Fall 2019 18.600 Final Exam Solutions

1. (10 points) A school has 60 students (30 boys and 30 girls) and these students are randomly divided into Classroom 1 and Classroom 2, with thirty students assigned to each class. Assume that each of the $\binom{60}{30}$ possible divisions is equally likely. A parent asks the principal, "Does Classroom 1 have at least 29 girls?" and is told that (surprisingly) the answer is yes. The following conversation ensues:

> **Parent zero:** Given what we have just learned, I wonder how likely it is that *all 30 students* in Classroom 1 are girls.
>
> **Parent one:** Well, any given child has a 1/2 chance to be a girl. So I'd say that if 29 are girls, the chance that the 30th is also a girl is 1/2.
>
> **Parent two:** No, that's silly. If we know the class has 29 girls, then there is only one girl left among the remaining 31 students, so the conditional probability is 1/31.
>
> **Parent three:** You are both formulating this the wrong way. The actual conditional probability is even lower than that.

You are summoned to resolve the dispute.

(a) Let A be the event that all 30 students in Classroom 1 are girls. Let B be the event that Classroom 1 has at least 29 girls. Compute the quantities $P(A)$ and $P(B)$. **ANSWER:** Only one of the $\binom{60}{30}$ divisions assigns all girls to Classroom 1, so $P(A) = 1/\binom{60}{30}$. To get exactly 29 girls, one has 30 choices for the girl not present, and 30 choices for the boy who is present—so the probability of exactly 29 girls is $900/\binom{60}{30}$ and $P(B) = 901/\binom{60}{30}$.

(b) Now compute the conditional probability $P(A|B)$. Based on this calculation, which (if any) of parents one, two and three is correct? **ANSWER:**

$$P(A|B) = P(AB)/P(B) = P(A)/P(B) = 1/901$$

so parent three is correct. **REMARK:** You can also consider a variant where there are 69 children and only 5 girls, and Classroom 1 has 5 students. Then the question "What is the probability Classroom 1 has 5 girls given it has at least 4?" is the same as the question "What is the probability you match all five regular Powerball balls given that you match at least four?" Again, it is somewhat surprising how much less likely 5 is than 4.

(c) Let G be the number of girls in Classroom 1 and B the number of boys in Classroom 1. Compute the expectation $E[GB]$. **ANSWER:** Let G_i (resp. B_i) be indicator function for event that ith girl (resp. boy) is in Classroom 1. Then

$$E[GB] = E\Big[\big(\sum_{i=1}^{30} B_i\big)\big(\sum_{i=1}^{30} G_i\big)\Big] = \sum_{i=1}^{30}\sum_{j=1}^{30} E[G_iB_j] = 900P(G_1 = B_1 = 1) = 900 \cdot \frac{1}{2} \cdot \frac{29}{59}$$

2. (10 points) A three-year-old child is vying for acceptance at two selective preschools. At each school, acceptance is determined by an entrance exam that measures important preschool skills but is not very reliable. The child's scores on the two exams take the form $S_1 = A + 2B_1$ and $S_2 = A + 2B_2$ where A, B_1, and B_2 are independent normal random variables, each with mean zero and variance one. Informally, A is the student's "entrance exam ability" while B_1 and B_2 are independent noise terms (encoding chance fluctuations).

(a) Compute the expectation $E[S_1 S_2]$. **ANSWER:**

$$E[S_1 S_2] = E[A^2 + 2B_1 A + 2B_2 A + 4B_1 B_2] = E[A^2] + E[2B_1 A] + E[2B_2 A] + E[4B_1 B_2].$$

Latter three terms are zero (by independence) so answer is $E[A^2] = 1$.

(b) Compute $\mathrm{Var}(S_1)$ and $\mathrm{Var}(S_2)$ and the correlation coefficient $\rho(S_1, S_2)$. **ANSWER:** By addivity of variance (for independent random variables) we have

$$\mathrm{Var}(S_1) = \mathrm{Var}(A) + \mathrm{Var}(2B_1) = \mathrm{Var}(A) + 4\mathrm{Var}(B_1) = 5.$$

Similarly $\mathrm{Var}(B_2) = 5$, and $\rho(S_1, S_2) = \frac{\mathrm{Cov}(S_1,S_2)}{\sqrt{\mathrm{Var}(S_1)\mathrm{Var}(S_2)}} = 1/5$.

(c) Compute the conditional expectation $E[S_2|S_1]$ in terms of S_1. That is, express the random variable $E[S_2|S_1]$ as a function of the random variable S_1. (Hint: if it helps, you can argue that $2B_1$ agrees in law with $\sum_{i=1}^4 Y_i$ where the Y_i are independent normal random variables, each with mean zero and variance one.) **ANSWER:** By independence, $E[A + 2B_2|A + 2B_1] = E[A|A + 2B_1]$. Following hint and writing $A = Y_0$, we can write this as $E[Y_0|S]$ where $S = \sum_{i=0}^4 Y_i$. We know that $S = E[S|S] = E[\sum_{i=0}^4 Y_i|S] = \sum_{i=0}^4 E[Y_i|S] = \sum_{i=0}^4 E[Y_i|\sum_{i=0}^4 Y_i]$ and since the Y_i are i.i.d. the latter five terms are all the same, hence equal to $S/5$. So answer is $S/5$.

3. Jill is an enthusiastic guitarist. Every song she plays has exactly 16 measures. During each measure, independently of all others, she randomly plays of one the five chords she knows:

1. **A major** with probability 1/2
2. **F# minor** with probability 1/8
3. **D major** with probability 1/8
4. **B minor** with probability 1/8
5. **E major** with probability 1/8

Let $X = (X_1, X_2, \dots, X_{16})$ be the sequence of chords associated to one of Jill's songs.

(a) Compute the entropy $H(X_1)$, i.e., the entropy involved in choosing the first chord. **ANSWER:** $\frac{1}{2}(-\log(\frac{1}{2})) + 4 \cdot \frac{1}{8}(-\log(\frac{1}{8})) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 3 = 2$.

(b) Compute the entropy $H(X)$, i.e., the entropy involved in choosing an entire song. **ANSWER:** Since all 16 entries are independent, answer is $2 \cdot 16 = 32$.

(c) Describe a strategy for asking yes-or-no questions to determine X, such that the expected number of questions asked is as small as possible. What is the expected number of questions in this case? **ANSWER:** Ask "Is first chord A major?" If no, ask "Is first chord either F# minor or D major?" and "Is first chord either F# minor or B minor?" This determines first chord in 1 question half the time and 3 questions half the time. Repeat for all 16 chords. Expected number of question is $2 \cdot 16 = H(X)$ which is the smallest possible.

(d) Let $S = \{i : X_i = \text{A major}\}$. In other words, S is the collection of times at which an A major chord is played. Compute $H(S)$ and $H(S,X)$ and $H_S(X)$. **ANSWER:** $H(S) = 16$ (since S encodes 16 i.i.d. fair coin tosses) and $H(S,X) = H(X) = 32$ (since X determines S). By a general identity, $H(S,X) = H(S) + H_S(X)$ so $H_S(X) = 16$.

4. A certain small technical university has only five majors: Humanities, Science, Engineering, Business, and Math. During each given week, each student at this university is assigned to exactly one of the five majors. The students at this university change their majors frequently, but they tend to stay in Math a little longer than they do in the other majors. Here is how that works:

> If a student is majoring in Math one week, then the next week she stays in Math with probability 1/2 and transitions to each of the other majors with probability 1/8.
>
> If a student is majoring in any major other than Math one week, then the next week she stays in that major with probability 1/5 and also transitions to each of the other majors with probability 1/5.

Now answer the following:

(a) Represent the major transition process as a Markov chain and write out the five-by-five transition matrix. **ANSWER:** If first row and column correspond to Math, then the transition matrix is

$$\begin{pmatrix} 1/2 & 1/8 & 1/8 & 1/8 & 1/8 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix}$$

(b) Suppose a student is majoring in Math during the first week; compute the probability that he or she will be majoring in Business two weeks later. **ANSWER:** The probability of transition to Math followed by transition to Business is $\frac{1}{2} \cdot \frac{1}{8}$. The probability of transition to *some* other state followed by transition to Business is $\frac{1}{2} \cdot \frac{1}{5}$. So overall answer is $\frac{1}{2} \cdot (\frac{1}{5} + \frac{1}{8}) = \frac{1}{2} \cdot \frac{13}{40} = \frac{13}{80}$.

(c) Over the long haul, what fraction of the time does a student spend in each of the five majors? **ANSWER:** We can solve

$$\begin{pmatrix} \pi_1 & \pi_2 & \pi_3 & \pi_4 & \pi_5 \end{pmatrix} \cdot \begin{pmatrix} 1/2 & 1/8 & 1/8 & 1/8 & 1/8 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 & \pi_4 & \pi_5 \end{pmatrix}$$

We know that $\pi_2 = \pi_3 = \pi_4 = \pi_5 = c$ for some constant c and $\pi_1 = (1 - 4c)$ so we can write this as

$$\begin{pmatrix} 1-4c & c & c & c & c \end{pmatrix} \cdot \begin{pmatrix} 1/2 & 1/8 & 1/8 & 1/8 & 1/8 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{pmatrix} = \begin{pmatrix} 1-4c & c & c & c & c \end{pmatrix}$$

Using just first column for above product, we get $\frac{1}{2}(1-4c)+\frac{4}{5}c=(1-4c)$ and solving gives $c=5/28$. So over long haul a student spends a $5/28$ fraction of time in each major other than Math and a $(1-4c)=8/28=2/7$ fraction in Math.

5. (10 points) Let X be an exponential random variable with parameter $\lambda=1$. For each real number K write $C(K)=E[\max\{X-K,0\}]$.

(a) Compute $C(K)$ as a function of K for $K\geq 0$. **ANSWER:**

$$\int_0^\infty e^{-x}\max\{x-K,0\}dx=\int_K^\infty e^{-x}(x-K)dx=\int_0^\infty e^{-(x+K)}xdx=e^{-K}\int_0^\infty e^{-x}xdx=e^{-K}$$

(b) Compute the derivatives C' and C'' on $[0,\infty)$. **ANSWER:** $C(x)=e^{-x}$ so $C'(x)=-e^{-x}$ and $C''(K)=e^{-x}$. Alternatively, if one simply remembers that the second derivative of the call function is the density function of X, and the first derivative is F_X-1, then one can do this part without computation (and use it to derive part (a)).

(c) Compute the expectation $E[X^3+3X^2+3X+1]$. **ANSWER:** Recall that $E[X^k]=k!$ so additivity of expectation gives the answer $3!+3\cdot 2!+3\cdot 1!+1=6+6+3+1=16$.

6. (10 points) In a certain political party, there are 100 voters actively involved in the process of selecting a nominee. The are 23 candidates, which we number from 1 to 23. At any given time, each voter supports exactly one of these candidates. At the initial time $t=0$

1. Candidates 1-15 are "lower tier candidates." Each has the support of exactly 2 voters.
2. Candidate 16-20 are "mid-tier candidates." Each has the support of exactly 3 voters.
3. Candidate 21 has the support of 11 voters.
4. Candidate 22 has the support of 21 voters
5. Candidate 23 has the support of 23 voters.

During each unit interval of time (after time $t=0$) two voters are chosen (uniformly at random from the set of all $\binom{100}{2}$ possible pairs) to have a discussion with each other. If the two voters support the same candidate before the discussion, then the discussion changes nothing; but if they support *different* candidates then *one* of the two voters (chosen by a fair coin toss) switches support to the *other* voter's candidate. This continues until the time T at which all of the voters support the same candidate, at which point that candidate is declared the nominee. (This happens eventually with probability 1, but you don't have to prove that.)

(a) Let $A_i(t)$ be the number of voters supporting candidate i at time t. Is it true that for each $i\in\{1,2,\ldots,23\}$, the quantity $A_i(t)$ is a martingale? Explain why or why not. **ANSWER:** Yes, it is a martingale. The only way $A_i(t)$ can change is if two voters line up where one supports the ith candidate and one doesn't. When this happens, $A_i(t)$ goes up by one with probability $1/2$ and down by one with probability $1/2$, so the expected change (conditioned on everything known up to stage t) is 0.

(b) Compute the probability that for some t with $0 < t < T$ we have $A_{23}(t) = 1$ (so that the initially-leading candidate has lost the support of all but one voter). **ANSWER:** Since initial value of 23 is a $(100 - 23)/99 = 7/9$ fraction of the way from 100 toward 1, the answer (by shortcut from lecture) is 7/9.

(c) Let N be the number of candidates who at *some* point before time T only have the support of a single voter. That is, N is the number of i values for which there is some t such that $0 < t < T$ and $A_i(t) = 1$. Compute the expectation $E[N]$. (Hint: you can use the fact that $\sum_{i=1}^{23} A_i(0) = 100$ and hence $\sum_{i=1}^{23}\big(100 - A_i(0)\big) = 2200$.) **ANSWER:** The probability for the ith candidate is $(100 - A_i(0))/99$. Adding these up to get the expectation, we have $2200/99 = 200/9$.

(d) Compute the probability that the candidate who wins the election is somebody who at some time $t > 0$ only had the support of a single voter. (This scenario is called an *epic comeback*.) **ANSWER:** The ith candidate has a $\frac{100-A_i(0)}{99} \cdot \frac{1}{100}$ of being an epic comeback winner. Since only one candidate do this, we can just add these probabilities to get $\frac{2200}{9900} = \frac{2}{9}$. That might seem surprisingly high: epic comebacks are not so unusual after all. (This is similar to the islander problem from the spring 2019 exam.) Note that the answer does not actually depend on the initial support levels for the candidates, as long as they are all at least one. You could also formulate the problem in terms of a betting market on which prices were known to be martingales changing by increments of ± 1.

7. (10 points) Suppose that the pair (X, Y) is uniformly distributed on the circle $\{(x, y) : x^2 + y^2 \leq 1\}$.

(a) Compute the joint probability density $f_{X,Y}(x, y)$ and the marginal density function $f_X(x)$. **ANSWER:** $f_{X,Y}(x, y)$ is $1/\pi$ in the unit circle, 0 elsewhere and $f_X(x) = \frac{1}{\pi} 2\sqrt{1 - x^2}$.

(b) Compute the probability $P(0 < X < Y)$. **ANSWER:** The intersection of the disk, the plane where $x > 0$, and the plane where $x < y$ together forms a wedge of angle $\pi/4$, which is one eighth of total disk, so answer is 1/8.

(c) Write $Z = X^2 + Y^2$ and work out the density function f_Z. **ANSWER:** First work out

$$F_Z(a) = P(Z \leq a) = P(X^2 + Y^2 < a) = P(\sqrt{X^2 + Y^2} < \sqrt{a}) = \pi(\sqrt{a})^2/\pi = a.$$

Differentiating we find that f_Z is equal to 1 on $[0, 1]$ so that Z is a uniform random variable.

8. (10 points) Sally the Spammer sends millions of emails every day encouraging the recipients (using various rationales) to grant her unrestricted access to their bank accounts. The times $X_1, X_2, X_3, \ldots$ (measured in years from some initial time) at which Sally is *granted* access to such a bank account form a Poisson point process with parameter $\lambda = 3$. So on average three people per year give Sally access to their bank accounts.

(a) Write $Y = X_3$ and compute the density function f_Y. **ANSWER:** The nth term in a Poisson point process with parameter λ (which is a sum of n independent exponentials with parameter λ) is a Gamma random variable with parameters λ and n. In this case X_3 is Gamma with parameters $n = 3$ and $\lambda = 3$ so answer is $f_Y(y) = 3 \cdot (3y)^2 e^{-(3y)}/2$.

(b) What is the probability that Sally is granted bank acount access at least three times during her first year of operation? **ANSWER:** Number of accesses during 1 year is Poisson with parameter $1 \cdot \lambda = 3$. So answer is $1 - \sum_{k=0}^{2} e^{-3}3^k/k!$

(c) Write $Y_0 = 0$ and $Y_k = X_k - \frac{k}{3}$. Is the sequence $Y_0, Y_1, \dots$ a martingale? Why or why not? **ANSWER:** Yes. The increments $I_k = X_k - X_{k-1}$ are independent exponential random variables with expectation $1/3$. So the increments $Y_k - Y_{k-1}$ are independent random variables with expectation 0. Alternatively, write $E[Y_{k+1}|\mathcal{F}_k] = E[Y_k + I_{k+1} - 1/3|\mathcal{F}_k] = Y_k$.

9. (10 points) There are 3 people, each of whom has 4 hats. All 12 hats are tossed into a bin and random divided evenly among the 3 people (so each person gets 4 hats back, with all ways of doing this being equally likely). Let A_i be the event that the ith person gets *all four* of his or her own hats back. Let N be the number of people who get *all four* of their own hats back.

(a) Compute the quantities $a = P(A_1)$ and $b = P(A_1A_2)$. **ANSWER:** There are $\binom{12}{4}$ possible collections of hats for person one, and only of them is right. So $a = P(A_1) = 1/\binom{12}{4}$. Similarly $b = P(A_1A_2) = (1/\binom{12}{4})(1/\binom{8}{4})$.

(b) Compute $E[N]$ and $\mathrm{Var}(N)$. (You can use the a and b from the previous part in your answer, if that helps.) **ANSWER:** $E[N] = \sum_{i=1}^{3} P(A_i) = 3a$. $E[N^2] = \sum_{i=1}^{3}\sum_{j=1}^{3} P(A_iA_j) = 3a + 6b$. So $\mathrm{Var}(N) = E[N^2] - E[N]^2 = 3a + 6b - 9a^2$.

(c) Compute $P(N > 0)$. (Again, you can use a and b in your answer, if that helps.) **ANSWER:** Note that $P(A_1A_2A_3) = P(A_1A_2)$. Then inclusion exclusion gives $P(N > 0) = 3a - 3b + b = 3a - 2b$.

10. (10 points) Suppose that $X_1, X_2, \dots$ are i.i.d. random variables, each equal to 0 with probability $1/8$, 1 with probability $3/4$ and 2 with probabiliy $1/8$. Write $S_n = \sum_{i=1}^{n} X_i$ and $A_n = S_n/n$.

(a) Compute the moment generating functions $M_{S_{50}}(t)$ and $M_{A_{50}}(t)$. **ANSWER:** $M_{X_1}(t) = E[e^{tX_1}] = \frac{1}{8} + \frac{3}{4}e^t + \frac{1}{8}e^{2t}$ and $M_{S_{50}}(t) = (\frac{1}{8} + \frac{3}{4}e^t\frac{1}{8}e^{2t})^{50}$ and $M_{A_{50}}(t) = (\frac{1}{8} + \frac{3}{4}e^{t/50}\frac{1}{8}e^{2t/50})^{50}$

(b) Compute $E[X_1]$ and $\mathrm{Var}(X_1)$. **ANSWER:** $E[X_1] = 1$ and $\mathrm{Var}[X_1] = 1/4$.

(c) Use the central limit theorem to approximate $P(S_{100} > 110)$. You may use the function $\Phi(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}dx$ in your answer. **ANSWER:** We have $E[S_{100}] = 100$ and $\mathrm{Var}(S_{100}) = 25$ and $\mathrm{SD}(S_{100}) = 5$. And chance to be at least two standard deviations above the mean is about $1 - \Phi(2) = \Phi(-2)$.

(d) Compute the correlation coefficient $\rho(S_{50}, S_{200})$. **ANSWER:** Using bilinearity of covariance we have $\mathrm{Cov}(S_{50}, S_{200}) = \mathrm{Cov}(S_{50}, S_{200} - S_{50}) + \mathrm{Cov}(S_{50}, S_{50})$. First term is zero (since it is a covariance of two independent random variables) and second term is $\mathrm{Var}(S_{50}) = 50/4$. So answer is $\frac{50/4}{\sqrt{(50/4)(200/4)}} = 1/2$.