Name: ____________________________

Directions: Work only on this sheet (on both sides, if needed); do not turn in any supplementary sheets of paper. There is actually plenty of room for your answers, as long as you organize yourself BEFORE starting writing. In order to get full credit, SHOW YOUR WORK.

1. Consider the following model of a cache. Its capacity is m lines. In any memory reference, each of the b blocks in memory is equally likely to be referenced, and successive references are independent.[1] Line replacement is done by an LRU policy.

We will be viewing everything from the point of view of one particular block in memory which I will call B. The state X_n of our system at time n (an instant after the n^{th} memory reference) is the rank of B in terms of recent use: $X_n = k$ means that B is the k^{th} most-recently used among the blocks currently in the cache. The value of k can range from 1 to m, inclusive. If B is not in the cache, then define the state to be m+1.

For example say m = 4 and B is originally not in the cache, so our state is 5. Then suppose we have two references to B followed by five different references to other blocks, then one to B. These eight references would take us to states 1, 1, 2, 3, 4, 5, 5 and 1, respectively. Note that the sixth of these references resulted in B being *evicted* from the cache.

Suppose references are i.i.d., with the probability that block B is requested being p.

(a) (10) Find the (one-step) transition probabilities p_{ij}, i,j = 1,2,...,m,m+1.

(b) (10) Find closed-form expressions for the steady-state probabilities π_i.

In the remaining parts, express your answers in terms of the π_i and p. Do not substitute the expressions you found in part (b) above.

(c) (10) What proportion of the time will B be in the cache?

(d) (15) What proportion of references will result in evictions of B?

(e) (15) What will be the mean time between evictions of B?

2. (20) Consider a long data tape. Tapes are broken into *blocks*, analogous to disk sectors. Suppose we are storing data items of lengths D_1, D_2, ..., with D_{i+1} being stored contiguously after D_i. (A special character will be used to demarcate the end of one item and the beginning of the next.) Data items are definitely allowed to cross block boundaries.

Suppose the block size is large enough that we can scale units so that the D_i are continuous random variables and the block size is 1.0. Suppose also that the D_i are i.i.d. and that each is bounded by 1.0.

Within any block, there will be parts of at least two data items. Let L denote the position within the block at which the first piece ends. For example, suppose a data item begins at position 0.88 in one block and ends at 0.28 in the next block. Then L = 0.28 for that second block (and D = 0.40).

Find the long-run average value of L.

3. (20) Consider an M/G/1 queue, and let S, N and R be the service time, number in the system and residence time as in our notes, respectively. Let l_Y denote the Laplace transform of any continuous nonnegative random variable, and let g_Z denote the generating function of any nonnegative integer-value random variable. Let S denote the service time. Show that for $w > 0$,

$$l_R(w) = \frac{1}{wES} g_N[l_S(w)][1 - l_S(w)]$$

Solutions:

1.a I was rather liberal in grading this problem, as some people imposed additional structure, assuming that all b blocks had probability 1/b of being accessed. I will do so here too.

For i = 1,2,...,m:

[1] I am using the term *line* to refer to a slot in the cache, to be filled by one *block* from memory. Many books use the term *block* to refer to both entities.

$$p_{i,i+1} = 1 - \frac{m}{b}$$

$$p_{ii} = \frac{m-1}{b}$$

For all i, $p_{i1} = \frac{1}{b}$.

And

$$p_{m+1,m+1} = 1 - \frac{1}{b}$$

1.c $1 - \pi_{m+1}$

1.d $\pi_m(1 - \frac{m}{b})$ (or $\pi_m(1-p)$)

1.e Think of an alternating renewal process in which the "on" times consist of the time slots in which B is newly evicted. The long-run proportion of such slots is the answer to part (d) above, and of course each slot lasts 1 unit of time. The "off" times are those in which B is either not in the cache, or in the cache but not currently being evicted. By renewal theory, the mean on+off time is then 1 divided by the proportion of "on" slots, i.e. the reciprocal of the answer to (d). (Note that the answer is not the reciprocal of π_{m+1}, since that reciprocal includes loops from state m+1 to itself.

2. We can use renewal theory here. "Time" is distance along the tape, and a "renewal" is the start of a new data item. At each tape position $i \cdot 1.0$, i = 1,2,3,..., L is the forward recurrence "time." Thus

$$EL = \int_0^\infty t \frac{1 - F_D(t)}{ED} \, dt = \frac{E(D^2)}{2ED}$$

(recall that in our discussion of the M/G/1 queue, we derived the mean forward/backward recurrence time).

3. As in our discussion of the M/G/1 queue,

$$R = S_{1,resid} + S_2 + ... + S_N + S_{self}$$

Following the steps in that discussion, we get

$$l_R(w) = l_{S_{1,resid}}(w) E[l_S^N(w)] \tag{1}$$

The first factor in (1) is equal to

$$\int_0^\infty e^{-wt} \frac{1 - F_S(t)}{ES} \, dt$$

which, after again following steps in our Laplace transform derivation for the M/G/1 queue, is equal to

$$\frac{1}{wES}[1 - l_S(w)]$$

The second factor in (1) is equal to $g_N[l_S(w)]$.