**COMPSCI 330: Design and Analysis of Algorithms** March 10, 2016

# Random Processes

*Lecturer: Debmalya Panigrahi* *Scribe: Tianqi Song*

## 1 Overview

We introduce random processes by the following examples: Coupon Collector, Birthday Paradox, Shared Resource Contention and Balls in Bins.[1]

## 2 Preliminary

### 2.1 Geometric Random Variables

Suppose we have a coin which, when flipped, lands on heads with probability $p$ (otherwise, it lands on tails). We flip this coin repeatedly until we get a heads. How many flips will we perform, in expectation? Consider any sequence of flips:

$$T \quad T \quad T \quad H \quad H \quad T \quad \cdots \tag{1}$$

What is the probability that it takes $i$ trials to see the first head? For that occur, we must *fail* the first $(i-1)$ trials, and *succeed* exactly on the $i$th. Let $X$ be a random variable equal to the first $i$ with a heads. Since the trials are independent:

$$\Pr(X = i) = \Pr(\text{First } H \text{ is on trial } i) = (1-p)^{i-1}(p) \tag{2}$$

We call variables distributed in this way (independent trials until success)*geometric* random variables. We say that $X$ is *drawn* from a geometric distribution with parameter $p$.

$$X \sim \text{Geo}(p) = \text{Geo}(1-p) \tag{3}$$

Using the definitions, one can show that the expectation of a geometric random variable is:

$$\mathbb{E}[X] = \Sigma_{i=1}^{\infty} i(1-p)^{i-1}p \tag{4}$$

We have

$$\mathbb{E}[X] - (1-p)\mathbb{E}[X] = \Sigma_{i=0}^{\infty}(1-p)^{i}p = p\frac{1}{p} \tag{5}$$

Therefore

$$\mathbb{E}[X] = \frac{1}{p} \tag{6}$$

The expected number of flips before we see a heads is $1/p$.

[1] Some materials are from a note by Roger Zou for this class in Fall 2014 and a note by Allen Xiao for COMPSCI 532 in Fall 2015.

# 3 Coupon Collector Problem

## 3.1 Problem Statement

There are $n$ types of coupons. In each draw, a coupon is picked at random and each type of coupon has the same probability to be picked. What is the expected number of draws to get all $n$ types of coupons ?

## 3.2 Analysis

Let $X_i$ be the random variable representing the number of draws to get the $i^{th}$ distinct coupon after we have got $(i-1)$ distinct coupons. We have $X_i$ follows a geometric distribution and the probability of success for each trial is:

$$p_i = \frac{n-i+1}{n} \tag{7}$$

The expectation of $X_i$ is:

$$\mathbb{E}[X_i] = \frac{1}{p_i} \tag{8}$$

$$= \frac{n}{n-i+1} \tag{9}$$

Let $X$ be the number of draws to get all $n$ types of coupons. We have:

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] \tag{10}$$

$$= \sum_{i=1}^{n} \frac{n}{n-i+1} \tag{11}$$

$$= n\left(\frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{2} + \frac{1}{1}\right) \tag{12}$$

$$= \Theta(n \log n) \tag{13}$$

# 4 The Birthday Paradox

## 4.1 Problem Statement

There are $k$ people and each person has a random birthday from $n$ days. How large should $k$ be such that the chance that two people have the same birthday is at least $\alpha$?

## 4.2 Analysis

Let $X$ be the event that every person has a distinct birthday. The $i$th person has $(n-i)$ options to have a distinct birthday and then the probability to have a distinct birthday is $\frac{n-i}{n}$, where $i \in \{0,1,2,...,k-1\}$. We have:

$$Pr[X] = \frac{n}{n}(\frac{n-1}{n})(\frac{n-2}{n})\dots(\frac{n-(k-1)}{n}) \tag{14}$$

$$= 1(1-\frac{1}{n})(1-\frac{2}{n})\dots(1-\frac{k-1}{n}) \tag{15}$$

Because $e^{-x} \approx 1-x$ when $x$ is small:

$$Pr[X] \approx 1 \cdot e^{-\frac{1}{n}} \cdot e^{-\frac{2}{n}} \dots e^{-\frac{k-1}{n}} \tag{16}$$

$$\approx e^{-\frac{\frac{(1+(k-1))(k-1)}{2}}{n}} \tag{17}$$

$$= e^{-\frac{k(k-1)}{2n}} \tag{18}$$

We just need to solve $Pr[X] \approx e^{-\frac{k(k-1)}{2n}} \leq 1-\alpha$. For example, when $\alpha = 0.5$ and $n = 365$, we have $k \geq 23$.

# 5 Shared Resource Contention

## 5.1 Problem Statement

We have $n$ processes $P_1, P_2, ..., P_n$ that can access a shared resource $S$. Each process tosses a random coin with success probability $\frac{1}{n}$. If the coin returns true for process $P_i$, then $P_i$ tries to access the resource. If $P_i$ is the only process that is trying to access the resource, then $P_i$ accesses the resource. If there is contention, then no process accesses the resource. The goal is to determine the expected number of rounds required for all processes to access the resource.

## 5.2 Analysis

Let $X_i$ be the number of rounds needed for the $i^{th}$ process to access $S$, after $(i-1)$ processes have accessed $S$. Let $p_i$ be the probability that a new process accesses the resource in a round, after $(i-1)$ processes have accessed $S$. Then we have:

$$p_i = (n-i+1)\frac{1}{n}\left(1-\frac{1}{n}\right)^{n-1} \tag{19}$$

Note that $\frac{1}{n}\left(1-\frac{1}{n}\right)^{n-1}$ is the probability that an $i$-th process returns true with the rest being false, and $(n-i+1)$ is the number of possible processes.

**Fact 1.**

$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^{n} = \lim_{n\to\infty}\left(1-\frac{1}{n}\right)^{n-1} = \frac{1}{e} = \Theta(1) \tag{20}$$

Therefore:

$$p_i = \frac{n-i+1}{n}\Theta(1) \tag{21}$$

$X_i$ is a geometric random variable, then we have:

$$\mathbb{E}[X_i] = \frac{1}{p_i} \tag{22}$$

$$= \frac{n}{n-i+1}\Theta(1) \tag{23}$$

Again, similar to the coupon collector problem, let $X$ be the number of rounds:

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] \tag{24}$$

$$= \Theta(n \log n) \tag{25}$$

# 6 Balls in Bins

## 6.1 Problem Statement

Suppose we have $n$ balls and $m$ bins, and we toss the balls (uniformly at random) into the bins. This is called a *balls and bins* process, and is a popular model in practice. For example, hashing algorithms will often use a balls and bins process for analyzing collisions. Let $X_i$ be a random variable for the number of balls in the $i$th bin. We ask:

1. How large should $n$ be before every bin is filled in expectation ($\mathbb{E}[X_i] \geq 1$)?
2. What is $\mathbb{E}[X_i]$?
3. When $m = n$, what is the maximum load over all the bins? If $X = \max_i X_i$, what is $\mathbb{E}[X]$?

## 6.2 Analysis

(1) is actually a restatement of the coupon collector problem, so the answer is $n = \Theta(m \log m)$. For (2), notice that the expectation is symmetric across $i$.

$$\mathbb{E}[X_1] = \mathbb{E}[X_2] = \cdots = \mathbb{E}[X_n] \tag{26}$$

Now, since we throw $n$ balls,

$$\sum_{i=1}^{m} \mathbb{E}[X_i] = n \tag{27}$$

It follows that:

$$\mathbb{E}[X_i] = \frac{n}{m} \tag{28}$$

Finally, for (3), the probability a fixed bin has more than $k$ balls is at least the probability that, for $k$ of the $n$ balls, they all landed in this bin. These trials are independent, so we can say:

$$\Pr(X_i \geq k) \leq \binom{n}{k} \left(\frac{1}{n}\right)^k \tag{29}$$

We can apply an inequality known as *Stirling's approximation* to give a tight bound on $\binom{n}{k}$.

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k \tag{30}$$

This gives us:

$$\Pr(X_i \geq k) \leq \left(\frac{en}{k}\right)^k \left(\frac{1}{n}\right)^k = \left(\frac{e}{k}\right)^k \tag{31}$$

The definition of $X$ gives us:

$$\Pr(X \geq k) = \Pr(\exists i \mid X_i \geq k) = \Pr\left(\bigcup_{i=1}^{n} (X_i \geq k)\right) \tag{32}$$

Applying the union bound:

$$\Pr(X \geq k) \leq \sum_{i=1}^{n} \Pr(X_i \geq k) \tag{33}$$

$$\leq \sum_{i=1}^{n} \left(\frac{e}{k}\right)^k \tag{34}$$

$$= n\left(\frac{e}{k}\right)^k \tag{35}$$

We know that $X \leq n$. We can divide the outcomes of $X$ into cases where no $X_i \geq k$, and when at least one $X_i \geq k$. Taking the expectation of $X$:

$$\mathbb{E}[X] \leq k\Pr(X < k) + n\Pr(X \geq k) \tag{36}$$

$$\leq n \cdot n\left(\frac{e}{k}\right)^k + k \tag{37}$$

We will choose $k$ in order to remove the first half of the sum. Asymptotically, this will be the same as if we had tried to optimize for $k$.

$$n^2\left(\frac{e}{k}\right)^k \leq 1 \tag{38}$$

$$\implies \left(\frac{e}{k}\right)^k \leq \frac{1}{n^2} \tag{39}$$

$$\implies \left(\frac{k^k}{e^k}\right) \geq n^2 \tag{40}$$

$$\implies k^k \sim poly(n) \tag{41}$$

$$\implies k\log k = O(\log n) \tag{42}$$

How can we solve for $k$ now? It turns out one natural solution to this expression is $k = O(\log n/\log\log n)$.

$$k\log k = \frac{\log n}{\log\log n} \cdot \log\left(\frac{\log n}{\log\log n}\right) \tag{43}$$

$$= \frac{\log n}{\log\log n} \cdot (\log\log n - \log\log\log n) \tag{44}$$

The rightmost term is vanishingly small, so:

$$k \log k = \Theta\left(\frac{\log n}{\log \log n} \cdot \log \log n\right) \tag{45}$$

$$= \Theta(\log n) \tag{46}$$

Similar to how $\log n$ solves $2^k = poly(n)$, $k^k = poly(n)$ is solved by $k = O(\log n / \log \log n)$. Using that choice of $k$:

$$\Pr\left(X_i \geq \frac{e \log n}{\log \log n}\right) \leq \left(\frac{\log \log n}{\log n}\right)^{\frac{e \log n}{\log \log n}} \leq \frac{1}{n^2} \tag{47}$$

This makes $\mathbb{E}[X]$:

$$\mathbb{E}[X] = n \cdot \frac{n}{n^2} + k = O\left(\frac{\log n}{\log \log n}\right) \tag{48}$$

In fact, this is tight; there is a lower bound example which matches this. Additionally, notice for $X$:

$$\Pr\left(X \geq \frac{e \log n}{\log \log n}\right) \leq \frac{1}{n} \tag{49}$$

When some event fails with probability $\Omega(1/n)$, we typically say it succeeds *with high probability*. Here, we say that with high probability $X = O((\log n)/(\log \log n))$. These types of statements are *tail bounds*, and provide a formal notion of controlling a distribution. Here, we used a tail bound to obtain a bound on the expectation. Often, we want to do the reverse: use expectation to obtain a tail bound ("value stays close to the mean with high probability").