Assignment #3 (Due May 3)

1. Consider the loading dock with \( N = 8 \) items from Assignment #2, Problem 2, but with the following changes. (a) The conveyor is subject to occasional breakdowns. In the absence of breakdowns, the time to transport an item to the loading dock is a deterministic constant \( Q \) (i.e., an item would arrive every \( Q \) time units). When a breakdown occurs, the conveyor immediately stops and a repair starts. As soon as the repair is complete, the conveyor starts transporting the item from wherever the item was when the breakdown occurred. Thus if \( Q = 20 \) minutes and a breakdown occurs 3 minutes after the conveyor started to transport the item, then the item will arrive on the dock 17 minutes after the repair is completed (assuming no further breakdowns). (b) The trucks still drive to the loading dock independently and in parallel, but the time for each truck \( i \) to arrive at the dock now is distributed as a random variable \( A_i \), having an Erlang(2,\( \lambda \)) distribution, i.e., is distributed as the sum of two \( \exp(\lambda) \) random variables; thus the arrival times for the different trucks are identically distributed and mutually independent. (c) The successive times between a conveyor repair and the next breakdown are i.i.d. as a random variable \( L \) having a symmetric triangular distribution on the interval \([0, V]\), i.e., with the peak of the density function occurring at \( V/2 \). (c) The successive repair times are i.i.d. as a random variable \( R \) uniformly distributed on the interval \([\text{min}, \text{max}]\). Define the state of the system at time \( t \) as 

\[
X(t) = (Y_1(t), Y_2(t), \ldots, Y_8(t), M(t), I(t)),
\]

where \( Y_i(t) = 1 \) iff truck \( i \) has arrived at the loading dock in the interval \([0, t]\) and \( Y_i(t) = 0 \) otherwise, \( M(t) \) is the number of items that arrive on the loading dock in the interval \([0, t]\), and \( I(t) = 1 \) iff the conveyor is broken at time \( t \) and \( I(t) = 0 \) otherwise.

   a) Specify \( \{X(t): t \geq 0\} \) as a GSMP with event set \( E = \{e_1, \ldots, e_{11}\} \), where \( e_i = \text{“arrival of truck } i \text{ at loading dock”} \) for \( 1 \leq i \leq 8 \), \( e_9 = \text{“arrival of item at loading dock”} \), \( e_{10} = \text{“breakdown of conveyor”} \), and \( e_{11} = \text{“completion of conveyor repair”} \). Give your specification generically in terms of \( Q, A_i, L, R \).

   b) In terms of the process \( \{X(t): t \geq 0\} \), precisely specify the three performance measures given in Problem 2(b) of Assignment #2, along with the expected average number of trucks waiting at the loading dock during the interval \([0, T]\), where \( T \) is the time at which the last truck departs from the dock. [Your answers will be very similar to the corresponding answers in Assignment #2.]

   c) Now assume that \( Q = 20 \) minutes, \( 1/\lambda = 30 \) minutes, \( V = 1 \) hour, \( \text{min} = 30 \) minutes, and \( \text{max} = 45 \) minutes. Using the general algorithm for simulating GSMPs given in class, estimate the four performance measures from part (b) to within \( \pm 1\% \) with probability 99%.

2. For each of the following distributions, derive formulas for the MLE’s of the indicated parameters, based on i.i.d. observations \( X_1, X_2, \ldots, X_n \) from the distributions in question. (i) Uniform[0,b], MLE for \( b \), (ii) Uniform[a,b], joint MLEs for \( a \) and \( b \), (iii) Normal(\( \mu, \sigma^2 \)), joint MLEs for \( \mu \) and \( \sigma \).

3. (Quasi-empirical distributions) Suppose that we have \( n \) real-valued nonnegative observations \( X_1, X_2, \ldots, X_n \) from an unknown distribution \( F \) with \( F(0)=0 \), ordered so that \( X_1 < X_2 < \cdots < X_n \). We
can model the the distribution $F$ by using the empirical cdf: $\hat{F}_n(x) = \frac{i}{n}$ for $X_i \leq x < X_{i+1}$ and $i = 0, 1, \ldots, n$. (Take $X_0 = 0$ and $X_{n+1} = \infty$.) The problem with this approach is that the only values that will ever be generated from $\hat{F}_n$ are $X_1, X_2, \ldots, X_n$. One solution to this problem is to fit a piecewise-linear cdf to the first $n - k$ observations and fit an exponential tail to the remaining $k$ observations. The piecewise linear smoothing ensures that we generate a continuous range of values, and fitting the tail allows for the generation of values larger than those observed. The fitted cdf is

$$\tilde{F}_n(t) = \begin{cases} 
\frac{i}{n} + \frac{(t - X_i)}{[n(X_{i+1} - X_i)]} & \text{for } X_i \leq t \leq X_{i+1}, \quad i = 0, 1, \ldots, n - k - 1, \\
1 - c \exp(-\theta(t - X_{n-k})) & \text{for } t > X_{n-k}
\end{cases}$$

where $c$ is chosen to make the cdf continuous and $\theta$ is chosen to make the expected value of the fitted distribution equal to the sample mean of $X_1, X_2, \ldots, X_n$. The fitted cdf might look as follows for $n - k = 2$:

![Graph showing a piecewise-linear cdf with exponential tail]

a) Show that $c = k/n$ and give a formula for $\theta$ in terms of $n, k,$ and $X_1, X_2, \ldots, X_n$. You may use the fact that, with $\tilde{f}_n$ denoting the pdf corresponding to $\tilde{F}_n$,

$$\int_0^\infty \tilde{f}_n(t) \, dt = \sum_{i=0}^{n-k-1} \int_{X_i}^{X_{i+1}} \tilde{f}_n(t) \, dt + \frac{k}{n} \int_0^\infty \theta e^{-\theta t} \, dt$$

$$= \sum_{i=0}^{n-k-1} \frac{t}{n(X_{i+1} - X_i)} \int_{X_i}^{X_{i+1}} \tilde{f}_n(t) \, dt + \frac{k}{n} \int_0^\infty \theta e^{-\theta t} \, dt$$

$$= \sum_{i=0}^{n-k-1} \frac{X_{i+1} - X_i}{2n} + \frac{k}{n} \int_0^\infty \theta e^{-\theta t} \, dt + \frac{k}{n} \int_0^\infty \theta e^{-\theta t} \, dt$$

b) Give an algorithm for generating samples from $\tilde{F}_n$.

4. Consider the following density functions (with $0 \leq a < b < c$):

$$f_{Y_1}(x) = \begin{cases} 
\frac{2(x-a)}{(b-a)^2}, & a \leq x \leq b \\
0, & \text{otherwise}
\end{cases} \quad \text{and} \quad f_{Y_2}(x) = \begin{cases} 
\frac{2(c-x)}{(c-b)^2}, & b \leq x \leq c \\
0, & \text{otherwise}
\end{cases}$$
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 \( f_{\max} \). Give an alternative algorithm for generating a sample from \( f_{\max} \) using inversion. [Hint: \( \max(Z_1, Z_2) \leq x \) if and only if \( Z_1 \leq x \) and \( Z_2 \leq x \).]

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 \( f_{\min} \). Give an alternative algorithm for generating a sample from \( f_{\min} \) using inversion.

5. Provide an intuitive motivation for the definition of MLE’s in the continuous case by going through steps (a) — (c) below. The observed data are i.i.d. observations \( X_1, X_2, \ldots, X_n \) of a random variable \( X \) with density \( f_0 \). Bear in mind that the \( X_i \)'s have already been observed, so are to be regarded as fixed numbers.

a) Let \( \varepsilon \) be a small positive real number, and define the phrase “getting a value of \( X \) near \( X_i \)” to be the event \( \{X_i - \varepsilon < X < X_i + \varepsilon\} \). Use the mean value theorem from calculus to argue that \( P(\text{getting a value of } X \text{ near } X_i) \approx 2\varepsilon f_0(X_i) \) for any \( i = 1, 2, \ldots, n \).

b) Define the phrase “getting a sample of \( n \) i.i.d. values of \( X \) near the observed data” to be the event \( \{\text{getting a value of } X \text{ near } X_1, \text{ getting a value of } X \text{ near } X_2, \ldots, \text{ getting a value of } X \text{ near } X_n\} \). Show that \( P(\text{getting a sample of } n \text{ i.i.d. values of } X \text{ near the observed data}) \approx (2\varepsilon)^n f_0(X_1)f_0(X_2)\cdots f_0(X_n) \), and note that this is proportional to the likelihood function \( L(\theta) \).

c) Argue that the MLE \( \hat{\theta} \) is the value of \( \theta \) that maximizes the approximate probability of getting a sample of \( n \) i.i.d. values of \( X \) near the observed data, and in this sense “best explains” the data that were actually observed.