Stat 202C Homework #2 (10 points)

Due: June 10th on CCLE.

Problem 1. Consider the Markov kernel for the five families living in a pacific island, that we studied in HW1.

$$K_0 = \begin{pmatrix} 0.0, & 0.7, & 0.3, & 0.0, & 0.0 \\ 0.2 & 0.0, & 0.6, & 0.0, & 0.2 \\ 0.1, & 0.4, & 0.0, & 0.5, & 0.0 \\ 0.0, & 0.3, & 0.4, & 0.0, & 0.3 \\ 0.0, & 0.0, & 0.3, & 0.7, & 0.0 \end{pmatrix}$$

This transition matrix defines a directed graph $G =< V, E >$ where $V = \{1, 2, 3, 4, 5\}$ is the set of states, and $E = \{e = (x, y) : K(x, y) > 0\}$ is a set of directed edges. You know invariant probability $\pi(x)$ for the five states $x \in \{1, 2, 3, 4, 5\}$; and λ_{slem} in HW1.

Now let's try to verify the bounds of λ_{slem} by the following two concepts that we studied in class – bottleneck and conductance. Since we have only 5 states, we can calculate the two quantities by enumerating all the paths and subsets in a brute-forth way.

1. Which edge $e = (x, y)$ is the bottleneck of G? (you may make a guess based on the graph connectivity first, and then calculate by its definition); and calculate the Bottleneck κ of the graph G. Verify the Poincare inequality:
$$\lambda_{\text{slem}} \leq 1 - \frac{1}{\kappa}.$$
2. Calculate the Conductance h of the graph G. Verify the Cheeger's inequality:
$$1 - 2h \leq \lambda_{\text{slem}} \leq 1 - \frac{h^2}{2}.$$
3. Now, since we know π, we can design the "dream" matrix K^* that converges in one step. Then $\lambda_{\text{slem}} = 0$ for K^*. Rerun your code above to calculate the Conductance h for K^*. Verify the Cheeger's inequalities.

Problem 2. For problem 1 above, we have the invariant probability π and the dream matrix K^*. Now we design a Metropolised Gibbs sampler for π. Each time, it proposes 4 possible states to move, excluding its current state. The proposal probability for each candidate state i is proportional to it probability $\pi(i)$, and then the proposal is accepted by a Metropolis step. Calculate the new transition matrix K_{MGS}.

Check whether K_{MGS} *dominates* K_2 *in the Pushin order*:

$$K_{\text{MGS}}(x, y) \geq K^*(x, y), \forall x \neq y.$$

I.e. the off-diagonal elements of K_{MGS} are no less than that of K^*. Simulate 500 samples $X(1), ..., X(500)$ from K^* and K_{MGS} respectively.

1. Calculate, plot and compare the auto-correlations $\mathrm{Corr}(\tau)$ from the two sequences above for $\tau = 1, 2, 3, 4, ..., 10$. Which transition matrix has lower auto-correlation? [The auto-correlation is the correlation of two variables $X(i)$ and $X(i+\tau)$ with i being a moving index.]

2. Suppose we estimate the expectation $\theta = \sum_i \pi(x)x^2$ using the 500 samples from each Markov chain simulation respectively, which transition matrix yields better estimate?

Problem 3 In the riffle shuffling of cards, we mentioned two bounds: 7 and 11 as the expected number of shuffles to make the 52 cards random. Before proving bounds, it is often a good idea to empirically plot the convergence curve.

Suppose we label the 52 cards as 1,2,...,52 and start with a deck (or state) X_0 which is sorted from 1 to 52. Then we simulate the following riffle shuffling process iteratively from X_{t-1} to X_t.

Simulate 52 independent Bernoulli trials with probability 1/2 to be 0 or 1. Thus we obtain a binary vector 0,1,1,0....0. Suppose there are n zero's and $52 - n$ one's. We take the top n cards from deck X_{t-1} and sequentially put them in the positions of the zero's and the remaining $52 - n$ cards are sequentially put in the positions of one's.

Let's check whether the deck X_t is random as t increases. You may design your own methods to test randomness. Below is a default method if you don't have better ideas.

We always start with a sorted deck X_0 and repeat the shuffling process K times. Thus at each time t we record a population of K decks: $\{X_t^k : k = 1, 2, ..., K\}$.

For each card position $i = 1, 2..., 52$, we calculate the histogram (marginal distribution) of the K cards at position i in the K decks. Denote it by $H_{t,i}$ and normalize it to 1. This histogram has 52 bins, so we may choose $K = 52 \times 10$. Then we compare this 52-bin histogram to a uniform distribution by TV-norm and average them over the 52 positions as a measure of randomness at time t

$$\mathrm{err}(t) = \frac{1}{52} \sum_{i=1}^{52} ||H_{t,i} - \mathrm{uniform}||_{\mathrm{TV}}.$$

Plot $\mathrm{err}(t)$ over time t to verify the convergence steps. Based on your plot, how many times do we really need to shuffle the cards?