Math 547: Mathematical Foundations of Statistical Learning Theory Fall 2017

Lecture 18 — October 2

Instructor: Stas Minsker *Scribe: Mose Wintner*

18.1 Applications of the symmetrization inequality

Let $T \subset \mathbb{R}^n$ be finite and let $t = (t_1, \dots, t_n)$. Consider

$$X(t) = \sum_{j=1}^{n} \varepsilon_j t_j$$

for ε_j iid Rademacher variables. Then $x(t) \in SG(||t||_2^2)$. Let $R(T) = \sup_{t \in T} |x(t)|$. Then

$$\mathbb{E}R(T) \le \sqrt{2} \max_{t \in T} ||t||_2 \sqrt{\log 2 \,\mathrm{card}(T)}.$$

Definition 1. Let $F \subset S$ be a finite set (a collection of points in S), and let $\mathcal{C}$ be a collection of subsets of S. The *shattering number*

$$\Delta^{\mathcal{C}}(F) := \mathrm{card}\{\{C \cap F\}, \, C \in \mathcal{C}\}.$$

Example 1. Let $F = \{x_1, \dots, x_n\} \subset \mathbb{R}$ and let $\mathcal{C} = \{(-\infty, t] \,:\, t \in \mathbb{R}\}$. Assume WLOG that the x_i are ordered. Then $\Delta^{\mathcal{C}}(F) = n + 1$ (we include the empty set).

If $\Delta^{\mathcal{C}}(F) = 2^{\mathrm{card}(F)}$, then we say that F is shattered by $\mathcal{C}$. Let $X_1, \dots, X_n$ be iid and consider

$$\mathbb{Z}_n(I_C) = \frac{1}{n} I\{X_j \in C\} - Pr(X \in C)$$

for $C \in \mathcal{C}$. Then $\{\mathbb{Z}_n(I_C), \, C \in \mathcal{C}\}$ is a stochastic process. By the symmetrization inequality,

$$\mathbb{E} \sup_{C \in \mathcal{C}} |\mathbb{Z}_n(I_C)| \le 2\mathbb{E}_X \mathbb{E}_{\varepsilon|X} \sup_{C \in \mathcal{C}} \left| \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j I\{X_j \in C\} \right|.$$

This term can be put in the form above with $t_j = I\{X_j \in C\}/n$,

$$T = \{\frac{1}{n} I\{X_j \in C\}, j = 1, \dots, n, \, C \in \mathcal{C}\}.$$

We have that

$$\sup_{t \in T} ||t||_2 = \sup_{C} \sqrt{\sum (I\{X_j \in C\}/n)^2} = 1/\sqrt{n},$$

when all the indicator functions are 1. We have

$$\mathbb{E}_{\varepsilon|X} \sup_{C \in \mathcal{C}} \left| \frac{1}{n} \sum \varepsilon_j I\{X_j \in C\} \right| \le \sqrt{2} \sqrt{\frac{1}{n}} \sqrt{\log 2\Delta^{\mathcal{C}}(X_1, \dots, X_n)}$$

because $\text{card}\{(I_C(X_1), \dots, I_C(X_n) :, C \in \mathcal{C}\} = \Delta^{\mathcal{C}}(X_1, \dots, X_n)$. Now taking expectations with respect to X and multiplying, we get the estimate

$$\mathbb{E} \sup_{C \in \mathcal{C}} |\mathbb{Z}_n(I_C)| \le 2\mathbb{E}_X \left[\sqrt{\frac{2}{n} \log 2\Delta^{\mathcal{C}}(\{X_1, \dots, X_n\})} \right].$$

Theorem 1. Let $X_1, \dots, X_n$ be iid from P and let $\mathcal{C}$ be a collection of subsets. Then

$$\mathbb{E} \sup_{c \in \mathcal{C}} |P_n(C) - P(C)| \le 2\sqrt{\frac{2}{n}} \mathbb{E} \sqrt{\log 2\Delta^{\mathcal{C}}(\{X_1, \dots, X_n\})}.$$

Corollary 1. Borel-Cantelli theorem. $X_1, \dots, X_n \in \mathbb{R}$ and $\mathcal{C}$ the set of intervals $(-\infty, t]$. If $X_j \sim F$, then

$$\mathbb{E} \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \le 2\sqrt{\frac{2}{n}} \sqrt{\log 2n + 2}.$$

In particular, as $n \to \infty$, this expectation converges to 0. In fact, it can be shown that this convergence can be improved to $O(1/\sqrt{n})$.

18.2 Shattering numbers and Vapnik-Chervonenkis (VC) dimension

Jensen's inequality implies that $\mathbb{E}\sqrt{X} \le \sqrt{\mathbb{E}X}$ for $X \ge 0$ a.s. Therefore

$$\mathbb{E}\sqrt{\log 2\Delta^{\mathcal{C}}(X_1, \dots, X_n)} \le \sqrt{\mathbb{E} \log 2\Delta^{\mathcal{C}}(X_1, \dots, X_n)}.$$

Theorem 2. Let $X_1, \dots, X_n$ be iid and let $\mathcal{C}$ be a collection of subsets of S. Then there exists a constant K for which

$$\mathbb{E}||P_n - P||_{\mathcal{C}} \le K \sqrt{\sup_{C \in \mathcal{C}} P(C)} \mathbb{E}\sqrt{\log 2\Delta^{\mathcal{C}}(X_1, \dots, X_n)/n} \vee K\mathbb{E} \log 2\Delta^{\mathcal{C}}(X_1, \dots, X_n)/n.$$

In particular, the rate of convergence is not faster than $1/n$. *Proof:* As before, we introduce a Rademacher process. Recall that our T was defined as collections of $\{I_C(X_j)/n\}$ for $j = 1 \dots, n$ and $C \in \mathcal{C}$. Therefore

$$\sup_{t \in T} ||t||_2 = \frac{1}{\sqrt{n}} \sqrt{\sum \frac{1}{n} I_C(X_j)} = \frac{1}{\sqrt{n}} \sup_{C \in \mathcal{C}} \sqrt{P_n(C)}.$$

$$
\begin{aligned}
\mathbb{E}||P_n - P||_{\mathcal{C}} &\le 2\mathbb{E}\sup_{C\in\mathcal{C}}\left|\frac{1}{n}\sum \varepsilon_j I\{X_j \in C\}\right| \\
&\le 2\sqrt{2}\mathbb{E}\left[\frac{1}{\sqrt{n}}\sup_C \sqrt{P_n(C)}\sqrt{\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)}\right] \\
&= \frac{2\sqrt{2}}{\sqrt{n}}\mathbb{E}\sup_C \sqrt{(P_n-P)(C)+P(C)}\sqrt{\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)} \\
&\le \frac{2\sqrt{2}}{\sqrt{n}}\mathbb{E}\sup_C \left(\sqrt{(P_n-P)(C)}+\sqrt{P(C)}\right)\sqrt{\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)} \\
&\le \frac{2\sqrt{2}}{\sqrt{n}}\sqrt{\sup_C P(C)}\mathbb{E}\sqrt{\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)} \\
&\qquad + \frac{2\sqrt{2}}{\sqrt{n}}\mathbb{E}\sup_C\left(\sqrt{(P_n-P)(C)}\right)\sqrt{\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)} \\
&\le \frac{2\sqrt{2}}{\sqrt{n}}\sqrt{\sup_C P(C)}\mathbb{E}\sqrt{\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)} \\
&\qquad + \frac{2\sqrt{2}}{\sqrt{n}}\mathbb{E}^{1/2}||P_n-P||_{\mathcal{C}}\mathbb{E}^{1/2}\sqrt{\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)}.
\end{aligned}
$$

We have

$$W = \mathbb{E}^{1/2}||P_n - P||_{\mathcal{C}} \le 2\sqrt{\frac{2}{n}}\sqrt{\sup_C P(C)}\mathbb{E}\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n) + 2\sqrt{\frac{2}{n}}\sqrt{W}\mathbb{E}^{1/2}\log 2\Delta^{\mathcal{C}}(X_1,\ldots,X_n)$$

$W \le \alpha_1 + \sqrt{W}\alpha_2 \le 2\max(\alpha_1, \sqrt{W}\alpha_2)$

$W \le \max(2\alpha_1, 4\alpha_2^2)$.