1. **(5 pt.) Sampling *Without* Replacement**

   Suppose there are n total balls, of which m are red. We sample k of the balls uniformly *without replacement.*[1] Let Z be the random variable denoting how many of the k balls are red. In this problem, you will show that Z is concentrated around its mean.

   (a) **(1 pt.)** Show that $\mathbb{E}[Z] = \frac{km}{n}$.

   (b) **(4 pt.)** When $k \geq 1$, show that $\Pr[|Z - \mathbb{E}[Z]| \geq \lambda] \leq 2e^{-\lambda^2/(2k)}$ for any $\lambda > 0$.
   [**HINT:** *Try applying the Azuma-Hoeffding tail bound to a Doob martingale. When applying Azuma-Hoeffding to a martingale $\{Z_t\}$, feel free to provide a short/intuitive explanation for why $|Z_i - Z_{i-1}| \leq c_i$ rather than a rigorous proof.*]

   (c) [**Optional: this won't be graded.**] When k is close to n, a tighter bound than that from part (b) holds.

      i. **(0 pt.)** When $k = n$, explain why $\Pr[Z = \mathbb{E}[Z]] = 1$.

      ii. **(0 pt.)** When $1 \leq k \leq n-1$, show that $\Pr[|Z - \mathbb{E}[Z]| \geq \lambda] \leq 2e^{-\lambda^2/(2v)}$ where v is defined as
      $$v := \sum_{i=1}^{k} \left(1 - \frac{k-i}{n-i}\right)^2.$$

      iii. **(0 pt.)** Show that $v \leq O(k(n-k)/n)$. This shows that the bound from part (c), ii is tighter than the bound from part (b) when k is close to n.

SOLUTION:

(a) Let $X_i \in \{0, 1\}$ be the indicator variable for whether the i^{th} ball is red. Since $Z = \sum_{i=1}^{k} X_i$, we can apply linearity of expectation to conclude
$$\mathbb{E}[Z] = \sum_{i=1}^{k} \mathbb{E}[X_i] = k \cdot \frac{m}{n}$$

(b) For each $t = 0, \ldots, k$, we define the Doob martingale,
$$Z_t = \mathbb{E}[Z \mid X_1, \ldots X_t].$$
This is a martingale with respect to $\{X_t\}$. Then, $Z_0 = \mathbb{E}[Z]$ with probability 1 and $Z_k = Z$. We aim to apply the Azuma-Hoeffding equality to bound $|Z_k - Z_0|$. Intuitively, revealing whether a single ball is red can only affect the number of final red balls by 1, so $|Z_t - Z_{t-1}| \leq 1$ for all $t = 1, \ldots, k$. For a more careful derivation that accounts for the correlations between whether each ball is red see the solution to part (c), ii.

With the bound $|Z_t - Z_{t-1}| \leq 1$, the desired result follows from the Azuma-Hoeffding inequality.

[1] Note that this only makes sense when $k, m \leq n$.

(c) i. When $k = n$, we sample all of the n balls. Since this is done without replacement, we are guaranteed to select all m red balls.

ii. For each $t = 0, \dots, k$, we define the Doob martingale,

$$Z_t = \mathbb{E}[Z \mid X_1, \dots X_t].$$

This is a martingale with respect to $\{X_t\}$. Then, $Z_0 = \mathbb{E}[Z]$ with probability 1 and $Z_k = Z$. We aim to apply the Azuma-Hoeffding inequality to bound $|Z_k - Z_0|$, which requires upper bounding $|Z_t - Z_{t-1}|$ for each $t = 1, \dots, k$.
Fix a realization for $X_1, \dots, X_t$, and let R be the number of red balls already selected ($R = \sum_{j=1}^{t} X_j$). There are two possible realizations for X_{t+1}: If $X_{t+1} = 1$, then we will have selected a total of $R + 1$ balls in the first $t + 1$ steps and there will be $k - R - 1$ remaining red balls for steps $t + 2, \dots, k$. Applying part (a) to bound the expectation of $X_{t+2} + \cdots + X_k$,

$$\begin{aligned} Z_{t+1} &= \mathbb{E}[Z \mid X_1, \dots, X_{t+1}] \\ &= (R + 1) + \mathbb{E}\left[\sum_{j=t+2}^{k} X_j \,\middle|\, X_1, \dots, X_{t+1} \right] \\ &= R + 1 + \frac{(k - t - 1)(m - R - 1)}{n - t - 1}. \end{aligned}$$

The other possible realization is that $X_{t+1} = 0$. In this case, there will be $k - r$ remaining red balls beginning at time step $t + 2$. Using similar logic to the first case,

$$\begin{aligned} Z_{t+1} &= \mathbb{E}[Z \mid X_1, \dots, X_{t+1}] \\ &= R + \mathbb{E}\left[\sum_{j=t+2}^{k} X_j \,\middle|\, X_1, \dots, X_{t+1} \right] \\ &= R + \frac{(k - t - 1)(m - R)}{n - t - 1}. \end{aligned}$$

Since $\{Z_t\}$ is a martingale with respect to $\{X_t\}$, it must be the case that $\mathbb{E}[Z_{t+1} \mid X_0, \dots, X_t] = Z_t$, which means that Z_t is between the minimum and maximum possible values for Z_{t+1} conditioned on $X_0, \dots, X_t$. Therefore,

$$\begin{aligned} |Z_{t+1} - Z_t| &\le \left(R + 1 + \frac{(k - t - 1)(m - R - 1)}{n - t - 1} \right) - \left(R + \frac{(k - t - 1)(m - R)}{n - t - 1} \right) \\ &= 1 - \frac{k - t - 1}{n - t - 1}. \end{aligned}$$

The desired result follows from the Azuma-Hoeffding bound.

iii. We rewrite v as

$$
\begin{aligned}
v &= \sum_{i=1}^{k} \left(1 - \frac{k-i}{n-i}\right)^2 \\
&= \sum_{i=1}^{k} \left(\frac{n-k}{n-i}\right)^2 \\
&= (n-k)^2 \cdot \sum_{i=1}^{k} \left(\frac{1}{n-i}\right)^2 \\
&= (n-k)^2 \cdot \sum_{j=n-k}^{n-1} \frac{1}{j^2}
\end{aligned}
$$

For any integers $a \leq b$,

$$
\begin{aligned}
\sum_{j=a}^{b} \frac{1}{j^2} &\leq \int_{a-1}^{b} 1/x^2 dx \\
&= \frac{1}{a-1} - \frac{1}{b}.
\end{aligned}
$$

We apply this to our bound for v,

$$
\begin{aligned}
v &= (n-k)^2 \cdot \sum_{j=n-k}^{n-1} \frac{1}{j^2} \\
&\leq (n-k)^2 \cdot \left(\frac{1}{(n-k)^2} + \sum_{j=n-k+1}^{n} \frac{1}{j^2}\right) \\
&\leq (n-k)^2 \cdot \left(\frac{1}{(n-k)^2} + \frac{1}{n-k} - \frac{1}{n}\right) \\
&= 1 + (n-k)^2 \cdot \left(\frac{k}{(n-k)n}\right) \\
&= 1 + \frac{k(n-k)}{n} = O\left(\frac{k(n-k)}{n}\right).
\end{aligned}
$$

2. **(11 pt.) Reaching Consensus**

This question considers a simple and fairly natural model of the dynamic of how opinions shift over time in a group.

Suppose there is an undirected graph $G = (V, E)$ whose vertices represent the group members and a pair of members are friends if and only if they are connected by an edge. For simplicity, we assume that G contains none of the following: 1) self-loops, 2) multiple edges connecting the same pair of vertices, or 3) isolated vertices, i.e., vertices with no edge on them. Let S be the set of possible "opinions" on some topic, and lets suppose that each person has one and only one opinion on the topic at a time. (For concreteness, think of $S = \{A, B, C, ...\}$.) We can represent the opinions of the group members by a mapping $\sigma : V \to S$ where the group member corresponding to vertex v has opinion $\sigma(v)$.

The opinions σ of the group members evolve due to discussions between friends. We model the evolution of σ by the following time-homogeneous Markov chain: starting from the initial opinion σ_0, σ changes from σ_{t-1} to σ_t at step t as follows. Independently for every vertex v, we flip a fair coin. If the outcome is "heads", $\sigma_t(v)$ remains the same as $\sigma_{t-1}(v)$; otherwise, $\sigma_t(v)$ becomes $\sigma_{t-1}(v')$ for a uniformly random neighbor v' of v. In short, every group member keeps their own opinion with probability 1/2, and takes one of their friends' opinion with the remaining 1/2 probability.

In this problem, we will determine the likelihood that the group members reach a certain consensus, given their initial opinions.

(a) **(1 pt.)** If G is disconnected and $|S| > 1$, show that there exist initial opinions σ_0 of the members for which consensus is never reached.

(b) **(3 pt.)** If G is connected, show that consensus is eventually reached almost surely. That is, show that as the number of steps goes to infinity, the probability that consensus has been reached approaches 1.

(c) **(2 pt.)** Let X_t be the number of group members who have some opinion, say $A \in S$ after step t. Give an example where $(X_t)_{t\geq 0}$ is *not* a martingale with respect to $(\sigma_t)_{t\geq 0}$. The example should be one specific tuple (G, S, σ_0).

(d) **(3 pt.)** Let Y_t be the sum of the degrees of the vertices v corresponding to the group members with opinion A after step t. Prove that $(Y_t)_{t\geq 0}$ is a martingale with respect to $(\sigma_t)_{t\geq 0}$.

(e) **(2 pt.)** Assume that G is connected. What is the probability that all members of the group end up with opinion A (ie after some time, everyone has opinion, $A \in S$, for the rest of time)? Express your answer in terms of G and the initial opinion σ_0 of the group members.

[**HINT:** *Try applying the martingale stopping theorem to the martingale* $(Y_t)_{t\geq 0}$.]

SOLUTION:

(a) Let A, B denote two different opinions in S. If the group members in one connected component of G all agree on A initially, and the group members in another component all agree on B initially, there is no way to reach overall consensus.

(b) Suppose G is connected, and suppose that some vertex has opinion A. Then, for each vertex that has opinion A or has a neighbor with opinion A, there is at least a $1/2n$ probability they will have opinion A next round. Therefore there is at least a $(1/2n)^n$ chance that all vertices that have opinion A keep opinion A and all vertices that have a neighbor with opinion A adopts opinion A. If this happens n times in a row, then the entire graph will have opinion A, since the longest path between any two nodes is at most n. Therefore, with probability at least $p = (1/2n)^{n^2} > 0$, the entire graph will have the same opinion after n timesteps. Hence the probability that consensus does not occur in kn timesteps is at most $(1-p)^k$, since we can split it up into k blocks of n timesteps, and we reach consensus in each block with probability p. This goes to 0 as $k \to \infty$, so consensus is reached almost surely.

(c) Suppose the graph of opinions at time 0 is A−B−A. Then, each vertex switches opinion with probability exactly 1/2, so $\mathbb{E}[X_1|\sigma_0] = 3/2$, but $X_0 = 2$.

(d) It is clear that Y_t is determined by σ_t and is finite, so only the third condition remains to be checked. Let d_v denote the degree of vertex v, a_v^t denote the number of neighbors of v with opinion A at time t, and A^t denote the set of vertices with opinion A. Then, we have

$$Y_t := \sum_{v \in A^t} d_v = \sum_v a_v^t.$$

The second inequality follows since the left side counts all outgoing edges from vertices with opinion A and the right side counts all incoming edges to vertices with opinion A. Then, since a_v^t is determined by σ_t, we have

$$\begin{aligned}
\mathbb{E}[Y_{t+1}|\sigma_0, ..., \sigma_t] &= \sum_v d_v P(v \in A^{t+1}|\sigma_t) \\
&= \sum_{v \in A^t} d_v P(v \in A^{t+1}|\sigma_t) + \sum_{v \notin A^t} d_v P(v \in A^{t+1}|\sigma_t) \\
&= \sum_{v \in A^t} d_v (\frac{1}{2} + \frac{a_v^t}{2d_v}) + \sum_{v \notin A^t} d_v \frac{a_v^t}{2d_v} \\
&= \sum_{v \in A^t} \frac{d_v}{2} + \sum_v \frac{a_v^t}{2} \\
&= Y_t.
\end{aligned}$$

(e) Consider the stopping time $T = \min\{t : Y_t = 0 \vee Y_t = \sum_v d_v\}$. This is the first time either nobody has opinion A or everybody has opinion A. Note that once $Y_t = 0$ or $Y_t = \sum_v d_v$, it will not change, and the event of reaching consensus on answer A is the same as the event $Y_T = \sum_v d_v$. From part (b), we know that $T < \infty$ almost surely, and so $Y_T = 0$ or $Y_T = \sum_v d_v$ almost surely. So,

$$\mathbb{E}[Y_T] = P(Y_T = \sum_v d_v) \sum_v d_v.$$

By the stopping theorem,

$$\mathbb{E}[Y_T] = \mathbb{E}[Y_0] = \sum_{v \in A^0} d_v.$$

Therefore,

$$P(Y_T = \sum_v d_v) = \frac{\sum_{v \in A^0} d_v}{\sum_v d_v}.$$