Assignment #3 (Due February 27)

1) (**Computing problem**) This problem explores the sensitivity of a simulated performance measure to various modeling assumptions about the arrival process to a service system, and also introduces a method for modeling autocorrelation of interarrival times. Consider a GI/G/1 queue in which the service time distribution is symmetric and triangular, taking values in [0, 1.98] and having a mean of 0.99. (All times are in minutes.) The performance measure μ of interest is the expected value of the average number of jobs in the system over the interval [0, 500]. Using the GSMP simulation algorithm, write a program that simulates this queueing system, assuming an arrival at time 0 to an empty system. Your program should be able to estimate μ for the five arrival processes listed below. Use 10,000 repetitions for each arrival process, which will allow estimation of μ to within less than roughly $\pm 1\%$ error with 95% probability. Use different random numbers for each of Parts (a)–(e). Give an intuitive explanation of the results that you observe in terms of "clumping behavior" of the arriving jobs. The arrival process are

   a) (Baseline) A Poisson process with rate $\lambda = 1$. That is, successive interarrival times are i.i.d. according to an exponential distribution with mean 1 and standard deviation 1.

   b) A process in which successive interarrival times are i.i.d according to a Weibull distribution with scale parameter $\lambda = 0.8856899$ and shape parameter $\alpha = 2.1013491$. This distribution has mean = 1 and standard deviation = 0.5. Recall that the cdf of a Weibull distribution is $F(x) = 1 - e^{-(\lambda x)^\alpha}$ for $x \geq 0$ and $F(x) = 0$ if $x < 0$, so that the exponential distribution can be viewed as a special case of the Weibull distribution with $\alpha = 1$.

   c) A process in which successive interarrival times are i.i.d according to a Weibull distribution with scale parameter $\lambda = 1.7383757$ and shape parameter $\alpha = 0.5426926$. This distribution has mean = 1 and standard deviation = 2.0

   d) A process in which each interarrival time has an exponential distribution with mean = 1, but successive interarrivals time are correlated in the following way. Let $(Z_n : n \geq 0)$ be a sequence of i.i.d. $N(0,1)$ random variables, and set $Y_n = c(Z_n - Z_{n-1})$ for $n \geq 1$, where $c = 1/\sqrt{2}$. Finally, set $X_n = -\log(\Phi(Y_n))$ for $n \geq 1$, where Φ is the cdf of the $N(0,1)$ distribution. What is $E[Y_n]$ and what is $\mathrm{Cov}[Y_n, Y_{n-j}]$ for $j = 0,1,2,\ldots$? What is the distribution of each Y_n? Explain why each X_n has a unit exponential distribution. [Hint: Use a result from Assignment #2.] Note: When programming the simulation in Python, you can use the norm.cdf() function from scipy.stats.
   To generate normal random variates, use the Box-Muller method as with Assignment #1.

   e) Same questions, and simulation, as part (d), except that now $Y_n = c(Z_n + Z_{n-1})$.

2) For each of the following distributions, derive the maximum likelihood estimates for the indicated parameters:

a) $U(0,b)$, MLE for b
b) $U(a,b)$, joint MLEs for a and b
c) $N(\mu,\sigma^2)$, joint MLEs for μ and σ

3) Consider the following density functions (with $0 \le a < b < c$):

$$f_{Y_1}(x) = \begin{cases} \dfrac{2(x-a)}{(b-a)^2} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad f_{Y_2}(x) = \begin{cases} \dfrac{2(c-x)}{(c-b)^2} & \text{if b} \le x \le c \\ 0 & \text{otherwise} \end{cases}$$

a) Let Z_1, Z_2 be two independent uniform random variables on $[a,b]$. Prove that the random variable $\max(Z_1,Z_2)$ has density f_{Y_1}. Give an alternative algorithm for generating a sample from f_{Y_1} using inversion. [Hint: $\max(Z_1,Z_2) \le x$ if and only if $Z_1 \le x$ and $Z_2 \le x$.]

b) Let Z_1, Z_2 be two independent uniform random variables on $[b,c]$. Prove that the random variable $\min(Z_1,Z_2)$ has density f_{Y_2}. Give an alternative algorithm for generating a sample from f_{Y_2} using inversion.

4) (Extra credit, to be done individually) Suppose that we are trying to fit a distribution to data that appears to have two modes (i.e., peaks in the histogram of the data). Based on prior knowledge, we think that a good model for the underlying distribution is a mixture of two normal distributions. Each normal has a variance equal to 1, so that the overall pdf is

$$f(x;\alpha,\mu_1,\mu_2) = \alpha \frac{1}{\sqrt{2\pi}} e^{-(x-\mu_1)^2/2} + (1-\alpha)\frac{1}{\sqrt{2\pi}} e^{-(x-\mu_2)^2/2}.$$

That is, for each $j = 1,2,\ldots,n$, we effectively generate X_j from a $N(\mu_1,1)$ distribution with probability α and from a $N(\mu_2,1)$ distribution with probability $1-\alpha$. Our goal is to estimate the unknown parameter vector $\Theta = (\alpha,\mu_1,\mu_2)$.

a) First suppose that the data consist of observations $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and labels $\mathbf{L} = (L_1, L_2, \ldots, L_n)$, where $L_j = i$ if X_j was generated from the normal distribution with mean μ_i ($i = 1,2$). Derive formulas for the MLE estimates of α,μ_1,μ_2. Write these formulas in the form $\hat{\alpha} = f_1(N_1,S_1,S_2)$, $\hat{\mu}_1 = f_2(N_1,S_1,S_2)$, and $\hat{\mu}_2 = f_3(N_1,S_1,S_2)$ for appropriate functions f_1, f_2, and f_3, where $N_1 =$ the random number of the X_j's generated according to $N(\mu_1,1)$ and $S_i = \sum_{j\in A_i} X_j$ for $i = 1,2$. Here $A_i = \{j : L_j = i\}$, so that N_1 is the number of elements in A_1. [Hint: The sets A_1 and A_2 are useful when writing out the likelihood function.]

b) In practice, we won't get to see the labels $L_1, L_2, \ldots, L_n$, so the straightforward approach of part (a) won't quite work. We can use the following iterative approach, however.

i) Fix an initial guess $\hat{\Theta}^{(0)} = (\hat{\alpha}^{(0)}, \hat{\mu}_1^{(0)}, \hat{\mu}_2^{(0)})$ for the parameter values and set $m = 0$
ii) Compute $\hat{N}_1^{(m)} = E[N_1 \mid \hat{\Theta}^{(m)}, \mathbf{X}]$, $\hat{S}_1^{(m)} = E[S_1 \mid \hat{\Theta}^{(m)}, \mathbf{X}]$, $\hat{S}_2^{(m)} = E[S_2 \mid \hat{\Theta}^{(m)}, \mathbf{X}]$
iii) Compute $\hat{\Theta}^{(m+1)}$ via $\hat{\alpha}^{(m+1)} = f_1(\hat{N}_1^{(m)}, \hat{S}_1^{(m)}, \hat{S}_2^{(m)})$, $\hat{\mu}_1^{(m+1)} = f_2(\hat{N}_1^{(m)}, \hat{S}_1^{(m)}, \hat{S}_2^{(m)})$, and $\hat{\mu}_2^{(m+1)} = f_3(\hat{N}_1^{(m)}, \hat{S}_1^{(m)}, \hat{S}_2^{(m)})$, where f_1, f_2, and f_3 are as in part (a).
iv) If estimates have not converged, set $m \leftarrow m+1$ and go to step (ii)

Fill in the missing details of the algorithm by giving explicit formulas for $\hat{N}_1^{(m)}$, $\hat{S}_1^{(m)}$, and $\hat{S}_2^{(m)}$ in step (ii). [Hint: write $N_1 = \sum_{j=1}^n I(L_j = 1)$, so that, using the independence of the X_j's,

$$\begin{aligned} E[N_1 \mid \hat{\Theta}^{(m)}, \mathbf{X}] &= E\left[\sum_{j=1}^n I(L_j = 1) \middle| \hat{\Theta}^{(m)}, \mathbf{X}\right] \\ &= \sum_{j=1}^n E[I(L_j = 1) \mid \hat{\Theta}^{(m)}, \mathbf{X}] \\ &= \sum_{j=1}^n P(L_j = 1 \mid \hat{\Theta}^{(m)}, \mathbf{X}) \\ &= \sum_{j=1}^n P(L_j = 1 \mid \hat{\alpha}^{(m)}, \hat{\mu}_1^{(m)}, \hat{\mu}_2^{(m)}, X_j). \end{aligned}$$

Use Bayes' Theorem to write the quantity $P(L_j = 1 \mid \hat{\alpha}^{(m)}, \hat{\mu}_1^{(m)}, \hat{\mu}_2^{(m)}, X_j)$ as a ratio that involves normal density functions; when applying Bayes' Theorem, you can treat $\hat{\alpha}^{(m)}$, $\hat{\mu}_1^{(m)}$, and $\hat{\mu}_2^{(m)}$ as constants so that

$$\begin{aligned} &P(L_j = 1 \mid \hat{\alpha}^{(m)}, \hat{\mu}_1^{(m)}, \hat{\mu}_2^{(m)}, X_j) \\ &\quad = \frac{P(L_j = 1 \mid \hat{\alpha}^{(m)}, \hat{\mu}_1^{(m)}, \hat{\mu}_2^{(m)}) P(X_j \mid \hat{\alpha}^{(m)}, \hat{\mu}_1^{(m)}, \hat{\mu}_2^{(m)}, L_j = 1)}{P(X_j \mid \hat{\alpha}^{(m)}, \hat{\mu}_1^{(m)}, \hat{\mu}_2^{(m)})} \\ &\quad = g(\hat{\alpha}^{(m)}, \hat{\mu}_1^{(m)}, \hat{\mu}_2^{(m)}, X_j) \end{aligned}$$

for an appropriate function g. Similarly, we can write $S_i = \sum_{j=1}^n I(L_j = i) X_j$ for $i = 1, 2$.]

c) Run the algorithm (using R, Matlab/Octave, Excel, Python, etc.) to fit parameters to the dataset HW3.dat that is posted on the class web page.