Lecture Notes CS:5350 Coupon Collector, Markov's Inequality, and Max Cut
Lecture 18: Oct 24, 2019
Scribe: Andrew Marmaduke

1 The Coupon Collectors Problem

In the Coupon Collectors Problem we have a cereal company who is placing coupons in their cereal boxes. In particular there are

- n coupons (these are distinct coupons),
- where each cereal box has a coupon chosen uniformly at random.

Our goal is to collect the minimum number of cereal boxes so that we have all the coupons. Let X denote the number of cereal boxes collected until we have all n coupons. In Figure 1 we see a picture of a general series of events when trying to collect all n coupons.

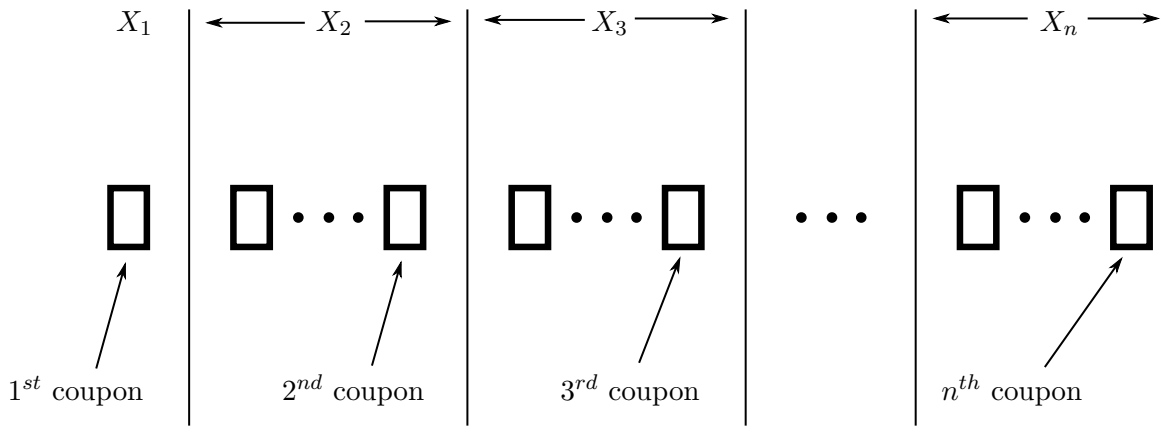

Figure 1: Picture of a general series of events when collecting coupons.

The figure shows that we can break down the variable X into n variables X_i for $1 \leq i \leq n$ that denote the number of cereal boxes bought after coupon $i-1$ was collected until coupon i is collected. Thus, X_1 denotes the number of cereal boxes collected until the first new coupon (trivially this is equal to 1 because the first box has the first new coupon). Then, X_2 denotes the number of cereal boxes collected until the second new coupon, not recounting the number boxes needed to collect the previous new coupon, and so forth.

This gives us the following decomposition: $X = \sum_{i=1}^{n} X_i$.

Now, by linearity of expectation $\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i]$. In order to compute this we can now focus on $\mathbb{E}[X_i]$. Notice that, the very first box contains the first unique coupon, thus $X_1 = 1$. Now we consider X_2. The probability that it will take one additional box to get the second coupon is $\frac{n-1}{n}$ because there are $n-1$ unique coupons left of the n total coupons. The probability that it will

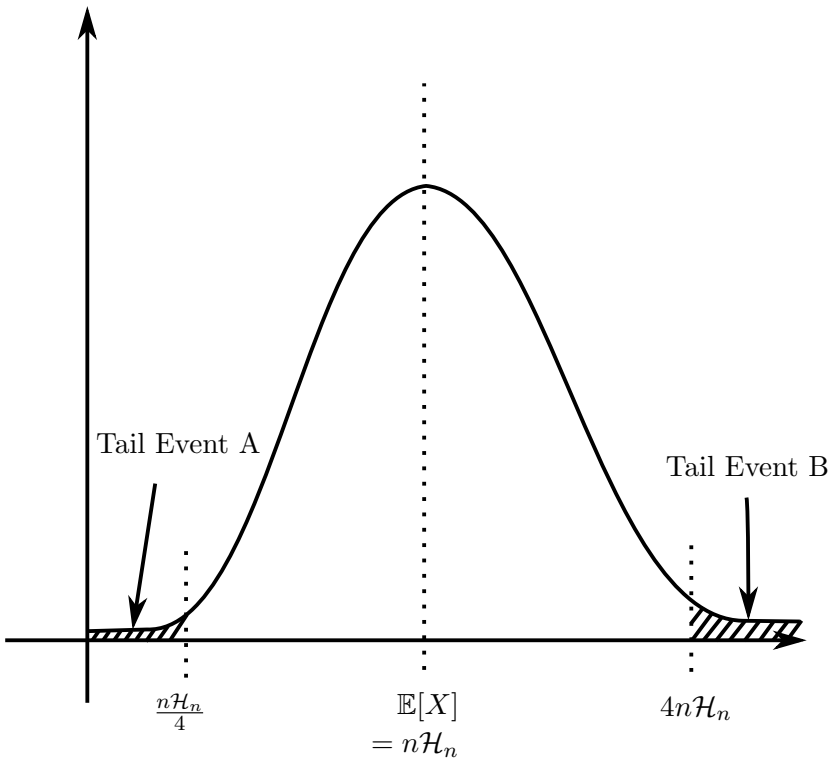

Figure 2: A sketch of the coupon collector distribution.

take two additional boxes to get the second coupon is $\frac{n-1}{n}(1-\frac{n-1}{n})$. In general, the probability that it takes j boxes to get the second coupon is $\frac{n-1}{n}(1-\frac{n-1}{n})^{j-1}$. Thus, X_2 follows a geometric distribution with success probability $\frac{n-1}{n}$ which is written: $X_2 \sim Geom(\frac{n-1}{n})$.

Moreover, for any X_i it has the probability that the *next* additional box is $\frac{n-i+1}{n}$ where $n-i+1$ are the remaining unique coupons of the n total number of coupons. As in the X_2 case, the probability that it will take two additional boxes is $\frac{n-i+1}{n}(1-\frac{n-i+1}{n})$. This generalizes like before such that the probability that it will take j additional boxes is $\frac{n-i+1}{n}(1-\frac{n-i+1}{n})^{j-1}$ Thus, all X_i follow a geometric distribution with success probability $\frac{n-i+1}{n}$, denoted $X_i \sim Geom(\frac{n-i+1}{n})$.

Now we have $\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \frac{n}{n-i+1}$. We can re-index the sum by $k = n-i+1$ to get $\sum_{k=1}^{n} \frac{n}{k} = n\sum_{k=1}^{n} \frac{1}{k}$ and finally notice that the sum $\sum_{k=1}^{n} \frac{1}{k}$ is the n^{th} Harmonic number, $\mathcal{H}_n$. Therefore, $\mathbb{E}[X] = n\mathcal{H}_n$ which we know is $\Theta(n \log n)$.

2 Tail Events

Although we can determine how many cereal boxes in expectation we need in order to collect all n coupons it may be the case that the distribution of the number of cereal boxes has an undesirable tail. An example of a potential distribution is depicted in Figure 2. In this picture we mark the expectation with the one we computed for the Coupon Collectors Problem.

However, we want to know what is the probability that the value X takes on is greater than, say $4n\mathcal{H}_n$? Although this upper bound for us is chosen rather arbitrarily it gets to the point: the shaded regions in the figure are called Tail Events and they represent potential values of X whose probability we would like to either bound or understand in some way.

In particular, in Figure 2 we do not care about Tail Event A because this means that collect fewer boxes then we expected. In some sense, this is a *good* tail event. To the contrary, the Tail Event B shows that there is some chance that we could collect many more boxes than what we expect, but how likely is it?

3 Markov's Inequality

To accomplish our previous goal, we first need tools to bound the probability of ranges of values for a random variable, one such tool is Markov's Inequality.

Theorem. ***Markov's Inequality***. *Let X be a nonnegative integer random variable. Then $Pr[X \geq c] \leq \dfrac{\mathbb{E}[X]}{c}$ or $Pr[X \geq c \cdot \mathbb{E}[X]] \leq \frac{1}{c}$.*

Notice that the second form is a consequence of the first. We will prove this theorem two different ways.

Proof. First. We unravel the definition of expectation and split the sum at the arbitrary value c:

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x \cdot Pr[X = x] = \sum_{x=0}^{c-1} x \cdot Pr[X = x] + \sum_{x=c}^{\infty} x \cdot Pr[X = x]$$

Now, we pick the smallest value of x for each sum independently, this is 0 for the first sum and c for the second, this gives us a lower bound for the expectation:

$$\begin{aligned}\mathbb{E}[X] &\geq \sum_{x=0}^{c-1} 0 \cdot Pr[X = x] + \sum_{x=c}^{\infty} c \cdot Pr[X = x] \\ &= c \sum_{x=c}^{\infty} Pr[X = x] \\ &= c \cdot Pr[X \geq c].\end{aligned}$$

Now we have $Pr[X \geq c] \leq \dfrac{\mathbb{E}[X]}{c}$ as desired. □

Proof. Second. Let

$$Y = \begin{cases} 0 & \text{if } X < c \\ c & \text{if } X \geq c \end{cases}$$

Note that Y is a random variable. Unraveling the definition of expectation of Y we have $\mathbb{E}[Y] = 0 \cdot Pr[X < c] + c \cdot Pr[X \geq c] = c \cdot Pr[X \geq c]$. Finally, notice that $Y \leq X$ by definition and thus $\mathbb{E}[Y] \leq \mathbb{E}[X]$. Therefore, $Pr[X \geq c] \leq \frac{\mathbb{E}[X]}{c}$ □

With this new tool we can see that, for the Coupon Collector Problem, that $Pr[X \geq 4n\mathcal{H}_n] \leq \frac{1}{4}$ using the second form of Markov's Inequality.

4 Maximum Cut

Maximum cut is another problem whose tail events we can attempt to bound using Markov's Inequality. The input of Maximum Cut is a unweighted undirected graph $G = (V, E)$. The output is a global cut $(S, V \setminus S)$ such that the number of edges crossing cut is maximized. Note that the Maximum Cut Problem is NP-hard.

Here is a suggested randomized algorithm for Maximum Cut:

input : a graph $G = (V, E)$
output: a global cut $(S, V \setminus S)$
1 **foreach** $v \in V$ **do**
2 | with probability $\frac{1}{2}$ put v in S otherwise put v in $V \setminus S$
3 **end**
4 **return** $(S, V \setminus S)$

Algorithm 1: Maximum Cut Algorithm

Now we can ask, how good is the cut that this algorithm returns? Let X be the random variable denoting the size of the cut we return. Let X_e be 1 if e crosses $(S, V \setminus S)$ and be 0 otherwise. This setup is depicted in Figure 3.

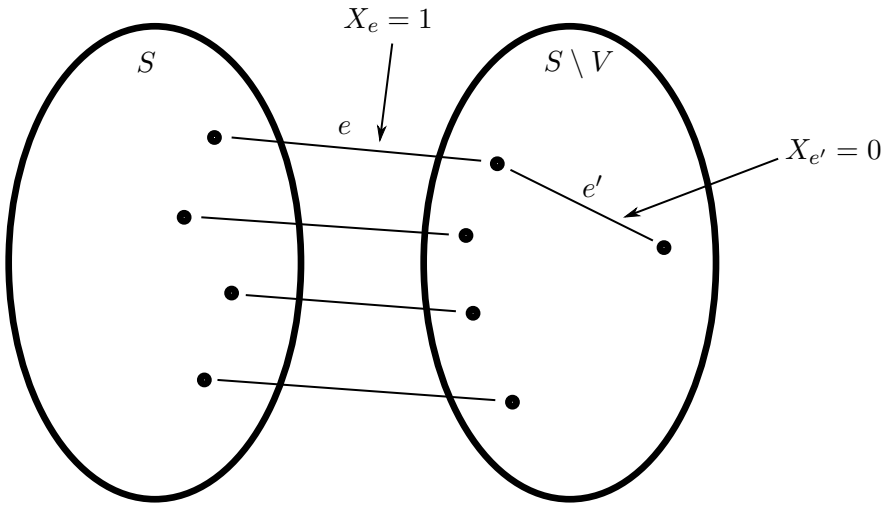

Figure 3: A particular maximum cut returned from the algorithm.

Thus, with this decomposition we have $X = \sum_{e \in E} X_e$. Now by linearity of expectation and definition of characteristic variables we have $\mathbb{E}[X] = \sum_{e \in E} \mathbb{E}[X_e] = \sum_{e \in E} Pr[X_e = 1]$. Now we need to compute $Pr[X_e = 1]$.

Let $e = \{u, v\}$. Notice that $X_e = 1$ if and only if $(v \in S \wedge u \in V \setminus S) \vee (v \in V \setminus S \wedge u \in S)$. Now, we use disjointedness of the possibilities and independence of the coin flips to break this event into pieces:

$$Pr[X_e = 1] = Pr[v \in S] \cdot Pr[u \in V \setminus S] + Pr[v \in V \setminus S] \cdot Pr[u \in S]$$

Now, because the coin flip is at probability $\frac{1}{2}$ we have that $Pr[X_e = 1] = \frac{1}{2}$. Next, we compute the expectation of X, $\mathbb{E}[X] = \sum_{e \in E} \frac{1}{2} = \frac{m}{2}$, where m is the number of edges.

In the next class we will show using Markov's Inequality that $Pr[X \leq \frac{m}{4}] \leq \frac{2}{3}$ and use it to get and approximation algorithm for Maximum Cut that produces at $\frac{1}{4}$-approximation with high probability.