IDS.160 – Mathematical Statistics: A Non-Asymptotic Approach

Lecturer: A Rakhlin | Lecture 26
Scribe: A. Rakhlin | May 12, 2020

1. TIME SERIES

Suppose we observe a sequence

$$\boldsymbol{x}_{t+1} = f^*(\boldsymbol{x}_t) + \eta_t, \quad t = 1, \ldots, n$$

where $\boldsymbol{x}_t \in \mathbb{R}^d$ and η_t are independent zero mean vectors. The function f^* is unknown, but we assume it is a member of a known class $\mathcal{F}$. Let us treat this problem as a fixed-design regression problem, except that the outcomes are now vectors rather than reals, and the sequence $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is a sequence of *dependent* random variables.

Consider the least squares solution:

$$\widehat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \|\boldsymbol{x}_{t+1} - f(\boldsymbol{x}_t)\|_2^2,$$

where the norm is the euclidean norm. This is a natural generalization of least squares to vector-valued regression. As before, we denote

$$\|f - g\|_n^2 = \frac{1}{n} \sum_{t=1}^{n} \|f(\boldsymbol{x}_t) - g(\boldsymbol{x}_t)\|_2^2$$

The basic inequality can now be written as (exercise):

$$\left\|\widehat{f} - f^*\right\|_n^2 \leq 2 \frac{1}{n} \sum_{t=1}^{n} \langle \eta_t, \widehat{f}(\boldsymbol{x}_t) - f^*(\boldsymbol{x}_t) \rangle.$$

Choosing the offset-style approach covered in previous lectures, we have

$$\left\|\widehat{f} - f^*\right\|_n^2 \leq \sup_{g \in \mathcal{F} - f^*} \frac{1}{n} \sum_{t=1}^{n} 4 \langle \eta_t, g(\boldsymbol{x}_t) \rangle - \|g(\boldsymbol{x}_t)\|^2.$$

Up until now, the statement is conditional on $\{\eta_1, \ldots, \eta_n\}$. What happens if we take expectations on both sides? On the left-hand side we have a denoising guarantee on the sequence. On the right-hand side, we have a "dependent version" of offset Gaussian/Rademacher complexity where $\boldsymbol{x}_t$ is measurable with respect to $\sigma(\eta_1, \ldots, \eta_{t-1})$. To analyze this object, we first need to understand the simpler $\mathbb{R}$-valued version without the offset: what is the behavior of

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\boldsymbol{x}_t)$$

where $\boldsymbol{x}_t$ is $\sigma(\epsilon_1, \ldots, \epsilon_{t-1})$-measurable, $\mathcal{F}$ is a class of real-valued functions $\mathcal{X} \to \mathbb{R}$, and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables.

2. SEQUENTIAL COMPLEXITIES

We choose to study the random process generated by Rademacher random variables for several reasons. First, just as in the classical case, conditioning on the data will lead to a simpler object (binary tree) and, second, other noise processes can be reduced to the Rademacher case, under moment assumptions on the noise. The development here is based on [3], and we refer also to [2] for an introduction.

Let us elaborate on the first point. Note that $\boldsymbol{x}_t$ being measurable with respect to $\sigma(\epsilon_1, \ldots, \epsilon_{t-1})$ simply means $\boldsymbol{x}_t$ is a function of $\epsilon_1, \ldots, \epsilon_{t-1}$ (in other words, it's a predictable process). Note that the collection $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ can be "summarized" as a depth-n binary tree decorated with elements of $\mathcal{X}$ at the nodes. Indeed, $\boldsymbol{x}_1 \in \mathcal{X}$ is a constant (root), $\boldsymbol{x}_2 = \boldsymbol{x}_2(\epsilon_1)$ takes on two possible values depending on the sign of ϵ_1 (left or right), and so forth. It is useful to think of $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ as a tree, even though it doesn't bring any more information into the picture. We shall denote the collection of n functions $\boldsymbol{x}_i : \{\pm 1\}^{i-1} \to \mathcal{X}$ as $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and call it simply as an $\mathcal{X}$-valued *tree*. We shall refer to $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ as a *path* in the tree. We will also talk about $\mathbb{R}$-valued trees, such as $f \circ \boldsymbol{x}$ for $f : \mathcal{X} \to \mathbb{R}$.

Given a tree $\boldsymbol{x}$, we shall call

$$\widehat{\mathcal{R}}^{\mathrm{seq}}(\mathcal{F}, \boldsymbol{x}) = \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\boldsymbol{x}_t(\epsilon_1, \ldots, \epsilon_{t-1}))$$

the *sequential Rademacher complexity* of $\mathcal{F}$ on the tree $\boldsymbol{x}$.

Comparing to the classical version,

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(x_t)$$

where $x_1, \ldots, x_n$ are constant values, we see that it is a special case of a tree with constant levels $\boldsymbol{x}_t(\epsilon_1, \ldots, \epsilon_{t-1}) = x_t$. Hence, sequential Rademacher complexity is a generalization of the classical notion.

To ease the notation, we will write $\boldsymbol{x}_t$ without explicit dependence on ϵ, or for brevity write $\boldsymbol{x}_t(\epsilon)$ even though $\boldsymbol{x}_t$ only depends on the prefix $\epsilon_{1:t-1}$.

Observe that for any $f \in \mathcal{F}$, the variable

$$\nu_f = \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\boldsymbol{x}_t)$$

is zero mean. Moreover, it is an average of martingale differences $\epsilon_t f(\boldsymbol{x}_t)$, and so we expect $1/\sqrt{n}$ behavior from Azuma-Hoeffding's inequality. It should be clear that, say, for $\mathcal{F}$ consisting of a finite collection of $[-1, 1]$-valued functions on $\mathcal{X}$, we have

$$\mathbb{E} \max_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\boldsymbol{x}_t) \leq \sqrt{\frac{2 \log \operatorname{card}(\mathcal{F})}{n}}$$

Given that there is no difference with the classical case, one may wonder if we can just reduce everything to the classical Rademacher averages. The answer is no, and the differences already start to appear when we attempt to define covering numbers.

More precisely, since any tree $\boldsymbol{x}$ is defined by $2^n - 1$ values, one might wonder if we could define a notion of pseudo-distance between f and f' as an ℓ_2 distance on these $2^n - 1$ values.

It is easy to see that this is a huge overkill. Perhaps one of the key points to understand here is: what is the equivalent of the projection $\mathcal{F}|_{x_1,\ldots,x_n}$ for the tree case? Spoiler: it's not $\mathcal{F}|_{\boldsymbol{x}}$. The following turns out to be the right definition:

Definition: A set V of $\mathbb{R}$-valued trees is an 0-cover of $\mathcal{F}$ on a tree $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ if

$$\forall f \in \mathcal{F}, \epsilon \in \{\pm 1\}^n, \exists \boldsymbol{v} \in V \quad \text{s.t.} \quad f(\boldsymbol{x}_t(\epsilon_{1:t-1})) = \boldsymbol{v}_t(\epsilon_{1:t-1}) \quad \forall t \in [n]$$

The size of the smallest 0-cover of $\mathcal{F}$ on a tree $\boldsymbol{x}$ will be denoted by $\mathcal{N}(\mathcal{F}, \boldsymbol{x}, 0)$.

The key aspect of this definition is that $\boldsymbol{v} \in V$ can be chosen based on the sequence $\epsilon \in \{\pm 1\}^n$. In other words, in contrast with the classical definition, for the same function f different elements $\boldsymbol{v} \in V$ can provide a cover on different paths. This results in the needed reduction in the size of V.

As an example, take a set of 2^{n-1} functions that take a value of 1 on one of the 2^{n-1} leaves of $\boldsymbol{x}$ and zero everywhere else. Then the projection $\mathcal{F}|_{\boldsymbol{x}}$ is of size 2^{n-1} but the size of the 0-cover is only 2, corresponding to our intuition that the class is simple (as it only varies on the last example). Indeed, the size of the 0-cover is the analogue of the size of $\mathcal{F}|_{x_1,\ldots,x_n}$ in the binary-valued case.

For real-valued functions, consider the following definition.

Definition: A set V of $\mathbb{R}$-valued trees is an α-cover of $\mathcal{F}$ on a tree $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ with respect to ℓ_2 if

$$\forall f \in \mathcal{F}, \epsilon \in \{\pm 1\}^n, \exists \boldsymbol{v} \in V \quad \text{s.t.} \quad \frac{1}{n}\sum_{i=1}^{n} (f(\boldsymbol{x}_t(\epsilon_{1:t-1})) - \boldsymbol{v}_t(\epsilon_{1:t-1}))^2 \le \alpha^2$$

The size of the smallest α-cover of $\mathcal{F}$ on a tree $\boldsymbol{x}$ with respect to ℓ_2 will be denoted by $\mathcal{N}_2(\mathcal{F}, \boldsymbol{x}, \alpha)$.

A similar definition can be stated for cover with respect to ℓ_p.

The following is an analogue of the chaining bound:

Theorem: For any class of $[-1, 1]$-valued functions $\mathcal{F}$,

$$\widehat{\mathcal{R}}^{\text{seq}}(\mathcal{F}, \boldsymbol{x}) \le \inf_{\alpha \ge 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{\log \mathcal{N}_2(\mathcal{F}, \boldsymbol{x}, \varepsilon)} d\varepsilon \right\}$$

Recall the definition of VC dimension and a shattered set. Here is the right sequential analogue:

Definition: Function class $\mathcal{F}$ of $\{\pm 1\}$-valued functions shatters a tree $\boldsymbol{x}$ of depth d if

$$\forall \epsilon \in \{\pm 1\}^d, \ \exists f \in \mathcal{F}, \quad \text{s.t.} \quad \forall t \in [d], \qquad f(\boldsymbol{x}_t(\epsilon)) = \epsilon_t$$

The largest depth d for which there exists a shattered $\mathcal{X}$-valued tree is called the *Littlestone dimension* and denoted by $\mathrm{ldim}(\mathcal{F})$.

To contrast with the classical definition, the path on which the signs should be realized is given by the path itself. But it's clear that the definition serves the same purpose: if $\boldsymbol{x}$ is shattered by $\mathcal{F}$ then $\widehat{\mathcal{R}}^{\mathrm{seq}}(\mathcal{F}, \boldsymbol{x}) = 1$. It is also easy to see that $\mathrm{vc}(\mathcal{F}) \leq \mathrm{ldim}(\mathcal{F})$, and the gap can be infinite.

The following is an analogue of the Sauer-Shelah-Vapnik-Chervonenkis lemma.

Theorem: For a class of binary-valued functions $\mathcal{F}$ with Littlestone dimension $\mathrm{ldim}(\mathcal{F})$,

$$\mathcal{N}(\mathcal{F}, \boldsymbol{x}, 0) \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^{d}$$

Scale-sensitive sequential versions are defined as follows:

Definition: Function class $\mathcal{F}$ of $\mathbb{R}$-valued functions shatters a tree $\boldsymbol{x}$ of depth d at scale α if there exists a witness $\mathbb{R}$-valued tree $\boldsymbol{s}$ such that

$$\forall \epsilon \in \{\pm 1\}^{d},\ \exists f \in \mathcal{F}, \quad \text{s.t.} \quad \forall t \in [d], \qquad \epsilon_t(f(\boldsymbol{x}_t(\epsilon)) - \boldsymbol{s}_t(\epsilon)) \geq \alpha/2$$

The largest depth d for which there exists an α-shattered $\mathcal{X}$-valued tree is called sequential scale-sensitive dimension and denoted $\mathrm{ldim}(\mathcal{F}, \alpha)$.

We note that the above definitions reduce to the classical ones if we consider only trees $\boldsymbol{x}$ with constant levels.

Theorem: For any class of $[-1, 1]$-valued functions $\mathcal{F}$ and $\mathcal{X}$-valued tree $\boldsymbol{x}$ of depth n

$$\mathcal{N}_{\infty}(\mathcal{F}, \boldsymbol{x}, \alpha) \leq \left(\frac{2en}{\alpha}\right)^{\mathrm{ldim}(\mathcal{F}, \alpha)}$$

Finally, it is possible to show an analogue of symmetrization lemma: for any joint distribution of $(X_1, \ldots, X_n)$,

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}[f(X_t) | X_{1:t-1}] - f(X_t) \leq 2 \sup_{\boldsymbol{x}} \widehat{\mathcal{R}}^{\mathrm{seq}}(\mathcal{F}, \boldsymbol{x})$$

If the sequence $(X_1, \ldots, X_n)$ is i.i.d., the left-hand side is the expected supremum of the empirical process. The present version provides a martingale generalization. Furthermore, if we take supremum over all joint distributions on the left-hand-side, then the lower bound is also matching the upper bound, up to a constant.

The offset Rademacher complexity has been analyzed in [1].

3. ONLINE LEARNING

Consider the following online classification problem. On each of n rounds $t = 1, \dots, n$, the learner observes $x_t \in \mathcal{X}$, makes a prediction $\widehat{y}_t \in \{\pm 1\}$, and observes the outcome $y_t \in \{\pm 1\}$. The learner models the problem by fixing a class $\mathcal{F}$ of possible models $f : \mathcal{X} \to \{\pm 1\}$, and aims to predict nearly as well as the best model in $\mathcal{F}$ in the sense of keeping *regret*

$$\mathrm{Reg}(\mathcal{F}) = \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\left\{\widehat{y}_t \neq y_t\right\}\right] - \inf_{f \in \mathcal{F}} \left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\left\{f(x_t) \neq y_t\right\}\right] \tag{3.1}$$

small for any sequence $(x_1, y_1), \dots, (x_n, y_n)$. At least visually, this looks like oracle inequalities for misspecified models. The distinguishing feature of this online framework is that (a) data arrives sequentially, and (b) we aim to have low regret for any sequence without assuming any generative process.

It is also worth noting that in the above protocol there is no separation of training and test data: the online nature of the problem allows us to first test our current hypothesis by making a prediction, then observe the outcome and incorporate the datum in to our dataset.

The expectation on the first term in (3.1) is with respect to learner's internal randomization. More specifically, let Q_t be the distribution on $\{\pm 1\}$ that the learner uses to predict $\widehat{y}_t \sim Q_t$. Let $q_t = \mathbb{E}\widehat{y}_t$ be the (conditional) mean of this distribution. In other words, $q_t = 0$ would correspond to the learner tossing a fair coin.

A note about the protocol. The results below hold even if the sequence is chosen based on learner's past predictions. However, in this case, y_t may only depend on q_t but not on the realization $\widehat{y}_t$. To simplify the presentation, let us just assume that the sequence $(x_1, y_1), \dots, (x_n, y_n)$ is fixed in advanced (this turns out not to matter).

We will answer the following question: what is the best achievable $\mathrm{Reg}(\mathcal{F})$ for a given $\mathcal{F}$ by any prediction strategy?

Let us first rewrite $\mathbf{1}\left\{\widehat{y}_t \neq y_t\right\} = (1 - \widehat{y}_t y_t)/2$ and do the same for the oracle term. Cancelling $1/2$, we have

$$\begin{aligned} 2\mathrm{Reg}(\mathcal{F}) &= \frac{1}{n}\sum_{t=1}^{n} -q_t y_t - \inf_{f \in \mathcal{F}} \left[\frac{1}{n}\sum_{t=1}^{n} -y_t f(x_t)\right] && (3.2) \\ &= \sup_{f \in \mathcal{F}} \left[\frac{1}{n}\sum_{t=1}^{n} y_t f(x_t)\right] - \frac{1}{n}\sum_{t=1}^{n} q_t y_t && (3.3) \end{aligned}$$

Now, consider a particular stochastic process for generating the data sequence: fix any $\mathcal{X}$-valued tree $\boldsymbol{x}$ of depth n, and on round t let $x_t = \boldsymbol{x}_t(y_1, \dots, y_{t-1})$ and $y_t = \epsilon_t$ be an independent Rademacher random variable. This defines a stochastic process with 2^n possible sequences $(x_1, y_1), \dots, (x_n, y_n)$. Now, clearly

$$\sup_{(x_1,y_1),\dots,(x_n,y_n)} 2\mathrm{Reg}(\mathcal{F}) \geq 2\mathbb{E}_\epsilon \mathrm{Reg}(\mathcal{F}).$$

Observe that $q_t = q_t(\epsilon_1, \dots, \epsilon_{t-1})$ and thus

$$\mathbb{E}_\epsilon \left[\frac{1}{n}\sum_{t=1}^{n} q_t \epsilon_t\right] = 0.$$

Hence,

$$\mathbb{E}_\epsilon \text{Reg}(\mathcal{F}) = \mathbb{E} \sup_{f\in\mathcal{F}} \left[\frac{1}{n} \sum_{t=1}^{n} \epsilon_t f(\boldsymbol{x}_t) \right]. \tag{3.4}$$

Since the argument holds for any $\boldsymbol{x}$, we have proved that the optimal value of $\text{Reg}(\mathcal{F})$ is lower bounded by half of

$$\bar{\mathcal{R}}^{\text{seq}}(\mathcal{F}) = \sup_{\boldsymbol{x}} \widehat{\mathcal{R}}^{\text{seq}}(\mathcal{F}, \boldsymbol{x}).$$

It turns out that this lower bound is within a factor of 2 from optimal. Define the minimax value

$$\mathcal{V} = \min_{\text{Algo}} \max_{\{(x_t,y_t)\}_{t=1}^n} \text{Reg}(\mathcal{F})$$

Theorem: For a binary-valued class $\mathcal{F}$,

$$\frac{1}{2}\bar{\mathcal{R}}^{\text{seq}}(\mathcal{F}) \leq \mathcal{V} \leq \bar{\mathcal{R}}^{\text{seq}}(\mathcal{F})$$

Similar results also holds for absolute value and other Lipschitz loss functions. For square loss, the sequential Rademacher averages are replaced by offset sequential Rademacher averages (again, as both upper and lower bounds).

In short, sequential complexities in online learning play a role similar to the role played by i.i.d. complexities as studied in this course. However, quite a large number of questions still remains open. But that's a topic for a different course.

References

[1] A. Rakhlin and K. Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264, 2014.

[2] A. Rakhlin and K. Sridharan. On martingale extensions of vapnik–chervonenkis theory with applications to online learning. In *Measures of Complexity*, pages 197–215. Springer, 2015.

[3] A. Rakhlin, K. Sridharan, and A. Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1-2):111–153, 2015.