CS 281B (Spring 2006) **Lecture: 28**

Better rates through "variance information"

Lecturer: Ambuj Tewari *Scribe: Ben Blum*

Our setting is classification, with input space $\mathcal{X}$; the function class of interest is $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$. Recall that with probability at least $1-\delta$,

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

In particular, if $\hat{f} \in \mathcal{F}$ minimizes $\hat{\mathbb{E}}f$, then with probability at least $1-\delta$,

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

so that the risk of the empirical risk minimizer converges at rate $\sqrt{n}$. However, if $\mathrm{Var}(\ell_{\hat{f}})$ is small, much smaller than $\sup_{f\in\mathcal{F}}\mathrm{Var}(\ell_f)$ (where here by ℓ_f we mean the function $(x,y)\mapsto \ell(f(x),y)$), then we might expect to achieve faster rates.

We first give some intuition for why this might be so. Fix a function f and the 0-1 loss function. The Hoeffding inequality gives us $\Pr\{\hat{L}(f) - L(f) \geq \epsilon\} \leq \exp(-2\epsilon^2 n)$. But the Central Limit Theorem tells us that the tails of $\hat{L}(f) - L(f)$ should behave like $\exp(-\frac{\epsilon^2 n}{2\sigma^2})$. CLT is not uniform over $\mathcal{F}$, so does not directly apply to the question we're interested in. But it does give us a hint that σ matters.

We'll give several concentration inequality theorems in this lecture without providing proofs; for proofs of these results and others like them, please see Boucheron, Bousquet, & Lugosi, 2004.

Theorem 1 (Bernstein's Inequality)**.** For independent $X_1, X_2, \ldots, X_n$ with $\mathbb{E}X_i = 0$, $|X_i| < c$, $\sigma_i^2 = \mathrm{Var}(X_i)$, $\sigma^2 = \frac{1}{n}\sum\limits_{i=1}^{n}\sigma_i^2$, we have

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

This inequality is for sums of random variables, like the Hoeffding inequality. We'd like to generalize to more general functions of random variables, in the manner of the Bounded Differences Inequality. To this end, we have the following theorem.

Theorem 2 (Boucheron, Lugosi, and Massart)**.** Suppose $g : \mathcal{X}^n \to \mathbb{R}$ has the following "self-bounding" property: there exist functions $g_k : \mathcal{X}^{n-1} \to \mathbb{R}$ for $k = 1, 2, \ldots, n$ such that for all $\mathbf{X} = (X_1, X_2, \ldots, X_n) \in \mathcal{X}^n$,

(i) $0 \leq g(\mathbf{X}) - g_k(\mathbf{X}^{(-k)}) \leq 1$, where $\mathbf{X}^{(-k)} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n)$.

(ii) $\sum\limits_{k=1}^{n}(g(\mathbf{X}) - g_k(\mathbf{X}^{(-k)})) \leq g(\mathbf{X})$.

Then $Z = g(X_1, X_2, \ldots, X_n)$ has $\Pr\{Z - \mathbb{E}Z \geq t\} \leq \exp\left(\frac{-t^2}{2\mathbb{E}Z + 2t/3}\right)$.

For the proof, see Boucheron et al., 2004. Note that condition (ii) acts like a variance constraint, because $0 \le g(\mathbf{X}) - g_k(\mathbf{X}^{(-k)}) \le 1$ so $\sum_k (g(\mathbf{X}) - g_k(\mathbf{X}^{(-k)}))^2 \le \sum_k (g(\mathbf{X}) - g_k(\mathbf{X}^{(-k)})) \le g(\mathbf{X})$.

Example 1. Set $g(X_1, X_2, \ldots, X_n) = \sup\limits_{f \in \mathcal{F}} \sum\limits_{i=1}^{n} f(X_i)$, with $f : \mathcal{X} \to [0,1]$. We will now see that g has the self-bounding property. Define $g_k = \sup\limits_{f \in \mathcal{F}} \sum\limits_{i \neq k} f(X_i)$. Then

(i)

$$\begin{aligned} g(\mathbf{X}) - g_k(\mathbf{X}^{(-k)}) &= \underbrace{\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i)}_{\text{achieved by } f^* \in \mathcal{F}} - \underbrace{\sup_{f \in \mathcal{F}} \sum_{i \neq k} f(X_i)}_{\text{achieved by } f_k^* \in \mathcal{F}} \\ &= \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i) - \sum_{i \neq k} f_k^*(X_i) \\ &\ge \sum_{i=1}^{n} f_k^*(X_i) - \sum_{i \neq k} f_k^*(X_i) \\ &= f_k^*(X_k) \\ &\ge 0, \end{aligned}$$

and similarly,

$$\begin{aligned} g(\mathbf{X}) - g_k(\mathbf{X}^{(-k)}) &= \sum_{i=1}^{n} f^*(X_i) - \sup_{f \in \mathcal{F}} \sum_{i \neq k} f(X_i) \\ &\le \sum_{i=1}^{n} f^*(X_i) - \sum_{i \neq k} f^*(X_i) \\ &= f^*(X_k) \\ &\le 1. \end{aligned}$$

We have therefore established condition (i). Now we establish condition (ii).

(ii)

$$\begin{aligned} (n-1)g(\mathbf{X}) &= (n-1) \sum_{i=1}^{n} f^*(X_i) \\ &= \sum_{k=1}^{n} \sum_{i \neq k} f^*(X_i) \\ &\le \sum_{k=1}^{n} (\sup_{f \in \mathcal{F}} \sum_{i \neq k} f(X_i)) \\ &= \sum_{k=1}^{n} g_k(\mathbf{X}^{(-k)}). \end{aligned}$$

Therefore, the BLM theorem gives us, with $t = \epsilon n$ and $nZ = \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i)$,

$$\Pr\{nZ - n\mathbb{E}Z \ge n\epsilon\} \le \exp\left(-\frac{\epsilon^2 n}{2\mathbb{E}Z + 2\epsilon/3} \right).$$

We can remove the factor of n in the left hand side to get a bound on $Z - \mathbb{E}Z$. Note that if ϵ is on the same order as $\mathbb{E}Z$, then one ϵ cancels and we get a deviation of order $1/n$.

Example 2. We'll now see another example of a "self-bounding" g. Now we take $\mathcal{F} \in [-1,1]^{\mathcal{X}}$. Define

$$g(X_1, X_2, \ldots, X_n) = \mathbb{E}[\sup_{f\in\mathcal{F}} \sum_i \epsilon_i f(X_i) | X_1, \ldots, X_n].$$

Note that this is simply $n\hat{R}_n(\mathcal{F})$, n times the empirical Rademacher complexity of $\mathcal{F}$. Next we will see that g satisfies the two self-bounding conditions, choosing, as one might expect,

$$g_k(\mathbf{X}^{(-k)}) = \mathbb{E}[\sup_{f\in\mathcal{F}} \sum_{i\neq k} \epsilon_i f(X_i) | X_1^n].$$

By X_1^n, we mean $X_1, X_2, \ldots, X_n$. We now establish the two conditions.

(i) We first make use of the fact that the ϵ_i's are independent of each other, so that $\mathbb{E}[\epsilon_k | \epsilon^{(-k)}] = 0$, to write

$$\begin{aligned}
g_k(\mathbf{X}^{(-k)}) &= \mathbb{E}[\sup_{f\in\mathcal{F}} \mathbb{E}[\sum_{i=1}^n \epsilon_i f(X_i) | \epsilon^{(-k)}] | X_1^n] \\
&\leq \mathbb{E}[\mathbb{E}[\sup_{f\in\mathcal{F}} \sum_{i=1}^n \epsilon_i f(X_i) | \epsilon^{(-k)}] | X_1^n] \\
&= \mathbb{E}[\sup_{f\in\mathcal{F}} \sum_{i=1}^n \epsilon_i f(X_i) | X_1^n] \\
&= g(\mathbf{X}),
\end{aligned}$$

so $g(\mathbf{X}) - g_k(\mathbf{X}^{(-k)}) \geq 0$. Also, $g(\mathbf{X}) - g_k(X^{(-k)}) \leq \sup_{f\in\mathcal{F}} |f(X_k)| \leq 1$.

(ii) We establish the second condition in the same style as in Example 1.

$$\begin{aligned}
(n-1)g(\mathbf{X}) &= \mathbb{E}[\sup_{f\in\mathcal{F}} \sum_{k=1}^n \sum_{i\neq k} \epsilon_i f(X_i) | X_1^n] \\
&\leq \mathbb{E}[\sum_{k=1}^n \sup_{f\in\mathcal{F}} \sum_{i\neq k} \epsilon_i f(X_i) | X_1^n] \\
&= \sum_{k=1}^n \mathbb{E}[\sup_{f\in\mathcal{F}} \sum_{i\neq k} \epsilon_i f(X_i) | X_1^n] \\
&= \sum_{k=1}^n g_k(\mathbf{X}^{(-k)}).
\end{aligned}$$

Then the BLM theorem tells us that

$$\Pr\{\hat{R}_n(\mathcal{F}) \geq R_n(\mathcal{F}) + t\} \leq \exp\left(-\frac{t^2 n}{2R_n(\mathcal{F}) + 2t/3}\right).$$

The variance doesn't appear explicitly, but it has an influence in the form of the $R_n(\mathcal{F})$ in the denominator of the exponent. Note that this bound isn't useful in its direct form, since we don't know the value of $R_n(\mathcal{F})$ in order to insert it on the right. We have to do a little bit of further algebraic work if we want to apply the bound. We won't go over that here.

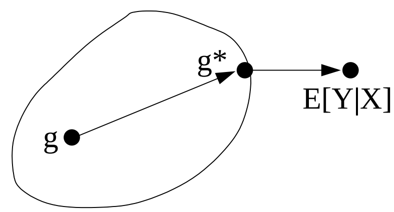

Figure 1: Dot-product space of prediction functions

Another route to variance-dependent rates is the following theorem.

Theorem 3 (Talagrand's functional Bernstein's Inequality)**.** For $f : \mathcal{X} \to [-b, b]$, $\mathbb{E}f = 0$ for all $f \in \mathcal{F}$, $Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(X_i)$, $\sigma^2 = n \sup_{f\in\mathcal{F}} \mathrm{Var}(f(X_i))$, $X_1, X_2, \ldots, X_n$ i.i.d,

$$\Pr\{Z - \mathbb{E}Z \geq t\} \leq \exp\left(-\frac{t^2}{2(\sigma^2 + b\mathbb{E}Z + bt/3)}\right).$$

We won't prove this here. In order apply the result, or similar ones, we must bound the variance of $f \in \mathcal{F}$ somehow. Earlier, we saw that if we have $f : \mathcal{X} \to [0, 1]$, then $\mathrm{Var}(f) \leq \mathbb{E}f^2 \leq \mathbb{E}f$. It turns out that we can similarly bound the variance by the mean for the excess loss class of a convex function class.

Example 3. Let $\mathcal{G} \subseteq [-1, 1]^{\mathcal{X}}$ be convex. The excess loss class is $\mathcal{F} = \{\ell_g - \ell_{g^*} : g \in \mathcal{G}\}$, where $\ell_g(X, Y) = \ell(g(X), Y)$ and $\ell(\hat{y}, y) = (\hat{y} - y)^2$, for $y \in [-1, 1]$. Here $g^* = \mathrm{argmin}_{g\in\mathcal{G}} \mathbb{E}\ell_g$. For any $g \in \mathcal{G}$, we will see that $\mathbb{E}(\ell_g - \ell_{g^*})^2 \leq 16\mathbb{E}(g - g^*)^2 \leq 16\mathbb{E}(\ell_g - \ell_{g^*})$.

We begin by establishing the first inequality. We have $f = \ell_g - \ell_{g^*}$ satisfies

$$\begin{aligned}
\mathbb{E}f^2 &= \mathbb{E}[((Y - g(X))^2 - (Y - g^*(X))^2]^2 \\
&= \mathbb{E}[(2Y - g(X) - g^*(X))(g(X) - g^*(X))]^2 \\
&\leq 4^2\mathbb{E}[g(X) - g^*(X)]^2.
\end{aligned}$$

In the last step we make use of the fact that $0 \leq 2Y - g(X) - g^*(X) \leq 2$.

We now establish the second inequality, $\mathbb{E}(g - g^*)^2 \leq \mathbb{E}[\ell_g - \ell_{g^*}]$.

$$\begin{aligned}
\mathbb{E}(\ell_g - \ell_{g^*}) &= \mathbb{E}[(Y - g(X))^2 - (Y - g^*(X))^2] \\
&= \mathbb{E}[(g^*(X) - g(X))(2Y - g(X) - g^*(X)] \\
&= \mathbb{E}[(g^*(X) - g(X))^2 + 2(g^*(X) - g(X))(Y - g^*(X))] \\
&= \mathbb{E}(g^*(X) - g(X))^2 + 2\mathbb{E}_X[(g^*(X) - g(X))(\mathbb{E}[Y|X] - g^*(X))].
\end{aligned}$$

In suffices now to show that the term $\mathbb{E}_X[(g^*(X) - g(X))(\mathbb{E}[Y|X] - g^*(X))]$ is at least zero. We will only give an intuitive proof by picture here. Consider the usual dot product $f \cdot g \triangleq \mathbb{E}fg$. $\mathbb{E}[Y|X]$ is the global best estimate of Y under squared error, so g^* is closer to $\mathbb{E}[Y|X]$ in this space than g. Then by the convexity of $\mathcal{G}$, $(g^*(X) - g(X))$ and $(\mathbb{E}[Y|X] - g^*(X))$ "point" in the same direction, so their dot product, $\mathbb{E}_X[(g^*(X) - g(X))(\mathbb{E}[Y|X] - g^*(X))]$, is positive. The picture is as in Figure 1. We conclude that $\mathbb{E}(\ell_g - \ell_g^*) \geq \mathbb{E}(g^* - g)^2$ and hence that $\mathbb{E}(\ell_g - \ell_{g^*})^2 \leq 16\mathbb{E}(\ell_g - \ell_{g^*})$.

References

Boucheron, S., Bousquet, O., & Lugosi, G. (2004). Concentration inequalities. In *Advanced Lectures in Machine Learning*, pp. 208–240. Springer.