Math 4740: Homework 5

Due Friday, March 4 in class.

1. Let (X_n) be an irreducible Markov chain on a finite state space $\mathcal{X}$ with stationary distribution π. In class we proved that if $N_n(x) = \#\{1 \leq i \leq n : X_i = x\}$, then for all $x \in \mathcal{X}$,

$$\lim_{n \to \infty} \frac{N_n(x)}{n} = \pi(x) \quad \text{with probability 1.}$$

The goal of this problem is to prove the "Markov chain ergodic theorem": If $f : \mathcal{X} \to \mathbf{R}$ is a function, then

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} f(X_i) = \sum_{y \in \mathcal{X}} \pi(y) f(y) \quad \text{with probability 1.} \tag{1}$$

(a) For fixed $x \in \mathcal{X}$, let $\mathbf{1}_x$ be the function from $\mathcal{X}$ to $\mathbf{R}$ given by $\mathbf{1}_x(x) = 1$ and $\mathbf{1}_x(y) = 0$ for all $y \neq x$. Prove equation (1) when $f = \mathbf{1}_x$.

(b) Now let f be any function from $\mathcal{X}$ to $\mathbf{R}$. For each x, let $c_x = f(x)$. Explain why $f(y) = \sum_{x \in \mathcal{X}} c_x \mathbf{1}_x(y)$.

(c) Use parts (a) and (b) to prove equation (1) for general $f : \mathcal{X} \to \mathbf{R}$.

2. Consider the random walk on the state space $\{0, 1, \ldots, N\}$ defined as follows. At each time step the random walker moves left with probability $2/3$ and right with probability $1/3$. If the walker is currently at 0 and attempts to move left, it stays at 0; if the walker is currently at N and attempts to move right, it stays at N.

(a) Use the detailed balance equations to find a formula for the stationary distribution of the random walk.

(b) Now consider the random walk on the infinite state space $\{0, 1, 2, \ldots\}$ that moves left with probability $2/3$ and right with probability $1/3$ (unless the walker is currently at 0, in which case it stays put with probability $2/3$). Find a formula for the stationary distribution of this Markov chain, and verify by direct computation that it satisfies $\sum_{i=0}^{\infty} \pi(i) P(i, j) = \pi(j)$.

The remaining problems consider the model of English text as a Markov chain on the state space $\{A, B, C, \ldots, Z, \text{space}\}$. Peter Norvig, head of research at Google, helpfully supplies us with some statistics from the Google Books n-grams raw data set containing 3,563,505,777,820 letters (not including spaces) that make up 743,842,922,321 words. An n-gram is a sequence of n consecutive characters. See:

http://norvig.com/mayzner.html

3. Use Norvig's data to find $\pi(\text{space})$, $\pi(F)$, and $\pi(Z)$. *Hint:* Use the total number of letters and words to deduce $\pi(\text{space})$. Recall the assumptions from class that each word is followed by a space and that $P(\text{space}, \text{space}) = 0$.

4. Use the data to find the transition probabilities $P(F, O)$, $P(O, R)$, $P(\text{space}, F)$, and $P(R, \text{space})$.

5. (a) Write down formulas for $\pi(\alpha)$ and $P(\alpha, \beta)$ in terms of the quantities

$$\begin{aligned} f(\alpha) &= \text{ total number of occurrences of the 1-gram } \alpha \text{ in the data set,} \\ f(\alpha, \beta) &= \text{ total number of occurrences of the 2-gram } \alpha\beta \text{ in the data set.} \end{aligned}$$

(b) If π and P are defined by these formulas, will they satisfy $\pi P = \pi$?

6. This problem explores ways English text is a lot more complicated than a Markov chain.

(a) Write down a 3-gram that is at least 1000 times more likely to occur in the Markov model than in standard English text. Explain how you know this is the case.

(b) According to the Markov model, how many times would you expect to encounter the 5-gram $\{\text{space}, F, O, R, \text{space}\}$ in Google's data set? How many times does it actually occur in the data set?