10-716: Advanced Machine Learning	**Spring 2019**

Lecture 6: February 5

Lecturer: Lecturer: Pradeep Ravikumar Scribes: Scribes: Aleksandr Podkopaev, Jeffrey Li, Ruogu Lin

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.*

6.1 Sub-Gaussian Random variables

Definition 6.1 (Sub-Gaussian Random variable) *A random variable X with mean μ is called Sub-Gaussian with parameter σ ($X \sim SG(\sigma)$) if:*

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

By applying Chernoff argument, it translates into:

$$\mathbb{P}\left(|X-\mu| \geq t\right) \leq 2e^{-\frac{t^2}{2\sigma^2}}, \ \forall t \geq 0$$

Example: Consider Rademacher random variable $X \in \{-1, +1\}$ and $\mathbb{P}(X = 1) = \frac{1}{2}$. One can show that:

$$\mathbb{E}\left[e^{\lambda X}\right] \leqslant e^{\frac{\lambda^2}{2}}, \forall \lambda \in \mathbb{R}$$

implying that $X \sim SG(1)$.

Example: Consider Normal random variable $X \sim N\left(\mu, \sigma^2\right)$. One has:

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] = e^{\frac{\lambda^2\sigma^2}{2}}, \ \forall \lambda \in \mathbb{R}$$

implying that $X \sim SG(\sigma)$.

Lemma 6.2 (Bounded random variables) *Consider zero-mean random variable X, which is bounded: $X \in [a, b]$. One has:*

$$\mathbb{E}\left[e^{\lambda X}\right] \leqslant e^{\frac{\lambda^2(b-a)^2}{2}}, \forall \lambda \in \mathbb{R}$$

implying that $X \sim SG(b-a)$.

Proof: The idea of the proof is related to so-called symmetrization argument and introducing a ghost sample. Introduce a random variable X' that has the same distribution as random variable X. Since $\mathbb{E}\left[X'\right] = \mathbb{E}\left[X\right] = 0$:

$$\mathbb{E}\left[e^{\lambda X}\right] = \mathbb{E}\left[e^{\lambda X - \lambda \mathbb{E}\left[X'\right]}\right]$$

Using Jensen's inequality:

$$\mathbb{E}\left[e^{\lambda X - \lambda \mathbb{E}\left[X'\right]}\right] \leqslant \mathbb{E}_X \mathbb{E}_{X'}\left[e^{\lambda\left(X - X'\right)}\right]$$

Introduce a Rademacher random variable $\varepsilon \in \{-1, +1\}$ here. Then $\varepsilon(X - X')$ has the same distribution as $X - X'$ and, thus:

$$\mathbb{E}_X \mathbb{E}_{X'} \left[e^{\lambda \left(X - X' \right)} \right] = \mathbb{E}_X \mathbb{E}_{X'} \mathbb{E}_\varepsilon \left[e^{\lambda \varepsilon \left(X - X' \right)} \right]$$

$\mathbb{E}_\varepsilon \left[e^{\lambda \varepsilon \left(X - X' \right)} \right]$ is the MGF (moment generation function) of Rademacher random variable ε:

$$\mathbb{E}_X \mathbb{E}_{X'} \mathbb{E}_\varepsilon \left[e^{\lambda \varepsilon \left(X - X' \right)} \right] \leqslant \mathbb{E}_X \mathbb{E}_{X'} \left[e^{\frac{\lambda^2 \left(X - X' \right)^2}{2}} \right]$$

Since $X \in [a, b]$, $|X - X'| \leqslant b - a$:

$$\mathbb{E}_X \mathbb{E}_{X'} \left[e^{\frac{\lambda^2 \left(X - X' \right)^2}{2}} \right] \leqslant e^{\frac{\lambda^2 (b-a)^2}{2}}$$

■

Sub-Gaussian random variables satisfy the following additive properties:

1. $X_1 \sim SG(\sigma_1), X_2 \sim SG(\sigma_2)$, X_1, X_2 are independent. Then $X_1 + X_2 \sim SG\left(\sqrt{\sigma_1^2 + \sigma_2^2} \right)$

   Proof:

$$\mathbb{E} \left[e^{\lambda (X_1 + X_2)} \right] = \mathbb{E} \left[e^{\lambda X_1} \right] \mathbb{E} \left[e^{\lambda X_2} \right] \leqslant e^{\frac{\lambda^2 \sigma_1^2}{2}} + e^{\frac{\lambda^2 \sigma_2^2}{2}} = e^{\frac{\lambda^2 \left(\sqrt{\sigma_1^2 + \sigma_2^2} \right)^2}{2}}$$

■

2. Denote $\sigma = \begin{bmatrix} \sigma_1 & \sigma_2 & \cdots & \sigma_n \end{bmatrix}$ and assume $X_i \sim SG(\sigma_i), i = 1, \ldots, n$. Then $\sum\limits_{i=1}^n X_i \sim SG(\|\sigma\|_2)$

Hoeffding's Bound: Assume that $X_1, X_2, \ldots, X_n$ are independent random variables with means $\mu_1, \mu_2, \ldots, \mu_n$ and Sub-Gaussian with parameters $\sigma_1, \sigma_2, \ldots, \sigma_n$. Then:

$$\mathbb{P} \left[\sum_{i=1}^n (X_i - \mu_i) \geqslant t \right] \leqslant \exp \left\{ -\frac{t^2}{2 \sum_{i=1}^n \sigma_i^2} \right\}, \ \forall t \geqslant 0$$

Hoeffding's inequality: Assume that $X_1, X_2, \ldots, X_n$ are independent random variables with means $\mu_1, \mu_2, \ldots, \mu_n$ and Sub-Gaussian with parameters $\sigma_1, \sigma_2, \ldots, \sigma_n$. Then:

$$\mathbb{P} \left[\left| \frac{1}{n} \sum_{i=1}^n (X_i - \mu_i) \right| \geqslant t \right] \leqslant 2 \exp \left\{ -\frac{n^2 t^2}{2 \sum_{i=1}^n \sigma_i^2} \right\}, \ \forall t \geqslant 0$$

Note that if $\sigma_i^2 = \sigma^2$, then the inequality simplifies to:

$$\mathbb{P} \left[\left| \frac{1}{n} \sum_{i=1}^n (X_i - \mu_i) \right| \geqslant t \right] \leqslant 2 \exp \left\{ -\frac{n t^2}{2 \sigma^2} \right\}, \ \forall t \geqslant 0$$

Theorem 6.3 *Assume that a random variable X is centered: $\mathbb{E}X = 0$. Then the following statements are equivalent:*

1. $\mathbb{E} e^{\lambda X} \leq e^{\frac{\lambda^2 \sigma^2}{2}}, \forall \lambda \in \mathbb{R}$

2. $\exists c, b > 0 : \mathbb{P}(|X| > t) \le c\mathbb{P}(|Z| > t)$ *where* $Z \sim \mathcal{N}(0, b^2)$
3. $\forall k \in \mathbb{N} \hookrightarrow \mathbb{E}X^{2k} \le \frac{(2k)!}{2^k k!}\theta^{2k}$ *for some* θ

In fact, the second characterization gives rise to the name Sub-Gaussian random variable.

6.2 Sub-Exponential Random variables

Definition 6.4 (Sub-Exponential Random variable) *Random variable X is called Sub-Exponential with parameters $\nu, \alpha > 0$ ($X \sim SE(\nu, \alpha)$) if:*

$$\mathbb{E}e^{\lambda(X-\mu)} \le e^{\lambda^2\nu^2/2}, \forall \lambda \in [0, \frac{1}{\alpha}]$$

Note that $X \sim SG(\sigma) \implies X \sim SE(\sigma, 0)$.

Example: Consider $Z \sim \mathcal{N}(0,1)$. Then $X = Z^2 \sim \chi^2(1)$. Also $\mathbb{E}X = \mathbb{V}(Z) = 1$. One has:

$$\mathbb{E}e^{\lambda(X-\mu)} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\lambda(z^2-1)}e^{-\frac{z^2}{2}}dz = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}, \ \forall \lambda : |\lambda| < \frac{1}{2}$$

So, the moment generating function is not defined over the whole real line, implying that X is not Sub-Exponential. On the other hand, if $\lambda < \frac{1}{4}$, then one can show that:

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \le e^{2\lambda^2} = e^{4\lambda^2/2}$$

implying that $X \sim SE(2, 4)$.

Theorem 6.5 (Tail bounds for Sub-Exponential random variables) *Assume that $X \sim SE(\nu, \alpha)$. Then:*

$$\mathbb{P}(X - \mu \ge t) \le \begin{cases} e^{-\frac{t^2}{2\nu^2}}, & t \le \frac{\nu^2}{\alpha} \\ e^{-\frac{t}{2\alpha}}, & t > \frac{\nu^2}{\alpha} \end{cases}$$

Notice that only for sufficiently small t ($t \le \frac{\nu^2}{\alpha}$) one observes Sub-Gaussian tail behavior while for larger values the decay is slower.

Proof: Using the same argument as before:

$$\mathbb{P}(X - \mu > t) = \mathbb{P}\left(e^{\lambda(X-\mu)} > e^{\lambda t}\right) \le e^{-\lambda t}\mathbb{E}e^{\lambda(X-\mu)} \le e^{-\lambda t + \frac{\lambda^2\nu^2}{2}}, \ \forall \lambda \in [0, \frac{1}{\alpha}]$$

To obtain the tightest bound one needs to find:

$$g^*(t) = \inf_{\lambda \in (0, \frac{1}{\alpha}]}\left(-\lambda t + \frac{\lambda^2\nu^2}{2}\right) = \inf_{\lambda \in (0, \frac{1}{\alpha}]} g(\lambda, t)$$

To do so, notice, firstly, that unconstrained minimum occurs at $\lambda^* = t/\nu^2 > 0$. Consider two cases:

1. If $\lambda^* < 1/\alpha \Leftrightarrow t \le \frac{\nu^2}{\alpha}$, then unconstrained minimum appears to be also constrained minimum and by plugging in this value one obtains the bound describing sub-Gaussian behavior.

2. If $\lambda^* > 1/\alpha \Leftrightarrow t > \nu^2/\alpha$, then notice that the function $g(\lambda, t)$ is decreasing in λ in the interval $\lambda \in (0, \frac{1}{\alpha})$. Thus, the constrained minimum occurs at the boundary:

$$\lambda^*_{constrained} = \frac{1}{\alpha}$$

and since $\frac{\nu^2}{\alpha} \leq t$:

$$g(\lambda^*_{constrained}, t) = -\frac{t}{\alpha} + \frac{1}{2\alpha}\frac{\nu^2}{\alpha} \leq -\frac{t}{2\alpha}$$

■

Lemma 6.6 (Properties of Sub-Exponential random variables) *Assume that $X_1, \ldots, X_n$ are independent sub-exponential random variables: $X_i \sim SE(\nu_i, \alpha_i)$. Then:*

$$\sum_{i=1}^{n} X_i \sim SE(\nu^*, \alpha^*)$$

where $\nu^* = \sqrt{\sum_{i=1}^{n} \nu_i^2}, \alpha^* = \max_i \alpha_i$

The proof is straightforward and uses two facts:

- MGF of a sum of independent random variables is a product of the individual MGFs.
- Within range $[0, \frac{1}{\alpha^*}]$ the moment generating function (MGF) for each X_i can be upper bounded as usually.

that within range $[0, \alpha^*]$ MGF for each X_i can be upper bounded and

Example: Consider $X_i = Z_i^2, Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Since $X_i \sim SE(2, 4)$. Then by the lemma 6.6:

$$\sum_{i=1}^{n} X_i \sim SE(2\sqrt{n}, 4)$$

Considering the tail bound, but for the average:

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - 1) > t\right) \leq \begin{cases} e^{-\frac{nt^2}{8}}, & t \in [0, 1] \\ e^{-\frac{nt}{8}}, & t > 1 \end{cases}$$

6.2.1 Different characterizations

Theorem 6.7 *Assume that a random variable X has mean μ. Then the following statements are equivalent statements for X being sub-exponential:*

1. $\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\frac{\lambda^2\nu^2}{2}}, \forall \lambda \in [0, \frac{1}{\alpha}]$
2. $\exists c_0 > 0 : \mathbb{E}e^{\lambda x} < \infty, \forall |\lambda| < c_0$

The second characterization is particularly remarkable as it implies that any random variable with finite MGF in some open region around $\lambda = 0$ is a sub-exponential random variable for some ν, α.

6.3 Martingales

Our previous results for Sub-Gaussian random variables allowed for us to naturally provide tail bounds for sums of independent random variables, each of which are Sub-Gaussian.

However, we often would like to more generally analyze sums of non-independent Sub-Gaussian random variables. One example of why we may want to do this is to apply tail bounds to ERM, where we are trying to minimize the empirical risk, the sample average of a random variable, and want that to be close to its expectation. Looking at this more generally, we can consider a sample of random variables $X = (X_1, ..., X_n)$ and a function $f(X)$. We want to analyze

$$f(X) - \mathbb{E}[f(X)]$$

The problem is for general f's outside of just sample averages, we cannot use our previous results about Sub-Gaussian random variables. Even if the X_i's are independent and individually Sub-Gaussian, we cannot say whether $f(\boldsymbol{X})$ is Sub-Gaussian in any way. Assumptions about f are needed. Instead, we can try to write our expression as a sum of random variables, which we can analyze.

Let us define

$$\begin{aligned}
Y_n &= f(X) && \text{(random variable)} \\
Y_k &= \mathbb{E}[f(X)|X_1, ..., X_k] && \text{(still random, but less variability)} \\
Y_0 &= \mathbb{E}[f(X)] && \text{(constant)}
\end{aligned}$$

Then we can rewrite the quantity we are interested in as a sum of differences

$$Y_n - Y_0 = \sum_{k=1}^{n}(Y_k - Y_{k-1}) = \sum_{k=1}^{n} D_k$$

Each D_k is not independent of the others but we still want to reason about how far these are from their expectations. It turns out we cannot reason about arbitrary differences between random variables like this, but we can for a class of sequences called Martingales.

Definition 6.8 (Martingales) *A sequence of random variables $\{Y_k\}_{k=1}^{\infty}$ is said to be a Martingale sequence with respect to some other sequence of random variables $\{X_k\}_{k=1}^{\infty}$ if*

1. *Y_k is a measurable function of $X_1, ..., X_k$*
2. *$\mathbb{E}[Y_{k+1}|X_1, ..., X_k] = Y_k$, i.e. Y_{k+1} is centered around Y_k.*
3. *$\mathbb{E}[|Y_k|] < \infty$*

Example (Applying definition to our above example): In our previous example we indeed have a Martingale sequence provided $\mathbb{E}[|f(x)|] < \infty$.

1. We can assume that the first condition is satisfied:

$$Y_k = \mathbb{E}[f(X)|X_1, ..., X_k]$$

2. We can use the Law of Total Expectation to get

$$\mathbb{E}[Y_{k+1}|X_1, ..., X_k] = \mathbb{E}\left[\mathbb{E}[f(X)|X_1, ..., X_{k+1}]|X_1, ..., X_k\right] = \mathbb{E}[f(X)|X_1, ..., X_k] = Y_k$$

3. We have by Jensen's Inequality and the Law of Total Expectation:

$$\mathbb{E}|Y_k| = \mathbb{E}\Big||\mathbb{E}\left[f(x)|X_1, ..., X_k\right]|\Big| \leq \mathbb{E}\left[\mathbb{E}\left[\Big|f(x)\Big||X_1, ..., X_k\right]\right] = \mathbb{E}[|f(x)|] < \infty$$

   Here the last inequality is actually something we need to assume about $f(x)$ in order to satisfy the conditions for a Martingale sequence.