Advanced Probability Modeling and Techniques
MATH 4960/6960 – Fall 2007
Homework 4

Due Friday, December 7 at 2 PM
Absolute Late deadline: Tuesday, December 11 at 5 PM

This homework has 225 points for MATP 6960 (285 points for MATP 4960) available but, as always, homeworks are graded out of 100 points. Full credit will generally be awarded for a solution only if it is both correctly and efficiently presented using the techniques covered in the lecture and readings, and if the reasoning is properly explained. If you used software or simulations in solving a problem, be sure to include your code, simulation results, and/or worksheets documenting your work. If you score more than 100 points, the extra points do count toward your homework total.

1 MATP 4960 Problems

1.1 Split Particle Correlation (15 points)

Consider a particle of mass M which splits into two subparticles whose mass still adds up to M. Suppose that the probability distribution for the amount of mass in the lighter subparticle is absolutely continuous with probability density p_L; the situation of splitting into two subparticles of exactly equal mass can just be considered as a limiting case of zero probability and does not require special handling. Calculate the correlation coefficient between the masses of the lighter and heavier of the two subparticles.

1.2 Waiting for Non-Exponential Time (10 points)

Suppose the time T to wait for a chemical reaction is given by the following gamma distribution

$$p_T(t) = \begin{cases} t\tau^{-2}\mathrm{e}^{-t/\tau} & \text{for } t > 0, \\ 0 & \text{for } t < 0. \end{cases}$$

Compute the conditional probability distribution and expectation for the random variable T, conditioned on the event that $T > a$ where a is a nonnegative constant.

1.3 Bad Cars (20 points)

Suppose that cars are produced in lots of 100, and that in each lot, each car has a certain probability P of being defective, independently of the defect status of other cars. The defect probability P is constant within a given lot of 100 cars, but fluctuates unpredictably between lots. (These overall variations may be due to different workmanship quality in different shifts, or different calibrations of the machines.) Suppose the defect probability for a lot is modeled in advance to be uniformly distributed between 0 and 0.1. If 20 cars are tested within a lot and found to be free of defect, then how should the probability distribution for the underlying defect probability P in this lot be adjusted in light of this information?

1.4 Poisson Process with Uncertain Rate (15 points)

Consider a Poisson process with uncertain rate $R = 1/\tau$, where τ is the average time between events. Suppose that the rate R is modeled to take one of two values r_1 and r_2 (say corresponding to a normal and abnormal situation, respectively), with probabilities p and $1 - p$, respectively. Calculate the moment generating function, mean, and standard deviation for the number of events observed over a time interval of length t (during which the uncertain rate remains constant at one of the values r_1 or r_2; we just don't know which).

2 MATP 4960/6960 Problems

2.1 Birth Control (15 points)

Consider a population in which each child born has a probability 1/2 of being a boy and a probability 1/2 of being a girl. Suppose that this population, for whatever reason, prefers children of a certain sex, and they plan their families according to the following strategy: Each couple has children until a boy is born, then have no further children. How will this birth control strategy affect the proportion of boys and girls born? For full credit, provide a precise calculation based on the laws of probability theory.

2.2 Coordination Counts (15 points)

Suppose a signal is broadcast from a set of N antennae, all with frequency ν. A receiver adds the signals it receives from each antenna:

$$R(t) = \sum_{j=1}^{N} A_j \cos(2\pi\nu t + \phi_j),$$

where A_j is the amplitudes and $\{\phi_j\}_{j=1}^N$ is the phase of the signal at the receiver which originated from the jth antenna. Define the strength of the received signal as

$$S = \sqrt{\frac{1}{T}\int_t^{t+T} (R(t'))^2 \,\mathrm{d}t'};$$

where $T = 1/\nu$ is the period of the signal. The strength of the signal can be shown to be independent of the choice of t defining the beginning of the integration, and represents a typical amplitude of the total signal received.

a. (**5 points**) Suppose the amplitudes $\{A_j\}_{j=1}^N$ are deterministic and the phases $\{\phi_j\}_{j=1}^N$ are all equal to a common deterministic value ϕ. This corresponds to signals which are all in phase. Calculate the signal strength S in terms of the amplitudes and phases. Also provide an expression for the special case where all the amplitudes $\{A_j\}_{j=1}^N$ are equal to a common value A.

b. (**10 points**) Suppose again the amplitudes $\{A_j\}_{j=1}^N$ are deterministic, but now the phases $\{\phi_j\}_{j=1}^N$ are modeled as independent random variables uniformly distributed on $[0, 2\pi]$. This is the so-called "random phase approximation" which is often used when the phases are not controlled (perhaps because the receiver is mobile). Calculate the mean value of the signal strength S according to this model. Also provide an expression for the special case where all the amplitudes $\{A_j\}_{j=1}^N$ are equal to a common value A.

2.3 Minimal Influence to the Max! (20 points)

Consider a probability model consisting of a finite collection $\{X_j\}_{j=1}^n$ of independent, identically distributed real-valued random variables with arbitrary probability distribution. Suppose the minimum value of these n random variables is observed; how does this affect the probability distribution for the maximum value of these n random variables? For full credit, formalize your answer as explicitly as possible in terms of conditioning with respect to a coarse-grained σ-algebra. For 10 points, it suffices to present your answer in conventional applied form (along with an explained derivation).

2.4 Condition for Caution (30 points)

This problem illustrates concretely why one must be careful with how one understands conditioning with respect to a continuous random variable. Consider two independent random variables X_1 and X_2, each of which have an exponential distribution with mean 1. Define $Y = X_1 - X_2$ and $Z = X_1/X_2$.

a. (**10 points**) Compute the conditional expectation $\mathbb{E}[X_1|Y]$.

b. (**10 points**) Compute the conditional expectation $\mathbb{E}[X_1|Z]$.

c. (**10 points**) Compare $\mathbb{E}[X_1|Y = 0]$ and $\mathbb{E}[X_1|Z = 1]$; these should be different if correctly computed. Yet, the events $Y = 0$ and $Z = 1$ are identical! Provide an explanation for why these conditional expectations differ.

2.5 Gaussianity of Random Sums of Random Variables (40 points)

Consider a sum of random variables:

$$Z = \sum_{j=1}^{N} X_j$$

where the X_j are independent, identically distributed random variables, and N is either a fixed deterministic constant $N = M$ or a random variable with Poisson distribution with mean M. The aim of this question is to investigate how well Z can be modeled by a Gaussian random variable when N is deterministic or random.

We will restrict attention to the case in which the (common) probability distribution of the $\{X_j\}_{j=1}^{N}$ is symmetric (so that $P(X_j > a) = P(X_j < -a)$ for any real a). Then the random variable Z readily can be shown to be symmetric. A standard measure for the deviation of a symmetric probability distribution from a Gaussian shape is through the *kurtosis*

$$K = \frac{\langle\langle Z^4 \rangle\rangle}{\sigma_Z^4},$$

where σ_Z is the standard deviation of Z. (Note some sources, including Grigoriu's *Stochastic Calculus* use a slightly different definition which just differs by an additive constant of 3, as explained in Wikipedia.)

a. (**15 points**) Calculate the kurtoses of Z for both the deterministic and Poisson random models of N when the $\{X_j\}$ are drawn from a common Gaussian distribution with mean zero. Plot the kurtosis in each case as a function of M.

b. (**15 points**) Choose some simple, symmetric, non-Gaussian probability distribution for the $\{X_j\}_{j=1}^N$, and then calculate the kurtoses of Z for both the deterministic and Poisson random models of N. Plot the kurtosis in each case as a function of M.

c. (**10 points**) Discuss your above results.

Such statistical studies have relevance in fields such as finance, where one wishes to know whether the statistical models can be reasonably modeled by Gaussian random variables. Asset prices have both Gaussian and non-Gaussian features, depending on how one collects the statistics. Some analysts consider price fluctuations over a fixed number of transactions while others consider price fluctuations over a fixed time period. The above results can be connected to such data analysis if we think of N as the number of transactions and X_j as the price change due to the jth transaction.

2.6 Just Where Do I Think You're Going? (30 points)

One can imagine various scenarios in which an animal or juvenile is released into the wilderness or neighborhood, and is observed to return at a later time. The handler or parent/guardian was not observing where the organism went while it was away, but only knows the time at which it returned. She may wonder, given this information, a few things, such as:

- How far away did the wandering organism go during its journey?
- Did the wandering organism visit a special place (mating grounds, friend with bad influence, etc.) while it was away?

As a crude model for this situation, we consider a symmetric random walker on the one-dimensional integers $\mathbb{Z}$, with location after n epochs given by $S_0 = 0$ and $S_n = \sum_{j=1}^n X_j$ for $n \geq 1$, where X_j are independent, identically distributed random variables with probability 1/2 to take each of the values ± 1. This would be best suited for a rather dazed juvenile or an animal with no tracking or orienteering skills whatsoever.

We now pose the following model-based versions of the above questions:

a. (**10 points**) Compute the probability distribution for the maximum distance from the origin which the random walker traveled between the starting time $n = 0$ and its first return, given the information about the epoch of the first return to the origin.

b. (**10 points**) Calculate the probability that the random walker visited the site r (an arbitrary nonzero integer) during its journey from the starting point until its first return, given the information about the epoch of the first return to the origin.

c. (**10 points**) You may have answered the above questions in terms of conventional conditional probability. If you have not already done so, express your conditional probabilities, as explicitly as possible, in terms of random variables measurable with respect to a coarse-grained σ-algebra in the original probability space describing the random walk. Be sure to identify and describe this coarse-grained σ-algebra as explicitly as possible.

2.7 Polymer Statistics (20 points)

This is a simplified version of a calculation for the end-to-end distance of a polymer. Suppose the polymer is represented as an open *two-dimensional* chain consisting of N beads, with bead j connected to bead $j + 1$ by a rigid bond of length a for $1 \leq j \leq N - 1$. Molecular chemistry puts strong constraints on the angles between bonds, so let us say that successive bonds must make an angle of α with each other. (α is generally an obtuse angle). Suppose that each polymer conformation subject to these geometric constraints is equally likely (which neglects long-range interactions). Let L_N be the end-to-end distance of the polymer, meaning the distance between atom 1 and atom N. Calculate $\langle L_N^2 \rangle$ explicitly.

Many reasonable attempts will probably get you stuck on this problem, so I offer some hints on how to make progress. If you find another way to calculate the answer, that's fine; you don't have to do it this way.

- Define the orientation of each bond by the angle θ_j it makes relative to the x axis.
- Express L_N in terms of $\{\cos\theta_j, \sin\theta_j\}_{j=1}^{N-1}$.
- Next, carefully express $\langle \cos\theta_j \cos\theta_k \rangle$ and $\langle \sin\theta_j \sin\theta_k \rangle$ for $j < k$ in terms of $\langle \cos^2\theta_j \rangle$ and $\langle \sin^2\theta_j \rangle$. I don't know how to calculate explicit expressions for these latter two quantities, but you don't need to!

2.8 Raindrops Keep Getting in My Way (25 points)

Consider the following model for calculating the attentuation of light as it passes through a cloud of water droplets or various other particulate suspensions in the atmosphere or outer space. We are completely neglecting refraction and reflection, and are grossly simplifying the actual physics.

Suppose we have a suspension of spherical "obstacles," each of radius ρ, which allow light to pass straight through, but diminish the amplitude of the light. We model the suspension by generating a Poisson point process with intensity λ to define the centers of the obstacles. Technically, the obstacles should not overlap but we won't worry about that. (If the suspension is sufficiently dilute, this simplification won't make too much of a difference). We now imagine a light ray starting from some arbitrary point and traveling along a straight line for a distance L through this suspension (which we can think of as the geometric distance between the point of entry and exit of light from the suspension).

We define an "optical depth" for the suspension as the total distance D the light ray travels through an obstacle. Your calculation should simplify by taking the convention in our model that when a light ray is simultaneously passing through k overlapping obstacles, then that distance traveled will count k times toward the optical depth D. (So the optical depth would then simply be the added effects of the influence of each obstacle on the light ray, even if the obstacles overlap.)

Suppose the attenuation of the light is given by $A = \mathrm{e}^{-D}$. Calculate the probability distribution for the attentuation factor A.

2.9 Simulated Finishing (30 points)

Recall the example from class in which we had a server handling two requests, arriving at times 0 and a, with the requests requiring independent random service times X_1 and X_2, given by exponential distributions with mean τ. The server could only handle one request at a time, and we computed the CDF for the time T to complete the second request as

$$F_T(t) = \begin{cases} 1 - \mathrm{e}^{-(t-a)/\tau} - \frac{(t-a)}{\tau}\mathrm{e}^{-t/\tau}, & \text{for } t > a, \\ 0 & \text{for } t \leq a. \end{cases} \tag{1}$$

a. (**10 points**) Using knowledge of the model underlying the CDF, devise an efficient algorithm to simulate numerically the random variable T. Show through numerical plots that your simulated CDF agrees well with the theoretical one (1) (and/or point out an algebra mistake I made in the lecture notes!)

b. (**20 points**) On the other hand, suppose the CDF (1) appears out of the blue and you would like to simulate it numerically. Without using special knowledge about the relation of the CDF to the simple underlying probability model, devise an efficient numerical algorithm to simulate a random variable with the CDF (1). Show through numerical plots that your simulated CDF agrees well with the theoretical formula (1).