36-755: Advanced Statistical Theory **Fall 2016**

Lecture 4: September 12

Lecturer: Alessandro Rinaldo *Scribe: Xiao Hui Tai*

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.*

4.1 Last time

4.1.1 Sub-exponential random variables

The class of sub-exponential variables is larger than sub-Gaussian, and captures variables with longer tails. For example the χ^2 distribution has a left tail that is Gaussian-like, but a longer right tail.

4.1.2 Bernstein's inequality

We can get better bounds using Bernstein's inequality. Hoeffding's is sharp only if the variance is maximal. For example, if $|X - \mathbb{E}[X]| \leq b$ a.e. and $\sigma^2 = \mathbb{V}[X]$, we get

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq 2e^{-\frac{t^2}{2(\sigma^2+bt)}} \text{ using Bernstein's, and}$$
$$\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq 2e^{-\frac{t^2}{2b^2}} \text{ using Hoeffding's.}$$

If $t \ll \sigma^2$, Bernstein gives sub-Gaussian tails with parameter σ^2, as opposed to the parameter b^2 using Hoeffding's. Since $\sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] \leq b^2$, we get an improvement in some range of values of t.

In general we use Hoeffding's because of its simplicity, unless we need more refined bounds. We should use Bernstein's inequality especially if $\mathbb{V}[X] \leq c\mathbb{E}[X]$, and in the literature this is sometimes referred to as Bernstein's condition (note that this is different from the Bernstein's condition introduced in the previous lecture).

For an example of Bernstein giving better bounds than Hoeffding, refer to Theorem 7.1 in [GKKW02].

Remark If $X \in SG(\sigma^2)$, then $X^2 \in SE(\nu^2, \alpha)$, where $\nu = \alpha = 16\sigma^2$. This is proven in Homework 2 Problem 4.

4.2 Applications

4.2.1 Maxima

4.2.1.1 Expectation

Let $X_1, ..., X_n$ be random variables, not necessarily independent, such that $\log(\mathbb{E}[e^{\lambda X_i}]) \leq \psi(\lambda)$, where $\lambda \in [0, b)$, $0 < b \leq \infty$. Then

$$\mathbb{E}[\max_i X_i] \leq \inf_{\lambda \in [0,b)} \frac{\log n + \psi(\lambda)}{\lambda}.$$

Proof:

$$\begin{aligned} e^{\lambda \mathbb{E}[\max_i X_i]} &\leq \mathbb{E}[e^{\lambda \max_i X_i}] \text{ by Jensen's inequality} \\ &= \mathbb{E}[\max_i e^{\lambda X_i}] \\ &\leq \sum_{i=1}^{n} \mathbb{E}[e^{\lambda X_i}] \\ &\leq n e^{\psi(\lambda)} \text{ by assumption.} \end{aligned}$$

Taking logs, we have $\mathbb{E}[\max_i X_i] \leq \frac{\log n + \psi(\lambda)}{\lambda}$. We then pick λ to minimize the RHS. ■

Example If $\psi(\lambda) = \frac{\lambda^2 \sigma^2}{2}$ (sub-Gaussianity), then

$$\begin{aligned} \mathbb{E}[\max_i X_i] &\leq \frac{\log n}{\lambda} + \frac{\lambda \sigma^2}{2} \\ &\leq \sqrt{2\sigma^2 \log n}, \end{aligned}$$

where the second inequality is obtained by noting that $\frac{\log n}{\lambda}$ is decreasing in λ and $\frac{\lambda\sigma^2}{2}$ is increasing, so we balance them by setting $\frac{\log n}{\lambda} = \frac{\lambda\sigma^2}{2}$ and solving for λ. This gives $\lambda = \sqrt{\frac{2\log n}{\sigma^2}}$. Notice that the bound is on the order of $\sqrt{\log n}$, so the maximum does not grow very fast even if we have many observations.

The following is one way of obtaining a general result in the case of non-sub-Gaussian random variables. Such results can be used, for example, for sub-exponential random variables. This result is for reference and the proof is not stated here, but can be found in Section 2.5 of [BLM13]. In general, for any $u > 0$,

$$\inf_{\lambda \in (0,b)} \left\{ \frac{u + \psi(\lambda)}{\lambda} \right\} = \inf\{t \geq 0 : \psi^*(t) > u\},$$

where $\psi^*(t) = \sup_{\lambda \in (0,b)} \{\lambda t - \psi(\lambda)\}$. Using this, if $\psi(\lambda) = \frac{\lambda^2\sigma^2}{2(1-\lambda b)}$, where $\lambda \in (0, \frac{1}{b})$, $b > 0$, then

$$\mathbb{E}[\max_i X_i] \leq \sqrt{2\sigma^2 \log n} + b \log n$$

because $\psi^{*-1}(u) = \sqrt{2\sigma^2 u} + bu$, $u > 0$.

Example If $X_1, ..., X_n \sim \chi_p^2$, then $\mathbb{E}[\max_i X_i - p] \leq \sqrt{2p \log n} + 2 \log n$.

4.2.1.2 Probability

If $X_1, ..., X_n$ are random variables and $X_i \in SG(\sigma^2)$,

$$\begin{aligned}\mathbb{P}[\max_i X_i > t] &= \mathbb{P}[\bigcup_{i=1}^{n}\{X_i \geq t\}] \\ &\leq \sum_{i=1}^{n} \mathbb{P}[X_i \geq t] \\ &\leq ne^{-\frac{t^2}{2\sigma^2}} \\ &= e^{-\frac{t^2}{2\sigma^2}+\log n},\end{aligned}$$

where the first inequality is obtained using the union bound and the second using the sub-Gaussian property of X_i. If we want a t^* such that $\mathbb{P}[\max_i X_i > t^*] \leq \delta \in (0,1)$, then setting the RHS of the inequality above to δ, we get

$$t^* = \sqrt{2\sigma^2(\log\frac{1}{\delta} + \log n)}.$$

On the other hand, if we consider the X_i's individually, we get $\mathbb{P}[X_i > \sqrt{2\sigma^2 \log \frac{1}{\delta}}] \leq \delta$, and the two only differ by the $\log n$ term. The dimension of the problem only enters the bound logarithmically, which means that dealing with the maximum is only almost as difficult as dealing with individual random variables. In particular if we take $\delta = \frac{1}{n}$, we get $t^* = \sqrt{4\sigma^2 \log n}$, where the additional $\log n$ term only affects the constant.

Remark In the above derivation, the X_i's do not need to be independent. If we want to use independence we can do so using de Morgan's law instead of the union bound, but in this case the exponential decay is so strong that there is not much of a difference between the two.

4.2.2 Quadratic Forms

If X is a random vector of length d, and A is a d x d symmetric matrix, X^TAX is known as a quadratic form in X, and $\mathbb{E}[X^TAX] = \text{tr}(A\Sigma) + \mu^TA\mu$, where $\mu = \mathbb{E}[X]$ and $\Sigma = \mathbb{V}[X]$.

Now, we assume that A is a d x d symmetric positive definite matrix, $\mu = 0$, and (without loss of generality[1]) $\Sigma = I$ (i.e. $X \sim N_d(0, I_d)$). We are interested in the concentration behavior of the quadratic form X^TAX. Now, we have

$$\begin{aligned}X^TAX &= X^T\Gamma\Lambda\Gamma^TX \\ &\overset{d}{=} Z^T\Gamma Z \\ &= \sum_{i=1}^{d}\lambda_i Z_i^2, \text{ where } \lambda_i \text{ is the } i\text{th eigenvalue of } A.\end{aligned}$$

[1] If X does not have covariance matrix I, we can standardize it by pre-multiplying by $\Sigma^{-\frac{1}{2}}$:

$$\begin{aligned}X^TAX &= X^T\Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}X \\ &\overset{d}{=} Z^TBZ, \text{ where } B = \Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}} \text{ and } Z \sim N(0, I).\end{aligned}$$

The first equality is obtained using the spectral theorem, which tells us that if A is symmetric positive definite, we can write $A = \Gamma\Lambda\Gamma^T$, where Λ is a diagonal matrix with the eigenvalues of A, and Γ is an orthogonal matrix. The second equality uses the property of rotational invariance of standard normal variables: if a standard normal random variable is pre-multiplied by an orthogonal matrix, the result is still standard normal, i.e. $\Gamma^T X \overset{d}{=} Z$. In 2D, we can think of this property as rotating a contour plot of a bivariate standard normal density, where after rotation the result is still standard normal.

Now, we note that $Z_i^2 \sim \chi_1^2$, so we have reduced a potentially complicated quadratic form to a weighted sum of χ_1^2 random variables. Hence we look at concentration bounds for $W = \sum_{i=1}^d \lambda_i(Y_i - 1)$, where $Y_i \sim \chi_1^2$. Then, $\forall t > 0$,

$$\mathbb{P}[W \geq 2\|\lambda\|\sqrt{t} + 2\|\lambda\|_\infty t] \leq e^{-t} \text{ and } \mathbb{P}[W \leq -2\|\lambda\|\sqrt{t}] \leq e^{-t},$$

where $\|\lambda\| = \sqrt{\sum \lambda_i^2} = |||A|||_F$, the Frobenius norm, and $\|\lambda\|_\infty = \max_i \lambda_i = |||A|||_{op}$, the operator norm. The operator norm is also equal to $\sup_{\{x:\|x\|=1\}} x^T Ax$. For details on the derivation see Example 2.12 in [BLM13] and Lemma 1 in [LM00].

The following result is an extension to sub-Gaussian random variables, and is known as the Hanson-Wright inequality. The proof is in [RV13] and Homework 2 Problem 5.

If $X_1, ..., X_d$ are independent and $X_i \in SG(\sigma^2)$,

$$\mathbb{P}[|X^T AX - \mathbb{E}[X^T AX]| \geq t] \leq 2e^{-c\min\left\{\frac{c_1 t^2}{|||A|||_F}, \frac{c_2 t}{|||A|||_{op}}\right\}},$$

where c_1 and c_2 depend on σ^2.

4.3 Bounded Difference Inequality

The bounded difference inequality allows us to get bounds on a function of independent random variables, where the function could be more complicated than just a sum. Let $Z = f(X_1, ..., X_n)$, where $X_1, ..., X_n$ are independent. Let $Y_0 = \mathbb{E}[Z]$ and for $k = 1, ..., n$ let $Y_k = \mathbb{E}[Z|X_1, ..., X_k]$. In particular, $Y_n = Z$. Then, by telescoping,

$$\begin{aligned} Z - \mathbb{E}[Z] &= Y_n - Y_0 \\ &= \sum_{k=1}^n (Y_k - Y_{k-1}) \\ &= \sum_{k=1}^n D_k, \end{aligned}$$

where $D_k = Y_k - Y_{k-1}$. We have expressed Z, a possibly complicated function of the X_i's, as a sum of k random variables. However, these D_k's are not independent, so we will need several more tools.

Remark The ordering of X_i's is not important!

4.3.1 Martingales

Let $(\Omega, \mathcal{F})$ be a probability space. $\mathcal{F}_0 = \{\phi, \Omega\}$ is the trivial σ-field. A filtration is a sequence $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2...$ of sub-σ-fields of $\mathcal{F}$.

The sequence $\{Y_k\}_{k=1,2,...}$ is adapted to the filtration if Y_k is $\mathcal{F}_k$-measurable $\forall\, k$.

This sequence is a martingale if $\mathbb{E}[|Y_k|] < \infty$ and $\mathbb{E}[Y_{k+1}|\mathcal{F}_k] = Y_k \;\forall\, k$.

4.3.2 Doob Construction

Let $Z = f(X_1, ..., X_n)$ be integrable, and $\mathcal{F}_k = \sigma(X_1, ..., X_k)$ for $k = 0, ..., n$. Then $Y_k = \mathbb{E}[Z|\mathcal{F}_k]$ is a martingale. This is known as the Doob martingale or Levy martingale.

If $(Y_k, \mathcal{F}_k)_{k=0,1,...}$ is a martingale, then $(D_k, \mathcal{F}_k)_{k=1,2,...}$, where $D_k = Y_k - Y_{k-1}$, is a martingale difference sequence. It is adapted to the filtration, and $\mathbb{E}[D_k] = 0 \,\forall\, k$.

4.3.3 Concentration bounds for martingale difference sequences

Theorem 4.1 *Let $(D_k, \mathcal{F}_k)_{k=1,2,...}$ be a martingale difference such that*

$$\mathbb{E}[e^{\lambda D_k}|\mathcal{F}_{k-1}] \leq e^{\frac{\lambda^2 \nu_k^2}{2}} \;\forall\, |\lambda| < \frac{1}{\alpha_k},$$

where $\nu_k, \alpha_k > 0$.

Then

1. $\sum_{k=1}^n D_k \in SE(\nu_*, \alpha_*)$*, where* $\nu_*^2 = \sum_{k=1}^n \nu_k^2$ *and* $\alpha_* = \max_k \alpha_k$*.*
2. $\mathbb{P}[|\sum_{k=1}^n D_k| \geq t] \begin{cases} 2e^{-\frac{t^2}{2\nu_*^2}} & \textit{if } 0 \leq t \leq \frac{\nu_*^2}{\alpha_*} \\ 2e^{-\frac{t^2}{2\alpha_*}} & \textit{otherwise} \end{cases}.$

Proof:

$$\begin{aligned} \mathbb{E}[e^{\lambda \sum_{k=1}^n D_k}] &= \mathbb{E}[\mathbb{E}[e^{\lambda \sum_{k=1}^n D_k}|\mathcal{F}_{n-1}]] \\ &= \mathbb{E}[e^{\lambda \sum_{k=1}^{n-1} D_k} \mathbb{E}[e^{\lambda D_n}|\mathcal{F}_{n-1}]] \\ &\leq \mathbb{E}[e^{\lambda \sum_{k=1}^{n-1} D_k}] e^{\frac{\lambda \nu_n^2}{2}}, \text{ if } |\lambda| < \tfrac{1}{\alpha_n} \\ &\leq e^{\frac{\lambda^2 \sum_{k=1}^n \nu_k^2}{2}}, \text{if } |\lambda| < \frac{1}{\max_k \alpha_k}, \end{aligned}$$

where the second equality is because $e^{\lambda \sum_{k=1}^{n-1} D_k}$ is $\mathcal{F}_{n-1}$-measurable, and the last inequality is obtained by iterating this procedure.

Point 2 follows directly from point 1.

■

Remark This result is the same as for independent sub-exponential variables.

References

[GKKW02] László Györfi, Michael Kohler, Adam Krzyżak and Harro Walk, *A Distribution-Free Theory of Nonparametric Regression*, Springer, 2002.

[BLM13] S. Boucheron, G. Lugosi and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, 2013.

[LM00] B. Laurent and P. Massart, “Adaptive estimation of a quadratic functional by model selection,” *Annals of Statistics*, 2000, 28(5), 1302-1338.

[RV13] M. Rudelson and R. Vershynin, “Hanson-Wright inequality and sub-gaussian concentration,” *Electron. Commun. Probab.*, 2013, 18(82), 1–9.