Math 547: Mathematical Foundations of Statistical Learning Theory Fall 2017

Lecture 15 — September 25

Instructor: Stas Minsker *Scribe: Mose Wintner*

15.1 Metric entropy and Dudley's theorem

$\{x(t),\, t \in T\}$ stochastic process with sub-Gaussian increments.

Theorem 1. (Dudley)

$$\mathbb{E}\sup_{t\in T}|x(t)-x(t_0)| \le c\int_0^\infty \sqrt{H(\varepsilon)}\,d\varepsilon.$$

Proof: Let D be the diameter of T, so the integral above can be considered as an integral from 0 to D. Let T_n be a smallest $D2^{-n}$ net for T, meaning that $T_n = \{t_i\}$ such that

$$T \subset \bigcup B(t_i, D2^{-n}).$$

The cardinality of T_n is $N(T, d, D2^{-n}) = N(D2^{-n})$. Let $\pi_n t$ be the projection of t onto T_n. $x(t) - x(t_0) = \sum \pi_{n+1}t - \pi_n t$. Note that T_n is not necessarily a subset of T.

$$\sup_{t\in T}|x(t)-x(t_0)| \le \sum_{n\ge 0}\sup_{t\in T}|x(\pi_{n+1}t)-x(\pi_n t)|.$$

$$\mathbb{E}\sup_{t\in T}|x(t)-x(t_0)| \le \sum_{n\ge 0}\mathbb{E}\sup_{t\in T}|x(\pi_{n+1}t)-x(\pi_n t)|.$$

Then

$$\mathbb{E}\sup_{t\in T}|x(\pi_{n+1}t)-x(\pi_n t)| = \mathbb{E}\max_{\substack{t_1\in T_n\\ t_2\in T_{n+1}\\ d(t_1,t_2)\le 3D2^{-(n+1)}}}|x(t_1)-x(t_2)|$$

$$\begin{aligned}
d(t_1,t_2) \le d(t_1,t) + d(t,t_2) &\le D2^{-n} + D2^{-(n+1)} = 3D2^{-n-1}.\\
\operatorname{card}(\{(t_1,t_2)|\, t_1\in T_n,\, t_2\in T_{n+1}\}) &\le N(D2^{-n})N(D2^{-n-1})\\
\log(N(D2^{-n})N(D2^{-n-1})) &\le 2H(D2^{-n-1}).
\end{aligned}$$

Putting this together and using the assumption that $x(t)-x(s)\in SG(d^2)$,

$$\mathbb{E}\sup_{t\in T}|x(t)-x(t_0)| \le 12\sqrt{2}\sum_{n\ge 0}D2^{-n-1}\sqrt{H(D2^{-n-1})} \le 12\sqrt{2}\int_0^\infty\sqrt{H(\varepsilon)}\,d\varepsilon$$

where the last inequality comes from a Riemann sum bound.

Theorem 2. (Sudakov's minoration)
Let $\{x(t)\}$ be the Gaussian process with $d(t,s) = \sqrt{\mathrm{Var}(x(t)-x(s))}$. Then

$$\mathbb{E}\sup_{t\in T}|x(t)-x(t_0)| \geq c\sup_{\varepsilon>0}\varepsilon\sqrt{H(T,d,\varepsilon)}.$$

Theorem 3. (Talagrand)
There exists a numerical constant $c>0$ such that

$$\frac{1}{c}\gamma_2(T,d) \leq \mathbb{E}\sup_{t\in T}(x(t)-x(t_0)) \leq c\gamma_2(T,d).$$

Recall that the generic chaining complexity γ_2 is defined by

$$\gamma_2(T,d) := \inf_{T_n}\sup_{t\in T}\sum_{n\geq 0}2^{n/2}d(t,T_n).$$

15.1.1 Concentration inequalities

Recall Hœffding's inequality: let $x_1,\ldots,x_n$ be iid such that $\mathbb{E}x_j=0$ and $a_j \leq x_j \leq b_j$. Then

$$Pr\left(|\sum x_j| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum(b_j-a_j)^2}\right).$$

Next, we will prove Bernstein's inequality, which is much better when

$$\mathrm{Var}(\sum X_j) << \sum(b_j-a_j)^2.$$

Theorem 4. (Bernstein's inequality)
Let $X_1,\ldots,X_n$ be iid random variables with $\mathbb{E}X_j=0$, and $|X_j|\leq M$ almost surely, and let $B_n^2 = \sum \mathrm{Var}X_j$. Then for all $t>0$,

$$Pr\left(|\sum X_j| \geq t\right) \leq 2\exp\left(-\frac{t^2/2}{B_n^2+Mt/3}\right).$$

Proof: Case (a): $B_n^2 \geq Mt/3 \Leftrightarrow t \leq 3B_n^2/M$. When the X_j are iid, $t \leq 3n\sigma^2/M$, so

$$Pr\left(|\sum X_j| \geq t\right) \leq 2\exp\left(\frac{-t^2}{4B_n^2}\right)$$

by Hœffding's inequality. Case (b): $B_n^2 < Mt/3 \Leftrightarrow t > 3B_n^2/M$, then

$$Pr\left(|\sum X_j| \geq t\right) \leq 2\exp\left(\frac{-3t}{4M}\right).$$

Putting these together,

$$Pr(|X_j| \geq t) \leq 2\max\left(e^{-t^2/4B_n^2},\, e^{-3t/M}\right).$$