X_n$ be independent, $\mathbb{E}X_j = 0$, $|X_j| \le M$ almost surely for $j = 1, \dots, n$. Let $B_n^2 = \sum_{j=1}^n \mathrm{Var} X_j$. Then

$$P\left(\left|\sum X_j\right| > t\right) \le 2\exp\left(\frac{-t^2/2}{B_n^2 + Mt/3}\right).$$

Proof: We will prove Bennett's inequality (slightly stronger). Let $\lambda > 0$,

$$P(S_n > t) = P(\lambda S_n > \lambda t) = P(e^{\lambda S_n}) > e^{\lambda t}) \le \inf_{\lambda > 0} \frac{\mathbb{E}e^{\lambda S_n}}{e^{\lambda t}}$$

for all $t > 0$, by Chebyshev's inequality. Then also

$$P(S_n < -t) = P(\sum -X_j > t) \le \inf_{\lambda > 0} \frac{\mathbb{E}e^{\lambda - S_n}}{e^{\lambda t}}.$$

Note that $\mathbb{E}e^{\lambda S_n} = \mathbb{E}e^{n\lambda X_1}$ and that

$$\mathbb{E}e^{\lambda X_1} = \mathbb{E}\sum_{j=1}^{\infty} \frac{(\lambda X_1)^j}{j!} = 1 + \frac{\lambda^2}{2}\mathbb{E}X^2 + \cdots + \frac{\lambda^k \mathbb{E}X^k}{k!} + \cdots$$

Since $\mathbb{E}X^k \le \mathbb{E}(|X|^{k-2}|X|^2) \le M^{k-2}\mathbb{E}X^2$, the above becomes

$$\begin{aligned}
&= 1 + \frac{\lambda^2}{2}\mathbb{E}X^2 + \lambda M \frac{\lambda^2}{3!}\mathbb{E}X^2 + \cdots + \frac{\lambda^2 \mathbb{E}X^2}{k!} M^{k-2}\lambda^{k-2} \\
&= 1 + \lambda^2 \mathbb{E}X^2 \left(\frac{1}{2!} + \frac{\lambda M}{3!} + \cdots + \frac{M^{k-2}\lambda^{k-2}}{k!} + \cdots\right) \\
&= 1 + \lambda^2 \mathbb{E}X^2 \left(\frac{e^{\lambda M} - 1 - \lambda M}{\lambda^2 M^2}\right) \mathbb{E}e^{\lambda X_j} \qquad \le 1 + \frac{\sigma_j^2}{M^2}\left(e^{\lambda M} - \lambda M - 1\right) \\
&\le \exp\left(\frac{\sigma_j^2}{M^2}\left(e^{\lambda M} - \lambda M - 1\right)\right)
\end{aligned}$$

since $1 + x \le e^x$ for $x \ge 1$. Finally,

$$\mathbb{E}e^{\lambda S_n} \le \exp \frac{B_n^2}{M^2}\left(e^{\lambda M} - \lambda M - 1\right).$$

It remains to minimize

$$F(\lambda) = \frac{B_n^2}{M^2}\left(e^{\lambda M} - \lambda M - 1\right) - \lambda t$$

over $\lambda > 0$. Setting $F' = 0$, we get

$$\lambda^* = \frac{1}{M} \log\left(1 + \frac{tM}{B_n^2}\right).$$

Plugging in this value, we get

$$F(\lambda^*) = -\frac{B_n^2}{M^2} H\left(\frac{tM}{B_n^2}\right),$$

where $H(u) := (1+u)\log(1+u) - u$ is related to the Kullback-Leibler divergence. This is Bennett's inequality; Bernstein's follows since

$$H(u) \geq \frac{u^2/2}{1 + u/3}.$$

16.1.1 Empirical and Rademacher processes

(X_j, Y_j) some training data for $j = 1, \ldots, n$. Let $\mathcal{F}$ be a collection of binary classifiers and let

$$\hat{f}_n = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum I\{Y_j \neq f(X_j)\}.$$

Now let

$$\overline{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E} I\{Y \neq f(X)\} = \operatorname*{argmin}_{f \in \mathcal{F}} Pr(Y \neq f(X)).$$

Let $f_*(x) = \text{sign}(\eta(x))$ be the Bayes classifier. Recall the excess risk

$$\mathcal{E}(f) = P(Y \neq f(X)) - P(Y \neq f_*(X)).$$

What is $\mathcal{E}(\hat{f}_n)$?

$$\begin{aligned}
\mathcal{E}(\hat{f}_n) &= \mathbb{E} I\{Y \neq \hat{f}_n(x)\} - P(Y \neq f_*(x)) \\
&= \mathbb{E} I\{Y \neq \hat{f}_n(x)\} - \mathbb{E} I\{Y \neq \overline{f}(x)\} + \mathcal{E}(\overline{f}),
\end{aligned}$$

where the last term is determined and called the approximation error. We will add and subtract $\frac{1}{n}\sum I\{Y_j \neq \overline{f}(X_j)\}$ and $\frac{1}{n}\sum I\{Y_j \neq \hat{f}(X_j)\}$. We know that

$$\frac{1}{n}\sum I\{Y_j \neq \hat{f}_n(X_j)\} - \frac{1}{n}\sum I\{Y_j \neq \overline{f}(X_j)\} < 0.$$

Combining this with what we had, we get

$$\begin{aligned}
\mathcal{E}(\hat{f}_n) &\leq \mathcal{E}(\overline{f}) + \mathbb{E} I\{Y \neq \hat{f}_n(x)\} - \frac{1}{n}\sum I\{Y_j \neq \hat{f}_n(X_j)\} \\
&\quad + \frac{1}{n}\sum I\{Y_j \neq \overline{f}(X_j)\} - \mathbb{E} I\{Y \neq \overline{f}(X) \\
&\leq \mathcal{E}(\overline{f}) + 2 \sup_{f \in \mathcal{F}} \left| \frac{1}{n}\sum I\{Y_j \neq f(X_j)\} - P(Y \neq f(X) \right|.
\end{aligned}$$

Our goal here is to study this random variable. P_n is an empirical measure based on $X_1, \dots X_n$ (here these are any observation – like (X_i, Y_i)). Define

$$P_n = \frac{1}{n} \sum \delta_{X_j}$$
$$P_n f = \int f \, dP_n = \frac{1}{n} \sum f(X_j).$$

If $X_1 \sim P$, then $Pf = \int f \, dP = \mathbb{E} f(X_1)$.

Definition 1. The empirical process indexed by the class $\mathcal{F}$ is

$$\mathbb{Z}_n(f) := \sqrt{n}(P_n f - Pf).$$

Our question: what is the "size" of

$$\sup_f |\mathbb{Z}_n(f)| = \sqrt{n} \sup_f |P_n f - Pf| =: \sqrt{n} ||P_n - P||_{\mathcal{F}}$$

One such class is

$$\mathcal{H} = \{I_f = I\{y \neq f(x)\}.$$