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

Lecture 5: September 17

Lecturer: Alessandro Rinaldo *Scribes: Pratik Patil*

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

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

5.1 Last class

Recall that in the last lecture, we covered:

- Definition of sub-exponential random variables
- Some properties of sub-exponential random variables
- Sub-exponential tail bound
- A sufficient condition (Bernstein condition) that implies that a random variable is sub-exponential and corresponding concentration inequality (Bernstein inequality)

In this lecture, we will cover:

- Additivity property of sub-exponential random variables
- Examples of useful bounds resulting from the additivity property
- Bernstein versus Hoeffding inequalities
- Statistical applications that make use of the Bernstein inequality

5.2 Additivity property of sub-exponential random variables

As in the case of sub-Gaussian random variables, the sub-exponential property is preserved under addition of (not necessarily independent) sub-exponential random variables. The result in summarized the following lemma.

Lemma 5.1 (Additivity property) *If $X_i \in SE(\nu_i^2, \alpha_i)$, then*

$$\sum_{i=1}^n X_i - \mathbb{E}X_i \in \begin{cases} SE\left(\sum\limits_{i=1}^n \nu_i^2, \max\limits_{i=1,\dots,n} \alpha_i\right) & \textit{if } X_i \textit{ are independent} \\ SE\left(\left(\sum\limits_{i=1}^n \nu_i\right)^2, \max\limits_{i=1,\dots,n} \alpha_i\right) & \textit{if } X_i \textit{ are not independent} \end{cases}$$

We can specialize the sub-exponential tail bound for a normalized sum of exponential random variables as summarized the following lemma. For simplicity, we assume independence, but similar bound can be obtained for the general case using the appropriate transformed parameters.

Lemma 5.2 (Tail bound) *Assume X_is are independent, sub-exponential with parameters (ν_i^2, α_i), then we have this sub-exponential tail bound:*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X_i \geq t\right) \leq \begin{cases} \exp\left\{-\frac{nt^2}{2\nu_\star^2}\right\} & \text{if } 0 \leq nt \leq \frac{\nu_\star^2}{\alpha_\star} \text{ (small deviation regime, SG behaviour)} \\ \exp\left\{-\frac{nt}{2\alpha_\star}\right\} & \text{if } nt > \frac{\nu_\star^2}{\alpha_\star} \text{ (large deviation regime, SE behaviour)} \end{cases}$$

where $\nu_\star = \sqrt{\sum\limits_{i=1}^{n} \nu_i^2}$, $\alpha_\star = \max\limits_{i=1,\ldots,n} \alpha_i$.

We give some examples to illustrate the use of the sub-exponential tail bound.

Example 5.3 (χ^2-variables) *Let $Z_i \sim \mathcal{N}(0,1)$ be independent random variables. We can then obtain a tail bound for chi-squared random variable $\sum_{i=1}^n Z_i^2$ as*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i^2 - 1 \geq t\right) \leq \begin{cases} \exp\left\{-\frac{nt^2}{8}\right\} & \text{if } t \in (0,1) \\ \exp\left\{-\frac{nt}{8}\right\} & \text{if } t > 1 \end{cases}$$

Note that there are sharper results available for bounding deviations of a chi-squared random variable. We give one such result below from [LM00].

Lemma 5.4 (Laurent-Massart) *Let $a_1, ..., a_n$ be nonnegative reals. Define $||a||_\infty = \sup\limits_{i=1,\ldots,n} |a_i|$ and $||a||_2^2 = \sum\limits_{i=1}^{n} a_i^2$. Let $Z_i \sim \mathcal{N}(0,1)$ be independent random variables and $X = \sum\limits_{i=1}^{n} a_i(Z_i^2 - 1)$. Then, for any positive t:*

$$\mathbb{P}(X \geq 2||a||_2^2\sqrt{t} + 2||a||_\infty t) \leq \exp\{-t\}$$
$$\mathbb{P}(X \leq -2||a||_2\sqrt{t}) \leq \exp\{-t\}$$

Using this result, we obtain the following bound on chi-squared random variables.

Corollary 5.5 (Sharper bound) *Let X be chi-square with n degrees of freedom. Then, for any positive t:*

$$\mathbb{P}(X - n \geq 2\sqrt{nt} + 2t) \leq e^{-t}$$
$$\mathbb{P}(X - n \leq -2\sqrt{nt}) \leq e^{-t}$$

Finally, we consider concentration of high dimensional sub-Gaussian vectors.

Example 5.6 (Concentration of high-dimensional sub-Gaussian vectors) *Let $X = (X_1, ..., X_d) \in \mathbb{R}^d$ with independent SG(σ^2) random variables X_i such that $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[X_i^2] = 1$. Then, we have that $||X||_2 = \sqrt{\sum_{i=1}^d X_i^2}$ concentrates around $\sqrt{d}$.*

To show this result, we note that $X_i^2 - 1$ is SE with parameters $\nu_i^2 = K\sigma^4$ and $\alpha = K'\sigma^2$, for some constants K and K'. From the tail bound, we have that

$$\mathbb{P}\left(\left|\frac{||X||_2^2}{d} - 1\right| \geq t\right) \leq 2\exp\left\{-\frac{d}{2}\min\left\{\frac{t^2}{\nu^2}, \frac{t}{\alpha}\right\}\right\}. \tag{5.1}$$

Using the fact $|z-1| \geq c \implies |z^2 - 1| \geq \max\{c, c^2\}$ for $z \geq 0$ and $c < 0$, we get

$$\mathbb{P}\left(\left|\frac{||X||_2}{\sqrt{d}} - 1\right| \geq t\right) \leq \mathbb{P}\left(\left|\frac{||X||_2^2}{d} - 1\right| \geq \max\left\{t, t^2\right\}\right) \leq 2\exp\left\{-K''dt^2\right\}, \tag{5.2}$$

for some K'' that depends on K and K'. Letting $\mu = t\sqrt{d}$, we arrive at the required concentration inequality

$$\mathbb{P}\left(\left|||X||_2 - \sqrt{d}\right| \geq \mu\right) \leq 2\exp\left\{-K''\mu^2\right\}. \tag{5.3}$$

5.3 Bernstein versus Hoeffding inequalities

Assume $|X| < b$ *a.e.* and $Var[X] = \sigma^2$. Assume $\mathbb{E}[X] = 0$. We can apply both Hoeffding-type bound and Bernstein-type bound to obtain

$$\mathbb{P}[|X| \geq t] \leq \begin{cases} 2\exp\left\{-\frac{t^2}{2b^2}\right\} & \text{Hoeffding inequality} \\ 2\exp\left\{-\frac{t^2}{2(\sigma^2 + bt)}\right\} & \text{Bernstein inequality} \end{cases}$$

Bernstein is always sharper than Hoeffding because $\sigma^2 = \mathbb{E}[X^2] \leq b^2$. It is substantially better if $\sigma^2 \ll b^2$. Notice that Hoeffding always assumes the worst variance by using the range to bound the second moment. However, if we don't know any information about the variance, we might only be able to use Hoeffding.

Remark 5.7 *Even if sharper than Hoeffding, Bernstein is not the sharpest for a bounded random variable. Bennett's inequality can be used to provide sharper control on the tails [B62].*

5.4 Statistical applications that use Bernstein inequality

5.4.1 Sample splitting in nonparametric regression

Suppose we have a dataset $\mathcal{D}_n = \{(X_i, Y_i),\ i = 1, ..., n\}$, where $(X_i, Y_i) \in \mathbb{R}^{d+1}$, $Y_i = m(X_i) + \epsilon_i$, $m : \mathbb{R}^d \to \mathbb{R}$, ϵ_i are mutually independent and independent of X_i, and $\mathbb{E}[\epsilon_i] = 0$. Let $\hat{m}_h$ be some estimator of m that depends on $\mathcal{D}_n$ and a tuning parameter h. The question is how do we choose h?

Given a collection $\mathcal{Q}_n$ of candidate values for h, an oracle estimator can choose $\hat{h}^\star$ to be the minimizer of $\int_{\mathbb{R}^d}(\hat{m}_h(x) - m(x))^2 d\mu(x)$, where μ is distribution of X, over $\mathcal{Q}_n$. To get a data-driven choice, we can use sample splitting as follows:

1. Split the sample into training data $\{(X_i, Y_i),\ i = 1, ..., \lfloor \frac{n}{2} \rfloor\}$ and test data $\{(X_i, Y_i),\ i = \lfloor \frac{n}{2} \rfloor + 1, ..., n\}$.
2. Fit models $\{\hat{m}_h, h \in \mathcal{Q}_n\}$ only on the training data.
3. Chose $\hat{h}$ to be the minimizer of the quantity $\frac{2}{n}\sum_{i=\lfloor \frac{n}{2} \rfloor + 1}^{n}(\hat{m}_h(X_i) - Y_i)^2$.

We can ask how well does this data-driven procedure compare with the oracle procedure? The following theorem quantifies such comparison.

Theorem 5.8 ([GKKW06]) *Assume that* $|Y| \leq L$ *and* $\max_{h \in \mathcal{Q}_n} ||\hat{m}_h||_\infty \leq L$ *a.e. Then, for all* $\delta > 0$*,*

$$\mathbb{E}\left[\int_{\mathbb{R}^d} (\hat{m}_{\hat{h}}(x) - m(x))^2 d\mu(x)\right] \leq (1+\delta)\mathbb{E}\left[\int_{\mathbb{R}^d} (\hat{m}_{\hat{h}^\star}(x) - m(x))^2 d\mu(x)\right] + C(\delta, L)\frac{1 + \log|\mathcal{Q}_n|}{n/2}, \tag{5.4}$$

where $C(\delta, L) = L^2(\frac{16}{\delta} + 35 + 19\delta)$*.*

The proof uses Bernstein inequality (which retrieves n under $1 + \log|\mathcal{Q}_n|$ while $\sqrt{n}$ with Hoeffding) and a majorization argument.

5.4.2 Distribution-free confidence intervals

Suppose we have tail bound of the form

$$\mathbb{P}\left(|X - \mu| \geq t\right) \leq a \exp\left\{\frac{-nbt^2}{c + dt}\right\}. \tag{5.5}$$

Then, we can invert the bound to obtain a high probability statement that

$$|X - \mu| \leq \sqrt{\frac{c}{nb}\log\frac{a}{\delta} + \frac{d}{nb}\log\frac{a}{\delta}} \tag{5.6}$$

which holds with probability at least $(1 - \delta)$ for some $\delta \in (0, 1)$, typically a small number.

5.5 Next class

In summary, in this lecture, we covered:

- Additivity property of sub-exponential random variables
- Examples of useful bounds resulting from the additivity property
- Bernstein versus Hoeffding inequalities
- Statistical applications that make use of the Bernstein inequality

In the next lecture, we will cover:

- Maximal inequalities
- Review of martingales
- Azuma-Hoeffding inequality
- Bounded difference inequality

References

[LM00] B. Laurent and P. Massart ”Adaptive estimation of a quadratic functional by model selection,” *Annals of Statistics*, 2000.

[B62] G. Bennett, ”Probability inequalities for the sum of independent random variables,” *Journal of the American Statistical Association*, 1962.

[GKKW06] L. Gyorfi, M. Kohler, A. Krzyzak, and H. Walk, ”A distribution-free theory of non-parametric regression,” *Springer Science and Business Media*, 2006.