Problem Set 4 Andrew Koulogeorge, Eric Richardson
COSC 34: Randomized Algorithms Date: December 20, 2024

Collaboration Statement: Andrew Koulogeorge and Eric Richardson

Problem 4: Birthday Paradox with Non-Uniform Bins.

Solution.
Let us consider the balls and bins paradigm (m balls n bins) where each ball now has a probability of p_i of landing in the ith bin $\big(\sum_{i=1}^{n} p_i = 1\big)$. The sample space Ω in this random process contains tuples ω of length m where each entry $\omega_i \in [n]$. Each ball is still thrown independently.

(a) Let us define $Z_{i,j}$ to be a Bernoulli Random Variable which equals 1 if the ith and the jth ball land in the same bin. Since $Z_{i,j}$ is Bernoulli, $\mathbf{E}[Z_{i,j}] = \mathbb{P}[Z_{i,j} = 1]$. Reasoning off the definition of the sample space, for a given sample $\omega \in \Omega$, $Z_{i,j} = 1 \iff \omega_i = \omega_j$. There are n bins in which the ith and jth ball can share, and since the probability of landing in a given bin is non-uniform, we must consider each bin individually. For the two balls to collide, they can both land in bin 1 or bin 2 or ... or bin n:

$$\mathbf{E}[Z_{i,j}] = \mathbb{P}[Z_{i,j} = 1] = \sum_{i=1}^{n} p_i^2$$

As a sanity checker, note that when we are in the uniform balls and bins case, $p_i = \frac{1}{n} \,\forall\, n$ our formula produces that $\mathbb{P}[Z_{i,j} = 1] = \frac{1}{n}$ which is indeed the probability two balls thrown u.a.r land in the same bin.

(b) Define $X^{(m)} = \sum_{1 \le i < j \le m} Z_{i,j}$ where i and j consider all pairs of balls (i, j) thrown. We can apply linearity of expectation:

$$\mathbf{E}\left[X^{(m)}\right] = \sum_{1 \le i < j \le m} \mathbf{E}[Z_{i,j}] = \binom{m}{2} \sum_{i=1}^{n} p_i^2$$

where the last equality follows from there being $\binom{m}{2}$ total pairs.

(c) Let E_m be the event that no balls landed in the same bin after m balls were thrown. Note that if no balls land in the same bin, $Z_{i,j} = 0 \,\forall\, (i, j)$ pairs and it follows that:

$$\mathbb{P}[E_m] = \mathbb{P}\left[X^{(m)} = 0\right]$$

(d) Now that we have a way to connect the event of getting no collisions on m balls thrown to our random variable $X^{(m)}$ taking on a particular value, we will apply Markov's inequality to bound $\mathbb{P}[E_m]$ to at least $\frac{3}{4}$ when $m = \frac{c_1}{||p||}$ where $c_1 = \frac{\sqrt{2}}{2}$ and $||p||$ is the $L2$ norm of the vector of probabilities $p = (p_1, p_2, ..., p_n)$

$$\mathbb{P}\left[X^{(m)} = 0\right] + \mathbb{P}\left[X^{(m)} \geq 1\right] = 1 \implies \mathbb{P}\left[X^{(m)} = 0\right] = 1 - \mathbb{P}\left[X^{(m)} \geq 1\right]$$

Thus, in order to show $\mathbb{P}\left[X^{(m)} = 0\right] \geq \frac{3}{4}$ we need to show that $\mathbb{P}\left[X^{(m)} \geq 1\right] \leq \frac{1}{4}$.

$$\mathbb{P}\left[X^{(m)} \geq 1\right] \leq \mathbf{E}\left[X^{(m)}\right] = \binom{m}{2}||p||^2 \leq \frac{m^2}{2}||p||^2$$

where the above follows from Markov's inequality. Thus, it suffices to show that:

$$\frac{m^2}{2}||p||^2 \leq \frac{1}{4}$$

$$\implies m \leq \frac{1}{\sqrt{2}}\frac{1}{||p||}$$

Note that when p is the uniform vector, we see that:

$$||p|| = \sqrt{\sum_{i=1}^{n} \frac{1}{n^2}} = \frac{1}{\sqrt{n}}$$

$$\implies \frac{1}{||p||} = \sqrt{n}$$

$$\implies m \leq \frac{\sqrt{2}}{2}\sqrt{n}$$

where $\frac{\sqrt{2}}{2} < 1$ which aligns with our intuition. In other words, if we throw a less than a small constant times $\sqrt{n}$ balls at n bins, the probability we no getting no collisions is somewhat high.

(e) Note that $X^{(m)}$ is the sum of pairwise independent random variables $X_{i,j}$. This was shown in the quiz; knowing that two balls land in the same bin does not tell you any information about where another ball is going to land, because knowing that two balls landed in the same bin does not tell you which bin they landed in. Also note that these random variables are **not** mutually independent, since if you knew that $X_{i,j} = 1$ and $X_{j,k} = 1$, then you would know for certain that $X_{i,k} = 1$.
Recall that we wish to upper bound $\mathbb{P}[E_m] = \mathbb{P}\left[X^{(m)} = 0\right]$. Let $m = \frac{c_2}{||p||}$ where $c_2 = \sqrt{2 \ln 4}$. Recall that $X^{(m)}$ is the sum of $\binom{m}{2}$ pairwise independent Bernoulli

random variables where $X_{i,j} = 1$ if the ith and jth ball land in the same bin. Observe that $X^{(m)} = 0 \iff X_{i,j} = 0 \,\forall\, (i,j)$ pairs:

$$\mathbb{P}\left[X^{(m)} = 0\right] = \mathbb{P}\left[X_{i,j} = 0 \,\forall\, (i,j)\right]$$

Note that due to pairwise independence, we can express the right side as a product of probabilities since knowing that any collection of balls (i,j) do not land in the same bin does not change our belief about where another pair of balls will land, even if that other pair of balls has 1 ball in common with (i,j)

$$\mathbb{P}\left[X_{i,j} = 0 \,\forall\, (i,j)\right] = \prod_{(i,j)} \mathbb{P}\left[X_{i,j} = 0\right]$$

Now, fix a pair of balls (i,j). We shall compute the probability that (i,j) do not fall in the same bin by computing 1 minus the probability they do fall in the same bin. There are n bins in which the two balls can land into such that they share a bin, and the probability of landing in the ith bin is given by p_i. Thus,

$$\mathbb{P}\left[X_{i,j} = 1\right] = \sum_{i=1}^{n} p_i^2 = ||p||^2$$

$$\implies \mathbb{P}\left[X_{i,j} = 0\right] = 1 - ||p||^2$$

$$\prod_{(i,j)} \mathbb{P}\left[X_{i,j} = 0\right] = \left(1 - ||p||^2\right)^{\binom{m}{2}} \leq \left(1 - ||p||^2\right)^{\frac{m^2}{2}}$$

Now we will use our assumption that $m = \dfrac{c_2}{||p||}$ and apply everyone favorite TCS inequality! $1 - x \leq \exp -x \,\forall\, x$:

$$\leq \exp\left(-||p||^2\right)^{\frac{m^2}{2}} = \exp\frac{-c_2^2}{2} = \exp\frac{-2\ln 4}{2} \leq \frac{1}{4}$$

When plugging in the uniform distribution for the probability vector p, we see that $m = \sqrt{2\ln 4}\sqrt{n}$ where $\sqrt{2\ln 4} \approx 2$. This follows our intuition learned from class; if we want to throw m balls into n bins and get a collision with somewhat (at least $\frac{3}{4}$) high chance, we need to throw a constant times $\sqrt{n}$ balls. In this case, the constant is larger then it was in the case where we were trying to avoid collisions. Major take away is that most of the probability mass relating to collisions occurs around $\sqrt{n}$ □

Problem 9: Number of Empty Bins.

Solution. Suppose that we threw n balls into n bins uar and let X be a random variable denoting the number of empty bins at the end of the experiment. Note that our sample space Ω contains length n tuples ω where each element of the tuple ω_i represents the number of balls in the ith bin. X is counting the number entries $\omega_i \in \omega$ which equal 0

(a) We can express X as a sum of indicator random variables where $X_i = 1$ if the ith bin is empty after the experiment is over. Note that X_i is Bernoulli random variable where $\mathbb{P}\left[X_i\right]$ is the probability that the ith bin is empty after n balls are thrown which occurs when all n independently thrown balls land in a different bin than bin i:

$$X = \sum_{i=1}^{n} X_i$$

$$\mathbf{E}\left[X\right] = \sum_{i=1}^{n} \mathbf{E}\left[X_i\right] = n\mathbf{E}\left[X_i\right]$$

$$\mathbf{E}\left[X_i\right] = \mathbb{P}\left[X_i = 1\right] = \left(1 - \frac{1}{n}\right)^n$$

$$\implies \mathbf{E}\left[X\right] = n * \left(1 - \frac{1}{n}\right)^n$$

Note that as $n \to \infty$, $\left(1 + \frac{x}{n}\right)^n \to e^x \implies \mathbf{E}\left[X\right] \to \frac{n}{e}$ so $\mathbf{E}\left[X\right] = cn$ where $c = \frac{1}{e}$ □

(b) Now that we have shown the the expected number of empty bins after n throws is around $\frac{n}{e}$ as $n \to \infty$, we also want to show that the probability mass of X is dense around its expected value. That is, we wish to upper bound the $\mathbb{P}\left[X \geq \frac{n}{2}\right]$ and $\mathbb{P}\left[X \leq \frac{n}{3}\right]$. Let $\mathcal{E}_1$ be the event that $X \geq \frac{n}{2}$ and $\mathcal{E}_2$ be the event that $X \leq \frac{n}{3}$. We can express $\mathbb{P}\left[\mathcal{E}_1\right]$ and $\mathbb{P}\left[\mathcal{E}_2\right]$ by defining 2 binary function f_1 and f_2 that take in input the load random variables $L_1, L_2, ..., L_n$. $f_1(L_1, ..., L_n) = 1$ if at least $\frac{1}{2}$ of the load random random variables are 0 and $f_2(L_1, ..., L_n) = 1$ if at most $\frac{1}{3}$ of the load variables are 0:

$$\mathbb{P}\left[\mathcal{E}_1\right] = \mathbb{P}\left[f_1(L_1, ..., L_n) = 1\right]$$
$$\mathbb{P}\left[\mathcal{E}_2\right] = \mathbb{P}\left[f_2(L_1, ..., L_n) = 1\right]$$

Now we wish to show that each of the events $\mathcal{E}_1$ and $\mathcal{E}_2$ are monotone as a function of m; that is, we wish to show that the the probability of $\mathcal{E}_1$ and $\mathcal{E}_2$ occurring as a function of the number of balls thrown is monotone. Indeed, this is true for both functions. Consider $\mathcal{E}_1$ where $f_1(L_1, ..., L_n) = 1$ when at least $\frac{1}{2}$ of the load vectors are 0, i.e, at least half of the bins are empty. Suppose we create a new experiment where we threw more than m balls. The probability that at least $\frac{1}{2}$ of the bins are empty can only decrease with the number of balls thrown. Every time a new ball is thrown, either it lands in a bin that already has at least 1 ball in it, or it can land in an empty bin, making the function closer to having less than $\frac{n}{2}$ bins be empty. Thus, the probability of $\mathcal{E}_1$ occurring is monotonically decreasing with m. The same reasoning applies to f_2 except $\mathcal{E}_2$ is monotonically increasing with m; as we throw more balls, we are more likely to have the balls fill an empty bin and thus there will be more nonzero load random variables, which will make it more likely to have at most $\frac{1}{3}$ empty bins after m throws.

Now that we have shown these two events are monotonic, we can apply the Poisson approximation theorem for balls and bins with monotone events:

$$\mathbb{P}\left[\mathcal{E}_1\right] = \mathbb{P}\left[f_1(L_1, ..., L_n) = 1\right] \leq 2\mathbb{P}\left[f_1(Z_1, ..., Z_n) = 1\right]$$

$$\mathbb{P}\left[\mathcal{E}_2\right] = \mathbb{P}\left[f_2(L_1, ..., L_n) = 1\right] \leq 2\mathbb{P}\left[f_2(Z_1, ..., Z_n) = 1\right]$$

where Z_i is a Poisson random variable with mean $\frac{n}{n} = 1$. Now we wish to compute the probability that at least $\frac{1}{2}$ of the independent Poisson random variables are 0 and the probability that at most $\frac{1}{3}$ are 0 respectively. The problem with taking the sum of our Poisson random variables and applying Chernoff bound on that new random variable is that the information about which bins were empty is lost. Therefore, we define a new indicator random variable Y_i which equals 1 when $Z_i = 0$ and 0 otherwise. This way $Y = \sum_{i=1}^{n} Y_i$ is the number of Poisson random variables which are zero which is exactly related to our indicator functions f_1 and f_2!

$$\mathbb{P}\left[\mathcal{E}_1\right] = \mathbb{P}\left[f_1(L_1, ..., L_n) = 1\right] \leq 2\mathbb{P}\left[f_1(Z_1, ..., Z_n) = 1\right] = 2\mathbb{P}\left[Y \geq \frac{n}{2}\right]$$

$$\mathbb{P}\left[\mathcal{E}_2\right] = \mathbb{P}\left[f_2(L_1, ..., L_n) = 1\right] \leq 2\mathbb{P}\left[f_2(Z_1, ..., Z_n) = 1\right] = 2\mathbb{P}\left[Y \leq \frac{n}{3}\right]$$

From here, we can apply the vanilla Chernoff bound on Y using the value of $\mathbf{E}\left[Y\right]$:

$$\mathbf{E}\left[Y\right] = \sum_{i=1}^{n} \mathbf{E}\left[Y_i\right] = n\mathbf{E}\left[Y_i\right] = n\mathbb{P}\left[Y_i = 1\right] = n\mathbb{P}\left[Z_i = 0\right] = \frac{n}{e}$$

For bounding $\mathbb{P}\left[Y \geq \frac{n}{2}\right]$ let $\varepsilon_1 = \frac{e}{2} - 1$ and for bounding $\mathbb{P}\left[Y \leq \frac{n}{3}\right]$ let $\varepsilon_2 = 1 - \frac{e}{3}$, where $\varepsilon_1, \varepsilon_2 \in (0, 1)$. It follows from Chernoff that:

$$\mathbb{P}\left[Y \geq \frac{n}{2}\right] \leq \exp -\frac{n\varepsilon_1^2}{3e} = \exp -O(n)$$

$$\mathbb{P}\left[Y \leq \frac{n}{3}\right] \leq \exp -\frac{n\varepsilon_2^2}{2e} = \exp -O(n)$$

Thus, we have bounded the probability mass slightly away from the mean!

$$\implies \mathbb{P}\left[X \geq \frac{n}{2}\right] = \mathbb{P}\left[\mathcal{E}_1\right] \leq 2\exp -\frac{n\varepsilon_1^2}{3e}$$

$$\implies \mathbb{P}\left[X \leq \frac{n}{3}\right] = \mathbb{P}\left[\mathcal{E}_2\right] \leq 2\exp -\frac{n\varepsilon_2^2}{2e}$$

Problem 10: Coupon Collector's Lower Bound.

Solution.
First, observe that the problem can be reduced to the balls-and-bins paradigm. We wish to determine the probability that we receive all n coupons in at most m days, where $m = n\ln n - cn$. This is equivalent to the event that, after throwing $n\ln n - cn$ balls, each of the n bins contains at least one ball. Let $L_i^{(m)}$ denote the number of balls that land in bin i. With $m = n\ln n - cn$, we have that $\sum_{i=1}^{n} L_i^{(m)} = m$. Now, let $\mathcal{E}$ denote the event that no bin is empty after throwing m balls. That is, $\mathcal{E} := \forall i \in [n]$, $L_i^m >= 1$. Now, let f be a function on the load vector $L^{(\overrightarrow{m})} = (L_1^{(m)}, L_2^{(m)}, ..., L_n^{(m)})$ such that $f(L^{(\overrightarrow{m})}) = 1$ if $\forall i \in [n]$, $L_i^m >= 1$ and 0 otherwise. Now, observe that this function is monotonically non-decreasing. That is, for some fixed n, the probability that no bin is empty can only increase as we throw more balls. Thus, we have that

$$\mathbf{Pr}[\mathcal{E}] = \mathbf{Pr}[f(L^{(\overrightarrow{m})}) = 1] \leq 2 \cdot \mathbf{Pr}[f(Z_1, Z_2, ..., Z_n) = 1]$$

where each Z_i is an independent, identical Poisson random variable defined via $Z_i \sim \text{Pois}(\frac{m}{n})$. Since $m = n\ln n - cn$, it follows that $\frac{m}{n} = \ln n - c$, so we can say that each $Z_i \sim \text{Pois}(\ln n - c)$. Since each Z_i is independent, we have

$$\begin{aligned}
\mathbf{Pr}[f(Z_1, Z_2, ..., Z_n) = 1] &= \prod_{i=1}^{n} 1 - \mathbf{Pr}[Z_i = 0] \\
&= \prod_{i=1}^{n} 1 - e^{-(\ln n - c)} \\
&= \left(1 - e^{-(\ln n - c)}\right)^n \\
&\leq \left(e^{-e^{-(\ln n - c)}}\right)^n \\
&= e^{-ne^{-\ln n + c}}
\end{aligned}$$

Hence, we wish to show that

$$2e^{-ne^{-\ln n + c}} \leq 2e^{-e^c}$$

From here, we have

$$\begin{aligned}
e^{-ne^{-\ln n + c}} &\leq e^{-e^c} \\
-ne^{-\ln n + c} &\leq -e^c \\
ne^{-\ln n + c} &\geq e^c \\
\ln\left(ne^{-\ln n + c}\right) &\geq c \\
\ln n + \ln\left(e^{-\ln n + c}\right) &\geq c \\
\ln n - \ln n + c &\geq c \\
c &\geq c
\end{aligned}$$

Therefore, the inequality holds for all $c > 0$. This means that **Pr**$[X \leq n \ln n - cn] \leq 2e^{-e^c}$ for all $c > 0$.