18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet Lecture 3
Scribe: James Hirst Sep. 16, 2015

1.5 Learning with a finite dictionary

Recall from the end of last lecture our setup: We are working with a finite dictionary $\mathcal{H} = \{h_1, \dots, h_M\}$ of estimators, and we would like to understand the scaling of this problem with respect to M and the sample size n. Given $\mathcal{H}$, one idea is to simply try to minimize the empirical risk based on the samples, and so we define the empirical risk minimizer, $\hat{h}^{\mathrm{erm}}$, by

$$\hat{h}^{\mathrm{erm}} \in \operatorname*{argmin}_{h \in \mathcal{H}} \hat{R}_n(h).$$

In what follows, we will simply write $\hat{h}$ instead of $\hat{h}^{\mathrm{erm}}$ when possible. Also recall the definition of the oracle, $\bar{h}$, which (somehow) minimizes the true risk and is defined by

$$\bar{h} \in \operatorname*{argmin}_{h \in \mathcal{H}} R(h).$$

The following theorem shows that, although $\hat{h}$ cannot hope to do better than $\bar{h}$ in general, the difference should not be too large as long as the sample size is not too small compared to M.

Theorem: The estimator $\hat{h}$ satisfies

$$R(\hat{h}) \leq R(\bar{h}) + \sqrt{\frac{2\log(2M/\delta)}{n}}$$

with probability at least $1 - \delta$. In expectation, it holds that

$$\mathbb{E}[R(\hat{h})] \leq R(\bar{h}) + \sqrt{\frac{2\log(2M)}{n}}.$$

Proof. From the definition of $\hat{h}$, we have $\hat{R}_n(\hat{h}) \leq \hat{R}_n(\bar{h})$, which gives

$$R(\hat{h}) \leq R(\bar{h}) + [\hat{R}_n(\bar{h}) - R(\bar{h})] + [R(\hat{h}) - \hat{R}_n(\hat{h})].$$

The only term here that we need to control is the second one, but since we don't have any real information about $\bar{h}$, we will bound it by a maximum over $\mathcal{H}$ and then apply Hoeffding:

$$[\hat{R}_n(\bar{h}) - R(\bar{h})] + [R(\hat{h}) - \hat{R}_n(\hat{h})] \leq 2 \max_j |\hat{R}_n(h_j) - R(h_j)| \leq 2\sqrt{\frac{\log(2M/\delta)}{2n}}$$

with probability at least $1 - \delta$, which completes the first part of the proof.

To obtain the bound in expectation, we start with a standard trick from probability which bounds a max by its sum in a slightly more clever way. Here, let $\{Z_j\}_j$ be centered random variables, then

$$\mathbb{E}\left[\max_j |Z_j|\right] = \frac{1}{s}\log\exp\left(s\mathbb{E}\left[\max_j |Z_j|\right]\right) \leq \frac{1}{s}\log\mathbb{E}\left[\exp\left(s\max_j |Z_j|\right)\right],$$

where the last inequality comes from applying Jensen's inequality to the convex function $\exp(\cdot)$. Now we bound the max by a sum to get

$$\leq \frac{1}{s}\log\sum_{j=1}^{2M}\mathbb{E}\left[\exp(sZ_j)\right] \leq \frac{1}{s}\log\left(2M\exp\left(\frac{s^2}{8n}\right)\right) = \frac{\log(2M)}{s} + \frac{s}{8n},$$

where we used $Z_j = \hat{R}_n(h_j) - R(h_j)$ in our case and then applied Hoeffding's Lemma. Balancing terms by minimizing over s, this gives $s = 2\sqrt{2n\log(2M)}$ and plugging in produces

$$\mathbb{E}\left[\max_j |\hat{R}_n(h_j) - R(h_j)|\right] \leq \sqrt{\frac{\log(2M)}{2n}},$$

which finishes the proof. □

2. CONCENTRATION INEQUALITIES

Concentration inequalities are results that allow us to bound the deviations of a function of random variables from its average. The first of these we will consider is a direct improvement to Hoeffding's Inequality that allows some dependence between the random variables.

2.1 Azuma-Hoeffding Inequality

Given a filtration $\{\mathcal{F}_i\}_i$ of our underlying space $\mathcal{X}$, recall that $\{\Delta_i\}_i$ are called *martingale differences* if, for every i, it holds that $\Delta_i \in \mathcal{F}_i$ and $\mathbb{E}\left[\Delta_i|\mathcal{F}_i\right] = 0$. The following theorem gives a very useful concentration bound for averages of bounded martingale differences.

Theorem (Azuma-Hoeffding): Suppose that $\{\Delta_i\}_i$ are margingale differences with respect to the filtration $\{\mathcal{F}_i\}_i$, and let $A_i, B_i \in \mathcal{F}_{i-1}$ satisfy $A_i \leq \Delta_i \leq B_i$ almost surely for every i. Then

$$\mathbb{P}\left[\frac{1}{n}\sum_i \Delta_i > t\right] \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n \|B_i - A_i\|_\infty^2}\right).$$

In comparison to Hoeffding's inequality, Azuma-Hoeffding affords not only the use of non-uniform boundedness, but additionally requires no independence of the random variables.

Proof. We start with a typical Chernoff bound.

$$\mathbb{P}\left[\sum_i \Delta_i > t\right] \leq \mathbb{E}\left[e^{s\sum\Delta_i}\right]e^{-st} = \mathbb{E}\left[\mathbb{E}\left[e^{s\sum\Delta_i}|\mathcal{F}_{n-1}\right]\right]e^{-st}$$

$$= \mathbb{E}\left[e^{s\sum^{n-1}\Delta_i}\mathbb{E}[e^{s\Delta_n}|\mathcal{F}_{n-1}]\right]e^{-st} \le \mathbb{E}[e^{s\sum^{n-1}\Delta_i}\cdot e^{s^2(B_n-A_n)^2/8}]e^{-st},$$

where we have used the fact that the Δ_i, $i < n$, are all $\mathcal{F}_n$ measureable, and then applied Hoeffding's lemma on the inner expectation. Iteratively isolating each Δ_i like this and applying Hoeffding's lemma, we get

$$\mathbb{P}\left[\sum_i \Delta_i > t\right] \le \exp\left(\frac{s^2}{8}\sum_{i=1}^n \|B_i - A_i\|_\infty^2\right)e^{-st}.$$

Optimizing over s as usual then gives the result. □

2.2 Bounded Differences Inequality

Although Azuma-Hoeffding is a powerful result, its full generality is often wasted and can be cumbersome to apply to a given problem. Fortunately, there is a natural choice of the $\{\mathcal{F}_i\}_i$ and $\{\Delta_i\}_i$, giving a similarly strong result which can be much easier to apply. Before we get to this, we need one definition.

Definition (Bounded Differences Condition): Let $g : \mathcal{X} \to \mathbb{R}$ and constants c_i be given. Then g is said to satisfy the bounded differences condition (with constants c_i) if

$$\sup_{x_1,\ldots,x_n,x_i'} |g(x_1,\ldots,x_n) - g(x_1,\ldots,x_i',\ldots,x_n)| \le c_i$$

for every i.

Intuitively, g satisfies the bounded differences condition if changing only one coordinate of g at a time cannot make the value of g deviate too far. It should not be too surprising that these types of functions thus concentrate somewhat strongly around their average, and this intuition is made precise by the following theorem.

Theorem (Bounded Differences Inequality): If $g : \mathcal{X} \to \mathbb{R}$ satisfies the bounded differences condition, then

$$\mathbb{P}\left[|g(X_1,\ldots,X_n) - \mathbb{E}[g(X_1,\ldots,X_n)| > t\right] \le 2\exp\left(-\frac{2t^2}{\sum_i c_i^2}\right).$$

Proof. Let $\{\mathcal{F}_i\}_i$ be given by $\mathcal{F}_i = \sigma(X_1,\ldots,X_i)$, and define the martingale differences $\{\Delta_i\}_i$ by

$$\Delta_i = \mathbb{E}\left[g(X_1,\ldots,X_n)|\mathcal{F}_i\right] - \mathbb{E}\left[g(X_1,\ldots,X_n)|\mathcal{F}_{i-1}\right].$$

Then

$$\mathbb{P}\left[|\sum_i \Delta_i| > t\right] = \mathbb{P}\left[\left|g(X_1,\ldots,X_n) - \mathbb{E}[g(X_1,\ldots,X_n)\right| > t\right],$$

exactly the quantity we want to bound. Now, note that

$$\Delta_i \le \mathbb{E}\left[\sup_{x_i} g(X_1,\ldots,x_i,\ldots,X_n)|\mathcal{F}_i\right] - \mathbb{E}\left[g(X_1,\ldots,X_n)|\mathcal{F}_{i-1}\right]$$

$$= \mathbb{E}\left[\sup_{x_i} g(X_1,\ldots,x_i,\ldots,X_n) - g(X_1,\ldots,X_n)|\mathcal{F}_{i-1}\right] =: B_i.$$

Similarly,

$$\Delta_i \geq \mathbb{E}\left[\inf_{x_i} g(X_1,\ldots,x_i,\ldots,X_n) - g(X_1,\ldots,X_n)|\mathcal{F}_{i-1}\right] =: A_i.$$

At this point, our assumption on g implies that $\|B_i - A_i\|_\infty \leq c_i$ for every i, and since $A_i \leq \Delta_i \leq B_i$ with $A_i, B_i \in \mathcal{F}_{i-1}$, an application of Azuma-Hoeffding gives the result.

□

2.3 Bernstein's Inequality

Hoeffding's inequality is certainly a powerful concentration inequality for how little it assumes about the random variables. However, one of the major limitations of Hoeffding is just this: Since it only assumes boundedness of the random variables, it is completely oblivious to their actual variances. When the random variables in question have some known variance, an ideal concentration inequality should capture the idea that variance controls concentration to some degree. Bernstein's inequality does exactly this.

Theorem (Bernstein's Inequality): Let $X_1,\ldots,X_n$ be independent, centered random variables with $|X_i| \leq c$ for every i, and write $\sigma^2 = n^{-1}\sum_i \mathrm{Var}(X_i)$ for the average variance. Then

$$\mathbb{P}\left[\frac{1}{n}\sum_i X_i > t\right] \leq \exp\left(-\frac{nt^2}{2\sigma^2 + \frac{2}{3}tc}\right).$$

Here, one should think of t as being fixed and relatively small compared to n, so that strength of the inequality indeed depends mostly on n and $1/\sigma^2$.

Proof. The idea of the proof is to do a Chernoff bound as usual, but to first use our assumptions on the variance to obtain a slightly better bound on the moment generating functions. To this end, we expand

$$\mathbb{E}[e^{sX_i}] = 1 + \mathbb{E}[sX_i] + \mathbb{E}\left[\sum_{k=2}^\infty \frac{(sX_i)^k}{k!}\right] \leq 1 + \mathrm{Var}(X_i)\sum_{k=2}^\infty \frac{s^k c^{k-2}}{k!},$$

where we have used $\mathbb{E}[X_i^k] \leq \mathbb{E}[X_i^2|X_i|^{k-2}] \leq \mathrm{Var}(X_i)c^{k-2}$. Rewriting the sum as an exponential, we get

$$\mathbb{E}[e^{sX_i}] \leq s^2\mathrm{Var}(X_i)g(s), \quad g(s) := \frac{e^{sc} - sc - 1}{c^2s^2}.$$

The Chernoff bound now gives

$$\mathbb{P}\left[\frac{1}{n}\sum_i X_i > t\right] \leq \exp\left(\inf_{s>0}[s^2(\sum_i \mathrm{Var}(X_i))g(s) - nst]\right) = \exp\left(n \cdot \inf_{s>0}[s^2\sigma^2 g(s) - st]\right),$$

and optimizing this over s (a fun calculus exercise) gives exactly the desired result. □

3. NOISE CONDITIONS AND FAST RATES

To measure the effectiveness of the estimator $\hat{h}$, we would like to obtain an upper bound on the excess risk $\mathcal{E}(\hat{h}) = R(\hat{h}) - R(h^*)$. It should be clear, however, that this must depend significantly on the amount of noise that we allow. In particular, if $\eta(X)$ is identically equal to $1/2$, then we should not expect to be able to say anything meaningful about $\mathcal{E}(\hat{h})$ in general. Understanding this trade-off between noise and rates will be the main subject of this chapter.

3.1 The Noiseless Case

A natural (albeit somewhat naïve) case to examine is the completely noiseless case. Here, we will have $\eta(X) \in \{0,1\}$ everywhere, $\mathrm{Var}(Y|X) = 0$, and

$$\mathcal{E}(h) = R(h) - R(h^*) = \mathbb{E}[|2\eta(X) - 1| \mathbb{1}(h(X) \neq h^*(X))] = \mathbb{P}[h(X) \neq h^*(X)].$$

Let us now denote

$$Z_i = \mathbb{1}(\bar{h}(X_i) \neq Y_i) - \mathbb{1}(\hat{h}(X_i) \neq Y_i),$$

and write $\bar{Z}_i = Z_i - \mathbb{E}[Z_i]$. Then notice that we have

$$|Z_i| = \mathbb{1}(\hat{h}(X_i) \neq \bar{h}(X_i)),$$

and

$$\mathrm{Var}(Z_i) \leq \mathbb{E}[Z_i^2] = \mathbb{P}[\hat{h}(X_i) \neq \bar{h}(X_i)].$$

For any classifier $h_j \in \mathcal{H}$, we can similarly define $Z_i(h_j)$ (by replacing $\hat{h}$ with h_j throughout). Then, to set up an application of Bernstein's inequality, we can compute

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(Z_i(h_j)) \leq \mathbb{P}[h_j(X_i) \neq \bar{h}(X_i)] =: \sigma_j^2.$$

At this point, we will make a (fairly strong) assumption about our dictionary $\mathcal{H}$, which is that $h^* \in \mathcal{H}$, which further implies that $\bar{h} = h^*$. Since the random variables Z_i compare to $\bar{h}$, this will allow us to use them to bound $\mathcal{E}(\hat{h})$, which rather compares to h^*. Now, applying Bernstein (with $c = 2$) to the $\{\bar{Z}_i(h_j)\}_i$ for every j gives

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \bar{Z}_i(h_j) > t\right] \leq \exp\left(-\frac{nt^2}{2\sigma_j^2 + \frac{4}{3}t}\right) =: \frac{\delta}{M},$$

and a simple computation here shows that it is enough to take

$$t \geq \max\left(\sqrt{\frac{2\sigma_j^2 \log(M/\delta)}{n}}, \frac{4}{3n}\log(M/\delta)\right) =: t_0(j)$$

for this to hold. From here, we may use the assumption $\bar{h} = h^*$ to conclude that

$$\mathbb{P}\left[\mathcal{E}(\hat{h}) > t_0(\hat{j})\right] \leq \delta, \quad h_{\hat{j}} = \hat{h}.$$

However, we also know that $\sigma_{\hat{j}}^2 \leq \mathcal{E}(\hat{h})$, which implies that

$$\mathcal{E}(\hat{h}) \leq \max\left(\sqrt{\frac{2\mathcal{E}(\hat{h})\log(M/\delta)}{n}}, \frac{4}{3n}\log(M/\delta)\right)$$

with probability $1-\delta$, and solving for $\mathcal{E}(\hat{h})$ gives the improved rate

$$\mathcal{E}(\hat{h}) \leq 2\frac{\log(M/\delta)}{n}\,.$$