Homework 2 - Solutions

Problem 1

Let $X_n$ be the random variable for the number of 6’s after $n$ dice rolls. Thus, we know that $X_n \sim \text{Binomial}(n, \frac{1}{6})$.

a. Markov: $P(X_n > \frac{n}{4}) < \frac{E[X_n]}{\frac{n}{4}} = \frac{\frac{n}{6}}{\frac{n}{4}} = \frac{2}{3}$

b. Chebyshev: $P(|X_n - E[X_n]| > \frac{n}{12}) < \frac{\text{Var}(X_n)}{(\frac{n}{12})^2} = \frac{\frac{n}{6} \cdot \frac{5}{6}}{(\frac{n}{12})^2} = \frac{20}{n}$

c. Chernoff: $P(X_n > \frac{n}{4}) < \frac{E[e^{tX_n}]}{e^{t\frac{n}{4}}} = \frac{(\frac{1}{6} e^t + \frac{5}{6})^n}{e^{nt\frac{1}{4}}} \leq \frac{e^{\frac{t}{6}(e^t-1)}}{e^{nt\frac{1}{4}}}$

So choosing, $t^* = \ln \frac{3}{2}$ gives us a bound of $\approx e^{-0.018n}$. Using the Chernoff bound from the book is also acceptable.

Problem 2

a. Let our estimate be $Y_t = \sum_{i=1}^t X_i / t$. Take $t = \frac{r^2}{\epsilon^2 \delta}$. Note that $E[Y_t] = E[\frac{\sum_{i=1}^t X_i}{t}] = \frac{t E[X]}{t} = E[X]$ and $\text{Var}(Y_t) = \text{Var}(\frac{\sum_{i=1}^t X_i}{t}) = \frac{1}{t^2} \sum_{i=1}^t \text{Var}(X_i) = \frac{\text{Var}(X)}{t}$.

Thus, we get $P(|Y_t - E[X]| > \epsilon E[X]) = P(|Y_t - E[Y_t]| > \epsilon E[Y_t]) \leq \frac{\text{Var}(Y_t)}{\epsilon^2 (E[Y_t])^2} = \frac{\text{Var}(X) \epsilon^2 \delta}{\epsilon^2 (E[X])^2} = \delta$

b. Using part a) and letting $\delta = \frac{1}{4}$, we get that $\frac{4r^2}{\epsilon^2}$ samples are enough for a weak estimate.

c. Let $Y_t$ be the $t^{th}$ weak estimate and $Z_i = 1$ if $|Y_t - E[X]| \leq \epsilon E[X]$ and 0 otherwise, and let $Y$ be the median of $N$ weak estimates. Note that $E[Z_i] = P(Z_i = 1) \geq \frac{3}{4}$ by the definition of a weak estimate. Note that $|Y - E[X]| \geq \epsilon E[X]$ (which is the event that $Y$ is not within $\epsilon E[X]$ of $E[X]$) implies that $\sum_{i=1}^N Z_i \leq \frac{N}{2}$ (if the median if out of range, then at least half of the weak estimates used to produce that median must also be out of range).

Let us take $N = 24 \ln \frac{1}{\delta}$ weak estimates. By Chernoff bound for sum of IID Poisson trials

$$P(|Y - E[X]| \geq \epsilon E[X]) \leq P\left(\sum_{i=1}^N Z_i \leq \frac{N}{2}\right) = P\left(\sum_{i=1}^N Z_i \leq \left(1 - \frac{1}{3}\right) \frac{3N}{4}\right) \leq e^{-\frac{3N}{4} \left(\frac{1}{2}\right)^2} = e^{-\frac{N}{2\pi}} \leq \delta$$

Thus, $O\left(\frac{r^2}{\epsilon^2} \ln \frac{1}{\delta}\right)$ samples are sufficient.
Problem 3

a. Consider a path from root to leaf. Let it contain $k$ good nodes. The number of elements left in the leaf is at most $n \left(\frac{3}{2}\right)^k$ which must be at least 1. Thus, we have $k \leq \left(\log_2 \frac{3}{2}\right)^{-1} \log_2 n < 2 \log_2 n$. So, there can be no more than $\left\lfloor 2 \log_2 n \right\rfloor$ good nodes on any path from root to leaf.

b. One way to approach this problem is to start by considering the following related problem. Suppose we keep flipping a fair coin until we see 100 heads. What is the probability that we flip the coin more than 1000 times? We would like to use the Chernoff bounds we’ve derived, but it seems like we can’t, because the number of flips is the sum of 100 geometric random variables, not 0-1 random variables. (The distribution of the number of flips is also called the negative binomial distribution.)

The key observation that allows us to turn this into a Chernoff bounds problem (without having to derive Chernoff bounds for sums of geometric random variables) is that the probabilities of the following two events are equal:

- Event 1 (the event in the problem): We keep flipping a fair coin until we see 100 heads, and we flip the coin more than 1000 times
- Event 2: We flip a fair coin exactly 1000 times, and the number of heads is less than 100

Event 2 is something we can analyze with the Chernoff bounds we’ve derived.

Now, consider how the coin flipping problem relates to the original problem. In part a, we showed that any path has to end after at most $c \log n$ good nodes—for now, let’s just say it ends after exactly $c \log n$ good nodes. For a particular path—in other words, the path corresponding to a particular element of the input array—at each step, the set containing that element is divided in either a good way or a bad way, with a good division occurring with probability $\frac{1}{3}$.

In the coin flipping problem, suppose heads represents a good division, and we bias the coin so it comes up heads with probability $\frac{1}{3}$. In the original problem, we keep dividing the set containing the particular element (cf. keep flipping biased coins) until we get $c \log n$ heads (cf. 100 good nodes). We want to find the probability that the path contains more than $c' \log n$ nodes (cf. more than 1000 flips). By considering Event 1 and Event 2 in the coin flipping problem, we see that this equals the probability of getting less than $c \log n$ heads in exactly $c' \log n$ flips.

Let $X_i$ be an indicator random variable for whether or not the $i$th flip is heads. Recall that $E[X_i] = \frac{1}{3}$. Setting $c = 2$ (from part a), and $c' = 24$, we get that

$$P\left(\sum_{i=1}^{c' \log_2 n} X_i < c \log_2 n \right) = P\left(\sum_{i=1}^{c' \log_2 n} X_i < \left(1 - \frac{3c}{c'}\right) \frac{c' \log_2 n}{3}\right)$$

$$\leq e^{-\frac{c' \log_2 n}{3} \left(1 - \frac{3c}{c'}\right) \frac{1}{2}} = e^{-8 \log_2 n \left(\frac{3}{2}\right)^{\frac{1}{2}}} \leq \frac{1}{n^2}$$

Thus, with probability greater than $1 - \frac{1}{n^2}$ there are at most $24 \log_2 n$ nodes on any path from root to leaf.
A few notes:

- The key observation of relating Event 1 and Event 2 is similar to an idea we used in the packet routing problem, when we converted part of the analysis into a coin flipping problem. This starts on the last paragraph of page 84 of the textbook (“We claim that the probability that the active packets cross edges of $P$ more than $30n$ times...”), and is also on the slides for Chapter 4-a, slide 17.

- In part of this solution, in order to simplify the presentation, we assumed that a path ends after exactly $c \log n$ good nodes, rather than at most $c \log n$ good nodes. It might be good to think about how the argument would change (slightly) with at most $c \log n$ good nodes.

- One common approach was something like the following: for $0 \leq i \leq c' \log n$, let $X_i$ be an indicator random variable for whether the node at depth $i$ is good. Then we can bound the probability that $\sum_{i=0}^{c' \log n} X_i \leq c \log n$ using Chernoff bounds. One problem is that it’s unclear what this really means unless we’re guaranteed that the path must have at least $c' \log n$ nodes (otherwise, what does $X_{c' \log n}$ refer to?). But this is the event whose probability we were trying to bound in the first place.

There is a way to tweak this idea to make it work. For $0 \leq i \leq c' \log n$, let $X_i$ be an indicator random variable for whether the node at depth $i$ is good; however, if there is no node at depth $i$, let $X_i$ be a Bernoulli random variable that takes value 1 with probability $1/3$. This is essentially another way to view the relationship between Event 1 and Event 2 in the coin flipping problem.

c. Let $E_i$ be the event that the $i^{th}$ path is longer than $24 \log_2 n$ nodes. Since there is a leaf for every entry of the original list, there are $n$ paths from root to leaf. We know by part b), $P(E_i) \leq \frac{1}{n^2}$. By the union bound, $P(\bigcup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} P(E_i) \leq n \frac{1}{n^2} = \frac{1}{n}$. Thus, with probability greater than $1 - \frac{1}{n}$, there are no paths from root of leaf of length greater than $24 \log_2 n$.

d. The number of comparisons needed to place an element at the correct index is the length of the path from the root to the element’s leaf in this quicksort graph. There are $n$ leafs, and with probability greater than $1 - \frac{1}{n}$, all paths to these leaves have length at most $24 \log_2 n$, so the amount of comparisons is at most $n \cdot 24 \log_2 n \in O(n \log_2 n)$ with probability greater than $1 - \frac{1}{n}$.