36-710: Advanced Statistical Theory	**Fall 2018**

Lecture 3: September 5

Lecturer: Alessandro Rinaldo *Scribes: Heejong Bong*

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

3.1 Review: sub-Gaussian random variables

A random variable X is sub-Gaussian with a variance factor σ^2, denoted $X \in \mathcal{SG}(\sigma^2)$, if

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\lambda^2\sigma^2/2}, \forall \lambda \in \mathbb{R}$$

where $\mu = \mathbb{E}[X]$.

Remark 3.1

1) We always center X.

2) If $X \in \mathcal{SG}(\sigma^2)$ then mgf of $X - \mu$ is uniformly bounded by mgf of $\mathcal{N}(0, \sigma^2)$.

3) $X \in \mathcal{SG}(\sigma^2) \iff -X \in \mathcal{SG}(\sigma^2)$

4) $X \in \mathcal{SG}(\sigma^2) \iff \mathbb{P}(|X - \mu| \leq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$ (: another way to define $\mathcal{SG}$)
: interesting if t is large.

3.2 Properties of $\mathcal{SG}$

1) $X \in \mathcal{SG}(\sigma^2) \implies Var[X] \leq \sigma^2$
In fact, $Var[X] \leq \sigma^2(X)$ where $\sigma^2(X) := \inf\{\sigma^2 : \mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\lambda^2\sigma^2/2}\}$
Proof:
By Taylor expansion and dominated convergence theorem,

$$\mathbb{E}[e^{t(X-\mu)}] \leq e^{\lambda^2\sigma^2/2}, \forall \lambda,$$

and hence

$$1 + \lambda\mathbb{E}[X - \mu] + \frac{\lambda^2}{2}\mathbb{E}[(X - \mu)^2] + o(\lambda^2) \leq 1 + \frac{\lambda^2\sigma^2}{2} + o(\lambda^2), \forall \lambda.$$

As $\mathbb{E}[X - \mu] = 0$,

$$Var[X] \leq \sigma^2.$$

■

2) If $a \leq X - \mu \leq b$ a.e. for $-\infty < a < b < \infty$ then $X \in \mathcal{SG}((\frac{b-a}{2})^2)$.
   i.e., bounded random variables are sub-Gaussians.
   Proof:
   w.l.o.g., assume $\mu = 0$. We will show that

$$\psi(\lambda)(:= \log \mathbb{E}[e^{\lambda X}]) \leq \frac{(b-a)^2\lambda^2}{8}, \forall \lambda \in \mathbb{R}\text{: Hoeffding's bound}$$

   First, notice that $Var[Z] \leq \frac{(b-a)^2}{2}$.
   Next, for any $\lambda \in \mathbb{R}$, define a new random variable Z_λ with probability distribution P_λ. s.t.

$$\frac{dP_\lambda}{dP_X}(z) = e^{\lambda z} e^{-\psi(\lambda)}, z \in [a, b]$$

   Note that $\frac{dP_\lambda}{dP_X}$ is a density.
   Now, $a \leq Z_\lambda \leq b$ a.e., and

$$Var[Z_\lambda] = \psi''(\lambda)$$

   Hence,

$$\psi''(\lambda) \leq (\frac{b-a}{2})^2, \forall \lambda \in \mathbb{R}.$$

   Since $\psi(0) = \psi'(0)(= \mathbb{E}[X]) = 0$,

$$\begin{aligned}\psi(\lambda) &= \int_0^\lambda \psi'(\mu)d\mu = \int_0^\lambda \int_0^\mu \psi''(\omega)d\omega d\mu \\ &\leq \int_0^\lambda \int_0^\mu (\frac{b-a}{2})^2 d\omega d\mu \leq \frac{(b-a)^2}{4}\frac{\lambda^2}{2} = \frac{(b-a)^2\lambda^2}{8}\end{aligned}$$

$\blacksquare$

3) $X \in \mathcal{SG}(\sigma^2) \implies \alpha X \in \mathcal{SG}(\alpha^2\sigma^2), \forall \alpha \in \mathbb{R}$.
4) $X \in \mathcal{SG}(\sigma^2), Y \in \mathcal{SG}(\tau^2) \implies X + Y \in \mathcal{SG}((\sigma+\tau)^2)$.
   Moreover, if $X \perp Y$, $X + Y \in \mathcal{SG}(\sigma^2 + \tau^2)$.
   Proof:
   If $X \perp Y$, trivial.
   If not, assume $\mathbb{E}[X] = \mathbb{E}[Y] = 0$.

$$\begin{aligned}\mathbb{E}[e^{\lambda(X+Y)}] &= \mathbb{E}[e^{\lambda X}e^{\lambda Y}] \\ &\leq \mathbb{E}[e^{\lambda X\frac{\sigma+\tau}{\sigma}}]^{\frac{\sigma}{\sigma+\tau}}\mathbb{E}[e^{\lambda Y\frac{\sigma+\tau}{\tau}}]^{\frac{\tau}{\sigma+\tau}}\text{by Hölder's inequality} \\ &\leq (\exp[\frac{1}{2}\lambda^2(\sigma+\tau)^2])^{\frac{\sigma}{\sigma+\tau}}(\exp[\frac{1}{2}\lambda^2(\sigma+\tau)^2])^{\frac{\tau}{\sigma+\tau}} \\ &= \exp[\frac{1}{2}\lambda^2(\sigma+\tau)^2]\end{aligned}$$

$\blacksquare$

   In H.W. 1, we will show that a similar result applies to $\sum_{i=1}^n X_i$ where $X_i \in \mathcal{SG}(\sigma_i^2)$: not necessarily independent.
   : use generalized Hölder's inequality.

3.3 Hoeffding's inequality

Let $X_1, \cdots, X_n$ be independent random variables s.t.

$$X_i \in \mathcal{SG}(\sigma_i^2), \forall i = 1, \cdots, n.$$

Then,

$$\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i)| \geq t) \leq 2\exp[-\frac{n^2t^2}{2(\sum_i \sigma_i^2)}]$$

because $\sum(X_i - \mu_i) \in \mathcal{SG}(\sum_i \sigma_i^2)$.

If $\sigma_i^2 = \sigma^2, \forall i$ then

$$\mathbb{P}(|\frac{\sum_i(X_i - \mu_i)}{n}| \geq t) \leq 2\exp[-\frac{nt^2}{2\sigma^2}].$$

Without independence,

$$\mathbb{P}(|\frac{\sum_i(X_i - \mu_i)}{n}| \geq t) \leq 2\exp[-\frac{n^2t^2}{2(\sum_i \sigma_i)^2}].$$

Example 3.2 *Suppose that $X_1, \cdots, X_n$ are independent, and $X_i \sim$ Bernoulli$(p_i), p_i \in (0,1)$.*

Then, $X_i \in \mathcal{SG}(\frac{1}{4}), \forall i$.

By Hoeffding's inequliaty,

$$\mathbb{P}(|\bar{X}_n - \bar{p}_n| \geq t) \leq 2\exp[-2nt^2]$$

where $\bar{X}_n = \frac{1}{n}\sum_i X_i, \bar{p}_n = \frac{1}{n}\sum_i p_i$.

With probability at least $1-\delta, \delta \in (0,1)$,

$$|\bar{X}_n - \bar{p}_n| \leq \sqrt{\frac{1}{2n} log(\frac{1}{\delta})}.$$

In particular, setting $\delta = \frac{1}{n^c}$, for some $c > 0$,

$$log(\frac{1}{\delta}) = c\log n.$$

Then,

$$|\bar{X}_n - \bar{p}_n| \leq O(\sqrt{\frac{\log n}{n}}), w.p. \geq 1 - \frac{1}{n^c}.$$

By CLT, $\bar{X}_n - \bar{p}_n = O_p(\frac{1}{\sqrt{n}})$ where $X_n = O_p(r_n)$ if $\forall \epsilon > 0, \exists M = M(\epsilon) and n_o = n_o(\epsilon, M)$ s.t.

$$\mathbb{P}(|X_n| \geq Mr_n) \leq \epsilon$$

for $n \geq n_0$.

Remark 3.3 *(Warning)*

Hoeffding's inequality is a great off-the-shell concentration inequality, and it can be sharp in some cases.

e.g., Rademacher random variable:

$$X = \begin{cases} -1 & w.p.\ \frac{1}{2} \\ 1 & w.p.\ \frac{1}{2} \end{cases}$$

Then,

$$Var(X) = 1 = \text{(variance of factor)}.$$

However, in most possible cases with constraints, it is no longer sharp. If you can, use Chernoff bound instead.

e.g., for $X_1, \cdots, X_n$: independent Bernoulli(p_i), you may want to use a multiplicative Chernoff bound:

$$\begin{cases} \mathbb{P}(\sum_i X_i \geq (1+\epsilon)\sum_i p_i) & \leq \begin{cases} \exp[-\frac{1}{3}\epsilon^2 \sum_i p_i], & \epsilon \in (0,1] \\ \exp[-\frac{1}{2}\epsilon^2 \sum_i p_i], & \epsilon > 1 \end{cases} \\ \mathbb{P}(\sum_i X_i \leq (1-\epsilon)\sum_i p_i) & \leq \exp[-\frac{\epsilon^2}{2}\sum_i p_i], \epsilon \in (0,1) \end{cases}$$

If $X_1, \cdots, X_n \overset{iid}{\sim}$ Bernoulli(p),

$$\text{Hoeffding: } \mathbb{P}(p - \frac{\sum_i X_i}{n} \geq t) \leq exp[-2nt^2] \qquad \to p - \bar{X}_n \leq \sqrt{\frac{1}{2n}\log(\frac{1}{\delta})}$$
$$\text{Multi. Chernoff: } \mathbb{P}(p - \frac{\sum_i X_i}{n} \geq \epsilon p) \leq exp[-np\epsilon^2/2] \qquad \to p - \bar{X}_n \leq \sqrt{\frac{2p}{n}log(\frac{1}{\delta})}$$

w.p. $\geq 1 - \delta$.

i.e., Chernoff bound is better if $p < \frac{1}{4}$ and better in terms of rates if $p \to 0$ as $n \to \infty$.

***Take home message**: Use Hoeffding if nothing else works; however, if there is more information which you can leverage, use other than Hoeffding bounds.*

3.4 Equivalent characterization of $\mathcal{SG}(\sigma^2)$

TFAE:

1) $\mathbb{E}[e^{\lambda X}] \leq e^{\lambda^2\sigma^2/2}, \forall \lambda \in \mathbb{R}$
2) $\mathbb{P}(|X| \geq t) \leq \sqrt{8}e\mathbb{P}(|Y| \geq t)$ where $Y \sim \mathcal{N}(0, 2\sigma^2)$
3) $\mathbb{E}[e^{a(\sigma)X^2}] \leq 2$ for some $a(\sigma)$ dependent to σ.

For more, see Vershynin's book or David Pollard's notes (to be posted).