CS 388R: Randomized Algorithms Fall 2015

Lecture 7 — September 21, 2015

Prof. Eric Price *Scribe: Manu Agarwal, Surbhi Goel*

1 Overview

In the last lecture we computed the expected number of missing coupons after collecting n coupons as

$$E\left[\#missing\right] = \left(1 - O\left(\frac{1}{n}\right)\right)\frac{n}{e}$$

Since the variables were no longer independent, we questioned whether they concentrate. Using the intuition that the probability of finding a new coupon having previously found another coupon should be lower, ideally they should concentrate better than if they were independent.

In this lecture we discuss the concept of *Negative Association* to help us prove concentration properties for variables such as the above mentioned. We also discuss the properties a set of variables must satisfy to be negatively associated and see some examples of such set of variables. We end with a brief analysis of the *Balls in Bins* problem.

2 Negative Association

2.1 Definition

Let $X = \{X_1 \dots X_n\}$. When is X said to be negatively associated?

One possible definition could be as follows:

Definition 2.1.1. *X is negatively correlated if*

$$E\left[X_i X_j\right] \leq E\left[X_i\right] E\left[X_j\right] \;\; \forall i, j \in [n] \tag{1}$$

Unfortunately, this does not lead to good concentration properties, so let's try to find a better definition. What would we like our definition to have?

- It should hold for both the discrete as well as the continuous case.
- X_i's should concentrate as well as if they were independent.
- Subsets of negatively associated variables should also be negatively associated.
- It should hold good for independent variables.
- It should be easy to prove.

- It should also satisfy composition rules.

Note that the first definition is not too strong for it to satisfy all the desired properties. Hence, we propose the following definition of negative association:

Definition 2.1.2. *X is negatively associated (NA) if $\forall I, J \subset [n]$ that are disjoint and $\forall$ monotonic f, g (both increasing or both decreasing),*

$$E\left[f(X_I)g(X_J)\right] \leq E\left[f(X_I)\right] E\left[g(X_J)\right] \tag{2}$$

where X_I, X_J are subsets of X indexed by I, J respectively.

2.2 Example

Suppose $X = \{X_1, \ldots, X_n\}$ is negatively associated and each X_i is σ_i subgaussian. We show that $Z = \sum_{i=1}^n X_i$ is $\sqrt{\sum_{i=1}^n \sigma_i^2}$ subgaussian. To do so, we prove the first property of subgaussians. We have,

$$\begin{aligned}
E\left[e^{\lambda Z}\right] &= E\left[e^{\lambda(X_1+\ldots+X_n)}\right] \\
&= E\left[e^{\lambda X_n} e^{\lambda(X_1+\ldots+X_{n-1})}\right] \\
&\leq E\left[e^{\lambda X_n}\right] E\left[e^{\lambda(X_1+\ldots+X_{n-1})}\right] \\
&\leq \prod_{i=1}^n E\left[e^{\lambda X_i}\right] \\
&\leq \prod_{i=1}^n e^{\frac{\lambda^2\sigma_i^2}{2}} \\
&= e^{\frac{\lambda^2 \sum_{i=1}^n \sigma_i^2}{2}}
\end{aligned}$$

Here, the first inequality follows from the definition of NA variables with $f(x) = g(x) = e^{\lambda x}$ (both f, g are both monotonically increasing), second follows by inducting the previous inequality, and third follows from the subgaussian property of each X_i. Thus, Z is also subgaussian with parameter $\sqrt{\sum_{i=1}^n \sigma_i^2}$.

2.3 Properties

Property 2.3.1. *If $X = \{X_1, \ldots, X_n\}$ is NA and $Y = \{Y_1, \ldots, Y_n\}$ is NA independent of X, then $\{X_1, \ldots, X_n, Y_1, \ldots, Y_n\}$ is NA.*

Property 2.3.2. *Let $I_1, \ldots, I_m \subset [n]$ be disjoint and $f_1, \ldots, f_n$ be all monotonically increasing or decreasing functions. If $X = \{X_i, \ldots, X_n\}$ is NA then $Y = \{Y_i = f_i(X_{I_i})\}$ is NA.*

For example, consider a matrix of NA variables X of size $m \times n$. Let $Z_i = \max_j X_{ij}$. Since max is a monotonically increasing function and the rows form disjoint subsets of the variables in the matrix, by property 2.3.2 $Z = \{Z_1, \ldots, Z_m\}$ is NA.

2.4 Zero-One Rule

Rule 2.4.1. *If $X_1, X_2, \ldots, X_n \in \{0,1\}$ and $\sum X_i = 1$, then X is NA.*

Proof. Let f, g be monotonic and $I, J \subset [n]$ be disjoint. Without loss of generality, assume $f(\vec{0}) = 0$ and $g(\vec{0}) = 0$ (we can subtract a constant (value at 0) from each value to get the same). This means either $f(X) \geq 0$ and $g(X) \geq 0$ simultaneously or $f(X) \leq 0$ and $g(X) \leq 0$ simultaneously. Also,

$$E\left[f(X_I)g(X_J)\right] = 0 \leq E\left[f(X_I)\right] E\left[g(X_J)\right]$$

The equality follows from the fact that one of the vectors X_I and X_J must be zero (since each X_i i either 0 or 1 and the sum is 1, only one of the X_is is 1 rest 0).

Let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be constants and let $\sigma_1, \sigma_2, \ldots, \sigma_n \in [n]$ be distinct and uniformly chosen. Then, $X_i = \alpha_{\sigma_i}$ is negatively associated. This requires a more involved proof but the intuition is that if one set has larger number than the other will have smaller numbers since the numbers belong to $[n]$.

3 Coupon Collector Revisited

Let's get back to the question we started with. We sample n coupons from $[n]$. How many are missing after this sampling?

Let $X_{t,i}$ be the event that the coupon sampled at t is i, Then we have $X_t = \{X_{t,1}, \ldots, X_{t,n}\}$ is NA by rule 2.4.1 (zero-one rule) since only one of the $X_{t,i}$ is one and the remaining 0. Since X_t are independent, by property 2.3.1 we have that the matrix X formed by having rows X_t is NA.

Now, let $Y_i = \sum_t X_{t,i}$, that is, the number of times we sample coupon i. Since the columns of X are disjoint and summation of non-negative values is monotonically increasing, the set of Y_i's is NA by property 2.3.2.

Finally, we want to find the number of coupons missing so we define $Z_i = (Y_i \geq 1)$, that is, Z_i is 0 if coupon i is missing after the sampling and 1 otherwise. It is easy to see that the set of Z_i is NA. This implies that $\sum_{i=1}^n Z_i$ concentrate as well as if they were independent.

We have,

$$Pr[Z_i = 1] = 1 - \left(1 - \frac{1}{n}\right)^n = 1 - \frac{1}{e} + O\left(\frac{1}{n}\right)$$

This implies that $E[Z_i] = 1 - \frac{1}{e} + O\left(\frac{1}{n}\right)$ and $E\left[\sum_{i=1}^n Z_i\right] = n\left(1 - \frac{1}{e}\right) + O(1)$. Now using Chernoff's inequality with $t = n\left(\frac{1}{2} - \frac{1}{e}\right)$, we get

$$\begin{aligned} Pr\left[\sum_{i=1}^n Z_i \leq \frac{n}{2}\right] &= Pr\left[\sum_{i=1}^n Z_i \leq E\left[\sum_{i=1}^n Z_i\right] - n\left(\frac{1}{2} - \frac{1}{e}\right)\right] \\ &\leq e^{-\frac{2\left(n\left(\frac{1}{2}-\frac{1}{e}\right)\right)^2}{n}} = e^{-\Omega(n)} \end{aligned}$$

Thus, with high probability we have found more than half the coupons in the sampling.

4 Balls in Bins

We throw n balls into n bins. Let X_i denote the number of balls in bin i. We have $E[X_i] = 1$ and

$$Pr\left[X_i = k\right] \;=\; \binom{n}{k}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{n-k}$$

Using the fact that $\left(1-\frac{1}{n}\right)^{n-k} < 1$, we have

$$Pr\left[X_i = k\right] < \;\binom{n}{k}\left(\frac{1}{n}\right)^k$$

We know that $\left(\frac{n}{k}\right) \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k$. Substituting the same, we get

$$Pr\left[X_i = k\right] \;\leq\; \left(\frac{en}{k}\right)^k\left(\frac{1}{n}\right)^k = \left(\frac{e}{k}\right)^k$$

We want to bound this probability by some small value in terms of n, say we want $Pr[X_i \geq k] < \frac{1}{n^{10}}$

$$\begin{aligned}
E[\max_i X_i] &= E[\max_i X_i | \max_i X_i < k]Pr[\max_i X_i < k] \\
&\quad + E[\max_i X_i | \max_i X_i \geq k]Pr[\max_i X_i \geq k] \\
&\leq k.1 + n.\frac{1}{n^{10}} \\
&= k + \frac{1}{n^9}
\end{aligned}$$

Now, we want

$$\begin{aligned}
\left(\frac{e}{k}\right)^k &< \frac{1}{n^{11}} \\
e^{-k\log k + k} &< e^{-11\log n} \\
k\log k - k &> 11\log n
\end{aligned}$$

It is easy to see by substituting that

$$\sqrt[2]{\log n} < k < \log n$$

Taking log both sides, we have

$$\frac{1}{2}\log\log n < \log k < \log\log n$$

This means

$$log\ k \in (\frac{1}{2}\log\log\ n, \log\log n)$$

In other words,

$$k = \Theta(\frac{\log n}{\log\log n})$$

The last equation follows since $k\log k - k > 11\log n$. We ignore the k, so we get $k\log k > 11\log n$ or $k > \frac{11\log n}{\log k}$.

References

[MR] Rajeev Motwani, Prabhakar Raghavan Randomized Algorithms. *Cambridge University Press*, 0-521-47465-5, 1995.