STA2311 (Fall 2023) - Homework 1

Due October 31, 2023

Instructions: This homework is to be completed using R Markdown. Each question is worth 25 marks and includes both theoretical and coding aspects. For each question, you must first derive the relevant mathematical formulas and typeset them in LaTeX using notation consistent with that from lecture, carefully justifying all of your steps (using complete sentences) and making references to relevant material from lecture or elsewhere (with appropriate citations). You must then implement your work in `R` in order to numerically solve the specific problem asked in the question, thoroughly commenting your code and formatting it with appropriate indentations and whitespace. Aside from base packages, you must *not* use any `R` packages except for the Tidyverse package (which is optional).

Formatting and Submission: In R Markdown, your output must be in .pdf format, rather than HTML or Microsoft Word (ugh). While the numerical inputs and outputs of your code should appear within the main text of your paper (i.e., the main text should include sentences along the lines of "We initialize the algorithm at $\alpha^{(0)} = 0.5$" and "The algorithm converges after 39493 iterations, yielding the final estimate $\hat{\alpha} = 0.043$"), your code should appear in an appendix at the end of your paper; see here for instructions. You must submit a hard copy of your .pdf in class on October 31, and you must submit the .Rmd file which generates your paper using the virtual assignment dropbox on Quercus. Someone running your .Rmd on another machine should be able to reproduce your document *exactly*, so remember to set seeds whenever they are appropriate.

Collaboration: While you may discuss the homework with your peers, the work you submit should be entirely your own.

1. Consider the sample of size $n = 10,000$ in the file `banana.txt` which we model using the distribution

$$
\begin{aligned}
f(\boldsymbol{y}|\boldsymbol{\theta}) = & \frac{1}{(2\pi)^2\theta_1\theta_2} \\
& \times \exp\left(-\frac{1}{2}\left[\frac{y_1^2}{\theta_1^2} + \frac{y_2^2}{\theta_2^2}\right]\right)\exp\left(-\frac{1}{2}\left\{\left[\frac{y_3 - \theta_3(y_1^2 - \theta_1)}{\theta_1}\right]^2 + \left[\frac{y_4 - \theta_4(y_2^2 - \theta_2)}{\theta_2}\right]^2\right\}\right),
\end{aligned}
$$

   where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_4)$ with $\theta_1, \theta_2 > 0$ and $\theta_3, \theta_4 \in \mathbb{R}$, and $\boldsymbol{y} = (y_1, \ldots, y_4)$.

   (a) Find the MLE for $\boldsymbol{\theta}$ using both gradient descent (with backtracking line search) and stochastic gradient descent.

   (b) Compare the computation costs for the two methods.

2. The `cities.txt` includes the geographical (x, y) coordinates of 20 city centers within a small rectangle-shaped state. Use the simulated annealing algorithm to determine where to build three hospitals in order to minimize the the sum of the distances to the cities that are nearest.

3. Consider a probit regression model: for $i = 1, \ldots, n$, given a vector of covariates $\boldsymbol{x}_i \in \mathbb{R}^p$,

$$
\begin{aligned}
z_i &= \boldsymbol{\beta}^\top \boldsymbol{x}_i + \epsilon_i \\
y_i &= \mathbb{1}_{z_i > 0},
\end{aligned}
$$

   where $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, and only the y_i's are observed. Using the $(y_i, \boldsymbol{x}_i)$ data in the `probit.txt` file (with $n = 10,000$ and $p = 5$), devise an EM algorithm to estimate $\boldsymbol{\beta}$, and produce confidence intervals for your estimator using the SEM algorithm.

4. Consider the sample of size $n = 5,000$ in the file `mixture.txt` which we model using a K-mixture of exponential distributions:
$$f(y \mid \boldsymbol{\pi}, \boldsymbol{\lambda}) = \sum_{k=1}^{K} \pi_k \cdot \lambda_k e^{-\lambda_k y}$$
where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ is such that each $\pi_i > 0$ and $\sum_{k=1}^{K} \pi_k = 1$, and $\lambda_1, \ldots, \lambda_K > 0$. In this question we are Bayesians, and we impose a Dirichlet$(1, \ldots, 1)$ prior on $\boldsymbol{\pi}$ and an Exp(1) prior on each λ_k. Derive the mean-field approximation of the posterior distribution $p(\boldsymbol{\pi}, \boldsymbol{\lambda} \mid \boldsymbol{y})$ and compare the results for $K = 2, 3, 4, 5$ (note that the number of mixture components K is a "known" parameter).