Math 547: Mathematical Foundations of Statistical Learning Theory Fall 2017

Lecture 12 — September 18

Instructor: Stas Minsker *Scribe: Mose Wintner*

12.1 Sub-Gaussian random variables, continued

$X \in SG(\sigma^2)$ if the MGF of X admits the bound

$$\mathbb{E}e^{\lambda X} \le e^{\lambda^2\sigma^2/2}.$$

Example 1. Assume that $a \le x \le b$ almost surely and assume that $\mathbb{E}X = 0$. Then $X \in SG\left(\frac{(b-a)^2}{4}\right)$.

Proof:

$$\mathbb{E}e^{\lambda X} \le \frac{b}{b-a}e^{\lambda a} + \frac{-a}{b-a}e^{\lambda b}.$$

Let $p = \frac{b}{b-a}$, $(1-p) = \frac{-a}{b-a}$, noting that $\mathbb{E}X = 0$ forces $a \le 0 \le b$. Let $h = \lambda(b-a)$. Then

$$\mathbb{E}e^{\lambda X} \le pe^{-h(1-p)} + (1-p)e^{ph} = e^{ph}((1-p) + pe^{-h}) =: e^{F(h)},$$

where

$$F(h) = ph + \log(1 - p + e^{-h}).$$

We have that

$$\begin{aligned}
F'(h) &= p + \frac{-pe^{-h}}{1-p+pe^{-h}}, \quad F'(0) = 0 \\
F''(h) &= \frac{pe^{-h}(1-p+pe^{-h}) - (pe^{-h})^2}{(1-p+pe^{-h})^2} \\
&= \frac{pe^{-h}}{1-p+pe^{-h}} - \left(\frac{pe^{-h}}{1-p+pe^{-h}}\right)^2 \\
&= \frac{pe^{-h}}{1-p+pe^{-h}}\left(1 - \frac{pe^{-h}}{1-p+pe^{-h}}\right) \\
&= z(1-z) \\
&\le 1/4
\end{aligned}$$

for $z \in [0,1]$. Therefore from Taylor's,

$$F(h) \le \frac{h^2}{8} = \frac{\lambda^2(b-a)^2}{8},$$

so

$$e^{F(h)} \le e^{\frac{\lambda^2(b-a)^2}{2\cdot 4}}.$$

Theorem 1. Assume that $X \in SG(\sigma^2)$. Then

$$\begin{aligned} P(X > t) &\leq e^{-t^2/2\sigma^2} \\ P(X < -t) &\leq e^{-t^2/2\sigma^2} \end{aligned}$$

for any $t \geq 0$.
Proof:

$$\begin{aligned} P(X > t) &= P(\lambda X > \lambda t) \\ &= P(e^{\lambda X} > e^{\lambda t}) \\ &\leq \mathbb{E} e^{\lambda X} e^{-\lambda t} \end{aligned}$$

Letting $Z = e^{\lambda X}$, $P(Z > s) \leq \mathbb{E}Z/s$ by Chebyshev's inequality for nonnegative r.v. Z. Then

$$P(X \geq t) \leq e^{\lambda^2\sigma^2/2 - \lambda t}.$$

Minimizing $\lambda^2\sigma^2/2 - \lambda t$ with calculus gives $\lambda_* = t/\sigma^2$. Then $P(X > t) \leq e^{-t^2/2\sigma^2}$.

Corollary 1. For X sub-Gaussian, $P(|X| > t) \leq 2e^{-t^2/2\sigma^2}$. Use that $-X$ is also sub-Gaussian. Exercise: if this is true and $\mathbb{E}X = 0$ then X is sub-Gaussian with parameter $c\sigma^2$.

Lemma 1. Let $X_1 \in SG(\sigma^2)$, $X_n \in SG(\sigma_n^2)$ be independent. Then $\sum X_j \in SG(\sum \sigma_j^2)$.
Proof:

$$\mathbb{E} e^{\lambda \sum X_j} = \mathbb{E} \prod e^{\lambda X_j} = \prod \mathbb{E} e^{\lambda X_j} \leq e^{\frac{\lambda^2}{2} \sum \sigma_j^2}.$$

Theorem 2. (Wassily) Hœffding's theorem. Let $X_1, \ldots X_n$ be independent random variables such that $a_j \leq X_j - \mathbb{E}X_j \leq b_j$ almost surely for all j. Then

$$P\left(|\sum_j (X_j - \mathbb{E}X_j)| > t \right) \leq e^{-\frac{2t^2}{\sum (b_j - a_j)^2}}.$$

Follows from lemma and corollary: $\sum X_j - \mathbb{E}X_j \in SG(\Sigma^2)$, where $\Sigma^2 = \sum (b_j - a_j)^2/4$.

Let $X_1, \ldots X_n$ be a sequence of random variables.

Problem 1. How to estimate $\mathbb{E} \max X_j$?

Lemma 2. Let $X_1 \in SG(\sigma_1^2), \ldots, X_n \in SG(\sigma_n^2)$ (not necessarily independent). Then there exist random variables for which this bound is sharp:

$$\mathbb{E} \max_{j=1,\ldots,n} X_j \leq \sqrt{2 \log n}\sigma,$$

where $\sigma = \max \sigma_j$.
Proof: Let $\lambda > 0$. $e^{\lambda t}$ is convex, so we can apply Jensen's inequality, which says for a convex function

$$f(\mathbb{E}X) \leq \mathbb{E}f(X).$$

Therefore

$$e^{\lambda \mathbb{E} \max X_j} \leq \mathbb{E} e^{\lambda \max X_j} = \mathbb{E} \max e^{\lambda X_j} \leq \mathbb{E} \sum e^{\lambda X_j} \leq \mathbb{E} \sum_{j=1}^{n} e^{\lambda^2 \sigma_j^2/2} \leq n e^{\lambda^2 \sigma^2/2}.$$

Taking logarithms,

$$\mathbb{E} \max X_j \leq \frac{\log n}{\lambda} + \frac{\lambda \sigma^2}{2}.$$

Minimizing over $\lambda > 0$, we get

$$\lambda_* = \frac{1}{\sigma} \sqrt{2 \log n},$$

which we plug in to get the bound

$$\sqrt{\log n} \frac{\sigma}{\sqrt{2}} + \sqrt{2 \log n} \frac{\sigma}{2} = \sqrt{2 \log n} \sigma.$$

Corollary 2. $\mathbb{E} \max |X_j| = \mathbb{E} \max(\max_j X_j, \max_j -X_j) \leq \sigma \sqrt{2 \log 2n}$.

Problem 2. What if n is very large or ∞ (for example, Brownian motion)?
Sharp answer: "chaining method", i.e. approximate it by a mesh plus some difference.