Homework 1

Justin Alvin & Prakhar Sharma

1. For our unbiased coin tossing experiment, we can use the Chernoff Bound to obtain an upper bound on the probability of more than $\frac{n}{2}$ flips being heads. More specifically, we want to find a value of n such that

$$\Pr[> \frac{n}{2} \text{ flips are heads}] < 0.001\,.$$

We first define the indicator random variable

$$X_i = \begin{cases} 1 & \text{if the } i\text{th coin flip is heads} \\ 0 & \text{otherwise} \end{cases}$$

and

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

Given that $\Pr(\text{Heads}) = \frac{1}{3}$ and $\Pr(\text{Tails}) = \frac{2}{3}$, the expected value of X is

$$\mathbf{E}[X] = \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = \frac{n}{3}\,.$$

By applying the Chernoff Bound we have

$$\Pr[X \geq (1+\delta)\mu] = \Pr\left[X \geq \frac{n}{2}\right] \leq e^{-\frac{\mu\delta^2}{3}} = 0.001\,.$$

Setting $\mu = \mathbf{E}[X]$, we can solve for δ

$$\frac{n}{2} = (1+\delta)\mu = (1+\delta)\frac{n}{3}$$

and get $\delta = \frac{1}{2}$. Substituting this back into our Chernoff Bound expression gives us

$$0.001 = e^{-\frac{\mu\delta^2}{3}} = e^{-\frac{(n/3)(1/4)}{3}} = e^{-\frac{n}{36}}\,.$$

And finally, solving for n we get

$$n \geq \lceil 36\, \ln(1000) \rceil = 249\,.$$

Using the Chernoff Bound, we see that with $n = 249$, the probability that more than half of the coin flips come out heads is less that 0.001.

2. (a) Let R denote a random variable that is 1 when the coin is heads (I win) and -1 when the coin is tails up (friend wins). The expression for the expectation in one single trial is:

$$E[R] \;=\; 1 * \frac{1}{2} + (-1) * \frac{1}{2} = 0$$

Hence the expected payoff is 0*100 = 0

(b) In the case of an unbiased coin, following from the solution to the previous part, the expectation expression becomes:

$$1 * (0.3) + (-1) * (0.7) = -0.4$$

Hence the expected value of R for 100 tries then becomes:

$$E[R] = -0.4 * 100 = -40$$

(c) Since the random variables we considered in the last section can take on a negative value (-1), we will modify the problem a little bit to ensure we are dealing with random variables strictly greater than 0. Let Y be a random varibale that takes the value 1 when the coin turns up tails. For n coin tosses, the payoff would be [no. of tails - number of heads] wich is $n - (100 - n)$. Hence for a payoff of $50:

$$P[\text{payoff} \geq 50] = P[Y \geq 75] = \frac{E[Y]}{75} = \frac{14}{15}$$

3. (a) Let A be a random variable that takes integer values of numbers on the face of a die $1, 2, 3, 4, 5, 6$. The expected value of A :

$$E[A] = \frac{1}{6} * (1 + 2 + 3 + 4 + 5 + 6) = 21/6$$

and expected value of A^2:

$$E[A^2] = \frac{1}{6} * (1 + 4 + 9 + 16 + 25 + 36) = 91/6$$

The variance of A can be calculated as:

$$Var[A] = E[A^2] - E[A]^2 = \frac{91}{6} - \frac{49}{4} = \frac{182 - 147}{12} = 35/12$$

As the events are pairwise independent:

$$Var[X] = Var[\sum_{i=0}^{100} A_i] = \sum_{i=1}^{100} Var[A_i] = 875/3$$

This result follows from the linearity of variance.
According to Chebyshev's inequality,

$$P(|X - E[X]| \geq \lambda) \leq \frac{Var[X]}{\lambda^2}$$

Setting $\lambda = 50$, It follows from the above inequality that

$$P(|X - E[X]| \geq 50) \leq \frac{875}{3 * 50^2} \leq 0.116$$

(b) Consider the Markov inequality:

$$P[Y > t\ E[Y]] \leq \frac{1}{t}$$

where $\lambda = tE[Y]$ Similarly,

$$P[Y > t^k\ E[Y]] \leq \frac{1}{t^k}$$

Since k is a positive integer and Y has to be positive for the markov inequality to hold,

$$P[Y^{\frac{1}{k}} > t\ E[Y]^{\frac{1}{k}}] < \frac{1}{t^k}$$

Now Let $Y^{\frac{1}{k}} = X - E[X]$ then following from the equation above,

$$P[|X - E[X]| \geq t(E[(X - E[X])^k]])^{\frac{1}{k}} < \frac{1}{t^k}$$

<u>Q4:</u> Let X_j^i be an indicator variable.

$$X_j^i = \begin{cases} 1 & \text{if ball falls in bin } i \ (j^{th} \text{ ball}) \\ 0 & \text{otherwise.} \end{cases}$$

Here, $1 \le j \le n$ & $1 \le i \le m$

Now, $E[X_j^i] = \sum X_j^i P(X_j^i) = P(X_j^i = 1)$ (as $P(X_j^i = 0) \cdot X_j^i$ is zero).

$P(X_j^i = 1) = \frac{1}{n}$ [n bins] Let B_i be total Number of balls that land in i^{th} bin : $B_i = \sum_{j=1}^{m} X_j^i$

$$\rightarrow \quad E[B_i] = E\left[\sum_{j=1}^{m} X_j^i\right] = \sum_{j=1}^{m} E[X_j^i]$$

$$\boxed{E[B_i] = \sum_{j=1}^{m} \frac{1}{n} = \frac{m}{n}}$$

<u>4(a)</u>: Using chernoff bound on Variable B_i

$$P(|B_i - E[B_i]| > 25\ln n) = P\left(|B_i - E[B_i]| > \frac{100 n \ln n}{4 \cdot n}\right)$$

Notice that $\frac{100 n \ln n}{n}$ is $E[B_i]$

$$\rightarrow P\left(|B_i - E(B_i)| > \frac{E[B_i]}{4}\right) \le 2e^{-\frac{100 \ln n}{4^2 \cdot 3}} \le \frac{2}{n^2}$$

Since this is a prob. dist., reversing the inequality leads to

$$P(|B_i - E[B_i]| \le 25\ln n) \ge \left(1 - \frac{2}{n^2}\right)$$

4. (a)

Now, $P(\exists i \in [1,n] : |B_i - E[B_i]| > 25\ln n) \leq \sum_{i=1}^{n} P[|B_i - E[B_i]|]$

Since B_i is a random variable which is independent from other B_i's [each ball equally likely to go into any bin], (I)

$$P(|B_i - E[B_i]| > 25\ln n) \leq n \cdot \frac{1}{n^2} \leq \frac{1}{n} \text{ (overall } i\text{)}$$

Reversing inequality, $P(\forall i \in [1,n] |B_i - E[B_i]| \leq 25\ln n) \leq 1 - \frac{1}{n}$

This shows that the maximum difference between two bins could be $50\ln n$ (constant factor) as maximum overload with probability $\left(1-\frac{1}{n}\right)$ can be $25\ln n$.

4(b) Since $E[B_i] = \frac{m}{n}$ (part a), $\frac{m}{n} \pm O\left(\sqrt{\frac{m}{n}\ln n}\right) = E[B_i] \pm k\sqrt{\frac{m}{n}\ln n}$ where k is a constant factor.

To prove: $P\left(|B_i - E[B_i]| \geq k\sqrt{\frac{m}{n}\ln n}\right)$ is bounded by $\left(1-\frac{1}{n}\right)$

Rearranging terms on LHS (prob. term above)

$A = P\left(|B_i - E[B_i]| \geq k \cdot \frac{m}{n}\sqrt{\frac{n}{m}\ln n}\right)$ This expression can be converted to $P(|B_i - E[B_i]| \geq \frac{m}{n}\delta)$ where $\delta = k\sqrt{\frac{n}{m}\ln n}$

Applying chernoff bound to this: $A \leq 2e^{-\frac{m}{n}(9)\left(\frac{n}{m}\right)\left(\frac{\ln n}{3}\right)}$.

$$\Rightarrow A \leq \frac{2}{n^3} \leq \frac{1}{n^2} \quad [\text{when } k=3]$$

Applying union bound: $\boxed{P\left(\forall i \in [1,n] |B_i - E[B_i]| \leq 3\sqrt{\frac{m}{n}\ln n}\right) \geq 1 - \frac{1}{n}}$

from (I) in previous part (top of this page)

Because $P\left(\forall i \in [1,n] |B_i - E[B_i]| > 3\sqrt{\frac{m}{n}\ln n}\right) \leq n \cdot \frac{1}{n^2} = \frac{1}{n}$

(b)

(c) We can express the probability that bin i has at least k balls in it as

$$Pr(\geq k \text{ balls in bin } i) \leq \binom{n}{k}\left(\frac{1}{n}\right)^k .$$

Using Stirling's approximation, we get

$$Pr(\geq k \text{ balls in bin } i) \leq \left(\frac{ne}{k}\right)^k = \left(\frac{e}{k}\right)^k = e^{\ln(\frac{e}{k})^k} = e^{k(1-\ln\, k)} .$$

If we set the height of the heaviest bin to be $k = \frac{\ln\, n}{\ln\, \ln\, n}$, we get

$$e^{k(1-\ln\, k)} = e^{(\frac{\ln\, n}{\ln\, \ln\, n})(1-\ln\, (\frac{\ln\, n}{\ln\, \ln\, n}))} \leq e^{-(\frac{\ln\, n}{\ln\, \ln\, n})(\frac{\ln\, \ln\, n}{2})} = \sqrt{n} .$$

Finally, we can express this as $o(1)$. By union bound, we know that the probability that there is a bin containing at least k balls is $1 - \sqrt{n} = 1 - o(1)$.

5. (a) For our initial algorithm, we can show that if the total stream size is m, any item that has frequency $> \frac{m}{k+1}$ is returned.

Consider an item, j with observed frequency $\hat{f}_j$ and true frequency $f_j > \frac{m}{k+1}$. If our total stream size is m, we know that there can be at most $k-1$ such items. Additionally, we know that $f_j \geq \hat{f}_j$ because we only ever increment the frequency for item j when it is observed in the stream.

If item j is never deleted from our list, then $f_j = \hat{f}_j$ because we always update the frequency for j. There are therefore two events we must consider that lead to $\hat{f}_j < f_j$.

Event 1: Item j arrives and is not in our list yet, but our list is already full. In this case, item j is not added to our list and its frequency is not recorded at this step.

Event 2: Item i arrives and is not in our list yet, item j is in our list, and our list is already full. In this case, the frequency of item j is decremented.

In both events, the observed frequency of item j becomes one less than the true frequency. Additionally, in both events, whenever item j or i arrives, all k items in our list have their counters decremented. For this to occur, we must have already seen at least k items, plus the current item at this step. Hence, these events can occur in total at most $\frac{m}{k+1}$ times and we have $\hat{f}_j > f_j - \frac{m}{k+1}$. Because $\hat{f}_j > 0$, we therefore know that $f_j > \frac{m}{k+1}$. Thus any item with frequency $> \frac{m}{k+1}$ will be returned by our algorithm.

(b) We can use a Count-Min Sketch (CMS) data structure, along with a min-heap to solve the ϵ-approximate heavy hitters problem. For our Twitter dataset, this allows us to return all hashtags with frequency at least $0.002n$, where n is the total size of the dataset. This also means that any hashtag returned has frequency $0.001n$.

Our CMS implementation returned the set of hashtags [*31minutos, blanco, duckdynasty, jaibrooksfollowspree, job, jobs, love, lt, marchwish, meteoalarm, nowplaying, np, oomf, rt, spikersmarchwish, tweetmyjobs, viña2013,* 地震], compared to our algorithm in part (a), which returned [*31minutos, blanco, duckdynasty, jaibrooksfollowspree, job, jobs, love, lt, meteoalarm, nowplaying, np, oomf, rt, tweetmyjobs, viña2013,* 地震]. As can be seen, our CMS algorithm returns more hashtags (specifically, [*marchwish, spikersmarchwish*]) than the algorithm in part (a), which represents the true frequencies of items in our dataset. However, the CMS implementation requires much less space and only needs a single pass over our data stream, with space usage $\tilde{O}(k)$, where n is the total size of the stream. The algorithm in part (a) uses $O(k(\log n + \log m))$ space, where m is the maximum value in the data stream.