AUBURN UNIVERSITY

Stat 7650 - Computational Statistics

In-Class Component - Midterm March 25, 2025

Duration: 75 minutes

Instructions:

1. This is the in-class component of the MT exam. If you want to ask a question, please raise your hand, but no answer is guaranteed unless there is an ambiguity or typo in the question.
2. Calculators are permitted but must be supplied by you.
3. Work all the problems. Please submit only the solutions you want to be grade (i.e., do not submit any scratch work).
4. You need to start each problem on a new page. Clearly label each problem and write your name at top right of each page.
5. You must always **explain your answers** and **show your work** to receive **full credit** or proper **partial credit**. If necessary, you can use both sides of the sheets, but make sure you have indicated doing so.
6. There are 4 questions. Please make sure you have the pages with all 4 questions.

QUESTIONS

Q1. Let $X_1, X_2, \ldots, X_k$ be a random sample from a Binomial distribution (i.e. $X_i \overset{iid}{\sim} \text{Bin}(n, p)$, for $i = 1, \ldots, k$, where n is fixed and "Bin" denotes Binomial distribution) with pmf:

$$f(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n, \quad 0 \leq p \leq 1.$$

(a) Letting $y = \sum_{i=1}^{k} x_i$, derive the **likelihood function** $L(p)$ and the **log-likelihood function** $\ell(p)$.

(b) Derive the **maximum likelihood estimator (MLE)** of p using calculus.

(c) Suppose a random sample of $k = 100$ is observed from the $\text{Bin}(n = 20, p)$ distribution and let $y = \sum_{i=1}^{k} x_i$. (i) $y = 500$; (ii) $y = 1000$; and (iii) $y = 1500$.
For each case in Q1(b), compute the MLE estimate for p based on the given sample.

(d) Derive the **Fisher Information** $I(p)$ for n Bernoulli observations.

Q2. Can you write each of Newton, Secant, and Fisher Scoring methods as special cases of the Fixed Point Iteration Algorithm. If so write the corresponding $F(x)$ whose fixed point would yield the solution. If not explain why not. (**Hint:** This question is asking whether you can write the updating equation for these algorithms as $x_{t+1} = F(x_t)$ for some function F.)

Q3. Let $Y_1, Y_2 \overset{iid}{\sim} \text{Poisson}(\lambda)$ with $y_1 = 3$ observed and y_2 missing.

The probability mass function (PMF) of the Poisson(λ) distribution is

$$f_Y(y|\lambda) = P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad y = 0, 1, 2, \ldots$$

(a) Show that the complete data log-likelihood function is

$$\log L(\lambda|\mathbf{y}) = -2\lambda + (y_1 + y_2)\log\lambda - \log(y_1!) - \log(y_2!).$$

(b) Since y_2 is unobserved, we apply the EM algorithm.
E-step: Compute the expected complete data log-likelihood, i.e.,

$$Q(\lambda|\lambda^{(t)}) = \mathbb{E}[\log L(\lambda|\mathbf{y}) \mid y_1, \lambda^{(t)}].$$

(c) **M-step:** Maximize $Q(\lambda|\lambda^{(t)})$ with respect to λ to obtain $\lambda^{(t+1)}$.

(d) Derive the EM updating equation $\lambda^{(t+1)}$ and determine the limiting value as $t \to \infty$.

Q4: Consider maximizing a real-valued function $g(\mathbf{x})$, where $\mathbf{x}$ is a vector in $\mathbb{R}^p$. Two methods presented in class—and used in the homework—are Newton's method and the method of steepest ascent. Discuss the pros and cons of each method.

AUBURN UNIVERSITY

Stat 7650 - Computational Statistics

Take-Home Component - Midterm   March 25 2025

Due 9:30 am on March 27, 2025

Instructions:

1. This is the take-home exam component of the MT exam. If you want to ask a question, please email me, but no answer is guaranteed unless it is to clarify the question.
2. Work all the problems. Please submit only the solutions you want me to grade (i.e., do not submit any scratch work).
3. You need to start each problem on a new page. Clearly label each problem and write your name at top right of each page.
4. You must always **explain your answers** and **show your work** to receive **full credit** or proper **partial credit**. If necessary, you can use both sides of the sheets, but make sure you have indicated doing so.
5. There are 2 questions in this part. Please make sure you have the sheet with the 2 questions.

QUESTIONS

Q5. Consider a random sample of size 20 from a t-distribution with 5 degrees of freedom (df), i.e., $X_i \overset{iid}{\sim} t_5$.

(a) Write down the probability density function (pdf) of a t-distribution with ν degrees of freedom.

(b) Now, assume that df ν were not available. Formulate the likelihood function for the given data and derive the log-likelihood function (for ν). Based on the log-likelihood you obtained, discuss whether an analytical solution for the MLE of ν is feasible or if numerical methods are required.

(c) Generate (i.e. simulate) a dataset of size 20 from a t-distribution with 5 degrees of freedom (df) (use `set.seed(123)`), i.e.,

$$X_1, X_2, \ldots, X_{20} \sim t_5.$$

Using this sample as your dataset, plot the log-likelihood as a function of ν and estimate the MLE for ν, the degrees of freedom. **Hint:** You may use the built in `digamma` and `trigamma` functions in `R` for the derivative of the gamma function, $\Gamma(x)$.

(d) Is your MLE estimate close to the true value? If not, would increasing n help? Why or why not?

Q6. Consider a random sample of size 20 from a Binomial distribution with unknown parameters n and p, i.e.,

$$X_i \overset{iid}{\sim} \text{Bin}(n, p).$$

(a) Write down the probability mass function (pmf) of a Binomial distribution with parameters n and p.

(b) Now, assume that both n and p are unknown. Formulate the likelihood function for the given data and derive the log-likelihood function for n and p Discuss whether an analytical solution for the MLE of n and p is feasible or if numerical methods are required.

(c) Set `seed(123)` and simulate 20 observations from a Binomial($n = 10, p = 0.4$) distribution. Using this sample, apply the **BFGS** method (e.g., `optim(..., method = "BFGS")`) to estimate the MLE of (n, p). You may use the built-in `digamma` and `trigamma` functions in `R` for the derivatives of the gamma function. Since n is discrete, treat n as continuous during optimization and then round the final estimate to the nearest integer.