1. (30 points) **(Designing a Tournament)** This question will help you understand the applications of Chernoff bound and Union bound more, as you'll deal with tuning parameters to optimize the efficiency of the system. You'll also get an understanding of why NBA tournaments are designed the way they are, with more repeated matches being played as we get closer to the finals. For example, in the NBA, the early rounds are *best-of-five*, and the later rounds become *best-of-seven*. You will understand why, and also learn how to design tournaments with n participants, for large n.

   Suppose there are n teams, and they are totally ranked. That is, there is a well-defined best team, second ranked team and so on. It's just that we (the tournament designer) don't know the ranking. Moreover, assume that for any given match between two players, the better ranked team will win the match with probability $p = \frac{1}{2} + \delta$, *independent of all other matches between these players and all other players also.* Here δ is a small positive constant. All your answers below will depend on the value of δ.

   (a) (5 points) Let n be a power of two, and fix an *arbitrary tournament tree* starting with $n/2$ matches, then $n/4$ matches and so on. That is, initially, team 1 plays team 2, team 3 plays team 4, and so on (each series is only 1 game). The winners advance, and pair up and play each other once, and so on until one team remains. What is the probability that the best team wins the tournament?

   Solution:

   Proof. The best team wins if it wins all the matches that it plays. Since there are n teams in total. A team that wins will have to play a total of $\log n$ matches. The probability that the best team wins a match is always $\frac{1}{2} + \delta$ irrespective of the other teams as its the best team. As matches are independent of each other. The probability that the best team wins $\log n$ matches is $(\frac{1}{2} + \delta)^{\log n}$. □

   (b) (10 points) Now change the tournament to make each series as a *best-of-k* series. How large should k be, and so how many games do you end up conducting in total to get a $1 - \epsilon$ probability of the best team winning the overall series?

   Solution:

   Proof. We first analyze for one series, Probability that the best team loses this round is if it loses in less than equal to $k/2$ rounds. Let X be the random variable that is the number of rounds the best team won in one series. Let X_i

be 1 if the outcome of the ith match in the series is a win and zero otherwise.

$$\forall i, Pr(X_i = 1) = \frac{1}{2} + \delta$$
$$E[X_i] = \frac{1}{2} + \delta$$
$$Var[X_i] = E[{X_i}^2] - E[X]^2$$
$$Var[X_i] = \frac{1}{2} + \delta - (\frac{1}{2} + \delta)^2$$
$$Var[X_i] = \frac{1}{4} - \delta^2$$
$$X = \sum_{i=1} X_i$$
$$E[X] = \sum_{i=1} E[X_i]$$
$$E[X] = n(\frac{1}{2} + \delta)$$

$$Var[X] = \sum_{i=1} Var[X_i] \qquad \text{as random variables are independent}$$

$$Var[X] = n(\frac{1}{4} - \delta^2)$$
$$Pr(X \leq \mu(1-\Delta)) \leq e^{\frac{-\mu\Delta^2}{2}}$$

We want to calculate the probability that X is greater than $k/2$. Equating $\mu(1-\Delta)$ with $k/2$ we find the value of Δ.

$$\Delta = 1 - \frac{k}{2\mu}$$
$$\Delta = 1 - \frac{k}{k + 2\delta k}$$
$$\Delta = 1 - \frac{1}{1 + 2\delta}$$
$$\Delta = \frac{2\delta}{1 + 2\delta}$$

Putting the value of Δ in our chernoff bound equation.

$$Pr(X \leq k/2) \leq e^{(-\frac{k}{2})(\frac{1}{2}+\delta)(\frac{2\delta}{1+2\delta})^2}$$
$$Pr(X \leq k/2) \leq e^{\frac{(-k)(\delta^2)}{1+2\delta})}$$
$$Pr(X > k/2) \geq 1 - e^{\frac{(-k)(\delta^2)}{1+2\delta})}$$
$$Pr(\text{success in } \log n \text{ rounds}) \geq (1 - e^{\frac{(-k)(\delta^2)}{1+2\delta}})^{\log n}$$

If we want the success probability to be greater than $1-\epsilon$. The bound that we have calculated must be greater than $1-\epsilon$.

$$(1-e^{\frac{(-k)(\delta^2)}{1+2\delta}})^{\log n} \geq 1-\epsilon$$

From here we take all terms with k in one side and simplify and find the expression for the number of rounds.

$$k \geq (\frac{1+2\delta}{\delta^2})(\ln \frac{1}{1-(1-\epsilon)^{\frac{1}{\log n}}})$$

□

(c) (15 points) Can you get better dependence on n if you allow different number of games in each round: example, try having k_1 games in the first series, k_2 games in the next series, etc. and optimize for these values to get a total of $O_\epsilon(n)$ games which still retains $1-\epsilon$ probability of the best team winning eventually. Here $O_\epsilon(n)$ means $O(n)$ for all constant $\epsilon > 0$.

Solution:

Proof. Here each round has a different k with it hence each the probability of success formula changes as follows:

$$Pr(X > k/2) \geq 1-e^{\frac{(-k)(\delta^2)}{1+2\delta}})$$

$$Pr(\text{success in } \log n \text{ rounds}) \geq \prod_{s=1}^{\log n} 1-e^{\frac{(-k_s)(\delta^2)}{1+2\delta}})$$

We also have the added constraint that the total number of rounds must be linear, for this constraint to hold good, the following equation must be satisfied.

$$\sum_{s=1}^{\log n} \frac{(n)(k_s)}{2^s} = O_\epsilon(n)$$

Choose your k_s to be $(2^s)(\alpha)/\log n$. Where α is a function of ϵ. From here we get the fact that.

$$Pr(\text{success in } \log n \text{ rounds}) \geq \prod_{s=1}^{\log n} 1-e^{\frac{(-2^s\alpha)(\delta^2)}{(1+2\delta)(\log n)}})$$

We want that this inequality is greater than $1-\epsilon$. To get this, we want the bound on alpha, we observe that when s is smaller then the term $1-e^{\frac{(-2^s\alpha)(\delta^2)}{(1+2\delta)(\log n)}})$

is small, and when s is bigger then this term is bigger, if we proved our term to be bigger than some $(1-\epsilon)^{(1/\log n)}$ for small s, then the product of all terms must be greater than $(1-\epsilon)$. The smallest s possible is 1, hence we need to show the following now:-

$$\forall s, 1 - e^{\frac{(-2^s\alpha)(\delta^2)}{(1+2\delta)(\log n)})} \geq (1-\epsilon)^{(1/\log n)}$$

□

2. (15 points) **(Who wins the election? Exit Poll Design)** Imagine there are only two parties standing in the national election, and you have access to sampling and calling up uniformly random people from the electorate to find out who they're going to vote for. If the total population size of India is N and an unknown $\frac{1}{4} \leq p \leq \frac{3}{4}$ fraction prefer BJP and $(1-p)$ prefer Congress, approximately how many people do you need to sample in order to estimate p upto an additive error of ϵ? Don't try too hard to optimize the constants, so feel free to use whichever concentration bound you see fit.

Solution:

Proof. Lets assume the cardinality of the set of people that we pick to be k. Let X be the random variable that denotes the number of people who support BJP in this set. The probability of BJP supporters that we assume from this set is X/k. We want this estimated probability to be within an error of ϵ, hence we want the condition that $|(X/k) - p| \leq \epsilon$. Applying chebychev bound on this random variable and proceding. Let X_i be the random variable such that the ith person in our set supports BJP.

$$\begin{aligned}
Pr(X_i = 1) &= p \\
E[X_i] &= p \\
Var[X_i] &= E[X_i{}^2] - E[X]^2 \\
Var[X_i] &= p - p^2 \\
Var[X_i] &= p(1-p) \\
X &= \sum_{i=1} X_i \\
E[X] &= \sum_{i=1} E[X_i] \\
E[X] &= kp
\end{aligned}$$

$$Var[X] = \sum_{i=1} Var[X_i] \qquad \text{as random variables are independent}$$

$$Var[X] = kp(1-p)$$
$$Pr(|X-\mu| \geq t) \leq \frac{\sigma^2}{t^2}$$
$$Pr(|X-kp| \geq t) \leq \frac{kp(1-p)}{t^2}$$
$$Pr(|\frac{X}{k}-p| \geq \frac{t}{k}) \leq \frac{kp(1-p)}{t^2}$$
$$Pr(|\frac{X}{k}-p| \geq \epsilon) \leq \frac{p(1-p)}{\epsilon^2 * k}$$
$$p(1-p) \leq \frac{1}{4}$$
$$Pr(|\frac{X}{k}-p| \geq \epsilon) \leq \frac{1}{\epsilon^2 * 4k}$$
$$Pr(failure) \leq \frac{1}{\epsilon^2 * 4k}$$

As you increase k the bound becomes better and better for failure, choosing the value of k depends on how much precision you want. If you want the probability of failure to be at max $1/sqrt(n)$ which seems a nice bound for a large population, choose your k as $sqrt(n)/(4 * \epsilon^2)$. For the Indian population of 1 billion if we want a max error of 0.05, it means we would have to talk to approximately 32 lakh people. □