18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet
Scribe: Cheng Mao

Lecture 4
Sep. 21, 2015

In this lecture, we continue to discuss the effect of noise on the rate of the excess risk $\mathcal{E}(\hat{h}) = R(\hat{h}) - R(h^*)$ where $\hat{h}$ is the empirical risk minimizer. In the binary classification model, noise roughly means how close the regression function η is from $\frac{1}{2}$. In particular, if $\eta = \frac{1}{2}$ then we observe only noise, and if $\eta \in \{0, 1\}$ we are in the noiseless case which has been studied last time. Especially, we achieved the fast rate $\frac{\log M}{n}$ in the noiseless case by assuming $h^* \in \mathcal{H}$ which implies that $\bar{h} = h^*$. This assumption was essential for the proof and we will see why it is necessary again in the following section.

3.2 Noise conditions

The noiseless assumption is rather unrealistic, so it is natural to ask what the rate of excess risk is when the noise is present but can be controlled. Instead of the condition $\eta \in \{0, 1\}$, we can control the noise by assuming that η is uniformly bounded away from $\frac{1}{2}$, which is the motivation of the following definition.

Definition (Massart's noise condition): The noise in binary classification is said to satisfy Massart's condition with constant $\gamma \in (0, \frac{1}{2}]$ if $|\eta(X) - \frac{1}{2}| \geq \gamma$ almost surely.

Once uniform boundedness is assumed, the fast rate simply follows from last proof with appropriate modification of constants.

Theorem: Let $cE(\hat{h})$ denote the excess risk of the empirical risk minimizer $\hat{h} = \hat{h}^{\mathrm{erm}}$. If Massart's noise condition is satisfied with constant γ, then

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

with probability at least $1 - \delta$. (In particular $\gamma = \frac{1}{2}$ gives exactly the noiseless case.)

Proof. Define $Z_i(h) = \mathbb{1}(\bar{h}(X_i) \neq Y_i) - \mathbb{1}(h(X_i) \neq Y_i)$. By the assumption $\bar{h} = h^*$ and the definition of $\hat{h} = \hat{h}^{\mathrm{erm}}$,

$$\begin{aligned}
\mathcal{E}(\hat{h}) &= R(\hat{h}) - R(\bar{h}) \\
&= \hat{R}_n(\hat{h}) - \hat{R}_n(\bar{h}) + \hat{R}_n(\bar{h}) - \hat{R}_n(\hat{h}) - \big(R(\bar{h}) - R(\hat{h})\big) \qquad (3.1) \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \big(Z_i(\hat{h}) - \mathbb{E}[Z_i(\hat{h})]\big). \qquad (3.2)
\end{aligned}$$

Hence it suffices to bound the deviation of $\sum_i Z_i$ from its expectation. To this end, we hope to apply Bernstein's inequality. Since

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

we have that for any $1 \le j \le M$,

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

Bernstein's inequality implies that

$$\mathbb{P}\big[\frac{1}{n}\sum_{i=1}^{n}(Z_i(h_j) - \mathbb{E}[Z_i(h_j)]) > t\big] \le \exp\big(-\frac{nt^2}{2\sigma_j^2 + \frac{2}{3}t}\big) =: \frac{\delta}{M}.$$

Applying a union bound over $1 \le j \le M$ and taking

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

we get that

$$\frac{1}{n}\sum_{i=1}^{n}(Z_i(h_j) - \mathbb{E}[Z_i(h_j)]) \le t_0(j) \tag{3.3}$$

for all $1 \le j \le M$ with probability at least $1 - \delta$.

Suppose $\hat{h} = h_{\hat{j}}$. It follows from (3.2) and (3.3) that with probability at least $1 - \delta$,

$$\mathcal{E}(\hat{h}) \le t_0(\hat{j}).$$

(Note that so far the proof is exactly the same as the noiseless case.) Since $|\eta(X) - \frac{1}{2}| \ge \gamma$ a.s. and $\bar{h} = h^*$,

$$\mathcal{E}(\hat{h}) = \mathbb{E}[|2\eta(X) - 1|\mathbb{1}(\hat{h}(X) \neq h^*(X))] \ge 2\gamma\mathbb{P}[h_{\hat{j}}(X) \neq \bar{h}(X)] = 2\gamma\sigma_{\hat{j}}^2.$$

Therefore,

$$\mathcal{E}(\hat{h}) \le \max\big(\sqrt{\frac{\mathcal{E}(\hat{h})\log(M/\delta)}{\gamma n}}, \frac{2\log(M/\delta)}{3n}\big), \tag{3.4}$$

so we conclude that with probability at least $1 - \delta$,

$$\mathcal{E}(\hat{h}) \le \frac{\log(M/\delta)}{\gamma n}.$$

□

The assumption that $\bar{h} = h^*$ was used twice in the proof. First it enables us to ignore the approximation error and only study the stochastic error. More importantly, it makes the excess risk appear on the right-hand side of (3.4) so that we can rearrange the excess risk to get the fast rate.

Massart's noise condition is still somewhat strong because it assumes uniform boundedness of η from $\frac{1}{2}$. Instead, we can allow η to be close to $\frac{1}{2}$ but only with small probability, and this is the content of next definition.

Definition (Tsybakov's noise condition or Mammen-Tsybakov noise condition): The noise in binary classification is said to satisfy Tsybakov's condition if there exists $\alpha \in (0,1)$, $C_0 > 0$ and $t_0 \in (0, \frac{1}{2}]$ such that

$$\mathbb{P}[|\eta(X) - \frac{1}{2}| \leq t] \leq C_0 t^{\frac{\alpha}{1-\alpha}}$$

for all $t \in [0, t_0]$.

In particular, as $\alpha \to 1$, $t^{\frac{\alpha}{1-\alpha}} \to 0$, so this recovers Massart's condition with $\gamma = t_0$ and we have the fast rate. As $\alpha \to 0$, $t^{\frac{\alpha}{1-\alpha}} \to 1$, so the condition is void and we have the slow rate. In between, it is natural to expect fast rate (meaning faster than slow rate) whose order depends on α. We will see that this is indeed the case.

Lemma: Under Tsybakov's noise condition with constants α, C_0 and t_0, we have

$$\mathbb{P}[h(X) \neq h^*(X)] \leq C\mathcal{E}(h)^\alpha$$

for any classifier h where $C = C(\alpha, C_0, t_0)$ is a constant.

Proof. We have

$$\begin{aligned}
\mathcal{E}(h) &= \mathbb{E}[|2\eta(X) - 1| \mathbb{1}(h(X) \neq h^*(X))] \\
&\geq \mathbb{E}[|2\eta(X) - 1| \mathbb{1}(|\eta(X) - \frac{1}{2}| > t) \mathbb{1}(h(X) \neq h^*(X))] \\
&\geq 2t\mathbb{P}[|\eta(X) - \frac{1}{2}| > t, h(X) \neq h^*(X)] \\
&\geq 2t\mathbb{P}[h(X) \neq h^*(X)] - 2t\mathbb{P}[|\eta(X) - \frac{1}{2}| \leq t] \\
&\geq 2t\mathbb{P}[h(X) \neq h^*(X)] - 2C_0 t^{\frac{1}{1-\alpha}}
\end{aligned}$$

where Tsybakov's condition was used in the last step. Take $t = c\mathbb{P}[h(X) \neq h^*(X)]^{\frac{1-\alpha}{\alpha}}$ for some positive $c = c(\alpha, C_0, t_0)$ to be chosen later. We assume that $c \leq t_0$ to guarantee that $t \in [0, t_0]$. Since $\alpha \in (0,1)$,

$$\begin{aligned}
\mathcal{E}(h) &\geq 2c\mathbb{P}[h(X) \neq h^*(X)]^{1/\alpha} - 2C_0 c^{\frac{1}{1-\alpha}} \mathbb{P}[h(X) \neq h^*(X)]^{1/\alpha} \\
&\geq c\mathbb{P}[h(X) \neq h^*(X)]^{1/\alpha}
\end{aligned}$$

by selecting c sufficiently small depending on α and C_0. Therefore

$$\mathbb{P}[h(X) \neq h^*(X)] \leq \frac{1}{c^\alpha} \mathcal{E}(h)^\alpha$$

and choosing $C = C(\alpha, C_0, t_0) := c^{-\alpha}$ completes the proof. □

Having established the key lemma, we are ready to prove the promised fast rate under Tsybakov's noise condition.

Theorem: If Tsybakov's noise condition is satisfied with constant α, C_0 and t_0, then there exists a constant $C = C(\alpha, C_0, t_0)$ such that

$$\mathcal{E}(\hat{h}) \leq C\big(\frac{\log(M/\delta)}{n}\big)^{\frac{1}{2-\alpha}}$$

with probability at least $1 - \delta$.

This rate of excess risk parametrized by α is indeed an interpolation of the slow ($\alpha \to 0$) and the fast rate ($\alpha \to 1$). Futhermore, note that the empirical risk minimizer $\hat{h}$ does not depend on the parameter α at all! It automatically adjusts to the noise level, which is a very nice feature of the empirical risk minimizer.

Proof. The majority of last proof remains valid and we will explain the difference. After establishing that

$$\mathcal{E}(\hat{h}) \leq t_0(\hat{j}),$$

we note that the lemma gives

$$\sigma_{\hat{j}}^2 = \mathbb{P}[\hat{h}(X) \neq \bar{h}(X)] \leq C\mathcal{E}(\hat{h})^\alpha.$$

It follows that

$$\mathcal{E}(\hat{h}) \leq \max\big(\sqrt{\frac{2C\mathcal{E}(\hat{h})^\alpha \log(M/\delta)}{n}}, \frac{2\log(M/\delta)}{3n}\big)$$

and thus

$$\mathcal{E}(\hat{h}) \leq \max\Big(\big(\frac{2C\log\frac{M}{\delta}}{n}\big)^{\frac{1}{2-\alpha}}, \frac{2\log(M/\delta)}{3n}\Big).$$

□

4. VAPNIK-CHERVONENKIS (VC) THEORY

The upper bounds proved so far are meaningful only for a finite dictionary $\mathcal{H}$, because if $M = |\mathcal{H}|$ is infinite all of the bounds we have will simply be infinity. To extend previous results to the infinite case, we essentially need the condition that only a finite number of elements in an infinite dictionary $\mathcal{H}$ really matter. This is the objective of the Vapnik-Chervonenkis (VC) theory which was developed in 1971.

4.1 Empirical measure

Recall from previous proofs (see (3.1) for example) that the key quantity we need to control is

$$2\sup_{h\in\mathcal{H}}\big(\hat{R}_n(h) - R(h)\big).$$

Instead of the union bound which would not work in the infinite case, we seek some bound that potentially depends on n and the complexity of the set $\mathcal{H}$. One approach is to consider some metric structure on $\mathcal{H}$ and hope that if two elements in $\mathcal{H}$ are close, then the quantity evaluated at these two elements are also close. On the other hand, the VC theory is more combinatorial and does not involve any metric space structure as we will see.

By definition

$$\hat{R}_n(h) - R(h) = \frac{1}{n}\sum_{i=1}^{n}\big(\mathbb{1}(h(X_i) \neq Y_i) - \mathbb{E}[\mathbb{1}(h(X_i) \neq Y_i)]\big).$$

Let $Z = (X, Y)$ and $Z_i = (X_i, Y_i)$, and let $\mathcal{A}$ denote the class of measurable sets in the sample space $\mathcal{X} \times \{0, 1\}$. For a classifier h, define $A_h \in \mathcal{A}$ by

$$\{Z_i \in A_h\} = \{h(X_i) \neq Y_i\}.$$

Moreover, define measures μ_n and μ on $\mathcal{A}$ by

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(Z_i \in A) \quad \text{and} \quad \mu(A) = \mathbb{P}[Z_i \in A]$$

for $A \in \mathcal{A}$. With this notation, the slow rate we proved is just

$$\sup_{h\in\mathcal{H}} \hat{R}_n(h) - R(h) = \sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \leq \sqrt{\frac{\log(2|A|/\delta)}{2n}}.$$

Since this is not accessible in the infinite case, we hope to use one of the concentration inequalities to give an upper bound. Note that $\mu_n(A)$ is a sum of random variables that may not be independent, so the only tool we can use now is the bounded difference inequality.

If we change the value of only one z_i in the function

$$z_1, \ldots, z_n \mapsto \sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|,$$

the value of the function will differ by at most $1/n$. Hence it satisfies the bounded difference assumption with $c_i = 1/n$ for all $1 \leq i \leq n$. Applying the bounded difference inequality, we get that

$$\Big| \sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| - \mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|]\Big| \leq \sqrt{\frac{\log(2/\delta)}{2n}}$$

with probability at least $1 - \delta$. Note that this already precludes any fast rate (faster than $n^{-1/2}$). To achieve fast rate, we need Talagrand inequality and localization techniques which are beyond the scope of this section.

It follows that with probability at least $1 - \delta$,

$$\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \leq \mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

We will now focus on bounding the first term on the right-hand side. To this end, we need a technique called symmetrization, which is the subject of the next section.

4.2 Symmetrization and Rademacher complexity

Symmetrization is a frequently used technique in machine learning. Let $\mathcal{D} = \{Z_1, \ldots, Z_n\}$ be the sample set. To employ symmetrization, we take another independent copy of the sample set $\mathcal{D}' = \{Z'_1, \ldots, Z'_n\}$. This sample only exists for the proof, so it is sometimes referred to as a ghost sample. Then we have

$$\mu(A) = \mathbb{P}[Z \in A] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(Z'_i \in A)] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(Z'_i \in A)|\mathcal{D}] = \mathbb{E}[\mu'_n(A)|\mathcal{D}]$$

where $\mu_n' := \frac{1}{n}\sum_{i=1}^n \mathbb{1}(Z_i' \in A)$. Thus by Jensen's inequality,

$$\begin{aligned}
\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] &= \mathbb{E}\big[\sup_{A\in\mathcal{A}} \big|\mu_n(A) - \mathbb{E}[\mu_n'(A)|\mathcal{D}]\big|\big] \\
&\le \mathbb{E}\big[\sup_{A\in\mathcal{A}} \mathbb{E}[|\mu_n(A) - \mu_n'(A)|\,|\mathcal{D}]\big] \\
&\le \mathbb{E}\big[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu_n'(A)|\big] \\
&= \mathbb{E}\big[\sup_{A\in\mathcal{A}} \big|\frac{1}{n}\sum_{i=1}^n \big(\mathbb{1}(Z_i \in A) - \mathbb{1}(Z_i' \in A)\big)\big|\big].
\end{aligned}$$

Since $\mathcal{D}'$ has the same distribution of $\mathcal{D}$, by symmetry $\mathbb{1}(Z_i \in A) - \mathbb{1}(Z_i' \in A)$ has the same distribution as $\sigma_i\big(\mathbb{1}(Z_i \in A) - \mathbb{1}(Z_i' \in A)\big)$ where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rad($\frac{1}{2}$), i.e.

$$\mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma_i = -1] = \frac{1}{2},$$

and σ_i's are taken to be independent of both samples. Therefore,

$$\begin{aligned}
\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] &\le \mathbb{E}\big[\sup_{A\in\mathcal{A}} \big|\frac{1}{n}\sum_{i=1}^n \sigma_i\big(\mathbb{1}(Z_i \in A) - \mathbb{1}(Z_i' \in A)\big)\big|\big] \\
&\le 2\mathbb{E}\big[\sup_{A\in\mathcal{A}} \big|\frac{1}{n}\sum_{i=1}^n \sigma_i \mathbb{1}(Z_i \in A)\big|\big]. \qquad (4.5)
\end{aligned}$$

Using symmetrization we have bounded $\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|]$ by a much nicer quantity. Yet we still need an upper bound of the last quantity that depends only on the structure of $\mathcal{A}$ but not on the random sample $\{Z_i\}$. This is achieved by taking the supremum over all $z_i \in \mathcal{X} \times \{0, 1\} =: \mathcal{Y}$.

Definition: The Rademacher complexity of a family of sets $\mathcal{A}$ in a space $\mathcal{Y}$ is defined to be the quantity

$$\mathcal{R}_n(\mathcal{A}) = \sup_{z_1,\ldots,z_n\in\mathcal{Y}} \mathbb{E}\big[\sup_{A\in\mathcal{A}} \big|\frac{1}{n}\sum_{i=1}^n \sigma_i \mathbb{1}(z_i \in A)\big|\big].$$

The Rademacher complexity of a set $B \subset \mathbb{R}^n$ is defined to be

$$\mathcal{R}_n(B) = \mathbb{E}\big[\sup_{b\in B} \big|\frac{1}{n}\sum_{i=1}^n \sigma_i b_i\big|\big].$$

We conclude from (4.5) and the definition that

$$\mathbb{E}[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] \le 2\mathcal{R}_n(\mathcal{A}).$$

In the definition of Rademacher complexity of a set, the quantity $\big|\frac{1}{n}\sum_{i=1}^n \sigma_i b_i\big|$ measures how well a vector $b \in B$ correlates with a random sign pattern $\{\sigma_i\}$. The more complex B is, the better some vector in B can replicate a sign pattern. In particular, if B is the full hypercube $[-1, 1]^n$, then $\mathcal{R}_n(B) = 1$. However, if $B \subset [-1, 1]^n$ contains only k-sparse

vectors, then $\mathcal{R}_n(B) = k/n$. Hence $\mathcal{R}_n(B)$ is indeed a measurement of the complexity of the set B.

The set of vectors to our interest in the definition of Rademacher complexity of $\mathcal{A}$ is

$$T(z) := \{(\mathbb{1}(z_1 \in A), \ldots, \mathbb{1}(z_n \in A))^T, A \in \mathcal{A}\}.$$

Thus the key quantity here is the cardinality of $T(z)$, i.e., the number of sign patterns these vectors can replicate as A ranges over $\mathcal{A}$. Although the cardinality of $\mathcal{A}$ may be infinite, the cardinality of $T(z)$ is bounded by 2^n.

18.657 Mathematics of Machine Learning
Fall 2015