STAT 7650: Homework 4
(Due: Tuesday, 03/18/2025)

Note: Show all your work for the necessary steps to receive full credit.

Please turn in the HW on paper, hand-written and/or typed. For computational problems, return only the relevant parts of the output with comments/annotations. No need to turn in your code file or send it via email or print long lists of generated samples, but you are welcome to email it to me. However, from the code you are using to answer the problems, turn in the relevant code and output with explanation and justification, and the figures (if requested), preferably printed from the output.

Please disclose any us of AI in your solutions. Regardless of you use it or not, make sure you submit your own work, not copy from other source(s). Any suspicion of AI use will result in automatic 0 or substantial point loss in any question.

Q1. **Trivial Example - 2 (from slides)** Let $Y = (X, Z)$, where X, Z are i.i.d. from $N(\theta, 1)$, but Z is missing. Observed data MLE is $\hat{\theta} = X$.

(a) Show that the Q function in the E-step is (effectively) given by:

$$Q(\theta|\theta^{(t)}) = -\frac{1}{2}\left[(\theta - X)^2 + (\theta - \theta^{(t)})^2\right].$$

(b) Find the **M-step Update** — what should happen as $t \to \infty$?

(c) For $X = 5$, use an AI-based tool to simulate the M-step update process as $t \to \infty$. Compare this with the theoretical expectations and discuss any insights or anomalies observed.

Q2. **Extension of Trivial Example - 2 (from slides)** Let $\mathbf{Y} = (\mathbf{X}, \mathbf{Z})$, where $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_m)$ with both X_i and Z_j are jointly independent and i.i.d. from $N(\theta, 1)$, but $\mathbf{Z}$ is missing. Observed data MLE $\hat{\theta} = \bar{X}$.

(a) Derive the Q function in the E-step.

(b) Find the updating equation doing the maximization in the **M-step**. What should happen as $t \to \infty$?

(c) What do you think is the effect of relative magnitude of m and n. Try experiments with $(m, n) = (50, 50)$, $(m, n) = (25, 75)$, and $(m, n) = (75, 25)$. Use $\theta = 1$ for the generation of each of the observed sample $\mathbf{X}$. Use `set.seed(123)` for each of the samples.

Q3. **Example 2: Probit Regression (from class slides)** Recall the **probit regression model**: $X_i \sim \text{Ber}(\Phi(\mathbf{u}_i^T\boldsymbol{\theta}))$. The EM algorithm facilitates obtaining the MLE of $\boldsymbol{\theta}$.
Complete Data Representation: $Y = (Y_1, \ldots, Y_n)$, where $Y_i \sim N(\mathbf{u}_i^T\boldsymbol{\theta}, 1)$, and X_i is defined as:

$$X_i = \begin{cases} 1 & \text{if } Y_i > 0 \\ 0 & \text{if } Y_i \leq 0 \end{cases}$$

Verify that X_i, defined in this way, has the same distribution as that given by the probit model.

Q4. Consider the *Pima Indians Diabetes Dataset* named as `pima` from R's `faraway` package). Extract the following variables from this dataset: `- test` (1 = diagnosed with diabetes, 0 = no diabetes), `- glucose` (Plasma glucose concentration), `- bmi` (Body Mass Index), `- age` (Age in years), and `- diastolic` (Blood pressure).

(a) Using this dataset, implement the **EM algorithm** to estimate $\boldsymbol{\theta}$ for the probit regression model. **Hint:** Treat the variable `test` as the response, and the other variables as predictors, $\mathbf{u}_i$.

(b) Compare your estimated $\boldsymbol{\theta}$ in part (a) with estimates from the probit regression model using the built-in functions in `R` for the probit regression.

(c) Compare your estimated $\boldsymbol{\theta}$ in parts (a) and (b) with estimates from a logistic regression model. (You may use built-in functions in `R` for the logistic regression.)

(d) **Optional:** Plot the estimated decision boundary and assess the classification performance of the probit model.