10-755: Advanced Statistical Theory I	**Fall 2017**
Lecture 23: Nov. 27	
Lecturer: Alessandro Rinaldo	*Scribes: Boyan Duan*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

23.1 Supreme of sub-Gaussian process

Theorem 23.1 (One step discretization bound) *Assume $\{X_\theta, \theta \in \mathbb{T}\}$ is a sub-Gaussian process w.r.t. d, then*

$$E[sup_{\theta,\theta' \in \mathbb{T}}|X_\theta - X_{\theta'}|] \leq inf_{\delta \in (0,D]}(2E[sup_{\gamma,\gamma' \in \mathbb{T}, d(\gamma,\gamma') \leq \delta}|X_\gamma - X_{\gamma'}|] + 4D\sqrt{logN(\mathbb{T},\delta)}).$$

Example 23.2

- *For $\mathbb{T} \subseteq \mathbb{R}$, then for radermacher of Gaussian complexities, we have $X_\gamma = \epsilon^T \delta$, then*

  $$E[sup_{\gamma,\gamma' \in \mathbb{T}, d(\gamma,\gamma') \leq \delta}|X_\gamma - X_{\gamma'}|] \leq \sqrt{n}\delta.$$

  So $\mathcal{R}_n(\tilde{\mathbb{T}}(\delta)) \lesssim min_{\delta \in (0,D]}\{\delta\sqrt{n} + D\sqrt{logN(\mathbb{T},\delta)}\}$

- *Non-parametric regression:*
  $\mathcal{F}_L = \{f : [0,1] \to \mathbb{R}, f(0) = 0, |f(x) - f(y)| \leq L(x-y)\forall x, y \in [0,1]\}$. Need to handle $\mathcal{G}_n(\mathcal{F}(x_1^n)) = E[sup_{f \in \mathcal{F}_L} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)]$, where $\mathcal{F}(x_1^n) = \{(f(x_1), \cdots, f(x_n)), f \in \mathcal{F}_L\} \subseteq \mathbb{R}^n$.
  Canonical metric is $||f - g||_n = \sqrt{\frac{1}{n}\sum(f(x_i) - g(x_i))^2}$. Here $X_f = \frac{1}{\sqrt{n}}\sum \epsilon_i f(x_i)$.
  $logN_2(\frac{\mathcal{F}(x_1^n)}{\sqrt{n}}, \delta) \leq logN_\infty(\mathcal{F}(x_1^n), \delta) \leq logN_\infty(\mathcal{F}_L, \delta)$. (Because $||f - g||_n \leq max_i|f(x_i) - g(x_i)|$).
  Note that $logN_\infty(\mathcal{F}_L, \delta) \asymp \frac{L}{\delta}, 0 < \delta \leq \delta_0$. Then $\mathcal{G}_n(\mathcal{F}(x_1^n)) \lesssim \frac{1}{\sqrt{n}}(\delta\sqrt{n} + \sqrt{\frac{L}{\delta}}), \forall \delta < \delta_0$.
  Balance the terms by choosing $\delta \asymp n^{-\frac{1}{3}}$, $\mathcal{G}_n(\mathcal{F}(x_1^n)) \lesssim n^{-\frac{1}{3}}$.
  Remark: this is not sharp. The optimal rate for non-parametric least squares is $n^{-\frac{2}{3}}$.

- *Wasserstein distance:*
  $E[W_1(P_n, P)] \lesssim n^{-\frac{1}{3}}$. Because in this case, $E[sup_{d(f,g)<\delta}|X_f - X_g|] \leq 2\delta$, the bound is $\delta + \sqrt{\frac{L}{\delta n}}$.
  Remark: this is also sub-optimal, one sharper rate is $n^{-\frac{1}{2}}$.

Remark *In one step discretization bound, we bound $E[max_i|X_{\theta_i} - X_{\theta_1}|]$ by $2\sqrt{D^2 logN(\mathbb{T},\delta)}$. In the following chaining bound, we obtain a sharper bound as $2\int_\delta^D \sqrt{logN(\mathbb{T},\mu)}d\mu$.*

Theorem 23.3 (Chaining bound) *Let $\{X_\theta, \theta \in \mathbb{T}\}$ is zero-mean, sub-Gaussian, separable process w.r.t. d on $\mathbb{T}$, then*

$$E[sup_{\theta,\theta' \in \mathbb{T}}|X_\theta - X_{\theta'}|] \leq 2E[sup_{\gamma,\gamma' \in \mathbb{T}, d(\gamma,\gamma') \leq \delta}|X_\gamma - X_{\gamma'}|] + 16\int_{\delta/4}^D \sqrt{logN(\mathbb{T},\mu)}d\mu$$

.

Remark

- *Constraint is not sharp.*
- *if $\delta = 0$, the integral term is $\int_0^D \sqrt{logN(\mathbb{T}, \mu)}d\mu$. (Known as Dudley's bound).*

Proof: Let $\mathbb{U} = \{\theta_1, \cdots, \theta_N\}$ be a minimal δ-cover of $\mathbb{T}$ w.r.t. d. Then,

$$sup_{\theta,\theta' \in \mathbb{T}} |X_\theta - X_{\theta'}| \leq 2sup_{\gamma,\gamma' \in \mathbb{T}, d(\gamma,\gamma') \leq \delta}(X_\gamma - X_{\gamma'}) + 2max_i|X_{\theta_i} - X_{\theta_1}|.$$

Now we bound the second terms:
For $m = 1, \cdots, L$ to be specified, let $\mathbb{U}_m$ be a $\delta_m = \frac{D}{2^m}$ minimal cover of $\mathbb{U}$ (where we are allowed to choose points for $\mathbb{U}_m$ from $\mathbb{T}$, not just from $\mathbb{U}$. Then $|\mathbb{U}_m| \leq N(\mathbb{T}, \delta_m)$.
Since $\mathbb{U}$ is finite, $\exists L$ s.t. $|\mathbb{U}_L| = |\mathbb{U}|$ and in this case: $\mathbb{U}_L = \mathbb{U}$. So L is the smallest integer s.t. $|\mathbb{U}_L| = |\mathbb{U}|$.
For each $m = 1, \cdots, L$, let $\pi_m : \mathbb{T} \to \mathbb{U}_m$ s.t. $\pi_m(\theta) = argmin_{\beta \in \mathbb{U}_m} d(\theta, \beta)$.
So for each $\theta \in \mathbb{U}$, we get a sequence $(\gamma_1, \cdots, \gamma_L)$, where $\gamma_L = \theta, \gamma_j = \pi_j(\gamma_{j+1}), j = 1, \cdots, L-1$.
*Remark: $(\gamma_1, \cdots, \gamma_L)$ can be viewed as a chain, with finer resolution to be close to θ.
Note for $\theta \in \mathbb{U}$, we can write $X_\theta - X_{\gamma_1} = \sum_{m=2}^{L}(X_{\gamma_m} - X_{\gamma_{m-1}})$. Thus,

$$|X_\theta - X_{\gamma_1}| \leq \sum_{m=2}^{L} max_{\beta \in \mathbb{U}_m} |X_\beta - X_{\mathbb{T}_{m-1}(\beta)}|.$$

Given another $\theta' \in \mathbb{U}$, we can construct another chain $(\gamma_1', \cdots, \gamma_L')$ and get the same bound for $|X_{\theta'} - X_{\gamma_1'}|$.
By triangle inequality:

$$|X_\theta - X_{\theta'}| \leq max_{\gamma,\gamma' \in \mathbb{U}_1} |X_\gamma - X_{\gamma'}| + 2\sum_{m=2}^{L} max_{\beta \in \mathbb{U}_m} |X_\beta - X_{\mathbb{T}_{m-1}(\beta)}|.$$

(To be continued) ■

References

[W17] M. Wainwright, "High-dimensional statistics: A Non-asymptotic Viewpoint. (Draft)," Chapter 5. 2017.