Problem 1: (Practice with Chebyshev and Chernoff bounds)

When using concentration bounds to analyze randomized algorithms, one often has to approach the problem in different ways depending on the specific bound being used. Typically, Chebyshev is useful when dealing with more complicated random variables, and in particular, when they are pairwise independent; Chernoff bounds are usually used along with the union bound for events which are easier to analyze. We'll now go back and look at a few of our older examples using both these techniques.

Part (a)

(Number of collisions) Recall we showed that if we throw m balls in n bins, the average number of collisions $X_{m,n}$ to be $\mu_{m,n} = \binom{m}{2}\frac{1}{n}$. Use Chebyshev's inequality to show that:

$$\mathbb{P}[|X_{m,n} - \mu_{m,n}| \geq c\sqrt{\mu_{m,n}}] \leq 1/c^2.$$

Next suppose we choose $m = 2\sqrt{n}$, then $\mu_{m,n} \leq 1$. Use Chernoff bounds plus the union bound to bound the probability that no bin has more than 1 ball. Compare this to the more exact analysis you did in homework 1.

Solution: Let $\sigma_{m,n}$ be the standard deviation of $X_{m,n}$. As we did before, for every pair of balls (i,j), let $Y_{i,j}$ be the indicator that the two balls collide. Note that since $Y_{i,j}$ are $\{0,1\}$ valued r.vs, we have $\mathbb{E}[Y_{i,j}^2] = \mathbb{E}[Y_{i,j}]$, and thus $Var(Y_{i,j}) \geq \mathbb{E}[Y_{i,j}]$. Moreover, observe that $Y_{i,j}$ are *pairwise independent*, i.e., for any $(i,j) \neq (i',j')$, the r.v.s $Y_{i,j}$ and $Y_{i',j'}$ are independent (in particular, note that $Y_{i,j}, Y_{i,k}$ and $Y_{j,k}$ are pairwise independent – however they are not mutually independent). Now we have $X_{m,n} = \sum_{i,j} Y_{i,j}$, and thus $\mu_{m,n} = \sum_{i,j}\mathbb{E}[Y_{i,j}]$ and $Var(X_{m,n}) = \sum_{i,j} Var(Y_{i,j}) \leq \sum_{i,j}\mathbb{E}[Y_{i,j}] = \mu_{m,n}$. Thus, by Chebyshev's inequality, we have:

$$\mathbb{P}\left[|X_{m,n} - \mu_{m,n}| \geq c\sqrt{\mu_{m,n}}\right] \leq 1/c^2.$$

To use Chernoff bounds, we instead consider the number of balls in each bin. Let $Z_{b,i}$ be the indicator that ball i fell in bin b – these are now mutually independent. Further, let $B_b = \sum_i Z_{b,i}$ be the number of balls in bin b – then we know that $\mathbb{E}[B_b] = m/n$, and moreover using our standard Chernoff bound, we have:

$$\begin{aligned}
\mathbb{P}\left[B_b \geq 2\right] &= \mathbb{P}\left[B_b \geq (1 + (2n/m - 1))\,\mathbb{E}[B_b]\right] \\
&\leq \exp\left(-\frac{(2n/m-1)^2(m/n)}{2 + (2n/m-1)}\right) \quad \left(\text{Using } \mathbb{P}[X \geq (1+\epsilon)\mu] \leq \exp\left(\frac{-\epsilon^2\mu}{2+\epsilon}\right)\right) \\
&= \exp\left(-\frac{(2n-m)^2}{n(m+2n)}\right)
\end{aligned}$$

Finally, by the union bound, we have $\mathbb{P}[\text{Some bin has } \geq 2 \text{ balls}] \leq n\exp\left(-\frac{(2n-m)^2}{n(m+2n)}\right)$. Now if $m = 2\sqrt{n}$, we get $\exp\left(-\frac{(2n-m)^2}{n(m+2n)}\right) = \Theta(1)$, and thus the bound on $\mathbb{P}[\text{Some bin has } \geq 2 \text{ balls}]$ is not

useful (as it grows with n). On the other hand, in the first homework, we did a direct calculation to show that $\mathbb{P}[\text{No collisions}] \leq \exp(-\mathbb{E}[X_{m,n}])$, and thus if $m = 2\sqrt{n}$, we have $\mathbb{P}[\text{No collisions}] \leq 1/2$. Thus, the Chernoff bound does not give us the tightest scaling in this case.

Part (b)

(Coupon collector) For n bins, recall that we defined T_i to be the first time when i unique bins were filled, and used these random variables to show that T_n, i.e., the number of balls we need to throw before every bin has at least one ball, satisfies $\mathbb{E}[T_n] = nH_n = \Theta(n \log n)$. Using the same random variables, show that $\mathbb{P}[|T_n - \mathbb{E}[T_n]| \geq c\mathbb{E}[T_n]] \leq \frac{\pi^2}{6c^2 {H_n}^2}$.

Next, suppose we throw in $m = n \log n + cn$ balls – using Chernoff bounds plus the union bound, choose c such that no bin is empty with probability greater than $1 - \delta$.
Hint: Use $\sum_{i=1}^{\infty} 1/i^2 = \pi^2/6$

Solution: First, using the independence of T_is let's calculate $Var[T_n]$:

$$Var[T_n] = \sum_{i=0}^{n-1} Var[T_i] = \sum_{i=0}^{n-1} \frac{1-p_i}{{p_i}^2} = \sum_{i=0}^{n-1} \frac{ni}{(n-i)^2} = \sum_{i=1}^{n} \frac{n(n-i)}{i^2} \leq n^2 \sum_{i=1}^{n} \frac{1}{i^2} \leq \frac{\pi^2 n^2}{6}.$$

Now, using Chebyshev's inequality, we get:

$$\mathbb{P}[|T_n - \mathbb{E}[T_n]| \geq c\mathbb{E}[T_n]] = \mathbb{P}[|T_n - \mathbb{E}[T_n]| \geq cnH_n] \leq \frac{Var[T_n]}{(cnH_n)^2} \leq \frac{\pi^2}{6c^2 {H_n}^2}.$$

Next, let $m = n \log n + cn$, and let B_b be the number of balls in bin b. We know $\mathbb{E}[B_b] = \log n + c$. Then by the Chernoff bound, we have:

$$\mathbb{P}[B_b \leq 0] = \mathbb{P}[B_b \leq (1-1)\mathbb{E}[B_b]] \leq \exp\left(\frac{-\mathbb{E}[B_b]}{2}\right) =\leq \exp\left(\frac{-(\log n + c)}{2}\right)$$

Finally, using the union bound, we have $\mathbb{P}[\text{Some bin is empty}] \leq n\mathbb{P}[B_b \leq 0] = \exp(\log n - \frac{(\log n + c)}{2})$, and we can set $c = \log n + 2\log(1/\delta)$ to get this less than δ.

Here, using a direct calculation is better than the Chernoff bound. In particular, we have:

$$\mathbb{P}[B_b \leq 0] = \left(1 - \frac{1}{n}\right)^m \leq e^{-m/n} = e^{-c}/n$$

By the union bound, we have $\mathbb{P}[\text{Some bin is empty}] \leq e^{-c}$, and thus we need $c = \log(1/\delta)$ to ensure this is less than δ.

The main takeaway again is that Chernoff bounds are fine when probabilities are small and we do not want the tightest bounds, but may be weak when probabilities are larger and we want tighter bounds.

Problem 2: (The Hoeffding Extension)

In class, we saw that for a r.v. $X_i \sim$Bernoulli(p_i), we have:

$$\mathbb{E}[e^{\theta X}] = 1 - p + pe^{\theta}$$

Plugging this into the Chernoff bound and optimizing over θ, we obtained a variety of bounds – in particular, for independent r.vs $\{X_i\}$ with $X = \sum_i X_i$ and $\mu = \mathbb{E}[X] = \sum_i p_i$, we showed for any $\epsilon > 0$:

$$\mathbb{P}[|X - \mu| \geq \epsilon\mu] \leq 2\exp\left(\frac{-\epsilon^2\mu}{2+\epsilon}\right)$$

We now extend this to more general bounded r.vs.

Part (a)

First, for any θ, argue that the function $f(x) = e^{\theta x}$ for $x \in [0,1]$ is bounded above by the line joining $(0,1)$ and $(1, e^{\theta})$. Using this, find constants α, β such that $\forall\, x \in [0,1]$:

$$e^{\theta X} \leq \alpha x + \beta$$

Solution: Recall that $f(x) = e^{\theta x}$ is a convex function. Now, note that $f(0) = 1$ and $f(1) = e^{\theta}$. Therefore, for $x \in [0,1]$, it is bounded above by the line joining $(0,1)$ and $(1, e^{\theta})$, which means:

$$e^{\theta x} \leq (e^{\theta} - 1)x + 1.$$

Part (b)

Next, for any random variable X_i taking values in $[0,1]$ such that $\mathbb{E}[X_i] = \mu_i$, show that:

$$\mathbb{E}[e^{\theta X_i}] \leq 1 - \mu_i + \mu_i e^{\theta}.$$

Using this, for independent r.vs X_i taking values in $[0,1]$ with $\mathbb{E}[X_i] = \mu_i$, and defining $X = \sum_i X_i, \mu = \sum_i \mu_i$, show that:

$$\mathbb{P}[|X - \mu| \geq \epsilon\mu] \leq 2\exp\left(\frac{-\epsilon^2\mu}{2+\epsilon}\right)$$

(Note: You can directly use the inequality for Bernoulli r.v.s – no need to show the optimization.)

Solution: Let's take the expectations of both sides of the inequality in the solution of the part a):

$$\mathbb{E}[e^{\theta X_i}] \leq \mathbb{E}[(e^{\theta} - 1)X_i + 1] = (e^{\theta} - 1)\mathbb{E}[X_i] + 1 = (e^{\theta} - 1)\mu_1 + 1 = 1 - \mu_i + \mu_i e^{\theta}.$$

Now, let $X = \sum_i X_i, \mu = \sum_i \mu_i$. Using exactly the same method is we did in class for Bernoulli r.vs, we get:

$$\mathbb{P}[|X - \mu| \geq \epsilon\mu] \leq 2\exp\left(\frac{-\epsilon^2\mu}{2+\epsilon}\right)$$

Part (c) (Optional)

Next, for any random variable X_i taking values in $[a_i, b_i]$ such that $\mathbb{E}[X_i] = \mu_i$, a similar bounding technique as above can be used to show:

$$\mathbb{E}\left[e^{\theta(X_i-\mu_i)}\right] \leq \exp\left(\frac{1}{8}\theta^2(b-a)^2\right)$$

(This is sometimes referred to as Hoeffding's lemma – for the proof, see the wikipedia article)

Now consider independent r.vs X_i taking values in $[a_i, b_i]$ with $\mathbb{E}[X_i] = \mu_i$, and as before, let $X = \sum_i X_i, \mu = \sum_i \mu_i$. Using the above inequality, optimize over θ to show that:

$$\mathbb{P}[(X-\mu) \geq \epsilon\mu] \leq \exp\left(\frac{-2\epsilon^2\mu^2}{\sum_{i=1}^n (b_i-a_i)^2}\right)$$

Solution: As in the standard Chernoff inequality, we have:

$$\begin{aligned}
\mathbb{P}[X-\mu \geq \epsilon\mu] &= \mathbb{P}\left[e^{\theta(X-\mu)} \geq e^{\theta\epsilon\mu}\right] \\
&\leq \min_{\theta>0} \frac{\mathbb{E}[e^{\theta(X-\mu)}]}{e^{\theta\epsilon\mu}} \\
&= \min_{\theta>0} \frac{\Pi_i \mathbb{E}[e^{\theta(X_i-\mu_i)}]}{e^{\theta\epsilon\mu}} \\
&\leq \min_{\theta>0} \exp\left(\frac{\theta^2}{8}\sum_i (b_i-a_i)^2 - \theta\epsilon\mu\right)
\end{aligned}$$

To optimize, we choose $\theta^* = 4\epsilon\mu/\sum_i (b_i-a_i)^2$, to get:

$$\mathbb{P}[X-\mu \geq \epsilon\mu] \leq \exp\left(-\frac{2\epsilon^2\mu^2}{\sum_i (b_i-a_i)^2}\right)$$

Problem 3: (A Weaker Sampling Theorem: Adapted from MU Ex 4.9)

In class we saw the following 'sampling theorem' for estimating the mean of a $\{0,1\}$-valued random variable: In order to get an estimate within $\pm\epsilon$ with confidence $1-\delta$, we need $n \geq \frac{2+\epsilon}{\epsilon}\ln\frac{2}{\delta}$. A crucial component in this proof was using the Chernoff bound for Bernoulli(p) random variables. Suppose instead we want to estimate a more general random variable X (for example, the average number of hours of TV watched by a random person) – we may not be able to use a Chernoff bound if we do not know the moment generating function We now show how to to get a similar sampling theorem which only uses knowledge of the mean and variance of aX.

We want to estimate a r.v. X with mean $\mathbb{E}[X]$ and variance $Var[X]$, given i.i.d samples $X_1, X_2, \ldots$. Let $r = \sqrt{Var[X]}/\mathbb{E}[X]$ – we now show that we can estimate it up to accuracy $\pm\epsilon\mathbb{E}[X]$ and confidence $1-\delta$ using $O\left(\frac{\epsilon^2}{r^2}\ln\frac{1}{\delta}\right)$ samples.

Part (a)

Given n samples $X_1, \ldots, X_n$, suppose we use the estimator $\widehat{X} = \left(\sum_{i=1}^n X_i\right)/n$. Show that $n = O(r^2/\epsilon^2\delta)$ is sufficient to ensure:

$$\mathbb{P}\left[|\widehat{X} - \mathbb{E}[X]| \geq \epsilon\mathbb{E}[X]\right] \leq \delta$$

Solution: By linearity of expectation, we have that $\mathbb{E}[\widehat{X}] = \sum_{i=1}^n \mathbb{E}[X_i]/n = \mathbb{E}[X]$. Moreover, since X_is are i.i.d., we have that $Var\widehat{X} = \sum_{i=1}^n Var(X_i)/n^2 = Var(X)/n = r^2\mathbb{E}[X]^2/n$. Now, using Chebyshev's inequality we get:

$$\begin{aligned}\mathbb{P}\left[|\widehat{X} - \mathbb{E}[X]| \geq \epsilon\mathbb{E}[X]\right] &\leq \frac{Var(\widehat{X})}{\epsilon^2\mathbb{E}[X]^2} \\ &= \frac{r^2}{n\epsilon^2}\end{aligned}$$

To ensure $\mathbb{P}\left[|\widehat{X} - \mathbb{E}[X]| \geq \epsilon\mathbb{E}[X]\right] \leq \delta$, we can choose n such that $\frac{r^2}{n\epsilon^2} \leq \delta$. Using $n = O(r^2/\epsilon^2\delta)$ samples is thus sufficient.

Part (b)

We say an estimator is a *weak estimator* if it satisfies that $\mathbb{P}\left[|\widehat{X} - \mathbb{E}[X]| \geq \epsilon\mathbb{E}[X]\right] \leq 1/4$ – using part (a), show, that we need $O(r^2/\epsilon^2)$ samples to obtain a weak estimator. Now suppose we are given m weak estimates $\widehat{X}_1, \widehat{X}_2, \ldots, \widehat{X}_m$, and we define a new estimator $\widetilde{X}$ to be the *median* of these weak estimates. Show that using $m = O(\ln(1/\delta)$ weak estimates gives us an estimate $\widetilde{X}$ that satisfies:

$$\mathbb{P}\left[|\widetilde{X} - \mathbb{E}[X]| \geq \epsilon\mathbb{E}[X]\right] \leq \delta$$

What could go wrong if we used the mean of $\widehat{X}_1, \widehat{X}_2, \ldots, \widehat{X}_m$ instead of the median?
NOTE: I got the definition of weak estimator inverted in the Homework - this is the correct definition. Essentially, we want the probability of samples being close to the mean to be $> 1/2$.

Solution: If we take $\delta = \frac{1}{4}$ in part a), we see that $\widehat{X}$ is a weak estimator if we use $n = O(r^2/\epsilon^2)$ samples. Now we want to show that we need $O(\log 1/\delta)$ such weak estimators to ensure that the median of these estimators is within $(1 \pm \epsilon)\mathbb{E}[X]$.

Let us introduce new r.vs Y_i s.t. $Y_i = 1$, if $|\widehat{X}_i - \mathbb{E}[X]| \leq \epsilon\mathbb{E}[X]$, and $Y_i = 0$, if $|\widehat{X}_i - \mathbb{E}[X]| \geq \epsilon\mathbb{E}[X]$. $Y = \sum_{i=i}^m Y_i$ is the number of weak estimates for which $\mathbb{P}\left[|\widehat{X} - \mathbb{E}[X]| \leq \epsilon\mathbb{E}[X]\right] \leq \delta$; to ensure that the median is also within $(1 \pm \epsilon)\mathbb{E}[X]$, we need $Y > m/2$. In other words, we have:

$$\mathbb{P}\left[|\widetilde{X} - \mathbb{E}[X]| \geq \epsilon\mathbb{E}[X]\right] = \mathbb{P}\left[Y \leq m/2\right].$$

Moreover, $\mathbb{P}[Y_i = 0] \leq 3/4$, and hence $\mathbb{E}[Y] \geq 3m/4$. Now for independent random variables $Z_i \sim$ Bernoulli(p_i), with $Z = \sum_{i=1}^{n} Z_i$, we saw the following Chernoff Bound:

$$\mathbb{P}[Z \leq (1-\epsilon)\mathbb{E}[Z]] \leq \exp\left(-\frac{\epsilon^2}{2}\mathbb{E}[Z]\right)$$

Using this, we can write:

$$\begin{aligned}\mathbb{P}\left[Y \leq m/2\right] &= \mathbb{P}\left[Y \leq (1-1/3)3m/4\right] \\ &\leq \mathbb{P}[Y \leq (1-1/3)\mathbb{E}[Y]] \\ &\leq e^{\frac{-(1/9)\mathbb{E}[Y]}{2}} \leq e^{-m/18}.\end{aligned}$$

Now, if we take $m \geq 18\ln(1/\delta) = O(\log 1/\delta)$, we get $\mathbb{P}\left[|\widehat{X} - \mathbb{E}[X]| \leq \epsilon\mathbb{E}[X]\right] \leq \delta$.

Note though that we only needed to count the weak estimators that fell outside $(1 \pm \epsilon)\mathbb{E}[X]$ – we did not need to assume they are bounded, or have bounded variance, etc. If instead we had used the mean of the weak estimators, we could have a problem if these bad estimates happened to take very large values. Using the median made our estimate *robust* to such outliers.

Problem 4: (Randomized Set-Cover)

In this problem, we'll look at *randomized rounding*, which is a very powerful technique for solving large-scale combinatorial optimization problems. The main idea is that given a problem which can be written as an optimization problem with integer constraints, we can sometimes solve the *relaxed problem* with non-integer constraints, and then *round* the solutions to get a good assignment. We will highlight this technique for the *Minimum Set-Cover* problem.

We are given a collection of m subsets $\{S_1, S_2, \ldots, S_m\}$ which are subsets of some large set U of n elements, such that $\bigcup_i S_i = U$. The Minimum Set-Cover problem is that of selecting the smallest number of sets $\mathcal{C}$ from the collection $\{S_1, S_2, \ldots, S_m\}$ such that they cover U, i.e., such that each element in U lies in at least one of the sets in $\mathcal{C}$.

Part (a)

Argue that the minimum-set cover problem is equivalent to the following integer program:

$$\begin{aligned}&\underset{x}{\text{Minimize}} && \sum_i x_i \\ &\text{subject to} && \sum_{i|e\in S_i} x_i \geq 1,\ e \in U \\ & && x_i \in \{0,1\},\ i \in \{1,2,\ldots,m\}\end{aligned}$$

Let the solution to this problem, i.e., the minimum set-cover, be denoted OPT.

Next, argue that if we solve the same problem, but now change the last constraint to $x_i \in [0,1]$ for all i, then the resulting solution OPT_{LP} of this *relaxed* problem obeys $OPT_{LP} \leq OPT$. Note that the relaxed problem is an LP and hence can be solved efficiently.

Solution: For each set S_i, we associate a variable $x_i \in \{0,1\}$ that indicates of we want to choose S_i or not. We can then write the solutions for *Minimum Set-Cover* problem as a vector $x \in \{0,1\}^m$. The objective is clearly to minimize the sum of these variables – moreover, since the sets we select must cover each element in U, we need one constraint to ensure this for each element.

If we replace each constraint $x_i \in \{0,1\}$ with $x_i \in [0,1]$ for all i, then the resulting problem can be solved in a polynomial time as it is a Linear Program (LP). However, by replacing integer constraints with continuous domains, we have expanded the set of feasible solutions – thus, the resulting minimization problem must give a smaller value (as all feasible integer solutions are within our new feasible region), and hence we have that $OPT_{LP} \leq OPT$.

Part (b)

Given a solution z to the relaxed LP, we now round the values to obtain a feasible solution for the original minimum set-cover problem. For each set S_i, we generate $k = c\log n$ i.i.d Bernoulli(z_i) random variables $X_{i,1}, X_{i,2}, \ldots, X_{i,k}$ – if any of them is 1, then we set $x_i = 1$, i.e., we add S_i to our cover $\mathcal{C}$. Prove that the resulting set-cover obeys $\mathbb{E}[|\mathcal{C}|] \leq c\log n \cdot OPT$.

Solution: First, from the definition of the rounding process, for any $j \in \{1,2,\ldots,k\}$, we have:

$$\sum_i \mathbb{E}[X_{i,j}] = \sum_i \mathbb{P}[X_{i,j} = 1] = \sum_i z_i = OPT_{LP}$$

Therefore, by linearity of expectation, we have:

$$\mathbb{E}[|\mathcal{C}|] \leq \sum_i \sum_{j=1}^{c\log n} \mathbb{E}[X_{i,j}] = c\log n \cdot OPT_{LP} \leq c\log n \cdot OPT.$$

Part (c)

Finally, choose c to ensure that the probability that the resulting set-cover $\mathcal{C}$ does not cover any element $e \in \mathcal{U}$ is less than $1/n^2$.

Solution: For any element $e \in U$, and for any $j \in \{1,2,\ldots,c\log n\}$, let $Y_{e,j} = \sum_{i|e\in S_i} X_{i,j}$, and let $Y_e = \sum_{j=1}^{c\log n} Y_{e,j}$. Note that for the sets S_i containing element e, if any of the $X_{i,j} = 1$, then e is covered – in other words, $\mathbb{P}[e \text{ is covered}] = \mathbb{P}[Y_e \geq 1]$. Moreover, $\mathbb{E}[Y_{e,j}] = \sum_{i|e\in S_i} \mathbb{E}[X_{i,j}] = \sum_{i|e\in S_i} z_i$. Let $k_e = |\{i|e \in S_i\}|$; since e is fractionally covered in the LP, we have $z_1 + \cdots + z_{k_e} \geq 1$, and thus $\mathbb{E}[Y_e] \geq c\log n$. Now using our basic Chernoff bound, we have:

$$\begin{aligned}\mathbb{P}[Y_e \leq 0] &= \mathbb{P}\left[Y_e \leq (1-1)\mathbb{E}[Y_e]\right] \\ &\leq e^{\frac{-\mathbb{E}[Y_e]}{2}} \leq e^{\frac{-c\log n}{2}} = \frac{1}{n^{c/2}}\end{aligned}$$

Thus $\mathbb{P}[e \text{ is not covered}] \leq n^{-c/2}$. Moreover, by the union bound, we have that:

$$\mathbb{P}[\text{Any element } e \in U \text{ is not covered}] \leq n^{1-c/2}.$$

If $c \geq 6$, we get $1 - c/2 = -2$, and hence $\mathbb{P}[\text{Every element } e \in U \text{ is covered}] \geq 1 - n^{-2}$.

(Note: You can actually get a tighter bound of $c \geq 3$ using the following argument: First, we need to observe that probability of e being covered is minimized when z_i are all equal, i.e., $z_1 = \cdots = z_{k_e} = 1/k_e$ (this is not obvious - try to prove it...). Thus we get for any j:

$$\mathbb{P}\left[\sum_{i|e\in S_i} Y_{i,j} = 0\right] \leq (1-z_1)\cdots(1-z_{k_e}) \leq \left(1 - \frac{1}{k_e}\right)^{k_e} \leq \frac{1}{e}.$$

Therefore, each element is covered with probability at least $1 - 1/e$ *if we draw only one sample of each* X_i. Since we draw $c \log n$ samples, the probability that element e is not covered is:

$$\mathbb{P}[e \text{ is not covered}] \leq \left(\frac{1}{e}\right)^{c\log n} = \frac{1}{n^c}.$$

Now via the union bound, we see that if we take $c = 3$, then we'll have:

$$\mathbb{P}[\text{Any element } e \in U \text{ is not covered}] \leq \frac{1}{n^2}.$$

Problem 5: (Papadimitrou's 2SAT Algorithm) (Optional)

The 2SAT problem (Wikipedia entry), and more generally, the boolean satisfiability (SAT) problem (Wikipedia entry) are one of the cornerstones of theoretical algorithms, and also a very useful modeling tool for a variety of optimization problems. In the general SAT problem, we want to find a satisfying assignment for a given a Boolean expression in n Boolean (i.e., $\{0, 1\}$, or FALSE/TRUE) variables $\{X_1, X_2, \ldots, X_n\}$ typically involving *conjunctions* (i.e., logical AND, denoted as $\wedge$), *disjunctions* (i.e., logical OR, denoted as $\vee$) and *negations* (logical NOT; typically $\overline{X}$ denotes the negation of a variable X).

In 2SAT, the expression is restricted to being a conjunction (AND) of several *clauses*, where each clause is the disjunction (OR) of two *literals* (either a variable or its negation). For example, the expression $(X_1 \vee X_2) \wedge (\overline{X_1} \vee X_3) \wedge (X_2 \vee \overline{X_3})$ is a 2SAT formula, which has several satisfying assignments including $(X_1 = 1, X_2 = 1, X_3 = 1)$. Note that in order to find a satisfying assignment, we need to set the variables such that each clause in the formula has at least one literal which is TRUE. Although the general SAT problem is known to be NP-complete, 2SAT can be solved in polynomial time – we will now see a simple randomized algorithm that demonstrates this fact:
Papapdimitrou's 2SAT Algorithm): Given a 2CNF formula F involving n Boolean variables, and an arbitrary assignment τ, we check if τ satisfies F. If not, we pick an arbitrary unsatisfied clause, pick one of its literals *uniformly at random*, and flip it to get a new assignment τ'. We then repeat this until we find a satisfying assignment.

Part (a)

Assume that F has a unique satisfying assignment τ^*, and for any assignment τ, let $N(\tau)$ be the number of literals in τ which agree (i.e., have the same value) as the corresponding literal in

τ. Argue that each time we execute an iteration of Papadimitrou's 2SAT algorithm with input assignment τ, the new assignment τ' satisfies:

$$N(\tau') = \begin{cases} N(\tau)+1 & \text{with probability at least } \frac{1}{2} \\ N(\tau)-1 & \text{with probability at most } \frac{1}{2} \end{cases}$$

Solution: Given any unsatisfied clause, we know that in τ^* at least one of the literals is flipped (it could be both are flipped). Since we choose to flip a uniform random literal out of the two, we are guaranteed that we increase the agreement between our guess τ and τ^* by 1 with probability at least 1/2.

Part (b)

Based on the above, argue that the running time T_n of the algorithm is upper bounded by the *first time* that a symmetric random walk starting from 0 hits n or $-n$ (equivalently, $T_n = \arg\min_{K>0} \left\{ |\sum_{i=1}^{K} X_i| = n \right\}$, where X_i are i.i.d Rademacher random variables). Next, show that $\mathbb{E}[T_n] = n^2$, and thus, prove that after $O(n^3)$ iterations, Papadimitrou's algorithm terminates with probability greater than $1 - 1/n$.

Solution: See this lecture (and the next) from Tim Roughgarden's algorithms course for a beautiful exposition of this proof.

One thing you should note: although running the algorithm till $O(n^3)$ steps gives the desired probability, one can get much better bounds if we instead *stop after $O(n^2)$ steps, and then restart the process*. Markov's tells us that after $2n^2$ steps, we find the solution with probability at least $1/2$ – now from previous assignments, you know that doing $O(\log n)$ independent trials is sufficient to amplify the probability to $1 - 1/n$. However, the analysis works for any starting point – so this does suggest that $O(n^2 \log n)$ steps of Papadimitrou's algorithm was sufficient...

Is this surprising? Not really – basically what this says is that the failure probability does follow a 'Chernoff-style' exponential decay. We could not show this directly as the random variables were not independent – however, Markov's inequality still works, and we can exploit that in a clever way to get our result.