Lecture 6 — Sep 16, 2015

Prof. Eric Price *Scribe: Rishab Goyal*

1 Overview

In the last lecture we introduced subgaussian, subexponential and subgamma distributions, and also discussed few equivalent sufficient conditions for each of them.

In this lecture we first recall those definitions. Then we discuss Johnson-Lindenstrauss (JL) lemma and provide a proof for the same. After that we revisit the problem of coupon collector, and finally we provide a Bernstein-type inequality for a bounded range random variable.

2 Preliminaries

Definition 1. *(Subexponential random variable) A random variable X is said to be subexponential with parameter σ iff $X - \mathbb{E}[X]$ satisfies any of the following three properties:*

$$\mathbb{P}[|X| \geq t] \leq 2e^{-t/(2\sigma)} \quad \forall\ t > 0, \text{ or}$$
$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\lambda^2\sigma^2/2} \quad \forall\ |\lambda| < 1/\sigma, \text{ or}$$
$$\mathbb{E}\left[|X|^k\right]^{1/k} \leq \sigma \cdot k.$$

If properties (1) and (3) hold for X, then they also hold for $X - \mathbb{E}[X]$, so X is subexponential.

Definition 2. *(Subgamma random variable) A random variable X is said to be subgamma with parameters σ and B if $X - \mathbb{E}[X]$ satisfies the following:*

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\lambda^2\sigma^2/2} \quad \forall\ |\lambda| < B.$$

This additionally implies that:

$$\mathbb{P}[|X| \geq t] \leq 2 \cdot \max\left(e^{-t^2/(2\sigma^2)}, e^{-Bt/2}\right).$$

Lemma 3. *If $X_1, X_2, \ldots, X_n$ are independently distributed subgamma variables with parameters (σ_i, B_i) for $i \in [n]$, then $X = \sum_{i=1}^n X_i$ is also subgamma with parameters $(\sqrt{\sum_{i=1}^n \sigma_i^2}, \min_{i\in[n]} B_i)$.*

Lemma 4. *If $X_1, X_2, \ldots, X_n$ are independent standard normal variables ($\sim N(0,1)$), then X^2 is subgamma with parameters $(\sqrt{n}, 1)$, where $X^2 = \sum_{i=1}^n X_i^2$.*

Proof. Since $X_i \sim N(0,1)$, we know that $\mathbb{P}\left[|X_i| \geq t\right] \leq e^{-t^2/2}$ (for all $t > 0$). This implies that $\mathbb{P}\left[X_i^2 \geq t\right] \leq e^{-t/2}$ (for all $t > 0$). Therefore, each X_i^2 is a subexponential with parameter 1 $(= O(1))$, and X_i^2 is subgamma with parameters $(1,1)$ (as per Defintion 2).

Now, using Lemma 3, we can conclude that X^2 is subgamma with parameters $(\sqrt{n}, 1)$. □

3 Johnson-Lindenstrauss Lemma

During lecture, the following question was posed which captures the essence of JL lemma.

Question. Let $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$. Find $Y_1, Y_2, \ldots, Y_n \in \mathbb{R}^m$ such that for all i, j:

$$(1-\epsilon)\|x_i - x_j\|_2 \leq \|y_i - y_j\|_2 \leq (1+\epsilon)\|x_i - x_j\|_2.$$

How large should m be?

For completeness, we state the formal statement of the JL lemma:

Theorem 5. *(JL Lemma) If $X \subseteq \mathbb{R}^d$ such that $|X| = n$, then for every $0 < \epsilon < 1/2$, there exists a linear map $A : \mathbb{R}^d \to \mathbb{R}^m$ such that for all $x_i, x_j \in X$:*

$$(1-\epsilon)\|x_i - x_j\|_2 \leq \|Ax_i - Ax_j\|_2 \leq (1+\epsilon)\|x_i - x_j\|_2$$

and $m = \Theta\left(\dfrac{\log n}{\epsilon^2}\right)$.

Proof. Consider the following linear map A such that $A_{i,j} \sim N(0, 1/m)$.

Let $x \in \mathbb{R}^d$ and $y = Ax$. Observe that $(Ax)_i = \sum_{j=1}^{d} A_{i,j} x^{(j)}$. This is a sum of independent gaussians, and since the sum of gaussians is itself a gaussian with variance equal to the sum of individual variances, we have that $(Ax)_i \sim N\Big(0, \sum_{j=1}^{d} (x^{(j)})^2\Big)$. Therefore, $(Ax)_i \sim N(0, \|x\|^2/m)$.

Next, observe that $\|Ax\|^2 = \sum_{i=1}^{m} (Ax)_i^2$. Therefore, $\mathbb{E}\left[\|Ax\|^2\right] = \sum_{i=1}^{m} \mathbb{E}\left[(Ax)_i^2\right]$, using linearity of expectation. Since, $(Ax)_i \sim N(0, \|x\|^2/m)$, this implies that $\mathbb{E}\left[(Ax)_i^2\right] = Var((Ax)_i) = \|x\|^2/m$. So, $\mathbb{E}\left[\|Ax\|^2\right] = \|x\|^2$.

Let $z_i = \dfrac{\sqrt{m}}{\|x\|} \cdot (Ax)_i$ for $i \in [m]$, and $\|z\|^2 = \sum_{i=1}^{m} z_i^2 = \dfrac{m}{\|x\|^2} \cdot \|Ax\|^2$. It is straightforward to verify that $z_i \sim N(0,1)$, and z_i's are all independent. Therefore, using Lemma 4, we can claim that $\|z\|^2$ is subgamma with parameters $(\sqrt{m}, 1)$. Since $\|z\|^2$ is subgamma, we could write the following (as implied by Definition 2):

$$\mathbb{P}\Big[\big|\|z\|^2 - m\big| \geq t\Big] \leq 2 \cdot \max(e^{-t^2/(2m)}, e^{-t/2}).$$

Substituting t with $2\epsilon m$, we get

$$\mathbb{P}\Big[\|z\|^2 \geq (1+2\epsilon)m \vee \|z\|^2 \leq (1-2\epsilon)m\Big] \leq 2 \cdot \max(e^{-2\epsilon^2 m}, e^{-\epsilon m}).$$

Since $0 < \epsilon < 1/2$, we can write that $\max(e^{-2\epsilon^2 m}, e^{-\epsilon m}) = e^{-2\epsilon^2 m}$. Also, $(1+2\epsilon)^{1/2} \geq (1+\epsilon)$ and $(1-2\epsilon)^{1/2} \leq (1-\epsilon)$, so we could rewrite the above inequality as:

$$\mathbb{P}\left[\|z\| \geq (1+\epsilon)\sqrt{m} \vee \|z\| \leq (1-\epsilon)\sqrt{m}\right] \leq 2e^{-2\epsilon^2 m}.$$

Substituting $\|z\|$ with $\dfrac{\sqrt{m}}{\|x\|} \cdot \|Ax\|$, we get

$$\mathbb{P}\left[\|Ax\| \geq (1+\epsilon)\|x\| \vee \|Ax\| \leq (1-\epsilon)\|x\|\right] \leq 2e^{-2\epsilon^2 m}.$$

Note that above inequality holds for all $x \in \mathbb{R}^d$, therefore substituting x with $x_i - x_j$, and Ax with $A(x_i - x_j) = y_i - y_j$, we get for all x_i, x_j

$$\mathbb{P}\left[\|y_i - y_j\| \notin \big((1-\epsilon)\|x_i - x_j\|, (1+\epsilon)\|x_i - x_j\|\big)\right] \leq 2e^{-2\epsilon^2 m}.$$

Therefore, using union bound, we can say that

$$\mathbb{P}\left[A \text{ preserves distance between all pairs}\right] \geq 1 - n^2 e^{-2\epsilon^2 m}.$$

So, for $m = c\dfrac{\log n}{\epsilon^2}$,

$$\mathbb{P}\left[A \text{ preserves distance between all pairs}\right] \geq 1 - \frac{1}{n^{2c-2}} \geq 0.$$

This concludes the proof as we have proved that there exists a linear map A which reduces the dimensionality of data and preserves its structure (distances) *with high probability.* $\square$

4 Coupon Collector Problem

In the n coupon collector problem, T_i denoted the number of draws to get the i^{th} new coupon, and $T = \sum_{i=1}^n T_i$ denoted the total number of draws made to draw each coupon at least once.

Recall that T_i's are all independent of each other, and $T_i \sim Geom\Big(\dfrac{n+1-i}{n}\Big)$, $\mathbb{E}[T] = nH_n$, and $Var(T) = O(n^2)$. Using Markov's inequality for concentration, we get

$$\mathbb{P}\left[T \geq \frac{nH_n}{\delta}\right] \leq \delta.$$

Similarly, Chebyshev gives

$$\mathbb{P}\left[T \geq nH_n + \frac{O(n)}{\sqrt{\delta}}\right] \leq \delta.$$

Now, we will try to get similar bounds using results from subgamma distributions. First, note that

$$\mathbb{P}[T_i \geq t] \leq \left(1 - \frac{n+1-i}{n}\right)^t \leq e^{-(\frac{n+1-i}{n})t}.$$

Therefore, T_i is subexponential with parameter $\sigma_i = \dfrac{n}{n+1-i}$. Hence, T is subgamma with mean $\mu = nH_n$ and parameters $(n\sqrt{\sum_{i=1}^n \big(\dfrac{1}{n+1-i}\big)^2}, min_i\ \sigma_i^{-1}) = (O(n), \dfrac{1}{n})$. Since T is subgamma, we could write the following:

$$\mathbb{P}[T \geq \mu + t] \leq 2 \cdot \max\left(e^{-t^2/(2n^2)}, e^{-t/(2n)}\right) \leq 2e^{-t/(2n)}.$$

Substituting t with $O(n\log(1/\delta))$, we get

$$\mathbb{P}\Big[T \geq nH_n + O(n\log(1/\delta))\Big] \leq \delta.$$

Note that all the above analysis was done using different concentration inequalities. Below we discuss how to obtain the same bound as obtained with subgamma analysis of T using only union bound.

Let $F_{i,T}$ denote the event that coupon i is not found after T steps. Therefore, we can say that

$$\mathbb{P}[F_{i,T}] = \Big(1 - \frac{1}{n}\Big)^T \leq e^{-T/n}.$$

Substituting T with $n\log(n/\delta)$, we get (for all $i \in [n]$)

$$\mathbb{P}[F_{i,n\log(n/\delta)}] \leq \frac{\delta}{n}.$$

So, using union bound, we can conclude that

$$\mathbb{P}[\text{Any coupon not found in } n\log(n/\delta) \text{ steps}] = \mathbb{P}\Big[\bigvee_{i\in[n]} F_{i,n\log(n/\delta)}\Big] \leq \delta.$$

The above analysis gives an upper bound on the probability that all coupons are not found after T steps. It should be interesting to find whether this bound is tight, or if we could do better. For that purpose, we start with lower bounding the probability of event $F_{i,T}$ as

$$\mathbb{P}[F_{i,T}] = \Big(1 - \frac{1}{n}\Big)^T \geq e^{-\frac{T}{n}(1-O(\frac{1}{n}))} \geq \frac{\delta}{n}\Big(1 - O(\frac{\delta}{n})\Big),$$

where we substituted T with $n\log(n/\delta)$. If we assume that the events $\neg F_{i,T}$ (coupon i is found after T steps) are *independent*, then using the above inequality we could write the following

$$\begin{aligned}\mathbb{P}[\text{Any coupon not found in } n\log(n/\delta) \text{ steps}] &= 1 - \mathbb{P}[\text{All coupons found after } n\log(n/\delta) \text{ steps}]\\ &= 1 - \mathbb{P}\Big[\bigwedge_{i\in[n]} \neg F_{i,n\log(n/\delta)}\Big]\\ &\geq 1 - \Big(1 - \frac{\delta}{n}\Big(1 - O(\frac{\delta}{n})\Big)\Big)^n \approx \delta\end{aligned}$$

This would suggest that our bound is tight. But it assumes independence of events $\neg F_{i,T}$ which is incorrect, so we could only consider the above lower bound as a mild indicator of tightness. However, despite lack of independence, we can expect much better concentration because intuitively if some subset $I \subset [n]$ is drawn within first T draws, then the event that some coupon $j \notin I$ is also drawn in first T draws is less likely. Looking ahead, the key fact will be that $\neg F_{i,T}$ are negatively correlated.

Next class. First, observe that $\mathbb{E}[\text{Number of coupons not found after } n \text{ steps}] = n\left(1 - \frac{1}{n}\right)^n \approx n/e$. In the next lecture, we will try to bound the probability that more than $n/2$ coupons not found after n draws ($\mathbb{P}[\text{Number of coupons not found after } n \text{ steps} \geq n/2]$).

5 Bernstein-Type Inequality

Recall that while discussing Chernoff bound, we observed that its main limitation is that it does not depend on the variance of each individual X_i. In the same vein, we try to improve the concentration bound for bounded range random variables. Below we prove that each bounded range variable is also subgamma. Since, sum of subgamma variables is also subgamma (from Lemma 3), therefore it has interesting concentration properties. Formally, we state the theorem below.

Theorem 6. *If X is a random variable with range $[0, 1]$ and variance s^2, then it is also subgamma with parameters $(\sqrt{2}s, 1/2)$.*

Proof. Let $Y = X - \mathbb{E}[X]$. This implies that $\mathbb{E}[Y^2] = Var(X) = s^2$ and $|Y| \leq 1$. Since $|Y| \leq 1$, we can also conclude that $\mathbb{E}\left[|Y|^k\right] \leq s^2$ for all $k > 2$. Next we will try to bound the moment generating function of Y.

$$\begin{aligned}
\mathbb{E}[e^{\lambda Y}] &= \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{\lambda^k Y^k}{k!}\right] \\
&\leq 1 + \sum_{k=2}^{\infty} \frac{\lambda^k}{k!} s^2 \\
&\leq 1 + \left(\sum_{k=0}^{\infty} \lambda^k\right) \frac{\lambda^2 s^2}{2}
\end{aligned}$$

Note that the geometric series on the right converges only if $\lambda < 1$. So, for $\lambda < 1$, we could write

$$\mathbb{E}[e^{\lambda Y}] \leq 1 + \frac{\lambda^2 s^2}{2(1-\lambda)} \leq e^{(\lambda^2 s^2)/(2(1-\lambda))}.$$

Now if we choose λ such that $\frac{\lambda^2}{(1-\lambda)} \leq 2\lambda^2$, then the exponent in the previous bound would agree with that of a subgamma variable. Concretely, if we keep $\lambda \leq 1/2$, then we could write the following

$$\mathbb{E}[e^{\lambda Y}] \leq e^{(2\lambda^2 s^2)/2} = e^{\lambda^2(\sqrt{2}s)^2/2}.$$

Therefore, X is subgamma with parameters $(\sqrt{2}s, 1/2)$. This concludes the proof of Theorem 6. $\square$

Biased coin toss. Let $X_1, \ldots, X_n$ be n i.i.d. Bernoulli random variables ($X_i \sim Ber(p)$), and $X = \sum_{i=1}^n X_i \quad (\sim Bin(n, p))$. Applying Theorem 6, we get that each X_i is subgamma with parameters $(\leq \sqrt{2p}, 1/2)$. Applying Lemma 3, we get that X is also subgamma with parameters $(\sqrt{2pn}, 1/2)$. Therefore,

$$\mathbb{P}[X \geq np + t] \leq 2 \cdot \max(e^{-t^2/(4np)}, e^{-t/4}).$$

If $t \leq np$, then the controlling term would be $e^{-t^2/(4np)}$, else it would be $e^{-t/4}$. Therefore, $\mathbb{P}[X \geq 2np] \leq e^{-np/4}$ and $\mathbb{P}[X \geq np + t] \leq e^{-t^2/(4np)}$ for $t \in [0, np]$.

References

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

[2] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *CoRR*, abs/1011.3027, 2010.