ECE901 Spring 2004 Statistical Regularization and Learning Theory **Lecture: 12**

Complexity Regularization in Regression

Lecturer: Rob Nowak *Scribe: David Winters*

1 Complexity Regularization in Regression

Recall the classification problem. In Lecture 6, where we assumed that $\min_{f\in\mathcal{F}} R(f) = 0$, we obtained the PAC bound $\forall f \in \mathcal{F}$

$$\mathcal{P}\{R(\widehat{f}_n) > \epsilon\} \leq |\mathcal{F}|e^{-n\epsilon}.$$

From Corrolary 1 in Lecture 6,

$$E[R(\widehat{f}_n)] \leq \frac{1+\log|\mathcal{F}|}{n}.$$

In Lectures 7 and 8, we dropped the assumption that $\min_{f\in\mathcal{F}} R(f) = 0$ and obtained, $\forall f \in \mathcal{F}$

$$\mathcal{P}\{R(\widehat{f}_n) > \epsilon\} \leq |\mathcal{F}|e^{-2n\epsilon^2}.$$

This led to

$$E[R(\widehat{f}_n) - \min_{f\in\mathcal{F}} R(f)] \leq \sqrt{\frac{\log|\mathcal{F}| + \log n + 2}{n}}.$$

Hoeffding's inequality was central to our analysis of learning under bounded loss functions. In many regression and signal estimation problems it is natural to consider squared error loss functions (rather than 0/1 or absolute error). In such cases, we will need to derive bounds using different techniques.

Example 1 *To illustrate the distinction between classification and regression, consider a simple, scalar signal plus noise problem. Consider $Y_i = \theta + W_i$, $i = 1, \ldots, n$, where θ is a fixed unknown scalar parameter and the W_i are independent, zero-mean, unit variance random variables. Let $\bar{Y} = 1/n\sum_{i=1}^n Y_i$. Then, according to the Central Limit Theorem, $\bar{Y}$ is distributed approximately $N(\theta, 1/n)$. A simple tail-bound on the Gaussian distribution gives us*

$$P(\bar{Y} - \theta > \epsilon) = P(W > \epsilon) \ \leq \ \frac{1}{2}e^{-n\epsilon^2/2},$$

which implies that

$$P(|\bar{Y} - \theta|^2 > \epsilon) \ \leq \ e^{-n\epsilon/2}$$

This is a bound on the deviations of the squared error $err^2 = |\bar{Y} - \theta|^2$. Notice that the exponential decay rate is a function of ϵ rather than ϵ^2, as in Hoeffding's inequality. The squared error concentration inequality implies that $E[|\bar{Y} - \theta|^2] = O(\frac{1}{n})$ (just write $E[err^2] = \int_0^\infty P(err^2 > t)dt$). Therefore, in regression with a squared error loss, we can hope to get a rate of convergence as fast as n^{-1} instead of $n^{-1/2}$. The reason is simply because we are using an squared error loss instead of the 0/1 or absolute error loss.

To begin our investigation into regression and function estimation, let us consider the following. Let $\mathcal{X} = \mathbf{R}^d$ and $\mathcal{Y} = \mathbf{R}$. Take $\mathcal{F}$ such that $f \in \mathcal{F}$ is a map $f : \mathbf{R}^d \mapsto \mathbf{R}$. We have training data $\{X_i, Y_i\}_{i=1}^n \overset{i.i.d.}{\sim} P_{XY}$. As our loss function, we take the squared error, *i.e.,*

$$l(f(X_i), Y_i) = (f(X_i) - Y_i)^2.$$

The risk is then the MSE:

$$R(f) = E[(f(X) - Y)^2].$$

We know that the function f^* that minimizes the MSE is just the conditional expectation of Y given X:

$$f^* = E[Y|X = x].$$

Now let $R^* = R(f^*)$. We would like to select an $\widehat{f}_n \in \mathcal{F}$ using the training data $\{X_i, Y_i\}_{i=1}^n$ such that the *excess risk*

$$E[R(\widehat{f}_n)] - R^* \geq 0$$

is small. Let's consider the difference between the empirical risks:

$$\widehat{R}(f) - \widehat{R}(f^*) = \frac{1}{n}\sum_{i=1}^n (f(X_i) - Y_i)^2 - \frac{1}{n}\sum_{i=1}^n (f^*(X_i) - Y_i)^2.$$

Note that $E[\widehat{R}(f) - \widehat{R}(f^*)] = R(f) - R(f^*)$. Hence, by the SLLN, we know that

$$\widehat{R}(f) - \widehat{R}(f^*) \to R(f) - R(f^*)$$

as $n \to \infty$. But how fast is this convergence?

We will derive a PAC style bound for the difference $\widehat{R}(f) - \widehat{R}(f^*) - (R(f) - R(f^*))$. The following derivation is from Barron 1991. The excess risk and it empirical counterpart will be denoted by

$$r(f, f^*) = R(f) - R(f^*)$$
$$\widehat{r}(f, f^*) = \widehat{R}(f) - \widehat{R}(f^*)$$

Note that $\widehat{r}(f, f^*)$ is the sum of independent random variables:

$$\widehat{r}(f, f^*) = -\frac{1}{n}\sum_{i=1}^n U_i,$$

where $U_i = -(Y_i - f(X_i))^2 + (Y_i - f^*(X_i))^2$. Therefore, $r(f, f^*) - \widehat{r}(f, f^*) = \frac{1}{n}\sum_{i=1}^n (U_i - E[U_i])$.

We are looking for a PAC bound of the form

$$\mathcal{P}(r(f, f^*) - \widehat{r}(f, f^*) > \epsilon) < \delta.$$

If the variables U_i are bounded, then we can apply Hoeffding's inequality. However, a more useful bound for our regression problem can be derived if the the variables U_i satisfy the following moment condition:

$$E[|U_i - E[U_i]|^k] \leq \frac{var(U_i)}{2}\ k!\ h^{k-2} \tag{1}$$

for some $h > 0$.

The moment condition can be difficult to verify in general, but it does hold, for example, for bounded random variables. If (1) holds, then the Craig-Bernstein (CB) inequality (Craig 1933) states:

$$\mathcal{P}\left(\frac{1}{n}\sum_{i=1}^n (U_i - E[U_i]) \geq \frac{t}{n\epsilon} + \frac{n\epsilon\ var(\frac{1}{n}\sum U_i)}{2(1-c)}\right) \leq e^{-t},$$

for $0 < \epsilon h \leq c < 1$ and $t > 0$. This shows that the tail decays exponentially in t, rather than exponentially in t^2. Recall Hoeffding's inequality:

$$\mathcal{P}\left(\frac{1}{n}\sum_{i=1}^n (Z_i - E[Z_i]) \geq \frac{t}{n}\right) \leq e^{\frac{-2t^2}{n}}.$$

If $\frac{t}{n} \ll 1$, then $\frac{t^2}{n} \ll t$, which implies $e^{\frac{-2t^2}{n}} \gg e^{-t}$. This indicates that the CB inequality may be much tighter than Hoeffding's. To use the CB inequality, we need to bound the variance of $\frac{1}{n}\sum_{i=1}^n U_i$. Note that

$$var(U_i) = var(-(Y_i - f(X_i))^2 + (Y_i - f^*(X_i))^2).$$

Assumption 1 *The support of Y and the range f(X) is in a known interval of length b.*

Proposition 1 *With the above assumption, (1) holds with* $h = \frac{2b^2}{3}$.

Proposition 2 *Again, with the above assumption, it may be shown that*

$$var(U_i) \leq 5b^2 r(f, f^*) \tag{2}$$

Proof 1 *You can write* U_i *as*

$$\begin{aligned} U_i &= 2Y_i f(X_i) - 2Y_i f^*(X_i) + f^*(X_i)^2 - f(X_i)^2 \\ &= 2Y_i f(X_i) - 2Y_i f^*(X_i) + 2f^*(X_i)^2 - f^*(X_i)^2 - f(X_i)^2 + 2f(X_i)f^*(X_i) - 2f(X_i)f^*(X_i) \\ &= 2\left(Y_i - f^*(X_i)\right)\left(f(X_i) - f^*(X_i)\right) - \left(f(X_i) - f^*(X_i)\right)^2 \end{aligned}$$

Note that the variance of U_i *is upper-bounded by its second moment. Also note that the covariance of the two terms above is zero:*

$$\begin{aligned} E[2\left(Y_i - f^*(X_i)\right)\left(f(X_i) - f^*(X_i)\right)\left(f(X_i) - f^*(X_i)\right)^2] &= E[T_1 T_2] \\ &= E_X[E_{Y|X}[T_1 T_2]] \\ &= E_X[T_2 E_{Y|X}[T_1]] \\ &= E_X[T_2 * 0] \\ &= 0 \end{aligned}$$

This is evident when you recall that $f^*(X_i) = E[Y|X = X_i]$. *Now we can bound the second moments of* T_1 *and* T_2 :

$$\begin{aligned} E[T_1] &= 4E[((Y_i - f^*(X_i))(f(X_i) - f^*(X_i)))^2] \\ &= 4E[(Y_i - f^*(X_i))^2 (f(X_i) - f^*(X_i))^2] \\ &\leq 4E[b^2 (f(X_i) - f^*(X_i))^2] \\ E[T_2] &= E[(f(X_i) - f^*(X_i))^4] \\ &= E[\left(f(X_i) - f^*(X_i)\right)^2 \left(f(X_i) - f^*(X_i)\right)^2] \\ &\leq E[b^2 \left(f(X_i) - f^*(X_i)\right)^2] \end{aligned}$$

So $var(U_i) \leq 5b^2 E[(f(X_i) - f^*(X_i))^2]$. *The final step is to see that*

$$r(f, f^*) = E[U_i] = E_X[E_{Y|X}[U_i]] = E[(f(X_i) - f^*(X_i))^2].$$

■

Thus, $n\ var(\frac{1}{n}\sum_{i=1}^n U_i) \leq 5b^2 r(f, f^*)$. And therefore, we can say that, with probability at least $1 - e^{-t}$,

$$r(f, f^*) - \widehat{r}(f, f^*) \leq \frac{t}{n\ \epsilon} + \frac{5\epsilon\ b^2\ r(f, f^*)}{2(1-c)}.$$

In other words, with probability at least $1-\delta$ (where $\delta = e^{-t}$),

$$r(f, f^*) - \widehat{r}(f, f^*) \leq \frac{\log \frac{1}{\delta}}{n\ \epsilon} + \frac{5\epsilon\ b^2\ r(f, f^*)}{2(1-c)}. \tag{3}$$

Now, suppose we have assigned positive numbers $c(f)$ to each $f \in \mathcal{F}$ satisfying the Kraft inequality:

$$\sum_{f \in \mathcal{F}} 2^{-c(f)} \leq 1.$$

Note that (3) holds $\forall \delta > 0$. In particular, we let δ be a function of f:

$$\delta(f) = 2^{-c(f)}\delta.$$

So we can use this δ along with the procedure introduced in Lecture 9 (*i.e.,* Union of events bound followed by the Kraft inequality) to obtain the following. For all $f \in \mathcal{F}$, $\forall \delta > 0$,

$$r(f, f^*) - \widehat{r}(f, f^*) \leq \frac{c(f)\log 2 + \log \frac{1}{\delta}}{n\ \epsilon} + \frac{5\epsilon\ b^2\ r(f, f^*)}{2(1-c)} \tag{4}$$

with probability at least $1-\delta$. Now set $c = \epsilon\ h = \frac{2b^2\ \epsilon}{3}$ and assume $\epsilon < \frac{6}{19b^2}$. Then define

$$\alpha = \frac{5\epsilon\ b^2}{2(1-c)} < 1.$$

Now, after using α and rearranging terms, we have:

$$(1-\alpha) r(f, f^*) \leq \widehat{r}(f, f^*) + \frac{c(f)\log 2 + \log \frac{1}{\delta}}{\epsilon\ n}.$$

We want to choose f to minmize this upper bound. So take

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \left\{ \widehat{R}_n(f) + \frac{c(f)\log 2}{n\epsilon} \right\}.$$

So, with probability at least $1-\delta$,

$$\begin{aligned} (1-\alpha) r(\widehat{f}_n, f^*) &\leq \widehat{r}(\widehat{f}_n, f^*) + \frac{c(\widehat{f}_n)\log 2 + \log \frac{1}{\delta}}{\epsilon\ n} \\ &\leq \widehat{r}(f_n^*, f^*) + \frac{c(f_n^*)\log 2 + \log \frac{1}{\delta}}{\epsilon\ n} \end{aligned} \tag{5}$$

where $f_n^* = \arg\min_{f \in \mathcal{F}} \left\{ R(f) + \frac{c(f)\log 2}{n\epsilon} \right\}$.

Now we use the Craig-Bernstein inequality to bound the difference between $\widehat{r}(f_n^*, f^*)$ and $r(f_n^*, f^*)$: With probability at least $1-\delta$,

$$\widehat{r}(f_n^*, f^*) \leq r(f_n^*, f^*) + \alpha\ r(f_n^*, f^*) + \frac{\log(\frac{1}{\delta})}{n\epsilon}. \tag{6}$$

Now we can again use the union bound to combine (5) and (6): With probability at least $1-2\delta$, $\forall \delta > 0$,

$$r(\widehat{f}_n, f^*) \leq \frac{1+\alpha}{1-\alpha} r(f_n^*, f^*) + \frac{c(f_n^*)\log 2 + 2\log 1/\delta}{n\epsilon}.$$

Now set $\delta = e^{\frac{-n\epsilon t}{2}}$, then we have

$$\mathcal{P}(r(\widehat{f}_n, f^*) - \frac{1+\alpha}{1-\alpha} r(f_n^*, f^*) + \frac{c(f_n^*)\log 2}{n\epsilon} \geq t) \leq 2e^{\frac{-n\epsilon t}{2}}.$$

Integrating, we get

$$\begin{aligned} E\left[r(\widehat{f}_n, f^*) - \frac{1+\alpha}{1-\alpha} r(f_n^*, f^*) + \frac{c(f_n^*)\log 2}{n\epsilon}\right] &\leq \int_0^\infty \mathcal{P}(\text{ " } \geq t)\, dt \\ &\leq \int_0^\infty 2e^{\frac{-n\epsilon t}{2}} \\ &= \frac{4}{n\epsilon} \end{aligned}$$

To Sum up, we have shown that for $\epsilon < \frac{6}{19b^2}$,

$$E[r(\widehat{f}_n, f^*)] \leq \left(\frac{1+\alpha}{1-\alpha}\right) r(f_n^*, f^*) + \frac{c(f_n^*)\log 2 + 4}{n\epsilon},$$

or,

$$E[r(\widehat{f}_n, f^*)] \leq \left(\frac{1+\alpha}{1-\alpha}\right) \min_{f\in\mathcal{F}} \left\{ r(f, f^*) + \frac{c(f)\log 2}{n\epsilon} \right\} + \frac{4}{n\epsilon},$$

since $\alpha < 1$. Or, in expanded form:

$$E[R(\widehat{f}_n)] - R(f^*) \leq \left(\frac{1+\alpha}{1-\alpha}\right) \min_{f\in\mathcal{F}} \left\{ R(f) - R(f^*) + \frac{c(f)\log 2}{n\epsilon} \right\} + \frac{4}{n\epsilon}$$

Notice that if $f^* \in \mathcal{F}$ and if $c(f^*)$ is not too large (e.g., $c(f^*) \approx \log n$), then we have $E[R(\widehat{f}_n)] - R(f^*) = O(n^{-1}\log n)$, within a logarithmic factor of the parametric rate of convergence!