STA402L: HOMEWORK 3

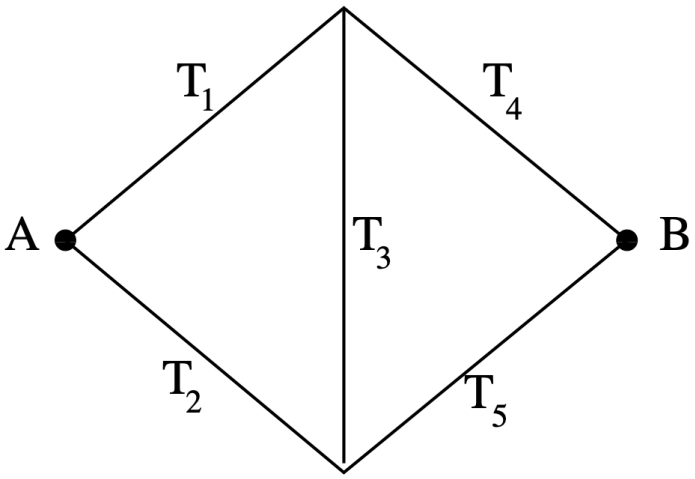
DUE: 11:59 PM ON FRIDAY, FEBRUARY 13

Instructions. Solutions must be submitted to Gradescope as a single PDF. Programming exercises must be completed in R, should be clearly presented, and include all R code. Lab questions are restated here for convenience, but you should refer to the lab itself for details.

Total points. Book exercises: 24; Lab exercises 20; Overall: 44.

BOOK EXERCISES

B1. (4 points) This problem is about the network pictured below.



Here, $T_1, \ldots, T_5$ are independent random variables representing traversal times along the five edges of the network. The traversal times are uniformly distributed as follows:

$$T_1 \sim \text{Unif}(0,1),\ T_2 \sim \text{Unif}(0,2),\ T_3 \sim \text{Unif}(0,3),\ T_4 \sim \text{Unif}(0,1),\ T_5 \sim \text{Unif}(0,2).$$

Let X denote the minimum time it takes to get from A to B, i.e.,

$$X = \min\{T_1 + T_4, T_1 + T_3 + T_5, T_2 + T_5, T_2 + T_3 + T_4\}.$$

(a) (2 points) Write down an integral expression for $\mathbb{E}[X]$.

(b) (2 points) Using $S = 10{,}000$ samples, estimate $\mathbb{E}(X)$ and $\text{Var}(X)$ via Monte Carlo.

B2. (12 points) *When independence fails: Random walks, the Law of Large Numbers, and the Central Limit Theorem.* Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(0,1)$ and define the *random walk* $(S_0, \ldots, S_n)$ where

$$S_0 = 0 \quad \text{and, for } k > 0, \quad S_k = \sum_{i=1}^{k} X_i.$$

(a) (1 point) Suppose X_k represents the profit your business makes in month k, measured in standardized units, i.e., mean zero, variance 1. In this context, what does S_k represent?

(b) (1 point) Set $n = 10{,}000$ and plot 5 trajectories of S_k as a function of $k \in \{0, \ldots, n\}$.

(c) (1 point) Compute $\text{Var}(S_k)$.

(d) (2 points) Compute $\text{Cov}(S_j, S_k)$ for arbitrary $j \neq k$. Are S_j and S_k independent?

(e) (3 points) Define the sample mean

$$\bar{S}_n = \frac{1}{n} \sum_{k=1}^{n} S_k.$$

Compute $\text{Var}(\bar{S}_n)$. Explain why the Law of Large Numbers does not hold for $\bar{S}_n$.

Hint. The following identities may be useful:

- For any sequence of random variables $Z_1, \ldots, Z_n$,

$$\text{Var}(\sum_{i=1}^{n} Z_i) = \sum_{i=1}^{n} \text{Var}(Z_i) + \sum_{i \neq j} \text{Cov}(Z_i, Z_j).$$

- $\sum_{1 \leq j < k \leq n} \min(j,k) = \sum_{j=1}^{n-1} j(n-j) = \frac{n(n-1)(n+1)}{6}$.

(f) (2 points) Set $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ and let $n = 1000$. Plot histograms of $\sqrt{n}\bar{X}_n$ and $\sqrt{n}\bar{S}_n$ on separate subplots by generating $t = 1000$ distinct trajectories $X_1^{(t)}, \ldots, X_n^{(t)}$.

(g) (2 points) How do your results in Part (e) relate to the Law of Large Numbers and the Central Limit Theorem?

B3. (2 points) Hoff 4.1.

B4. (6 points) Hoff 4.2.

(a) (2 points)

(b) (2 points)

(c) (2 points)

Lab exercises

L1. (2 points) Plot a histogram of the death counts.

L2. (2 points) Use the function `mcmc_areas()` to plot the smoothed posterior distribution for λ with a 90% Highest Posterior Density region. Changing the `prob` parameter in `mcmc_areas()` will change the amount of probability mass that is highlighted. The highlighted region will begin in regions of highest density, and will move towards regions of lower density as `prob` gets larger. The line in the plot represents the posterior mean, and the shaded area represents 90% of the posterior density. How does the posterior mean for λ compare to the sample mean?

L3. (2 points) Generate posterior predictive samples using the posterior values of λ and store the result in a `length(lambda_draws)` by `n` matrix called `y_rep`.

L4. (2 points) Based on these PPCs, does this model appear to be a good fit for the data?

L5. (2 points) Using the code provided for the simple Poisson model, simulate draws from the posterior density of λ with the Poisson hurdle model. The Stan file you'll need to use is called `lab-03-poisson-hurdle.stan`. Store the resulting object in a variable called `fit2`.

L6. (2 points) Use the code given above to produce the same PPC visualizations as before for the new results. Comment on how this second model compares to both the observed data and to the simple Poisson model.

L7. (2 points) Which model performs better in terms of prediction?

L8. (2 points) Why are PPCs important?

L9. (2 points) Was the second model a good fit for the data? Why or why not?

L10. (2 points) If someone reported a single LOOCV error to you, would that be useful? Why or why not?