Mathematical Foundations of Infinite-Dimensional Statistical Models

3.4 First Applications of Talagrand's Inequality

Gyuseung Baek

December 14, 2018

Introduction

- (A few) important results of Talagrand's Inequality related to EPs and U-statistics

- Moment Inequatlies for EPs
- Data-Driven Inequalities
- Inequality for U-statistics

Exercise 3.3.4

- $S_n = \sup_{f \in \mathcal{F}} |\sum_{k=1}^n f(X_k)|$ with X_i indep., $\mathcal{F}$: countable, $\forall f \in \mathcal{F}, ||f||_\infty \leq U/2$.
- (Note) $\mathcal{V}_n = 2UES_n + \sup_{f \in \mathcal{F}} \sum_{k=1}^n Ef^2(X_k)$
  Then

$$||S_n||_p \leq (1+\tau)ES_n + N_p^{1/2}(1+\delta)^{1/2}\mathcal{V}_n^{1/2} + \left[\frac{N_p^{2/p}(1+\delta)}{\tau} + 2E_p^{1/p}(1+\delta^{-1})\right] U \tag{1}$$

  for all $p > 1$ and $\delta, \tau > 0$, N_p, E_p : only related to p.

- For example, taking $\delta = \tau = 1$, we obtain

$$||S_n||_p \leq 2ES_n + \left(\frac{9p}{2}\right)^{1/(2p)} \sqrt{\frac{2p}{e}\mathcal{V}_n} + (9p)^{1/p}\frac{4}{e}pU \tag{2}$$

Goal

- Extend Exercise 3.3.4 to the classes with *unbounded envelope*
    - Combine Ex. 3.3.4 with Hoffmann-Jørgensen's Inequality (Thm 3.1.15)
      For each $p > 0$, if $Y_i, i \leq n\infty$ are indep., symmetric $SBC(T)$ processes, and if t_0 is defined as

      $$t_0 = \inf\{t > 0 : Pr(\{||S_n||_T > t\} \leq 1/8\},$$

      then

      $$||S_n||_p \leq 2^{(p+2)/p}(p+1)^{(p+1)/p}$$

- Similar result (very sharp)
  ξ_i : indep. centred r.v.s. then, for all $p \geq 2$ there exist $C, K < \infty$ s.t.

$$E\left|\sum_{i=1}^{n} \xi_i\right|^p \leq CK^p \left[p^p E \max_{i \leq n} |\xi_i|^p + p^{p/2}\left(\sum_{i=1}^{n} E\xi_i^2\right)^{p/2}\right] \quad (3)$$

Theorem 3.4.1 (Statement)

- $\mathcal{F}$: countable collection of measurable functions on $(S, \mathcal{S})$
- X_i : indep. S-valued variables s.t. $\mathcal{V}_n := \sup_{f \in \mathcal{F}} \sum_{i=1}^n Ef^2(X_i) < \infty$ and $Ef(X_i) = 0$ for all $i, f \in \mathcal{F}$.
- Set $F(\cdot) := \sup_{f \in \mathcal{F}} |f(\cdot)|$ and

$$S_n = \left|\left|\sum_{i=1}^n f(X_i)\right|\right|_{\mathcal{F}} \quad \text{and} \quad S_{n,M} = \left|\left|\sum_{i=1}^n \left(f(X_i) I_{F(X_i) \le M} - Ef(X_i) I_{F(X_i) \le M}\right)\right|\right|_{\mathcal{F}}$$

where $M > 0$ is a positive constant.
Then, for any $n \in \mathbb{N}$ and any $p > 1$,

$$\begin{aligned} ||S_n||_p \le 2ES_{n,M_p} &+ \left(\frac{9p}{2}\right)^{1/(2p)} \sqrt{\frac{2p}{e}\mathcal{V}_n} \\ &+ \left(\frac{4}{e}(72p)^{1/p} + 16(4p)^{1/p}\right) p|| \max_i F(X_i)||_p \end{aligned} \tag{4}$$

where $M_p^p = 8E \max_i F^p(X_i)$.

Theorem 3.4.1 (interpretation)

- In concrete situations, as with metric entropy expectation bounds for VC classes of functions, one may have as good an estimate for $ES_{n,M}$ as for ES_n.
- In general, $ES_{n,M} \leq 2ES_n$
    - If $f(X_i)$ are symmetric, then $ES_{n,M} \leq ES_n$

- (Remark 3.4.2) the coefficient 2 for ES_{n,M_p} can be replaced by $1+\delta$ at the expense of increasing other two summands from the bound for $||S_n||_p$.
- (4) can simplifires a bit by using the bound $p^{1/p} \leq e^{1/e}$
- In i.i.d. case, one can do a little better bound.

Note: Talagrand's Inequality

- Talagrand's Inequality bound consists of ES_n(center), σ^2(2nd moment), U(upper bound of function space).
- It gives an essentially best-possible rate, whereas, in general, the available bounds are much less precise.
- **It would be much more useful if these quantities could be replaced by data-dependent surrogates (or estimates).**

- σ^2 can be bounded by U and usually by much smaller quantities (c.f. density estimation).
- In this subsection, we replace ES_n by random surrogates, namely

$$\left\|\sum_{i=1}^{n} \epsilon_i f(X_i)\right\|_{\mathcal{F}} \quad \text{or} \quad E_\epsilon \left\|\sum_{i=1}^{n} \epsilon_i f(X_i)\right\|_{\mathcal{F}} \tag{5}$$

These are sometimes called *Rademacher complexities*.

Theorem 3.4.3 (state)

- $\mathcal{F}$: countable collection of m'sble ftns on $(S, \mathcal{S})$ w/ abs. bounded by $1/2$.
- $X_i, i \in \mathbb{N} \sim P$, i.i.d., S-valued.
- $\epsilon_i, i \in \mathbb{N}$: Rademacher seq. indep. from $\{X_i\}$ and $\sigma^2 \geq \sup_{f\in\mathcal{F}} Pf^2$.
  Then, for all $n \in \mathbb{N}$ and $x \geq 0$,

$$Pr\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Pf)\right\|_{\mathcal{F}} \geq 3\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right\|_{\mathcal{F}} + 4\sqrt{\frac{2\sigma^2 x}{n}} + \frac{70}{3n}x\right\} \leq 2e^{-x} \tag{6}$$

Theorem 3.4.3 (Proof)

- Set $S_n = \left\| \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Pf) \right\|_{\mathcal{F}}$ and $\tilde{S}_n = \left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right\|_{\mathcal{F}}$.
- Apply Talagrand's Inequality to both S_n and $\tilde{S}_n$
- For $\tilde{S}_n$, use the Klein-Rio version (3.111)
- For S_n, use Theorem 3.3.7
- Different δ produce different coefficients. ((6) - set $\delta = 1/5$)

- (Remark 3.4.4) Since Rademacher complexities are celf-bounding (Exercise 3.3.6), if we use $E_\epsilon \tilde{S}_n$ instead of $\tilde{S}_n$ then achieve better bound.

$$Pr\left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Pf) \right\|_{\mathcal{F}} \geq 3E_\epsilon \left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right\|_{\mathcal{F}} + 4\sqrt{\frac{2\sigma^2 x}{n}} + \frac{12}{n} x \right\} \leq 2e^{-x} \tag{7}$$

Theorem 3.4.5

- Same assumption with theorem 3.4.2 except $U = 1$, not $1/2$.
  Then, for all $n \in \mathbb{N}$ and $x \geq 0$,

$$Pr\left\{ \left\| \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Pf) \right\|_{\mathcal{F}} \geq 2\left\| \frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i) \right\|_{\mathcal{F}} + 3\sqrt{\frac{2x}{n}} \right\} \leq 2e^{-x} \tag{8}$$

- (Proof) If a class of functions $\mathcal{F}$ is bounded by 1, then when one replace X_i in $||\sum_{i=1}^{n}(f(X_i) - Pf)/n)||_{\mathcal{F}}$, the variable changes by at most $2/n$.
  It means these r.v. have bounded differences with constant $c^2 = 4/n$ and the same is true for $||\sum_{i=1}^{n}\epsilon_i f(X_i)/n||_{\mathcal{F}}$.
- Use theorem 3.3.14

Coparison between Thm 3.4.3 and Thm 3.4.5

- The smaller lower bound term, the better the inequality.
- Let $\mathcal{F}_h = \{y \to K((x-y)/h) : x \in \mathbb{R}\}$, where $K \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$. and a probability measure $dP(x) = f(x)dx$, where f is bounded and continuous. Then $U = ||K||_\infty$ and $\sigma^2 \leq ||f||_\infty ||K||_{L^2}^2 h \to 0$ as $h \to 0$.
  In this case, Theorem 3.4.3 is more adequate than Theorem 3.4.5.

U-Statistics

- X_i : indep. r.v.s in $(S, \mathcal{S})$ with repective laws P_i.
- $h_{ij} : S^2 \to \mathbb{R}$ s.t. $E|h_{ij}(X_i, X_j)| < \infty$ for all i, j.
  U_n is called *U-statistic* of order 2 if U_n has a form

$$U_n = \sum_{1 \le i < j \le n} h_{ij}(X_i, X_j) \tag{9}$$

- U-statistic is *canonical* if for all i, j and $x, y \in S$,

$$Eh_{ij}(X_i, y) = Ej_{ij}(x, X_j) = 0$$

U-Statistics

- (Hoeffding decomposition) If U_n is not canonical, it decomposes into a 'linear' term and a canonical U-statistic.
  $h_{ij} = h, h(x,y) = h(y,x)$ and X_i : i.i.d. then

$$2(U_n - EU_n) = \sum_{i \neq j} [h(X_i, X_j) - E_X(h(X, X_j) - E_X(h(X_i, X) + Eh(X_i, X_j)] + 2(n-1) \sum_{i=1}^{n} [E_X h(X_i, X) - Eh(X_i, X_j)]. \tag{10}$$

- The second term is a sum of independent r.v.s, and its tail probabilities assuming that h is bounded are well understood.

- Thus, to achieve a tail probability ineq. of U-statistics, we only need to know a tail probability ineq. of *canonical* U-statistics

Four parameters for tail inequality of canonical U-statistic

- Whereas Bernstein's ineq. is in terms of supreme norm and variance, for canonical U-statistics we need two more parameters about the matrix (h_{ij}).

$$A := \max_{i,j} ||h_{ij}||_\infty, \qquad C^2 := \sum_{j=2}^{n}\sum_{i=1}^{j-1} Eh_{ij}^2(X_i, X_j),$$

$$B^2 := \max\left\{ \max_j \left\| \sum_{i=1}^{j-1} E_i h_{ij}^2(X_i, x) \right\|_\infty, \max_i \left\| \sum_{j=i+1}^{n} E_j h_{ij}^2(x, X_j) \right\|_\infty \right\},$$

$$D := \sup\left\{ \sum_{j=2}^{n}\sum_{i=1}^{j-1} E(h_{ij}(X_i, X_j)\xi_i(X_i)\xi_j(X_j)) : \sum_{i=1}^{n-1} E\xi_i^2(X_i) \le 1, \sum_{j=2}^{n} \xi_j^2(X_i) \le 1 \right\}. \tag{11}$$

- If h is symmetric and X_i's are i.i.d,

$$A = ||h||_\infty, \quad C^2 = \frac{n(n-1)}{2} Eh^2(X1, X1), \quad B^2 = (n-1)||E_1 h^2(X_1, x)||_\infty$$

$$D := \frac{n}{2} \sup\left\{ E(h(X_1, X_2)\xi(X_1)\xi(X_2)) : E\xi^2(X_1) \le 1, \xi^2(X_1) \le 1 \right\} = \frac{n}{2} ||h||_{L^2 \to L^2} \tag{12}$$

Notations

- Let U_n be a canonical U-statistic. we can write U_n as

$$U_n = \sum_{j=2}^{n} \left(\sum_{i=1}^{j-1} h_{ij}(X_i, X_j) \right) =: \sum_{j=2}^{n} Y_j. \tag{13}$$

- Note that $E_j Y_j := E(Y_j | X_1, \cdots, X_{j-1}) = 0$, hence $\{U_k : k \geq 2\}$ is a **martingale** relative to the σ-algebras $\mathcal{G} = \sigma(X_1, \cdots, X_k), k \geq 2$
  - The martingale can be extended to $n = 0$ and $n = 1$ by taking $U_0 = U_1 = 0$ and $\mathcal{G}_0 = \{\emptyset, \Omega\}, \mathcal{G}_1 = \sigma(X_1)$.

Theorem 3.4.8

- U_n : canonical U-statistic, h_{ij} : uniformly bounded.
- A, B, C, D : defined on (11)
- For $\epsilon > 0$, define

$$\kappa(\epsilon) = 3/2 + 1/\epsilon, \qquad \eta(\epsilon) = \sqrt{2}(2 + \epsilon + \epsilon^{-1}),$$
$$\beta(\epsilon) = e(1 + \epsilon^{-1})^2 \kappa(\epsilon) + [\eta(\epsilon) \vee (1 + \epsilon)^2/\sqrt{2}], \tag{14}$$
$$\gamma(\epsilon) = [e(1 + \epsilon^{-1})^2 \kappa(\epsilon)] \vee (1 + \epsilon)^2/3.$$

Then, for all $\epsilon, u > 0$,

$$Pr\{U_n \geq 2(1 + \epsilon)^{3/2} C\sqrt{u} + \eta(\epsilon) Du + \beta(\epsilon) Bu^{3/2} + \gamma(\epsilon) Au^2\} \leq e^{1-u} \tag{15}$$

Lemma 3.4.6

- $(U_n, \mathcal{G}_n), n \geq 0$: martingale w.r.t. $\mathcal{G}_n$ s.t. $U_0 = U_1 = 0$.
- For each $n \geq 1, k \geq 2$, define the 'angle brackets' $A_n^k = A_n^k(U)$ by

$$A_n^k = \sum_{i=1}^{n} E[(U_i - U_{i-1})^k | \mathcal{G}_{i-1}]$$

  (and note $A_1^k = 0$ for all k).
- Suppose that for $\lambda > 0$ and all $i > 1$, $Ee^{\lambda |U_i - U_{i-1}|} < \infty$. Then

$$\left(\mathcal{E}_n := e^{\lambda U_n - \sum_{k=2}^{\infty} \lambda^k A_n^k / k!}, \mathcal{G}_n\right), n \in \mathbb{N} \tag{16}$$

  is a supermartingale.
- In particular, $E\mathcal{E}_n \leq E\mathcal{E}_1 = 1$, so, if $A_n^k \leq w_n^k$ for constants $w_n^k \geq 0$, then

$$Ee^{\lambda U_n} \leq e^{\sum_{k \geq 2} \lambda^k w_n^k / k!}$$

Lemma 3.4.6 for U-statistic

- If U_n is a canonical U-statistic, we have

$$A_n^k = \sum_{j=2}^{n} E_j \left[\sum_{i=1}^{j-1} h_{ij}(X_i, X_j) \right]^k \leq V_n^k = \sum_{j=2}^{n} E_j \left| \sum_{i=1}^{j-1} h_{ij}(X_i, X_j) \right|^k \qquad (17)$$

- Then, by duality (Exercise 3.4.1),

$$(V_n^k)^{1/k} = \sum_{\xi_j \in L^{k/(k-1)}(P): \sum_{j=2}^{n} E|\xi_j(X_j)|^{k/(k-1)}=1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E_j(h_{ij}(X_i, X_j)\xi_j(X_j)). \qquad (18)$$

- Thus, if we set suitable $\mathbf{X}_i$ and $\mathcal{F}$, we have

$$(V_n^k)^{1/k} = \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n-1} f(\mathbf{X}_i) \right|$$

Lemma 3.4.6 for U-statistic (Continue)

- Therefore, by Talagrand's Inequality, we obtain

$$Pr\left\{(V_n^k)^{1/k} \geq (1+\epsilon)E(V_n^k)^{1/k} + \sqrt{2\mathcal{V}_k x} + \kappa(\epsilon) b_k x\right\} \leq e^{-x} \tag{19}$$

for

$$\mathcal{V}_k = \sup_{\sum_{j=2}^n E|\xi_j(X_j)|^{k/(k-1)}=1} \sum_{i=1}^{n-1} E\left[\sum_{j=i+1}^{n} E_j(h_{ij}(X_i, X_j)\xi_j(X_j))\right]^2 \tag{20}$$

and

$$b_k = \sup_{\sum_{j=2}^n E|\xi_j(X_j)|^{k/(k-1)}=1} \max_i \sup_x |E_j(h_{ij}(X_i, X_j)\xi_j(X_j))| \tag{21}$$

Lemma 3.4.7

- For every $u \geq 0$, with $\mathcal{V}_k$ and b_k defined by (20) and (21), respectively, we have

$$Pr \bigcup_{k=2}^{\infty} \left\{ (V_n^k)^{1/k} \geq (1+\epsilon) E(V_n^k)^{1/k} + \sqrt{2\mathcal{V}_k j u} + \kappa(\epsilon) b_k k u \right\} \leq \frac{1+\sqrt{5}}{2} e^{-u}. \tag{22}$$