Math 4740: Homework 5 Solutions

1. (a) When $f = \mathbf{1}_x$, $\sum_{i=1}^n f(X_i) = N_n(x)$ and $\sum_{y\in\mathcal{X}} \pi(y)f(y) = \pi(x)$. Therefore the desired statement reduces to

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

which was shown in class.

(b) Fix $y \in \mathcal{X}$. By the definition of $\mathbf{1}_x$, $\sum_{x\in\mathcal{X}} c_x \mathbf{1}_x(y) = c_y = f(y)$.

(c) First use (b) to expand f as a linear combination of the $\mathbf{1}_x$, then bring the limit inside the linear combination and use (a).

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n f(X_i) &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \sum_{x\in\mathcal{X}} c_x \mathbf{1}_x(X_i) = \sum_{x\in\mathcal{X}} c_x \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbf{1}_x(X_i) \\ &= \sum_{x\in\mathcal{X}} c_x \pi(x) = \sum_{x\in\mathcal{X}} \pi(x) f(x) \quad \text{with probability 1.}\end{aligned}$$

2. (a) For each $0 \le i \le N-1$, $\pi(i)P(i,i+1) = \pi(i+1)P(i+1,i)$. Since $P(i,i+1) = 1/3$ and $P(i+1,i) = 2/3$, this means $\pi(i) = 2\pi(i+1)$. Therefore for any $0 \le j \le N$, $\pi(j) = 2^{-j}\pi(0)$. We compute

$$1 = \sum_{j=0}^N \pi(j) = \sum_{j=0}^N 2^{-j}\pi(0) = (2 - 2^{-N})\pi(0).$$

Therefore $\pi(0) = 1/(2-2^{-N}) = 2^N/(2^{N+1}-1)$ and $\pi(j) = 2^{N-j}/(2^{N+1}-1)$.

(b) Again, each $\pi(j) = 2^{-j}\pi(0)$, and

$$1 = \sum_{j=0}^\infty \pi(j) = 2\pi(0),$$

so $\pi(0) = 1/2$ and $\pi(j) = 2^{-j-1}$. For each $j > 0$, $P(i,j) = 0$ except when $i = j \pm 1$, when we have $P(j-1,j) = 1/3$ and $P(j+1,j) = 2/3$. Therefore,

$$\sum_{i=0}^\infty \pi(i)P(i,j) = \frac{1}{3}\pi(j-1) + \frac{2}{3}\pi(j+1) = \frac{1}{3}\cdot\frac{1}{2^j} + \frac{2}{3}\cdot\frac{1}{2^{j+2}} = \frac{1}{2^{j+1}} = \pi(j).$$

When $j = 0$ we have

$$\sum_{i=0}^{\infty} \pi(i)P(i,0) = \pi(0)P(0,0) + \pi(1)P(1,0) = \frac{1}{2}\cdot\frac{2}{3} + \frac{1}{4}\cdot\frac{2}{3} = \frac{1}{2} = \pi(0).$$

3. Under the assumptions, the total number of spaces equals the total number of words. Therefore the total number of characters (i.e. letters plus spaces) equals the number of letters plus the number of words. So:

$$\begin{aligned}
\pi(\text{space}) &= \frac{\#\text{ spaces}}{\#\text{ characters}} = \frac{743842922321}{743842922321 + 3563505777820} = 0.1727,\\
\pi(F) &= \frac{\#\text{ letters F}}{\#\text{ characters}} = \frac{85635440629}{743842922321 + 3563505777820} = 0.0199,\\
\pi(Z) &= \frac{\#\text{ letters Z}}{\#\text{ characters}} = \frac{3205398166}{743842922321 + 3563505777820} = 0.0007.
\end{aligned}$$

4. We compute

$$\begin{aligned}
P(F,O) &= \frac{\#\text{ 2-grams FO}}{\#\text{ letters F}} = \frac{13753006196}{85635440629} = 0.1606,\\
P(O,R) &= \frac{\#\text{ 2-grams OR}}{\#\text{ letters O}} = \frac{35994097756}{272276534337} = 0.1322.
\end{aligned}$$

$P(\text{space}, F)$ is the probability that a word starts with F, while $P(R, \text{space})$ is the probability that a given letter R occurs at the end of its word.

$$\begin{aligned}
P(\text{space}, F) &= \frac{\#\text{ words starting with F}}{\#\text{ words}} = \frac{29952197540}{743842922321} = 0.0403,\\
P(R, \text{space}) &= \frac{\#\text{ words ending with R}}{\#\text{ letters R}} = \frac{43881791800}{223767519675} = 0.1961.
\end{aligned}$$

5. (a) We have

$$\pi(\alpha) = \frac{f(\alpha)}{\sum_\gamma f(\gamma)}, \qquad P(\alpha,\beta) = \frac{f(\alpha,\beta)}{f(\alpha)},$$

where the sum in the denominator of $\pi(\alpha)$ is over all characters γ.

(b) To verify that $\pi P = \pi$, we must check that $\sum_\alpha \pi(\alpha)P(\alpha,\beta) = \pi(\beta)$ for all characters β. Let $Z = \sum_\gamma f(\gamma)$. Then

$$\sum_\alpha \pi(\alpha)P(\alpha,\beta) = \sum_\alpha \frac{f(\alpha)}{Z} \cdot \frac{f(\alpha,\beta)}{f(\alpha)} = \frac{1}{Z}\sum_\alpha f(\alpha,\beta),$$

which will equal $\pi(\beta) = f(\beta)/Z$ as long as $\sum_\alpha f(\alpha,\beta) = f(\beta)$. This is true because every time the character β appears, it is preceded by some other character, so if we sum over all possible preceding characters α, we get the total number of occurrences of β.

The alert reader may notice a minor problem with the reasoning above. We are treating the Google corpus as a long string of words, each one followed by a space. What if β_0 is the very first letter in the corpus? Then $f(\beta_0)$ is the total number of occurrences of β_0, while $\sum_\alpha f(\alpha,\beta_0)$ is the total number of times that β_0 occurs after some other character. So

$$f(\beta_0) = 1 + \sum_\alpha f(\alpha,\beta_0),$$

while for $\beta \neq \beta_0$, $f(\beta) = \sum_\alpha f(\alpha,\beta)$ as desired. The solution is to declare that if β_0 is the first letter in the corpus, then $f(\text{space},\beta_0)$ should not denote the total number of occurrences of the 2-gram (space,β_0), but rather the total number of words beginning with β_0, which is the number of occurrences of (space,β_0) plus one.

In fact, this modification is necessary for the definition of P given in part (a) actually to give a valid transition matrix. When we declared that $P(\alpha,\beta) = f(\alpha,\beta)/f(\alpha)$, we were asserting that for each character α, $\sum_\beta f(\alpha,\beta) = f(\alpha)$, in order for each row of P to sum to 1. Since $f(\alpha)$ is the total number of occurrences of α and $\sum_\beta f(\alpha,\beta)$ is the number of times that α is followed by some other character, this is true except when α is the very last character in the corpus, which we are assuming is a space. That is, using the unmodified definition of $f(\alpha,\beta)$, we have $f(\alpha) = \sum_\beta f(\alpha,\beta)$ for all $\alpha \neq (\text{space})$, but $f(\text{space}) = 1 + \sum_\beta f(\text{space},\beta)$. When we increase $f(\text{space},\beta_0)$ by 1, this makes $f(\text{space}) = \sum_\beta f(\text{space},\beta)$, so that P is a valid transition matrix.

If you look at the computation of $P(\text{space},F)$ in problem 4, you will see that we have implicitly followed this rule.

6. (a) There are many solutions to this problem. For example, the probability that a string of 3 consecutive characters in the Markov model equals BLL is

$$\begin{aligned}
&\mathbf{P}_\pi(X_k = B, X_{k+1} = L, X_{k+2} = L) \\
&= \pi(B)P(B,L)P(L,L) \\
&= \frac{52905544693}{743842922321 + 3563505777820} \cdot \frac{6581097936}{52905544693} \cdot \frac{16257360474}{144998552911} \\
&= 0.0001713.
\end{aligned}$$

Since the whole corpus has $743842922321 + 3563505777820 = 4307348700141$ characters, there are $4307348700141 - 2 = 4307348700139$ strings of 3 consecutive characters, each of which has probability 0.0001713 of being BLL. Hence the expected number of occurrences of BLL in a body of text produced by the Markov model with the same length as the Google corpus is $0.0001713 \cdot 4307348700139 = 7.379 \cdot 10^8$. The actual number of occurrences is zero!

Note: To make the reasoning above rigorous, let $(X_0, \ldots, X_{N-1})$ be the text produced by the Markov model, where $N = 4307348700141$. For $0 \le k \le N-3$, define random variables $\mathbf{1}_k$ to be 1 if $(X_k, X_{k+1}, X_{k+2}) = (B, L, L)$ and 0 otherwise. The variables $\mathbf{1}_k$ are not independent, but by linearity of expectation,

$$\begin{aligned}
&\mathbf{E}_\pi[\#\text{ occurrences of } BLL] = \mathbf{E}_\pi\left[\sum_{k=0}^{N-3} \mathbf{1}_k\right] = \sum_{k=0}^{N-3} \mathbf{E}_\pi[\mathbf{1}_k] \\
&= \sum_{k=0}^{N-3} \mathbf{P}_\pi(X_k = B, X_{k+1} = L, X_{k+2} = L) = (N-2)\pi(B)P(B,L)P(L,L).
\end{aligned}$$

(b) The word FOR (preceded and followed by spaces) appears $6545282031 \approx 6.5 \cdot 10^9$ times in the data set. To figure out how many times it would be expected to appear from the Markov model, there are two approaches which yield essentially the same answer.

First approach: In a body of text with 743842922321 words produced by the Markov model, how many of them would be the word FOR? The probability

that any individual word is FOR is

$$\begin{aligned}&\mathbf{P}_{\text{space}}(X_1 = F, X_2 = O, X_3 = R, X_4 = \text{space})\\&= P(\text{space}, F)P(F, O)P(O, R)P(R, \text{space})\\&= 0.0001676,\end{aligned}$$

using the computations from problem 4. Therefore the expected number of words FOR is $743842922321 \cdot 0.0001676 = 1.2470 \times 10^8$.

Second approach: In a body of text with 4307348700141 characters produced by the Markov model, what is the expected number of occurrences of the 5-gram $(\text{space}, F, O, R, \text{space})$? The probability that any particular string of 5 consecutive characters is $(\text{space}, F, O, R, \text{space})$ equals

$$\begin{aligned}&\mathbf{P}_\pi(X_k = \text{space}, X_{k+1} = F, X_{k+2} = O, X_{k+3} = R, X_{k+4} = \text{space})\\&= \pi(\text{space})P(\text{space}, F)P(F, O)P(O, R)P(R, \text{space})\\&= 0.00002895.\end{aligned}$$

The number of strings of 5 consecutive characters is $4307348700141 - 4 = 4307348700137$. Therefore, the expected number of occurrences of the 5-gram is $4307348700137 \cdot 0.00002895 = 1.2470 \times 10^8$.

Using either approach, the word FOR appears about 52 times as often in the data set as the Markov model would predict.

The answers provided by the two approaches are almost but not quite equal. The expectation given by the first approach is

$$(\#\text{ words})P(\text{space}, F)P(F, O)P(O, R)P(R, \text{space}),$$

while the expectation given by the second approach is

$$(\#\text{ characters} - 4)\pi(\text{space})P(\text{space}, F)P(F, O)P(O, R)P(R, \text{space}).$$

Since

$$\pi(\text{space}) = \frac{\#\text{ words}}{\#\text{ characters}},$$

the only difference is the "minus 4" in the second expression. The ratio between the answers given by the two approaches is

$$\frac{\#\text{ characters}}{\#\text{ characters} - 4} = 1.00000000000093.$$