Probability Theory 2 – Proposed solution of model exam

1. In the solution of this question we will use the following version of Chernoff's inequality which was proved in the lecture:

   Theorem 1 *Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, each satisfying $Pr(X_i = 1) = Pr(X_i = -1) = 1/2$. Then*

   $$Pr\left(\sum_{i=1}^{n} X_i > t\right) < e^{-t^2/(2n)}$$

   for every $t > 0$.

   Let Y be the random variable which determines the location of the particle after 100 steps. Let $X_1, \ldots, X_{100}$ be mutually independent random variables such that $Pr(X_i = 1) = Pr(X_i = -1) = 1/2$ for every $1 \leq i \leq 100$. Then $Y = \sum_{i=1}^{100} X_i$. Using Theorem 1 we conclude that

   $$Pr(Y > 20) = Pr\left(\sum_{i=1}^{100} X_i > 20\right) < e^{-20^2/(2 \cdot 100)} = e^{-2}.$$

2. The algorithm does the following:

   (a) For every $1 \leq i \leq 3001$ sample a coordinate of $(x_1, \ldots, x_n)$ uniformly at random with replacement. Denote the sampled value by y_i.

   (b) If $\sum_{i=1}^{3001} y_i \geq 1501$, then output "$(x_1, \ldots, x_n)$ is large". Otherwise output "$(x_1, \ldots, x_n)$ is small".

   It is evident that the algorithm runs in constant time (as usual we assume that sampling one element from a set of size n takes constant time). It remains to prove that, for every vector $(x_1, \ldots, x_n) \in F$, it outputs the correct answer with high probability. In our analysis we will make use of the following version of Chernoff's inequality which was stated in the lecture:

   Theorem 2 *Let $X_1, \ldots, X_n$ be independent and identically distributed random variables whose values lie in the segment $[0, 1]$ and let $X = \sum_{i=1}^{n} X_i$. Then*

   $$Pr\left(X < \mathbb{E}(X) - t\right) < e^{-2t^2/n}$$

   for every $t > 0$.

   Assume first that the input vector $(x_1, \ldots, x_n)$ is large. Since the sampling is uniform and with replacement, for every $1 \leq i \leq 3001$, the probability that $y_i = 1$ is at least $2/3$. It follows that $\mathbb{E}(Y) > 2000$, where $Y = \sum_{i=1}^{3001} y_i$. Since $y_1, \ldots, y_{3001}$ are independent, we can apply Theorem 2 to deduce that the probability that the algorithm erroneously outputs "$(x_1, \ldots, x_n)$ is small" is at most

   $$Pr(Y \leq 1500) \leq Pr(Y < \mathbb{E}(Y) - 500) < e^{-2 \cdot 500^2/3001} \leq 2^{-100}.$$

An analogous argument shows that if "$(x_1, \ldots, x_n)$ is small", then the probability that the algorithm erroneously outputs "$(x_1, \ldots, x_n)$ is large" is at most 2^{-100}. We conclude that, for any input, the probability that the algorithm's outputs is correct is at least $1 - 2^{-100}$ as required.

3. By definition, the entropy of the pair (X, Y) is $H(X,Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} Pr(X = x_i, Y = y_j) \log_2 Pr(X = x_i, Y = y_j)$. Since X and Y are independent, for every $1 \leq i \leq n$ and $1 \leq j \leq m$ we have $Pr(X = x_i, Y = y_j) = p_i q_j$. Hence

$$\begin{aligned} H(X,Y) &= -\sum_{i=1}^{n} \sum_{j=1}^{m} p_i q_j \log_2(p_i q_j) = -\sum_{i=1}^{n} p_i \left[\sum_{j=1}^{m} q_j (\log_2 p_i + \log_2 q_j) \right] \\ &= -\sum_{i=1}^{n} p_i \left[\sum_{j=1}^{m} q_j \log_2 p_i + \sum_{j=1}^{m} q_j \log_2 q_j \right] = -\sum_{i=1}^{n} p_i \left[\log_2 p_i \cdot \sum_{j=1}^{m} q_j - H(Y) \right] \\ &= -\sum_{i=1}^{n} p_i \log_2 p_i + H(Y) \cdot \sum_{i=1}^{n} p_i = H(X) + H(Y), \end{aligned}$$

where the fifth equality holds since $\sum_{j=1}^{m} q_j = 1$ and the sixth equality holds since $\sum_{i=1}^{n} p_i = 1$.

4. Let $H \sim B(n, n, 1/2)$. Denote its parts by X and Y and its edge set by E. In order to prove that H is connected, it suffices to prove that it satisfies the following two properties:

   (a) Every two vertices in the same part have a common neighbour, i.e., for every $u, v \in X$ (respectively, $u, v \in Y$) there exists $w \in Y$ (respectively, $w \in X$) such that $uw, vw \in E$.

   (b) No vertex is isolated (i.e. each vertex is incident to at least one edge).

   Indeed, let $u, v \in X \cup Y$ be two arbitrary vertices. If $u, v \in X$ or $u, v \in Y$, then by (a) there is a path of length 2 between u and v in H. If on the other hand $u \in X$ and $v \in Y$, then by (b) there exists a vertex $w \in Y$ such that $uw \in E$ and by (a) there is a path of length 2 between v and w in H. This yields a path of length at most 3 between u and v in H. Either way, there is a path between u and v in H. Since u and v are two arbitrary vertices, this means that H is connected.

   Therefore, in order to prove the claim, it suffices to show that the probability that H satisfies both (a) and (b) tends to 1 as n tends to infinity. The probability that H does not satisfy (b) is at most $2n(1 - 1/2)^n = n \cdot 2^{1-n}$ and the probability that H does not satisfy (a) is at most

$$2\binom{n}{2}\left(1 - (1/2)^2\right)^n < n^2 \cdot (3/4)^n.$$

   It is evident that

$$\lim_{n \to \infty} Pr(H \text{ is connected}) = 1 - \lim_{n \to \infty} n \cdot 2^{1-n} - \lim_{n \to \infty} n^2 \cdot (3/4)^n = 1.$$