CS 388R: Randomized Algorithms Fall 2017

Lecture 5 — 09/14/2017

Prof. Eric Price *Scribe: Satyanarayana, Ajil Jalal*

NOTE: THESE NOTES HAVE NOT BEEN EDITED OR CHECKED FOR CORRECTNESS

1 Overview

In this lecture we study the Coupon Collector problem and calculate expected and high probability estimates for the problem. We also study the famous Balls in Bins problem and negative dependence among random variables.

2 Coupon Collector's Problem

Suppose you go to a coupon shop which has coupons of n different colors. When you buy a coupon, the shopkeeper picks a color (coupon) uniformly at random and sells it. Your goal is to keep on buying coupons until you collect all the n varieties of coupons.

(a) Let T be the random variable describing the total number of coupons you buy to collect all the n colors (coupons). What is $E[T]$?

(b) How well does T concentrate about $E[T]$?

Solution (a) Let Z_i be the random variable describing the number of coupons you need to buy to collect a new color after collecting $(i-1)$ colors.

So, $T = \sum\limits_{i=1}^{n} Z_i$ and hence $E[T] = \sum\limits_{i=1}^{n} E[Z_i]$.

After collecting $i-1$ colors, the probability that the shopkeeper picks a new color for the next coupon is $\frac{n-i+1}{n}$. Let this be p_i.

We know that, $Pr(Z_i = k) = (1-p_i)^{k-1} p_i$

$E[Z_i] = \sum\limits_{k=1}^{\infty} k(1-p_i)^{k-1} p_i$

Let $S = \sum\limits_{k=1}^{\infty} k(1-p_i)^{k-1}$.

$$\begin{aligned}
S &= 1.(1-p_i)^0 + 2.(1-p_i)^1 + 3.(1-p_i)^2 + 4.(1-p_i)^3 + \ldots \\
(1-p_i)S &= \qquad 0 \qquad + 1.(1-p_i)^1 + 2.(1-p_i)^2 + 3.(1-p_i)^3 + \ldots \\
\text{On subtracting them,} \quad p_i.S &= (1-p_i)^0 + (1-p_i)^1 + (1-p_i)^2 + (1-p_i)^4 + \ldots \\
p_i.S &= \frac{1}{p_i} \\
S &= \frac{1}{p_i^2}
\end{aligned}$$

Hence, $E[Z_i] = p_i.S = \frac{1}{p_i} = \frac{n}{n-i+1}$.
$E[T] = \sum\limits_{i=1}^{n} E[Z_i] = \sum\limits_{i=1}^{n} \frac{n}{n-i+1} = \sum\limits_{j=1}^{n} \frac{n}{j} = nH_n \leq n(\ln n + 1)$

Solution (b) By Markov's inequality, we know that

$$\begin{aligned}
Pr(T \geq n^2) &\leq \frac{E[T]}{n^2} \\
&\leq \frac{\ln n + 1}{n}
\end{aligned}$$

Hence, for large values of n, T does not exceed n^2 with high probability. Now, let's try to get a better bound with chebysev's inequality $Pr(|T - E[T]| > t) \leq \frac{Var(T)}{t^2}$, where $Var(T)$ is the variance of T.

$$\begin{aligned}
Var(T) &= Var(\sum_{i=1}^{n} Z_i) \\
&= \sum_{i=1}^{n} Var(Z_i) \\
&= \sum_{i=1}^{n} \frac{1-p_i}{p_i^2} \\
&= \sum_{i=1}^{n} n \frac{i-1}{(n-i+1)^2} \\
&= n \sum_{i=1}^{n} \frac{n-i}{i^2} \\
&\leq \sum_{i=1}^{n} \frac{n^2}{i^2} \\
&\leq n^2 . \frac{\pi^2}{6} = \Theta(n^2)
\end{aligned}$$

Hence, $Pr(|T - E[T]| > t) \leq \Theta(\frac{n^2}{t^2})$. So, T lies between $n.H_n - O(n)$ and $n.H_n + O(n)$ with high probability.

Note: To get the tightest bound possible, use the fact that each Z_i is a sub-exponential random variable, and hence T is a sub-gamma variable. This gives $O(n \log n) - O(n \log \frac{1}{\delta}) \leq T \leq O(n \log n) + O(n \log \frac{1}{\delta})$ with probability $1 - \delta$.

3 Balls and Bins

Given n balls thrown uniformly and independently at random into n bins. Let X_i be the random variable denoting the number of balls that land in bin i.

(a) Find $E[X_i]$.

(b) Find a good Upper bound on $E[\max_i X_i]$.

(c) Find a good Lower bound on $E[\max_i X_i]$.

(d) Find the expected number of empty bins and concetration bounds for it.

Solution (a) Let X_{ij} be the indicator random variable denoting whether ball j falls into bin i. So, $X_i = \sum\limits_{j=1}^{n} X_{ij}$. We know that, $E[X_{ij}] = \frac{1}{n}$. Hence $E[X_i] = \sum\limits_{j=1}^{n} E[X_{ij}] = 1$

Solution (b) Let $Y = \max_i X_i$. In order to compute $E[Y]$, we use the following result.

$$E[Y] = \sum_k Pr(Y \geq k)$$

Hence it suffices to compute $Pr(\max_i X_i \geq k)\ \forall k$.
We know that,

$$Pr(X_i = k) = \binom{n}{k}(\frac{1}{n})^k.(1 - \frac{1}{n})^{n-k}$$

By Sterling's approximation which states that $(\frac{n}{k})^k \leq \binom{n}{k} \leq (\frac{en}{k})^k$,

$$\begin{aligned} Pr(X_i = k) &\leq (\frac{en}{k})^k.(\frac{1}{n})^k.1 \\ &= (\frac{e}{k})^k \end{aligned} \tag{1}$$

and since $(1 - \frac{1}{n})^{n-k} \geq \frac{1}{e}$,

$$\begin{aligned} Pr(X_i = k) &\geq (\frac{n}{k})^k.(\frac{1}{n})^k.(1 - \frac{1}{n})^{n-k} \\ &\geq (\frac{n}{k})^k.(\frac{1}{n})^k.(\frac{1}{e}) \\ &= \frac{1}{e.k^k} \end{aligned} \tag{2}$$

Now let's compute $Pr(\max_i X_i \geq k)$, which is equal to the probability that any of the X_i's exceed k. This can be calculated using union bound as follows.

$$
\begin{aligned}
Pr(\max_i X_i \geq k) &= Pr(\bigcup_{i=1}^{n} X_i \geq k) \\
&\leq \sum_{i=1}^{n} Pr(X_i \geq k) \quad \text{Using union bound} \\
&= n.Pr(X_i \geq k) \quad \text{for any } i \\
&\leq n. \sum_{k' \geq k} Pr(X_i = k') \\
&\leq n. \sum_{k' \geq k} (\frac{e}{k'})^{k'} \quad \text{From Equation 1} \\
&\leq n. \sum_{k' \geq k} (\frac{e}{k})^{k'} \\
&\leq n. \frac{(\frac{e}{k})^k}{1 - \frac{e}{k}} \quad \left(\text{As } \sum_{k' \geq k} p^{k'} \leq \sum_{k'=k}^{\infty} p^{k'} = \frac{p^k}{1-p}\right) \\
&\leq 2n.(\frac{e}{k})^k \text{ for } k \geq 6
\end{aligned}
\tag{3}
$$

As stated above,

$$
\begin{aligned}
E[\max_i X_i] &= \sum_{k=1}^{n} Pr(\max_i X_i \geq k) \\
&\leq \sum_{k=1}^{n} min(1, 2n(\frac{e}{k})^k) \quad \text{From Inequality 3}
\end{aligned}
$$

Let k^* be the minimum value of k such that $2n\left(\frac{e}{k}\right) < 1$. Then,

$$
\begin{aligned}
E[\max_i X_i] &\leq \sum_{k=1}^{n} min(1, 2n(\frac{e}{k})^k) \quad \text{From Inequality 3} \\
&\leq (k^* - 1) + \sum_{k=k^*}^{n} 2n(\frac{e}{k})^k \\
&\leq (k^* - 1) + 2\left(2n(\frac{e}{k^*})^{k^*}\right) \\
&\leq k^* - 1 + 2 \\
&\leq k^* + 1
\end{aligned}
$$

Now let's compute the value of k^*.

$$
2n(\frac{e}{k^*})^{k*} < 1 \implies (\frac{e}{k^*})^{k*} < \frac{1}{2n} \implies k^* \log(\frac{k^*}{e}) < \log 2n
$$

It's easy to see that $k^* = \Theta\left(\frac{\log n}{\log \log n}\right)$ satisifies the above inequality. Hence, $E[\max_i X_i] \leq \Theta\left(\frac{\log n}{\log \log n}\right)$.

Solution (c) Using Stirling's approximation, a lower bound on $\mathbb{P}(X_i \geq k)$ is,

$$\begin{aligned}\mathbb{P}(X_i \geq k) &\geq \left(\frac{n}{k}\right)^k \frac{1}{n^k}\left(1-\frac{1}{n}\right)^{n-k}, \\ &\geq \left(\frac{1}{k}\right)^k e^{-1}.\end{aligned}$$

For $k = \frac{\log n}{3 \log\log n}$, this gives $\mathbb{P}(X_i \geq k) \geq \frac{1}{en^{1/3}}$.

The variables X_i, $i \in [n]$ are negatively associated (formally defined in the next section). One implication of negative association among random variables is:

$$\mathbb{P}(X_i \geq k \mid X_1 < k, \cdots, X_{i-1} < k) \geq \mathbb{P}(X_i \geq k) \, \forall \, i \in [n]$$

Hence the probability that all X_i are less than k is given by,

$$\begin{aligned}\mathbb{P}(\cap_{i=1}^n X_i < k) =& \mathbb{P}(X_1 < k) \, \mathbb{P}(X_2 < k | X_1 < k) \cdots \mathbb{P}(X_n < k | X_1 < k, X_2 < k, \cdots, X_{n-1} < k), \\ \leq& \left(1 - \frac{1}{en^{1/3}}\right)^n = e^{-\Omega(n^{2/3})} \quad \square.\end{aligned}$$

From part (b) and (c), we conclude that with high probability, the maximum number of balls in a bin is within a constant factor of the expected maximum number of balls.

3.1 Number of Empty Bins

An extension to the above problem is to ask how many bins are empty in expectation, and how well they concentrate around the expectation.

Let $Z_i = 1_{X_i=0}$ indicate the event that bin i is empty, and $Z = \sum_{i=1}^n Z_i$ denote the number of empty bins. In this case,

$$\mathbb{P}\left[Z_i = 1\right] = \left(1 - \frac{1}{n}\right)^n \leq \frac{1}{e} \Rightarrow \mathbb{E}\left[Z\right] \leq \frac{n}{e}.$$

If X_i are independent variables, then Z_i would also be independent Bernoulli variables, and we could apply a Chernoff bound to obtain a high probability estimate of Z. However, negative association of $\{Z_i\}_{i=1}^n$ allows for application of Chernoff-Hoeffding bounds on Z, which takes the form,

$$\mathbb{P}\left[|Z - \mathbb{E}\, Z| \geq t\right] \leq 2e^{-2t^2/n}.$$

Hence $Z \in \left[\mathbb{E}\, Z - O\left(\sqrt{n \log \frac{2}{\delta}}\right), \mathbb{E}\, Z + O\left(\sqrt{n \log \frac{2}{\delta}}\right)\right]$ with probability $1 - \delta$.

4 Negatively Associated Random Variables

All material for this section can be found in [DD96].

Let $\mathbf{X} := (X_1, X_2, \cdots, X_n)$ be a vector of random variables.

Definition 1. *The random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ are negatively associated, if for every two disjoint index sets, $I, J \subseteq [n]$,*

$$\mathbb{E}\left[f(\mathbf{X}_I)g(\mathbf{X}_J)\right] \leq \mathbb{E}\left[f(\mathbf{X}_I)\right]\mathbb{E}\left[g(\mathbf{X}_J)\right],$$

for all non-decreasing functions f and g.

Negative association implies the following properties:

1. Negative correlation: $\mathbb{E}\left[\prod_{i\in I} x_i\right] \leq \prod_{i\in I}\mathbb{E}\left[x_i\right]$.

   Proof: Let $X_I = (X_{i_1}, X_{i_2}, \cdots, X_{i_k})$ Choose $f(X_{I\setminus i_k}) = \prod_{i\in I, i\neq i_k} x_i$ and $g(X_{i_k}) = x_{i_k}$. This gives $\mathbb{E}\left[\prod_{i\in I} x_i\right] \leq \mathbb{E}\left[\prod_{i\in I, i\neq i_k} x_i\right]\mathbb{E}\left[x_{i_k}\right]$. Apply induction over all k variables.

2. Negative orthants: $\mathbb{P}\left[X_i \geq t_i \forall i \in I\right] \leq \prod_{i\in I}\mathbb{P}\left[X_i \geq t_i\right]$.

   Proof: Similar to negative correlation, choose f, g to be indicator functions and apply induction.

3. Chernoff-Hoeffding Bounds: $\mathbb{E}\left[e^{\lambda\sum_i x_i}\right] \leq \prod_i \mathbb{E}\left[e^{\lambda x_i}\right]$. (Equality holds if all variables are independent)

   Proof: Similar to negative correlation, choose $f(X_{I\setminus i_k}) = \prod_{i\in I, i\neq i_k} e^{\lambda x_i}, g(X_{i_k}) = e^{\lambda x_{i_k}}$ and apply induction.

4. If $\mathbf{X}$ and $\mathbf{Y}$ are negatively associated individually, and $\mathbf{X}, \mathbf{Y}$ are independent, then $(\mathbf{X}, \mathbf{Y})$ is jointly negatively associated.

5. For disjoint index sets $I_j \subseteq [n]$, let $Y_j = f_j(X_{I_j})$, where f_j are all non-decreasing functions. Then $\{Y_j\}$ are negatively associated.

4.1 Negative Association in Balls and Bins

Lemma 2. *Zero-One Lemma: If $X_1, X_2, \cdots, X_n \in \{0, 1\}$ and $\sum_{i=1}^n X_i = 1$, then $X_1, X_2, \cdots, X_n$ are negatively associated.*

Proof. Without loss of generality, assume $f(\overrightarrow{0}) = 0, g(\overrightarrow{0}) = 0$ (if this is not true, you can add appropriate constants). For disjoint index sets $I, J \subseteq [n]$, either $f(X_I) = 0$ or $g(X_J) = 0$, since the index corresponding to the non zero value cannot be in I and J simultaneously. Hence,

$$\begin{aligned}\mathbb{E}\left[f(X_I)g(X_J)\right] =& 0,\ \mathbb{E}\left[f(X_I)\right] \geq 0,\ \mathbb{E}\left[g(X_J)\right] \geq 0,\\ \Rightarrow \mathbb{E}\left[f(X_I)g(X_J)\right] \leq& \mathbb{E}\left[f(X_I)\right]\mathbb{E}\left[g(X_J)\right].\end{aligned}$$

□

For the balls and bins problem, let $Y_{ij} := 1$ if ball i lands in bin j, and $Y_{ij} = 0$ otherwise. By the Zero-One lemma, $\{Y_{ij}\}_{j=1}^n$ are negatively associated. Property (4) from the previous subsection implies that $\{Y_{ij}\}_{i=1,j=1}^{i=n,j=n}$ are negatively associated.

Since $X_j = \sum_{i=1}^n Y_{ij}$, property (5) from the previous subsection implies that $\{X_j\}_{j=1}^n$ are negatively associated. This also implies that the indicator random variables $Z_j = 1_{X_j=0}$ are negatively associated.

References

[MR95] Rajeev Motwani and Prabhakar Raghavan *Randomized Algorithms* Cambridge University Press, New York, 1995.

[DD96] Devdatt Dubhashi and Desh Ranjan *Balls and Bins: A Study in Negative Dependence* BRICS Report Series, University of Aarhus, 1996.