CS281B/Stat241B (Spring 2003) Statistical Learning Theory **Lecture: 21**

Bernstein's inequality, and generalizations

Lecturer: Peter L. Bartlett *Scribe: XuanLong Nguyen*

Let $F \subseteq [-1,1]^{\mathcal{X}}$. Recall that with probability $\geq 1-\delta$ we have

$$\sup_{f\in F} \mathbb{E}f - \hat{\mathbb{E}}f \leq 2R_n(F) + c\sqrt{\frac{\log(1/\delta)}{n}}$$

where $R_n(F)$ is the Rademacher average of F, and $\hat{\mathbb{E}}$ denotes the empirical expectation. Hence, if $\hat{f} \in F$ minimizes $\hat{\mathbb{E}}f$, we have

$$\mathbb{E}\hat{f} \leq \inf_{f\in F} \mathbb{E}f + 2R_n(F) + c\sqrt{\frac{\log(1/\delta)}{n}}.$$

Since $R_n(F)$ is generally in $O(1/\sqrt{n})$, the RHS of above inequality is in $O(1/\sqrt{n})$. Applying to classes of loss function $\{l_f | f \in F\}$, we obtained a generalization bound for our learning algorithms. In some cases, we can obtain error rate better than $O(1/\sqrt{n})$. For instance, for the class of perceptron studied earlier in the course, the error rate is $O(1/n)$. On the other hand, if $\mathrm{Var}(l_{\hat{f}}) \approx \sup_{f\in F} \mathrm{Var}(l_f)$, the bound is tight, as illustrated by the following theorem:

Theorem 0.1. For $F \subseteq [-1,1]^{\mathcal{X}}$, $\sigma_F^2 = \sup_{f\in F} \mathrm{Var}(f)$. If $\sigma_F^2 \geq 1/n$, then for some constant c

$$\mathbb{E}\sup_{f\in F} |\mathbb{E}f - \hat{\mathbb{E}}f| \geq \frac{c\sigma_F}{\sqrt{n}}.$$

To prove this result, we will need a concentration inequality that is stronger than Hoeffding's inequality, which is loose in case of small variance.

Theorem 0.2. (Beinstein's inequality) Suppose that $X_1, \ldots, X_n$ are independent random variables with mean 0. $|X_i| < c$ almost surely. Let $\sigma_i^2 = \mathrm{Var}(X_i)$ and $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$. Then,

$$P(\frac{1}{n}\sum_{i=1}^n X_i \geq \epsilon) \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2 + 2c\epsilon/3}\right).$$

Proof. The proof is similar to that of Chernoff's bound, but we need a slightly different bound for the moment that accounts for variance information. We have, as before, for any $s > 0$:

$$P(\sum X_i \geq t) \leq e^{-st}\prod_{i=1}^n \mathbb{E}e^{sX_i}.$$

Now,

$$\begin{aligned}
\mathbb{E}e^{sX_i} &= 1 + \mathbb{E}sX_i + \sum_{j=2}^{\infty} \frac{s^j \mathbb{E}X_i^j}{j!} \\
&\le 1 + \frac{\sigma_i^2}{c^2} \sum_{j=2}^{\infty} \frac{s^j c^j}{j!} \text{ since } \mathbb{E}X_i^j \le c^{j-2}\sigma_i^2 \\
&= 1 + \frac{\sigma_i^2}{c^2}(e^{sc} - 1 - sc) \\
&\le \exp\left(\frac{\sigma_i^2}{c^2}(e^{sc} - 1 - sc)\right)
\end{aligned}$$

Hence,

$$\begin{aligned}
P(\sum_{i=1}^{n} X_i \ge t) &\le \inf_{s>0} \exp(-st + \frac{n\sigma^2}{c^2}(e^{sc} - 1 - sc)) \\
&\le \exp \frac{-t^2}{2n\sigma^2 + 2ct/3}
\end{aligned}$$

by choosing $s = \frac{1}{c}\log(1 + \frac{tc}{n\sigma^2})$. □

Now we prove the first theorem using Bernstein's inequality.

Proof. Assume that $\sigma_F^2 = \sup_{f\in F} \mathrm{Var}(f)$ is attained at some $f \in F$ (otherwise, an extra limiting argument will suffice). Set $Y = \sum_{i=1}^{n}(f(X_i) - \mathbb{E}f)$, and $v = \mathbb{E}Y^2 = n\sigma_F^2 \ge 1$.

We will show that $|Y| \approx \sqrt{v}$ with high probability. Indeed, for any $k \ge 1$,

$$\mathbb{E}Y^2 1_{[|Y|\ge k\sqrt{v}]} = \sum_{m=k}^{\infty} \mathbb{E}Y^2 1_{[m\sqrt{v}\le|Y|\le(m+1)\sqrt{v}]} \le v \sum_{m=k}^{\infty} (m+1)^2 P[|Y| \ge m\sqrt{v}].$$

Note that by Bernstein's inequality,

$$\begin{aligned}
P[|Y| \ge m\sqrt{v}] &\le P\left[\sum_{i=1}^{n}(f(X_i) - \mathbb{E}f) \ge m\sqrt{v}\right] \\
&\le \exp\left(-\frac{m^2 v}{2(v + m\sqrt{v}/3)}\right) \\
&\le \exp\left(-\frac{m^2 v}{2(4mv/3)}\right) = e^{-3m/8}.
\end{aligned}$$

Hence, with some sufficiently large constant k, we have $\mathbb{E}Y^2 1_{[|Y|\ge k\sqrt{v}]} \le v/4$. So,

$$\begin{aligned}
v = \mathbb{E}Y^2 &= \mathbb{E}Y^2 1_{[|Y|\le\sqrt{v}/2]} + \mathbb{E}Y^2 1_{[\sqrt{v}/2\le|Y|\le k\sqrt{v}]} + v/4 \\
&\le v/2 + k^2 v P(|Y| \ge \sqrt{v}/2)
\end{aligned}$$

So, $P(|Y| \ge \sqrt{v}/2) \ge 1/2k^2$. So, $P(\sup_{f\in F} |\mathbb{E}f - \hat{\mathbb{E}}f| \ge \sqrt{v}/2n) \ge 1/2k^2$. By Markov's inequality, this implies that

$$\mathbb{E}\sup_{f\in F} |\mathbb{E}f - \hat{\mathbb{E}}f| \ge \frac{\sigma_F}{4k^2\sqrt{n}}.$$

□

The proof above uses Bernstein's inequality, which is stronger than Hoeffdling's inequality when the variance information is known. From now until the rest of the lecture, we will present a number of Bernstein style inequalities and their application. To draw the parallel, we first list the familiar Hoeffding's inequality, and its consequences, and then Bernstein's inequality, and its consequences.

Lemma 0.3. (Hoeffding's inequality) Let $f : \mathcal{X} \to [a, b]$, $Z = \sum_{i=1}^n f(X_i)$. Then,

$$P[Z - \mathbb{E}Z \geq t] \leq \exp\left(-\frac{2t^2}{n(b-a)^2}\right)$$

Lemma 0.4. (Bounded sequence inequality) if $|g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_n)| \leq c_k$ for all $x_1, \ldots, x_n$, then:

$$P[g(X_1, \ldots, X_n) - \mathbb{E}g(X_1, \ldots, X_n) \geq t] \leq \exp\frac{-2t^2}{\sum_{i=1}^n c_i^2}.$$

Consequence to empirical processes:

Lemma 0.5. Consider $f : \mathcal{X} \to [a, b]$, $Z = \sup_{f \in F} \sum_{i=1}^n (\mathbb{E}f - f(X_i))$, then:

$$P[Z - \mathbb{E}Z \geq t] \leq \exp -\frac{2t^2}{n(b-a)^2}.$$

Lemma 0.6. (Bernstein's inequality) Consider $f : \mathcal{X} \to [-c, c], Z = \sum_{i=1}^n f(X_i), \sigma_i^2 = \mathrm{Var} f(X_i), \sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$, then:

$$P[Z - \mathbb{E}Z \geq t] \leq \exp -\frac{t^2}{2n\sigma^2 + 2ct/3}.$$

The counter-part to bounded difference inequality is concentration inequality for functions with self-bounding property (due to Boucheron, Lugosi and Massart).

A function $g : \mathcal{X}^n \to \mathbb{R}$ has **self-bounding property** iff there exist $g_k : \mathcal{X}^{n-1} \to \mathbb{R}$ for $k = 1, \ldots, n$ such that for all $x_i \in \mathcal{X}$, $0 \leq g(x) - g_k(x^{(-k)}) \leq 1$, and $\sum_{k=1}^n (g(x) - g_k(x^{(-k)})) \leq g(x)$.

Lemma 0.7. (Inequality of self-bounding functions) If g is has self-bounding property, then $Z = g(X_1, \ldots, X_n)$ (with X_i are i.i.d) satisfies:

$$P[Z - \mathbb{E}Z \geq t] \leq \exp -\frac{t^2}{2\mathbb{E}Z + 2t/3}.$$

The following inequality for empirical processes is called Talagrand's inequality. This version is, however, due to Bousquet:

Lemma 0.8. Consider $f : \mathcal{X} \to [-1, 1]$. Let $Z = \sup_{f \in F} \sum f(X_i)$. Assume that $\mathbb{E}f = 0$, and $\sup_{f \in F} \mathrm{Var} f(X_i) \leq \sigma^2$, then:

$$P[Z - \mathbb{E}Z \geq t] \leq \exp\frac{-t^2}{2v + 2t/3},$$

with $v = n\sigma^2 + 2\mathbb{E}Z$.

If $\mathbb{E}Z \approx n\sigma^2, t = \alpha\mathbb{E}Z + \epsilon n$, $(0 < \alpha < 1)$, then an application of above inequality gives:

$$P[\frac{1}{n}Z \geq (1+\alpha)\frac{\mathbb{E}Z}{n} + \epsilon] \leq \exp\frac{-\alpha\epsilon n}{7}.$$

Before ending the lecture, we give an application of Boucheron, Lugosi and Massart's inequality for self-bounding functions. Consider $g(X_1, \ldots, X_n) = \sup_{f \in F} \sum_{i=1}^n f(X_i)$, where $f \in F$ has $f : \mathcal{X} \to [0,1]$, then g has self-bounding property.

Indeed, define $g_k(x^{(-k)}) = \sup_{f \in F} \sum_{i \neq k} f(x_i)$. Then we have:

$$\begin{aligned} g(x) - g_k(x^{(-k)}) &= \sup_{f \in F} \sum_{i=1}^n f(x_i) - \sup_{f \in F} \sum_{i \neq k} f(x_i) \\ & \quad \text{suppose the first sup is attained at } f^*\text{, the second at } f_k^* \\ &= \sup_{f \in F} \sum_{i=1}^n f(x_i) - \sum_{i \neq k} f_k^*(x_i) \\ &\geq f_k^*(x_i) \geq 0. \end{aligned}$$

Similarly, $g(x) - g_k(x^{(-k)}) = \sum_{i=1}^n f^*(x_i) - \sup_{f \in F} \sum_{i \neq k} f(x_k) \leq f^*(x_k) \leq 1$. Also,

$$\begin{aligned} (n-1)g(x) &= (n-1) \sum_{i=1}^n f^*(x_i) \\ &= \sum_{k=1}^n \sum_{i \neq k} f^*(x_i) \leq \sum_{k=1}^n \sup_{f \in F} \sum_{i \neq k} f(x_i) \\ &= \sum_{k=1}^n g_k(x^{(-k)}). \end{aligned}$$

Rearraining gives $\sum_{k=1}^n (g(x) - g_k(x^{(-k)})) \leq g(x)$. Therefore, we can apply Boucheron, Lugosi and Massart's inequalty: If $Z = \sup_{f \in F} \frac{1}{n} \sum f(X_i)$, then:

$$P[Z - \mathbb{E}Z \geq t] \leq \exp \frac{-n\epsilon^2}{2\mathbb{E}Z + 2\epsilon/3}.$$