Midterm Exam, Fall 2023: CS 215

Write your name and roll number on the answer sheet. Attempt all questions. You have 2 hours for this exam. Clearly mark out rough work. No calculators or phones are allowed (or required). You may directly use results/theorems we have stated or derived in class, unless the question explicitly mentions otherwise. **Avoid writing lengthy answers.**

Useful Information

1. Markov's inequality: For a non-negative random variable X, we have $P(X \geq a) \leq E(X)/a$ where $a > 0$.
2. Chebyshev's inequality: For a r.v. X with mean μ and variance σ^2, we have $P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$.
3. Gaussian PDF: If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$, MGF $\phi_X(t) = e^{\mu t + \sigma^2 t^2/2}$
4. Poisson PMF: $P(X = i) = \frac{e^{-\lambda}\lambda^i}{i!}$, MGF $\phi_X(t) = e^{\lambda e^t - 1}$
5. Integration by parts: $\int udv = uv - \int vdu$
6. Gaussian tail bound: If $X \sim \mathcal{N}(0, 1)$, then $P(X > x) \leq \frac{e^{-x^2/2}}{x\sqrt{2\pi}}$.

---

1. In a certain town, there exist 100 rickshaws out of which 1 is red and 99 are blue. A person XYZ observes a serious accident caused by a rickshaw at night and remembers that the rickshaw was red in color. Hence, the police arrest the driver of the red rickshaw. The driver pleads innocence. Now, a lawyer decides to defend the hapless rickshaw driver in court. The lawyer ropes in an opthalmologist to test XYZ's ability to differentiate between the colors red and blue, under illumination conditions similar to those that existed that fateful night. The opthalmologist suggests that XYZ sees red objects as red 99% of the time and blue objects as red 2% of the time. What will be the main argument of the defense lawyer? (In other words, what is the probability that the rickshaw was really a red one, when XYZ observed it to be red?) [10 points]

   Solution: Let R_R, R_B be the events that the rickshaw was red, blue respectively. Let X_R, X_B be the events that XYZ perceived a rickshaw to be red, blue respectively. We have $P(X_R|R_R) = 0.99, R(X_R|R_B) = 0.02, P(R_R) = 0.01, P(R_B) = 0.99$. We need to evaluate $P(R_R|X_R) = P(X_R|R_R)P(R_R)/P(X_R)$. $P(X_R) = P(X_R|R_R)P(R_R) + P(X_R|R_B)P(R_B) = 0.99 \times 0.01 + 0.02 \times 0.99 = 0.99 \times 0.03$. Hence $P(R_R|X_R) = \frac{0.99 \times 0.01}{0.99 \times 0.03} = 1/3$. In other words, the probability that the rickshaw was red when XYZ observed it to be red is only 1/3. In other words, it is more probable that the rickshaw was a blue one, based on the available data!

   Marking scheme: 3 points for writing the correct probabilities, $P(X_R|R_R) = 0.99, R(X_R|R_B) = 0.02, P(R_R) = 0.01, P(R_B) = 0.99$, 2 points for correct Bayes formula and 5 points for computing $P(R_R|X_R)$ correctly.

2. If X and Y are two independent continuous random variables with PDFs f_X and f_Y respectively, then prove that the PDF of $Z = XY$ is given by $f_Z(z) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z/x) \frac{1}{|x|} dx$. [10 points]

   Solution: The CDF of Z is given by:

$$F_Z(z) = P(Z \leq z) = P(XY \leq z, X \geq 0) + P(XY \leq z, X \leq 0) \tag{1}$$

$$= P(Y \leq z/X, X \geq 0) + P(Y \geq z/X, X \leq 0) \tag{2}$$

$$= (*) \int_0^{+\infty} f_X(x) \int_{-\infty}^{z/x} f_Y(y) dy dx + \int_{-\infty}^{0} f_X(x) \int_{z/x}^{\infty} f_Y(y) dy dx \tag{3}$$

$$= \int_0^{+\infty} f_X(x)(F_Y(z/x) - 0) dx + \int_{-\infty}^{0} f_X(x)(1 - F_Y(z/x) dx, \tag{4}$$

where the step marked (*) follows due to independence. Taking derivatives w.r.t. z, we obtain the PDF of Z as follows:

$$f_Y(z) = \int_0^{+\infty} f_X(x) f_Y(z/x)/x dx - \int_{-\infty}^{0} f_X(x) f_Y(z/x)/x dx \tag{5}$$

$$= \int_0^{+\infty} f_X(x) f_Y(z/x)/|x| dx + \int_{-\infty}^{0} f_X(x) f_Y(z/x)/|x| dx \tag{6}$$

$$= \int_{-\infty}^{+\infty} f_X(x) f_Y(z/x)/|x| dx \tag{7}$$

Marking scheme: 5 points for the CDF expression, 1 point for marking out the step where independence was used, and 4 points for the PDF. Partial marks for conceptually right answers if there are just some calculation errors. A maximum of 5 out of 10 to be awarded if the possible negativity of X is ignored.

3. (a) Prove that the sum of n independent Gaussian random variables with means $\mu_1, \mu_2, ..., \mu_n$ and variances $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$ respectively is also a Gaussian random variable even when n is small (thus, please do not use the central limit theorem). What is its mean and variance? [4+1=5 points]
(b) If $X \sim \mathcal{N}(0,1)$, then derive the CDF, PDF and mode of $Z = X^2$. [3+1+1=5 points]
Solution: (a) The MGF of the ith Gaussian X_i is $\phi_{X_i}(t) = \exp(\mu_i t + \sigma_i^2 t^2/2)$. The MGF of $S = \sum_{i=1}^n X_i$ is given $\phi_S(t) = \prod_{i=1}^n \phi_{X_i}(t) = \exp(t\sum_{i=1}^n \mu_i + t^2/2 \sum_{i=1}^n \sigma_i^2)$ due to independence of $\{X_i\}_{i=1}^n$. This is the MGF of a Gaussian with mean $\sum_{i=1}^n \mu_i$ and variance $\sum_{i=1}^n \sigma_i^2$, and hence the result follows.
Marking scheme: 3 points for proper substitution of the MGFs. 1 point for mention of independence. 0.5 points each for the mean and variance of the result.
(b) We have $Z = X^2$. Hence $P(Z \leq z) = P(X^2 \leq z) = P(-\sqrt{z} \leq X \leq \sqrt{z}) = F_X(\sqrt{z}) - F_X(-\sqrt{z})$. This yields the CDF of Z in terms of the CDF of X, i.e. $F_X(.)$. The PDF of Z is given by $f_Z(z) = \frac{f_{Z_1}(\sqrt{z})}{2\sqrt{z}} + \frac{f_Z(-\sqrt{z})}{2\sqrt{z}} = \frac{f_Z(\sqrt{z})}{\sqrt{z}} = \frac{e^{-z/2}}{\sqrt{2\pi z}}$. This is the PDF of Z. The mode is clearly at $z = 0$ and the value of the PDF at this point is infinite. As z increases, the numerator decreases and the denominator decreases. Note that z must be non-negative.
Marking scheme: 3 points for CDF, 1 point for PDF and 1 points for mode. It is okay if PDF is in terms of f_Z. No marks out of 1 for an answer containing derivative-based mode which will give -1.

4. Let $X_1, X_2, ..., X_n$ be a sequence of n independent Bernoulli random variables with success probability p. We have seen how to obtain the maximum likelihood estimate of p in class. Suppose now we have the additional knowledge that $p \in \{1/2, 1\}$. Define $Y = \sum_{i=1}^n X_i$. Consider the following estimator U for p:

$$U = \begin{cases} 1, \text{ if } Y = n \\ 0.5 \text{ if } Y < n \end{cases} \tag{8}$$

Now answer the following questions [2.5 + 2.5 + 2.5 + 2.5 = 10 points]:

(a) Derive $E(U)$ if $p = 1$ and if $p = 1/2$.

(b) Is U is a biased estimator for finite n? Is it biased as $n \to \infty$?

(c) Derive the MSE for U and show that it goes to 0 as $n \to \infty$.

(d) Is U a maximum likelihood estimator for p? Why (not)?

Solution: (a) If $p = 1$, then we must have $Y = n$ and hence $U = 1$. That is, if $p = 1$, then $P(U = 1) = 1$, and hence $E(U) = 1$ trivially. If $p = 1/2$, then $E(U) = 1 \times P(Y = n) + 1/2 \times P(Y < n) = 1 \times (1/2)^n + 1/2 \times (1 - (1/2)^n) = 1/2 + (1/2)^{n+1}$. **Marking scheme:** 1 point for the $p = 1$ case and 1.5 points for the $p = 1/2$ case.
(b) If $p = 1$, then $E(U) = 1 = p$ and hence it is always unbiased. However if $p = 1/2$, then we have $E(U) \geq p$. Clearly U is biased. The value of the bias is $(1/2)^{n+1}$ which is 0 as $n \to \infty$. **Marking scheme:** 0.5 points for the bias of the $p = 1$ case, 1.5 points for the bias of the $p = 1/2$ case, and 0.5 points for concluding that

it is asymptotically unbiased.
(c) The MSE for U is 0 when $p = 1$. When $p = 1/2$, the MSE is given by $(1 - 1/2)^2 P(Y = n) + (1/2 - 1/2)^2 P(Y < n) = (1/2)^2 (1/2)^n = (1/2)^{n+2}$. Clearly, the MSE goes to 0 as $n \to \infty$. **Marking scheme:** 0.5 points for the MSE in the $p = 1$ case. 1.5 points for the MSE in the $p = 1/2$ case and 0.5 points for stating that MSE is 0 when n becomes larger.
(d) Yes, U is a maximum likelihood estimator for p. The likelihood function is given as $L(p|\{x_i\}_{i=1}^n) = p^y(1-p)^{n-y}$ where $y = \sum_{i=1}^n x_i$. Note that $p \in \{1/2, 1\}$, and hence $L(1/2|\{x_i\}_{i=1}^n) = (1/2)^n$ both when $p = 1$ and when $p = 1/2$. Also $L(1|\{x_i\}_{i=1}^n) = 1^y(1-1)^{n-y} = 1$ if $y = n$ and $L(1|\{x_i\}_{i=1}^n) = 0$ if $y < n$. Hence the likelihood is maximized by U as it sets $p = 1$ when $Y = n$ (note that when $Y = n$, then $L(1|\{x_i\}_{i=1}^n) > L(1/2|\{x_i\}_{i=1}^n)$) and $p = 1/2$ when $Y < n$ (note that when $Y < n$, we have $L(1|\{x_i\}_{i=1}^n) < L(1/2|\{x_i\}_{i=1}^n)$). **Marking scheme:** No marks without justification. 1.5 marks for the $p = 1$ case and 1 mark for the $= 1/2$ case. Both times, there should be a clear explanation that U is maximizing the likelihood.

5. Suppose some n individuals have arrived at a lab for RTPCR testing. An RTPCR test involves extracting the nasal mucus of the individual and testing it within an RTPCR machine. Suppose that the probability that any individual is infected is p. Instead of individually testing each person, we follow the two-step procedure described below to save on the number of tests: (1) We divide the people into n/g pools, each of size g where we assume that g divides n. Small, equal-volume portions of the mucus samples of all individuals belonging to the same pool are mixed together. This mixture is tested, thus leading to n/g independent tests, one per pool. (2) If the mixture tests negative (non-infected), then all pool members are declared negative. If the mixture tests positive (infected), then each member of the pool is individually tested in a second round of tests. This procedure is called Dorfman pooling. In addition, suppose we assume that each test on a pool containing at least one infected sample has a fixed probability u of correctly returning a positive result, and that each test on a pool of entirely non-infected samples has a fixed probability v of correctly returning a negative result, both independently of the outcomes of all other tests, and both independently of the size of the pool (including a pool-size of 1). As per this protocol, we declare an individual to be infected if and only if both his/her individual test and pooled test are positive. Now answer the following, and **provide some steps/calculations for your answers**:

   (a) What is the expected total number of tests? Note that an individual test counts as one test, and the test of a mixture also counts as one test. [3 points]

   (b) What is the expected number of false negatives for Dorfman pooling? How does this compare to the expected number of false negatives of individual testing? [1+1+1 = 3 points]

   (c) What is the expected number of false positives for Dorfman pooling? How does this compare to the expected number of false positives of individual testing? [3+1+1=5 points]

   Solution: (a) The number of tests in the first round is n/g, one per pool of size g each. The positive pools are tested in the second round. A pool will be reported as positive for one of two reasons: (1) at least one member of the pool is genuinely positive and the pooled test yields a positive result, OR (2) no member of the pool is genuinely positive, but the pooled test yields a positive result falsely. The probability of (1) is $(1 - (1-p)^g)u$ since $1 - (1-p)^g$ is the probability that at least one member of the pool is positive. The probability of (2) is $(1-p)^g(1-v)$ since $(1-p)^g$ is the probability that no member of the pool tested positive and $1 - v$ is the probability that the pooled test did not yield a negative result. Hence the expected number of tests in the second round is $n[(1 - (1-p)^g)u + (1-p)^g(1-v)]$. Hence the expected number of total tests is $n[1/g + (1 - (1-p)^g)u + (1-p)^g(1-v)]$. **Marking scheme:** 0.5 points for the number in the first round, 2.5 points for the number in the second round. In the second round cases, the two cases (1) and (2) need to be clearly explained. In the absence of clear explanation, 1.5 points to be deducted. If one of the cases (i.e. (1) or (2)) is dropped, then 1.5 points are to be deducted.
   (b) The expected number of false negatives in the case of individual testing of n samples is $(1 - u)pn$ as pn is the expected number of genuine positives, of which a fraction $1 - u$ will be falsely reported to be negative. For Dorfman's method, the expected number of individuals reported to be positive will be u^2pn as there are pn infected individuals on average and an individual is reported to be infected only if his/her individual and pooled test are both positive. Hence the total expected number of false negatives is $(1 - u^2)pn$. Thus the ratio of the expected number of false negatives of Dorfman to individual testing is $1 + u$ which less than or

equal to 2. That is with the Dorfman method, there may be a larger number of false negatives. **Marking scheme:** 1 point for the number of false negatives with Dorfman, 1 point for the number of false negatives with individual testing, 1 point for the comparison.
(c) The expected number of false positives in case of individual testing is $(1-v)(1-p)n$ as $(1-p)n$ individuals are non-infected in expectation, and a fraction $1-v$ will be reported as positive (falsely). The expected number of false positives in case of Dorfman testing is as follows: A non-infected person is reported to be positive (falsely) if his/her pooled test is positive and his individual test is also positive. The latter happens with probability $(1-v)(1-p)$. The former happens if either (1) the pool has a genuinely infected person and a positive result was reported, OR (2) the pool has no infected individuals and yet a false positive result was reported. The probability of (1) is $(1-(1-p)^{g-1})u$, and the probability of (2) is $(1-p)^{g-1}(1-v)$. The probability of a false positive report is thus $[(1-(1-p)^{g-1})u+(1-p)^{g-1}(1-v)](1-v)(1-p)$, and so the expected number of false positives is $[(1-(1-p)^{g-1})u+(1-p)^{g-1}(1-v)](1-v)(1-p)n$. If u, v are sufficiently large (more than 0.9) and p is tiny (say 0.1), we see that the expected number of false positives is much lower in Dorfman's testing than individual testing. **Marking scheme:** 1 point for the number of false positives with individual testing. 3 points for the number of false positives with Dorfman, which should be split up as follows: 0.5 for the case when individual test is positive, 1 point for case (1) and 1.5 points for case (2). 1 point for the final comparison.

6. Pooled testing of a different form than Dorfman's method from the previous question has been advocated as a method for saving resources in COVID-19 testing. Instead of testing the samples of n people separately (one sample per person), we create $m < n$ pools and test the m pools. Each pool is obtained by mixing fixed, small portions of a randomly chosen subset of the n samples. Let $\boldsymbol{y}$ be a binary vector of m elements, where $y_i = 1$ if the i^{th} pool is positive (i.e. at least one of the contributing samples is infected) and 0 if the pool is negative (i.e. non-infected which means that none of the contributing samples are infected). Let $\boldsymbol{x}$ be the binary vector of n elements where $x_j = 1$ if the j^{th} sample is infected and 0 otherwise. We have the relationship $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ where $\boldsymbol{A}$ is a $m \times n$ pooling matrix, where $A_{ij} = 1$ if the j^{th} sample contributed to the i^{th} pool and 0 otherwise. For any i, we have $y_i = A_{i1}x_1 \vee A_{i2}x_2 \vee ... \vee A_{in}x_n$ where $\vee$ is the logical OR operator. The aim of pooled testing is to determine $\boldsymbol{x}$ directly from $\boldsymbol{y}$ and $\boldsymbol{A}$. It turns out that such a procedure is feasible if the number of infected people k is small compared to n. Note that we have no knowledge of k beforehand. This question seeks to explore a technique to estimate k directly from $\boldsymbol{y}$ and $\boldsymbol{A}$, assuming that the entries of $\boldsymbol{A}$ are drawn independently from Bernoulli(0.5) (i.e. $P(A_{ij} = 0) = P(A_{ij} = 1) = 0.5$). To this end, answer the following questions: [2.5+2.5+2.5+2.5=10 points]

   (a) Let d_i be the number of entries for which A_{ij} and x_j are both unequal to 0, where $1 \leq j \leq n$. What is the distribution of d_i, if the number of non-zero elements of $\boldsymbol{x}$ is k, since the entries of $\boldsymbol{A}$ are drawn independently from Bernoulli(0.5)?

   (b) Prove that $P(y_i = 0) = P(d_i = 0)$.

   (c) Let H be a random variable for the number of non-zero elements in $\boldsymbol{y}$. Then what is the distribution of H, if the number of non-zero elements of $\boldsymbol{x}$ is k?

   (d) Express k in terms of $P(d_i = 0)$ and hence write the maximum likelihood estimate of k given $\boldsymbol{y}$.

   Solution: (a) d_i is a random variable that counts the number of indices i at which $x_j = 1$ and $A_{ij} = 1$. Now, $\boldsymbol{x}$ has k ones. Since the elements A_{ij} are Bernoulli distributed, we can consider d_i to be binomial distributed with k trials and success probability 1/2. **Marking scheme:** 2.5 points for the statement that is a binomial random variable with appropriate number of trials and appropriate success probability.
   (b) $P(y_i = 0) = \sum_{j=0}^{k} P(y_i = 0|d_i = j)P(d_i = j) = P(d_i = 0)$ because $P(y_i = 0|d_i = 0) = 1$ and $P(y_i = 0|d_i > 0) = 0$. **Marking scheme:** 2.5 points for the proof.
   (c) H counts the number of ones in $\boldsymbol{y}$ which has m independent elements. These can be regarded as m independent Bernoulli trials with success probability $1 - P(y_i = 0) = 1 - P(d_i = 0)$. **Marking scheme:** 2.5 points for correctly identifying this as a binomial random variable with the correct success probability.
   (d) Given the distribution of d_i, we know that $P(d_i = 0) = C(k, 0)(1/2)^0(1/2)^{k-0} = (1/2)^k$. But we know that $P(d_i = 0) = P(y_i = 0)$. Hence $(1/2)^k = P(y_i = 0)$, due to which $k = -\log_2 P(y_i = 0)$. But the MLE of $P(y_i = 0)$ is given as $\hat{P}(y_i = 0) = 1 - \#\text{number of non-zero entries in } \boldsymbol{y}/m$. Thus $k =$

$-\log_2(1 - \#\text{number of non-zero entries in } \boldsymbol{y}/m)$. **Marking scheme:** 1 point for the relation that $P(y_i = 0) = (1/2)^k$. 1.5 points for plugging in the MLE of $P(y_i = 0)$ from $\boldsymbol{y}$.