Mathematical foundations of infinite-dimensional statistical models

Chapter 3.5 Metric entropy bounds for suprema of empirical processes

Chapter 3.5.1 Bracketing 1: The expectation bound

Presented by Ilsang Ohn

December 28, 2018

Bracketing numbers

Definition (The first definition of the bracketing number). For any $\epsilon > 0$, the $L^p(\mathrm{P})$-bracketing number $N_{[]}(\mathcal{F}, L^p(\mathrm{P}), \epsilon)$ of $\mathcal{F} \subset L^p(\mathrm{P})$ is defined as the smallest cardinality of any partition $B_1, \ldots, B_N$ of $\mathcal{F}$ such that

$$\mathrm{P}\left[\left(\sup_{f,g\in B_i} |f-g|\right)^*\right]^p \leq \epsilon^p \text{ for every } i = 1, \ldots, N.$$

g^* denotes a measurable cover of a nonnegative, not necessarily measurable function g. Proposition 3.7.1 guarantees its existence.

Bracketing numbers

Definition (The second definition of the bracketing number). For any $\epsilon > 0$, the $L^p(\mathrm{P})$-bracketing number $N_{[]}(\mathcal{F}, L^p(\mathrm{P}), \epsilon)$ of $\mathcal{F} \subset L^p(\mathrm{P})$ is defined as the smallest cardinality of any pairs of the functions (f_i^L, f_i^U), $i = 1, \ldots, N$ with $f_i^L \leq F_i^U$ and $\mathrm{P}(f_i^U - f_i^L)^p \leq \epsilon^p$ such that for any $f \in \mathcal{F}$, there is $i \in \{1, \ldots, N\}$ such that $f_i^L \leq f \leq f_i^U$.

Proposition. The two definitions are equivalent:

$$N_{[]}^{2\mathrm{nd}}(\mathcal{F}, L^p(\mathrm{P}), 2\epsilon) \leq N_{[]}^{1\mathrm{st}}(\mathcal{F}, L^p(\mathrm{P}), \epsilon) \leq N_{[]}^{2\mathrm{nd}}(\mathcal{F}, L^p(\mathrm{P}), \epsilon).$$

Proof Let $B_i = [f_i^L, f_i^U]$. Then $B_1, \ldots, B_N$ is a partition of $\mathcal{F}$ with $\mathrm{P}\left[\left(\sup_{f,g \in B_i} |f - g|\right)^*\right]^p = \mathrm{P}(f_i^U - f_i^L)^p \leq \epsilon^p$.
Let $(f_i^L, f_i^U) = f_i \pm \sup_{f,g \in B_i} |f - g|$ for $f_i \in B_i$. Then
$\mathrm{P}(f_i^U - f_i^L)^p = 2^p \mathrm{P}(\sup_{f,g \in B_i} |f - g|)^p \leq (2\epsilon)^p$ □

Bracketing numbers and covering numbers

Proposition. For any $p \in [0, \infty)$,

$$N(\mathcal{F}, L^p(\mathrm{P}), \epsilon) \leq N_{[]}^{\text{1st}}(\mathcal{F}, L^p(\mathrm{P}), \epsilon) \leq N_{[]}^{\text{2nd}}(\mathcal{F}, L^p(\mathrm{P}), \epsilon) \leq N(\mathcal{F}, L^\infty(\mathrm{P}), \epsilon/2).$$

Proof Let $f_i = f_i^L$. Then for any f, there is f_i such that $\mathrm{P}\|f - f_i\|^p \leq \mathrm{P}\|f_i^U - f_i^L\|^p \leq \epsilon^p$.
Let $(f_i^L, f_i^U) = f_i \pm \epsilon/2$ where f_i, $i = 1, \ldots, N(\mathcal{F}, L^\infty(\mathrm{P}), \epsilon/2)$ is the minimal $\epsilon/2$-covering set. $\square$

Maximal inequality with the bracketing number

Theorem 3.5.13 Let P be a probability measure on $(S, \mathcal{S})$ and for any $n \in \mathrm{N}$, and let $X_1, \dots, X_n$ be an independent sample of size n from P. Let $\mathcal{F}$ be a class of measurable functions on S that admits a P-square integrable envelope F and satisfies the $L^2(\mathrm{P})$-bracketing condition

$$\int_0^2 \sqrt{\log N_{[]}(\mathcal{F}, L^2(\mathrm{P}), \tau \|F\|_{L^2(\mathrm{P})})} d\tau < \infty$$

Set $\sigma^2 := \sup_{f \in \mathcal{F}} \mathrm{P} f^2$ and

$$a(\delta) := \frac{\delta}{\sqrt{32 \log(2N_{[]}(\mathcal{F}, L^2(\mathrm{P}), \delta/2))}}.$$

Then for any $\delta > 0$

$$\begin{aligned} \mathrm{E}\left\|\sum_{i=1}^n (f(X_i) - \mathrm{P}f)\right\|_{\mathcal{F}}^* \leq & 56\sqrt{n} \int_0^{2\delta} \sqrt{\log(2N_{[]}(\mathcal{F}, L^2(\mathrm{P}), \tau))} d\tau \\ & + 4n\mathrm{P}(F \mathbb{1}(F > \sqrt{n} a(\delta)) \\ & + \sqrt{n\sigma^2 \log(2N_{[]}(\mathcal{F}, L^2(\mathrm{P}), \delta))} \end{aligned} \tag{1}$$

Compared to Theorem 3.5.4

$$\nu_n(f) = \sqrt{n}(\mathrm{P}_n - \mathrm{P})f = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i) - \mathrm{P}f)$$

Theorem 3.5.4 (Remark 3.5.5)

$$\mathrm{E}\|\nu_n\|_{\mathcal{F}}^* \lesssim \|F\|_{L^2(\mathrm{P})}\int_0^1 \sup_{Q:\text{finitely discrete}} \sqrt{\log\left\{2N\left(\mathcal{F}, L^2(Q), \tau\|F\|_{L^2(Q)}\right)\right\}}\mathrm{d}\tau$$

Theorem 3.5.13 (Remark 3.5.14)

$$\begin{aligned}\mathrm{E}\|\nu_n\|_{\mathcal{F}}^* &\lesssim \|F\|_{L^2(\mathrm{P})}\int_0^1 \sqrt{\log\left\{2N_{[]}\left(\mathcal{F}, L^2(\mathrm{P}), \tau\|F\|_{L^2(\mathrm{P})}\right)\right\}}\mathrm{d}\tau \\ &= \int_0^{\|F\|_{L^2(\mathrm{P})}} \sqrt{\log\left\{2N_{[]}\left(\mathcal{F}, L^2(\mathrm{P}), \tau\right)\right\}}\mathrm{d}\tau\end{aligned}$$

The two bounds are incomparable in general.

Proof of Theorem 3.5.13

Sketch of proof. First we divide the function f into two parts $f\mathbb{1}(F \leq \sqrt{n}a(\delta))$ and $f\mathbb{1}(F > \sqrt{n}a(\delta))$. The second, we can obtain

$$\mathrm{E}\|\nu_n \mathbb{1}(F > \sqrt{n}a(\delta))\|_{\mathcal{F}}^{*} \leq 2\sqrt{n}\mathrm{P}(F\mathbb{1}(F > \sqrt{n}a(\delta))).$$

We can now assume that every $f \in \mathcal{F}$ is bounded by $\sqrt{n}a(\delta)$. We combine two devices: a chaining argument and maximal inequalities for finite maxima.

A chaining argument Define two indicator functions $A_k f$ and $B_k f$ and decompose f as

$$f - \pi_q f = \sum_{k=q+1}^{\infty} (f - \pi_k f) B_k f + \sum_{k=q+1}^{\infty} (\pi_k f - \pi_{k-1} f) A_{k-1} f$$

where $\mathrm{P}|f - \pi_k f|^2 \leq (2^{-k})^2$.

Proof of Theorem 3.5.13

Lemma 3.5.12 (Maximal inequality for finite maxima). Let X, $X_i, i = 1, \dots, n$, be independent S-valued random variables with common probability law P, and let $f_1, \dots, f_N$ be measurable real functions on S such that $\max_{1\le r\le N} \|f_r - \mathrm{P}f_r\|_\infty \le c < \infty$ and $\sigma^2 = \max_{1\le r\le N} \mathrm{var}(f_r(X))$. Then

$$\mathrm{E}\left[\max_{1\le r\le N}\left|\sum_{i=1}^n (f_r(X_i) - \mathrm{P}f_r)\right|\right] \le \sqrt{2n\sigma^2\log(2N)} + \frac{c}{3}\log(2N)$$

Applying Lemma 3.5.12 Let $N_k := \log N_{[]}(\mathcal{F}, L^2(\mathrm{P}), 2^{-k})$

$$\begin{aligned}
\mathrm{E}\left\|\sum_{k=q+1}^\infty \nu_n\left((f-\pi_k f)B_k f\right)\right\|_{\mathcal{F}}^* &\le \sum_{k=q+1}^\infty \mathrm{E}\left\|\nu_n\left((f-\pi_k f)B_k f\right)\right\|_{\mathcal{F}}^* \\
&\lesssim \sum_{k=q+1}^\infty \left[a_{k-1}\log(2N_k) + 2^{-k}\sqrt{\log(2N_k)} + 2^{-2k+2}/a_k\right] \\
&\lesssim \sum_{k=q+1}^\infty 2^{-k}\sqrt{\log(2N_k)} \\
&\le 2\int_0^{2^{-(q+1)}} \sqrt{\log\left(2N_{[]}(\mathcal{F}, L^2(\mathrm{P}), \epsilon)\right)}d\tau.
\end{aligned}$$

We can also bound $\sum_{k=q+1}^\infty (\pi_k f - \pi_{k-1} f)A_{k-1}$ and $\pi_q f$.

If the class $\mathcal{F}$ is uniformly bounded, then the bound in Theorem 3.5.13 can be improved.

Theorem 3.5.15 Assume that $\|F\|_\infty < \infty$ and $\mathrm{P}f^2 \leq \delta$ for any $f \in \mathcal{F}$. Then

$$\mathrm{E}\,\|\nu_n\|_{\mathcal{F}}^* \leq J_{[]}(\delta, \mathcal{F}, L^2(\mathrm{P}))\left(1 + \frac{J_{[]}(\delta, \mathcal{F}, L^2(\mathrm{P}))}{\delta^2\sqrt{n}}\|F\|_\infty\right) \tag{2}$$

where we denote

$$J_{[]}(\delta, \mathcal{F}, L^2(\mathrm{P})) = \int_0^{2\delta} \sqrt{\log(2N_{[]}(\mathcal{F}, L^2(\mathrm{P}), \tau))}\mathrm{d}\tau.$$

Example: Monotone functions

Proposition 3.5.17. Let $\mathcal{F}$ be the class of monotone functions $f : \mathbb{R} \to [a, b]$. Then there is an universal constant $A > 0$ such that

$$\log N_{[]}(\mathcal{F}, L^p(\mathrm{P}), \epsilon) \leq A\epsilon^{-1},$$

for every $p \geq 1$, $\epsilon > 0$ and probability measure P on $\mathbb{R}$.

Application to density estimation (Example 3.4.5 of [3]). Suppose that the observations are sampled from a nonincreasing density on a compact interval in the real line and let $\mathcal{P}$ be the collection of such densities. Then the MLE $\hat{p}$ over suitable sieves satisfies

$$\sup_{p \in \mathcal{P}} \mathrm{E}\|\hat{p} - p\|_{L^2(p)} \lesssim n^{-1/3}.$$

Example: Smooth functions

Corollary 2.7.2 of [3]. Let $\mathcal{X}$ be a bounded, convex subset of $\mathbb{R}^d$ with nonempty interior. Let $\mathcal{H}^{\beta,M}(\mathcal{X})$ be the class of Hölder β-smooth functions whose Hölder norms are less than or equal to M. Then there is an universal constant $A > 0$ such that

$$\log N_{[]}(\mathcal{H}^{\beta,M}(\mathcal{X}), L^p(\mathrm{P}), \epsilon) \leq A\epsilon^{-d/\beta},$$

for every $p \geq 1$, $\epsilon > 0$ and probability measure P on $\mathbb{R}^d$

Application to binary classification [1]. Assume that $(\mathbf{x}, y) \sim \mathrm{P}$ where P is a distribution on $[0,1]^d \times \{-1,1\}$. Let $\eta(\mathbf{x}) = \mathrm{P}(y = 1|\mathbf{x})$. Then, the ERM classifier $\hat{f}$ over suitable sieves satisfies

$$\sup_{\eta \in \mathcal{H}^{\beta,M}([0,1]^d)} \mathrm{E}\left[\mathrm{P}(y\hat{f}(\mathbf{x}) < 0) - \mathrm{P}(y\eta(\mathbf{x}) < 0)\right] \lesssim n^{-1/(2+d/\beta)}.$$

The peeling method to derive rates

For a given loss function $\ell : \mathcal{F} \times \mathcal{X} \to \mathbb{R}_+$, we let

$$\mathcal{E}(f, f') := \mathrm{P}\{\ell(f, X) - \ell(f', X)\}, \quad \mathcal{E}_n(f, f') := \mathrm{P}_n\{\ell(f, X_i) - \ell(f', X_i)\}.$$

We assume that a function $f^\star := \operatorname{argmin}_f \mathrm{P}\ell(f, X)$ lies on the sieves $\mathcal{F}_n$ (i.e., no approximation error). Also assume that for any $f, f' \in \mathcal{F}_n$ (see [2]),

$$d^2(f, f^\star) \le \mathcal{E}(f, f^\star)$$
$$\mathrm{Var}(\ell(f, X) - \ell(f', X)) \le d^2(f, f').$$

We let

$$\mathcal{F}_{n,j} := \left\{ f \in \mathcal{F}_n : 2^{j-1}\epsilon_n \le d(f, f^\star) < 2^j \epsilon_n \right\}.$$

For the ERM (or ML) estimator $\hat{f}_n$, we have that

$$\begin{aligned}
\mathrm{P}\left(d(\hat{f}_n, f^\star) \ge \epsilon_n\right) &\le \mathrm{P}\left(\sup_{f : d(f, f^\star) \ge \epsilon_n} \mathcal{E}_n(f^\star, f) \ge 0\right) \\
&\le \sum_{j=1}^{\infty} \mathrm{P}\left(\sup_{f \in \mathcal{F}_{n,j}} \mathcal{E}_n(f^\star, f) - \mathcal{E}(f^\star, f) \ge \mathcal{E}(f, f^\star)\right) \\
&\le \sum_{j=1}^{\infty} \mathrm{P}\left(\sup_{f \in \mathcal{F}_{n,j}} \mathcal{E}_n(f^\star, f) - \mathcal{E}(f^\star, f) \ge 4^{j-1}\epsilon_n^2\right) \\
&\le \sum_{j=1}^{\infty} \frac{1}{4^{j-1}\epsilon_n^2} \mathrm{E}\left[\sup_{f \in \mathcal{F}_{n,j}} (\mathcal{E}_n(f^\star, f) - \mathcal{E}(f^\star, f))\right]
\end{aligned}$$

The peeling method to derive convergence rates

Since $\mathrm{Var}(\ell(f,X) - \ell(f',X)) \leq d^2(f,f')$,

$$\mathrm{E}\left[\sup_{f\in\mathcal{F}_{n,j}} \left(\mathcal{E}_n(f^\star,f) - \mathcal{E}(f^\star,f)\right)\right] \leq \frac{1}{\sqrt{n}} \int_0^{2^j\epsilon_n} \sqrt{\log\left\{2N_{[]}\left(\mathcal{L}_{n,j}, L^2(\mathrm{P}), \tau\right)\right\}}\mathrm{d}\tau$$

where $\mathcal{L}_{n,j} = \{\ell(f^\star) - \ell(f) : f \in \mathcal{F}_{n,j}\}$, and moreover,

$$N_{[]}(\mathcal{L}_{n,j}, L^2(\mathrm{P}), \tau) \leq N_{[]}(\mathcal{F}_{n,j}, L^2(\mathrm{P}), C\tau) \leq N_{[]}(\mathcal{F}, L^2(\mathrm{P}), C\tau).$$

Assume that

$$\log N_{[]}(\mathcal{F}, L^2(\mathrm{P}), \tau) \lesssim \tau^{-\rho}$$

for some $0 < \rho < 2$. Then

$$\begin{aligned}\int_0^{2^j\epsilon_n} \sqrt{\log\left\{2N_{[]}\left(\mathcal{L}_{n,j}, L^2(\mathrm{P}), \tau\right)\right\}}\mathrm{d}\tau &\lesssim \int_0^{2^j\epsilon_n} \tau^{-\rho/2}\mathrm{d}\tau \\ &= (2^j\epsilon_n)^{1-\rho/2}\end{aligned}$$

Hence,

$$\mathrm{P}\left(d(\hat{f}_n, f^\star) \geq \epsilon_n\right) \lesssim \epsilon_n^{-\rho/2-1}/\sqrt{n}$$

which yields the convergence rate

$$\epsilon_n \geq n^{-1/(2+\rho)} \log n$$

References I

[1] Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

[2] Pascal Massart. Concentration inequalities and model selection. 2007.

[3] Aad W Van Der Vaart and Jon A Wellner. *Weak convergence and empirical processes*. Springer, 1996.