CS365 Written Assignment 1
Khoa Cao
Due

Question 1

Proof. First, since we pick the largest number from the k numbers we sampled, if at least one of the k numbers is larger than the median, then the largest number we picked must be larger than the median.
Let E be the event that all k numbers we examined are smaller than or equal to the median of the n numbers. Then, the complement of E, denoted E^c, is the event that at least one of the k numbers we examined is larger than the median.
Since there are $\lceil n/2 \rceil$ numbers that are smaller than or equal to the median, the probability that any single number we randomly select is smaller than or equal to the median is:

$$\frac{1}{2}$$

Then, we can give a bound on the probability of E and E^c:

$$\begin{aligned}
Pr(E) &= \left(\frac{\lceil n/2 \rceil}{n}\right)\left(\frac{\lceil n/2 \rceil - 1}{n}\right)\cdots\left(\frac{\lceil n/2 \rceil - k}{n}\right) && \text{sampling with replacement} \\
&\leq \left(\frac{1}{2}\right)^k && \text{upper bound on the expression above} \\
\Rightarrow Pr(E) &\leq \left(\frac{1}{2}\right)^k \\
\Rightarrow Pr(E^c) &= 1 - Pr(E) \geq 1 - \left(\frac{1}{2}\right)^k && \text{complement rule} \\
Pr(E^c) &\geq \frac{2^k - 1}{2^k} && \text{simplifying}
\end{aligned}$$

Therefore, the probability that the largest number we picked is larger than the median of the n numbers is at least:

$$\frac{2^k - 1}{2^k}$$

□

Problem 2

Proof. Let X be the RV representing the number of successes we get. Then,

$$\begin{aligned}
X &= \sum_{i=1}^{n} X_i && \text{definition of } X \\
X &\sim Binomial(n, p) && \text{since each coin flip is i.i.d. Bernoulli trial}
\end{aligned}$$

Examining $\bar{X}$, the average number of successes, we have:

$$\begin{aligned}
\bar{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i && \text{definition of } \bar{X} \\
&= \frac{1}{n}X && \text{definition of } X
\end{aligned}$$

Using the linearity of expectation, we can compute the expected value of $\bar{X}$:

$$\begin{aligned}
E(\bar{X}) &= E(\frac{1}{n}X) && \text{definition of } \bar{X} \\
&= \frac{1}{n}E(X) && \text{linearity of expectation} \\
&= \frac{1}{n}(np) && X \sim Binomial(n, p) \\
&= p && \text{simplifying}
\end{aligned}$$

Next, we can compute the variance of $\bar{X}$:

$$\begin{aligned}
Var(\bar{X}) &= Var(\frac{1}{n}X) && \text{definition of } \bar{X} \\
&= \left(\frac{1}{n}\right)^2 Var(X) && \text{property of variance} \\
&= \frac{1}{n^2}(np(1-p)) && X \sim Binomial(n, p) \\
&= \frac{p(1-p)}{n} && \text{simplifying}
\end{aligned}$$

Using Chebyshev's inequality, we can give an upper bound on how likely the value of $\bar{X}$ differs from the bias of the coin p by at least $\frac{p}{10}$:

$$\begin{aligned}
Pr(|\bar{X} - E(\bar{X})| \geq \frac{p}{10}) &\leq \frac{Var(\bar{X})}{(p/10)^2} && \text{Chebyshev's inequality} \\
\Rightarrow Pr(|\bar{X} - p| \geq \frac{p}{10}) &\leq \frac{Var(\bar{X})}{(p/10)^2} && \text{substituting in } E(\bar{X}) \\
&= \frac{\frac{p(1-p)}{n}}{(p/10)^2} && \text{substituting in } Var(\bar{X}) \\
&= \frac{100(1-p)}{np} && \text{simplifying}
\end{aligned}$$

Therefore:

$$Pr(|\bar{X} - p| \geq \frac{p}{10}) \leq \frac{100(1-p)}{np}$$

$\square$

Problem 3

Claim. *Let f and g be valid probability distributions defined over the same domain, S. Then, the convex combination of f and g*

$$h := \lambda f + (1 - \lambda) g$$

for some lambda $\in [0, 1]$ is also a valid probability distribution.

Proof. To show that h is a valid probability distribution, we need to show that:

(i) $h(x) \geq 0 \forall x \in S$

(ii) $\sum_{x \in S} h(x) = 1$

Since the domains of f and g are the same, h is also defined over the same domain, S.
For (i), we have:

$$\begin{aligned}
h(x) &= \lambda f(x) + (1 - \lambda) g(x) && \text{definition of } h \\
\sum_{x \in S} h(x) &= \sum_{x \in S} \lambda f(x) + (1 - \lambda) g(x) && \text{applying summation} \\
&= \sum_{x \in S} \lambda f(x) + \sum_{x \in S} (1 - \lambda) g(x) && \text{linearity of summation} \\
&= \lambda \sum_{x \in S} f(x) + (1 - \lambda) \sum_{x \in S} g(x) && \text{linearity of summation} \\
&= \lambda(1) + (1 - \lambda)(1) && \text{since } f \text{ and } g \text{ are valid probability distributions} \\
&= \lambda + 1 - \lambda && \text{simplifying} \\
&= 1 && \text{simplifying}
\end{aligned}$$

Thus, h satisfies (i).
For (ii), we have:

$$\begin{aligned}
\lambda f(x) &\geq 0 && \text{since } \lambda \in [0, 1] \text{ and } f(x) \geq 0 \\
(1 - \lambda) g(x) &\geq 0 && \text{since } (1 - \lambda) \in [0, 1] \text{ and } g(x) \geq 0 \\
\Rightarrow h(x) = \lambda f(x) + (1 - \lambda) g(x) &\geq 0 + 0 && \text{adding the two inequalities} \\
&= 0 && \text{simplifying}
\end{aligned}$$

Thus, h satisfies (ii).
Since h satisfies both (i) and (ii), h is a valid probability distribution. $\square$

Problem 4

To find the MLE for p^* and μ^*, we will optimize the log-likelihood function:

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} \left(\log \pi_j + \log f(x_i; \vec{\theta_j}) \right)$$

$$s.t. \sum_{j=1}^{k} \pi_j = 1$$

Using the Langrangian multiplier method, we have:

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} \left(\log \pi_j + \log f(x_i; \vec{\theta_j}) \right) - \lambda \left(\sum_{j=1}^{k} \pi_j - 1 \right) \\
&\frac{\partial}{\partial \vec{\theta_w}} \left[\sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} \left(\log \pi_j + \log f(x_i; \vec{\theta_j}) \right) - \lambda \left(\sum_{j=1}^{k} \pi_j - 1 \right) \right] = 0 && \text{differentiating w.r.t. } \vec{\theta_w} \\
&= \sum_{i=1}^{n} \gamma_{iw} \frac{\partial}{\partial \vec{\theta_w}} \log \left(Pr[x_i | \theta_w] \right) && \text{derived in lecture}
\end{aligned}$$

For the Geometric distribution, we have:

$$Pr[x_i | p] = (1 - p)^{x_i - 1} p$$

$$
\begin{aligned}
&\sum_{i=1}^{n} \gamma_{iw} \frac{\partial}{\partial \vec{\theta_w}} \log\left(Pr[x_i|\theta_w]\right) && \text{above} \\
&= \sum_{i=1}^{n} \gamma_{iw} \frac{\partial}{\partial p} \log\left((1-p)^{x_i-1}p\right) && \text{substituting in } Pr[x_i|p] \\
&= \sum_{i=1}^{n} \gamma_{iw} \frac{\partial}{\partial p} \left((x_i-1)\log(1-p) + \log(p)\right) && \text{log property} \\
&= \sum_{i=1}^{n} \gamma_{iw} \left(-(x_i-1)\frac{1}{1-p} + \frac{1}{p}\right) && \text{differentiating} \\
&= \sum_{i=1}^{n} \gamma_{iw} \left(\frac{(1-p)-(x_i-1)p}{p(1-p)}\right) && \text{combining the two fractions} \\
&= \sum_{i=1}^{n} \gamma_{iw} \left(\frac{1-px_i}{p(1-p)}\right) && \text{simplifying} \\
&= \frac{1}{p(1-p)} \sum_{i=1}^{n} \gamma_{iw}(1-px_i) && \text{factoring out } \frac{1}{p(1-p)} \\
&\frac{1}{p(1-p)} \sum_{i=1}^{n} \gamma_{iw}(1-px_i) = 0 && \text{setting equal to 0} \\
&\Rightarrow \sum_{i=1}^{n} \gamma_{iw}(1-px_i) = 0 && \text{multiplying both sides by } p(1-p) \\
&\Rightarrow \sum_{i=1}^{n} \gamma_{iw} - \sum_{i=1}^{n} \gamma_{iw} p x_i = 0 && \text{distributing the summation} \\
&\Rightarrow \sum_{i=1}^{n} \gamma_{iw} = p \sum_{i=1}^{n} \gamma_{iw} x_i && \text{moving the second term to the right side} \\
&\Rightarrow p^* = \frac{\sum_{i=1}^{n} \gamma_{iw}}{\sum_{i=1}^{n} \gamma_{iw} x_i} && \text{dividing both sides by } \sum_{i=1}^{n} \gamma_{iw} x_i
\end{aligned}
$$

For the Borel distribution, we have:

$$
Pr[x_i|\mu] = \frac{e^{-\mu x_i}(\mu x_i)^{x_i-1}}{x_i!}
$$

$$
\begin{aligned}
&\sum_{i=1}^{n} \gamma_{iw} \frac{\partial}{\partial \vec{\theta_w}} \log\left(Pr[x_i|\theta_w]\right) && \text{above} \\
&= \sum_{i=1}^{n} \gamma_{iw} \frac{\partial}{\partial \mu} \log\left(\frac{e^{-\mu x_i}(\mu x_i)^{x_i-1}}{x_i!}\right) && \text{substituting in } Pr[x_i|\mu] \\
&= \sum_{i=1}^{n} \gamma_{iw} \frac{\partial}{\partial \mu} \left(-\mu x_i + (x_i - 1)\log(\mu x_i) - \log(x_i!)\right) && \text{log property} \\
&= \sum_{i=1}^{n} \gamma_{iw} \left(-x_i + (x_i - 1)\frac{1}{\mu}\right) && \text{differentiating} \\
&\sum_{i=1}^{n} \gamma_{iw} \left(-x_i + (x_i - 1)\frac{1}{\mu}\right) && \text{setting equal to 0} \\
&\Rightarrow \sum_{i=1}^{n} \gamma_{iw}(x_i - 1)\frac{1}{\mu} - \sum_{i=1}^{n} \gamma_{iw} x_i = 0 && \text{splitting the summation and rearranging} \\
&\Rightarrow \sum_{i=1}^{n} \gamma_{iw}(x_i - 1)\frac{1}{\mu} = \sum_{i=1}^{n} \gamma_{iw} x_i && \text{moving the second term to the right side} \\
&\Rightarrow \frac{1}{\mu} = \frac{\sum_{i=1}^{n} \gamma_{iw} x_i}{\sum_{i=1}^{n} \gamma_{iw}(x_i - 1)} && \text{dividing both sides by } \sum_{i=1}^{n} \gamma_{iw}(x_i - 1) \\
&\Rightarrow \mu^* = \frac{\sum_{i=1}^{n} \gamma_{iw}(x_i - 1)}{\sum_{i=1}^{n} \gamma_{iw} x_i} && \text{taking the reciprocal of both sides}
\end{aligned}
$$

Therefore, the MLEs are:

$$
p^* = \frac{\sum_{i=1}^{n} \gamma_{iw}}{\sum_{i=1}^{n} \gamma_{iw} x_i}
$$

$$
\mu^* = \frac{\sum_{i=1}^{n} \gamma_{iw}(x_i - 1)}{\sum_{i=1}^{n} \gamma_{iw} x_i}
$$

Problem 5

Proof. Let X be the RV representing the total number of successes we get after repeating the experiement n times. Since each coin flip is i.i.d., we have: $X \sim Binomial(N, p)$.
Here, $N = nm$, since we have n experiements, each with m coin flips. Let x_{ij} be the outcome of the j^{th} coin flip in the i^{th} experiement, then:

$$x_{ij} = \begin{cases} x_{ij} = 1 & \text{if the } j^{th} \text{ coin flip in the } i^{th} \text{ experiement is success} \\ x_{ij} = 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned}
x_{ij} &\sim Bernoulli(p) \text{each coin flip is i.i.d. Bernoulli trial} \\
\Rightarrow X_i = \sum_{j=1}^{m} x_i j & \qquad \text{definition of } X_i \\
\Rightarrow X_i \sim Binomial(m, p) & \qquad \text{since each } x_{ij} \text{ is i.i.d. Bernoulli trial} \\
\Rightarrow X = \sum_{i=1}^{n} X_i & \qquad \text{definition of } X \\
\Rightarrow X = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} & \qquad \text{substituting in } X_i \\
\Rightarrow X \sim Binomial(nm, p) & \qquad \text{since each } X_i \text{ is i.i.d. Binomial trial}
\end{aligned}$$

Let k be the total number of successes we observed:

$$k = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}$$

Then, we can use the likelihood function to derive the MLE for p:

$$
\begin{aligned}
\mathcal{L}(p) &= Pr[X = k|p] && \text{definition of likelihood function} \\
&= \binom{nm}{k} p^k (1-p)^{nm-k} && X \sim Binomial(nm, p) \\
\mathcal{LL}(p) &= \log\left(\binom{nm}{k} p^k (1-p)^{nm-k}\right) && \text{taking the log of the likelihood function} \\
&= \log\left(\binom{nm}{k}\right) + k\log(p) + (nm-k)\log(1-p) && \text{log property} \\
\frac{\partial}{\partial p}\mathcal{LL}(p) &= \frac{\partial}{\partial p}\left[\log\left(\binom{nm}{k}\right) + k\log(p) + (nm-k)\log(1-p)\right] && \text{differentiating} \\
&= 0 + k\frac{1}{p} - (nm-k)\frac{1}{1-p} && \text{differentiating} \\
&= k\frac{1}{p} - (nm-k)\frac{1}{1-p} && \text{simplifying} \\
k\frac{1}{p} - (nm-k)\frac{1}{1-p} &= 0 && \text{setting equal to 0} \\
&\Rightarrow \frac{k}{p} = \frac{nm-k}{1-p} && \text{moving the second term to the right side} \\
&\Rightarrow \frac{1-p}{p} = \frac{nm-k}{k} && \text{cross-multiplying} \\
&\Rightarrow \frac{1}{p} - 1 = \frac{nm-k}{k} && \text{rewriting the left side} \\
&\Rightarrow \frac{1}{p} = \frac{nm-k}{k} + 1 && \text{adding 1 to both sides} \\
&\Rightarrow \frac{1}{p} = \frac{nm}{k} && \text{combining the right side} \\
&\Rightarrow p^* = \frac{k}{nm} && \text{taking the reciprocal of both sides}
\end{aligned}
$$

Therefore, the MLE for p is:

$$p^* = \frac{k}{nm}$$

where k is the total number of successes we observed after repeating the experiement n times, each with m coin flips. $\square$

Extra Credit

Proof. Suppose we use our Monte-carlo algorithm from Problem 1, `pick_only_see_k(pts, k)` to find the median of n numbers by repeatedly sampling k numbers and picking the largest number from the k numbers we sampled. We found that the probability we sample a number larger than the median of the n numbers is at least:

$$\frac{2^k - 1}{2^k}$$

Let X be the RV representing the number of times we have to repeat the sampling process until we pick a number larger than the median. Then,

$$X \sim Geometric(p)$$

where p is the probability of success with each call of `pick_only_see_k(pts, k)`. The amount of times we are expected to call `pick_only_see_k(pts, k)` is:

$$\begin{aligned}
E(X) &= \frac{1}{p} && X \sim Geometric(p) \\
p &\geq \frac{2^k - 1}{2^k} && \text{from problem 1} \\
\Rightarrow \frac{1}{p} &\leq \frac{1}{\frac{2^k-1}{2^k}} && \text{taking the reciprocal of both sides} \\
\Rightarrow E(X) &\leq \frac{1}{\frac{2^k-1}{2^k}} && \text{substituting in } E(X) \\
\Rightarrow E(X) &\leq \frac{2^k}{2^k - 1} && \text{simplifying}
\end{aligned}$$

Therefore, the expected number of times we have to call `pick_only_see_k(pts, k)` until we pick a number larger than the median is at most $\frac{2^k}{2^k-1}$. $\square$